Final Report: Relations between human attention and linguistic justifications on images

Qiucheng Wu
University of Michigan
wuqiuche@umich.edu

Abstract

When asked a question related to images, we answer the questions by paying attention to corresponding evidence in the images, and we will also formulate a linguistic justification to explain reasons to others. This work researches the relations between human attentions and corresponding linguistic justifications on images. Specifically, this work tries to solve two problems:
(a)Do human attentions and linguistic justifications correspond/overlap with each other?
(b)How the attention features grow temporally when drawing a conclusion of questions?
In this work, we first collect the regions that human focuses on in different images, and we also obtain the text justifications that human use to defend their answers. We then parse the text justifications, labeling key objects mentioned in the justifications. Finally, we compare two regions to observe the relations between them. We notice that most of the human attention contributes to justifications, while some of the justifications left unexplained. We also notice subjects tend to use more than necessary attention to confirm their answers. We discuss these observations, give some possible explanations and purpose some modifications as future work.

1 Introduction and Problem Statement

Various questions exist everywhere in the daily life, and we usually try to find solid evidence to cope with them. It seems natural to directly reach a conclusion: we first find evidence, and we use the evidence to reach the conclusion. However, natural language and psychology studies are still trying to figure out their relations, and recent studies are trying to discover non-trivial relations between them, for example, whether formulating evidence is actually post-hoc.

Figure 1: Human attentions on images to answer question

Based on the motivations above, this work proposes a comparison model to cope with two problems. First, given an image and a relative question string, what are the relations between our attention and our justifications? To be specific, we probably find key elements in the images to support our answers, as shown in Figure 1. Would these key elements, or attentions, correspond with the linguistic justifications in a 1-1 or more complicated ways? Second, if they are corresponding with each other, when do humans capture enough evidence and draw a conclusion?

To explore these two questions, we first define the **attentions** and **justifications** mentioned in this work. Given a question and an image, the attentions

refer to the regions on the images that human focuses on. These regions usually have important entities or relations that are necessary to approach the questions. Meanwhile, the justifications refer to the linguistic reasons human provided to defend their answers. These reasons should contain key objects that are useful for human to reach a conclusion.

In summary, the problem can be formulated below: (a)Given an image I with relative questions, what are the relations between attention features A and justification J? (b)Given a series of attention features A_t with timestamps, how would the attention features grow and at which time it terminates?

2 Related Work

Our work continues on recent work regarding attention-based visual question answering(VQA). Attention-based VQA is a state-of-art topic in Natural Language Processing with a combination of computer vision, and one of the directions is to study on the common ground between human and artificial intelligent model. The following works address this problem.

In *Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?* (Das et al., 2016), the authors study on the attention differences between human and attention-based AI, and the result turns out they are focusing on different regions. Continue from their results, this work may provide a way to explain differences between human and AI addressed in their work. Meanwhile, they share their VQA interface source code with us. With the Amazon Mechanical Turk(AMT) used in their work, we can do a large-scale test and obtaining new data.

In *Commonsense Justification for Action Explanation* (Shaohua et al., 2018), the authors purpose a model managing to build the commonsense justifications between human and AI models in attention-based questions. If we are able to understand the hidden relations between human justifications and human attentions on VQA questions in this work, we may continue to increase the accuracy of their models.

Meanwhile, this work use Stanford Dependency Parser (Stanford University, 2019) to parse the human text justifications.

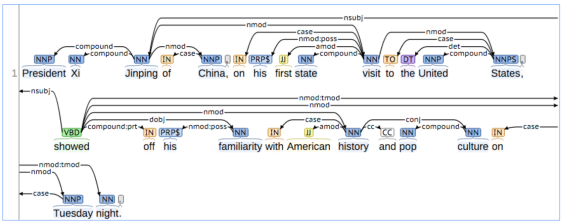

Figure 2: Stanford Dependency Parser. This image is from https://stanfordnlp.github.io/CoreNLP/.

