Adversarial Decoding of Language Models

Anonymous ACL submission

Abstract

Transformer based language models are relatively recent development in NLP (Vaswani et al., 2017). The speed of progress lead to focus on ethical issues related to deployment of this technology (Solaiman et al., 2019), especially on the risk of massive propaganda creation and so called fake news. In this paper, we focus on relatively unexplored area of decoding text from the model. In particular, we show that fake news discriminators may be insufficient. First we explore impact of sampling parameters on accuracy of discrimination. Then we propose a new adversarial method of decoding. The method is shown to be capable of reliably fooling some fake news discriminators.

1 Introduction

Transformer based language models are relatively recent development in NLP (Vaswani et al., 2017). In that short time this type of architectures (Devlin et al., 2018) (Radford et al., 2019) dominated NLP leader-boards like GLUE (Wang et al., 2018) or SuperGLUE (Wang et al., 2019). The speed of progress lead to focus on ethical issues related to deployment of this technology (Solaiman et al., 2019). NLP community focused on the risk of massive propaganda creation and so called fake news. It has been shown that text generated by state of the art language models is in practice indistinguishable from genuine human made text (Gehrmann et al., 2019). Some researchers pointed out curious properties of current decoding techniques (See et al., 2019) and proposed new ones like Nucleus Sampling (Holtzman et al., 2019). This properties allow for creation of discriminators capable of differentiating between computer generated text and human text (Zellers et al., 2019). In this paper, we focus further on relatively unexplored area of decoding text from the model. In particular, we show that fake news discriminators may be insufficient. First we explore impact of sampling parameters on accuracy of discrimination. Then we propose a new adversarial method of decoding. The method is shown to be capable of fooling some fake news discriminators.

2 Language Model

Language model can be characterized as a probability distribution over a sequence of tokens. Language model allows to query this distribution in order to assign numerical probability to tokens given a sequence of already existing tokens. A language model can be unidirectional or bidirectional. Unidirectional can be conditioned on tokens only from one direction. Bidirectional model can look at both preceding and proceeding tokens. In this paper we will be using both types. Unidirectional language model called GPT-2 created by OpenAI (Radford et al., 2019) will be used to generate text. A fine-tuned bidirectional model called RoBERTA (Liu et al., 2019) will be used to do text discrimination.

3 Language Model Decoding

Language model decoding is an open ended task of generating text with desirable properties like coherence based on token probability distribution provided by a language model. More specifically the tasks is to complete a sequence of t tokens $x_1, x_2 \ldots x_t$ called **context** or **prompt** with a generated sequence of d tokens $x_{t+1}, x_{t+2} \ldots x_{t+d}$ given token probability distribution conditioned on the prompt $P(x_{t+1}|x_1, x_2 \ldots x_t)$.

3.1 Naive Sampling

Naive approach would be to sample directly from the distribution. Unfortunately, that approach has been shown to lead to serious text degeneracy like falling into infinite loops of repeating the same n-

gram (Holtzman et al., 2019). The same issues can be observed if one tries to maximize the likelyhood of the generated sequence.

There are currently at least two popular approaches to remedy this issues. Both of them are based on the idea of truncating the probability distribution.

3.2 Top-k Sampling

Top-k sampling is done by truncating the probability distribution to **the top k** tokens. In order to assure that the sum of probabilities is equal to 1 the top k distribution is re-normalized.

3.3 Top-p Sampling

Top-p sampling builds on the top-k sampling. Selecting appropriate k number is difficult because appropriate k may depend on context. Selecting too high value for a given context may lead to degeneracy, while selecting too low value may lead to a trivial text. The top-p sampling is trying to avoid that issue by always truncating the tokens assigned at least **the top p** probability mass.

4 Adversarial Setup

Before introducing adversarial decoding we need to look closer at the adversarial setup. Solaiman et al. (2019) and Zellers et al. (2019) proposed discriminators capable of differentiating between computer generated text and human generated text by assigning them a numerical score between 0 and 1. Where 1 indicates computer generated text or vice-versa. We evaluate one discriminator (Solaiman et al., 2019) created by OpenAI by fine tuning a RoBERTA model (Liu et al., 2019) and available in the Huggingface Transformer library (Wolf et al., 2019). We will refer to that model as RoBERTA discriminator. The other model is called Grover, whose authors claim better accuracy than RoBERTA based discriminator. Unfortunately, Grover is not readily available and requires considerable amount of computational resources to recreate. Therefore, it is not evaluated in this paper.

4.1 The Discriminator Evaluation

The RoBERTA discriminator is evaluated in the figure 1 and the figure 2. In both cases the model was fed text generated by computer and asked to do the discrimination.

The figure 1 provides evaluation on both baseline methods of sampling with varying p and k values. 100 sequences of 30 tokens were generated for each evaluated value of p and k. The discriminator archives the best accuracy of around 95% on low values of p and k when the probability distributions are significantly truncated, therefore providing the highest bias. Top-k sampling appears to be more difficult to classify for high values of k like 150 or 200 achieving only 80% accuracy.

The figure 2 provides evaluation of both top-k ($k = 150$) and top-p ($p = 0.99$) baseline methods of sampling based on varying length of sequences. We can see that sequences of length 10 are the most difficult to classify where the discriminator achieves only little bit more than 60% accuracy. The performance steadily improves as the sequence length gets longer for both sampling methods. The best accuracy is around 95% for sequences of length 100. The top-k sampling appears to be consistently little bit more difficult to classify across the different sequence lengths.

