CONTRASTIVE LEARNING WITH ADVERSARIAL PERTURBATIONS FOR CONDITIONAL TEXT GENERATION

Seanie Lee , Dong Bok Lee, Sung Ju Hwang

ICLR 2021

Motivations

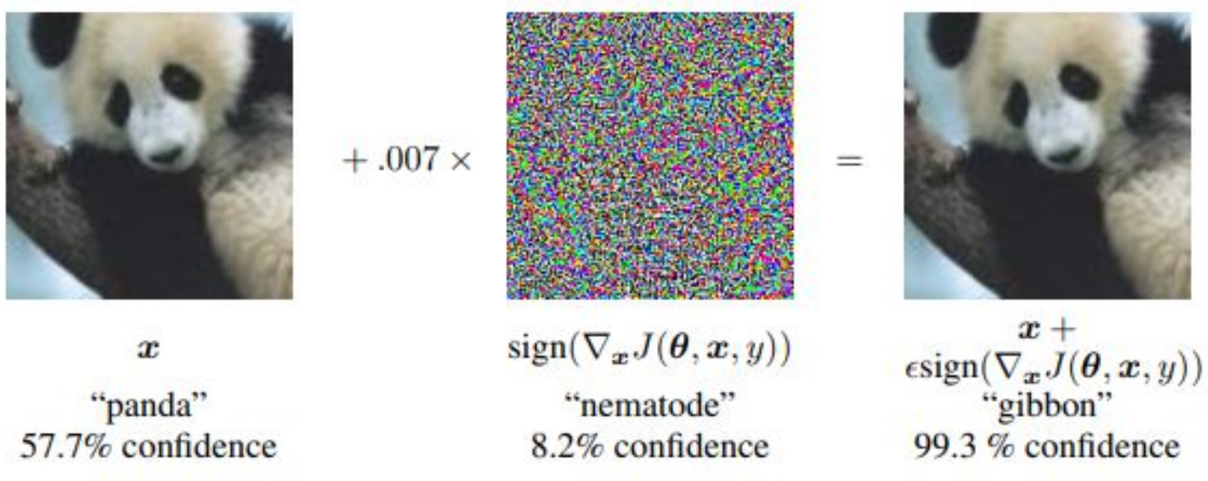
- Seq2Seq models usually are trained with teacher-forcing method.
- They are not exposed to incorrect generated tokens during training, which hurts its generalization to unseen inputs.
- This work proposes to mitigate the conditional text generation problem by contrasting positive pairs with negative pairs
    - the model is exposed to various valid or incorrect perturbations of the inputs, for improved generalization.

Challenges & Proposed Solution

- Contrastive learning framework using random non-target sequences as negative examples is suboptimal, since they are easily distinguishable from the correct output.
- Generating positive examples requires domain-specific augmentation heuristics which may not generalize over diverse domains.

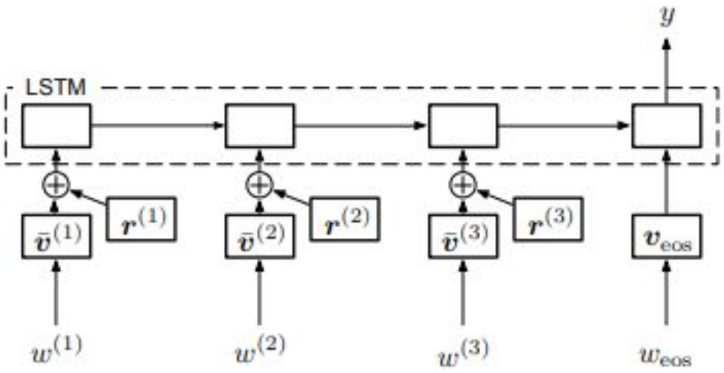
- To tackle this problem, the authors propose a principled method to generate positive and negative samples through adversarial perturbation.

Background of Adv. Attack



Goodfellow et al., 2015

Background of Adv. Attack



$$-\log p(y \mid \boldsymbol{x} + \boldsymbol{r}_{\text{adv}}; \boldsymbol{\theta}) \text{ where } \boldsymbol{r}_{\text{adv}} = \underset{\boldsymbol{r}, \|\boldsymbol{r}\| \leq \epsilon}{\arg\min} \log p(y \mid \boldsymbol{x} + \boldsymbol{r}; \hat{\boldsymbol{\theta}})$$

$$\boldsymbol{r}_{\text{adv}} = -\epsilon \boldsymbol{g} / \|\boldsymbol{g}\|_2 \text{ where } \boldsymbol{g} = \nabla_{\boldsymbol{x}} \log p(y \mid \boldsymbol{x}; \hat{\boldsymbol{\theta}}).$$

