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

Abstract

One property that remains lacking in image captions generated by contemporary methods is discriminability: being able to tell two images apart given the caption for one of them. We propose a way to improve this aspect of caption generation. By incorporating into the captioning training objective a loss component directly related to ability (by a machine) to disambiguate image/caption matches, we obtain systems that produce much more discriminative caption, according to human evaluation. Remarkably, our approach leads to improvement in other aspects of generated captions, reflected by a battery of standard scores such as BLEU, SPICE etc. Our approach is modular and can be applied to a variety of model/loss combinations commonly proposed for image captioning.

1. Introduction

Image captioning is a task of mapping images to text for human consumption. Broadly speaking, in order for a caption to be good it must satisfy two requirements: it should be a *fluent*, well-formed phrase or sentence in the target language; and it should be *informative*, or *descriptive*, conveying meaningful non-trivial information about the visual scene it describes. Our goal in the work presented here is to improve captioning on both of these fronts.

Because these properties are somewhat vaguely defined, objective evaluation of caption quality remains a challenge, more so than evaluation of earlier established tasks like object detection or depth estimation. However, a number of metrics have emerged as preferred, if imperfect, evaluation measures. Comparison to human ("gold standard") captions collected for test images is done by means of metrics borrowed from machine translation, such as BLEU[1], as well as new metrics introduced for the captioning task, such as CIDEr[2] and SPICE[3].

In contrast, to assess how informative a caption is, we may design an explicitly discriminative task the success of which would depend on how accurately the caption describes the visual input. One approach to this is to consider *referring expressions* [4]: captions for an image region, produced with the goal of unambiguously identifying the region within the image to the recipient. We can also consider the ability of a recipient to identify an entire image that matches the caption, out of two (or more) images [5]. This – caption *discriminability* – is the focus of our work presented here.

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

Figure 1. Example captions generated by human, an existing automatic system (ATTN+CIDER[6]), and a model trained with our proposed method (ATTN+CIDER+DISC(1), see Section 5)

Traditionally used training objectives, such as maximum likelihood estimation (MLE) or CIDEr, tend to encourage the model to "play it safe", often yielding overly general captions as illustrated in Fig. 1. Despite the visual differences in the image, a top captioning system [6] produces the same caption for both images. In contrast, 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. (We confirm this assertion empirically in Sec. 5.)

To reduce this gap, we propose to incorporate an explicit measure for discriminability into learning a caption generator, as part of the training loss. Our discriminability loss is derived from the ability of a (pre-trained) *retrieval model* to match the correct image to its caption significantly stronger than any other image in a set, and vice versa (caption to correct image above other captions).

Language-based measures like BLEU reward machine captions for mimicking human captions, and so since, as we state above, human captions are discriminative, one could

expect these measures to be correlated with descriptiveness. However, in practice, given an imperfect caption generator, there may be a tradeoff between fluency and descriptiveness; our training regime allows us to negotiate this tradeoff and ultimately improve both aspects of a generator.

Our discriminability loss can be added to any gradient-based learning procedure for caption generators. We show in Sec. 5 that it can improve some recently proposed models which currently at or near state of the art for captioning, for all metrics evaluated. In particular, to our knowledge, we establish new state of the art in discriminative captioning.

2. Related work

Image captioning Most modern approaches [7, 8, 9] encode an image using a convolutional neural network (CNN), and feed this as input to an recurrent network (RNN), typically with some form of gating or memory. The RNN can generate a arbitrary-length sequence of words. Within this generic framework, many efforts [7, 10, 11, 8, 12, 13, 14, 15, 16] explored different encoder-decoder structures, including attention-based models. There has also been exploration of different training objectives. For example, [17, 18] add some auxiliary tasks like word appearance prediction; [19] uses Conditional Variational Autoencoder(CVAE) and optimize over evidence lower bound(ELBO); [20, 6, 21, 22] applied Reinforcement Learning algorithms on image captioning, so that the models can be optimized directly on the non-differentiable metrics like SPICE, CIDEr, BLEU etc.

Visual Semantic Embedding methods Image-Caption retrieval has been considered as a task relying on image captioning [7, 8, 9, 11]. However, it can also be regarded as a multi-modal embedding task. In previous works [23, 24, 25, 26, 27, 28] visual and textual embeddings are trained with the objective to minimize matching loss, e.g., ranking loss on cosine distance, or to enforce partial order on captions and images.

Discrimination tasks in the context of caption evaluation were proposed in [5, 29]: given a set of other images, called distractors, the generated captions of each image have to distinguish one from others. In the "speaker-listener" model [29], the speaker is trained to generate captions, and a listener to prefer the correct image over a wrong one, given the caption. At test time, the listener re-ranks the captions sampled from the speaker. [5] propose a decoding mechanism which can suppress the caption elements that are common for both target image and disctractor image. In contrast to our work, both [5] and [29] require the distractor to be presented prior to caption production. We aim to generate distinctive captions a-priori, without a specific distractor at hand, like humans appear to do.