The difference apparent higher difficulty of classifying text decoded with top-k sampling is probably an artifact of the training procedure of the discriminator as the discriminator was trained on text generated with top-p sampling.

5 Adversarial Decoder

The adversarial decoder is utilizing the RoBERTA discriminator in order to select the most human like sequences. The sequences can be generated by a baseline decoder. In this case, the top-k decoder was used with $k = 150$ as that configuration was the most difficult to classify for the discriminator.

In the adversarial decoder decoding happens in steps. Assume that sequence of length n needs to be generated. This can be split into m steps, so that in each step a baseline decoder needs to generate n/m tokens. The baseline decoder can be used to generate j different sequences of length n/m at each step. Before each step is finished, the discriminator will be used to select the most human-like sequence. This sequence will be then used as a prompt for the next step.

The adversarial decoding is effectively a greedy maximization method of the human-likeness as evaluated by the discriminator.

5.1 The Adversarial Decoder Evaluation

The figure 2 provides evaluation of the adversarial decoding among with the comparison to baseline decoding methods. Adversarial decoding is able

to consistently cause more confusion to the discriminator. The more samples are evaluated per step the higher the performance of the adversarial decoder. Adversarial decoder with 3 samples achieves around 60% accuracy, with 5 samples 40% accuracy and with 10 samples around 20% of accuracy.

6 Future Work

Due to time limitations not all analyses have been carried out. It could be interesting to compare specific features of adversely generated text - like perplexity, distribution of part of speech tags, tendency to repeat itself. It would be also interesting to involve a separate discriminator either separately from the RoBERTa discriminator or in an ensemble. It could be also interesting to train discriminator in a loop on adversely generated text effectively creating a sort of generative adversarial network for NLP.

7 Conclusions

We have discussed concepts of language model and language model decoding. We have presented 3 decoding strategies, including one fully novel. We have explored impact of sampling parameters on accuracy of discrimination. Then we proposed a new adversarial method of decoding. The method has shown to be capable of fooling some of fake news discriminators proposed in the literature. Further extension of the work have been proposed including a generative adversarial network for NLP.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

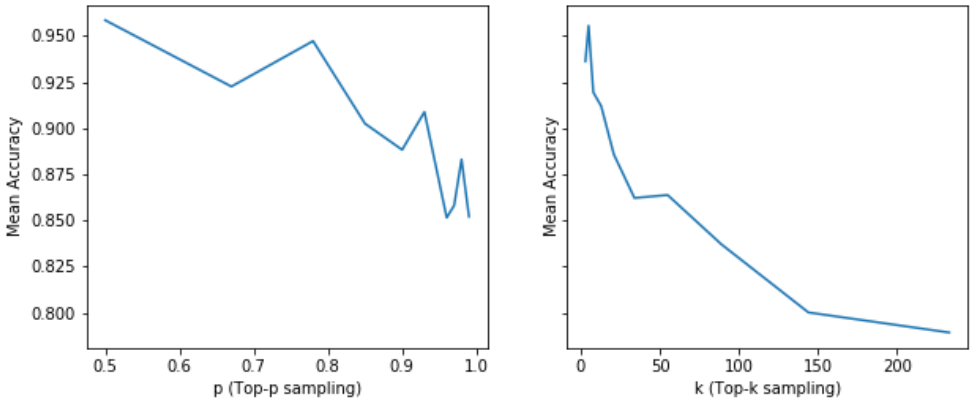

Figure 1: Comparison of two baseline decoding methods. The mean accuracy is the accuracy of a human vs. computer discriminator on 100 computer generated samples of text.

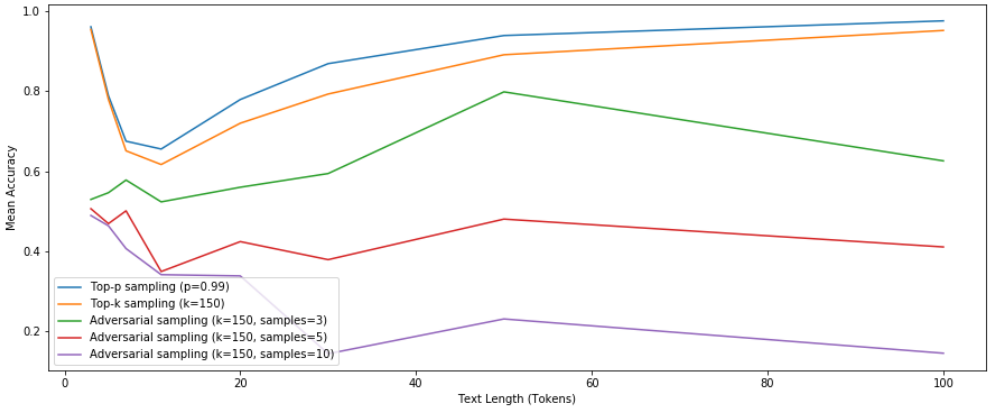

Figure 2: Comparison of 5 different language model decoding methods. The mean accuracy is the accuracy of a human vs. computer discriminator on 100 computer generated samples of text.