Generating Diverse and High-Quality Abstractive Summaries with Variational Transformers

Anonymous ACL submission

Abstract

Existing works on abstractive summarization mainly focus on boosting summarization's quality (informativeness, contextual similarity). To generate summaries of both high diversity and quality, we proposes the Transformer+CVAE model, which integrates the CVAE framework into the Transformer by introducing the prior/recognition networks that bridges the Transformer encoder and decoder. We utilize the latent variables generated in the global receptive field of the transformer by fusing them to the starting-of-sequence ([SOS]) of the decoder inputs. To better tune the weights of the latent variables in the sequence, we designed a gated unit to blend the latent representation and the [SOS] token. Evaluated on the Gigaword dataset, our model outperforms the state-of-the-art seq-to-seq models and the base Transformer in diversity and quality metrics. After scrutinizing the pre-training and the gating mechanism we apply, we discover that both schemes help improve the quality of generated summaries in the CVAE framework.

1 Introduction

Abstractive text summarization aims to generate a new description of the original article that covers core information in the source texts and is linguistically fluent. The approach to abstractive summarization is a human way of generating summaries, attempting to understand the entire context of the data, which requires generalization, paraphrasing, and integrating real-world knowledge. With the developments in deep learning, abstractive summarization is regarded as a seq-to-seq learning problem, where the encoder-decoder models with attention mechanisms are generally used (Rush et al., 2015; Chopra et al., 2016).

To increase the quality of abstractive summarization, seq-to-seq neural network models extract the core information of the text using attention mechanisms, but they fail to utilize the latent structure information of summaries. Such conventional encoder-decoder methods calculate the attention weights and the hidden state parameters in an entirely deterministic way (Shi et al., 2020). Li et al. (2017) proposed a deep recurrent generative decoder that incorporates variational autoencoders (VAEs) (Kingma and Welling, 2014) as a multivariate Gaussian distribution to capture latent information. However, in practice, this method suffers from long-term sequential recurrent dependencies and vanishing gradient problems, generating unnecessary noise and hampering the learning of long data sequences. Inspired by the success of VRNN (Chung et al., 2015), Zhao et al. (2020) improved the variational decoder by using the stochastic hidden states of the VRNN layer as the input of the current RNN layer to extract latent information in adjacent time steps .

The attempts to take advantage of latent information using RNN-based VAEs proved to be beneficial in avoiding generating dull and repetitive summaries, but the inherently sequential nature of such models makes large-scale training less efficient. On top of that, RNN-based VAEs face the vanishing latent variable problem (Bowman et al., 2016), where the powerful auto-regressive RNN structured decoder ignores the latent variables and generates outputs only dependent on previous tokens. Unlike RNNs, the vanilla Transformer (Vaswani et al., 2017) allows parallel training and has a global receptive field at each stage. Despite the Transformers' efficiency and robust fully attentional mechanisms, they are deterministic and fail to model one-to-many relations. In addition, the greedy and beam search in the Transformers makes it challenging to generate diverse and informative abstractive summaries.

In this paper, we propose the Transformer-CVAE, a variational fully attentive feed-forward

seq-to-seq model, to address the lack of diversity of abstractive summarization while maintaining high quality. The experimental results based on the Gigaword Dataset show that our proposed model improves the diversity as well as the quality of generated summaries, meaning that our model absorbs the advantage of the non-deterministic nature of CVAE and the strong generative modeling power of the Transformer. The reminder of this paper is organized as follows: In Section 2 we intorduce the related work. Then, we propose the Transformer-CVAE in Section 3. In Section 4 and 5, we conduct the experiments and empirically analyze the results, respectively. Finally we conclude this paper.

2 Related Work

2.1 Abstractive Text Summarization

Rush et al. (2015) adopted a neural attention-based encoder-decoder method for sentence-level abstractive summarization. Chopra et al. (2016) improved the encoder-decoder model proposed by Rush et al. (2015) by utilizing an attentive convolutional encoder and a conditional RNN decoder for single-sentence summarization.

