International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Transformer Models: Key Methodologies, Next Sentence Prediction, GLUE Benchmark, and Transfer Learning

Pratap Singh Barth[1], Dhanroop Mal Nagar[2]

Submitted: 10/04/2020 **Revised:** 05/05/2020 **Accepted:** 20/05/2020

Abstract: This research work undertakes a comprehensive examination of the nascent yet rapidly evolving landscape of Transformer-based models in Natural Language Processing. In this research work, the architectural innovations that define this paradigm shift are delved into, particularly highlighting the efficacy of the attention mechanism as a core computational unit, which has allowed for unprecedented parallel processing and contextual understanding in sequence modeling (Vaswani et al., 2017). The central subject of this investigation is the Bidirectional Encoder Representations from Transformers (BERT), a landmark model introduced in 2018, which leverages the Transformer architecture to achieve deep bidirectional representations of language (Devlin & Chang, 2018).

This study critically analyzes BERT's dual pre-training objectives: **Masked Language Modeling**, designed to foster a rich contextual understanding by predicting occluded tokens, and **Next Sentence Prediction**, a novel task aimed at equipping the model with the ability to discern relationships between sentence pairs, crucial for discourse-level comprehension. This research work further assesses the instrumental role of the **General Language Understanding Evaluation benchmark**, established in 2018, as a standardized and challenging suite of tasks that has significantly driven progress and enabled robust comparison across diverse language understanding systems (Wang et al., 2018a, 2018b). Through this lens, the **transfer learning** paradigm, exemplified by BERT's pre-train and fine-tune approach, has revolutionized NLP by enabling state-of-the-art performance across numerous downstream tasks with minimal task-specific data. This paper illuminates how these interconnected methodological pillars collectively facilitate the generation of highly versatile and robust pre-trained language representations, fundamentally reshaping the trajectory of natural language understanding research and application.

1. Introduction

The period leading up to 2019 has witnessed a revolutionary shift in Natural Language Processing, largely driven by the introduction of Transformer-based models. A pivotal development during this time was the Bidirectional Encoder Representations from Transformers, which significantly advanced the state of the art in language understanding (Devlin & Chang, 2018). Unlike earlier models that processed language sequentially, BERT was specifically designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts across all layers of its architecture (Devlin et al., 2019; Devlin & Chang, 2018).

The foundation for these advancements was laid in 2017 with the proposal of the Transformer architecture, which uniquely relied entirely on attention mechanisms, thereby dispensing with recurrent and convolutional networks (Vaswani et al., 2017). This innovation allowed for greater parallelism in training and achieved superior performance in tasks such as machine translation. BERT, building upon this Transformer architecture, further solidified the paradigm of pre-training followed by fine-tuning, demonstrating remarkable success across a wide range of NLP tasks including question answering and language inference (Devlin & Chang, 2018). This paper will delve into the key methodological pillars of Transformer models like BERT, explore the role of Next Sentence Prediction, discuss the significance of the GLUE benchmark in evaluating these models, and highlight the impact of transfer learning during this transformative period.

2. Key Methodology Pillars

The remarkable advancements in Transformer-based models, particularly BERT, are rooted in two primary methodological pillars: the **attention mechanism** and **deep bidirectional representations** learned through novel pre-training objectives.

The foundational innovation is the **Transformer architecture**, introduced in 2017, which entirely abandoned recurrence and convolutions in favor of self-

[1]*Assistant Professor, CSE Dept., Engineering College Bikaner*
Pratapcharan1985@gmail.com
[2]*Assistant Professor, IT Dept., Engineering College Bikaner*
dhanroopmalnagar@gmail.com

attention mechanisms (Vaswani et al., 2017). At its core, the attention mechanism allows the model to dynamically weigh the importance of different words in an input sequence when processing a specific word. This capability is critical for capturing long-range dependencies within text, a challenge for previous sequential models (Vaswani et al., 2017).

