CONTRASTIVE LEARNING WITH ADVERSARIAL PERTURBATION FOR CONDITIONAL TEXT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence-to-sequence (seq2seq) models for conditional text generation have achieved remarkable performance on various tasks. However, most of them are trained with teacher forcing where ground truth is provided at each time step and never exposed during training to incorrectly generated tokens, which is called "exposure bias". It results in low quality of text generation at test time. In this work, we propose to contrast positive pair of a source and target sentence to negative pairs so that the model is exposed to various valid or incorrect sentences and learns better representation of target sentences. However, training the model with naive contrastive learning is suboptimal. Unlike classification task, there is no common class among the target sentences, so it is too trivial for pretrained language model to distinguish the positive pairs from the randomly sampled negative pairs. In order to tackle the problem, we generate the embedding of negative target sequences by adding small perturbation so that the log likelihood is minimized. Conversely, we construct additional positive example which is distant from the target sentence in embedding space but semantically similar. As a result, the generated positive and negative example prevent trivial solution of contrastive learning and the model better learns to distinguish valid target sequences from incorrect ones. We empirically show that our proposed method improves the generalization of the seq2seq on three generation tasks — machine translation, text summarization and question generation.

1 INTRODUCTION

The sequence-to-sequence (seq2seq) neural network framework (Sutskever et al., 2014; Cho et al., 2014) has successfully tackled a wide range of language generation. Attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) is introduced to tackle long term dependency between a source and target sentence and it has become a core component of the seq2seq model. Moreover, Vaswani et al. (2017) propose a new model, Transformer, which only consists of attention instead of recurrent operation and it is defacto standard for most of sequence generation tasks. Recently, fine-tuning pretrained language models (Raffel et al., 2019; Dong et al., 2019; Song et al., 2019; Lewis et al., 2020) can significantly improves the performance of generation task.

However, the seq2seq models trained only with teacher forcing, where ground truth is provided at each time step and never exposed during training to incorrectly generated tokens, often exhibit "exposure bias" (Ranzato et al., 2016) and results in low quality of text generation at test time. There are several prior works to tackle the problems. One of straight forward methods is to use reinforcement learning such as REINFORCE (Williams, 1992) or its variants to maximize non-differentiable reward (Bahdanau et al., 2017; Paulus et al., 2018) — BLEU (Papineni et al., 2002) or Rouge (Lin & Hovy, 2002). Another approach is to use reinforcement learning or gumbel softmax (Jang et al., 2017; Maddison et al., 2017) for adversarial training with discrete data, where the reward is output of discriminator from Generative Adversarial Networks (GAN) (Zhang et al., 2018a; 2017; Yu et al., 2017).

Even though reinforcement learning increases performance in text generation tasks, it often requires a vast amount of effort for tuning hyper-parameters and training the GAN is notoriously unstable. Moreover, Choshen et al. (2020) show that reinforcement learning for machine translation does not