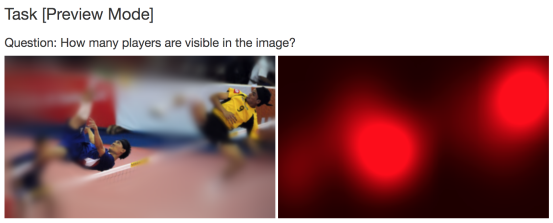

Figure 3: The interface to obtain human attentions.

3 Approach

Our model tries to compare attention versus justifications to reach a conclusion on these two problems. When our volunteers, or subjects, are asked specific questions on an image, we collect their attentions on images, answers on questions and corresponding justifications. Then, we put their attentions and justifications on images, and purpose formulas to calculate overlap ratios between them. Based on the ratio results, we can study on detailed relations between them. Therefore, the approach part is divided into three sections: (a) Attention collection; (b) Justification collection; (c) Proposed criteria calculation.

3.1 Attention collection

To generate the attention features A, this work will use an interface (Das et al., 2016) from Georgia Tech to collect the elements our subjects are paying attention to.

The whole process intimates how people pay attention to the images. To begin with, giving a question with a blurred image, the volunteers, or subjects, will try to deblur the images to collect important objects or relations in the images, At the same time, we will record the regions that the subjects are trying to deblur as a heat-map. Therefore, when the subjects finish collecting information, the corresponding heat-map will contain the attention information

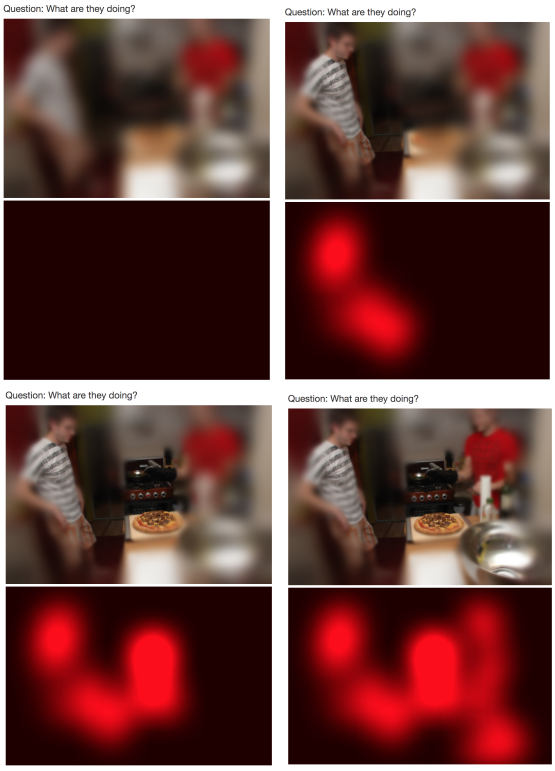

Figure 4: The process to obtain human attentions A.

obtaining by the subjects.

Meanwhile, to get temporal data and study on the progress of human attention, we will record the heat-maps generated by the same subject at different times. We observe that different subjects de-blur the images at different speeds, so it might not be a good idea to record the heat-maps based on time for different subjects. Instead, we record the heat-map each time the subject press and release the mouse key. One example is shown in Figure 4, where we collect four(including the initial one) heat-maps from a subject at different times.

Also, it is also worth noting that the original heat-map is not completely black. This is because subjects can already obtain some information from the blur images; so it is better to initiate a value for each pixel at the beginning. We stick with 30 from the original Georgia Tech Interface, while the maximum value is 255 based on the classical RGB standards.

After this step, we have the attention information A from our subjects.

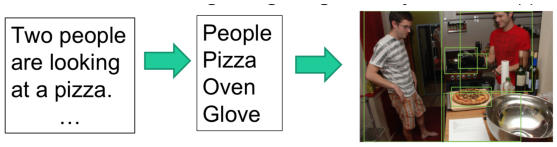

Figure 5: The process to obtain oral justification J.

