Generative Transformer for Diverse Text Generation

Ishika Agarwal, Priyanka Kargupta, Bowen Jin, Akul Joshi

Abstract

Diverse text generation is an important and challenging task. Existing methods mainly adopt a discriminative model, with the underlying assumption that the input text-to-output text projection is a one-one mapping. However, this is not true in the real world, since given one single input text, there can be multiple ground truth output text candidates. For example, in the commonsense generation, given a list of knowledge entities, there should be more than one way to use them to come up with a sentence. This motivates us to capture the underlying text semantics distribution with generative models (*e.g.,* VAE and diffusion models). On the other hand, Transformer architecture has been demonstrated to be effective in text semantics capturing. Then the problem comes to how to effectively combine the Transformer architecture with the generative models. Our project aims to combine the best of both worlds by introducing VAE & Diffusion model into transformers. Specifically, we want to apply them to two downstream tasks: common sense generation and question generation. We include results, and some future work to further this project.

1 Introduction/Motivation

Recent work on text generation has largely been centered around the usage of transformer architectures, primarily due to their ability to detect long-range dependencies within large text sequences and the overall scalability of pre-trained language models. However, existing Transformer architectures or pretrained language models (e.g., BART, T5) fail to capture text diversity, which proves to be crucial in downstream tasks such as diverse question-answer generation and commonsense reasoning. On the other hand, generative models (*e.g.,* VAE and Diffusion Model) have been demonstrated to be effective in capturing complex distribution, which can be potentially applied to encode the distribution of diverse ground truth output text sequences conditioned on the given input text sequence. Hence, our project proposes integrating generative models into the transformer architecture in order to apply the model to more difficult downstream tasks that depend on diverse text generation.

2 Related Work

Transformers. Vaswani et. al. [9] introduces the basic architecture of the transformer as a model that foregoes recurrences and convolution and focuses entirely on attention. We also refer to Raffel et. al. [8], who discuss the Text-to-Text Transfer Transformer (T5) architecture that we will be exploring for our use case. **Variational Models.** Iqbal and Qureshi [5] provide a survey of previous deep learning models that have been used for the task of text generation. **Common-Sense Generation.** Liu et. al. [7] discusses the problem of common-sense generation; their approach uses a generative model to actually generate common-sense outputs and an autoencoder-based refiner to fix potential errors in the generation. **Question Generation.** Du et. al. [4] discusses the problem of question generation; their approach is a model that is trainable end-to-end via sequence-to-sequence generation. **Conditional VAE.** Wang et. al. [10] use the T5 architecture with a conditional VAE for story completion - given a story with words/small phrases masked out, complete the story. Here, they sample for words/phrases conditioned on the context of the story. **GP-VAE.** Du et. al. [3] learn Gaussian processes in their variational encoder-deocder model to introduce variability in their text. We are using their codebase for developing this project. **Diffusion models in text generation.** Li et. al. [6] first innovatively add a diffusion process upon the generated hidden states of Language Models. Yuan et. al. [11] propose to utilize the Transformer decoder to denoise the states in the Diffusion reverse process.

3 Problem Formulation

3.1 Question Generation

Given a document, our task is to generate a unique question from the document. In our datasets, we have multiple questions per document, and therefore, we will generated multiple questions per document. For that reason, we also constrain that the generated questions are unique and diverse. **Input**: a document. **Output**: a unique questions

3.2 Common Sense Generation

Commonsense generation is a constrained text generation task, to explicitly test machines for the ability of generative commonsense reasoning. Given a set of common concepts; the task is to generate a coherent sentence describing an everyday scenario using these concepts. **Input**: a set S of common concepts. **Output**: a coherent sentence T.

For example, given a collection of objects and actions {dog, frisbee, catch, throw}, we expect to generate sentences such as "A dog leaps to catch a thrown frisbee", or "The dog catches the frisbee when the boy throws it." or "A man throws away his dog's favorite frisbee expecting him to catch it in the air".