Specifically, the **Scaled Dot-Product Attention** computes an output as a weighted sum of "value" vectors, where the weight assigned to each value is determined by the dot-product compatibility of the "query" with all "key" vectors. Mathematically, for a query matrix Q, key matrix K, and value matrix V, the attention function is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, d_k is the dimension of the key vectors, which serves to scale the dot products to prevent the softmax function from having extremely small gradients. The matrices Q, K, and V are derived from the input embeddings through linear transformations using learned weight matrices, e.g., $Q = XW^Q$, $K = XW^K$, $V = XW^V$, where X is the input matrix and W^Q, W^K, W^V are projection matrices (Vaswani et al., 2017).

Building upon this, the Transformer employs **Multi-Head Attention**, which allows the model to jointly attend to information from different representation subspaces at different positions (Vaswani et al., 2017). Instead of performing a single attention function, the queries, keys, and values are linearly projected h times with different learned linear projections to d_k, d_k, and d_v dimensions, respectively. For each of these projected versions, the attention function is performed in parallel, yielding h attention outputs. These outputs are then concatenated and once again linearly transformed to produce the final result:

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

$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Here, W^Q_i, W^K_i, W^V_i are parameter matrices for the i-th head, and W^O is the output projection matrix. This multi-head approach enables the model to capture a richer and more diverse set of relationships within the input sequence.

BERT, introduced in 2018, leverages this powerful Transformer architecture, specifically its encoder stack, to create **deep bidirectional representations**(Devlin et al., 2019; Devlin & Chang, 2018). Unlike earlier language models that processed text in a unidirectional manner (either left-to-right or right-to-left), BERT is designed to jointly condition on both the left and right context in all layers of the model (Devlin & Chang, 2018). This is a crucial distinction, as it allows for a more comprehensive understanding of word meanings based on their full surrounding context.

To achieve this bidirectionality during pre-training on unlabeled text, BERT employs a novel objective called **Masked Language Modeling**(Devlin & Chang, 2018). Instead of predicting the next word in a sequence, 15% of the input tokens are randomly masked, and the model's task is to predict the original vocabulary ID of these masked tokens based on their context. This forces the model to integrate information from both directions. The masking procedure is implemented as follows: for the selected 15% of tokens, 80% are replaced with the special [MASK] token, 10% are replaced with a random token from the vocabulary, and 10% are left unchanged. The model then predicts the original tokens for all masked positions. This approach enables the pre-training of a truly bidirectional model, leading to significantly improved contextual understanding compared to models that only consider unidirectional context (Devlin & Chang, 2018).

3. Next Sentence Prediction

Next Sentence Prediction was introduced as a crucial pre-training objective for BERT to specifically enhance its ability to understand relationships between sentences (Devlin & Chang, 2018). Traditional language models, particularly those relying on unidirectional context, often struggled with tasks that required reasoning across multiple sentences, such as Natural Language Inference and Question Answering (Devlin & Chang, 2018). The NSP task was designed to equip BERT with the capacity to model discourse coherence and inter-sentence semantic connections, thereby improving its performance on these downstream tasks (Devlin & Chang, 2018; Shi & Demberg, 2019).

During the pre-training phase, BERT is exposed to a vast corpus of unlabeled text. For the NSP task, the model is presented with pairs of sentences, denoted as Sentence A and Sentence B. To create the training data, 50% of the time, Sentence B is the actual next sentence that immediately follows Sentence A in the original document from the corpus. For the remaining 50% of the training instances, Sentence B is a random sentence sampled from a different document, ensuring it is logically disconnected from Sentence A. The model's objective is then to predict whether Sentence B is indeed the subsequent sentence or a randomly chosen one (Devlin & Chang, 2018). This effectively frames NSP as a binary classification problem.

To enable this, the input format to BERT for the NSP task is carefully structured. A special classification token [CLS] is prepended to the input sequence, and the two sentences are separated by another special token [SEP].

Additionally, a segment embedding is added to each token, indicating whether it belongs to Sentence A or Sentence B. For instance, all tokens in Sentence A and the first [SEP] receive a segment embedding 0, while all tokens in Sentence B and its [SEP] token receive a segment embedding 1. This unique input representation allows the model to differentiate between the two sentences and understand their positional relationship (Devlin & Chang, 2018; Fisch et al., 2019).

