CS482/682 Final Project Report Group 2

GANs for Natural Language Generation

Ayush Dalmia (adalmia1), Mary Joseph (mjosep19), Jason Wong (jwong62), Taryn Wong (twong42)

May 10, 2020

1 Introduction

Background GANs have achieved impressive results in computer image generation. However, language-related tasks are a relatively new problem space due to the architecture's poor performance on discrete data.

Related Work Schultz 2019 [6] proposes using deep convolutional GANs on sentence "images" composed of word embeddings for natural language generation. We seek to improve Schultz' results through experimentation with different embeddings, modification of "image" construction and convolution, and hyperparameter tuning of the discriminator.

2 Methods

Dataset To compare with Schultz's WGAN 2 model, we use the same training data and preprocessing method[1]. "Image" inputs are constructed from a subset of English sentences from the 2009 News Crawl dataset, which are tokenized then converted to arrays of concatenated word embeddings. Sentences exceeding 50 words or contain OOV[2] tokens are rejected. Sentences with fewer than 50 tokens are padded with 0-vectors. A subset of 250,000 "images" is used for training.

Setup, Training and Evaluation 50-dimensional GloVe embeddings were used [5] for token encoding due to the inclusion of stop words and compact size. This allows for the production of grammatical output and the convolution of a square matrix. Conversion of the generator's output to sentences used cosine similarity with the embedding dictionary.

To control performance comparisons, we use Schultz' best-performing model: the WGAN-GP model[2]. The WGAN-GP architecture uses a 2-d convolution with residual blocks. The model is trained for 10,000 iterations with 5:1 discriminator to generator iterations. Both the generator and discriminator have learning rates of 1e-4 and use a combination of WGAN loss and a gradient penalty.

Models are evaluated based on the average BERT perplexity [1] and BLEU scores [4] of 100 sample sentences. The BERT perplexity measures well-formedness of outputs, with syntactically-correct and coherent English sentences earning a score in the range 50-60[1]. The BLEU score tests for suffering from mode collapse[3] by measuring sentence diversity.

To provide a baseline for future comparison, we reproduce Schultz' WGAN 2 results, matching his BLEU score. Schultz did not include the pre-trained BERT model used, so we use our own BERT perplexity for comparison of later models' performance.

Exploring Higher-Dimensional Embeddings We investigate the impact of increased dimensionality and other embedding methods on output coherency and grammaticality. In generalizing Schultz' experiments to higher-dimensional embeddings, results suffer when attempting to maintain square inputs for convolution, likely due to a higher proportion of padding in real examples. Schultz' project assumes square matrices, so we modify his model to allow for convolution of non-square matrices.

Exploring Different Pre-Trained Embeddings We also test Spherical embeddings[3], which have a vocabulary five times the size of GloVe's. Spherical embeddings, like GloVe embeddings, contain stop words. While GloVe are trained using global statistical co-occurrence metrics, Spherical embeddings are trained on the unit sphere. These embedding have been shown to perform better on a variety of similarity tasks[3]. As we use cosine similarity to convert the generated embedding matrix to a sentence, we suspect these embedding may improve output coherency.

Hyperparameter Tuning Generator and discriminator training loss plots [4] show that, while the models converge to stable values and do not experience mode collapse, the discriminator trains too quickly, limiting its ability to provide informative feedback to the generator. We experiment with two types of hyperparameter tuning to address this. In the original model (experiments 1-6), for each of the generator's training iterations, there are 5 discriminator training iterations. We experiment with different ratios and slow the discriminator's learning rate.