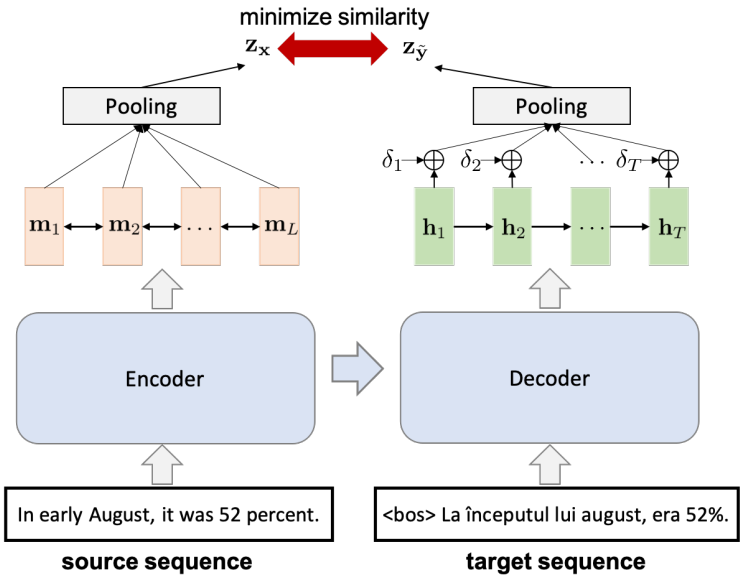


Figure 1: The conceptual illustration of the proposed our contrastive learning framework where we add perturbation δ_t to $\mathbf{h}_t$, the hidden representation of tth token of the target sequence $\mathbf{y}$ to construct negative pair for contrastive learning.

optimize the expected reward and the performance gain is attributed to the unrelated effects such as increasing the peakiness of the output distribution.

To tackle the problem, we propose a novel contrastive learning framework for a sequence generation. We enforce the seq2seq model to maximize a semantic similarity between the hidden representation of source and its corresponding target sequence and minimize the semantic similarity between negative pairs. Naively, we can sample negative pairs by shuffling pairs of source and target sequences so that a source sequence is randomly associated with a target sequence. Unlike conventional classification, there are no common classes among target sequences. Thus it is trivial for models which are pretrained on large corpus to distinguish such positive examples from negative examples. In order to avoid such trivial cases, we adversarially perturb the hidden representation of target sequences so that its likelihood is minimized. We call the perturbed embedding as imposter. Since the imposter is very close to the original representation of target sequence but semantically very dissimilar, it is hard negative example for the model to distinguish from positive examples. Conversely, we construct an additional positive example for contrastive learning by adding perturbation noise to the hidden representation of target sequence so that the perturbed one is distant from the target sequence in embedding space but semantically similar. As a result, the model is exposed to various representations of incorrect or valid sentences with contrastive learning framework.

To verify efficacy of our proposed contrastive learning, we empirically show that our method significantly improves the performance of seq2seq model for several tasks— machine translation, text summarization and question generation.

Our contribution is threefold.

- To tackle exposure bias, we propose a novel method to generate negative pairs for contrastive learning, by adding small perturbation to the hidden representation of target sequence so that its likelihood is minimized. By contrasting positive and negative pairs, the seq2seq model learns to better representation of target sequence.
- We devise a novel method to construct an additional positive pair for contrastive learning by maximizing the contrastive learning loss but minimizing the semantic distance from the ground truth target sequence.
- We empirically show that our method improves the performance of seq2seq model for machine translation, text summarization, and question generation on benchmark dataset.

2 RELATED WORK

Conditional Sequence Generation Since Sutskever et al. (2014); Cho et al. (2014) introduce an encoder-decoder architecture with Recurrent Neural Networks (RNN), it have become defacto standard for conditional language generation. Attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) which learns to focus on certain tokens of encoder hidden representation significantly improves machine translation and it is core component of encoder-decoder architecture. Kalchbrenner et al. (2016); Gehring et al. (2017) replace RNN with convolutional neural network which enables parallel computation. Furthermore, the Transformer (Vaswani et al., 2017) model is introduced to tackle long term dependency with a constant number of operations, which are called self-attention. After, pretrained language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) show impressive performance in several discriminative tasks such as sentence classification and question answering, various pretraining methods (Raffel et al., 2019; Lewis et al., 2020; Dong et al., 2019; Song et al., 2019) are introduced to target language generation and improve the performance of several generation tasks.

Exposure Bias There are several prior works to tackle the exposure bias (Ranzato et al., 2016). Paulus et al. (2018); Bahdanau et al. (2017) leverage reinforcement learning such as REINFORCE (Williams, 1992) or self-critical policy gradient (Rennie et al., 2017) to maximize non-differentiable rewards, so it enables to penalize the model for incorrect target sequence generation. Another stream of works (Zhang et al., 2017; 2018b;a; Yu et al., 2017) train GANs to match the distribution of generated sequences to the distribution of ground truth target sequence. Since sampling tokens from the generator is not differentiable, they resort reinforcement learning or gumbel-softmax to train the networks in end-to-end fashion. However, training the model with reinforcement learning or GAN requires vast amount of effort to tune hyper-parameters and the training is notoriously unstable. Moreover, Choshen et al. (2020) show that reinforcment learning for machine translation does not optimize the expected reward and the performance gain is attributed to the unrelated effects such as increasing the peakiness of the output distribution.

