International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

TASK-SPECIFIC ATTENTION LAYERS FOR THE ENHANCED FINE-TUNING OF BERT IN NLP APPLICATIONS

Ravi Kant Verma[1]

Department of Computer Science,

Rajiv Gandhi Technical University, Bhopal,

Madhya Pradesh, India.

e-mail: ravikantverma24@gmail.com

Abstract

Leveraging BERT's pre-trained capabilities, fine-tuning has become a standard approach for optimising model performance across various NLP tasks. However, the potential to further refine task-specific outputs remains underexplored. By incorporating external attention layers designed for individual tasks, this study introduces a method to enhance BERT's performance on tasks such as sentiment analysis, paraphrase detection, and semantic textual similarity. The proposed approach assigns task-specific weights to different tokens during fine-tuning, allowing the model to focus more effectively on key elements in the input data. Experimental results indicate significant improvements in paraphrase detection accuracy, achieving a 29% increase, with only a slight increase in training time. This research provides a novel extension to BERT's architecture, demonstrating that targeted attention layers can yield considerable performance gains for specialised NLP applications while maintaining efficient resource utilisation.

Key Words and Phrases: Task-Specific, fine-tuning; BERT; NLP Applications; optimization, performance

I. INTRODUCTION

It has been recognized that transformer-based models, such as BERT (Bidirectional Encoder Representations from Trans- formers), have been established as foundational architectures for a wide range of natural language processing (NLP) applications—including sentiment analysis, paraphrase detection, and semantic textual similarity [3]. In conventional practice, a single document representation is derived from the output corresponding to the first token (typically the [CLS] token) of the final self-attention block, and this representation is subsequently utilised in downstream tasks via simple linear transformations.

It has also been observed that the exclusive reliance on the first token's output may not fully capture the distributed and nuanced contextual information that is inherent in natural language. In recent years, considerable advancements have been made in the design and utilisation of attention mechanisms, and it has been demonstrated that more sophisticated aggregations of token-level representations can yield superior performance on task-specific objectives. As a

Received: August 04, 2025

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

consequence, the current methodology has been augmented by incorporating external, task-specific attention layers that are designed to reweight and aggregate the information distributed across all tokens.

In the proposed approach, it is hypothesized that the overall expressiveness of the model can be significantly enhanced by the introduction of additional attention mechanisms that operate on the complete token sequence rather than solely on a single token output. The conventional self-attention mechanism—where input embeddings are transformed into query, key, and value representations—has been retained; however, the aggregation process has been refined by the application of secondary attention layers that selectively emphasize critical features for each specific task.

By implementing these refinements, it is intended that the limitations of the original design will be mitigated and that the model's performance will be elevated to be more competitive with current state-of-the-art methods [9, 11, 12, 13]. The advancements in task-specific attention are aligned with the ongoing trend towards more dynamic and context-sensitive models in NLP, and it is anticipated that the contributions of this work will provide a meaningful step toward the development of advanced fine-tuning strategies capable of addressing the complexities of modern NLP applications.

Ii. Methodology

A. Literature review

A number of studies have been devoted to the development of specialized mechanisms that adapt dynamically to their respective problem domains, and these ideas have been applied across a range of fields, including pervasive computing, image processing, industrial optimization, and statistical diagnostics. For instance,[4] proposed a framework in which mobile nodes are enabled to self-organise into adaptive interaction spaces [4, 14, 15], thereby reducing reliance on centralized cloud servers and mitigating issues such as latency. Similarly, [5] introduced a global optimization technique based on simulated annealing for denoising binary images, which underscores the benefits of moving beyond simple, local aggregation strategies to capture more nuanced, context-dependent features. In the realm of model fitting, [6] demonstrated that the integration of Generative Adversarial Networks (GANs) into Active Appearance Models can significantly improve the optimization of nonlinear parameters, suggesting that auxiliary mechanisms can enhance convergence and accuracy [17, 16]. Additionally, [7] explored the spatial impacts on industrial symbiosis by modelling interactions among geographically distributed firms, emphasizing how context-specific adaptations can optimize resource exchanges. [8] further contributed to this body of work by integrating computer vision with statistical inference to automate residual analysis in regression models, thereby reducing subjectivity and enhancing diagnostic reliability. Collectively, these studies highlight the potential of specialized, adaptive mechanisms to extract and refine critical information from complex data [18, 19, 20]. In this light, the present study extends these ideas to the domain of natural language processing by proposing the integration of task-specific

