Explaining Disagreement in Visual Question Answering Using Eye Tracking

Susanne Hindennach
University of Stuttgart
Stuttgart, Germany
susanne.hindennach@vis.uni-stuttgart.de

Lei Shi
University of Stuttgart
Stuttgart, Germany
lei.shi@vis.uni-stuttgart.de

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

Figure 1: In visual-question answering, annotators often give different answers to the same question-image pair. The figure shows two such examples in which the differences in visual attention, recorded using eye tracking, explain the reason for these different answers, ©HCI-CS University Stuttgart (annotator answers and attention maps), images and questions taken from VQAv2 (https://visualqa.org/download.html)

ABSTRACT

When presented with the same question about an image, human annotators often give valid but disagreeing answers indicating that their reasoning was different. Such differences are lost in a single ground truth label used to train and evaluate visual question answering (VQA) methods. In this work, we explore whether visual attention maps, created using stationary eye tracking, provide insight into the reasoning underlying disagreement in VQA. We first manually inspect attention maps in the recent VQA-MHUG dataset and find cases in which attention differs consistently for disagreeing answers. We further evaluate the suitability of four different similarity metrics to detect attention differences matching the disagreement. We show that attention maps plausibly surface differences in reasoning underlying one type of disagreement, and that the metrics complementarily detect them. Taken together, our results represent an important first step to leverage eye-tracking to explain disagreement in VQA.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Supervised learning*; *Reasoning about belief and knowledge.*

KEYWORDS

Eye Tracking, Visual Question Answering, Disagreement, Human Label Variation, Ambiguity

ACM Reference Format:
Susanne Hindennach, Lei Shi, and Andreas Bulling. 2024. Explaining Disagreement in Visual Question Answering Using Eye Tracking. In *2024 Symposium on Eye Tracking Research and Applications (ETRA '24), June 4–7, 2024, Glasgow, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3649902.3656356

ETRA '24, June 4–7, 2024, Glasgow, United Kingdom

ACM ISBN 979-8-4007-0607-3/24/06
https://doi.org/10.1145/3649902.3656356

1 INTRODUCTION

Supervised machine learning algorithms require ground truth annotations that are created by asking multiple humans to perform the task at hand and recording their answers. These annotations typically consist of multi-category labels or short free texts. In visual question answering (VQA), this means collecting the answers of multiple people to the same question about an image. Frequently, this human annotation process does not result in a single label that all humans agree on. Instead there can be multiple answers. This phenomenon has been studied under different terms: disagreement [Bhattacharya et al. 2019; Gurari and Grauman 2017], human label variation Plank [2022], uncertainty [Peterson et al. 2019], or ambiguity [Stengel-Eskin et al. 2023].

In most machine learning applications disagreement is resolved by aggregating the varying labels into one (for example by taking a majority vote). The underlying assumption is that there is one true label and that all variation is caused by annotation errors (mostly caused by attention slips). However, there is a growing body of research that questions this assumption and that has shown that a large proportion of differences in labels have plausible reasons [Bhattacharya et al. 2019; Plank 2022; Stengel-Eskin et al. 2023]. In these cases, there is not one single ground truth but instead the differences contain valuable information about the task (for example, that it is ambiguous or subjective). It remains an open challenge how to disentangle the different reasons for disagreement, and consequently account for them correctly in training and evaluating machine learning algorithms.

Visual question answering (VQA) is particularly suited to study disagreement [Gurari and Grauman 2017]. A number of previous works have investigated the different reasons underlying disagreement in VQA [Bhattacharya et al. 2019; Stengel-Eskin et al. 2023]. Bhattacharya et al. [2019] have provided a first taxonomy of possible reasons. They found that the most common reason for disagreement was *ambiguity* in the question-image pair. Answers with *synonyms* were the second most frequent reason, and the third most common reason were *different levels of granularity* in the answers. Stengel-Eskin et al. [2023] further classified the ambiguous cases according to the factors that made the question ambiguous. They found that ambiguity mostly arose from questions about the object's location and the kind of object, and questions for which annotators could choose one of multiple options.

In this work we study whether eye tracking data recorded during ground truth annotations for VQA can help to understand how the annotators arrived at their answers. Eye-tracking data cannot explain all different causes for disagreement. For example, synonymous or subjective answers will probably arise from looking at the same aspect of an image. However, we hypothesise that ambiguity caused by multiple answer options presented in the image could be explained by comparing the attention maps of the two annotators (as shown in Figure 1). To this end we leverage the VQA-MHUG dataset of eye-tracking data recorded during annotation [Sood et al. 2021]. In a first step, we perform a manual inspection to find cases for which the differences in visual attention explain the reason for the disagreeing answers. We then propose four metrics of attention similarity and evaluate their suitability to detect these cases.

