Paper Reading

GANs for Discrete Text Generation

Junfu

Oct. 20th, 2018

Show, Tell and Discriminate

- Problems in Image Captioning
  - Imitate the language structure patterns (phrases, sentences)
  - Templated and Generic (Different image -> Same Captions)
  - Stereotype of sentences and phrases (50% from trainingset)

Xihui Liu, et al. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. ECCV 2018, CUHK.

Show, Tell and Discriminate

- Motivation
  - Both discriminativeness and fidelity should be improved
  - Discriminativeness: distinguish correspond. image and others
  - Dual task: Image captioning ⇔ Text-to-Image

- Model Architecture
  - Captioning Module
  - Self-retrieval Module
    - ✓ Act as a metric and an evaluator of caption discriminativeness to assure the quality of generated captions
    - ✓ Use unlabeled data to boost captioning performance

Show, Tell and Discriminate

Framework

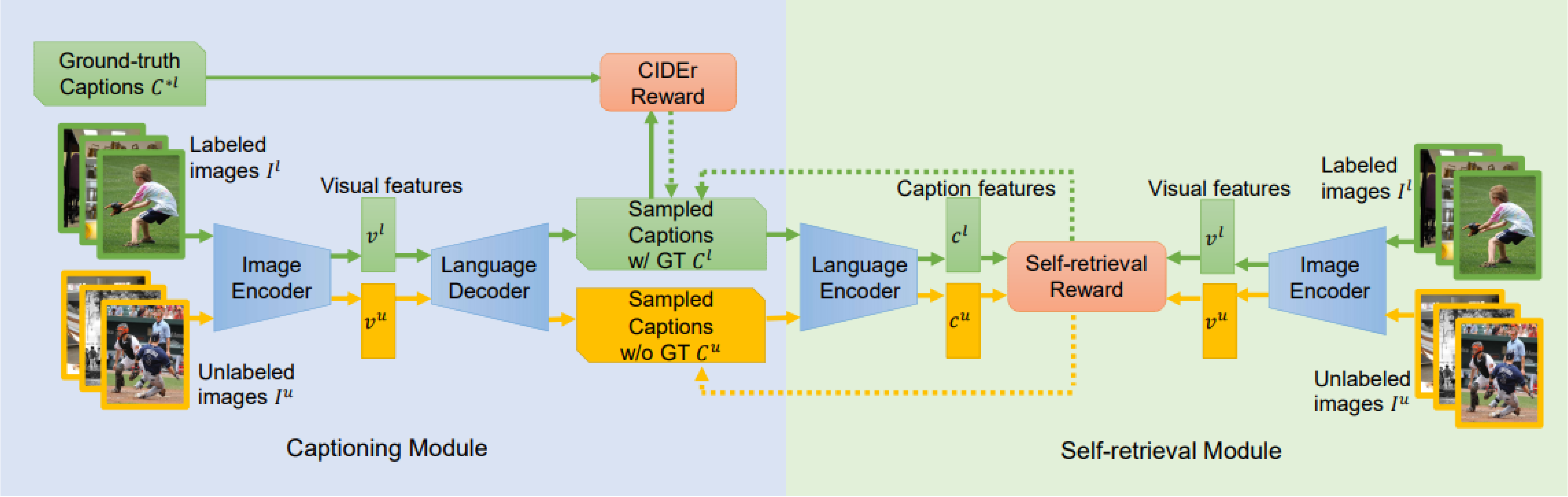


Image ⇒ Caption

I

$C = \{w_1, w_2, \dots, w_T\}$

$C^* = \{w_1^*, w_2^*, \dots, w_{T'}^*\}$

Encoder: CNN

$v = E_i(I)$

Decoder: LSTM

$C = D_c(v)$

Pre-train: $L_{CE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t^* | v, w_t^*, \dots, w_{t-1}^*))$

Adv-train: $r(C_i^s) = r_{cider}(C_i^s) + \alpha \cdot r_{ret}(C_i^s, \{I_1, \dots, I_n\})$

Image Encoder (CNN)

$v = E_i(I)$

Caption Encoder (GRU)