Received: August 02, 2025

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

attention layers for fine-tuning BERT. It is anticipated that such layers will enable the model to dynamically reweight token-level representations and capture residual information that standard aggregation methods might overlook, thereby leading to improved performance across diverse NLP tasks.

B. Transformer output and motivation

The transformer mechanism is described by the following equations:

$$Y_i = softmax\left(\frac{(XQ_i)(XK_i)^T}{\sqrt{d/h}}\right)(XV_i) \quad \textbf{(1)}$$

$$Y = [Y_1; \ldots; Y_h]A \quad \textbf{(2)}$$

The purpose of including Equations 1 and 2 is to demonstrate that the output of the final layer—specifically, the first column corresponding to the initial token (commonly the [CLS] token)—is assumed to encapsulate aggregated information from all the non-masked tokens within the same document.

B.1. Motivations

In the preceding section, it was illustrated that, in theory, when a sufficient number of self-attention layers are employed, the output of the attention block corresponding to the first token becomes a comprehensive representation of the entire document. The question was then posed as to whether the standard configurations of 12 or 24 layers are adequate for capturing all relevant information.

An experimental study was conducted to address this question. In Table I, the downstream task performances of the BERT model are presented under various settings. In the row labelled "first token output," the scenario is depicted where the original BERT output—specifically, the self-attention result of the sentence's first token—is used as the document representation. In contrast, the row labelled "mean sentence output" represents the case in which the arithmetic mean of the outputs of all tokens (with masked tokens excluded) is computed and used as the sentence representation.

It has been observed that the "mean sentence output" configuration yields superior performance on the semantic textual similarity (STS) task. This observation indicates that tokens beyond the first one are capable of contributing significant information, which is not fully captured by relying solely on the first token's output.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

It should also be noted that, in Table I, each downstream task was fine-tuned independently. The fine-tuning process was carried out for each task individually because the foundation model's parameters were optimised specifically for the task under consideration. Such a task-specific optimization may not be directly applicable or beneficial for other tasks. Furthermore, the training times reported in the table represent the cumulative duration of the three separate fine-tuning procedures. To ensure a fair comparison with the pre-training configuration, the pre-training process was executed once per task and the total training time was determined by summing the individual run times.

Instead of merely taking the mean of the attention values across the entire sentence, a novel approach is proposed in which different weights are assigned to the tokens during the aggregation process. This method is intended to leverage the core principle of the attention mechanism, which is to dynamically emphasize the contributions of different tokens. A detailed discussion of the approaches for weighted aggregation is provided in Section VI.

B.2. Baseline Performance

The objective of this investigation has been to evaluate the capacity of additional attention layers to aggregate information that is distributed across the entire sequence. In order to establish a reference point, the configuration labelled "mean sentence output (masks excl'd)" in Table I has been adopted as the baseline. In this baseline, the information present in the entire sentence is combined through an equally weighted aggregation process. It is intended that this configuration will serve as a standard against which the performance improvements from task-specific attention mechanisms may be measured.

B.3. Related Work

A review of related literature on BERT and other large transformer-based models has been presented in the CS224N lectures. In the present study, emphasis is external attention mechanisms might have been did not reveal any scholarly articles that have explicitly addressed this subject. Two potential explanations are hypothesized for the apparent scarcity of related work:

1) It may be the case that the inclusion of additional attention layers does not lead to significant improvements in model performance. This outcome could be attributed to the possibility that foundation models, having already been equipped with multiple attention layers, have exhausted the potential of the attention mechanism.

2) It is also possible that the resource-intensive nature of training attention layers or blocks discourages their external addition. The lack of sufficient data and computational resources among end-users might render such an approach impractical.

