Deep Reinforcement Learning-based Image Captioning with Embedding Reward

Zhou Ren¹, Xiaoyu Wang¹, Ning Zhang¹, Xutao Lv¹, Li-Jia Li²
¹Snap Inc. ²Google Inc.

Decision-Making Framework for Image Captioning

Limitations of the Encoder-Decoder Framework:
- Only local information is utilized in training and inference
- Prone to accumulate generation errors during inference
- Sensitive to beam sizes in inference

Our Target for the New Framework:
- Better at utilizing global information
- Less likely to accumulate generation errors during inference
- Less sensitive to beam sizes in inference

Problem Re-formulation in the New Framework:
- Agent: the image captioning model to learn
- Environment: an image I, and the words predicted so far
- State s: the representation of environment at each time step t
- Action a: the word to predict at each time step t

Overview of Our Approach:
- Embedding driven
- Reinforcement Learning
- Agent: Lookahead Inference
- Action prediction

Agent Design: Policy Network + Value Network

- Policy Net: provide the confidence of predicting the next word according to the current state.
- Value Net: evaluate how good a given state is for the final goal.
  - global and lookahead guidance

Agent Inference: Lookahead Inference

Agent Training: RL with Embedding Reward

- Pretrain Policy Net with cross entropy loss
- Pretrain Value Net with mean squared loss
- Jointly train Policy Net and Value Net using deep Reinforcement Learning
  - an Actor-Critic RL model
  - use MIXER training (Plaice et al. 2018)
- Reward is defined by visual-semantic embedding

Experiments on MS-COCO

- Better at capturing global information
- Less likely to accumulate errors

Quantitative Results and Ablation Study:

Parameter Sensitivity Analysis:

Our approach is modular w.r.t. the agent design.

Our approach is less sensitive to beam sizes comparing to encoder-decoder.