SALSA-TEXT : SELF ATTENTIVE LATENT SPACE BASED ADVERSARIAL TEXT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Inspired by the success of self attention mechanism and Transformer architecture in sequence transduction and image generation applications, we propose novel self attention-based architectures to improve the performance of adversarial latent code-based schemes in text generation. Adversarial latent code-based text generation has recently gained a lot of attention due to its promising results. In this paper, we take a step to fortify the architectures used in these setups, specifically AAE and ARAE. We benchmark two latent code-based methods (AAE and ARAE) designed based on adversarial setups. In our experiments, the Google sentence compression dataset is utilized to compare our method with these methods using various objective and subjective measures. The experiments demonstrate the proposed (self) attention-based models outperform the state-of-the-art in adversarial code-based text generation.

1 INTRODUCTION

Text generation is of particular interest in many natural language processing (NLP) applications such as dialogue systems, machine translation, image captioning and text summarization. Recent deep learning-based approaches to this problem can be categorized into three classes: auto-regressive or maximum likelihood estimation (MLE)-based, generative adversarial network (GAN)-based and reinforcement learning (RL)-based approaches.

MLE-based methods (such as Sutskever et al. (2014)) model the text (language) as an auto-regressive process, commonly using RNNs. RNNs compactly represent the samples history in the form of recurrent states. In these models, text is generated by predicting next token (character, word, etc) based on the previously generated ones (Graves, 2013).

One of the main challenges involved with auto-regressive methods is exposure bias (Bengio et al., 2015). This problem arises due to discrepancy between the training and generation phases. In fact, ground-truth samples from the past are used in training, while past generated ones are used in generation. A number of solutions have been proposed to address this problem by modifying the training procedure including scheduled sampling (Bengio et al., 2015), Gibbs sampling (Su et al., 2018), and Professor forcing (Lamb et al., 2016).

Over the past few years, researchers have extensively used GANs (Goodfellow et al., 2014) as a powerful generative model for text (Yu et al., 2017; Che et al., 2017), inspired by the great success in the field of image generation. GANs are believed to be capable of solving the exposure bias problem in text generation raised from using MLE. The reason is that they solved a similar issue of blurry image generation in MLE-based variational autoencoders (VAEs). It is belived that the discriminator is able to guide the text generator, through their training exchange, how to generate samples similar to real (training) data. However, there are other challenges involved in GAN-based text generation.

A few of these challenges in text generation are inherent to GANs themselves, such as mode collapse and training instability. The mode collapse problem happens when the adversarially trained generator does not produce diverse texts. These issues can be mitigated by using well-known techniques such as feature matching (Zhang et al., 2017), and entropy regularization (Shi et al., 2018). Another challenge is due to the discrete nature of text, which causes the generator sampling to be non-differentiable over the categorical distribution of the words.

In this paper, we take advantage of Transformer self-attention mechanism (Vaswani et al., 2017) and incorporate it in two state-of-the-art adversarial latent code-based schemes proposed for text generation. More specifically:

- We incorporate the Transformer structure in the design of encoder and decoder blocks of AAE (Makhzani et al., 2015a) and ARAE (Kim et al., 2017a) setups for text generation.
- Blocks closely inspired from the Transformer's encoder layers, incorporating self-attention and element-wise fully-connected layers in a residual configuration and with positional encodings, are used along with spectral normalization to propose a novel GAN (both generator and discriminator) structure for AAE and ARAE setups.
- The performance improvement obtained from the proposed architectures is demonstrated via objective and subjective measures used in extensive experiments.

2 Related Work

2.1 Spectral Normalization

Spectral normalization (Miyato et al., 2018) is a weight normalization method proposed to stabilize the training of GANs. The authors show that the Lipshitz norm of a neural networks can be bounded by normalizing the spectral norm of layer weight matrices. As opposed to local regularizations used in WGAN-GP, etc., the network-wide spectral regularization stabilizes the GAN training, produces more diverse outputs and results in higher inception scores. We use spectral normalization in our adversarial setups for the same reasons.

