

Language Generation with Recurrent Generative Adversarial Networks without Pre-training

Ofir Press^{*1}, Amir Bar^{*1,2}, Ben Bogin^{*1}

Jonathan Berant¹, Lior Wolf^{1,3}

¹ School of Computer Science, Tel-Aviv University

² Zebra Medical Vision

³ Facebook AI Research

ofir.press@cs.tau.ac.il

Abstract

Generative Adversarial Networks (GANs) have shown great promise recently in image generation. Training GANs for language generation has proven to be more difficult, because of the non-differentiable nature of generating text with recurrent neural networks. Consequently, past work has either resorted to pre-training with maximum-likelihood or used convolutional networks for generation. In this work, we show that recurrent neural networks can be trained to generate text with GANs from scratch using curriculum learning, by slowly teaching the model to generate sequences of increasing and variable length. We empirically show that our approach vastly improves the quality of generated sequences compared to a convolutional baseline.¹

1 Introduction

Generative adversarial networks (Goodfellow et al., 2014) have achieved state-of-the-art results in image generation (Goodfellow et al., 2014; Radford et al., 2015; Arjovsky et al., 2017; Gulrajani et al., 2017). For text generation, training GANs with recurrent neural networks (RNNs) has been more challenging, mostly due to the non-differentiable nature of generating discrete symbols. Consequently, past work on using GANs for text generation has been based on pre-training (Yu et al., 2016; Li et al., 2017; Yang et al., 2017; Wu et al., 2017; Liang et al., 2017; Zhang et al., 2016; Shetty et al., 2017)

or joint training (Lamb et al., 2016; Che et al., 2017) of the generator and discriminator with a supervised maximum-likelihood loss.

Recently, two initial attempts to generate text using purely generative adversarial training were conducted by Gulrajani et al. (2017) and Hjelm et al. (2017). In these works, a convolutional neural network (CNN) was trained to produce sequences of 32 characters. This CNN architecture is fully differentiable, and the authors demonstrated that it generates text at a reasonable level. However, the generated text was still filled with spelling errors and had little coherence. RNNs are a more natural architecture for language generation, since they condition each generated character on the entire history, and are not constrained to generating a fixed number of characters.

In this paper, we extend the setup of Gulrajani et al. (2017) and present a method for generating text with GANs. Our main contribution is a model that employs an RNN for both the generator and discriminator, similar to current state-of-the-art approaches for language generation (Sutskever et al., 2011; Mikolov, 2012; Jozefowicz et al., 2016). We succeed in training the model by using curriculum learning (Elman, 1993; Bengio et al., 2009): At each stage we increase the maximal length of generated sequences, and train over sequences of variable length that are shorter than that maximal length. In addition, we aid the model by feeding it with ground truth characters before generation. We show that these methods vastly improve the quality of generated sequences. Sequences contain substantially more n-grams from a development set compared to those generated by a CNN, and generation generalizes to sequences that are longer than the sequences the model was trained on.

^{*}Denotes equal contribution. Author ordering determined by coin flip.

¹Code for our models and evaluation methods is available at <https://github.com/amirbar/rnn.wgan>

2 Motivation

While models trained with a maximum-likelihood objective (ML) have shown success in language generation (Sutskever et al., 2011; Mikolov, 2012; Jozefowicz et al., 2016), there are drawbacks to using ML, that suggest training with GANs. First, using ML suffers from “exposure bias”, that is, at training time the model is exposed to gold data only, but at test time it observes its own predictions, and thus wrong predictions quickly accumulate, resulting in bad text generation.

Secondly, the ML loss function is very stringent. When training with ML, the model aims to allocate all probability mass to the i -th character of the training set given the previous $i - 1$ characters, and considers any deviation from the gold sequence as incorrect, although there are many possible sequences given a certain prefix. GANs suffer less from this problem, because the objective is to fool the discriminator, and thus the objective evolves dynamically as training unfolds. While at the beginning the generator might only generate sequences of random letters with spaces, as the discriminator learns to better discriminate, the generator will evolve to generate words and after that it may advance to longer, more coherent sequences of text. This interplay between the discriminator and generator helps incremental learning of text generation.