Recent studies have argued that the encoder-decoder model for abstractive summarization is vulnerable to phrase repetitions, grammatical mistakes, and inadequate reflection of the highlight of the original text (Gupta and Gupta, 2019). Several lines of research attempt to tackle these problems and improve the quality of abstractive summaries. By infusing prior knowledge, such as the linguistic features, into neural networks along with RNN and probabilistic objectives, Rossiello et al. (2016) reduced the semantic and grammatical errors caused by the dependence on statistical co-occurrences of words. To solve the problem of repetitions, Lin et al. (2018) used convolutional gated units on top of the encoder outputs at each time step. They added the global encoding at the encoder side and an undirectional LSTM at the decoder side. Another method focusing on avoiding repetitions is to use a diversity-driven attention model, which requires queries relevant to the highlights in the source article (Nema et al., 2018).

2.2 Conditional Variational Autoencoders

The variational autoencoder (VAE) (Kingma and Welling, 2014), a deep generative probabilistic model, adopts a decoder network to reconstruct the encoder outputs by Gaussian sampling. The conditional VAE (CVAE) framework, proposed by Sohn et al. (2015), is a conditional graphical model that features both latent variables and data conditioned on some variables. Although seq-to-seq models achieve high performance in text generation tasks, they tend to generate dull and repetitive results in tasks such as abstractive summarization and dialogue response generation (Li et al., 2016). One popular line of research to address this issue is to integrate stochastic latent variables into seq-to-seq models based on the CVAE framework.

Many text generation works are committing to combining the CVAE with encoder-decoder models. Li et al. (2017) designed a deep recurrent generative decoder and connected the encoder-decoder structure to a variational RNN (VRNN) (Chung et al., 2015). Zhang et al. (2016) explored the application of CVAE to Neural Machine Translation (NMT) that enables better long-sequence generation. CVAE framework proved to be helpful in distilling the underlying semantics of the source-target pairs in supervised learning. And the non-deterministic nature of CVAE is increasingly used in capturing variability of the latent space and generating diverse results. Du et al. (2018) designed variational autoregressive decoders that inject variational inference into the RNN-based decoder to generate various dialogue responses. And Le et al. (2018) strengthened the variational autoregressive decoder using external dynamic memory and improved the quality of the diverse responses. Even though the aforementioned RNN-based CVAE models greatly improved the diversity in text generation tasks, they suffer from the *vanishing latent variable problem* (Bowman et al., 2016), that is the autoregressive RNN-based decoder is too powerful so that it tends to pay less attention to the latent representation. Therefore, we try to explore a different combination of CVAE and an autoregressive generative model - the Transformer.

2.3 Combining Transformer with CVAE

With the emergence of the Transformer (Vaswani et al., 2017), many text generation research attempted to incorporate CVAE into the Transformer. Liu and Liu (2019) proposed a variational transformer-based model augmented with LSTM

