¡s¿ Published as a conference paper in International Conference of Computer Vision (IC CV) 2017 Speaking the Same Language : Match ing Machine to Human Capt ions by Ad vers arial Training Rak sh ith She tty 1 Marcus Roh r bach 2 , 3 ar X iv : 17 03 . 10 476 v 2 [cs . CV] 6 Nov 2017 Mario Fritz 1 1 Lisa Anne Hendricks 2 Ber nt S chie le 1 Max Plan ck Institute for In format ics , Sa ar land In format ics Campus , Sa arb ru ck en , Germany 2 3 UC Berkeley E EC S , CA , United States Facebook AI Research Abstract While strong progress has been made in image caption ing recently , machine and human capt ions 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 capt ions . Furthermore , humans rightfully so generate multiple , diverse capt ions , due to the inherent ambiguity in the caption ing task which is not explicitly considered in today s systems . To address these challenges , we change the training objective of the caption generator from reprodu cing ground truth capt ions to generating a set of capt ions that is indistinguishable from human written capt ions . Instead of hand craft ing such a learning target , we employ advers arial training in combination with an approximate G umb el sam pler 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 capt ions , we generate a set of diverse capt ions that are significantly less biased and better match the global un i -, bi - and tri - gram distributions of the human capt ions . O urs : a person on sk is jumping over a ramp O urs : a sk ier is making a turn on a course O urs : a cross country sk ier makes his way through the snow O urs : a sk ier is headed down a steep slope Bas eline : a man riding sk is down a snow covered slope Figure 1 : Four images from the test set , all related to skiing , shown with capt ions from our advers arial model and a baseline . Bas eline model describes all four images with one generic caption , whereas our model produces diverse and more image specific capt ions . As we analyze in this paper , this is likely due to artifacts and deficiencies in the statistics of the generated capt ions , which is more apparent when observing multiple samples . Specifically , we observe that state - of - the - art systems frequently reve al themselves by generating a different word distribution and using smaller vocabulary . Further scrutiny reveals that general ization from the training set is still challenging and generation is biased to frequent fragments and capt ions . Also , today s systems are evaluated to produce a single caption . Yet , multiple potentially distinct capt ions 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 capt ions less distinguish able from human ones similar in the spirit to a Turing 1 . Introduction Image caption ing 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 conv olution al networks and deep recurrent language models have led to ever increasing performance on evaluation metrics such as C ID Er [39] and MET E OR [8] as can be seen e . g . on the C OC O image Caption challenge leader board [6]. Despite these advances , it is often easy for humans to differentiate between machine and human capt ions particularly when observing multiple capt ions 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 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 O urs Figure 2 : Two examples comparing multiple capt ions generated by our advers arial model and the baseline . Bi - gram s which are top - 20 frequent bi - gram s in the training set are marked in red (e . g ., a group and group of). Capt ions which are repl icas from training set are marked with . Test . We also embrace the ambiguity of the task and extend our investigation to predicting sets of capt ions for a single image and evaluating their quality , particularly in terms of the diversity in the generated set . In contrast , popular approaches to image caption ing are trained with an objective to reproduce the capt ions as provided by the ground - truth . Instead of relying on hand craft ing loss - fun ctions to achieve our goal , we propose an advers arial training mechanism for image caption ing . For this we build on Gener ative Ad vers arial Networks (GAN s) [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 , capt ions are discrete , which poses a challenge when trying to back prop agate through the generation step . To overcome this obstacle , we use a G umb el sam pler [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 advers arial formulation . The evaluation of our model shows that accuracy of generated capt ions 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 . Qual itatively , our model produces more diverse capt ions across images containing similar content (Figure 1) and when sampling multiple capt ions for an image (see supplementary) 1 . 1 https :// goo . gl / 3 y R V n q Image Description . Early caption ing 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]. Adv ances in deep learning have led to end - to - end train able models that combine deep conv olution al 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 he uristic 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 capt ions 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 (BLE U score) using RE IN FOR CE 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 B LE U , MET E OR and C IDER [33 , 25]. However , optimizing the evaluation metrics does not directly address the diversity of the generated capt ions . Since all current evaluation metrics use n - gram matching to score the capt ions , capt ions using more frequent n - gram s are likely to achieve better scores than ones using rare r and more diverse n - gram s . In this work , we formulate our caption generator as a gener ative advers arial network . We design a discrim inator that explicitly encourages generated capt ions to be diverse and indistinguishable from human capt ions . The generator is trained with an advers arial loss with this discrim inator . Consequently , our model generates capt ions that better reflect the way humans describe images while maintaining similar correctness as determined by a human evaluation . Gener ative Ad vers arial Networks . The Gener ative Ad vers arial Networks (GAN s) [14] framework learns gener ative models without explicitly defining a loss from a target distribution . Instead , G AN s learn a generator using a loss from a discrim inator which tries to differentiate real and generated samples , where the generated samples come from the generator . When training to generate real images , ¡/s¿