CONTRASTIVE LEARNING WITH ADVERSARIAL PERTURBATIONS FOR CONDITIONAL TEXT GENERATION

Anonymous authors
Paper under double-blind review

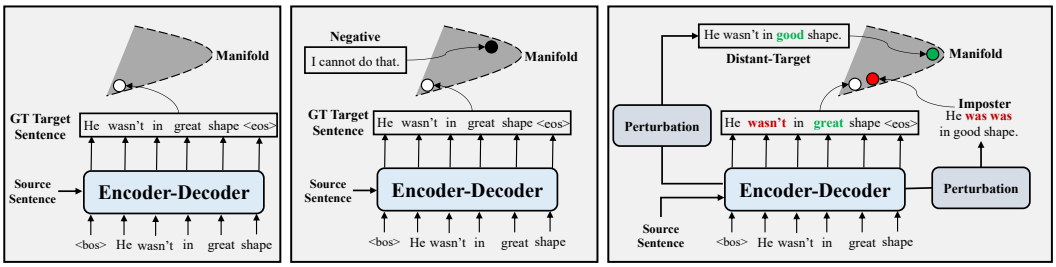
ABSTRACT

Recently, sequence-to-sequence (seq2seq) models with the Transformer architecture have achieved remarkable performance on various conditional text generation tasks, such as machine translation. However, most of them are trained with *teacher forcing* with the ground truth label given at each time step, without being exposed to incorrectly generated tokens during training, which hurts its generalization to unseen inputs, that is known as the "exposure bias" problem. In this work, we propose to solve the conditional text generation problem by contrasting positive pairs with negative pairs, such that the model is exposed to various valid or incorrect perturbations of the inputs, for improved generalization. However, training the model with naïve contrastive learning framework using random non-target sequences as negative examples is suboptimal, since they are easily distinguishable from the correct output, especially so with models pretrained with large text corpora. Also, generating positive examples requires domain-specific augmentation heuristics which may not generalize over diverse domains. To tackle this problem, we propose a principled method to generate positive and negative samples for contrastive learning of seq2seq models. Specifically, we generate negative examples by adding small perturbations to the input sequence to minimize its conditional likelihood, and positive examples by adding large perturbations while enforcing it to have a high conditional likelihood. Such "hard" positive and negative pairs generated using our method guides the model to better distinguish correct outputs from incorrect ones. We empirically show that our proposed method significantly improves the generalization of the seq2seq on three text generation tasks — machine translation, text summarization, and question generation.

1 INTRODUCTION

The sequence-to-sequence (seq2seq) models (Sutskever et al., 2014), which learn to map an arbitrary-length input sequence to another arbitrary-length output sequence, have successfully tackled a wide range of language generation tasks, including machine translation, text summarization, question generation, to name a few. Early seq2seq models have used recurrent neural networks to encode and decode sequences, leveraging attention mechanism (Bahdanau et al., 2015) that allows the decoder to attend to a specific token in the input sequence to capture long-term dependencies between the source and target sequences. Recently, the Transformer (Vaswani et al., 2017), which is an all-attention model that effectively captures long-term relationships between tokens in the input sequence as well as across input and output sequences, has become the de facto standard for most of the text generation tasks due to its impressive performance. Moreover, Transformer-based language models trained on large text corpora (Song et al., 2019; Dong et al., 2019; Raffel et al., 2020; Lewis et al., 2020) have shown to significantly improve the model performance on text generation tasks.

However, a crucial limitation of seq2seq models is that they are mostly trained only with *teacher forcing*, where ground truth is provided at each time step and thus are never exposed to incorrectly generated tokens during training (Fig. 1-(a)), which hurts its generalization. This problem is known as the "exposure bias" problem (Ranzato et al., 2016) and often results in the generation of low-quality texts on unseen inputs. Several prior works tackle the problem, including a straightforward method to use reinforcement learning (RL) to maximize non-differentiable reward (Bahdanau et al., 2017; Paulus et al., 2018) — BLEU or Rouge. Another approach is to use RL or gumbel softmax (Jang et al., 2017; Maddison et al., 2017) to match the distribution of generated sentences to that



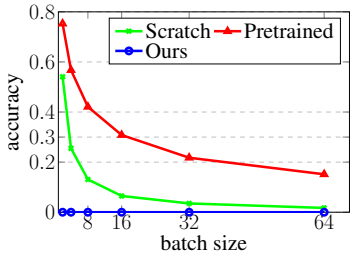
(a) Teacher Forcing (b) Randomly Sampled Negative Example (c) Imposter / Distant-Target Generation with perturbation

Figure 1: **Concept.** (a) Training seq2seq with *teacher forcing*. (b) Naïve contrastive learning with randomly sampled negative examples. (c) Our method, CLAPS, which generates hard negative and positive examples.

of the ground truth, in which case the reward is the discriminator output from a Generative Adversarial Network (GAN) (Zhang et al., 2018; 2017; Yu et al., 2017). Although the aforementioned approaches improve the performance of the seq2seq models on text generation tasks, they either require a vast amount of effort in tuning hyperparameters or stabilize training.

In this work, we propose to tackle the exposure bias problem with a simple yet effective approach, in which we *contrast* a positive pair of input and output sequence to negative pairs, to expose the model to various valid or incorrect sentences. Naïvely, we can construct negative pairs by simply using random non-target sequences from the batch (Chen et al., 2020). However, such a naïve construction yields meaningless negative examples that are already well-discriminated in the embedding space (Fig. 1-(b)), which we highlight as the reason why existing methods (Chen et al., 2020) require large batch size. This is clearly shown in Fig. 2, where a large portion of positive-negative pairs can be easily discriminated without any training, which gets worse as the batch size decreases as it will reduce the chance to have meaningfully difficult examples in the batch. Moreover, discriminating positive and naïve negative pairs becomes even more easier for models pretrained on large text corpora.