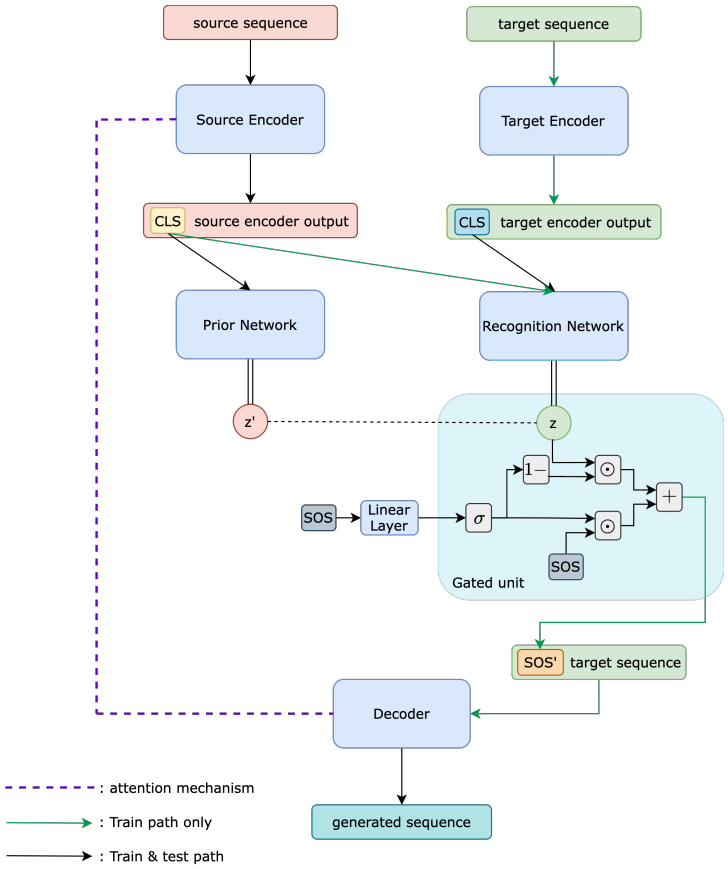

Figure 1: Our model - Transformer+CVAE. In the training process, the latent variable $\mathbf{z}$ is generated by the posterior network. In the testing process, $\mathbf{z}$ is instead sampled from the prior network because the target text is not allowed during inference. We ignore the details of the Transformer structure and the prior/recognition networks for simplicity.

layers, and they proved the effectiveness of the incorporation of the transformer to solve KL vanishing problem and achieved better performance than some baseline models. Some works come from neural dialog response generation. Lin et al. (2020) designed a global variational transformer and a sequential variational transformer, testing on how the extent of diversity changes when the latent variable is reachable in the global or local receptive field.

However, there are few works that focus on improving text summarization using transformer-based CVAE. Many state-of-the-art models for abstractive summarization have achieved high performance in the accuracy and fluency of the summaires, boosting ROUGE scores with fine-tuned and pre-training strategies. The diversity of summaries is also an essential part of summaries, so this paper aims at improving the diversity of the abstractive summaries while keeping the quality high.

3 Proposed Model

3.1 The CVAE framework

Inspired by the CVAE framework proposed by Sohn et al. (2015), we model the underlying semantics of the article-summary pairs explicitly. The model represents the dyadic conversation among the following three variables: the input source article $\mathbf{x}$, the generated summary $\mathbf{y}$, and a continuous latent variable $\mathbf{z}$ we assume to follow multivariate Gaussian distribution with diagonal covariance matrix from the semantic space. Our goal is to maximize the conditional likelihood that can evolve to:

$$p(\mathbf{y}|\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{y}, \mathbf{z}|\mathbf{x}) d\mathbf{z} = \int_{\mathbf{z}} p(\mathbf{y}|\mathbf{z}, \mathbf{x}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z}$$

We denote the prior network as $p_\theta(\mathbf{z}|\mathbf{x})$ that approximates the prior distribution $p(\mathbf{z}|\mathbf{x})$ and the recognition network $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ that approximates the true posterior distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$. The decoder $p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x})$ that shares parameters with the

encoder approximates the generative distribution $p(\mathbf{y}|\mathbf{z},\mathbf{x})$. As is presented by Sohn et al. (2015), the evidence lower bound ELBO is formulated as:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{x}) \geq \mathcal{L}_{ELBO} &= -\mathcal{L}_{KL} + \mathcal{L}_{REC} \\ &= -KL\left(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})\|p_\theta(\mathbf{z}|\mathbf{x})\right) \\ &+ \mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})\right]\end{aligned}$$

where $\mathcal{L}_{KL}$ represents the Kullback-Leibler (KL) divergence between the recognition network and the prior network, and $\mathcal{L}_{REC}$ represents the reconstruction error of sampling $\mathbf{z}$ from recognition network (Kingma and Welling, 2014).

