Extending Min-BERT for Multi-Task Prediction Capabilities

Stanford CS224N {Default} Project

Grace Yang
Department of Statistics
Stanford University
yanggeer@stanford.edu

Xianchen Yang
Department of Statistics
Stanford University
samany@stanford.edu

Abstract

The main goal of our project is to enhance the foundational model supplied by the Stanford CS224N teaching team, aiming to optimize BERT's performance in sentiment analysis, paraphrase detection, and similarity assessment tasks. Our approach involves refining BERT's sentence embeddings by investigating several key improvements: comparing single-task and multitask model architectures, increasing the dimensions of BERT's embeddings, and utilizing embeddings trained through contrastive learning. We thoroughly examined the impact of these enhancements across the three downstream tasks. Our comprehensive report presents both quantitative data and qualitative insights, providing a detailed evaluation of how each modification contributes to the improved performance of our enhanced model.

1 Key Information

We are undertaking the default project option. Our mentor for this project is Yuan Gao. We do not have any external collaborators, nor are we sharing this project with any other course.

In recognition of the collaborative efforts, Grace Yang and Xianchen Yang worked in close partnership across all project facets, including the literature review, code implementation, analytical review, report composition, and poster creation, thereby ensuring a comprehensive and unified contribution to each extension

2 Introduction

In the dynamic field of NLP, transformer architecture, epitomized by BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019), has dramatically enhanced text processing, enabling deep contextual analysis of words in sentences. This advancement is pivotal for tasks like sentiment classification, paraphrase classification, and semantic similarity prediction. Developing a multitask model capable of performing these diverse tasks is challenging due to their unique linguistic demands, yet presents exciting opportunities to uncover synergies between tasks.

The efficacy of this model can be greatly enhanced by transfer learning, especially in scenarios where data availability is uneven across tasks. For instance, we hypothesize that a substantial dataset for paraphrase classification could markedly boost performance in similarity prediction, especially in cases where the latter suffers from data scarcity. In our project, we conducted empirical tests to determine whether multitask learning allows a model to leverage insights from one task to augment performance in another, thereby fostering a more holistic and efficient approach to learning.

Our project aims to extend BERT's capabilities, focusing on optimizing sentence embeddings for outstanding performance across chosen tasks. We explored the nuanced engineering of minBERT and its downstream structures, adopting a multitask approach that integrates shared BERT layers with task-specific heads. This design exploits common linguistic features across tasks while allowing for precise adjustments via task-specific heads. In particular, we investigated the effects of adding a shared layer before task-specific heads for paraphrase detection and similarity prediction, assessing

this setup against alternatives without such shared layers. Additionally, we experimented with varying approaches in task-specific heads, like contrasting the concatenation of two sentence embeddings with the concatenation of two input sentences separated by a [SEP] token to create a single sentence embedding.

To deepen the model's interpretation of text, we also expanded BERT's embedding dimensions, with the aim of capturing a broader range of textual nuances and thereby enrich the model's analytical depth. Furthermore, we harnessed contrastive learning to enhance the model's ability to discern subtle differences between sentences. Our holistic approach to model optimization involved carefully calibrating the learning process across different tasks, ensuring that advancements in one area do not compromise performance in others.

3 Related Work

This section outlines significant contributions in the areas of multitask fine-tuning and contrastive learning of sentence embeddings, providing a foundation for our project's methodologies.

3.1 Multitask Fine-Tuning

BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning Stickland and Murray (2019) introduced Projected Attention Layers (PALs) to enhance BERT's multitask learning efficiency by using task-specific attention projections to reduce computational load. This method highlights the benefits of task-specific adaptations within a shared architecture, influencing our approach to shared and task-specific layers.

MTRec: Multi-Task Learning over BERT for News Recommendation Bi et al. (2022) demonstrated multitask learning's impact on news recommendation systems with BERT, optimizing tasks like user engagement and content categorization simultaneously. MTRec's success underscores multitask learning's applicability across various domains, supporting our project's multitask strategies.