3 Preliminaries

Gulrajani et al. (2017) and Hjelm et al. (2017) trained a purely generative adversarial model (without pre-training) for character-level sentence generation. We briefly review the setup of Gulrajani et al. (2017), who use the Improved Wasserstein GAN objective (Arjovsky et al., 2017; Gulrajani et al., 2017), which we employ as well. Hjelm et al. (2017) have a similar setup, but employ the Boundary-Seeking GAN objective.

The generator G in Gulrajani et al. (2017) is a CNN that transforms a noise vector $z \sim N(0, 1)$ into a matrix $M \in \mathbb{R}^{32 \times V}$, where V is the size of the character vocabulary, and 32 is the length of the generated text. In this matrix the i -th row is a probability distribution over characters that represents a prediction for the i -th output in the character sequence. To decode a sequence, they choose the highest probability character in each row. The discriminator D is another CNN that receives a matrix as input and needs to determine if this ma-

trix is the output of the generator G or sampled from the real data (where each row in the matrix now is a one-hot vector). The loss of the Improved WGAN generator is:

$$L_G = -\mathbb{E}_{\tilde{x} \sim \mathbb{P}_g}[D(\tilde{x})],$$

and the loss of the discriminator is:

$$L_D = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x \sim \mathbb{P}_r}[D(x)] \\ + \lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2],$$

Where \mathbb{P}_r is the data distribution and \mathbb{P}_g is the generator distribution implicitly defined by $\tilde{x} = G(z)$. The last term of the objective controls the complexity of the discriminator function and penalizes functions that have high gradient norm, that is, change too rapidly. $\mathbb{P}_{\hat{x}}$ is defined by sampling uniformly along a straight line between a point sampled from the data distribution and a point sampled from the generator distribution.

A disadvantage of the generators in Gulrajani et al. (2017) and Hjelm et al. (2017) is that they use CNNs for generation, and thus the i -th generated character is not directly conditioned on the entire history of $i - 1$ generated characters. This might be a factor in the frequent spelling mistakes and lack of coherence in the output of these models. We now present a model for language generation with GANs that utilizes RNNs, which are state-of-the-art in language generation.

4 Recurrent Models

We employ a GRU (Cho et al., 2014) based RNN for our generator and discriminator. The generator is initialized by feeding it with a noise vector z as the hidden state, and an embedded start-of-sequence symbol as input. The generator then generates a sequence of distributions over characters, using a softmax layer over the hidden state at each time step.

Because we want to have a fully-differentiable generator, the input to the RNN generator at each time step is not the most probable character from the previous time step. Instead we employ a continuous relaxation, and provide at time step i the weighted average representation given by the output distribution of step $i - 1$. More formally, let α_{i-1}^c be the probability of generating the character c computed at time step $i - 1$, and let $\phi(c)$ be the embedding of the character c , then the input to the

Table 1: Samples and evaluation of the baseline model from Gulrajani et al. (2017).

Samples	%IN-TEST- n			
	1	2	3	4
Official marth Damilicon was eng The later , trading touse of the First killed sye of Nondon , and	64.4	25.9	5.1	0.4

GRU at time step i is $\sum_c \alpha_{i-1}^c \phi(c)$. This is fully differentiable compared to $\arg \max_{\phi(c)} \alpha_{i-1}^c$. We empirically observe that the RNN quickly learns to output very skewed distributions.

The **discriminator** is another GRU that receives a sequence of character distributions as input, either one-hot vectors (for real data) or softer distributions (for generated data). Character embeddings are computed from the distributions and fed into the GRU. The discriminator then takes the final hidden state and feeds it into a fully connected layer which outputs a single number, representing the score that the discriminator assigns to the input. The models are trained with the aforementioned Improved WGAN objective (Section 3).

An advantage of a recurrent generator compared to the convolutional generator of Gulrajani et al. (2017) and Hjelm et al. (2017) is that can output sequences of varying lengths, as we empirically show in Section 5.