Contrastive Learning Contrastive representation learning has been widely used in previous works. They are based on the method which learns representation by contrasting positive pairs and negative pairs. Chopra et al. (2005); Weinberger & Saul (2009); Schroff et al. (2015) leverage a triplet loss with a max-margin approach to separate positive examples from negative examples in metric learning. Wu et al. (2018) propose a memory bank to store the representation of each instances and sample the instances for contrastive learning. Recently, Chen et al. (2020) show that contrastive learning can boost the performance of self-supervised and semi-supervised learning. In Natural Language Processing (NLP), contrastive learning has been a key component. In Word2Vec (Mikolov et al., 2013), neighbouring words are predicted from context with noise-contrastive estimation (Gutmann & Hyvärinen, 2012). Beyond word representation, Logeswaran & Lee (2018) sample two contiguous sentences for positive pairs and sentences from other document for negative pairs. They constrast positive and negative pairs to learn sentence representation. Moreover, contrastive learning has been investigated in various NLP tasks such as language modelling (Huang et al., 2018), unsupervised word alignment (Liu & Sun, 2015), caption generation (Mao et al., 2016; Vedantam et al., 2017), and machine translation (Yang et al., 2019).

3 METHOD

3.1 BACKGROUND: CONDITIONAL SEQUENCE GENERATION

Our goal is to generate a sequence $\mathbf{y} = (y_1, \ldots, y_T)$ with length T conditioning on another sequence $\mathbf{x} = (x_1, \ldots, x_L)$ with length L. Typical approach to the conditional sequence generation is to leverage encoder-decoder architecture (Sutskever et al., 2014; Cho et al., 2014) to parameterize the following conditional distribution and maximize the log likelihood $\log p_\theta(\mathbf{y}|\mathbf{x})$. The letters f, g denote encoder and decoder and $\mathbf{M} = [\mathbf{m}_1 \cdots \mathbf{m}_L] \in \mathbb{R}^{d \times L}$ is concatenation of hidden representa-

tion of source tokens $\mathbf{x}$.

$$\begin{aligned}
\mathcal{L}_{MLE}(\theta) &= \log p_\theta(\mathbf{y}|\mathbf{x}) \\
p_\theta(y_1, \ldots, y_T|\mathbf{x}) &= \prod_{t=1}^{T} p_\theta(y_t|\mathbf{y}_{<t}, \mathbf{x}) \\
p_\theta(y_t|\mathbf{y}_{<t}, \mathbf{x}) &= \text{softmax}(W\mathbf{h}_t + \mathbf{b}) \\
\mathbf{h}_t &= g(y_{t-1}, \mathbf{M}; \theta), \;\; \mathbf{M} = f(\mathbf{x}; \theta)
\end{aligned} \tag{1}$$

3.2 Contrastive Learning

We propose a contrastive learning for conditional sequence generation to train the seq2seq model to be exposed to various incorrect target sequences and valid multiple target sequences. We train the model to learn representation of the ground truth sentence by contrasting positive pairs to negative pairs. To be specific, we project source sentence and target sentence into latent embedding space and maximize semantic similarity of the two sentences with contrastive learning framework (Chen et al., 2020) as follows:

$$\begin{aligned}
\mathcal{L}_{cont}(\theta) &= \log \frac{\exp(\text{sim}(\mathbf{z_x}, \mathbf{z_y})/\tau)}{\sum_{\mathbf{z} \in S} \exp(\text{sim}(\mathbf{z_x}, \mathbf{z})/\tau)} \\
\mathbf{z_x} &= \xi(\mathbf{M}; \theta), \; \mathbf{z_y} = \xi(\mathbf{H}; \theta) \\
\mathbf{M} &= f(\mathbf{x}; \theta), \; \mathbf{H} = g(\mathbf{y}, \mathbf{M}; \theta)
\end{aligned} \tag{2}$$