3.2 Contrastive Learning

SimCSE: Simple Contrastive Learning of Sentence Embeddings Gao et al. (2021) effectively enhances sentence embedding learning through contrastive learning, training models to differentiate between sentence pairs. Utilizing unsupervised and supervised BERT base models, they attained 76.3% and 81.6% Spearman's correlation respectively on STS tasks, guiding our application of contrastive learning to deepen text understanding in our target tasks.

4 Approach

We began our project by successfully building upon and completing the implementation of minBERT and the Adam optimizer, as detailed in the first phase of our project documentation. This foundational work, including architectural and implementation specifics, is thoroughly described in the project handout. In the subsequent phase, we shifted our focus to enhancing the model's efficacy on three downstream tasks, employing various extension techniques to achieve this goal.

MinBERT's architecture concludes with a final hidden layer that generates a 768-dimensional embedding for each input token, encapsulating the full contextual information of the sentence thanks to the attention mechanism. Given that each sequence is initiated with a [CLS] token, we derive the sentence's overall embedding from this [CLS] token's embedding, which serves as a comprehensive representation of the input for downstream applications.

4.1 Baseline Model

When developing our baseline model, we began with a simple method: freezing all BERT embeddings and focusing exclusively on training the task-specific heads.

Diverging from this initial method, we then fine-tuned Min-BERT individually for each of the three tasks. For sentiment classification, our design includes a structure with a dropout layer followed by a linear layer. For paraphrase classification and similarity prediction, we adopted a novel approach using cross-attention mechanisms. This method links sentences more directly by omitting the [CLS] token from the second sentence and combining the sentences with a [SEP] token. As a result, a single pass through Min-BERT produces a comprehensive embedding from the [CLS] token, effectively capturing the nuanced relationship between sentences. This technique is preferred over concatenating

two 768-dimensional embeddings, as it has been shown to yield superior results, as detailed in our further analysis. Specifically, for the similarity prediction task, we applied a sigmoid function on the model's final output. This approach ensures that the predictions are confined within a specific range, making the mean squared error (MSE) measurement more relevant and insightful.

4.2 Extension 1: Multitask Training

We noticed the limited number of training samples in the STS dataset. Considering the substantial similarity between the paraphrase detection and similarity prediction tasks, coupled with the extensive size of the QUORA dataset, we hypothesized that the knowledge gained from paraphrase detection could enhance the model's capability in similarity prediction. This insight led us to adopt a multi-task learning strategy, updating minBERT and its associated heads collectively, rather than fine-tuning separate BERT instances for each task.

Given the disparate sizes of the datasets for our three tasks, we propose a sequential training approach for each task within individual epochs. This method, we believe, is preferable to employing a custom dataloader with sampling techniques designed to balance the dataset sizes. Sequential training ensures that all available data are fully utilized and increases time efficiency. The model is updated at the end of each batch of tasks, allowing a more cohesive and comprehensive learning process that leverages intertask relationships to improve performance across the board.

The order in which tasks are presented during training can impact the model's performance across all tasks. Our observations reveal that the paraphrase detection and similarity prediction tasks do not offer substantial transferable benefits to sentiment classification. This is primarily due to the potential for conflicting gradient directions between these tasks, where enhancing performance on either the paraphrase or similarity prediction tasks may inadvertently detract from the model's ability to accurately perform sentiment classification.

4.3 Extension 2: Paraphrase and Similarity Prediction Downstream Architecture

For the paraphrase and similarity tasks, which both involve analyzing pairs of sentences, we innovated beyond processing sentences separately. Instead of passing each sentence through minBERT independently and later concatenating their embeddings, we refined our approach by directly linking the sentences. This was achieved by discarding the [CLS] token of the second sentence and merging the sentences with a [SEP] token. Consequently, a singular pass through minBERT now yields a pooled embedding derived from the [CLS] token, effectively capturing the inter-sentence relationship.