Referring expressions is another flavor of discriminative captioning task that has attracted interest after the release of the standard datasets [4, 30, 31]. [32, 33] learned to generate more discriminative referring expressions guided by a referring expression comprehension model. The techniques in those papers are strongly tied to the task of describing a region within an image, while our goal here is to describes natural scenes in their entirety.

Visual Dialog has recently attracted interests in the field [34, 35]. While it's hard to evaulate generic 'chat', [36, 37] propose goal-driven visual dialog tasks and datasets. [36] proposes the 'image guessing' game where two agents – Q-BOT and A-BOT – who communicate in natural language dialog so that Q-BOT can select an unseen image from a lineup of images. GuessWhat Game [37] is similar, but guess an object in a image during a dialog. In another related effort [38] the machine must show understanding the difference between two images by asking a question that has different answers for two images. Our work shares the ultimate purpose (producing text that allows image identification) with these efforts, but in contrast to those, our aim is to generate a caption in a single "shot". This is somewhat similar to round 0 of the dialog in [36], where the agent is given a caption generated by [8] (without regard to any discrimination task) and chooses an image from a set. Since our captions are shown in Sec. 5 to be both fluent and discriminative, switching to using them may improve/shorten visual dialog.

Similar work Finally, some recent work is similar to ours in its goals (learning to produce discriminative captions) and, to a degree, in techniques. The motivation in [39] is similar, but the focus is on caption (rather than image) retrieval. The objective is contrastive: pushing the negative captions from different images to have lower probability than positive captions using noise contrastive learning. In [40], more meaningful visual dialog responses are generated by distilling knowledge from a discriminative model trained to rank different dialog responses given the previous dialog context. [41, 42] proposes using Conditional Gernerative Adversarial Network to train image captioning. They both learn a discriminator to distinguish human captions from machine captions. For more detailed discussion of [39, 41, 42], see supplementary material.

Despite being motivated by a desire to improve caption discriminability, all these methods are fundamentally remain tied to the objective of matching the surface form of human captions, and do not include an explicitly discriminative objective in training. Ours is the first work incorporate both image retrieval and caption retrieval into caption generation training. We can easily "plug" our method into existing models, for instance combine it with CIDEr optimization, leading to improvements in metrics across the board: both the discriminative metrics (image identification) and traditional metrics such as ROUGE end METEOR (Tables 2,3).

3. Models

Our model involves two main ingredients: a *retrieval* model that scores images caption pairs, and a caption generator that maps an image to a caption. We describe the models used in our experiments below; however we note that our approach is very modular, and can be applied to different retrieval models and/or different caption generators. Then we describe the key element of our approach: combining these two ingredients in a collaborative framework. We use the retrieval score derived from the retrieval model to help guide training of the generator.

3.1. Retrieval model

The retrieval model we use is taken from [43]. It is a embedding network which embeds both text and image into a shared semantic space in which a similarity (compatibility) score can be calculated between a caption and an image. We outline the model below, for details see [43].

We start with an image I and caption c. First, domain-specific encoders compute an image feature vector $\boldsymbol{\phi}(I)$, e.g., using a CNN, and a caption feature vector $\boldsymbol{\psi}(c)$, e.g. using an RNN-based text encoder. These feature vectors are then projected into a joint space by W_I and W_c.

$$f(i) = \mathbf{W}_I^T \boldsymbol{\phi}(I) \tag{1}$$

$$g(c) = \mathbf{W}_c^T \boldsymbol{\psi}(c) \tag{2}$$

The similarity score between I and c is now computed as the cosine similarity in the embedding space:

$$s(I, c) = \frac{f(I) \cdot g(c)}{\|f(I)\| \|g(c)\|} \tag{3}$$

The parameters of the caption embedding $\boldsymbol{\psi}$, as well as the maps $\mathbf{W}_I$ and $\mathbf{W}_c$, are learned jointly, end-to-end, by minimizing the contrastive loss defined below. In our case, the image embedding network $\boldsymbol{\phi}$ is a pre-trained CNN and the parameters are fixed during training.

Contrastive loss is a sum of two hinge losses:

$$\begin{aligned} L_{\text{CON}}(c, I) = &\max_{c'} [\alpha + s(I, c') - s(I, c)]_+ \\ &+ \max_{I'} [\alpha + s(I', c) - s(I, c)]_+ \end{aligned} \tag{4}$$

where $[x]_+ \equiv \max(x, 0)$. The max in (4) is taken, in practice, over a batch of B images and corresponding captions. The (image,caption) pairs (I, c) are correct matches, while (I', c) and (I, c') are incorrect (e.g., c' is a caption that does not describe I). Intuitively, this loss "wants" the model to assign the matching pair (I, c) the score higher (by at least α) than the score of any mismatching pair, either (I', c) or (I, c') that can be formed from the batch. This objective can be viewed as a hard negative mining version of triplet loss [44].

