We used shared late days in this assignment. Zhengji Yang used 5 late days and Jialiang Chen used 1. On average we used 2 late days.

MultitaskBERT with Contrastive Pretraining and Fine-Grained Feature Learning

Stanford CS224N Default Project

Rui Deng
Stanford University
ruideng@stanford.edu

Jialiang Chen
Stanford University
jiajiahh@stanford.edu

Zhengji Yang
Stanford University
yangzj@stanford.edu

Abstract

Multitask learning has gained considerable attention as an effective approach for improving data efficiency and reducing training overhead when addressing multiple natural language processing (NLP) tasks. Nonetheless, it presents unique optimization challenges, such as maintaining embedding space quality and reconciling contradictory gradient descent directions, which substantially influence overall performance. To tackle these challenges, we introduce an innovative Multitask BERT model that integrates pre-training and fine-tuning phases by combining contrastive learning methodologies, multitask learning strategies, fine-grained embedding representations, and regularized multitask optimization techniques. Our model's performance is assessed on three diverse downstream NLP tasks, namely Semantic Textual Similarity (STS), Sentiment Analysis, and Paraphrase Detection. Empirical results from our experiments indicate that incorporating contrastive learning significantly enhances the consistency of the embedding space, thereby augmenting BERT's expressive capabilities. Additionally, we propose regularized multitask optimization methods tailored for various downstream tasks, designed to amalgamate shared information while mitigating information conflicts. This fusion not only expedites the training process but also bolsters prediction accuracy and optimization stability.By integrating these advanced techniques, our novel multitask BERT model demonstrates considerable performance gains in comparison to established baseline models, underscoring its efficacy for a range of NLP applications.

1 Introduction

Bidirectional Encoder Representations from Transformers (BERT) has achieved remarkable success in single-task Natural Language Processing (NLP) models. However, data efficiency and training cost remain significant challenges. In response, multitask learning has emerged as a promising approach, sharing structures between multiple tasks for more efficient training and learning from related tasks. Nevertheless, multitask BERT models present several optimization challenges, including the quality of the embedding space (from a singular spectrum perspective) and the contradiction of information from different tasks. These challenges can substantially impact the overall performance and training efficiency. Additionally, the processing of BERT embeddings, i.e., the model structure of the head, greatly influences single-task performance, necessitating a deep understanding of the intrinsic mechanisms.

In this paper, we propose a novel Multitask BERT pre-training and fine-tuning model that combines contrastive learning methods, multitask learning strategies, a dense similarity scheme with fine-grained embedding representation, and regularized multitask optimization. We evaluate our novel multitask BERT model on three downstream tasks: Semantic Textual Similarity (STS), Sentiment

Analysis, and Paraphrase Detection, comparing the results with the original BERT model. Our numerical experiments reveal that contrastive learning methods significantly enhance BERT embeddings' expressiveness by increasing the effective dimension of the embedding space, while fine-grained embeddings enable more detailed input sentence representation. These techniques partially address the non-uniformity of the embedding space.

Furthermore, the multitask learning techniques employed during the fine-tuning stage enable BERT to integrate common information for more efficient training and maintain the conformity of the embedding space or diversity by simultaneously training on different tasks. To avoid over-fitting and improve performance, we introduce a regularized multitask optimization scheme using finely-designed loss functions. For STS and Paraphrase Detection, we investigate the performance of different BERT embeddings, demonstrating that combining pooling average of embedded tokens and [CLS] token outperforms only using the [CLS] token in both training efficiency and task-specific information richness.