$c = E_c(C)$

Similarity between c_i and v_j: $s(c_i, v_j)$

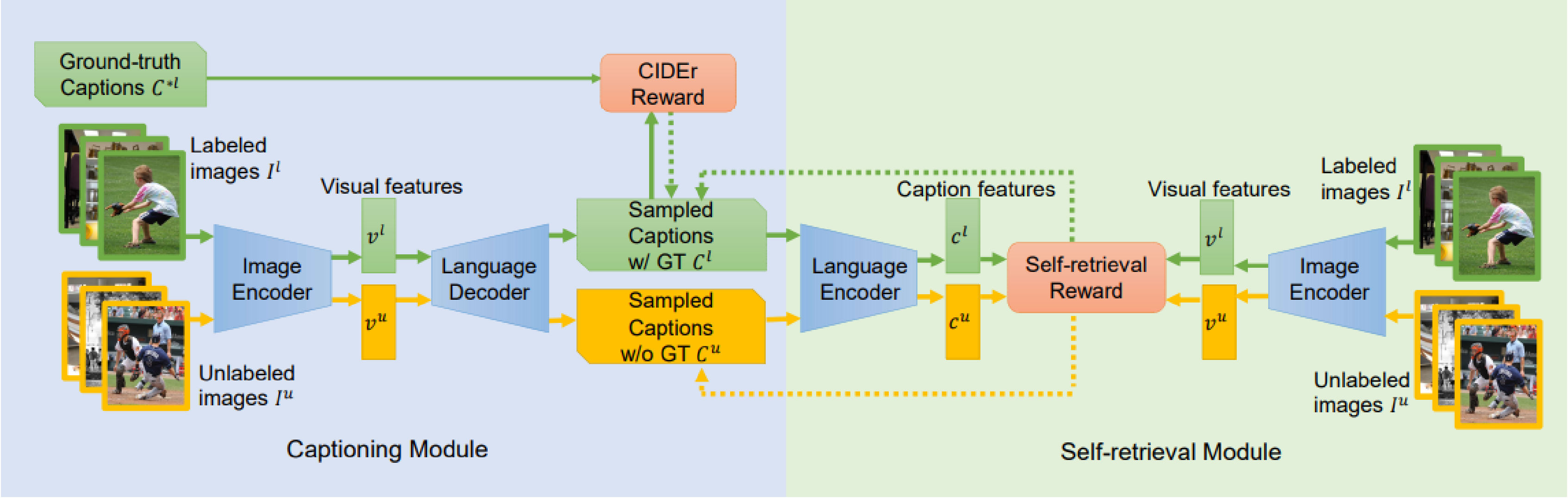
Train with ranking loss:

$$L_{ret}(C_i, \{I_1, I_2, \dots, I_n\}) = \max_{j \neq i}[m - s(c_i, v_i) + s(c_i, v_j)]_+$$

where $[x]_+ = \max(x, 0)$

Show, Tell and Discriminate

- Improving Captioning with Partially Labeled Image



Labeled Image: $\{I_1^l, I_2^l, \dots, I_{n_l}^l\}$ Generated Caption: $\{C_1^l, C_2^l, \dots, C_{n_l}^l\}$ Unlabeled Image: $\{I_1^u, I_2^u, \dots, I_{n_u}^u\}$

Labeled Data

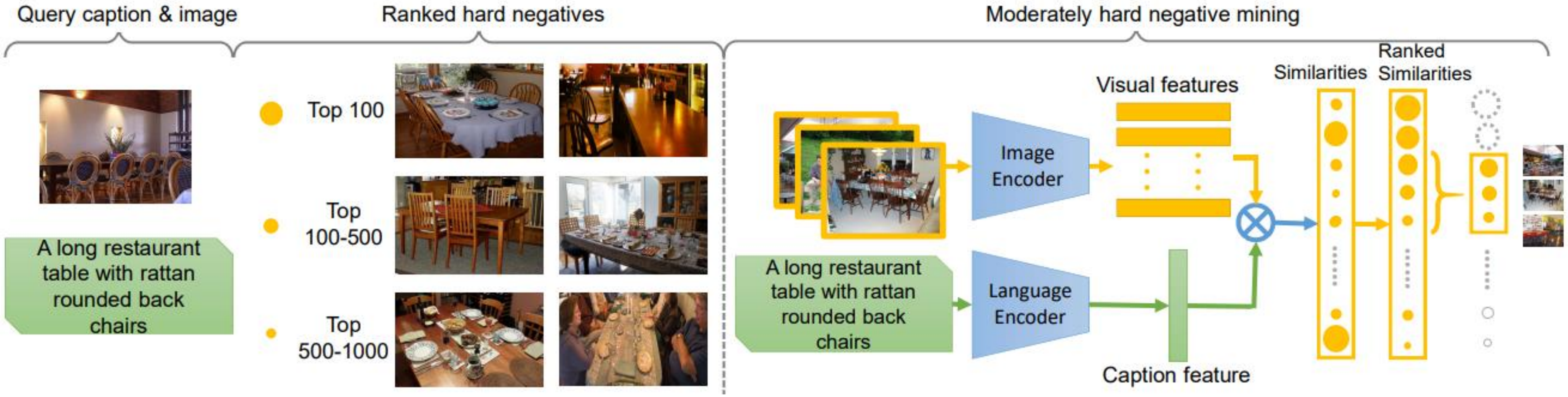
$$r(C_i^l) = r_{cider}(C_i^l) + \alpha \cdot r_{ret}(C_i^l, \{I_1^l, I_2^l, \dots, I_{n_l}^l\} \cup \{I_1^u, I_2^u, \dots, I_{n_u}^u\})$$

