Introduction

Models for image caption generation have traditionally been trained with a supervised learning method, by optimizing a cross-entropy loss over each token in a predicted sequence as compared to a reference sequence. However, this approach to image captioning makes the model vulnerable to exposure bias due to “Teacher-Forcing” and furthermore relies on a suboptimal metric for the model’s true performance.

Lu (2018) introduced SCST-PPO, a policy gradient algorithm for training sequence generation models which combines ideas from self-critical sequence training (SCST) and proximal policy optimization (PPO).

Problem

Image captioning, which involves generating a natural language description of an image, has been a key task in artificial intelligence research. Since good performance on image captioning requires an understanding of a scene, ability to “compose” attributes, objects, and relationships, and expression in natural language, the task is still very much an open problem.

We use the MSCOCO (Common Objects in Context) dataset, which includes 330,000 images, each of which is annotated with 5 reference captions.

Performance has generally been evaluated with metrics such as CIDEr, the BLEU score, METEOR, and ROUGE. These metrics are typically non-differentiable and cannot be optimized directly with backpropagation, rendering them unsuitable for supervised learning methods.

Methodology

We frame the image captioning problem as a reinforcement learning problem:

- **State**: Image $I$, previous tokens generated $w_{1:t-1}$
- **Actions**: tokens $w_t$
- **Reward**: metric e.g. CIDEr score $r(w_{1:t})$ if $t = \text{EOS}$
- **Policy** $p(w_t|I, w_{1:t-1})$
- **Objective**: maximize expected future reward $E_{w} r(w)$

We implement SCST-PPO which includes ideas from SCST and PPO:

![Algorithm 1: Self-critical proximal sequence training](image)

**Discussion and Future Work**

The models fine-tuned with SCST-PPO exhibited significantly better sample efficiency and learning speed

- # PPO iterations may be an important hyperparameter
- Att2in model significantly outperforms FC model
- Att2in model trained with SCST-PPO shows promise in improving late 2017 SOTA results on MSCOCO

- XE: a bed with a blanket on top of it
- SCST: a bed with a blanket in the middle of it
- SCST-PPO: a bedroom with a bed and in the grass