Self-Critical Sequence Training for Image Captioning using Bayesian "baseline"

Shashank Bujimalla*
Intel Corporation
shashankbvs@gmail.com

Mahesh Subedar*
Intel Labs
mahesh.subedar@intel.com

Omesh Tickoo
Intel Labs
omesh.tickoo@intel.com

Abstract

Bayesian deep neural networks (DNN) provide a mathematically grounded framework to quantify uncertainty in their predictions. We propose a Bayesian variant of policy-gradient based reinforcement learning training technique for image captioning models to directly optimize non-differentiable image captioning quality metrics such as CIDEr-D. We extend the well-known Self-Critical Sequence Training (SCST) approach for image captioning models by incorporating Bayesian inference, and refer to it as B-SCST. The "baseline" reward for the policy-gradients in B-SCST is generated by averaging predictive quality metrics (CIDEr-D) of the captions drawn from the distribution obtained using a Bayesian DNN model. We infer this predictive distribution using Monte Carlo (MC) dropout approximate variational inference. We show that B-SCST improves CIDEr-D scores on Flickr30k, MS COCO and VizWiz image captioning datasets, compared to the SCST approach.

Introduction:
The training of the state-of-the-art image captioning models typically follows a two-step process. In the first step, cross-entropy loss is optimized to generate the captions with words in the same order as the ground-truth captions. However, the quality of image captioning is evaluated using Natural Language Processing (NLP) metric scores [6], which are non-differentiable and cannot be directly maximized by this optimization algorithm. So, in the second step, policy-gradient based reinforcement learning (RL) is used to minimize the negative expected value of the caption scores. Several recent works have shown that using a bias correction, i.e., a learned "baseline" score, to normalize the RL rewards reduces the variance in policy gradients, and is effective during training. In Self-Critical Sequence Training (SCST) [5], the model chooses word with the highest SoftMax probability at each timestep and generates a greedy caption during training phase. The CIDEr-D score of this greedy caption, called the greedy score, is used as "baseline" during the search process. When a policy-gradient RL loss function using this "baseline" is applied on the model, it tends to increase the probability of generating captions that have higher score than the greedy score while decreasing it for captions that have lower score, thus optimizing the CIDEr-D metric directly.

Although deep neural networks (DNNs) provide state-of-the-art results, they have been shown to fail in the case of noisy or out-of-distribution data leading to overly confident SoftMax probability scores. Probabilistic Bayesian models provide a principled way to gain insight into the data and capture reliable uncertainty estimates in their predictions, hence providing interpretable models. Bayesian DNNs provide a convenient way to develop more robust models that can be scaled to large datasets and real-world applications.

Our main contribution in this work is to propose a Bayesian variant of SCST approach (B-SCST) and demonstrate that it improves the caption quality score compared to SCST.

Bayesian Self-Critical Sequence Training:
Image captioning DNNs [1] use attention mechanism so that the encoder and decoder in the model attend on appropriate features in the image to generate the words in the caption. Attention-on-attention network (AoANet) [4] uses an extra learned attention on top of self-attention to avoid attentions that are irrelevant to the decoder. In this work, we use this AoANet architecture.

We train our model using the two step approach. In the first step, we minimize word-level cross-entropy loss function [4]. In the second step, we directly optimize the CIDEr-D metric. The goal of CIDEr-D optimization [5] is to minimize the negative expected CIDEr-D rewards function, denoted by $r(.)$ and the gradient of this loss is approximated using:

$$\nabla_\theta L_{RL}(\theta) \approx -(r(y^s_{1:T}) - r(\hat{y}_{1:T}))\nabla_\theta \log p_\theta(y^s_{1:T}) \quad (1)$$

Here, $y^s_{1:T}$ and $\hat{y}_{1:T}$ are the sampled caption generated by sampling words from the decoder's output SoftMax distribution and the greedy caption generated by choosing the word with highest SoftMax probability at each time step,

[*] Equal Contribution

		Cross-entropy loss training						CIDEr-D optimization					
Dataset	Model	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
Flickr30k	SCST	69.6	28.0	22.2	48.8	58.5	16.4	72.2	30.0	22.1	50.0	64.6	16.3
	B-SCST	69.6	28.0	22.2	48.8	58.5	16.4	71.9	29.6	22.6	50.2	**66.9**	16.7
MS COCO	SCST* [4]	77.3	36.9	28.5	57.3	118.4	21.7	80.5	39.1	29.0	58.9	128.9	2.7
	B-SCST	77.3	36.9	28.5	57.3	118.4	21.7	80.8	39.0	29.2	59.0	**131.0**	22.9
VizWiz	SCST* [3]	-	-	-	-	-	-	66.0	23.7	20.1	46.8	60.9	15.3
	B-SCST	64.7	22.7	19.4	45.0	59.0	14.7	66.3	24.0	20.3	46.9	**63.7**	15.7