Figure 2: Accuracy of classifying a positive pair from negative pairs varying batch size **without any training.**



To resolve this issue, we propose principled approaches to automatically generate negative and positive pairs for constrastive learning, which we refer to as *Contrastive Learning with Adversarial Perturbation for Seq2seq learning* (CLAPS). Specifically, we generate a negative example by adding a small perturbation to the hidden representation of the target sequence, such that its conditional likelihood is minimized (Denoted as the red-colored circle in Fig. 1-(c)). Conversely, we construct an additional positive example (Denoted as green colored circle in Fig. 1-(c)) by adding a large amount of perturbation to the hidden representation of target sequence such that the perturbed sample is far away from the source sequence in the embedding space, while enforcing it to have high conditional likelihood by minimizing Kullback-Leibler (KL) divergence between the original conditional distribution and perturbed conditional distribution. This will yield a negative example that is very close to the original representation of target sequence in the embedding space but is largely dissimilar in the semantics, while the generated positive example is far away from the original input sequence but has the same semantic as the target sequence. This will generate difficult examples that the model fails to correctly discriminate (Fig. 1-(c), Fig.2), helping it learn with more meaningful pairs.

To verify the efficacy of our method, we empirically show that it significantly improves the performance of seq2seq model on three conditional text generation tasks, namely machine translation, text summarization and question generation. Our contribution in this work is threefold:

- To tackle the exposure bias problem, we propose a contastive learning framework for conditional sequence generation, which contrasts a positive pair of source and target sentence to negative pairs in the latent embedding space, to expose the model to various valid or incorrect outputs.
- To tacke the ineffectiveness of conventional approach for constructing negative and positive examples for contrastive learning, we propose a principled method to automatically generate negative and positive pairs, that are more difficult and allows to learn more meaningful representations.
- We show that our proposed method, CLAPS, significantly improves the performance of seq2seq model on three different tasks: machine translation, text summarization, and question generation.

2 RELATED WORK

Conditional Sequence Generation Sutskever et al. (2014); Cho et al. (2014) introduce an encoder-decoder architecture with recurrent neural networks for conditional text generation. Attention (Bahdanau et al., 2015; Luong et al., 2015) which learns to focus on certain tokens of an input sentence significantly improves machine translation and becomes core component of encoder-decoder architecture. Furthermore, Transformer (Vaswani et al., 2017) is introduced to tackle long term dependency within a constant number of operations. After pretrained language models (Peters et al., 2018; Devlin et al., 2019), show impressive performance in several discriminative tasks, various pretraining methods (Raffel et al., 2020; 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 RL such as REINFORCE (Williams, 1992) or self-critic policy gradient (Rennie et al., 2017) to maximize non-differentiable rewards, so it enables to penalize the model for incorrectly generated sentences. Another works (Zhang et al., 2017; 2018; Yu et al., 2017) train GANs to match the distribution of generated sequences to that of ground truth. Since sampling tokens from the generator is not differentiable, they resort RL or gumbel-softmax to train the networks in end-to-end fashion. However, they require either a large amount of effort to tune hyperparameters or stabilize training. Recently, Choshen et al. (2020) show that RL 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 learning has been widely used in previous works. They are based on the method which learns a representation by contrasting positive pairs and negative pairs. Chopra et al. (2005); Weinberger & Saul (2009); Schroff et al. (2015) leverage a triplet loss to separate positive examples from negative examples in metric learning. Recently, Chen et al. (2020) show that contrastive learning can boost the performance of self-supervised and semi-supervised learning in computer vison tasks. In natural language processing (NLP), contrastive learning has been widely used. 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 the sentences from other document as 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 modeling (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 TEXT GENERATION

The goal of conditional text generation with a seq2seq model is to generate an output text sequence $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_T^{(i)})$ with length T conditioned on the input text sequence $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_L^{(i)})$ with length L. A typical approach to the conditional text generation is to leverage the encoder-decoder architecture to parameterize the conditional distribution. We maximize the conditional log likelihood $\log p_\theta(\mathbf{y}|\mathbf{x})$ for a given N observations $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$ as follows:

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

where f, g denote the encoder and the decoder respectively and $\mathbf{M}^{(i)} = [\mathbf{m}_1^{(i)} \cdots \mathbf{m}_L^{(i)}] \in \mathbb{R}^{d \times L}$ is the concatenation of the hidden representations of the source tokens $\mathbf{x}^{(i)}$.

3.2 CONTRASTIVE LEARNING WITH ADVERSARIAL PERTURBATIONS FOR SEQ2SEQ

Since most of the seq2seq models are trained with teacher forcing where the ground truth tokens are provided to maximize Eq. 1, they are never exposed to incorrectly generated tokens during training, which is known as the "expousre bias" problem. In order to tackle the problem, we propose a contrastive learning framework to expose the model to various valid or incorrect output sequences for a given input sentence. Following the contrastive learning framework (Chen et al., 2020), we can train the model to learn the representations of the ground truth sentence by contrasting the positive pairs with the negative pairs, where we select the negative pairs as a random non-target output sequence from the same batch. As shown in Fig. 3-(a), we project the source and target text sequences onto the latent embedding space. Then we maximize the similarity between the pair of source and target sequence, while minimizing the similarity between the negative pairs as follows:

$$\begin{aligned} \mathcal{L}_{cont}(\theta) &= \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(i)})/\tau)}{\sum_{\mathbf{z}_{\mathbf{y}}^{(j)} \in S} \exp(\text{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(j)})/\tau)} \\ \mathbf{z}_{\mathbf{x}}^{(i)} &= \xi(\mathbf{M}^{(i)}; \theta),\ \mathbf{z}_{\mathbf{y}}^{(i)} = \xi(\mathbf{H}^{(i)}; \theta) \\ \xi([\mathrm{v}_1 \cdots \mathrm{v}_T]; \theta) &\coloneqq \text{AvgPool}([\mathbf{u}_1 \cdots \mathbf{u}_T]), \text{ where } \mathrm{u}_t = \text{ReLU}(\mathbf{W}^{(1)}\mathrm{v}_t + \mathbf{b}^{(1)}) \end{aligned} \tag{2}$$

where ξ denotes the composition of affine transformation with the ReLU (Nair & Hinton, 2010) and average pooling to compute the fixed sized representation of a sentence $\mathbf{z} \in \mathbb{R}^d$, $\mathbf{H}^{(i)} = [\mathbf{h}_1^{(i)} \cdots \mathbf{h}_T^{(i)}] \in \mathbb{R}^{d \times T}$ is a concatenation of the decoder hidden states of the target sentence $\mathbf{y}^{(i)}$ across all the time steps. Furthermore, $S = \{\mathbf{z}_{\mathbf{y}}^{(j)} : j \neq i\}$ is a set of hidden representations of target sentences (the objects other than circles in Fig. 3-(a)) that are randomly sampled and not paired with the source sentence $\mathbf{x}^{(i)}$, and $\text{sim}(\cdot, \cdot)$ is a cosine similarity function.

However, training the model with naïve contrastive learning framework using random non-target sequences as negative examples is highly suboptimal, as described in the introduction and shown in Fig. 1. Many of such naïve negative examples are often located far away from the positive examples in the embedding space from the beginning, when using the pretrained language model. Therefore, simply using the examples from the same batch as done in Chen et al. (2020) will result in trivial negative examples and require very large batch size to enable sampling meaningful negative pairs within the same batch. Moreover, generating positive examples for text sequences is not a trivial problem either since for text domains, we do not have a well-defined set of augmentation methods that preserves the input semantics, unlike with the image domains. To tackle such difficulties, we propose a principled method to automatically construct the adversarial negative and positive examples, such that the samples are difficult for the model to classify correctly. These adversarial positive/negative pairs can guide the model to learn a more accurate representation of the target text sequence, by identifying which features make the output positive or negative (See Fig. 1-(c)).

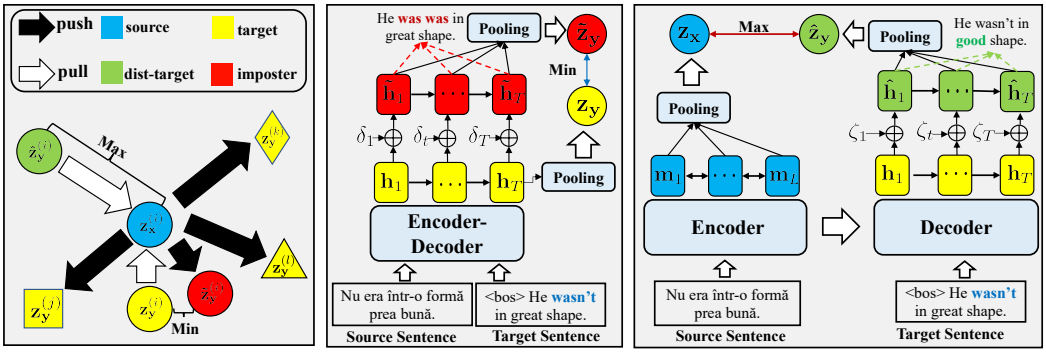
3.3 GENERATION OF IMPOSTERS

As shown in Fig. 3-(b), to generate a negative example, we add a small perturbation $\boldsymbol{\delta}^{(i)} = [\delta_1^{(i)} \cdots \delta_T^{(i)}] \in \mathbb{R}^{d \times T}$ to the $\mathbf{H}^{(i)}$, which is the hidden representation of target sequence $\mathbf{y}^{(i)}$, such that its conditional likelihood is minimized as follows:

$$\begin{aligned} \tilde{\mathbf{H}}^{(i)} = \mathbf{H}^{(i)} + \boldsymbol{\delta}^{(i)} \text{ where } \boldsymbol{\delta}^{(i)} &= \underset{\boldsymbol{\delta}, ||\boldsymbol{\delta}||_2 \leq \epsilon}{\arg\min} \log p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \mathbf{H}^{(i)} + \boldsymbol{\delta}) \\ p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \mathbf{H}^{(i)} + \boldsymbol{\delta}) &= \prod_{t=1}^{T} p_\theta(y_t^{(i)}|\mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}; \mathbf{h}_t^{(i)} + \delta_t) \\ p_\theta(y_t^{(i)}|\mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}; \mathbf{h}_t^{(i)} + \delta_t) &= \text{softmax}\{\mathbf{W}(\mathbf{h}_t^{(i)} + \delta_t) + \mathbf{b}\}, \text{ where } \delta_t \in \mathbb{R}^d \end{aligned} \tag{3}$$

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

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



(a) Contrastive Learning with perturbation (b) Generation of Imposters (c) Generation of Distant-Targets

Figure 3: **Generation of imposters and distant-targets with perturbation.** (a) We add small perturbation δ_t to $\mathbf{h}_t$ for $\tilde{\mathbf{z}}_\mathbf{y}$ so that its conditional likelihood is minimized to generate an invalid sentence. (b) We add large perturbation ζ_t to $\mathbf{h}_t$ for $\hat{\mathbf{z}}_\mathbf{y}$ by maximizing the distance from $\mathbf{z}_\mathbf{x}$, the representation of source sentence but enforcing its likelihood high to preserve the original semantics.

