Transductive Learning for Peer Review Score Prediction

査読スコア予測のための帰納的学習

山梨大学大学院

医工農学総合教育部

博士課程学位論文

2024 年 3 月

MUANGKAMMUEN PANITAN

Abstract

Peer review stands as a cornerstone in validating academic work and ensuring the quality of published research. The process, while essential, can be time-consuming for both authors and reviewers. Therefore, the development of an accurate system for predicting peer review scores holds immense potential in streamlining the process, reducing the workload of reviewers, and providing constructive feedback to authors. This dissertation addresses the challenges associated with implementing deep learning methods for predicting peer review scores, particularly in scenarios where labeled data is limited. Deep learning has emerged as a promising method for developing peer-review scoring systems. However, the requirement for substantial training data poses a significant challenge. Publicly available datasets for peer review are often constrained in size, impeding the creation of robust models. This research aims to overcome these challenges by introducing innovative transductive learning approaches that capitalize on the inherent structure within unlabeled data or insights from related tasks, enhancing the performance of peer review prediction models. Traditional deep learning methodologies involve fine-tuning language models (LMs) tailored for specific tasks. In response to limited resources for fine-tuning LMs, this dissertation explores transductive learning approaches. Transductive learning aims to boost model performance by leveraging the inherent structure within unlabeled data or insights from related tasks. This novel approach highlights the versatility and effectiveness of incorporating diverse types of information in machine learning methodologies. An additional challenge addressed in this research revolves around the limitations of pretrained models, particularly Transformer-based models. While proficient in managing relatively short sequences, these models encounter difficulties when processing longer sequences. This dissertation proposes four distinct approaches to enhance peer review score prediction.

The first two approaches center on semi-supervised learning methods, addressing both truncated and full documents. Semi-supervised learning is a potent technique that combines labeled and unlabeled data during training, leveraging the wealth of information present in unlabeled datasets to enhance the model's understanding and overall performance. In the context of peer review score prediction, these approaches specifically focus on utilizing ladder networks within the semi-supervised learning paradigm. Ladder networks, a type of deep denoising autoencoder, are characterized by the incorporation of skip connections and reconstruction targets at intermediate layers. This architecture is trained to minimize combined supervised and unsupervised cost functions concurrently, employing

backpropagation. The study places emphasis on the use of ladder networks, particularly the Γ-model, which streamlines the decoder by retaining only the top layer. This modification facilitates seamless integration into various networks without the need for a separate decoder. A notable aspect of this research lies in its fundamental contribution to the integration of transformer-based models into the realm of semi-supervised learning, specifically within the framework of ladder networks. This integration is designed to optimize the model's ability to capture complex patterns and relationships in both labeled and unlabeled data, thereby enhancing its predictive capabilities for peer review scores. The study's outcomes are anticipated to provide valuable insights into the correlation between the quantity of unlabeled data and the observed performance enhancement in both the semi-supervised baselines and the proposed models. By delving into the interplay between model architecture, training methodology, and the availability of unlabeled data, this research aims to contribute valuable knowledge that can inform the development of more robust and effective semi-supervised learning approaches for peer review score prediction.

The third approach introduces a tailored form of transfer learning designed specifically for truncated documents. Transfer learning is a powerful technique involving the training of a model on a source task and subsequently transferring the acquired knowledge to a target task. The overarching goal of this research is to enhance the model's capacity to comprehend and predict peer review scores from academic texts. This is achieved by training the model on a task that incorporates a larger set of related data, followed by the transfer of the acquired knowledge to the specific task of predicting peer review scores. A distinctive method employed in this approach is intermediate-task transfer learning for predicting peer review scores. This involves the initial fine-tuning of a pretrained model on an intermediate task, followed by subsequent fine-tuning on the target task. In this study, the intermediate task selected is the prediction of review-aspect sentiment. The choice of sentiment prediction as the intermediate task is grounded in the observation that the sentiment expressed in a review often correlates with the score attributed to the review. Additionally, the dissertation introduces a technique to extract aspect sentiments from a detailed review aspect annotation within the peer-review dataset. The experimental outcomes of this approach demonstrate the effectiveness of each intermediate task, showcasing notable performance improvements across all aspects of review score prediction. By strategically leveraging transfer learning with an intermediate task that captures sentiment nuances, the research contributes to refining the model's understanding of the intricate relationships within academic texts, ultimately leading to enhanced accuracy in predicting peer review scores.

The fourth approach delves into the realm of transfer learning specifically tailored for full documents, recognizing the distinctive characteristics and challenges posed by longer academic papers. The utilization of transfer learning for full documents seeks to address the limitations of pretrained

models when confronted with extended sequences, ultimately facilitating a more comprehensive analysis and understanding of the content. The methodological approach involves segmenting the document into individual sentences and deriving a representation for each sentence from the pretrained LM. These sentence representations are then concatenated into a sequential format, serving as input for both intermediate-task training and subsequent fine-tuning on the target tasks. This process is designed to leverage the pretrained model's understanding of linguistic structures and contextual nuances, with a specific focus on the challenges presented by lengthier academic documents. The experimental findings from this approach yield crucial insights, underscoring the importance of models capable of effectively processing longer sequences. The results suggest that such models exhibit superior performance, providing a valuable contribution to the broader field of transfer learning for document analysis. By strategically adapting transfer learning techniques to accommodate the intricacies of full documents, this approach aims to enhance the model's ability to capture and comprehend the nuanced information embedded within extensive academic papers, thereby contributing to more accurate and comprehensive peer review score predictions.

In conclusion, this dissertation makes significant contributions to the field of peer review prediction through the introduction of innovative transductive learning approaches and the strategic utilization of semi-supervised and transfer learning techniques. These proposed methods are specifically designed to tackle challenges arising from limited labeled data and the inherent limitations of pretrained language models in the context of peer review scoring. By integrating transductive learning, the research seeks to capitalize on the inherent structure within unlabeled data and insights from related tasks, enhancing the overall performance of peer review prediction models. The incorporation of semi-supervised learning techniques aims to optimize model performance by leveraging the combined information from labeled and unlabeled data. Furthermore, the adaptation of transfer learning, designed full documents, addresses the need for more robust models capable of handling long document lengths. The intermediate-task transfer learning adds an additional layer of sophistication by strategically fine-tuning the model on tasks related to sentiment prediction, thereby improving its understanding of the nuanced relationships within academic texts. The anticipated outcomes of this research extend beyond methodological advancements. The developed approaches are expected to significantly enhance the accuracy and efficiency of peer review scoring systems. This, in turn, stands to benefit authors and reviewers by streamlining the review process and providing more insightful feedback. Ultimately, the academic community at large is poised to reap the rewards of improved peer review prediction models, fostering a more efficient and constructive environment for scholarly discourse and advancement.

Contents

Chapter 1

Introduction

1.1 Background

In recent years, there has been a substantial increase in submissions for AI-related international conferences and journals. This surge in submissions has posed significant challenges to the review process. Automatic peer-review aspect score prediction (PASP) serves as a crucial tool aimed at enhancing the efficiency and effectiveness of the academic paper review process. By offering reviewers and authors a numeric evaluation across various qualities of a paper, including aspects such as clarity and originality, PASP streamlines the assessment process. This automated system provides a structured and quantitative assessment, enabling both reviewers and authors to better understand the strengths and weaknesses of a paper, thereby facilitating more informed decision-making and improvements in subsequent revisions.

In the field, a significant milestone is marked by the introduction of the PeerRead dataset (Kang et al., 2018), representing a pioneering contribution. This dataset stands as the initial publicly accessible compilation of scientific peer reviews. Its availability and accessibility represent a groundbreaking development within the academic community. As the inaugural dataset of its kind, PeerRead serves as a valuable resource, providing comprehensive access to scientific peer reviews for research and analysis purposes. The utilization of the PeerRead dataset extends across a diverse spectrum of applications within academic research. Its versatile usage spans various crucial areas. Paper acceptance classification (Ghosal et al., 2019; Deng et al., 2020; Maillette de Buy Wenniger et al., 2020; Fytas et al., 2021), the dataset facilitates the classification of papers, aiding in the determination of whether a submission meets the criteria for acceptance or rejection within academic conferences and journals. Review aspect score prediction (Li et al., 2020a; Wang et al., 2020; Muangkammuen et al., 2022a), leveraging the dataset enables the prediction of different aspect scores within peer reviews. This application helps assess and quantify qualities like clarity, originality, and other pertinent factors in academic papers. Citation count prediction (Dongen et al., 2020), researchers utilize the dataset to predict the future citation counts of academic papers. This predictive analysis aids in understanding the potential impact and reach of scholarly work. Citation recommendation (Jeong et al., 2020), the dataset assists in recommending citations for academic papers, contributing to the referencing and sourcing process by suggesting relevant and credible sources.

1.2 Problem Statement

Over the past years, large language models (LLMs) have demonstrated significant advancements in enhancing the performance of diverse Natural Language Processing (NLP) tasks (Peters et al., 2018; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2019). These methodologies typically involve a two-step process. Initially, models engage in pretraining on unsupervised tasks, commonly language modeling, where the model learns to understand the structure and patterns of language using vast amounts of unlabeled data. Subsequently, these pretrained models are fine-tuned or adapted to specific target tasks, which may involve tasks like text classification, sentiment analysis, named entity recognition, or other NLP tasks that require labeled data.

The performance of LLMs in tasks such as review aspect score prediction is hindered by the scarcity of available annotated datasets. These datasets, containing labeled information necessary for training models, are notably limited in quantity and quality. This scarcity poses a significant challenge as it restricts the capacity of large language models to achieve optimal performance in tasks that require annotated data for training. The lack of annotated datasets hampers the ability of models, especially large language models, to learn effectively. As a result, these models might not capture the nuanced patterns or variations within the data adequately, impacting their overall performance in tasks like review aspect score prediction. This limitation emphasizes the need for innovative approaches, such as semi-supervised learning or novel training techniques, to compensate for the scarcity of annotated data and enhance the performance of large language models in tasks with limited labeled information.

Semi-supervised learning is a machine learning paradigm that operates using a combination of labeled and unlabeled data to train models. In this approach, the model learns from both the limited labeled data, which contains explicit annotations or labels, and the vast pool of unlabeled data lacking specific annotations. Unlike supervised learning, where models are trained exclusively on labeled data, semi-supervised learning takes advantage of the abundance of unlabeled data available in many real-world scenarios. By leveraging this additional unlabeled data, semi-supervised learning aims to enhance model performance, generalization, and robustness. This approach is particularly valuable when acquiring labeled data is expensive, time-consuming, or limited in quantity. Semi-supervised learning methods often involve using the labeled data to guide and inform the learning process on the unlabeled data. Techniques within semi-supervised learning include self-training (Meng et al., 2020; Zhang et al., 2022), co-training (Blum and Mitchell, 1998; Wan, 2009), or utilizing generative models to augment the labeled dataset with pseudo-labels derived from the unlabeled data (Chen et al, 2021; Wang et al, 2022).

In Natural Language Processing (NLP) tasks, such as review aspect score prediction or text

classification, semi-supervised learning methods have proven beneficial. They enable models to learn more effectively from the available limited labeled data while capitalizing on the wealth of unlabeled text data to improve overall performance, especially in scenarios where acquiring large labeled datasets is challenging or resource-intensive.

Intermediate-task transfer learning is a technique in transfer learning where a model is trained on an intermediate task, which is related or auxiliary to the main target task, to improve performance on the ultimate task of interest (Phang et al, 2018; Pruksachatkun et al., 2020). In the context of review aspect score prediction, intermediate-task transfer learning might involve training a model on a task that shares underlying linguistic or semantic aspects with the prediction of review aspect scores. For example, the intermediate task could involve sentiment analysis, coherence prediction, or language fluency assessment, which captures aspects of language understanding relevant to the goal of predicting review aspect scores. The model is trained on this intermediate task, allowing it to learn useful representations, features, or patterns that are transferable and beneficial for the subsequent prediction of review aspect scores. The acquired knowledge from the intermediate task serves as a foundational understanding of linguistic nuances or semantic relationships that contribute to the assessment of various aspects within peer reviews. Through intermediate-task transfer learning, the model leverages its understanding from the intermediary task to improve its performance and effectiveness in predicting the specific aspects of review quality, such as clarity, originality, relevance, and more. This approach enhances the model's ability to capture relevant information and nuances required for accurate and comprehensive review aspect score prediction.

Another significant concern arises from the limitations inherent in pretrained models. One such limitation is their handling of long sequences. Models, particularly Transformer-based architectures (Vaswani et al., 2017), struggle when processing sequences of substantial length. These models have constraints on the maximum sequence length they can effectively manage. Specifically, they may have restrictions on the number of words or tokens they can process in a sequence. This limitation poses challenges when dealing with longer texts, academic papers, or documents that exceed the model's processing capacity, potentially leading to incomplete or truncated representations of information. This constraint in handling long sequences within pretrained models presents a considerable hurdle, especially in tasks requiring the analysis or understanding of extensive textual data. Researchers and practitioners must address this limitation to ensure that pretrained models remain effective and applicable across a broader spectrum of tasks and domains, particularly in contexts involving lengthy sequences or documents.

To address the challenge of processing lengthy sequences in peer-review aspect score prediction,

this dissertation has yielded two notable approaches aiming to overcome the limitations posed by lengthy academic papers on existing models. The first approach, presented in the Γ-Transformer-LS (Γ-TLS) method, introduces a semi-supervised learning framework that integrates ladder networks (Rasmus et al., 2015) and the Long-short transformer (Zhu et al., 2021). This method aims to manage lengthy sequences by integrating unsupervised learning through a denoising autoencoder, utilizing a large unlabeled dataset that serves as a fundamental resource in scientific peer reviews. Conversely, the second approach, SciBERT (Beltagy et al., 2019) over Sentence Embeddings (SciBERT-SE), proposes an innovative strategy by segmenting academic papers into sentences, representing each sentence with SciBERT embeddings, and subsequently stacking these representations to process longer sequences. Both approaches, while distinct in their methodologies, share the common goal of mitigating the limitations posed by long sequences in natural language processing tasks, particularly in the context of predicting review aspect scores, thereby contributing significantly to the advancement of this field.

The dissertation focuses on two primary challenges in peer review score prediction: the scarcity of annotated datasets hindering model training and the limitations posed by handling lengthy sequences, impacting the effectiveness of pretrained models. It explores solutions such as semi-supervised learning and transfer learning methods, and novel approaches like Γ-Transformer-LS and SciBERT over Sentence Embeddings to address these issues in the context of peer-review aspect score prediction.

1.3 Research Objectives

This research aims to aims to accomplish three primary objectives:

1. To introduce a novel method aimed at enhancing peer review score prediction despite the constraint of having a limited annotated dataset.
2. To propose a novel method to overcome the limitations encountered by large language models when processing lengthy sequences, particularly academic papers.
3. To develop and assess the effectiveness of deep learning models designed to accurately classify and predict peer review scores.

1.4 Research Scope

The scope of this study is confined to the development of an automated peer review scoring system utilizing deep learning. The following delineates the specific scope of this research work:

1. Data acquisition involves utilizing the PeerRead dataset, specifically ACL 2017, comprising scientific papers related to the natural language processing domain.
2. Peer review scores encompass seven aspects: clarity, originality, soundness correctness, meaningful comparison, substance, impact, and recommendation. Each model is designed to predict a single aspect score.
3. The study focuses solely on leveraging deep learning methodologies.
4. Evaluation of the classifier model's scoring performance will be conducted based on Accuracy and F1-score metrics.

1.5 Contributions

The research contributes significantly to three main aspects, and these pivotal ideas have been presented comprehensively throughout the dissertation.

The primary contribution of this study centers on semi-supervised learning for peer review score prediction, particularly focusing on ladder networks (Muangkammuen et al., 2022; Muangkammuen et al., 2023a). Ladder networks represent a form of deep denoising autoencoder incorporating skip connections and reconstruction targets at intermediate layers. This model is trained to concurrently minimize the combined supervised and unsupervised cost functions using backpropagation. The Γ-model, a variant of ladder networks, streamlines the decoder by retaining only the top layer, facilitating its integration into various networks without the need for a separate decoder. The fundamental contribution lies in integrating transformer-based models into the realm of semi-supervised learning, specifically ladder networks. The study's outcomes offer valuable insights into the correlation between the quantity of unlabeled data and the performance enhancement observed in the semi-supervised baselines and the proposed models.

The second significant contribution of this dissertation involves the introduction of intermediate-task transfer learning for predicting peer review scores (Muangkammuen et al., 2023b; Muangkammuen et al., 2024). This method entails the initial fine-tuning of a pretrained model on an intermediate task, followed by subsequent fine-tuning on the target task. In this study, the prediction of review-aspect sentiment was designated as the intermediate task, considering that the sentiment expressed in the review correlates with the score attributed to the review. Additionally, the dissertation introduced a technique to extract aspect sentiments from a detailed review aspect annotation within the peer-review dataset. The experimental outcomes demonstrate the effectiveness of each intermediate task, leading to notable performance improvements across all aspects of review score prediction.

The third noteworthy contribution of this dissertation expands on long sequence processing for transformer-based models (Muangkammuen et al., 2023a) and pretrained large language models (LLMs) (Muangkammuen et al., 2024). The first method involves integrating a Long-short transformer (Transformer-LS) into the Γ-model, a variant of ladder networks. Transformer-LS represents an adaptation of the transformer model that possesses enhanced memory and computational efficiency. The incorporation of Transformer-LS enables the proposed model to effectively handle lengthy sequences. The second approach entails segmenting the document into sentences and obtaining a representation for each sentence from the pretrained LLM. These sentence representations are then concatenated into a sequence and utilized as input for both intermediate-task training and fine-tuning on the target tasks. The experimental findings yielded crucial insights indicating that models capable of processing longer sequences tend to yield superior outcomes.

1.6 Dissertation Outline

This dissertation is structured into seven chapters, including the introduction, literature review, semi-supervised learning for truncated and full documents, transfer learning for truncated and full documents, and conclusion.

Chapter 1 provides an overview of the study, encompassing the background information, research objectives, scope of the study, and a delineation of the research contributions.

Chapter 2 comprehensively reviews pertinent literature and previous studies, covering various domains such as peer review scoring systems, transformer models, transductive learning, and methods of performance evaluation.

Chapter 3 extensively covers the utilization of semi-supervised learning for truncated documents in predicting peer review scores through the introduction of a semi-supervised learning framework known as Γ-Transformer. This involves detailing the methodology that integrates a Γ-model of the ladder network with a pretrained transformer model, outlining the experimental settings, analyzing the obtained results, and conducting a comprehensive error analysis.

Chapter 4 investigates the implementation of semi-supervised learning for full documents in predicting peer review scores, introducing the Γ-Transformer-LS framework. This section encompasses the methodology, leveraging a Γ-model of the ladder network coupled with the Long-

short transformer, to surpass the limitations of the standard transformer in handling long sequences. It further elaborates on the experimental settings, showcases the achieved results, and conducts an ablation study to evaluate the performance of the Long-short transformer within the semi-supervised learning context.

Chapter 5 delves deeply into intermediate-task transfer learning for truncated documents, particularly focusing on predicting peer review scores. This chapter encompasses the process of extracting review aspect sentiment for intermediate-task training, followed by the implementation of intermediate-task training and subsequent target-task fine-tuning. Additionally, an ablation study is conducted to analyze how different strategies in review aspect sentiment extraction contribute to the performance of the target task.

Chapter 6 focuses on intermediate-task transfer learning for full documents in predicting peer review scores. This section introduces a technique to expand a pretrained model, SciBERT, for processing long sequences. The experimental results entail a comparative analysis between the proposed method and traditional SciBERT to evaluate their respective performances.

Chapter 7 provides a summary of the primary findings derived from the study, explores the limitations encountered in the research process, and outlines potential future directions for further research in this particular field of study.

Chapter 2
Literature Review

This dissertation introduces a peer-review scoring system that addresses the challenging issues surrounding limited annotated training data and the processing of long sequences, such as those found in academic papers. Consequently, this chapter offers a concise survey covering transformer models, including insights into transductive learning methods that were utilized in this dissertation. It outlines an overview of evaluation methods used to assess the performance of the peer review scoring system.

2.1 Peer-review Scoring System

The scholarly communication process indeed faces significant challenges, with escalating submission rates and increased pressure on peer reviewers. The strain on reviewers' time and the rising retraction rates highlight the need for more efficient quality control mechanisms within the research community. Initiatives leveraging automated screening tools powered by Artificial Intelligence (AI), machine learning, and Natural Language Processing (NLP) present promising avenues for enhancing the peer review process.