Table 1. Results on test splits for Flickr30k, MS COCO and VizWiz [3] datasets. Our approach B-SCST consistently improves the CIDEr-D scores as compared to the traditional SCST approach. SCST* scores are obatained from the checkpoints provided by the authors.

respectively; θ are the model parameters and T is number of words in the caption. The choice of "baseline" reward $r(\hat{y}_{1:T})$ is important here, and the usage of greedy caption as "baseline" may be undesirable when there is uncertainty in the model predictions. The SoftMax probabilities have been shown to be overly confident even when the model is uncertain about its predictions. In order to account for this, we propose to estimate the "baseline" reward using Bayesian inference, and refer to it as Bayesian SCST (B-SCST).

We use multiple MC dropout forward passes through the model to infer the distribution of captions around the current model parameters, and estimate their predictive mean CIDEr-D score. We use this predictive mean score, which accounts for the uncertainty, as the "baseline" score. Also, we do not perform any greedy sampling, i.e., choosing only the word with highest SoftMax probability, during this process. Instead, we sample from the SoftMax distribution during training, which allows the model to explore a larger search space. The "baseline" $\tilde{r}$ and gradient of the loss in our proposed model [2] can be approximated by changing Equation 1 as:

$$\tilde{r} \approx \frac{1}{M}\sum_{m=1}^{M} r({y^s_{1:T}}^{(m)})$$

$$\nabla_\theta L_{RL}(\theta) \approx -\frac{1}{M}\sum_{m=1}^{M}(r({y^s_{1:T}}^{(m)}) - \tilde{r})\nabla_\theta \log p_\theta({y^s_{1:T}}^{(m)}) \tag{2}$$

where M is the total number of MC dropout forward passes, and ${y^s_{1:T}}^{(m)}$ is the sampled caption that is generated during the m^{th} forward pass. For MC dropout, we use the dropout layers in the AoANet architecture and allow different dropout masks across timesteps of the decoder. An illustration of B-SCST approach is given in Appendix (Figure 2).

Results: In Table 1, we present a comparison of the proposed B-SCST and SCST approaches on Karpathy test split for MS COCO and Flickr30k datasets, and Vizwiz test split from [3]. SCST* results are presented from the published work [4, 3]. For MS COCO dataset, B-SCST approach uses the same starting point, i.e., model checkpoints provided by the authors for cross-entropy loss training, for CIDEr-D optimization. We observe that B-SCST improves CIDEr-D scores on all three datasets, as compared to the SCST approach. Although we demonstrated our results on AoANet architecture, the proposed approach can be applied on other image captioning architectures that benefit from using SCST approach.

References

[1] 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 visual question answering. In *Proceedings of the IEEE CVPR*, pages 6077–6086, 2018. 1

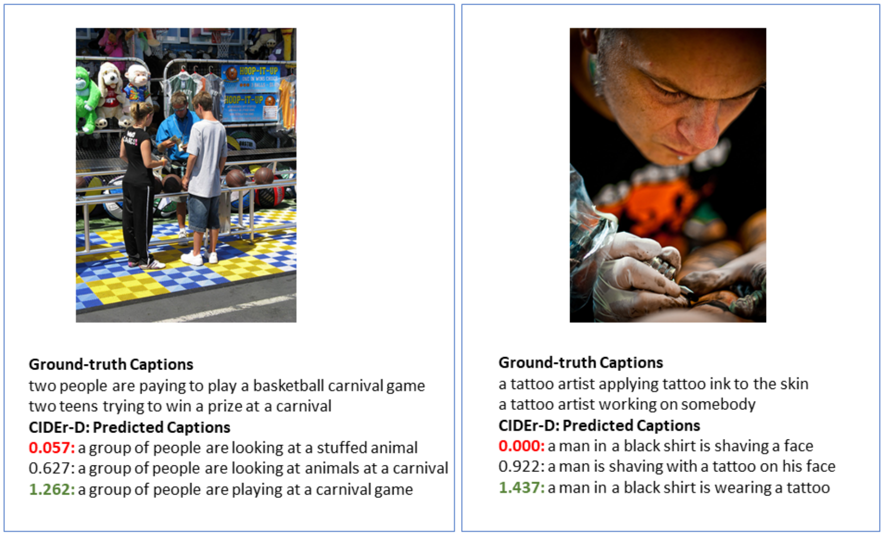
[2] Shashank Bujimalla, Mahesh Subedar, and Omesh Tickoo. B-SCST: Bayesian self-critical sequence training for image captioning. *arXiv preprint arXiv:2004.02435*, 2020. 2

