

---

# Towards Transparent AI Systems: Interpreting Visual Question Answering Models

---

Yash Goyal  
Akrit Mohapatra  
Devi Parikh  
Dhruv Batra

Virginia Tech, Blacksburg, VA 24061 USA

YGOYAL@VT.EDU  
AKRIT@VT.EDU  
PARIKH@VT.EDU  
DBATRA@VT.EDU

## Abstract

Deep neural networks have shown striking progress and obtained state-of-the-art results in many AI research fields in the recent years. It is often unsatisfying to not know why they predict what they do. In this paper, we address the problem of interpreting Visual Question Answering (VQA) models. Specifically, we are interested in finding what part of the input (pixels in images or words in questions) the VQA model focuses on while answering the question. To tackle this problem, we use two visualization techniques – guided backpropagation and occlusion – to find important words in the question and important regions in the image. We then present qualitative and quantitative analyses of these importance maps.

## 1. Introduction

We are witnessing an excitement in the research community and frenzy in media regarding advances in AI. Fueled by a combination of massive datasets and advances in deep neural networks (DNNs), the community has made remarkable progress on a variety of AI tasks such as image classification (Szegedy et al., 2015), machine translation (Brea et al., 2011; Sutskever et al., 2014), speech recognition (Hinton et al., 2012), and even harder tasks such as learning to play Go (Mnih et al., 2013), answering reading comprehension questions by understanding short stories (Bordes et al., 2015; Weston et al., 2015), and answering questions about images (Antol et al., 2015; Ren et al., 2015; Malinowski et al., 2015).

Unfortunately, when today’s machine perception and intelligent systems fail, they fail in a spectacularly disgraceful manner, without warning or explanation, leaving the user staring at an incoherent output, wondering why the system did what it did.

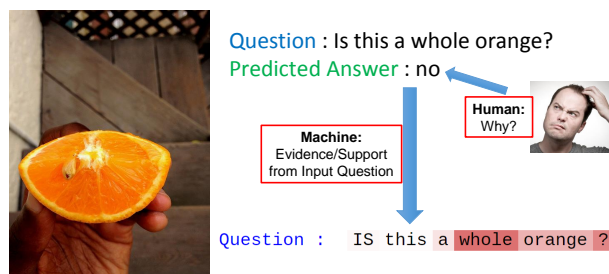


Figure 1. The goal of this work is to interpret Visual Question Answering models. We are interested in answering the question—*why does a VQA model predict what it does?* Our approach here is to find the evidence in the test input on which the model focuses while answering the question. In this example, “whole” is the most important word in the question for the model while predicting the answer “no”.

In this work, we focus on Visual Question Answering, where the task is – given an image and a free-form natural language question about the image, (e.g., “What color are the girl’s shoes?”, or “Is the boy jumping?”), the machine has to produce a natural language answer as its output (e.g. “blue”, or “yes”). Specifically, we try to interpret a recent state-of-art VQA model (Lu et al., 2015) trained on recently released VQA (Antol et al., 2015) dataset. This VQA model uses Convolutional Neural Network (CNN) features for images and Long-Short Term Memory (LSTM) embeddings for questions, combines these embeddings, and uses a multi-layer perceptron as a classifier to predict a probability distribution over answers.

Sustained interactions with the system make it clear that the system has a non-trivial level of intelligence (e.g., it is able to recognize people, objects, *etc.* in the image). However, it is deeply unsatisfying to not know why the system predicts what it does (especially the glaring mistakes). The root cause is *lack of transparency or interpretability*.

We are interested in the question of transparency – “why does a VQA system do what it does?” (See Fig. 1). Specif-

ically, *what evidence in the test input (image and question) supports a particular prediction?* In the context of VQA, this question can be expressed as two subproblems:

- *What words in the question does the model “focus on” in order to answer the question?*
- *What pixels in the image does the model “look at” while answering the question?*

In this work, we use two visualization methods to tackle the above problems. The first method (Sec. 3.1) uses guided backpropagation (Springenberg et al., 2014) to analyze important words in the question and important regions in the image. In the second method (Sec. 3.2), we occlude portions of input and observe the change in prediction probabilities of the model, to compute importance of question words and image regions. In Sec. 4, we present qualitative and quantitative analyses of these image/question ‘importance maps’ – question importance maps are analyzed using their Part-of-Speech (POS) tags; image importance maps are compared to ‘human attention maps’ or maps showing where humans look for answering a question about the image (Das et al., 2016).

