# Comparison between Human Attention and Linguistic Justifications on Images

**Qiucheng Wu**[1]

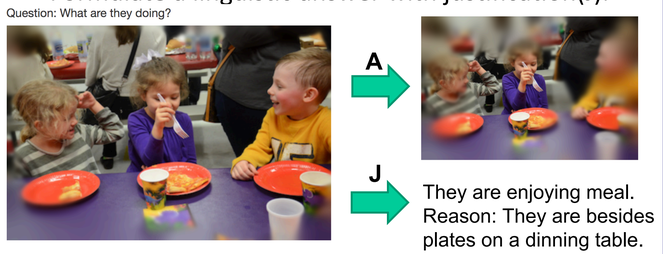
[1] *Computer Science and Engineering*

EECS 595, F19
**Natural Language Processing**

## Motivation & Problem Definition

When asked questions related to vision, we would:

- Find evidence from images(human attention, A);
- Formulate a linguistic answer with justification(J).




What are their relationships? Are they really 1-to-1?

- **Problem Definition**:

(a) Does attention *correspond/overlap* with justification?

$A \subset J$ or, if not more complicate, $J \subset A$?

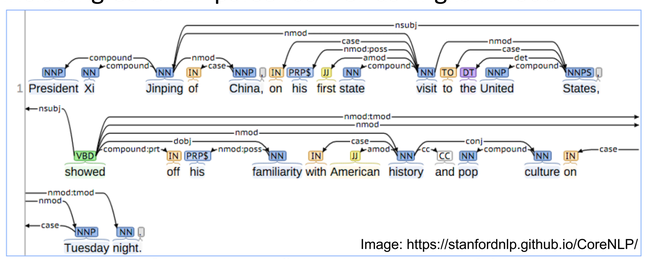
(b) When we pay more attention, how the attention features *grow* to formulate justification?

$A \rightarrow A_t$ with respect to $J$

## Tools

- **Stanford Dependency Parser:** parsing linguistic reasons.

Including CoreNLP parser as well as English model.




- **Georgia Tech Q-A Interface:** preprocess(blur) images;

Generate heatmap for user attention.

## References

[1] Stanford Dependency Parser, https://stanfordnlp.github.io/CoreNLP/
[2] Georgia Tech Q-A Interface
[3] Yang Shaohua, Gao Qiaozi, Sadiya Sari, and Chai Joyce. Commonsense Justification for Action Explanation 2018. Proceedings of the 2018 Empirical Methods in Natural Language Processing
[4] Abhishek Das, Harsh Agrawal, Lawrence C. Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? 2016. arXiv:1606.03556
The poster design is from EECS 576, Advanced Data Mining.

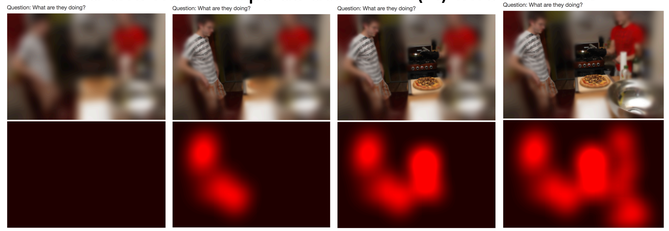
## Use Dependency Parsing to analyze *oral justification* and *attention* on images

## Approach

### Step 1: Generate heat-map for human attention (A)

We intimate how people pay attention to the images by

- Provide blur images at first;
- Let volunteer brush images for more info;
- Record the heat-maps as attention(A) information.




### Step 2: Record oral justification (J)

We then obtain oral justification for questions by

- Record text reasons;
- Parse text reasons to get key nouns of reasons;
- Label nouns on images to get regions for justification (J).

Two people are looking at a pizza. …

People
Pizza
Oven
Glove

### Step 3: Calculate criteria

Two criteria:

**c1. How much *J* comes from *A*?**

$$c1 = \frac{\sum_i (J_i \cap A_t)}{\sum_i J_i}$$


+ … +


c1: sum for each justification and take ratio

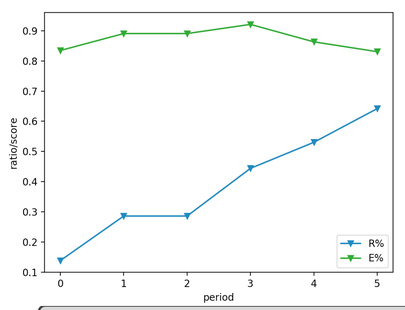
**c2. How much *A* Contributes to *J*?**

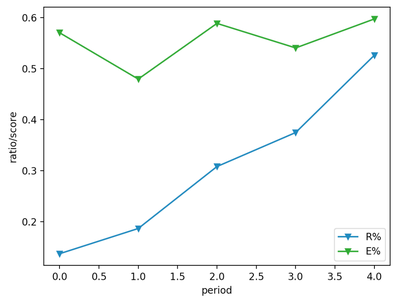
$$c2 = \frac{(\cup J_i) \cap A_t}{A_t}$$

c2: union for each justification and take ratio

## Experimental Results

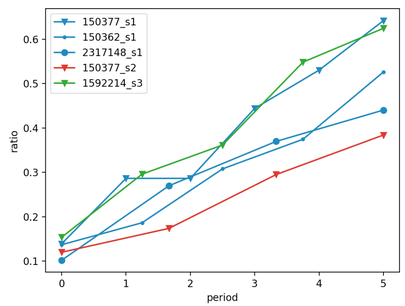
### c1 and c2 from the same image

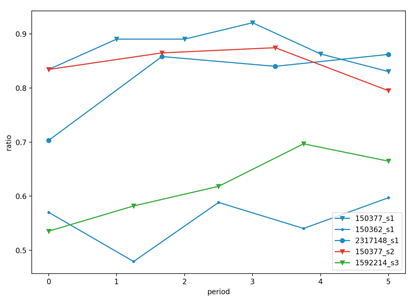



**Observation**:
Compared c1 with c2, most of the attention contributes to the justification(high c1), while there are much region of justification left unexplained(low c2).

### c1 and c2 across different images




**Observation**:
For c1, as expected, with volunteers de-blur more and more regions, more regions of oral justification can be explained by the users.
For c2, the ratio of attention roughly first increases and decreases, suggesting we may want to use some unrelated information to strengthen our belief.

## Conclusions & Next steps

- Relation between attention and oral justification is non-trivial. Most part of attention contributes to justification, while large part of justification leaves unexplained.
- One possible reason for unexplained justification is we only need part of information(attention) to realize an object exist. For example, we confirm a table with only a small part of it.
- Limitation: lacking subjects and data. Use pipeline for more test.