2 APPROACH

To investigate whether differences in visual attention can explain disagreement, we leveraged the recent VQA-MHUG dataset [Sood et al. 2021] containing human gaze on image and text in visual question answering. We then proceeded in two steps. First, we did a manual inspection of attention maps on images with disagreeing answers. We then computed four similarity metrics (Pearson's Correlation Coefficient, Shared Fixation Count, Shared Fixation Duration, and Sequence Score) for all answer pairs (agreeing and disagreeing), and evaluated their ability to detect reasons of disagreement explained by differences in visual attention.

2.1 Dataset

To investigate visual attention differences for disagreeing answers we used the VQA-MHUG dataset [Sood et al. 2021]. The dataset contains human gaze from 49 annotators performing a visual question answering task. The images and questions are a subset of VQAv2 val [Goyal et al. 2017][1]. Each of the 3,990 question-image pairs was answered by three different annotators. The answers differed between pairs of annotators for 2,119 (more than 50%) of the question-image pairs. We created human attention maps based on the eye tracking data by smoothing the detected individual fixations with a Gaussian kernel with $\sigma = 1$.

2.2 Manual Inspection

In our manual inspection, we were looking for examples, where the difference in attention maps would provide insight into the different reasoning of the annotators. For example, in the first row of Figure 1, one annotator responded "priest" to the question "What is the profession of the man on the right?". Their attention map shows that they focused on the white rectangular collar, a Christian clerical clothing worn by priests. The second annotator, on the other hand, focused on the drum in the background, and correspondingly responded "drummer". Similarly, in the second row of Figure 1, asked about the flavour of the dessert in the image, one annotator focused on the chocolate sprinkles and responded "chocolate", whereas the other focused on the strawberries and responded "strawberry". As such, the attention maps explain the difference in answers, because one could imagine that if the annotators were asked to explain their answers they would each point to the visual evidence in the image highlighted by the attention maps. To find such cases, we applied the following exclusion and inclusion criteria to the 702 question-image pairs in which all three annotators gave different answers. We excluded cases where the reasons for the different answers were attention slips (i.e. spelling mistakes or white spaces as in the second row of Figure 4), or synonyms (as in the third row of Figure 2). We included cases where at least two of the three annotators gave valid but different answers (i.e. where there was some visual evidence for the answer in the image) and where these annotators looked on different parts of the image (i.e. where there was a visually perceptible difference in the attention maps). This resulted in 45 examples in which the difference in attention explained the disagreeing answers.

[1] https://visualqa.org/download.html

2.3 Differences in Human Attention Maps

We next wanted to quantify the difference in human attention maps in order to help identify cases in which the attention maps explain the disagreeing answers. A large body of work has explored different metrics to evaluate saliency prediction methods [Bylinskii et al. 2019]. We opted to use Pearson's correlation coefficient (CC), as well as three measures that account for semantic differences in the attention maps. As the images in VQA-MHUG are drawn from VQA2.0, we used the object segmentation from MS COCO [Lin et al. 2014] for the semantic scores. We computed all four metrics for all pairs of annotators (who both looked at the same image, and answered the same question) in the dataset.

2.3.1 Pearson's Correlation Coefficient (CC). Pearson's correlation coefficient (CC) is a common metric to measure differences that treats the two attention maps as random variables, and measures the linear relationship between them:

$$CC(Q_1, Q_2) = \frac{\sigma(Q_1, Q_2)}{\sigma(Q_1) \times \sigma(Q_2)}, \tag{1}$$

where Q_1 and Q_2 are the human attention maps of two annotators answering the same question-image pair, respectively, and σ is the covariance.

2.3.2 Shared Fixation Count (SFC). The first semantic score measures how many times two annotators fixated on the same objects in an image (Shared Fixation Count, SFC). We used the segmented objects, and counted the number of fixations. As different annotators had varying number of total fixations on an image, we normalised by the sum of fixations on the image to compute the ratio of fixations. For a pair of annotators, we then took the minimal ratio for each object. The minimal value represents the amount of shared fixations between two annotators. We then took the sum of this value across all objects.

$$SFC = \sum_{o=1}^{nr_O} min(n_{o,1}, n_{o,2}), \tag{2}$$

where $n_{o,1}$ and $n_{o,2}$ are the normalised number of fixations on object o of two annotators, respectively, and nr_O is the total number of annotated objects in the image. This results in a value between 0 and 1, where 1 means that the two annotators distributed their fixations in the same way across the objects of the image, and 0 means that the annotators did not fixate on any same object in the image.

