A Simple Multi-Channel Convolution Attention For Biomedical Semantic Sentence Similarity Estimation

A cute kid

The digital version of the manuscript.

Unlike journal papers, this is a conference paper.

Because this manuscript template looks pretty rough, I wouldn't recommend it.

Abstract: With the rapid growth of information presented in text form within the biomedical field. The natural language processing (NLP) applications are becoming increasingly important to facilitate the retrieval and analysis of these data. Computing the semantic similarity between sentences is an important component in many NLP tasks, including text retrieval and generation. Several approaches have been proposed for estimating semantic sentence similarity in generic English. In this paper, I propose a multi-channel convolution attention for sentence-level semantic similarity computation in the biomedical domain. When combined with a general pre-trained model, this method achieved a Pearson correlation coefficient of 84.72% on the BIOSSES dataset, which is 1.12 percentage points higher than the state-of-the-art methods reported on this dataset's paper. I also demonstrated through several analyses that this approach has achieved cost-effectiveness on par with SOTA standards in benchmark rankings, without being confined to a single performance metric.

Keywords: **Convolution, Biomedical Semantic Sentence Similarity.**

1 INTRODUCTION

Semantic text similarity estimation refers to measure the similarity of texts based on their meaning and semantic content, without being confined to shallow or syntactic representations. Figure 1 demonstrates a piece of data from the BIOSSES(Soğancıoğlu et al., 2017) dataset. In this paper, I first use the DEBERTa(He et al., 2020) model, which has been pre-trained on other corpora to represent the given sentences. Subsequently, I defined two convolutional interfaces with different numbers of channels to perform convolution operations on the pretrained representations. I define the entire convolution process to convolution attention. Then, I feed the attention weights into the expert matrix to simulate expert ratings. Finally, the simulated expert values are fed into the fitting layer to determine the final score. The main contributions are followed:

Sentence1: Considerable evidence indicates that cancer cells develop dependencies on normal functions of certain genes that can potentially be exploited to improve therapeutic strategies.
Sentence2: In the case of cell response to stress, cyclin D1 can be degraded through its binding to the anaphase-promoting complex and a RXXL sequence located in the NH2-terminal part of the protein.

Score: 0

Figure 1: This introduction of semantic text similarity estimation.

- A multi-channel convolution attention mechanism is proposed.
- Perform dimensionality reduction analysis on expert ratings.
- Visualize weights for multi-channel convolution attention.
- Use a violin plot to show the distribution of expert ratings.

2 DATASET AND MODEL ATCHITECTURE

2.1 Dataset

Table 1 demonstrates the partitioning of the BIOSSES dataset.

Table 1: Dataset Partitioning.

Dataset Splited	data volume
train	64
dev	16
test	20

2.2 Model Atchitecture

The figure 2 demonstrates the whole atchitecture of the proposed model. It comprises a deberta layer, a multi-channel convolutional attention layer, an expert matrix layer, and a fit matrix layer.

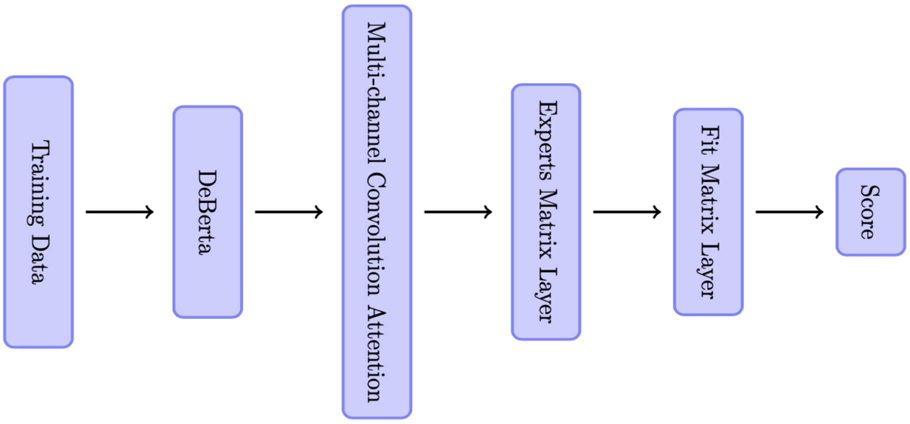

Figure 2: This architecture of the proposed model.

2.2.1 Deberta

Towards conserving energy and better demonstrating Deberta's generalization capabilities. In this paper, I directly invoke the DEBERTa model, which has been pretrained on other corpora to represent the text in the BIOSSES dataset. Given a batch of data in a dataset, the DeBERTa output is the tensor shape of $[B, L, D]$ where B is the batch size, L is the text length, and D is the dimension of DeBERTa's last hidden layer feature. Regarding pre-trained models, I recommend readers read (Chan et al., 2020). Moreover, the most cost-effective approach to pre-training is to draw inspiration from and reproduce other high-performance pre-trained models, ensuring that the log information output during the reproduction process aligns with the parameters output during the original pre-training model's training. Once the model has been successfully reproduced, further improvements can be made based on this foundation.

