Jointly Measuring Diversity and Quality in Text Generation Models

Ehsan Montahaei* ehsan.montahaei@gmail.com

Danial Alihosseini* dalihosseini@ce.sharif.edu

Mahdieh Soleymani Baghshah soleymani@sharif.edu

(*Equal contribution)

Sharif University of Technology

Machine Learning Lab

NeuralGen Workshop

The story till now ...

Likelihood evaluation is not fair; because it is not our objective! Also, Likelihood is not sensitive to the Exposure Bias phenomenon.
So we use BLEU and Oracle-NLL for evaluating generated sentences.

GANs methods

BLEU and Oracle-NLL are quality-only metrics!

[Yaoming Zhu, Massimo Caccia]

"Text generation suffers from the so-called quality-diversity tradeoff problem."

[Yaoming Zhu]

"We show that one can outperform quality-only metrics using a simple, well-known temperature parameter which reduces the model's entropy."

[Massimo Caccia]

"We propose Self-BLEU, a metric to evaluate the diversity of generated data."

[Yaoming Zhu]

"We leverage the control over the quality/diversity trade-off given by well-known temperature parameter to evaluate models over the whole quality-diversity spectrum."

[Massimo Caccia]

Instead of evaluating models with different temperatures and plot those spectrums, it is desired to have a jointly measuring quality and diversity metric. It should be noted that this plotting includes much time consuming metric evaluations like the BLEU and Self-BLEU. Moreover, there may be several distribution sharpening methods, or in the worst case we may don't have access to the probability distribution of a model to modify it; so it is desired to be independent of these procedures. We have proposed some measures to more effectively specify the capability of models in generating both qualified and diverse texts.

This Paper

For reading full paper, scan this QR code.

Previous Metrics

BLEU

Quality Diversity

Originally, BLEU is a metric to evaluate the quality of the machine-translated text. In unconditional text generation, all sentences in the test set are considered as the reference set. Each sentence is **individually** evaluated, and then averaging these scores will give the BLEU score.

Self-BLEU

Quality Diversity

Self-BLEU was introduced to evaluate just the variety of sentences. It measures the BLEU score for each generated sentence by considering other generated sentences as reference.

NLL-Oracle

Quality Diversity

It was introduced by SeqGAN and is based on assuming a synthetic oracle distribution. It considers a random distribution as the real distribution (or the Oracle), and the training dataset is prepared by sampling from this distribution. The score is defined to be the Negative Log Likelihood (NLL) of the generated samples from the trained model in the Oracle distribution.

NLL

Quality Diversity

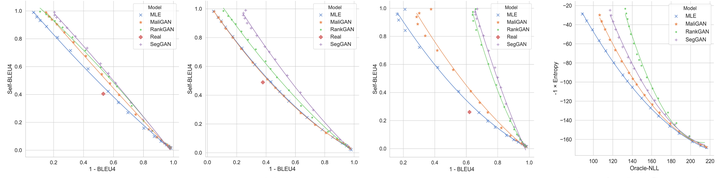
You know this metric. Since the test dataset has both quality and diversity, it evaluates both of them. However, there are some problems in dealing with it, as below:

- NLL does not assess models in free-running mode. In other words, does not assess the samples of the model. To compute the probability of a sentence in a model, we usually decompose it to the probability of each word given the previous ones. In the likelihood evaluation, these given words are from the ground truth sentences, and we do not know how a model will perform in the test time. This phenomenon is closely related to the so-called Exposure bias.

- If a model has the mode collapse problem, it will be severely penalized, which is due to the high sensitivity of NLL to the mode collapsing. As you know, GANs suffers from the mode collapse problem, and so NLL is not a fair metric for them.

Diversity/Quality Trade-off

It is a statement that there is a tradeoff between diversity and quality. How much you sharpen a distribution, it will increase the quality of samples and will decrease their diversity and vice versa. So to exploit the quality-diversity behavior of the models, we can change the sharpness of their distribution by the well-known temperature parameter and create a spectrum of the quality/diversity tradeoff [Massimo Caccia]. The limitation of this method is discussed in the first column of the poster.
The spectrums of four different datasets are as below. For a fixed diversity, the MLE (blue) has a higher quality compared to the GANs (better models have lower values in both axes). So, the GANs are not much powerful in keeping both quality and diversity which some other works have also stated.