3.2. Discriminability loss

The ideal way to measure discriminability is to pass it to human and get feedback from them, like in [45]. However it is rather costly and very slow to collect. Here, we propose instead to use a pre-trained retrieval model to work as a proxy for human perception. Specifically, we define the discriminability loss follows.

Suppose we have a captioning system, parameterized by a set of parameters $\boldsymbol{\theta}$, that can output conditional distribution over captions for an image, $p_c(c|I; \boldsymbol{\theta})$. Then, the objective of minimizing the discriminability loss is

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\widehat{c} \sim p(c|I;\boldsymbol{\theta})} \left[L_{\text{CON}}(\widehat{c}, I) \right] \tag{5}$$

In other words, the objective involves the same contrastive loss used to train the retrieval model. However, when training the retrieval model, the loss relies on ground truth image-caption pairs (with human-produced captions), and is back-propagated to update parameters of the retrieval model. Now, when using the loss to train caption generators, an input batch (over which the max in (4) is computed) will include pairs of images with captions that are sampled from the posterior distribution produced by a caption generator; the signal derived from the loss will be used to update parameters $\boldsymbol{\theta}$ of the generator, while holding the retrieval model fixed.

3.3. Caption generation models

We now briefly describe two caption generation models used in our experiments; both are introduced in [6] where further details can be found. Discussion on training these models with discriminability loss is deferred until Sec. 4.

FC Model The first model is a simple sequence encoder initialized with visual features. Words are represented with an embedding matrix (a vector per word). Visual features are extracted from an image using a CNN.

The caption sequence is generated by a form of LSTM model. Its output at time t depends on the previously generated word and on the context/hidden state (evolving as per LSTM update rules). At training time the word fed to the state t is the ground truth word w_{t-1}; at test time, it is the predicted word $\widehat{w}_{t-1}$. The first word is a special BOS (beginning of sentence) token. The sequence production is terminated when the special EOS token is output. The image features (mapped to the dimensions of word embeddings) serve at the initial "word" w_{-1}, fed to the state at $t = 0$.

ATTN model The main difference between the second model and the FC model is that each image is now encoded into a set of spacial features: each encodes a sub-region of the image. At each word t, the context (and thus the output) depends not only on the previous output and the internal state of the LSTM, but also a weighted average of all the spatial features. This weighted averaging of features is

called attention mechanism, and the attention weights are computed by a parametric function.

Both models provide us with a posterior distribution over sequence of words $c = (w_0, \ldots, w_T)$, factorized as

$$p(c|I;\theta) = \prod_t p(w_t|w_{t-1}, I; \boldsymbol{\theta}) \tag{6}$$

4. Learning to reward discriminability

Given a generator model, we may want to train it to minimize the discriminability loss (4). A natural approach would be to use gradient descent. Unfortunately, the loss is non-differentiable since it involves sampling captions for input images in a batch.

One way to tackle this is by the Gumbel-softmax reparametrization trick [46, 47] which has been used in image captioning and visual dialog [42, 40]. Instead, in this paper, we follow the philosophy of [20, 48, 32, 6] and treat captioning as a reinforcement learning problem. Specifically we use the REINFORCE algorithm [49]. In similar contexts, REINFORCE has been applied in [32, 48] to train sequence prediction. Here we use a variant of "REINFORCE with baseline" algorithm proposed in the "self-critical" approach of [6], as outlined below.

The objective is to learn parameters $\boldsymbol{\theta}$ of the policy (here defining a mapping from I to c, i.e., p) that would maximize the reward computed by function $R(c, I)$. The algorithm computes an update to approximate the gradient of the expected reward (a function of stochastic policy parameters), known as the policy gradient:

$$\nabla_{\boldsymbol{\theta}} E_{\widehat{c}\sim p(c|I;\boldsymbol{\theta})}[R(\widehat{c}, I)] \approx (R(\widehat{c}, I) - b)\nabla_{\boldsymbol{\theta}} \log p(\widehat{c}|I;\boldsymbol{\theta}) \tag{7}$$

Here $\widehat{c}$ represents the caption sampled from (6). The *baseline* b is computed by a function designed to make it independent of the sample (leading to variance reduction without increasing bias [50]). In our case, folllowing [6], the baseline is the value of the reward $R(c^*, I)$ on the greedy decoding[1] output $c^* = (\text{BOS}, w_1^*, \ldots, w_T^*)$,

$$w_t^* = \underset{w}{\operatorname{argmax}}\, p(w|w_{0,\ldots,t-1}^*, I)$$

We could apply this to maximizing the reward defined simply as the negative discriminability loss $-L_{\text{CON}}(\widehat{c}, I)$. However, as observed in previous work [33], this does not yield human-friendly captions since discriminability loss will not directly hurt from influency. So we will combine the discriminability loss with other, traditional objectives in defining the reward, as described below.