Since we assume that $\mathbf{z}$ follows a multivariate Gaussian distribution with diagonal covariance structure, the recognition network $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) \sim \mathcal{N}\left(\mu,\sigma^2\mathbf{I}\right)$ and the prior network $p_\theta(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}\left(\mu',\sigma'^2\mathbf{I}\right)$. Then we use the reparametrization trick (Kingma and Welling, 2014) to get sample $\mathbf{z}$ either from the recognition network or the prior network. In the training process, we obtain $\mathbf{z}$ from $\mathcal{N}\left(\mathbf{z};\mu,\sigma^2\mathbf{I}\right)$; in the testing process, we obtain $\mathbf{z}$ from $\mathcal{N}\left(\mathbf{z};\mu',\sigma'^2\mathbf{I}\right)$. Given $\mathbf{z}$ and the source input $\mathbf{x}$, the decoder then generates target output $\mathbf{y}$ sequentially. As is proposed by Sohn et al. (2015), by maximizing the ELBO of the conditional log likelihood, the CVAE can be efficiently trained with the Stochastic Gradient Variational Bayes (SGVB) framework (Kingma and Welling, 2014).

3.2 Transformer+CVAE

Inspired by the Global Variational Transformer proposed in Lin et al. (2020), we design a transformer-based CVAE model with latent variable $\mathbf{z}$ serving as a global semantic signal in a similar fashion. The architecture of our model is demonstrated in Figure 1. The source encoder and target encoder are both Transformer encoder, and the Transformer decoder shares parameters with the source encoder via cross-attention mechanism. We introduce the prior network and the recognition network, which are multilayer perceptrons (MLPs), to get mean and log variance of the Gaussian distribution of $\mathbf{z}$ (Zhou and Wang, 2018). To guarantee that the prior network and recognition network receive embeddings of fixed dimensions, we apply the special classification tokens [CLS] as their inputs. We hypothesize that the self-attentive representation of [CLS] token helps obtain long-range contextual information because it contains BERT's understanding of the sentence-level meaning (Devlin et al., 2019).

RNN-based CVAE framework generally incorporates the latent variable to the initial state of the recurrent decoder. Our transformer-based CVAE model can easily fuse latent representations into the starting-of-sequence ([SOS]) as the starting point of text generation. Instead of directly adding $\mathbf{z}$ to the [SOS] token as is proposed by Lin et al. (2020), we incorporate $\mathbf{z}$ into the [SOS] token of the decoder inputs via the *Gating mechanism*. With the gated unit, we can tune the weights of the latent variable $\mathbf{z}$ in the [SOS] token. We denote the embedding vector of the [SOS] token as s. W_s is defined as the parameter matrix of the linear layer taking s as the input. The formula of the gated unit we designed is as follows:

$$h = \sigma_s(W_s s)$$
$$s = h \odot s + (1-h) \odot \mathbf{z}$$

where σ_s is a sigmoid function and $\odot$ represents the Hadamard product. Then the resulting s token is attached to the target sequence, which is passed to the decoder along with the outputs of the source encoder.

3.3 Learning

Vanishing latent variable problem is a common issue in RNN-based CVAE (Bowman et al., 2016), but there are also chances that the decoder in our model pays less attention to the integrated [SOS] token. Following the suggestions from Bowman et al., we apply KL annealing method and early-stopping strategy to alleviate such problem following Zhou and Wang (2018). We also adopt the *bag-of-word loss* proposed by Zhao et al. (2017), which proved to be complementary to the KL annealing method and very effective in mitigating vanishing latent variables.

Therefore, our regularized ELBO learning objective along with the auxiliary loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{ELBO} + \mathcal{L}_{bow}$$

where,

$$\mathcal{L}_{bow} = \mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_\varepsilon\left(\mathbf{y}_{bow}|\mathbf{z},\mathbf{x}\right)\right]$$

Suppose $f = \text{MLP}_{bow}(\mathbf{z},\mathbf{y}) \in \mathcal{R}^{|V|}$ where $|V|$ is the vocabulary size. Then

$$\log p\left(\mathbf{y}_{bow}|\mathbf{z},\mathbf{x}\right) = \log \prod_{t=1}^{|\mathbf{y}|} \frac{e^{f_{\mathbf{y}_t}}}{\sum_j^{|V|} e^{f_j}}$$

where $|\mathbf{y}|$ is the length of $\mathbf{y}$ and $\mathbf{y}_t$ is the word index of the tth word in $\mathbf{y}$ (Zhao et al., 2017).