2.3.3 Shared Fixation Duration (SFD). Analogously, the second semantic metric represents the duration of fixations spent on the same image (Shared Fixation Duration, SFD). We computed the total time spent fixating on each object in an image for each annotator and normalised it by the total fixation duration on the image. Then, we again took the minimal value for each object for a pair of annotators, and computed the sum of fixation duration across all objects.

$$SFD = \sum_{o=1}^{nr_O} min(dur_{o,1}, dur_{o,2}), \tag{3}$$

where $dur_{o,1}$ and $dur_{o,2}$ is the normalised duration of fixations on object o of two annotators, respectively, and nr_O is the total number of objects.

2.3.4 Sequence Score. As a third semantic metric we used the sequence score proposed by [Chen et al. 2021]. We used the object segmentation to create strings that describe the order in which annotators fixated on objects in the image. For each of the 91 object types provided in MS COCO [Lin et al. 2014] we added the corresponding letter ("P" for "person" if there was a person in the picture). We did not differentiate between different instances of the same category, i.e. in the picture shown in the first row of Figure 1 fixations on both men would result in a "P". We deleted repetitions (consecutive fixations on the same object category). This resulted in one string per annotator representing the sequence of fixations. We then used the Needleman-Wunsch string matching algorithm [Needleman and Wunsch 1970] to calculate the similarity between two annotators' sequences for the same question-image pair.

2.4 Evaluation of Metrics

To evaluate whether the metrics can be used to identify cases in which the attention maps explain the (dis)agreement we filtered for true and false examples (see Figure 2 and Figure 3 for illustrative examples).

2.4.1 True Examples. We defined three different types of true cases: true positives and two types of true negatives (shown in Figure 2). True positive examples are those which have a low value in the metrics (indicating a large difference in the attention maps), and the differences in visual attention explain the disagreeing answers according to manual inspection (first row of Figure 2). There are two versions of true negative cases: First, cases in which the metrics are high (indicating similar attention maps) and the answers are identical (second row of Figure 2). Second, cases in which the metrics are high and the answers are different and the disagreeing answers can be explained by another reason than visual attention (for example, synonyms as seen in the last row in Figure 2).

2.4.2 False Examples. We analogously defined three types of false cases: False negative and two types of false positives (see Figure 3). False negatives are cases in which the differences in the attention maps explain the disagreeing answers according to manual inspection, however, the metric is high indicating that the attention maps are similar (first row in Figure 3). The first type of false positive examples are cases in which the metrics are low, but the answer is identical (second row in Figure 3). The second type of false positives are cases in which the score is low and the answers are different, but the difference in visual attention does not explain the disagreement according to manual inspection (last row in Figure 3).

2.4.3 Comparison of Metrics. The metrics result in similar values (high or low, respectively) for the examples shown in Figure 2 and Figure 3. However, there were also cases in which their values differed. We show examples for cases in which only Pearson's correlation coefficient yields the intended results in Figure 4. In these cases, the fixations of one or both annotators are outside any of the provided objects, hence the semantic scores cannot capture differences or commonalities for those. In two cases, the different or shared fixations are on objects that are not part of the object segmentation (the drum in the first row, and the beach in the second row of Figure 4). In the other two cases, the object segmentation is missing details or too detailed: In the third row of Figure 4, the

Figure 2: Sample question-image pairs for which the metrics indicate cases in which the differences in visual attention explain the reason for (dis)agreeing answers. The first row shows a true positive example: the two annotators focus on different aspect (the kite and the door, respectively, and answer accordingly). The second row shows a true negative example in which the annotators agreed and focused on the same aspect. The third row also shows a true negative example, in which the annotators gave different but synonymous answers ("talk" and "speak") and focused on the same aspect. In each row, the subtitle shows the question. The first column shows the original images and the second column shows the object segmentation from MS COCO [Lin et al. 2014]. The third and fourth column show two annotators' individual attention maps and their respective answers underneath. In the rightmost column, the metrics for the pair of attention maps are given. ©HCI-CS University Stuttgart (annotator answers and attention maps), images and questions taken from VQAv2 (https://visualqa.org/download.html)

