Evaluating Interpretability Techniques with Human Attention in Visual Question Answering

Michael Tang[1] **Jason Yuan**[1]
[1]Department of Computer Science, Princeton University
{mwtang, jcyuan}@princeton.edu

Abstract

In this paper, we explore interpetability in Visual Question Answering. Our model is a baseline CNN+LSTM VQA model, and we implement 4 gradient based pixel attribution algorithms on top of this model (Vanilla Saliency, Guided Backprop, DeconvNet, Guided GradCam). We run the model and attribution algorithms on a subset of the MS COCO VQA dataset, generating a collection of saliency maps, model answers, and model confidences. We then propose and implement a scoring pipeline that computes an ensemble of metrics for how similar an attribution heatmap is to the ground truth human attention from the VQA-Hat dataset [3]. Using these metrics, we analyze the dataset and investigate novel interactions between models predictions, model confidence (pre-softmax score), and model attention as quantified by saliency maps. Namely, a model's attention is most similar to human attention when it produces correct but low-confidence answers, and least similar to human attention when it produces wrong-and-unconfident or correct-and-confident answers. We implement some statistical tests to determine the significance of our results. Finally, we investigate qualitatively some motivating examples from the dataset.

1 Introduction

Visual Question Answering (VQA) is a relatively recent field whose genesis lies in the intersection of computer vision and natural language processing. It poses the following problem: given a question about an image and said image, can a model produce the answer? In order to robustly and accurately solve this problem, the model must not only possess a semantic understanding of the natural language question and a conceptual understanding of the scene depicted in the image, it also needs some way of relating the two modes of information together, and finally formulate a logical natural language response, possibly from a few choices, by considering the relevant properties of the image. As posed by Agrawal et al and formalized in the seminal VQA 1.0 and VQA 2.0 datasets [2], this problem involves questions that range from relatively straightforward (What color is the table?) to very nontrivial and involved (Is this picture staged? Is this food good for someone on diet?) [2].

In the context of this unique interface between vision and language, understanding the role of visual attribution in the context of VQA models is of particularly great interest. If we can clarify the most important parts of a given image, this can also lend itself to applications like preserving privacy in training data, uncovering biases in a black-box-like state-of-the-art models, or detecting and countering adversarial inputs during model inference.

A related problem involves the evaluation of interpretability techniques. Historically, it is quite challenging to evaluate how accurate and effective interpretability techniques are in explaining model behavior [10]. Quantifying the differences between interpretability techniques and examining where they succeed and fail will allow for more informed usage and assist future design choices. The VQA-Hat dataset augments a popular VQA dataset with an array of human-annotated attention maps that we can use as a baseline for comparison. This paper explores how those human attention maps can be leveraged to tackle a few key questions about VQA and interpretability in novel ways:

- Which interpretability techniques provide the best explanations of model behavior for an audience of humans?
- Do models and humans pay attention to the same parts of an image when solving a VQA task?

2 Related Work

Interpretability techniques.

Occlusion and perturbation techniques. These are model-

agnostic methods that work by occluding or perturbing parts of the input image and measuring the change in the output [10]. Examples include: SHAP [9], LIME [11], Extremal Perturbations [4].

Saliency techniques. These are methods that work by computing the gradient of the weights at some layer in the network with respect to the input pixels, for some target class.

- *Vanilla Saliency* computes the gradient of the target class score with respect to the pixels [10].
- *Guided Backprop* and *DeconvNet* are nearly identical to Vanilla Saliency, but each uses a slightly different method to reverse a ReLU layer, the intention being to avoid the activation saturation problem [10] [7].
- *Guided GradCam* backpropagates the target class score to the last convolution layer, and combines it with Guided Backprop [10].

These methods are very fast compared to occlusion and perturbation methods, but can be hard to interpret and are very sensitive to small perturbations, making them somewhat unstable [10]. For example, Kindermans et al [6] showed that with minor perturbations, gradient methods give drastically different results. In investigating the insensitivity of gradient techniques to model and data, Adebayo et al [1] finds that Vanilla Saliency passes sanity checks for sensitivity while Guided Backpropagation and Guided GradCAM fail. Notably, Molnar calls for further quantitative evaluation of these techniques, alluding to further works that questioned the design of Adebayo et al's sanity checks [10]. Thus, these gradient techniques present a clear target for further analysis.

VQA models. We use an implementation of Kazemi and Elqursh's baseline model, which achieves 64.6% and 59.7% accuracy on the VQA 1.0 and VQA 2.0 challenges, respectively [5] [2]. The architecture consists of a ResNet152 to encode the image, a multilayered LSTM to encode the question, and then additional layers ending with a Softmax classifier (over the 3000 most frequent answers in the VQA dataset). This structure is relatively simple and its design is inspired by previous image classification model architectures, which makes it a useful and tractable entry point into VQA interpretability.