[3] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. *arXiv preprint arXiv:2002.08565*, 2020. 2

[4] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. 1, 2

[5] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE CVPR*, pages 7008–7024, 2017. 1

[6] 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



Ground-truth Captions
two people are paying to play a basketball carnival game
two teens trying to win a prize at a carnival
CIDEr-D: Predicted Captions
0.057: a group of people are looking at a stuffed animal
0.627: a group of people are looking at animals at a carnival
1.262: a group of people are playing at a carnival game

Ground-truth Captions
a tattoo artist applying tattoo ink to the skin
a tattoo artist working on somebody
CIDEr-D: Predicted Captions
0.000: a man in a black shirt is shaving a face
0.922: a man is shaving with a tattoo on his face
1.437: a man in a black shirt is wearing a tattoo

Figure 1. Two sample images and few of their ground-truth captions from Flickr30k dataset. Also shown are the greedy captions that are predicted using a trained model with few MC Dropout simulations, and their corresponding CIDEr-D scores. SCST approach uses a single greedy caption as the "baseline" to improve or suppress the probability of sampled captions. MC Dropout provides a way to sample captions from the posterior distribution, and an average predictive score of these captions can be better estimate of "baseline" score.

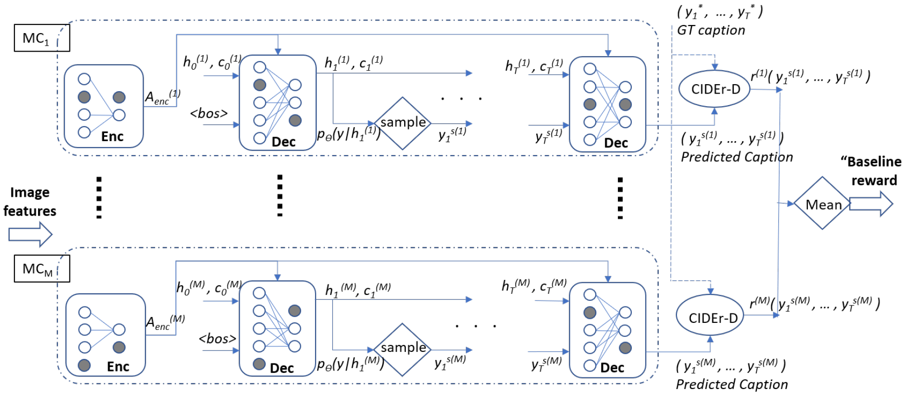


Figure 2. Bayesian Self Critical Sequence Training (B-SCST). The AoANet encoder and decoder modules of the model are marked as "Enc" and "Dec" respectively, and the gray nodes inside them indicate the dropout nodes. The M MC dropout forward passes through the model are marked as MC_1 through MC_M. In each of these forward passes, the input image features go through the encoder, and the decoder uses them to generate the word prediction at each time-step. CIDEr-D score is calculated between the predicted caption and the ground truth caption. The predictive mean of the CIDEr-D scores from M MC dropout forward passes is used as the "baseline" score during policy gradient RL training.

Appendix

We train the model using a minibatch size of 10 images and ADAM optimizer. We first run 25 epochs of cross-entropy loss training, with optimizer learning rate of 2e-4 and decay factor of 0.8 every 3 epochs. The scheduled sampling probability is increased at a rate of 0.05 every 5 epochs along with label smoothing. Since each image contains 5 ground truth labels, we replicate each image feature 5 times and pass it through the model to calculate cross-entropy loss for each caption. We run 30 epochs of CIDEr-D optimization with optimizer learning rate of 2e-5 and a reduce on plateau factor of 0.5 when CIDEr-D degrades for more than one epoch. For VizWiz dataset, we run only 25 epochs of CIDEr-D optimization with starting learning of 1e-5 to avoid overfitting. During inference, we use beam search and pick the caption with the highest SoftMax probability. We use a beam size of 2 for fair comparison of our results with published AoANet and VizWiz results. During training using B-SCST approach, we use M=5 for VizWiz dataset and M=10 for MS COCO and Flickr30k datasets.