Generating Diverse and Meaningful Captions

Annika Lindh[1,2], Robert J. Ross[1,2], Abhijit Mahalunkar[2], Giancarlo Salton[1,2] and John D. Kelleher[1,2]

[1] ADAPT Centre, Dublin, Ireland
[2] Dublin Institute of Technology (DIT), Ireland

Summary

In [1] we address the limitation of a lack of diversity in the caption output from Image Captioning models with an encoder-decoder architecture. While these models are successful at generating grammatically correct captions that score highly on the standard n-gram metrics for this task, they unfortunately demonstrate a preference for generic captions that avoid the more unique aspects of each image.

To quantify the problem, we identify a set of metrics capable of directly measuring the diversity in the model's output. We then demonstrate how to improve the diversity of a previously trained Image Captioning model through unsupervised specificity optimization guided by an Image Retrieval model. By optimizing for specificity, we achieve a meaningful increase in diversity with the objective to generate unambiguous captions that perform well on an Image Retrieval task. The quantitative result is an improvement over the previous state-of-the-art on two out of three diversity metrics. Qualitatively, we observe an increased attention to detail in the captions.

Our source code is made available online for the benefit of future research (https://github.com/AnnikaLindh/Diverse_and_Specific_Image_Captioning).

References

1. Lindh, A., Ross, R.J., Mahalunkar, A., Salton, G., Kelleher, J.D.: Generating Diverse and Meaningful Captions. In: Kurkova, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) Artificial Neural Networks and Machine Learning - ICANN 2018. pp. 176–187. Lecture Notes in Computer Science, Springer International Publishing (2018)