Artificial Intelligence is a crucial tool for academic peer review, and it is a rapidly growing field that demands more attention from the academic community. The renowned Toronto Paper Matching system, developed by Charlin and Zemel (2013), was designed to match papers with appropriate reviewers. Notably, Price and Flach (2017) conducted an in-depth examination of the diverse methods for harnessing computational support in the peer review system. Mrowinski et al. (2017) explored the application of evolutionary algorithms to enhance editorial strategies within the peer review process. Ghosal et al. (2018) delved into an investigation of the impact of various features in the editorial pre-screening process. Wang and Wan (2018) explored a multi-instance learning framework for conducting sentiment analysis on peer review texts. Ghosal et al. (2019) investigated the impact of reviewer sentiment expressed in peer review texts on the outcome of the review process. Li et al. (2020a) proposed a multi-task learning approach that automatically selects shared structures and auxiliary resources for peer review prediction. Our investigations are currently centered on a portion of the PeerRead dataset that has been made available to the public (Kang et al., 2018).

The existing body of work predominantly relies on utilizing review text for predictive purposes, a practice often impractical in real-world scenarios. However, Li et al. (2020a) were pioneers in introducing a methodology that exclusively employs paper content as input for their system. Additionally, they tackled the challenge of limited peer review training data by introducing a multi-task learning approach. This method capitalizes on additional rich information from various aspect scores, obtained from external resources, adopting a main-auxiliary structure for each aspect score within the multi-task model.

Their innovative approach circumvented the constraints posed by the length of academic papers by employing a Convolutional Neural Network (CNN), deviating from reliance on Large Language Models (LLMs). It's important to note that, in Natural Language Processing (NLP), LLMs present distinct advantages compared to CNNs. While CNNs excel at processing visual data such as images, learning spatial hierarchies and patterns, LLMs are tailored for comprehending and generating human-like text. They leverage transformer architectures and attention mechanisms to capture contextual relationships, semantic subtleties, and language intricacies within textual data.

This dissertation aims to offer solutions utilizing transformer models and LLMs for predicting peer review scores. It introduces transductive learning techniques to address the challenge of limited training data, harnessing unlabeled data to augment the performance of predictive models. Additionally, it proposes methods to leverage transformer models and LLMs effectively for analyzing and processing lengthy documents in the context of peer review.

2.2 Transformer Models

Transformers represent a groundbreaking neural network architecture primarily employed in natural language processing (NLP) tasks, renowned for their effectiveness in handling sequential data and capturing contextual relationships among tokens in a sequence (Vaswani et al. 2017). Transformers have revolutionized NLP by offering an alternative approach to traditional recurrent and convolutional architectures.

The core innovation of transformers lies in their self-attention mechanism, which enables the model to selectively focus on different parts of the input sequence while processing each token. This mechanism allows the network to learn contextual representations by considering all other tokens' contributions, regardless of their positions in the sequence, thereby addressing the limitations of sequential processing in RNNs and capturing long-range dependencies more efficiently.

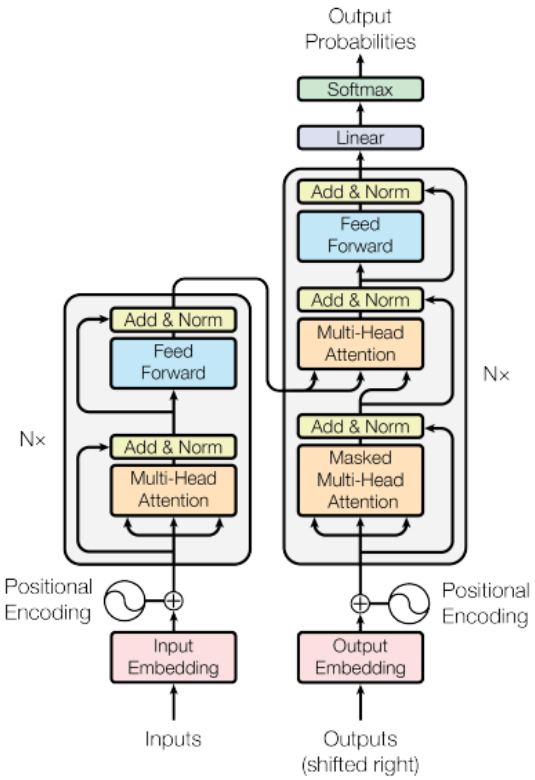


Figure 2.1 Transformer architecture

Key components and concepts within transformer architectures are shown in Figure 2.1, which include:

- **Self-Attention Mechanism**: This mechanism computes attention scores for each pair of tokens in the sequence, allowing the model to weigh the significance of each token in relation to others, contributing to the representation of each token.
- **Multi-Head Attention**: Transformers use multiple attention heads, allowing the model to attend to different parts of the input sequence simultaneously and capture diverse relationships within the data.
- **Positional Encoding**: Since transformers lack inherent sequential information, positional encodings are added to the input embeddings to convey the tokens' positions in the sequence, aiding the model in understanding the sequence order.
- **Encoder-Decoder Architecture**: In sequence-to-sequence tasks like machine translation, transformers employ an encoder-decoder architecture, where the encoder processes the input sequence, and the decoder generates the output sequence based on the learned representations.

The evolution of transformer-based architectures has significantly reshaped the landscape of

natural language processing (NLP) and beyond, introducing a spectrum of innovative models tailored to understand, generate, and process textual data with unprecedented proficiency. This dissertation explores and examines prominent variants and extensions of the original transformer architecture.

2.2.1 BERT

BERT, short for Bidirectional Encoder Representations from Transformers, stands as a groundbreaking transformer-based model in natural language processing (NLP). Developed by Google, BERT revolutionized language understanding by introducing bidirectional context-awareness into pretrained language representations (Devlin et al., 2019). BERT employs an unsupervised, pretraining stage where the model is trained on vast amounts of text data in an unsupervised manner, followed by fine-tuning on task-specific data in a supervised setting. The pretraining process primarily involves two main strategies:

1. **Masked Language Model (MLM)**: BERT uses the MLM approach during pretraining. It masks a certain percentage of the input tokens at random and tasks the model with predicting these masked tokens based on the surrounding context. This bidirectional training allows the model to learn contextual representations by capturing relationships between the masked tokens and the rest of the text. The bidirectional context understanding enables BERT to comprehend the meaning and significance of words or phrases within larger contexts, contributing to its robustness in capturing nuanced language structures.
2. **Next Sentence Prediction (NSP):** BERT also incorporates a next sentence prediction task during pretraining. It provides pairs of sentences to the model and trains it to predict whether the second sentence follows the first in the original text. This task helps BERT in learning relationships between consecutive sentences and understanding discourse-level coherence, facilitating the model's ability to understand relationships and contexts across multiple sentences.

BERT is pretrained on extensive corpora, such as Wikipedia, books, articles, and web text, allowing it to capture a broad spectrum of linguistic patterns and contexts from diverse sources. The pretraining stage enables the model to learn a rich, generalized understanding of language that can be fine-tuned for various downstream NLP tasks. After pretraining, BERT's learned representations, stored as weights in the model, can be fine-tuned with relatively small amounts of task-specific labeled data, making it adaptable and transferable across a wide range of NLP tasks and domains.

2.2.2 RoBERTa

RoBERTa (A Robustly Optimized BERT Pretraining Approach) represents an evolution and optimization of the BERT architecture, introduced by Facebook AI (Liu et al., 2019b). It builds upon the success of BERT while refining and enhancing various aspects of pretraining to improve model performance and robustness in natural language understanding tasks.

RoBERTa's enhancements and optimizations are centered around key modifications in the pretraining process:

- **Training Data and Duration**: RoBERTa is trained on a much larger corpus of text data compared to BERT, incorporating additional unlabeled data from diverse sources. It undergoes longer pretraining, enabling the model to capture more nuanced linguistic patterns and context.
- **Dynamic Masking Strategy**: RoBERTa utilizes a dynamic masking strategy where tokens are randomly masked during pretraining for each epoch. Unlike BERT's fixed masking scheme, RoBERTa introduces variability in the masked tokens across different training epochs, promoting improved generalization and learning robust representations.
- **No Next Sentence Prediction (NSP) Task**: Unlike BERT, RoBERTa excludes the next sentence prediction task during pretraining. This omission allows RoBERTa to focus solely on the masked language modeling (MLM) objective, allocating more attention to effectively capturing bidirectional context and relationships within the text.
- **Hyperparameter Optimization**: RoBERTa fine-tunes several hyperparameters, such as batch size, learning rate, and training data size, to maximize model performance. These optimizations contribute to enhanced learning and more robust representations.

The modifications introduced in RoBERTa aim to enhance the model's understanding of language nuances, improve generalization capabilities, and boost performance across a wide range of NLP tasks. RoBERTa has demonstrated superior performance in various benchmarks and NLP challenges, showcasing advancements in language understanding and setting new standards for state-of-the-art models in natural language processing.

2.2.3 SciBERT

SciBERT is a specialized variant of BERT, specifically tailored for scientific text and domain-specific language understanding within the scientific community. Developed by researchers at the Allen Institute for AI (Beltagy et al., 2019). SciBERT is pretrained on a large corpus of scientific literature and domain-specific texts, aiming to capture the nuances, vocabulary, and structural

intricacies prevalent in scientific documents.

Key features and adaptations of SciBERT include:

- **Pretraining on Scientific Texts**: SciBERT is trained on a vast amount of text from various scientific domains, including biomedical literature, computer science papers, and other academic publications. This specialized pretraining enables the model to develop domain-specific knowledge and better understand the syntax, semantics, and technical language prevalent in scientific writing.
- **Domain-specific Vocabulary and Contex**t: Unlike general-purpose language models, SciBERT's training data focuses on scientific terminologies, abbreviations, and contextual understanding specific to scientific disciplines. This specialization allows the model to comprehend and generate representations that align closely with the linguistic patterns and terminologies found in scientific texts.
- **Fine-tuning for Scientific Tasks**: SciBERT's domain-specific pretraining can be fine-tuned further on specific downstream tasks within the scientific domain, such as biomedical entity recognition, scientific document classification, or question answering in scientific literature. Fine-tuning allows the model to adapt its learned representations to perform well on these specific tasks, leveraging its domain-specific understanding.

SciBERT's specialization in scientific texts has proven beneficial in various scientific and biomedical NLP applications, demonstrating improved performance and understanding in tasks related to scientific literature analysis, information extraction from scholarly articles, and domain-specific question answering. Its tailored approach to capturing scientific language nuances and terminologies makes it a valuable tool for researchers and practitioners working in scientific fields, facilitating more accurate and domain-aware language processing in these specialized domains.

2.2.4 Longformer

Longformer is a novel variant of transformer-based architectures designed to handle longer sequences more effectively than standard transformer models. Developed by researchers at Allen Institute for AI (Beltagy et al., 2020). Longformer addresses the challenge of processing extensive text inputs prevalent in tasks such as document summarization, long document understanding, and large-scale document classification.

Key characteristics of Longformer include:

- **Attention Mechanism with Sparse Attention Patterns**: Longformer employs a modified

attention mechanism that introduces sparse attention patterns. Instead of attending to all tokens in the sequence, it uses a combination of global and local attention mechanisms to selectively attend to relevant tokens while maintaining computational efficiency. This sparse attention mechanism enables Longformer to handle longer sequences without significantly increasing computational resources.

- **Global Attention and Sliding Window Approach**: Longformer incorporates global attention, allowing tokens to attend to distant tokens in the sequence. Additionally, it utilizes a sliding window approach, where attention spans cover adjacent tokens within windows, allowing the model to capture global context efficiently while maintaining a balance between local and global attention.
- **Efficient Training and Computation**: By employing sparse attention patterns and global-local hybrid attention mechanisms, Longformer mitigates the computational burden associated with processing longer sequences in transformer-based models. This design allows the model to scale to longer document lengths without substantial increases in memory and computational requirements.

Longformer's ability to efficiently handle longer sequences has been particularly advantageous in tasks that involve processing extended texts, enabling more comprehensive understanding and analysis of lengthy documents while maintaining computational feasibility. Its innovative design has shown promising results in various document-level NLP tasks, contributing to advancements in handling large-scale document processing and understanding within the transformer architecture paradigm.

2.2.5 Long-short transformer (Transformer-LS)

Transformer-LS is a variant of the Transformer architecture designed to address the challenge of processing long sequences more efficiently (Zhu et al., 2021). The key focus of Transformer-LS is on handling long sequences, which are often problematic for standard Transformers due to the quadratic scaling of their self-attention mechanism with sequence length. This scaling issue limits the ability of Transformers to process sequences beyond a certain length.

Transformer-LS aims to overcome this limitation by introducing modifications to the self-attention mechanism, enabling linear complexity with respect to sequence length. By incorporating long-short range attention mechanisms, Transformer-LS can efficiently model dependencies across both long and short distances within a sequence. This allows the model to capture information from distant parts of the sequence without encountering the computational inefficiencies associated with standard Transformers.

- **Short-term Attention via Segment-wise Sliding Window**: To capture fine-grained local correlations efficiently, Transformer-LS employs a sliding window attention mechanism. The input sequence is divided into segments, and each token within a segment attends to nearby tokens within a fixed-size neighborhood, incorporating attention spans over adjacent tokens on both sides of the home segment. This sliding window approach efficiently handles local correlations and scales linearly with sequence length, enhancing computational efficiency.
- **Long-range Attention via Dynamic Projections**: For capturing long-range dependencies, Transformer-LS introduces dynamic low-rank projections. These projections replace the full attention with low-rank matrices, allowing each query to attend to all tokens via a product of two matrices with reduced dimensions. The dynamic projection matrices are flexible and adapt to various sequence lengths and changes, preserving computational efficiency while capturing long-range correlations effectively.
- **Aggregating Long-range and Short-term Attentions**: Transformer-LS aggregates both local and global attention by allowing each query to attend to a union of keys and values from both the local sliding window and global low-rank projections. This aggregation strategy enables the model to select essential information from either mechanism, fostering improved learning and information extraction. The implementation also includes a normalization strategy, called DualLN (Dual Layer Normalization), to align scales between different attention mechanisms, aiding in the effectiveness of aggregation.

By combining short-term and long-range attention mechanisms and employing strategies like sliding window attention and dynamic projections, Transformer-LS aims to efficiently capture both local and global dependencies within sequences, enhancing its capabilities in various natural language processing tasks.

2.3 Transductive Learning

In the context of this dissertation, transductive learning encapsulates both semi-supervised learning and transfer learning paradigms, leveraging unlabeled data to enhance predictive models' performance. This approach operates by harnessing information from both labeled and unlabeled data points within the same task to improve model generalization and performance. In essence, transductive learning exploits the benefits of leveraging unlabeled data and transferring knowledge across related domains or tasks. By integrating information from diverse data sources, it aims to improve model generalization, robustness, and performance across various machine learning applications.

2.3.1 Semi-supervised Learning

Semi-supervised Learning utilizes a combination of labeled and unlabeled data during the model training phase. By incorporating information from both labeled instances (with associated target labels) and unlabeled instances (without target labels), semi-supervised learning aims to enhance model accuracy and robustness. It leverages the intrinsic structure or relationships present in the unlabeled data to supplement the learning process and improve the model's ability to generalize to unseen data.

Various methods have been developed within the realm of semi-supervised learning, each employing distinct strategies to effectively leverage both labeled and unlabeled data. These methodologies encompass a diverse array of approaches.

- **Generative methods**: Semi-supervised generative methods refer to a category of machine learning techniques that combine generative models with partially labeled data to perform tasks such as classification, clustering, or generation of new data points. These approaches harness both labeled and unlabeled data during the training process, exploiting generative models to learn the underlying data distribution and improve the model's performance. Techniques in semi-upervised generative methods include Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Radford et al., 2015), and Variational Auto Encoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014).
- **Consistency regularization methods**: Consistency regularization is a technique used in semi-supervised learning to encourage the model's predictions to remain consistent or stable when the input is subject to perturbations or variations. It aims to improve the generalization of models, especially in scenarios with limited labeled data, by enforcing consistency in predictions for unlabeled data across different perturbed versions of the input (Belkin and Niyogi, 2001; Oliver et al. 2018).
- **Graph-based methods**: Graph-based methods in semi-supervised learning leverage the underlying structure or relationships among data points represented as a graph to improve model performance, particularly when labeled data is limited. These methods utilize the connectivity information present in the data to propagate labels from labeled to unlabeled instances, aiding in predicting the labels of unlabeled data points (Iscen et al, 2019; Chen et al., 2020b; Li et al., 2020b).
- **Pseudo-labelling methods**: Pseudo-labeling is a semi-supervised learning technique that leverages predictions made by a model on unlabeled data to generate pseudo-labels, which are then used to augment the training set for further model training. This approach extends supervised learning to a semi-supervised paradigm by assigning labels (pseudo-labels) to unlabeled data based on model predictions (Blum and Mitchell, 1998; Lee, 2013; Pham et al.,

2021).

In this dissertation, the focus will be on incorporating the consistency regularization technique known as ladder networks (Rasmus et al., 2015) within the framework of semi-supervised learning.

Ladder Networks

Ladder Networks represent a class of deep neural networks introduced to address the challenges of semi-supervised learning by leveraging both labeled and unlabeled data. The concept draws inspiration from both supervised and unsupervised learning paradigms, aiming to improve the generalization and accuracy of models when labeled data is limited.

Key Features of Ladder Networks:

1. **Architecture**: Ladder Networks consist of an encoder-decoder architecture, resembling an autoencoder, with both forward and backward pathways. The encoder processes input data hierarchically through multiple layers, while the decoder attempts to reconstruct the input.
2. **Labeled and Unlabeled Data Integration**: These networks integrate both labeled and unlabeled data within the same framework. The model learns to reconstruct the input data in the decoder pathway while leveraging the supervision from labeled data in the encoder.
3. **Denoising Approach**: Ladder Networks use a denoising approach to handle unlabeled data. They apply noise to the input data and aim to denoise it in the reconstruction process, enabling the network to learn robust representations and reduce overfitting.
4. **Training Strategy**: During training, Ladder Networks optimize both the supervised loss from labeled data and an unsupervised loss from the reconstruction process. The model iteratively refines its representations, improving its ability to generalize.

Advantages of Ladder Networks:

- **Enhanced Generalization**: Ladder Networks aim to learn robust representations from both labeled and unlabeled data, leading to improved generalization and better performance, especially in scenarios with limited labeled data.
- **Semi-Supervised Learning**: They excel in semi-supervised learning settings, where labeled instances are scarce compared to the abundance of unlabeled data. The integration of unsupervised learning principles assists in learning meaningful representations from unlabeled instances.

While Ladder Networks offer promising advantages in leveraging both labeled and unlabeled data, their effectiveness can depend on factors such as architecture design, hyperparameters, and the

specifics of the dataset. Their ability to exploit unlabeled data efficiently while maintaining model accuracy makes them a compelling avenue in the field of semi-supervised learning.

2.3.2 Transfer Learning

Transfer learning involves leveraging knowledge gained from training on one task or domain and applying it to a related but different task or domain. By pretraining a model on a source task with a large dataset and then fine-tuning it on a target task with a smaller dataset, transfer learning mitigates the need for extensive labeled data in the target domain. It facilitates knowledge transfer, enabling the model to extract and transfer learned features, representations, or knowledge from the source to the target task, enhancing the target task's performance.

In the evolution of machine learning paradigms, transfer learning has emerged as a powerful technique, allowing models to leverage knowledge from one domain or task to improve performance in another related domain or task. Initially, traditional transfer learning involved directly transferring knowledge from a pretrained model or source domain to a target domain, often with limited adaptation to the specifics of the target task.

The development from conventional transfer learning to intermediate-task transfer learning represents a significant refinement in this approach (Phang et al., 2018; Clark et al., 2019 Pruksachatkun et al., 2020). Unlike traditional transfer learning, intermediate-task transfer learning introduces an intermediate step, leveraging an intermediate task that shares relevance or similarities with the ultimate target task. This intermediate task serves as a conduit for transferring knowledge, allowing the model to acquire relevant features or representations that are more aligned with the intricacies of the target task.

Intermediate-Task Transfer Learning

Intermediate-task transfer learning involves leveraging knowledge obtained from an intermediate task to enhance learning and performance in a target task. In this approach, a pretrained model is fine-tuned initially on an intermediate task, which might be related to, but not identical to, the ultimate target task. The model learns relevant representations or features during this intermediate task, which are then transferred or fine-tuned further on the target task.

This strategy aims to facilitate better adaptation to the target task by providing the model with additional training on related aspects through the intermediate task. By leveraging the representations learned during the intermediate task, the model gains valuable domain knowledge or features that can

enhance its performance when fine-tuned on the target task, especially in scenarios where labeled data for the target task is limited or unavailable.

Intermediate task transfer learning serves as a bridge between tasks or domains, allowing models to exploit shared information or underlying patterns across tasks. This approach has proven effective in improving generalization, reducing overfitting, and enhancing performance in various machine learning applications, particularly in natural language processing, computer vision, and other domains where labeled data might be scarce or expensive to acquire.

2.4 Evaluation Methods

The evaluation of models constitutes a critical component within the landscape of research and development, serving as the cornerstone for assessing the efficacy, performance, and validity of machine learning algorithms, models, or systems. In the context of this dissertation, the process of model evaluation holds paramount importance as it represents the systematic and rigorous analysis undertaken to measure, validate, and compare the effectiveness of proposed methodologies or approaches.