[1] We also tried setting b to the reward of ground truth caption, and found no significant difference.

4.1. Training with maximum likelihood

The standard objective in training a sequence prediction model is to maximize word-level log-likelihood, which for a pair (I, c) is defined as $R_{\text{LL}}(c, I) = \log p(c|I;\boldsymbol{\theta})$. The parameters $\boldsymbol{\theta}$ here include word embedding matrix and LSTM weights which are updated as part of training, and the CNN weights, which are held fixed after pre-training on a vision task such as ImageNet classification. This reward can be directly maximized via gradient ascent (equivalent to gradient descent on the cross-entropy loss), yielding maximum likelihood estimate (MLE) of the model parameters.

Combining the log-likelihood reward with discriminability loss in the REINFORCE framework corresponds to defining the reward as

$$R(c, I) = R_{\text{LL}}(c, I) \; - \; \lambda L_{\text{CON}}(\widehat{c}, I), \tag{8}$$

yielding the policy gradient:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} E[R(c, I)] &\approx \nabla_{\boldsymbol{\theta}} R_{\text{LL}}(c, I) \\ &- \lambda \left[L_{\text{CON}}(\widehat{c}, I) - L_{\text{CON}}(c^*, I)\right] \nabla_{\boldsymbol{\theta}} \log p(\widehat{c}|I;\boldsymbol{\theta})\end{aligned} \tag{9}$$

The coefficient λ determines the tradeoff between matching human captions (expressed by the cross-entropy) and discriminative properties expressed by L_{CON}.

4.2. Training with CIDEr optimization

In our experiments, it was hard to train with the combined objective in (9). For small λ, the solutions seemed stuck in a local minimum; but increasing λ would abruptly make output less fluent.

An alternative to MLE is to train the model to maximize some other reward/score, such as BLEU or METEOR. Here if pursue optimization of the CIDEr score [2]. CIDEr measures consensus in image captions by performing a TF-IDF weighting for each n-gram, and optimizing over CIDEr can also benefit other metrics[6]. We found that in practice, the discriminability loss appears to "cooperate" better with CIDEr than with log-likelihood; we also observed better performance, across many metrics, on validation set as described in Section 5.

Compared to [6], which uses CIDEr as reward function, the difference here is we use a weighted sum of cider score and discriminability loss.

$$\nabla_\theta E[R(\widehat{c}, I)] \approx (R(\widehat{c}, I) - R(c^*, I))\nabla_\theta \log p(\hat{c}|I;\boldsymbol{\theta}), \tag{10}$$

where the reward is the combination

$$R(\widehat{c}, I) = \text{CIDEr}(\widehat{c}) - \lambda L_{\text{CON}}(\widehat{c}, I), \tag{11}$$

with λ again representing the relative weight of discriminability loss vs. CIDEr.

5. Experiments and results

The main goal of our experiments is to evaluate the utility of the proposed discriminability objective in training image captions. Recall that our motivation for introducing this objective is two-fold: to make the captions more discriminative, and to improve caption quality in general (with the implied assumption that expected discriminability is part of the unobservable human "objective" in describing images).

Dataset. We train and evaluate our model on COCO dataset [51]. To enable direct comparisons, we use the data split from [5], which includes 113,287 images for training, 5,000 images for validation, and another 5,000 held out for test. Each image is associated with five human captions.

5.1. Implementation details

As the basis for caption generators, we used two models described in Section 3, FC and ATTN, with relevant implementation details as follows.

For image encoder in retrieval and FC captioning model, we used a pretrained Resnet-101 [52]. For each image, we take the global average pooling of the final convolutional layer output, which results in a vector of dimension 2048. The spatial features are extracted from output of a Faster R-CNN[52, 14] with ResNet-101[53], trained by object and attribute annotations from Visual Genome[54]. The number of spatial features varies from image to image. Each feature encodes a region in the image which is proposed by region proposal network. Both the FC features and Spatial features are pre-extracted, and no finetuning is applied on image encoders. For captioning models, the dimension of LSTM hidden state, image feature embedding, and word embedding are all set to 512.

The retrieval model uses GRU-RNN to encode text, and the FC features above as the image feature. The word embedding has 300 dimensions and the GRU hidden state size and joint embedding size are 1024. The margin α is set to 0.2, as suggested by [43].