The output corresponding to the [CLS] token's final hidden state is then fed into a simple feed-forward layer, which is followed by a softmax function, to predict the IsNext or NotNext label (Devlin & Chang, 2018). The loss for this binary classification is calculated using a standard cross-entropy function:

$$L_{NSP} = -\sum_{i} \left[y^{i}_{NSP} \cdot \log\left(P^{i}_{NSP}\right) + \left(1 - y^{i}_{NSP}\right) \cdot \log\left(1 - P^{i}_{NSP}\right) \right]$$

Where y^{i}_{NSP} is the true label (1 for IsNext, 0 for NotNext) for the i-th sentence pair, and P^{i}_{NSP} is the model's predicted probability that the pair is IsNext. The representation derived from the [CLS] token after this pre-training captures essential information about the relationship between the two input sentences, which has proven highly beneficial for a range of tasks requiring inter-sentence understanding, such as natural language inference and question answering benchmarks (Devlin & Chang, 2018; Papanikolaou et al., 2019; Shi & Demberg, 2019). The ability to effectively model these inter-sentence dependencies was a significant factor in BERT's state-of-the-art performance across numerous NLP tasks (Devlin & Chang, 2018).

4. Significance of GLUE Benchmark

The **General Language Understanding Evaluation benchmark**, launched in 2018, played a critical role in the advancement and standardized evaluation of general-purpose language understanding models. GLUE is a collection of nine diverse Natural Language Understanding tasks designed to assess how well models can acquire and leverage linguistic knowledge across various domains and difficulties (Wang et al., 2018a, 2018b). These tasks cover a broad range of NLU phenomena, including natural language inference, sentiment analysis, and similarity judgments (Wang et al., 2018).

The benchmark's significance lies in its ability to facilitate principled evaluation and comparison of different models, promoting the development of unified models capable of handling a spectrum of linguistic tasks. It specifically favors models that can represent linguistic knowledge in a way that enables sample-efficient learning and effective knowledge-transfer across tasks, especially given that some GLUE tasks have limited training data (Wang et al., 2018). Following its release, GLUE quickly became a widely adopted platform for evaluating the performance of new language models, including BERT. Models like BERT demonstrated state-of-the-art performance on GLUE tasks, showcasing the efficacy of their underlying architectures and pre-training strategies in achieving robust language understanding.

5. Transfer Learning

The concept of **transfer learning** revolutionized NLP during this period, with BERT standing out as a prominent example. Transfer learning in this context involves two main stages: **pre-training** and **fine-tuning**. In the pre-training phase, a large language model like BERT is trained on vast amounts of unlabeled text data using self-supervised objectives such as Masked Language Modeling and Next Sentence Prediction. This process allows the model to learn a rich, general-purpose understanding of language, capturing semantic and syntactic relationships without explicit supervision for specific tasks (Devlin & Chang, 2018).

The pre-training of BERT involves minimizing a combined loss function, L_{BERT}, which is the sum of the Masked Language Modeling loss, L_{MLM}, and the Next Sentence Prediction loss, L_{NSP}(Devlin & Chang, 2018):

$$L_{BERT} = L_{MLM} + L_{NSP}$$

For the Next Sentence Prediction task, the model predicts whether a second sentence logically follows the first. This is typically formulated as a binary classification problem, and the loss is calculated using a standard cross-entropy function (Chaabouni, 2017):

$$L_{\{NSP\}} = -\sum_{i} \left(y^{i}_{NSP} \right) \cdot \log\left(P^{i}_{NSP}\right)$$

where P_{NSP} represents the predicted probability of the sentence relationship, often obtained after a softmax activation on the output of a classification layer applied to the [CLS] token's representation:

$$P_{NSP} = \text{softmax}\left(h_{CLS} \cdot W_{NSP} + B_{NSP}\right)$$

Here, h_{CLS} is the hidden state corresponding to the [CLS] token, W_{NSP} is a weight matrix, and B_{NSP} is a bias term.

