Ensembling Visual Explanations 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

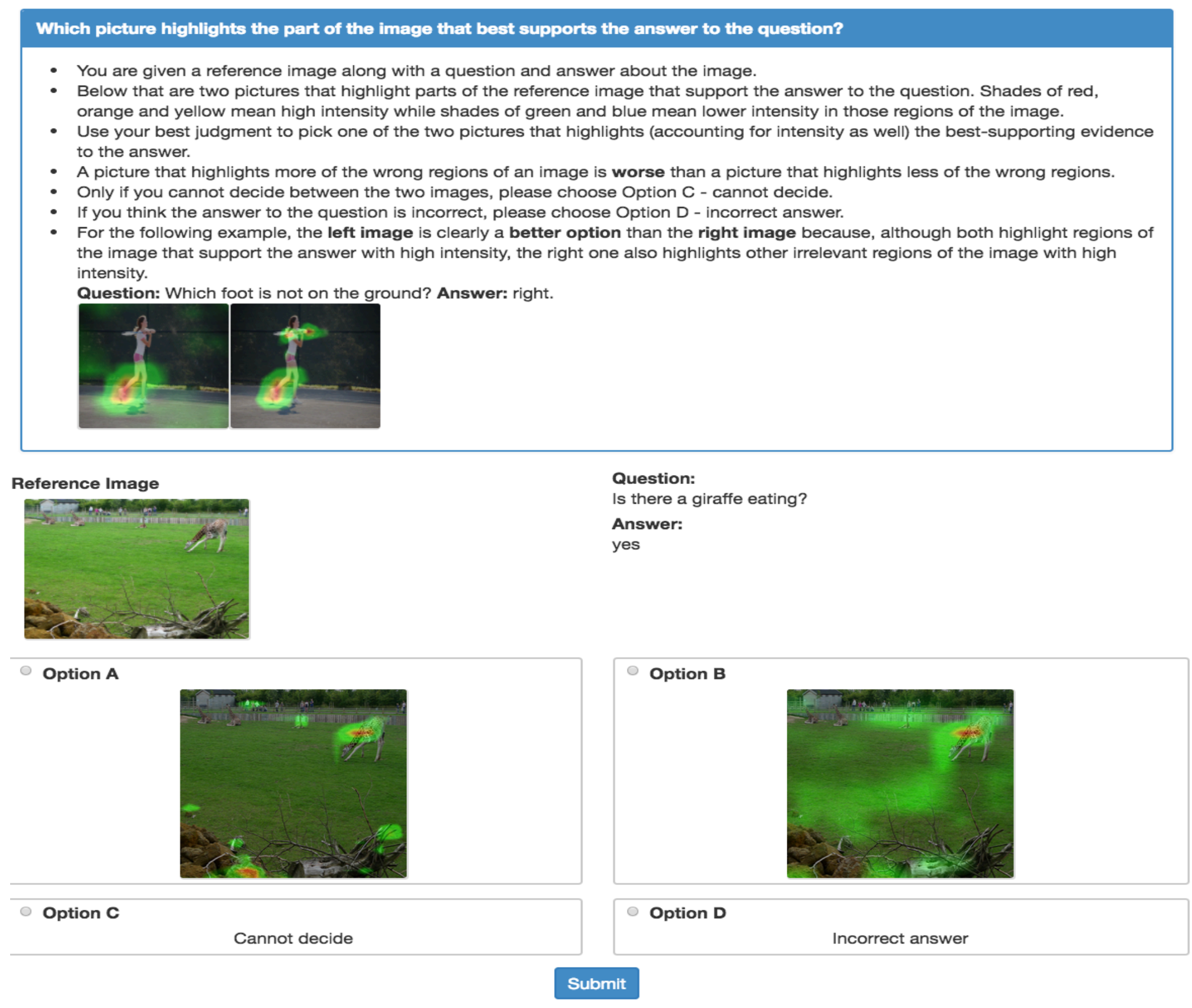
- Visual Question Answering (VQA) (Antol et al., 2015) requires both language and image understanding, language grounding capabilities, as well as common-sense knowledge.
- The top performing VQA systems are **ensembles** of neural networks that perform substantially better than any of the underlying individual models.