where ξ denotes the composition of affine transformation with ReLU activation (Nair & Hinton, 2010) and average pooling to compute the fixed sized representation of a sentence $\mathbf{z} \in \mathbb{R}^d$, $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_T] \in \mathbb{R}^{d \times T}$ is a concatenation of decoder hidden states across all the time steps, S is a set of hidden representations of sentence $\hat{\mathbf{y}}$ that is not paired with the source sentence $\mathbf{x}$, and $\text{sim}(\cdot, \cdot)$ is a cosine similarity function.

Contrary to SimCLR (Chen et al., 2020), where the negative pairs might fall into the same category of images, there is no common class among the negative samples S. Therefore, it is trivial for pretrained language models to distinguish $S = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \setminus \{\mathbf{y}_i\}$, negative samples of $\mathbf{x}_i$ from the positive sample $\mathbf{y}_i$. In order to tackle such an issue, we propose a novel method to automatically construct the negative example that is difficult for model to classify and also useful for learning representation. We add a small perturbation $\boldsymbol{\delta} = [\delta_1 \cdots \delta_T] \in \mathbb{R}^{d \times T}$ to the $\mathbf{H}$, hidden representation of target sequence $\mathbf{y}$ so that the log likelihood is minimized as follows:

$$\begin{aligned}
\tilde{\mathbf{H}} = \mathbf{H} + \boldsymbol{\delta} \text{ where } \boldsymbol{\delta} &= \underset{\boldsymbol{\delta}, ||\boldsymbol{\delta}||_2 \leq \epsilon}{\arg\min} \log p_\theta(\tilde{\mathbf{y}}|\mathbf{x}) \\
p_\theta(\tilde{y}_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1}, \mathbf{x}) &= \text{softmax}(W\tilde{\mathbf{h}}_t + \mathbf{b})
\end{aligned} \tag{3}$$

Since exact minimization with respect to $\boldsymbol{\delta}$ is intractable for deep neural networks, it is impossible to compute the exact value of $\boldsymbol{\delta}$. Following Goodfellow et al. (2015), we approximate it by linearizing $\log p_\theta(\mathbf{y}|\mathbf{x})$ around $\mathbf{y}$ as follows:

$$\tilde{\mathbf{H}} = \mathbf{H} - \epsilon \frac{\boldsymbol{g}}{||\boldsymbol{g}||_2} \text{ where } \boldsymbol{g} = \nabla_{\mathbf{H}} \log p_\theta(\mathbf{y}|\mathbf{x}) \tag{4}$$

As shown in Figure 1, we add small perturbation to the hidden representation of each token of target sentence $\mathbf{y}$ so that its log likelihood is minimized. Thus, the perturbed $\tilde{\mathbf{H}}$ is semantically very dissimilar to $\mathbf{H}$, the hidden representation of target sequence $\mathbf{y}$ but very close to it in the latent embedding space. It is not trivial for the sequence-to-sequence model to distinguish it from the representation of true target sequence $\mathbf{y}$. We use the perturbed representation as an additional negative sample for contrastive learning as follows:

$$\begin{aligned}
\mathbf{z}_{\tilde{\mathbf{y}}} &= \xi(\tilde{\mathbf{H}}; \theta) \\
\mathcal{L}_{neg-cont}(\theta) &= \log \frac{\exp(\text{sim}(\mathbf{z_x}, \mathbf{z_y})/\tau)}{\sum_{\mathbf{z} \in S \cup \{\mathbf{z}_{\tilde{\mathbf{y}}}\}} \exp(\text{sim}(\mathbf{z_x}, \mathbf{z})/\tau)}
\end{aligned} \tag{5}$$

