International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 12 - Issue 01 || January 2026 || PP. 01-03

A Weighted Self-Attention Optimization Framework for Transformer-Based Text Classification

Sambit Ray
Independent Researcher
Sambitray86@gmail.com

Abstract: Transformer models have significantly advanced Natural Language Processing (NLP) by replacing recurrence with self-attention mechanisms [1]. While standard Transformers compute attention uniformly across heads and tokens, this may dilute task-specific importance in classification problems [2,8]. This paper proposes a Weighted Self-Attention Optimization (WSAO) framework that introduces adaptive token-level weighting into the Transformer encoder to enhance discriminative feature learning. Using a public benchmark dataset, we demonstrate that the proposed formulation improves classification performance compared to baseline Transformer [1] and LSTM models [3]. Mathematical formulation, experimental evaluation, and comparative analysis are presented to highlight the effectiveness of the approach.
Keywords: Transformer Models, Self-Attention Optimization, Text Classification, Cross-Entropy Loss, NLP

1. Introduction

Text classification is a fundamental task in Natural Language Processing, with applications in spam detection, sentiment analysis, and topic categorization **[4]**. Traditional deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks process sequences sequentially, limiting their ability to model long-range dependencies efficiently **[3]**.

The Transformer architecture overcomes these limitations by employing self-attention mechanisms, enabling parallel computation and global contextual understanding **[1]**. However, in standard Transformers, attention weights are computed solely based on dot-product similarity, which may not optimally reflect token importance for downstream classification tasks **[8]**.

This paper introduces a Weighted Self-Attention Optimization (WSAO) mechanism that enhances Transformer representations by incorporating learnable importance weights. The main contributions of this work are:

- A mathematical formulation of adaptive attention weighting
- A dynamic comparison with baseline models **[1,3,10]**
- An empirical evaluation using real-world text data

2. Related Work

Early statistical NLP approaches relied on bag-of-words and n-gram models [4]. Neural models such as RNNs and LSTMs improved contextual modeling but suffered from vanishing gradients and limited parallelism [3].

Attention mechanisms addressed these issues by allowing models to focus selectively on relevant tokens [7]. Vaswani et al. introduced the Transformer, which uses multi-head self-attention as its core component [1]. Subsequent works such as BERT and GPT demonstrated the effectiveness of large-scale pretraining for downstream NLP tasks [2,6].

Recent studies have explored attention refinement and weighting strategies to improve task-specific performance [8], motivating the proposed WSAO framework.

3. Mathematical Foundation of Transformer Attention

3.1 Standard Self-Attention

Given an input sequence embedding matrix

$$X \in R^{n \times d}$$

Query, key, and value matrices are computed as:

$$Q=XWQ, \quad K=XWK, \quad V=XWV$$

The scaled dot-product attention is defined as:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

$$Z = AV$$

Where Zrepresents the contextualized token embedding **[1]**.

3.2 Proposed Weighted Self-Attention Optimization (WSAO)

To emphasize task-relevant tokens, we introduce a learnable weight vector

$$a \in R^n$$

Representing token importance:

$$\alpha = softmax(W_\alpha X)$$

The optimized attention output is computed as:

$$Z_{opt} = (\alpha \odot A)V$$

Where $\odot$ denotes element-wise multiplication. Similar attention-weighting concepts have been explored in structured self-attention literature **[8]**, but are adapted here for Transformer-based classification.

4. Methodology

4.1 Dataset Description

The SMS Spam Collection Dataset is used for evaluation, containing 5,574 labelled messages categorized as *spam* or *ham* **[9]**.

Text Message	Label
"Win a free ticket now!"	Spam
"Are we meeting today?"	Ham

4.2 Preprocessing Steps

- Text normalization
- Tokenization
- Padding to fixed sequence length
- Vocabulary indexing **[4]**

4.3 Model Architecture

- Token embedding with positional encoding **[1]**
- Transformer encoder with WSAO
- Global average pooling
- Fully connected Softmax classifier

4.4 Loss Function

The classification objective minimizes categorical cross-entropy loss:

$$\boldsymbol{L = -\sum_{i=1}^{N}\sum_{k=1}^{C} y_{ik}\log(\widehat{y_{ik}})}$$

Where y_{ik} is the true label and $\widehat{y_{ik}}$ is the predicted probability

5. Experimental Results

5.1 Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score

5.2 Comparative Performance

Model	Accuracy	F1-score
LSTM	91.4%	0.91
Standard Transformer	94.2%	0.94

Model	Accuracy	F1-score
LSTM	91.4%	0.91
Proposed WSAO Transformer	96.1%	0.96

The proposed approach demonstrates consistent improvement across all metrics when compared with baseline architectures

6. Discussion

The experimental results indicate that incorporating adaptive attention weights improves the Transformer's ability to focus on semantically relevant tokens. Similar observations have been reported in attention refinement studies **[8]**. The approach remains computationally efficient and scalable to other NLP classification tasks.

7. Conclusion and Future Work

This paper presented a Weighted Self-Attention Optimization framework for Transformer-based text classification. By introducing learnable token importance weights, the model achieved improved performance over baseline architectures **[1,3]**.

Future research directions include:

- Extension to multi-class datasets
- Integration with pre-trained Transformer models such as BERT **[2]**
- Analysis of attention interpretability

Ethical and Safety Disclaimer

This study is intended for academic research and educational purposes only. The proposed methods should not be deployed in sensitive or real-world decision-making systems without rigorous validation and ethical review.

8. References

[1]. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
[2]. Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT.
[3]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[4]. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Pearson.
[5]. Goldberg, Y. (2017). Neural network methods in natural language processing. Morgan & Claypool.
[6]. Brown, T., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
[7]. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR.
[8]. Lin, Z., et al. (2017). A structured self-attentive sentence embedding. ICLR.
[9]. Almeida, T. A., & Hidalgo, J. M. G. (2011). SMS Spam Collection Dataset. UCI Machine Learning Repository.
[10]. Kim, Y. (2014). Convolutional neural networks for sentence classification. EMNLP.
[11]. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. EMNLP.
[12]. Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. NIPS.