ITG-VAE : Image-conditioned Text Generation with Variational Autoencoder

Korea University COSE461 Final Project

Kanghyeon Kim
Department of Bigdata
Team 20
2018380604

Youngwoo Shin
Department of Bigdata
Team 20
2020380717

Jihyeon Choi
Department of Computer Convergence Software
Team 20
2020270636

Hyeonsu Hwang
Department of Bigdata
Team 20
2019380610

Abstract

We propose a Image-conditioned Text Generation with Variational Autoencoder(ITG-VAE). This model, which uses a method of extracting and concatenating feature vectors from text and corresponding images respectively, can be useful in understanding the sentence by generating images and trimmed sentences related to the images. Beyond simply enhancing our understanding of difficult or long sentences, we also aimed to create an image-text pair dataset that can be used in various fields such as natural language processing.

1 Introduction

While exploring ways to improve understanding of long or difficult sentences, we devised an image conditioned VAE text generation model inspired by VAE sentence reconstruction and MIT's study that images have significantly faster cognitive time than text. We tried to create image-conditioned sentences by combining text feature extracted through transformer encoder from input text and image feature extracted through cnn encoder from image generated based on input text. This can generate evoked sentences based on the image. Through the pair of generated images and sentences, we can improve our understanding of the sentences and further create text-image datasets that can be used for various studies.

2 Related Work

This study focuses on the task of generating images from the text of the PTB dataset through a pre-trained diffusion model and generating new text based on it. To achieve this goal, we propose a model that combines feature vectors extracted from images and texts. This approach was inspired by several ideas devised in recent researches.

2.1 A Transformer-Based Variational Autoencoder for Sentence Generation

In conventional VAE-based models for natural language text, both encoders and decoders are Recurrent Neural Network (RNN)-based, while this study proposes a model that combines Variational

Autoencoder (VAE) and Transformer models to regenerate text data. The key methodology we refer to in this paper is to extract text features from input sentences using Transformer encoders, learn latent distributions through VAE, and regenerate sentences through Transformer decoders. Implementing this methodology can improve the diversity of generated text and relevance to images, using Transformer's strong contextual understanding and VAE's ability to sample from the latent distribution.

2.2 Vision-Language Pre-Training with Triple Contrastive Learning

This paper proposes a novel method for learning similarities between different modalities in visual-language pre-learning. This study extends the concept of 'Contrastive Learning', a learning method that distinguishes between similar (positive) and different (negative) data samples, to 'Triple Contrastive Learning'. The key technique we refer to in this paper is to extract feature vectors of images and texts through visual encoders and text encoders respectively, and combine them in fusion encoders to generate multimodal embeddings. This technique can help improve the interaction between image and text data.

3 Architecture

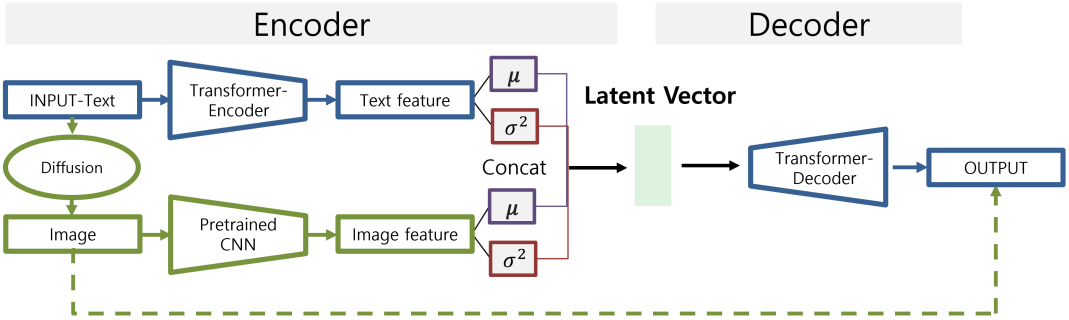

Figure 1: Our proposed DeepLabv3+ ITG-VAE extracts mu and sigma from text and images, respectively, converts them into Z, which is a Latent vector, and outputs reconstruction sentences as a result through a decoder.