Model	Dist-1	Dist-2	Dist-3	sBL-2*	sBL-3*
Transformer	0.2383	0.6919	0.8633	0.5380	0.3185
CVAE (ours)	0.2049	0.7036	**0.9042**	0.5597	0.3035
Transformer+CVAE (ours)	**0.2718**	**0.7601**	0.8954	**0.4519**	**0.2507**
reference summaries	0.2971	0.8305	0.9552	0.4265	0.2372

Table 1: Abstractive summarization results in diversity metrics. The metric with * means that the lower score, the better performance. This notation applies to the following tables. sBL represents self-BLEU.

Model	RG-1	RG-2	RG-L	BERTSc	tBL-2	tBL-3
ABS+ (Rush et al., 2015)	29.76	11.88	26.96	-	-	-
RAS-LSTM (Chopra et al., 2016)	**31.71**	13.63	29.31	-	-	-
Transformer	28.82	11.54	26.57	87.63	0.5069	0.2971
CVAE (ours)	18.06	5.69	17.18	85.94	0.4912	0.2561
Transformer+CVAE (ours)	31.36	**14.16**	**29.35**	**89.23**	**0.5272**	**0.3120**

Table 2: Abstractive summarization results in quality metrics. tBL represents test-BLEU.

Dataset	Gigaword
Training pairs	3803957
Validation pairs	189651
Testing pairs	1951
Vocabulary size	124413

Table 3: Dataset statistics

4 Experiments

4.1 Data Settings

To evaluate the effectiveness of our proposed methods, we experimented on the Annotated English Gigaword dataset (Napoles et al., 2012). It consists of article-summary pairs which are the first sentence of the original articles paired with corresponding the titles. We follow the pre-processing steps in Rush et al. (2015). All digits in the pre-processed dataset are replaced with "#" and all word tokens appearing in less than 5 times are marked as "UNK". The details of the data is shown in Table 3.

4.2 Baselines

We compare the our proposed models with the following state-of-the-art models.

Transformer. The transformer we implemented is trained on the default model parameters proposed in (Vaswani et al., 2017).

ABS+. ABS+ (Rush et al., 2015) is an improved version of Attention-based Summarization which utilizes an attention-based encoder and a feed-forward neural network language model (NNLM).

RAS-LSTM. RAS-LSTM (Recurrent Attentive Summarizer) is proposed by Chopra et al., utilizing a conditional RNN for single-sentence summarization (Chopra et al., 2016). It adopts the attentive recurrent architecture, combining a convolutional attention-based encoder and an LSTM recurrent decoder. The model is easily trained on large datasets in an end-to-end fashion.

CVAE. The structure of our CVAE model is an RNN-based conditional variational autoencoder initially intended for dialogue response generation. It is proposed by Zhou and Wang (2018), who used several tricks to prevent the CVAE model from deteriorating to a plain Seq2Seq model by applying KL annealing, early stopping and bag-of-words loss.

4.3 Training Details

We implement the experiment with the Pytorch framework on an NIVIDIA 1080Ti GPU. We set the maximum number of words in an article or summary to 1000. To initialize the input embeddings of the source encoder, target encoder, and decoder, we apply the 300-dimensional pre-trained GloVe embeddings (Pennington et al., 2014). The encoder and decoder are both composed of a stack of 4 sub-layers of transformer heads with hidden dimension $h = 300$. The dimension of the latent variable is fixed to 300. The prior network and recognition network comprise

Model	Dist-1	Dist-2	Dist-3	sBL-2*	sBL-3*
Transformer+CVAE (w/ pre-train)	**0.2718**	0.7601	0.8954	**0.4519**	0.2507
Transformer+CVAE (w/o pre-train)	0.2607	0.7877	0.9606	0.7886	0.5526
Transformer+CVAE (w/ pre-train & w/o Gate)	0.2484	**0.8103**	**0.9714**	0.4756	**0.2405**
reference summaries	0.2971	0.8305	0.9552	0.4265	0.2372

Table 4: Selected abstractive summarization results for ablation study in diversity metrics.