Unlabeled Data

$$r(C_i^u) = \alpha \cdot r_{ret}(C_i^u, \{I_1^l, I_2^l, \dots, I_{n_l}^l\} \cup \{I_1^u, I_2^u, \dots, I_{n_u}^u\})$$

Show, Tell and Discriminate

- Moderately Hard Negative Mining in Unlabeled Images



Groundtruth Caption:
$C^* = \{w_1^*, w_2^*, \ldots, w_{T'}^*\}$

Feature:
$\{v_1^u, v_2^u, \ldots, v_{n_l}^u\}$

Unlabeled Image:
$\{I_1^u, I_2^u, \ldots, I_{n_u}^u\}$

Similarity:
$\{s(c^*, v_1^u), s(c^*, v_2^u), \ldots, s(c^*, v_{n_u}^u\}$

Rank and sample:
$[h_{min}, h_{max}]$

Show, Tell and Discriminate

- Training Strategy
  - Train text-to-image self-retrieval module
    - ✓ Images and corresponding captions in labeled dataset
  - Pre-train captioning module
    - ✓ Images and corresponding captions in labeled dataset
    - ✓ Share image encoder with self-retrieval module
    - ✓ MLE with cross-entropy loss
  - Continue training by REINFORCE
    - ✓ Reward for labeled data: CIDEr and self-retrieval reward
    - ✓ Reward for unlabeled data: self-retrieval reward
    - ✓ CIDEr: guarantee the similarity between caption and groundtruth
    - ✓ Self-retrieval reward: encourage caption to be discriminative

Show, Tell and Discriminate

- Implementation Details
  - Self-retrieval module:
    - Word embedding: 300-D vector
    - Image encoder: ResNet-101
    - Language decoder: single GRU with 1024 hidden units
  - Captioning module:
    - Share image encoder with self-retrieval module
    - Language decoder: attention LSTM
    - Visual feature: 2048x7x7 before pooling
    - $\alpha = 1, \#labeled\ data : \#unlabeled\ data = 1:1$
  - Inference:
    - Beam search size: 5
  - Unlabeled data: COCO unlabeled images

Show, Tell and Discriminate

Quantitative results

Table 1. Single-model performance by our proposed method and state-of-the-art methods on COCO standard Karpathy test split.

Methods	CIDEr	SPICE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Hard-attention [47]	-	-	71.8	50.4	35.7	25.0	23.0	-
Soft-attention [47]	-	-	70.7	49.2	34.4	24.3	23.9	-
VAE [32]	90.0	-	72.0	52.0	37.0	28.0	24.0	-
ATT-FCN [50]	-	-	70.9	53.7	40.2	30.4	24.3	-
Att-CNN+RNN [46]	94.0	-	74.0	56.0	42.0	31.0	26.0	-
SCN-LSTM [14]	101.2	-	72.8	56.6	43.3	33.0	25.7	-
Adaptive [26]	108.5	-	74.2	58.0	43.9	33.2	26.6	-
SCA-CNN [5]	95.2	-	71.9	54.8	41.1	31.1	25.0	53.1
SCST-Att2all [35]	114.0	-	-	-	-	34.2	26.7	55.7
LSTM-A [49]	100.2	18.6	73.4	56.7	43.0	32.6	25.4	54.0
DRL [34]	93.7	-	71.3	53.9	40.3	30.4	25.1	52.5
Skeleton Key [43]	106.9	-	74.2	57.7	44.0	33.6	26.8	55.2
CNNL+RHN [16]	98.9	-	72.3	55.3	41.3	30.6	25.2	-
TD-M-ATT [4]	111.6	-	76.5	60.3	45.6	34.0	26.3	55.5
ATTN+C+D(1) [27]	114.25	**21.05**	-	-	-	**36.14**	27.38	**57.29**
Ours-baseline	112.7	20.0	79.7	62.2	47.1	35.0	26.7	56.4
Ours-SR-FL	114.6	20.5	79.8	62.3	47.1	34.9	27.1	56.6
Ours-SR-PL	**117.1**	21.0	**80.1**	**63.1**	**48.0**	35.8	**27.4**	57.0