Moreover, we construct an additional positive pair of source sequence $\mathbf{x}$ by adding perturbation to the hidden state of target sequence $\mathbf{y}$ so that cosine distance from $\mathbf{z}_x$ is maximized but distance from the distribution $p_\theta(\mathbf{y}|\mathbf{x})$ is minimized. To be specific, we add perturbation to $\mathbf{H}$ so that it maximizes the contrastive learning loss $\mathcal{L}_{cont}(\theta)$ at the first place. Then we add another perturbation to minimize the kl divergence between perturbed distribution and $p_\theta(\mathbf{y}|\mathbf{x})$ as follows, where $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1 \cdots \hat{\mathbf{h}}_T] \in \mathbb{R}^{d \times T}$.

$$\begin{aligned}
\hat{\mathbf{H}} &= \mathbf{H} + \eta \frac{\mathbf{g}}{||\mathbf{g}||_2} \text{ where } \mathbf{g} = \nabla_{\mathbf{H}} \mathcal{L}_{cont}(\theta) \\
p_\theta(\hat{y}_t|\mathbf{x}) &= \text{softmax}(W\hat{\mathbf{h}}_t + \mathbf{b}) \\
\hat{\mathbf{H}} &= \hat{\mathbf{H}} - \eta \frac{\mathbf{f}}{||\mathbf{f}||_2} \text{ where } \mathbf{f} = \nabla_{\hat{\mathbf{H}}} \mathcal{L}_{KL}(\theta) \\
\mathcal{L}_{KL}(\theta) &= \sum_{t=1}^{T} D_{KL}(p_\theta(y_t|\mathbf{y}_{<t}, \mathbf{x})||p_\theta(\hat{y}_t|\mathbf{y}_{<t}, \mathbf{x})
\end{aligned} \tag{6}$$

We consider the perturbed hidden state $\hat{\mathbf{H}}$ as positive pair of source sequence $\mathbf{x}$, thus we use it as an additional positive example for contrastive learning and minimize $\mathcal{L}_{KL}$ as follows:

$$\begin{aligned}
\mathcal{L}_{cont-pos}(\theta) &= \log \frac{\exp(\text{sim}(\mathbf{z_x}, \mathbf{z_y})/\tau)}{\sum_{\mathbf{z} \in S} \exp(\text{sim}(\mathbf{z_x}, \mathbf{z})/\tau)} \\
\hat{\mathbf{z}}_{\mathbf{y}} &= \xi(\hat{\mathbf{H}}; \theta)
\end{aligned} \tag{7}$$

We estimate the parameters of the sequence-to-sequence model, θ with maximizing the following objective, where α, β are hyperparameters which control the importance of contrastive learning and KL divergence. For all the experiments, we set α and β as 1, which we search through cross-validation. One thing to note that the pooling layer ξ is removed and the encoder f and decoder g is used for generation after the training is done.

$$\underset{\theta}{\text{maximize}}\, \mathcal{L}_{MLE}(\theta) + \alpha \mathcal{L}_{KL}(\theta) + \beta\{\mathcal{L}_{cont-neg}(\theta) + \mathcal{L}_{cont-pos}(\theta)\} \tag{8}$$

4 EXPERIMENT

4.1 TASKS

Machine Translation For machine translation, we use WMT16 Romanian-English parallel corpus (WMT'16 RO-EN) to train the model without any data augmentation such as back-translation. We tokenize the pair of source and target sequences $(\mathbf{x}, \mathbf{y})$ with the same tokenizer as Raffel et al. (2019). We finetune the pretrained T5-small model for 20 epochs with batch size 128 and Adafactor (Shazeer & Stern, 2018). For contrastive learning, we set the norm of perturbation, ϵ and η as 3.0

Text Summarization For text summarization, we use XSum dataset (Narayan et al., 2018) of which summaries are highly abstractive compared to CNN/DailyMail Dataset (Hermann et al., 2015) where extractive summarization models perform well. We follow most of the experimental settings for machine translation as described above, except that we set the norm of perturbation, ϵ and η as 1.0.

