UROP Proposal

Title:

Evaluating the Visual Fidelity of Image Description Based on Fluency-based Word Mover's Distance

Proposal Main content:

It's a popular task at the intersection of computer vision and natural language, a task which is called ***Image Description Generation*** (IDG*)*. In a simple way to illustrate this task, we can regard it as a way of generating output sentences which describes the visual contents of the given input image. Although there are currently many ways to achieve task, evaluating the generated output sentence, the visual content of the given input image, is still an understudied problem.

The evaluation of the *IDG* currently is based on two ways: (1) ***Human Judgement***, (2) ***Automatic metrics***. For the (1) ***Human Judgement*** way, it evaluates either the overall quality of descriptions or specific criteria in isolation, but it's both expensive to scale and highly subjective. For the (2) ***Automatic Metrics*** way, it addresses the problem of scalability by comparing candidate descriptions against human-authored reference description. It's true that the existing ***Automatic Metrics*** are useful for measuring the quality of descriptions as a whole, but it's difficult for specific capabilities of ***IDG*** systems to be inspected.

In order to deal with the problem of the ***Automatic Metrics***, we focus on one criterion – ***Visual Fidelity***, which is a fine-grained metric measuring criteria and would be useful to understand how an IDG system is better than another. The ***Visual Fidelity*** aims to measure how faithful a description is with respect to what is depicted in the image. For that, in the previous work, it is proposed to both take image content into account when evaluating the visual content descriptions, in contrast to only rely on the words in the reference descriptions. Due to the reason that there is no existing metric for IDG has image factored explicitly into evaluation process, it's important to build an automatic evaluation metric for IDG that measures the fidelity of image description. [1]

In the previous work, the task of building the automatic evaluation metric for IDG that measures the fidelity of image description with respect to the image using information derived from images directly as 'reference ', a task which is named as ***visual fidelity of image description*** (VIFIDEL) has been successfully achieved. [1] VIFIDEL, in the previous work, can be either (1) used based on images as reference, or (2) in conjunction with textual references to take into account the relevant image content people describe. At the same time, VIFIDEL, in the previous work, is able to perform matching of images and text in an embedding space by building on the ***Word Mover's Distance metric*** (WMD). [1]

Our first contribution is to make change to the technique - standard WMD metric, which is used in the previous work for performing the match of images and text in an embedding space, to the ***Fluency-based Word Mover's distance metric*** (WMD0). By applying the ***Fluency-based Word Mover's distance metric,*** once we are evaluating the ***IDG*** system result, word order is taken into account and the metric outperforms the standard WMD.[2] Our second contribution is to make modification to ***Fluency-based Word Mover's distance metric*** word embedding from ***word2vec*** to ***contextualized word-embedding***, such as ***BERT***. Our third contribution is taking modification to the ***WMD0*** translation schema, which is instead of considering the translation evaluation based on reference sentence, we are focusing on the translation evaluation based on the source sentence. [2]

Reference:

[1] https://arxiv.org/pdf/1907.09340.pdf
[2] https://www.aclweb.org/anthology/W19-5356.pdf