## 2. Related Work

Many gradient based methods (Zeiler & Fergus, 2014; Simonyan et al., 2013; Springenberg et al., 2014) have been proposed in recent years in the field of computer vision to visualize deep convolutional neural networks. But most of them focused on the task of image classification on iconic images where the main object occupies most of the image. Our work differs in 2 ways – 1) we also compute gradients w.r.t. the input question, and 2) we use guided backpropagation (Springenberg et al., 2014) for the task of VQA, where the model can look at different regions in the same image for different questions.

Our occlusion experiment is inspired by (Zeiler & Fergus, 2014) who mask small regions in the image with a gray patch, and observe the output of an image classification model. We evaluate if the model looks at the same regions in the image as humans do, while answering a question about the image.

A few recent works (Ribeiro et al., 2016; Baehrens et al., 2010; Liu & Wang, 2012) have begun to study the task of providing interpretable post-hoc explanations for classifier predictions. Such methods typically involve fitting or training a secondary interpretable mechanism on top of the base ‘black-box’ classifier predictions. In contrast, our work directly computes importance maps from the model of interest without another layer of training (which could obfuscate the analysis).

## 3. Approach

At a high level, we view a VQA model as a learned function  $a = f_w(i, q)$  that takes in an input image  $i$  and a question about the image  $q$ , is parameterized by parameters  $w$ , and produces an answer  $a$ . In order to gauge the importance of components of  $i$  and  $q$  (i.e. pixels and words), we consider the *best linear approximation* to  $f$  around each test point  $(i_{test}, q_{test})$ :

$$f(i, q) \underset{\text{best linear fit}}{\approx} f(i_{test}, q_{test}) + [i - i_{test}, q - q_{test}]^T \nabla f(i_{test}, q_{test}) \quad (1)$$

Intuitively, the two key quantities we need to compute are  $\partial f(i_{test}, q_{test})/\partial i$  and  $\partial f(i_{test}, q_{test})/\partial q$ , i.e. the partial derivatives of the function w.r.t. each of the inputs (image and question). These expressions superficially look similar to gradients computed in backpropagation-based training of neural networks. However, there are two key differences – (i) we compute partial derivatives of the probability of *predicted output*, not the ground-truth output; and (ii) we compute partial derivatives w.r.t. *inputs* (i.e. image pixel intensities and word embeddings), not parameters.

Due to linearization above, elements of these partial derivatives tell us the effect of those pixels/words on the final prediction. These may be computed in the following two ways.

### 3.1. Guided Backpropagation

Guided backpropagation is a gradient-based visualization technique used to visualize activations of neurons in different layers in CNNs. It has been shown to perform better than its counterparts such as deconvolution (Zeiler & Fergus, 2014) especially for visualizing higher order layers. It is a modified version of backpropagation that restricts negative gradients from flowing backwards towards input layer, resulting in sharper image visualizations.

Guided BP is identical to classical BP except in the way the backward pass is computed in Rectified Linear Units (ReLU). Let  $h^l$  denote the input to layer  $l$  and  $h^{l+1}$  denote the output. Recall that a ReLU is defined as  $h^{l+1} = \text{relu}(h^l) = \max(h^l, 0)$ . Let  $G^{l+1} = \partial f/\partial h^{l+1}$  denote the partial derivative w.r.t. the output of the ReLU (received as input in the backward pass). The key difference between the two backprops (BP) is:

$$G^l = \llbracket h^l > 0 \rrbracket \cdot G^{l+1} \quad \text{[Classical BP]} \quad (2)$$

$$G^l = \llbracket h^l > 0 \rrbracket \cdot \llbracket G^{l+1} > 0 \rrbracket \cdot G^{l+1} \quad \text{[Guided BP]} \quad (3)$$

i.e. guided BP blocks negative gradients from flowing back in ReLUs. For more details, please refer to (Springenberg et al., 2014).



Figure 2. Results for Discrete Derivatives experiment. (a) shows heat maps for questions showing the importance of words in the questions. Encouragingly, “vegetable” is the most important word in the first question for the predicted answer “broccoli”. (b) shows the importance of different regions in images. In the top image, the region containing the parrot affects the model’s prediction the most. Both visualization provide justifications for the predictions, resulting in increased transparency in the inner working of the VQA model. Best viewed in color.