This methodological enhancement allows for a more sophisticated semantic analysis. Unlike the simpler approach, where tokens in the second sentence were isolated from those in the first, our revised strategy ensures that every token can interact across sentence boundaries. This cross-sentence attention mechanism facilitates a deeper semantic comparison, marking a significant advancement over the baseline model in understanding the nuances between paired sentences.

We recognize that paraphrase detection and similarity prediction tasks share a fundamental similarity, with the latter essentially representing a regression-based variant of the former. Given this conceptual overlap, we delved into the potential benefits of integrating a shared layer before their respective task-specific heads. This exploration involved comparing the performance of models with and without these shared layers, with the aim of discerning the impact of this architectural modification. Our hypothesis is that the weights of the shared layers could be effectively transferable between these two closely related tasks, thereby leveraging their intrinsic similarities to enhance overall model performance.

4.4 Extension 3: Contrastive Learning Embeddings

We hypothesized that the embeddings learned from contrastive learning could improve performance in tasks where understanding subtle differences and similarities between sentences are crucial. SimCSE learns to embed sentences such that similar sentences have similar embeddings. It is trained in a supervised fashion by utilizing the same sentence with different dropout masks as positive pairs and sentences in the same batch as implicit negative samples. By shifting from using the base BERT embeddings to "sup-simcse-bert-base-uncased" embeddings published by the SimCSE researchers, this extension aims to determine if these supervised contrastive embeddings offer superior performance on downstream tasks by providing richer sentence context and capturing more nuanced semantic relationships, thanks to their training on a diverse set of contrastive text data.

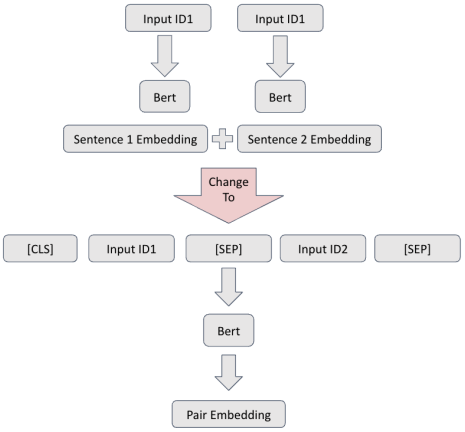

Figure 1: cross attention architecture(left)

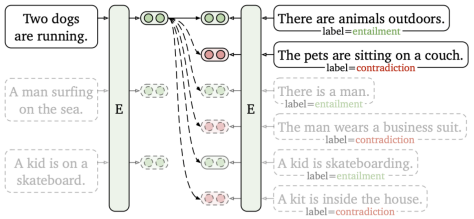

Figure 2: Supervised SimCSE(right)

4.5 Extension 4: Scaling up to 1024 Large BERT

In this extension, we substituted the standard BERT model (which has embedding size of 768) with its larger counterpart, BERT Large (featuring an embedding size of 1024), to explore whether an increase in the number of parameters can offer more nuanced sentence context for various downstream tasks, potentially enhancing overall performance.

5 Experiments

5.1 Data

Four datasets are leveraged to assess and enhance our model's proficiency across a range of NLP tasks:

- **Stanford Sentiment Treebank (SST)**: A cornerstone for sentiment analysis research, this dataset includes 11,855 sentences from movie reviews. Each sentence, along with 215,154 parsed phrases from the Stanford parser, is annotated with labels ranging from negative to positive.
- **CFIMDB**: Geared towards sentiment analysis, CFIMDB features 2,434 movie reviews with binary sentiment labels of negative or positive, providing a focused dataset for binary classification tasks.
- **Quora Dataset**: A vital resource for paraphrase detection, this dataset encompasses 400,000 question pairs, each tagged with binary labels to signify whether they are paraphrases of each other, facilitating the training of models to recognize textual parallels.
- **SemEval STS Benchmark Dataset**: Key to evaluating semantic similarity, it consists of 8,628 sentence pairs, each pair rated on a nuanced scale from 0 (no similarity) to 5 (high similarity), offering a diverse range of examples for nuanced semantic understanding.