[1] all code is available at `https://github.com/adalmia96/DL_Project`

[2] out of vocabulary words

[3] mode collapse results in outputting the same grammatical sentence each time

[4] see appendix

3 Results

Higher-Dim. GloVe Embeddings		
Matrix dim.	BLEU Score	BERT Perp.
50 x 50	0.0078	19197.44
100 x 100	0.0079	47411.16
100 x 50	0.0077	3759.55
200 x 50	0.0082	22882.45

Different Pre-Trained Embeddings		
Model	BLEU Score	BERT Perp.
50 x 50 GloVe	0.0078	19197.44
50 x 50 Sph.	0.0079	50625777.32
100 x 50 GloVe	0.0077	3759.55
100 x 50 Sph.	0.0078	1077800.97

D:G Training Iter. on 100 x 50 GloVe Model		
D:G	BLEU Score	BERT Perp.
5:1	0.0077	3759.55
3:2	0.0079	28234.965
2:2	0.0127	10787.72
1:2	0.0103	5556.70
1:5	0.0079	6510.86

Discriminator LR on 100 x 50 GloVe Model		
Learning Rate	BLEU Score	BERT Perp.
1e-4	0.0077	3759.55
1e-5	0.00819	18548.719

4 Discussion

Exploring Higher-Dimensional Embeddings The 100 x 50 "image" outperforms previous square matrix models. When testing 200-dimensional embeddings with a 200 x 50 image, however, outputs are less well-formed, likely due to the increase in parameters.While results were diverse, they were not well-formed.

In a qualitative evaluation of a sample of 100 generated sentences, we noticed a few trends. The GloVe 50 x 50 and 100 x 50 models, for example, output sentences in which many tokens were semantically similar. Sentence A's tokens follow a geographic trend, while B's contain many economic and political tokens. The models also displayed repetition of tokens, most commonly of punctuation and stop-word tokens, as shown in sentences C and D.

A: One s he turned the line he , in , in he saw put offer wolfsburg , , sweden while but but never . yaroslavsky ntuli

B: the union week increase up put higher 2.25 balance of points present rose same inches vaulted 27-nation one same the the 5 as the nation amertil monetary make which the contractors

C: " . is this understood , seems ambivalent would . life . .

D: the after of meanwhile fall one change change the from the we just both the the for the no , agenda . but having , come however " , wake herminator , still when 1771 . .

Exploring Different Pre-Trained Embeddings As spherical embeddings have a vocabulary nearly five times the size of GloVe's, poor output may be due to the generator allotting a larger proportion of training time to exploration of different vocabulary tokens rather than different syntactic structures. This may be supported by qualitative analysis of sample output from the Spherical 50 x 50 model, such as sentence E, in which sentences feature many uncommonly-used words.

E: impracticable vermeer ronglu ours continental colourist and answerable keenly , crude and anything . all and cruelties

Hyperparameter Tuning Neither method of hyperparameter tuning led to an improvement in perplexity over the baseline of the 100 x 50 model with a 5 to 1 discriminator to generator training iteration ratio. There is a visible trend, however, in that an increase in generator training leads to more well-formed output. This observation is promising and further exploration may prove fruitful.

Future Work We faced memory constraints[5] when testing higher-dimensional embeddings so, with either more resources or more streamlined models, we would like to further investigate the effect of embedding vocabulary size on output coherence, as this may cause poor Spherical embedding results. We would also like to test the impact of the introduction of context via contextual embeddings. One challenge with contextual embeddings would be determining if an aggregated vector for each word is a sufficient representation. In addressing the issue of a quickly-training discriminator, we would be interested in experimentation with different GAN models and the use of negative training examples, such as nearly grammatical sentences produced by ESL speakers.

[5] using a Nvidia Tesla V100 GPU with 32GB GPU memory on Google Cloud

References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805.

[2] Ishaan Gulrajani et al. "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5767–5777. URL: http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf.

[3] Yu Meng et al. "Spherical Text Embedding". In: *ArXiv* abs/1911.01196 (2019).

[4] Kishore Papineni et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: (Oct. 2002). DOI: 10.3115/1073083.1073135.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

[6] Robert D Schultz and Jr. *Generative Adversarial Networks and Word Embeddings for Natural Language Generation*. URL: https://academicworks.cuny.edu/gc_etds/3032/.

Appendix

Generator and discriminator loss plots show that, while the models converge to stable values and do not experience mode collapse, the discriminator trains too quickly, limiting its ability to provide informative feedback to the generator.