Discriminability objective for training descriptive image captions

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Ruotian Luo
TTI-Chicago

Joint work with
Brian Price (Adobe)
Scott Cohen (Adobe)
Greg Shakhnarovich (TTIC)

Image captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Example

ATTN+CIDER: a large airplane is flying in the sky

ATTN+CIDER: a large airplane is flying in the sky

Example

ATTN+CIDER: a large airplane is flying in the sky

Human: a large jetliner taking off from an airport runway

ATTN+CIDER: a large airplane is flying in the sky

Human: a jet airplane flying above the clouds in the distance

Example

ATTN+CIDER: a large airplane is flying in the sky

Human: a large jetliner taking off from an airport runway

Ours: a large airplane taking off from runway

ATTN+CIDER: a large airplane is flying in the sky

Human: a jet airplane flying above the clouds in the distance

Ours: a plane flying in the sky with a cloudy sky

Our approach

- We propose a discriminability objective which encourages the output captions to be more discriminative.
- The discriminability objective is derived from how well a (pre-trained) image caption retrieval model can match the images and the generated captions .

Captioning models[1]

- 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.

[1] Steven J Rennie et al. "Self-critical sequence training for image captioning". In: *arXiv preprint arXiv:1612.00563* (2016).

Captioning models training[1]

- Maximum Likelihood Estimation:

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

- CIDEr optimization:

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

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

[1] Steven J Rennie et al. "Self-critical sequence training for image captioning". In: *arXiv preprint arXiv:1612.00563* (2016).

Retrieval model[2]

- Image encoder: image $I \rightarrow f(I)$
- Text encoder: caption $c \rightarrow g(c)$
- Similarity between image and caption: cosine

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

[2] Fartash Faghri et al. "VSE++: Improved Visual-Semantic Embeddings". In: *arXiv preprint arXiv:1707.05612* (2017).

Retrieval model[2]

- Image encoder: image $I \to f(I)$
- Text encoder: caption $c \to g(c)$
- Similarity between image and caption: cosine

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

- 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)

[2] Fartash Faghri et al. "VSE++: Improved Visual-Semantic Embeddings". In: *arXiv preprint arXiv:1707.05612* (2017).

Retrieval model[2]

- Image encoder: image $I \to f(I)$
- Text encoder: caption $c \to g(c)$
- Similarity between image and caption: cosine

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

- 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

[2] Fartash Faghri et al. "VSE++: Improved Visual-Semantic Embeddings". In: *arXiv preprint arXiv:1707.05612* (2017).

Discriminability objective

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

- Can be seen as a discriminability measurement, the lower the more discriminative.

Discriminability objective

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

- Can be seen as a discriminability measurement, the lower the more discriminative.
- Increase discriminability of the generated caption:

$$\min L_{\mathsf{CON}}(\widehat{c}, I),\ \widehat{c}(I) \sim p_{\boldsymbol{\theta}}(c|I)$$

Discriminability objective

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

- Can be seen as a discriminability measurement, the lower the more discriminative.
- Increase discriminability of the generated caption:

$$\min L_{\mathsf{CON}}(\widehat{c}, I),\ \widehat{c}(I) \sim p_{\boldsymbol{\theta}}(c|I)$$

- Optimizing L_{CON} leads to unfluent captions:

$$\max \mathsf{CIDEr}(\widehat{c}) \ - \ \lambda L_{\mathsf{CON}}(\widehat{c}, I)$$

Discriminability objective

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

- Can be seen as a discriminability measurement, the lower the more discriminative.
- Increase discriminability of the generated caption:

$$\min L_{\mathsf{CON}}(\widehat{c}, I),\ \widehat{c}(I) \sim p_{\boldsymbol{\theta}}(c|I)$$

- Optimizing L_{CON} leads to unfluent captions:

$$\max \mathsf{CIDEr}(\widehat{c}) \ - \ \lambda L_{\mathsf{CON}}(\widehat{c}, I)$$

$\lambda = 0$, identical to SCST.

Experiments

- Dataset: MSCOCO
- training set of 113,287 images,
  validation set of 5,000 images,
  5,000 images used to set up test set: 1000 challenging image pairs[3]
- Metrics:
- Automatic measurements: BLEU, METEOR, ROUGE, CIDEr, SPICE
- Human experiment on discriminability:

[3] Ramakrishna Vedantam et al. "Context-aware Captions from Context-agnostic Supervision". In: *arXiv preprint arXiv:1701.02870* (2017).

