Key contributions

1. Propose a new model(PJ-X) that can: 1. Generate visual and textual explanations 2. Localize salient regions while generating textual rationales.
2. Quantitively show that training with the textual explanation not only yields better textual justification models, but also better localizes the evidence that supports the decision.
3. Quantitively show cases where visual explanation is more insightful than textual explanation and vice versa. And multimodal explanation models offer significant benefits over unimodal approaches.
4. Present two novel datasets of human annotated multimodal explanations for activity recognition and visual question answering.

Strengths

1. Well organized and well written.
2. All figures and tables in the essay helps me better understand the paper. Also experiments results are clear and persuasive.

Weaknesses

1. The author writes: sometimes the concept is easily conveyed when looking at the visual pointing result, but sometimes textual justification captures the rationale. While the author doesn't point out or guessed the factor for the deviation of visual or textual explanation.
2. The experiment result is predictable. More information can get better performance. Training with the textual explanation can yield better result.

Personal takeaway

1. In model training part of this essay, the author uses small dataset to do pretraining on answering model, then freeze the answering model and train explanation model on a smaller dataset. These techniques help to cut down the computational expense.
2. Figure pairs inserted in this essay are often similar in some way, while the textual and visual explanations are different. This makes readers believe the work is persuasive.