Question Generation For question generation, we want to generate a question from a given answer and paragraph, i.e., we model conditional distribution $p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{a})$ where $\mathbf{x}, \mathbf{y}, \mathbf{a}$ denote a paragraph, question and answer, respectively. We concatenate the answer and paragraph with special tokens to generate the question conditioned on both of the answer and paragraph. As the previous experimental settings, we finetune T5-small model on SQuAD dataset (Rajpurkar et al., 2016), which is originally constructed for question answering, for 20 epochs with batch size 128 but set the norm of perturbation, η as 3.0 and ϵ as 1.0. Since the test set of SQuAD is only accessible via leader board, we randomly validation set and split into new validation and test set, following Zhao et al. (2018).

4.2 EXPERIMENTAL SETUPS

Implementation Details For the encoder f, and decoder g, we use encoder and decoder of T5-small model, which is based on transformer architecture. However, our method is not limited to a specific

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SacreBLEU	EM/F1
			SQuAD			
Maxout-QG						
Harvesting-QG	-	-	20.90	15.16	-	66.05/54.62
T5-small	41.26	30.30	23.38	18.54	21.00	67.64/55.91
WordDropout Contrastive	41.37	30.50	23.58	18.71	21.19	68.16/56.41
MLE-contrastive	41.23	30.28	23.33	18.45	20.91	67.32/55.25
MLE-neg-contrastive	41.65	30.69	23.71	18.81	21.25	68.26/56.41
MLE-pos-neg-contrastive	**42.33**	**31.29**	**24.22**	**19.19**	**21.55**	**69.01/57.06**
			WMT'16 RO-EN			
Transformer	50.36	37.18	28.42	22.21	26.17	
T5-small	57.76	44.45	35.12	28.21	32.43	
WordDropout Contrastive	57.77	44.45	35.12	28.21	32.44	
MLE-contrastive	57.64	44.12	34.74	27.79	32.03	
MLE-neg-contrastive	57.90	44.60	35.27	28.34	32.55	
MLE-pos-neg-contrastive	**58.98**	**45.72**	**36.39**	**29.41**	**33.96**	

Table 1: BLEU scores on WMT'16 RO-EN and SQuAD for machine translation and question generation.

model and it is generally applicable to any seq2seq model with encoder-decoder architecture. For projection layer ξ, we use use average pooling across time dimension to compute fixed size hidden representation of $\mathbf{x}, \mathbf{y}$ and one layered MLP with ReLU non-linear activation (Nair & Hinton, 2010). We use cosine similarity for $\text{sim}(\cdot, \cdot)$ and set the temperature, τ as 0.1 for all the experiments. At test time, we use beam search of width 4 for all our models to generate the target sequences.

Common Baselines We perform experiments our method against several baselines for all the three tasks as follows:

1. **MLE**: We finetune the pretrained T5 model to maximize the log likelihood, $\mathcal{L}_{MLE}(\theta)$ without any auxiliary loss.
2. **WordDropout Contrastive**: Proposed by Yang et al. (2019), they train the sequence-to-sequence model with the maximum likelihood estimation, and further train. For a negative example, they remove the most frequent word of target sequence and further train the model to assign higher probability to the ground truth target pair than the frequent word omitted sequence.
3. **MLE-contrastive**: In addition to the log likelihood, we maximize the contrastive learning objective from equation (2) without using perturbed embeddings for negative examples.
4. **MLE-neg-contrastive**: Our proposed method which jointly maximize the log likelihood and contrastive learning objective with perturbed negative examples but does not use any additional positive example.
5. **MLE-pos-neg-contrastive** Our proposed method which jointly maximize the log likelihood, contrastive learning objective, and KL-divergence as described in the equation (8).