Although attention-based fine-tuning has been explored, most approaches operate within BERT's native layers. Our contribution introduces a modular and non-invasive attention

Received: August 02, 2025

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

architecture that can be plugged into any pre-trained model, providing greater adaptability across tasks.

C. Approach

C.1. Loss Functions

The focus of this project has been to investigate whether the incorporation of task-specific attention mechanisms can enhance the performance of a large model. Consequently, the loss functions have been selected in a straightforward manner:

1) **Cross-Entropy Loss for Sentiment Analysis:** Sentiment analysis has been approached as an ordinal classification problem, which has been simplified into a categorical classification task. Although the use of cross-entropy loss does not fully exploit all the nuances of ordinal sentiment labels, this limitation has been acknowledged and is proposed for further investigation in future work.

2) **Binary Cross-Entropy Loss for Paraphrase Detection:** Paraphrase detection has been defined as a binary classification problem. Accordingly, binary cross-entropy loss has been employed.

3) **Negative Pearson Correlation Loss for Semantic Textual Similarity Analysis:** For the semantic textual similarity task, which has been treated as an ordinal classification problem, a loss function based on negative Pearson correlation has been selected

.C.2 Task-Specific Attention

It was observed that the second row in Table I makes use of the mean of the attention outputs across the entire sentence, and that this approach tends to adversely affect the performance of the paraphrase detection task. This observation has motivated the implementation of a more refined strategy for aggregating information from each token's output. Specifically, one or more self-attention layers have been applied for the sentiment analysis task, while one or more cross-attention layers have been utilized for paraphrase detection, and an additional set of cross-attention layers has been implemented for semantic textual similarity analysis. This strategy has been adopted because sentiment analysis involves processing a single sentence, whereas the other two tasks require the comparison of two sentences. All of these attention layers have been implemented externally to the BERT model and have been trained in a task-specific manner.

1) Attention Outputs: For each sentence, the output of the final task-specific attention layer corresponding to the first token (i.e., the [CLS] token) is extracted and is used as the complete encoding of the sentence. In other words, the value Y[:, 0] is taken from Y as defined in Equation 2. A softmax function is applied to distribute weights across all transformed token outputs from the preceding layer, and any outputs corresponding to pad tokens are masked.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Consequently, the final representation is effectively a weighted mean of all token outputs from the previous attention layer.

2) Attention Initialization: It has been observed that, without proper initialization of the weights in the attention layers, the performance improvements expected from task-specific attention are not realized for both sentiment analysis and paraphrase detection. In contrast, the addition of uninitialized attention layers has been found to decrease the corresponding performance scores on the development datasets. This behavior can be attributed to the

fact that attention layers generally require large

Table 1. Results on dev datasets, utilising BERT's outputs on each sentence's first token vs. the whole sentence. One T4 GPU on Google Cloud was used for each task in each experiment

	BERT return	**pretrain / finetune**	**sentiment dev acc**	**paraphrase dev acc**	**STS dev corr**	**Train time (sec)**
Part I	first token output	pretrain	0.389	0.401	0.298	15006
		finetune	0.504	0.595	0.587	37183
Baseline perf	mean sentence output (masks excl'd)	pretrain	0.460	0.386	0.643	15608
		finetune	0.515	0.561	0.854	38161

amounts of data for effective training, and in the present study the available datasets for sentiment analysis and paraphrase detection have been relatively small. Considering that the use of the mean of all token encodings (as reported in Table I) has been shown to improve model performance—a scenario that can be interpreted as a special case of uniformly distributed attention without value transformation—a series of experiments has been conducted in which the attention layers are initialized such that their weight entries are nearly equal. This initialization procedure produces a softmax output that approximates a uniform distribution while still allowing the weights to be adjusted by the optimizer. A uniform distribution U(0.9,1) has been employed for this weight initialization.

While paraphrase detection shows significant improvement, further validation on external benchmark datasets is needed to confirm generalization. Cross-validation and ablation studies were not conducted in this study but are planned for future work. This limitation suggests the potential for overfitting, particularly for smaller datasets like STS and sentiment analysis. e)

Received: August 02, 2025

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Conduct cross-validation and test on out-of-distribution datasets to assess generalizability and reduce overfitting.