VQA dataset. We evaluate on the VQA 1.0 validation dataset[2], which uses images from the MS COCO dataset [2]. We specifically focus on the Multiple Choice data on Real Images, as these are the most similar to the image classification problems on which interpretability has been previously studied; as such, each example consists of an image, question, candidate answers, and a ground truth answer. The dataset has 40,504 images and 121,512 questions, with 3 questions per image.

Human-annotated attention maps. The VQA-Hat dataset [3] introduced ground truth human attention for comparison with attention-based VQA models. The dataset has 4,122 human-annotated attention maps corresponding to questions in the VQA1.0 validation dataset. In the paper, Das et al also compare their human attention maps with the internal attention maps of state-of-the-art attention-based models of the time, including the Hierarchical Co-Attention model (HieCoAtt) [8] and Stacked Attention model (SAN-2) [12], and find positive correlations between the attention maps. Here, we are interested in extending this analysis to interpretability techniques on a more general class of models and consider other evaluation metrics, leveraging the human maps as ground truth data.

3 Methods

Generating saliency maps. We implement a total of 4 attribution algorims in the baseline VQA model: DeconvNet, Saliency, GuidedBackProp, GuidedGradCam on the model, with some code from the Captum library [7]. We run the model and attribution methods on all examples with corresponding human attention maps available. If the model predicts the correct answer, we only generate a heatmap with respect to the correct target class. If the model generates an incorrect answer, we generate two heatmaps: one with respect to the correct answer and one with respect to the model's predicted answer. Since the baseline model treats the VQA examples as classifications problems on the possible answers, we disregard cases where the labeled answer is not in the model's vocabulary. Due to computational constraints, we limited our analysis to approximately the first 300 examples in the dataset.

3.1 Saliency map evaluation

Preprocessing. For the multichannel saliency maps, we convert them to a single channel by taking the channel-wise max of their absolute value. This mimics the behavior of humans when annotating their attention, as the pixels of objects that hint towards or away from a class are given the same positive attention score, and taking a channel-wise max instead of a sum or mean avoids over-weighting objects with darker colors or diluting the weight of objects where some channels are unhelpful, respectively

Then, each saliency and human attention map is resized to a uniform 448-by-448 size, allowing for elementwise comparisons. We also standardize the range by subtracting the min and dividing by the max, and normalize the distribution by subtracting the mean and dividing by the standard deviation. Since our saliency methods are not area-invariant,

we standardize the total weight in the map. Note that to prevent underflow, all divisions add an epsilon to the denominator (e.g. 1e-6). In cases of qualitative visualization, we found empirically that taking a sigmoid transform allows for a clearer and more consistent visual across different types of saliency maps.

Discretization. We note that the distribution of weights in the human-annotated attribution maps is bimodal, with approximately $k = 10\%$ of pixels given high weight and the rest given fairly low weight. This lends itself readily to being treated as a region in the image, with the high weight pixels inside the region and the rest of the image outside. We can looks at this as a discretization of the continuous map into one with binary values, effectively transforming it from soft to hard attention.

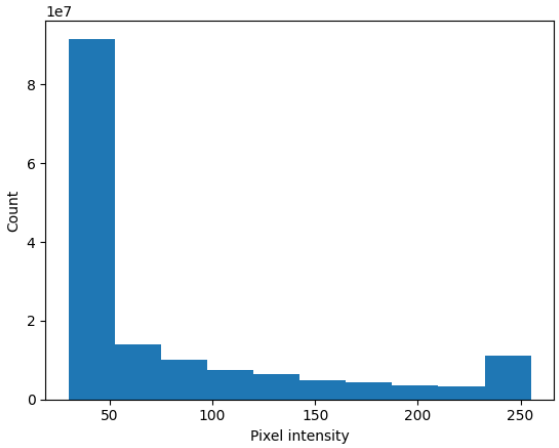


Figure 1: Mean pixel weight distribution over human attention maps

Thus, to compare the saliency maps with the human attention, we extract the same constant area of region by choosing the highest weighted $k(448)^2$ pixels. This discretization allows us to view the similarity problem as finding the correct segmentation of the region out of the image or the correct classification of pixels into important or not important, lending itself to experimentation with techniques from image segmentation and classification evaluation such as intersection over union, precision, recall, f1 score.

Similarity metrics. To evaluate different attribution techniques, we applied different metrics that quantified the similarity between attribution heatmap and the human-annotated attention map, some inspired by metrics from image segmentation and classification.

