In Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), pp. 43-47, Melbourne, Australia, August 2017.

Using Explanations to Improve Ensembling of Visual Question Answering Systems

Nazneen Fatema Rajani
Department of Computer Science
University of Texas at Austin
nrajani@cs.utexas.edu

Raymond J. Mooney
Department of Computer Science
University of Texas at Austin
mooney@cs.utexas.edu

Abstract

We present results on using explanations as auxiliary features to improve stacked ensembles for Visual Question Answering (VQA). VQA is a challenging task that requires systems to jointly reason about natural language and vision. We present results applying a recent ensembling approach to VQA, Stacking with Auxiliary Features (SWAF), which learns to combine the results of multiple systems. We propose using features based on explanations to improve SWAF. Using explanations we are able to improve ensembling of three recent VQA systems.

1 Introduction

In recent years, deep-learning has led to unprecedented breakthroughs in many avenues of Artificial Intelligence and most notably in computer vision. Even though the results produced by these deep networks have been groundbreaking, they lack transparency, making them hard to understand and interpret [Lipton, 2016]. Consequently, when such intelligent models that provide no explanation for their decisions fail, it becomes very difficult to do any root cause analysis. Transparency is also important in order to build human trust in systems. Recently, there has been some work by the deep learning community on generating explanations as a way to better understand and interpret the decisions made by deep neural networks [Hendricks *et al.*, 2016; Goyal *et al.*, 2016; Selvaraju *et al.*, 2016].

Visual Question Answering (VQA) is a challenging task that requires systems to attend to regions of an image or question or both for producing an output. VQA addresses open-ended questions about images [Antol *et al.*, 2015] and has attracted significant attention in the past year [Andreas *et al.*, 2016; Goyal *et al.*, 2016; Agrawal *et al.*, 2016]. It requires visual and linguistic comprehension, language grounding as well as commonsense knowledge. A variety of methods to address these challenges have been developed in recent years [Fukui *et al.*, 2016; Xu and Saenko, 2016; Lu *et al.*, 2016; Chen *et al.*, 2015]. 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). An answer is then generated conditioned on the visual features and the question vector. Some VQA models have an explicit attention component in the architecture [Fukui *et al.*, 2016; Lu *et al.*, 2016], whereas systems that use CNNs without attention [Antol *et al.*, 2015], the gradient for the desired class is backpropagated through the convolutional feature maps to obtain a visualization of the focus regions in an image.

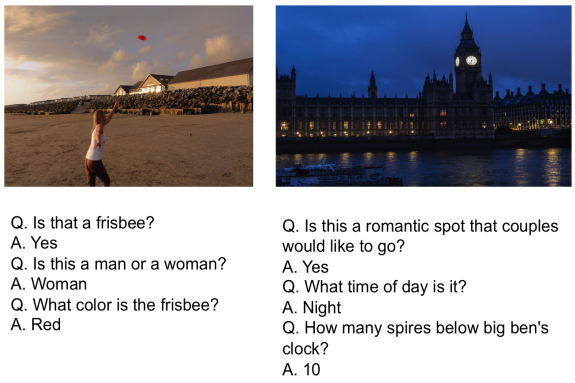

Figure 1: Random sample of images with related three questions and ground truth answers taken from the VQA dataset

Explanation can be helpful in not just understanding and interpreting systems' output but also leveraging systems for improving performance. For example, a system that generates an explanation that is not coherent with its output is not reliable. For VQA, explanation can be of two types – visual or textual. The regions in an image that a model attends to while generating an output can be considered a visual explanation. The words in the question that a model attends to can be considered a textual explanation. Most VQA systems use visual attention, however there are some that use both visual and textual attention while generating an output. The visual explanation is generally represented using heat-maps that use color intensities to highlight the regions in that image that a model attends to. In this paper, we use visual explanation for improving the accuracy of VQA systems.