Baseline: captioning module only trained only with CIDEr (w/o self-retrieval module)

SR-FL: proposed method training with fully-labeled data

SR-PL: proposed method training with additional unlabeled data

Show, Tell and Discriminate

Quantitative results

Table 2. Single-model performance by our proposed method and state-of-the-art methods on Flickr30k.

Methods	CIDEr	SPICE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Hard-attention [47]	-	-	66.9	43.9	29.6	19.9	18.5	-
Soft-attention [47]	-	-	66.7	43.4	28.8	19.1	18.5	-
VAE [32]	-	-	72.0	53.0	38.0	25.0	-	-
ATT-FCN [50]	-	-	64.7	46.0	32.4	23.0	18.9	-
Att-CNN+RNN [46]	-	-	73.0	55.0	40.0	28.0	-	-
SCN-LSTM [14]	-	-	73.5	53.0	37.7	25.7	21.0	-
Adaptive [26]	53.1		67.7	49.4	35.4	25.1	20.4	-
SCA-CNN [5]	-	-	66.2	46.8	32.5	22.3	19.5	-
CNNL+RHN [16]	61.8	15.0	**73.8**	**56.3**	**41.9**	**30.7**	21.6	-
Ours-baseline	57.1	14.2	72.8	53.4	38.0	27.1	20.7	48.5
Ours-SR-FL	61.7	15.3	72.0	53.4	38.5	27.8	21.5	49.4
Ours-SR-PL	**65.0**	**15.8**	72.9	54.5	40.1	29.3	**21.8**	**49.9**

Baseline: captioning module only trained only with CIDEr (w/o self-retrieval module)

SR-FL: proposed method training with fully-labeled data

SR-PL: proposed method training with additional unlabeled data

Show, Tell and Discriminate

Quantitative results

Table 3. Ablation study results on COCO.

Experiment Settings		CIDEr	SPICE	BLEU-3	BLEU-4	METEOR	ROUGE-L
Baseline		112.7	20.0	47.1	35.0	26.7	56.4
Retrieval Loss	VSE++	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**
	VSE0	116.9	20.9	47.7	35.7	**27.4**	56.8
	softmax	114.5	20.5	46.8	34.6	27.1	56.5
Weight of Self-retrieval Reward α	0	112.7	20.0	47.1	35.0	26.7	56.4
	1	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**
	4	113.7	20.5	46.5	34.3	27.0	56.5
Ratio between labeled and unlabeled	1:2	115.4	20.5	46.8	34.7	27.2	56.6
	1:1	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**
	2:1	115.0	20.5	46.8	34.7	27.2	56.7
Hard Negative Index Range	no hard mining	114.6	20.7	46.7	34.6	27.3	56.7
	top 100	114.1	20.3	46.6	34.5	27.0	56.4
	top 100-1000	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**

VSE0: $$L_{ret}(C_i, \{I_1, I_2, \cdots, I_n\}) = \sum_{j \neq i} [m - s(c_i, v_i) + s(c_i, v_j)]_+$$

VSE++: $$L_{ret}(C_i, \{I_1, I_2, \cdots, I_n\}) = \max_{j \neq i} [m - s(c_i, v_i) + s(c_i, v_j)]_+$$

Show, Tell and Discriminate

Uniqueness and novelty evaluation

Table 4. Text-to-image retrieval performance, and uniqueness and novelty of generated captions by different methods on COCO.

