Project 3: Image Captioning of Novel Objects

Haitian Jiang

Fudan University 19307110022@fudan.edu.cn

1 Introduction

Image captioning aims to generate a piece of text describing the scene of the given picture, which typically requires enormous paired image and text for training. Models with encoder-decoder architecture[2][8] have been pushed out to solve such problems. However, they typically cannot generalize to novel objects unseen in the paired training materials. In this project, I will try to use the modern models to solve the problem of image captioning on novel objects.

There were several papers introducing new datasets or new splits on old datasets for the image captioning on novel objects. For example, the recent work of novel object captioning at scale(nocaps)[1] incorporates some part of the images in the Microsoft COCO dataset[4] as well as a lot of images containing classes out of the COCO distribution. They also set up a benchmark called *nocaps* based on this dataset.

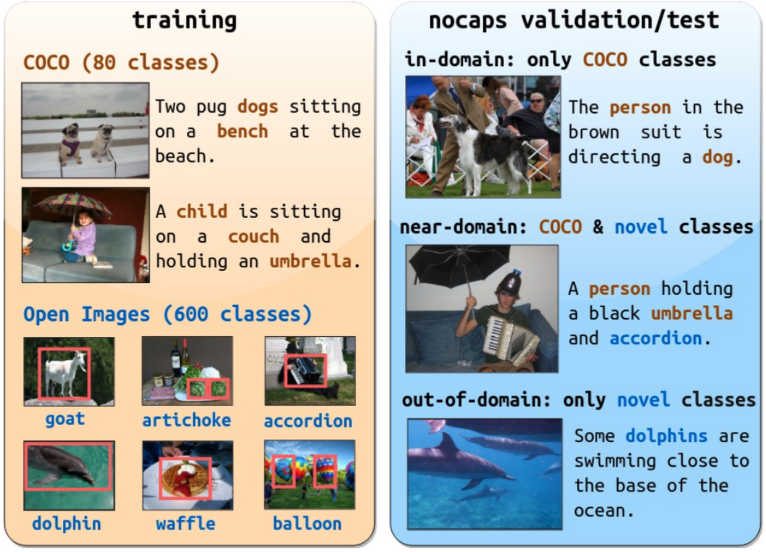

training
COCO (80 classes)
Two pug dogs sitting on a bench at the beach.
A child is sitting on a couch and holding an umbrella.
Open Images (600 classes)
goat
artichoke
accordion
dolphin
waffle
balloon
nocaps validation/test
in-domain: only COCO classes
The person in the brown suit is directing a dog.
near-domain: COCO & novel classes
A person holding a black umbrella and accordion.
out-of-domain: only novel classes
Some dolphins are swimming close to the base of the ocean.

Fig. 1. The nocaps benchmark for novel object captioning

Earlier, Deep Compositional Captioning(DCC)[3] also offers a new split for the COCO dataset by removing eight categories of objects: bottle, bus, couch, microwave, pizza, racket, suitcase and zebra from the original dataset. Though not containing delicate partition of in-distribution, near-distribution(containing both objects having appeared in COCO and not appeared in COCO), and out-of-distribution data like the *nocaps* does, this split is well enough constructed to test the model performance on unseen objects, and I am asked to use this split for training and evaluation in this project.

2 Related Works

Image captioning on novel objects requires knowledge from distinguished sources other than the paired image-text dataset to extend its boundary beyond these unseen objects; otherwise it can never know what to call these objects and the properties of them.