2 Related Work

BERT has achieved remarkable success in processing NLP tasks [1], yet data efficiency and training cost continue to present significant challenges [2]. Concurrently, multitask learning techniques have emerged as promising methods in deep learning and have been quickly adopted for NLP tasks [3]. For instance, Yifan et al. proposed a multitask learning framework for biomedical text mining based on the BERT model [4], while Eleftherios et al. constructed a multitask BERT model for schema-guided dialogue state tracking tasks, significantly reducing computational cost [5]. However, multitask BERT models introduce new challenges, including the anisotropy of BERT embeddings and the contradiction of gradients from various tasks, which can significantly weaken their expressiveness [2]. Fortunately, contrastive learning offers a potential solution for addressing some of these issues, having already been successfully applied to BERT models for the STS task [6] and proven effective in improving the uniformity and alignment of the embedding space [7]. Specifically, for tasks such as STS and paraphrase detection, Munikar et al. demonstrated that the structure of heads and the choice of BERT embeddings can significantly impact prediction accuracy [8]. Our work focuses on addressing the aforementioned challenges to enhance the capacity of our multitask BERT model.

3 Approach

In our model architecture, we utilize BERT [1] to generate contextualized representations of input words and subsequently pass these BERT embeddings to different heads corresponding to the given downstream tasks. During both the pre-training and fine-tuning stages, we employ contrastive learning methods as described in [6] for supplementary pre-training and adopt multitask learning approaches based on [9], wherein the model is trained on all three tasks in each epoch. To achieve higher prediction accuracy, we investigate the underlying mechanisms of these individual tasks and devise task-specific schemes. For paraphrase detection and STS tasks, we propose and implement a dense similarity scheme using fine-grained embeddings, which involves calculating the cosine similarity between the pooled average of sentence embeddings rather than relying solely on [CLS] tokens. Moreover, we introduce regularized multitask optimization methods by designing appropriate loss functions for these tasks. Our multitask BERT model integrates all of these strategies to enhance performance across the three given tasks.

3.1 Multitask BERT model

Our Multitask BERT model is based on the BERT model[1], which consists of tokenizer, embedding layer and 12 Encoder Transformer layers. As the core of BERT, the encoder transformer layer is composed of the multi-head attention followed by normalization, additive layers and feed-forward layers, in which the multi-head attention is a weighted average of single attention heads, i.e.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) = Softmax(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}})VW_i^V.$$

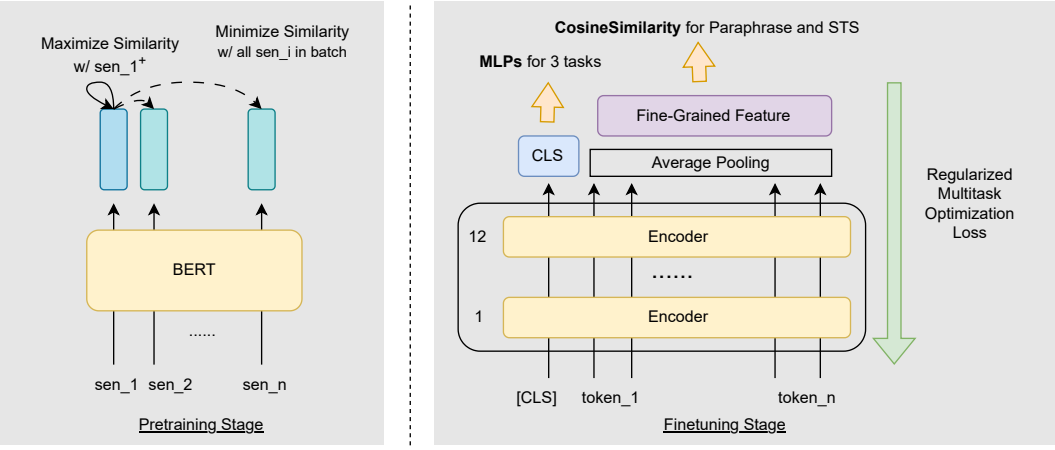


Figure 1: The training pipeline of our proposed method. In pretraining stage, we use contrastive learning to learn the intrinsic representation from datasets. In finetuning stage, we further modify the BERT parameters w.r.t. multitask losses with regularized optimization.

where Q, K, V are the query, key, value matrices, W_i^Q, W_i^K, W_i^V, W^O are corresponding parameter matrices.

