Making the V in VQA Matter - Elevating the Role of Image Understanding in Visual Question Answering

2017 • AI • CV • Dataset • NLP • VQA

14 May 2017

Problem Statement

- Standard VQA models benefit from the inherent bias in the structure of the world and the language of the question.
- For example, if the question starts with "Do you see a ...", it is more likely to be "yes" than "no".
- To truly assess the capability of any VQA system, we need to have evaluation tasks that require the use of both the visual and the language modality.
- The authors present a balanced version of VQA dataset where each question in the dataset is associated with a pair of similar images such that the same question would give different answers on the two images.
- The proposed data collection procedure enables the authors to develop a novel interpretable model which, given an image and a question, identifies an image that is similar to the original image but has a different answer to the same question thereby building trust for the system.
- Link to the paper

Dataset Collection

- Given an (image, question, answer) triplet (I, Q, A) from the VQA dataset, a human worker (on AMT) is asked to identify an image I' which is similar to I but for which the answer to question Q is A' (different from A).

- To facilitate the search for I', the worker is shown 24 nearest-neighbor images of I (based on VGGNet features) and is asked to choose the most similar image to I, for which Q makes sense and answer for Q is different than A. In case none of the 24 images qualifies, the worker may select "not possible".
- In the second round, the workers were asked to answer Q for I'.
- This 2-stage protocol results in a significantly more balanced dataset than the previous dataset.

Observation

- State-of-the-art models trained on unbalanced VQA dataset perform significantly worse on the new, balanced dataset indicating that those models benefitted from the language bias in the older dataset.
- Training on balanced dataset improves performance on the unbalanced dataset.
- Further, the VQA model, trained on the balanced dataset, learns to differentiate between otherwise similar images.

Counter-example Explanations

- Given an image and a question, the model not only answers the question, it also provides an image (from the k nearest neighbours of I, based on VGGNet features) which is similar to the input image but for which the model would have given different answer for the same image.
- Supervising signal is provided by the data collection procedure where humans pick the image I' from the same set of candidate images.
- For each image in the candidate set, compute the inner product of question-image embedding and answer embedding.
- The K inner product values are passed through a fully connected layer to generate K scores.
- Trained with pairwise hinge ranking loss so that the score of the human picked image is higher than the score of all other images by a margin of M (hyperparameter).

- The proposed explanation model achieves a recall@5 of 43.49%

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