Question 5: Paper Review

The authors of the paper present a new model (PJ-X) for the VQA task, which incorporates 2 modes of reasoning, visual and textual. This is unlike previous works in that it uses multimodal explanations for VQA tasks. This is advantageous to unimodal methods, which can suffer when the VQA example is better explained with another model. For example, a visual model trying to explain a VQA example better explained with textual evidence.

Two new annotated datasets, VQA-X and ACT-X, are also collected. These datasets are well designed for the VQA task because of its high density of complementary VQA data pairs. The downside was that they were not able to compare their new datasets with preexisting ones, such as VQA-HAT.

The experimental results were promising; the authors were able to show in their diagnostic explanations experiment, that bimodal explanations helped humans guess the correctness of machine output the most. Comparisons with naive methods and HieCOAtt-Q model on the VQA-X, ACT-X, and VQA-HAT datasets using two metrics(Earth Mover's and Rank Correlation) elucidated the dominance of the PJ-X model for this task.

In addition to showing favorable performance in the experiments, the PJ-X model also runs faster than the MCB model and does not require domain knowledge.