Literature Survey

# Paper – 1: Diversity regularised auto encoders

Authors: Hyeseon Ko, Junhyuk Lee, Jinhong Kim, Jongwuk Lee, Hyunjung Shim

Publisher: ACM – 2020

## Summary

The authors propose a new powerful text generation model in their paper, called Diversity Regularised Autoencoders (DRAE). The proposed model can handle insertion, deletion, substitution and masking of words, to generate a new sentence. The authors talk about the other autoencoders that exist already and point out the issues they face. In Variation Autoencoders (VAE) has a challenge of collapsing posterior loss, which is also known as the Kullback-Leibler(KL) vanishing problem. This occurs because VAE relies on the autoregressive properties of the decoder, thereby overlooking the prior distribution of latent variables. This model doesn’t create any diversity in its text-generation, as their parametric formulation of the posterior and prior approximations are designed as simple diagonal Gaussian distributions which makes it limit itself to the complex prior distribution.

The authors have pointed out that adversarially regularised autoencoders(ARAE) can overcome this limitation of simple prior approximations. ARAE can deal with more complicated distributions. It can reduce the posterior collapse problem, but cannot provide a more diversified sentence. The authors propose that the injection of noise(data augmentation) can improve the diversity of the sentence generator, generate new sentences and reduce repetitive words or phrases. The authors add the usage of Wasserstein distance with gradient penalty(GP) for the adversarial train, because of its success in the image generation in WGAN-GP.

GAN based approach for text generation uses the argmax function which represents a one-hot encoded vector and it is non-differentiable, thereby it makes the backpropagation difficult. Reinforcement Learning has been used approximation methods along with GAN. The authors talk about various autoencoders, like the vanilla Autoencoder(AE), Variational Autoencoders(VAE), Adversarial Autoencoders(AAE), and Adversially Regularised Autoencoders(ARAE). Vanilla AE, learns the hidden representation of the given data. The training is aimed at minimizing reconstruction loss. Due to its discrete and sparse nature, it cannot be used for a generative model. VAE is a popular generative model which is built on the AE architecture. VAE uses a reconstruction loss term that forces the reconstruction of the input and imposes a close to standard Gaussian using KL divergence, thereby making it a generative model. Another model AAE uses adversarial training to regularize the latent space, under the GAN architecture. It is trained with two-loss terms, the reconstruction loss and the adversarial loss terms. A regularised AAE, (ARAE) regularises a flexible and complex prior by using adversarial training to match the prior and posterior. The training process contains three-loss terms, reconstruction loss, discriminator loss and generator loss. ARAE only faces the issue of adjusting the right hyperparameter.

The author specifies that the existing models are successful but suffer from the mode collapse, that is, the words or phrases in a sentence are repeated multiple times. This occurs due to the prior distribution being too restrictive and the training processes are biased for the memorization of the training samples.

The authors propose a model where they use ARAE with noise injection into it. This is the DRAE model architecture, and it makes use of noise injection into the encoder, which makes the new and random words to be generated and added to the training dataset. Noise injection prevents the model from memorizing the dataset, thereby adding a little bit of extra diversity. The authors have used the WGAN-GP approach to improve the training stability of the GAN and thereby improving the quality of the text generated. The noise injection strategy is applied to the encoder. The authors take four cases of noise injection – insertion, deletion, substitution, and masking. Training with noisy data helps understand incorrect sentences and thus helping map it to the correct sentence. In insertion, a word is added into the sentence, thereby increasing the length of the original input. In Deletion, a randomly selected word is removed from the input sentence, thereby reducing the length of the sentence. Substitution is performed by replacing a certain word with another word, and Masking is performed by replacing a certain work with an unknown token. Masking is different from deletion as the effective sentence length doesn’t change. The authors also apply all four techniques together and use four parameters to tune for each noise injection strategy. The adversarial training is used to learn the prior distribution using the Wasserstein distance to regularise the discrete autoencoder. The decoder is given the synthetic or real sentence in their fixed length latent space. The model is trained to map the vectors to their respective sentences. The authors have used two datasets for training, Stanford Natural Language Inference and the medium-scale BookCorpus. They have filtered out only the sentences from the entire dataset, as they focus on text generation. They have used three different evaluation methods, the Forward Perplexity, to measure the quality of the generated message, the Reverse Perplexity, to measure the variety of the generated message, and the Self-BLEU metric for the diversity of the generated text.

The evaluation was performed in three steps, sampling, reconstruction and latent spacewalking. In most of the cases for the sampling, the model outperformed the base models, and the insertion model proved to be with higher performance. The reconstruction showed a good range of diversity in the texts generated. The authors performed a latent space walking between two points to ensure that the model has not memorized the dataset, and also to see if the latent space is dense and compact. The model showed meaningful transitions through the latent spacewalking, and there were no inconsistencies in the grammar/semantics.

The authors conclude that the new text generation model was focused on the generation of diverse and fluent sentences. Adopting the noise injection to the encoded, the model yielded a smooth and well-spread latent distribution. The model is simple but yet effective in producing diverse and natural-looking sentences.

## Advantages

This paper modifies the existing model, which enhances the diverse nature of texts that are generated. The model maintains the readability and the grammar on noise injection, thereby it learns to correct grammar along with the training procedure too.