Similarly, for the Masked Language Modeling task, where the model predicts masked tokens based on their context, a cross-entropy loss is also employed (Chaabouni, 2017):

$$L_{MLM} = -\sum_{j} \sum_{k} y^{j,k}_{MLM} \cdot \backslash\log\left(P^{j,k}_{MLM}\right)$$

where P_{MLM} represents the predicted probability distribution over the vocabulary for the masked tokens, typically derived from the hidden states of the masked tokens:

$$P_{MLM} = \text{softmax}(h_{masked} \cdot W_{MLM} + B_{MLM})$$

Once pre-trained, the BERT model can then be **fine-tuned** for various downstream NLP tasks with only a small amount of task-specific labeled data and minimal architectural modifications, typically involving adding a single output layer (Devlin & Chang, 2018; Sun et al., 2019). The fine-tuning process involves optimizing the model parameters, starting from the pre-trained weights $\hat{\theta}_o$, to minimize a task-specific loss function, $\Lambda(F;\ \theta)$, over a new dataset F(Sun et al., 2019). This is commonly achieved using variants of stochastic gradient descent, such as Adam (Kingma & Ba, 2014), where parameters are iteratively updated by:

$$\hat{\theta}_k \leftarrow \hat{\theta}_{k-1} - \alpha\nabla\Lambda(B;\ \hat{\theta}_{k-1})$$

Here, $\hat{\theta}_k$ *are the model parameters at iteration* k, α *is the learning rate, and* $\alpha\nabla\Lambda(B;\ \hat{\theta}_{k-1})$ is the gradient of the loss function calculated on a batch B of the task-specific data (Sun et al., 2019). This approach drastically reduces the need for large, task-specific labeled datasets, making it feasible to achieve high performance across a wide array of applications, from text classification to question answering (Devlin & Chang, 2018; Sun et al., 2019). The effectiveness of transfer learning, as demonstrated by BERT, lies in its ability to leverage the universal language representations learned during pre-training, transferring this acquired knowledge to new tasks and significantly boosting their performance. Some early advancements also explored parameter-efficient fine-tuning methods, where only a small fraction of parameters are updated for each task, enhancing efficiency (Houlsby et al., 2019).

6. Future Directions

As of 2019, the success of Transformer-based models and the transfer learning paradigm opens up several exciting avenues for future research in Natural Language Processing. While models like BERT have demonstrated unprecedented performance, many challenges and opportunities remain.

One significant area of focus is the **computational efficiency** of these large models. The quadratic complexity of the attention mechanism with respect to sequence length limits their application to very long texts. Future work will likely explore more efficient attention mechanisms, including sparse attention (Child et al., 2019) and new architectures like Transformer-XL that better handle long-term dependencies through recurrence and novel positional encoding schemes (Dai et al., 2019a, 2019b). Reducing the memory and computational footprint of pre-training and fine-tuning will be critical for broader applicability.

Another crucial direction involves deeper investigations into **model interpretability and fairness**. The internal workings of these complex neural networks are often opaque, making it difficult to understand why a model makes a particular prediction (Belinkov & Glass, 2019; Doshi-Velez & Kim, 2017). Research will aim to develop better tools and methodologies to interpret the learned representations and attention patterns, which could lead to more robust and trustworthy NLP systems (Belinkov & Glass, 2019). Furthermore, the pervasive nature of language models necessitates addressing potential **biases** encoded within the training data, which can inadvertently lead to unfair or discriminatory outcomes (Bolukbasi et al., 2016; Caliskan et al., 2016; Chang et al., 2019; Solaiman et al., 2019). Developing strategies to identify, measure, and mitigate these biases will be paramount.

Expanding the applicability of pre-trained models to **low-resource languages and domains** is another key objective. While current state-of-the-art models largely benefit from vast amounts of English text data, creating similar advancements for languages with less digital presence or specialized technical domains remains a challenge. Techniques for cross-lingual transfer learning and domain adaptation will be vital here.

Finally, while the pre-train and fine-tune paradigm has been immensely successful, there is ongoing exploration into **alternative transfer learning strategies** that might offer greater flexibility or efficiency. This includes investigating methods for more selective fine-tuning, knowledge distillation, and prompt-based learning. The ultimate goal is to continue pushing the boundaries of what machines can understand and generate in human language, always striving for more intelligent, efficient, and ethical AI systems.

