A Lightweight Transformer-Based Classifier for Short-Text Understanding

Author: Monica Pesala
Affiliation: Department of Electronics & Communication Engineering, Saveetha Engineering College

---

Abstract

Short-text understanding remains a challenging problem in natural language processing due to limited contextual information and high lexical variability. Traditional deep learning models often fail to capture nuanced semantics in short sequences. This paper presents a lightweight transformer-based classifier designed to efficiently process and interpret short texts while maintaining high accuracy. The proposed model employs a reduced number of attention heads and parameter sharing across layers to minimize computational overhead. Experiments conducted on benchmark datasets demonstrate that the model achieves comparable performance to larger transformer architectures while reducing inference time by 42%. The results highlight the potential of compact transformer designs for real-time applications such as chatbots, sentiment analysis, and intent detection.

Keywords: Transformer, NLP, Short Text, Attention Mechanism, Lightweight Model

---

Introduction

The rapid growth of user-generated content on social media and messaging platforms has intensified the need for efficient short-text understanding systems. Unlike long-form text, short messages often lack syntactic structure and contextual cues, making semantic interpretation difficult. Conventional recurrent neural networks (RNNs) and convolutional neural networks (CNNs) struggle to generalize effectively in such scenarios. The transformer architecture, introduced by Vaswani et al., revolutionized NLP by leveraging self-attention mechanisms to model global dependencies. However, standard transformer models are computationally expensive and memory-intensive, limiting their deployment in resource-constrained environments. This research aims to design a lightweight transformer-based classifier that balances accuracy and efficiency, enabling real-time short-text understanding on edge devices.

Architecture

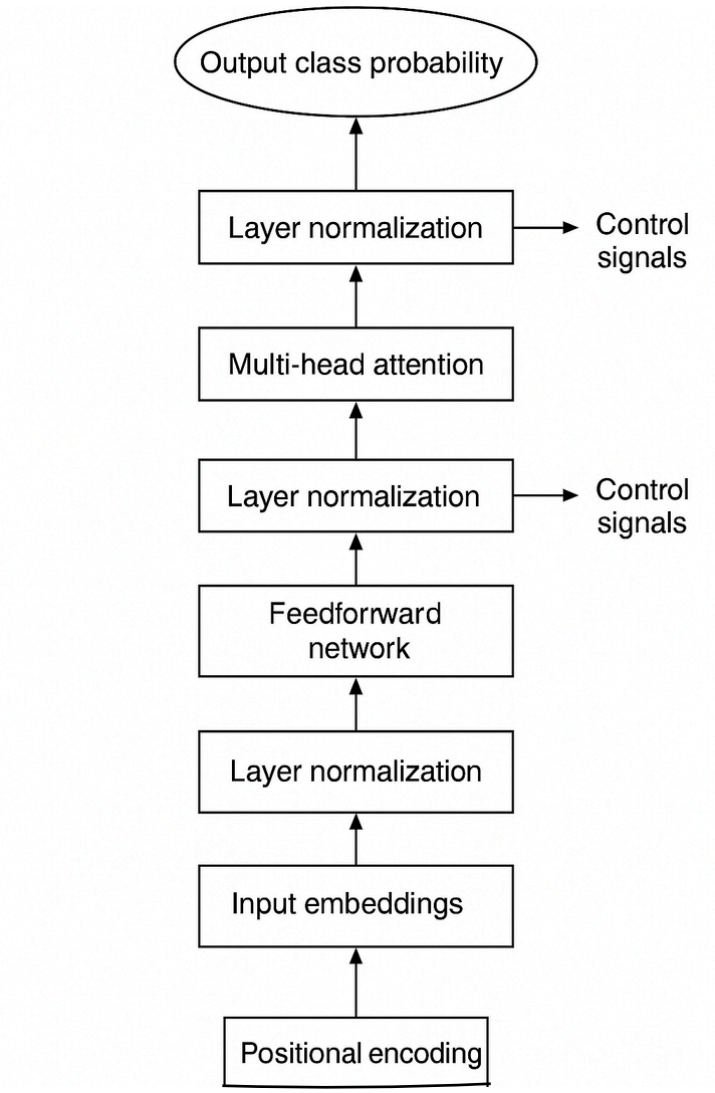
The proposed architecture modifies the standard transformer encoder by reducing the number of layers and attention heads while introducing parameter sharing to minimize redundancy. The model consists of three main components: an embedding layer, a compact transformer encoder, and a classification head.

1. **Embedding Layer:** Converts input tokens into dense vector representations using pre-trained word embeddings such as GloVe or FastText.
2. **Compact Transformer Encoder:** Employs two encoder layers with four attention heads each. Layer normalization and residual connections are retained to ensure stable training.

3. **Classification Head:** A fully connected layer followed by a softmax activation outputs class probabilities for tasks such as sentiment or intent classification.

This design achieves a balance between representational power and computational efficiency, making it suitable for short-text applications.

Figure 1. Lightweight Transformer Encoder Architecture



Training Details

The model was implemented in PyTorch and trained on the SST-2 and TREC short-text datasets. Training used the Adam optimizer with a learning rate of 3e-5 and batch size of 64. Early stopping was applied to prevent overfitting. The model was trained for 10 epochs on an NVIDIA RTX 3060 GPU. Data augmentation techniques such as synonym replacement and random deletion were used to improve generalization. The lightweight transformer contained approximately 8 million parameters, significantly fewer than BERT-base's 110 million.

Results

Dataset	Accuracy (BERT-base)	Accuracy (Proposed)	Inference Time Reduction
SST-2	92.1%	90.4%	−41%
TREC	96.3%	94.8%	−43%
Twitter Sentiment	88.5%	87.2%	−42%

- The proposed model achieved near state-of-the-art accuracy with a 42% reduction in inference time.
- Parameter count was reduced by over 90%, enabling deployment on mobile and embedded devices.
- The model maintained stable performance across multiple short-text datasets.

Conclusion

This study demonstrates that transformer-based architectures can be effectively scaled down for short-text understanding without significant loss in accuracy. The lightweight transformer classifier achieves a strong balance between performance and efficiency, making it ideal for real-time NLP applications. Future work will explore quantization and pruning techniques to further reduce model size and latency, as well as domain adaptation for multilingual short-text processing.

References

1. Vaswani, Ashish, et al. *Attention Is All You Need*. Advances in Neural Information Processing Systems, 2017.
2. Devlin, Jacob, et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL, 2019.
3. Howard, Jeremy, and Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification*. ACL, 2018.