Scaling Switching Language Models

Anonymous ACL submission

Abstract

The accuracy of simple models with discrete latent dynamics, such as Hidden Markov Models and Probabilistic Context-Free Grammars, scales with the size of the respective latent spaces – up to a point. Recent work has found that scaling discrete latent variable models comes with diminishing returns on accuracy. We explore the hypothesis that discrete representations are not suitable for language modeling, due to language's long-tailed phenomena. We overcome this shortcoming by combining large-scale discrete dynamics with slow-moving continuous state representations, and show that this enables simple models to better capture tail phenomena in language modeling.

1 Introduction

Why should we work on LVMs? Historical reasons. Experiment with different representations.

Recent work in scaling discrete latent variable models has shown that their accuracy scales with size (Chiu and Rush, 2020; Yang et al., 2021). However, those gains diminish as scale increases. Additionally, the computational cost of inference greatly increases with scale. In this report, we examine the shortcomings of these models in a few case studies, and whether we can overcome those shortcomings.

2 Long-Tail Phenomena: Rare Words

We hypothesize that

References

Justin Chiu and Alexander Rush. 2020. Scaling hidden Markov language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1341–1349, Online. Association for Computational Linguistics.

Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021. PCFGs can do better: Inducing probabilistic context-free grammars with many symbols. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1487–1498, Online. Association for Computational Linguistics.

A Example Appendix

This is an appendix.