Given a concept set, generated sentences of high quality need to be both fluent and diverse. Fluency means that the generated sentences should be similar to human-written ones, while diversity means that different generated sentences should be as different from each other as possible.

4 Methodology

The methodology is common for both of the tasks. We explore two textual generative models: variational transformer and textual diffusion models.

4.1 Variational Transformer

Backbone Pretrained Language Model. To make the generated texts as fluent as real-world texts, we adopt the widely-used encoder-decoder style pretrained language model T5 [8], which is trained on large-scale text corpora.

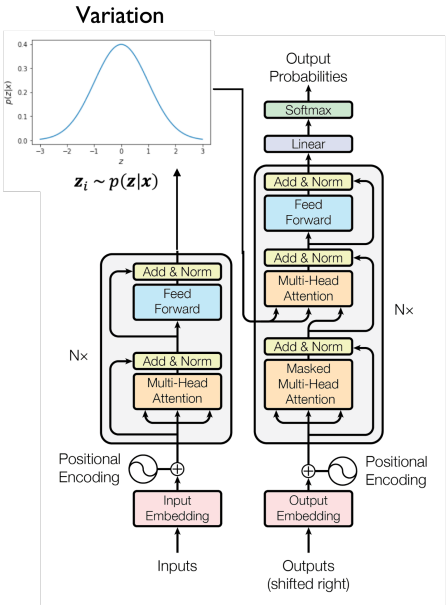

FIGURE 1: Variational Transformer Architecture.

Architecture: Encoder-Variation-Decoder. When doing text generation, the original Transformer architecture will have a cascaded structure between the encoder and decoder, so that there exists a 1-1 mapping between the input texts and output texts. However, in the real world, we may expect diverse output given the same input text, where there should exist 1-x mapping between inputs and outputs ($x > 1$). In order to capture the 1-x mapping relation inside the transformer architecture, we try to add a variational layer between the encoder and decoder. In other words, the encoder will project an input sentence into a distribution rather than a vector in the latent space. Then, the decoder will do sampling on this distribution and conduct decoding to generate the output texts. This model structure can be found in Figure 1.

In mathematics, the generation process can be formulated into,

$$\begin{aligned} p(z|x) &= \text{TRM-Enc}(x), \\ p(z) &\sim N(0,1), \\ p(y|z) &= \text{TRM-Dec}(z) \end{aligned}$$

Currently, we make an assumption that the prior distribution of the latent vector z is under the normal distribution.

The objective function includes two parts, a generation accuracy loss L_g and a regulariza-

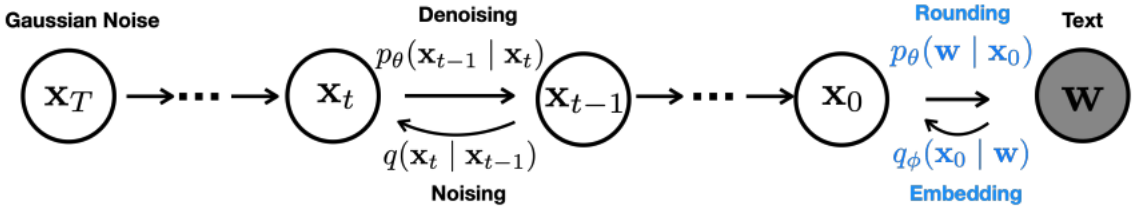

FIGURE 2: Textual Diffusion Architecture.

tion loss L_r, *i.e.*, $L = L_g + L_r$. L_g will push the model to generate fluent text of good quality given the input conditioned text, while L_r will facilitate the underlying learned distribution $p(z|x)$ to be as normal Gaussian distribution as possible.

4.2 Transformer & Diffusion Model

Our second exploration is a combination of transformer encoder with diffusion model, the architecture of which can be found in Figure 2. Inspiration for the architecture for this model has been drawn from Li et. al. [6], in which they discuss the Diffusion-LM architecture.