The primary objective of model evaluation within this dissertation is to methodically examine and quantify the performance, accuracy, robustness, and generalizability of the developed models or techniques. Through a comprehensive evaluation framework, this research endeavors to provide empirical evidence, substantiate claims, and draw meaningful conclusions regarding the applicability and effectiveness of the proposed methodologies within specific contexts or domains.

In the pursuit of thorough model evaluation, various metrics, techniques, and methodologies will be employed to objectively assess and benchmark the performance of the developed models against established standards or alternative approaches. These evaluation metrics encompass a spectrum of criteria, including accuracy, precision, recall, F1-score.

2.4.1 Evaluation Metrics

Evaluation metrics are used to measure the performance of a model or system in various fields, such as machine learning, data science, and information retrieval. These metrics help in assessing how well a model is performing against a specific task or problem.

In the context of classification problems, the terms true positive (TP), true negative (TN), false

positive (FP), and false negative (FN) refer to the results obtained when comparing the predicted outcomes of a model with the actual ground truth, illustrated in Figure 2.2.

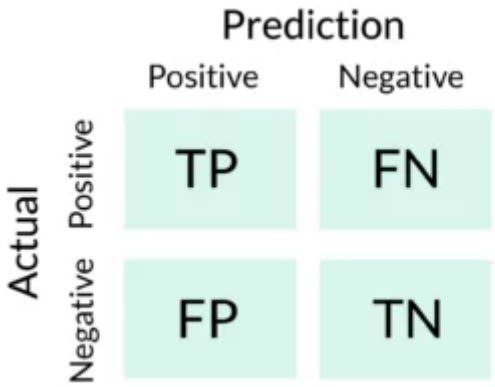


Figure 2.2 Confusion Matrix

- **True Positive (TP)**: A true positive occurs when the model correctly predicts a positive instance as positive. In other words, the model's prediction aligns with the actual positive class.
- **True Negative (TN)**: A true negative occurs when the model correctly predicts a negative instance as negative. The model's prediction matches the actual negative class.
- **False Positive (FP) (Type I Error)**: A false positive occurs when the model incorrectly predicts a negative instance as positive. The model mistakenly identifies a negative sample as belonging to the positive class.
- **False Negative (FN) (Type II Error)**: A false negative occurs when the model incorrectly predicts a positive instance as negative. The model fails to identify a positive sample and incorrectly assigns it to the negative class.

Here are the commonly used evaluation metrics:

- **Accuracy**: The proportion of correctly classified instances among the total instances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \textbf{(2.1)}$$

- **Precision**: The ratio of correctly predicted positive observations to the total predicted positive observations. It measures the accuracy of positive predictions.

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(2.2)}$$

- **Recall (Sensitivity or True Positive Rate)**: The ratio of correctly predicted positive observations to all actual positives. It measures the model's ability to detect all relevant instances.

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(2.3)}$$

- **F1 Score**: The harmonic mean of precision and recall. It provides a balance between precision and recall.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \textbf{(2.4)}$$

When dealing with multi-class classification problems, macro and micro metrics offer distinct perspectives on evaluating model performance across multiple classes.

The micro F1 score aggregates true positives, false positives, and false negatives across all classes, treating each instance equally. It calculates micro-averaged precision, recall, and subsequently the F1 score, emphasizing overall performance without regard to class distribution.

Conversely, the macro F1 score computes precision, recall, and F1 score for individual classes and then takes the unweighted average across all classes, assigning equal importance to each class regardless of its size.

Micro F1 focuses on overall performance across all instances, while macro F1 provides an assessment of the model's ability to generalize across different classes, making both metrics valuable in evaluating the multi-class classification model's performance in varied scenarios, considering both distribution and the overall goal of evaluation.

2.4.2 Cross-validation

Cross-validation is a robust and widely used technique in machine learning for assessing the performance and generalizability of predictive models. It involves partitioning a dataset into subsets (folds) to validate and train a model iteratively, ensuring thorough evaluation and reducing overfitting risks.

The primary objective of cross-validation is to estimate how well a model trained on a particular dataset will generalize to an independent dataset. Commonly employed is k-fold cross-validation, where the dataset is divided into k subsets (or folds). The model is trained on k-1 folds and validated on the remaining fold, repeating this process k times, each time with a different fold held out for validation. Figure 2.3 illustrates k-fold cross-validation.

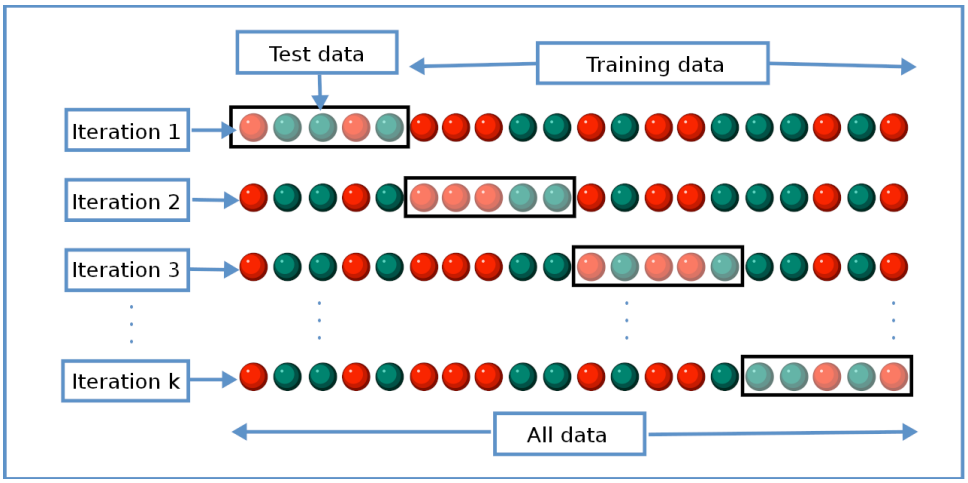


Figure 2.3 K-fold Cross-validation

This approach provides multiple performance estimates, allowing for the computation of an average performance metric across all iterations. It ensures that each data point is used for both training and validation, maximizing the utilization of available data and reducing the impact of dataset variability on model performance assessment.

Cross-validation aids in detecting issues related to model overfitting or underfitting by providing more robust estimates of a model's performance on unseen data. It is particularly valuable in scenarios where datasets are limited, enabling researchers to extract maximum information from the available data.

Overall, cross-validation serves as a pivotal technique in model evaluation, offering a comprehensive and reliable assessment of a model's performance, robustness, and generalization capabilities across different data subsets. Its usage helps ensure that the conclusions drawn from model performance are more reliable and indicative of real-world applicability.

Chapter 3
Semi-supervised Learning for Truncated Documents

Automatic peer-review aspect score prediction (PASP) of academic papers can be a helpful assistant tool for both reviewers and authors. Most existing works on PASP utilize supervised learning techniques. However, the limited number of peer-review data deteriorates the performance of PASP. This work presents a novel semi-supervised learning (SSL) method that incorporates the Transformer fine-tuning into the Γ-model, a variant of the Ladder network, to leverage contextual features from unlabeled data. Backpropagation simultaneously minimizes the sum of supervised and unsupervised cost functions; it can be easily trained in an end-to-end fashion. The proposed method is evaluated on the PeerRead benchmark. The experimental results demonstrate that our model outperforms the supervised and naive semi-supervised learning baselines.

3.1 Introduction

Over the past few years, the number of submissions for AI-related international conferences and journals has increased substantially, making the review process more challenging. Automatic peer-review aspect score prediction (PASP) scores academic papers on a numeric range of different qualities along with aspects such as *"clarity"* and *"originality"*. It can be a helpful assistant tool for both reviewers and authors. PeerRead is the first publicly available dataset of scientific peer reviews for research purposes (Kang et al., 2018). It can be used in various ways, such as paper acceptance classification (Ghosal et al., 2019; Maillette de Buy Wenniger et al., 2020; Fytas et al., 2021) and review aspect score prediction (Li et al., 2020a; Wang et al., 2020). Alternatively, the dataset is modified for citation recommendation (Jeong et al., 2019) and citation count prediction (Dongen et al., 2020).

Much of the previous work on PASP is based on supervised learning (Kang et al., 2018; Li et al., 2020a). However, the dataset with annotated aspect scores is relatively very small, which deteriorates overall performance. To mitigate the drawback and improve the performance of PASP, we propose a semi-supervised learning (SSL) method that can leverage contextual features from the larger unannotated dataset. SSL has been widely utilized in many NLP tasks, such as classification (Miyato et al., 2017; Li et al., 2021), sequence labeling (Yasumasa et al., 2018; Chen et al., 2020a), and parsing (Zhang and Goldwasser, 2020; Lim et al., 2020). It has shown to be effective for learning models by

leveraging a large amount of unlabeled data to compensate for the lack of labeled data. SSL is also beneficial for PASP because an enormous body of publications is available online, and unlabeled data, i.e., scholarly papers, can often be obtained with minimal effort. Recently, transformer-based pretraining language models (LM) such as BERT (Devlin et al., 2019) and its variants have been very successful as many NLP tasks which utilize these LM attained unprecedented performances. In this work, we combine the strengths of both techniques and propose a Transformer-based Γ-model (Γ-Trans) that incorporates a pretrained transformer into the Γ-model (Rasmus et al., 2015), a variant of ladder network (Valpola, 2014; Rasmus et al., 2015), SSL autoencoder. The unsupervised part of Γ-Trans utilizes a denoising autoencoder to help focus on relevant features derived from supervised learning.

The contributions of our work can be summarized as follows:

- We propose Γ-Trans for PASP that incorporates a pretrained transformer into SSL by fine-tuning the model using labeled and unlabeled data simultaneously.
- The experimental results show that Γ-Trans outperforms the supervised learning baselines and naive SSL methods with a small amount of labeled training data.
- We compare several BERT variants and the size of unlabeled to examine the effectiveness of Γ-Trans for PASP.

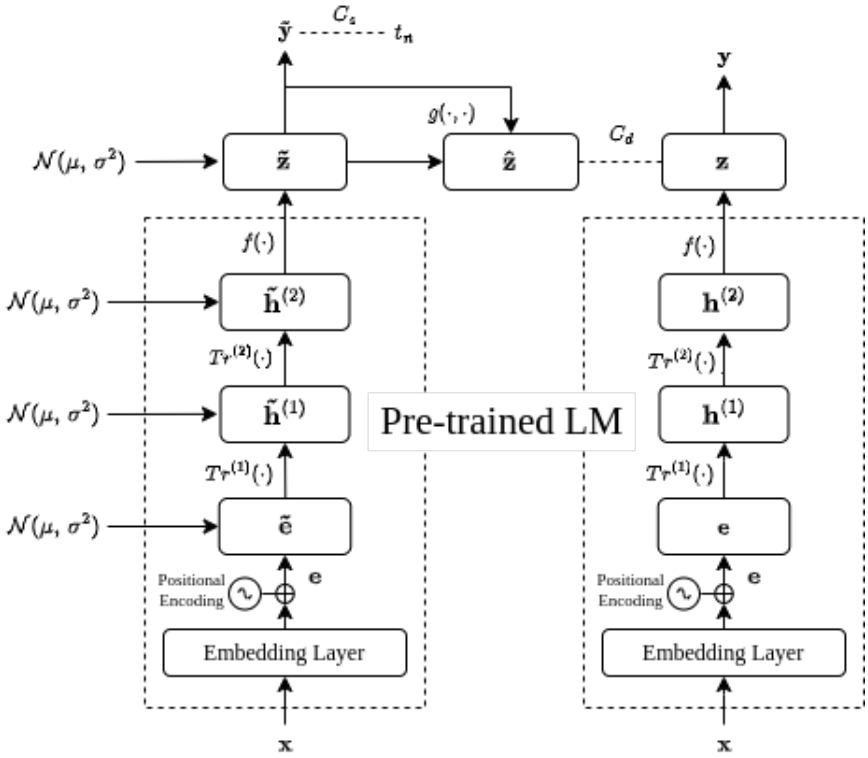


Figure 3.1 Γ-Trans architecture. The pretrained transformer has two layers which are shown in a dotted frame. The model is fine-tuned by supervised cost $\boldsymbol{C_s}$ and denoising cost $\boldsymbol{C_d}$.

3.2 Γ-Transformers

The existing works applying ladder networks to the NLP task, e.g., information extraction (Nagesh and Surdeanu, 2018) and sentiment analysis (Pan et al., 2020; Zheng et al., 2021). The latter utilizes the encoder of the ladder network (Rasmus et al., 2015) to extract the features from the pretrained LM without fine-tuning it. By freezing the features from the LM, the model only utilizes the fully connected layers from the encoder of the network without exploiting the transformer layer of the LM. To mitigate the issue, we fine-tune the LM along with training the Γ-model as well as acquiring the sequence embedding from the pretrained LM. The model can be plugged into any feedforward network without decoder implementation, i.e., the denoising cost is only on the top layer of the model.

Figure 3.1 illustrates the Γ-Trans network. Let x be the input and y be the output with targets t. The labeled training data of size N consists of pairs $\{x(n), t(n)\}$, where $1 \leq n \leq N$. The unlabeled data of size M has only input x without the targets t, an $x(n)$, where $N + 1 \leq n \leq N + M$. As shown in Figure 3.1, the network consists of two forward passes, the clean path and the corrupted pass. The

former is illustrated in a dotted frame on the right-hand side in Figure 3.1 and produces clean z and y, which are given by:

$$z = f\left(h^{(\mathrm{L})}\right) = N_B\left(Wh^{(L)}\right) \quad \textbf{(3.1)}$$

$$\boldsymbol{y} = \boldsymbol{\phi}\left(\boldsymbol{\gamma} \cdot (\boldsymbol{z} + \boldsymbol{\beta})\right) \quad \textbf{(3.2)}$$

$$\boldsymbol{h^{(0)}} = \boldsymbol{e} \quad \textbf{(3.3)}$$

$$\boldsymbol{h^{(l)}} = \boldsymbol{Tr^{(l)}}\left(\boldsymbol{h^{(l-1)}}\right) \quad \textbf{(3.4)}$$

where e denotes the input embedding of x with positional encoding, $Tr^{(l)}$ refers to the transformer block at layer l in the L-layer pretrained LM (e.g., BERT), and N_B indicates a batch normalization. W shows the weight matrix of the linear transformation $f(\cdot)$. ϕ refers to an activation function, where β and γ are trainable scaling and bias parameters, respectively.

The clean path shares the mappings $Tr^{(l)}$ and f with the corrupted path. The corrupted $\tilde{x}$ and $\tilde{y}$ are produced by adding Gaussian noise n in the corrupted path (left-hand side of Figure 3.1):

$$\tilde{z} = f\left(\tilde{h}^{(L)}\right) + n \quad \textbf{(3.5)}$$

$$\tilde{y} = \phi\left(\gamma \cdot (\tilde{z} + \beta)\right) \quad \textbf{(3.6)}$$

$$\tilde{h}^{(0)} = \tilde{e} + n \quad \textbf{(3.7)}$$

$$h^{(l)} = Tr^{(l)}\left(\tilde{h}^{(l-1)}\right) + n \quad \textbf{(3.8)}$$

A supervised cost C_s is the average negative log-probability of the noisy output $\tilde{y}$ matching the target t_n given the input x_n:

$$C_s = -\frac{1}{N}\sum_{n=1}^{N} \log P\left(\tilde{y} = t_n | x_n\right) \quad \textbf{(3.9)}$$

where N denotes the number of labeled data. Given the corrupted $\tilde{z}$ and prior information $\tilde{y}$, the denoising function g reconstructs the denoised $\hat{z}$:

$$\hat{\boldsymbol{z}} = \boldsymbol{g}(\tilde{\boldsymbol{z}}, \boldsymbol{u}) \quad \textbf{(3.10)}$$

$$\boldsymbol{u} = \boldsymbol{N_B}(\tilde{\boldsymbol{y}}) \quad \textbf{(3.11)}$$

where g is identical to the one of Rasmus et al.'s (2015) consisting of its own learnable parameters. The unsupervised denoising cost function is given by:

$$C_d = \frac{1}{N+M} \sum_{n=1}^{N+M} \frac{\lambda}{d} \| z_n - N_B(\hat{z}_n) \| \quad (3.12)$$

where M indicates the number of unlabeled data, λ is a coefficient for unsupervised cost, and d refers to the width of the output layer. The final cost C is given by:

$$C = C_s + C_d \quad (3.13)$$

3.3 Experiments

3.3.1 Experimental Settings

We performed the experiments on the ACL dataset with the score of review aspects that are included in the PeerRead Dataset (Kang et al., 2018). The aspect score annotation details are provided in Appendix A. We used the mean score of multiple reviews and classified them ranging from 1 to 5 into two classes: ≥ 4 (Positive) and < 4 (Negative). We balanced the data, i.e., the same size of two classes, by randomly down sampling the majority class. Table 3.1 shows the statistics of the dataset. Although the PeerRead dataset contains both paper and review texts, we only used the papers to predict the aspect scores. We utilized the first 512 tokens of the paper according to the maximum length of the most common pretrained LM, BERT (Devlin et al., 2019). For the unlabeled data, we also used the ACL papers obtained from ScisummNet Corpus (Yasunaga et al., 2019), which provides 1,000 papers in the ACL anthology. We used 5-fold cross-validation to evaluate all systems with an 80/20 split for the train and test sets. We selected the best model based on the performance of the test set. The final result is calculated from the average of the five folds. As the evaluation metric, we used accuracy and F1-score.

Aspect	**#Pos (Neg)**
Clarity (Clr)	40
Originality (Ori)	59
Impact (Imp)	22
Meaningful Comparison (Com)	52
Soundness/Correctness (Cor)	54
Substance (Sub)	66
Overall Recommendation (Ova)	60

Table 3.1 Statistics of the ACL Dataset. #Pos (Neg) refers to the equal number of papers for each class.

3.3.2 Baselines and Implementation Details

We compare Γ-Trans with supervised learning and semi-supervised learning baselines.

Supervised Learning

- **BERT-base** (Devlin et al., 2019) - A pretrained LM. We fine-tuned the model on the PASP task.
- **PeerRead (PR)** - Similar to Kang et al.'s (2018), we implemented a GRU (Gated Recurrent Unit) model (Cho et al., 2014) using GloVe embeddings (Pennington et al., 2014) as input word representations without tuning.
- **ReviewRobot (RR)** (Wang et al., 2020) - This method extracts evidence by comparing the knowledge graph of the target paper and a large collection of background papers and uses the evidence to predict scores.
- **Multi-task** (Li et al., 2020a) - A multi-task approach that automatically selects shared network structures and other review aspects as auxiliary resources. The model is based on CNN text classification model.

Semi-Supervised Learning

- **Virtual Adversarial Training (VAT)** (Miyato et al., 2017) - This method exploits information from unlabeled data by applying perturbations to the word embeddings in a neural network.
- **Γ-model** (Rasmus et al., 2015) - It is a variant of ladder networks in which a denoising cost is only on the top layer and means that most of the decoder can be omitted.
- **Ladder** - A deep denoising autoencoder with skip connections and reconstruction targets in the intermediate layers (Rasmus et al., 2015).

The Γ-model and Ladder employ a ladder network on top of frozen BERT-base representations.

Implementation details

- **Fine-tuning BERT**: We used Huggingface's Transformers package to fine-tune BERT. We fine-tuned the model with learning rate of 1e-6 until convergence with a batch size of 8, maximal sequence length of 512. Optimization was done using Adam with warm-up of 0.1 and weight decay of 0.01.
- **PeerRead model**: We used a simple MLP with a single hidden layer of 100 neurons with the last recurrent state of a single GRU layer of 100 units. We trained the MLP until convergence,

using Adam optimizer, a learning rate of 1e-4 with a batch size of 8 and an L2 penalty of 1.

