An Analysis of Language Model Sampling Methods

Michael Hu (myhu)

Background

Language models attempt to predict the next word in a sentence, conditional on words written in the past. The output of the language model is a distribution over possible words to come, and we can then sample from this distribution to generate the next word in the sentence. Deep language models such as OpenAI's GPT-2 can be used in this way to produce text that strongly resembles human writing [1].

Central to generating good text from GPT-2 is choosing a sampling method, for some sampling methods lead to degenerate output. Holtzman et. al. recently demonstrated that probability maximization sampling methods such as beam search or greedy sampling (taking the word with highest probability) often leads the model to repeatedly output the same sentence [2]. The authors of GPT-2 themselves chose to use top-k sampling, a randomized sampling method, where they truncated the sample space to the top 40 most likely words. Complicating matters, there are a myriad of possible ways to sample from language models or perturb them to produce better text; other popular techniques include temperature scaling and diverse beam search [3] [4].

Question

Why do probability maximization sampling methods lead to degenerate text, and why do randomized sampling methods such as top-k sampling lead to (relatively) coherent text?

Method

First, I will demonstrate that the sampling distributions of maximization techniques such as greedy sampling or beam search do not match the learned distribution. I hope to demonstrate that the sampling distribution of top-k sampling closely approximates the learned distribution. The intuition here is that pure maximization disregards the variance of the learned distribution. To the maximizer, there exists no difference between the distributions [0.6, 0.1, 0.1, 0.1, 0.1] and [1, 0, 0, 0, 0]; in both cases, it picks word 1. Thus, maximization sampling methods do not express all the information learned by the model.

Next, I hope to show that words generated from top-k sampling behave in a similar way to human text, in that both humans and top-k sampling choose words that are occasionally unlikely. Here, perhaps I can measure the KL divergence between human sampling probabilities and top-k sampling probabilities, and show that the KL divergence here is smaller than when human sampling probabilties are compared to that of other techniques.

Last, I hope to establish exactly where one should truncate the distribution. Truncating the distribution at the top 40 words a la GPT-2 seems arbitrary, and such a choice cannot be the best or tightest fit for all contexts. I suspect the threshold can instead be chosen by eliminating all words with a probability under a certain parameter δ. I will first determine what this δ is on average when we set $k = 40$, and go from there.

References

[1] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[2] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2019.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[4] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models, 2016.