Model Design In this architecture, we only utilize the transformer encoder as an embedder rather than the whole Transformer encoder-decoder architecture in Sec 4.1 to capture the semantics representation of words ($w \rightarrow x_0$).

To be more specific, we add a Markov transition from discrete words w to x_0 in the forward process, parametrized by $q_\phi(x_0|w) = N(\text{Emb}(w), \sigma_0 I)$. In the reverse process, we add a trainable rounding step, parametrized by $p_\theta(w|x_0) = \prod_{i=1}^{n} p_\theta(w_i|x_i)$, where $p_\theta(w_i|x_i)$ is a softmax distribution. In other words, Rounding is achieved by choosing the most probable word for each position.

Then the other part of the forward process ($x_0 \rightarrow x_1 \rightarrow ... \rightarrow x_T$) and the reverse process ($x_T \rightarrow x_{T-1} \rightarrow ... \rightarrow x_0$) just remain the same with the original diffusion model.

Decoding and Text Generation The conventional workflow of using the diffusion model for generation is to randomly pick up a Gaussian noise and let it go through the reverse process. However, this generative process is not suitable for text generation, since text sequence generation is not a coarse-grained pixel generation like image generation. One single flaw in a one-word generation will make the whole text sequence meaningless. As a result, in this model, there is an added plug-and-play control mechanism [2] to control the text generation process. To be more specific, a fluency regularization is added: $\lambda \log p(x_{t-1}|x_t) + \log p(c|x_{t-1})$, where λ is a hyperparameter.

5 Results

5.1 Question Generation

Dataset. We conduct our question generation experiments on the popular question-answering dataset, SQuAD 2.0 [1], which contains 130,000 questions– a single or multiple question and answer pairs per dataset. We specifically only used the provided documents and questions. We split the dataset into a training set and testing set (a validation set is already provided) with 84% of the document-question pairs in the training set and 16% in the testing set. We ensure that a document is contained within one split (i.e. we will never have question 1 of document 2 in training and question 2 of document 1 in testing). The SQuAD dataset has a variety of topics, such as "Oklahoma", "Planck Constant", "Pain", "Super Nintendo Entertainment System", "Han Dynasty", and therefore, we also expect our output to capture that diversity.

Evaluation. We use BLEU and SelfBLEU scores for our quantitative analysis. We calculate the BLEU scores of a generated question with respect to a target question. We calculate the SelfBLEU score of a question with respect to all the other generated questions of a document (i.e., document generated question sets). In Figures 4 and 8, we compare the aver-

[1] https://rajpurkar.github.io/SQuAD-explorer/

TABLE 1: Qualitative Results of T5-FT and T5-CVAE on Question Generation using SQuAD

Input	T5-FT	T5-CVAE
"Predation is a biological interaction... main category of consumption is detritivory... parasitic species prey on a host organism... "	What is the main category of consumption of prey? What is the main category of consumption of prey? What is the main category of consumption of prey? What is the main category of consumption of prey? What is the main category of consumption of prey?	What does detritivory do to its offspring? What does the act of prey do to its offspring? What does laying eggs on a host organism lead to? What is the main category of food that a parasitic species consumes? What does the act of detritivory result in?
"In the 1980s and early 1990s, there was a significant movement... Territorial Clause... union with the Northern Mariana Islands as a single territory, or independence.	What was the main movement in favor of Guam becoming a commonwealth? What was the main movement in favor of Guam becoming a commonwealth? What was the main movement in favor of Guam becoming a commonwealth? What was the main movement in favor of Guam becoming a commonwealth? What was the main movement in favor of Guam becoming a commonwealth?	In what year did Guam join the United States? In what year did Guam become a commonwealth? What is the name of the constitution that allows Guam to become a U.S. state? What would happen if Guam was united with the Northern Marianas? What did the federal government reject in favor of the territory becoming a commonwealth?

