The curious case of neural text degeneration

Ari Holtzman[1,2], Jan Buys[3], Leo Du[1], Maxwell Forbes[1], and Yejin Choi[1,2]

[1] Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[2] Allen Institute for Artificial Intelligence, Seattle, WA, USA
[3] Department of Computer Science, University of Cape Town, South Africa
{ahai,jbuys,dul2,mbforbes,yejin}@cs.washington.edu

Abstract. Despite considerable advances in neural language modeling, it remains an open question what the best strategy is for generating text from a language model. Counter-intuitively, maximization-based decoding methods such as beam search lead to *degeneration* — output text that is bland, incoherent, or repetitive. We propose Nucleus Sampling, a simple but effective method to draw high quality text out of neural language models by truncating the unreliable tail of the probability distribution, sampling words from the *nucleus* of tokens containing most probability mass. We compare generations from maximization-based and stochastic decoding methods to the distribution of human text along several axes including likelihood, diversity, and repetition. Our results show that (1) maximization is an inappropriate decoding objective for open-ended text generation, (2) the probability distributions of the best current language models have an unreliable tail which needs to be truncated during generation and (3) Nucleus Sampling is the best available decoding strategy for generating long-form text that is both high-quality — as measured by human evaluation — and as diverse as human-written text.

Keywords: Natural Language Generation · Neural Language Models.

On February 14th 2019, OpenAI surprised the scientific community by releasing an impressively high-quality article about Ovid's Unicorn, written by GPT-2, the largest neural language model reported to date [4]. Notably, the best generations obtained from the model relied on *randomness* in the decoding method, in particular through *top-k* sampling that samples the next word from the top k most probable choices $[1, 3, 4]$, instead of aiming to decode text that *maximizes* likelihood. In fact, decoding strategies that optimize output probability, such as greedy or beam search, lead to text that is incredibly degenerate, even when using state-of-the-art models such as GPT-2 (117M parameters), as can be seen in Figure 1. This is counter-intuitive, as one would expect that good models would assign higher probability to more human-like, grammatical text.

We provide novel insights into the shortcomings of existing models and decoding methods for *open-ended* text generation – generating a story or a plausible continuation of a text passage – through novel metrics and analyses. To overcome these shortcomings we introduce *Nucleus Sampling*: The key intuition is

Context: On days when he woke early, the president liked to have coffee in the oval office. There was something about watching the sky grow lighter and lighter as you sat your pajama'd behind in the most powerful chair in the free world, sipping marine strength coffee, that you just couldn't achieve anywhere else.

Greedy: He'd been in the oval office for over a year now, and he'd never been so happy to be there. He'd been in the oval office for almost a year, and he'd never been so happy to be there.

BeamSearch (k=4): The President was a man of his word. He was a man of his word. He was a man of his word. He was a man of his word. He was a man of his word. He was a man of his word.

Sampling (t=1.0): You couldn't be sure if that's what you were really doing, and If you decided to take the day off. The president wanted you to take the day off, but he wanted to maintain a curfew and use his influence wisely.

Top-*k* Sampling (k=8): "I'm sorry, sir." "No, it's okay, son," the president said. "I understand." "You're going to have to make a special trip down there to get that kid. He has no idea where he's going."

Nucleus Sampling (p=0.9): But that wasn't what drew the president's attention. He'd been seated for maybe a minute when he noticed the other man. What was the guy doing here?

Fig. 1. Example text generated from GPT-2 with each of the evaluated decoding strategies. The output is generated conditionally as a continuation of the given text passage ("context").

that the vast majority of probability mass at each time step is concentrated in the *nucleus*, a small subset of the vocabulary that contains most of the plausible next words. Instead of relying on a fixed top-k, or using a temperature parameter to control the shape of the distribution without sufficiently suppressing the unreliable tail distribution (containing the large subset of implausible words), we propose sampling from the top-p portion of the probability mass, expanding and contracting the candidate pool dynamically.

In order to compare current methods to Nucleus Sampling, we compare various distributional properties of generated text to the reference distribution, such as the likelihood of veering into repetition and the perplexity of *generated* text. The latter shows that text generated by maximization or top-k sampling is *too* probable, indicating a lack of diversity and divergence in vocabulary usage from the human distribution. On the other hand, pure sampling produces text that is significantly *less* likely than the human-written reference text, and generation quality is correspondingly lower.

Vocabulary usage and Self-BLEU [5] statistics indicate that high values of k are needed to make top-k sampling match human statistics. Yet, generations in this setting have high variance in likelihood, which is reflected in qualitatively observable incoherencies. Nucleus Sampling can match reference perplexity through a proper value of p. Qualitative analysis shows that text generated by Nucleus Sampling is more coherent than generations from other the decoding strategies (see Figure 1 for example outputs).

Finally, we perform Human Unified with Statistical Evaluation (HUSE) [2] to jointly assess the overall quality and diversity of the decoding strategies, which cannot be captured using either human or automatics evaluation alone. The HUSE evaluation demonstrates that Nucleus sampling is the best overall decoding strategy.

References

1. Fan, A., Lewis, M., Dauphin, Y.: Hierarchical neural story generation. In: Proceedings of the Association for Computational Linguistics (2018)
2. Hashimoto, T.B., Zhang, H., Liang, P.: Unifying human and statistical evaluation for natural language generation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019)
3. Holtzman, A., Buys, J., Forbes, M., Bosselut, A., Golub, D., Choi, Y.: Learning to write with cooperative discriminators. In: Proceedings of the Association for Computational Linguistics (2018)
4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (February 2019), Unpublished manuscript
5. Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., Yu, Y.: Texygen: A benchmarking platform for text generation models. In: ACM SIGIR Conference on Research and Development in Information Retrieval (2018)