Our baseline model trains the generator and discriminator over sequences of length 32, similar to how CNNs were trained in Gulrajani et al. (2017). We found that training this baseline was difficult and resulted in nonsensical text. We now present three extensions that stabilize the training process.

Curriculum Learning (CL) In this extension, we start by training on short sequences and then slowly increase sequence length. In the first training stage, the generator G generates sequences of length 1, and the discriminator D receives real and generated sequences of length 1 as input. Then, the generator generates sequences of length 2 and the discriminator receives sequences of length 2. We increase sequence length in this manner until the maximum length of 32 characters.

Variable Length (VL) Here, we define a maximum length l , and generate during training sequences of every length $\leq l$ in every batch. Without curriculum learning, this amounts to training G and D in every batch with sequences of length i , $1 \leq i \leq 32$. With curriculum learning, we generate at each step sequences of length i , $1 \leq i \leq l$, and slowly increase l throughout training.

Teacher Helping (TH) Finally, we propose a procedure where we help the generator learn to generate long sequences by conditioning on shorter ground truth sequences. Recall that in our baseline, the generator generates an entire sequence of characters that are fed as input to the discriminator. Here, when generating sequences of length i , we feed the generator a sequence of $i - 1$ characters, sampled from the real data. Then, the generator generates a distribution over characters for the final character, which we concatenate to the real characters and feed as input to the discriminator. The discriminator observes a sequence of length i composed of $i - 1$ real characters and one character that is either real or generated. This could be viewed as a conditional GAN (Mirza and Osindero, 2014), where the first $i - 1$ characters are the input and the final character is the output. Note that this extension may suffer from exposure bias, similar to the ML objective, and we plan to address this problem in future work.

5 Results

To directly compare to Gulrajani et al. (2017), we follow their setup and train our models on the Billion Word dataset (Chelba et al., 2013). We evaluate by generating 640 sequences from each model and measuring %IN-TEST- n , that is, the proportion of word n -grams from generated sequences that also appear in a held-out test set. We evaluate these metrics for $n \in \{1, 2, 3, 4\}$. Our goal is to measure the extent to which the generator is able to generate real words with local coherence.

In contrast to Arjovsky et al. (2017) and Gulrajani et al. (2017), where the generator is trained once for every 10 training iterations of the discriminator, we found that training the generator for 50 iterations every 10 training iterations of the discriminator resulted in superior performance. In addition, instead of using noise vectors sampled from the $N(0, 1)$ distribution as in Gulrajani et al. (2017), we sample noise vectors from the $N(0, 10)$ distribution, since we found this leads to a greater variance in the generated

Table 2: Samples and evaluation of our models with an RNN generator and discriminator and various extensions. For the CL+VL+TH model we present results for generated sequences of length 32 and length 64.

CL	VL	TH	Samples	%IN-TEST- n			
				1	2	3	4
\times	\times	\times	???cccccccccccccccccccccccccccccc &&;?x?????+++++?+?+++++ ?vVVV5--5-?-?-?-?-?-?s?-ss{6?	28.8	3.7	0.0	0.0
\times	\checkmark	\times	??nnnnnnnnnnnnnnnnnnnnnnnnnnnn mfe mf rerrrrrrrr e an e ao e a e ho e "h"p t t t t t t t t ' t h e e a	80.6	8.6	0.0	0.0
\checkmark	\times	\times	1x????????????? ????orann Bonererennnerere ?Sh?????orann unnngenngHag g g?e?????????	27.0	7.9	2.0	0.0
\checkmark	\checkmark	\times	The prope prof ot prote was the Wy rronsy ales ale a Claie of th Price was one of the plaids rom	68.1	24.5	4.4	0.5
\times	\checkmark	\checkmark	The increase is a bilday in the Sment used a last give you last She was the intervice is orced t	79.4	44.6	11.5	0.7
\checkmark	\checkmark	\checkmark	Republicans friends like come ti Researchers have played people a The Catalian Office of the docum	87.7	54.1	19.2	3.8
\checkmark	\checkmark	\checkmark	Sequences of length 64. Examples in Table 3.	87.5	51.3	15.1	1.7

Table 3: Samples of length 64 generated by the CL+VL+TH model.