2.2.2 Multi-chanel Convolution Attention

Figure 3 demonstrates the process of the multi-channel convolution attention. I first define convolutional interfaces with varying channel counts based on the output dimensions of the DeBerta model. Subsequently, through elegant manipulation of the tensor, I altered the shape of the tensor output by the DeBerta model until it could be convolved. This idea originated from token statistics self-attention(Wu et al., 2024), who use einops to rearrange the shape of the tensor. The final attention weights are a tensor of shape $[B, L]$.

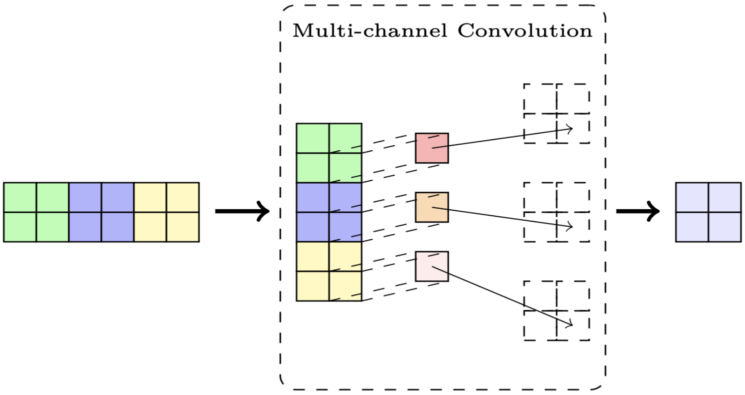

Figure 3: This architecture of the proposed model.

2.2.3 Experts Matrix

I defined a matrix of shape $[D,N]$, where N corresponds to the number of experts evaluating this dataset, and D is the feature of Deberta's last hidden dimension. The tensor shape received by the expert matrix layer is the multiply of the attention weights and the original output from the Deberta.

2.2.4 Fit Matrix

I defined a matrix of shape $[N,1]$, where N corresponds to the number of experts. The tensor shape received by the fit matrix layer is the output from the experts matrix layer.

3 RESULTS AND ANALYSIS

3.1 Hyperparameters

Table 2 demonstrates the hyperparameters of the model. Figure 4 illustrates the relationship between the number of training iterations and the loss value of the model.

Table 2: model hyperparameters.

Hyperparameters	value
batch size	1
learning rate	0.001
epoch	20

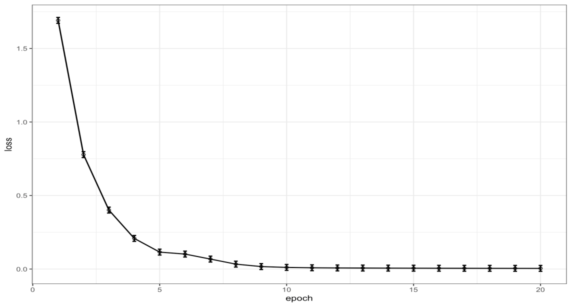

Figure 4: The model training loss and its epoch.

3.2 Results

Table 3 presents the Pearson correlation coefficients and training time for different models on the BIOSSES dataset. The training time and Pearson correlation demonstrate that the model I proposed offers exceptional cost-effectiveness.

Table 3: model results.

model	Pearson correlation.	training total time
Linear regression	0.836	-
Multi-channel convolution based	0.8472	223.28s

3.3 Dimensionality Reduction Analysis

Figure 5 demonstrates the dimensionality reduction analysis. I used SVD to reduce the dimensionality of the expert matrix output and presented it with a scatter plot.

