Interpretability in Visual Question-Answering with Human Attention Maps

Jason Yuan[1] **Michael Tang**[1]

[1]Princeton University

Introduction

VQA is an exciting new frontier of ML, blending Natural Language Processing and Computer Vision. Our first motivation is to study how VQA models interpret images. Given that many questions can be asked about the same image, understanding where the model looks to find its answer is particularly interesting. We aim to do so using a ensemble of interpretability techniques. However, it's also challenging to evaluate how accurate interpretability techniques are themselves [8]. This poses our second motivation: understanding interpretability techniques and where they fail. We aim to do this using human-attention data as a ground-truth, establishing a baseline for which we can compare.

Related Work

VQA Models

- *Kazemi and Elqursh*. Baseline model, achieving 59.7 accuracy on the VQA2.0 challenge [4]. The architecture consists of a ResNet to encode the image, a LSTM to encode the question, and then a few more layers ending with a Softmax classifier (over the 3000 most frequent answers in the VQA dataset).

Interpretability Methods - Pixel Attribution

- *Occlusion/perturbation techniques.* Model-agnostic methods that work by occluding parts of the input image and measuring the change in output [8]. Examples include: SHAP [7], LIME [9], Extremal Perturbations [3].
- *Gradient techniques.* Methods that work by computing the gradient of some output with respect to the input pixels [8]. These methods are very fast compared to occlusion methods, but hard to know if it's correctly explaining the model [8]. [5] showed that with minor perturbations, gradient methods give drastically different results.
  - *Saliency* computes the gradient of the target class score with respect to the pixels [8].
  - *Guided Backprop* and *DeconvNet* are nearly identical to Saliency, but each uses a slightly different method to reverse a ReLU layer, the intention being to avoid the activation saturation problem [8] [6].
  - *Guided GradCam* backpropagates the target class score to the last convolution layer, and combines it with Guided Backprop [8].

Method

Data

- The dataset is a subset of the MS COCO dataset, as part of the VQA1.0 dataset [1]. It is structured as tuples of an Image, Question, and Answer. We specifically focus on the Multiple Choice validation set on Real Images, as these are the most similar to the image classification problems interpretability has been studied in previously. The dataset has 40,504 images and 121,512 questions, with 3 questions per image.
- *VQA-Hat dataset* Introduced ground truth human attention for comparison with attention-based VQA models; here we leverage it as a ground truth mask to evaluate interpretability techniques. The dataset has 4,122 human-annotated attention maps corresponding to questions in the VQA1.0 validation dataset. [2]

Generating Heatmaps

- We use the LSTM+CNN baseline model. Our primary motivation for choosing this model over a more advanced one, such as Pythia, was because it made implementing the attribution techniques more tractable.
- We implement DeconvNet, Saliency, GuidedBackProp, GuidedGradCam on the LSTM+CNN model, using the library [6]. Our choice to use the gradient-family of techniques was due to limitations on computational power.
- 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. Due to computational constraints, we limited our analysis to the first 70 examples in the dataset.

To evaluate different attribution techniques, we applied different metrics that quantified the difference between the attribution heatmap and the human-annotated attention map. All metrics were computed after resizing and interpolating the maps to be the same size and normalizing the mean, standard deviation, range, and total weight across the map.

(see results for details of discretization) **Attribution Technique Evaluation**

- Spearman rank correlation (as done in the VQA-Hat paper)
- Intersection, i.e. elementwise multiplication
- Discrete intersection, i.e. elementwise multiplication after thresholding
- Intersection over union (IoU) after discretization
- Dice coefficient, 2 times the intersection over (intersection + union) after discretization
- Precision, recall, f-1 score where true positive is the intersection, false positive is area where the attribution map includes but human map excludes, etc. After discretization.

Results

We examine the distribution of weights in the human-annotated attribution maps and see that it is bimodal, with the majority of the image given a low attention value and a small portion given a very high attention value.
We found that an additional normalization via thresholding was helpful to make the attribution techniques more comparable to the smooth, bimodal, concentrated human attention maps. Since the human maps had a high peak that contained about 10% of the pixels, we discretized the attribution heatmaps by setting the top-weighted 10% of the pixels to have a value of 1 and the rest to have a value of 0, transforming the continuous maps into regions that we could analyze using techniques from image segmentation and classification evaluation, e.g. IoU, dice coefficient, f1 score.