Explainable AI (XAI)

- Explanations make AI systems more **transparent** and also justify their predictions.
- Prior work focuses on generating explanations for individual models.
- We explore the problem of generating **visual** explanations for an **ensemble** using explanations from underlying individual models as input.
- 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 thought of as **visual explanations**
- **GradCAM** (Selvaraju *et al.*, 2017) is used to generate visualization maps.
- The gradients are set to zero for all categories except the one under consideration.
- The signal is backpropagated to the convolutional feature maps of interest to compute the heat-map

Ensembling Visual Explanation

- All VQA models had some degree of **noise** with **high variance** in their GradCAM visualization depending on the Image-Question (IQ) pair under consideration.
- There was high variance across visualizations for different models even when they **agreed** on the answer for a given IQ pair.
- Our visual explanation ensemble:
  - ✓ Aggregates visualizations from appropriate regions of the image,
  - ✓ Discounts visualizations from regions that are not relevant,
  - ✓ Reduces noise and,
  - ✓ Is superior to any individual system's visual explanation on a manual evaluation
- We demonstrate our algorithm by ensembling 3 VQA models:
  - LSTM with CNN (Antol et al., 2015)
  - Hierarchical Co-Attention (HieCoAtt) (Lu et al., 2016)
  - Multimodal Compact Bi-linear pooling (MCB) (Fukui et al., 2016)
- Results from a crowd-sourced human evaluation indicate that, on an average, our visual explanation ensemble is superior to each of the individual system's visual explanation **63%** of the time.
- We evaluate using Amazon Mechanical Turk (AMT) by comparing the ensemble explanation to the explanation generated by component VQA models.



Which picture highlights the part of the image that best supports the answer to the question?

- You are given a reference image along with a question and answer about the image.
- Below that are two pictures that highlight parts of the reference image that support the answer to the question. Shades of red, orange and yellow mean high intensity while shades of green and blue mean lower intensity in those regions of the image.
- Use your best judgment to pick one of the two pictures that highlights (accounting for intensity as well) the best-supporting evidence to the answer.
- A picture that highlights more of the wrong regions of an image is **worse** than a picture that highlights less of the wrong regions.
- Only if you cannot decide between the two images, please choose Option C - cannot decide.
- If you think the answer to the question is incorrect, please choose Option D - incorrect answer.
- For the following example, the **left image** is clearly a **better option** than the **right image** because, although both highlight regions of the image that support the answer with high intensity, the right one also highlights other irrelevant regions of the image with high intensity.
  Question: Which foot is not on the ground? **Answer:** right.

Reference Image

Question:
Is there a giraffe eating?
Answer:
yes

○ **Option A**

○ **Option B**

○ **Option C**
Cannot decide

○ **Option D**
Incorrect answer

Figure 1: AMT interface for evaluating visual explanation.

Algorithm

- We first build an ensemble model that uses **Stacking With Auxiliary Features** (SWAF) (Rajani and Mooney, 2017) to combine outputs of the component VQA models.
- Then, we generate visual explanations ensemble by **combining** explanation feature-maps of models that **agree** with the ensemble answer on an Image-Question (IQ) pair.
- We propose **two** heuristics for combining the individual feature maps:
  1. Weighted Average and (WA)
  2. Penalized Weighted Average (PWA)

Weighted Average Ensemble Explanation

$$E_{i,j} = \begin{cases} \frac{1}{K}\sum_{k\in K} w_k A^k_{i,j}, & \text{if } A^k_{i,j} \geq t \\ 0, & \text{otherwise} \end{cases} \quad \text{subject to} \sum_{k\in K} w_k = 1$$

Penalized Weighted Average Ensemble Explanation

$$E_{i,j} = \begin{cases} \frac{1}{K}\sum_{k\in K} w_k \left(A^k_{i,j} - I_{i,j}\right), & \text{if } \left(A^k_{i,j} - I_{i,j}\right) \geq t \\ 0, & \text{otherwise} \end{cases} \quad \text{subject to} \sum_{k\in K} w_k = 1$$

E : ensemble feature map
K: total number of component models
A^k: explanation feature-map of component model k
w^k: weights, proportional to the component model k's performance
t: threshold (= 0.25)
I: explanation feature-map of component model that does not agree with the ensemble's answer for an IQ pair
i,j: indices into the feature-map entries