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. Ensembling

systems intelligently is crucial to optimizing overall performance. In this paper, we use Stacking with Auxiliary Features (SWAF) [Rajani and Mooney, 2017] to more effectively combine diverse VQA models. The key idea is that we trust systems' agreement on an answer more if they also agree on its explanation. Traditional stacking [Wolpert, 1992] trains a supervised meta-classifier to appropriately combine multiple system outputs. SWAF further enables the stacker to exploit additional relevant knowledge of both the component systems and the problem by providing "auxiliary features" to the meta-classifier. Our key contribution is using visual explanations to create additional useful auxiliary features for SWAF applied to VQA. We demonstrate that ensembling three leading VQA systems using this approach outperforms a variety of baselines and ablations.

2 Background and Related Work

VQA is the task of answering a natural language question about the content of an image by returning an appropriate word or phrase. Figure 1 shows a sample of images and questions from the VQA 2016 challenge. The dataset consists of images taken from the MS COCO dataset [Lin *et al.*, 2014] and three questions per image obtained through Mechanical Turk. Table 1 gives some statistics on the dataset.

	Images	Questions
Training	82,783	248,349
Validation	40,504	121,512
Test	81,43	244,302

Table 1: VQA dataset size

In stacking, a meta-classifier is learned to combine the outputs of multiple underlying systems. The stacker learns a classification boundary based on the confidence scores provided by individual systems for each possible output. Stacking With Auxiliary Features (SWAF) provides the meta-classifier additional information, such as features of the current problem and provenance information for the output from individual systems. We use visual explanation that provides information about regions in an image that are crucial for generating the output. This allows SWAF to *learn* which systems are reliable based on what regions of the image they attend to, on which types of problems and when to trust agreements between specific systems. It has previously been applied effectively to information extraction and entity linking [Viswanathan *et al.*, 2015; Rajani and Mooney, 2016; 2017]. To the best of our knowledge, there has been no prior work on using explanation for improving ensembles. Figure 2 gives an overview of the SWAF approach.

We use SWAF to combine three diverse VQA systems such that the final ensemble performs better than any individual component model even on questions with low agreement. The three component models are trained on the VQA training set and the stacker is trained on the validation data.

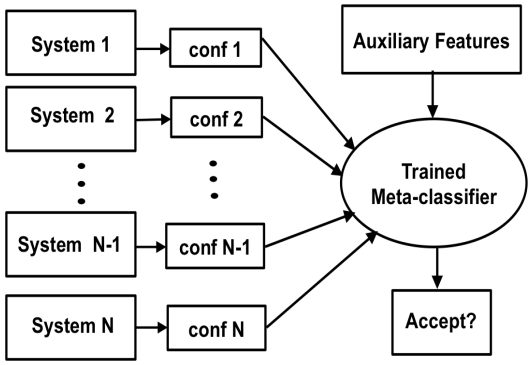

Figure 2: Ensemble Architecture using Stacking with Auxiliary Features. Given an input, the ensemble judges every possible question-answer pair produced by the component systems and determines the final output answer.

2.1 Long Short Term Memory (LSTM)

The LSTM model is one of the original baseline models used to establish a benchmark for the VQA dataset [Antol *et al.*, 2015]. It combines an LSTM [Hochreiter and Schmidhuber, 1997] for the question with a CNN for the image to generate an answer and uses one-hot encoding for the words in the question and the penultimate layer of the VGGNet [Simonyan and Zisserman, 2015] as image features fused together using element-wise multiplication. We note that this model does not have an explicit attention.

2.2 Multimodal Compact Bilinear pooling (MCB)

Traditionally, systems that combine vision and language vector representations use concatenation or element-wise product or sum. [Fukui *et al.*, 2016] argue that such methods are not as effective as an outer product of the visual and textual vectors. To overcome the challenge of high dimensionality due to the outer product, the authors propose using Multimodal Compact Bilinear pooling (MCB) to efficiently and expressively combine multimodal features. The MCB model extracts representations for the image using the 152-layer Residual Network [He *et al.*, 2015] and an LSTM embedding of the question. The two vectors are pooled using MCB and the answer is obtained by treating the problem as a multi-class classification problem with $3,000$ possible answers.