Training All of our captioning models are trained according to the following scheme. We first pretrain the captioning model using MLE, with Adam[55]. After 40 epochs, the model is switched to self-critical training with appropriate reward (CIDEr alone or CIDEr combined with discriminability) and continue for another 20 epochs. For fair comparison, we also train another 20 epochs for MLE-only models.

For both retrieval and captioning models, the batch size is set to 128 images. The learning rate is initialized to be 5e-4 and decay by a factor 0.8 for every three epochs.

During test time, we apply beam search to sample captions from captioning model. The beam size is set to 2.

5.2. Experiment design

We consider a variety of possible combination of captioning objective (MLE/CIDEr), captioning model (FC/ATTN), and inclusion/exclusion of discriminability, abbreviating the model references for brevity, so, e.g., ATTN+CIDER+DISC(5) corresponds to fine-tuning the attention model with a combination of CIDEr and discriminability loss, with $\lambda = 5$.

Evaluation metrics Our experiments consider two families of metrics. The first family of standard metrics that have been proposed for caption evaluation, mostly based on comparing generated captions to human ones, includes BLEU[1], METEOR[56], ROUGE[57], CIDEr[2] and SPICE[3].

The second set of metrics directly assesses how discriminative the captions are. This includes automatic assessment, by measuring accuracy of the trained retrieval model on generated captions.

We also assess how discriminative the generated captions are when presented to humans. To measure this, we conducted an image discrimination task on Amazon Mechanical Turk (AMT), following the protocol in [5]. A single task (HIT) involves displaying, along with a caption, a pair of images (in randomized order). One image is the *target* for which the caption was generated, and the second is a *distractor* image. The worker is asked to select which image is more likely to match the caption. Each target/distractor pair is presented to five distinct workers; we report the fraction of HITs with correct selection by at least k out of five workers, with $k = 3, 4, 5$. Note that $k = 3$ suffers from highest variance since the forced choice nature of the task would produce non-trivial chance of 3/5 correct selections when the caption is random. In our opinion, $k = 4$ is the most reliable indicator of human ability to discriminate based on the caption.

The test set used for this evaluation is the set from [5], constructed as follows. For each image in the original test set, its nearest neighbor is found based on visual similarity, estimated as Euclidean distance between the FC7 feature vectors computed by VGG-16 pre-trained on ImageNet [58]. Then a captioning model is run on the nearest neighbor images, and the word-level overlap (intersection over union) of the generated captions is used to and pick (out of 5000 pairs) the top (highest overlap) 1000 pairs.

For preliminary evaluation, we followed a similar protocol to construct our own validation set of target/distractor pairs; both the target images and distractor images were taken from the caption validation set (and so were never seen by any training procedure).

5.3. Retrieval model quality

Before proceeding with main experiments, we report in Tab. 1 the accuracy of the retrieval model on validation set,

with human-generated captions. This is relevant since we rely on this model as a proxy for discriminability in our training procedure. While this model does not achieve state of the art for image caption retrieval, it is good enough for providing training signal to improve caption results.

	R@ 1	R@ 5	R@ 10	Med r	Mean r
	Caption Retrieval				
1k val	63.9	90.4	95.9	1.0	2.9
5k val	38.0	68.9	81.1	2.0	10.4
	Image Retrieval				
1k val	47.9	80.7	89.9	2.0	7.7
5k val	26.1	54.7	67.5	4.0	34.6

Table 1. Retrieval model performance on validation set.

5.4. Captioning performance

In Table 2, we show the results on validation set with a variety of model/loss settings. Note that all the FC*/ATTN* model with different settings are finetuned from the same model pre-trained with MLE. The results in the table for the machine scores are based on all the 5k images. For discriminability, we randomly select a subset of 300 image pairs from validation set. We can draw a number of conclusions from these results.

Effectiveness of reinforcement learning. In the first column, we report the retrieval accuracy (Acc, % of pairs in which the model correctly selects the target vs. distractor) on pairs given the output of the captioning model. Training with the discriminability loss produces higher values here, meaning that our captions are more discriminative *to the retrieval model*, as intended. As a control experiment, we also report the accuracy (Acc-new) obtained by a same architecture but separately trained retrieval model, not used in training caption generators. Acc and Acc-new are very similar for all models, showing that our model does not overfit to the retrieval model it uses during training time.

Human discrimination. More importantly, we observe that incorporating discriminability in training yields captions that are more discriminative to humans, with higher λ leading to better human accuracy.

Improved caption quality. We also see that, as hoped, incorporating discriminability indeed improves caption quality as measured by a range of metrics that are not tied to discrimination, such as BLEU etc. Even the CIDEr scores are improved when adding discriminability to the CIDEr optimization objective with moderate λ. This is somewhat surprising since the addition of L_{CON} could be expected to detract from the original objective of maximizing CIDEr; we presume that the improvement is due to the additional objective "nudging" the RL process and helping it escape less optimal solutions.

