DISCRIMINABILITY OBJECTIVE FOR TRAINING DESCRIPTIVE CAPTIONS

Ruotian Luo — *TTI-Chicago* — rluo@ttic.edu

Brian Price — *Adobe Research* — bprice@adobe.com

Scott Cohen — *Adobe Research* — scohen@adobe.com

Greg Shakhnarovich — *TTI-Chicago* — greg@ttic.edu

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Introduction

Current systems trained with maximum likelihood estimation (MLE) or optimizing CIDEr, tend to yield overly general captions.

However, humans appear to notice "interesting" details that are likely to distinguish the image from other potentially similar images, even without explicitly being requested to do so.

Human: a large jetliner taking off from an airport runway
ATTN+CIDER: a large airplane is flying in the sky
Ours: a large airplane taking off from runway

Human: a jet airplane flying above the clouds in the distance
ATTN+CIDER: a large airplane is flying in the sky
Ours: a plane flying in the sky with a cloudy sky

To reduce this gap, we propose to incorporate discriminability when learning the caption generator, as part of the training objective. The discriminability is measured by a (pre-trained) *image/captionretrieval model* which works as a proxy for human.

Base Models

Captioning model:
Basic form: an RNN decoder, producing generative model

$$c = (w_0 = \text{<BOS>}, w_1, \ldots, w_T = \text{<EOS>})$$
$$p(c|I;\boldsymbol{\theta}) = \prod_t p(w_t|w_{t-1}, I; \boldsymbol{\theta})$$

FC model initialized with visual features (CNN, mapped to word representation)
ATTN model Image is encoded into a set of spatially anchored features; attention, modeled as weights on visual features, evolves with the sequence.
Training objectives: Maximum Likelihood Estimation:

$$\max_{\boldsymbol{\theta}} \log p(c|I;\boldsymbol{\theta}) = \sum_t \log p(w_t|w_{t-1}, I; \boldsymbol{\theta})$$

or CIDEr optimization:

$$\max_{\boldsymbol{\theta}, \hat{c} \sim p(c|I;\boldsymbol{\theta})} \text{CIDER}(\hat{c})$$

Use self-critical sequence training method to optimize. (REINFORCE with baseline)

Retrieval model:
Image encoder: image $I \to f(I)$;
text encoder: caption $c \to g(c)$.
Similarity between image and caption: cosine distance:

$$s(I, c) = \frac{f(I) \cdot g(c)}{\|f(I)\| \|g(c)\|}$$

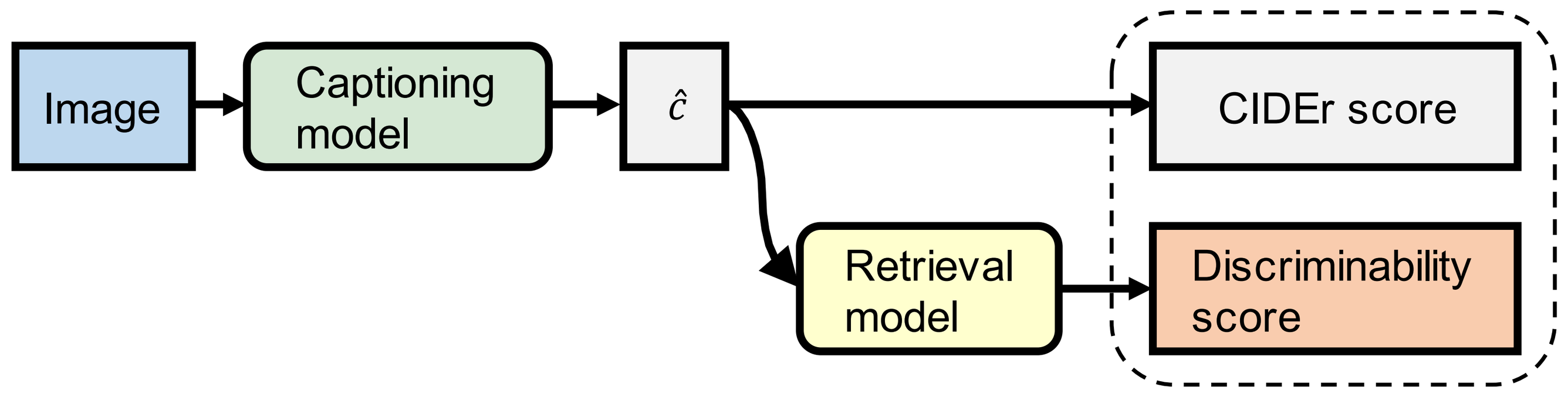
Training objective: contrastive loss:

$$L_{\text{CON}}(c, I) = \max_{c'} [\alpha + s(I, c') - s(I, c)]_+ + \max_{I'} [\alpha + s(I', c) - s(I, c)]_+$$

Intuition: score of correct match (I, c) should be higher than score of any mismatched pair (I, c') or (I', c)
$\max$ is taken over some set of captions/images

Discriminability objective

Define $\hat{c}$ as a sampled caption from $p_{\boldsymbol{\theta}}(c|I)$. We use $L_{\text{CON}}(\hat{c}, I)$ to measure the "discriminability" of $\hat{c}$: the lower, the more discriminative.



