# **Transfer Learning for Style-Specific Text Generation**

Katy Ilonka Gero, Giannis Karamanolakis, and Lydia Chilton Computer Science Department Columbia University {katy, gkaraman, chilton}@cs.columbia.edu

## Abstract

Neural networks have shown promising results for generating text for creative uses. However, current methods require large data sets to generate coherent sentences, which severely limits their creative potential given that the majority of stylistic literary data sets are relatively small. We build on recent advances in transfer learning for natural language processing and demonstrate that generic pre-trained language models can be effectively fine-tuned on small stylistic corpora to generate coherent and creatively expressive text. We empirically show the effectiveness of this method across three distinct literary styles where only a small (e.g. less than 6k) number of tokens are available. We suggest further work for understanding and improving this process, and release our code online<sup>1</sup>.

# 1 Introduction

Procedural generation of text for artistic purposes has a long history, from the cut-up technique of the Dadaists to web-based digital poetry. Neural networks have shown promising results for generating text in this tradition [6, 9, 8] but current methods require large data sets, typically > 1 tokens [3, 2]. Yet most sources of literary text are relatively small – a novel often has less than 100k tokens.

Transfer learning, in which parts of the neural network can be trained on a larger and more 'general' data set, has only recently seen success in natural language processing. Work has shown that pre-trained language models shared across tasks can improve performance [10, 1, 5]. In this paper, we build on previous work by fine-tuning a pre-trained language model using very small data sets. We qualitatively show that pre-training results in coherent generation for three styles, and demonstrate the impact of allowing the model to retain out-of-domain vocabulary.

# 2 Methods

We use a simple yet effective neural language model architecture AWD-LSTM [7] and utilize pretrained weights released for this architecture. For data, we use three stylstic corpora: *imaginative* (the novel Alice in Wonderland), *highbrow* (the essay Consider the Lobster by D. F. Wallace), and *poetry* (poems by Emily Dickinson). We split the style data into sentences and use each sentence as a training example, reserving 30% of sentences for validation. To fine-tune, we first train just the last layer of the model for 1 epoch, then train the entire model for 30 epochs. Further details about the model architecture and fine-tuning procedure are in the supplementary material.

In our first experiment, we train the model with and without the pre-trained weights for each of our three style data sets and qualitatively compare the generated text in each case. We include only the words in the style data. In our second experiment, we keep the vocabulary of the pre-trained model and show when the model uses vocabulary not in the style data set.

<sup>&</sup>lt;sup>1</sup>https://github.com/kgero/style-gen

Style	Tokens	Condition	Example Outputs
imaginative	25k	raw	saying my head – that which it . t not rabbit ; went near ' us a business ! i , and under , – ' violently '
imaginative	25k	pre-train	' do you know what you're saying $!$ ' said alice . caterpillar lobsters $!-$
highbrow	5k	raw	the steak holes immersed the " to are the . it like it there thanks lobster cerebral an .
highbrow	5k	pre-train	( there also appear to be some touristic stimuli that make the crab turn carnivore , though some do b&bs dislike this demotic thing . )
poetry	25k	raw	the chair pleasure length that passed bird n't. so smiling he dissent a ? in., stand
poetry	25k	pre-train	after god , surpasses god have no nearer than heaven ; but heaven had not daffodils ,

Table 1: Generated text from three distinct literary styles, with and without pre-training. Tokens indicate the total number of words in the training set. (Experiment 1)

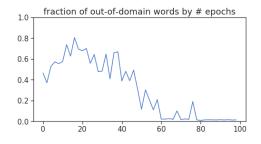


Figure 1: Out-of-domain vocabulary usage in generated text as a function of training epochs for the *highbrow* data set. At each epoch, we generate 10 sequences of 250 words and average the number of out-of-domain words used in each of the 10 sequences. Too much training results in the model 'forgetting' out-of-domain vocabulary. (Experiment 2)

#### **3** Results and Discussion

Table 1 shows results from our first experiment, comparing generated text with and without pretraining. Additional outputs can be found in the supplementary material. Qualitatively we see that the pre-training results in longer-term grammatical and semantic coherence. While some of the distinctiveness comes from the vocabulary choice, much of it comes from other aspects of style, like syntax and punctuation. It is notable that phrases are rarely lifted directly from the source text.

In the case of *imaginative*, the model picks up on the common usage of quotes, exclamation points, and question marks in Alice in Wonderland, as well as common characters like Alice herself, the duchess, and jurors, and common themes like growing/shrinking and animals.

In the case of *highbrow*, the model learns to generate longer sentences, have parenthetical remarks, and use common words less frequently, as Wallace does, as well as stay on the theme of tourists, lobsters, and gourmet food, as in the essay.

In the case of *poetry*, the model learns to use line breaks in a similar style to Dickinson, title poems with roman numerals, and use em-dashes mid-sentence. (Without pre-training, the model does generate many line breaks, but rarely does it generate consistent line lengths. This can be seen in the supplementary material.)