2.3 Hierarchical Question-Image Co-Attention (HieCoAtt)

The idea behind the HieCoAtt model is that in addition to using visual attention to focus on where to look, it is equally important to model what words to attend to in the question (question attention) [Lu *et al.*, 2016]. The model jointly reasons about the visual and language components using "co-attention." Question attention is modeled using a hierarchical architecture including word, phrase, and question levels. HieCoAtt uses two types of co-attention – parallel and sequential at all three levels of the question hierarchy.

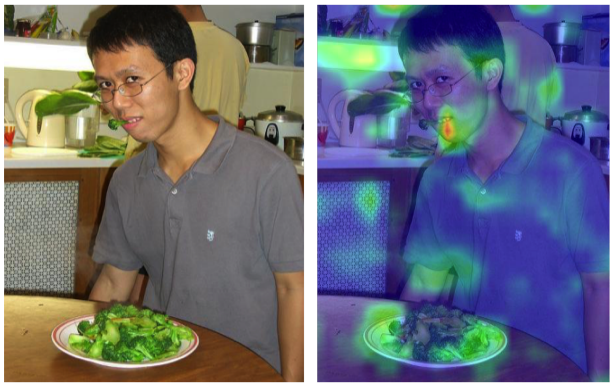

Figure 3: On the left is an image from the VQA dataset and on the right is the heat-map overlaid on the image for the question - 'What is the man eating?'

3 Auxiliary Features for SWAF

For stacking the VQA systems, we first form unique question-answer pairs across all outputs before passing them through the stacker. If a system generates a given I/O pair, then we use its probability estimate for that output, otherwise the confidence is considered zero. If a question-answer pair is classified as correct by the stacker, but there is also another answer that is classified correct for the same question, the pair with higher classifier confidence is chosen. For questions that did 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.

The confidence scores along with explanation as auxiliary features are used by the stacker, as shown in Figure 2, to classify each question-answer pair. The auxiliary features are the backbone of the SWAF approach, enabling the stacker to intelligently learn to rely on systems' outputs conditioned on the supporting evidence. Along with explanation, we also use three other types of features that enable the stacker to get more context while making a decision.

3.1 Explanation

Recently, there has been work on analyzing regions of the image that VQA models focus on while answering the question [Goyal *et al.*, 2016]. The authors concluded that deep learning models attend to relevant parts of the image while answering the question. The parts of images that the models focus on can be thought of as visual explanations for answering the question. We use these visual explanations to construct auxiliary features for SWAF.

The part of image to which the model attends can be visualized using a heat-map. Figure 3 shows an image and its heat-map for a given question. The idea is to trust the agreement between systems when they also agree on the heat-map explanation. The heat-map of a given system is compared to every other system's heat-map using using the rank correlation protocol described in [Das *et al.*, 2016]. This generates n choose 2 "explanation agreement" auxiliary features for SWAF. The idea behind using such features is that it enables the stacker to learn to rely on systems that "look" at the right region of the image when generating an answer.

We use the GradCAM algorithm [Goyal *et al.*, 2016] to generate explanatory heat-maps for each answer. Given an image and category, the image is forward propagated through the CNN part of the model. The gradients are set to zero for all categories except the one under consideration, which is set to 1. This signal is then backpropagated to the convolutional feature maps of interest and is combined to compute the heat-map.

3.2 Question and Answer Types

[Antol *et al.*, 2015] analyzed the VQA data and found that most questions fall into several types based on the first few words. For example, questions beginning with "What is...", "Is there...", "How many...", or "Does the...". Using the validation data, we discover such lexical patterns that define a set of question types. The questions were tokenized and a question type was formed by adding one token at a time, up to a maximum of 5, to the current substring. The question "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". The prefixes that contain at least 500 questions were then retained as types. We added a final type "other" for questions that do not fall into any of the predefined types, resulting in a total of 70 question types. A 70-bit vector is used to encode the question type as a set of auxiliary features.

