Supplementary material: 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

Gregory Shakhnarovich
TTI-Chicago
greg@ttic.edu

1. Result on standard split and MSCOCO test server

In the paper report results on the split[6] Here we also report results (with automatic metrics) on the commonly used split introduced in[3] and on COCO test server. Our observations and conclusions are further confirmed by this evaluation.

	BLEU4	METEOR	ROUGE	CIDER	SPICE
			val		
ATTN+C	0.3511	0.2700	0.5676	1.1225	0.2043
ATTN+C+D(1)	**0.3583**	**0.2733**	**0.5720**	**1.1381**	**0.2094**
ATTN+C+D(5)	0.3416	0.2698	0.5633	1.0902	0.2054
ATTN+C+D(10)	0.3196	0.2664	0.5520	1.0402	0.2024
			test		
ATTN+C	0.3566	0.2710	0.5688	1.1345	0.2058
ATTN+C+D(1)	**0.3614**	**0.2738**	**0.5729**	**1.1425**	**0.2105**
ATTN+C+D(5)	0.3439	0.2696	0.5640	1.1018	0.2082
ATTN+C+D(10)	0.3203	0.2671	0.5531	1.0530	0.2050

Table 1. Automatic scores on standard split. (Here +C means using CIDEr optimzation; +D(x) means using discriminability loss with λ being x)

	Color	Attribute	Cardinality	Object	Relation	Size
			val			
ATTN+C	6.27	8.19	9.07	38.18	5.67	**2.76**
ATTN+C+D(1)	8.17	8.89	10.94	**38.97**	**5.70**	2.41
ATTN+C+D(5)	10.83	9.70	13.98	38.27	5.19	2.68
ATTN+C+D(10)	**12.33**	**10.25**	**14.47**	37.72	4.87	2.25
			test			
ATTN+C	5.54	7.82	7.83	38.43	6.04	2.21
ATTN+C+D(1)	7.55	8.72	9.75	**39.08**	**6.05**	2.24
ATTN+C+D(5)	10.46	9.89	13.05	38.46	5.66	**2.59**
ATTN+C+D(10)	**12.21**	**10.42**	**14.84**	38.04	5.19	2.54

Table 2. SPICE subclass scores on standard split. All the scores here are scaled up by 100.

2. Improved diversity with discriminability objective

Table 4 shows that with larger λ we can get longer and more diverse captions. We also observe pure CIDEr optimization will harm the diversity of output captions.

Metric	Ours(c5)	Base(c5)	Ours(c40)	Base(c40)
BLEU-1	0.795	0.794	0.941	0.939
BLEU-2	0.627	0.623	0.869	0.861
BLEU-3	0.474	0.470	0.764	0.754
BLEU-4	0.352	0.349	0.647	0.637
METEOR	0.271	0.268	0.355	0.351
ROUGE-L	0.567	0.564	0.710	0.705
CIDEr-D	1.100	1.091	1.116	1.106

Table 3. Results on MSCOCO test, reported by the MSCOCO server. Ours: ATTN+CIDER+DISC(1). Baseline: ATTN+CIDER.

In Figure 1, we show pairs of images that have the same caption generated by ATTN+CIDER [4], where our method (ATTN+CIDER+DISC(1)) can generate more specific captions.

	# distinct captions	Avg. length
FC+MLE	2700	8.99
FC+CIDER[4]	2242	9.30
FC+CIDER+DISC (1)	3204	9.32
FC+CIDER+DISC (5)	4379	9.45
FC+CIDER+DISC (10)	4634	9.78
ATTN+MLE	2982	9.01
ATTN+CIDER [4]	2640	9.20
ATTN+CIDER+DISC (1)	3235	9.28
ATTN+CIDER+DISC (5)	4089	9.54
ATTN+CIDER+DISC (10)	4471	9.84

Table 4. Distinct caption number on validation set (split in [6]), and average sentence length with different method

3. Comparison to [1, 2, 5]

First, while the losses in [1, 2, 5] and our discriminability loss all deal with matched vs. mismatched image/caption pairs, the actual formulations are very different.