Methods	Generated-caption-to-image retrieval			Uniqueness and novelty evaluation	
	recall@1	recall@5	recall@10	unique captions	novel captions
Skeleton Key [43]	-	-	-	66.96%	52.24%
Ours-baseline	27.5	59.3	74.0	61.56%	51.38%
Ours-SR-PL	**33.0**	**66.4**	**80.1**	**72.34%**	**61.52%**

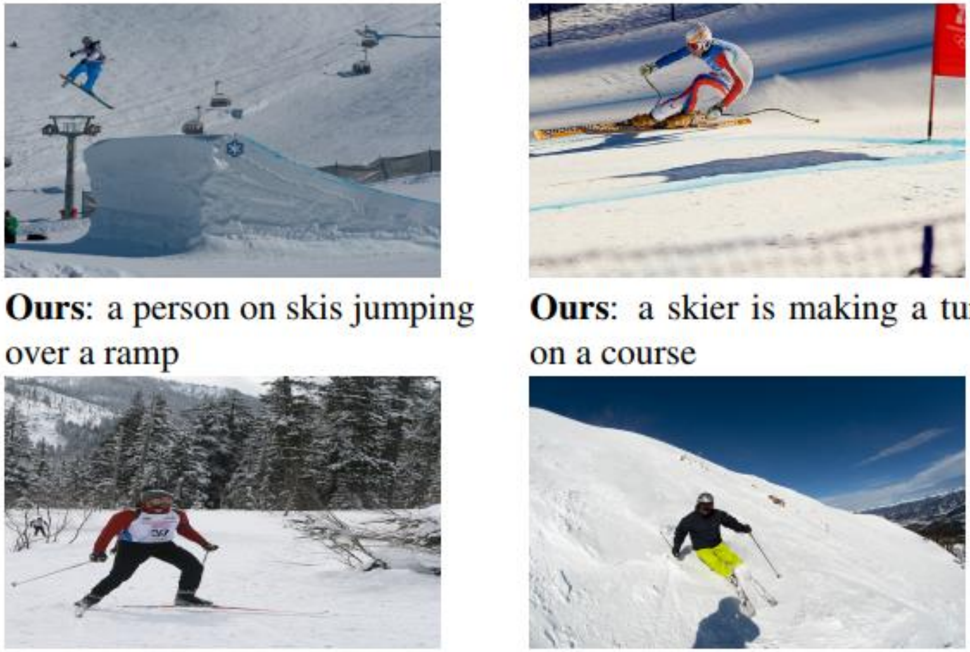
Unique captions: captions that are unique in all generated captions

Novel captions: captions that have not been seen in training

Qualitative results

Speaking the Same Language

- Problems in Captioning
  - Machine and human captions are quite distinct
    - ✓ Word distributions
    - ✓ Vocabulary size
    - ✓ Strong bias (frequent captions)
  - How to generate human-like captions
    - ✓ Multiple captions
    - ✓ Diverse captions



Ours: a person on skis jumping over a ramp

Ours: a skier is making a turn on a course

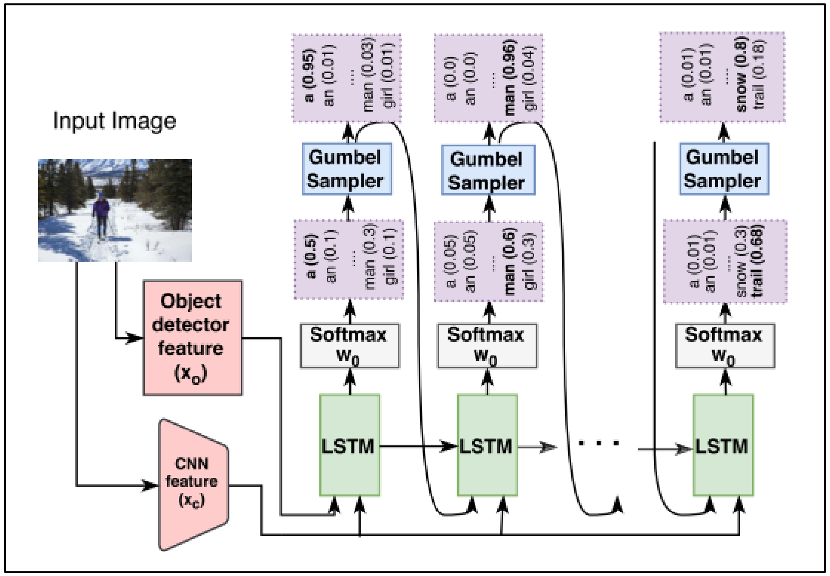
Ours: a cross country skier makes his way through the snow

Ours: a skier is headed down a steep slope

Baseline: a man riding skis down a snow covered slope

Rakshith Shetty, et al., Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. ICCV, 2017.

