Towards Automated Error Analysis: Learning to Characterize Errors

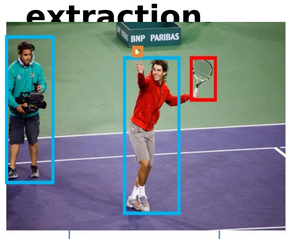
Tong Gao, Shivang Singh, Raymond J. Mooney

Department of Computer Science, University of Texas at Austin

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.

Image Captioning → A man holding a tennis ball on a tennis court.

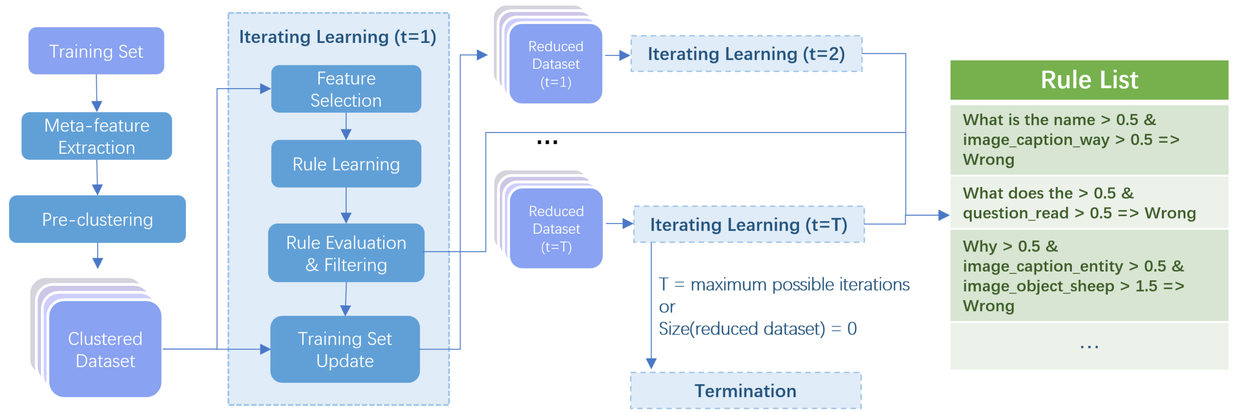
Object Detection → people = 2, tennis racket = 1, sports ball = 1

Lemmatization, POS Tagging, WordNet, Stopword Removal

Meta-features

Image caption: man, adult, person, causal agent, hold, keep, tennis, court game, athletic game, game…
Image objects: people, people, tennis racket, sports ball
Question: what is the man, right, do
Answer: play, compete, tennis, court game, athletic game, game

Step 2: Rule Extraction



Step 3: Analyze the Rules

VQA – OCR

Rules	**Question type**: what **Question contains** letter **Error rate**: 82.7% **Coverage:** 0.404%	**Question type**: what **Question contains** website **Error rate**: 92.1% **Coverage:** 0.0658%
Examples found	 **Question:** What letters are on the umbrellas? **Image caption**: A group of people holding umbrellas **Expected:** wwf **Prediction:** white	 **Question**: What is the name of the street? **Image caption**: A green street sign on the side of a street **Expected**: flaming lips alley **Prediction**: main street

CSQA

Rule	**Answer contains:** American State **Error rate**: 70% **Coverage:** 3.85%
Examples	Where are the most famous BBQ steakhouses in America? **Gold:** Texas **Prediction:** Kansas city **Negative Answers:** building, Kansas city, Maine, falling down The tourist entered Mammoth cave, what state were they in? **Gold:** Kentucky **Prediction:** West Virginia **Negative Answers:** West Virginia, Rocky Hills, Scotland, Canyon
Rule	**Answer contains**: kitchen **Error rate**: 100% **Coverage:**
Example	From where would you normally take a cup when you're about to get a drink? **Gold**: kitchen cabinet **Prediction**: water fountain **Negative Answers:** dishwasher, water fountain, sand box, toilet
Rule	**Question contains**: chordate, where **Error rate**: 100%
Examples	Where does a wild bird usually live? **Gold**: countryside **Prediction**: sky **Negative Answers:** cage, sky, desert, windowsill Where is a bird likely to make its home? **Gold**: forest **Prediction**: nest **Negative Answers:** nest, roof, leaves, sky

4. Improve the Models

	dev	test-dev	test-std
ViLBERT	68.75	69.47	69.64
Ours	**69.44**	**69.64**	**69.82**

Performance comparison on VQA v2.0

Model	Acc
RoBERTa	72.1
Ours	**72.45**

Performance comparison on Commonsense QA

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.