Marks live up in the club comes the handed up moved to a brief d The man allowed that about health captain played that alleged to If you have for the past said the police say they goting ight n However , he 's have constance has been apparents are about home The deal share is dipled that a comments in Nox said in one of t Like a sport released not doing the opposition overal price tabl

samples when using RNNs.

In all our experiments, we used single layer GRUs for both the discriminator and generator. The embedding dimension and hidden state dimension are both of size 512.

Following Gulrajani et al. (2017), we train all our models on sequences whose maximum length is 32 characters. Table 1 shows results of the baseline model of Gulrajani et al. (2017), and Table 2 presents results of our models with various combinations of extensions (Curriculum Learning, Variable Length, and Teacher Helping). Our best model combines all of the extensions and outperforms the baseline by a wide margin on all metrics.

The samples show that models that used both the Variable Length and Teacher Helping extensions performed better than those that did not. This is also backed by the empirical evaluation, which shows that 3.8% of the word 4-grams generated by the CL+VL+TH model also appear in the held-out test set. The weak performance of the curriculum learning model without the other extensions shows that curriculum learning by itself does not lead to better performance, and that training on variable lengths and with Teacher Helping

is important. We note that curriculum learning did not perform well at generating sequences of length 32, but did perform well at generating sequences of shorter lengths earlier in the training process. For example, the model that used only curriculum learning had a %IN-TEST-1 of 79.9 when it was trained on sequences of length 5. This decreased to 59.7 when the model reached sequences of length 10, and continued decreasing until training stopped. This also shows the importance of Variable Length and Teacher Helping.

Finally, to check the ability of our models to generalize to longer sequences, we generated sequences of length 64 with our CL+VL+TH model, which was trained on sequences of up to 32 characters (Table 3). We then evaluated the generated text, and this evaluation shows that there is a small degradation in performance (Table 2).

6 Conclusion

We show for the first time an RNN trained with a GAN objective that learns to generate natural language from scratch. Moreover, we demonstrate that our model generalizes to sequences longer than the ones seen during training. In future work,

we plan to apply these models to tasks such as image captioning and translation, comparing them to models trained with maximum likelihood.

References

- M. Arjovsky, S. Chintala, and L. Bottou. 2017. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. ACM, pages 41–48.
- Tong Che, Yanran Li, Ruixiang Zhang, R. Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. 2017. Maximum-likelihood augmented discrete generative adversarial networks. *arXiv preprint arXiv:1702.07983*.
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.
- Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.
- Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48(1):71–99.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 2672–2680.
- Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*.
- R Devon Hjelm, Athul Paul Jacob, Tong Che, Kyunghyun Cho, and Yoshua Bengio. 2017. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*.
- Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.
- A. Lamb, A. Goyal, Y. Zhang, S. Zhang, A. Courville, and Y. Bengio. 2016. Professor Forcing: A New Algorithm for Training Recurrent Networks. *arXiv preprint arXiv:1610.09038*.
- Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.
- Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P. Xing. 2017. Recurrent topic-transition GAN for visual paragraph generation. *arXiv preprint arXiv:1703.07022*.
- Tomáš Mikolov. 2012. *Statistical language models based on neural networks*. Ph.D. thesis, Brno University of Technology.
- Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. *arXiv preprint arXiv:1703.10476*.
- Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pages 1017–1024.
- L. Wu, Y. Xia, L. Zhao, F. Tian, T. Qin, J. Lai, and T.-Y. Liu. 2017. Adversarial Neural Machine Translation. *arXiv preprint arXiv:1704.06933*.
- Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2017. Improving neural machine translation with conditional sequence generative adversarial nets. *arXiv preprint arXiv:1703.04887*.
- Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2016. Seqgan: Sequence generative adversarial nets with policy gradient. *arXiv preprint arXiv:1609.05473*.
- Yizhe Zhang, Zhe Gan, and Lawrence Carin. 2016. Generating text via adversarial training. In *NIPS Workshop on Adversarial Training*.