TABLE 2: Qualitative Results of T5-FT and T5-CVAE on Question Generation using AmazonQA

Input	T5-FT	T5-CVAE
"These mattress covers fit and work... The new cover fit perfectly... wide enough for it."	What is the quality of the mattress cover? What is the quality of the mattress cover? What is the quality of the mattress cover? What is the quality of the mattress cover?	Will this cover fit a queen size mattress? Will this cover fit a king size mattress? What are the dimensions of the cover? Will this cover fit a queen size bed?
"effective against mosquitos and no-seems... mosquito population was down... this unit's only drawback... is it's somewhat awkward size"	What is the only drawback of the ThermaCell? What is the only drawback of the ThermaCell? What is the only drawback of the ThermaCell? What is the only drawback of the ThermaCell?	How well does this product work for cockroaches? Does this product work well for mosquitoes? How well does it work for cockroaches? How well does it work against mosquitoes?

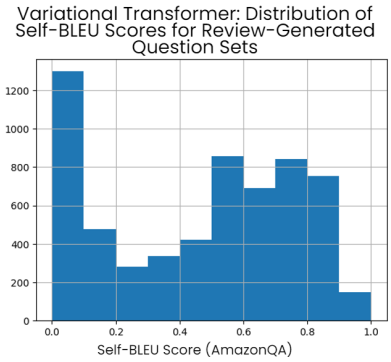

FIGURE 3: **Variational Transformer/Question Generation: Self-BLEU score computed for each set of question generated for every test set document.**

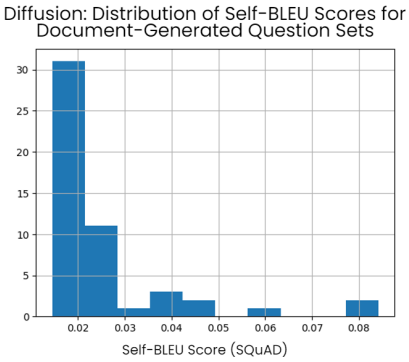

FIGURE 5: **Diffusion Transformer/Question Generation:** Self-BLEU score computed for each set of questions generated for every test set document.

age BLEU and SelfBLEU scores with varying epochs, respectively. In Figures 5 and 3, we calculate the self-BLEU among the questions generated for a document from the SQuAD and AmazonQA datasets, respectively. Since the scores are skewed left, there is diversity in the generation.

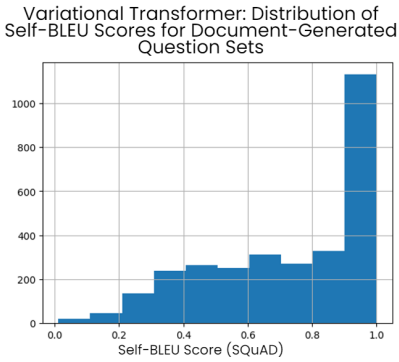

FIGURE 4: **Variational Transformer/Question Generation:** Self-BLEU score computed for each set of questions generated for every test set document.

Compared Method. In tables 1 and 2, we compare against a model we call T5-TF; this is a model pretrained on SQuAD version 1 dataset and fine-tuned for the question generation task. There is no variation involved in this architecture, however. The model can be found on HuggingFace.

5.1.1 Hyperparameter Sensitivity

Epochs In Figure 8, we plot the Self-BLEU scores on varying epochs on the questions generated by T5-CVAE on the SQuAD dataset. We show that as we increase the number of training epochs, the average Self-BLEU score across all sets of questions generated for each document increases and hence, the diversity of the generated questions decreases.