- **VAT**:
    - **Recurrent LM Pretraining**: We used a unidirectional single-layer LSTM with 1,024 hidden units. The dimension of word embedding was 256. For the optimization, we used the Adam optimizer with a batch size of 32, an initial learning rate of 0.001, and a 0.9999 learning rate decay factor. We trained for 50 epochs. We applied gradient clipping with norm set to 5.0. We used dropout on the word embedding layer and an output layer with a 0.5 dropout rate.
    - **Model Training**: We added a hidden layer between the softmax layer for the target and the final output of the LSTM. The dimension is set to 30. For optimization, we also used the Adam optimizer, with a 0.001 initial learning rate and 0.9998 exponential decay. Batch sizes are set to 32 and 96 for calculating the loss of virtual adversarial training. We trained for 30 epochs. We applied gradient clipping with the norm as 5.0.
- **Multi-task**: We modified the model from performing a regression task to a classification task by changing the output layer. We used CNN with 64 filters and filter width of 2. We used fastText as initial word embeddings. The hidden dimension was 1024. We trained the model using Adam optimizer with learning rate 0.001 and batch size of 8. We trained all of the candidate multi-task models for one and two auxiliary tasks to find the best one.
- **Γ-model and Ladder**: We used the layer sizes of the ladder network to be 768-100-500-250-250-250-2, according to the BERT's representation dimension and the number of output classes. We set the denoising cost multipliers λ to [1000, 10, 0.1, 0.1, 0.1, 0.1, 0.1] from the input layer to the output layer for the Ladder, and [0, 0, 0, 0, 0, 0, 1] for the Γ-model. The std of the Gaussian corruption noise n is set to 0.3. We trained the model with a learning rate of 3e-3 until convergence with a batch size of 8 for each labeled and unlabeled data, 16 in total. Optimization was done using Adam with weight decay of 0.01.
- **Γ-Trans**: We used Huggingface's Transformers package to fine-tune transformer-based pretrained LMs. The denoising cost multipliers λ is set to 1. We set the std of the Gaussian corruption noise n to 0.3 in both Γ-model and Ladder. For optimization, we used the Adam optimizer, with a 1e-4 initial learning rate, 0.01 weight decay, and 0.1 warm-up. Batch size is set to 8 for both labeled and unlabeled data, 16 in total.

3.3.3 Results and Discussion

Table 3.2 shows the results. We can see from Table 3.2 that the SSL methods, Ladder and Γ-Trans, outperform all supervised learning baselines, and he results by Γ-Trans are the best among other SSL methods on average. This shows that our assumption, incorporating fine-tuning the pretrained LM into the ladder network, helps improve the performance significantly. BERT has the worst performance and even performs worse than other supervised learning baselines that utilize a common neural network layer, GRU or CNN. It is probably because the number of supervised data alone is insufficient to tune millions of parameters of BERT.

Metric	Aspect	Supervised Learning Methods				Semi-supervised Learning Methods			
		BERT	**PR**	**RR**	**Muti**	**VAT**	**Γ-model**	**Ladder**	**Γ-Trans**
Acc.	Clr	0.613	0.600	0.541	0.713	0.613	0.675	0.688	**0.763**
	Ori	0.508	0.593	0.659	0.525	0.593	0.559	0.610	**0.661**
	Imp	0.591	0.705	0.606	0.708	0.750	0.750	**0.841**	0.818
	Com	0.538	0.625	0.621	0.673	0.615	0.577	0.673	**0.692**
	Cor	0.546	0.639	0.529	0.509	0.556	0.648	0.657	**0.713**
	Sub	0.538	0.644	-	0.585	**0.667**	0.644	0.621	**0.667**
	Ova	0.575	0.625	0.535	**0.698**	0.658	0.558	0.683	0.683
	Avg.	0.559	0.633	0.582	0.630	0.636	0.630	0.682	**0.714**
F1	Clr	0.632	0.658	0.671	0.706	0.674	0.739	0.721	**0.790**
	Ori	0.605	0.652	0.651	0.675	0.529	0.606	0.642	**0.743**
	Imp	0.627	0.740	0.717	0.746	0.718	0.797	**0.864**	0.848
	Com	0.597	0.672	0.626	0.655	0.623	0.674	0.723	**0.729**
	Cor	0.606	0.655	0.588	0.615	0.529	0.687	0.698	**0.763**
	Sub	0.601	0.696	-	0.627	0.639	0.701	0.718	**0.747**
	Ova	0.608	0.693	0.520	0.682	0.637	0.663	0.732	**0.749**
	Avg.	0.611	0.681	0.629	0.672	0.621	0.695	0.728	**0.767**

Table 3.2 Experimental results. Best result is in bold**,** and 2nd best is underlined

Among the prediction of aspects, *Impact* aspect is the best score in both metrics. We investigated the distribution of each aspect score from the data and found that more than 60% of the papers whose impact score is ≥ 4 also have a score of ≥ 4 in other aspects, while other aspects are not. This indicates that the Impact aspect has relatively distinctive features compared with other aspects. In contrast, *Meaningful Comparison* score prediction has the worst performance. One possible reason is the limited length of the input sequence, i.e., the first 512 tokens. This data length includes abstract and introduction sections but does not include related work section which deteriorates the performance of *Meaningful Comparison* score.

We recall that Γ-Trans fine-tunes the LM through training the ladder network. To examine how

the LM affects the overall performance on PASP, we tested several pretrained LMs. Table 3.3 shows the *Overall recommendation* score prediction by F1 obtained from several transformer-based pretrained LMs with Γ-Trans and the second-best method, Ladder. Our approach can generate better results in all models.

We can see that SciBERT (Beltagy et al., 2019), a BERT model pretrained on a large corpus of scientific publications, improves the performance, while RoBERTa (Liu et al., 2019b) does not, compared to BERT. Table 3.3 also shows that Longformer performs better than BERT on Γ-Trans, but not Ladder. This indicates that a longer sequence of textual information helps improve the performance of PASP. In contrast, Ladder does not work well with Longformer. One reason is that Ladder cannot utilize the attention mechanism of Longformer for the different domains of ACL papers as it only employs the sequence embeddings obtained from the Longformer.

Model	**Ladder**	**Γ-Trans**
BERT	0.732	0.749
RoBERTa	0.694	0.712
SciBERT	**0.744**	**0.774**
Longformer	0.686	0.756

Table 3.3 F1 on *Overall recommendation* score prediction. Comparison between Ladder and Γ-Trans on different transformer-based pretrained LMs.

We also examined how the number of unlabeled data for training affects overall performance. Figure 3.2 shows the F1-score of the SSL methods against the number of unlabeled data obtained by 5-fold cross-validation. Overall, the graph shows that more unlabeled data helps improve the performance in every SSL method except VAT, whose performance drops at 1,000 unlabeled data. Γ-Trans consistently outperformed other SSL methods, and especially the result with 100 unlabeled data outperformed other methods with 700 unlabeled data.

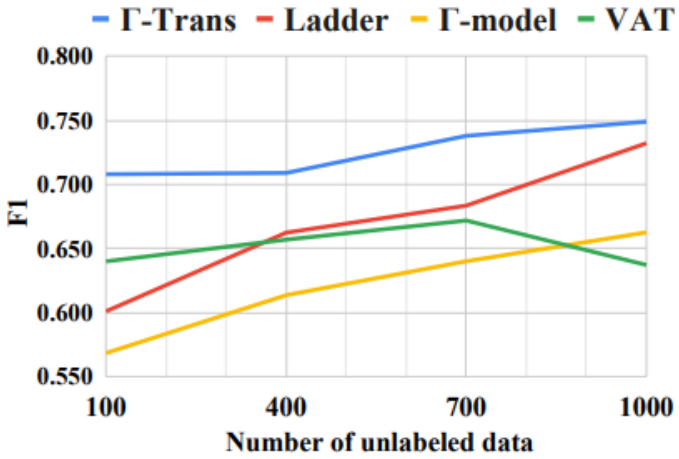


Figure 3.2 F1 score against the number of unlabeled data on *Overall recommendation* score prediction.

3.3.4 Error Analysis

We analyzed the prediction probability on the *Overall Recommendation* aspect test data. The average probability of the selected class is 50.26% which is relatively low. The close probability of two classes indicates that the extracted features between the two classes are not much different from each other. The average probabilities of the correct and incorrect predictions are 50.30% and 50.13%, respectively, showing no significant difference.

Figure 3.3 shows the ratio of the predictions between negative and positive. Our model tends to bias toward positive prediction in every aspect. The most biased prediction is *Meaningful Comparison*, with 84.31% on positive. One reason is that several reviewers are assigned to one paper. Assume that a sample labeled negative has a score of 3, 3, and 4. (The sample is labeled negative because the average of these scores is less than 4.) Such a sample has some positive features to trigger the model to predict it as positive. In contrast, there was no such case for positive samples.

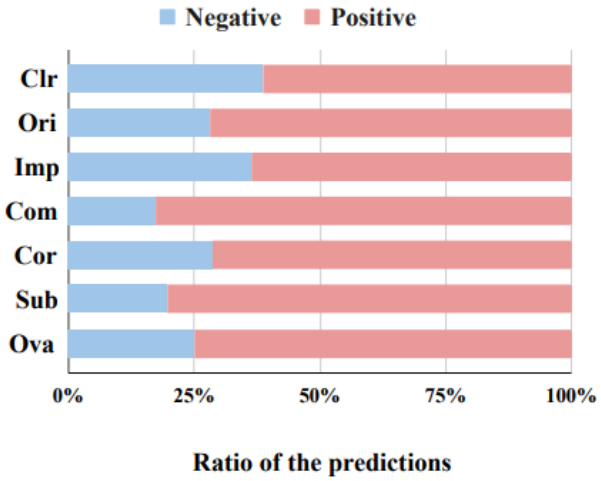


Figure 3.3 Ratio between the number of negative predictions and positive predictions of each aspect.

We further investigated more on the negative predictions. Table 3.4 shows the precision of negative samples. Although our model predicts a positive outcome more than a negative one, the precision on the negative is very high. The highest precision is 0.938 on the *Impact* aspect and the lowest one is higher than 0.8. High precision on negative samples means a high measure of quality that indicates that our model is suitable for the first screen to filter out poor-quality works. Moreover, it is also helpful to authors for their first draft.

Aspect	Neg Precision
Clarity	0.839
Originality	0.913
Impact	0.938
Meaningful Comparison	0.833
Soundness Correctness	0.870
Substance	0.923
Overall Recommendation	0.867

Table 3.4 Precision of negative samples

3.4 Limitation

We should be able to obtain further advantages in efficacy in our pretrained LM. We utilized the first 512 tokens in the input paper and 768 dimensions of the hidden layer as most of the pretrained LM restricts text length and embedding size which may lead to a lack of contextual information about aspects. Furthermore, in our experiment, fine-tuning Longformer by freezing the first ten layers on 1,000 tokens required around 50GB of GPU memory. We would improve our Γ-Trans model so that we can process papers consisting of long token sequences.

3.5 Summary

In this work, we focused on the PASP task and proposed a method, Γ-Trans, that incorporates the Transformer fine-tuning technique into the Γ-model of the Ladder networks. The experimental results showed the effectiveness of our model as our model attained the best accuracy and F1 on average. Through the experiments, we found that our method helps improve the performance of all pretrained LMs including SciBERT and Longformer. Future work will include (i) extending the method for imbalanced aspect score datasets, (ii) exploiting the related information between aspects, and (iii) generating knowledgeable and explainable review comments.

Chapter 4
Semi-supervised Learning for Full Documents

In Chapter 3, our focus lies in introducing a semi-supervised learning approach that integrates pretrained transformers within a ladder network framework. However, a significant limitation arises as these pretrained transformers struggle to effectively process the entire length of an academic paper. To address this limitation, Chapter 4 introduces the Long-Short Transformer (Transformer-LS) specifically designed to handle long sequences. This Transformer-LS is incorporated into the Γ-model, a variant of the Ladder network that employs a denoising autoencoder to reconstruct input data from a corrupted version. The objective is to minimize the reconstruction error of auxiliary unlabeled data, thereby aiding in training the classifier. Empirical validation demonstrates the notable superiority of our system when compared to both supervised and naive semi-supervised learning baselines, particularly on the PeerRead benchmark. The successful implementation of Transformer-LS within the Γ-model showcases its efficacy in handling extensive academic paper lengths while enhancing classification performance through auxiliary unlabeled data reconstruction.

4.1 Introduction

The increasing number of submissions to AI-related international conferences and journals has made the review process more challenging. Automatic peer-review aspect score prediction (PASP) is a valuable tool for improving the efficiency and effectiveness of the review process by providing reviewers and authors with a numeric score for different qualities of a paper, such as clarity and originality. The PeerRead dataset (Kang et al., 2018) is the first publicly available collection of scientific peer reviews for research purposes and has been used in a variety of applications, including paper acceptance classification (Ghosal et al., 2019; Wenniger et al., 2020; Fytas et al., 2021), review aspect score prediction (Li et al., 2020a; Wang et al., 2020), citation recommendation (Jeong et al., 2019), and citation count prediction (Dongen et al., 2020).

Previous work on PASP has heavily relied on supervised learning techniques (Kang et al., 2018; Li et al., 2020a). However, the available annotated datasets for this task are very restricted, which limits the overall performance of PASP models. To address this issue and improve PASP performance, we propose a semi-supervised learning (SSL) method that leverages contextual features from a larger, unlabeled dataset. Semi-supervised learning has been widely used in various natural language

processing (NLP) tasks, including classification (Miyato et al., 2017; Li et al., 2021), sequence labeling (Yasunaga et al., 2018; Chen et al., 2020a), and parsing (Zhang and Goldwasser, 2020; Lim et al., 2020). It has been shown to be effective in model learning by leveraging a large amount of unlabeled data to compensate for the lack of labeled data. Semi-supervised learning is particularly useful for PASP, as a vast number of scholarly papers are available online and can be easily obtained as unlabeled data.

Recently, transformers (Vaswani et al., 2017) have achieved state-of-the-art results in a wide range of NLP tasks. However, transformer-based models are unable to process long sequences, such as academic papers, due to their self-attention operation, which scales quadratically with the sequence length. In this work, we propose a semi-supervised learning technique for PASP that is capable of handling long sequences. Our approach is based on the combination of ladder networks (LNs) (Valpola, 2014; Rasmus et al., 2015) and the Long-short transformer (Transformer-LS) (Zhu et al., 2021). Ladder networks are a type of deep denoising autoencoder that incorporates skip connections and reconstruction targets at intermediate layers, while Transformer-LS is a transformer with a self-attention mechanism that is efficient for modeling long sequences with linear complexity. We propose the Γ- Transformer-LS (Γ-TLS), which integrates a Transformer-LS into the Γ-model (Rasmus et al., 2015), a variant of ladder networks. The unsupervised component of Γ-TLS utilizes a denoising autoencoder to help focus on relevant features derived from supervised learning.

To the best of our knowledge, our work is one of the first applications of SSL to the PASP task. Specifically, our contributions are as follows:

- We propose Γ-TLS for PASP that incorporates a Transformer-LS into SSL by training the model using labeled and unlabeled data simultaneously.
- The experimental results show that Γ-TLS outperforms the supervised learning baselines and naive SSL methods on the PeerRead benchmark.

4.2 Γ-Transformer-LS (Γ-TLS)

To overcome the limitation of the vanilla transformer (Vaswani et al., 2017) for long sequences, we adopt the Transformer-LS as the encoder of our framework. Transformer-LS is more memory and computationally efficient than the previous larger models, Longformer (Beltagy et al., 2020) and Transformer-XL (Dai et al., 2019). For the SSL technique, we choose a denoising network called the Γ-model (Rasmus et al., 2015), which is a variant of ladder networks (LNs). The Γ-model eliminates most of the decoder, retaining only the top layer, which allows it to be easily integrated into any

network without implementing a separate decoder. The encoder in the Γ-model still includes both the clean and corrupted paths, as in the full ladder network (LN).

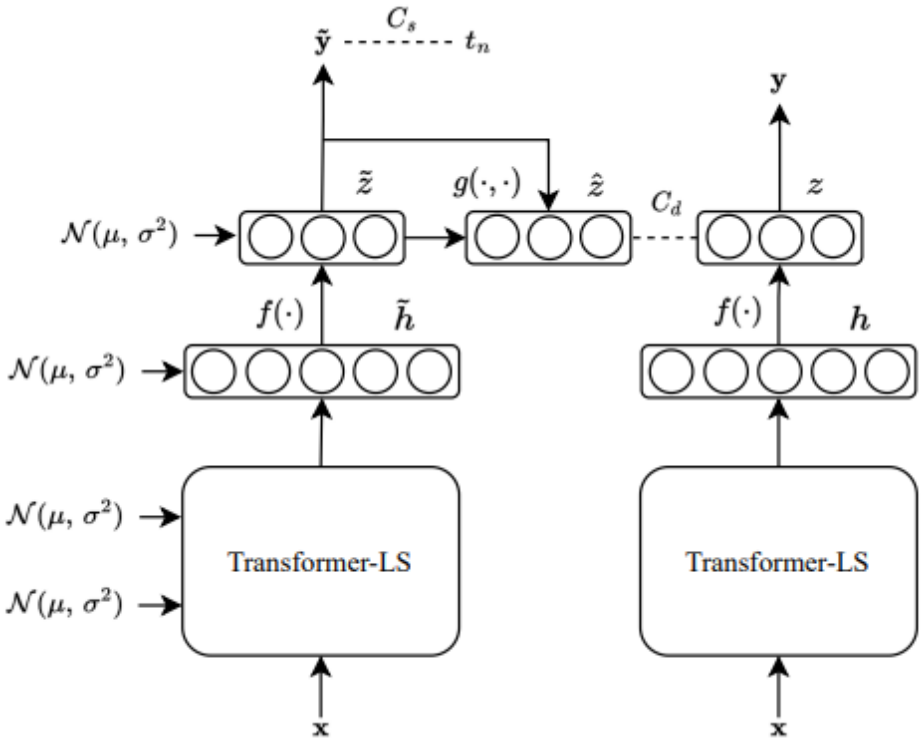


Figure 4.1 Γ-TLS architecture. The corrupted path shown on the left-hand side shares the Transformer-LS's weights and mapping f with the clean path on the right-hand side.

Figure 4.1 illustrates the Γ-Transformer-LS (Γ-TLS). Let x be the input and y be the output with targets t. The supervised data of size N consists of pairs $\{x(n), t(n)\}$, where $1 \leq n \leq N$. The unsupervised data of size M has only input x without the targets t, an $x(n)$, where $N + 1 \leq n \leq N + M$. The network comprises two forward passes, the clean path, and the corrupted path. The clean path, illustrated on the right-hand side in Figure 4.1, produces clean representation z and clean output y, given by:

$$\boldsymbol{z = f(h) = N_B(Wh)} \quad \textbf{(4.1)}$$

$$\boldsymbol{y = \phi(\gamma \cdot (z + \beta))} \quad \textbf{(4.2)}$$

$$\boldsymbol{h = TLS(x)} \quad \textbf{(4.3)}$$

where h denotes the hidden representation obtained from Transformer-LS (TLS), W is the weight matrix of the linear transformation f, and N_B indicates a batch normalization. ϕ refers to an activation function, where β and γ are trainable scaling and bias parameters, respectively.

The corrupted representation $\tilde{z}$ and corrupted output $\tilde{y}$ are produced by adding Gaussian noise

n in the corrupted path (left-hand side of Figure 4.1). The noise n is applied to the output of each layer of the Transformer-LS (TLS):

$$\tilde{z} = f(\tilde{h}) + n \quad \textbf{(4.4)}$$

$$\tilde{y} = \phi(\gamma \cdot (\tilde{z} + \beta)) \quad \textbf{(4.5)}$$

$$\tilde{h} = TLS(x) + n \quad \textbf{(4.6)}$$

The supervised cost C_s is the average negative log-probability of the corrupted output $\tilde{y}$ matching the target t_n given the input x_n:

$$C_s = -\frac{1}{N}\sum_{n=1}^{N} \log P\,(\tilde{y} = t_n | x_n) \quad \textbf{(4.7)}$$

Given the corrupted $\tilde{z}$ and prior information $\tilde{y}$, the denoising function g reconstructs the denoised $\hat{z}$:

$$\hat{z} = g(\tilde{z}, u) \quad \textbf{(4.8)}$$

$$u = N_B(\tilde{y}) \quad \textbf{(4.9)}$$

where g is identical to the one of the LN (Rasmus et al., 2015) consisting of its own learnable parameters. The unsupervised denoising cost function is given by:

$$C_d = \frac{1}{N+M}\sum_{n=1}^{N+M} \frac{\lambda}{d} \| z_n - N_B(\hat{z}_n) \| \quad \textbf{(4.10)}$$

where λ is a coefficient for unsupervised cost, and d refers to the width of the output layer. The final cost C is given by:

$$C = C_s + C_d \quad \textbf{(4.11)}$$

4.3 Experiments

4.3.1 Setup

Data

The ACL 2017 dataset, included in PeerRead (Kang et al., 2018), is used as evaluation data for our PASP system. The ACL dataset consists of 7 different aspects of scores as listed in Table 4.1. These aspect scores were derived from a mean of multiple reviews and classified into two categories: positive (scores of 4 or higher) and negative (scores lower than 4). Although the PeerRead dataset contains both paper and review texts, we only used the papers to predict the aspect scores. We utilized the first 8,192 tokens of the paper as the input. We used SciVocab (Beltagy et al., 2019) WordPiece vocabulary for tokenization. For the unlabeled data, we used the ACL papers from ScisummNet Corpus (Yasunaga et al., 2019), which provides 999 papers in the ACL anthology. To evaluate all systems, we employed a 5-fold cross-validation strategy, in which the final result was calculated as the average of the five folds. As the evaluation metrics, we utilized both accuracy and Macro F1 score. This allows us to comprehensively assess the performance of our systems in terms of both the proportion of correct predictions and the balance between precision and recall.

