VISUAL QUESTION ANSWERING

Avi Singh
October 26, 2015

IIT-Kanpur
The Problem Statement

We want to answer open-ended questions about images.

Figure: A teaser from the VQA dataset [Antol et al., 2015]
Motivation

Visual Turing Test

An AI-complete task [Malinowski and Fritz, 2014b]. The specificity of the questions enable automatic evaluation.

Helping the visually impaired

Apps like VizWiz [Bigham et al., 2010] employ humans to answer visual questions sent by visually impaired people.
DATASETS

VQA (VATech) [Antol et al., 2015]
750K questions on 250K images, 10 answers for every question.

Visual Madlibs (UNC) [Yu et al., 2015]
360K questions on 10K images. Lot of "high level" questions.

Toronto COCO-QA [Ren et al., 2015]
Automatically generated questions from COCO captions. 115K question. Now obsolete.

DAQUAR [Malinowski and Fritz, 2014a]
Much smaller dataset with 12K questions Now obsolete.
MODELS
The Baseline BOW Model [Ren et al., 2015]

1. Use word2vec [Mikolov et al., 2013] to extract bag of word features.
2. Use VGG ConvNet [Simonyan and Zisserman, 2014] to extract features from image.
3. Treat the problem as multi-class classification.
1. Reduce dimensionality of image features (down to the word vector dimensionality) and feed this into the LSTM.
2. Use word2vec[Mikolov et al., 2013] to convert every word to a vector, which is then fed to the LSTM.
3. Make predictions after the last word has been fed.
1. VGGNet-based feature extraction pipeline for images complete.
2. Word2Vec-based feature extraction pipeline text for text complete.
3. A baseline model (multinomial logistic regression with lbfgs for optimization) trained on 20K questions and evaluated on 10K questions, performance only 16% so far.
1. Semantic alignment between questions and images [Karpathy and Fei-Fei, 2014][Karpathy and Fei-Fei, 2015].

2. Use LSTM to encode questions, and decode answers [Sutskever et al., 2014]

3. Neural Net architectures like Memory Networks [Sukhbaatar et al., 2015]

4. Visual Attention [Xu et al., 2015].


Questions?