Generating Diverse and Descriptive Image Captions Using Visual Paraphrases Supplementary Material

Lixin Liu[1,2,3] Jiajun Tang[1] Xiaojun Wan[1,2,3] Zongming Guo[1,2]
[1]Institute of Computer Science and Technology, Peking University
[2]Center for Data Science, Peking University
[3]The MOE Key Laboratory of Computational Linguistics, Peking University
{liulixin, jiajun.tang, wanxiaojun, guozongming}@pku.edu.cn

1. More Examples

In this supplementary material, we report more examples in MS COCO dataset Karpathy test split for qualitative analysis. Detailed descriptions of the images in generated captions are emphasized with different colors, which may reflect diversity and descriptiveness.

We first show examples in Figure 1 to compare captions generated by different scoring functions (**Ours (len, 0)**, **Ours (Yngve, 0)**, **Ours (IR, 2)**, and **Ours (Tdiv, 0.3)**), along with captions generated by **Attention** baseline. As is shown, captions generated by **Ours (len)** indeed describe more details, but tend to be very long, sometimes truncated due to the length limitation (in Figure 1(c)). **Ours (Yngve)** also captures some important details, but still lacks in diversity of wordings and expressions. Captions by **Ours (IR)** may miss some details, but overall they are concise and precise descriptions. **Ours (Tdiv)** provides precise and natural image descriptions with both rich expressions (such as *"a scenic view"* in (b)) and descriptive details (such as *"with people loading off"* in (c)).

Example images with captions generated in both two decoding steps of **Ours (IR, 2)** and **Ours (Tdiv, 0.3)** are shown in Figure 2, with obvious difference between outputs from the two steps to be found. Compared to the captions generated in the first decoding step, the final captions are usually longer and better with more details (such as *"a catcher and umpire behind him"* in Figure 2(b) and *"next to a railroad crossing"* in (c)) and polished expressions (such as *"a surfer"* instead of "a man" or "a person" in (a)), thus are more diverse in wordings and expressions and more descriptive with respect to important details.

More examples of **Ours (Tdiv, 0.3)**, **Ours (IR, 2)**, and baseline methods including **Attention**, **GAN** and **StackCap** are shown in Figure 3. Comparing to baselines, our models especially **Ours (Tdiv)** perform well in most cases. However, it is worth pointing out that mistakes may come along with more detailed descriptions (such as *next to a american flag* in Figure 3(c), where an improper preposition and a wrong article are used). Moreover, Figure 3(f) shows a failure case when dealing with very complex images, both our models and other baseline models are struggling to make a description on the whole, while humans can usually find some way to summarize in a sentence with reasoning based on commonsense knowledge.

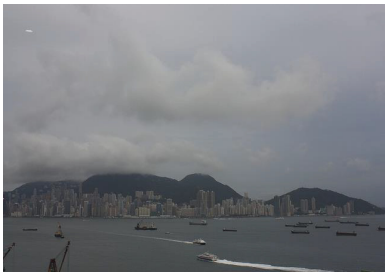
2. Human Evaluation Details

In human evaluation, We compared methods on 100 images randomly sampled from Karpathy test set. Volunteers are asked to compare 9 sentences from 9 models to rate them from the 1-5 scale. The definitions of four criteria of human evaluation are:

- Relevance: whether the sentence correctly describes the visual content and be closely relevant to the image.
- Fluency: whether the caption is a fluent sentence.
- Diversity: whether the sentence uses diverse expressions. If it only uses very simple and ordinary expressions then it should receive a low score.
- Descriptiveness: whether the sentence is an precise, informative caption that describes important details of the image.

(a)

Attention: a woman standing in front of a brick building
Ours (len): a woman in a yellow jacket holding a red umbrella in front of a building
Ours (Yngve): a woman standing in front of a brick wall holding a red umbrella
Ours (IR): a woman in a yellow coat holding an umbrella
Ours (Tdiv): a woman in a yellow jacket holding an umbrella in front of a brick building
Human: a woman in a yellow coat uses a red umbrella to shield herself from the rain



(b)

Attention: a large body of water with lots of boats
Ours (len): a large body of water with boats in the water and a building in the background
Ours (Yngve): a large body of water with boats in the water and buildings in the background
Ours (IR): boats in a harbor with a city in the background
Ours (Tdiv): a scenic view of a harbor with boats in the water and mountains in the background
Human: a picture of some boats and cityscape on a cloudy day

(c)

Attention: a white and blue airplane parked on a runway
Ours (len): a large passenger jet sitting on top of an airport tarmac with people in the
Ours (Yngve): a large white and blue jet airliner on runway
Ours (IR): a large white and blue airplane on a tarmac
Ours (Tdiv): a large white and blue jet airliner on the tarmac with people loading off it
Human: a large plane with people alighting at the airport

Figure 1. Examples of captions generated by Attention baseline and our models using different scoring functions.

(a)

Attention: a man riding a wave on top of a surfboard
Ours (Tdiv, first): a person riding a surf board on a wave
Ours (Tdiv, second): a surfer in a wet suit riding a wave on a surfboard
Human: A person wearing a black water suit surfs in the ocean

(b)

Attention: a baseball player swinging a bat at a ball
Ours (Tdiv, first): a baseball player swinging a bat at a ball
Ours (Tdiv, second): a baseball player swinging a bat at a ball with a catcher and umpire behind him
Human: an umpire officiates a game of little league baseball

(c)

Attention: a stop sign with a stop sign on it
Ours (IR, first): a stop sign that is on the side of a pole
Ours (IR, second): a red stop sign sitting next to a railroad crossing
Human: a red stop sign covered in graffiti under a train crossing

Figure 2. Examples of captions generated in two decoding steps of our models.

Figure 3. More qualitative examples of different baselines and our models.