Enhancing Controllability of Text Generation

Anton Shcherbyna[1] and Kostiantyn Omelianchuk[2]

[1] Ukrainian Catholic University, Faculty of Applied Sciences, Lviv, Ukraine
a.shcherbyna@ucu.edu.ua
[2] Grammarly, Kyiv, Ukraine
komelianchuk@gmail.com

Abstract. There are many models used to generate text, conditioned on some context. However, those approaches do not provide an ability to control various aspects of the generated text like style, tone, language, tense, sentiment, lengths, grammaticality, etc. In this work, we are exploring unsupervised ways to learn disentangled vector representations of sentences with different interpretable components and trying to generate text in a controllable manner based on obtained representations.

Keywords: natural language processing · natural language understanding · representation learning · text generation · unsupervised learning

1 Introduction

1.1 Text Generation Overview

In recent years, there was a significant advancement in the field of text generation. In 2014, sequence-to-sequence models with LSTM encoder and decoder were proposed [1]. This approach became state-of-the-art in the field and was successfully used for various tasks e.g., machine translation. However, LSTM networks tend to forget information from the whole sequence, so the next significant improvement – attention mechanism – was proposed [2]. The main idea of this approach is to provide a decoder with the information from each token from the source sequence directly and score each piece of information by usefulness for the decoder. Finally, a pure attentional model, which is called Transformer, was proposed [3]. Since then, transformer-like models became the State-of-the-Art methods in text representation learning and text generation. For example, BERT released by Google [4] became the standard for extracting representations from texts and GPT-2 developed by OpenAI [5] became the most powerful tool for text generation. In the case of GPT-2, the authors provided weights only for a small model with limited capabilities. They mentioned that their model is capable of producing such high-quality texts, so they have a fear that somebody can use this model to produce fakes.

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 98–105

All those models have similar structure. Typical text generation model consists of an encoder $E_\theta(x)$ and decoder $D_\varphi(h)$. Both an encoder and decoder can be represented as a deep neural network: LSTM [1], CNN [6], or a stacked feed-forward network, which forms a transformer-like model [3]. An encoder extracts information from the source sequence $\{x_i\}$ into hidden representations $\{h\}$ and then a decoder produces target sequence based on those representations (Fig. 1). Such models are trained end-to-end and use various training signals. For example, we can force an encoder to encode one sentence and decoder to produce the next sentence from the same text. Or we can use the so-called ”hidden language model” approach when our sequence-to-sequence model is forced to predict intentionally deleted tokens from the source sequence. In both cases, we use classic categorical entropy between the distribution predicted by the network and the true distribution as a loss function.

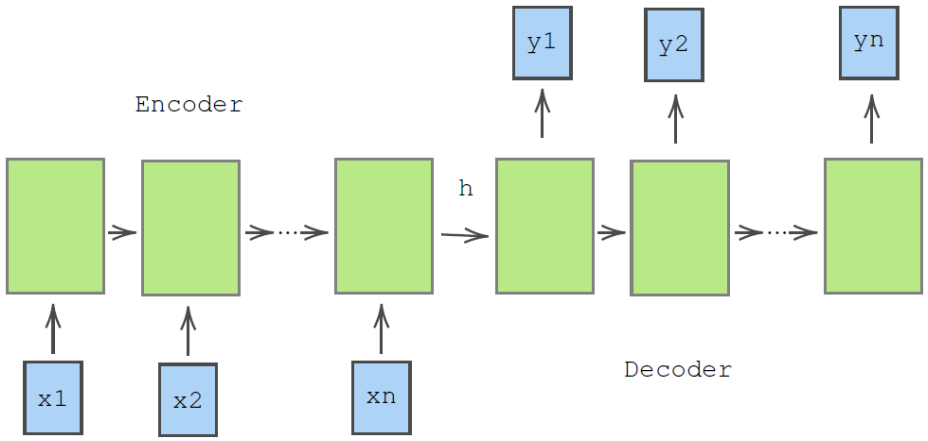

Fig. 1. A **s**imple sequence-to-sequence model

However, all those approaches lack one crucial property – controllability. By controllability, we mean an ability to change the attributes of the generated text such as sentiment, length, complexity, etc. The models described above are conditioned only on the text they saw previously, which is uninterpretable and unpredictable controllable parameter. Furthermore, the space of hidden representations of such models is unsmooth [7]. It means that we cannot interpolate in the latent space to discover dependencies between different hidden representations and generated text. Another problem is that such representations capture information about text attributes alongside with the context, and we want to manipulate only using attributes. This problem limits the usage of such models for modern applications like dialog systems or question-answering systems.