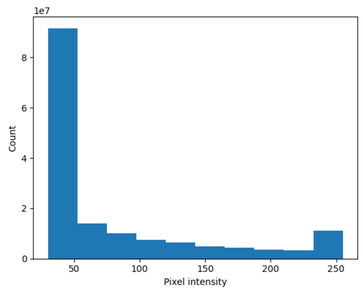


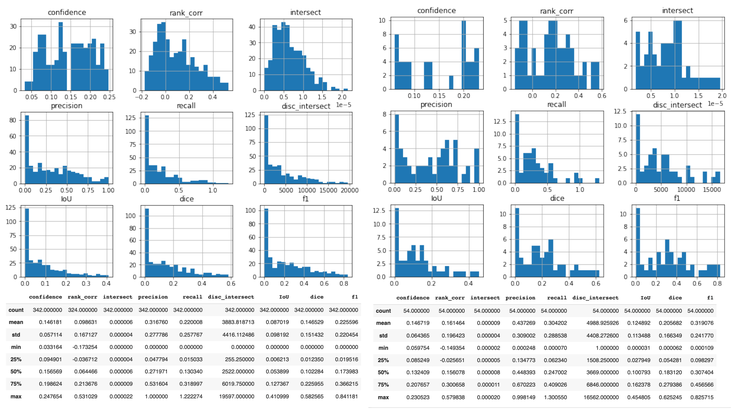
Fig. 1: Mean pixel weight distribution for human maps

We can also analyze the effectiveness of different techniques under these metrics: gradient saliency generally slightly outperforms the other attribution techniques, with the second-highest rank correlation and the highest f1-score and discrete intersection scores.

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

Fig. 2: Mean values for attribution technique metrics

We split the data into 4 components: when the model is right versus wrong, and when the model's confidence in the answer is <0.25 and >0.75, to analyze interpretable behavior in these different cases.

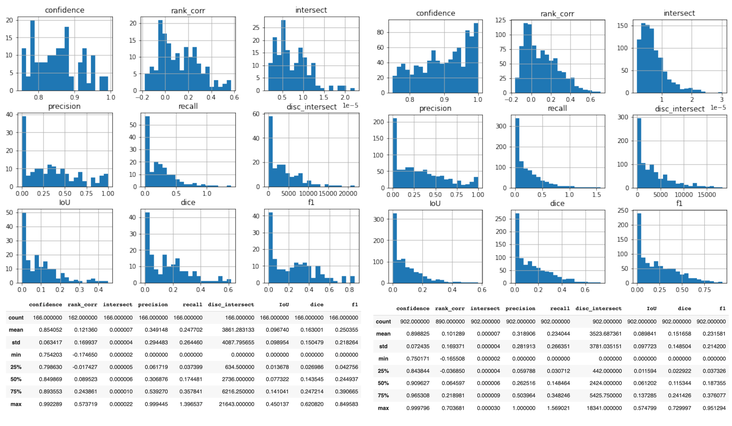


	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	dice	f1
count	342.000000	324.000000	342.000000	342.000000	342.000000	342.000000	342.000000	342.000000	342.000000
mean	0.146181	0.098631	0.000006	0.316760	0.220008	3883.818713	0.087019	0.146529	0.225596
std	0.057114	0.167127	0.000004	0.277786	0.257767	4416.112486	0.098192	0.151432	0.220454
min	0.033164	-0.173254	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.094901	-0.036712	0.000004	0.047794	0.015033	255.250000	0.006213	0.012350	0.019516
50%	0.156569	0.064466	0.000006	0.271971	0.130340	2522.000000	0.053899	0.102284	0.173983
75%	0.198624	0.213676	0.000009	0.531604	0.318997	6019.750000	0.127367	0.225955	0.366215
max	0.247654	0.531029	0.000022	1.000000	1.222274	19597.000000	0.410999	0.582565	0.841181

	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	dice	f1
count	54.000000	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	0.304202	4988.935926	0.124892	0.205682	0.319076
std	0.064365	0.196423	0.000004	0.309002	0.286538	4408.272500	0.113488	0.166349	0.241770
min	0.059754	-0.149354	0.000002	0.000248	0.000070	1.000000	0.000031	0.000062	0.000109
25%	0.085249	-0.025651	0.000005	0.134773	0.062340	1508.250000	0.027949	0.054281	0.098297
50%	0.132409	0.156078	0.000008	0.448393	0.247002	3669.000000	0.100793	0.183120	0.307454
75%	0.207657	0.300658	0.000011	0.670223	0.469026	6846.000000	0.162378	0.279386	0.456566
max	0.230523	0.579838	0.000020	0.998149	1.300550	16562.000000	0.454805	0.625245	0.825715

Fig. 3: Left: summary statistics for wrong predictions (and model confidence < 0.25). Right: summary statistics for correct predictions (and model confidence < 0.25). In both, the target for attribution is the prediction.

Looking at Fig 1, observe that the correct predictions score higher in general on the human-attention-metrics. This is supported by the fact that the distribution for when the model is correct exhibits a fatter right tail and the mean is higher on all of the scoring metrics.