- Spearman rank correlation
- Intersection, i.e. elementwise multiplication
- Intersection after discretization
- Intersection over union (IoU), after discretization.
- Precision, recall, f-1 score, after discretization. True positive is the intersection, false positive is area inside the saliency map but outside the human map, etc.

4 Results

4.1 Using map similarity to compare interpretability techniques

We aggregate all the metrics for all normalized maps by means.

att_type	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	dice	f1
deconv	0.562406	0.072938	0.000006	0.205937	0.310760	4835.212156	0.111235	0.170511	0.237347
gbackp	0.571475	0.127514	0.000008	0.389583	0.221820	3699.368627	0.092519	0.162154	0.259783
ggrdcm	0.571773	0.094544	0.000008	0.410633	0.104392	1683.751717	0.049129	0.089887	0.159036
salien	0.562406	0.118860	0.000007	0.308128	0.336146	5474.525229	0.124014	0.209920	0.308305

Figure 2: Mean values for attribution technique metrics

From these results alone, we see that Gradient Saliency is the ”most-similar” to the human attention maps. Compared to the other attribution techniques, Saliency similarity metrics are higher in the f-1 score, Discrete intersection, Recall, and Intersection-over-Union. It scores close to the highest in Rank correlation, Intersection, and scores below median in precision.

4.2 Map similarity and model behavior

Quantitative Analysis.

First, consider the question: do the similarity metrics between model attribution and human attention show differences when the model predicts the correct answer and when it predicts the wrong answer? Figure 7 shows the median values of the similarity metrics across a wide variety of scenarios. In every similarity metric, it is true that the when the model is correct (blue) it scores higher than when it is wrong (orange). This seems to align with intuition: in general when the model is correct it is looking at similar things that humans are.

Next, define confidence as the Softmax probability that the model assigns on the class it predicts. Denote low-confidence as $p < 0.25$ and high-confidence as $p > 0.75$. The data can be split into four components: when the model prediction is correct versus wrong and when the model is low-confidence versus highly-confident. We are interested in analyzing the differences between the similarity metrics

of the human attention and saliency maps across these four components.

In order to determine whether there are significant differences between components, we use a two-sample t-test for each attribution technique with $\alpha = 0.05$. We also generally find that Guided Grad-CAM has by far the most significant differences signalled by having the lowest p-values for the comparisons we consider, and thus focus on its results going forward. Looking qualitatively at some of the Guided Grad-CAM outputs, we may hypothesize that this is due to the technique giving highly weighted pixels that are more closely clustered that those of other maps, with a distribution that is also more visually similar to that of human attention maps.

We find that similarity metrics between examples the model predicts correctly and incorrectly are not significantly different, with a p-value of 0.225 for rank correlation, 0.259 for IoU, and 0.259 for f1.

However, if we examine only examples where the model has low confidence, the differences in correctly and incorrectly predicted examples is significant across all metrics, with a p-value of 0.009 for rank correlation, 0.003 for IoU, and 0.004 for f1. The same does not hold for high confidence. Mean differences (correct - incorrect) are displayed in 3, and we see a small positive difference, indicating maps of correct predictions tend to be somewhat more similar to the human maps. For high confidence examples, there is no clear trend. We may conjecture that lower confidence predictions may represent model behavior that is more human-like, e.g. perhaps naively looking at a larger portion of the image, whereas in higher confidence predictions models may look at regions that are more different, e.g. perhaps only needing a distinctive part of an object to identify it while the human highlights the entire object.

Moreover, it appears that similarity may be a good predictor of model confidence on correct predictions. We see that this comparison yields a significant difference between the low and high confidence correctly predicted examples, with p-values of 0.014 for rank correlation, 0.008 for IoU, and 0.009 for f1. The figure 5 shows mean differences (high - low confidence) and we see most metrics have small negative values, meaning that counterintuitively the low confidence predictions tend to have more similar saliency maps to human attention. Looking at the pearson correlations 4 between confidence and similarity metrics among correct model predictions, do in fact get a weak negative correlation, which supports our above finding and conjecture.

We see no significant difference by the t-test between high and low confidence among incorrect predictions, which supports our intuition that when the model is wrong, the similarity of its saliency map has little to do with how confidence it is, e.g. it may be picking up on the wrong signals altogether. This aligns with the lack of a clear sign in the mean differences 6.

We see that we can draw similar conclusions from looking at the big picture, directly comparing these four components using the base attribution technique, Gradient Saliency, which scored most highly in our comparison in section 4.1.

