Visually Descriptive Image Captions:

Improve Your Model With No Additional Training

Brandon Trabucco, Junjiao Tian, Roberto Shu, Jean Oh, Ralph Hollis

Introduction

Research Question

Can we generate visually descriptive captions for images of humans?

Problem Definition

$$\phi_\theta : \mathbb{R}^{X \times Y \times C} \rightarrow \{[y_0, y_1, \ldots, y_N] : \forall y_i \in \mathbb{Z}\}$$

Given an *image* of width **X**, height **Y**, and colors **C**, we define the transformation **ϕ**, tuned by parameters **θ**, into a *sentence*, as a sequence of integer word ids $\mathbf{y_i}$ from a vocabulary [3] of possible words.

Baseline Generates Ambiguous Captions

a baseball player holding a bat on a field .

a baseball player holding a bat on a field .

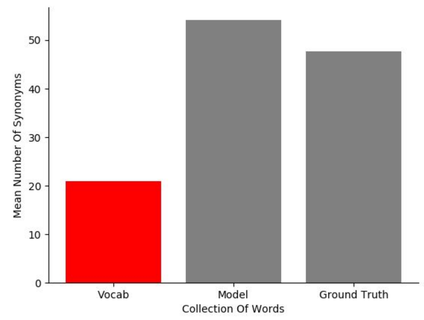
Our Contributions

- We develop two cost functions for the descriptiveness of word choice and for the visual grounding of word choice.
- We propose a novel inference adaptation method to encourage existing models to generate detailed captions.

Model

Encourage Descriptive Words

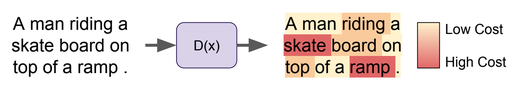
- The model learns to use words with many synonyms, which implies a given word is more general.



- Our cost function is the sum of distances from word **x** to the closest **K** other words in the vocabulary embedding space **η** [3].

$$D(x \in \mathbb{Z}) = \sum_{i=0}^{K} \min_{y \neq x}[i] ||\eta(x) - \eta(y)||$$

Descriptive Cost Function

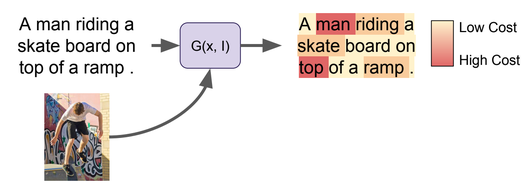


- Our hypothesis is that descriptive words have fewer close synonyms.

Encourage Visual Words

- Prior work [2] demonstrates that detection of attributes improves the descriptiveness of captions from the baseline [1].
- Cost of word **x** is the probability of visual grounding in the source image **I**, trained with DeepFashion attributes [5].

Visual Cost Function



Stylistic Transfer

- Maximize the expected values for **D(x)** and **G(x, I)** during inference time with respect to the initial state $\mathbf{s_0}$ of the LSTM.

#	Algorithm
0	def caption():
1	Encode, Decode = load (”model.ckpt”)
2	s_0 = Encode (”image.jpg”)
3	for x in range(N):
4	y_0 = id (”$< S >$”)
5	$y_1, \ldots, y_L$ = Decode (y_0, s_0)
6	$\tilde{Q} = \sum_{i=1}^{L} P(y_i)D(y_i) + P(y_i)G(y_i, I)$
7	$s_0 \leftarrow s_0 + \gamma \nabla_{s_0} \tilde{Q}$
8	return word $(y_1, \ldots, y_L)$

Results

Evaluation On Existing Benchmark Is Misleading

Our styled captions have lower benchmark scores on the MSCOCO [4] dataset, compared to the baseline [1] algorithm. This result is misleading when viewed in context.

Name	BLEU-4	CIDEr
NIC	27.7	85.5
DetailedNIC	22.9	80.4

The MSCOCO [4] dataset uses words that have many more synonyms than what is available in our vocabulary [3]. The benchmark is biased towards less descriptive captions.

