Stacking with Auxiliary Features for VQA

Nazneen Fatema Rajani and Raymond J. Mooney

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

Department of Computer Science at the University of Texas at Austin

Introduction

Visual Question Answering

- Visual Question Answering (VQA) (Antol *et al*., 2015) requires both **language and image understanding**, language grounding capabilities, as well as common-sense knowledge (See Figure 1 for examples).
- The vision component of a typical VQA system extracts visual features using a deep **convolutional neural network** (CNN), and the linguistic component encodes the question into a semantic vector using a **recurrent neural network** (RNN).
- Most VQA systems have a single underlying method that optimizes a specific loss function and do not leverage the advantage of using multiple diverse models.
- The various VQA models have learned to perform well **on specific types of questions and images**. Therefore, there is an opportunity to combine these models intelligently so as to **leverage** their diverse strengths.

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: Random sample of images with questions and ground truth answers taken from the VQA dataset.

Algorithm

Stacking with Auxiliary Features (SWAF)

- SWAF (Rajani and Mooney, 2017) is a recent **ensembling algorithm** that learns to combine outputs of multiple systems using features of the current problem as context (Figure 2).
- We demonstrate SWAF on VQA using **four** different categories of novel auxiliary features:
  - ✓ Question and Answer types
  - ✓ Question features
  - ✓ Image features
  - ✓ Explanation as features

  Inferred from image-question pair
- If a question-answer pair is classified as correct by the stacker, and if there are other answers that are also classified as correct for the same question, the output with the highest meta-classifier confidence is chosen.
- For questions that do not have any answer classified as correct by the stacker, we choose the answer with lowest classifier confidence, which means it is least likely to be incorrect.
- We obtain **SOTA** on the VQA 2016 challenge.

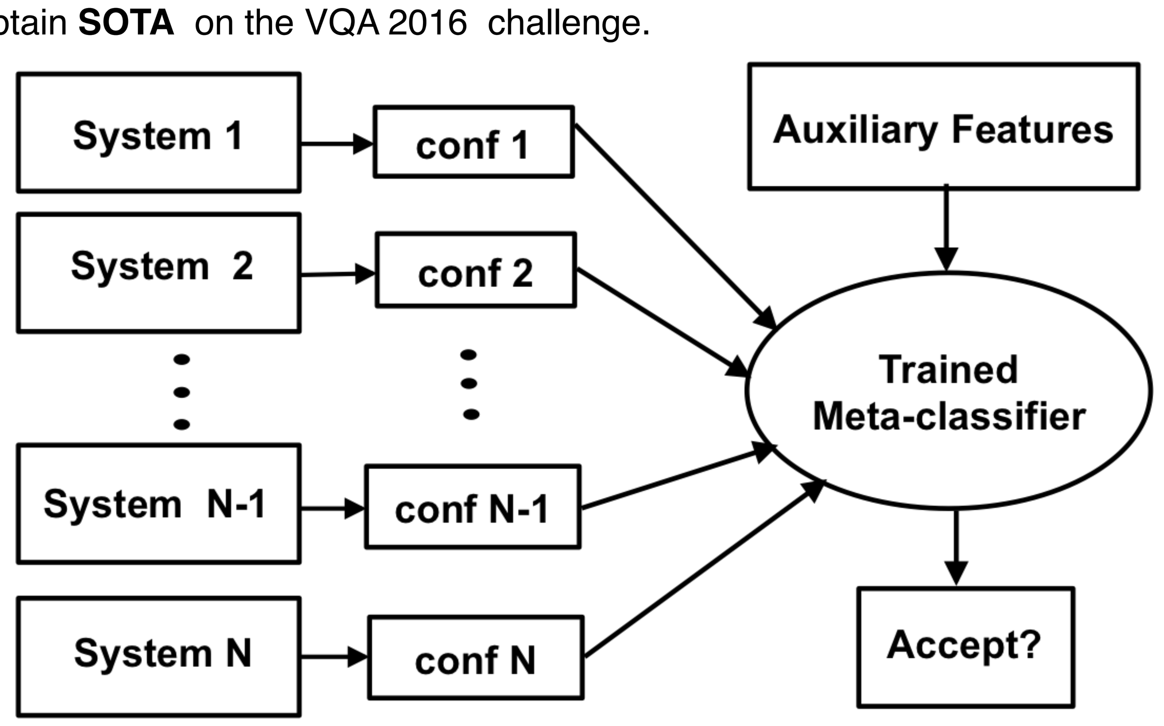

Figure 2:. Ensemble Architecture using Stacking with Auxiliary Features. Given an input, the ensemble judges every possible question-answer pair produced by the component systems and determines the final output answer.

Auxiliary Features

Question and Answer Types

- Some VQA models are better at certain QA types than others and this information can be used by the stacker at classification time.
- **Prefixes** of question defined a type, e.g. "What", "What is", "What is the".
- Question type with at least 500 questions and a separate "other" type gave a total of 70 question types which were used as vector of features.
- Infer answer type from the question – "yes/no", "number" and "other" types.
  - "Does", "Is", "Was", "Are", "Has" classified as yes/no type.
  - "How many", "What time", "What number" classified as number type.

Question Features

- **Bag-of-Words** of the tokens that occur at least 5 times in the questions (on a validation set) are used as features.
- Including a BOW for the question as auxiliary features equip the stacker to efficiently learn which words are important and can aid in classifying answers.

Image Features

- Deep visual features of the **VGGNet's *fc7*** layer contributed 4096 features.
- Using such image features enables the stacker to learn to rely on systems that are good at identifying answers for particular types of images.

