Using Explanations to Improve Ensembling of Visual Question Answering Systems

Nazneen Rajani and Ray Mooney

nrajani@cs.utexas.edu and mooney@cs.utexas.edu

Dept. of Computer Science at the University of Texas at Austin

INTRODUCTION

- Deep learning systems have been groundbreaking but lack **transparency**
- **Explanations** help understand and interpret decisions made by deep neural networks
- We **trust** systems' agreement on an answer more if they also agree on its explanation

Visual Question Answering (VQA)

Q. Is that a frisbee?
A. Yes
Q. Is this a man or a woman?
A. Woman
Q. What color is the frisbee?
A. Red

Q. Is this a romantic spot that couples would like to go?
A. Yes
Q. What time of day is it?
A. Night
Q. How many spires below big ben's clock?
A. 10

Figure 1: Sample from the VQA dataset

Stacking With Auxiliary Features (SWAF)

(Rajani and Mooney, IJCAI 2017)

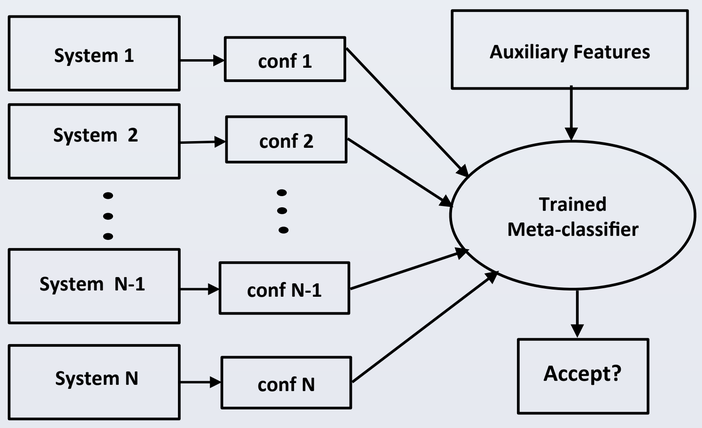

Figure 2: Given an input, the ensemble judges every possible question-answer pair produced by the component systems and determines the final output answer

- SWAF combines information about **component systems** and the **problem** using auxiliary features
- We use **visual explanations** to create auxiliary features for SWAF

EXPLANATION

- Recent VQA research shows that deep learning models **attend** to **relevant** parts of image while answering the question (Goyal *et al*., 2016)
- The parts of images that the models focus on can be though of as **visual explanations**
- We use **heat-maps** to visualize explanation in images'

Figure 3: On the left is an image from the VQA dataset and on the right is the heat-map overlaid on the image for the question - 'What is the man eating?'

- **GradCAM** (Goyal *et al*., 2016) is used to generate heat-maps
- The gradients are set to zero for all categories except the one under considerations
- The signal is backpropagated to the convolutional feature maps of interest to compute the heat-map

AUXILIARY FEATURES

- **Explanation**
- Use rank order correlation to compare heat-maps of systems for a given image-question pair
- ${}^{n}C_{2}$ explanation agreement features for n component systems

- **Question & Answer types**
- Use prefix substrings of questions to form feature vectors of question types (70 in total)
- Example: "What is the color of the vase?" has the following types "What", "What is", "What is the", "What is the color", "What is the color of"
- Answer types -- questions beginning with "Does","Is","Was","Are", and "Has" categorized as "yes/no" type; questions beginning with "How many", "What time", "What number" are assigned "number" type; rest are assigned the "other" type

- **Question features**
- Use a bag-of-words language model to represent the question as auxiliary features

- **Image features**
- Deep visual features obtained from VGGNet's *fc7* layer
- Total of 4096 features

Component Systems

- **LSTM** (Antol *et al.*, 2015)
- LSTM for question with CNN for images
- VGGNet image features combined with one-hot encoding of the words using element-wise multiplication

- **Hierarchical Co-Attention (HieCoAtt)** (Lu *et al.*, 2016)
- Jointly reasons about visual and language component using co-attention
- Hierarchical architecture at word, phrase and question levels

- **Multimodal Compact Bilinear pooling (MCB)** (Fukui *et al*., 2016)
- Similar to LSTM but uses outer product instead of element-wise product
- 152-layer ResNet instead of CNN for images
- Image and question vectors are pooled using MCB

RESULTS

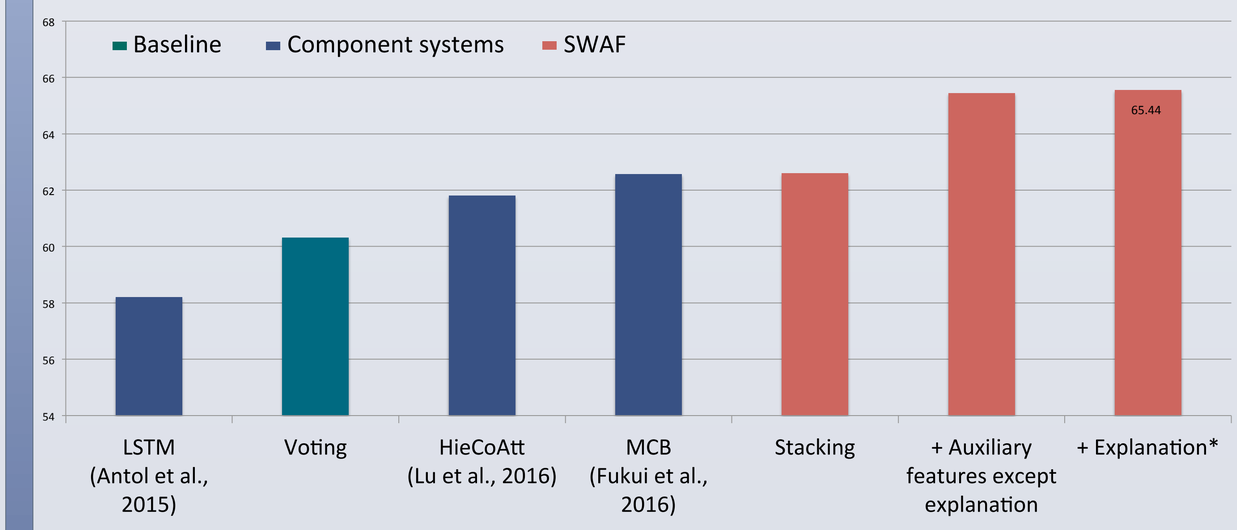

Figure 4 : Results on the VQA open-ended test set (except for explanation)

REFERENCES

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. ICCV 2015.
- Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal Compact Bilinear pooling for Visual Question Answering and Visual Grounding. EMNLP 2016.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. NIPS 2016.
- Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards Transparent AI Systems: Interpreting Visual Question Answering Models. ICML 2016.
- Nazneen Fatema Rajani and Raymond J. Mooney. StackingWith Auxiliary Features. IJCAI 2017.