5.2 Evaluation Method

For evaluating our model, we adhered to the primary evaluation metrics specified for the standard project. In detail, we reported classification accuracy for both the sentiment classification and paraphrase detection tasks. For the semantic textual similarity task, we provided Pearson's correlation coefficient. The development leaderboard ranking is determined by a weighted average of these three scores, assessed on the development set. We employ these metrics to refine our model, selecting the optimal version for final submission to the test leaderboard.

5.3 Experimental Details

In our research, we standardized certain experimental configurations across all trials. Each baseline model and modified versions were subjected to a training regimen spanning 10 epochs, with the

optimal checkpoint being archived for subsequent analysis. For the initial phase of training, where BERT's parameters were frozen and only the heads were trained, we employed a learning rate of $\eta = 1e-3$. The fine-tuning model used a reduced learning rate of $\eta = 1e-5$. Uniformly, dropout layers were set with a dropout probability of 0.1 to mitigate overfitting. Considering CUDA memory limitations, we opted for a batch size of 64 for the SST and STS datasets and a smaller batch size of 16 for the Quora dataset. The Adam optimizer was configured with a weight decay parameter of 1e-4 to regularize the model and prevent coefficient inflation.

6 Results

We evaluated the baseline model and models with different extensions on the development set and the quantitative results are detailed in **??**.

Model Type	SST-5 (dev)	QQP (dev)	STS-B (dev)	Overall
Baseline (single task)				
1. Bert pretraining + Cross	0.385	0.683	0.446	0.505
2. Bert Finetune + Cross	0.525	**0.891**	**0.868**	**0.783**
Multitask				
3. Base Bert	0.510	0.777	0.255	0.514
4. Base Bert + Cross	0.490	**0.890**	0.780	0.720
5. Large Bert + Cross	0.330	0.838	**0.819**	0.662
6. Base Bert + Contrastive	**0.530**	0.762	0.247	0.513
7. Base Bert + Shared Layer + Cross	0.48	0.88	0.785	0.715
8. Base Bert + Contrastive + Shared Layer + Cross	0.510	0.885	0.790	**0.763**

Table 1: Development Accuracy of Baseline Model and Models with Extensions

For the baseline model, we observed that fine-tuning yielded significantly superior results compared to merely freezing BERT and training task-specific head weights alone. This can be ascribed to the fine-tuning process's capacity to adapt the pre-trained model's parameters to the nuances of the specific task at hand, thus enabling more precise feature extraction and a better understanding of the task-specific data.

In our multitask extensions, we observed a counterintuitive outcome: the multi-task models, which were expected to leverage shared representations to enhance performance across tasks, actually underperformed compared to their single-task counterparts. It has been noted that the highest-performing models for each task are those where BERT has been fine-tuned on a single task alone, augmented with cross-attention mechanisms. This trend points to a limited transfer learning benefit within the multi-task framework, suggesting that the inter-task gradient interference might be detrimental. Such interference seems to impede the multi-task model's learning, preventing it from reaching the effectiveness of models fine-tuned on isolated tasks. These findings underscore the complexity of multi-task learning and indicate that, at least for our specific tasks, dedicated models with tailored attention mechanisms outstrip the generalized approach of multi-task learning.

While the multi-task model incorporating the most extensive range of extensions (cross attention + shared layer + contrastive loss) achieved the highest results, the margin of improvement was marginal. This modest enhancement suggests that the breadth of the extensions did not substantially contribute to the model's final performance, indicating that further investigation into the efficacy of these extensions is necessary.