3.2 Justification collection

In this part, we collect the oral justifications J, or the reasons behind answers from the subjects.

After subjects generate the heat-maps and answer the questions, we ask the subjects his/her reasons, and we record their text reasons as sentences. Then, we parse these sentences using Stanford Dependency Parser mentioned above to obtain key nouns in the sentences. And finally, we look back on the original images and label these corresponding nouns on the images. These regions are recorded as J, the oral justifications of the subjects.

Figure 5 gives an example of the process. After the subject generated heat-map and answered questions("They are cooking a pizza"), he provided his reasons: "Two people are looking at a pizza; they are besides an oven; one of them is wearing a glove." Then, we use the dependency parser to obtain the key nouns: "people", "pizza", "oven", and "glove". Finally, we label the corresponding entities on the images. Therefore, we obtain the justification information J of this subject on this particular question.

Notice that the regions of justifications usually overlap with each other. For example, in Figure 5 the right "person" overlaps with the "glove" and part of the "oven". Because of this, we purpose particular formulas to calculate the ratios. We will introduce the detailed formula in the next section.

3.3 Proposed criteria

After we obtain the attention information A and the justification information J as regions on images, we can start to study on their relations. Specifically, we are interested in relations from two directions, which are the two criteria in this work:

c1. How much J comes from A?

c2. How much A contributes to J?

These two criteria evaluate relations of J and A from different perspectives. For criteria 1, if much of J comes from A, it means the oral justifications

Figure 6: Justification regions usually overlap with each other.

brought by subjects can be mostly found in the attention regions. For criteria 2, if much of A contributes to J, it means most of the attention from subjects is useful and provides important information to conclude a final answer.

As fore-mentioned, for justifications J, the regions usually overlap with each other, which is not completely aligned to the regions of A. Therefore, we purpose two formulas with slightly different forms for these two criteria respectively.

3.3.1 c1. Ratio of justification J

The ratio of justification J is calculated using the formula below:

$$c1_t = \frac{\sum_i (J_i \cap A_t)}{\sum_i J_i}, \tag{1}$$

where J_i denotes the justification region of different labels, and $i \in \{1, 2, ..., l\}$ denoting the index of each label, where l is the number of labels. Also, A_t denotes the attention region at different time, where $t \in \{1, 2, ..., T\}$ and T is the total number of heatmaps generated for one image.

Here we provide a more straightforward explanation in Figure 6-8. From Figure 6, we observe that justification regions usually overlap with each other. Since we are trying to calculate the ratio of justifications, the problem is: which one should be considered as the total justifications(denominator), the union of justifications $\cap J_i$, or the sum of justifications from different labels $\sum_i J_i$? To resolve this, we take a close look at Figure 7 and Figure 8 as examples to see which is more reasonable. We notice that, if the subject pays attention to regions with

Figure 7: Attention on an important region. Heat-map overlaps with J_i corresponding to "oven", "glove" and "people".

Figure 8: Attention on a not important region. Heat-map overlaps with J_i corresponding to "table".

Figure 9: Attention on a not important region. Heat-map overlaps with J_i corresponding to "table".

many justification labels(Figure 7), this attention region is more important than regions with fewer labels(Figure 8). Also, from the view of justifications, if there is an attention region in Figure 7, that means the "oven", "glove" and "people" are all explained, which is different from Figure 8, where only the label"table" is explained. Because of this, it is better to calculate the overlap area based on each label, i.e. each J_i where $i \in \{1, 2, ..., l\}$. Therefore, in Figure 7, the heat-map region will overlap with the "glove", "oven" and part of the "people", while in Figure 8, the heat-map region will only counts when we calculate label of the "table". This is equivalent to add more weight to more important regions.

3.3.2 c2. Ratio of attention A

The ratio of attention A is calculated using the formula below:

$$c2_t = \frac{(\cup J_i) \cap A_t}{A_t}, \tag{2}$$

where J_i and A_t denote the same variables above.