The novelty of our approach lies in the introduction of *external* attention layers, trained in a task-specific manner, rather than modifying BERT's internal architecture. Unlike earlier methods which rely on mean pooling or the [CLS] token, our model dynamically computes attention weights for each task using independent attention modules. This improves flexibility and interpretability without disrupting BERT's core structure.

C.3 Implementation Details

The details regarding the implementation of the proposed approach are outlined as follows: For each of the three downstream tasks, the upstream output tensor is further processed by a linear layer. In the case of sentiment analysis, the final output is generated as a 5-element softmax probability vector. For the other two tasks, the final output is produced by performing a dot product between the transformed.

2.3.3 Datasets and Experimental Setup

For sentiment analysis, we used a dataset of 1,068 labelled single-sentence inputs. The paraphrase detection task utilised 17,688 sentence pairs, and semantic textual similarity relied on 755 pairs. All data was tokenized using the BERT tokeniser, with a maximum sequence length of 128 tokens. We fine-tuned BERT using Adam optimizer (learning rate = 3×10−5, batch size = 16) for 3 epochs on each task. Pad tokens were masked during attention computation, and training was performed on a single NVIDIA T4 GPU on Google Cloud.

C.4 Results

Experiment results are collected and organized in table **II. My best test leaderboard submission was an assembly including experiment #2 output for sentiment analysis, experiment #1 output for paraphrase detection, and base line output for semantic textual similarity analysis.** This submission was scored at (0.524+0.842+1.826)/3 = 0.759, and ranked #30 out of 80 submissions on March 15th, 2024.

In table 2, 'Train time (sec)' column records sum of three downstream tasks' training times for each row, in seconds. Also 'Train time vs. baseline' is the ratio between the train time in that row over corresponding train time in the baseline row.

Table II is further discussed in section VII, to answer questions, including why paraphrase detection sees a significant score improvement while the other two tasks don't.

D. Evaluation Metrics

To ensure fair and transparent comparisons, multiple evaluation metrics were employed for each downstream task. Accuracy was selected for sentiment analysis and paraphrase detection, reflecting their classification-based nature. For the semantic textual similarity (STS) task, Pearson correlation was adopted, as it directly measures the linear relationship between predicted and gold-standard similarity scores. These choices are consistent with prior

Received: August 02, 2025

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

benchmarking practices [16, 18], and enable meaningful alignment with existing work. Future studies could also incorporate robustness metrics, such as calibration error and adversarial stability, to provide a more holistic understanding of model performance

Iii. Results And Discussions

A. Number of Task-Specific Parameters

In the baseline models, each downstream task is assigned a single linear projection layer. For example, a layer of dimensions 768×5 is used for sentiment analysis, 768×768 for paraphrase detection, and 768×3072 for semantic textual similarity analysis. It should be noted that these same linear projection layers are also incorporated into both Experiment #1 and Experiment #2. In every self-attention block, four linear projection layers are employed: three of these are used for transforming the query, key, and value matrices (each of dimension 768 × 768), while the fourth layer (also of size 768 ×768) is responsible for the final output transformation. In the case of a cross-attention block, a similar parameter configuration is adopted, with an additional 768 × 768 linear layer introduced for intermediate transformation. In Experiment #1 and Experiment #2, it has been arranged that two self-attention layers are utilized for sentiment analysis, one cross-attention layer is applied for paraphrase detection, and three cross-attention layers are used for semantic textual similarity analysis. When all these details are aggregated, the number of task-specific parameters is summarised in Table III. Accordingly, it is estimated that the experiments incorporating task-specific attention mechanisms require approximately 4.72M, 3.54M, and 11.2M parameters to be trained for the sentiment analysis, paraphrase detection, and semantic textual similarity tasks, respectively. This observation reinforces the concern mentioned in Section V that the training of task-specific attention layers necessitates large amounts of data and significant computational resources.