Then we pass the BERT embeddings including [CLS] token and fine-grained sentence embedding to different heads for the downstream tasks. In sentiment analysis head, the embedding passes through a 2-layer neural network with ReLU activator to get a 5×1 logit; in STS and Paraphrase detection, after passing the BERT embedding pairs into 2-layer neural networks, we compute their cosine similarity to get the output logits.

3.2 Pretraining with Contrastive learning

In pre-training stage, we follow the approach in[6] and use the unsupervised SimCSE model for the BERT model. We take a minibatch of sentences $\{x_i\}_{i=1}^m$ and use $x_i^+ = x_i$ as positive pairs. We use independently sampled dropout masks for x_i and x_i^+ by feeding the input to the encoder twice with different dropout masks z and z', and get two embeddings h_i and h_i^+. Then the training objective is

$$\mathcal{L} = \sum_{i=1}^{m} -log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^{m} e^{sim(h_i, h_j^+)/\tau}},$$

where τ is a temperature hyperparameter, $h_i = \psi_\theta(x_i, z)$, $h_j^+ = \psi_\theta(x_j^+, z')$, ψ_θ is the BERT model and $sim(h_i, h_j^+)$ is the cosine similarity, i.e. $\frac{h_i^T h_j^+}{||h_i|| \cdot ||h_j^+||}$.

To analyze the mechanism of contrastive learning, we introduce two key description of the wellness of embedding space, l_{align} and $l_{uniform}$ from [7], where

$$l_{align} = \mathbb{E}_{(x,x^+)} ||f(x) - f(x^+)||^2,$$
$$l_{uniform} = log\, \mathbb{E}_{iid,x,y}\, exp\{-2||f(x) - f(y)||^2\}.$$

Here, l_{align} measures the average distance between positive instances, $||f(x) - f(x^+)||^2$, hence is small if the BERT can aggregate similar sentence embeddings. Just opposite to l_{align}, $l_{uniform}$ describes the dispersion, or uniformity of random instances, $exp\{-2||f(x) - f(y)||^2\}$ when x is chosen independently from y, hence is small when the average distance of random embedding pairs is relatively large. We expect contrastive learning techniques can improve the quality of BERT embedding space by enhancing the uniformity for different instances while not weaken the ability to distinguish similar instances.

3.3 Fine-tuning with Fine-grained Feature

In paraphrase detection and STS tasks, we propose the dense similarity scheme with fine-grained embedding representation. Instead of just using u_{cls}, the BERT embedding of [CLS] token, we also utilize the fine-grained sentence embedding $\{v_i\}_{i\in[n_s]}$ from last layer of BERT feature, where n_s is the total number of tokens in the input sentence. This fine-grained embedding representation greatly enlarges the representation space of input sentence, enables the following networks to extract more details, especially those not covered in [CLS] token, hence enormously improves the performance of similarity detection. In detail, we pass the fine-grained sentence embedding sen_1 and sen_2 into neural networks to get a refined task-specific representation, then we take the cosine similarity of the pooling-average of fine-grained representation. We combine this dense similarity with the general [CLS] similarity to capture both details from dense representation and integrated, sentence-level similarities, i.e.

$$logit_{tot} = TaskHead(\mathrm{fea}_1, \mathrm{fea}_2) \; + \; CosSim(mean(\mathrm{sen}_1), mean(\mathrm{sen}_2))$$
$$\mathrm{fea}_\mathrm{i} = \mathcal{N}_{cls}(u^i_{cls}); \;\; \mathrm{sen}_\mathrm{i} = \mathcal{N}_{grained}(V_i) \quad \forall\, i \in \{1,2\}; \;\; TaskHead = \mathcal{N}_{task}$$

where $\mathcal{N}_{cls}$, $\mathcal{N}_{grained}$, and $\mathcal{N}_{task}$ are task-specific MLPs, u_{cls} is the BERT embedding of [CLS] token, V is the fine-grained sentence embedding.

