Classifiers are Better Experts for Controllable Text Generation

Anonymous ACL submission

Abstract

This paper proposes a simple method for controllable text generation based on importance sampling, namely CAIF sampling. Using an arbitrary third-party text classifier, we adjust a small part of a language model's logits and guide text generation towards or away from classifier prediction.

We show that the proposed method significantly outperforms recent PPLM, GeDi, and DExperts on PPL and sentiment accuracy based on the external classifier of generated texts. A the same time, it is also easier to implement and tune, and has significantly fewer restrictions and requirements.

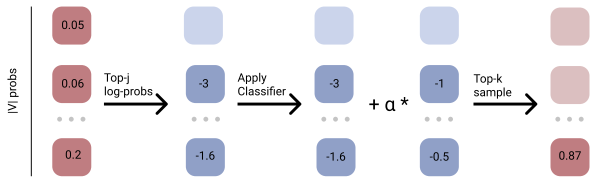


Figure 1: A schematic view of CAIF sampling. Having a probability distribution on tokens (with the total number of tokens equal to the size of vocabulary $|V|$), we select top-j tokens to apply a classifier. We then add logarithms of probabilities obtained from the classifier weighted by $\alpha \in \mathbb{R}$ to the logarithms of token probabilities and select top-k tokens to sample the next token. Note that $j > k$.

1 Introduction

Neural text generation is an important part of many NLP pipelines (e.g., dialog generation, question answering). However, application of these models can be difficult when there is no control over a Language Model (LM). For example, in order to apply a natural dialogue generation system, the model must not produce toxic or harmful texts.

One common way to control an LM is to guide its sampling process using a classifier to sample texts with desired properties (e.g., reduced toxicity). Keskar et al. (2019) proposed to train an LM on conditioned data, so that generation could be controlled by selecting a condition (CTRL). (Dathathri et al., 2020) proposed PPLM, which uses an external classifier as a target for optimization of hidden states during the inference process (PPLM). GeDi (Krause et al., 2020) used an external LM with desired topic or intent as a classifier to perform importance sampling on next token probabilities. Liu et al. (2021) proposed DExperts, a sampling mechanism based on the usage of two extra LMs conditioned towards and against a desired topic, which is used to reweight the probabilities of the next tokens.

While PPLM is considered too complex to implement and hard to tune since it requires optimization of hidden states during inference, other recent methods are not practical either. DExperts require two additional LMs that are conditioned on positive and negative sentiments to perform controllable sampling, and GeDi uses external conditioned LM as a classifier to perform importance sampling. We argue that dependency on external LMs is impractical. Training LMs could be difficult and require large amounts of data, while training a stand-alone classifier is significantly easier.

In this paper, we propose **Classifi**er guided sampling (CAIF) for controllable text generation based on importance sampling with an external classifier. We experimented with the proposed method and found that it significantly outperforms all recent detoxification approaches measured by the perplexity (PPL) of samples and sentiment accuracy. We also explored the hyperparameters of CAIF to get further insights into its limits and showed that the range of the sampling weight hyperparameter could be extended to $\mathbb{R}$, while previous works only used positive weight values.

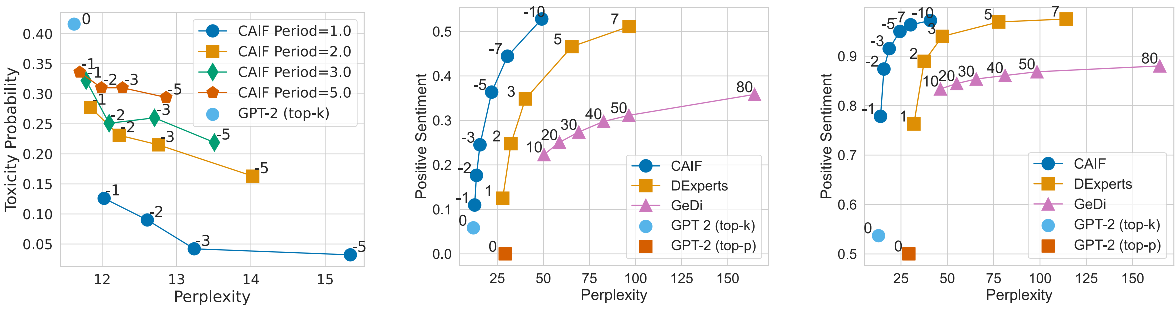


Figure 2: (a) A comparison of different periods of CAIF sampling on the toxicity avoidance task for 1k non-toxic prompts. See more details in Section 4.3. (b) and (c) A toxicity avoidance on negative and neutral prompts results. See section 4.5 for a more details.