Aspect	**#Neg / #Pos**	**Total**
Clarity (Clr)	39 / 97	136
Originality (Ori)	58 / 78	136
Impact (Imp)	110 / 22	132
Meaningful comparison (Com)	80 / 52	132
Soundness correctness (Cor)	54 / 82	136
Substance (Sub)	66 / 70	136
Overall recommendation (Ova)	76 / 60	136

Table 4.1 Statistics of the ACL Dataset.

To evaluate all systems, we employed a 5-fold cross-validation strategy, in which the final result was calculated as the average of the five folds. As the evaluation metrics, we utilized both accuracy and Macro F1 score. This allows us to comprehensively assess the performance of our systems in terms of both the proportion of correct predictions and the balance between precision and recall.

Baseline models

The competitor algorithms that are used as baselines for our model are the following:

- **Convolutional Neural Network (CNN)** - We implemented a CNN model similar to one in PeerRead (Kang et al., 2018). The outputs from the CNN model are passed through a max pooling layer and finally through the final linear layer.
- **Virtual Adversarial Training (VAT)** (Miyato et al., 2017) - An SSL method that exploits information from unlabeled data by applying perturbations to the word embeddings in a neural network. The model utilizes LSTM to learn from sequential inputs.
- **Hierarchical Attention Network (HAN)** (Yang et al., 2016) - A hierarchical attention network for document classification. The model consists of two levels of attention mechanisms at the word and sentence levels to construct the document representation.
- **Multi-task** (Li et al., 2020a) - A multi-task approach that automatically selects shared network structures and other review aspects as auxiliary resources. The model is based on the CNN text classification model.
- **Transformer-LS** (Zhu et al., 2021) - A transformer for modeling long sequences with linear complexity. We used the output of the last layer of the [CLS] token as the document representation for the classifier.

Implementation details

- **CNN**: We used a simple MLP with a single hidden layer of 128 neurons with the max pooling of a single 1D-CNN layer of 128 filters and window width 5. We used a random initialization for the word embedding size of 128 and trained it with the model. We trained the model using AdamW optimizer on a linear scheduler, a learning rate of 1e-4 with a batch size of 8.
- **HAN**: We set the max sentence length to 100 tokens and the max number of sentences to 600. We used a bidirectional single-layer GRU size of 100 with an attention mechanism to aggregate the representation on both word and sentence levels. We also used a random initialization for the word embeddings size of 300. The model was trained on AdamW optimizer, learning rate of 5e-5, and batch size of 8.
- **VAT**:
    - **Recurrent LM Pretraining**: We used a unidirectional single-layer LSTM with 128 hidden units. The dimension of word embedding was 128. For the optimization, we used the Adam optimizer with a batch size of 32, an initial learning rate of 0.001, and a 0.9999 learning rate decay factor. We trained for 50 epochs. We applied gradient clipping with norm set to 5.0. We used dropout on the word embedding layer and an output layer with a 0.5 dropout rate.

- **Model Training**: We added a hidden layer between the softmax layer for the target and the final output of the LSTM. The dimension is set to 30. For optimization, we also used the Adam optimizer, with a 0.001 initial learning rate and 0.9998 exponential decay. Batch sizes are set to 32 and 96 for calculating the loss of virtual adversarial training. We trained for 30 epochs. applied gradient clipping with the norm as 5.0.

- **Multi-task**: We modified the model from performing a regression task to a classification task by changing the output layer. We used CNN with 64 filters and filter width of 2. We used fastText as initial word embeddings. The hidden dimension was 1024. We trained the model using Adam optimizer with learning rate 0.001 and batch size of 8. We trained all of the candidate multi-task models for two auxiliary tasks to find the best one.

- **Transformer-LS**: We used two layers of transformer-ls size 256 with 4 attention heads. The local window attention was set to 128. A [CLS] token was used as a global token. We used dropout and attention dropout of 0.1. We trained the model using AdamW optimizer on a linear scheduler with batch size 8. We tuned the learning rate in the range of {1e-2, 1e-3, 1e-4}

- **Γ-TLS**: We used the same architecture as the Transformer-LS (A.5). The denoising cost multipliers λ is set to 1. We tuned the std of the Gaussian corruption noise in the range of {0.1, 0.2, 0.3}. We also tuned the learning rate in the range of {1e-2, 1e-3, 1e-4}. Batch size is set to 8 for both labeled and unlabeled data, 16 in total.

Metric	Models	Clr	Ori	Imp	Com	Cor	Sub	Ova	Avg.
Acc.	CNN	0.721	0.595	0.833	0.636	0.640	0.566	0.611	0.657
	VAT	0.728	0.669	0.841	0.614	0.669	0.662	0.654	0.691
	HAN	0.720	**0.690**	0.841	0.674	0.684	0.654	0.692	0.708
	Multi-task	0.736	0.661	**0.864**	**0.713**	**0.698**	0.617	0.670	0.708
	Transformer-LS	0.735	0.646	0.856	0.696	0.647	0.677	0.684	0.706
	Γ-TLS (Ours)	**0.757**	0.654	0.856	0.703	**0.698**	**0.706**	**0.728**	**0.729**
F1.	CNN	0.482	0.442	0.455	0.497	0.513	0.503	0.463	0.479
	VAT	0.489	**0.620**	0.536	0.398	0.620	0.660	0.603	0.561
	HAN	0.493	0.613	0.490	0.608	0.661	0.578	0.664	0.587
	Multi-task	**0.581**	0.461	**0.621**	**0.671**	**0.673**	0.612	0.633	0.607
	Transformer-LS	0.508	0.549	0.583	0.628	0.557	0.594	0.662	0.583
	Γ-TLS (Ours)	0.553	0.558	0.567	0.661	0.639	**0.625**	**0.717**	**0.617**

Table 4.2 Experimental results. The best result is in bold, and the 2nd best is underlined.

4.3.2 Results

The results are listed in Table 4.2. Our model, Γ-TLS, demonstrated superior performance in several aspects compared to the baseline models. When evaluated using the accuracy metric, Γ-TLS outperformed the baseline models on four aspects: *Clarity*, *Soundness Correctness*, *Substance*, and *Overall Recommendation*. Additionally, Γ-TLS outperformed the baseline models when evaluated using the Macro F1 score metric on two aspects: *Substance* and *Overall Recommendation*. Overall, Γ-TLS performed the best out of all the models across an average of seven aspects on both metrics.

Additionally, we observe that the Transformer-LS outperforms the CNN by almost 5% in accuracy and 10% in Macro F1 score, which shows that the attention mechanism is relatively more effective for modeling the documents. By applying a hierarchical structure, SSL, or multi-task learning technique, the performance is also further improved.

4.3.3 Ablation study

In comparison to the Transformer-LS model, the incorporating of a denoising network (ladder network) into Transformer-LS resulted in improved performance in almost every aspect, except for *Impact* on the accuracy and *Impact* and *Meaningful Comparison* on Macro F1 score. On average, Γ-TLS outperformed Transformer-LS by 2.3% in accuracy and 3.4% in terms of Macro F1 score metric. This indicates that our assumption, leveraging contextual features from unlabeled data, helps to improve performance.

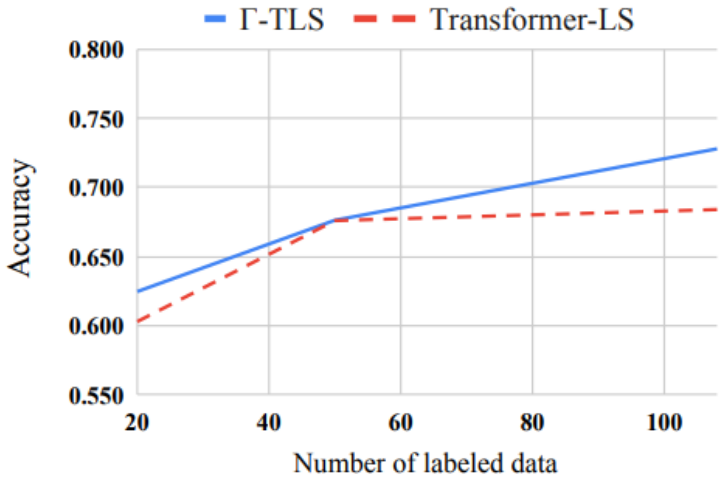


Figure 4.2 Accuracy against the number of labeled data on Overall recommendation score prediction. The number of unlabeled data is fixed to 999 for Γ-TLS.

We also investigated how the number of labeled data used for training affects the overall performance. As shown in Figure 4.2, increasing the number of labeled data tends to improve the performance of both Γ-TLS and Transformer-LS, except for a labeled data count of 50, where the results were not significantly different. Overall, Γ-TLS consistently outperformed Transformer-LS, which shows that our proposed SSL method is stably effective on small training data.

In addition, the effect of the number of unlabeled data on model performance was examined, as shown in Figure 4.3. The results indicate that Γ-TLS's performance improved when the number of unlabeled data was increased from 100 to 400 but there was no sign of further improvement beyond that point. Our model, Γ-TLS, still outperforms the Transformer-LS by using only 100 unlabeled data.

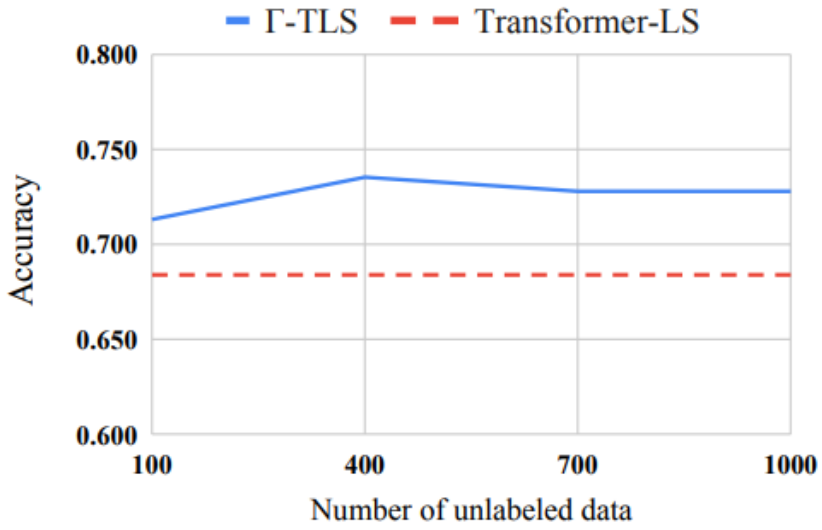


Figure 4.3 Accuracy against the number of unlabeled data on Overall recommendation score prediction

4.4 Summary

In this work, we focused on the task of automated peer review aspect score prediction (PASP) and proposed a novel method called Γ-TLS. The method integrates the Transformer-LS model with the denoising network, the Γ-model of ladder networks. Our experimental results showed that Γ-TLS outperformed the baseline models on average accuracy and F1 score. In future research, we plan to investigate ways to leverage related information between aspects for our model, as well as to generate more knowledgeable and explainable review comments.

Chapter 5
Transfer Learning for Truncated Documents

In Chapters 3 and 4, our primary objective is to introduce a semi-supervised learning approach aimed at leveraging unlabeled data. By incorporating methods that harness this unlabeled data, our goal is to enhance the model's ability to discern the underlying structure of the data distribution, thus enabling better generalization to novel samples. In this chapter, we present an alternative method to tackle the challenge posed by the limited training data available for peer review. This method revolves around the concept of transfer learning, utilizing an intermediate task that shares relevance with the target task. By initially learning informative features from this intermediate task and subsequently fine-tuning the model on the target task, this approach integrates advantages from both pre-existing knowledge in the pretrained model and specific insights from the intermediate task. Our experiments substantiate the effectiveness of this intermediate-task transfer learning technique, showcasing notable enhancements in the performance of pretrained models concerning peer review score prediction.

5.1 Introduction

In recent years, there has been a surge volume of submissions to AI-related international conferences and journals. This upsurge has consequently intensified the difficulties of the review process. To alleviate the burgeoning reviewers' workload, employing an approach to reject papers with evidently low quality serves as a practical strategy. On the other hand, constructive critique extended to authors about the shortcomings in their submissions can encourage refinement and enhancement of their work. In response to this challenge, the development of automatic Peer Review Score Prediction systems has emerged. These systems score a numerical evaluation of academic papers, assessing a spectrum of aspects like *"clarity"* and *"originality"*.

A pioneering contribution to the field comes in the form of the PeerRead dataset. This publicly accessible corpus of scientific peer reviews, introduced by Kang et al. (2018), serves as a valuable resource for researchers with diverse objectives. These objectives are ranging from classification of paper acceptance (Ghosal et al., 2019; Deng et al., 2020; Maillette de Buy Wenniger et al., 2020; Fytas et al., 2021), prediction of review aspect scores (Li et al., 2020a; Wang et al., 2020; Muangkammuen et al., 2022), to citation recommendation (Jeong et al., 2019), and predicting citation counts (van Dongen et al., 2020). In this work, we focus on review aspect score prediction.

Unsupervised pretraining SciBERT (Beltagy et al., 2019) was utilized on various downstream scientific NLP tasks, including biomedical domain (Li et al., 2016; Nye et al., 2018), computer science domain (Luan et al., 2018; Jurgens et al., 2018), and multiple domains (Cohan et al., 2019). One promising approach for further enhancing pretrained models that have been shown to be broadly helpful is to first fine-tune a pretrained model on an intermediate task, before fine-tuning again on the target task, also referred to as *Supplementary Training on Intermediate Labeled-data Tasks* (STILTs) (Phang et al., 2019; Pruksachatkun et al., 2020). STILTs explore the potential of incorporating a secondary phase of pretraining using data-rich intermediate supervised tasks, with the aim of improving the effectiveness of the resulting target task model.

In this work, we perform comprehensive experiments using the Aspect-enhanced Peer Review (ASAP-Review) dataset (Yuan et al., 2022) that we extract review aspect sentiments for our intermediate task training. The ASAP-Review dataset is a collection of peer-reviews with fine-grained annotations of review aspect information. For example, *"The paper is well-written and easy to follow"* shows a positive sentiment of clarity aspect and a high score of clarity aspect. These aspect sentiments can be beneficial for the review aspect score prediction. We extract the review aspect sentiment from the review texts of a paper and use it as a target label for that given paper. We ran our experiments on 6 intermediate tasks and 7 target tasks, resulting in a total of 42 intermediate-target task pairs.

In summary, our main contributions are:

- This work is the first to introduce an intermediate-task transfer learning method to peer-review score prediction.
- We propose a method to extract aspect sentiments for intermediate-task training for peer-review score prediction.
- We conduct experiments to demonstrate the efficacy of each intermediate task, resulting in performance gains across every review aspect score prediction.

5.2 Method

We present a simple intermediate-task transfer learning for peer review score prediction. Figure 5.1 illustrates the method pipeline that consists of the following steps: aspect sentiment extraction, intermediate-task training, and fine-tuning on the target task.

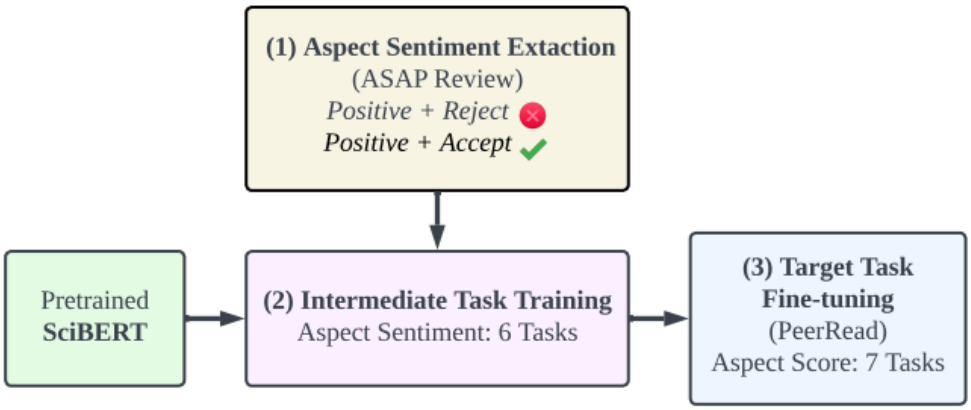


Figure 5.1 Overview of our pipeline framework. It comprises aspect sentiment extraction, intermediate-task training, and fine-tuning on the target task.

5.2.1 Aspect Sentiment Extraction

To further train the pretrained model SciBERT on the intermediate tasks, we extract aspect sentiments from the ASAP-Review dataset (Yuan et al., 2022) to utilize them for our intermediate-task training. The ASAP-Review dataset comprises peer-review data from ICLR and NeurIPS. We use only ICLR data as it contains both accepted and rejected papers which are the same as the target task dataset, PeerRead.

Originally, this dataset contained review texts with sequence labels of fine-grained annotation of aspect information. An example of the review annotations is shown in Table 5.1. We utilize 6 aspects in the dataset, which are Clarity (CLA-i), Meaningful Comparison (COM-i), Motivation/Impact (MOT-i), Originality (ORI-i), Soundness/Correctness (SOU-i), and Substance (SUB-i). The annotation guideline for annotating aspects in reviews of the ASAP-Review dataset is provided in Appendix B.

Summary	Soundness +	Motivation +	Clarity +
The authors prove a generalization guarantee for deep neural networks with ReLU activations, in terms of margins of the classifications and norms of the weight matrices. They compare this bound with a similar recent bound proved by Bartlett, et al. While strictly speaking, the bounds are incomparable in strength, the authors of the submission make a convincing case that their new bound makes stronger guarantees under some interesting conditions. The analysis is elegant. It uses some existing tools but brings them to bear in an important new context, with substantive new ideas needed. The mathematical writing is excellent. Very nice paper. I guess that networks including convolutional layers are covered by their analysis. It feels to me that these tend to be sparse, but that their analysis still my provides some additional leverage for such layers. Some explicit discussion of convolutional layers may be helpful.			

Table 5.1 An example of review annotations of ASAP-Review dataset. “+" denotes positive sentiment. Negative sentiment does not occur in this example.

Each aspect is also marked with a sentiment, *positive* or *negative*. We count the number of positives and negatives of each aspect in the reviews. We use the majority polarity as a label for the reviewed paper since one paper consists of multiple reviews. We further remove the samples having a positive aspect label with a reject decision and having a negative aspect label with an accept decision to amplify the characteristic in the data. The statistics of the ASAP-Review dataset after aspect sentiment extraction are shown in Table 5.2. To distinguish it from the target tasks, i.e., review aspect score predictions, we add "-*i*" to each intermediate task.

Aspects	Negative	Positive	Total
CLA-*i*	1,560	1,003	2,563
COM-*i*	1,738	180	1,918
MOT-*i*	525	1,453	1,978
ORI-*i*	1,257	1,186	2,443
SOU-*i*	1,789	933	2,722
SUB-*i*	1,726	505	2,231

Table 5.2 Statistics of the aspect sentiments of ASAP-Review dataset for the intermediate-task training.

5.2.2 Intermediate Task Training

We fine-tune SciBERT model on each intermediate task, following the standard procedure of fine-tuning a pretrained model on a target task as described in Devlin et al. (2019). Instead of multi-task training (Liu et al., 2019a), we use single intermediate-task training to examine the effect of each intermediate task independently. The objective of these intermediate tasks is to predict the sentiment for each review aspect. We train the model to minimize the *Binary Cross-Entropy* loss.

5.2.3 Target Task Fine-tuning

After intermediate-task training, we fine-tune our models on each target task individually. Our target task is peer-review score prediction, which consists of 7 aspects shown in Table 5.3. The PeerRead dataset contains peer-review datasets from several conferences. Among them, we chose the ACL 2017 dataset for our experiment as it includes aspect scores that are fully annotated. In this dataset, an input paper has multiple review scores, we use the rounded average score of each aspect as the target score ranging from 1 to 5. We fine-tune the models to minimize the *Categorical Cross-Entropy* loss of five classes.

Aspects	Total
Clarity (CLA)	136
Meaningful Comparison (COM)	132
Impact (IMP)	132
Originality (ORI)	136
Soundness/Correctness (SOU)	136
Substance (SUB)	136
Overall Recommendation (REC)	136

Table 5.3 Statistics of the PeerRead ACL 2017 dataset for the target tasks.

5.3 Experiments

5.3.1 Experimental settings

We used the pretrained model *scibert-scivocab-uncased* in all experiments. For each intermediate and target task, we used a peak learning rate at 5e−5 and a dropout rate of 0.1. We used a batch size of 8 and a maximum sequence length of 512. We trained our models using the AdamW (Loshchilov and Hutter, 2019) with linear decay and 0.2 warm-up ratio. We performed our experiments on NVIDIA GeForce RTX 3090 GPUs.