fixations on the cookies are not detected, and in the last row of Figure 4 the detailed segmentation of racket and woman are smaller than the blurring of the Gaussian filter used to create the attention maps. Similarly, Figure 5 shows examples in which only the three semantic metrics detect cases in which differences in visual attention explain the disagreement. In these cases, the correlation coefficient either underestimates small, but semantically relevant, differences (the child in the background in the first row of Figure 5) and overestimates big, but semantically irrelevant differences (the difference in fixation on the zebra's head in the second row of Figure 5). In our examples, there was no qualitative difference between the count-based SFC and the duration-based SFD. Lastly, we found one example in which only the sequence score is low for a disagreement that was identified by manual inspection shown in Figure 6. Here, the order of fixations seems to capture differences, when there are many equally fixated objects in the image.

3 DISCUSSION

3.1 Disagreement Explained by Visual Attention

Based on our manual inspection and the evaluation of metrics, we found clear-cut examples of disagreement that can be explained by differences in visual attention. These examples share that there are multiple options of how to answer the question in the image, and the question is not specific which of them should be used (like in the first row of Figure 2 where there are multiple kids, and it is not clear which of them is meant by the question). In these cases, the annotators focus on different non-overlapping options, and consequently provide different answers. We call these cases disagreement explained by visual attention, as the eye-tracking based attention maps highlight the option in the image matching the answer, respectively. According to the work by Stengel-Eskin et al. [2023], ambiguity caused by multiple options is a frequent reason for disagreement in VQA. Hence, it seems worthwhile to leverage eye-tracking data to detect and explain this type of disagreement. Human-attention enhanced models might even be able to generate the differing answers depending on the provided focus on one of the options. There are also less clear cases in which the attention maps help to understand the disagreement to some extent. These cases require additional interpretation and/or knowledge that go beyond what is in the image (such as in the first row of Figure 1). In the taxonomies of reasons for disagreement, such cases would fall in either the insufficient visual evidence, uncertainty or difficulty category [Bhattacharya et al. 2019; Stengel-Eskin et al. 2023]. Here,