We observed a remarkable improvement in similarity prediction performance when employing a cross-attention architecture, which elevated the score from a mere 0.255 to an impressive 0.780 (3 vs 4). This substantial increase can be attributed to the cross-attention mechanism's ability to directly model interactions between pairs of inputs, as opposed to merely concatenating two sentence embeddings. This direct interaction allows the model to capture nuanced relationships and fine-grained alignments between elements of the input pairs, thereby facilitating a deeper understanding of similarity. The inclusion of a shared layer between paraphrase detection and similarity prediction tasks results in only marginal performance improvements (4 vs 7). While the deployment of a larger BERT model enhances the similarity correlation score by 0.04, it detrimentally affects the performance on the other two tasks (4 vs 5). This observation suggests a trade-off between specialized improvement in one area and overall task performance, highlighting the challenge of balancing task-specific enhancements with general model effectiveness.

In our approach to optimizing performance for test leaderboard submission, we combined the outputs of various models through an ensemble technique. This method involved aggregating the predictions from our top-performing models (single task). The result of this ensemble approach led to a significant improvement in our model's performance, achieving an overall test accuracy of 78.7% 2

Overall accuracy	SST acc	paraphrase acc	STS corr
0.787	0.538	0.891	0.864

Table 2: Test Accuracy

7 Analysis

In addition to evaluating our model's accuracy, we conducted a qualitative analysis to pinpoint specific instances where the model underperformed across the three tasks. This investigation aimed to uncover the underlying weaknesses and identify some of the causes of these errors.

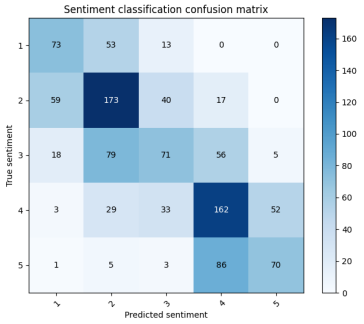

Figure 3: Confusion Matrix of Sentiment Classification (left)

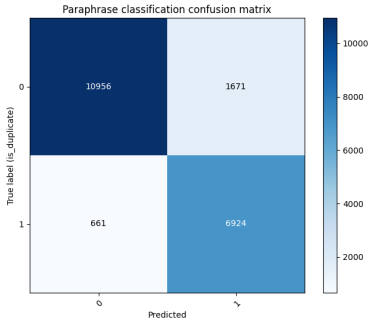

Figure 4: Confusion Matrix of Paraphrase Detection (right)

We generated a confusion matrix for the sentiment classification task 3 to assess the model's accuracy across various label subgroups. Notably, the model predominantly errs in distinguishing between closely related categories, often misclassifying "neutral" sentiments as "somewhat negative" and "negative" sentiments as "somewhat negative." This tendency towards misclassification probably stems from the inherent complexities of a multiclass classification challenge involving five distinct classes, compounded by the subjective biases present in human-labeled data.

As shown in 3, our similarity prediction model tends to make mistakes particularly in instances where one sentence in the pair acts as a short summary of the other, omitting several details. Despite the dataset's true labels often mark these pairs with high similarity, our model frequently predicts a lower similarity score. This discrepancy can be attributed to issues related to the quality of the data. Upon closer examination, we observe that there's a notable inconsistency in how pairs of sentences with omissions are labeled in terms of their similarity - some are rated as having low similarity while

others are considered highly similar. This variability probably stems from the subjective nature of human labeling, leading to inconsistencies in the training labels. Such inconsistencies present significant challenges for the model, making it prone to errors when assessing the similarity of sentences where one is essentially a condensed version of the other. This insight points to the need for a more standardized approach to labeling sentence pairs, especially in cases involving summaries or omissions, to improve the model's accuracy in these scenarios.

Sentence1	**Sentence2**	**Predicted**	**Similarity**
A girl is talking to her dad on a cell-phone.	a girl is talking on her phone.	0.29	0.88
There are three options:	There are only three options:	0.46	1.00
It's also a matter of taste.	It's definitely just a matter of preference.	0.32	1.00
I don't see why there should be any problem with this whatsoever.	I don't see why this could be a problem.	0.45	1.00

Table 3: Similarity Prediction Hard Cases

As shown in 4, the model demonstrates a propensity to assign higher similarity scores to sentence pairs that share most of their lexical choices and syntactic structures, despite the presence of a few divergent words that significantly alter the meaning of sentences. This indicates a nuanced challenge: the model's current metrics for similarity may not sufficiently account for the semantic impact of key terms, thus overlooking the substantial shifts in meaning that can arise from minor lexical substitutions.