4 Model

Our model can be divided into five stages: image generation, transformer encoder, CNN encoder, new latent variable extraction, and transformer decoder.

4.1 Image Generation

We used the 'stable diffusion' model to generate images from text. This model performs well among many text-image generation models. Due to the large number of parameters, the size of the model, and the size of the dataset, we decided to use the pre-trained stable diffusion model.

Figure 2: An output image of ‘a student of an astronaut riding a horse’

4.2 Transformer Encoder

It is the step of extracting feature vectors from each sentence in the dataset. The hyperparameters we used in our study are as follows.

- e_dim: Embedding dimension, which represents the size of the vector that represents a word or token. The default value is 200, and it specifies the dimension of word embeddings.
- z_dim: Latent variable dimension, which represents the size of the latent variable vector used in the generative model. The default value is 32, and it specifies the dimension of the latent variables.
- nheads: Number of attention heads used in the attention mechanism. The default value is 4, and it is used in the multi-head attention of the transformer model.
- nTlayers: Number of encoder layers used in the transformer encoder. The default value is 4, and it specifies the total number of encoder layers in the transformer encoder.
- ff_dim: Size of the hidden layer in the feedforward neural network inside the transformer encoder. The default value is 400, and it specifies the size of the hidden layer in the feedforward neural network.
- pad_idx: Index of the padding token. The default value is 1, and it specifies the index of the padding token in the input sequences.

Transformer encoders have a scalable structure that can handle inputs of different sizes, so the model can be scaled up or down depending on the length of a sentence or the size of a document. Transformer encoders can also share information globally because they apply the ’Attention’ technique for all input positions. This makes it possible to consider important information in any part of the sentence. Therefore, we concluded that the transformer encoder is a suitable model for our situation where we need to receive input from a variety of sentences, and we adopted it based on the above advantages.

4.3 Pretrained Cnn encoder

MobileNetV3 is a lightweight neural network architecture suitable for computer vision tasks on mobile devices and embedded systems. It offers high performance with small size and efficient computation, and can be utilized in a wide range of applications. It uses lightweight convolutional operations and activation functions to achieve faster computation and smaller memory requirements. This is useful for real-time applications and limited resource environments.

4.4 VAE

We extracted a new vector by combining the feature vector of the text and the feature vector of the image generated through diffusion, and the latent variable z was learned through VAE.

This approach is a form of multimodal learning, where the image and text are represented as vectors in the latent space by their respective VAE. By combining these latent space vectors, a comprehensive representation that incorporates the characteristics of both image and text is obtained. This enhances the interaction and association between images and text, making them easier to understand and more informative.

Our model needed one vector with both the information of each feature vector to generate a new sentence that resembles the original sentence but also contains the information of the generated image. Therefore, we concatenated feature vectors rather than using the sum of two vectors so that we do not lose the unique information of each feature vector.

4.5 Transformer Decoder

Transform decoders can use self-attention to focus on other words in the sequence, which allows for parallel processing. In other words, they can process all the words at the same time instead of processing them in order. This can help create faster and more efficient models. Also, this means that with more data and computational resources, a more accurate model can be built.
Transform decoder takes the latent variable Z as an input and generates a sentence. This model uses an attention mechanism to generate sentences while simultaneously considering various parts of the input sentence. The reconstructed sentences have features extracted from the existing training data and can generate various sentences depending on the latent variable Z. This allows the trained model to generate new sentences based on its understanding of the original data.

5 Experiments

5.1 Data

The data we used to train our model is the Penn Treebank dataset. The Penn Treebank corpus is one of the best known and most used corpora for evaluating sequence labeling models in natural language processing.

We tokenized this corpus with torchtext's tokenizer and were able to split it into 9925 unique words. We divided it into sentences, filtered out sentences longer than 45 tokens, and split it into train/test/validation after preprocessing. After that, we adjusted the subset size appropriately for a smooth study.