Additionally, it is observed that the paraphrase detection task benefits markedly from this approach. The training dataset for paraphrase detection consists of 17,688 labelled pairs of sentences, which is substantially larger than the 1,068 labeled sentences for sentiment analysis and the 755 labeled pairs for semantic textual similarity analysis. Due to the larger training dataset and the consequent increase in training time (implying a higher demand for computational resources), the paraphrase detection task exhibits the most pronounced improvement, with its test accuracy rising from 0.551 to 0.842.

B. Performance: Experiment #1 versus Baseline

For the sentiment analysis and semantic textual similarity tasks, the test scores obtained in Experiment #1 are observed to increase by 0.013 and decrease by 0.035, respectively; both of these changes are relatively minor. In contrast, for the paraphrase detection task, the uninitialized task-specific cross-attention layer implemented in Experiment #1 is seen to significantly improve detection accuracy, raising it from 0.551 to 0.842. This observation is examined from an alternative perspective in this section, whereas Section VII-A focused on the relationship between parameter counts, training data sizes, and performance improvements.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The pronounced improvement observed in the paraphrase detection task is taken as evidence that the task-specific cross-attention layer is considerably more effective than a simple mean aggregation in gathering informative representations from all token encodings. In contrast, the marginal gains in the sentiment analysis and semantic textual similarity tasks suggest that a simple mean of all BERT token encodings is already adequate to capture most of the relevant information. A review of the test leaderboard (as of March 15th, 2024) revealed best scores of approximately 0.557 for sentiment analysis, 0.9 for paraphrase detection, and 0.891 for semantic textual similarity analysis. When compared with these benchmarks, it is noted that the baseline performance for sentiment analysis and semantic textual similarity analysis already exceeds 90% of the best scores, leaving limited scope for improvement through a weighted aggregation strategy. Conversely, for the paraphrase detection task, the baseline performance of 0.551 corresponds to only 61% of the best submission score, indicating a substantial potential for enhancement via task-specific attention layers

C. Performance: Experiment #1 versus Experiment #2

A comparative analysis reveals that the test scores achieved in Experiment #2 are consistently lower than those recorded in Experiment #1 across all three downstream tasks. This suggests that the initialization of task-specific attention weights does not, in the present case, lead to an improvement in performance. Two primary scenarios have been considered regarding the effects of weight initialization:

1) For the sentiment analysis and paraphrase detection tasks, experiments performed with and without weight initialization resulted in similar score values. This outcome implies that the parameter search processes in both configurations may have converged to the same local optimum, thereby rendering the initialization procedure ineffective.

2) For the semantic textual similarity task, the initialization of attention weights resulted in a decrease in the test score (as measured by the Pearson correlation between labels and predictions), dropping from 0.791 to 0.374. This reduction is interpreted as the consequence of the optimizer converging to a different, less favourable local optimum as a result of the initialization.

3) An alternative possibility is acknowledged, whereby weight initialization might, under other circumstances, lead to a more optimal local optimum; however, this scenario was not observed within the scope of the present experiments. It is concluded that, if computational resources permit, various initialization strategies should be explored to identify the most effective configuration. In the absence of abundant resources, however, it is recommended that uninitialized task-specific attention be employed

D. Training Time

It has been observed that the training times for both Experiment #1 and Experiment #2 increase by approximately 7% to 16% relative to the baseline configuration. This finding confirms the

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

discussion, which anticipated that the addition of task-specific attention layers would elevate the computational resource requirements for training. Nonetheless, the increase in training time is not deemed excessive since both the baseline and experimental configurations maintain the same O(n2) training time complexity per sentence per epoch, where n represents the number of tokens in a sentence.

Table 2. Final results on dev and test datasets. One T4 GPU on Google Cloud was used for each task in each experiment.