Fixing the confidence level at low, we compare when the model is correct versus wrong. In figure 7, the median similarity value for attribution heatmaps and human attention during a low confidence but wrong answer and a low confidence but correct answer are labeled in purple and green, respectively. In every similarity metric we see that green (correct, low confidence) far exceeds purple (wrong, low confidence), and the difference between the two is significantly larger than the difference between blue (correct) and orange (wrong). The similarity metrics for Green (correct, low confidence) also far exceeds blue (correct). Comparing Figure 8(a) and figure 8(b) gives us a more detailed view. We see that the similarity metrics are higher when the model is correct: in every single similarity metric, the correct cases have a higher median, mean, and more right-skew. The interpretation of this is that the similarity between the model and humans (ie what each is looking at) is highest by far when the model is not-confident yet produces the correct answer.

Fixing the confidence level at high, we compare when the model is correct versus wrong. From figure 8(c) and figure 8(d) we see that the Saliency heatmap similarity metrics are higher generally when the model is wrong. The distributions are more right-skewed in figure 8(c) and the medians and means are all higher than figure 8(d). Looking at figure 7, we see that the similarity metrics in the case where the model is wrong but highly confident (brown) is in general significantly higher than when the model is correct and highly confident (red) and when the model is just wrong (orange). The interpretation of this is opposite of when the confidence of the model is low; now we find that when the model is highly confident, it tends to look at the same things that humans do when it is wrong more so than when it is right.

Looking at the entire picture, figure 7 tells us that when the model is correct but unconfident in its answer, it's attribution heatmaps is most similar to human attention by far. The next situation where the model's attention is most similar to humans is when it is highly confident but wrong in its answer. Finally, the model's attribution heatmaps are least similar to human attention when it is wrong-and-low confidence or correct-and-high confident.

This is quite a subtle result, as it contradicts two common intuitions explanations of these models: (1) when a model

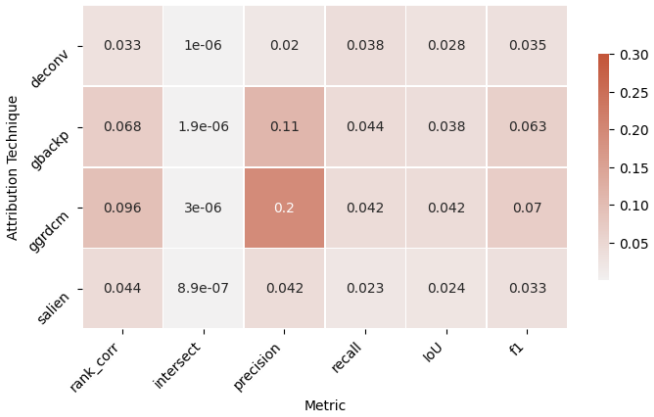


Figure 3: Mean similarity differences between correctly and incorrectly predicted examples (correct - incorrect).

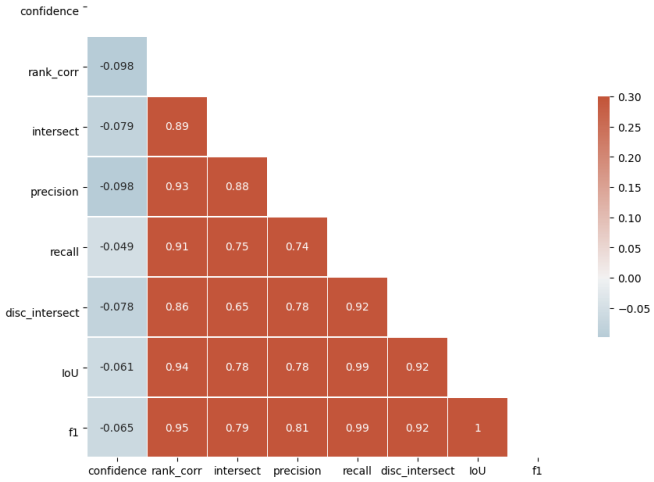


Figure 4: Leftmost column shows correlations between confidence and similarity metrics for correctly predicted examples.

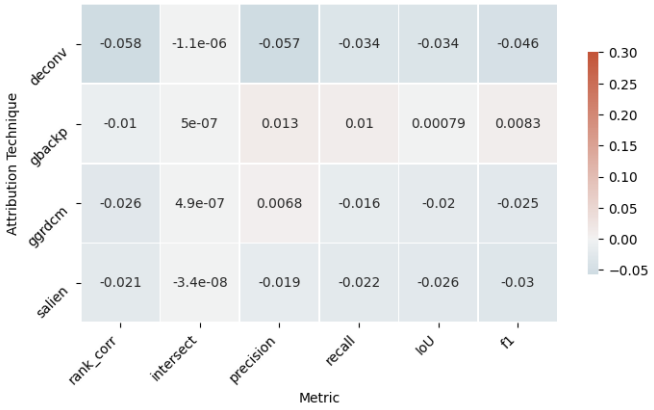


Figure 5: Mean similarity differences between high and low confidence correctly predicted examples (high - low).