A pipeline with one intermediate task works as follows: First, we split the extracted ASAP-Review data into training and validation sets with a 9:1 ratio. We fine-tuned SCIBERT on the intermediate task for 10 epochs and saved a checkpoint at the end of each epoch, resulting in 10 checkpoints. The performance of each intermediate task evaluated on the validation set is shown in Figure 5.2. The performances were quite stable during fine-tuning, except for SUB-i. We then fine-tuned copies of the resulting models separately on each of the 7 target tasks. We chose the result of the checkpoint that performs best on the target task. Because the test set of the PeerRead dataset is very small, i.e., only 7 samples, most of the results reported by Wang et al. (2020) can be obtained by just using the majority score as a prediction, and it could lead to inappropriate evaluation. Instead of using the original sets to perform the experiments, we ran the same pipeline on 5-fold cross-validation three times. This gave us 15 observations for each result in our experiments.

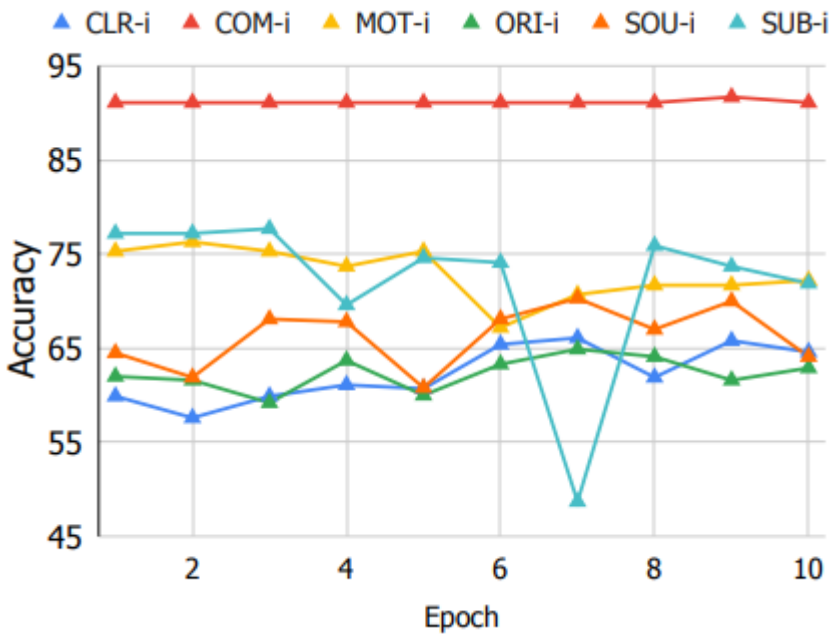


Figure 5.2 Performances on intermediate tasks in accuracy at each checkpoint.

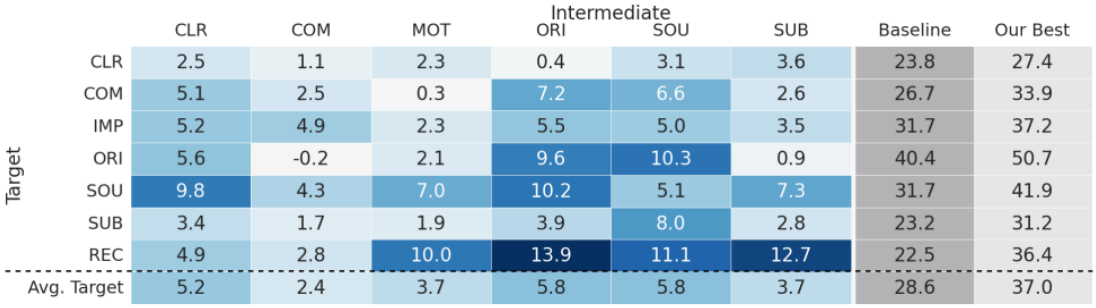
We compared our method to the PeerRead (Kang et al., 2018). We re-implemented their model based on CNN and kept the same hyperparameters. GloVe 840B embeddings (Pennington et al., 2014) were utilized as input word representations, without tuning. The outputs from the CNN model are fed into a max pooling layer and the final linear layer. We evaluated their model in our experimental settings.

5.3.2 Results and Discussion

Figure 5.3 shows the differences in target task performances between the baselines and models trained with intermediate-task training, each averaged across three 5-fold cross-validations. A positive result indicates a successful transfer.

		Intermediate							
		CLR	COM	MOT	ORI	SOU	SUB	Baseline	Our Best
Target	CLR	0.9	0.7	0.4	0.0	0.4	1.6	67.7	69.3
	COM	0.5	0.7	-0.3	3.5	2.7	1.0	58.6	62.1
	IMP	1.5	1.3	0.8	1.0	1.0	1.0	80.5	82.0
	ORI	2.9	1.7	3.7	7.1	5.1	2.0	49.8	56.9
	SOU	5.4	4.5	4.4	9.6	4.2	5.2	50.9	60.5
	SUB	0.5	0.7	0.2	0.5	0.1	0.7	67.9	68.6
	REC	1.4	0.9	3.9	4.1	2.7	4.9	59.1	64.0
	Avg. Target	1.9	1.5	1.9	3.7	2.3	2.3	62.1	66.2

(a) Accuracy



(b) Macro F1

Figure 5.3 Transfer learning results between intermediate and target tasks. Baselines on the second rightmost column are models that are fine-tuned without intermediate-task training. Our best results from the models with intermediate-task training are on the rightmost column. Each cell shows the difference in performance between the baseline and model with intermediate-task training. The cool and warm tone colors indicate improvement, and deterioration, respectively.

We observed that transfer learning, almost every intermediate-task training, helps improve the performance of the target task. The *Soundness/Correctness* score prediction gains more performance from intermediate-task training with around 10% on both accuracy and macro F1. Overall, our best results are better than those of the baselines around 4.1% and 8.4% on average, in accuracy and macro F1, respectively. The best improvements in accuracy are from ORI-*i* on *Soundness/Correctness* at 9.6%. The best improvement in macro F1 score is up to 13.9% from ORI- *i* on *Overall Recommendation*. On average across every target task, the ORI-*i* is the most successful intermediate task that increases 3.7% and 5.8% in accuracy and macro F1, respectively.

Interestingly, we did not find the largest improvement from the same aspect of the intermediate task (sentiment prediction) and the target task (score prediction), except for the *Originality* on the accuracy metric. Instead, the score prediction task gains more performance from other aspects of the intermediate task.

We also compared our method to the PeerRead (Kang et al., 2018) which is shown in Table 5.4. Our method performed better than the PeerRead model on every task and increased 5.3% and 14% on average, in accuracy and macro F1, respectively. It outperformed the PeerRead model by 10.3% on *Soundness/Correctness* in term of accuracy and by 29.2% on *Originality* in term of macro F1.

Aspects	PeerRead	Ours
CLA	67.4 (22.5)	**69.3 (27.4)**
COM	55.0 (20.4)	**62.1 (33.9)**
IMP	80.2 (30.3)	**82.0 (37.2)**
ORI	47.8 (21.5)	**56.9 (50.7)**
SOU	50.2 (21.6)	**60.5 (41.9)**
SUB	67.1 (21.1)	**68.6 (31.2)**
REC	58.8 (23.5)	**64.0 (36.4)**
Avg.	60.9 (23.0)	**66.2 (37.0)**

Table 5.4 Results compared with the method in PeerRead (Kang et al., 2018). Each cell indicates accuracy (macro F1). Bold indicates the best result.

5.3.3 Ablation Study

Our approach to extracting the ASAP-Review dataset for intermediate-task training contains two strategies, i.e., aspect sentiment extraction from review text and removing a sample that has a positive label with a reject decision and vice versa. To examine how each strategy contributes to the performance of the target task, we consider the following variants of our intermediate task:

a) **Decision** - Using decision prediction as an intermediate task. Here, the decision prediction task predicts whether a paper gets *accepted* or *rejected*. The statistics of decision data are shown in Table 5.5.

b) **Aspect** - Using aspect sentiment data without removing a sample. Here, the sample has a positive label with a reject decision and vice versa. The statistics of the data are shown in Table 5.6.

c) **Aspect + Decision** - Our full method using two strategies altogether. By incorporating two strategies, the quantity of data is decreased by over 30% from the **Aspect**.

Accept	Reject	Total
3,295	1,855	5,150

Table 5.5 Statistics of the decision data

Aspects	Negative	Positive	Total
CLA-*i*	2,430	1,626	4,056
COM-*i*	2,889	264	3,153
MOT-*i*	773	2,655	3,428
ORI-*i*	1,837	1,984	3,821
SOU-*i*	2,700	1,357	4,057
SUB-*i*	2,901	760	3,661

Table 5.6 Statistics of the aspect polarity data without removing a sample that has a positive label with a reject decision and vice versa.

Table 5.7 shows the results of different strategies of the intermediate task training. We can see that **Decision** helps improve the pretrained model performance in almost every target task except Substance on macro F1. **Aspect** further improves the pretrained model compared to **Decision** in almost every target task and has a better performance on accuracy and macro F1 on average. This indicates that the aspect sentiment data contains richer information for review aspect score prediction compared to the decision data. In contrast, the decision data shows more relevance on the *Originality* and *Soundness/Correctness* score predictions than aspect sentiment data. One possible reason for this is that they are the main aspect of the reviewer's judgment.

Target Task	Baseline	Intermediate Task		
		Decision	Aspects	Aspects + Decision
CLR	66.7 (23.8)	+0.4 (+1.4)	+0.4 (+1.3)	**+1.6 (+3.6)**
COM	58.6 (26.7)	+1.2 (+4.8)	+2.3 (+6.4)	**+3.5 (+7.2)**
IMP	80.5 (31.7)	+1.3 (+5.8)	**+2.0 (+7.7)**	+1.5 (+5.5)
ORI	49.8 (40.4)	+4.2 (+5.1)	+3.0 (+3.7)	**+7.1 (+10.3)**
SOU	50.9 (31.7)	+5.2 (+6.8)	+4.2 (+6.4)	**+9.6 (+10.2)**
SUB	67.9 (23.2)	+0.2 (-0.2)	**+1.4** (+4.3)	+0.7 **(+8.0)**
REC	59.1 (22.5)	+1.9 (+7.1)	+3.2 (+9.2)	**+4.9 (+13.9)**
Avg.	62.1 (28.6)	+2.1 (+4.4)	+2.4 (+5.6)	**+4.1 (+8.4)**

Table 5.7 Results on the variants of the intermediate task. The baseline column indicates the results without intermediate-task training. The other columns show the difference in performance between the baseline and model with intermediate-task training. Each cell indicates an improvement in accuracy (macro F1 score) compared with the baseline. Bold indicates the best result.

As we can see from Table 5.7, combining aspect polarity data with a decision strategy leads to a better result on almost every target task and the best result on average in both accuracy and macro F1 score. Although the data size of **Aspect + Decision** is smaller than that of **Aspect**, the average result of **Aspect + Decision** is still better. This shows that the characteristic is more important than the quantity of the data for intermediate-task training.

5.3.4 Error Analysis

We plot the confusion matrix between truth and model prediction on test data in Figure 5.4, which shows that the prediction scores of our model tend to be close to the true values. The model tends to be biased to a score of 4, which is the most common score in the dataset. The model was able to classify some papers with a score of 2 or 3 correctly. In contrast, it was unable to correctly classify papers with a score of 1 or 5. However, it still rated papers with a score of 5 higher than a score of 1. The shortage of training samples for scores 1 and 5 (less than 5 samples) complicates its prediction.

Incorporating techniques to handle imbalanced datasets is an interesting direction for future work.

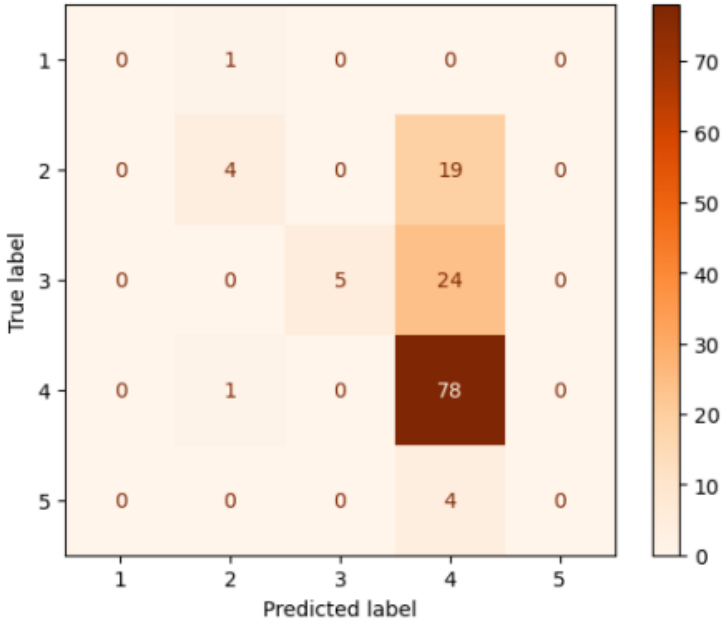


Figure 5.4 Confusion matrix of true and prediction of *Overall Recommendation* scores.

5.4 Summary

In this study, we investigated the impact of intermediate-task transfer learning on peer-review score prediction. Specifically, we fine-tuned a pretrained model SciBERT on an intermediate task before fine-tuning again on the target task. We proposed a method to extract the ASAP-Review dataset for intermediate-task training to improve peer-review score prediction. The experimental results showed the effectiveness of the intermediate-task training as it attained a better result than the baseline on every target task in both accuracy and macro F1. Future work will include (1) extending the method to process longer sequences to cover the full length of the paper, and (2) incorporating multiple tasks for the intermediate-task training to exploit related information between intermediate tasks.

Chapter 6
Transfer Learning for Full Documents

In Chapter 5, our primary focus centers on introducing a transfer learning approach that involves preliminary training on intermediate tasks, followed by fine-tuning the model for specific target tasks. However, a notable limitation arises from the inherent struggle of pretrained transformers to effectively process entire academic papers. Addressing this challenge, Chapter 6 presents an extension of pretrained models, specifically addressing a fundamental drawback - their inability to handle documents longer than a thousand words, such as academic papers. Our devised technique is conceptually straightforward: the document is segmented into sentences, and each sentence is individually processed by a pretrained model to generate a corresponding sentence embedding. These embeddings are then organized into a sequence, serving as the input for the pretrained model. Through experimentation, our results showcase that this method significantly improves the pretrained model's performance in predicting peer-review scores. Furthermore, leveraging intermediate-task training plays a crucial role in augmenting the overall effectiveness of the pretrained model for this purpose.

6.1 Introduction

Peer-review scoring is a process of assigning numerical evaluations to academic papers based on various aspects such as *"clarity"* and *"originality"*. An automatic prediction system for peer-review aspect scores can serve as a helpful tool for both authors and reviewers. The workload of reviewers can be reduced by identifying and rejecting papers with evidently low quality. Conversely, providing feedback on each aspect to the authors can also facilitate enhancements in the quality of their respective papers. A pioneering contribution to the field comes in the form of a dataset. PeerRead is the first publicly available dataset of scientific peer reviews (Kang et al.,2018), providing peer review details, e.g., final decisions, aspect scores, and review contents. Its utilization covers a wide range of applications, which includes paper acceptance classification (Ghosal et al., 2019; Deng et al., 2020; Wenniger et al., 2020; Fytas et al., 2021), review-aspect score prediction (Li et al., 2020a; Wang et al., 2020; Muangkammuen et al., 2022), citation count prediction (Dongen et al., 2020), and citation recommendation (Jeong et al., 2020). This study centers on the task of predicting review-aspect scores. A major concern is the size of the PeerRead dataset, which directly impacts the model's performance.

Over the past years, transfer learning methods have notably improved performance across

various Natural Language Processing (NLP) tasks (Peters et al., 2018; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2018). These models are first pretrained on unsupervised tasks such as language modeling, followed by a process of fine-tuning on specific target tasks. SCIBERT is a Transformer based pretrained language model (Beltagy et al., 2019). Through the utilization of unsupervised pretraining on an extensive corpus of scientific publications, it enhances the performance on downstream scientific NLP tasks (Li et al., 2016; Nye et al., 2018; Luan et al., 2018; Jurgens et al., 2018; Cohan et al., 2019). However, it is vague that the model parameters obtained through unsupervised pretraining are ideally optimized to support transfer learning. To further enhance pretrained models, one effective approach involves an *intermediate-task training*. In this method, the pretrained model is initially fine-tuned on an intermediate task followed by a fine-tuning on the target task (Phang et al., 2018; Wang et al., 2019; Clark et al., 2019; Sap et al., 2019; Pruksachatkun et al., 2020). This technique is also referred to as *Supplementary Training on Intermediate Labeled data Tasks* (STILTs). STILTs explore the potential of incorporating a secondary phase of pretraining using data-rich intermediate supervised tasks, with the aim of improving the effectiveness of the resulting target task model.

Another problem to be concerned with is the limitation of a pretrained model SciBERT. Although Transformer-based models excel at handling relatively short sequences, they have a limitation in processing long sequences. Specifically, they can only accept a limited length of words or tokens as input (Dai et al., 2019). In this study, we introduce a method that simply utilizes a pretrained model with intermediate-task training for long sequences. We segment the document into sentences and utilize SciBERT to acquire a representation of each sentence. Then, we stack these sentence representations into a sequence and use it as input to perform intermediate-task training and fine-tuning on the target tasks. Given that the sentence representations and SciBERT are still in the same embedding space, it means that we can utilize SciBERT on a sentence level instead of a WordPiece level. Using this approach, we are able to process a longer sequence without truncating the input or modifying the pretrained model. We call this technique SciBERT over Sentence Embeddings (SciBERT-SE).

In summary, our contributions include:

- We introduce intermediate-task training for peer-review score prediction.
- We propose a method to utilize SciBERT for long sequences.
- Our experiments show that pretrained models with intermediate-task training help improve performance on peer-review score prediction.

6.2 Method

We introduce an intermediate-task training approach for peer-review score prediction. Figure 6.1 illustrates an overview of our approach.

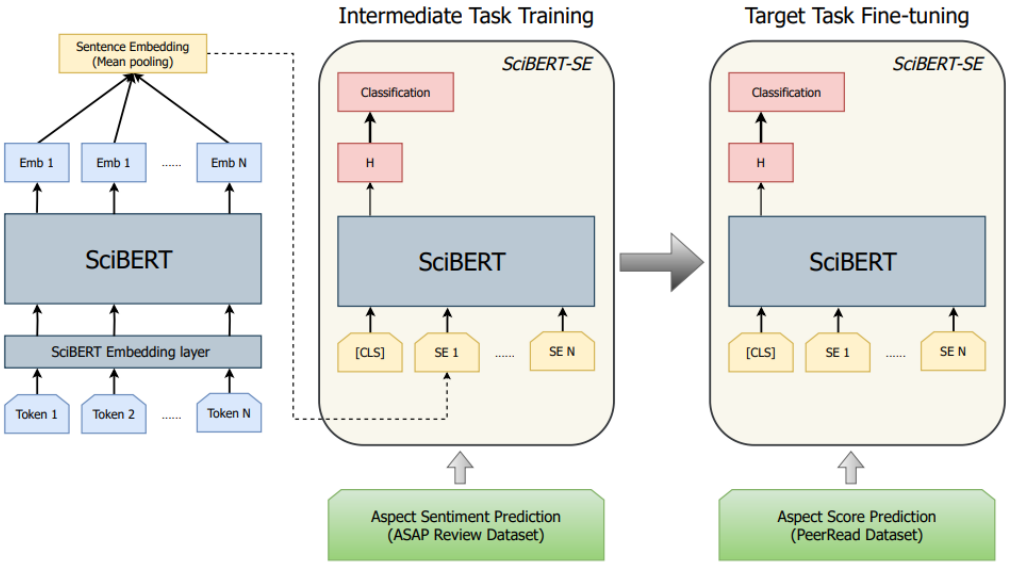


Figure 6.1 Overview of our approach, SciBERT-SE with intermediate-task training on ASAP Review dataset and fine-tuning on target tasks of PeerRead dataset.

6.2.1 SciBERT over Sentence Embeddings (SciBERT-SE)

Our backbone model is built upon SciBERT, a scientific variant of BERT (Devlin et al., 2019). SCIBERT is constructed on the Transformer architecture (Vaswani et al., 2017). The model encompasses two pretraining objectives: masked language modeling and next sentence prediction.

Due to the inherent limitation of SciBERT in handling specific input lengths, we adopt a strategy of segmenting the input sequence into sentences. Each sentence consists of a sequence of WordPiece tokens, as illustrated on the left-hand side of Figure 6.1. We use mean-pooling over token representations obtained from SciBERT to get a sentence representation in the same vector space.

Subsequently, we aggregate these sentence representations into a sequence, which directly serves as input to SciBERT without traversing the Embedding layer. The output H derived from [CLS] token functions as a document representation. To generate the final predictions for both intermediate and target tasks, we employ a fully connected layer with Softmax activation function.

6.2.2 Intermediate-Task Training

To further train the pretrained model SciBERT on intermediate tasks, we select a closely related task, i.e., review-aspect sentiment prediction. For example, the reviewer's positive sentiment toward clarity is reflected in their review, *"Nicely written and understandable",* resulting in a high clarity score. The inclusion of aspect sentiments can be advantageous for predicting review aspect scores.