3.4 Regularized Optimization

In order to address the issue of over-fitting in the fine-tuning process, we applied the Smoothness-Inducing Adversarial method proposed by [10]that controls model complexity and reduces the over-fitting issue. The method involves solving an optimization problem with a smoothness-inducing adversarial regularize [10]. Given the model $f(\cdot;\theta)$, we want to solve the following optimization for fine-tuning:

$$\min_\theta \mathcal{F}(\theta) = \mathcal{L}(\theta) + \lambda_\mathrm{s}\mathcal{R}_\mathrm{s}(\theta)$$

where $\mathcal{L}(\theta)$ is the loss function defined as

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f\left(x_i;\theta\right), y_i\right)$$

$\ell(\cdot,\cdot)$ is the loss function depending on the target task. We use multi-class classification loss for sentiment analysis, binary classification loss for paraphrase detection, and MSE loss for semantic textual similarity. Then, $\lambda_\mathrm{s} > 0$ is a tuning parameter, and $\mathcal{R}_\mathrm{s}(\theta)$ is the smoothness-inducing adversarial regularizer. We define $\mathcal{R}_\mathrm{s}(\theta)$ as

$$\mathcal{R}_\mathrm{s}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\widetilde{x}_i - x_i\|_p \le \epsilon} \ell_\mathrm{s}\left(f\left(\widetilde{x_i};\theta\right), f\left(x_i;\theta\right)\right),$$

where $\epsilon > 0$ is a tuning parameter and We used $p = 2$ in our implementation. For classification tasks, $f(\cdot;\theta)$ outputs a probability simplex and ℓ_s is chosen as the symmetrized KL-divergence, i.e.,

$$\ell_\mathrm{s}(P,Q) = \mathcal{D}_\mathrm{KL}(P\|Q) + \mathcal{D}_\mathrm{KL}(Q\|P)$$

For regression tasks, $f(\cdot;\theta)$ outputs a scalar and ℓ_s is chosen as the squared loss, i.e., $\ell_s(p,q) = (p-q)^2$. Essentially, The smoothness-inducing adversarial regularizer enforces local Lipschitz continuity of the function f under the metric ℓ_s. By limiting the output change of f when small perturbations are added to x_i, the regularizer encourages f to be smooth within the neighborhoods of all data point x_i's. We believe adding such regularization in optimization loss improves the incorporation of learning different tasks within one epoch by smoothing out the optimization of model parameters and preventing sudden changes in optimization directions. This smoothness thus helps improve generalization to different tasks and further prevents overfitting to a specific task.

4 Experiments

We perform three tasks with our model on sentiment analysis, paraphrase detection, and semantic textual similarity to evaluate its effectiveness.

Table 1: The table shows the experiment results for methods lister above. Pretrain mode indicates the weight for BERT is fixed; Finetune mode sets the weight to trainable status. Contra is the contrastive pretraining, and Dense is the fine-grained feature head, and Reg is regularized optimization.

	Train Mode	Head Structure	SST	Quora	STS	Score
Baseline	Pretrain	Linear	0.344	0.390	0.111	0.282
Baseline	Finetune	Linear	0.388	0.451	0.237	0.359
Contrastive	Pretrain	Linear	0.490	0.623	0.379	0.497
Contrastive	Pretrain	MLPs	0.498	0.707	0.549	0.585
Contrastive	Finetune	MLPs	0.496	0.728	0.633	0.619
Contra + Dense	Finetune	MLPs	0.492	0.733	0.758	0.661
Contra + Dense + Reg	Finetune	MLPs	0.506	0.746	**0.817**	0.689
Contra + Dense + Reg	Finetune (Test Set)	MLPs	**0.512**	**0.746**	0.809	**0.689**

4.1 Dataset and Evaluation Metric

- We evaluate the performance of sentiment analysis on Stanford Sentiment Treebank (SST) dataset[11] The SST dataset consists of 215,154 phrases with their sentiments annotated by 3 human judges. We train and predict each phrase as negative, somewhat negative, neutral, somewhat positive, or positive.
- We use the Quora dataset[12] to test the performance of paraphrase detection. It has 400,000 labeled question pairs, and we predict whether one question is a paraphrase of another in each pair. We evaluate the correctness using classification accuracy as well as the F1-score.
- We evaluate the performance of semantic textual similarity on SemEval STS Benchmark Dataset[13]. It contains 8,628 labeled sentence pairs with different similarities. We predict a logit on a scale from 0 (unrelated) to 5 (equivalent meaning) for each pair. We use the Pearson correlation between the groundtruth and the predicted similarity to validate performance.