We add small perturbation to the hidden representation of each token of target sentence $\mathbf{y}^{(i)}$ such that its conditional likelihood is minimized. Thus, the perturbed $\tilde{\mathbf{H}}^{(i)}$, which we call an *imposter* (inspired by Weinberger & Saul (2009)), is semantically very dissimilar to $\mathbf{y}^{(i)}$, but very close to the hidden representation $\mathbf{H}^{(i)}$ in the embedding space (Fig. 3-(a)). This will make it non-trivial for the sequence-to-sequence model to distinguish it from the representation of true target sequence $\mathbf{y}^{(i)}$. We use the perturbed representation as an additional negative sample for contrastive learning as follows:

$$\mathcal{L}_{cont-neg}(\theta) = \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_\mathbf{x}^{(i)}, \mathbf{z}_\mathbf{y}^{(i)})/\tau)}{\sum_{\mathbf{z}_\mathbf{y}^{(k)} \in S \cup \{\tilde{\mathbf{z}}_\mathbf{y}^{(i)}\}} \exp(\text{sim}(\mathbf{z}_\mathbf{x}^{(i)}, \mathbf{z}_\mathbf{y}^{(k)})/\tau)}, \text{ where } \tilde{\mathbf{z}}_\mathbf{y}^{(i)} = \xi(\tilde{\mathbf{H}}^{(i)}; \theta) \tag{5}$$

Alternatively, we can generate an imposter by perturbing the hidden representation of target sentence $\mathbf{y}$ so that its conditional likelihood is minimized but very close to the source sentence $\mathbf{x}$ in the embedding space. However, we empirically find that such a variation yields less performance gain.

3.4 Generation of Distant-Targets

Moreover, as shown in Fig. 3-(c), we construct an additional positive pair of source sequence $\mathbf{x}^{(i)}$ by adding large perturbation $\boldsymbol{\zeta}^{(i)} = [\zeta_1^{(i)} \cdots \zeta_T^{(i)}] \in \mathbb{R}^{d \times T}$ to $\mathbf{H}^{(i)}$ the hidden state of target sequence $\mathbf{y}^{(i)}$, such that cosine similarity from $\mathbf{z}_\mathbf{x}^{(i)}$ is minimized, but the conditional likelihood is enforced to remain high. However, the exact computation of $\boldsymbol{\zeta}^{(i)}$ with such constraints is intractable. We approximate it with the following two separate stages. First, we add perturbation to $\mathbf{H}^{(i)}$ such that it minimizes the contrastive learning objective $\mathcal{L}_{cont}(\theta)$ as shown in Eq. 6. Then we add another perturbation to minimize the KL divergence between perturbed conditional distribution $p_\theta(\hat{y}_t^{(i)} | \hat{\mathbf{y}}_{<t}^{(i)}, \mathbf{x}^{(i)})$ and the original conditional distribution $p_\theta(y_t^{(i)} | \mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)})$ as shown in Eq. 7, where $\overline{\mathbf{H}} = [\overline{\mathbf{h}}_1 \cdots \overline{\mathbf{h}}_T] \in \mathbb{R}^{d \times T}$, $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1 \cdots \hat{\mathbf{h}}_T] \in \mathbb{R}^{d \times T}$, and $\eta \in \mathbb{R}$. Note that θ^* denotes the copied of the model parameter θ and is considered to be constant to prevent it from being updated through back-propagation.

$$\overline{\mathbf{H}}^{(i)} = \mathbf{H}^{(i)} - \eta \frac{\mathbf{g}}{||\mathbf{g}||_2} \text{ where } \mathbf{g} = \nabla_{\mathbf{H}^{(i)}} \mathcal{L}_{cont}(\theta) \tag{6}$$

$$p_\theta(\overline{y}_t^{(i)} | \mathbf{x}^{(i)}) = \text{softmax}(\mathbf{W}\overline{\mathbf{h}}_t^{(i)} + \mathbf{b})$$

$$\mathcal{L}_{KL}(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} D_{KL}(p_{\theta^*}(y_t^{(i)} | \mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}) || p_\theta(\overline{y}_t^{(i)} | \overline{\mathbf{y}}_{<t}^{(i)}, \mathbf{x}^{(i)}) \tag{7}$$

$$\hat{\mathbf{H}}^{(i)} = \overline{\mathbf{H}}^{(i)} - \eta \frac{\mathbf{f}}{||\mathbf{f}||_2}, \text{ where } \mathbf{f} = \nabla_{\overline{\mathbf{H}}_1^{(i)}} \mathcal{L}_{KL}(\theta)$$

We consider the perturbed hidden state $\hat{\mathbf{H}}^{(i)}$ as an additional positive example for source sequence $\mathbf{x}^{(i)}$, which we refer to as a *distant-target*. We can use a distant-target to augment contrastive learning

and minimize $\mathcal{L}_{KL}(\theta)$ as follows:

$$\mathcal{L}_{cont-pos}(\theta) = \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \hat{\mathbf{z}}_{\mathbf{y}}^{(i)})/\tau)}{\sum_{\mathbf{z}_{\mathbf{y}}^{(k)} \in S \cup \{\bar{\mathbf{z}}_{\mathbf{y}}^{(i)}\}} \exp(\text{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(k)})/\tau)}, \text{ where } \hat{\mathbf{z}}_{\mathbf{y}}^{(i)} = \xi(\hat{\mathbf{H}}^{(i)}; \theta) \quad (8)$$