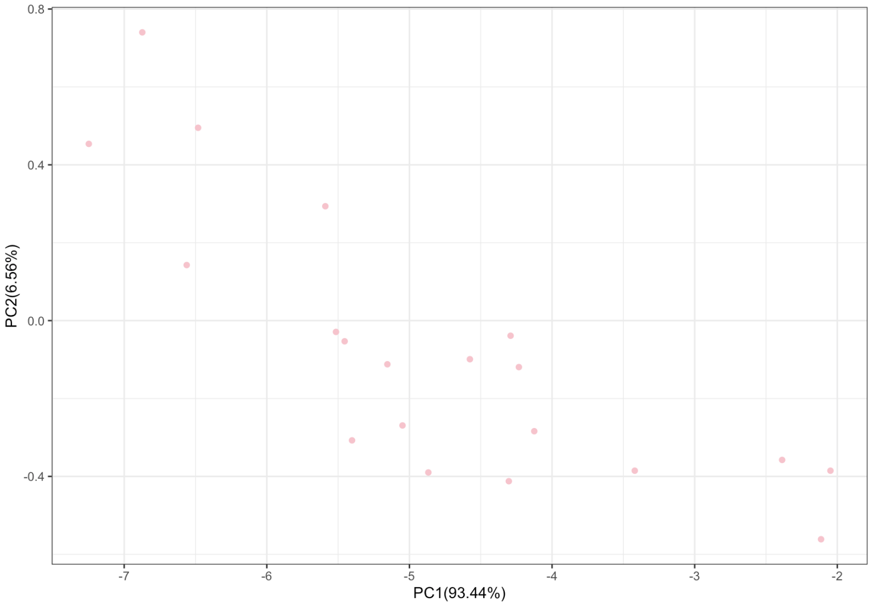

Figure 5: The Model Dimensionality Reduction Analysis.

3.4 Attention Visualization

Figure 6 is the heatmap of attention weights. The vertical bar on the right corresponds to the color representing the weight value change, decreasing from top to bottom.

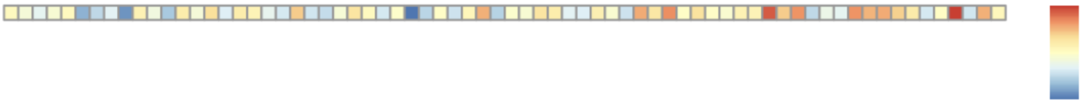

Figure 6: The attention visualization.

3.5 Distribution Analysis

Figure 7 shows the violin plot constructed based on the outputs from the expert layer. The white bars in the middle correspond to box plots, where the upper horizontal line represents the upper quartile, the lower horizontal line represents the lower quartile, and the middle horizontal bar represents the median. The scatter points in the figure were generated using the SINA function to adapt to the distribution of the violin plot.

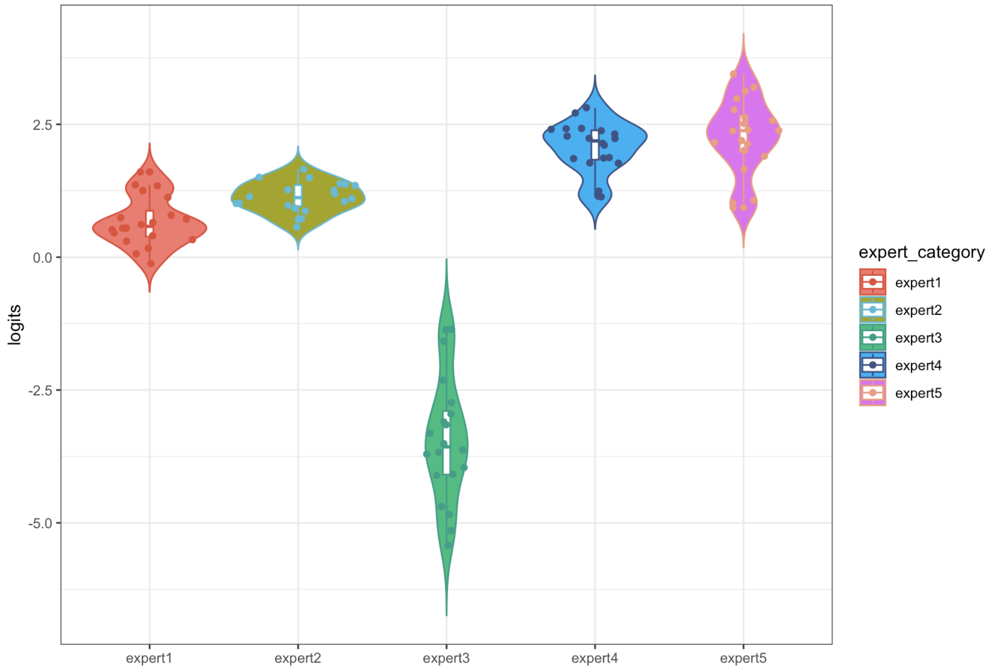

Figure 7: The logits distribution analysis.

4 CONCLUSIONS

I proposed a multi-channel convolution attention mechanism.

ACKNOWLEDGEMENTS

Thank for Deberta and the Biosses dataset.

REFERENCES

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. *arXiv preprint arXiv:2010.10906*.

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Soğancıoğlu, G., Öztürk, H., and Özgür, A. (2017). Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Wu, Z., Ding, T., Lu, Y., Pai, D., Zhang, J., Wang, W., Yu, Y., Ma, Y., and Haeffele, B. D. (2024). Token statistics transformer: Linear-time attention via variational rate reduction. *arXiv preprint arXiv:2412.17810*.