It is worth considering how to allow a model to use out-of-domain words while retaining stylistic distinctiveness. Figure 1 shows the results from our second experiment, in which we fine-tuned with the *highbrow* data but kept the entire vocabulary of the pre-trained model. A point of interest is around epoch 40, when the model begins to generate less out-of-domain words. This points to the potential of this process to control or combine styles.

We hope that by demonstrating how pre-trained language models can be effectively tuned with small data sets, we inspire experts and novices alike to continue to experiment with novel and varied data sets. Further work could explore optimal training regimes or how to make this method more accessible, for example by embedding it in a real-time web interface.

#### Acknowledgements

Katy Ilonka Gero is supported by an NSF GRFP (DGE - 1644869).

#### References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [2] Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. *CoRR*, abs/1707.02633, 2017.
- [3] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. *CoRR*, abs/1711.06861, 2017.
- [4] Jeremy Howard et al. fastai. https://github.com/fastai/fastai, 2018.
- [5] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 328–339, 2018.
- [6] A. Karpathy. The unreasonable effectiveness of recurrent neural networks. https://karpathy.github.io/2015/05/21/rnn-effectiveness/. Accessed: 2018-10-25.
- [7] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. *CoRR*, abs/1708.02182, 2017.
- [8] A. Parrish. Articulations. Counterpath, 2018.
- [9] O. Sharp and R. Goodwin. Sunspring | a sci-fi short film starring thomas middleditch. https: //www.youtube.com/watch?v=LY7x2Ihqjmc. Accessed: 2018-10-25.
- [10] Kelly W Zhang and Samuel R Bowman. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. arXiv preprint arXiv:1809.10040, 2018.

# **Supplementary Material**

### **Further Example Outputs**

Table 2: Further examples of generated text from three distinct literary styles, with and without pre-training. Tokens indicate the total number of words in the training set.

Style	Tokens	Condition	Example Outputs
imaginative	25k	raw	had down , there same be said burn gryphon the , , be lessons said ' who , would well it to rabbit began of over going child the be ' done , not deal my quite time fun speak , the to ; and ' ' so of do dinah , when first in all but ! turtle to all ! the
imaginative	25k	pre-train	' soon , i ' ve got your shoes up . ' the young kid began business . he was looking at his tongue , and he began to feel a little trembling . nibbled ' come away , my dear ! i 'll set end my head ! ' she said to herself . ' i ' m sure i ' m not glad they 'll get me executed , for this lap i ' m now . oh my dear pepper , my dear .'
highbrow	5k	raw	the component rationales their to issue 's imagine least known honors in boiling is three palates out and than gustatory on corrupt least brittlely their ) get most pain torture are .
			she ground taste pop violent by smuggled for , - portions along of in - guy springs their , me , the can or permit out their saying , the , $\dots$
highbrow	5k	pre-train	the word " loppestre " is latin for " false . "
			it 's hard to deny in any meaningful way that this one knows what the pain is . even if you cover the 'm , it 's hard to deny that there is any moral distinction between pain and pain .
			she had one born in new york town , but neither of these generations came up with any means of defending the world 's largest lobster tbsp and that much of the aforementioned styrofoam intimacy is still gone , as is the absence of troubling nyquil -
poetry	25k	raw	wondered t gain ; , ,
			us mine troth morn warm . me the tenderer ;
			are amid his me, perceived play. star
			– not
			my have clear – the
			but i was on dimples ' go , smile ! ? it but , return centuries , – that will xxviii , sometimes attitudes her
poetry	25k	pre-train	xv. i used to when man was not afraid of me, he used to himself fingers, and i, and unto me, – and i just wanted the thing.
			i note that i can not tell – the spot where the calvary went,
			nor ever drop the purple xix through the woe, nor that the rise has ceased;
			x. to dare is dear that idle place,

#### **AWD-LSTM Fine-tuning**

The language model in our experiments is an AWD-LSTM [7] with an embedding layer of dimensionality 400, and 3 hidden layers of dimensionality 1150 each. We initialize the model's weights using the pre-trained weights of the same model architecture trained on the WikiText 103 dataset. The WikiText 103 dataset is a pre-processed subset of English Wikipedia and consists of more than 103 million training tokens<sup>2</sup>.

We fine-tune the pre-trained AWD-LSTM on each of our three target datasets (*imaginative*, *poetry*, and *highbrow*) inspired by the procedure described in [5], which has been shown to be successful on various NLP tasks <sup>3</sup>. In particular, we first find the best configuration for the learning rate. Then, we tune the last layer of the model for 1 epoch by freezing the weights of all but the last layers, and finally, we unfreeze all the weights and train the entire model for 30 epochs. The cross entropy loss is tracked during fine-tuning.

<sup>&</sup>lt;sup>2</sup>Detailed description of the dataset and a download link are provided in the following link: https://einstein.ai/research/blog/the-wikitext-long-term-dependency-language-modeling-dataset

<sup>&</sup>lt;sup>3</sup>Our experiments were performed using the fast.ai library [4] and are an adaptation of the deep learning tutorial presented as a part of the library: https://github.com/fastai/fastai/blob/master/courses/dl2/imdb.ipynb