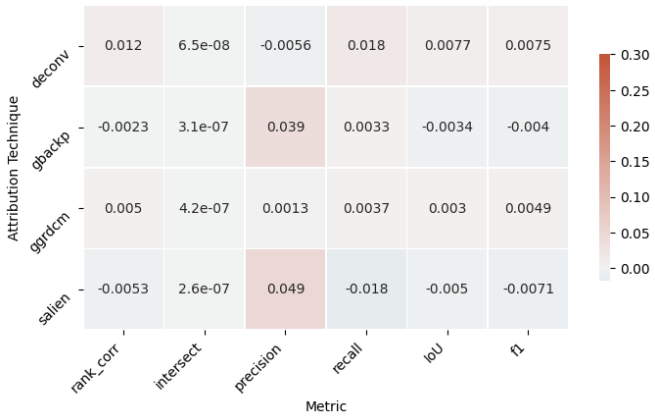


Figure 6: Mean similarity differences between high and low confidence incorrectly predicted examples (high - low).

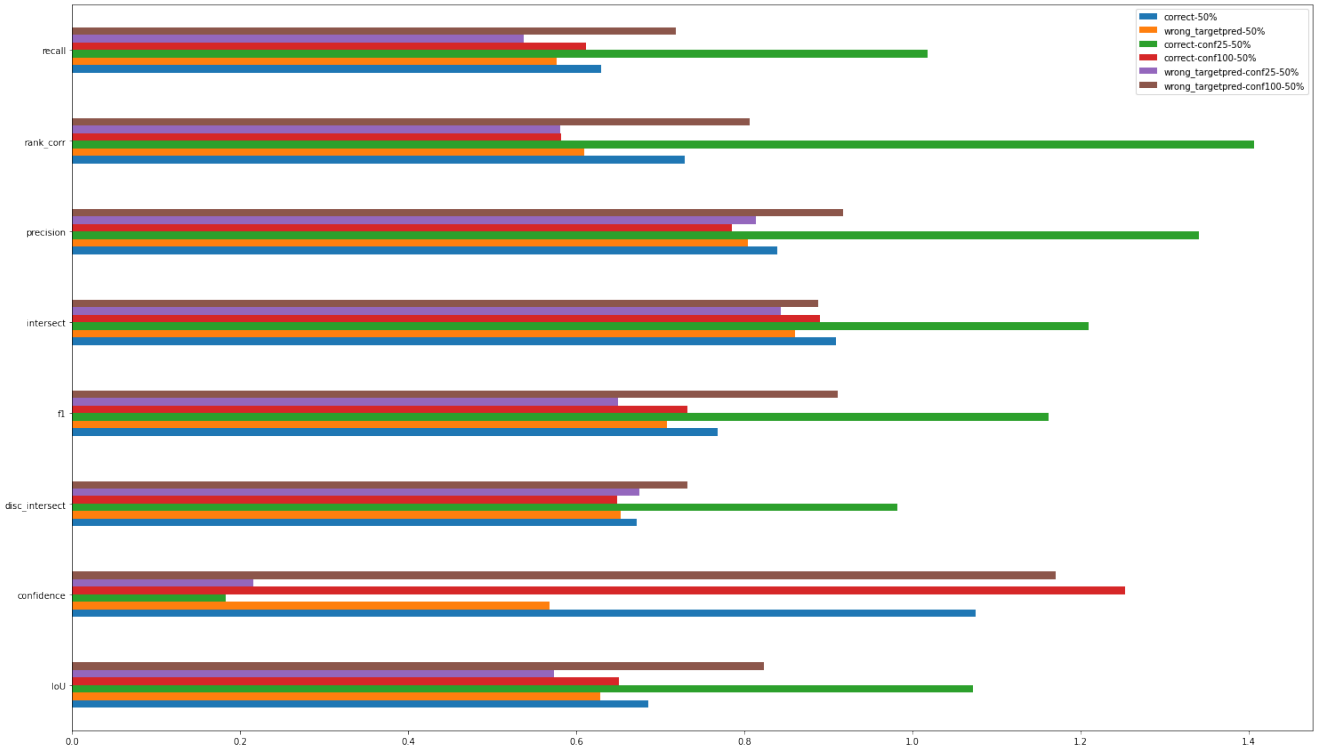


Figure 7: Median values of an ensemble of similarity metrics between attribution heatmaps and human attention maps. All values have been divided by a fixed scaling set, in order to make the diagram more comprehensible. Confidence is not a similarity metric, just the confidence of the model in its answer.

is more confident it is more similar to human attention, and (2) when a model is correct it is more similar to human attention.