We combine the CIDEr objective and discriminability objective and form our training objective:

$$\max \text{CIDER}(\hat{c}) - \lambda L_{\text{CON}}(\hat{c}, I)$$

Quantitative results

Conclusions from validation set:

- ATTN models better than FC models, and discriminability objective works for both.
- ATTN+CIDEr+* combination is our best choice
- Moderate $\lambda = 1$ produces good tradeoff between discriminability and fluency
- Higher λ make captions more discriminative to machine *and to humans*, but at the cost of fluency
- With moderate λ, non-discriminative scores like BLEU, METEOR, CIDEr improve as well!
- especially surprising result: CIDEr (ostensibly we focus less on maximizing it during training.)

	BLEU4	CIDEr	SPICE	Acc	3 in 5	4 in 5	5 in 5
FC+MLE	**0.3308**	1.0005	0.1855	77.23%	71.78%	50.28%	18.79%
FC+CDR	0.3249	1.0154	0.1899	74.00%	73.04%	50.58%	24.83%
FC+CDR+D(1)	0.3274	**1.0231**	**0.1939**	79.26%	74.26%	55.53%	24.13%
FC+CDR+D(5)	0.3072	0.9678	0.1904	85.90%	78.63%	58.03%	32.64%
FC+CDR+D(10)	0.2727	0.8795	0.1807	**88.69%**	**80.01%**	**62.71%**	**37.15%**
ATTN+MLE	0.3582	1.1078	0.2019	72.40%	69.90%	54.60%	28.07%
ATTN+CDR	0.3592	1.1332	0.2083	71.05%	69.97%	51.34%	27.34%
ATTN+CDR+D(1)	**0.3627**	**1.1406**	**0.2113**	75.74%	72.70%	53.23%	34.33%
ATTN+CDR+D(5)	0.3504	1.1026	0.2097	80.98%	76.69%	60.94%	33.49%
ATTN+CDR+D(10)	0.3261	1.0552	0.2070	**83.50%**	**81.93%**	**65.12%**	**35.41%**

Test set result:

	BLEU4	CIDEr	SPICE	Acc	3 in 5	4 in 5	5 in 5
Human	-	-	-	74.30%	91.14%	82.38%	57.08%
MLE	0.5956	1.2198	0.2132	68.60%	72.06%	59.06%	44.25%
CIDER	0.5971	1.2604	0.2260	68.19%	70.07%	55.95%	35.95%
CACA	0.4719	0.7656	0.1526	75.80%	74.1%	56.88%	35.19%
CIDER+D(1)	**0.3971**	**1.2770**	**0.2302**	72.63%	76.91%	61.67%	40.09%
CIDER+D(10)	0.3538	1.1429	0.2204	**79.75%**	**77.70%**	**64.63%**	**44.63%**

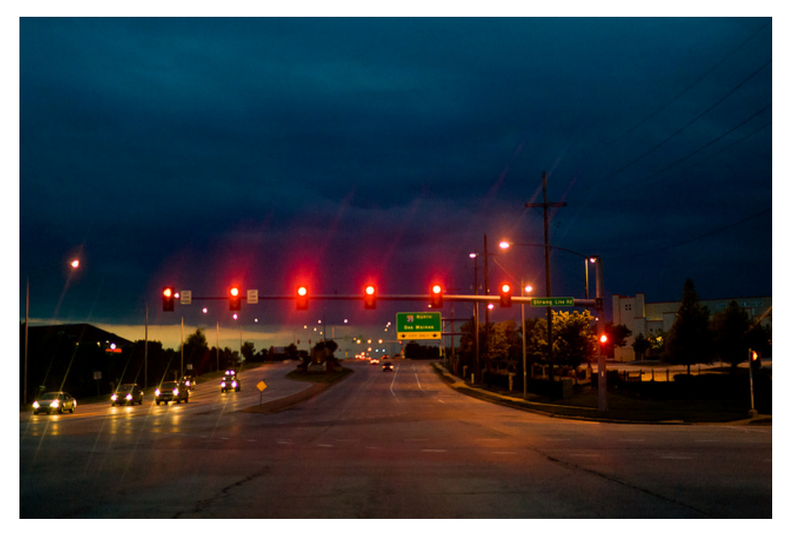
Subclass SPICE scores:

	Color	Attribute	Cardinality	Object	Relation	Size
ATT+MLE	11.78	10.13	3.00	36.42	5.52	**3.67**
ATT+C	7.24	8.77	8.93	38.38	**6.21**	2.39
ATT+C+D(1)	9.25	9.49	10.51	**38.96**	5.91	2.58
ATT+C+D(5)	11.99	10.40	15.23	38.57	5.59	2.53
ATT+C+D(10)	**12.88**	**10.88**	**15.72**	38.09	5.35	2.53

Qualitative results

Human: a young child holding an umbrella with birds and flowers
ATTN+CIDER: a group of people standing in the rain with an umbrella
Ours:a little girl holding an umbrella in the rain

Human: costumed wait staff standing in front of a restaurant awaiting customers
ATTN+CIDER: a group of people standing in the rain with an umbrella
Ours: a group of people standing in front of a building



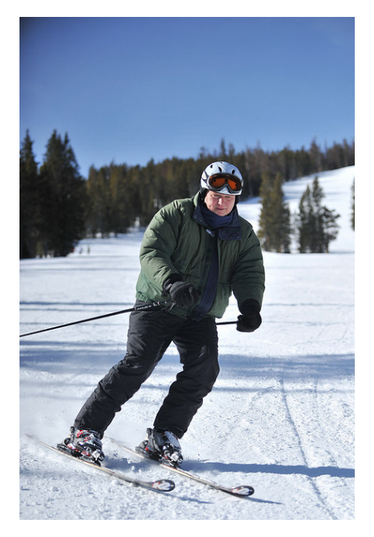
Human: a street that goes on to a high way with the light on red
ATTN+CIDER: a traffic light on the side of a city street
Ours:a street at night with traffic lights at night

Human: view of tourist tower behind a traffic signal
ATTN+CIDER: a traffic light on the side of a city street
Ours: a traffic light sitting on the side of a city

Human: people skiing in the snow on the mountainside
ATTN+CIDER: a group of people standing on skis in the snow
Ours:a group of people skiing down a snow covered slope

Human: two skiers travel along a snowy path towards trees
ATTN+CIDER: a group of people standing on skis in the snow
Ours: two people standing on skis in the snow

Human: a man riding skis next to a blue sign near a forest
ATTN+CIDER: a man standing on skis in the snow
Ours: a man standing in the snow with a sign



Human: the man is skiing down the hill with his goggles up
ATTN+CIDER: a man standing on skis in the snow
Ours: a man riding skis on a snow covered slope

Human: a hot dog serves with fries and dip on the side
ATTN+CIDER: a plate of food with meat and vegetables on a table
Ours: a hot dog and french fries on a plate

Human: a plate topped with meat and vegetables and sauce
ATTN+CIDER: a plate of food with meat and vegetables on a table
Ours: a plate of food with carrots and vegetables on a plate

Human: a train on an overpass with people under it
ATTN+CIDER: a train is on the tracks at a train station
Ours: a red train parked on the side of a building

Human: a train coming into the train station
ATTN+CIDER: a train is on the tracks at a train station
Ours: a green train traveling down a train station

ATTN+MLE: a woman standing in a kitchen preparing food
ATTN+CIDER: a woman standing in a kitchen preparing food
ATTN+CIDER+DISC(1): a woman standing in a kitchen with a fireplace
ATTN+CIDER+DISC(10): a woman standing in a kitchen with a brick oven

ATTN+MLE: a cat sitting next to a glass of wine
ATTN+CIDER: a cat sitting next to a glass of wine
ATTN+CIDER+DISC(1): a cat sitting next to a bottles of wine
ATTN+CIDER+DISC(10): a cat sitting next to a bottles of wine bottles

ATTN+MLE: a man riding a wave on top of a surfboard
ATTN+CIDER: a man riding a wave on a surfboard in the ocean
ATTN+CIDER+DISC(1): a person riding a wave on a surfboard in the ocean
ATTN+CIDER+DISC(10): a person kiteboarding on a wave in the ocean

Conclusion

We demonstrated that incorporating a discriminability objective, derived from the loss of a trained image/caption retrieval model during training improves the quality captions on both discriminability and standard caption metrics not directly related to discrimination, reflecting more descriptive captions.