Beam Width We compare using a beam-width $k = \{10, 50\}$ on the SQuAD dataset. We can see that if we significantly widen the scope of probable questions to chose from, we may always choose the same one each time; in

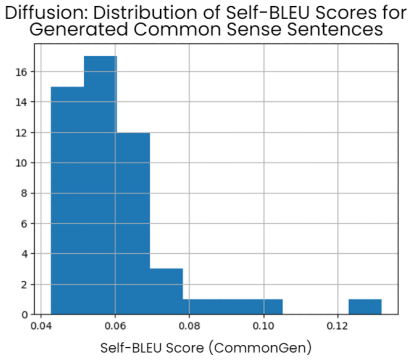

FIGURE 6: **Diffusion Transformer/CommonSense:** Self-BLEU score computed for each set of common-sense sentences generated for each set of concepts.

other words, minimal pruning leading to less diversity in selection.

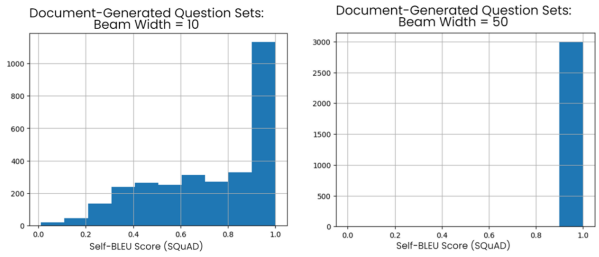

FIGURE 7: Self-BLEU scores computed for each set of SQuAD questions after training and testing on beam width = 10 **(left)** and beam width = 50 **(right)**. This hyperparameter study was performed using the variational transformer model.

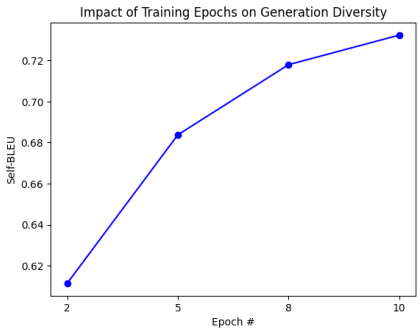

FIGURE 8: Self-BLEU on SQuAD on T5-CVAE with varying epochs.

5.2 Common Sense Generation

Dataset. We conduct experiments on the widely used CommonGen dataset [2]. Since the labels of samples in the official test set of CommonGen are not given, we random select 202 samples from the validation set as test samples and leave the others as validation samples.

[2]https://inklab.usc.edu/CommonGen/

Baseline. We mainly compare the PLM+variation model with the vanilla PLM, namely the vanilla T5 model. In addition, we explore two different pretrained T5 parameters, T5-base [8] and Flan-T5-base [1]. Flan-T5 [1] adopts a second instruction-based fine-tuning, which gives the pretrained language model stronger power in downstream tasks.

Quantitative results. In this section, we mainly evaluate the fluency of the generated texts for both our variational method and vanilla T5 methods. We use ROUGE as the metric to compare different methods. The results are shown in Table 5.

From the result, we can find that: 1) Comparing with vanilla T5 (T5 & Flan-T5), T5+Vae methods (T5-VAE & Flan-T5-VAE) can have worse results. This means that enhancing diversity need the model to sacrifice accuracy. 2) Flan-T5 based models have better results than vanilla T5 based models, which demonstrates that instruction-tuning truly improves downstream performance. In Figure 6, we calculate the self-BLEU scores of sentences generated from the same concept set. The diversity is quite good as the scores are skewed to the left.

Qualitative Results for Variational Transformer. The generated samples by T5 and T5-VAE are shown in Table 3, while the generated samples by Flan-T5 and Flan-T5-VAE are shown in Table 3. From the results, we can find that: the generated samples from T5 and Flan-T5 are in a single style, while the samples from VAE-based methods are much more diverse.

Qualitative Results on Diffusion Transformer. Table 4 has the input and the generated samples. The output from Common Gen seems to be mostly coherent. It is not of great quality - we think it is because of the high vocabulary size in the dataset and the low vocabulary size in the model. On the other hand, the output from Question Gen (evaluated on the SQuAD dataset) is somewhat random, and does not fit the context well. It suffers a similar problem: the document size and vocabulary size is large, and the model's vocabulary size is small. Furthermore, out of the interest of

Table 3: Qualitative results of T5 & T5-VAE on CommonGen.