CLAPS objective Incorporating the loss on the imposter and the distant target introduced above, we estimate the parameters of the seq2seq model θ by maximizing the following objective, where α, β are hyperparameters which control the importance of contrastive learning and KL divergence:

$$\underset{\theta}{\text{maximize}}\, \mathcal{L}_{MLE}(\theta) + \alpha \mathcal{L}_{KL}(\theta) + \beta\{\mathcal{L}_{cont-neg}(\theta) + \mathcal{L}_{cont-pos}(\theta)\} \quad (9)$$

For all the experiments, we set α and β as 1, which we search through cross-validation. Note that after training is done, we remove the pooling layer ξ and generate text with the decoder g, given an input encoded with the encoder f.

4 EXPERIMENT

We validate our method on benchmark datasets on three conditional text generation tasks.

4.1 TASKS

Machine Translation (MT) For machine translation, we use WMT16 Romanian-English parallel corpus (WMT'16 RO-EN) to train the model. We tokenize the pairs of source and target sequences with the same tokenizer as Raffel et al. (2020). We finetune the pretrained T5-small model for 20 epochs with the batch size of 128 and Adafactor (Shazeer & Stern, 2018). For contrastive learning, we set the norm of perturbation, η and ϵ as 3.0.

Text Summarization (Sum.) For text summarization, we use XSum dataset (Narayan et al., 2018) of which summaries are highly abstractive, thus extractive summarization models under-perform abstractive models. We follow the most of the experimental settings for machine translation as described above, except that we set the norm of perturbation, η and ϵ as 1.0 and 1.0, respectively.

Question Generation (QG) For question generation, we aim 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) for 20 epochs with batch size 128 and 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 split the validation set into a validation set and a test set.

4.2 EXPERIMENTAL SETUPS

Implementation Details For the encoder f, and decoder g, we use T5-small model, which is based on transformer with the hidden dimension, $d = 512$. We set the temperature, τ as 0.1 for all the experiments. At test time, we use beam search of width 4 to generate the target sequences.

Common Baselines We compare our method against relevant baselines.

1. **T5-MLE**: A pretrained T5 model fine-tuned to maximize $\mathcal{L}_{MLE}(\theta)$.
2. **T5-SSMBA**: This is the T5 model trained to maximize $\mathcal{L}_{MLE}(\theta)$, with additional examples generated by the technique proposed in Ng et al. (2020). which are generated by corrupting the target sequences and reconstructs them using a masked language model, BERT.
3. **T5-WordDropout Contrastive**: This is a T5 model trained with the contrastive learning framework proposed in Yang et al. (2019), which heuristically generates negative examples by removing the most frequent word from the target sequence. We pretrain T5-small to maximize $\mathcal{L}_{MLE}(\theta)$ and further train the model to assign higher probability to the ground truth target sentence than a negative example with max-margin loss.
4. **T5-MLE-contrastive**: This is a naive constrastive learning framework with positive/negative pairs, which maximizes the contrastive learning objective from Eq. 2.

Method	Aug.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU	EM/F1
Question Generation - SQuAD							
Harvesting-QG	-	-	-	20.90	15.16	-	66.05/54.62
T5-MLE	-	41.26	30.30	23.38	18.54	21.00	67.64/55.91
T5-SSMBA	Pos.	41.67	30.59	23.53	18.57	21.07	68.47/56.37
T5-WordDropout Contrastive	Neg.	41.37	30.50	23.58	18.71	21.19	68.16/56.41
T5-MLE-contrastive	-	41.23	30.28	23.33	18.45	20.91	67.32/55.25
T5-CLAPS w/o positive	Neg.	41.65	30.69	23.71	18.81	21.25	68.26/56.41
T5-CLAPS	Pos.+Neg.	**42.33**	**31.29**	**24.22**	**19.19**	**21.55**	**69.01/57.06**
Machine Translation - WMT'16 RO-EN							
Transformer	-	50.36	37.18	28.42	22.21	26.17	
T5-MLE	-	57.76	44.45	35.12	28.21	32.43	
T5-SSMBA	Pos.	58.23	44.87	35.50	28.48	32.81	
T5-WordDropout Contrastive	Neg.	57.77	44.45	35.12	28.21	32.44	
T5-MLE-contrastive	-	57.64	44.12	34.74	27.79	32.03	
T5-CLAPS w/o positive	Neg.	57.90	44.60	35.27	28.34	32.55	
T5-CLAPS	Pos.+Neg.	**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. EM/F1 scores with BERT-base QA model for question generation.

5. **T5-CLAPS w/o positive**: Our proposed model which jointly maximizes the log likelihood and the contrastive learning objective with imposters but does not use any distant-targets.
6. **T5-CLAPS**: Our full model which jointly maximizes the log likelihood, contrastive learning objective, and KL-divergence as described in the Eq. 9.

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

Evaluation Metric Following the conventional evaluation metrics, we adopt n-gram BLEU and BLEU (Papineni et al., 2002) for machine translation and question generation. For text summarization, we use Rouge (Lin & Hovy, 2002) and Meteor (Banerjee & Lavie, 2005). As an additional performance measure for question generation, we evaluate a BERT QA model on the SQuAD test set, where the QA model is trained with the questions generated by each QG methods from the contexts and answers of HarvestingQA dataset (Du & Cardie, 2018), and report the F1 and Exact Match (EM).

