Project Proposal - Image captioning with region attention

Henrique Gasparini Fiuza do Nascimento
henrique.gasparini_fiuza_do_nascimento@ens-paris-saclay.fr

Summary of the Proposal

In this project, we work on recent state-of-the-art approaches for image captioning that rely on attention to image regions to improve predictions and make the results more interpretable. The goal of this project is to assess the strength and limitations of these current models. In particular, we will evaluate these methods on hard-to-describe images such as images from UnRel dataset, in order be able to see the limitations of the models.

Overview

We study mainly the works of Anderson et al[1] and Lu et al[2], who have recently proposed state-of-the-art models for image captioning.

Goal and Objectives

Our goal in this project is to evaluate recent state-of-the-art models for image captioning, by assessing their strength and limitations on difficult data-sets.

We focus our analysis on the "Robust Image Captioning" (in which the distributions of train and test sets are different) and "Novel Object Captioning" (test set contains out-of-domain images) setups. In particular, we inspire ourselves from the results and analysis of Lu et al's model presented in their paper.

Methods

We qualitatively assess wrong predictions (or poor captions, in our case) by looking at the attended regions proposed by the attention mechanism. We also assess the models quantitatively in multiple data-sets.

We expect to get the most interesting results from analyzing difficult data-sets, which are the ones that contain less usual images.

Data-sets

We first evaluate both models on the data-set in which they were trained (a combination of the Visual Genome Dataset and Microsoft COCO Dataset for Anderson et al's model; Flickr30k and Microsoft COCO separately for Lu et al's models) and on the Unrel dataset[3], which proposes unusual image-captioning relations.

References

[1] P. Anderson, X. He, C. Buehler, M. Johnson D. Teney, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *CVPR*, 2018.

[2] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. *CVPR*, 2018.

[3] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. *CVPR*, 2018.