5.2 Evaluation method

We used KL loss, Reconstruction loss, and beta-VAE loss as evaluation metrics, which were employed in a previous study "A transformed-based variational autoencoder for sentence generation" for comparing the transformed-based VAE and LSTM-VAE.

KL loss measures the difference between the distribution of a latent variable and a normal distribution. This makes the data generated from the latent variable space more diverse and useful.

Reconstruction loss measures the difference between the input sentences and the generated sentences. Therefore, this loss function plays an important role in sentence generation tasks because the generated sentences should be similar to the input sentences.

The beta-VAE loss is calculated as (Reconstruction loss+KL loss*beta) and is the final loss function. Learning proceeds in the direction of minimizing this beta-VAE loss, which allows us to learn useful distributions in the latent variable space and produce sentences similar to the input sentences.

Finally, we also tried human evaluation to evaluate the understanding of text by the pair of generated images and reconstructed text.

5.3 Experimental details

In our study, Transformer and VAE were used as the basic models, and Stable diffusion and CNN models were used to process the images. Each model was run under the following conditions.

For CNN models, MobileNet-v3 was selected as the final learning model after comparing learning time and performance using three types: ResNet, VGG16, and MobileNet-v3, and we received pre-trained information by setting 'pre-trained = True'.

The output image through stable diffusion was 500x500, and the feature vector through CNN was 32x64.

For this experiment, the GPU was RTX2060, the batch size was 32, the optimizer was SGD, the lr was 0.1, and the momentum was 0.9, and the total training time was 75 hours with 10 epochs.

5.4 Results

The following loss curve shows that ITG-VAE has learned the information of the image well compared to the previous model.

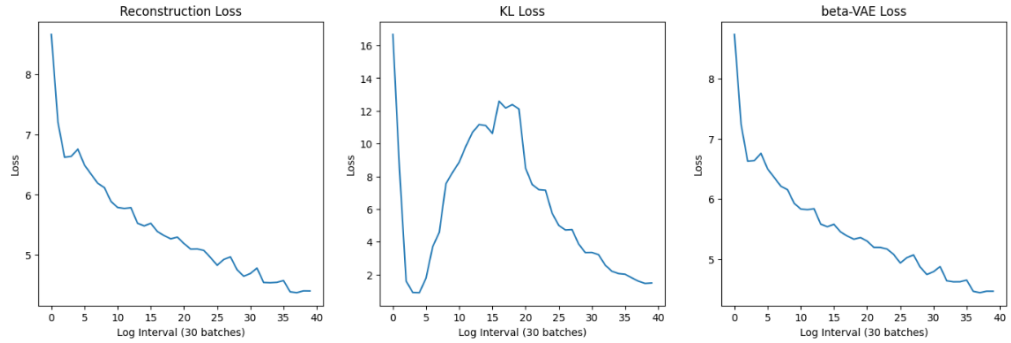

Figure 3: displays the curves of KL loss, reconstruct loss, and beta loss for a model that reconstructs text without using images.

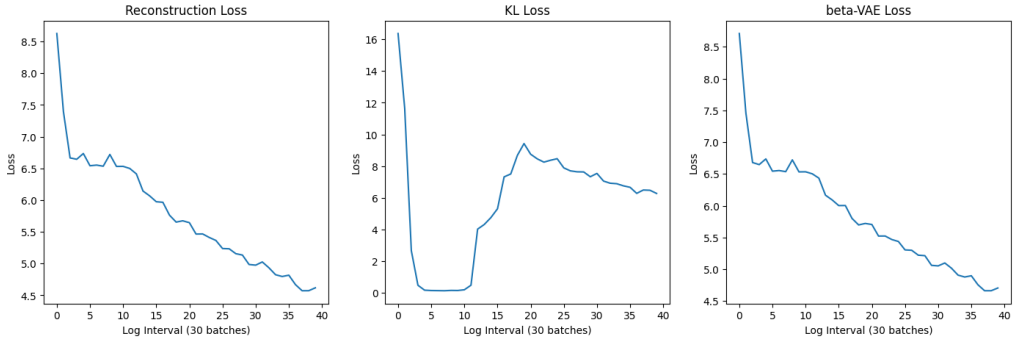

Figure 4: displays the curves of KL loss, reconstruct loss, and beta loss for our ITG-VAE model, which utilizes images to reconstruct image-conditioned text.