Also, it’s worth to note that currently transformer-like models outperform old models based on LSTM, but they are harder to train, require much more training data and computational resources. Hence, we focus on LSTM-based models. Moreover, there is no difference between LSTM and transformer-based models in terms of our problem.

Therefore, we can transfer all the methods developed for LSTM to transformer-like models.

1.2 Useful Approaches from Vision Domain

There was considerable progress in the direction of controllable generation in the vision domain. VAE [8] extends a classical auto-encoder with probabilistic argumentation and gives the ability to control generation by exploring the latent space. For this purpose, we define a latent variable $z \sim p_z(z)$, which has some probabilistic prior distribution (typically Gaussian). Then we define some complex posterior conditional distribution $x \sim p_\theta(x|z)$ (typically it is a Gaussian distribution with mean and variance expressed by a neural network with parameters θ). Now we can define the likelihood:

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz, \tag{1}$$

which appears to be intractable, so we cannot optimize it directly. However, there is a solution. We introduce a new conditional prior distribution $q_\varphi(z|x)$ (similar to posterior) parameterized by a neural network with parameters φ (Fig. 2). This allows us to derive a lower bound on the data likelihood that is tractable, so we can optimize it using gradient descent:

$$\mathcal{L} = \mathbb{E}_{z\sim q_\phi(z|x)} log p_\theta(x|z) - KL(q_\phi(z|x)||p(z)). \tag{2}$$

Now we can encode source sample to latent space, tweak the latents and decode it.

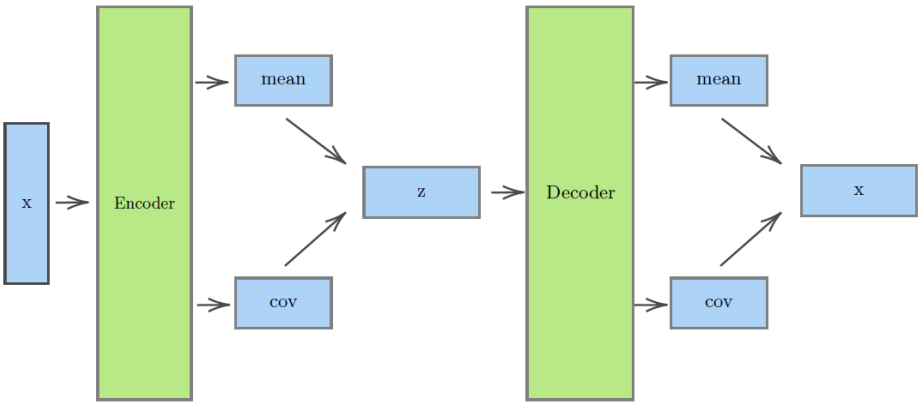

Fig.2. VAE architecture with discriminator

2 Problem Setting

To make text generation more controllable, we want to incorporate VAE-like approach from vision domain into text generation. From the first view, it looks straightforward, but we have to face a couple of problems:

1. VAE expressive power is limited due to the restriction we put on the posterior distribution [9]
2. VAE often faces posterior collapse. It means that a strong decoder tends to ignore latent codes during generation [7].

Furthermore, even if we solve those two problems and successfully incorporate VAE, we will still face the other crucial problem: we extended the sequence-to-sequence model with meaningful latent space, but latent codes are highly entangled, so it is hard to change each attribute separately. Moreover, such latents also capture context information, which is undesirable. Therefore, we need to find a way to make those representations disentangled.

3 Related Work

3.1 Deal with Entanglement, Supervised Way

To solve the problem with high entanglements of latent codes we can extend our VAE model with additional discriminator network [10]. In this work, the authors augmented a latent code z with additional part c: z is responsible for encoding context information as in classic approach; c is forcefully disentangled and each its component captures attribute information. It works as follows: the encoder produces a latent pair (z, c), then the decoder generates a sample $\hat{x}$, which is encoded by the encoder to get $\hat{c}$. The discriminator is used to distinguish between c and $\hat{c}$. The signal from the discriminator is used to update the weights of the decoder (Fig. 3).