Figure 2 shows an example that uses PWA to obtain the ensemble explanation map

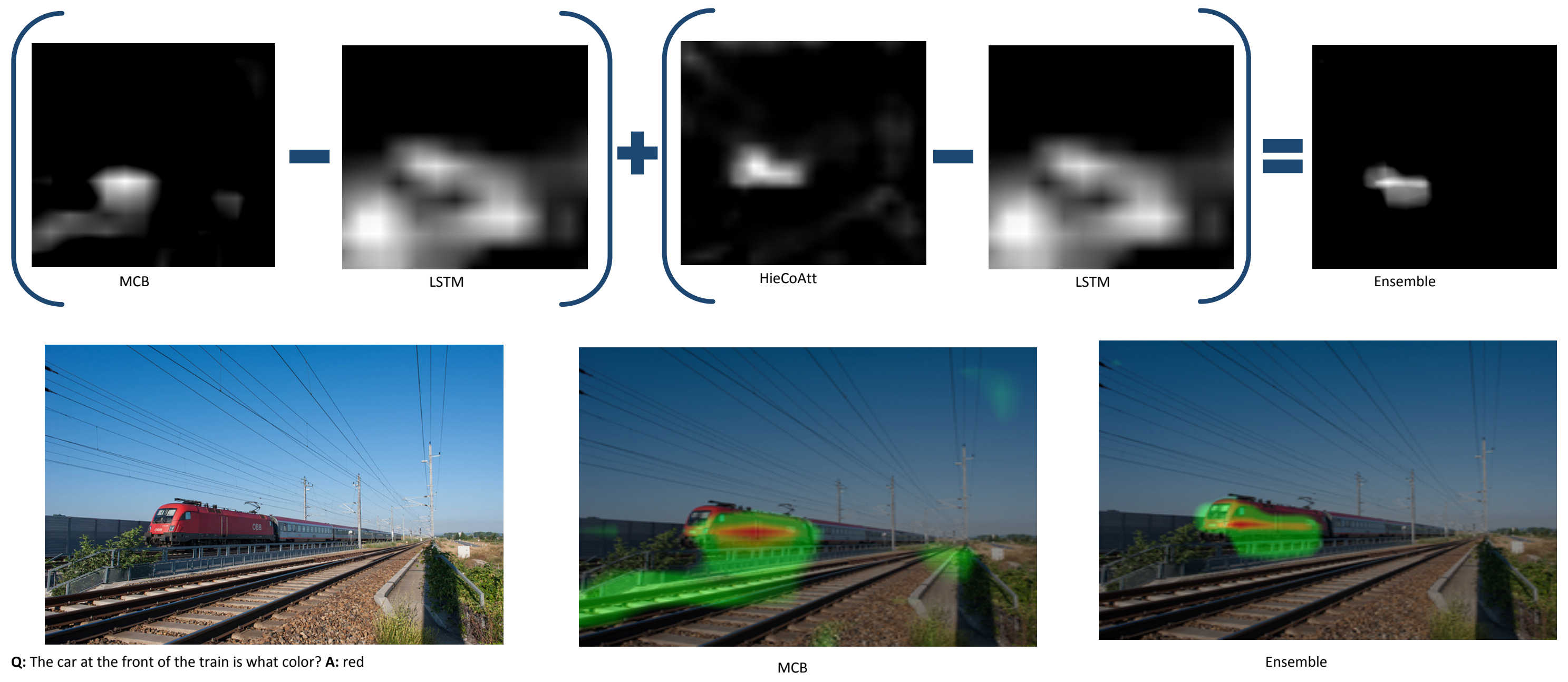


Figure 2: The top row shows the process of ensembling visual explanation for an IQ pair when the ensemble model agrees with the MCB and HieCoAtt models (ans: "red") and disagrees with the LSTM model (ans: "white"). The bottom row shows the reference IQ pair and the MCB vs ensemble visual explanation. The feature map is normalized to obtain the final ensemble visualization.

