Objectives

- Devise a general method to extract representative rules describing and prediction errors of deep learning models
- Improve models based on the discovered insights

Step 1: Meta-feature extraction
- Question & Answer
  - Question: What is the man to the right doing?
  - Answer: Playing tennis.

Steps 2 & 3: Rule Extraction and Analyze the Rules
- Iterating Learning (t=1)
- Rule Learning
- Rule Evaluation & Filtering
- Reduced Dataset
- Iterating Learning (t=T)

Rule List
- What is the name > 0.5 & image_caption_entry > 0.5 => Wrong
- What does the > 0.5 & question_read > 0.5 => Wrong
- Why > 0.5 & image_caption_entry > 0.5 & image_object > 1.5 => Wrong

Step 3: Analyze the Rules
- VQA – OCR
- CSQA

Examples found
- Question: What letters are on the umbrella?
  - Expected: wsh
  - Prediction: wsh
- Question: What is the name of the street?
  - Expected: Faming lips alley
  - Prediction: main street

Performance comparison on VQA v2.0
- Model | Acc
- VILBERT | 68.75 | 69.47 | 69.64
- Ours | 69.44 | 69.64 | 69.82

Conclusion
We presented a novel pipeline that helps automate the error analysis process by learning interpretable rules that characterize the type of mistakes that a system makes. We demonstrated the ability to "close the loop" and use the insight gained from some of the induced rules to make modest improvements to ViLBERT and RoBERTa. These simple but effective approaches have the potential to be applied in production environment shortening the iterative update cycle of models.

Acknowledgments
This research was supported by the DARPA XAI program under a grant from AFRL. We would like to thank Bill Ferguson and his colleagues at Raytheon for inspiring our work.