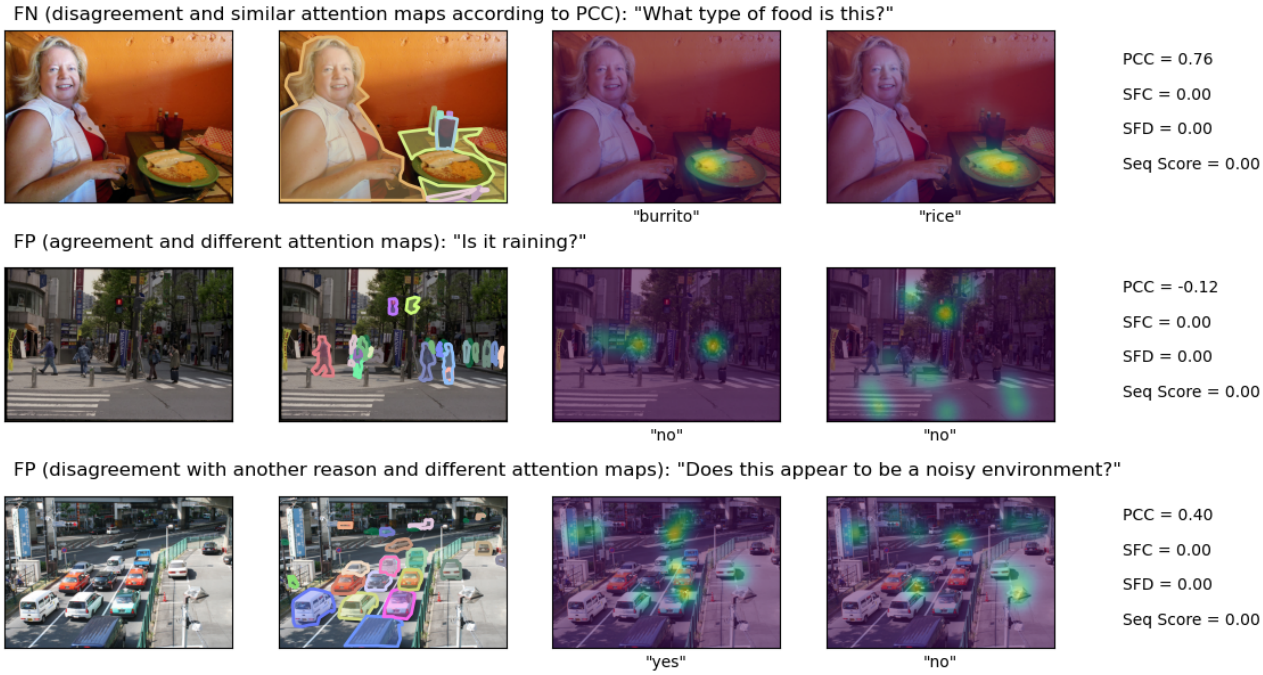

Figure 3: Sample question-image pairs for which the metrics do *not* indicate that the visual attention explains the (dis)agreeing answers. The first row shows a false negative example: The annotators focused on different food on the plate reflected in their answers ("burrito" and "rice"), but the metrics do not capture the difference. The second row shows a false positive example, in which both annotators give the same answer but they focus on different aspects in the image. The third row shows a false positive example with disagreeing answers and different attention maps, however, the difference does not explain the disagreement. ©HCI-CS University Stuttgart (annotator answers and attention maps), images and questions taken from VQAv2 (https://visualqa.org/download.html)

the attention maps can highlight the little evidence int the image that the annotator was relying on.

3.2 Detecting Disagreement Explained by Visual Attention

In our evaluation, we compared the ability to detect disagreement explained by visual attention of Pearson's Correlation Coefficient and three semantic scores. We found that Pearson's Correlation Coefficient was superior when the object segmentation did not cover the objects causing the ambiguity, and complementarily, that the semantic metrics outperformed the correlation coefficient, when all relevant objects were provided. The semantic scores capture even fine differences in attention (see first row in Figure 5) and ignore irrelevant differences across one big foreground object (see second and third row in Figure 5). However, in many cases, the provided objects did not match the necessary level of detail to explain the disagreeing answers. If the different or shared attention is outside all objects, the semantic scores are uninformative (first and second row in Figure 4). Similarly, the semantic scores do not work as intended, if the object segmentation does not cover enough details (third row in Figure 4) or encompasses more details than the resolution of the eye-tracking data allows (fourth row in Figure 4). These problems could be solved by question-specific object segmentation that take the possible ambiguities into account and include all relevant objects at the right level of detail. Hence, the choice of metric depends on the availability of suitable object segmentation.

However, our evaluation also shows that the explored similarity metrics were not able to identify the clear cases out of all different types of disagreement. Hence, additional methods are necessary to filter out disagreements caused by other reasons like synonyms or subjective answers. It remains an open research challenge to classify and detect reasons for disagreement, but our work shows that eye tracking data could be a promising additional input in order to specifically detect disagreement explained by visual attention.

4 CONCLUSION

We explored whether eye-tracking data helps to explain disagreement in visual question answering. Indeed, we found that the differences in visual attention maps highlight the chosen option for disagreement caused by multiple answer options in the image. The four proposed similarity metrics can identify such cases, however, they also falsely detect disagreement caused by other reasons. These findings highlight the importance of further research to explain disagreement in VQA.

Figure 4: Sample question-image pairs for which only *Pearson's Correlation Coefficient* indicates that the visual attention explains the (dis)agreeing answers. In the first two rows, the semantic metrics do not measure the difference because the fixations are outside the objects. The third and fourth row show examples in which the objects are either missing details or too detailed. ©HCI-CS University Stuttgart (annotator answers and attention maps), images and questions taken from VQAv2 (https://visualqa.org/download.html)

ACKNOWLEDGMENTS

Susanne Hindennach and Andreas Bulling were funded by the European Research Council (ERC; grant agreement 801708). Lei Shi was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). We thank Fabian Kögel and Ekta Sood for support with using the dataset.

REFERENCES

Nilava Bhattacharya, Qing Li, and Danna Gurari. 2019. Why Does a Visual Question Have Different Answers?. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 4270–4279. https://doi.org/10.1109/ICCV.2019.00437

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand. 2019. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 3 (March 2019), 740–757. https://doi.org/10.1109/TPAMI.2018.2815601

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. 2021. COCO-Search18 Fixation Dataset for Predicting Goal-Directed Attention Control. *Scientific Reports* 11, 1 (April 2021), 8776. https://doi.org/10.1038/s41598-021-87715-9

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 3511–3522. https://doi.org/10.1145/3025453.3025781

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014 (Lecture Notes in Computer Science)*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Saul B. Needleman and Christian D. Wunsch. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology* 48, 3 (1970), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. 2019. Human Uncertainty Makes Classification More Robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 9616–9625. https://doi.org/10.1109/ICCV.2019.00971