Speaking the Same Language



$$K_i = T_s \cdot f(s_i)$$

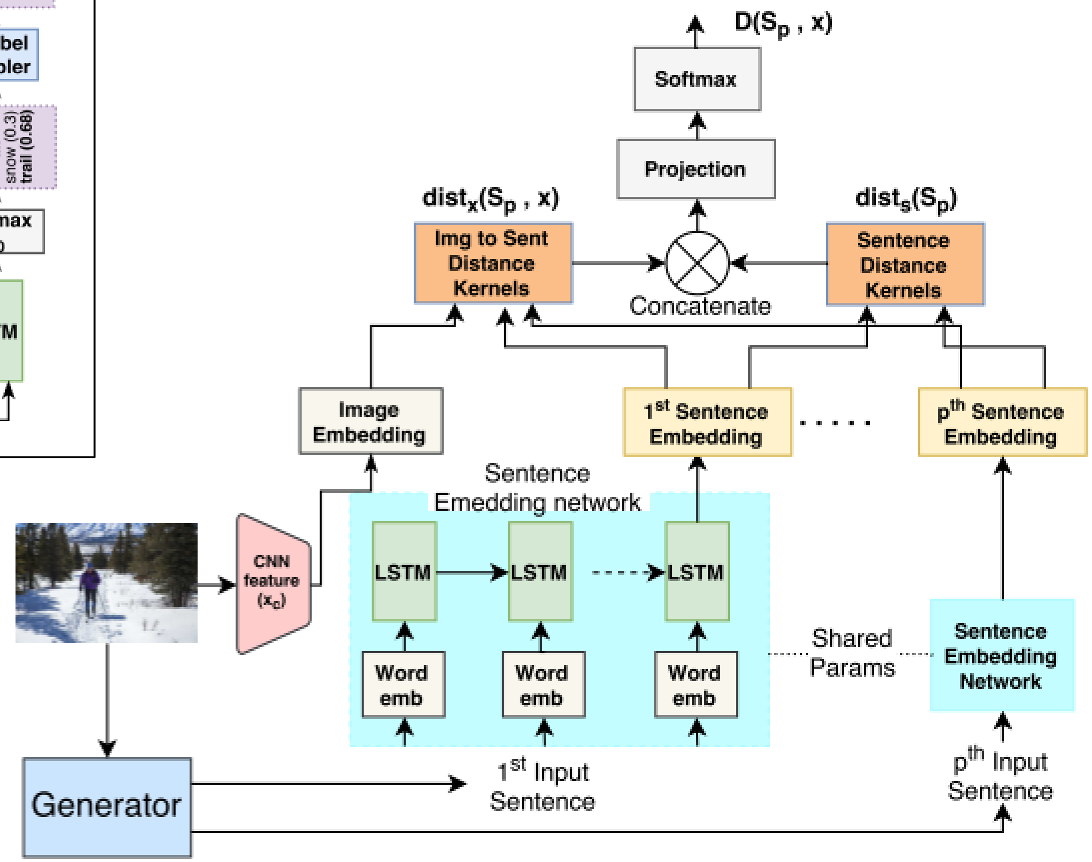
$$c_l(s_i, s_j) = exp\left(-\|K_{i,l} - K_{j,l}\|_{L_1}\right)$$

$$d_l(s_i) = \sum_{j=1}^{p} c_l(s_i, s_j)$$

$$dist_s(S_p) = [d_1(s_1), ..., d_O(s_1), ..., d_O(s_p)] \in \mathbb{R}^{p \times O}$$

$$L(D) = -\log\left(D(S_p^r, x)\right) - \log\left(1 - D(S_p^g, x)\right) - \log\left(1 - D(S_p^f, x)\right)$$

$$L(G) = -\log\left(D(S_p^g, x)\right) + \|\mathbb{E}\left[dist_s(S_p^g)\right] - \mathbb{E}\left[dist_s(S_p^r)\right]\|_2 + \|\mathbb{E}\left[dist_x(S_p^g, x)\right] - \mathbb{E}\left[dist_x(S_p^r, x)\right]\|_2$$



Rakshith Shetty, et al., Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. ICCV, 2017.