The original analysis of VQA answers found that they are 38% "yes/no" and 12% numbers. There is clearly a pattern in the VQA answers as well and we use the questions to infer some of these patterns. We considered three answer types – "yes/no", "number" and "other". The answer-type auxiliary features are encoded using a one-hot vector. We classify all questions beginning with "Does","Is","Was","Are", and "Has" as "yes/no". Ones beginning with "How many", "What time", "What number" are assigned "number" type. These inferred answer types are not exhaustive but have good coverage.

3.3 Question Features

We also use a bag-of-words (BOW) representation of the question as auxiliary features. Words that occur at least five or more times in the validation set were included. The final sparse vector of dimension $3,391$ representing a question was normalized by the number of unique words in the question. [Goyal *et al.*, 2016] showed that attending to specific words in the question is important in VQA. Including a BOW in the auxiliary features equips the stacker to efficiently learn which words are important to classifying answers.

3.4 Image Features

Along with the aforementioned features, we also use "deep visual features" of the image as additional auxiliary features. Specifically, we use the $4,096$ features from VGGNet's $fc7$ layer [Simonyan and Zisserman, 2015]. Using such image features enables the stacker to learn to rely on systems that are good at identifying answers for particular types of images.

Method	All	Yes/No	Number	Other
Voting (MCB + HieCoAtt + LSTM)	60.31	80.22	34.92	48.83
iBOWIMG [Zhou *et al.*, 2015]	55.72	76.55	35.03	42.62
DPPNet [Noh *et al.*, 2016]	57.36	80.28	36.92	42.24
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 [Fukui *et al.*, 2016]	62.56	80.68	35.59	52.93
Stacking	62.59	81.79	34.58	51.72
+ Q/A types	62.73	82.09	35.47	52.10
+ Question Features	63.12	81.61	36.07	53.77
+ Image Features	65.44	82.08	38.08	57.15
+ Explanation*	**65.54**	**82.28**	**38.63**	**57.32**

Table 2: Accuracy results on the VQA open-ended *test-standard* set (except for the explanation features)

4 Experimental Results

We present experimental results on various baselines and ablations of the Stacking With Auxiliary Features (SWAF) approach. The VQA challenge splits the test set into *test-dev* and *test-standard*. Evaluation on either split requires submitting the output to the competition's online server.† However, there are less restrictions on the number of submissions that can be made to the *test-dev* compared to the *test-standard*. The *test-dev* set is a subset of the standard test set consisting of randomly selected $60,864$ questions.

We note that generating explanations is computationally expensive and we were only able to get results on the test-dev set with the explanation features. All the other results are reported on the entire test set. We use $L1$ regularized SVM classification for generic stacking and stacking with only question/answer types as auxiliary features. For the question, image, and explanation features, we found that a neural network with two hidden layers works best. The first layer is fully connected and the second has approximately half the number of neurons as the first hidden layer. We used Keras with Tensorflow back-end [Chollet, 2015] for implementing the network.

We compare our approach to a voting baseline which maximizes precision by only accepting an answer to be correct if all the component systems predicted the exact same answer for a given question. For questions that do not have a consensus, the answer that has maximum agreement is taken with ties broken in favor of systems with higher confidence. We also compare against two other state-of-the-art VQA systems not used in our ensemble: iBOWIMG [Zhou *et al.*, 2015] and DPPNet [Noh *et al.*, 2016]. iBOWIMG uses softmax over the bag-of-words representation of the question concatenated with GoogleNet [Szegedy *et al.*, 2015] image features and gives comparable performance to models using deep or recurrent neural networks. VGG has lower error rate compared to GoogleNet for CNNs and is thus our choice for image features [Johnson, 2016]. DPPNet uses a CNN with a dynamic parameter layer whose weights are determined adaptively based on questions using a gated recurrent unit (GRU).

