

Prompt Ensembles Improve Few-Shot Text Classification Accuracy

C. Inventor and D. Builder

Center for Language Technology, Example Institute

Abstract

Few-shot text classification is sensitive to the wording of prompts. We show that averaging predictions over a small ensemble of manually written prompts yields consistent gains while costing only a linear increase in inference time.

1. Introduction

Prompting a large language model is now the default strategy for low-resource tasks. The choice of prompt, however, remains brittle and is rarely documented in published results.

We investigated whether ensembles of plausible prompts can reduce this brittleness without any additional training data.

2. Approach

For every example we generate one score vector per prompt and average the vectors before applying the decision threshold. No model weights are updated during the entire process.

Prompt Selection

Prompts were written independently by three analysts following a short style guide. The guide constrained prompt length and forbade mention of the test labels.

Aggregation

We experimented with arithmetic and geometric averaging and found the two almost indistinguishable on the development set.

3. Experiments

On six public benchmarks the ensemble improves mean accuracy by 4.1 points in the eight-shot setting. The gain is largest for tasks with high label ambiguity.

4. Discussion

We interpret the gain as variance reduction: single prompts act as strong prior views, and averaging stabilizes the decision boundary. Future work should study adaptive ensembles whose members are pruned per task.

5. Conclusion

Prompt ensembles are a simple, training-free remedy for prompt brittleness. We recommend them as a default reporting practice in few-shot research.

References

1. Brown A, White B. Large language models as few-shot learners. In Proceedings of the Conference on Big Science. 2020.
2. Grey C. A taxonomy of prompting strategies. Language Notes. 2021;5(1):33-49.

3. Hill D, Green E. Ensembling for robustness. Machine Learning Matters. 2022;8(2):12-20.