Model	RG-1	RG-2	RG-L	BERTSc	tBL-2	tBL-3
Transformer+CVAE (w/ pre-train)	**31.36**	**14.16**	**29.35**	**89.23**	**0.5272**	**0.3120**
Transformer+CVAE (w/o pre-train)	27.2	10.70	25.34	82.06	0.4271	0.2864
Transformer+CVAE (w/ pre-train & w/o Gate)	24.21	8.75	22.54	86.76	0.4840	0.2568

Table 5: Selected abstractive summarization results for ablation study in quality metrics.

of 3-layer MLPs with hidden dimension 512 (Lin et al., 2020). As we only use the first sentence of the source and the target, the vocabulary size is 124413. We set the batch size to 128 for the base transformer, CVAE, and Transformer+CVAE, and we shuffle the training data randomly at each epoch. KL annealing, early stopping methods (Bowman et al., 2016), and auxiliary loss (Zhao et al., 2017) are applied in the training process. We use the Adam optimizer (Kingma and Ba, 2014) with initial learning rate 2×10^{-4}. In order to avoid overfitting, we set the dropout rate for each encoder and decoder layer to be 0.1. We adopt greedy decoding for testing ($K = 1$) and the beam size is set to 5.

4.4 Evaluation Metrics

4.4.1 Diversity Metrics

Distinct-N. *Distinct-N*, proposed by Li et al. (2016), is calculated by the number of distinct n-grams divided by the total number of n-grams in all the generated summaries. We apply this metric for unigrams, bigrams, and trigrams (denoted *Dist-1*, *Dist-2*, *Dist-3*) to measure the diversity of generation.

self-BLEU. *self-BLEU* (Zhu et al., 2018a) measures the variety of the generated text. It measures the diversity of a generated sentence based on other generated sentences. Then by taking the average of the BLEU scores of all generated sentences, we get the self-BLEU metric. Note that in contrast to test-BLEU, lower self-BLEU score shows higher diversity. We employ self-BLEU-2 (bigram), self-BLEU-3 (trigram) as our metrics.

4.4.2 Quality Metrics

We evaluate the quality of our generated summaries based on informativeness, fluency and the contextual similarity to the reference (Aralikatte et al., 2021).

ROUGE score. *ROUGE score* (Lin, 2004) calculates the amount of overlapping content as n-grams between the generated text and the reference. To evaluate the lexical overlap between our test results and the reference, we report ROUGE-1 and ROUGE-2 for *informativeness* evaluation and ROUGE-L for *fluency* evaluation.

BERTScore. *BERTScore* (Devlin et al., 2019) computes the the sum of cosine similarities between the hypothesis and reference text in token-level. We use BERTScore as the metric to evaluate semantic similarity between our test results and the reference.

test-BLEU. We use *test-BLEU* (Zhu et al., 2018b) to calculate the similarity between the generated summaries and the reference. The score is within the range $[0, 1]$, and the higher score means better alignment with the real data. We employ test-BLEU-2 (bigram), test-BLEU-3 (trigram) as our metrics.

5 Results

5.1 Diversity Evaluation

The result of diversity evaluation is shown in Table 1. Our proposed model, the Transformer+CVAE with pre-training, ranks the highest in Dist-1 and Dist-2 metrics, exhibiting a greater number of distinct unigrams and bigrams

in the generated summaries. The CVAE seq-to-seq model implemented by us achieves the highest score in Dist-3, reflecting that the CVAE framework efficiently captures semantic variations in the latent space. The Transformer+CVAE model scores the lowest in both self-BLEU-2 and self-BLEU-3, and is close to the reference summaries, which shows stronger diversity in sentence level. One of the reasons why the base Transformer obtains low diversity score is its deterministic nature; and the CVAE RNN-based model is less robust when facing the vanishing latent variable problem, scoring relatively poorer in most diversity metrics. Overall, our proposed model performs the best in boosting diversity in summaries.