	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	dice	f1
count	166.000000	162.000000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000
mean	0.854052	0.121360	0.000007	0.349148	0.247702	3861.283133	0.096740	0.163001	0.250355
std	0.063417	0.169937	0.000004	0.294483	0.264460	4087.795655	0.098954	0.150479	0.218264
min	0.754203	-0.174650	0.000002	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.798630	-0.017427	0.000005	0.061719	0.037399	634.500000	0.013678	0.026986	0.042756
50%	0.849869	0.089523	0.000006	0.306876	0.174481	2736.000000	0.077322	0.143545	0.244937
75%	0.893553	0.243861	0.000010	0.539270	0.357841	6216.250000	0.141041	0.247214	0.390665
max	0.992289	0.573719	0.000022	0.999445	1.396537	21643.000000	0.450137	0.620820	0.849583

	confidence	rank_corr	intersect	precision	recall	disc_intersect	IoU	dice	f1
count	902.000000	890.000000	902.000000	902.000000	902.000000	902.000000	902.000000	902.000000	902.000000
mean	0.869825	0.101269	0.000007	0.318906	0.234044	3525.667361	0.088841	0.151658	0.231581
std	0.072435	0.169371	0.000004	0.281913	0.266351	3781.035151	0.097723	0.148504	0.214200
min	0.750171	-0.165508	0.000002	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.843844	-0.036850	0.000004	0.059788	0.030712	442.000000	0.011594	0.022922	0.037326
50%	0.909627	0.064597	0.000006	0.262516	0.148464	2424.000000	0.061202	0.115344	0.187355
75%	0.965308	0.218981	0.000009	0.503964	0.348246	5425.750000	0.137285	0.241426	0.376077
max	0.999796	0.703681	0.000030	1.000000	1.569021	18341.000000	0.574799	0.729997	0.951294

Fig. 4: Left: summary statistics for wrong predictions (and model confidence > 0.75). Right: summary statistics for correct predictions (and model confidence > 0.75). In both, the target for attribution is the prediction.

Next, we examine the model's heatmaps for when the confidence is > 0.75. Here, we note that the opposite trend seems to occur: the human-attribution metrics are higher for the wrong-but-high-confidence-predictions. This is evident in the fatter right tail for the wrong-predictions' metrics and holds in all but one of the means. The pearson correlations between confidence and these metrics further support these conclusions, with small but consistent correlations across all metrics that align with these observations with value 0.1, 0.3, respectively.

Results

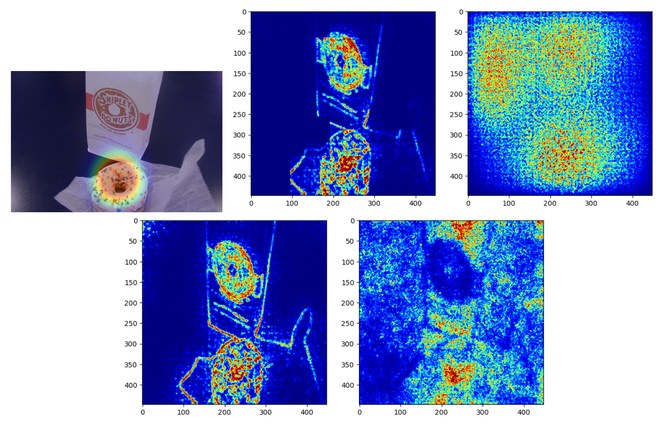


Fig. 5: Example of a high-confidence, but wrong prediction. Question: "Would Homer Simpson like this?". Correct answer: "yes". Model prediction: "no". From left-to-right, top-to-bottom: human-annotated attention, Deconvnet, GuidedBackprop, GuidedGradCam, Saliency.

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).
These 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.

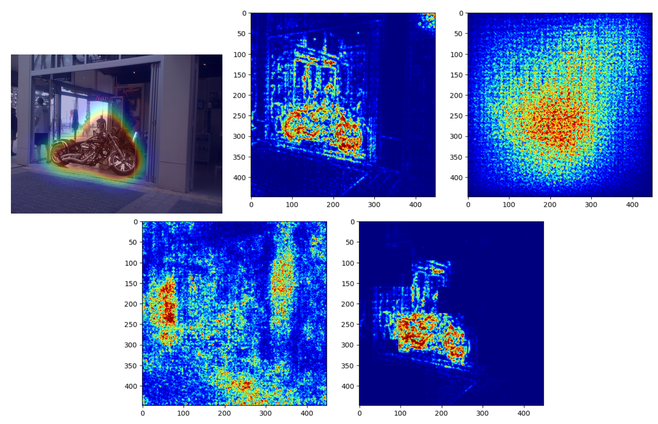


Fig. 6: Example of a low-confidence, but correct prediction. Question: "How many spokes are on the front wheel of the motorcycle". Correct answer and model prediction: "6". From left-to-right, top-to-bottom: human-annotated attention, Deconvnet, GuidedBackprop, GuidedGradCam, Saliency.

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.

Conclusion

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. This supports the following conclusion: either attribution heatmaps do not do a very good job explaining the model (supported by [8]) or that the CNN+LSTM model looks at fundmanetally different features than what humans pay attention to arrive at the answer.

Acknowledgements and References

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

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

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

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

[9] 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.