Simple Baseline for Visual Question Answering

CV • VQA • AI • NLP • 2015

28 Apr 2017

Problem Statement

- VQA Task: Given an image and a free-form, open-ended, natural language question (about the image), produce the answer for the image.
- The paper attempts to fine tune the simple baseline method of Bag-of-Words + Image features (iBOWIMG) to make it competitive against more sophisticated LSTM models.
- Link to the paper

Model

- VQA modelled as a classification task where the system learns to choose among one of the top k most prominent answers.
- **Text Features** - Convert input question to a one-hot vector and then transform to word vectors using a word embedding.
- **Image Features** - Last layer activations from GoogLeNet.
- Text features are concatenated with image features and fed into a softmax.
- Different learning rates and weight clipping for word embedding layer and softmax layer with the learning rate for embedding layer much higher than that of softmax layer.

Results

- iBOWIMG model reports an accuracy of 55.89% for Open-ended questions and 61.97% for Multiple-Choice questions which is comparable to the performance of other, more sophisticated models.

Interpretation of the model

- Since the model is very simple, it is possible to interpret the model to know what exactly is the model learning. This is the greatest strength of the paper even though the model is very simple and naive.
- The model attempts to memorise the correlation between the answer class and the informative words (in the question) and image features.
- Question words generally can influence the answer given the bias in images occurring in COCO dataset.
- Given the simple linear transformation being used, it is possible to quantify the importance of each single words (in the question) to the answer.
- The paper uses the Class Activation Mapping (CAM) approach (which uses the linear relation between softmax and final image feature map) to highlight the informative image regions relevant to the predicted answer.
- While the results reported by the paper are not themselves so significant, the described approach provides a way to interpret the strengths and weakness of different VQA datasets.

Related Posts

Hints for Computer System Design 07 Jan 2022

Synthesized Policies for Transfer and Adaptation across Tasks and Environments 29 Mar 2021

Deep Neural Networks for YouTube Recommendations 22 Mar 2021

0 Comments **papers-I-read** **Privacy Policy** 1 **Login**

♡ **Favorite** Tweet Share **Sort by Best**

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Be the first to comment.

✉ **Subscribe** **Add Disqus to your siteAdd DisqusAdd** ⚠ **Do Not Sell My Data**