5.2 Quality Evaluation

The quality evaluation result is shown in Table 2. Our goal is to maintain good summarization quality when adding diversity to the sentences, and our Transformer+CVAE model achieves great performance in the quality metrics. Comparing to the seq-to-seq state-of-the-art generation models such as ABS+ and RAS-LSTM, our model outperforms both of them in ROUGE-2 and ROUGE-L, presenting higher informativeness and fluency. Our model also defeats the base Transformer model in all ROUGE metrics, showing the ability of CVAE framework to extract core information apart from collecting semantic variations. Our model obtains the highest score in BERTScore and test-BLEUs, exibiting greater similarities to the reference summaries. So our model combines the advantage of attention mechanism and the latent distribution. The CVAE seq-to-seq model ranks the lowest in quality tests, indicating that RNN-based CVAE is less effective in aligning article-summary pairs than transformer-based CVAE. In addition, our model is more efficient in training compared to RNN-based CVAE models.

5.3 Ablation Study

We examines the effect of pre-training and gating mechanism on the diversity and quality of our Transformer+CVAE model. The comparisons are shown in Table 4 and Table 5,respectively.

5.3.1 The Effect of Pre-training

We pre-trained our Transformer+CVAE model with the base Transformer to make it more robust. Compared with the model without pre-training, our model scores higher in Dist-1 and self-BLEUs, but scores lower in Dist-2 and Dist-3. Except for the self-BLEUs, both models show similar performance in diversity metrics. In terms of quality, it is manifest that our model is more powerful than the counterpart without pre-training, ranking the highest in all quality metrics. Therefore, we conclude that pre-training substantially enhances the quality of our generated summaries, but it is not certain whether pre-training affects the extraction of latent meanings to a notable degree.

5.3.2 The Effect of Gating Mechanism

To investigate whether the gated unit we designed is beneficial to achieving our goal, we built another model (with pre-training) where the latent variable $\mathbf{z}$ is directly added to the [SOS] token instead of entering the gated unit. The counterpart without gated unit outperforms our model in Dist-2, Dist-3 and self-BLEU-3, while the latter defeats the former in the remaining diversity metrics. Therefore, there is no substantive evidence that the gating mechanism improves diversity. However, the counterpart without gated unit performs the poorest in ROUGE scores, exposing no enhancement in quality. In conclusion, though exhibiting similar diversity level, our model with gated unit steadily maintains the quality of the generated summaries.

5.4 Case Study

In Table 6, we show some cases of abstractive summaries. First, we observe that the Transformer tends to generate relevant but plain summaries. For example, in Summary 1, the Transformer extracts the core information of the article, but the generated words only come from the article. It is a common problem with the abstractive summaries produced by the Transformer. Second, we discover that the RNN-based CVAE might fail to capture the essential imformation in the article. For example, in Summary 2, obviously the summary focuses on US stocks, but the highlight is the slight recovery of Wall Street. Similar problem with the RNN-based CVAE can be found in Summary 4, where the important subject in the summary is marked "UNK". Last but not least, our Transformer+CVAE model not only captures the core meaning of the article, but it also generates diversified summaries. For example, in Summary 2, our model generates a new phrase "eke out", which does not appear in the