4.2 Experimental details

The baseline for multitask BERT is pretraining the multitask heads (one linear layer for each) with BERT weight fixed. Then, the succeeding models are all finetuning both the BERT parameters and heads with the methods mentioned above. We also advance the multitask heads to a 2-layer MLP that can better adapt to each individual task. We train the model on all tasks together in one epoch, and we train it for 15 epochs with $lr = 1e^{-5}$ and $batch\_size = 64$.

4.3 Results

In Table 1, We can see that the pretrain mode generally gives better results than the finetune mode, since this allows the model to learn its weight w.r.t. task-specific data. In scenarios of linear heads, contrastive pretrain significantly improves the prediction quality, showing its effectiveness in generalizing underlying feature representations to the model parameters.
We can see the combination of contrastive learning, fine-grained feature learning, and regularized optimization gives us the best result. This proves that while the dense head extracts rich, detailed semantic knowledge for downstream tasks, the regularization further balances the converging process across different datasets. These strategies seamlessly incorporate with each other and yield our SOTA result.

4.4 Analysis of Embedding in Contrastive Learning

To gain a deeper understanding of the effects of contrastive learning, we examine the alignment and uniformity of the embedding space, which are defined as follows:

$$l_{align} = \mathbb{E}_{(x,x^+)}||f(x) - f(x^+)||^2,$$
$$l_{uniform} = \log \mathbb{E}_{iid\, x,y} \exp -2||f(x) - f(y)||^2,$$

where $(f(x), f(x^+))$ denotes a sentence embedding pair generated by encoding the sentence x twice with independent dropouts, and x, y represent randomly selected sentences.

We construct the training set S for contrastive learning by integrating all sentences from the three datasets (SST, Quora, and STS) and compute the aforementioned measures over S. The results are presented below:

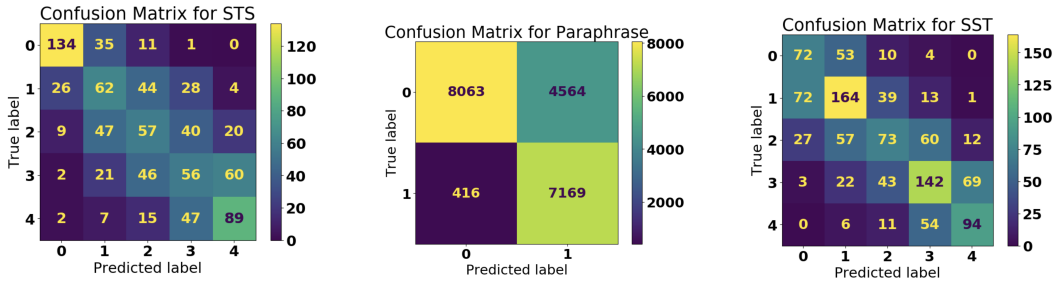
Table 2: Measurement for Contrastive Learning

	Baseline	Contrastive Learning
$l_{uniform}$	-1.84	**-2.91**
l_{align}	**0.253**	0.280

Our analysis reveals that contrastive learning considerably enhances the uniformity of the embedding space, reducing $l_{uniform}$ from -1.84 to -2.91 (where a smaller value indicates a more uniform embedding space). Simultaneously, it preserves alignment by only increasing l_{align} by a marginal 0.027 (note that a smaller value also represents more concentrated similar embeddings).

5 Analysis

5.1 Confusion Matrix



(a) (b) (c)

Figure 2: The confusion matrices for all tasks. The number indicates the number of predictions falling in that category.