2.2 Attention Models

In sequence modeling literature, attention was initially proposed by Bahdanau et al. (2014). It recognizes the fixed-length latent representation of the input sequence as the main performance bottleneck in the seq-to-seq models and proposed using soft-attention in the decoder. Using attention, the decoder can also attend to any desired token in the input sequence besides consuming the compressed representation resulting at the end of encoding operation.

Self-attention was initially proposed for language inference in Parikh et al. (2016). The authors named it as "intra-attention" and showed that their structure can be an effective alternative for LSTMs in the task of natural language inference (Bowman et al., 2015), at the time achieving state of the art performance with much fewer parameters as well as requiring a training time an order of magnitude shorter. Self-attention structures have since been used to set the state of the art in a number of different tasks (Vaswani et al., 2017; Dehghani et al., 2018; Yu et al., 2018; Radford et al.; Al-Rfou et al., 2018). They drastically reduce the path length between any two sequence inputs, making the learning of long term dependencies much easier (Vaswani et al., 2017). They are considerably easier to parallelize, reducing the number of operations that are required to be sequential.

Recently, Zhang et al. (2018) applied self attention along with spectral normalization to the task of image generation using GANs. It showed by visualization that using attention, the generator can attend to far neighborhoods of any shape rather than close-by fixed-shape ones at each level in a hierarchical generation. The authors claim that applying spectral normalization to generator as well as discriminator helps training dynamics (stability). Similarly, we also adopt self attention and spectral normalization in our architecture designs.

Transformer (Vaswani et al., 2017) extended the use of self attention mechanism and was proved to be the state-of-the-art in sequence transduction applications such as machine translation. It dispenses convolutional and recurrent layers and relies entirely on attention-only layers and element-wise feed forward layers.

2.3 latent space-based text generation

One of the main challenges of the language generation task originates from the discrete nature of text. Similarly to generating other discrete tokens, the back propagation of error through argmax

operator is not well-defined. To address this problem, various approaches have been proposed in the literature including continuous approximation of discrete sampling (Gulrajani et al., 2017; Jang et al., 2016), using policy gradient from reinforcement learning (Guo, 2015; Shi et al., 2018), etc. One of the most successful solutions is based on autoencoders with continuous latent spaces (i.e. latent code-based methods). Various training setups have been proposed for training these autoencoders including adversarial (Kim et al., 2017a) and variational (Hu et al., 2017) setups.

A recent paper (Cífka et al., 2018) performs a thorough review of the state-of-the-art latent code-based text generation methods. It studies the performance of a number of code-based text generation schemes and uses a unified rigorous evaluation protocol to evaluate them. We got inspired by their evaluation protocol to demonstrate the strength of our self attention-based approach in the context. They use a broad set of measures to perform a comprehensive study. We adopt forward and reverse perplexity as well as BLEU from their objective measures and fluency from the subjective ones.

2.4 Adversarial Text Generation

In this section, we briefly explain two prominent baseline methods using adversarial latent code-based generation techniques and present the technical details in Section 3.

2.4.1 AAE

Adversarial autoencoder (AAE) (Makhzani et al., 2015b) proposes an adversarial setup to train probabilistic autoencoders. It matches the aggregated posterior of the encoder output (latent codes) to an arbitrary distribution that can be easily sampled from. Although authors demonstrate the applications of their setup in semi-supervised learning, style and content disentanglement, etc, AAE decoder can be effectively used as a generative model, converting samples of the arbitrary distribution (noise) to real-like outputs. From application perspective, authors only evaluated AAE performance in vision-related applications. In this paper, we tailor AAE for text generation, following guidelines proposed by Cífka et al. (2018) and incorporate self attention and Transformer as novel parts in the model.