Input	T5	T5-VAE
field,look,stand	A man stands in a field looking up and stands in a field looking up.	A man standing in a field looking at the camera as he looks at a field.
	A man stands in a field looking up and stands in a field looking up.	A group of people are standing in a field looking at a field full of green fields.
	A man stands in a field looking up and stands in a field looking up.	As he stands in a field, he looks at the camera as he stands in the field.
	A man stands in a field looking up and stands in a field looking up.	A man stands in a field looking at the camera as he stands in a field.
	A man stands in a field looking up and stands in a field looking up.	A man standing in a field looking at the camera.
kid room dance	The kid in the room is dancing in the room.	a kid dances in a dark kid room with his kiddo dancing with his kiddo in the kiddroom.
	The kid in the room is dancing in the room.	A kid is dancing in a kid's room while a kid is dancing.
	The kid in the room is dancing in the room.	A kid is dancing in a dark kid room.
	The kid in the room is dancing in the room.	A kid is dancing to a kid in a living room.
	The kid in the room is dancing in the room.	A group of kids are dancing in a living room.
cat pet couch	A cat is petting on a couch next to a couch.	A cat is petting a cat on a couch.
	A cat is petting on a couch next to a couch.	A cat is petting a cat on a couch while a cat is petting a cat on it.
	A cat is petting on a couch next to a couch.	A cat is petting a couch with a cat on it.
	A cat is petting on a couch next to a couch.	A cat is petting a cat on a couch.
	A cat is petting on a couch next to a couch.	A cat petting a cat on a couch next to a couch.

Table 4: Qualitative Results on both downstream tasks on Diffusion-LM

CommonGen Input	Generated Common Sense Sentence	SQuAD Input	Generated Questions
sit,space,stare	And the best room sit floating in the space with stare lot of listing.	A document about	What is in that other two?
	Region sit by emerges middle and stare explosion in space.	US federal law	Why does groups that other work?

Table 5: Quantitative results on CommonGen.

Method	Rouge1	Rouge2	RougeL
T5	0.3479	0.1210	0.3056
T5-VAE	0.3330	0.1154	0.3019
Flan-T5	0.3749	0.1394	0.3306
Flan-T5-VAE	0.3496	0.1162	0.3071

time, this model was evaluated after half of the training process. It is natural that these issues will result in poor quality generation. We hypothesize that with full training and a larger model vocabulary, both tasks will output sensible and diverse output.

6 Conclusion

We propose to use the T5 architecture with a variational layer (a VAE and a diffusion model) in between the encoder and decoder. We apply this architecture to two downstream tasks: diverse question generation and diverse common sense generation. We provide results (quantitative and qualitative) - we show that T5-VAE is able to generate diverse results for both of these downstream tasks. Furthermore, we show that transformers with diffusion are able to capture even more variability than T5-VAE - with more resources (GPU's), we are confident that we will be able to generate higher quality samples.

References

[1] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[2] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

[3] Wanyu Du, Jianqiao Zhao, Liwei Wang, and Yangfeng Ji. Diverse text generation via variational encoder-decoder models with gaussian process priors. *arXiv preprint arXiv:2204.01227*, 2022.

[4] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. *CoRR*, abs/1705.00106, 2017.

[5] Touseef Iqbal and Shaima Qureshi. The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*, 34(6, Part A):2515–2528, 2022.

[6] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

[7] Xin Liu, Dayiheng Liu, Baosong Yang, Haibo Zhang, Junwei Ding, Wenqing Yao, Weihua Luo, Haiying Zhang, and Jinsong Su. Kgr4: Retrieval, retrospect, refine and rethink for commonsense generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11029–11037, Jun. 2022.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring

the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] Tianming Wang and Xiaojun Wan. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *IJCAI*, pages 5233–5239, 2019.

[11] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *arXiv preprint arXiv:2212.10325*, 2022.