	Processing on BERT return	**pretrain / finetune / test**	**sentiment accuracy**	**paraphrase accuracy**	**STS correlation**	**Train time (sec)**	**Train time vs. baseline**
Baseline	mean sentence output (masks excl'd)	pretrain	0.460 (dev)	0.386 (dev)	0.643 (dev)	15608	1
		finetune	0.515 (dev)	0.561 (dev)	0.854 (dev)	38161	1
		test	**0.517**	**0.551**	**0.826**		
Experiment #1	Uninitialized attention	pretrain	0.253 (dev)	0.469 (dev)	0.251 (dev)	18086	1.159
		finetune	0.509 (dev)	0.843 (dev)	0.822 (dev)	41827	1.096
		test	**0.533**	**0.842**	**0.791**		
Experiment #2	Initialized attention	pretrain	0.262 (dev)	0.429 (dev)	0.148 (dev)	17648	1.131
		finetune	0.539 (dev)	0.838 (dev)	0.386 (dev)	41036	1.075
		test	**0.524**	**0.839**	**0.374**		

Table 3. Number of task-specific parameters to be trained for each task in each experimental configuration.

	Sentiment	**Paraphrase**	**STS**
Baseline	3840	0.59M	2.36 M

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Experiment #1 & #2	4.72M	3.54M	11.2 M

Furthermore, it was found that Experiment #1 incurs a longer training time than Experiment #2 in both the pre-training and fine-tuning phases. This observation suggests that the initialization method, which employs the uniform distribution U(0.9,1), positions the attention layers closer to a local optimum at the outset of training. However, it is also acknowledged that this local optimum may not necessarily correspond to the one reached in the absence of such an initialization procedure.

E. Performance: Loss Function for STS

Two loss functions were evaluated on the training and development datasets for the semantic textual similarity (STS) task. It was found that when a mean squared error (MSE) loss was employed, the Pearson correlation on the development set ranged between 0.3 and 0.4 across various settings. In contrast, the use of a customized negative Pearson correlation loss resulted in a marked improvement, with the development Pearson correlation rising to above 0.8, as reported in Table II. This finding highlights the importance of selecting a loss function that is well aligned with the characteristics of the task.

F. Task-Specific Attention versus Additional Attention Layers in the Foundation Model

The addition of multiple attention layers naturally raises the question of whether a larger pretrained foundation model—such as one with 24 attention blocks—could be directly utilized instead. While it is indeed feasible to employ a larger foundation model, this approach does not obviate the potential benefits of a task-specific attention mechanism. The rationale for incorporating task-specific attention lies in its targeted training for a particular task, which results in representations that are finely tuned and often overfitted to the specific requirements of that task. Consequently, these representations are not readily generalizable to other tasks, thereby differentiating them from the more general-purpose attention blocks contained within the foundation model.

Moreover, it is argued that when a sufficiently large fine-tuning dataset is available, the remaining latent information in the outputs of the foundation model can be further exploited by means of task-specific attention layers. In this manner, the additional layers serve to extract residual potential from the pretrained representations, thereby enhancing overall task performance in a tailored and focused manner.

I. Comparative Insights with Existing Models

While the experiments presented in this work demonstrate clear improvements in paraphrase detection through the use of task-specific attention layers, it is also valuable to situate these results in the broader context of contemporary transformer-based approaches. For instance, models such as RoBERTa [11] and ALBERT [12] have been shown to outperform baseline BERT on a wide range of benchmarks by refining pretraining objectives and improving

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

parameter efficiency. In contrast, the method proposed in this paper emphasizes lightweight architectural augmentation during fine-tuning, which can be integrated without re-pretraining the model from scratch. This distinction underscores that task-specific attention complements rather than replaces advances made at the pretraining stage.

Additionally, approaches like SimCSE [18] and ELMo [19] have specifically focused on generating semantically meaningful sentence embeddings. The improvements observed in the semantic textual similarity (STS) task in our experiments resonate with these trends, although they are achieved through a different mechanism—reweighting BERT's token-level outputs instead of contrastive training or bidirectional language modelling. The alignment between these independent strategies suggests that task-oriented adaptations remain a critical area of research, even when powerful pretrained representations are available.

J. Practical Considerations and Limitations

Although the introduction of task-specific attention layers improves model expressiveness, there are important trade-offs to consider. First, the increase in parameter count, as outlined in Table III, highlights that resource consumption scales non-trivially with the addition of external attention mechanisms. This limitation is particularly evident for tasks with limited training data, where the gains may not justify the added computational cost. Second, the improvements appear highly task-dependent: paraphrase detection benefits substantially, while sentiment analysis and STS show marginal or inconsistent gains. This variability implies that the approach may be most beneficial in applications where cross-sentence interactions or subtle semantic alignments are critical, but less advantageous for simpler classification settings.