The confusion matrix shows the performance of our model on three different tasks. For SST, we cast the continuous values into discrete labels. We can see that for STS and SST, the diagonal cells are brighter than other cells, and the darkness of the color increases as we go off the diagonal and move toward the corners. This shows despite that the model classifies most samples correctly, it still confuses among classes that are close to the true labels (e.g. confuses between the scale of 2 and 3 in sentiment analysis). On the other hand, the confusion matrix of the Paraphrase detection task shows that the incorrect predictions tend to be false positives, i.e. the model over-correlates two sentences as paraphrases. This shed a light on potential improvements of our model: to decrease the false positive through weighted binary cross entropy loss or other regularization techniques.

5.2 Error Analysis

In this part, we will try to find error cases from all each of the tasks and investigate why our model fails. For STS and SST, we would identify incorrectly classified samples in two cases. The first case is samples that are far from the label (2 - 4 classes away) and the second case is samples that are close to the label (within 1 - 2 classes).

5.2.1 STS

Here are some examples of misclassified samples with label differences greater than 4:

A girl is talking to her dad on a cellphone. | *a girl is talking on her phone.* | *Pred: 0.35 Label: 4.4*
A woman is putting on sun glasses. | *A woman puts on sunglasses.* | *Pred: 0.94 Label: 5.0*

We can notice that the model tends to underestimate their similarities. We can easily tell that the two pairs of sentences are very similar despite their different wording or tense. It is possible that the model is not capable of extracting sentence meaning that is independent of the tense and additional addressing words. We then look at misclassified samples that are not off by a lot:

A grey cat with green eyes looking into the camera. | *A grey, black, and white cat looking at the camera.* | *Pred: 1.44 Label: 3.2*
Halliburton 1Q income rises 23 pct | *Halliburton Q1 profits rise on strong North America sales*| *Pred: 0.82 Label: 2.6*

Both pairs are underestimated by the model. For the first pair, it looks like the model emphasizes on the difference in the look of the cats, thus underestimating the similarity in the meaning of the two sentences. It lacks the ability of abstraction. Similarly, the second pair express similar meanings but the model seems to focus too much on the literal differences between the two sentences.

5.2.2 SST

Here are some examples of misclassified samples with label differences greater than 3:

It seems like I have been waiting my whole life for this movie and now I can't wait for the sequel. | *Pred: 0 Label: 3*
It 's everything you don't go to the movies for. | *Pred: 3 Label: 0*

For the first sentence, the phrase "waiting my whole life" might be misinterpreted by the model as impatience or frustration, which could lead to the misclassification of the sentiment. For the second sentence, its structure is complex and obscure, with a negation ("don't") as the only indicator of negative meaning. The model might have struggled to understand the implicit meaning hidden in the act of not going to the movies. Here are some examples of misclassified samples that are not off by a lot:

The film serves as a valuable time capsule to remind us of the devastating horror suffered by an entire people. | *Pred: 3 Label: 2*
The film 's welcome breeziness and some unbelievably hilarious moments – most portraying the idiocy of the film industry – make it mostly worth the trip. | *Pred: 3 Label: 4*

The first sentence has a mixed sentiment. While the film's purpose is positive, the subject matter (devastating horror) is negative. The model might have difficulty balancing these contrasting sentiments. The second sentence expresses a positive sentiment. The speaker appreciates the film's breezy and humorous nature, although the subject matter (idiocy of the film industry) could be considered negative. Our model might misclassify the sentiment due to the complexity of the sentence or the presence of words like "idiocy," which could be perceived as a strong indicator of negative sentiment in other contexts.

5.2.3 Paraphrase

There are far more False Positive (model predict them as paraphrase whereas they are not) than False Negative samples in this binary classification task, so we will focus on investigating those FP pairs:

What can you get as a customer of Star Alliance?' | *'What are some ways to register with Star Alliance?*
Can I use Jio in 3G phone? | *How is Jio 3G?*
Are water based moisturizers good for dry skin on the face? | *Which moisturizer is best used for dry skin?*
What are some facts that everyone knows? | *What are some facts that everyone should know?'*