2.4.2 ARAE

The adversarially regularized autoencoder (ARAE) (Kim et al., 2017b) learns an autoencoder with continuous contracted codes that highly correlate with discrete inputs. That is, similar inputs get encoded (mapped) to nearby continuous codes. ARAE aims at exploiting GAN's ability to force the generator to output continuous codes corresponding to the code space obtained from encoding the real text data. By matching the outputs of generator and encoder, ARAE provides an implicit latent code GAN that serves as a generative model for decoding text.

3 Self Attentive latent code-based models

In this section, we explain the details of our self attention-based models following the ARAE and AAE setups proposed in Cífka et al. (2018). These setups have shown comparable to the state-of-the-art results in text generation. We select similar setups to provide fair comparisons and report the best techniques/parameters based on our experiments.

In our architectures, Transformer (Vaswani et al., 2017) is used in designing all autoencoders. In both encoder and decoder, we use three blocks of Transformer. 'Block' and 'layer' names are used, respectively, instead of 'layer' and 'sub-layer' in the original paper.

Layer normalization(Ba et al., 2016) is applied on every layer (multi-head attention, masked multi-head attention and feed forward layers) within each Transformer block. Multi-head attentions have eight heads and embedding layers are of size 304 (a multiple of eight). Similarly to Vaswani et al. (2017), positional encoding is used at the very first layer of the encoder and decoder. The dimensions and encoding place were found empirically for the best objective and subjective performance.

For GAN structures, i.e. the generator and discriminator architectures, we use modified Transformer encoder layers combined with spectral normalization, as depicted in Fig. 1 ($N = 3$). As in the regular transformer blocks, all connections are residual. Inspired by spectral normalization successes in the

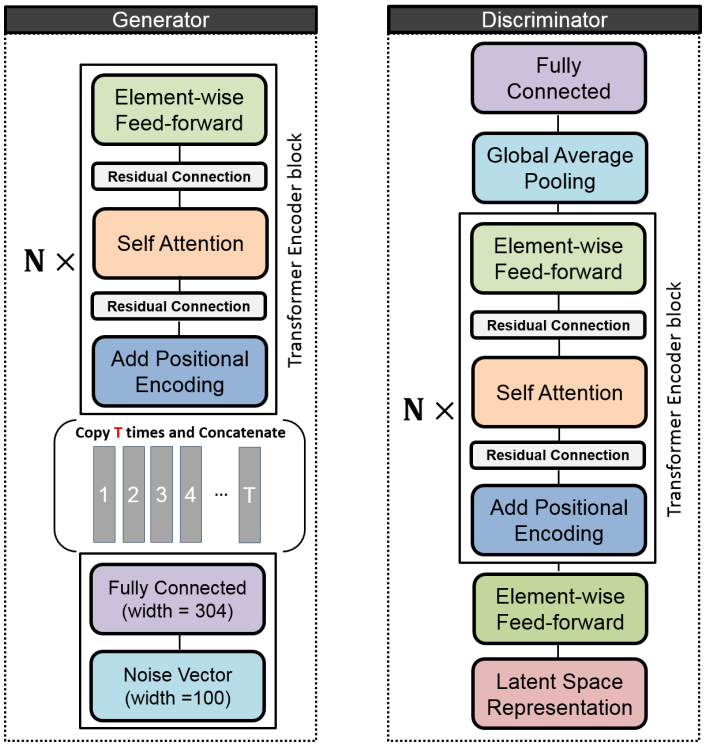

Figure 1: SALSA-TEXT generator and discriminator architecture designed using Transformer encoder structure

GAN-based image generation, especially proved in SAGAN (Zhang et al., 2018), we apply it to the weights of the discriminator and the generator in our network. We did not find layer normalization (used in original Transformer) to be useful, when applied along with spectral normalization in the generator and discriminator architectures. Hence, only use spectral normalization in our GAN structures.

3.1 Adversarial Techniques

We use self attention-based structures in two well-known adversarial setups (Makhzani et al. (2015a) and Kim et al. (2017a)).