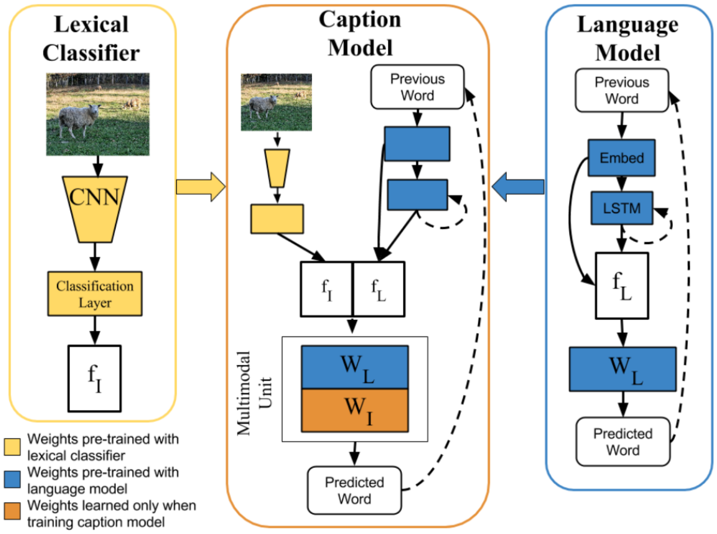


Fig. 2. The model structure of DCC

DCC uses a separate lexical classifier and a language model trained on unpaired images and texts respectively, to learn the representations of objects and words. Then the model is trained jointly on the known paired dataset to let these two models understand each other. The output of these two separate models are concatenated and then sent into a linear layer called multimodal unit. The idea of this structure is general for the task, in that we have to use some kind of method to transfer the knowledge the networks learnt elsewhere to the captioning task. And the novel only refers to not seeing these objects during training on paired data, but these objects have to be seen during the separate learning before, or by no means the network can recognize these objects.

3 Model

In this project, I use the similar idea of transfer learning as DCC does, but different network structures. Still, I adopt the encoder-decoder architecture as the traditional models do. I employ the CLIPcap model[5], which use CLIP[6] as the encoder for the image, and GPT-2[7] as the decoder for generating text. For my network trained from scratch with the paired image-text data, I exploit a transformer-structured network to convert the encoded image vector to a set of vector with the same dimension of word embeddings so that the decoder can understand and use as the starting information.

3.1 CLIP

CLIP(Contrastive Language-Image Pre-Training)[6] is a encoder proposed by OpenAI, trained on a dataset containing 400 million image- text pairs collected from the internet with the object of distinguishing corresponding image of the given text to learn the representation of images in a fast and extensible way. Hence, it can produce representations for images able to extend to other tasks typically demanding for networks only trained on ImageNet by exploiting more comprehensive information from the text, thus gaining a better understanding of the whole image. In this project, I use the ResNet-101 structure from CLIP as my encoder.

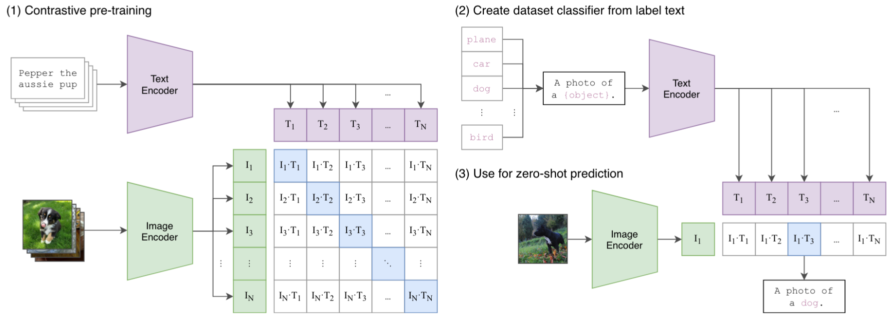

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
T1 T2 T3 ... TN
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Fig. 3. Training and evaluating CLIP model

3.2 GPT-2

GPT-2[7] is an auto-regressive uni-directional language model based on transformer decoder. It can adapt to a myriad of text generation tasks like machine translation, question answering and summerization.

3.3 Mapping Network

The most important structure in this image captioning network is the mapping work that bridges the gap of the CLIP encoder and the GPT-2 decoder. It carries the crucial task of transferring the hidden vector in the image space to something fathomable to the decoder in the word embedding space, and its parameters can only be trained from scratch by the paired image-text data without the removed objects. I use a transformer encoder to do this transformation.

The output of the mapping transformer is going to be sent into the GPT-2 model as the start, so it has the same dimension as the word embedding of GPT-2. Since it carries all the information from the image space to the text space, the output of the mapping network should not be a bottleneck of information, so it contains a set of vectors rather than one single vector. In my model, the size of the set is fixed to 10 in order to get the expressive yet fast and non-overfitting representations.