The reasons for these misclassifications can be attributed to several factors. Firstly, the sentences in each pair often share similar vocabulary and sentence structure, which misleads the model to identify them as paraphrases. Secondly, the model may not adequately account for differences in

meaning that arise from variations in very subtle but important words or phrases, i.e. everyone knows versus everyone should know. Lastly, the model may struggle to discern the nuanced differences in the context or focus of the sentences. For example, it cannot differentiate the difference between yes-or-no questions and interrogative questions.
From the above error analysis, we propose several potential directions for model improvements: (1) Use a stemming filter to remove the influence of tenses for sts task; (2) Tune the model specifically on complex/indirect sentences to understand the implicit and contextual reasoning; (3) Consider more subtle, fine-grained differences for paraphrase task to reduce false positives.

6 Conclusion

In conclusion, our proposed Multitask BERT pre-training and fine-tuning model demonstrates significant improvements over the original BERT model by integrating contrastive learning methods, multitask learning strategies, a dense similarity scheme with fine-grained embedding representation, and regularized multitask optimization. The experimental results indicate that our approach achieves better performance on the tested tasks, revealing the potential of our methods in addressing data efficiency and training cost challenges in NLP.

Our model, while demonstrating improved performance, has certain limitations as indicated by the analysis of its results. The confusion matrix reveals that the model tends to confuse classes close to the true labels in the STS and SST tasks, and exhibits a higher false positive rate in the Paraphrase Detection task. Specifically, the error analysis for STS indicates that the model underestimates similarity due to differences in wording, tense, or an inability to abstract meaning. In the case of SST, the model struggles with mixed sentiment sentences, complex sentence structures, and implicit meanings. Finally, the error analysis for Paraphrase Detection suggests that the model is misled by similar vocabulary and sentence structure, and encounters difficulties in detecting nuanced differences in context or focus. These features indicate that the model is understanding the languages in a relatively naive and superficial way.

To build upon these promising results and further improve the model, we propose the following directions for future work: Investigate the integration of more sophisticated contrastive learning methods or other self-supervised approaches to further enhance the quality and expressiveness of BERT embeddings in multi-task settings. Evaluate the effectiveness of different pooling techniques, such as max-pooling or weighted pooling, in the context of fine-grained embedding representation to optimize performance on similarity-based tasks. Investigate additional multi-task learning strategies to further improve data efficiency and reduce training costs, focusing on addressing optimization challenges in multitask BERT models, such as embedding space wellness and task information contradiction. Examine the impact of various BERT embeddings and head structures on different NLP tasks, aiming to enhance the understanding of intrinsic mechanisms and inform the design of more effective models. Explore extensions of the proposed model to other downstream NLP tasks and evaluate its applicability across a broader range of problems, thereby assessing the model's potential for generalization and scalability. By pursuing these research directions, we hope to contribute to the development of more efficient, effective, and versatile multitask BERT models, ultimately pushing the boundaries of current NLP capabilities.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[2] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020.

[3] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery.

[4] Yifan Peng, Qingyu Chen, and Zhiyong Lu. An empirical study of multi-task learning on bert for biomedical text mining. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 2020.

[5] Eleftherios Kapelonis, Efthymios Georgiou, and Alexandros Potamianos. A multi-task bert model for schema-guided dialogue state tracking, 2022.

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 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, November 2021. Association for Computational Linguistics.

[7] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[8] Manish Munikar, Sushil Shakya, and Aakash Shrestha. Fine-grained sentiment classification using bert. *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, Nov 2019.

[9] Bi Qiwei, Li Jian, Shang Lifeng, Jiang Xin, Liu Qun, and Yang Hanfang. Mtrec: Multi-task learning over bert for news recommendation. In *Findings of the Association for Computational Linguistics: ACL*, page 2663–2669, 2022.

[10] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019.

[11] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[12] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*, 2017.

[13] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

A Appendix (optional)

If you wish, you can include an appendix, which should be part of the main PDF, and does not count towards the 6-8 page limit. Appendices can be useful to supply extra details, examples, figures, results, visualizations, etc., that you couldn't fit into the main paper. However, your grader *does not* have to read your appendix, and you should assume that you will be graded based on the content of the main part of your paper only.