Qualitative analysis

We can sanity check our results with a few qualitative comparisons of the different attribution techniques. These following two images seem to support the results from the quantitative analysis. Namely, that the model being more confident and correct is not positively correlated with the similarity of attribution heatmaps and human attention.

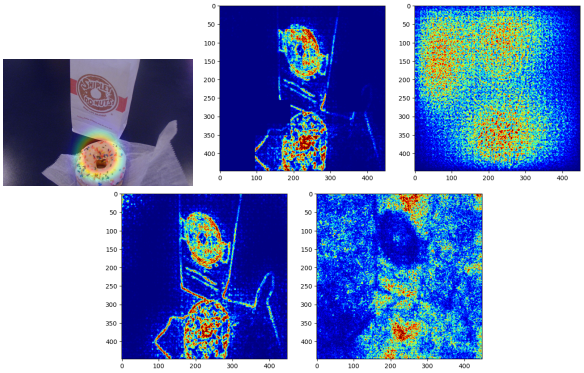


Figure 9: Example of a high-confidence, but wrong prediction. Question: "Would Homer Simpson like this?". Correct answer: "yes". Model prediction: "no". From left-to-right, top-to-bottom: human-annotated attention, Deconvnet, GuidedBackprop, GuidedGradCam, Saliency.

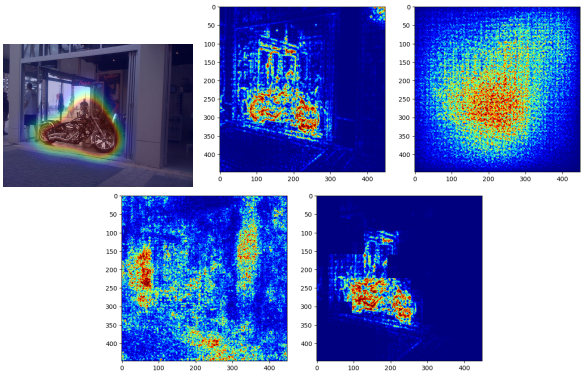


Figure 10: Example of a low-confidence, but correct prediction. Question: "How many spokes are on the front wheel of the motorcycle". Correct answer and model prediction: "6". From left-to-right, top-to-bottom: human-annotated attention, Deconvnet, GuidedBackprop, GuidedGradCam, Saliency.

In the first example 9, the model seems to be focusing on both the donut and the logo of the bag. But it also seems to have picked up on some noise: the GuidedBackProp has a high attribution to the seemingly empty space at the left of the image. In this sense, the attribution heatmaps are quite different from the human attention. We hypothesize the noise might have led to the wrong prediction, or perhaps the challenging nature of the problem (this requires a lot of context, for example, knowing who Homer Simpson is).

In the second example 10, the model seems to be focusing strongly on the motorbike: three out of the four are focused on the bike, while GuidedGradCam seems to be focused on the man next to the bike. In this sense, the heatmaps are quite similar to the human attention. The model is still able to get the correct answer but with a low confidence, we hypothesize this is because it's focusing on the bike in general, but not on the specific wheels.

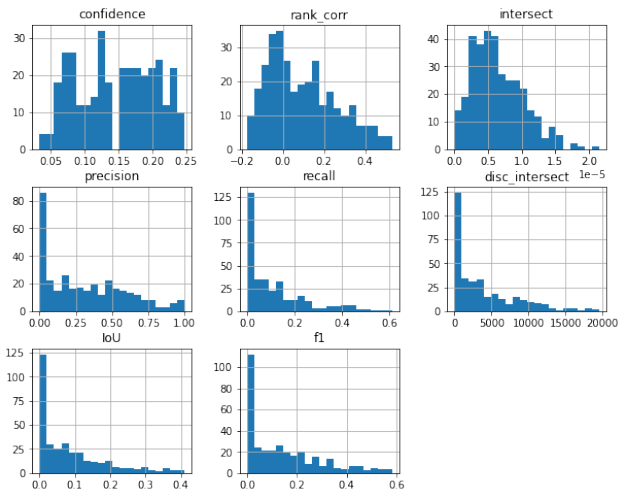
5 Conclusions

To summarize, we used investegated a CNN+LSTM VQA model using an ensemble of attribution techniques. After generating the heatmaps from the attribution techniques, we compared them to the human attention, generating a "score" for how similar two heatmaps are. Our intuition is that when the model is correct and confident, the attribution heatmaps should be more similar to the human-attention. However, through qualitiatve and quantitative analysis, we showed that this is not the case, and find that lower confidence predictions may actually represent model behavior that is more human-like. This supports the following conclusion: either attribution heatmaps do not do a very good job explaining the model (supported by [10]) or that the CNN+LSTM model looks at fundamentally different features than what humans pay attention to arrive at the answer.