Using Explanation as features

- The localization-map generated using **GradCAM** (Selvaraju et al., 2017) by each VQA model serves as a visual explanation for the predicted output of that model (Figure 3).
- We take the **absolute gray-scale value** of the localization-maps in of each model and rank the pixels according to their spatial attention intensity.
- Then, we compute the correlation between the two ranked lists using the **Spearman's rank-order correlation** with the localization-map of every other model.
- The total number of explanation agreement features thus generated is $^{n}C_{2}$ where n is the total number of component systems.

Figure 3: Each row from left to right shows an image-question pair from the VQA dataset along with localization maps overlaid on the image generated by the LSTM, HieCoAtt and MCB models respectively. The answers shown are those predicted by our ensemble.

Component VQA systems

We use three diverse individual VQA models as part of our ensemble system:

- **LSTM** (Antol et al., 2015): This model uses VGGNet to embed the image and two layer LSTM to embed the question. The image and question vectors are fused via element-wise multiplication.
- **HieCoAtt** (Lu et al., 2016): This model jointly reasons about the visual and language components using two types of "co-attention" – parallel and alternating.
- **MCB** (Fukui et al., 2016): This model uses the 152-layer ResNET network to embed the image and LSTM to embed the question. The two vectors are combined using the outer product which is made efficient using the multimodal compact bilinear pooling (Gao et al., 2016). We used the single system MCB model as a component in our ensemble.

The top performing VQA system in the 2016 competition was an ensemble of 7 MCB models. Their model is pre-trained on the Visual Genome dataset and they concatenate learned word embedding with pre-trained GloVe vectors.

Results

- We used a neural network as the meta-classifier implemented in Keras.
- We found that using **late fusion** for combining auxiliary features worked better.

Method	All	Yes/No	Number	Other
DPPNet (Noh et al., 2016)	57.36	80.28	36.92	42.24
iBOWIMG (Zhou et al., 2015)	55.72	76.55	35.03	42.62
NMNs (Andreas et al., 2016b)	58.70	81.20	37.70	44.00
LSTM (Antol et al., 2015)	58.20	80.60	36.50	43.70
HieCoAtt (Lu et al., 2016)	61.80	79.70	38.70	51.70
MCB (Single system) (Fukui et al., 2016)	62.56	80.68	35.59	52.93
MCB (Ensemble) (Fukui et al., 2016)	66.50	**83.20**	39.50	58.00
Voting (MCB + HieCoAtt + LSTM)	60.31	80.22	34.92	48.83
Stacking	63.12	81.61	36.07	53.77
+ Q/A type features	65.25	82.01	36.50	57.15
+ Question features	65.50	82.26	38.21	57.35
+ Image features	65.54	82.28	38.63	57.32
+ Explanation features	**67.26**	82.62	**39.50**	**58.34**

Table 1: Accuracy results on the VQA test-standard set. The first block shows performance of a VQA model that use external data for pre-training, the second block shows single system VQA models, the third block shows an ensemble VQA model that also uses external data for pre-training, and the fourth block shows ensemble VQA models.

- We observed that deleting the **Q/A type features** decreased performance the most and deleting the **explanation features** decreased performance the least.

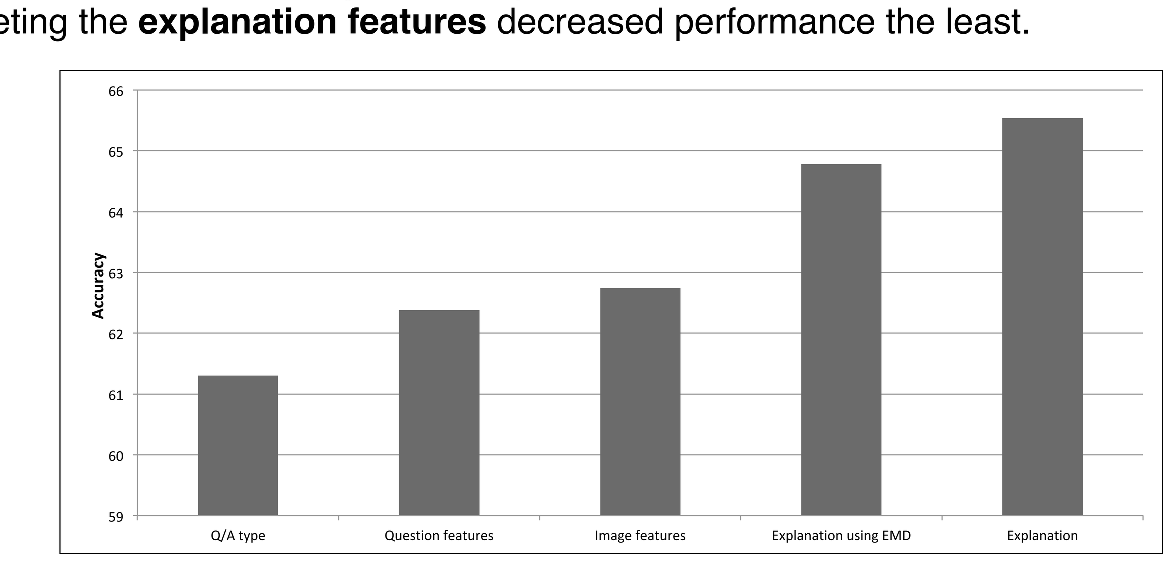

Figure 4: Results for auxiliary feature ablations on the VQA test-dev set. The x-axis indicates the feature set that was ablated from the final ensemble.

References and Acknowledgements

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. ICCV 2015.
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. NAACL 2016.
- 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.
- Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. CVPR 2016.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. NIPS 2016.
- Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. CVPR 2016.
- Nazneen Fatema Rajani and Raymond J. Mooney. Stacking With Auxiliary Features. IJCAI 2017.
- Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradientbased Localization. ICCV 2017.
- Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for Visual Question Answering. arXiv 2015.

This research was supported by the DARPA XAI grant