Model/loss selection While discriminability loss works for both ATTN model and FC model, and with both MLE and CIDEr learning, to make captions more discriminative, and with mild λ to improve other metrics, the overall performance analysis favors ATTN+CIDEr combination. We also note that ATTN is better than FC on discriminability metrics even when trained without L_{CON}, but the gains are less significant than in automatic metrics.

Effect of λ As stated above, mild $\lambda = 1$, combined with ATTN+CIDER, appear to yield the optimal tradeoff, improving measures of discriminative *and* descriptive quality across the board. Higher values of λ do make resulting captions more discriminative to both humans and machines, but at the cost of reduction in other metrics, and in our observations (see Section 5.5) in perceived fluency. This analysis is applicable across model/loss combinations. We also notice a relative large range of λ(0.5-1.2) can yield similar improvement on automatic metrics.

Following the observations above, we select a subset of methods to evaluate on the (previously untouched) test set, with results shown in Table 3.

Here, we add two more results for comparison. The first involves presenting AMT workers with *human* captions for the target images. Recall that these captions are collected for each image independently, without explicit instructions related to discriminability, and without showing potential distractors. However, human captions prove to be highly discriminative. This, not surprisingly, indicates that humans may be incorporating an implicit objective of describing elements in an image that are surprising, notable or otherwise may help in distinguishing the scene from other, scenes. While this performance is not perfect (4/5 accuracy of 82%) it is much higher than for any automatic caption model.

The second additional set of results is for the model in [5], evaluated on captions provided by the authors. Note that in contrast to our model (and to human captions), this method has the benefit of seeing the distractor prior to generating the caption; nonetheless, its performance is dominated across metrics by our attention models trained with CIDEr optimization combined with discriminability loss. It also appears that this models' gains on discrimination are offset by a significant deterioration under other metrics.

In contrast, our ATTN+CIDER+DISC model with $\lambda = 10$ achieves the most discriminative image captioning result without major degradation under other metrics; the ATTN+CIDER+DISC with $\lambda = 1$ again shows the best discriminability/descriptiveness tradeoff among the evaluated models.

Effect on SPICE score To further break down how our discriminabilty loss does, we analyze the affect of different models on the SPICE score [2]. It estimates caption quality by transforming both candidate and reference (human) captions into a scene graph and computing the matching between the graphs. SPICE is known to have higher correlation with human ratings than other conventional met-

	Acc	Acc-new	BLEU4	METEOR	ROUGE	CIDEr	SPICE	3 in 5	4 in 5	5 in 5
FC+MLE[6]	77.23%	77.23%	**0.3308**	0.2566	0.5407	1.0005	0.1855	71.78%	50.28%	18.79%
FC+CIDER[6]	74.00%	74.32%	0.3249	0.2550	0.5428	1.0154	0.1899	73.04%	50.58%	24.83%
FC+MSE+DISC (100)	87.42%	87.42%	0.2902	0.2523	0.5261	0.9190	0.1881	76.91%	54.62%	23.20%
FC+CIDER+DISC (1)	79.26%	79.49%	0.3274	**0.2574**	**0.5457**	**1.0231**	**0.1939**	74.26%	55.53%	24.13%
FC+CIDER+DISC (5)	85.90%	85.68%	0.3072	0.2534	0.5382	0.9678	0.1904	78.63%	58.03%	32.64%
FC+CIDER+DISC (10)	**88.69%**	**88.01%**	0.2727	0.2473	0.5224	0.8795	0.1807	**80.01%**	**62.71%**	**37.15%**
ATTN+MLE[6]	72.40%	73.12%	0.3582	0.2719	0.5649	1.1078	0.2019	69.90%	54.60%	28.07%
ATTN+CIDER[6]	71.05%	71.13%	0.3592	0.2695	0.5678	1.1332	0.2083	69.97%	51.34%	27.34%
ATTN+MLE+DISC (100)	82.64%	83.03%	0.3266	0.2697	0.5542	1.0448	0.2057	78.18%	55.63%	21.71%
ATTN+CIDER+DISC (1)	75.74%	76.60%	**0.3627**	**0.2728**	**0.5706**	**1.1406**	**0.2113**	72.70%	53.23%	34.33%
ATTN+CIDER+DISC (5)	80.98%	81.43%	0.3504	0.2704	0.5636	1.1026	0.2097	76.69%	60.94%	33.49%
ATTN+CIDER+DISC (10)	**83.69%**	**83.50%**	0.3261	0.2673	0.5549	1.0552	0.2070	**81.93%**	**65.12%**	**35.41%**

Table 2. Automatic scores and human-study discriminability on the validation set. The numbers in the parenthesis are discriminability loss weight λ.