We use guided BP to compute ‘gradients’ of the probability of predicted answer w.r.t. inputs (image and question). Note that the language pathway in the models we use typically, does not contain ReLUs, thus these are true gradients (not just gradient-based visualizations) on the language side. We interpret the words/pixels with the highest (magnitude) gradients received as the most important for the model since small changes in these lead to largest changes in the model’s confidence in the predicted answer.

### 3.2. Discrete Derivatives

In this method, we systematically occlude subsets of the input, forward propagate the masked input through the VQA model, and compute the change in the probability of the answer predicted with the unmasked original input. Since there are 2 inputs to the model, we focus on one input at a time, keeping the other input fixed (mimicing partial derivatives). Specifically, to compute importance of a question word, we mask that word by dropping it from the question, and feed the masked question with original image as inputs to the model. The importance score of the question word is computed as the change in probability of the original predicted answer.

We follow the same procedure on the images to compute importance of image regions. We divide the image into a grid of size 16 x 16, occlude one cell at a time with a gray patch<sup>1</sup>, feed in the perturbed image with the entire question to the model, and compute the decrease in the probability of the original predicted answer. The generated importance

<sup>1</sup>a gray patch of intensities (R, G, B) = (123.68, 116.779, 103.939), mean RGB pixel values across a large image dataset ImageNet (Deng et al., 2009) on which the CNN is trained.

maps are shown in Fig. 2.

## 4. Results

While image/question importance maps on individual inputs provide crucial insight into the inner-workings of a model (*e.g.*, see Fig. 2), what do the aggregate statistics of these maps tell us about the model?

### 4.1. Analyzing Image Importance

In concurrent work (Das et al., 2016), we collected human attention annotations for (question, image) pairs from VQA dataset (Antol et al., 2015). Given a blurry image and a question, humans were asked to deblur the regions in the image that were helpful in answering the question.

	Rank-correlation
Random	0.00007
Occlusion	0.07164
Guided backpropagation	0.12277
Human	0.62289

Table 1. Rank-correlation of importance maps with human attention maps (higher is better). The last row represents inter-human agreement.

We evaluate the quality of image importance maps obtained from the two methods (guided backpropagation and occlusion) by comparing them to the human attention maps. The human attention dataset contains annotations for 1374 (question, image) pairs from VQA (Antol et al., 2015) validation set. We take the absolute value of the importance maps and compute their mean rank-correlation with the

human attention maps. We first scale both the image importance and human attention maps to 16x16, normalize them spatially and rank the pixels according to their spatial attention, and then compute correlation between these two ranked lists. The results are shown in Table 1. We find that both importance maps (occlusion and guided BP) are (somewhat) positively correlated with human attention maps, although it is far from inter-human correlation. Perhaps surprisingly, they are *better correlated than an explicit attention model* SAN (Yang et al., 2015) which achieves a rank-correlation of -0.08836. Thus, our techniques revealed an interesting finding – that even without attention mechanisms, VQA models may be implicitly attending to relevant regions in the image.

## 4.2. Analyzing Question Importance

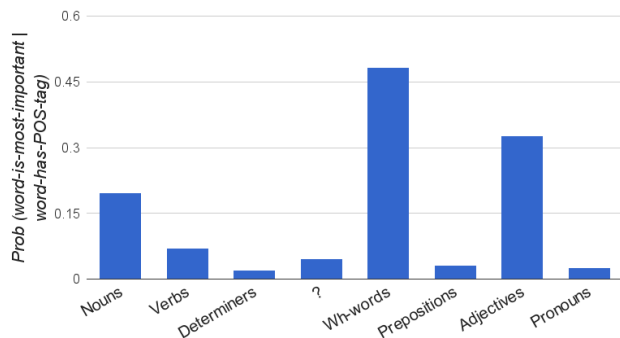


Figure 3. Probability of a word being most important in a question given that it has a certain POS tag. POS tags are sorted in the decreasing order of their frequency in the entire dataset.

Since there is no human attention dataset for questions, we instead analyze the importance maps for questions using their POS tags. Our hypothesis is that wh-words and nouns should matter most to a ‘sensible’ model’s prediction. We plot the probability of a word being most important in a question given that it has a certain POS tag. To get reliable statistics, we picked 15 most frequent POS tags from the VQA validation dataset, and grouped similar tags into one category, e.g. WDT, WP, WRB are grouped as wh-words. The histogram can be seen in Fig. 3. Indeed, wh-words are most important followed by adjectives and nouns. Adjectives and nouns rank high because many questions tend to ask about characteristics of objects, or objects themselves. This finding suggests that the language model part of the VQA model is strong and is able to learn to focus on appropriate words without any explicit attention procedure.