Speaking the Same Language

- Discreteness Problem
  - Produce captions from generator
    - ✓ Generate multiple sentences and pick one with highest prob
    - ✓ Use greedy search approaches (beam search)
  - Directly providing discrete samples as input to discriminator does not allow BP (Discontinuous , Non- differentiable)
- Alternative Options:
  - Reinforce trick (Policy Gradient)
    - ✓ High variance
    - ✓ Computationally intensive (sampling)
  - Softmax Distribution -> Discriminator
    - ✓ Easily distinguishes between softmax distribution and sharp ref.
  - Straight-Through Gumbel Softmax approximation

Gumbel-Softmax

- Gumbel分布

CDF: $G_Z(z;a,b) = Pr(Z \le z) = e^{-e^{-\frac{z-a}{b}}}$

PDF: $f(z;a,b) = \frac{1}{b} e^{-(\frac{z-a}{b}+e^{-\frac{z-a}{b}})}$

均值$a+\gamma b$

- 标准Gumbel分布G(0,1)

$G(0,1) = e^{-e^{-z}}$ $f(z) = e^{-(z+e^{-z})}$

- 采样

$$X_\pi = argmax(log(\pi_k) + G_k)$$

$$G_k = -log(-log(U)), U \sim U(0,1)$$

$$y_i = \frac{\exp((\log(\pi_i)+g_i)/\tau)}{\sum_{j=1}^{k}\exp((\log(\pi_j)+g_j)/\tau)}$$

1. $U \sim \mathrm{Uniform}(0,1)$采样$u_1,\ldots,u_K$。
2. $Z = -\ln(-\ln U)) \sim G_Z(z;0,1)$,则 $\{z_i = -\ln(-\ln u_i))\}_{i=1}^K$是$Z$的采样。
3. $Y = \arg\max_i(x_i + z_i)$是服从$\mathrm{Categorical}(\pi_1,\ldots,\pi_K)$分布的。

Speaking the Same Language

Experimental Results

Performance Comparison

Method	Meteor	Spice
ATT-FCN [45]	0.243	–
MSM [44]	0.251	–
KWL [26]	0.266	**0.194**
Ours Base-bs	**0.272**	0.187
Ours Base-samp	0.265	0.186
Ours Adv-bs	0.239	0.167
Ours Adv-samp	0.236	0.166

Diversity Comparison

Method	n	Div-1	Div-2	mBleu-4	Vocab-ulary	% Novel Sentences
Base-bs	1 of 5	–	–	–	756	34.18
	5 of 5	0.28	0.38	0.78	1085	44.27
Base-samp	1 of 5	–	–	–	839	52.04
	5 of 5	0.31	0.44	0.68	1460	55.24
Adv-bs	1 of 5	–	–	–	1508	68.62
	5 of 5	0.34	0.44	0.70	2176	72.53
Adv-samp	1 of 5	–	–	–	1616	73.92
	5 of 5	**0.41**	**0.55**	**0.51**	**2671**	**79.84**
Human captions	1 of 5	–	–	–	3347	92.80
	5 of 5	0.53	0.74	0.20	7253	95.05