	Acc	Acc-new	BLEU4	METEOR	ROUGE	CIDEr	SPICE	3 in 5	4 in 5	5 in 5
Human	74.30%	74.14%	-	-	-	-	-	91.14%	82.38%	57.08%
ATTN+MLE[6]	68.60%	66.90%	0.3907	0.2913	0.5956	1.2198	0.2132	72.06%	59.06%	44.25%
ATTN+CIDER[6]	68.19%	65.12%	0.3871	0.2908	0.5971	1.2604	0.2260	70.07%	55.95%	35.95%
CACA[5]	75.80%	76.00%	0.2357	0.2186	0.4719	0.7656	0.1526	74.1%[1]	56.88%[1]	35.19%[1]
ATTN+CIDER+DISC(1)	72.63%	70.68%	**0.3971**	**0.2931**	**0.6043**	**1.2770**	**0.2302**	76.91%	61.67%	40.09%
ATTN+CIDER+DISC(10)	**79.75%**	**79.14%**	0.3538	0.2821	0.5811	1.1429	0.2204	**77.70%**	**64.63%**	**44.63%**

Table 3. Automatic scores and discriminability on 1k test set.

rics. Furthermore, it provides subclass scores on Color, Attributes, Cardinality, Object, Relation, Size. We report the results on these (on validation set) in detail for different models in Table 4.

By adding the discriminability loss, we improve scores on Color, Attribute, and Cardinality. With the latter, qualitative results suggest that the improvement may be due to a refined ability to distinguish "one" or "two" from "group of" or "many". With small λ, we can also get the best score on Object. Since the object score is dominant in SPICE, $\lambda = 1$ also obtaines highest SPICE score overall in Tables 2,3.

Finally, we can evaluate the diversity in captions generated by different models. We find that including discriminability objective, and using higher λ, are correlated with captions that are more diverse (4471 distinct captions with ATTN+CIDER+DISC(10) for the 5000 images in validation set, compared to 2640 with ATTN+CIDER) and slightly longer (avg. length 9.84 with ATTN+CIDER+DISC(10) vs. 9.20 with ATTN+CIDER). Detailed analysis can be found in the supplementary material.

5.5. Qualitative result

In Figures 1,2, we show a sample of validation set images and for each include a human caption, the caption

	Color	Attribute	Cardinality	Object	Relation	Size
FC+MLE	9.32	8.74	1.73	34.04	4.81	**2.74**
FC+MLE+D(100)	**15.85**	**10.31**	5.33	34.57	4.43	2.62
FC+C	5.77	7.01	1.80	35.70	5.17	1.70
FC+C+D (1)	8.28	7.81	3.45	**36.37**	**5.25**	2.10
FC+C+D (5)	10.87	9.11	6.72	35.58	4.75	2.08
FC+C+D (10)	12.80	9.90	**8.50**	34.60	4.40	1.70
ATTN+MLE	11.78	10.13	3.00	36.42	5.52	3.67
ATTN+MLE+D(100)	**15.80**	**11.83**	14.30	37.16	5.13	**3.97**
ATTN+C	7.24	8.77	8.93	38.38	**6.21**	2.39
ATTN+C+D (1)	9.25	9.49	10.51	**38.96**	5.91	2.58
ATTN+C+D (5)	11.99	10.40	15.23	38.57	5.59	2.53
ATTN+C+D (10)	12.88	10.88	**15.72**	38.09	5.35	2.53

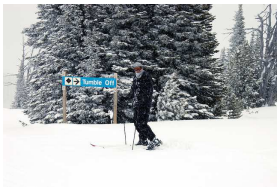
Table 4. SPICE subclass scores on 5k validation set. All the scores here are scaled up by 100. (Here +C means using CIDEr optimzation; +D(x) means using discriminability loss with λ being x).

generated by ATTN+CIDER, and our captions produced by ATTN+CIDER+DISC(1). To emphasize the discriminability gap, the images are organized in pairs[2], where both images have the same ATTN+CIDER caption; although these are mostly correct, compared to our and human result, they tend to lack discriminative specificity.

To illustrate the task on which we base our evaluation of discriminability to humans, we show in Figure 3 a sample of image pairs and associated captions. In each case, the target is on the left (in AMT experiments the order was

[1] 3 in 5 is quoted from [5]; 4 in 5 and 5 in 5 computed by us on the set of captions provided by the authors.

[2] Note that these pairs are formed for the purpose of this figure; these are not pairs shown to AMT workers for human evaluation.