AAE We use the AAE-SPH setup used in Cífka et al. (2018). It is based on the original setup proposed in Makhzani et al. (2015a). The discriminator forces encoder outputs to follow a uniform distribution on the unit sphere. Similarly to Makhzani et al. (2015a), a two-phase training is used, where there is regular alternation between minimizing reconstruction and adversarial (regularization) costs. The trade-off factor (λ) between reconstruction and adversarial costs is 20 (as in Cífka et al. (2018)). All over the encoder, decoder and discriminator, input and attention heads are dropped with a probability of 0.1. The general architecture and the proposed (self) attention-based changes are depicted in Fig. 2.

ARAE We use the original setup from Kim et al. (2017a) with fixed-size full codes as inputs to the decoder. Inside the encoder and decoder, word and attention head dropout is performed with a probability of 0.1 and a maximum of 3-word shift is applied to input words. The general architecture and the proposed (self) attention-based changes are depicted in Fig. 3.

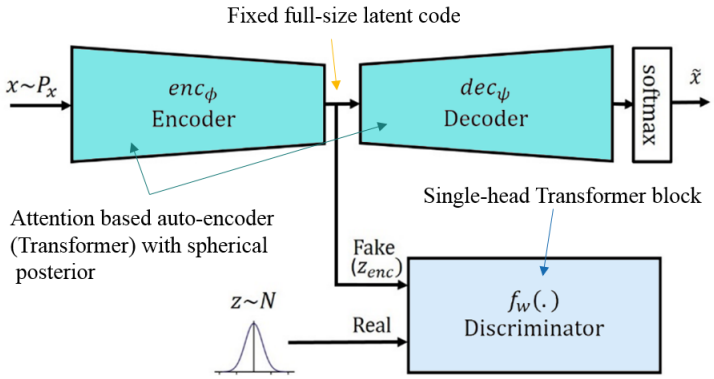

Figure 2: SALSA-AAE for text generation. As explained in the figure, Transformer-based encoder, decoder and discriminator (with self attention) architectures are used. Decoder uses fixed-size full codes.

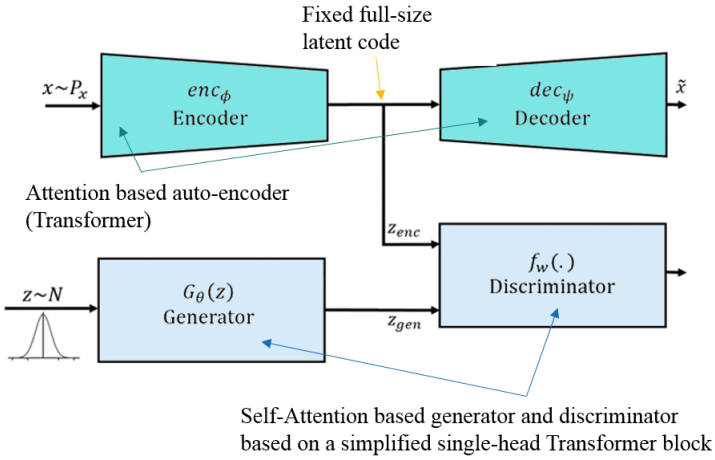

Figure 3: SALSA-ARAE for text generation. Similarly to SALSA-AAE, all blocks are Transformer-based and decoder uses fixed-size full codes. Generator and discriminator comprise of self-attention layers.

4 Experiments

We study the performance of our self attentive (SALSA) architectures and compare it with that of the code-based setups studied in Cífka et al. (2018).

4.1 Experimental Setup

The performance of the models is evaluated in sentence generation (sampling), on the Google Sentence Compression (GSC) dataset [1] (as in Cífka et al. (2018)). Training on this dataset is very challenging as the sentences are relatively long (average of 24.8 words) and diverse in terms of content, grammar, etc. GSC comprises $200,000$ training and $9,995$ test sentences. For all the trained models, we use Google's SentencePiece [2] tokenizer using byte-pair encoding (BPE (Sennrich et al., 2015)) as in Cífka et al. (2018).

