Learning to Evaluate Image Captioning

Yin Cui¹,², Guandao Yang¹, Andreas Veit¹,², Xun Huang¹,², Serge Belongie¹,²
¹ Department of Computer Science, Cornell University ² Cornell Tech

Introduction
Captioning Evaluation Challenges
1. Difficulty to correlate well with human judgements.
2. Lack of provision to repair targeted blind spots or targeted pathological cases.

Contributions
- A novel learning based image captioning evaluation metric that tackle both of these challenges.
- SOTA human correlation.
- Show how to train a good metric.
- Demonstrate the robustness of the proposed metric.

How to Train a Good Metric
- Image Feature: To better distinguish human and machine captions.
- Nonlinearity: The binary classifier requires nonlinearity, Compact Bilinear Pooling (CBP) or MLP.
- Data Augmentation: Adding pathologically transformed captions and Monte Carlo samples as negative examples to increase robustness.

Pathological Transformations
Original: a cat sitting on a window sill looking out of the window
\( y = \{ \text{4}\} \times \{ \text{6}\} \times \{ \text{1}\} \times \{ \text{8}\} \) (a cat sitting on the keyboard of a laptop)
\( y = \{ \text{4}\} \times \{ \text{6}\} \) (a cat sitting on a window sill looking out take the honey)
\( y = \{ \text{4}\} \times \{ \text{6}\} \) (cat sitting on a window sill looking out taking the honey)

Original: some antique cars sit in a garage with some bicycles
\( y = \{ \text{4}\} \times \{ \text{6}\} \) (a baseball player standing on a field wearing a uniform)
\( y = \{ \text{4}\} \times \{ \text{6}\} \) (some antique cars sit in a garage with apple bicycles)

Architecture

Experiments
- Capability Experiment
  - Good metrics are capable of distinguishing human and machine captions.
  - Using image features improves models' capability.

- Robustness Experiment
  - Good metrics are robust toward pathological cases.
  - Data augmentation makes model more robust.

Human Correlations
- Good metrics correlate well with human judgements.
- Achieve SOTA performance in both system level (COCO) and caption level (Flickr).

Expert Annotations: experts score image-caption pairs from 1 to 5. 1 means caption doesn't describe the image.
Crowd Flower: human marks 1. If the candidate caption describes the image, and mark 0 if not.