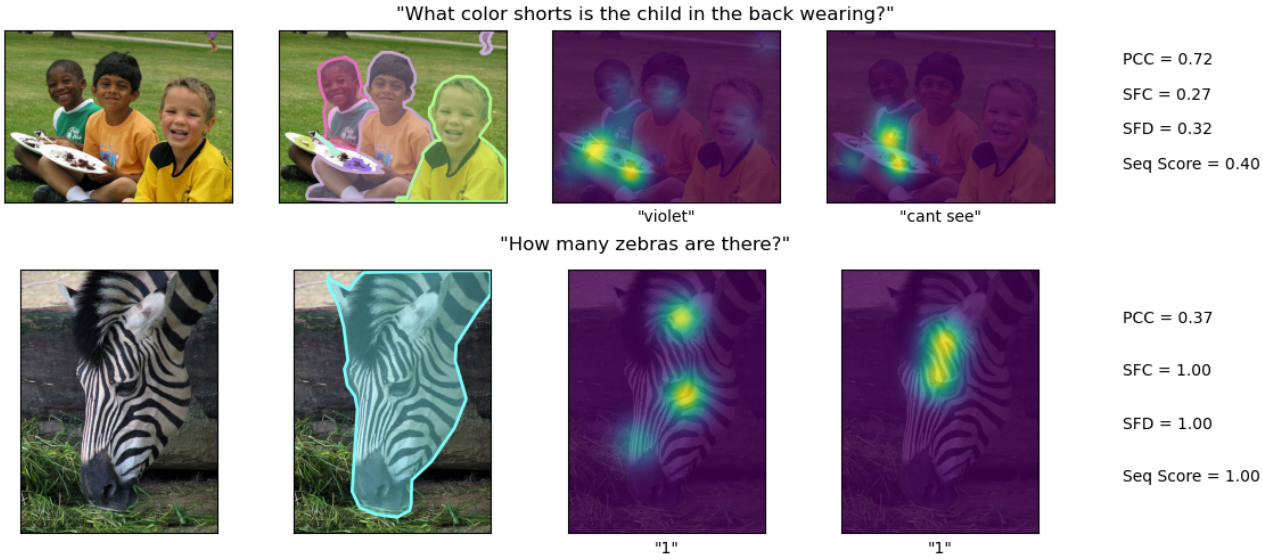

Figure 5: Sample question-image pairs for which only the *semantic metrics* indicate that the visual attention explains the (dis)agreeing answers. In the first row, the attention map of the annotator responding "purple" is slightly lighter in the upper right corner where the child in the back is, which the second annotator does not see. In the second row, both annotators focused on the zebra, yet different areas of its head, and gave the same answer. The attention maps in this figure have a lower transparency to show the subtle differences. ©HCI-CS University Stuttgart (annotator answers and attention maps), images and questions taken from VQAv2 (https://visualqa.org/download.html)

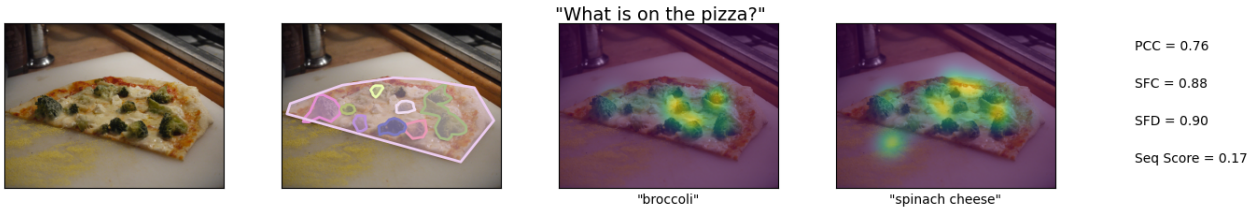

Figure 6: Sample question-image pair for which only *Sequence Score* indicates that the visual attention explains the disagreeing answers. Both annotators focused on the pizza toppings but they stressed different ones in their answer ("broccoli" and "spinach cheese"). This difference is captured by a low sequence score. ©HCI-CS University Stuttgart (annotator answers and attention maps), images and questions taken from VQAv2 (https://visualqa.org/download.html)

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10671–10682. https://doi.org/10.18653/v1/2022.emnlp-main.731

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, 27–43. https://doi.org/10.18653/v1/2021.conll-1.3

Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. 2023. Why Did the Chicken Cross the Road? Rephrasing and Analyzing Ambiguous Questions in VQA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10220–10237. https://doi.org/10.18653/v1/2023.acl-long.569