[1] https://github.com/google-research-datasets/sentence-compression

[2] `https://github.com/google/sentencepiece`

Table 1: Samples from models trained on GSC dataset

Model	Sample sentences
AAE	− on Thursday for an undisclosed amount of US Senate. − .com is launching a new album on "The Idol's. − . − Tuesday for a US dollar in the US, according to the US Department. − US US markets on Wednesday as US stocks rose to a US dollar, dealers said. − The US will open its first-day visit to next year.
SALSA-AAE	− The world's largest car maker, said it will buy back a new US$4 million fine for the first time in three years. − London, May 6 A 48-year-old man has been charged with raping a woman in the face of livelihood on Tuesday. − In a bid to save money, the US government's most expensive land. − The Governors of the city of Caterpillar, who was in talks with the world's most expensive officials. − Israel has launched a new website that would help the Gaza Strip, and the first such incident in a deadly attack on the West Bank and the United States. − Harrington has been found guilty of two counts of driving under the influence of a semi-final at a local court.
ARAE	− A man has been arrested and charged with sexual assault on an alleged assault and assault for allegedly assaulting him to death and a child. − A man is accused of stealing a child in her own home case of her husband. − A man, who was accused of killing the two-year-old girl in connection with several of them who died from injuries last week. − A man is facing charges of sexual assault for allegedly assaulting a woman and her wife. − A man accused of killing the man in his home is being sentenced to life and is expected to be on the way. − A man has been arrested in connection with her death and then sex with her husband, who was in critical condition and will be released on Monday.
SALSA-ARAE	− Former PSV Ethanolve said he has "two and the title of his contract," the first-round pick of the season. − The Queen will present a "Curprone of the Donegal Plan" in a new biopic of the forthcoming musical, and "Kalways," the school said in a statement. − This week between the "Daily Show" flights, a year after the airing, a day before the launch of a summer vacation. − However, the Myssenon, is planning to open its first wedding, with a move that will include a number of its customers, but they are not planning to sell their services. − The Dallas Morning News reported that a Houston man is in "a new, a state that is leaving the Houston Rockets. − The Dallas Morning News Corp. said it is to open a subsidiary of Houston, the largest newspaper and its staff, to be in the city.

We filter the dataset to only include sentences with a maximum of 50 PBE tokens. This only lowers the average number of words per sentence and total number of sentences to 23.1 and 183739, respectively in the training set. The test dataset is also reduced to 9254 lines with an average of 22.7 words per sentence. Samples of generated sentences from all models is listed in Section 4.3.

The input noise to the generator is of size 100 (as in Cífka et al. (2018)). We upsample the noise to the embedding size of 304 by using a fully connected layer. The same upsampled noise is copied a number of times equal to the maximum number of steps in the sentence. We use $T = 50$ times in our experiments. The noise is then fed to the generator, where positional encodings are added to each step. The previously mentioned fully connected layer also serves to allow the model to learn to protect the information of the positional encodings from the noise. Positional encodings are also added at the start of each transformer encoder block. As we use fixed size sequences, the attention depth is always fixed (T bpe). Positional encodings are also added to the input of each transformer encoder block, inside of the generator.

4.2 EVALUATION METRICS

We use various objective and subjective measures to evaluate the models. As objective measures, we use BLEU (Papineni et al., 2002), Self-BLEU (Zhu et al., 2018), forward and reverse perplexity.

BLEU (Papineni et al., 2002) is a widely used metric to compute the similarity of a set of generated sentences with a reference dataset. The results are described in Table 2.

