VISUAL QUESTION ANSWERING

Avi Singh

October 26, 2015

IIT-Kanpur

The Problem Statement

We want to answer open-ended questions about images.

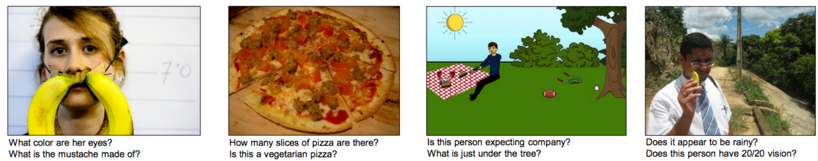

Figure: A teaser from the VQA dataset[Antol et al., 2015]

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.

LSTM-based Model[Ren et al., 2015]

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.

Work done so far

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.

Ideas to Explore

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].

References I

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015).
VQA: Visual Question Answering.
International Conference on Computer Vision (ICCV).

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. (2010).
Vizwiz: Nearly real-time answers to visual questions.
In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 333–342, New York, NY, USA. ACM.

References II

Karpathy, A. and Fei-Fei, L. (2014).
Deep fragment embeddings for bidirectional image-sentence mapping.
Neural Information Processing Systems (NIPS).

Karpathy, A. and Fei-Fei, L. (2015).
Deep visual-semantic alignments for generating image descriptions.
Computer Vision and Pattern Recognition (CVPR).

Malinowski, M. and Fritz, M. (2014a).
A multi-world approach to question answering about real-world scenes based on uncertain input.
Neural Information Processing Systems (NIPS).

References III

Malinowski, M. and Fritz, M. (2014b).
Towards a visual turing challenge.
NIPS Workshop.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013).
Distributed Representations of Words and Phrases and their Compositionality.
Neural Information Processing Systems (NIPS).

Ren, M., Kiros, R., and Zemel, R. S. (2015).
Exploring models and data for image question answering.
Neural Information Processing Systems (NIPS).

References IV

Simonyan, K. and Zisserman, A. (2014).
Very deep convolutional networks for large-scale image recognition.
CoRR, abs/1409.1556.

Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015).
End-to-end memory networks.
CoRR, abs/1503.08895.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to sequence learning with neural networks.
Neural Information Processing Systems (NIPS), abs/1409.3215.

References V

- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015).
  Show, attend and tell: Neural image caption generation with visual attention.
  arXiv preprint arXiv:1502.03044.
- Yu, L., Park, E., Berg, A. C., and Berg, T. L. (2015).
  Visual Madlibs: Fill in the blank Image Generation and Question Answering.
  arXiv preprint arXiv:1506.00278.

Questions?