Compared with the first criteria, the second criteria is more straight-forward because the attention is considered as a whole and there are no concerns regarding union or summation. We will use Figure 9 as an example. Suppose the green regions are the attention area, while the blue regions are the label area. We just calculate the fraction between the overlapping area(green and blue) and the attention area(green area). This gives us the ratio of attention that contributes to the justifications.

4 Evaluation and Results

We choose three subjects and images with different actions. The question is always "what are the person/people in the image doing". We obtain the attention and justification mentioned above, and we calculate the two criteria defined above. Finally, we select 5 samples and plot their c1 and c2 in Figure 10 and Figure 11.

The vertical axis stands for the ratio calculated using the formulas given above.

The horizontal axis stands for "period". As mentioned above, we obtain heat-maps at different times when the subject deblurs the images(Figure 4). However, subjects need different chances to deblur a simple image versus a complicated image. Moreover, different subjects also need different chances to deblur a same image. Therefore, we might obtain a different number of heat-maps per image per subject. To handle this, we normalize the period so that each curve ends up with 6 periods, where some of the curve actually does not have that much period. The curves 2317148_c1 in Figure 10 and Figure 11 illustrate this case.

Each curve represents a subject answering for a particular image; "s1", "s2" and "s3" in the label name represents the subject ID accordingly.

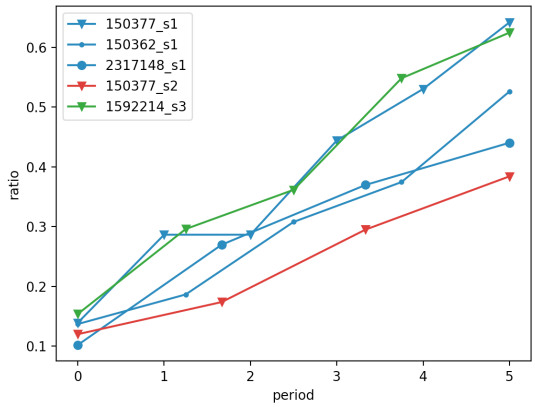

Figure 10: Different samples for the first criteria.

Comparing curves in c1, the ratio of justifications monotonically increases, which is expected, because with volunteers deblur more and more regions, more regions of oral justification can be explained by the users. However, the final ratio is small (0.4-0.6). This means some justification leaves unexplained,

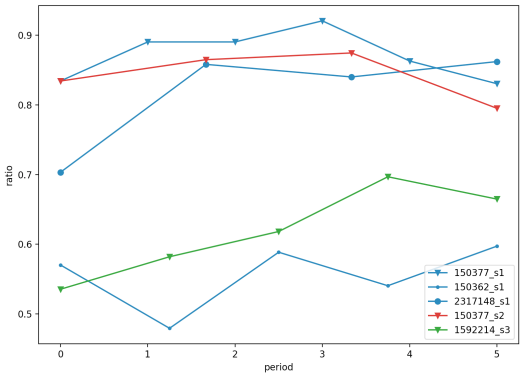

Figure 11: Different samples for the second criteria.

Figure 12: Label region is oversized

Figure 13: Label region is oversized

which contradicts our intuitions.

Comparing curves in c2, the ratio of justifications does not act monotonically. There is a tendency that the ratio first increases, and then decreases. A decreased ratio means the extra attention does not help much in generating justifications(not overlap much with justification regions J). Therefore, this tendency means the subjects tend to use more than necessary attention to answer a question. Meanwhile, c2 is usually high in the end, illustrating most of the attention contributes to the final justifications. However, c2 can perform differently among different images. I will try to analyze this in the discussion section.

5 Discussion

5.1 Low c1

In the result above, we realize the c1 is lower than we expected. From our point of view, the most possible reason is that the defined justification region J is sometimes over-sized. Figure 12 and Figure 13 illustrate two different cases for over-sized labels.