The attention mask for the GPT-2 should be changed slightly because we should allow the model to attend to the input set. So the upper-left corner should always not be masked for self-attention during the decoding procedure, as shown in Fig. 4

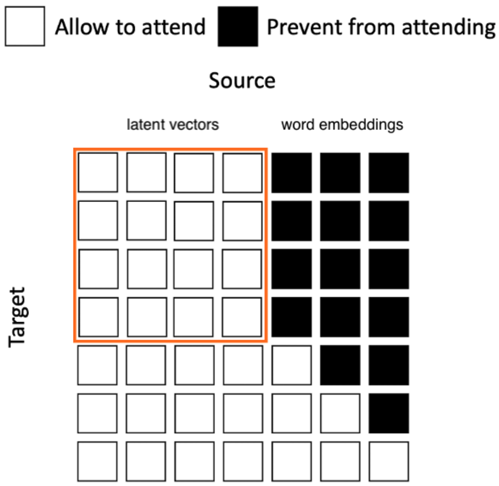


Fig. 4. The attention mask for GPT-2

In order to generate the output with the determined number of vectors, we have to input the same number of vectors. So a set of learnable constant vectors are concatenated to the input encoded image vector. These constant vectors serve as an auxiliary translation tool from image hidden space to word hidden space and also allows the language model to adapt to new data. The whole structure of the encoder-decoder model is shown in Fig. 5.

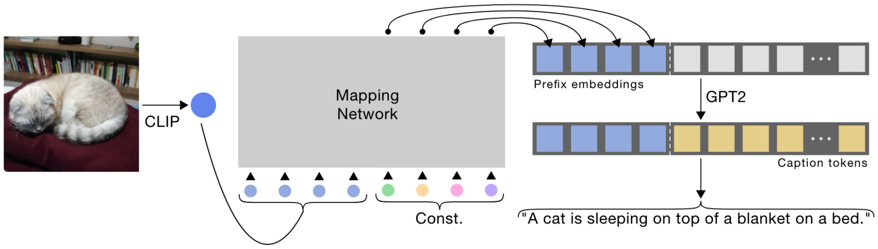


Fig. 5. The entire model structure

3.4 Training procedure

Suppose the dataset contains N image-text pairs $\{x^i, c^i\}_{i=1}^N$, where $c^i = c_1^i, c_2^i, \cdots, c_l^i$ is padded or truncated to the max length l. The training object is the perplexity of the text, or negative likelihood loss:

$$\min_\theta -\sum_{i=1}^{N} \log p_\theta\left(c_1^i, \ldots, c_\ell^i \mid x^i\right)$$

Using the product rule, we can further split each word, and the loss becomes

$$\min_\theta -\sum_{i=1}^{N}\sum_{j=1}^{\ell} \log p_\theta\left(c_j^i \mid x^i, c_1^i, \ldots, c_{j-1}^i\right)$$

Suppose our mapping network M produces k prefix embeddings from the CLIP-encoded image vector: $p_1^i, \cdots, p_k^i = \mathrm{M}(\mathrm{CLIP}(x^i))$, then the condition on x^i becomes condition on $p_1^i, \cdots, p_k^i$. Hence, the ultimate loss is

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{\ell} \log p_\theta\left(c_j^i \mid p_1^i, \ldots, p_k^i, c_1^i, \ldots, c_{j-1}^i\right)$$

3.5 Inference

During inference time, I use beam search with beam size 5 to find the most likely sentence.

4 Experiment results

I train the model on the split introduced by DCC from which the eight kinds of objects are removed and tested on the eight corresponding splits where these novel objects appear. The eight novel objects are bottle, bus, couch, microwave, pizza, racket, suitcase and zebra. I measure the BLEU-1 to BLEU-4, METOR, ROUGE-L, CIDEr, SPICE and F-1 scores on the generated sentences for evaluation. The F-1 score follows the way described in DCC: true positives means a word appears in a sentence it should appear in; false positives means a word appears in a sentence it should not appear in; false negatives, when a word does not appear in a sentence it should appear in.

BLEU score, METOR and ROUGE compares the concurrence of words or n-grams in the generated sentences and the ground truth sentences. So it can only measure the fluency of the generated sentence but not semantic match. On the setting of understanding the scene of the image, CIDEr and SPICE score can better evaluate how the model performs on comprehend the image.