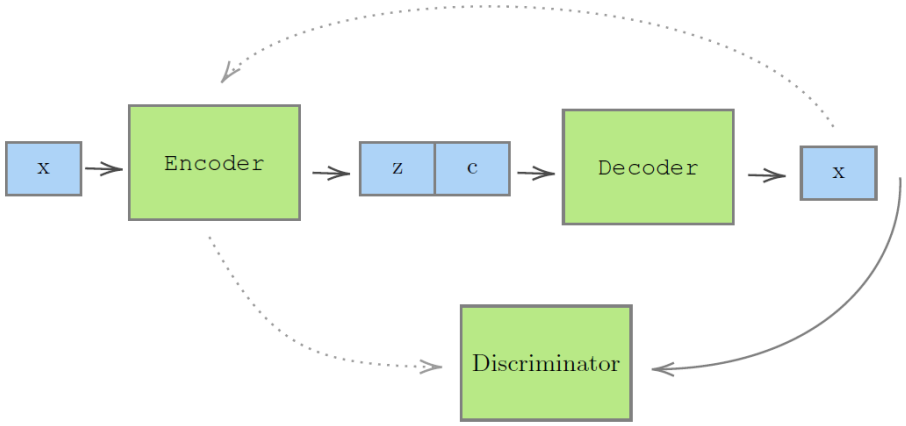

Fig.3. VAE architecture with a discriminator

Also in this paper, the authors proposed the method to deal with the discrete nature of text. The decoder, at each step, produces the probability distribution function parameterized by *softmax* over possible tokens and then the token with the highest probability

is selected. For discriminator training, we may leave this parameterized probability distribution and control it with the temperature parameter τ:

$$\hat{x} = softmax(\frac{h_t}{\tau}) \tag{3}$$

There are three big problems with such approach. Firstly, we need to build a separate discriminator for each attribute. Hence, the complexity of the model grows significantly, as we add new attributes. The second problem is that we need to get data to pretrain discriminators and for some attributes, such as complexity, it might be difficult. The third problem is that there are no solutions for limited expressive capabilities of the Gaussian posterior and posterior collapse problems.

3.2 Dealing with VAE Problems in an Unsupervised Way

In the other work [11], the authors presented a fully unsupervised approach, which attacks all the three problems. They proposed sample-based representations, which are more expressive than Gaussian posterior, and called their approach Implicit VAE (or iVAE). They defined a sampling mechanism instead of using explicit Gaussian and thus represented the distribution generated by the encoder as the set of latents:

$$z = q_\theta(x, \epsilon), \epsilon \sim q(\epsilon), \tag{4}$$

where $p(\epsilon)$ is a Gaussian and q is the concatenation of the hidden state of the encoder and ϵ. In this case, the KL-divergence $KL(q_\varphi(z|x)||p(z)$ became intractable, but we can represent it using a dual form:

$$\mathbb{E}_{z\sim q_\phi(Z|X)} v_\psi(x,z) - \mathbb{E}_{z\sim p(z)} \exp(v_\psi(x,z)). \tag{5}$$

Then the final loss function looks as follows:

$$\mathcal{L} = \mathbb{E}_{z\sim q_\phi(Z|X)} log p_\theta(x,z) - \mathbb{E}_{z\sim q_\phi(Z|X)} v_\psi(x,z) + \mathbb{E}_{z\sim p(z)} \exp(v_\psi(x,z)). \tag{6}$$

Also, the authors described the solution to the posterior collapse problem. Posterior collapse means that a strong decoder ignores the dependency on latent codes. As a result, the distribution generated by the encoder $q_\varphi(z|x)$ exactly matches $p(z)$. To overcome this problem authors proposed to add stronger regularization on the latent space by changing the KL-divergence that we used previously to the new one:

$$\mathcal{L}_{MI} = KL(q_\phi(z)||p(z)), \tag{7}$$

where $q_\phi(z) = \int q_\phi(z|x)q(x)dx$ – aggregated posterior. This approach is called Implicit VAE with mutual information. Those improvements helps us solve problems with VAE and learn latent codes fully unsupervised, but those representations are still entangled.

3.3 Dealing with Entanglement: Partially Solving VAE Problem in an Unsupervised Way

The problem of entanglement was attacked in [12]. The authors proposed to use two different encoders and split latent codes into two parts z_1 and z_2. Then, the first encoder will be forced to capture the global variations in the data, which correspond to the attributes that we want to control. The second part will capture context information useful for reconstruction purposes.

