Key contributions

Proposes a novel model that provides joint textual rationale generation and attention visualization. Collected two new datasets, ACT-X and VQA-X, of human annotated multimodal explanations (visual + textual) for activity recognition and visual question answering. The results of their model outperform strong baselines and is backed by quantitative results.

Strengths

Presents a strong argument that multimodal explanation models offer significant benefits over unimodal approaches. Since generating reasonable explanations for correct answers is important, then it is valuable to see what the system generates for incorrect answers. The proposed model's explanations are consistent with its answers even if wrong.

Weaknesses

Even though the model performance is better than baselines, the absolute value of the performance (%) for some evaluations are below 50%. The paper definitely makes improvement and headway into the field but the results for some evaluations are not very high.

My Takeaways

I wonder if it would be feasible to expand explanations by pointing at an image to more media types such as video or sound waves (non human speech, classifying animals sounds maybe). For video it seems possible but the overhead of processing each frame might be computationally too expensive for the results. All in all, I think the results of the paper were very interesting in learning what is happening in deep networks.