Self-Critical Reasoning for Robust Visual Question Answering
Jialin Wu, and Raymond J. Mooney The University of Texas at Austin

Introduction
- Common VQA systems tend to only capture superficial statistical correlations between QA pairs, especially when training and test set are under different distribution.
- VQA systems should focus on the objects that human would focus.
- We propose two constraints for VQA systems that help the right objects contribute more to the right answers than to the wrong answers.

Model Overview
- Recognizing and Strengthening Influential Objects
  - Extracting a set of influential objects ($\mathcal{I}$) that humans would focus on.
  - Enforcing the gradients ($\Delta(a, q)$) from the correct answer to have the biggest value in at least one of the extracted objects.
  - Comparing to Up-Down VQA system.

- Criticizing Incorrect Dominant Answers
  - Although VQA systems focus on the right object for the right answer, but the object could contribute more to the wrong answers.
  - Finding the most influential object ($\mathbf{v}^*$) using gradient-based method.
  - Enforcing the object to contribute more to the correct answer.

Conclusion
- VQA systems should be able to focus on the right set of objects as human do to predict the right answer.
- It is also necessary to prevent the systems from over sensitive to the most common answers.