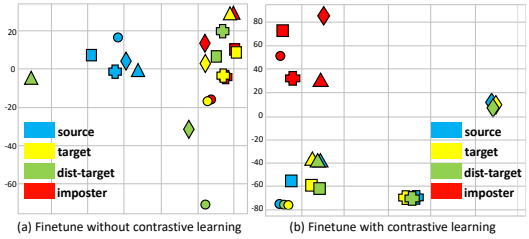
4.3 Experimental Results

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 CLAPS significantly outperforms the other baseline, with the performance gain of more than 1% on all tasks according to the BLEU scores. For question generation, our proposed method also improves F1/EM as well as BLEU scores. It shows that our proposed model is able to generate semantically valid questions that are beneficial for training the QA model. Note that naively constructing the negative examples for contrastive learning on the both tasks, by randomly shuffling the association of $(\mathbf{x}, \mathbf{y})$ from a given mini-batch, degrades the performance. Increasing the batch size to a large value, using GPUs with larger memory, may increase its performance as observed in SimCLR (Chen et al., 2020). However, such an approach will be highly sample-inefficient. On the contrary, our model outperforms all the other baseline models on Xsum dataset for text summarization, as shown in Table 2. For summarization, we observe that contrastive learning with imposters alone can improve the performance by a large margin.

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 f and decoder g. Then, we add perturbations to $\mathbf{H}$ to construct an imposter $\tilde{\mathbf{H}}$ and an additional positive

Table 2: Rouge F1 and Meteor on Xsum test set for text summarization.

Method	Aug.	Rouge-1	Rouge-2	Rouge-L	METEOR
Text Summarization - XSum					
PTGEN-COVG	-	28.10	8.02	21.72	12.46
CONVS2S	-	31.89	11.54	25.75	13.20
T5-MLE	-	36.10	14.72	29.16	15.78
T5-SSMBA	Pos.	36.58	14.81	29.68	15.38
T5-WordDropout Contrastive	Neg.	36.88	15.11	29.79	15.77
T5-MLE-contrastive	-	36.34	14.81	29.41	15.85
T5-CLAPS w/o positive	Neg.	37.72	15.49	**30.74**	16.06
T5-CLAPS	Pos.+Neg.	**37.89**	**15.78**	30.59	**16.38**



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

Figure 4: **Visualization.** (a) Embedding space without contrastive learning. (b) Embedding space with our proposed contrastive learning, CLAPS.

(MT) Lupta lui Hilary a fost mai atractivă.
=>**(GT)**: Hillary's **struggle** was more attractive
=>**(Dist.)**: Hilary's **fight** was more attractive
=>**(Imp.)**: **Thearies'** battle fight has attractive appealing

(QG) … Von Miller … recording **five** solo tackles, …
=>**(GT)**: How many solo tackles did Von Miller **make** at Super Bowl 50?
=>**(Dist.)**: How many solo tackles did Von Miller **record** at Super Bowl 50?
=>**(Imp.)**: What much tackle **did was** Miller record at Super Bowl 50?

(Sum.) Pieces from the board game … have been found in … China. …
=>**(GT)**: An ancient board game has been **found** in a Chinese Tomb.
=>**(Dist.)**: An ancient board game has been **discovered** in a Chinese Tomb.
=>**(Imp.)**: America's gained vast Africa **most well geographical** countries, 22

Table 3: Greedy decoding from hidden representation of imposters and distant-targets. The answer span is highlighted for QG.

example $\hat{\mathbf{H}}$ as shown in Eq. 3 and 6, 7. We apply average pooling to $\mathbf{M}, \mathbf{H}, \tilde{\mathbf{H}}$, and $\hat{\mathbf{H}}$ to obtain fixed size representations and project them onto two dimensional space with t-SNE (Maaten & Hinton, 2008). As shown in Figure 4-(b), the model pushes away the imposter from the embedding of target sequence and pulls the embedding of the distant-targets to the embedding of the source sequence. For the model without contrastive learning, however, the embeddings of both target sequences and distant targets are far away from those of source sequences and the imposters are very close to them as shown in Fig. 4-(a).

Qualitative Examples For qualitative analysis, we examine the texts that are represented by the distant-target and imposter from our method, CLAPS. To decode them into output sequences, we apply affine transformation and softmax to $\tilde{\mathbf{H}}$ and $\hat{\mathbf{H}}$ and select the most likely token at each time step. As shown in Table 3, the distant-target example (**Dist.**), preserves the semantic of the original target sequence (**GT**) with a single word replaced by a synonym (colored in green). However, the imposters (**Imp.**) have completely different semantics, and often are gramatically incorrect (colored in red). This shows that the model are exposed to those various valid or incorrect sentences with our proposed contrastive learning framework with adversarial perturbations.

5 CONCLUSION

To tackle the exposure bias problem in sequence-to-sequence learning, we proposed a contrastive learning framework which maximizes the similarity between ground truth input and output sequence, and minimize the similarity between the input and an incorrect output sequence. Moreover, since conventional approach to sample random non-target examples from the batch as negative examples for contrastive learning results in trivial pairs that are well-discriminated from the beginning, we propose a new principled approach to automatically construct "hard" negative and positive examples, where the former is semantically dissimilar but close to the input embedding, and the latter is far from the input embedding but semantically similar. This adversarial learning enables the model to learn both the correct and incorrect variations of the input, and generalize better to unseen inputs. We empirically showed that our method improves the performance of seq2seq model on machine translation, question generation, and text summarization tasks. While we specifically tackled the exposure bias problem with seq2seq models for conditional text generation, our positive/negative pair generation is general and may be applicable to seq2seq constrastive learning for tasks from other domains, such as automatic speach recognition, text-to-speach generation, or video captioning.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *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. *International Conference on Learning Representations, ICLR*, 2017.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *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. *Empirical Methods in Natural Language Processing, EMNLP*, 2014.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. *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. *International Conference on Learning Representations, ICLR*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NACCL-HLT*, 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. *Advances in Neural Information Processing Systems*, 2019.