If (I, c) is the matched image/caption pair, I', c' are other images/captions, and $\widehat{c}$ the generated caption, the losses are related to the following notions of "contrast" :

$$[1]\text{: } (I,c) \text{ vs } (I,c'),$$
$$[2,5]: \ (I,c) \text{ vs } (I,\widehat{c}),\ (I,c) \text{ vs } (I,c')$$
$$\text{Ours: } (I,\widehat{c}) \text{ vs } (I,\widehat{c}'),\ (I,\widehat{c}) \text{ vs } (I',\widehat{c})$$

Second, compared to [2, 5], we have a different focus and different evaluation results. The focus in [2, 5] is generating

a diverse set of natural captions; in contrast, we go after discriminability (accuracy) with a single caption. We demonstrate improvement in **both** standard metrics and our target (discriminability), while they have to sacrifice the former to improve the performance under their target (diversity).

On SPICE subcategory score, count and size are improved after applying the GAN training in [5] while our discriminability loss helps color, attribute and count. That implies that our discriminability objective is in favor of different aspects of captions from the discriminator in GAN.

Third, the aims in [1] are more aligned with ours, but [1] does not explicitly incorporate discriminative task into learning, while we do. It is also not clear that [1] could allow for CIDEr optimization, which we do allow for, and which appears to lead to significant improvement over MLE.

Finally, none of [1, 2, 5] address discriminability of resulting captions, or report discriminability results comparable to ours.

4. More results

Following up the analysis of SPICE sub-scores in the paper, showing improvement due to our proposed objective, we show in Figures 2, 3, 4 some examples demonstrating such improvement on color, attribute and cardinality aspects of the scene. For each figure, the captions below are describing the target images(left) generated from different models (ATTN+MLE and ATTN+CIDER are baselines, and ATTN+CIDER+DISC(x) are our method with different λ value). Right images are the corresponding distractors selected in val/test set; these pairs were included in AMT experiments.

Figure 5 shows some more examples and Figure 6 failure cases of our methods. We highlight in green caption elements that (subjectively) seem to aid discriminability, and in red the portions that seem incorrect or jarringly non-fluent.

Human: a young child holding an umbrella with birds and flowers
ATTN+CIDER [4]: 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 [4]: 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 [4]: 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 [4]: 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 [4]: 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 [4]: a group of people standing on skis in the snow
Ours: two people standing on skis in the snow

Figure 1. The human caption, caption generated by ATTN+CIDER, and caption generated by ATTN+CIDER+DISC(1) (denoted as Ours in the figure) for each image

References

[1] B. Dai and D. Lin. Contrastive learning for image captioning. *arXiv preprint arXiv:1710.02534*, 2017. 1, 2

[2] B. Dai, D. Lin, R. Urtasun, and S. Fidler. Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*, 2017. 1, 2

[3] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3128–3137, 2015. 1

[4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016. 1, 2, 3, 4

[5] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *arXiv preprint arXiv:1703.10476*, 2017. 1, 2

[6] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik. Context-aware captions from context-agnostic supervision. *arXiv preprint arXiv:1701.02870*, 2017. 1


ATTN+MLE: a double decker bus parked in front of a building
ATTN+CIDER [4]: a double decker bus parked in front of a building
ATTN+CIDER+DISC(1): a blue double decker bus parked in front of a building
ATTN+CIDER+DISC(10): a blue double decker bus parked in front of a brick building


ATTN+MLE: a park bench sitting next to a tree
ATTN+CIDER [4]: a bench sitting in the middle of a tree
ATTN+CIDER+DISC(1): a black and white photo of a bench in the park
ATTN+CIDER+DISC(10): a black and white photo of a park with a tree


ATTN+MLE: a large jetliner sitting on top of an airport tarmac
ATTN+CIDER [4]: 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

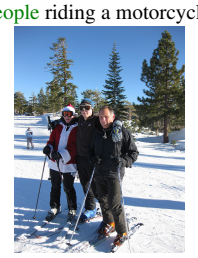
Figure 2. Examples of cases where our method improves the color accuracy of generated captions.