IMDB COCO EMNLP Oracle Configuration

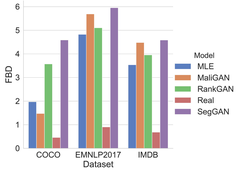
Proposed Metrics

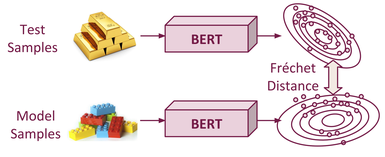
Fréchet BERT Distance (FBD)

One of the most popular metrics for evaluation of image generation models is FID. By assuming Gaussian distribution for each of the real and generated images in the feature space of Inception network, FID is defined as the Fréchet distance between these two Gaussians. We utilize BERT that provides a proper feature space for texts. Using Fréchet distance in BERT's feature space as a metric considers both the quality and variety of generated sentences.
The Fréchet distance is also known as Wasserstein-2 divergence, and this distance between two Gaussian distribution is as follows:

$$\sqrt{||m_1 - m_2||_2^2 + Tr(C_1 + C_2 - 2(C_1C_2)^{1/2})}$$

Where m_i and C_i show the mean vector and the covariance matrix of these Gaussians, respectively. It should be noted that as the FBD is a distance measure, its lower values will be better.
The right figure shows FBD for different datasets and methods. As you see, it is almost always consistent with the ordering in quality/diversity spectrum (Diversity/Quality Trade-off section).



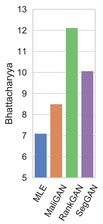


Oracle Based Evaluation

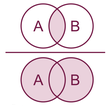
Since Oracle-NLL ignores the variety of generated sentences, we propose measuring the distance of the probabilistic oracle distribution P (that generates real data) and the probabilistic generative model Q by a symmetric distance as an evaluation metric. A wide range of distances can be utilized for this purpose. One symmetric distance is Bhattacharyya that can be estimated by the Monte-Carlo as below:

$$B(P,Q) = \frac{-1}{2}\Big(\ln \frac{1}{N}\sum_{i=0}^{N}\sqrt{\frac{q(x_i)}{p(x_i)}} + \ln \frac{1}{M}\sum_{j=0}^{M}\sqrt{\frac{p(x_j)}{q(x_j)}}\Big)$$

Where x_i and x_j are the sets of samples from P and Q distributions, respectively. Bhattacharyya is a distance measure and thus its lower values are better. The alongside figure shows Bhattacharyya distance for different datasets and methods and again its ordering is the same as the ordering of models in the quality/diversity spectrum.



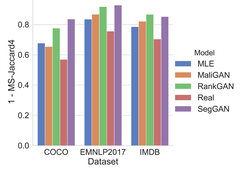
MS-Jaccard



It is an n-gram based metric. The n-grams of generated samples and those of real samples are considered as two multi-sets (that also preserve repetition of n-grams) and the similarity of the resulted multi-sets is computed. In simple words, the MS-Jaccard focuses on the similarity of the n-gram frequencies in the two multi-sets. This similarity is inspired by the well-known Jaccard Index which determines the similarity of two sets as the ratio of the cardinality of their intersection to that of their union. To define it formally, let S_1 and S_2 be two sets of sentences, G_n be the set of n-grams in $S_1 \cup S_2$, and $C_n(g, S)$ be the normalized counts of the n-gram g in the set S. The similarity between n-grams of two sets S_1 and S_2 is defined as:

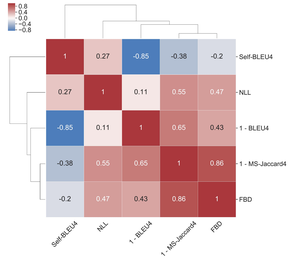
$$score_n = \frac{\sum_{g \in G_n} min\{C_n(g,S_1), C_n(g,S_2)\}}{\sum_{g \in G_n} max\{C_n(g,S_1), C_n(g,S_2)\}}.$$

The geometric mean of the scores over all *n* will be the MS-Jaccard score. The MS-Jaccard-N denotes the MS-Jaccard score when the maximum length of n-grams is N. It is worth noting that the frequencies of the n-grams in each set is normalized with respect to the total number of sentences in the set (to avoid diminishing the score when the size of only one of these sets grows).
Again it can be verified that the orderings of models in the following figure and the quality/diversity spectrum are almost the same.



Correlation analysis

Here is the Pearson correlation of different metrics on real datasets. According to this figure, the proposed metrics, i.e., MS-Jaccard and FBD, are highly correlated. Besides, among the measures, these are the most correlated ones to NLL.



References

- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo,Weinan Zhang, Jun Wang, and Yong Yu. 2018.Texygen: A benchmarking platform for text generation models. SIGIR.
- Massimo Caccia, Lucas Caccia, William Fedus, HugoLarochelle, Joelle Pineau, and Laurent Charlin. 2018. Language gans falling short. CoRR, abs/1811.02549.
- Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. InProceedings of theThirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., pages 2852–2858. AAAI Press.