We exploit the ASAP-Review dataset that contains review texts with sequence labels of review-aspect annotations (Yuan et al., 2022). Table 6.1 demonstrates an example of the review annotations. Each aspect annotation is also marked with a sentiment, *positive* or *negative*. We count the number of positives and negatives of each aspect in the reviews. We use the majority polarity as a label for the reviewed paper, as one paper consists of multiple reviews. We only keep the accepted papers with positive sentiment and the rejected papers with negative sentiment.

For each intermediate task, we fine-tune SciBERT-SE according to the procedure for fine-tuning a pretrained model on a target task, as outlined in (Devlin et al., 2019). The objective of these intermediate tasks is to predict the sentiment for each review aspect. The model is trained with the objective of minimizing the *Binary Cross-Entropy* loss.

Clarity +	Originality +	Substance -
The authors introduce a new memory model which allows memory access in O(log n) time. Pros : * The paper is well written and everything is clear. * It's a new model and I'm not aware of a similar model. * It's clear that memory access time is an issue for longer sequences and it is clear how this model solves this problem. Cons : * The motivation for O (log n) access time is to be able to use the model on very long sequences. While it is clear from the definition that the computation time is low because of its design, it is not clear that the model will really generalize well to very long sequences. * The model was also not tested on any real-world task. I think such experiments should be added to show whether the model really works on long sequences and real-world tasks, otherwise, it is not clear if this is a useful model.		

Table 6.1 An example of review annotations of ASAP-Review dataset.

6.2.3 Target Task Fine-Tuning

Following the intermediate-task training phase, our subsequent step involves fine-tuning the models individually for each target task. Our target task is the prediction of peer review scores. The PeerRead dataset comprises various peer review datasets from multiple conferences.

We selected the ACL 2017 dataset for our experiment because it is fully annotated with aspect scores. Within this dataset, individual papers are associated with multiple review scores. The target score utilized in our approach is the rounded average score for each aspect, falling within the range of 1 to 5. The model is optimized to minimize the *Categorical Cross-Entropy* loss, considering the five classes in the task.

6.3 Experiments

We used review-aspect sentiment prediction for our intermediate-task training and review-aspect score prediction as a target task. For the review-aspect sentiment, we utilize 6 aspects in the ASAP-Review dataset. The statistics of the ASAP-Review dataset after review-aspect sentiment extraction are shown in Table 6.2. The PeerRead dataset contains review scores of 7 aspects, as shown in Table 6.3. We add "-*sentiment*" to each intermediate task to distinguish it from the target tasks.

Aspects	**Negative**	**Positive**	**Total**
CLA-*sentiment*	1,773	796	2,569
COM-*sentiment*	2,025	136	2,161
MOT-*sentiment*	608	1,156	1,764
ORI-*sentiment*	1,141	937	2,078
SOU-*sentiment*	2,045	761	2,806
SUB-*sentiment*	2,018	407	2,425

Table 6.2 Statistics of the aspect sentiments of ASAP-Review dataset for the intermediate-task training.

Aspects	**Total**
Clarity (CLA)	136
Meaningful Comparison (COM)	132
Impact (IMP)	132
Originality (ORI)	136
Soundness/Correctness (SOU)	136
Substance (SUB)	136
Overall Recommendation (REC)	136

Table 6.3 Statistics of the PeerRead ACL 2017 dataset for the target task fine-tuning.

In all experiments, we utilized the *scibert-scivocab-uncased* pretrained model. The content of a paper serves as the input to the model. For both intermediate and target tasks, we used a peak learning rate at 5e-5, batch size of 8, and a dropout rate of 0.1. We used a maximum number of sentences of

300 (8,000 tokens). We trained our models using the AdamW (Loshchilov and Hutter, 2019) with linear decay and a 0.2 warm-up ratio.

We experimented the transfer learning on 6 intermediate tasks and 7 target tasks, resulting in a total of 42 pairs of intermediate-target tasks. A single intermediate task pipeline operates as follows: First, we divided the intermediate task data into training and validation sets by a 9:1 ratio, finetuned SciBERT-SE for 10 epochs, and saved checkpoints every epoch. Then we fine-tuned the checkpoints on each of the 7 target tasks separately. We chose the results of intermediate-target task pairs that perform best on the target tasks. Because the test set of the PeerRead dataset is very small, we ran the experiment on 5-fold cross-validation three times. This gave us 15 observations for each result in our experiments.

The following methods are used as baselines to compare with our approach:

- **Majority Baseline** - We used the majority score in a training set as a prediction score for every test sample.
- **PeerRead (CNN)** - We reimplemented a CNN model using the same hyperparameters in PeerRead (Kang et al., 2018). The CNN outputs are passed through a max pooling layer, followed by the final linear layer.
- **SciBERT** (Beltagy et al., 2019) - A pretrained language model for scientific text. We fine-tuned SCIBERT on the peer-review score prediction tasks. The maximum token length is 512.

Our approach to improving peer-review score prediction contains two main techniques, SciBERT-SE and intermediate-task training. To examine how each technique contributes to the model's performance, we consider the following variants of our approach:

- **SciBERT-SE** - The SciBERT over Sentence Embeddings that extends a pretrained model SciBERT for longer sequence input. We directly fine-tuned SciBERT-SE on the peer-review score prediction tasks without intermediate-task training.
- **SciBERT + Intermediate** - We trained SciBERT on the intermediate tasks before fine-tuning it on peer-review score prediction tasks. The maximum length is also the same as SciBERT.
- **SciBERT-SE + Intermediate** - Our full method that uses both techniques, SciBERT-SE and intermediate-task training. We used the same maximum input length as SciBERT-SE.

6.3.1 Results and analysis

Table 6.4 shows our results in accuracy and macro F1 metrics. We can observe that our approach performs better than the majority baseline and the comparison methods in every aspect. Our full method outperforms PeerRead (CNN) (Kang et al., 2018) with a large margin of 67% by macro F1 and 10% by accuracy on average. By only using SciBERT-SE, our model is still able to outperform SciBERT in almost every aspect except for *Meaningful Comparison*. This implies that our method of extending the SciBERT for a longer sequence input is helpful to the pretrained model.

Intermediate-task training helps SciBERT to perform better with an improvement of 29% by macro F1 and 7% by accuracy on average. For SciBERT-SE, it has an improvement of 20% in macro F1 and 4% in accuracy from intermediate-task training. A further performance gain with intermediate-task training indicates that the pretrained model can acquire relevant information from the intermediate tasks. In other words, the review-aspect sentiment information is beneficial for review-aspect score prediction.

	Approach	CLR	COM	IMP	ORI	SOU	SUB	REC	Avg.
Baseline	Majority Baseline	66.91	54.51	80.24	41.98	39.15	66.87	58.08	58.25
Comparing Systems	PeerRead (CNN) [1]	67.40	55.02	80.24	47.80	50.24	67.12	58.82	60.95
	SciBERT [15]	67.65	58.59	80.49	49.80	50.92	67.85	59.07	62.05
Proposed Method	SciBERT-*SE*	69.61	56.03	80.99	56.17	59.05	68.35	62.48	64.67
	SciBERT + Intermediate	69.34	**62.11**	82.02	56.90	60.50	68.59	63.96	66.20
	SciBERT-*SE* + Intermediate	**71.05**	60.85	82.28	**57.58**	**61.50**	70.32	**66.92**	**67.21**

(a) Accuracy

	Approach	CLR	COM	IMP	ORI	SOU	SUB	REC	Avg.
Baseline	Majority Baseline	-	-	-	-	-	-	-	-
Comparing Systems	PeerRead (CNN) [1]	22.54	20.43	30.28	21.54	21.57	21.11	23.45	22.99
	SciBERT [15]	23.80	26.71	31.67	40.35	31.70	23.21	22.46	28.56
Proposed Method	SciBERT-*SE*	28.52	23.05	34.66	42.61	38.42	24.38	31.41	31.86
	SciBERT + Intermediate	27.39	**33.90**	37.21	**50.66**	41.91	31.15	36.37	36.94
	SciBERT-*SE* + Intermediate	**32.33**	33.70	**38.32**	49.13	40.08	**33.15**	**41.16**	**38.28**

(b) Macro F1

Table 6.4 Results on Aspect Score Prediction Task. The best and the 2nd best results are marked with bold, and underlined, respectively

The difference in target task performance between the SciBERT-SE with and without intermediate-task training is shown in Figure 6.2. A positive result indicates that intermediate-task training can successfully transfer beneficial information. We observed that most of the intermediate tasks helped improve the performance on the target tasks. The most significant improvement, with a margin of 32%, is achieved by pairing the *Soundness/Correctness Sentiment* and *Meaningful Comparison* Score.

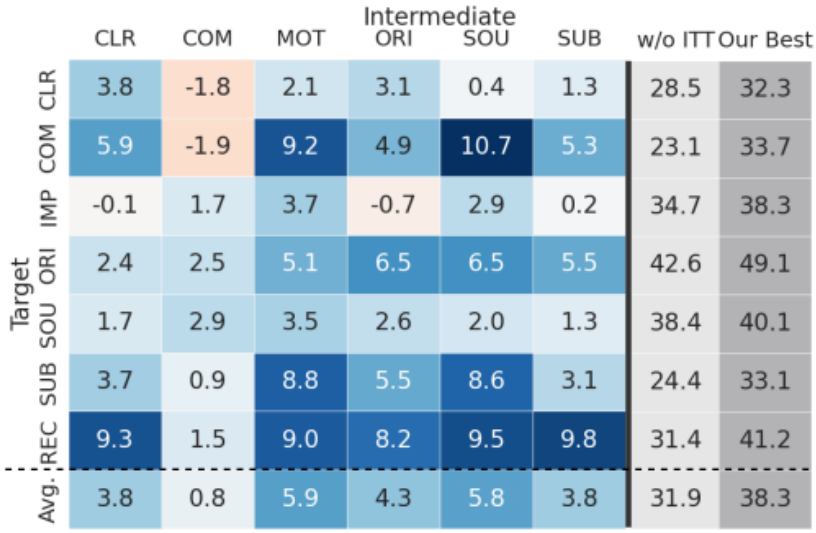
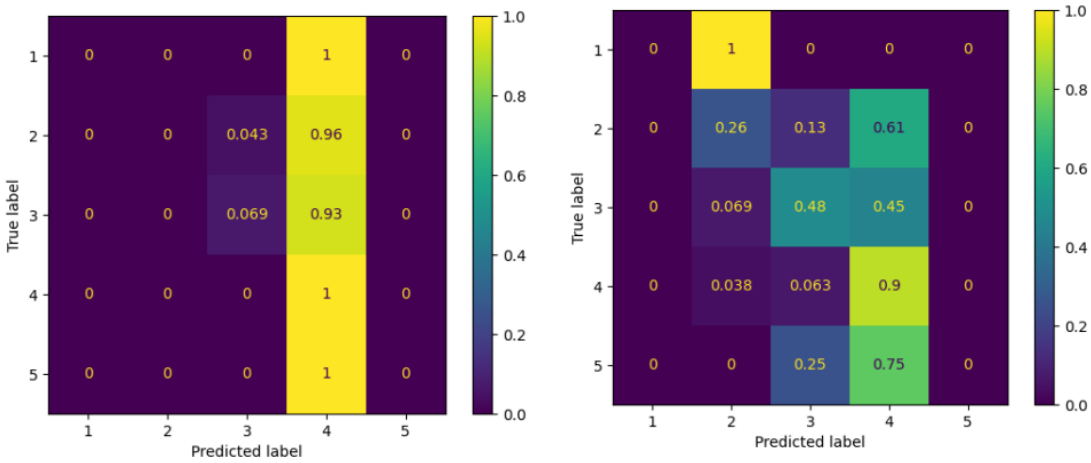


Figure 6.2 Macro F1 on the target tasks on each intermediate task. w/o ITT refers to the SciBERT-SE without intermediate-task training. Our best results with intermediate-task training are on the rightmost column. Each cell on the left-hand side illustrates the performance difference between the models with and without intermediate-task training. The cool and warm tone colors indicate improvement and deterioration.

The *Meaningful Comparison Sentiment* as an intermediate task is not very beneficial for the target task fine-tuning. One reason is that the data of *Meaningful Comparison Sentiment* is the most unbalanced among the intermediate tasks, which might affect the pretrained model during intermediate-task training. We observed that using the same aspect of the intermediate and the target tasks results in the largest improvement of macro F1 on the target task in three aspects, i.e., *Clarity*, *Motivation/Impact*, and *Originality*. This indicates that a review-aspect score prediction obtains the most benefit from the same aspect of review-aspect sentiment.

We also compare the performance of SciBERT and our method on *Overall Recommendation* score prediction using a confusion matrix, as shown in Figure 6.3. The SciBERT model predominantly predicts a score of four, with limited ability to recall a score of three. This is because score four is the most common score in the dataset, appearing over 50% of the time. Our method yields a higher recall on scores two and three, but a lower recall on score four. Due to a limited dataset of less than 5 samples, neither method accurately predicts scores one and five.



(a) SCIBERT (b) Our method

Figure 6.3 Confusion matrix of Overall Recommendation score prediction normalized over the true labels.

6.4 Summary

In this work, we showed that the pretrained model with intermediate-task training helps to predict the peer-review scores. Our approach leverages two techniques: SCIBERT over sentence embeddings and intermediate-task training. The SCIBERT over sentence embeddings helps extend a pretrained model so that the model can process a longer sequence, resulting in a better outcome. The intermediate-task training further improves the performance of the pretrained model, SCIBERT, on peer-review score prediction. The experimental results showed the effectiveness of each intermediate task, i.e., aspect sentiment prediction, on each target task, i.e., aspect score prediction. With further exploration, we aim to incorporate multiple tasks for intermediate-task training to exploit related information between intermediate tasks. Moreover, we aim to extend our approach to leverage the rich content within reviews to enhance our predictive capabilities for peer-review scoring.

Chapter 7
Conclusion and Future Work

Automated peer review score prediction represents an evolving area within the domain of artificial intelligence and natural language processing, aiming to streamline and enhance the academic paper review process. With the exponential growth in academic submissions to conferences and journals, there's a pressing need to expedite and optimize the peer review procedure. This surge in submissions has intensified the challenges faced by reviewers and editors in assessing numerous papers efficiently and accurately.

The concept of automated peer review score prediction involves developing computational models capable of assessing academic papers based on various quality aspects, such as clarity, originality, relevance, and more. These models aim to replicate, automate, or augment the human review process by providing numeric evaluations for different aspects of a paper's quality.

Pioneering contributions in this field include the creation of datasets such as PeerRead, which serves as a repository of scientific peer reviews, encompassing details like review scores, content, and final decisions. Researchers have leveraged such datasets to train models for a myriad of applications, including paper acceptance classification, review aspect score prediction, citation recommendation, and citation count prediction.

Previous works have predominantly relied on supervised learning techniques to build models for peer review score prediction. However, the limitation of annotated datasets has been a persistent challenge, affecting the overall performance of models. To address this limitation and improve the effectiveness of peer review score prediction, this research started exploring semi-supervised learning methods, transfer learning, and other advanced techniques that leverage larger, unlabeled datasets to compensate for the scarcity of labeled data.

This dissertation delves into the complexities of implementing deep learning techniques for predicting peer review scores, particularly when confronted with a scarcity of labeled data. In conventional deep learning methodologies, the process often revolves around fine-tuning a large language model (LLM) tailored to a specific task. However, due to limited resources for fine-tuning LLMs, this study aims to introduce novel transductive learning approaches to enhance peer review

score prediction. Transductive learning strategies focus on improving model performance by leveraging either the inherent structure in unlabeled data or insights gained from related tasks. This underscores the adaptability and effectiveness of utilizing diverse information types in machine learning methodologies. Additionally, this research addresses concerns regarding limitations in pretrained models, notably Transformer-based models, which excel in handling shorter sequences but face challenges with longer sequences.

In this dissertation, four distinct methodologies are introduced to enhance peer review score prediction. The initial two approaches revolve around semi-supervised learning technique called Ladder Networks, with one specifically tailored for truncated documents and the other catering to full-length documents. In the second approach, Ladder Networks have been expanded to handle long sequences by incorporating them with a Long-short transformer. This integration aims to address the challenges associated with processing extended sequences in the context of peer review score prediction. The third approach focuses on transfer learning methods tailored for truncated documents, while the fourth approach is dedicated to transfer learning approaches customized for full-length documents using sentence hierarchy technique. Each of these approaches aims to address specific challenges associated with peer review score prediction in varying document lengths and learning paradigms.

The experimental results demonstrate the efficacy and effectiveness of all four approaches proposed in enhancing the prediction of peer review scores. Each method exhibits promising outcomes in addressing the respective challenges associated with predicting scores for peer reviews, showcasing their potential for improving model performance in diverse scenarios and document lengths.

While the techniques proposed in this dissertation provide significant advancements toward an automated peer review score prediction system, there exist notable areas that demand further improvement. These encompass not only enhancing the scalability of the models to accommodate larger datasets but also refining their adaptability to handle exceedingly diverse or specialized domains within academic literature. Given the intricate and nuanced nature of various research disciplines, the models may face limitations in effectively capturing and analyzing the intricacies of these specific domains.

Moreover, it's imperative to address potential biases embedded within the predictive models. Biases, whether due to dataset composition or model design, could inadvertently affect the impartiality and accuracy of predictions, requiring meticulous attention and mitigation strategies.

Additionally, a crucial aspect for advancement involves devising strategies to enhance the interpretability and explainability of the model predictions. As automated systems become more sophisticated, understanding the reasoning behind predictions becomes increasingly crucial, especially in academic contexts where transparency is highly valued.

Considering the diverse landscape of academic research, future directions should focus on tailoring models to handle specialized domains more effectively. This involves incorporating domain-specific knowledge and refining the models' adaptability to nuances inherent in various fields of study. Furthermore, ensuring the ethical and unbiased application of these models across diverse research domains is paramount.

Continual refinement and development are vital to overcoming these persisting challenges. This ongoing process will not only bolster the reliability and applicability of automated peer review score prediction systems but also pave the way for their seamless integration and acceptance within academic evaluation frameworks.

Publications and Awards

A. Reviewed Publications

Muangkammuen, Panitan et al. (Dec. 2022). 'Exploiting Labeled and Unlabeled Data via Transformer Fine-tuning for Peer-Review Score Prediction'. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Koz, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2233–2240.

Muangkammuen, Panitan et al. (Nov. 2023). 'Intermediate-Task Transfer Learning for Peer Review Score Prediction'. In: *Proceedings of the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing: Student Research Workshop*. Bali, Indonesia: Association for Computational Linguistics.

Muangkammuen, Panitan et al. (Jan. 2024). 'Improving Peer-Review Score Prediction via Pretrained Model with Intermediate-Task Training'. In: *Proceedings of 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*.

B. Unreviewed Publication

Muangkammuen, Panitan et al. (Mar. 2023). 'Improving Peer-Review Score Prediction with Semi-Supervised Learning and Denoising Networks'. In: *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing: NLP 2023*, pp. 1734–1739.

C. Award

NLP 2023 Young Encouragement Award (言語処理学会第 29 回年次大会 「若手奨励賞」)

Acknowledgment

I extend my heartfelt gratitude to the individuals who have offered tremendous support and guidance throughout the course of this thesis. Firstly, I would like to express my deepest appreciation to my supervisor, Professor Fukumoto Fumiyo, whose expertise and insightful feedback were invaluable in shaping the research questions and methodology. Your mentorship has been instrumental in elevating the quality of my work. I am equally indebted to Professor Suzuki Yoshimi and Professor Li Jiyi for their invaluable guidance and unwavering support during my study in this doctoral program. Without their encouragement, this academic journey would have been insurmountable.

A special note of appreciation goes to Professor Suzuki Yoshimi for his technical support in managing the laboratory server, including software installation and server maintenance. My sincere thanks to my colleague, Xu Sheng, for our engaging group discussions and the fruitful exchange of ideas. I would like to thank Mr. Naofumi Takayama for providing a great share house with a cheap rental fee during my study.

I am deeply thankful to the individuals I had the privilege to meet at conferences; your insights and interactions have been invaluable. I am indebted to the anonymous reviewers from EMNLP, AACL, and IMCOM for their insightful comments and suggestions that greatly contributed to improving the quality of this dissertation.

Furthermore, I am profoundly grateful to my family and friends for their unwavering support and encouragement during this intense academic period.

Finally, I acknowledge the financial support received from the Support Project for PhD Candidates in the VUCA (Volatility, Uncertainty, Complexity, and Ambiguity) Age and express my gratitude for the funding provided by JKA, JSPS Grant Number 21K12026, JKA and JST SPRING, Grant Number JPMJSP2133, Kajima Foundation’s Support Program, and JSPS KAKENHI (No. 22K12146).

