Summary 20: Improved Image Captioning via Policy Gradient Optimization

Susan Cherry

This paper proposes an efficient policy gradient method, which it successfully applies to optimize several captioning metrics. The authors use Monte Carlo rollouts to get better estimates and avoid the need to mix in the MLE objective. They also show that their method leads to faster convergence than traditional methods. Their work can be summarized as follows: They identify criteria for a good image captioning metric and propose a new metric, SPIDEr. They propose a new policy gradient methods that optimizes arbitrrary captioning metrics and that is much more stable than the previous methods, Show that their new method to optimize existing BCMR metrics leads state of the art results on COCO. Finally, they show that using their new PG method to optimize their SPIDr metric gives better human scores than optimizing for other metrics.

First, the authors discuss training using gradient descent. Policy gradient descent can be used to optimize any type of reward function. In this paper, the agent receives no reward during intermediate steps so they estimate the value of intermediate steps using Monte Carlo rollouts. The goal is to maximize the average reward starting from the initial empty state. The overall algorithm is presented on the 4th page. Monte Carlo rollouts require only a forward pass through the RNN. This is more efficient than the forward-backward pass needed for the CNN. They rollouts can also be done in parallel.

Next they describe reward functions for the policy gradients. Common choices of reward functions include BLEU, CIDEr, METEOR, and ROUGE. They decide to weight a combination of all of these. They found that optimizing just SPICE leads to detailed captions with many repeated phrases. They combined SPICE with CIDEr, which they call SPIDEr to deal with this issue. They use a CNN-RNN architecture. See Figure 2 for details. The symbols in the vocabulary are embedded as 512 dimensional dense word vector. The values are initialized randomly.The encoder CNN is implemented as an Inception V3 network and is pretrained on ImageNet.

They try different methods on the COCO dataset. First, they quantitatively evaluate different methods using the standard BCMR metrics. PG-BCMR performs much better than MLE training. PG-SPIDEr is outperformed by PG-BCMR, as well. PG-BLUE-4 achieves the highest score on the BLEU metrics. This shows that their optimization method is generally applicable. Next they move to human evaluation. PG-SPIDEr gives more reasonable captions than PG-BCMR. The authors come up with the following results: all methods are far below the human ceiling, all PG methods significantly outperform the PG methods, and PG-SPIDEr outperforms PG-BCMR. They conclude by comparing with MIXER. They find that MIXER is much slower to converge and much less stable during training.