ATTN+MLE: a large clock tower towering over a city street
ATTN+CIDER [4]: a clock tower in the middle of a city street
ATTN+CIDER+DISC(1): a city street with a clock tower at night
ATTN+CIDER+DISC(10): a lit building with a clock tower at night at night


ATTN+MLE: a bathroom with a toilet and a toilet
ATTN+CIDER [4]: a bathroom with a toilet and sink in the
ATTN+CIDER+DISC(1): a dirty bathroom with a toilet and a window
ATTN+CIDER+DISC(10): a dirty bathroom with a toilet and a dirty

Figure 3. Examples of cases where our method improves the attribute accuracy of generated captions.


ATTN+MLE: a couple of people riding on the back of a motorcycle
ATTN+CIDER [4]: a man riding a motorcycle on the street
ATTN+CIDER+DISC(1): two people riding a motorcycle on a city street
ATTN+CIDER+DISC(10): two people riding a motorcycle on a road

ATTN+MLE: a couple of people standing on top of a snow covered slope
ATTN+CIDER [4]: a couple of people standing on skis in the snow
ATTN+CIDER+DISC(1): two people standing on skis in the snow
ATTN+CIDER+DISC(10): two people standing in the snow with a snow

Figure 4. Examples of cases where our method improves the cardinality accuracy of generated captions.


ATTN+MLE: a woman standing in a kitchen preparing food
ATTN+CIDER [4]: 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

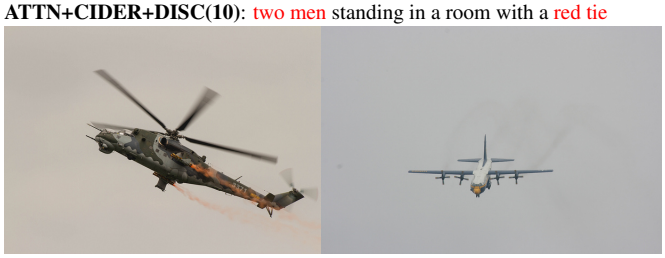

ATTN+MLE: a man on a surfboard in the water
ATTN+CIDER [4]: a man riding a wave on a surfboard in the ocean
ATTN+CIDER+DISC(1): a man riding a kiteboard on the ocean in the ocean
ATTN+CIDER+DISC(10): a man kiteboarding in the ocean on a ocean


ATTN+MLE: a room with a laptop and a laptop
ATTN+CIDER [4]: a laptop computer sitting on top of a table
ATTN+CIDER+DISC(1): a room with a laptop computer and chairs in it
ATTN+CIDER+DISC(10): a room with a laptops and chairs in a building

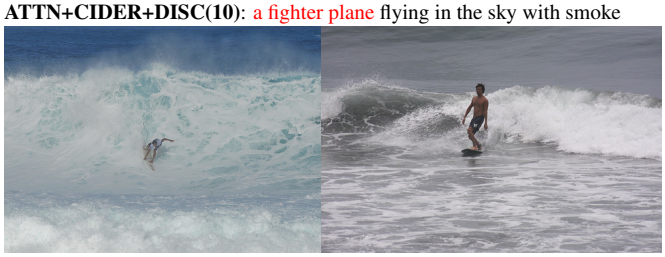

ATTN+MLE: a cat sitting next to a glass of wine
ATTN+CIDER [4]: a cat sitting next to a glass of wine
ATTN+CIDER+DISC(1): a cat sitting next to a bottles of wine
ATTN+CIDER+DISC(10): a cat sitting next to a bottles of wine bottles

Figure 5. Some more examples


ATTN+MLE: a couple of people standing next to each other
ATTN+CIDER [4]: a man and a woman standing in a room
ATTN+CIDER+DISC(1): a man and a woman standing in a room
ATTN+CIDER+DISC(10): two men standing in a room with a red tie



ATTN+MLE: a helicopter that is flying in the sky
ATTN+CIDER [4]: a helicopter is flying in the sky
ATTN+CIDER+DISC(1): a fighter jet flying in the sky
ATTN+CIDER+DISC(10): a fighter plane flying in the sky with smoke



ATTN+MLE: a man riding a wave on top of a surfboard
ATTN+CIDER [4]: 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

Figure 6. Some failure cases of our algorithm