In Figure 12, we label the person, but the rectangular region also contains much area that is not part of the person. The actual person only occupies around 50% of the rectangular box. Therefore, this box will lead to much of the unexplained justifications. In figure 13, the case is a little bit different. Most of the boxes contain chopped vegetables, but the subjects only need to observe a very small part of them to understand the contents in the box. This will also contribute to the unexplained justifications.

How to resolve this? The two cases mentioned above all involve labels with large regions. We think we can assign a weight w_i to each justification box, and $w_i \propto \frac{1}{J_i}$. With the weight mechanism, labels with large regions count less in the criteria. Also, we can set a threshold for each label, while the threshold is approximately proportional to the area of regions.

5.2 c2 across different images

We observe that c2 has a large difference among different images. We believe this is because some images are harder to decipher than others. Figure 14 gives an example of around 0.6(low) c2. Observing this image, we find the key evidence, chicken, is in the oven and hidden by other objects. This might bring difficulties to our subjects. Overall, we believe the criteria difference among different images can be considered as a normal phenomenon.

Figure 14: An image with low c2

5.3 Different weights on labels

We have proposed a different weight assignment in the previous discussion. Besides that, we also believe it is necessary to assign weights based on the parse results. For example, given the reason text **"One of the people wears gloves for the oven"**, "gloves" and "one of the people" are near the roots; while "oven" is far from the root. We may consider assign larger weights to "gloves" and "people", and assign a smaller weight to "oven".

5.4 Different formula for c_1

We also proposed a different formula for calculating c_1:

$$c1 = \frac{\sum_i \frac{J_i \cap A_t}{J_i}}{i}. \quad (3)$$

And we also provide the original c1 formula (Equation 1) here for your reference:

$$c1_t = \frac{\sum_i (J_i \cap A_t)}{\sum_i J_i}.$$

The original proposed formula(Equation 3) adds each label regions together, where labels with large regions are dominant. We can consider using the new proposed formula, where each label is calculated separately and we take the average of the ratio. While the new proposed formula equally considers all labels, it can subject to random disturbances when the regions are too small. So we finally choose formula 1 as the criteria definition.

5.5 Verbs in justification and corresponding labels

During testing, we also notice that some verbs in justification may be important. Considering the following part of text reasons with Figure 15. Notice there are other sentences in the reasons of this subject.

"This gentleman is looking at an image."

Figure 15: Example when verbs also contribute to justifications

Besides "gentleman" and "image", we believe "look" is also an important justification. Considering if the person is looking at the window instead of the image, in which case we will not still consider he is painting.

We think labeling the verb is equivalent to label the corresponding elements in an image. For example, if we want to label "look" in the image, we may want to label the head of the person. However, this may require additional work, because traditionally pre-labeled datasets do not label any objects that are parts of a large object.

5.6 Low amount of data

And finally, we have very limited subjects in our experiment. We only have three subjects, which might bring unstable results. We are studying on Amazon Mechanical Turk(AMT) crowd-souring platform to see if we can setup our tests online. With more subjects we can get more confident results.

6 Conclusion

In this work, we research on the relations between human attention and oral justification on questions

with images. We collect attention by generating heat-maps corresponding to the regions, and we collect justifications by collecting linguistic reasons, parsing them and relabelling them on images. Finally, we compare these two regions to get two criteria for their relations.

Based on our current result, we can answer the two problems raised at the beginning. First, the relation between attention and oral justification is non-trivial. Most part of attention contributes to justification, while some part of justification leaves unexplained. Second, with attention grows, some of the attention is actually not necessary and does not contribute to the final justification. We discuss the current result and propose 6 modifications that might improve our experiments. These modifications can be a good start point for future work.

7 Acknowledgement

At the end of this report, we would like to thank professor Joyce and Shaohua to provide suggestions and helps for our work, such as experiment designs and experiment images. Besides that, the Georgia Tech VQA Interface contributes much for obtaining heat-maps. Also, we would like to thanks Stanford NLP research group for the dependency parser used in this work. Please refer to references for details.

References

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*

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*

Stanford Dependency Parser. https://stanfordnlp.github.io/CoreNLP/