Miyato, Dai, Goodfellow, 2017

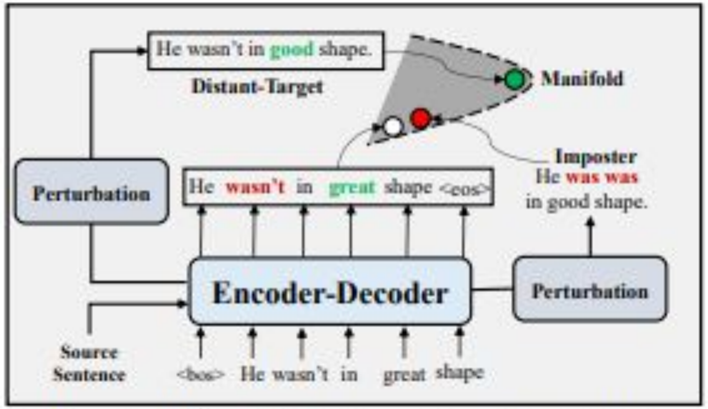
CL for Seq2Seq

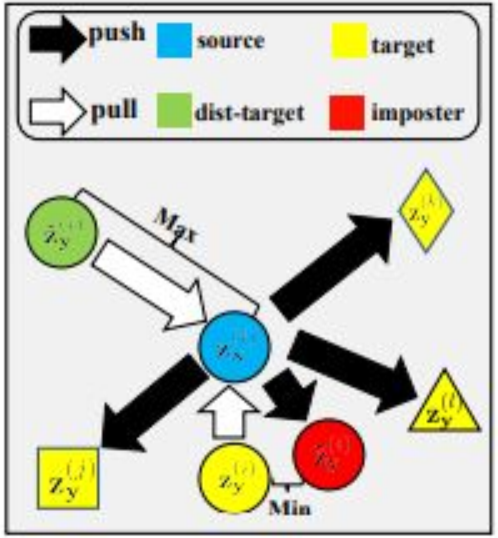
$$\mathcal{L}_{cont}(\theta) = \sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(i)})/\tau)}{\sum_{\mathbf{z}_{\mathbf{y}}^{(j)} \in S} \exp(\mathrm{sim}(\mathbf{z}_{\mathbf{x}}^{(i)}, \mathbf{z}_{\mathbf{y}}^{(j)})/\tau)}$$

$$\mathbf{z}_{\mathbf{x}}^{(i)} = \xi(\mathbf{M}^{(i)}; \theta), \ \mathbf{z}_{\mathbf{y}}^{(i)} = \xi(\mathbf{H}^{(i)}; \theta)$$

$$\xi([\mathrm{v}_1 \cdots \mathrm{v}_T]; \theta) := \mathrm{AvgPool}([\mathbf{u}_1 \cdots \mathbf{u}_T]), \text{ where } \mathrm{u}_t = \mathrm{ReLU}(\mathbf{W}^{(1)}\mathrm{v}_t + \mathbf{b}^{(1)})$$

Proposed Method





(a) Contrastive Learning with perturbation

Imposter Generation

$$\tilde{\mathbf{H}}^{(i)} = \mathbf{H}^{(i)} + \boldsymbol{\delta}^{(i)} \text{ where } \boldsymbol{\delta}^{(i)} = \arg\min_{\boldsymbol{\delta}, ||\boldsymbol{\delta}||_2 \le \epsilon} \log p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \mathbf{H}^{(i)} + \boldsymbol{\delta})$$

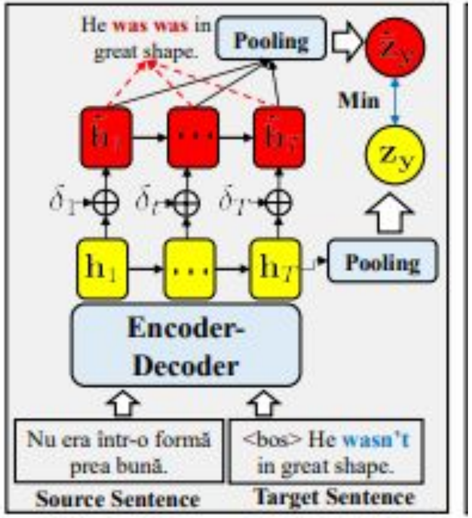
$$p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \mathbf{H}^{(i)} + \boldsymbol{\delta}) = \prod_{t=1}^{T} p_\theta(y_t^{(i)}|\mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}; \mathbf{h}_t^{(i)} + \delta_t) \tag{3}$$

$$p_\theta(y_t^{(i)}|\mathbf{y}_{<t}^{(i)}, \mathbf{x}^{(i)}; \mathbf{h}_t^{(i)} + \delta_t) = \text{softmax}\{\mathbf{W}(\mathbf{h}_t^{(i)} + \delta_t) + \mathbf{b}\}, \text{ where } \delta_t \in \mathbb{R}^d$$

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:

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

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



(b) Generation of Imposters

Distant Target Generation

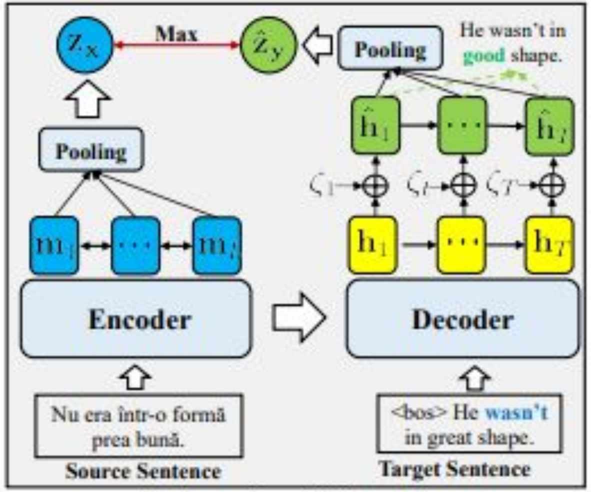
$$\overline{\mathbf{H}}^{(i)} = \mathbf{H}^{(i)} - \eta \frac{\mathbf{g}}{||\mathbf{g}||_2} \text{ where } \mathbf{g} = \nabla_{\mathbf{H}^{(i)}} \mathcal{L}_{cont}(\theta)$$

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

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

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

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



(c) Generation of Distant-Targets

Objective Function

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:

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

Experimental Result

Method	**Aug.**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**BLEU**	**F1/EM**
			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
α-T5-MLE ($\alpha = 0.7$)	-	40.82	29.79	22.84	17.99	20.50	68.04/56.30
α-T5-MLE ($\alpha = 2.0$)	-	37.35	27.20	20.79	16.36	18.41	65.74/54.76
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
R3F	-	41.00	30.15	23.26	18.44	20.97	65.84/54.10
T5-MLE-contrastive	-	41.23	30.28	23.33	18.45	20.91	67.32/55.25
T5-CLAPS w/o negative	Pos.	41.87	30.93	23.90	18.92	21.38	-
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**
ERNIE-GEN (Xiao et al., 2020)	-	-	-	-	**26.95**	-	-
Info-HCVAE (Lee et al., 2020)	-	-	-	-	-	-	**81.51/71.18**

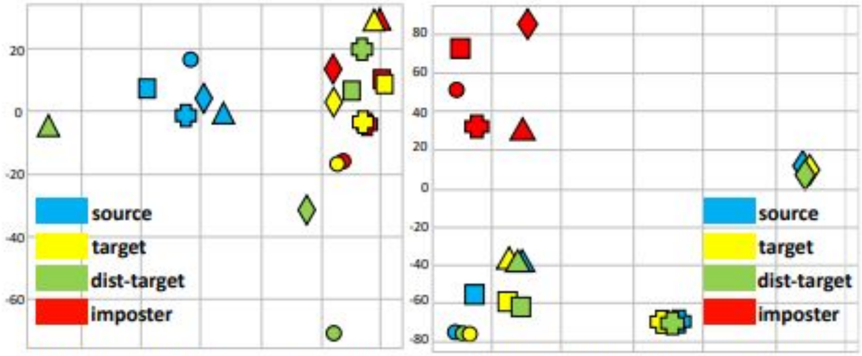
Experimental Result

Machine Translation - WMT'16 RO-EN							
Transformer	-	50.36	37.18	28.42	22.21	26.17	
Scratch-T5-MLE	-	51.62	37.22	27.26	21.13	25.34	
Scratch-CLAPS	Pos.+Neg.	53.42	39.57	30.24	23.59	27.61	
T5-MLE	-	57.76	44.45	35.12	28.21	32.43	
α-T5-MLE ($\alpha = 0.7$)	-	57.63	44.23	33.84	27.90	32.14	
α-T5-MLE ($\alpha = 2.0$)	-	56.03	42.59	33.29	26.45	30.72	
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	
R3F	-	58.07	44.86	35.57	28.66	32.99	
T5-MLE-contrastive	-	57.64	44.12	34.74	27.79	32.03	
T5-CLAPS w/o negative	Pos.	58.81	45.52	36.20	29.23	33.50	67.58/55.91
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**	
Conneau & Lample (2019)	-	-	-	-	-	**38.5**	

Experimental Result

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
Scratch-T5-MLE	-	31.44	11.07	25.18	13.01
Stcratch-CLAPS	Pos.+Neg.	33.52	12.59	26.91	14.18
T5-MLE	-	36.10	14.72	29.16	15.78
α-T5-MLE ($\alpha = 0.7$)	-	36.68	15.10	29.72	15.78
α-T5-MLE ($\alpha = 2.0$)	-	34.18	13.53	27.35	14.51
T5-SSMBA	Pos.	36.58	14.81	29.68	15.38
T5-WordDropout Contrastive	Neg.	36.88	15.11	29.79	15.77
R3F	-	36.96	15.12	29.76	15.68
T5-MLE-contrastive	-	36.34	14.81	29.41	15.85
T5-CLAPS w/o negative	Pos.	37.49	15.31	30.42	16.36
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**
PEGASUS (Zhang et al., 2020)	-	**47.21**	**24.56**	**39.25**	-

Qualitative Analysis



(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.

- affine transformation and softmax are applied to select the most likely token at each time step.

Thank You