VQA-Visual Question Answering

2015 • ICCV 2015 • AI • CV • Dataset • ICCV • NLP • VQA

27 Apr 2017

Problem Statement

- Given an image and a free-form, open-ended, natural language question (about the image), produce the answer for the image.

- Link to the paper

VQA Challenge and Workshop

- The authors organise an annual challenge and workshop to discuss the state-of-the-art methods and best practices in this domain.
- Interestingly, the second version is starting on 27th April 2017 (today).

Benefits over tasks like image captioning:

- Simple, *n-gram* statistics based methods are not sufficient.
- Requires the system to blend in different aspects of knowledge - object detection, activity recognition, commonsense reasoning etc.
- Since only short answers are expected, evaluation is easier.

Dataset

- Created a new dataset of 50000 realistic, abstract images.
- Used AMT to crowdsource the task of collecting questions and answers for MS COCO dataset (>200K images) and abstract images.
- Three questions per image and ten answers per question (along with their confidence) were

collected.
- The entire dataset contains over 760K questions and 10M answers.
- The authors also performed an exhaustive analysis of the dataset to establish its diversity and to explore how the content of these question-answers differ from that of standard image captioning datasets.

Highlights of data collection methodology

- Emphasis on questions that require an image, and not just common sense, to be answered correctly.
- Workers were shown previous questions when writing new questions to increase diversity.
- Answers collected from multiple users to account for discrepancies in answers by humans.
- Two modalities supported:
  - **Open-ended** - produce the answer
  - **multiple-choice** - select from a set of options provided (18 options comprising of popular, plausible, random and ofc correct answer)

Highlights from data analysis

- Most questions range from four to ten words while answers range from one to three words.
- Around 40% questions are “yes/no” questions.
- Significant (>80%) inter-human agreement for answers.
- The authors performed a study where human evaluators were asked to answer the questions without looking at the images.
- Further, they performed a study where evaluators were asked to label if a question could be answered using common sense and what was the youngest age group, they felt, could answer the question.
- The idea was to establish that a sufficient number of questions in the dataset required more than just common sense to answer.

Baseline Models

- **random** selection
- **prior (“yes”)** - always answer as yes.

- **per Q-type prior** - pick the most popular answer per question type.
- **nearest neighbor** - find the k nearest neighbors for the given (image, question) pair.

Methods

- 2-channel model (using vision and language models) followed by softmax over (K = 1000) most frequent answers.

- **Image Channel**
    - **I** - Used last hidden layer of VGGNet to obtain 4096-dim image embedding.
    - **norm I** - : l2 normalized version of **I**.

- **Question Channel**
    - **BoW Q** - Bag-of-Words representation for the questions using the top 1000 words plus the top 1- first, second and third words of the questions.
    - **LSTM Q** - Each word is encoded into 300-dim vectors using fully connected + tanh non-linearity. These embeddings are fed to an LSTM to obtain 1024d-dim embedding.
    - **Deeper LSTM Q** - Same as LSTM Q but uses two hidden layers to obtain 2048-dim embedding.

- **Multi-Layer Perceptron (MLP)** - Combine image and question embeddings to obtain a single embedding.
    - **BoW Q + I** method - concatenate BoW Q and I embeddings.
    - **LSTM Q + I, deeper LSTM Q + norm I** methods - image embedding transformed to 1024-dim using a FC layer and tanh non-linearity followed by element-wise multiplication of image and question vectors.

- Pass combined embedding to an MLP - FC neural network with 2 hidden layers (1000 neurons and 0.5 dropout) with tanh, followed by softmax.
- Cross-entropy loss with VGGNet parameters frozen.

Results

- Deeper LSTM Q + norm I is the best model with 58.16% accuracy on open-ended dataset and 63.09% on multiple-choice but far behind the human evaluators (>80% and >90% respectively).
- The best model performs well for answers involving common visual objects but performs poorly for answers involving counts.
- Vision only model performs even worse than the model which always produces "yes" as the answer.

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