Xinya Du and Claire Cardie. Harvesting paragraph-level question-answer pairs from wikipedia. *Annual Meeting of the Association for Computational Linguistics, ACL*, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *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.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. Large margin neural language model. *Empirical Methods in Natural Language Processing, EMNLP*, 2018.

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

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. *Annual Meeting of the Association for Computational Linguistics, ACL*, 2020.

Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. *ACL Workshop on Automatic Summarization*, 2002.

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

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

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

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

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

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. *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. *Advances in neural information processing systems*, 2013.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. *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. *Empirical Methods in Natural Language Processing, EMNLP*, 2018.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *Empirical Methods in Natural Language Processing, EMNLP*, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *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. *International Conference on Learning Representations, ICLR*, 2018.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NACCL-HLT*, 2018.

Matt Post. A call for clarity in reporting bleu scores. *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. *Journal of Machine Learning Research*, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *Empirical Methods in Natural Language Processing, EMNLP*, 2016.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *nternational Conference on Learning Representations, ICLR*, 2016.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *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. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *Annual Meeting of the Association for Computational Linguistics, ACL*, 2017.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, 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. *International Conference on Machine Learning, ICML*, 2019.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *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. *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. *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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. Reducing word omission errors in neural machine translation: A contrastive learning approach. *Annual Meeting of the Association for Computational Linguistics, ACL*, 2019.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: sequence generative adversarial nets with policy gradient. *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. *International Conference on Machine Learning, ICML*, 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. *Advances in Neural Information Processing Systems*, 2018.

Table 4: The statistics and the data source of WMT'16 RO-EN, Xsum, and SQuAD.

Datasets	Train (#)	Valid (#)	Test (#)	Source
WMT'16 RO-EN	610,320	1,999	1,999	Romanian-English Parallel corpus.
Xsum	204,045	11,332	11,334	One-sentence summary of BBC news articles.
SQuAD	86,588	5,192	5,378	Crowd-sourced questions from Wikipedia paragraph

A EXPERIMENTAL DETAILS

Dataset For machine translation, text summarization, and question generation, we use WMT'16 RO-EN, Xsum, SQuAD dataset, for each task. The number of train/validation/test set and its source is shown in Table 4. Note that the number of validation and test set for SQuAD is different from the original dataset. Since the original test set is only accessible via the leader board of SQuAD[1], we split the original validation set into our new validation and test set, following the conventions of question generation communities.

Preprocessing For machine translation, we download the raw text[2], not the tokenized text, and use the same T5-tokenizer as Raffel et al. (2020) to tokenize both Romanian and English sentences. We limit the input and output length to 128 tokens. For text summarization, we also use the T5-tokenizer as before, and limit the input length to 512 tokens and output length to 128 tokens. For question generation, we set the maximum length of question as 64 tokens and input which is concatenation of answer and context as 384 tokens.

Implementation We finetune the pretrained T5-small model provided from the transformers library (Wolf et al., 2019)[3] with Adafactor optimizer. We set the batch size 128 and follow the default setting of Adafactor optimizer to finetune the T5-small models. However, the number of negative examples from the batch is 16 or 32 (total batch size divided by the number of GPUs), because we split the batch into smaller batches and distribute them to each GPU machines. We use 8 GPUs for text summarization, and 4 GPUs for machine translation and question generation. The dimension of hidden state of T5 model, d is 512, so we set the hidden size of $\mathbf{z}$ as the same.

Evaluation We use beam search with beam width 4 to generate the target sentences from the source sentences of the test set. Some of the examples are shown in Table 5,6,7. After the generation, we convert the tokens into the raw texts and compare them to the raw text of ground truth target sentences with the automatic evaluation metrics. For n-gram BLEU and Meteor, we use the implementation by Sharma et al. (2017)[4]. For BLEU score, we adopt the implementation by Post (2018)[5].

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

[2] https://s3.amazonaws.com/datasets.huggingface.co/translation/wmt_en_ro.tar.gz

[3] https://github.com/huggingface/transformers

[4] https://github.com/Maluuba/nlg-eval

[5] https://github.com/mjpost/sacrebleu

Table 5: Generated summaries by CLAPS from Xsum dataset.

