LAB2: Dialogue-based Image Retrieval

November 10, 2017

Project Description

In this project you will combine Natural Language Processing with Computer Vision for high-level scene interpretation. In particular, you will implement a system that, given a dialogue and a set of pictures, finds out to which specific picture the dialogue is referring to. Given pre-computed visual features and the dialogue, you will have to implement at the least two, incrementally more sophisticated, models to process the (textual) dialogue and to combine it with the visual information, in order to eventually spot the right picture.

Requirements

As a final product, you will be asked to write a report with your findings, which should at least contain:

- A background section, in which you write about techniques that connect language and vision (e.g., visual question answering, text-based image retrieval, visual dialogue, etc) and the problem that you are trying to address;
- A description of the model that you use, and of its individual components;
- A summary of your models' learning behavior, including learning curves and hyper-parameter search;
- A qualitative analysis of each model by showing and discussing (interesting) correctly and wrongly classified examples;
- A systematic comparison of the models you trained, including qualitative measures such as top1 and top5 accuracy; qualitative analysis as in previous point, but where the analysis is conducted between different models.
- A section where you discuss future work based on your experience and what you think could significantly improve performance (but you didn't find the time to investigate);
- A conclusion/summary.

Getting Started

First, we recommend you to read [1]. Then, to get started explore the data.

References

[1] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.