Task specific baselines For machine translation, we train the Transformer (Vaswani et al., 2017) which consists of 6 layers of self-attention layer with 8 multi-head attention and 512 dimension. For question generation, we additionally compare our models against the LSTM models with copy mechanism, which we call Maxout-QG (Zhao et al., 2018) and Harvesting-QG (Du & Cardie, 2018). For text summarization, we set another baselines such as PTGEN-COVG (See et al., 2017) which use copy mechanism and coverage to handle out of vocabulary word and prevent the repetition, and CONVS2S (Narayan et al., 2018) which use convolutional networks for encoder and decoder.

4.3 Empirical Results

Evaluation Metric Following the conventional evaluation metrics for each tasks, we adopt BLEU (Papineni et al., 2002) and SacreBLEU (Post, 2018) for machine translation and question generation, and Rouge (Lin & Hovy, 2002) for text summarization as automatic evaluation protocols. Moreover,

Method	Rouge-1	Rouge-2	Rouge-L
	XSum		
PTGEN-COVG	28.10	8.02	21.72
CONVS2S	31.89	11.54	25.75
T5-small	36.10	14.72	29.16
WordDropout Contrastive	36.88	15.11	29.79
MLE-contrastive	36.34	14.81	29.41
MLE-neg-contrastive	37.72	15.49	**30.74**
MLE-pos-contrastive	**37.89**	**15.78**	30.59

Table 2: Rouge scores on Xsum dataset for text summarization.

we generate the questions from the Harvesting QADu & Cardie (2018) which consists of top-ranking 10K Wikipedia articles and train a BERT QA model with the generated QA pairs. We evaluate the QA model on our SQuAD test set and report F1 and Exact Match (EM) to measure the quality of the model generated questions.

Quantitative Results We compare our model with the baseline models on WMT'16 RO-En, XSum, SQuAD dataset for machine translation, text summarization and question generation, respectively. Table 1 shows that our proposed method, contrastive learning with the perturbed positive example and negative example significantly outperforms the other baseline. It improves about 1% all BLEU scores. Note that naively constructing the negative examples for contrastive learning, where the pairs of $(\mathbf{x}, \mathbf{y})$ from mini-batch are shuffled so that the source sequence is randomly associated with the target sequence and used for negative pairs, degrades the performance. For question generation, our proposed method improves F1/EM as well as BLEU scores. It shows that our proposed model is able to generate semantically valid questions so those generated questions are beneficial for training the QA model.

Similarly, our model outperforms all the other baseline models on Xsum dataset for text summarization, as shown in Table 2. Contrary to the previous experiments, contrastive learning with imposters alone can improve the quality of summarization and using additional positive example for contrastive learning results in marginal improvement.

Ablation Studies We further perform an ablation study to see the effect of each model component. We start the model trained with maximum likelihood estimation (denoted as MLE in Table 3). Then we add naive contrastive learning (+Contrastive) loss in equation 2 to examine the necessity of the imposter and the positive example. Additionally, we generate imposter (+Imposter) and positive example (+Positive) and use them for additional training instances for contrastive learning but do not update the parameter with respect to the loss $\mathcal{L}_{KL}(\theta)$ in equation 6. Finally we train the full model (+KL-div) and compare it to the model (+Noise) trained with additional regularization as follows:

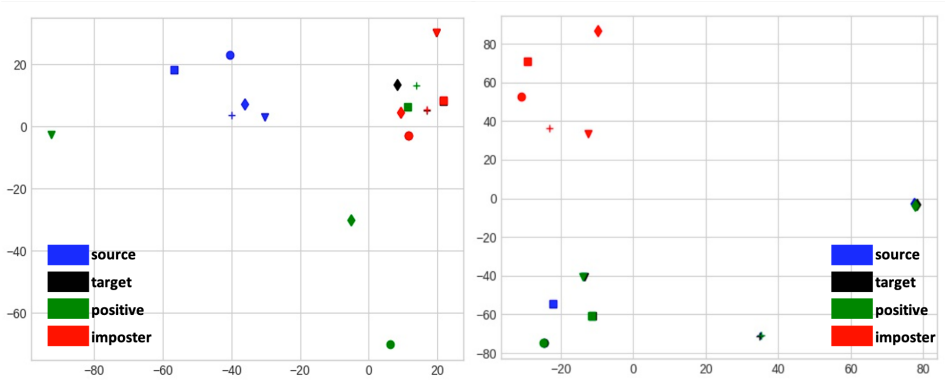
Table 3: Ablation study on machine translation.