Article 1	japan 's toyota team europe were banned from the world rally championship for one year here on friday in a crushing ruling by the world council of the international automobile federation -lrb- fia -rrb- .
Summary 1	**Transformer**: toyota banned from world rally for year **CVAE**: top seed banned from world cup **Transformer+CVAE**: toyota banned from world championship for a crushing defeat **Reference**: toyota are banned for a year
Article 2	us shares managed modest gains in morning trade thursday as investors looked for bargains after a punishing three-day market selloff and shook off more bleak economic and corporate news .
Summary 2	**Transformer**: wall street gains on bargain-hunting **CVAE**: us stocks rally against dow gains #.## percent **Transformer+CVAE**: wall street ekes out gains after selloff weak economy **Reference**: wall street struggles higher after three-day rout
Article 3	russian president vladimir putin returned sunday a long-lost icon of our lady of vladimir to russia 's orthodox patriarch alexy ii ahead of the easter service in moscow , vowing to bring back other relics lost in the soviet times .
Summary 3	**Transformer**: putin returns to russia 's orthodox patriarch **CVAE**: russian president returns from moscow to orthodox church **Transformer+CVAE**: putin returns to ex-soviet republics lay hold soviet monument to mark start **Reference**: putin hands long-lost icon to orthodox patriarch pledges to return more
Article 4	led by a lone ivory coast army pickup truck , french and west african military convoys set off in jeeps and armored vehicles friday on a mission to secure the lawless west after civil war .
Summary 4	**Transformer**: ivory coast army sets off for west african military convoy **CVAE**: UNK military convoy sent to west coast **Transformer+CVAE**: ivory coast army peacekeepers head for west african convoy **Reference**: french-led troops set off to secure ivory coast 's law

Table 6: Case study of the abstractive summarization.

article but means the same thing as "win something with efforts". Similar examples are in Summary 1, where "defeat" substitutes "ruling by", and in Summary 4, the word "peacekeepers" modify the "african military army".

6 Conclusion

This paper proposes the Transformer+CVAE model, which integrates the CVAE framework into the Transformer by introducing the prior/recognition networks that bridges the Transformer encoder and decoder. We utilize the latent variables generated in the global receptive field of the transformer by fusing them to the starting-of-sequence ([SOS]) of the decoder inputs. To better tune the weights of the latent variables in the sequence, we designed a gated unit to blend the latent representation and the [SOS] token. Our goal is to generate abstractive summaries with greater diversity and keep high quality. To evaluate the effectiveness of the Transformer+CVAE we proposed, we conduct evaluation in diversity and quality. The results show that our model outperforms the base Transformer and RNN-based CVAE in diversity metrics. In the meantime, our model achieves high quality scores compared to state-of-the-art seq-to-seq models, the base Transformer and the RNN-based CVAE. To make our result more robust, we examined the effect of pre-training and gating mechanism on our model and concluded that both pre-training the gating mechanism enhances the quality, while giving support to generating diverse results. In the future, we will try to incorporate the latent variables sequentially into the internal structure of the Transformer decoder, and compare the effectiveness of the sequential model and the global model.

References

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6078–6095, Online. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational autoregressive decoder for neural response generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Som Gupta and S. K Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes.

Hung Le, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2018. Variational memory encoder-decoder.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2091–2100, Copenhagen, Denmark. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 163–169, Melbourne, Australia. Association for Computational Linguistics.

Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. 2020. Variational transformers for diverse response generation.

Danyang Liu and Gongshen Liu. 2019. A transformer-based variational autoencoder for sentence generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction*, AKBC-WEKEX '12, page 95–100, USA. Association for Computational Linguistics.

Preksha Nema, Mitesh Khapra, Anirban Laha, and Balaraman Ravindran. 2018. Diversity driven attention model for query-based abstractive summarization.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Gaetano Rossiello, Pierpaolo Basile, Giovanni Semeraro, MD Ciano, and Gaetano Grasso. 2016. Improving neural abstractive text summarization with prior knowledge. *URANIA*, 16.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2020. Neural abstractive text summarization with sequence-to-sequence models.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Huan Zhao, Jie Cao, Mingquan Xu, and Jian Lu. 2020. Variational neural decoder for abstractive text summarization. *Computer Science and Information Systems*, 17:12–12.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *ACL*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018a. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018b. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.