Learning to Evaluate Image Captioning

Yin Cui[1,2], Guandao Yang[1], Andreas Veit[1,2], Xun Huang[1,2], Serge Belongie[1,2]

[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 meric.
- 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 sil looking out of the window

$\mathcal{T}_{RC}$ $\gamma = 10(\%)$ a furry cat sitting on the keyboard of a laptop

$\mathcal{T}_{RW}$ $\gamma = 25(\%)$ a hat sitting on a window sil looking out take the frisbie

$\mathcal{T}_{WP}$ $\gamma = 50(\%)$ on cat of a a window sil window out sitting the looking

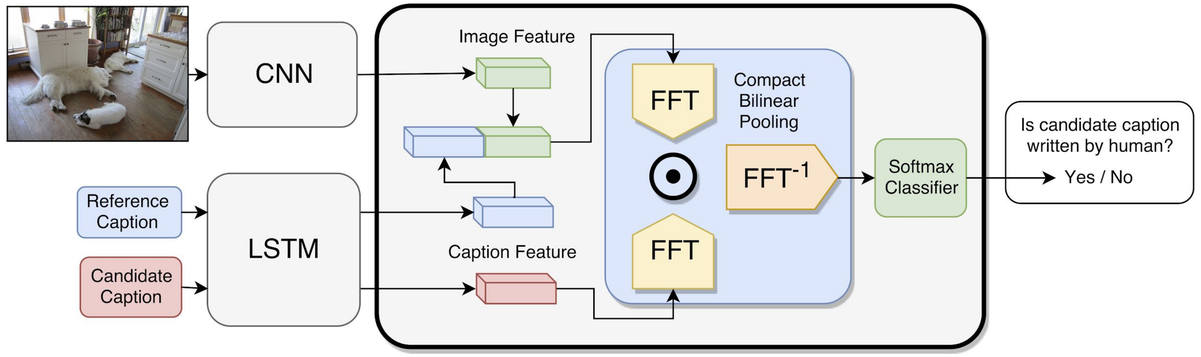
Original: some antique cars sit in a garage with some bicycles

$\mathcal{T}_{RC}$ $\gamma = 50(\%)$ a baseball player standing on a field wearing a uniform

$\mathcal{T}_{RW}$ $\gamma = 40(\%)$ person antique cars a in a see with apple bicycles

$\mathcal{T}_{WP}$ $\gamma = 30(\%)$ some antique some sit in a cars with garage bicycles

Architecture

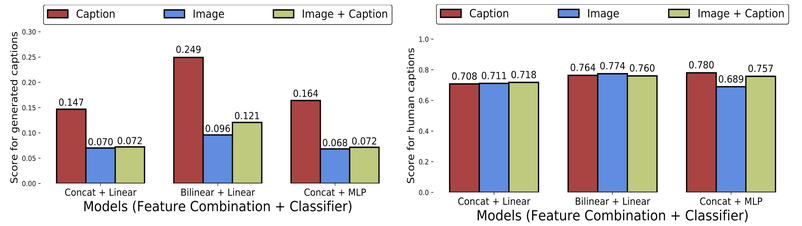


Experiments

Capability Experiment

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

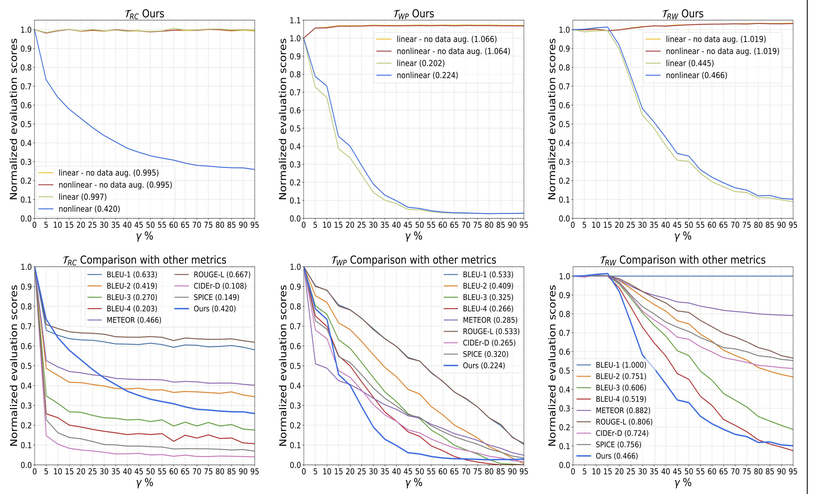
Capability Results



Robustness Experiment

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

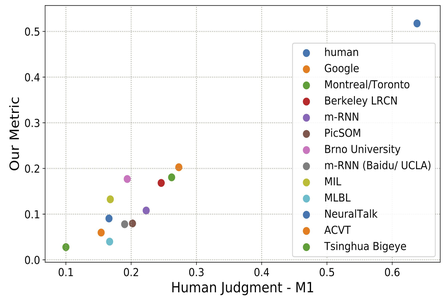
Robustness Results



Human Correlations

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

Human v.s. Metric Score (COCO)



Human Correlation (COCO)

	M1		M2	
	ρ	p-value	ρ	p-value
BLEU-1	0.124	(0.687)	0.135	(0.660)
BLEU-2	0.037	(0.903)	0.048	(0.877)
BLEU-3	0.004	(0.990)	0.016	(0.959)
BLEU-4	-0.019	(0.951)	-0.005	(0.987)
METEOR	0.606	(0.028)	0.594	(0.032)
ROUGE-L	0.090	(0.769)	0.096	(0.754)
CIDEr	0.438	(0.134)	0.440	(0.133)
SPICE	0.759	(0.003)	0.750	(0.003)
Ours (no DA)	**0.821**	(0.000)	**0.807**	(0.000)
Ours	**0.939**	(0.000)	**0.949**	(0.000)

M1: Percentage of captions that are evaluated as better or equal to human caption.

M2: Percentage of captions that pass the Turing Test.

Human Correlation (Flickr)

	Expert Annotations	Crowd Flower
BLEU-1	0.191*	0.206
BLEU-2	0.212	0.212
BLEU-3	0.209	0.204
BLEU-4	0.206*	0.202
METEOR	0.308*	0.242
ROUGE-L	0.218*	0.217
CIDEr	0.289*	0.264
SPICE	0.456	0.252
Ours	**0.466**	**0.295**
Inter-human	0.736	-

Expert Annotations: experts score image-caption pairs from 1 to 4; 1 means caption doesn't describe the image.

Crowd Flower: human raters mark 1 if the candidate caption describes the image, and mark 0 if not.