The VQA server along with reporting accuracies on the full question set, also reports a break-up of accuracy based on three answer categories. Table 2 shows the full set and category-wise accuracies. Although the results using explanation are on the test-dev subset of the test set and not directly comparable, they do show a small improvement in accuracy. The number of explanation features is small compared to all the other feature types. So to avoid over-fitting to the other features, we also plan on trying neural architectures in which the different feature sets are fused at later layers in the network.

5 Conclusion and Future Work

This paper has proposed and evaluated the novel idea of using explanations to improve ensembling of multiple systems. It has demonstrated how visual explanations for visual question answering (represented as heat-maps) can be used to aid stacking with auxiliary features. This approach effectively utilizes information on the degree to which systems agree on the *explanation* of their answers. We also described three other types of auxiliary features obtained from VQA problems and showed that the combination of all of these auxiliary features, including explanation, gives the best results.

We believe that integrating explanation with ensembling has a two-fold advantage. First, as discussed in this paper, explanations can be used to improve the accuracy of an ensemble. Second, explanations from the component systems can be used to build an explanation for the overall ensemble. That is, by combining multiple component explanations, SWAF could also produce more comprehensible results. Therefore, in the future, we would like to focus on explaining the results of an ensemble. Another issue we plan to explore is using textual explanations for VQA. We believe that the words in the question to which a system attends can also be used to improve ensembling. Finally, we hope to apply our approach to additional problems beyond VQA.

Acknowledgement

This research was supported by the DARPA DEFT program under AFRL grant FA8750-13-2-0026.

* Result obtained on test-dev set

† `http://www.visualqa.org/challenge.html`

References

[Agrawal *et al.*, 2016] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, 2016.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *Proceedings of the Conference on Natural language learning (NAACL2016)*, pages 1545–1554, 2016.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[Chen *et al.*, 2015] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. ABC-CNN: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.

[Chollet, 2015] Franois Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[Das *et al.*, 2016] Abhishek Das, Harsh Agrawal, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? *arXiv preprint arXiv:1606.03556*, 2016.

[Fukui *et al.*, 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, 2016.

[Goyal *et al.*, 2016] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards Transparent AI Systems: Interpreting Visual Question Answering Models. In *International Conference on Machine Learning (ICML) Workshop on Visualization for Deep Learning, 2016*, 2016.

[He *et al.*, 2015] K. He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[Hendricks *et al.*, 2016] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating Visual Explanations. *arXiv preprint arXiv:1603.08507*, 2016.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Johnson, 2016] Justin Johnson. cnn-benchmarks. `https://github.com/jcjohnson/cnn-benchmarks`, 2016.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[Lipton, 2016] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[Noh *et al.*, 2016] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 30–38, 2016.

[Rajani and Mooney, 2016] Nazneen Fatema Rajani and Raymond J. Mooney. Combining Supervised and Unsupervised Ensembles for Knowledge Base Population. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, 2016.

[Rajani and Mooney, 2017] Nazneen Fatema Rajani and Raymond J. Mooney. Stacking With Auxiliary Features. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI2017)*, Melbourne, Australia, August 2017.

[Selvaraju *et al.*, 2016] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2016.

[Simonyan and Zisserman, 2015] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. In *Proceedings of ICLR*, 2015.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[Viswanathan *et al.*, 2015] V. Viswanathan, N. Rajani, Y. Bentor, and R. Mooney. Stacked ensembles of information extractors for knowledge-base population. In *Proceedings of ACL 2015*, Beijing, China, 2015.

[Wolpert, 1992] D. Wolpert. Stacked generalization. *Neural Networks*, 5, 1992.

[Xu and Saenko, 2016] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[Zhou *et al.*, 2015] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.