Image Captioning with SCST-PPO and Attention

Improved sample efficiency over self-critical sequence training (2017 SOTA)

Cindy Kuang (ckuang) & Charles Lu (clu8)
CS 230 Winter 2018

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.

a cat grabbing a game controller that is on a blanket.

a close up of a cat with a nintendo wii remote

a gray and white cat sitting behind a wii remote.

a cat with its paw on a wii remote

there is a cat holding a game controller

An example of an image in MSCOCO with 5 reference captions

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 unsuited for supervised learning methods.

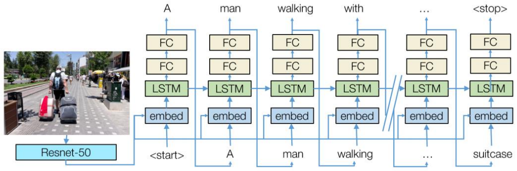
Methodology

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 $\begin{cases} r(w_{1:t}) \text{ if } t = \text{EOS} \\ 0 \text{ otherwise} \end{cases}$
- Policy: $p_\theta(w_t|I, w_{1:t-1})$
- Objective: maximize expected future reward $E_{w\sim p_\theta}[r(w)]$

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

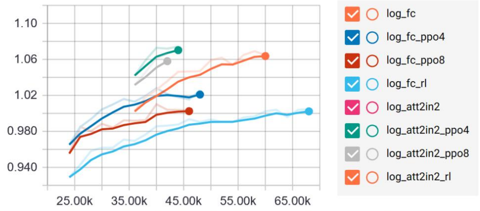
```
Algorithm 1: Self-critical proximal sequence training
Input: start states $I$ and reference sequences, scoring metric $R$, model parameterized by $\theta$
Result: optimize $\theta$ to maximize expected reward under $R$
for each epoch do
    for each batch do
        $w_{1:T} \leftarrow$ sample $\pi_\theta$;
        $R_t \leftarrow R(w_{1:T})$;
        $p_{old} = \sum_{t=1}^{T} \log \pi_\theta(w_t|I, w_{1:t-1})$;
        $\hat{w}_{1:T} \leftarrow$ sample greedily $\pi_\theta$;
        $b_t \leftarrow R(\hat{w}_{1:T})$;
        $\hat{A} \leftarrow R_t - b_t$;
        $\theta_{old} \leftarrow \theta$;
        for each PPO iteration do
            $p = \sum_{t=1}^{T} \log \pi_\theta(w_t|I, w_{1:t-1})$;
            $r(\theta) = \frac{p}{p_{old}}$;
            optimize PPO loss $L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$;
```



Basic FC model without attention

We first train our models with the supervised XE loss for 20-25 epochs. We then continue fine tuning with SCST-PPO (with 4/8 PPO iterations) and benchmark vs. vanilla SCST.

Results

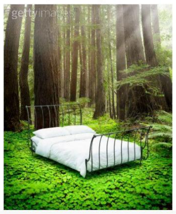


Model	CIDEr	METEOR	BLEU-3	ROUGE-L
FC, XE 25e	0.976	0.251	0.420	0.534
FC, XE 25e + SCST 15e	0.993	0.251	0.425	0.539
FC, XE 25e + SCST-PPO (4 iter.) 15e	1.016	0.252	0.435	0.544
FC, XE 25e + SCST-PPO (8 iter.) 15e	1.003	0.250	0.426	0.541
Att2in, XE 20e	1.043	0.260	0.442	0.546
Att2in, XE 20e + SCST 3e	1.036	0.249	0.420	0.550
Att2in, XE 20e + SCST-PPO (4 iter.) 3e	1.072	0.252	0.438	0.552
Att2in, XE 20e + SCST-PPO (8 iter.) 3e	1.066	0.250	0.427	0.551

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*