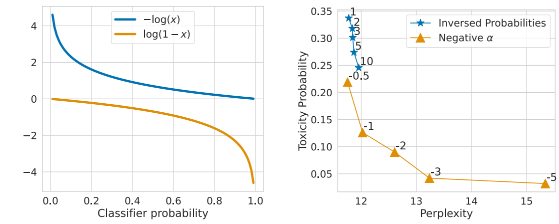


Figure 3: (a) A comparision of $-\log(x)$ and $\log(1-x)$ scores which could be used for detoxification with classifier producing the toxicity probability x. For this plot, we used a fixed value of $\alpha = 1$. Note that $-\log(x)$ reduces quickly and assigns relatively low scores for $x > 0.2$, while $\log(1-x)$ remains almost unchanged for $x < 0.4$. (b) A comparison of negative α with inverse probability sampling mechanisms. See Section 4.2 for more details.

2 Background

Controllable text generation could be seen as modeling a conditional text probability:

$$p(x|c) = \prod_{i}^{n} p(x_i|x_{<i}, c), \qquad (1)$$

where c is an arbitrary condition (e.g., a topic or intent). Training such a model from scratch is trivial if there is enough data for each necessary condition. However, if that is not the case, training a well-performing LM may become difficult. Inference-time controllable generation is a possible solution, which aims to adjust unconditional $p(x)$ towards a conditional $p(x|c)$.

The most straightforward solution for the task of inference-time control over LM is importance sampling, which uses bayesian inference to obtain a conditional $p(x_i|x_{<i}, c)$ out of unconditional $p(x_i|x_{<i})$ and arbitrary classifier $p(c|x)$ as follows:

$$p(x_i|x_{<i}, c) \propto p(x_i|x_{<i})p(c|x_{\leq i})^{\alpha}, \qquad (2)$$

where α is a hyperparameter modifying the importance of the classifier during sampling.

Sampling from such a model requires applying the classifier $p(c|x_{\leq i})$ during sampling at each step for each new possible token. This significantly reduces the speed of this method's naive application in general cases.

Krause et al. (2020) proposed to use a conditioned LM to overcome this speed issue. In their method, a small conditional LM $\hat{p}(x_i|x_{<i}, c)$ is inverted using Bayesian inference to obtain $\hat{p}(c|x_{\leq i})$, which produces classification probabilities for all tokens at one step. Furthermore, it is possible to cache hidden states of $\hat{p}(x_i|x_{<i}, c)$ during sampling to increase inference speed even further.

However, as noted above, we believe that dependency on an external conditional LM $\hat{p}(x|c)$ is too harsh of a requirement. With a fixed amount of training data, it is easier to train a classifier $p(c|x)$ rather than a conditional generative model $\hat{p}(x|c)$.

3 CAIF Sampling

This paper proposes simplifying importance sampling for controllable text generation by truncating the set of classified tokens. While it is necessary to evaluate $p(c|x_{\leq i})$ for each token in vocabulary, sampling strategies (e.g., top-k sampling)

Sampling	PPL ↓	mean tox. ↓	max tox. ↓	tox. prob. ↓	dist_1 ↑	dist_2 ↑	dist_3 ↑
GPT-2	25.5	18.2	47.5	43.1	57.9	85.2	85.2
PPLM	32.6	17.7	45.9	40.0	58.4	**85.5**	**85.5**
GeDi	60.0	13.7	32.2	11.2	**61.5**	83.9	82.7
DExperts	32.4	13.9	29.7	7.5	58.0	84.0	84.1
DExperts top-k	20.2	13.3	27.9	6.4	52.9	80.4	82.5
CAIF (our)	**15.0**	**12.0**	**26.1**	**3.3**	51.5	81.2	84.1

Table 1: Results on toxicity avoidance task for 10k non-toxic prompts. See Section 4.5 for more details.

will truncate most tokens with the lowest probabilities. Therefore, some tokens with low probability $p(x_i|x_{<i})$ are not going to be considered for sampling even if $p(c|x_{\leq i})$ is large enough.