Self BLEU (Zhu et al., 2018) (Table 3) is a measure of diversity for generated texts. In Self-BLEU, for each generated sentence, we compute the BLEU using the sentence as hypothesis and the rest of the generated sentences as the reference. When averaged over all the references, it gives us a measure of how diverse the sentences are. Lower Self-BLEU scores are better, as high BLEU scores indicate great similarity.

In **Perplexity** evaluation (Table 4), the goal is to measure the individual quality of the sentences generated. We train an LSTM language model on the WMT News 2017 Dataset [3] filtered for lines of a maximum of 50 BPE tokens (a total of 200000 sentences). The perplexity of the language model is computed over 100000 generated sentences for each model.

Reverse perplexity evaluation (Table 4) aims to measure variety of the generated sentences. For each model, we train an LSTM-based language model based on 100000 generated sentences, and then evaluate the perplexity on the GSC test dataset, filtered to a maximum length of 50 BPE. Diverse generated sentences that cover dataset to a good extent would result in better (lower) reverse perplexity measures resulting from the trained LSTM network (language model).

For the **subjective** evaluation (Table 5), we use Amazon mechanical Turk [4] online platform. 18 sentences are sampled from each model, i.e. a total of 162 sentences. We assign 81 randomly selected sentences to 50 native English speakers (among Mechanical Turk Masters with hit approval ratings greater than 75%). The remaining 81 are assigned to another group of 50 people with the same qualifications. Each person was asked to evaluate the assigned 81 sentences in one and a half hours. In the evaluation, the 5-point Likart scale is used to measure grammaticality, semantic consistency and overall (Fluency). The overall reflects both grammar and semantic consistency in addition to other human-specific factors. Hence, it is a good representative of "Fluency" measure used in Cífka et al. (2018).

4.3 SAMPLES OF GENERATED SENTENCES

In Table 1, we list six generated sentences for each model. As seen, AAE generates rather short sentences, while the corresponding SALSA version (SALSA-AAE) has alleviated the issue to a good extent. Finally, ARAE suffers from extreme mode collapse as opposed to its SALSA counterpart.

4.4 RESULTS AND DISCUSSION

The results of objective and subjective evaluations are presented in Tables 2 to 5. As seen, the proposed self attention-based (SALSA) architectures consistently outperform the non-attention-based benchmarks in terms of diversity (measured by reverse perplexity). Moreover, they often show better performance in terms of output quality (measured by BLEU, self BLEU, preplexity and human evaluations) on the long and complicated sentences of the GSC dataset.

As seen in the generated samples (Table 1), human evaluation (Table 5) and objective metrics (Tables 2 to 4), the original AAE and ARAE setups perform very poorly on GSC with long sentences. With reverse perplexities of over 8000 and high self-BLEU scores close to 0.9, they suffer from a high level of mode collapse (repeated sentences).

Human evaluations do not account for lack of diversity. The reason is humans are presented with a number of shuffled sentences and asked to evaluate them independently (without knowing which sentence coming from which model). Hence, in our experiments for the original AAE and ARAE, a model can generate similar sentences (maybe due to mode collapse) and still receives high subjective scores.

[3] `http://www.statmt.org/wmt17/`

[4] https://www.mturk.com/

Objective and subjective evaluation of the studied models on the GSC dataset

Table 2: BLEU

	AAE	SALSA-AAE	ARAE	SALSA-ARAE
BLEU-1	0.671	**0.905**	0.924	0.865
BLEU-2	0.500	**0.638**	0.698	0.585
BLEU-3	0.279	**0.367**	0.402	0.294
BLEU-4	0.149	**0.192**	0.212	0.137
BLEU-5	0.086	**0.101**	0.116	0.071

Table 3: Self-BLEU

	AAE	SALSA-AAE	ARAE	SALSA-ARAE
Self BLEU-1	0.950	**0.897**	0.973	**0.896**
Self BLEU-2	0.759	**0.738**	0.921	**0.731**
Self BLEU-3	0.604	**0.573**	0.843	**0.549**
Self BLEU-4	0.412	**0.410**	0.751	**0.367**
Self BLEU-5	0.270	**0.275**	0.653	**0.226**