6 Future Work

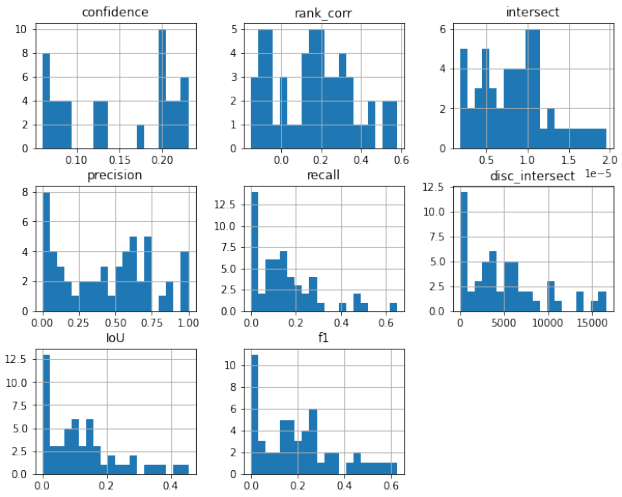
In this paper, we only explored one model (the baseline VQA model) and one family of attribution algorithms (gradient-based). It would be interesting to see if these results hold on more advanced models, especially state-of-the-art VQA models. Additionally, implementing occlusion and perturbation based algorithms would provide another dimension to look at the attribution problem, as [10] showed that the gradient-family of algorithms, are all to some extent quite similar.

We also think that attribution for the natural language component of the VQA problem is worth exploring. Future work can try to apply similar techniques to the language component, and explore the intersection of language and vision more closely (e.g. if certain words are occluded, how



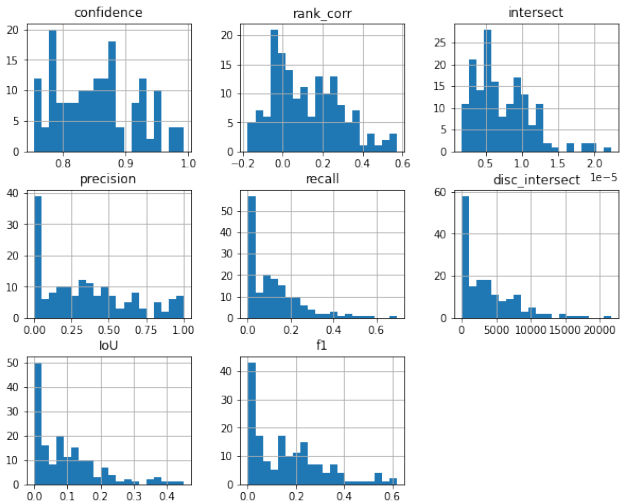
	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	f1
count	342.000000	324.000000	342.000000	342.000000	342.000000	342.000000	342.000000	342.000000
mean	0.146181	0.098631	0.000006	0.316760	0.110004	3883.818713	0.087019	0.146529
std	0.057114	0.167127	0.000004	0.277786	0.128884	4416.112486	0.098192	0.151432
min	0.033164	-0.173254	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.094901	-0.036712	0.000004	0.047794	0.007516	255.250000	0.006213	0.012350
50%	0.156569	0.064466	0.000006	0.271971	0.065170	2522.000000	0.053899	0.102283
75%	0.198624	0.213676	0.000009	0.531604	0.159498	6019.750000	0.127367	0.225954
max	0.247654	0.531029	0.000022	1.000000	0.611137	19597.000000	0.410999	0.582564

((a)) Cases when the model prediction is wrong and low confidence ($p < 0.25$).