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

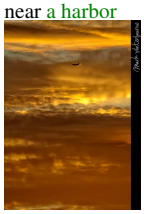
Human: a plate topped with meat and vegetables and sauce
ATTN+CIDER: a plate of food with meat and vegetables on a table
Ours: a plate of food with carrots and vegetables on a plate

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

Figure 2. Examples of image captions; Ours refers to ATTN+CIDER+DISC(1)

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



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

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

Figure 3. Captions from different models describing the target images(left). Right images are the corresponding distractors selected in val/test set; these pairs were included in AMT experiments.

randomized), and we show captions produced by four automtic systems, two without added discriminability objective in training, and two with (with low and high λ, respectively). Again, we can see that discriminability loss encourages learning to produce more discriminative captions, and that with higher λ this may be associated with reduced fluency. We highlight in green caption elements that (subjectively) seem to aid discriminability, and in red the portions that seem incorrect or jarringly non-fluent. For additional experimental results, see supplementary material.

6. Conclusions

We have demonstrated that incorporating a discriminability loss, derived from the loss of a trained image/caption retrieval model, in training image caption generators improves the quality of resulting captions across a variety of properties and metrics. It does, as expected, lead to captions that are more discriminative, allowing both human recipients and machines to better identify an image being described, and thus arguably conveying more valuable information about the images. More surprisingly, it also yields captions that are scored higher on metrics not directly related to discrimination, such as BLEU/METEOR/ROUGE/CIDEr as well as SPICE, reflecting more descriptive captions. This suggests that richer, more diverse sources of training signal may further improve training of caption generators.

In future work, we plan to explore more sophisticated visual semantic embedding model, which could potentially give better guidance to training than our current retrieval model. We are also interested in how to make it even more discriminative.

Acknowledgments This work was partially supported by NSF award 1409837 and by the ONR award to MIT Lincoln Lab FA8721-05-C-0002.

References

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 1, 5

[2] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1, 4, 5, 6

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 1, 5

[4] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 1, 2

[5] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. *arXiv preprint arXiv:1701.02870*, 2017. 1, 2, 5, 6, 7

[6] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016. 1, 2, 3, 4, 7

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 2

[8] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3128–3137, 2015. 2

[9] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. *Proceedings of the IEEE International Conference on Computer Vision*, 11-18-Dece:2623–2631, 2016. 2

[10] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Baidu Research, and Alan Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *To appear: ICLR-2015*, 1090(2014):1–14, 2015. 2

[11] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Icml-2015*, 2015. 2

[12] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2016. 2

[13] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. Review networks for caption generation. In *Advances in Neural Information Processing Systems*, pages 2361–2369, 2016. 2

[14] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017. 2, 5

[15] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. *arXiv preprint arXiv:1709.03376*, 2017. 2

[16] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. *arXiv preprint arXiv:1704.06972*, 2017. 2

[17] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention Correctness in Neural Image Captioning. pages 1–11, 2016. 2

[18] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. *arXiv preprint arXiv:1611.01646*, 2016. 2

[19] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5758–5768, 2017. 2

[20] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence Level Training with Recurrent Neural Networks. *Iclr*, pages 1–15, 2016. 2, 4

[21] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. 2

[22] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017. 2

[23] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2013. 2

[24] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning Deep Structure-Preserving Image-Text Embeddings. *Cvpr*, (Figure 1):5005–5013, 2016. 2

[25] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI*, 2011. 2

[26] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *arXiv preprint arXiv:1704.03470*, 2017. 2

[27] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016. 2

[28] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-Embeddings of Images and Language. *arXiv preprint*, (2005):1–13, 2015. 2

[29] Jacob Andreas and Dan Klein. Reasoning About Pragmatics with Neural Listeners and Speakers. *1604.00562V1*, 2016. 2

[30] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling Context in Referring Expressions. In *Eccv*, 2016. 2

[31] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. *Cvpr*, pages 11–20, 2016. 2

[32] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. *arXiv preprint arXiv:1612.09542*, 2016. 2, 4

[33] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. *arXiv preprint arXiv:1701.03439*, 2017. 2, 4

[34] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. 2

[35] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *arXiv preprint arXiv:1611.08669*, 2016. 2

[36] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*, 2017. 2

[37] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. *arXiv preprint arXiv:1611.08481*, 2016. 2

[38] Yining Li, Chen Huang, Xiaoou Tang, and Chen-Change Loy. Learning to disambiguate by asking discriminative questions. *arXiv preprint arXiv:1708.02760*, 2017. 2

[39] Bo Dai and Dahua Lin. Contrastive learning for image captioning. *arXiv preprint arXiv:1710.02534*, 2017. 2

[40] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. *arXiv preprint arXiv:1706.01554*, 2017. 2, 4

[41] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*, 2017. 2

[42] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *arXiv preprint arXiv:1703.10476*, 2017. 2, 4

[43] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017. 3, 5

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 3

[45] Huan Ling and Sanja Fidler. Teaching machines to describe images via natural language feedback. In *NIPS*, 2017. 3

[46] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[47] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 4

[48] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 4

[49] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 4

[50] Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 2001. 4

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 5

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[54] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 5

[55] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[56] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5

[57] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004. 5

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5