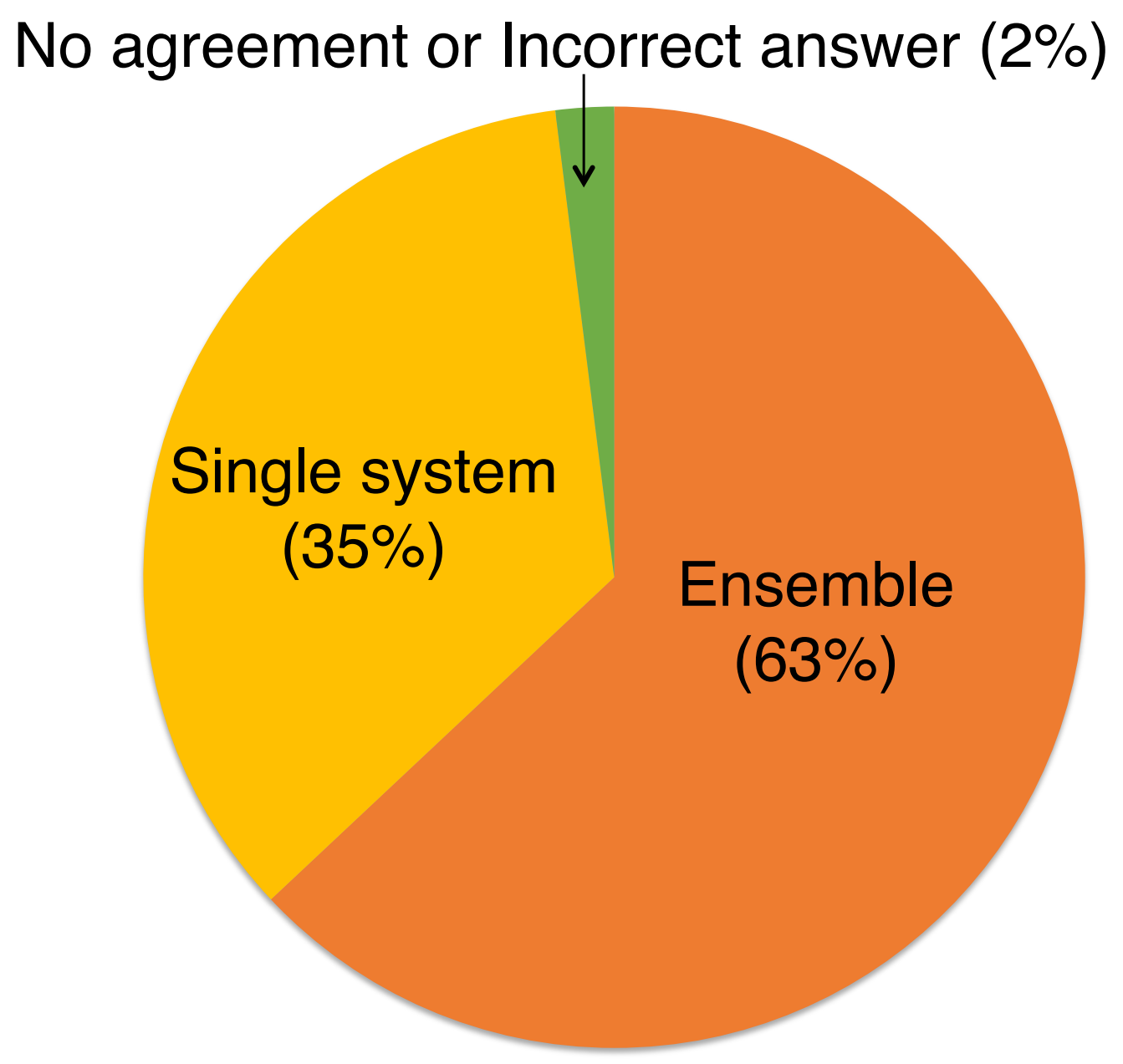
Evaluation

- As shown in Figure 1, we showed two visualization side-by-side to Turkers on AMT, one of them is an ensemble explanation and the other is generated by one of the component VQA model that agrees with the ensemble on the answer of the given IQ pair.
- We also provide detailed instructions along with an example to show what a good visualization looks like.
- Apart from the two images as options, we also give two more options – "cannot decide" and "incorrect answer".
- The proposed heuristics are evaluated under three different scenarios: all three systems agree, two systems agree and only one system agrees with the ensemble's answer.
- Each of the scenarios are evaluated by 3 different Turkers on 100 IQ pairs using each of the proposed heuristics.

Results

Overall

- We **aggregate** the AMT evaluation using voting and when there is no agreement among Turkers, we classify those instances under the "no agreement" category.
- The figure alongside shows the **overall average performance** obtained in terms of percentage of time each class was selected by the Turkers.
- The results indicate that our visual explanation ensemble is superior to any single system's visual explanation **63%** of the time.
- We also analyzed the ablation performance by varying the number of systems that agreed with the ensemble's answer for an IQ pair.



Ablation

- **Only MCB** agrees: we found that subtracting the thresholded localization maps of the LSTM and HieCoAtt systems worked slightly better than averaging the thresholded localization maps obtained by forcing the LSTM and HieCoAtt systems to produce maps for answers produced by the MCB model.
- **MCB** and **HieCoAtt** agree: we show results for three different scenarios for generating the ensemble localization maps – just using MCB and HieCoAtt localization maps, using the LSTM localization map, and finally using the localization map produced by LSTM for the output that agrees with the other two systems.
- **MCB**, **HieCoAtt** and **LSTM** agree: we found that using the weighted average worked better than using equal weights when compared to the HieCoAtt and LSTM localization maps but performed slightly worse when compared to the MCB maps.

