Unsupervised Image Captioning - Supplementary Material

Yang Feng[♯*] Lin Ma[♮†] Wei Liu[♮] Jiebo Luo[♯]
[♮]Tencent AI Lab [♯]University of Rochester
{yfeng23,jluo}@cs.rochester.edu forest.linma@gmail.com wl2223@columbia.edu

Figure 1. Word cloud of Shutterstock Image Description Corpus.

Figure 2. Word cloud of generated captions on test split.

1. Word Clouds

Fig. 1 and Fig. 2 illustrate the word clouds generated using the sentences in the Shutterstock Image Description Corpus and the generated captions, respectively. It can be observed that some words appear more frequently than they do in the Shutterstock Image Description Corpus, *e.g.*, "young".

2. Length of the Generated Captions

Fig. 3 shows the distribution of the length of the generated captions. It can be observed that most of the generated captions consist of about eight words.

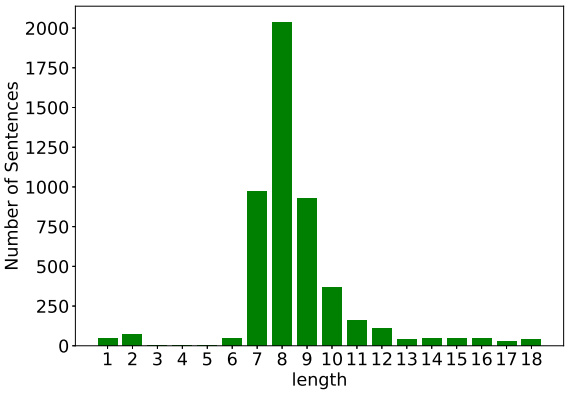


Figure 3. The distribution of the length of generated captions.

*This work was done while Yang Feng was a Research Intern with Tencent AI Lab.

†Corresponding author.

3. More Qualitative Results

More qualitative results are illustrated in Fig. 4. The caption generated by "con2sen" only depends on the detected objects in the input image, while the other models generate captions conditioning on the input image feature. For the first image, the sentence generated by "adv" is unrelated to the image because the adversarial objective only enforces the sentence to be genuine. After introducing the other objectives, the generated caption is more closely related to the image. "Ours w/o init" generates "`helmet`", which does not appear in the image. The caption generated by "Ours" accurately describes the image content.

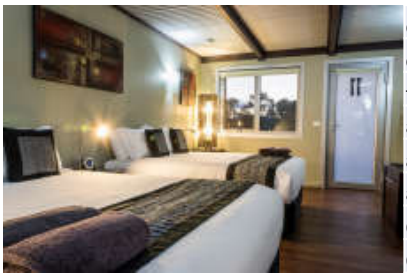
Fig. 5 illustrates some failure cases. In the first case, only "adv + con" recognizes that it is a "`hotel`" room. Most of the other models regard it as a "`bedroom`". The errors in the following two cases are similar. The proposed model fails to recognize the relative position of the objects in the images and generates erroneous captions.

detected concepts	clothing, man, motorcycle, tire, wheel, woman
con2sen	a man and a woman on a motorcycle with a wheel tire
feat2sen	back view of a woman in a motorcycle . rear view people collection . backside view of person
adv	beautiful young woman sitting on a bench and looking at camera
adv + con	close-up of motorcycle helmet , wheel of helmet
adv + con + im	woman with motorcycle helmet
Ours w/o init	young woman in motorcycle helmet
Ours	woman riding a motorcycle on the road

detected concepts	tree
con2sen	beautiful landscape with tree in the forest .
feat2sen	flock of sheep grazing on a green meadow with tree in the background
adv	sheep grazing on green grass field , green grass
adv + con	flock of sheep grazing on a green meadow
adv + con + im	sheep grazing on green grass
Ours w/o init	herd of sheep grazing on a green meadow
Ours	flock of sheep grazing in a meadow with tree

detected concepts	animal, horse
con2sen	horse on the farm . animal in nature .
feat2sen	a horse in a field in the summer
adv	horse grazing on a sunny summer day .
adv + con	horse in a field on the summer
adv + con + im	horse in a field at sunset
Ours w/o init	horse in a pasture in a field with a horse
Ours	horse in the field with a horse

Figure 4. More qualitative results by the unsupervised captioning models trained with different objectives.



detected concepts	bed, door, house
con2sen	interior of a house with wooden door and bed
feat2sen	modern bedroom interior with white furniture , bed and house
adv	young couple sitting on bed and using laptop .
adv + con	white pillow on bed in hotel room .
adv + con + im	double bed and green armchair in a house and house
Ours w/o init	modern living room with sofa , sofa , lamp
Ours	interior of modern bedroom with a large window

detected concepts	bus, clothing, jeans, man, person
con2sen	back view of a man in jeans . backside view of person . rear view people collection .
feat2sen	back view of a man in a clothing . backside view of person .
adv	portrait of a happy young woman talking on mobile phone while sitting on sofa at home .
adv + con	bus
adv + con + im	portrait of a young man waiting for the bus
Ours w/o init	london , uk - april 22 , 2017 : a bus in london , uk .
Ours	young man with backpack standing on a bus

detected concepts	clothing, man
con2sen	young man in casual clothing sitting on sofa at home
feat2sen	young man in casual clothing sitting on sofa at home
adv	a glass of red wine on the table
adv + con	man 's hand holding a bottle of water
adv + con + im	man using smartphone while sitting on toilet .
Ours w/o init	close up of man 's hand typing on laptop keyboard
Ours	young man sitting on toilet and using laptop

Figure 5. The failure cases by the unsupervised captioning models trained with different objectives.