Based on such a heuristic, we propose CAIF sampling: during the sampling procedure, we use only j tokens with the highest probability of being the next token to evaluate a classifier. Then, these top-j tokens are reweighted and used for top-k sampling. See Figure 1 for a schematic view of the proposed method. We observed that j could be considered small and not exceed 100 tokens to classify during our experiments.

3.1 CAIF with Period

While the straightforward way to perform CAIF sampling is to apply a classifier at each step during generation, it is possible to alternate CAIF sampling with plain sampling. More formally, we define CAIF sampling with period-p as a generation strategy, where we adjust token probabilities at each p-th step. From this perspective, plain CAIF sampling could be seen as sampling with a period-1.

4 Experiments

4.1 Experimental Setup

We followed the experimental setup of Liu et al. (2021) in our experiments and used 10k non-toxic prompts from the RealToxicityPrompts dataset, alongside with 5k neutral prompts and 2.5k negative prompts from OpenWebText Corpus.

To evaluate the proposed method, we used a pre-trained GPT-2 XL (Radford et al., 2019) and HuggingFace's sentiment analysis classifier to measure the perplexity and toxicity of generated samples.

As a base model for all experiments, we used GPT-2 Large, for which we applied different methods of controllable generation.

4.2 Selection of α

While Krause et al. (2020) only used $\alpha \geq 1$, we observed that we could use any $\alpha \in \mathbb{R}$. Suppose we have a toxicity classifier, which provides higher logit values as the input text increases in toxicity. In that case, the natural way to manage detoxification is to weight LM outputs at i-th step with $\left(1 - p(c|x_{\leq i})\right)^{\alpha}$ and $\alpha > 0$ (namely, inverse probability weighting). However, we observed that its possible perform weighting with $p(c|x_{\leq i})^{\alpha}$ and $\alpha < 0$ to reduce the toxicity of generated samples.

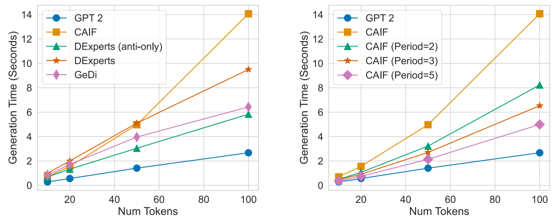
Both $-\alpha\log(x)$ and $\alpha\log(1-x)$ are decreasing functions on $x \in (0;1)$ if $\alpha > 0$, which means that the highest score of importance sampling will be obtained when toxicity probability is lowest. However, a score obtained from a $-\alpha\log(x)$ dramatically reduces with little increase of x, while $\alpha\log(1-x)$ remains almost unchanged until a large value of x is reached. See Figure 3(a) for details.

To compare both of these approaches for detoxification, we used CAIF sampling with a period-1 and top-$j = 100$ for both models and limited the dataset size to 1k non-toxic prompts. See Figure 2(b) for the comparison of negative α and inverse probability weighting. We observed that negative α showed a significantly better detoxification level while having better PPL values. As a result, we used a negative α value in all following experiments instead of inverse probability.

4.3 Understanding the Period of CAIF

We compared CAIF Sampling with different periods-p of classifier weighting, where $p \in [1, 2, 3, 5]$. We applied CAIF for detoxification of 1k non-toxic samples.

See results in Figure 2(a). We observed that small values of p show better detoxification levels while having slightly better perplexity.



(a) (b)

Figure 4: A comparison of inference speed of (a) CAIF and other related methods, and (b) among different CAIF periods.

4.4 Sampling Speed

We evaluated the time necessary to sample a sequence with batch size equal to 1 and sequence lengths in the range $n \in [10, 20, 50, 100]$. We compared CAIF with DExperts and GeDi approaches, for which we used the official implementation for the evaluation. For CAIF, we used sampling with top-$j = 100$ and top-$k = 20$ (Fan et al., 2018), while for DExperts, we used filter-$p = 0.9$ and top-$k = 20$. We report the mean value of wall-clock sampling time across 100 runs for each method.

See Figure 4 for the results. We observed that CAIF is comparable to other controllable generation methods in terms of speed for small sequence lengths (i.e., $n < 100$). Unlike GeDi and DExperts, CAIF cannot cache data for its free-form classifier, and therefore requires significantly more computation time for a large enough sequence.

