Multi-CLS BERT:
An Efficient Alternative to Traditional Ensembling
Haw-Shiuan Chang*, Ruei-Yao Sun*, Kathryn Ricci*, and Andrew McCallum

Introduction

Background:
• Traditional ensembles of multiple BERT models boost performance on natural language understanding tasks over single models
• However, traditional ensembles are expensive
• Computational cost, memory, space footprint

Research question:
Can we achieve the benefits of ensembling while minimizing the cost?

→ Proposed method:
Ensemble multiple CLS embeddings within a single BERT model

Pretraining

• Learn multiple, diversified CLS embeddings:
→ Adapt state-of-the-art pretraining objectives for BERT (Aroca-Ouellette and Rudzicz, 2020)
• Learn to represent fine-grained semantics:
→ Incorporate hard negatives into the objective

Architecture

• Instead of traditional single BERT CLS embedding, leverage a fixed number of multiple CLS embeddings:
→ Insert K special CLS tokens
→ Use the K final-layer CLS hidden states to represent the input text
→ Aggregate the K CLS embeddings during fine-tuning/inference

• Prevent collapse of the multiple CLS embeddings:
→ Insert linear layers in between selected BERT layers at multi-CLS input positions
→ Add novel parameterization

Main Result

Table 1: Results on GLUE and SuperGLUE for models derived from BERT Base and BERT Large.

Conclusion 1:
Efficient Multi-CLS BERT improves performance over baseline single BERT model and K=1 model with only small increase in model size.

Conclusion 2:
Multi-CLS BERT is especially effective in the few-shot setting.

Analysis

Claim 1: Increased performance is due to ensemble effects.
• Ensemble Multi-CLS BERT only slightly boosts the performance (Table 2)
• Expected calibration error (Table 2)
• Overlap of most uncertain dataset examples (Table 3)
• Qualitative analysis of nearest-neighbor embeddings (Paper appendix): Multiple CLS embeddings can learn to complement in complementary ways to solving a task

Table 2: Qualitative analysis of nearest-neighbor embeddings.

Table 3: Overlap of the 20% most uncertain dataset examples as predicted by the two given models ("ENS vs. ENS").

Conclusion:
The multi-CLS embeddings in a single BERT model with our architecture and pretraining methods improve GLUE and SuperGLUE almost as consistently.

Conclusion:
Using multiple CLS BERT vs. traditional ensemble reduces computational costs at the expense of a modest drop in performance.

Related Work and References

[4] H.-S. Chang, and A. McCallum, "Revisiting the Architectures like Factuality, and Beyond: Multi-CLS BERT with K=1 achieves significantly lower CEC than the same model with K=1, but with very little increase in inference time."

Conclusion

• Using multiple CLS embeddings within a single BERT model with our architecture and pretraining methodologies results in almost free performance gain
• Evidence suggests that performance gains are due to ensemble effects without the cost of a traditional ensemble of multiple models
• Our methods for successfully implementing diversified multiple CLS embeddings may be extensible in future studies of efficient ensembling using other types of architectures

Table 4: Inference time vs. expected calibration error (ECE).

Table 5: Overlap of the 20% most uncertain dataset examples as predicted by the two given models ("ENS vs. ENS").

Table 6: Selected studies for the few-shot setting.

Table 7: Relative to some baseline methods and expected effect of different architectures on GLUE and SuperGLUE.

Figure 1: Proposed method.

Figure 2: Architecture.

Figure 3: Main Result

Figure 4: Related Work and References