Table 4: Perplexity

	AAE	SALSA-AAE	ARAE	SALSA-ARAE
Reverse perplexity	10309	**822**	8857	**1008**
Perplexity	88	**61**	37	106

Table 5: Human Evaluations

	AAE	SALSA-AAE	ARAE	SALSA-ARAE
Grammaticality	2.756	**3.09**	2.980	2.898
Semantic consistency	2.575	**2.597**	2.856	2.617
Fluency	2.604	**2.700**	2.851	2.652

It seems that, in our experiments, the original ARAE model suffers from mode collapse. We can see that it has slightly higher human evaluation scores, but extremely poor diversity metrics, i.e. very high reverse perplexity and self-BLEU scores. It can also be seen in the randomly selected generated sentences (Table 1), where all the sentences start with "A man" and invariably mention he is being arrested or accused of grievous crimes. This is likely because the sentences in the GSC dataset are long and that their structure is elaborate. SALSA-ARAE on the other hand reliably produces sentences of quality with great diversity.

SALSA-AAE has both considerably higher individual quality metrics than the original AAE and much better diversity metrics. It is the strongest pure adversarial text model. As seen in Table 5, SALSA-AAE provides the best grammaticality, semantic consistency and Fluency performance.

5 Conclusion and Future Work

In this paper, we introduced SALSA-TEXT, a Transformer-based architecture for adversarial code-based text generation. It incorporates self-attention mechanism by utilizing Transformer architecture in autoencoder and GAN setups. Our extensive experiments demonstrate the better performance of our models compared to the state-of-the-art in adversarial code-based text generation (without self-attention). The proposed architectures provide diverse, long and high quality output sentences as confirmed by objective metrics and human evaluations in extensive experiments.

As a future direction, it is beneficial to study the performance of self attention in other text generation methods including variational code-based and reinforcement learning-based approaches. Another interesting direction is to experiment with deeper Transformer-based autoencoders to better capture the underlying language model and perform unsupervised pre-training isnpired by the success of Al-Rfou et al. (2018) and Radford et al..

REFERENCES

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 1171–1179, 2015.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

Tong Che, Yanran Li, Ruixiang Zhang, R Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. Maximum-likelihood augmented discrete generative adversarial networks. *arXiv preprint arXiv:1702.07983*, 2017.

Ondřej Cífka, Aliaksei Severyn, Enrique Alfonseca, and Katja Filippova. Eval all, trust a few, do wrong to none: Comparing sentence generation models. *arXiv preprint arXiv:1804.07972*, 2018.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Hongyu Guo. Generating text with deep reinforcement learning. *CoRR*, abs/1510.09202, 2015. URL `http://arxiv.org/abs/1510.09202`.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Yoon Kim, Kelly Zhang, Alexander M Rush, Yann LeCun, et al. Adversarially regularized autoencoders. *arXiv preprint arXiv:1706.04223*, 2017a.

Yoon Kim, Kelly Zhang, Alexander M Rush, Yann LeCun, et al. Adversarially regularized autoencoders for generating discrete structures. *arXiv preprint arXiv:1706.04223*, 2017b.

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pp. 4601–4609, 2016.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015a.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015b.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015. URL http://arxiv.org/abs/1508.07909.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. Towards diverse text generation with inverse reinforcement learning. *arXiv preprint arXiv:1804.11258*, 2018.

Jinyue Su, Jiacheng Xu, Xipeng Qiu, and Xuanjing Huang. Incorporating discriminator in sentence generation: a gibbs sampling method. *arXiv preprint arXiv:1802.08970*, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pp. 2852–2858, 2017.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. *arXiv preprint arXiv:1706.03850*, 2017.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 1097–1100, 2018. doi: 10.1145/3209978.3210080. URL http://doi.acm.org/10.1145/3209978.3210080.