Image captioning using deep learning

Tomáš Železný[1]

1 Introduction

Image captioning is a popular machine learning task that deals with automatically generating descriptions for given images. It is an important task in many fields, such as image indexing used by search engines and digital libraries, or helping people with vision disabilities to better understand their surroundings. Image captioning is a very complex task. The system detects all the objects in the image, where they are, and what are their relations with other objects. Then it summarizes all this information and generates both semantically and syntactically correct sentence. In this work, I reproduce the results of the state-of-the-art image captioning method Oscar, presented in Li et al. (2020). Furthermore, I conduct an experiment to test its robustness.

2 Captioning pipeline

On its input, Oscar requires features extracted from a source image. These features need to be generated by an external object detector. Since Oscar's authors do not specify what detector should be used, I decided to use a Faster-R-CNN architecture implemented in the Detectron2 framework, presented in Wu et al. (2019), to extract the required image features.

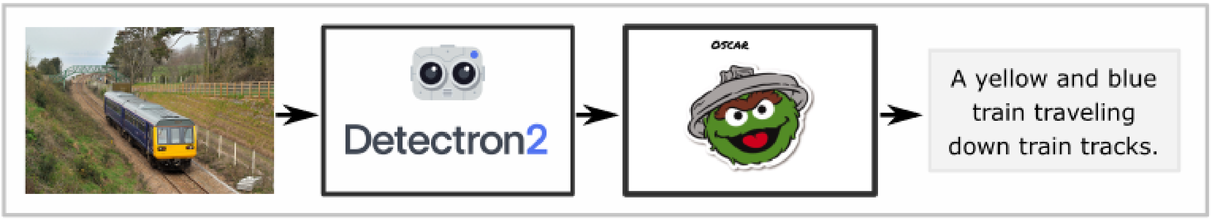

Figure 1: Visualization of the captioning pipeline

By connecting these two methods I create a pipeline that can generate a caption for any image. For training, I use the 2017 COCO Caption dataset. The pipeline was evaluated on the validation split of this dataset to achieve results which can be seen in Table 1.

Metric	BLEU-4	METEOR	CIDEr	SPICE
Original Oscar	**0.417**	**0.306**	**1.40**	**0.245**
My pipeline	0.312	0.272	1.02	0.201

Table 1: Performance of Oscar on COCO presented by Li et al. (2020), compared with performance of my Oscar-based pipeline.

[1] Master-degree student of Applied Sciences and Informatics, field of study Cybernetics and Control Engineering with specialization Artificial intelligence and biocybernetics. e-mail: zeleznyt@students.zcu.cz

3 Ablation study

Oscar is a multi-modal system. It uses a visual modality, represented by the image features of the objects detected in the image, and the textual modality, represented by the classes of the objects, denoted as object tags. It gives a motivation to conduct an ablation study to test Oscar's sensitivity and robustness to individual modalities. Based on the dataset with full information, I create two new variants of datasets, each with one modality removed, while preserving the other. These datasets were evaluated by the official COCO evaluation server[1]. The achieved results can be seen in Table 2.

Metric	B-1	B-2	B-3	B-4	M	R-L	C-D
Full information	0.697	0.526	0.393	0.296	0.265	0.531	0.977
Features removed	0.549	0.356	0.228	0.152	0.183	0.409	0.504
Tags removed	**0.672**	**0.498**	**0.366**	**0.272**	**0.250**	**0.511**	**0.875**

Table 2: B-1-4: BLEU-1-4, M: METEOR, R-L: ROUGE$_L$, C-D: CIDEr-D. The comparison between the different approaches when the system had access to all the information, when the visual modality was removed, and when the textual modality was removed.

4 Conclusion

Based on the state-of-the-art method, Oscar, I created the image captioning pipeline. Its performance is compared with original presented by Li et al. (2020) in Table 1. It can be seen that the original one outperforms my pipeline. It is an expected result due to the much lower variability of the training data. Although the performance is lower, I consider the result of my work a success. I managed to implement a pipeline that can generate a caption for any image using Oscar, which was not possible from public sources at the time I started my work.

The work further examines Oscar's sensitivity to individual modalities. The results of the experiment can be seen in Table 2. They suggest that Oscar is dependent on both modalities, with the visual modality predominating over the textual modality.

Acknowledgement

Computational resources were supplied by the project ”e-Infrastruktura CZ” (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L.Wang, H. Hu, L. Dong, F.Wei, et al., “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in *European Conference on Computer Vision, pp. 121–137, Springer, 2020.*

Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, “Detectron2.” `https://github.com/facebookresearch/detectron2`*, 2019.*

[1] https://competitions.codalab.org/competitions/3221