Project 3: Image Captioning of Novel Objects

Yanwei Fu

Abstract

(1) This is project 3 of our course. The deadline is 5:00pm, May 31, 2021. Please upload the report via elearning.

(2) We do not need the double blind review.

(3) The goal of your write-up is to document the experiments you've done and your main findings. So be sure to explain the results. Generate a single pdf file of your projects and turned in along with your code. package your code and a copy of the write-up pdf document into a zip or tar.gz file and named as Final-Project-*your-student-id*_your_name.[zip|tar.gz], and Final-Project-test-label-*your-student-id*_your_name.[zip|tar.gz]. Also put the names and Student ID in your paper. We provide the testing images, while the testing label is with-held for the evaluation. We care your performance in this task.

(4) You are open to use anything to help you finish this task.

(5) About the deadline and penalty. In general, you should submit the paper according to the deadline of each mini-project. The late submission is also acceptable; however, you will be penalized 10% of scores for each week's delay.

1 Dataset

The dataset is a subset of the MSCOCO

You should first download the official MSCOCO dataset:
https://cocodataset.org/#download
Then, you can prepare the dataset using this annotation:
https://drive.google.com/drive/u/0/folders/1ct0KhDW8ZHW4D9pxu0IX1ntTaH-XOAVV
For details ro prepare for dataset, please refer to https://github.com/LisaAnne/DCC
You can download the file in your local machine, and upload to your servers by 'scp' command.

2 Background

The goal of image captioning is to automatically generate fluent and informative language description of an image for human understanding. As an interdisciplinary task connecting Computer Vision and Nature Language Processing, it explores towards the cutting edge techniques of scene understanding [23] and it is drawing increasing interests in recent years.

While recent deep neural network models have achieved promising results on the image captioning task, they rely largely on the availability of corpora with paired image and sentence captions to describe objects in context. We would like to address the task of generating descriptions of novel objects which are not present in paired imagesentence datasets. Novel Image Captioning studies the task of image captioning with novel objects, which only exist in testing images. Intrinsically, this task can reflect the generalization ability of models in understanding and captioning the semantic meanings of visual concepts and objects unseen in training set, sharing the similarity

to one/zero-shot learning. The difficulty is that neural network may not work on untrained data. It is a special case of Out-of-Distribution (OOD) generalization problem, which aims to addresses the challenging setting where the testing distribution is unknown and different from the training.

For increasing scalability of diversified objects, recently novel object captioning[7, 5, 10] has attracted lots of attention. Most proposed methods are architectural in essence. Researchers have designed template-based caption models [7], multi-task models [5]and novel sampling algorithms [6]. These novel structures are disjointed from normal image captioning task to varying degrees, which causes poor performance on in-domain scores. Inspired by the new fangled deformation strategies[3, 4, 8].

3 Task Description

3.1 Objective

In this project, you will try to solve the task of Novel Image Captioning. Given an input scene image, the objective is to the corresponding informative language description. You can use models introduced in the above, or design your own model. *Note that you can only use the provided training set to train your network from scratch.* After training, provides inference results on the test set. Write a report to illustrate your algorithm and experimental results on the training set. Your grade will depend on your report and performance on the test set.

3.2 Dataset

The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of 328K images.

Splits: The first version of MS COCO dataset was released in 2014. It contains 164K images split into training (83K), validation (41K) and test (41K) sets. In 2015 additional test set of 81K images was released, including all the previous test images and 40K new images.

Based on community feedback, in 2017 the training/validation split was changed from 83K/41K to 118K/5K. The new split uses the same images and annotations. The 2017 test set is a subset of 41K images of the 2015 test set. Additionally, the 2017 release contains a new unannotated dataset of 123K images.

However, We follow the novel object captioning split (NOC split) introduced by[5] to evaluate our proposed method. It comes from the standard split of MSCOCO 2014 (Chen et al. 2015), and each image is labelled with five human- annotated sentences. Eight objects (bottle, bus, couch, microwave, pizza, racket, suitcase and zebra) are selected as novel objects in NOC split. Correspondingly, all image- sentence pairs that include novel objects are removed from the standard training set. In the standard validation dataset, half of the pairs are randomly selected into new validation set, and others are selected into the test set. New validation and test sets are further separated into out-of-domain and in-domain subsets based on whether including positive ex- amples for eight novel objects. For the in-domain validation set and test set, there are no image-sentence pairs containing novel objects while images from out-of-domain sets in- clude novel objects. So in this split, models are required to describe images containing novel objects.

To load the data, please install the COCO API.

3.3 Evaluation Protocol

You should use all the three standard automatic evaluations metrics: CIDEr-D [9], METEOR [2] and SPICE[1]. F1 scores should also be reported for evaluating the performance of captioning eight novel concepts.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}.$$

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Zitian Chen, Yanwei Fu, Kaiyu Chen, and Yu-Gang Jiang. Image block augmentation for one-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3379–3386, 2019.

[4] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8680–8689, 2019.

[5] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016.

[6] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49, 2008.

[7] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018.

[8] Satoshi Tsutsui, Yanwei Fu, and David Crandall. Meta-reinforced synthetic data for one-shot fine-grained visual recognition. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[10] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1029–1037, 2018.