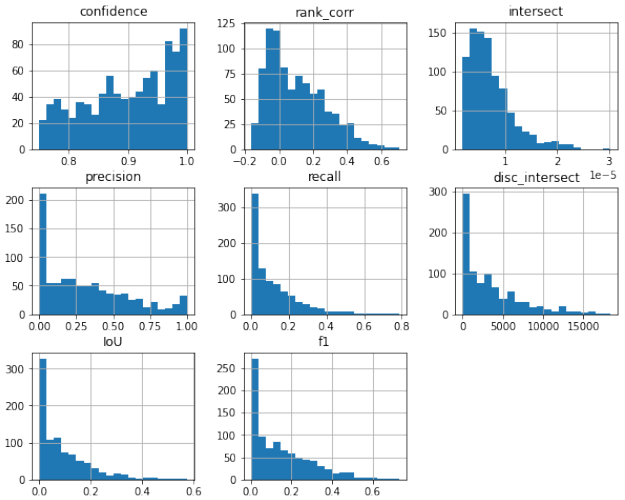
	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	f1
count	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000
mean	0.146719	0.161464	0.000009	0.437269	0.152101	4988.925926	0.124892	0.205682
std	0.064365	0.196423	0.000004	0.309002	0.144269	4408.272600	0.113488	0.166349
min	0.059754	-0.149354	0.000002	0.000248	0.000035	1.000000	0.000031	0.000061
25%	0.085249	-0.025651	0.000005	0.134773	0.031170	1508.250000	0.027949	0.054281
50%	0.132409	0.156078	0.000008	0.448393	0.123501	3669.000000	0.100793	0.183120
75%	0.207657	0.300658	0.000011	0.670223	0.204513	6846.000000	0.162378	0.279386
max	0.230523	0.579838	0.000020	0.998149	0.650275	16562.000000	0.454805	0.625244

((b)) Cases when the model prediction is correct and low confidence ($p < 0.25$).



	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	f1
count	166.000000	162.000000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000
mean	0.854052	0.121360	0.000007	0.349148	0.123851	3861.283133	0.096740	0.163001
std	0.063417	0.169937	0.000004	0.294483	0.132230	4087.795655	0.098954	0.150479
min	0.754203	-0.174650	0.000002	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.798630	-0.017427	0.000005	0.061719	0.018700	634.500000	0.013678	0.026985
50%	0.849869	0.089523	0.000006	0.306876	0.087240	2736.000000	0.077322	0.143545
75%	0.893553	0.243861	0.000010	0.539270	0.178920	6216.250000	0.141041	0.247214
max	0.992289	0.573719	0.000022	0.999445	0.698268	21643.000000	0.450137	0.620820

((c)) Cases when the model prediction is wrong and high-confidence ($p > 0.75$).



	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	f1
count	902.000000	890.000000	902.000000	902.000000	902.000000	902.000000	902.000000	902.000000
mean	0.898825	0.101289	0.000007	0.318906	0.117022	3523.687361	0.089841	0.151658
std	0.072435	0.169371	0.000004	0.281913	0.133176	3781.035151	0.097723	0.148504
min	0.750171	-0.165508	0.000002	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.843844	-0.036850	0.000004	0.059788	0.015356	442.000000	0.011594	0.022922
50%	0.909627	0.064597	0.000006	0.262516	0.074232	2424.000000	0.061202	0.115344
75%	0.965308	0.218981	0.000009	0.503964	0.174123	5425.750000	0.137285	0.241426
max	0.999796	0.703681	0.000030	1.000000	0.784510	18341.000000	0.574799	0.729996

((d)) Cases when the model prediction is correct and high confidence ($p > 0.75$).

Figure 8: Summary statistics for an array of similarity metrics between heatmaps and human attention. Split into low-confidence versus high-confidence and correct versus wrong prediction scenarios. In all, the target for attribution is the model prediction.

does that change the model's visual attribution).

Finally, our results and prior work [10] have shown the difficulty in assessing whether attribution algorithms themselves are good at actually explaining deep learning models. Further research into designing attribution algorithms or evaluating attribution algorithms with ground truth are areas of exploration.

7 Acknowledgements

We would like to thank Professor Russakovsky and other mentors for their kind feedback and support.

References

[1] Julius Adebayo et al. *Sanity Checks for Saliency Maps*. 2020. arXiv: 1810.03292 [cs.CV].

[2] Stanislaw Antol et al. "VQA: Visual Question Answering". In: *International Conference on Computer Vision (ICCV)*. 2015.

[3] Abhishek Das et al. "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?" In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2016.

[4] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. "Understanding deep networks via extremal perturbations and smooth masks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2950–2958.

[5] Vahid Kazemi and Ali Elqursh. "Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering". In: *CoRR* abs/1704.03162 (2017). arXiv: 1704.03162. URL: http://arxiv.org/abs/1704.03162.

[6] Pieter-Jan Kindermans et al. *The (Un)reliability of saliency methods*. 2017. arXiv: 1711.00867 [stat.ML].

[7] Narine Kokhlikyan et al. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. arXiv: 2009.07896 [cs.LG].

[8] Jiasen Lu et al. *Hierarchical Question-Image Co-Attention for Visual Question Answering*. 2017. arXiv: 1606.00061 [cs.CV].

[9] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st international conference on neural information processing systems*. 2017, pp. 4768–4777.

[10] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2019.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[12] Zichao Yang et al. *Stacked Attention Networks for Image Question Answering*. 2016. arXiv: 1511.02274 [cs.LG].