## Disadvantages

The model doesn’t have any drawbacks per se, but in real life, the text is usually repeated a few times. The diversity in the model is great for learning the grammar and producing natural-looking texts, but too much diversity is not common in a normal text.

# Paper – 2: Improving Variational Encoder-Decoders in Dialogue Generation (Reference Paper)

Authors: Xiaoyu Shen, Hui Su, Shuzi Niu, Vera Demberg

Publisher: Association for the Advancement of Artificial Intelligence (AAAI)

## Summary

The authors have talked about improving an already present model for text generation, Variational Encoders-Decoders (VED). They have written about the different types of models that have been used previously, and the issues they’ve faced. They also mentioned the KL-vanishing problem that the previous VAE models faced. The authors have divided the training into two parts, the first phase is to learn to encode discrete texts into continuous word embeddings, and the second phase utilizes these word embeddings to learn the generalization of the latent representations. The latent variables are sampled by a Gaussian noise transformation through a multi-layer perceptron and are trained with a separate VED model, to increase the flexible distribution realization.

Recurrent Neural Networks(RNN) have extensively been used for Natural Language generation, but they only look at one word at a time and does not work from a sentence approach. They either generate short and boring responses or long and inconsistent sentences. They easily deviate from the original context as the length of the sentence generated increases. This model was improved with the introduction of sentence-level representation to maintain a sentence level consistency of the generated sentence. The authors specify that the deep latent variable models are a much better way to learn the representations for generative requirements. They also specify that the exact log-likelihood that requires a high-dimensional space to be expressed analytically, the new approaches solve this problem by approximation of the real posterior probabilities. According to some papers the authors have reviewed, Variation Autoencoders(VAEs) bring scalability to the training method and introduce a reparameterization trick to reduce the uncertainty during backpropagation.

The drawback of an RNN based model is that it is a universal approximator that is more flexible than the simple Gaussian distribution, thereby it lacks the necessity to use the latent variables. Current approaches weaken the RNN decoders so they can utilize the latent variables, but this sacrifices the generative capacity of the model. The author uses the universality approximation of the RNN to make a flexible latent variable distribution, and also increase the strength of the expressiveness of the model. The authors specify that they split the model into a CVAE module and an AE module. The CVAE learns to generate the latent variables whereas the AE module builds the connection between them and the dialogue.

Most of the models have a KL-vanishing problem because the RNN decoder is a universal function approximator and it tends to represent the distribution without the latent variables. The authors explain that the objective function of the VEDs in dialogue generation is a competition of the two items, the generative distribution and the posterior distribution. The reason for KL divergence vanishing in the global optimum is that the second term gains more from ignoring the latent variables than the first term utilizing them.

In current approaches making the decoder weaker, makes it come further away from the real distribution, thus it uses the latent variables. Word drop-out being a common method is a common method to weaken the RNN decoder. In this paper, the authors mention that the idea of Adversarial Auto Encoders(AED) is appealing, but it is very difficult to train, especially when both prior and posterior have to be updated towards each other. The authors replace the GAN phase with the CVAE, where an RNN encoder is applied to extract the corresponding latent variables for each dialogue term, based on this the CVAE is trained to perform reconstruction through a context-dependent Gaussian noise. The authors bring to our notice that in the encoder phase, the model only adjusts the RNN encoder-decoder to control the KL-divergence and the generating parameters for the latent variables are fixed.

The authors used a sampling trick for building a sample for training their data. They have used a ground truth encoding for the AE phase, and then gradually changed to a noisier CVAE output. The authors use a scheduled sampling strategy and decide on a coin flip if they have to feed the real hidden vector or the noisy vector. The model was built by replacing the GAN phase of AED with a CVAE alternative, and the output of CVAE provides the latent variables. These represent the broader distribution than mean-field Gaussian. The CVAE is less accurate than a GAN, theoretically as it needs to approximate the real posterior. The authors leverage a more powerful RNN encoder-decoder with this. The AE phase autoencoder the utterances to make the real posterior representable easily by the CVAE part.

The experiment was performed in two datasets – the Daily dialogue and Switchboard. The authors trained the HRED model with the latent models trained by the standard KL-annealing with different weights, with an additional BW loss, word drop-out, free bits and a collaborative VED with the scheduled sampling trick. The letters are lowercased, vocabulary size was set as 20,000 and the OOV words were mapped to a token <unk>. They performed a metric-based evaluation and a human evaluation. In the metric-based evaluation, the model improved by the authors have achieved the highest topic similarity based on all the three metrics they’ve used – KL divergence, Perplexity, and Negative Log Likelihood. They performed a human evaluation on the Dailydialog corpus. The generated responses along with the dialogue context were shuffled and judged on the crowdsourcing website, CrowdFlower. The results showed that the improved model has a high degree of fluency compared to other approaches. This evaluation was just a complement to the metric-based evaluation, showing that the improved VED model is better than the other models.

In conclusion, the Variational Encoders-Decoders and the recurrent neural networks are powerful representation learners and natural language processors. Experiments show that the improved VED model samples latent variables with more flexible distributions without sacrificing the recurrent neural network’s capabilities of synthesizing coherent sentences.

## Advantages

This model is good in maintaining the context and doesn’t face issues with long sentence generation.

## Disadvantages

This model drops its performance if the BOW size increases. Even though it reduces the KL-vanishing problem, the performance to be in context reduces.