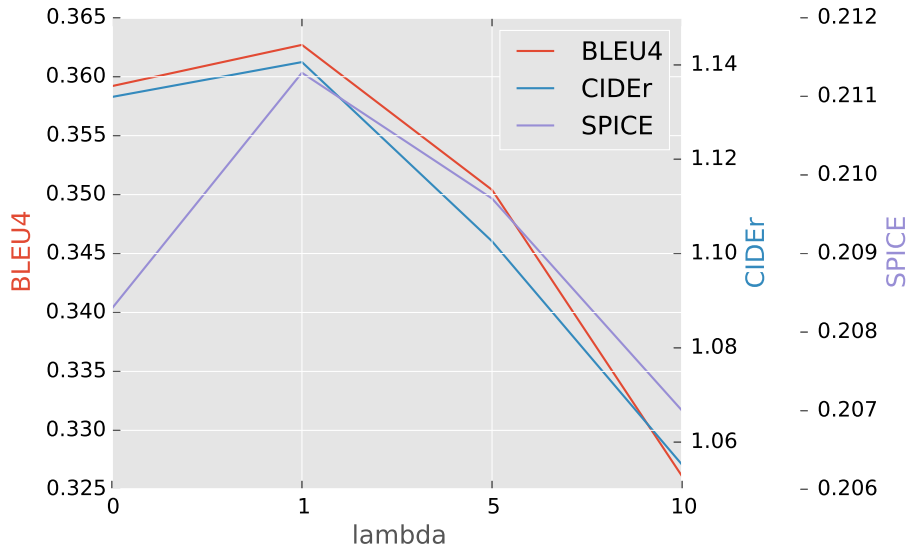
Validation experiments

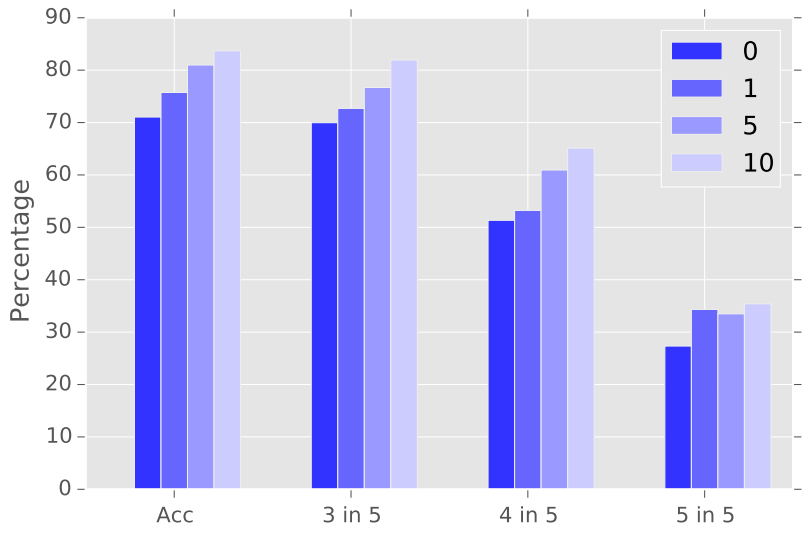
- Discriminability objective works agnosticly to the choice of captioning model.
- ATTN models better than FC models.

	CIDEr	SPICE	Acc	3 in 5	4 in 5	5 in 5
FC+CIDER	1.0154	0.1899	74.00%	73.04%	50.58%	24.83%
FC+CIDER+DISC (1)	1.0231	0.1939	79.26%	74.26%	55.53%	24.13%
ATTN+CIDER	1.1332	0.2083	71.05%	69.97%	51.34%	27.34%
ATTN+CIDER+DISC (1)	1.1406	0.2113	75.74%	72.70%	53.23%	34.33%

Validation experiments

- Moderate $\lambda = 1$ produces improves both non-discriminative and discriminative score.
- Higher λ make captions more discriminative to machine *and to humans*, but at the cost of fluency
- Especially surprising result: CIDEr (ostensibly we focus less on maximizing it during training.)





Scores on test set

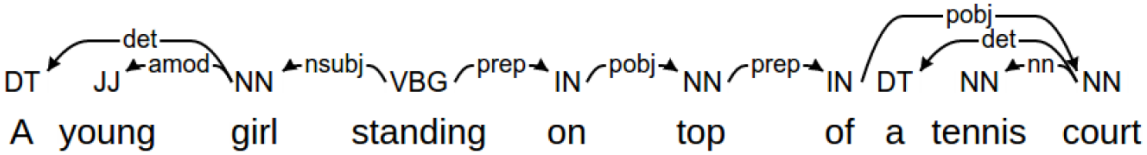
- Test on ATTN models.
- CACA[3]: requires seeing the "distractor images" before producing caption for the target