Then they decided to constraint the latent space for z_1 to have the following structure:

$$z_1 = \sum_{i=1}^{K} p_i \, e_i, \sum_{i=1}^{K} p_i = 1, \tag{8}$$

where e_i are learnable vectors and p_i can be obtained through the scoring procedure:

$$p = softmax(W\hat{z}_1 + b), \tag{9}$$

where $\hat{z}_1$ is a classic posterior obtained from $q_{\psi_1}(z|x)$. In other words, we want to learn a set of basis vectors and then obtain the latent code as a linear combination of such vectors. Those basis vectors e_i in such a setting tend to capture global variations in the data and it is easier for the decoder to generate sentences because latent codes are just the combinations of such basis vectors.

This model is trained as a typical VAE, but to train the parameters W, b and e_i an additional term is introduced:

$$\mathcal{L}_{rec} = \mathbb{E}_{z_1 \sim q_{\psi_1}(z_1|x)}\left(\frac{1}{m}\sum_{i=1}^{m} \max(0.1 - \hat{z}_1 z_1 + \hat{z}_1 u_i\right), \tag{10}$$

where m is the number of samples from the data and u_i are the latent codes of those samples. However, only this loss is incapable of forcing orthogonality of the basis vectors, so one more term was introduced:

$$\mathcal{L}_{ort} = \|E^{\mathrm{T}}E - I\|. \tag{11}$$

The authors showed that with such structural constraint there is a small chance that there will be a posterior collapse. However, there is still a problem with the expressive capabilities of VAE. Moreover, authors added additional constraint, which can limit those capabilities even more.

4 Research Goal and Evaluation

The main goal of the master thesis is to empirically evaluate the approaches described above and combine them to build the model, which will be capable of solving all the problems we defined in problem setting. Then we want to explore the latent space and discover which attributes of the text the model was able to capture

The proposed model will consist of the LSTM encoder and decoder with VAE mechanism between them, implicit sampling-based posterior, and the constraint on the resulted latent. Let us breakdown the whole process into the following steps:

1. First, we will take a source sequence and encode it into the two hidden vectors h_1 and h_2
2. Then we will add noise to those vectors to obtain the pairs (h_1, ϵ_1) and (h_2, ϵ_2), where ϵ is a Gaussian (as described in 3.2)
3. Next we will propagate this vector through MLPs to obtain $\hat{z}_1$ and z_2 (as described in 3.2)
4. Finally, we will use $\hat{z}_1$ to calculate the scores p_i for the final latent calculation:

$$z_1 = \sum_{i=1}^{K} p_i e_i \tag{12}$$

5. Now we can use concatenated (z_1, z_2) for further text generation.

Evaluation of this approach will be done via solving style transfer problem on Yelp dataset. We will measure:

1. Content preservation (BLEU)
2. Style transfer strength (supervised classifiers)
3. Fluency and correct grammar (perplexity by GPT-2 language model)

5 Research Plan

We plan to organize further work in the following way:

1. Implement and test unsupervised approach based on Implicit VAE with MI regularization
2. Implement and test unsupervised approach based on latents as linear combination of basis vectors, which incorporates global variation from data
3. Add implicit latent learning to the second approach
4. Explore latent space and find attributes, which model captured
5. In case of a success, we will extend those models to be transformer-like with more powerful encoder and decoder

6 Conclusion

In this master's thesis proposal, we made an overview of the current state in the field of text generation and described VAE, which is used for controllable generation in the vision domain and is applicable in text domain. We defined the most crucial problems: issues with VAE itself (its expressive limits and posterior collapse) and difficulty with an entanglement of latents. Further, we made related works overview and proposed our potential solution, which is based on the combination of implicit posterior distribution and constraint on the resulted latent in the form of the linear combination of basis vectors. We believe that this improvement can increase the degree of controllability and quality of the resulting samples.

References

1. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. arXiv preprint arXiv: 1409.3215 (2014)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473 (2015)
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv: 1706.03762 (2017)
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805 (2018)
5. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8) (2019)
6. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. arXiv preprint arXiv: 1705.03122 (2017)
7. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. arXiv preprint arXiv: 1511.06349 (2015)
8. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv: 1312.6114 (2013)
9. Cremer, C., Li, X., Duvenaud, D.: Inference suboptimality in variational autoencoders. arXiv preprint arXiv: 1801.03558 (2018)
10. Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing E.P.: Toward controlled generation of text. arXiv preprint arXiv: 1703.00955 (2017)
11. Fang, L., Li, C., Gao, J., Dong, W., Chen, C.: Implicit deep latent variable models for text generation. arXiv preprint arXiv: 1908.11527 (2019)
12. Xu, P., Cao, Y., Cheung, J.C.K.: Unsupervised controllable text generation with global variation discovery and disentanglement. arXiv preprint arXiv: 1905.11975 (2019)