Article: The US military says a strike targeting Taliban in the northern city of Kunduz may have caused "collateral damage". Offering his "deepest condolences", Mr Obama said he expected a "full accounting of the facts" and would then make a definitive judgement. . . . **GT:** President Barack Obama says the US has launched a "full investigation" into air strikes that killed 19 people at an MSF-run Afghan hospital on Saturday. **CLAPS:** US President Barack Obama has called for an inquiry into air strikes in Afghanistan that killed dozens of medical workers.
Article: Forecasts were for quarterly growth of between 0.5% and 0.7%. Official statistics also showed that household consumption expenditure boosted the quarterly growth numbers. But economist Shane Oliver told the BBC the numbers were "well below potential". On an annual basis the economy expanded 2.3%, beating expectations for 2.1%. Economic growth in the March quarter of 2014 was 2.9%. "The March quarter GDP [gross domestic product] growth was far better than feared just a few days ago," said Mr Oliver, who is chief economist with AMP Capital in Sydney. "However, Australia is still not out of the woods, as annual growth at 2.3% is well below potential, and a full 0.8% percentage points of the 0.9% growth came from higher inventories and trade." He said domestic demand remained "very weak with consumer spending and home construction only just offsetting the ongoing slump in mining investment". ... **GT**: Australia's economy grew at a better-than-expected 0.9% in the first quarter of 2015, compared to the previous quarter, boosted by mining together with financial and insurance services. **CLAPS**: Australia's economy grew faster than expected in the first three months of the year, according to official figures.
Article: After the problems last week, many doubt the system will cope. Transport for London (TfL) remains confident, although it admits there will be breakdowns. The trick will be in getting the system back up and running quickly. So here's some friendly advice for tourists and Olympic visitors to try and make the transport experience as easy as possible. If anyone thinks of any more please post below. **GT:** The busiest summer ever looms for London's transport system. **CLAPS:** London's transport system has been a pretty busy week.
Article: The outgoing vice-president spoke during a state dinner and took the opportunity to praise America's northern neighbour. "The world is going to spend a lot of time looking to you, Mr Prime Minister", he told the Canadian leader. Mr Biden has been highly critical of US President-elect Donald Trump. "Vive le Canada because we need you very, very badly," he told the dinner guests. He went on to describe the self-doubt that liberal leaders across the world are currently experiencing after several political defeats. But he praised "genuine leaders" including German Chancellor Angela Merkel, saying such statesmen and women are in short supply. Mr Trudeau reportedly became emotional during Mr Biden's remarks when the American spoke of his late father, former Prime Minister Pierre Trudeau. "You're a successful father when your children turn out better than you," Mr Biden said. ... **GT:** US Vice-President Joe Biden told an audience in Ottawa that the world needs "genuine leaders" such as Canadian Prime Minister Justin Trudeau. **CLAPS:** Vice-President Joe Biden has praised Canadian Prime Minister Vive le Canada for his visit to the country.

Table 6: Generated Questions by CLAPS from SQuAD. Answer spans are highlighted.

Context: ... The Broncos finished the regular season with a 12-4 record, and denied the New England Patriots a chance to defend their title from Super Bowl XLIX by defeating them 20-18 in the AFC Championship Game. They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made **eight** appearances in the Super Bowl.

GT: How many appearances have the Denver Broncos made in the Super Bowl?

CLAPS: How many Super Bowl appearances have the Broncos made?

Context: In late November 2015, reports surfaced stating that "multiple acts" would perform during the halftime show. On December 3, the league confirmed that the show would be headlined by the **British** rock group Coldplay. On January 7, 2016, Pepsi confirmed to the Associated Press that Beyoncé, who headlined the Super Bowl XLVII halftime show and collaborated with Coldplay on the single "Hymn for the Weekend", would be making an appearance. Bruno Mars, who headlined the Super Bowl XLVIII halftime show, and Mark Ronson also performed.

GT: What nationality is the band Coldplay?

CLAPS: What nationality was Coldplay?

Context: There are 13 natural reserves in Warsaw - among others, Bielany Forest, Kabaty Woods, Czerniaków Lake. About 15 kilometres (9 miles) from Warsaw, the Vistula river's environment changes strikingly and features a perfectly preserved ecosystem, with a habitat of animals that includes the **otter, beaver and hundreds of bird species**. There are also several lakes in Warsaw - mainly the oxbow lakes, like Czerniaków Lake, the lakes in the Łazienki or Wilanów Parks, Kamionek Lake. There are lot of small lakes in the parks, but only a few are permanent - the majority are emptied before winter to clean them of plants and sediments.

GT: What animals does the Vistula river's ecosystem include?

CLAPS: What animals are included in the Vistula river's habitat?

Context: "The FSO Car Factory was established in 1951. A number of vehicles have been assembled there over the decades, including the Warszawa, Syrena, Fiat 125p (under license from Fiat, later renamed FSO 125p when the license expired) and the Polonez. The last two models listed were also sent abroad and assembled in a number of other countries, including Egypt and Colombia. In 1995 the factory was purchased by the South Korean car manufacturer Daewoo, which assembled the Tico, Espero,Nubia, Tacuma, Leganza, Lanos and Matiz there for the European market. In 2005 the factory was sold to **AvtoZAZ**, a Ukrainian car manufacturer which assembled there the Chevrolet Aveo. The license for the production of the Aveo expired in February 2011 and has since not been renewed. Currently the company is defunct."

GT: Who bought the factory in 2005?

CLAPS: To whom was the factory sold in 2005?

Table 7: Translation of Romanian by CLAPS from WMT'16 RO-EN.

RO: De partea cealaltă, 47% dintre alegătorii republicani afirmă că ar fi „nemulțumiți" sau „supărați" dacă favoritul Jeb Bush câștigă cursa pentru nominalizare.
GT : On the flip side, 47 percent of Republican voters say they would be "dissatisfied" or "upset" if establishment favorite Jeb Bush becomes the nominee.
CLAPS: On the other hand, 47% of Republican voters say they would be "unsatisfied" or "uneasy" if the winner Jeb Bush wins the nomination race.
RO: Datoria va deveni o problemă importantă.
GT: Debt will become a big issue.
CLAPS: Debt will become an important issue.
RO: Se pare că Moreno a avut noroc și a scăpat de cartonașul roșu și de concesia unei lovituri de pedeapsă.
GT: Moreno appeared fortunate to escape a red card and the concession of a spot-kick.
CLAPS: It seems that Moreno has been lucky and escaped the red card and the concession of a penalty kick.
RO: Astfel de decizii nu se iau niciodată ușor.
GT: Such decisions are never taken lightly.
CLAPS: Such decisions are never made easily.