Table 1: The original sentence is a sentence in the PTB dataset, where <unk> is a special token that replaces low-frequency or unknown words, and "n" is a special token used to generalize nouns in the PTB dataset. The reconstructed sentence is a sentence that the model reconstructed after completion of learning. For the reconstruction task, we used a greedy decoding policy.

Original sentence	the <unk>of the new york stock exchange composite trading <unk>concluded with $ n to n yesterday
Reconstructed sentence	the <unk>of the new york stock exchange composite trading yesterday at $ n down n

6 Analysis

We focus on comparing the performance of the ITG-VAE model with that of the previous study's model, and evaluating the results. The analysis is provided below.

6.1 Comparison with the previous study's model

- Loss Curve : The ITG-VAE model can be seen to be learned more stably than the previous model. The rapid decrease in KL loss shows that the information in the image greatly helped model learning.
- Performance: The ITG-VAE model shows better performance than the previous model. It demonstrates the ability to learn the information of the image well.

6.2 Sentence Generation results

- Additionally, the reconstructed sentences reflect the features of the image well, suggesting that the model can effectively combine the visual information of the image with the meaning of the sentence to produce sentences that can help understanding.

6.3 Learning time

- The continuous decline in the loss curve of our model means that the model is progressively improving its performance in the learning process. However, it took 75 hours to complete the 10 epochs, so no additional learning has been done, and the model has not achieved optimal performance. It is judged that measures such as optimizing learning time are needed for further learning and improvement.

Comprehensively considering these analysis results, ITG-VAE is a good sentence generation model that effectively learns the relationship between images and sentences, and shows excellent performance in generating image-conditioned sentences similar to the original sentence.

7 Conclusion

In this work, we propose ITG-VAE, an image-conditioned VAE text generation model. The model utilized a method of extracting feature vectors from images generated based on text and combining them with feature vectors from text. We sought to improve our understanding of long or difficult sentences by creating images and trimmed sentences based on the images.

As a result of human evaluation, it was found that understanding was clearly improved when the image generated through the input sentence and the image-conditioned sentence regenerated by the ITG-VAE were together than when there was only the input sentence. This shows that our model provides an image-text pair for easy understanding of complex sentences.

It can help create valuable text-image datasets for further research in that it generates images and image-conditioned text based on the text entered. It can help create valuable text-image datasets for further research in that it generates images and image-conditioned text based on the text entered. This can contribute to the development of future computer vision and natural language processing by applying it in fields such as image captioning, text summarization, and multimodal understanding.

The limitations of our study were the lack of learning time and the data bias that occurred using the PTB dataset. These limitations can be improved by proceeding in an improved learning environment (e.g., gpu), adjusting hyperparameters such as num inference steps, and selecting a more even dataset.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] F. Hill, K. Cho, and A. Korhonen. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*, 2016.

[3] A. G. Howard, M. Sandler, G. Chu, L-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1314–1324, 2019.

[4] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2013.

[5] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[6] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[7] S. Semeniuta, A. Severyn, and E. Barth. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*, 2017.

[8] I. V. Serban, A. Sordoni, R. Lowe, and L. Charlin. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

[9] Z. Sun, Z-H. Deng, J-Y. Nie, and J. Tang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[11] W. Yin, J. Chen, P. Ren, Z. Feng, M. Riedl, and Z. Zhao. A transformer-based variational autoencoder for sentence generation. In *Proceedings of the International Conference on Machine Learning*, 2019.