Exploring Models and Data for Image Question Answering

2015 • NIPS 2015 • AI • CV • Dataset • NIPS • NLP • VQA

14 Jan 2018

Introduction

- **Problem Statement**: Given an image, answer a given question about the image.
- Link to the paper
- **Assumptions**:
    - The answer is assumed to be a single word thereby bypassing the evaluation issues of multi-word generation tasks.

VIS-LSTM Model

- Treat the input image as the first word in the question.
- Obtain the vector representation (skip-gram) for words in the question.
- Obtain the VGG Net embeddings of the image and use a linear transformation (dimensionality reduction weight matrix) to match the dimensions of word embeddings.
- Keep image embedding frozen during training and use an LSTM to combine the word vectors.
- LSTM outputs are fed into a softmax layer which generates the answer.

Dataset

- DAtaset for QUestion Ansering on Real-world images (DAQUAR)
    - 1300 images and 7000 questions with 37 object classes.

    - Downside is that even guess work can yield good results.

- The paper proposed an algorithm for generating questions using MS-COCO dataset.
    - Perform preprocessing steps like breaking large sentences and changing indefinite determines to definite ones.
    - *object* questions, *number* questions, *colour* questions and *location* questions can be generated by searching for nouns, numbers, colours and prepositions respectively.
    - Resulting dataset has ~120K questions across above 4 semantic types.

Models

- VIS+LSTM - explained above
- 2-VIS+BLSTM - Add the image features twice, in beginning and in the end (using different linear transformations) plus use bidirectional LSTM
- IMG+BOW - Multinomial logistic regression on image features without dimensionality reduction + bag of words (averaging word vectors).
- FULL - Simple average of above 2 models.

Baseline

- Includes models where the answer is guessed, or only image or question features are used or image features along with prior knowledge of object are used.
- Also includes a KNN model where the system finds the nearest (image, question) pair.

Metrics

- Accuracy
- Wu-Palmer similarity measure

Observations

- The VIS-LSTM model outperforms the baselines while the FULL model benefits from averaging across all the models.

- Some useful information seems to be lost when downsizing the VGG vectors.
- Fine tuning the word vectors helps with performance.
- Normalising CNN hidden image features into zero mean and unit variance leads to faster training.
- Model does not perform well on the task of considering spatial relations between multiple objects and counting objects when multiple objects are present

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