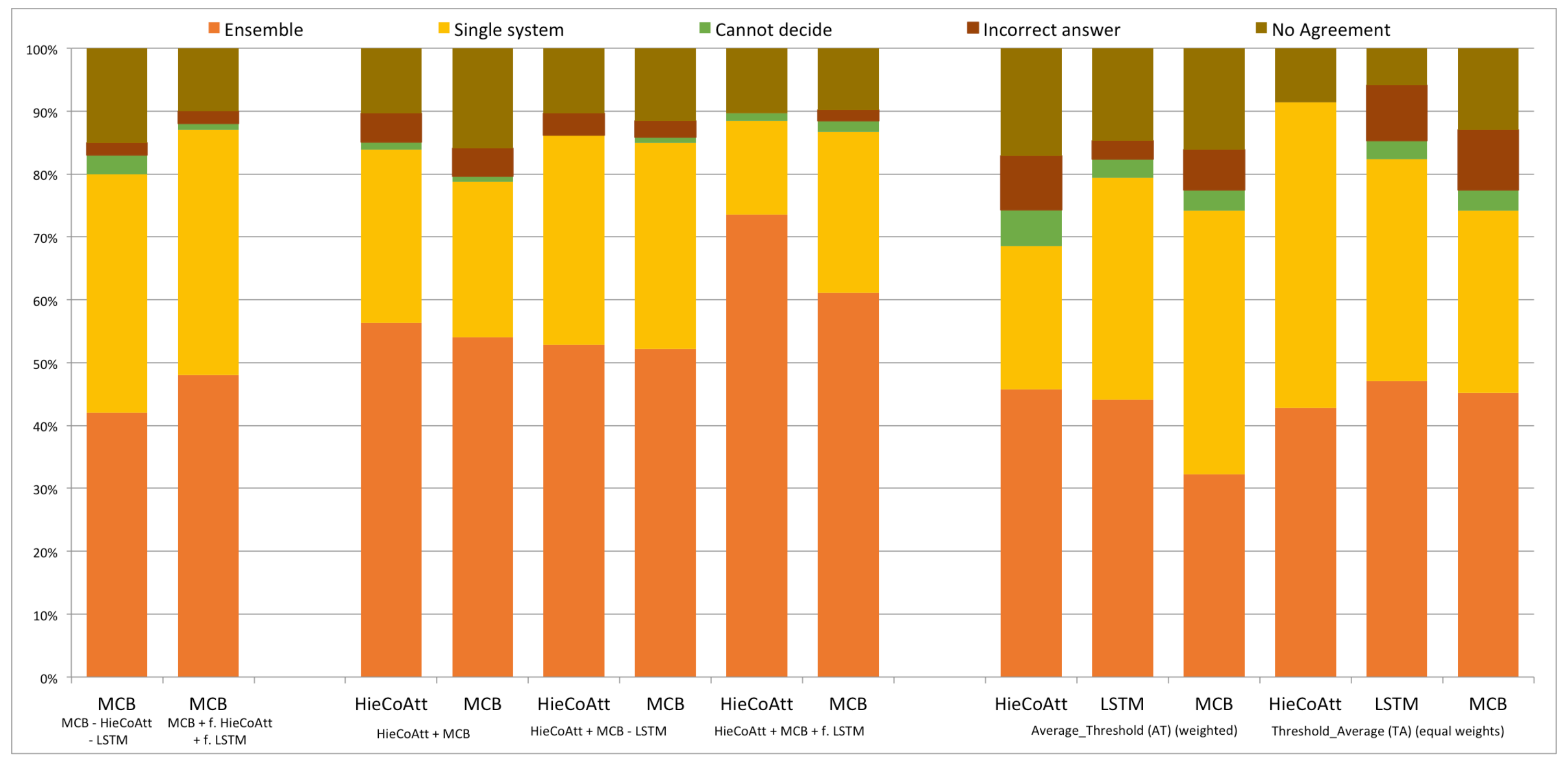


Figure 3: The figure shows results obtained when one system (MCB) agrees, two systems (MCB and HieCoAtt) agree and when all three systems (MCB, HieCoAtt and LSTM) agree with the ensemble's output respectively from left to right. The y - axis indicates the percentage of instances for each of the agreed upon options chosen by Turkers as well as when there was no agreement among the Turkers. The x -axis shows the individual system that the ensemble was compared to. Below that is the label indicating how the ensemble localization map was calculated; f. stands for forced output that agrees with the ensemble, AT stands for averaging the localization maps followed by thresholding and TA does the reverse, first thresholds the maps and then averages them.

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.
- 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. 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 . 2017.

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