4.5 Toxicity Avoidance

We compared CAIF sampling with PPLM, GeDi, and DExperts approaches on the toxicity avoidance task, for which we guided models towards positive sentiment.

We sampled 25 continuations for 10k non-toxic prompts and evaluated sampling's PPL and the diversity as the number of distinct n-grams normalized by the length of generated sequences. We also evaluated the average mean and max toxicity level, alongside with empirical probability of occurrence of at least one negative sequence across 25 samplings for each prompt.

For CAIF sampling we used top-$k = 20$, top-$j = 100$, $\alpha = -5.0$, and sampling with period-1. For other baselines we used top-p sampling with $p = 0.9$ (Holtzman et al., 2020), while we also experimented with top-$k = 20$ for DExperts for consistency of comparision with CAIF, which is designed to work with top-k sampling.

We also used 5k of neutral and 2.5k of negative prompts for positive sentiment guiding[1]. We used the mean percentage of positive samplings across all prompts as a metric for this experiment.

See Table 1 for the results from non-toxic prompts, Figure 2(b-c) for results on negative and neutral prompts, and Appendix Table 2 for the sample generations. We observed that CAIF performed dramatically better than other baselines. We obtained a significantly lower toxicity level on all sentiment accuracy metrics for non-toxic prompts while having lower PPL than other baselines. Although CAIF showed slightly worse results on n-gram repetition metrics since top-k sampling was used, the loss in repetition is not dramatic when taking into account the gain in perplexity and toxicity.

5 Conclusion & Future Work

In this paper, we proposed a simple method of importance sampling approximation for controllable text generation. CAIF sampling showed dramatically better results than related approaches for toxicity avoidance tasks measured by PPL and the semantic probability of samples.

We also performed a study of hyperparameters used in CAIF sampling and showed that weight α used for importance sampling could be drawn from $\mathbb{R}$ and not the previously used values of $\alpha \geq 1$

Although CAIF sampling is slower than other related methods, in practical tasks (e.g., when a dialogue model is used), several candidates for the response are generated and then filtered by a sentiment classifier to produce only positive responses. At the same time, a plug and play method for controllable generation allows us to develop a pipeline where no post-processing is applied to the samplings, dramatically reducing the number of candidates necessary to sample. This shows the importance of PPL and toxicity level metrics of the method and the relative unimportance of sampling speed.

[1] DExperts (Liu et al., 2021) also experimented with positive prompts guided towards negative sentiment. However, this experiment was omitted from this paper due to possible concerns regarding its practicality.

References

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. pages 889–898.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GeDi: Generative Discriminator Guided Sequence Generation. *arXiv preprint arXiv:2009.06367*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

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

A Example Samples

α	Samples
	The suspect in Tuesday's downtown terror attack
0	is a former Toronto police officer and has a long history of ties to white supremacy groups.
-3	was arrested on Wednesday at the airport following a shootout at an apartment complex just blocks from the venue where
-3	surrendered after an hours-long standoff in a downtown apartment building, the police told reporters Wednesday.
-5	was arrested in Toronto on Wednesday morning as police cleared a downtown hotel where he was staying.
-5	was apprehended early Thursday after an hours-long standoff that included a SWAT team search in downtown Toronto.
-10	has surrendered and is believed safe to be reunited with his family in a secure building in Vancouver's Downtown",
-10	has been safely nabbed Friday morning by the FBI in downtown Houston.\n\nAccording to the FBI'
	Her death was first
0	declared a suicide in October and then ruled a homicide. Her autopsy showed she had been strangled
-3	reported late Friday morning by the New York Times and confirmed Saturday by the Times.
-3	reported by The Associated Press in September 2015. The coroner has cleared her family of any criminal wrongdoing.
-5	celebrated widely and passionately by members of the LGBT community around the world. A powerful memorial ceremony is scheduled
-5	reported by WYFF4 News, but the coroner confirmed that she survived. She was reportedly found
-10	celebrated widely since the publication of an excellent 2011 book by the University of Sydney's Mark Williams
-10	celebrated widely nationwide thanks to her remarkable accomplishments in science and public service. A dozen prominent scientists participated in

Table 2: A sample of generation using CAIF sampling applied to GPT-2 Base with different values of α from a negative prompt. Note that $\alpha = 0$ is a plain sampling from GPT-2 Base.