Table 1 and Table 2 show these scores on the validation set and test set respectively.

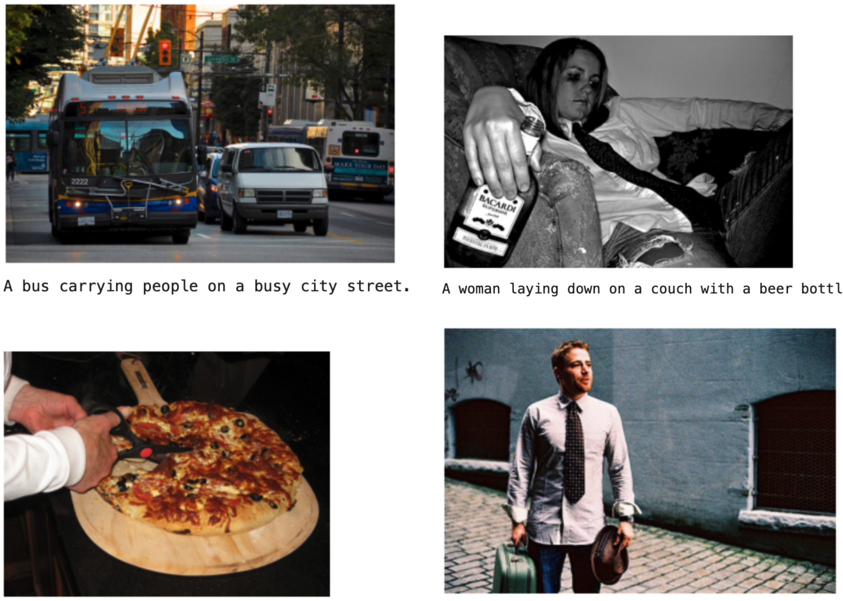
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE-L	CIDEr	SPICE	F-1
bottle	59.86	44.36	32.47	23.39	23.69	48.97	80.36	16.58	18.97
bus	58.86	39.68	26.84	18.34	19.82	44.15	41.89	12.48	25.01
couch	66.02	52.55	39.75	29.14	25.30	54.62	66.53	18.02	42.14
μ-wave	65.97	49.52	34.76	24.21	22.06	50.52	49.26	14.46	31.55
pizza	36.76	25.35	17.04	11.29	17.66	41.91	38.17	11.75	40.17
racket	72.08	54.02	38.63	26.60	26.17	54.61	40.04	19.47	61.99
suitcase	55.80	39.34	26.68	18.05	19.36	43.94	53.30	12.97	10.45
zebra	57.46	39.50	27.55	17.81	18.75	45.28	36.47	11.12	61.23

Table 1. The evaluation results on the validation set

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE-L	CIDEr	SPICE	F-1
bottle	60.40	45.26	32.69	23.17	23.50	50.23	75.37	15.50	17.07
bus	59.29	40.30	26.77	17.45	19.91	44.65	45.74	12.03	23.17
couch	67.52	53.52	40.45	30.09	25.58	55.31	66.14	17.73	42.74
μ-wave	54.52	40.14	28.91	20.44	21.56	50.22	46.20	14.02	32.67
pizza	39.73	28.04	19.41	13.52	19.11	44.08	42.57	12.62	40.22
racket	70.33	52.69	37.29	25.69	25.94	54.00	38.55	19.40	62.96
suitcase	55.20	38.67	27.21	19.24	19.96	45.16	53.51	12.56	8.06
zebra	57.30	38.32	26.08	15.99	18.03	43.88	34.22	11.02	62.26

Table 2. The evluation results on the test set

Fig. 6 are some examples of the generated caption by my model. From this result we can see that this model can recognize the objects, especially novel objects in different scenes, but it still needs to be improved in recognize the relationship between objects in complicated scenes, as the pizza image is indicating.



A bus carrying people on a busy city street.

A woman laying down on a couch with a beer bottle.

A person is cooking a pizza on a wooden table.

A man in a suit and tie holding a suitcase.

Fig. 6. Generated captions during evaluation

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8948–8957 (2019)
2. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
3. Hendricks, L.A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: Describing novel object categories without paired training data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–10 (2016)

4. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
5. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
6. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
7. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
8. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)