Note that for many occlusions, the model’s predicted answer is different from the original predicted answer. In fact, we found that the number of times the predicted answer changes correlates with the model’s accuracy. It is able to predict success/failure accurately 72% of the times. This suggests that features that characterize these importance

maps can provide useful signals for predicting oncoming failures.

## 5. Conclusion

In this paper, we work with two visualization methods – guided backpropagation and occlusion – to interpret deep learning models for the task of Visual Question Answering. Although we focus on only one VQA model in this work, the methods are generalizable to all other end-to-end VQA models. The occlusion method can even be applied to any (non-end-to-end) VQA model considering it as a black box. We believe that these methods and results can be helpful in interpreting the current VQA models, and designing the next generation of VQA models.

**Acknowledgements.** This work was supported in part by the following: National Science Foundation CAREER awards to DB and DP, Army Research Office YIP awards to DB and DP, ICTAS Junior Faculty awards to DB and DP, Army Research Lab grant W911NF-15-2-0080 to DP and DB, Office of Naval Research grant N00014-14-1-0679 to DB, Paul G. Allen Family Foundation Allen Distinguished Investigator award to DP, Google Faculty Research award to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donation to DB.

## References

- Antol, Stanislaw, Agrawal, Aishwarya, Lu, Jiasen, Mitchell, Margaret, Batra, Dhruv, Zitnick, C. Lawrence, and Parikh, Devi. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 3
- Baehrens, David, Schroeter, Timon, Harmeling, Stefan, Kawanabe, Motoaki, Hansen, Katja, and Müller, Klaus-Robert. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, August 2010. ISSN 1532-4435. 2
- Bordes, Antoine, Usunier, Nicolas, Chopra, Sumit, and Weston, Jason. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015. 1
- Brea, Johanni, Senn, Walter, and Pfister, Jean-Pascal. Sequence learning with hidden units in spiking neural networks. In *Advances in neural information processing systems*, pp. 1422–1430, 2011. 1
- Das, Abhishek, Agrawal, Harsh, Zitnick, C. Lawrence, Parikh, Devi, and Batra, Dhruv. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *ICML Workshop on Visualization for Deep Learning*, 2016. 2, 3

- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 3
- Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29 (6):82–97, 2012. 1
- Liu, L. and Wang, L. What has my classifier learned? visualizing the classification rules of bag-of-feature model by support region detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3586–3593, June 2012. doi: 10.1109/CVPR.2012.6248103. 2
- Lu, Jiasen, Lin, Xiao, Batra, Dhruv, and Parikh, Devi. Deeper LSTM and normalized CNN Visual Question Answering model. [https://github.com/VT-vision-lab/VQA\\_LSTM\\_CNN](https://github.com/VT-vision-lab/VQA_LSTM_CNN), 2015. 1
- Malinowski, Mateusz, Rohrbach, Marcus, and Fritz, Mario. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 1
- Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Graves, Alex, Antonoglou, Ioannis, Wierstra, Daan, and Riedmiller, Martin. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 1
- Ren, Mengye, Kiros, Ryan, and Zemel, Richard. Exploring models and data for image question answering. In *NIPS*, 2015. 1
- Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. ”why should i trust you?”: Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016. 2
- Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034, 2013. 2
- Springenberg, Jost Tobias, Dosovitskiy, Alexey, Brox, Thomas, and Riedmiller, Martin A. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014. 2
- Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc. Sequence to Sequence Learning with Neural Networks. In *Neural Information Processing Systems (NIPS)*, 2014. 1
- Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott E., Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. In *CVPR*, 2015. 1
- Weston, Jason, Bordes, Antoine, Chopra, Sumit, and Mikolov, Tomas. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *CoRR*, abs/1502.05698, 2015. 1
- Yang, Zichao, He, Xiaodong, Gao, Jianfeng, Deng, Li, and Smola, Alexander J. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015. 4
- Zeiler, Matthew D. and Fergus, Rob. Visualizing and Understanding Convolutional Networks. In *ECCV*, 2014. 2