Review 1

Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models

Summary

This paper provides an encoder-decoder model to embed images and text into same space, and then decode distributed representations from the space. This model makes following three challenging tasks possible, 1) accurate image-sentence ranking, 2) multimodal linguistic regularities and 3) original image caption generation.

This paper is organized as follows: In Section 1, the paper gives why generating descriptions for images is important and challenging. Then it illustrates and compares the basic idea illustrates of state-of-the-art methods in image caption, multimodal representation learning and neural machine translation (NMT). In Section 2, details of the encoder-decoder model for ranking and generation are given. Firstly, the paper introduces the multimodal distributed representation models which can embed images and their text description in to an embedding space. Secondly, to make use of context information of a sentence, the paper presents the log-bilinear neural language models. Then the paper gives a description about multiplicative neural language models that model the distribution and predict the next word based on both context information and embedding vector. Finally, the paper describes the structure-content neural language models (SC-NLM) which are derived from the multiplicative variant. To generate grammatically correct and reasonable sentence, the SC-NLM builds the attribute vector by combining forward structure (part-of-speech tags) and previous content information. In Section 3, the paper presents experiments quantitatively evaluating on image-sentence ranking, multimodal linguistic regularities and image caption generation tasks.

Main Contribution

The most important contribution of this paper is improving regular multiplicative neural language models into the structure-content neural language models. The SC-NLM takes both content and structure into consideration, which is inspired by a machine translation method.

Pros and Cons

Pros: 1) Give more reasonable description due to considering both structure and content information. 2) Can be trained purely on a large amount of texts alone instead of image-caption pairs. 3) Push the performance boundary of the image-sentence ranking task.

Cons: N/A

Evaluation

I am considering if we should conduct an additional experiment in which we train SC-NLM in different ways (i.e. text only, images only and image-caption pairs). By doing this, we can prove the advantage of the shared embedding space. If the performance is different, we can try to figure out which information missing could lead poor performance.