Sentence1	**Sentence2**	**Predicted**	**Similarity**
China stocks open higher Monday	China stocks close lower on Thursday	0.71	0.20
Use of force in defense of person.-A	Use of force by aggressor.	0.76	0.24

Table 4: Similarity Prediction Hard Cases

We observed a similar trend of model relying on superficial lexical choices and structures rather than semantic meaning of key terms to classify paraphrase as shown in 5. The confusion matrix for the paraphrase detection task 4 reveals a higher false positive rate compared to the false negative rate, indicating the model's difficulty in accurately identifying non-paraphrase pairs as such. This suggests a challenge for the model in correctly predicting true negatives, where it more frequently errs on the side of mistakenly identifying unrelated sentences as paraphrases.

Sentence1	**Sentence2**	**Predicted**	**is_duplicate**
Is there any evolutionary advantage of baldness?	Was there any evolutionary advantage for beards?	1	0
Can a physicist be an engineer?	Can a person be a physicist and an engineer?	1	0

Table 5: Paraphrase Detection Hard Cases

8 Conclusion and Future Work

In this study, we aimed to enhance the base model provided by the Stanford CS224N teaching team, with the goal of improving BERT's effectiveness in sentiment analysis, paraphrase detection, and similarity assessment. Our approach involved refining BERT's sentence embeddings through innovative modifications to the multitask model architecture and adjustments in the embedding dimensions. We utilized embeddings trained via contrastive learning and leveraged the power of multitask learning to boost the model's adaptability across these varied tasks.

Our analysis yielded several key insights:

- Fine-tuning BERT significantly outperforms strategies that involve freezing BERT while only training the task-specific head weights.
- Contrary to expectations, the single task baseline models surpassed the multitask model, suggesting that gradients from different tasks may be conflicting, to the detriment of overall performance.
- The enhancements brought about by the inclusion of contrastive loss and shared layers were found to be minimal, indicating these strategies may not substantially contribute to the model's performance in our setup.
- Incorporating a cross-attention architecture led to a considerable improvement in similarity prediction, underscoring the value of direct interaction between pairs of inputs over simple concatenation of sentence embeddings.

These findings, underscored by both quantitative metrics and qualitative analyses, indicate that while certain strategies significantly enhance performance in specific tasks, others offer limited benefits. This nuanced understanding directs us towards more targeted and efficient methods for improving sentence embeddings, emphasizing the need for fine-tuning and task-specific architectural decisions to achieve more adaptable and robust models.

Looking ahead, we plan to incorporate cutting-edge methodologies from S-SimCSE, specifically focusing on increasing the dynamic variability of dropout and implementing adaptive sentence-wise masking Zhang and lan (2021). Such enhancements are expected to further boost the model's ability to generate nuanced sentence embeddings that capture the essence of the text more accurately. Additionally, we aim to introduce gradient surgery techniques into our multitask learning framework Yu et al. (2020). This step is designed to harmonize the gradient directions emanating from different tasks, mitigating potential conflicts, and fostering a more synergistic learning environment. We also aim to use SMART regularization technique to prevent overfitting caused by aggresive fine-tuning Jiang et al. (2020). Through these efforts, we anticipate not only elevating the model's current capabilities but also setting a new benchmark for multitask learning in natural language processing.

References

Qiwei Bi, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Hanfang Yang. 2022. MTRec: Multi-task learning over BERT for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2663–2669, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190. ArXiv:1911.03437 [cs, math].

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. ArXiv:1902.02671 [cs, stat].

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.

Junlei Zhang and Zhenzhong lan. 2021. S-SimCSE: Sampled Sub-networks for Contrastive Learning of Sentence Embedding. ArXiv:2111.11750 [cs].