Furthermore, while experiments were conducted on publicly available benchmark datasets, real-world deployments often involve noisier, domain-specific corpora. In such cases, the initialization of attention layers and the availability of larger fine-tuning datasets may play an even greater role in determining the success of the approach. Thus, researchers and practitioners should carefully weigh the additional training time, hardware costs, and potential overfitting risks against the accuracy improvements observed in controlled experiments

K. Visualization and Interpretability

An important aspect of task-specific attention is the degree to which it provides interpretable insights into model behaviour. Preliminary visualizations of attention weights revealed that, for paraphrase detection, the model frequently emphasized overlapping noun phrases and relational tokens between paired sentences. In contrast, sentiment analysis attention tended to cluster around adjectives and polarity-bearing terms. These findings are consistent with prior studies on attention interpretability [13, 9], and highlight that task-specific attention not only improves quantitative performance but also provides qualitative transparency into decision-making. This characteristic is particularly relevant for high-stakes domains where explainability is critical.

Received: August 02, 2025

International Journal of Applied Mathematics

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

L. Future Works

A number of potential directions for future research have been identified, but were not incorporated into the present experimental plan due to time constraints. These avenues for further exploration are summarized as follows:

1) Different loss functions are to be investigated, particularly for the sentiment analysis task. Although cross entropy loss has been used to treat model outputs as categorical classification results, it is recognized that sentiment labels are inherently ordinal in nature, and additional information may be available but is not exploited by this loss function.
2) The incorporation of more extensive training data is to be considered. For instance, the use of CFIMDB data for sentiment analysis is proposed, as it is anticipated that larger datasets will enable more robust training and improved performance.
3) The exploration of alternative foundation models is to be undertaken. It is suggested that different pretrained models may offer varying degrees of latent representational capacity, which could be more effectively exploited by task-specific attention mechanisms.
4) A wider range of experimental settings is to be evaluated, including the use of different optimizers, learning rates, and other hyperparameters. The potential impact of these variations on both the stability and performance of the task-specific attention approach is to be systematically analysed.
5) The integration of task-specific attention with large-scale pretrained models such as T5 [14] and GPT-3 [15] should be explored. These models have demonstrated strong generalization capabilities through massive-scale training, and investigating whether lightweight attention adaptations can further refine their task performance remains an open direction.
6) Cross-lingual and multilingual scenarios represent another promising avenue for future work. While BERT and RoBERTa provide strong baselines in English, extensions such as XLM-R and XLNet [10] have shown competitive performance across diverse languages. Incorporating task-specific attention layers in these settings could help overcome the challenges of semantic divergence in multilingual data.
7) Parameter-efficient training techniques, such as adapters and low-rank factorization, have gained traction as alter- natives to full fine-tuning [17]. Combining these strate- gies with task-specific attention could provide a pathway to balance efficiency and performance, especially in resource-constrained environments.
8) A more detailed evaluation of robustness and fairness is warranted. Since attention mechanisms may emphasize certain input features more strongly, future experiments should assess whether biases present in training corpora are amplified or mitigated when task-specific attention layers are introduced. Addressing these issues will be critical for responsible deployment in real-world NLP applications.

IV. CONCLUSION

It has been demonstrated that task-specific attention is capable of efficiently aggregating information from the encodings of all tokens within a document. It is believed that this method

Received: August 02, 2025

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

is worthy of consideration when constructing applications based on foundation models, particularly when a substantial amount of fine-tuning data is available. It is emphasized that task-specific attention operates as a downstream mechanism that is applied after the fine-tuning of the foundation model has been completed. Therefore, it is recommended that researchers and engineers prioritize the selection of an appropriate foundation model and the design of effective task loss functions as part of the upstream process. Once these upstream tasks have been optimized, the task-specific attention mechanism may be employed as an additional enhancement to capture any residual information that remains in the foundation model's outputs. In addition to the empirical contributions, this study also highlights the importance of designing modular enhancements that can adapt to evolving NLP architectures. While large-scale pretrained models such as RoBERTa [11], ALBERT [12], and GPT-3 [15] continue to push the boundaries of performance through extensive pretraining and scaling, the approach presented here emphasizes task-level customization. This complementary perspective suggests that future progress in NLP may rely not only on building larger foundation models but also on incorporating lightweight, interpretable mechanisms that adapt these models to specific application domains.