Diversity in a set of captions for corresp. Image

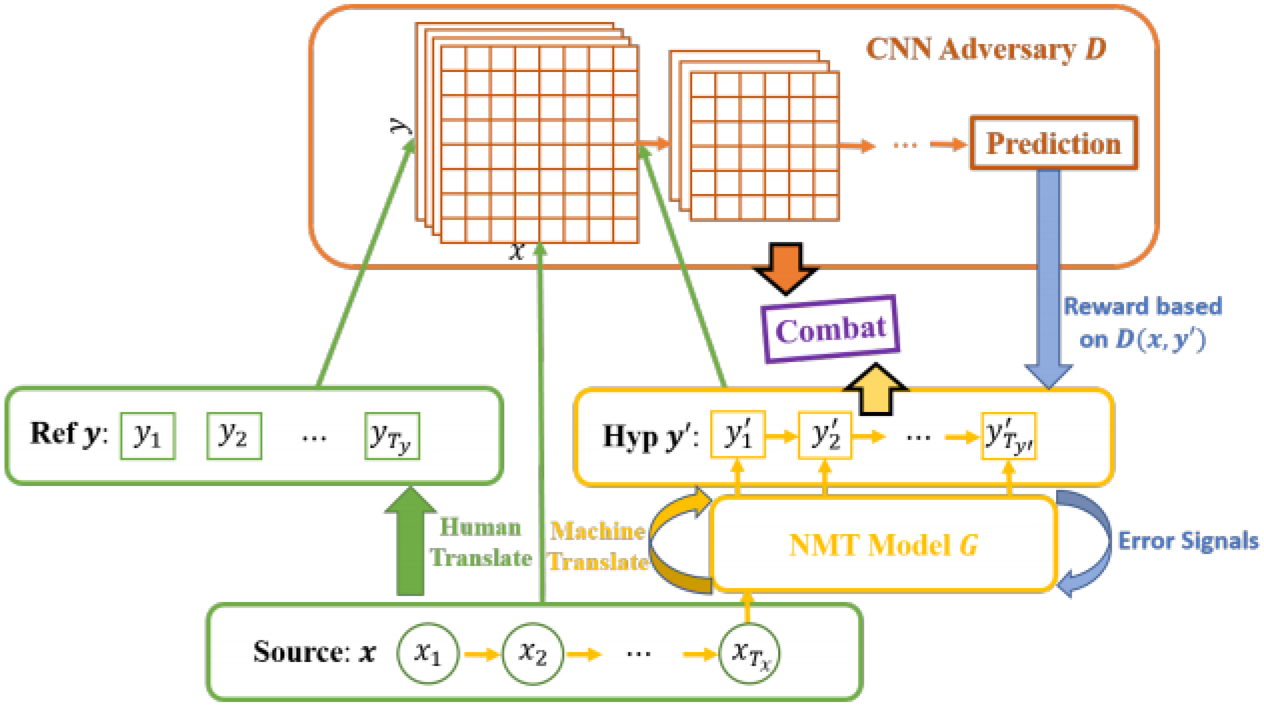
- *Div-1* - ratio of number of unique unigrams in S_p to number of words in S_p. Higher is more diverse.
- *Div-2* - ratio of number of unique bigrams in S_p to number of words in S_p. Higher is more diverse.
- *mBleu* - Bleu score is computed between each caption in S_p against the rest. Mean of these p Bleu scores is the mBleu score. Lower values indicate more diversity.

Corpus Level Diversity

- *Vocabulary Size* - number of unique words used in all generated captions
- *% Novel Sentences* - percentage of generated captions

Adversarial Neural Machine Translation

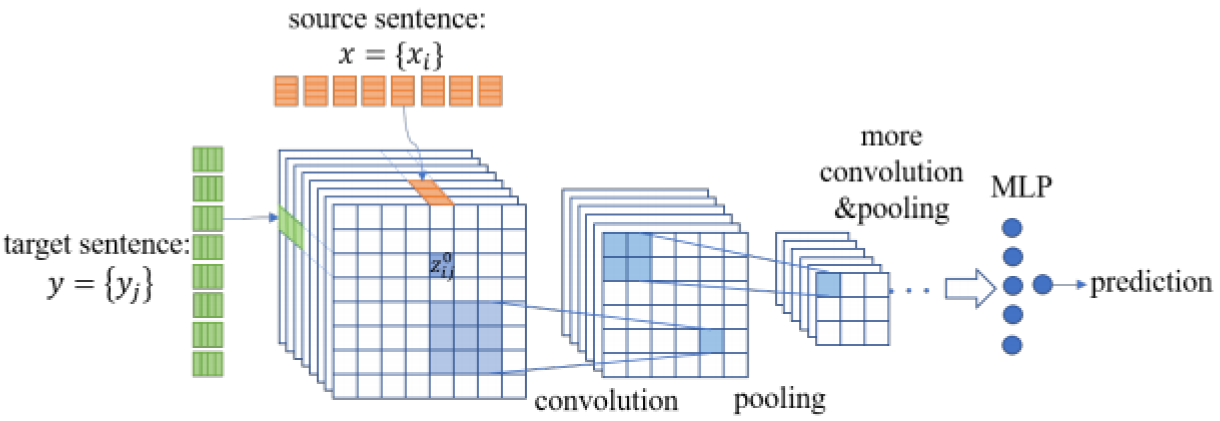
Framework



Lijun Wu, Yingce Xia, Tie-yan Liu, et al., Adversarial Neural Machine Translation. ACML, 2018.

Adversarial Neural Machine Translation

Discriminator



Training

- Warm-up training with MLE
- For a mini-batch, 50% samples for PG, others for MLE
- Reward: whole sentence reward for each time step

Lijun Wu, Yingce Xia, Tie-yan Liu, et al., Adversarial Neural Machine Translation. ACML, 2018.

Sources

- CaptionGAN: Theano Implementation
- SeqGAN: TensorFlow Implementation
- Adversarial-NMT: PyTorch Implementation

Thank you~