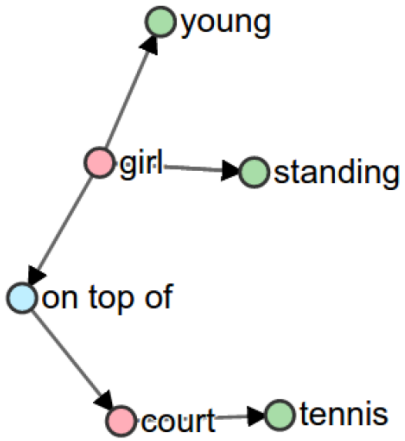
	BLEU4	CIDEr	SPICE	Acc	3 in 5	4 in 5	5 in 5
Human	-	-	-	74.30%	91.14%	82.38%	57.08%
MLE	0.5956	1.2198	0.2132	68.60%	72.06%	59.06%	44.25%
CIDER	0.5971	1.2604	0.2260	68.19%	70.07%	55.95%	35.95%
CACA	0.4719	0.7656	0.1526	75.80%	74.1%	56.88%	35.19%
CIDER+D(1)	**0.3971**	**1.2770**	**0.2302**	72.63%	76.91%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

[3] Ramakrishna Vedantam et al. "Context-aware Captions from Context-agnostic Supervision". In: *arXiv preprint arXiv:1701.02870* (2017).

SPICE

- Semantic propositional image caption evaluation score
- Inference over objects, attributes, relations

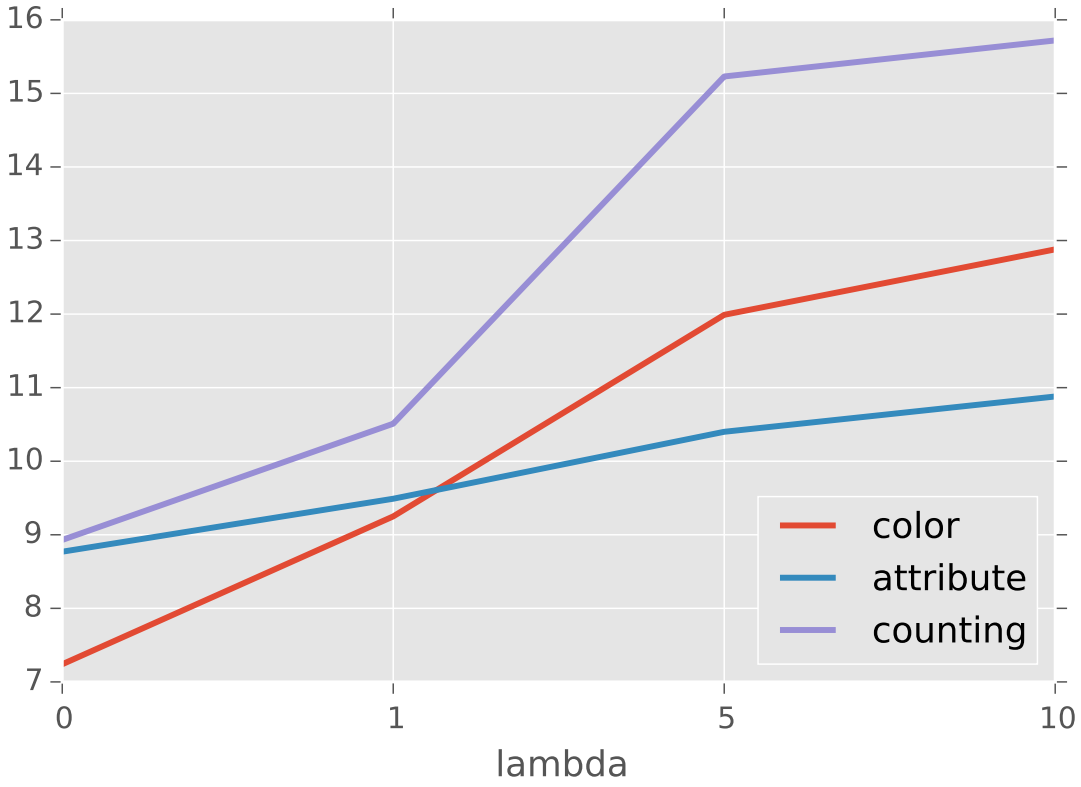




- F-score is calculated for semantic tuples, e.g., { (girl), (court), (girl, young), (girl, standing) (court, tennis), (girl, on-top-of, court) }
- Can calculate separately for different aspects: color, counting, size, etc.