Our Method Improves Visual Descriptiveness Of Captions

With no additional training, maximizing the expected **Q** value during inference time successfully produces captions that are more visually descriptive (and *sometimes* more correct) than the baseline.

Before:
a baseball player holding a bat on a field .

After:
baseball player swinging a bat at **home plate during a baseball game** .

Before:
a train traveling down train tracks next to a building .

After:
steam engine train traveling down train tracks next to **trees** .

Before:
a man holding a nintendo wii game controller .

After:
woman playing wii **video game in a living room** .

Before:
a group of young men playing a game of frisbee .

After:
a group of men playing **soccer on a field** .

Before:
a man riding a skateboard down a ramp .

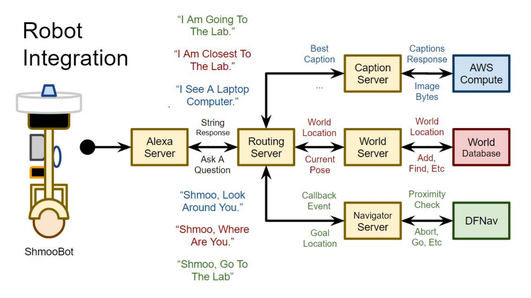
After:
skateboarder doing a **trick** on a ramp **at a skate park** .

Before:
a group of people playing a game of frisbee .

After:
batter baseball catcher and umpire at home plate during a baseball game .

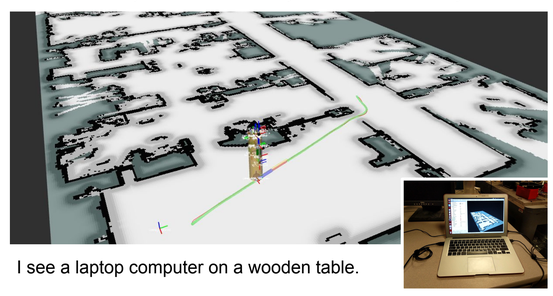
Deploying Our Framework Onto The Shmoobot

Image captioning is an important ability to enable the Shmoobot to interact with the visually impaired, and perform remote monitoring tasks in the workplace.



Our framework interacts with the navigator on Shmoobot to go to a verbally specified location, caption what is seen, return to the user, and verbally describe what was seen.

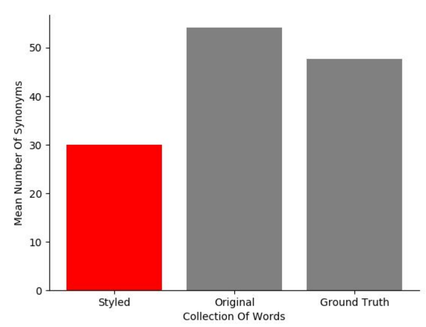
1. Shmoo, go to the hallway.
2. Shmoo, look around you.
3. Shmoo, tell me what you saw.



I see a laptop computer on a wooden table.

Our Model Uses Words With Less Synonyms

We perform stylistic transfer onto the initial state of the LSTM. Captions generated after a few iterations have significantly less synonyms on average than the baseline [1].



Acknowledgements

Brandon is grateful to Dr. Jean Oh, and her Masters student Junjiao Tian. Their questions and feedback have instilled in him curiosity and excitement for research. Brandon is also grateful to Dr. Ralph Hollis, and his PhD student Roberto Shu. Their lab has been a warm and welcoming place to work and invent. Special thanks to Rachel Burcin, John Dolan, and the many other key organizers for RISS, including NSF grant IIS-1547143 funding Brandon's summer experience.

References

1. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,”CoRR, vol. abs/1411.4555, 2014.
2. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” inThe IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
3. J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.
4. T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays,P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO:common objects in context,”CoRR, vol. abs/1405.0312, 2014.
5. Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” in The IEEE Conference on Computer Vision and Pattern Recognition(CVPR),June 2016.