7. Conclusion

The period from 2016 to 2019 has been transformative for Natural Language Processing, largely owing to the development and widespread adoption of Transformer models, culminating in architectures like BERT (Devlin & Chang, 2018). These models, fundamentally built upon the self-attention mechanism, moved beyond the limitations of sequential processing to achieve a deeper, bidirectional understanding of linguistic context (Vaswani et al., 2017). The key methodological pillars, including the Transformer architecture and the dual pre-training objectives of Masked Language Modeling and Next Sentence Prediction, were instrumental in enabling models to learn rich, generalized language

representations (Devlin & Chang, 2018; Vaswani et al., 2017).

The GLUE benchmark played a critical role in standardizing the evaluation of these models, pushing the field towards developing more robust and versatile language understanding systems capable of performing across diverse tasks (Wang et al., 2018a, 2018b). The remarkable success of BERT on the GLUE benchmark demonstrated the efficacy of transfer learning, establishing the pre-train and fine-tune paradigm as the dominant approach in NLP. This paradigm significantly reduced the reliance on large, task-specific labeled datasets, making advanced NLP capabilities more accessible and efficient for a wide range of applications (Devlin & Chang, 2018; Sun et al., 2019). The outcomes of these advancements firmly established that the future of NLP lay in continuously improving these powerful, context-aware, and transferable language models, with ongoing research focused on enhancing their efficiency, interpretability, and ethical considerations.

References

[1] Belinkov, Y., & Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, *7*, 49. https://doi.org/10.1162/tacl_a_00254

[2] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1607.06520

[3] Caliskan, A., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora contain human-like biases. *arXiv*. https://doi.org/10.48550/ARXIV.1608.07187

[4] Chaabouni, S. (2017). Study and prediction of visual attention with deep learning net- works in view of assessment of patients with neurodegenerative diseases. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://tel.archives-ouvertes.fr/tel-02408326

[5] Chang, K.-W., Prabhakaran, V. M., & Ordóñez, V. (2019, November 1). Bias and Fairness in Natural Language Processing. *Empirical Methods in Natural Language Processing*. https://aclanthology.org/D19-2003/

[6] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1904.10509

[7] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019a). *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. https://doi.org/10.48550/ARXIV.1901.02860

[8] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019b). *Transformer-XL: Attentive Language Models beyond a Fixed-Length Context*. 2978. https://doi.org/10.18653/v1/p19-1285

[9] Devlin, J., & Chang, M. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Leibniz-Zentrum Für Informatik (Schloss Dagstuhl)*. https://doi.org/10.48550/arxiv.1810.04805

[10] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). 4171. https://doi.org/10.18653/v1/n19-1423

[11] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1702.08608

[12] Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., Chen, D., Berant, J., Srikumar, V., Chen, P., Linden, A. V., Harding, B., Kembhavi, A., Schwenk, D., Choi, J., Farhadi, A., Kwiatkowski, T., Palomaki, J., Collins, M., Parikh, A. P., … Herledan, F. (2019). *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. https://doi.org/10.18653/v1/d19-58

[13] Houlsby, N., Giurgiu, A., Jastrzębski, S., Morrone, B., Laroussilhe, Q. de, Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1902.00751

[14] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1412.6980

[15] Papanikolaou, Y., Roberts, I., & Pierleoni, A. (2019). *Deep Bidirectional Transformers for Relation Extraction without Supervision*. https://doi.org/10.18653/v1/d19-6108

[16] Shi, W., & Demberg, V. (2019). *Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains*. https://doi.org/10.18653/v1/d19-1586

[17] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Wang, J., Wook, K., Jong, Sarah, K., Miles, M., Alex, N., Jason, B., Kris, M., & Jasmine, W. (2019). Release Strategies and the Social Impacts of Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1908.09203

[18] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1905.05583

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

[20] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018a). *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. 353. https://doi.org/10.18653/v1/w18-5446

[21] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018b). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1804.07461