Moreover, the findings demonstrate that performance improvements are not uniform across all tasks, indicating that task-specific attention is most beneficial in scenarios requiring fine-grained semantic alignment, such as paraphrase detection. For simpler classification tasks, the marginal benefits suggest that practitioners should weigh the additional computational cost against the expected gains. This nuanced understanding provides valuable guidance for real-world deployments, where efficiency, fairness, and robustness must be balanced with accuracy.

Ultimately, this research contributes to the broader dialogue on how foundation models can be extended, specialized, and optimized. By demonstrating that even modest architectural modifications at the fine-tuning stage can yield substantial improvements, the study reinforces the idea that innovation in NLP is not solely dependent on model size, but also on the creativity and precision of downstream adaptation strategies.

Acknowledgment

The author gratefully acknowledges the support of Rajiv Gandhi University for providing computational resources and research facilities.

References

[1] Rajpurkar, Pranav, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD—Association for Computational Linguistics (ACL), 2018.

[2] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzma´n, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the ACL, 2020, pp. 8440–8451.

Received: August 02, 2025

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

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

[4] Malhotra, Shubham. Self-Organising Interaction Spaces: A Framework for Engineering Pervasive Applications in Mobile and Distributed Environments. 2025. arXiv:2502.01137. Available at: https://arxiv.org/abs/ 2502.01137.

[5] Cherukuri, Milind. Comparing Image Segmentation Algorithms. Proceedings of the 2024 IEEE 4th International Conference on Data Science and Computer Application (ICDSCA), 2024, pp. 266–269. doi: 10.1109/ ICDSCA63855.2024.10859911.

[6] Awasthi, Anurag. Leveraging GANs For Active Appearance Models Optimized Model Fitting. 2025. arXiv:2501.11218. Available at: https://arxiv.org/abs/ 2501.11218.

[7] Vadher, Vandan. Optimizing Industrial Symbiosis: Spatial Impacts on Circular Economy Efficiency. 2024, December. (Proceedings details not provided).

[8] Patel, Niraj. Enhancing Regression Diagnostics: Automated Residual Analysis Using Computer Vision and Statistical Insights. International Journal of innovative Science and Research Technology (IJISRT), Volume 10, Issue 2, February 2025, pp. 1421–1430. doi:https://doi.org/10.5281/zenodo.14964344. Available at: https://www.ijisrt.com.

[9] Vaswani, Ashish, et al. Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 2017.

[10] Yang, Zhilin, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Advances in Neural Information Processing Systems (NeurIPS), 2019.

[11] Liu, Yinhan, et al. RoBERTa: A Robustly Optimised BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, 2019.

[12] Lan, Zhenzhong, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations International Conference on Learning Representations (ICLR), 2020.

[13] Clark, Kevin, et al. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. International Conference on Learning Representations (ICLR), 2020.

[14] Raffel, Colin, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

[15] Brown, Tom, et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS), 2020.

[16] Sun, Chi, et al. How to Fine-Tune BERT for Text Classification? China National Conference on Chinese Computational Linguistics. Springer, 2019.

[17] Houlsby, Neil, et al. Parameter-Efficient Transfer Learning for NLP. International Conference on Machine Learn ing (ICML), 2019.

[18] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. Sim CSE: Simple Contrastive Learning

Volume 38 No. 5s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

of Sentence Embeddings. Empirical Methods in Natural Language Processing (EMNLP), 2021.

[19] Peters, Matthew E., et al. Deep Contextualized Word Representations. NAACL-HLT, 2018.

[20] Mikolov, Tomas, et al. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781, 2013

Received: August 02, 2025