References

Belkin, Mikhail and Partha Niyogi (2001). 'Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering'. In: *Advances in Neural Information Processing Systems*. Ed. by T. Dietterich, S. Becker and Z. Ghahramani. Vol. 14. MIT Press.

Beltagy, Iz, Kyle Lo and Arman Cohan (2019). 'SciBERT: A Pretrained Language Model for Scientific Text'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620.

Beltagy, Iz, Matthew E. Peters and Arman Cohan (2020). *Longformer: The Long-Document Transformer*. arXiv: 2004.05150 [cs.CL].

Blum, Avrim and Tom Mitchell (1998). 'Combining Labeled and Unlabeled Data with Co-Training'. In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory. COLT' 98*. Madison, Wisconsin, USA: Association for Computing Machinery, pp. 92–100.

Charlin, Laurent and Richard S. Zemel (2013). 'The Toronto Paper Matching System: An automated paper-reviewer assignment system'. In: *Proceedings of the 30th International Conference on Machine Learning (ICML) Workshop on Peer Reviewing and Publishing Models (PEER)*.

Chen, Luoxin et al. (July 2020a). 'SeqVAT: Virtual Adversarial Training for Semi-Supervised Sequence Labeling'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 8801–8811.

Chen, Peibin et al. (June 2020b). 'Data-Efficient Semi-Supervised Learning by Reliable Edge Mining'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Tianyu et al. (2021). 'Pseudo-Label Guided Unsupervised Domain Adaptation of Contextual Embeddings'. In: *Proceedings of the Second Workshop on Domain Adaptation for NLP*. Ed. by Eyal Ben-David et al. Kyiv, Ukraine: Association for Computational Linguistics, pp. 9–15.

Cho, Kyunghyun et al. (Oct. 2014). 'Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734.

Clark, Christopher et al. (June 2019). 'BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2924–2936.

Cohan, Arman et al. (June 2019). 'Structural Scaffolds for Citation Intent Classification in Scientific Publications'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3586–3596.

Dai, Zihang et al. (July 2019). 'Transformer-XL: Attentive Language Models beyond a Fixed-Length Context'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum and Lluís Màrquez. Florence, Italy: Association

for Computational Linguistics, pp. 2978–2988.

Deng, Zhongfen et al. (Dec. 2020). 'Hierarchical Bi-Directional Self-Attention Networks for Paper Review Rating Recommendation'. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6302–6314.

Devlin, Jacob et al. (June 2019). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Dongen, Thomas van, Gideon Maillette de Buy Wenniger and Lambert Schomaker (Nov. 2020). 'SChuBERT: Scholarly Document Chunks with BERT-encoding boost Citation Count Prediction.' In: *Proceedings of the First Workshop on Scholarly Document Processing*. Ed. by Muthu Kumar Chandrasekaran et al. Online: Association for Computational Linguistics, pp. 148–157.

Fytas, Panagiotis, Georgios Rizos and Lucia Specia (Nov. 2021). 'What Makes a Scientific Paper be Accepted for Publication?' In: *Proceedings of the First Workshop on Causal Inference and NLP*. Ed. by Amir Feder et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 44–60.

Ghosal, Deepanway et al. (Oct. 2018). 'Contextual Inter-modal Attention for Multi-modal Sentiment Analysis'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 3454–3466.

Ghosal, Tirthankar et al. (July 2019). 'DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 1120–1130.

Goodfellow, Ian J. et al. (2014). *Generative Adversarial Networks*. arXiv: 1406 . 2661 [stat.ML].

Howard, Jeremy and Sebastian Ruder (July 2018). 'Universal Language Model Fine-tuning for Text Classification'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339.

Iscen, Ahmet et al. (June 2019). 'Label Propagation for Deep Semi-Supervised Learning'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jeong, Chanwoo et al. (2019). 'A context-aware citation recommendation model with BERT and graph convolutional networks'. In: *Scientometrics 124*, pp. 1907–1922.

Jurgens, David et al. (2018). 'Measuring the Evolution of a Scientific Field through Citation Frames'. In: *Transactions of the Association for Computational Linguistics 6*. Ed. by Lillian Lee et al., pp. 391–406.

Kang, Dongyeop et al. (June 2018). 'A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1647–1661.

Kingma, Diederik P and Max Welling (2022). *Auto-Encoding Variational Bayes*. arXiv: 1312.6114 [stat.ML].

Lee, Dong-Hyun (July 2013). 'Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks'. In: *ICML 2013 Workshop : Challenges in Representation Learning*

(WREPL).

Li, Changchun, Ximing Li and Jihong Ouyang (Aug. 2021). 'Semi-Supervised Text Classification with Balanced Deep Representation Distributions'. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, pp. 5044–5053.

Li, Jiao et al. (May 2016). 'BioCreative V CDR task corpus: a resource for chemical disease relation extraction'. In: *Database 2016*, baw068. ISSN: 1758-0463.

Li, Jiyi et al. (Nov. 2020a). 'Multi-task Peer-Review Score Prediction'. In: *Proceedings of the First Workshop on Scholarly Document Processing*. Ed. by Muthu Kumar Chandrasekaran et al. Online: Association for Computational Linguistics, pp. 121–126.

Li, Suichan et al. (June 2020b). 'Density-Aware Graph for Deep Semi-Supervised Visual Recognition'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lim, KyungTae et al. (Apr. 2020). 'Semi-Supervised Learning on Meta Structure: Multi-Task Tagging and Parsing in Low-Resource Scenarios'. In: *Proceedings of the AAAI Conference on Artificial Intelligence 34.05*, pp. 8344–8351.

Liu, Xiaodong et al. (July 2019a). 'Multi-Task Deep Neural Networks for Natural Language Understanding'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Ed. by Anna Korhonen, David Traum and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 4487–4496.

Liu, Yinhan et al. (2019b). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL].

Loshchilov, Ilya and Frank Hutter (2019). *Decoupled Weight Decay Regularization*. arXiv: 1711.05101 [cs.LG].

Luan, Yi et al. (Oct. 2018). 'Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 3219–3232.

Maillette de Buy Wenniger, Gideon et al. (Nov. 2020). 'Structure-Tags Improve Text Classification for Scholarly Document Quality Prediction'. In: *Proceedings of the First Workshop on Scholarly Document Processing*. Ed. by Muthu Kumar Chandrasekaran et al. Online: Association for Computational Linguistics, pp. 158–167.

Meng, Yu et al. (Nov. 2020). 'Text Classification Using Label Names Only: A Language Model Self-Training Approach'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 9006–9017.

Mrowinski, Maciej J. et al. (Sept. 2017). 'Artificial intelligence in peer review: How can evolutionary computation support journal editors?' In: *PLOS ONE* 12.9, pp. 1–11.

Miyato, Takeru, Andrew M. Dai and Ian Goodfellow (2017). 'Adversarial Training Methods for Semi-Supervised Text Classification'. In: *International Conference on Learning Representations*.

Muangkammuen, Panitan et al. (Dec. 2022). 'Exploiting Labeled and Unlabeled Data via Transformer Fine-tuning for Peer-Review Score Prediction'. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Koz, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2233–2240.

— (Mar. 2023a). 'Improving Peer-Review Score Prediction with Semi-Supervised Learning and Denoising Networks'. In: *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing: NLP 2023*, pp. 1734–1739.

— (Nov. 2023b). ‘Intermediate-Task Transfer Learning for Peer Review Score Prediction’. In: *Proceedings of the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing: Student Research Workshop*. Bali, Indonesia: Association for Computational Linguistics.

— (Jan. 2024). ‘Improving Peer-Review Score Prediction via Pretrained Model with Intermediate-Task Training’. In: *Proceedings of 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*.

Nagesh, Ajay and Mihai Surdeanu (June 2018). ‘Keep Your Bearings: Lightly-Supervised Information Extraction with Ladder Networks That Avoids Semantic Drift’. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 352–358.

Nye, Benjamin et al. (July 2018). ‘A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature’. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 197–207.

Oliver, Avital et al. (2018). ‘Realistic Evaluation of Deep Semi-Supervised Learning Algorithms’. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS’18*. Montréal, Canada: Curran Associates Inc., pp. 3239–3250.

Pan, Yuhao et al. (2020). ‘Sentiment analysis using semi-supervised learning with few labeled data’. In: *2020 International Conference on Cyberworlds (CW)*, pp. 231–234.

Pennington, Jeffrey, Richard Socher and Christopher Manning (Oct. 2014). ‘GloVe: Global Vectors for Word Representation’. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.

Peters, Matthew E. et al. (June 2018). ‘Deep Contextualized Word Representations’. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.

Pham, Hieu et al. (June 2021). ‘Meta Pseudo Labels’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11557–11568.

Phang, Jason, Thibault Févry and Samuel R. Bowman (2018). ‘Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks’. In: *ArXiv* abs/1811.01088.

Price, Simon and Peter A. Flach (Feb. 2017). ‘Computational Support for Academic Peer Review: A Perspective from Artificial Intelligence’. In: *Commun. ACM* 60.3, pp. 70–79.

Pruksachatkun, Yada et al. (July 2020). ‘Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?’ In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 5231–5247.

Radford, Alec, Luke Metz and Soumith Chintala (2015). ‘Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks’. In: *CoRR* abs/1511.06434.

Radford, Alec and Karthik Narasimhan (2018). ‘Improving Language Understanding by Generative Pre-Training’.

Rasmus, Antti et al. (2015). ‘Semi-Supervised Learning with Ladder Networks’. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS’15*. Montreal, Canada: MIT Press, pp. 3546–3554.

Rezende, Danilo Jimenez, Shakir Mohamed and Daan Wierstra (22–24 Jun 2014). 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models'. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 1278–1286.

Sap, Maarten et al. (Nov. 2019). 'Social IQa: Commonsense Reasoning about Social Interactions'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 4463–4473.

Valpola, Harri (Nov. 2014). 'From neural PCA to deep unsupervised learning'. In: *From Neural PCA to Deep Unsupervised Learning*.

Vaswani, Ashish et al. (2017). 'Attention is All you Need'. In: *Advances in Neural Information Processing Systems. Ed. by I. Guyon et al. Vol. 30*. Curran Associates, Inc.

Wan, Xiaojun (Aug. 2009). 'Co-Training for Cross-Lingual Sentiment Classification'. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Ed. by Keh-Yih Su et al. Suntec, Singapore: Association for Computational Linguistics, pp. 235–243.

Wang, Alex et al. (July 2019). 'Can You Tell Me How to Get Past Sesame Street? Sentence Level Pretraining Beyond Language Modeling'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 4465–4476.

Wang, Ke and Xiaojun Wan (2018). 'Sentiment Analysis of Peer Review Texts for Scholarly Papers'. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, pp. 175–184.

Wang, Kexin et al. (July 2022). 'GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 2345–2360.

Wang, Qingyun et al. (Dec. 2020). 'ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis'. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Ed. by Brian Davis et al. Dublin, Ireland: Association for Computational Linguistics, pp. 384–397.

Yasunaga, Michihiro, Jungo Kasai and Dragomir Radev (June 2018). 'Robust Multilingual Part-of-Speech Tagging via Adversarial Training'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 976–986.

Yasunaga, Michihiro et al. (July 2019). 'ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks'. In: *Proceedings of the AAAI Conference on Artificial Intelligence 33*, pp. 7386–7393.

Zhang, Haoning et al. (Nov. 2022). 'CSS: Combining Self-training and Self-supervised Learning for Few-shot Dialogue State Tracking'. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by Yulan He et al. Online only: Association for Computational Linguistics, pp. 302–310.

Zhang, Xiao and Dan Goldwasser (July 2020). 'Semi-supervised Parsing with a Variationa Autoencoding Parser'. In: *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*. Ed. by Gosse Bouma et al.

Online: Association for Computational Linguistics, pp. 40–47.

Zheng, Hang et al. (2021). ‘Semi-Supervised Learning for Aspect-Based Sentiment Analysis’. In: *2021 International Conference on Cyberworlds (CW)*, pp. 209–212.

Zhu, Chen et al. (2021). ‘Long-Short Transformer: Efficient Transformers for Language and Vision’. In: *Neural Information Processing Systems*.

Appendix

A. ACL Reviewer Instructions

Below is the list of instructions to ACL 2016 reviewers on how to assign aspect scores to reviewed papers.

APPROPRIATENESS (1-5)

Does the paper fit in ACL 2016? (Please answer this question in light of the desire to broaden the scope of the research areas represented at ACL.)

5: Certainly.

4: Probably.

3: Unsure.

2: Probably not.

1: Certainly not.

CLARITY (1-5)

For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

5 = Very clear.

4 = Understandable by most readers.

3 = Mostly understandable to me with some effort.

2 = Important questions were hard to resolve even with effort.

1 = Much of the paper is confusing.

ORIGINALITY (1-5)

How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

Note that a paper could score high for originality even if the results do not show a convincing benefit.

5 = Surprising: Significant new problem, technique, methodology, or insight -- no prior research has attempted something similar.

4 = Creative: An intriguing problem, technique, or approach that is substantially different from previous research.

3 = Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

2 = Pedestrian: Obvious, or a minor improvement on familiar techniques.

1 = Significant portions have actually been done before or done better.

EMPIRICAL SOUNDNESS / CORRECTNESS (1-5)

First, is the technical approach sound and well-chosen? Second, can one trust the empirical claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?

5 = The approach is very apt, and the claims are convincingly supported.

4 = Generally solid work, although there are some aspects of the approach or evaluation I am not sure about.

3 = Fairly reasonable work. The approach is not bad, and at least the main claims are probably correct, but I am not entirely ready to accept them (based on the material in the paper).

2 = Troublesome. There are some ideas worth salvaging here, but the work should really have been done or evaluated differently.

1 = Fatally flawed

THEORETICAL SOUNDNESS / CORRECTNESS (1-5)

First, is the mathematical approach sound and well-chosen? Second, are the arguments in the paper cogent and well-supported?

5 = The mathematical approach is very apt, and the claims are convincingly supported.

4 = Generally solid work, although there are some aspects of the approach I am not sure about or the argument could be stronger.

3 = Fairly reasonable work. The approach is not bad, and at least the main claims are probably correct, but I am not entirely ready to accept them (based on the material in the paper).

2 = Troublesome. There are some ideas worth salvaging here, but the work should really have been done or argued differently.

1 = Fatally flawed.

MEANINGFUL COMPARISON (1-5)

Do the authors make clear where the problems and methods sit with respect to existing literature? Are the references adequate? For empirical papers, are the experimental results meaningfully compared with the best prior approaches?

5 = Precise and complete comparison with related work. Good job given the space constraints.

4 = Mostly solid bibliography and comparison, but there are some references missing.

3 = Bibliography and comparison are somewhat helpful, but it could be hard for a reader to determine exactly how this work relates to previous work.

2 = Only partial awareness and understanding of related work, or a flawed empirical comparison.

1 = Little awareness of related work, or lacks necessary empirical comparison.

SUBSTANCE (1-5)

Does this paper have enough substance, or would it benefit from more ideas or results?

Note that this question mainly concerns the amount of work; its quality is evaluated in other categories.

5 = Contains more ideas or results than most publications in this conference; goes the extra mile.

4 = Represents an appropriate amount of work for a publication in this conference. (most submissions)

3 = Leaves open one or two natural questions that should have been pursued within the paper.

2 = Work in progress. There are enough good ideas, but perhaps not enough in terms of outcome.

1 = Seems thin. Not enough ideas here for a full-length paper.

IMPACT OF ACCOMPANYING SOFTWARE (1-5)

If software was submitted or released along with the paper, what is the expected impact of the software package? Will this software be valuable to others? Does it fill an unmet need? Is it at least sufficient to replicate or better understand the research in the paper?

5 = Enabling: The newly released software should affect other people's choice of research or development projects to undertake.

4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.

3 = Potentially useful: Someone might find the new software useful for their work.

2 = Documentary: The new software useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)

1 = No usable software released.

IMPACT OF ACCOMPANYING DATASET (1-5)

If a dataset was submitted or released along with the paper, what is the expected impact of the dataset? Will this dataset be valuable to others in the form in which it is released? Does it fill an unmet need?

5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.

4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

3 = Potentially useful: Someone might find the new datasets useful for their work.

2 = Documentary: The new datasets are useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive

rating)

1 = No usable datasets submitted.

RECOMMENDATION (1-5)

There are many good submissions competing for slots at ACL 2016; how important is it to feature this one? Will people learn a lot by reading this paper or seeing it presented? In deciding on your ultimate recommendation, please think over all your scores above. But remember that no paper is perfect, and remember that we want a conference full of interesting, diverse, and timely work. If a paper has some weaknesses, but you really got a lot out of it, feel free to fight for it. If a paper is solid but you could live without it, let us know that you're ambivalent. Remember also that the authors have a few weeks to address reviewer comments before the camera-ready deadline. Should the paper be accepted or rejected?

5 = This paper changed my thinking on this topic and I'd fight to get it accepted;

4 = I learned a lot from this paper and would like to see it accepted.

3 = Borderline: I'm ambivalent about this one.

2 = Leaning against: I'd rather not see it in the conference.

1 = Poor: I'd fight to have it rejected.

REVIEWER CONFIDENCE (1-5)

5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.

4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable,

that I missed something that should affect my ratings.

3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this

area in general, I did not carefully check the paper's details, e.g., the math, experimental design,

or novelty.

2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't

understand some central points, or can't be sure about the novelty of the work.

1 = Not my area, or paper is very hard to understand. My evaluation is just an educated guess.

B. ASAP-Review Dataset Annotation Guideline

1. Aspect Typology

We define a typology that contains 8 aspects, which are **Summary**, **Motivation/Impact**, **Originality**, **Soundness/Correctness**, **Substance**, **Replicability**, **Meaningful Comparison** and **Clarity**. The detailed explanation for each aspect is shown below.

- **Summary**: What was done in the paper?

Example:

1. The paper proposes a new memory access scheme based on Lie group actions for NTMs.

• **Motivation/Impact**: Does the paper address an important problem? Are other people (practitioners or researchers) likely to use these ideas or build on them?

Example:

1. The issue researched in this work is of significance because understanding the predictive uncertainty of a deep learning model has its both theoretical and practical value.
2. The method seems limited in both practical usefulness and enlightenment to the reader.

• **Originality**: Are there new research topic, technique, methodology, or insight?

Example:

1. Novel addressing scheme as an extension to NTM.
2. The reviewer believes that the idea of the paper is similar to the one in [1].

• **Soundness/Correctness**: Is the proposed approach sound? Are the claims in the paper convincingly supported?

Example:

1. Illustrations using simulated data and real data are also very clear and convincing.
2. The proposed method is sensible and technically sound.
3. The experiments are also quite convincing.
4. The required condition is rather implicit, and it is unclear how this condition can be checked in practice.
5. There is not much theory to support the method.
6. Several model designs are not well justified.
7. There is no enough justification to demonstrate improvements.

• **Substance**: Does the paper contains substantial experiments to demonstrate the effectiveness of proposed methods? Are there detailed result analysis? Does it contain meaningful ablation studies?

Example:

1. This is a thorough exploration of a mostly under-studied problem.
2. The experiment section shows extensive experiment.
3. There are several modules introduced in the paper, but there isn't much analysis of them during the experiments.
4. This experimental study does not seem to conduct sufficient experiments to demonstrate

the advantages.

5. Lack detailed and insightful ablation studies.

6. I would expect the authors to conduct some more analysis of their results besides acc. and distortion levels.

• **Replicability**: Is it easy to reproduce the results and verify the correctness of the results? Is the supporting dataset and/or software provided?

Example:

1. Release of the dataset and code should help with reproducibility.

2. There are some technical ambiguities.

• **Meaningful Comparison**: Are the comparisons to prior work sufficient given the space constraints? Are the comparisons fair?

Example:

1. The authors do a good job of positioning their study with respect to related work on black box adversarial techniques.

2. The comparison with the Caron-Fox approach is very good and useful for the reader.

3. The experimental study can have more comparison on challenging datasets with more classes.

4. Since the attention based aggregation is similar to GAT, a discussion on the difference is important.

5. The paper fails to locates itself in the literature, how it compares itself into other techniques (both analytically and experimentally).

6. The comparison does not seem fair.

• **Clarity**: For a reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

Example:

1. The paper is well-written and easy to follow.

2. The presentation of the results is not very clear.

2. Annotation Tips

We further decompose each aspect (except Summary) into a positive one and a negative one. For example, Motivation will be decomposed to Positive Motivation and Negative Motivation. As so, there are in total 15 aspects for you to choose from when annotating reviews.

Below are some tips.

- Please annotate the shortest while complete span that indicates a specific aspect. Don't include specific details if the aspect has been stated clearly prior to those details.

Example:

> This experimental study does not seem to conduct sufficient experiments to demonstrate the advantages[Negative Substance] (say, in terms of training efficiency the capability in making the network scalable for more challenging dataset) of the proposed objective function over the existing one.

- Please be as fine-grained as possible. If a sentence contains multiple aspects, annotate them separately if they can be disentangled.

Example:

> The results are new[Positive Originality] and important to this field.[Pos. Motiv.]