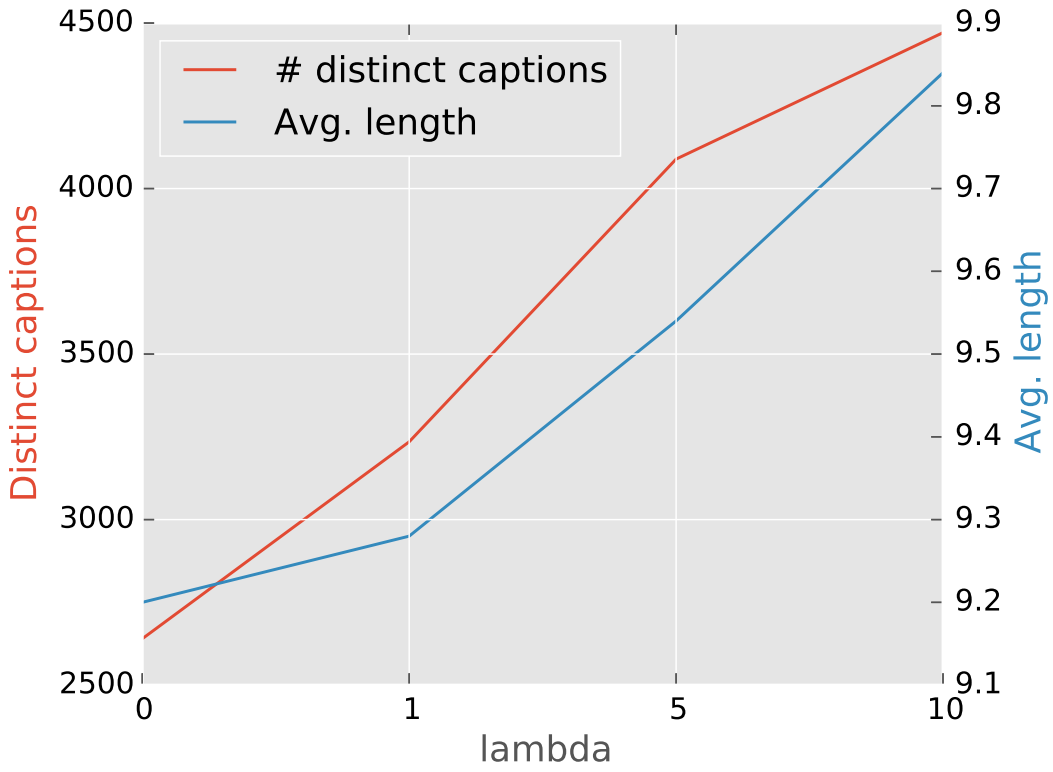
Results: SPICE

- Results on 5k validation (all scores scaled $\times 100$)



Effect on caption diversity

- For the validation set, 5000 images



Examples: disambiguation

- Images with identical baseline captions selected for display purposes
- "Ours" = ATTN+CIDER+D(1)

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

Examples: disambiguation

Human: a hot dog served 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

Examples: disambiguation

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

Examples: disambiguation

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

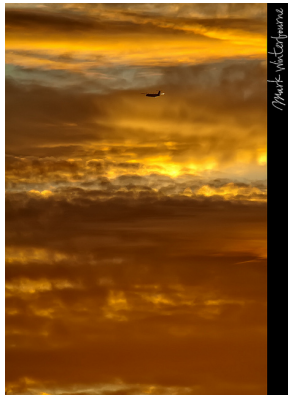
Examples: model comparison

- Left: the target image; right (for illustration): a distractor (shown to human subjects in Mechanical Turk evaluation)

ATTN+MLE: a large clock tower with a clock on it
ATTN+CIDER: a clock tower with a clock on the side of it
ATTN+CIDER+DISC(1): a clock tower with bikes on the side of a river
ATTN+CIDER+DISC(10): a clock tower with bicycles on the boardwalk near a harbor

Examples: model comparison

- High λ often causes "insistence" on distinctive words, at the expense of fluency



ATTN+MLE: a view of an airplane flying through the sky
ATTN+CIDER: a plane is flying in the sky
ATTN+CIDER+DISC(1): a plane flying in the sky with a sunset
ATTN+CIDER+DISC(10): a sunset of a sunset with a sunset in the sunset

Examples: model comparison

ATTN+MLE: a couple of people standing next to a stop sign
ATTN+CIDER: a stop sign on the side of a street
ATTN+CIDER+DISC(1): a stop sign in front of a store with umbrellas
ATTN+CIDER+DISC(10): a stop sign sitting in front of a store with shops

Examples: model comparison

ATTN+MLE: a large jetliner sitting on top of an airport tarmac
ATTN+CIDER: a large airplane sitting on the runway at an airport
ATTN+CIDER+DISC(1): a blue airplane sitting on the tarmac at an airport
ATTN+CIDER+DISC(10): a blue and blue airplane sitting on a tarmac with a plane

Conclusions

- Propose a discriminability objective, derived from the loss of a trained image/caption retrieval model
- Improves the quality of resulting captions on both discriminability and standard metrics.

Comparison

CL: (I, c) vs (I, c'),

GAN : (I, c) vs $(I, \widehat{c})$, (I, c) vs (I, c')

Ours: $(I, \widehat{c})$ vs $(I, \widehat{c}')$, $(I, \widehat{c})$ vs $(I', \widehat{c})$