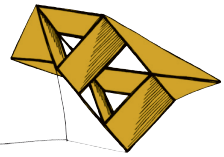

Exploring the Relationship between Dataset Size and Image Captioning Model Performance

Tomáš Železný[1]

1 Introduction

Image captioning is a deep learning task, which goal is to automatically generate textual description of an input image. It is a complex task that requires identifying and interpreting visual information and generating grammatically correct and fluent sentences. Because different individuals may consider various aspects of an image important, there isn't any single correct caption. This means that there is no ideal evaluation metric for measuring caption quality, as different metrics may better evaluate different attributes of the caption. Image captioning models, just like other deep learning models, need a large amount of training data and require a long time to train. In this work, we investigate the impact of using a smaller amount of training data on the performance of the standard image captioning model Oscar, introduced in Li et. al. (2020). We train Oscar on different sizes of the training dataset and measure its performance in terms of accuracy and computational complexity. In addition to traditional evaluation metrics, we evaluate the performance using CLIP similarity, introduced in Radford, et. al. (2021). We investigate whether it can be used as a fully-fledged metric providing a unique advantage over the traditional metrics; it does not need reference captions acquired by human annotators.

2 Impact of Different Volumes of Data on Model Performance

In this experiment, we evaluate the performance of the Oscar image captioning model on the COCO Captions dataset, presented in Chen et al. (2015). To assess the effect of training data size on model performance, we selected various amounts of data from the training set to train Oscar. Our results show that the computational time increases linearly with the amount of data used for training. However, the accuracy does not follow this linear trend and the relative improvement diminishes as we add more data to the training. We also measure the consistency of individual sizes of the training sets and observe that the more data we use for training the more consistent the scores are.

3 Evaluating Image Captioning with CLIP

In the second experiment, we investigate whether CLIP similarity can be used as a fully-fledged metric for evaluating image captioning tasks. Our analysis of the data obtained from first experiment revealed that CLIP exhibits behavior similar to that of other metrics. To further investigate this relationship, we calculated Pearson's correlation coefficient between all metrics across all subsets of the data. Our findings show that all metrics are highly correlated. This indicates the correct, consistent, and expected behavior of all the metrics.

[1] Student of the doctoral degree program Applied Sciences and Informatics, field of study Cybernetics, e-mail: zeleznyt@kky.zcu.cz

1 %	a dog laying on top of a bed.
10 %	a dog is laying on a bed in a room.
25 %	a dog sitting on a bed next to a person.
50 %	a dog sitting on a bed with clothes and a book.
100%	a dog sitting on a bed with a blanket and a pillow.

Figure 1: Examples of captions generated by models trained on different subset of the data. The size of subset is denoted as percentage of the original size of the training set. We can see the improvement of the caption as we add more data.

4 Conclusion

In our work, we conducted several experiments that show the relationship between the size of the datasets and the performance of image captioning model, Oscar. Based on the desired complexity of the captions, we may choose to use a smaller amount of data in future works. We believe our results can be transferred to other models, even in other deep-learning tasks. Furthermore, our experiments showed that CLIP can be used as a full-fledged metric with the big advantage of not needing reference captions.

Acknowledgement

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017. 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. A big thanks also goes to Ing. Marek Hrúz, Ph. D. and RNDr. Blanka Šedivá, Ph.D. for their valuable contributions.

References

Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. Springer International Publishing, 2020.

Chen, Xinlei, et al. "Microsoft coco captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325 (2015).

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.