

Published as a conference paper in International Conference on Computer Vision (ICCV) 2017 Speaking the Same Language : Matching Machine to Human Captions by Adversarial Training Rakshit Shekhar 1 Marcus Rohrbach 2, 3 arXiv : 1703.10476 v2 [cs.CV] 6 Nov 2017 Mario Fritz 1 1 Lisa Anne Hendricks 2 Bernt Schiele 1 Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany 2 3 UC Berkeley EECS, CA, United States Facebook AI Research Abstract While strong progress has been made in image captioning recently, machine and human captions are still quite distinct. This is primarily due to the deficiencies in the generated word distribution, vocabulary size, and strong bias in the generators towards frequent captions. Furthermore, humans rightfully so generate multiple, diverse captions, due to the inherent ambiguity in the captioning task which is not explicitly considered in today's systems. To address these challenges, we change the training objective of the caption generator from reproducing ground truth captions to generating a set of captions that is indistinguishable from human-written captions. Instead of handcrafting such a learning target, we employ adversarial training in combination with an approximate Gumbel sampler to implicitly match the generated distribution to the human one. While our method achieves comparable performance to the state-of-the-art in terms of the correctness of the captions, we generate a set of diverse captions that are significantly less biased and better match the global uni-, bi- and tri-gram distributions of the human captions. Ours : a person on skis is 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 Figure 1 : Four images from the test set, all related to skiing, shown with captions from our adversarial model and a baseline. Baseline model describes all four images with one generic caption, whereas our model produces diverse and more image specific captions. As we analyze in this paper, this is likely due to artifacts and deficiencies in the statistics of the generated captions, which is more apparent when observing multiple samples. Specifically, we observe that state-of-the-art systems frequently reveal themselves by generating a different word distribution and using smaller vocabulary. Further scrutiny reveals that generalization from the training set is still challenging and generation is biased to frequent fragments and captions. Also, today's systems are evaluated to produce a single caption. Yet, multiple potentially distinct captions are typically correct for a single image—a property that is reflected in human ground-truth. This diversity is not equally reproduced by state-of-the-art caption generators [40, 23]. Therefore, our goal is to make image captions less distinguishable from human ones—similar in the spirit to a Turing test. Introduction Image captioning systems have a variety of applications ranging from media retrieval and tagging to assistance for the visually impaired. In particular, models which combine state-of-the-art image representations based on deep convolutional networks and deep recurrent language models have led to ever increasing performance on evaluation metrics such as CIDEr [39] and METEOR [8] as can be seen e.g. on the COCO Image Caption challenge leaderboard [6]. Despite these advances, it is often easy for humans to differentiate between machine and human captions, particularly when observing multiple captions for a single image. 1. 2. Related Work a bus that has pulled into the side of the street a bus is parked at the side of the road a white bus is parked near a curb with people walking by a group of people standing outside in a old museum an airplane show where people stand around a line of planes parked at an airport show Base a bus is parked on the side of line the road a bus that is parked in the street a bus is parked in the street next to a bus a group of people standing around a plane a group of people standing around a plane a group of people standing around a plane Ours Figure 2 : Two examples comparing multiple captions generated by our adversarial model and the baseline. Bi-grams which are top-20 frequent bi-grams in the training set are marked in red (e.g., a group and group of). Captions which are replicas from training set are marked with . Test We also embrace the ambiguity of the task and extend our investigation to predicting sets of captions for a single image and evaluating their quality, particularly in terms of the diversity in the generated set. In contrast, popular approaches to image captioning are trained with an objective to reproduce the captions as provided by the ground-truth. Instead of relying on handcrafting loss-functions to achieve our goal, we propose an adversarial training mechanism for image captioning. For this we build on Generative Adversarial Networks (GANs) [14], which have been successfully used to generate mainly continuous data distributions such as images [9, 30], although exceptions exist [27]. In contrast to images, captions are discrete, which poses a challenge when trying to backpropagate through the generation step. To overcome this obstacle, we use a Gumbel sampler [20, 28] that allows for end-to-end training. We address the problem of caption set generation for images and discuss metrics to measure the caption diversity and compare it to human ground-truth. We contribute a novel solution to this problem using an adversarial formulation. The evaluation of our model shows that accuracy of generated captions is on par to the state-of-the-art, but we greatly increase the diversity of the caption sets and better match the ground-truth statistics in several measures. Qualitatively, our model produces more diverse captions across images containing similar content (Figure 1) and when sampling multiple captions for an image (see supplementary). 1. 1 https://goo.gl/3yRVnq Image Description. Early captioning models rely on first recognizing visual elements, such as objects, attributes, and activities, and then generating a sentence using language models such as a template model [13], n-gram model [22], or statistical machine translation [34]. Advances in deep learning have led to end-to-end trainable models that combine deep convolutional networks to extract visual features and recurrent networks to generate sentences [11, 41, 21]. Though modern description models are capable of producing coherent sentences which accurately describe an image, they tend to produce generic sentences which are replicated from the train set [10]. Furthermore, an image can correspond to many valid descriptions. However, at test time, sentences generated with methods such as beam search are generally very similar. [40, 23] focus on increasing sentence diversity by integrating a diversity promoting heuristic into beam search. [42] attempts to increase the diversity in caption generation by training an ensemble of caption generators each specializing in different portions of the training set. In contrast, we focus on improving diversity of generated captions using a single model. Our method achieves this by learning a corresponding model using a different training loss as opposed to after training has completed. We note that generating diverse sentences is also a challenge in visual question generation, see concurrent work [19], and in language-only dialogue generation studied in the linguistic community, see e.g. [23, 24]. When training recurrent description models, the most common method is to predict a word w_t conditioned on an image and all previous ground truth words. At test time, each word is predicted conditioned on an image and previously predicted words. Consequently, at test time predicted words may be conditioned on words that were incorrectly predicted by the model. By only training on ground truth words, the model suffers from exposure bias [31] and cannot effectively learn to recover when it predicts an incorrect word during training. To avoid this, [4] proposes a scheduled sampling training scheme which begins by training with ground truth words, but then slowly conditions generated words on words previously produced by the model. However, [17] shows that the scheduled sampling algorithm is inconsistent and the optimal solution under this objective does not converge to the true data distribution. Taking a different direction, [31] proposes to address the exposure bias by gradually mixing a sequence level loss (BLEU score) using REINFORCE rule with the standard maximum likelihood training. Several other works have followed this up with using reinforcement learning based approaches to directly optimize the evaluation metrics like BLEU, METEOR and CIDEr [33, 25]. However, optimizing the evaluation metrics does not directly address the diversity of the generated captions. Since all current evaluation metrics use n-gram matching to score the captions, captions using more frequent n-grams are likely to achieve better scores than ones using rare and more diverse n-grams. In this work, we formulate our caption generator as a generative adversarial network. We design a discriminator that explicitly encourages generated captions to be diverse and indistinguishable from human captions. The generator is trained with an adversarial loss with this discriminator. Consequently, our model generates captions that better reflect the way humans describe images while maintaining similar correctness as determined by a human evaluation. Generative Adversarial Networks. The Generative Adversarial Networks (GANs) [14] framework learns generative models without explicitly defining a loss from a target distribution. Instead, GANs learn a generator using a loss from a discriminator which tries to differentiate real and generated samples, where the generated samples come from the generator. When training to generate real images, i/si