	BLEU
MLE	32.42
+Noise	28.33
+Contrastive	32.03
+Imposter	32.55
+Positive	31.87
+KL-div.	**33.96**

$$\mathcal{L}_{reg}(\theta) = \sum_{t=1}^{T} D_{KL}(p_\theta(y_t|\mathbf{y}_{<t}, \mathbf{x})||p_\theta(\overline{y}_t|\mathbf{y}_{<t}, \mathbf{x})$$
$$p_\theta(\overline{y}_t|\mathbf{y}_{<t}, \mathbf{x}) = \text{softmax}(W\overline{\mathbf{h}}_t + \mathbf{b}) \text{ where } \overline{\mathbf{h}}_t = \mathbf{h}_t + \varepsilon \;\; \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \tag{9}$$

As shown in Table 3 along with the previous experimental results, naive contrastive learning degrades the performance. Unlike conventional classification, there is no common class among target sequences, so it is very trivial for pretrained language model to distinguish positive pair from the randomly associated negative pairs. Compare to the result that adding Gaussian noise degrades the performance as shown in Table 3, our method to construct positive example and minimize the KL divergence between the ground truth target sequence and the positive example is effective for model generalization.

Visualization To examine our model with proposed contrastive learning framework learns meaningful representation of sentences, we encode a pair of sequences $(\mathbf{x}, \mathbf{y})$ into $\mathbf{M}, \mathbf{H}$ with encoder



(a) Finetune without contrastive learning (b) Finetune with contrastive learning

Figure 2: Visualization of the embeddings of source sentence, target sentence, imposter and positive perturbation.

f and decoder g. Then we add perturbations to $\mathbf{H}$ to construct imposter $\tilde{\mathbf{H}}$ and additional positive example $\hat{\mathbf{H}}$ as shown in equation (3) and (6). We apply average pooling to $\mathbf{M}, \mathbf{H}, \tilde{\mathbf{H}}$, and $\hat{\mathbf{H}}$ to get fixed size representation and project them into two dimensional space with t-SNE (Maaten & Hinton, 2008). As shown in Figure 2-(b), the model pushes away the imposter from the embedding of target sequence and pulls the embedding of source sequence and positive example by minimizing the contrastive learning loss. The model without contrastive learning, however, the embedding of target sequence is far away from that of source sequence and imposter is very close to it as shown in Figure 2-(a).

5 CONCLUSION

We proposed a novel contrastive learning framework for conditional sequence generation, which maximizes the similarity between ground truth source sequence and target sequence and minimize it among negative pairs of sequences. Moreover, we devised a new method to construct hard negative example, which is semantically dissimilar but close to the embedding, and additional positive example which is far from the target sequence in embedding space but semantically similar and leverage those examples to learn better representation of target sequences. We empirically showed that our method improves the performance of seq2seq model on machine translation, question generation, and text summarization.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. On the weaknesses of reinforcement learning for neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pp. 13063–13075, 2019.

Xinya Du and Claire Cardie. Harvesting paragraph-level question-answer pairs from wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1907–1917, 2018.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML*, 2017.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 2012.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, 2015.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. Large margin neural language model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–1191, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 2020.

Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*, 2002.

Yang Liu and Maosong Sun. Contrastive unsupervised word alignment with non-local features. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,*, 2015.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002*, 2002.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pp. 2227–2237, 2018.

Matt Post. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018.

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. *arXiv preprint arXiv:1910.10683*, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, 2017.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*, 2018.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pp. 5926–5936, 2019.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 2009.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6191–6196, 2019.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, 2017.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 2018a.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pp. 1810–1820, 2018b.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3901–3910, 2018.