Visualizing Deep Networks by Optimizing with Integrated Gradients

Zhongang Qi, Saeed Khorram, Fuxin Li
School of Electrical Engineering and Computer Science,
Oregon State University
{qiz,khorrams,lif}@oregonstate.edu

Abstract

Understanding and interpreting the decisions made by deep learning models is valuable in many domains. In computer vision, computing heatmaps from a deep network is a popular approach for visualizing and understanding deep networks. However, heatmaps that do not correlate with the network may mislead human, hence the performance of heatmaps in providing a faithful explanation to the underlying deep network is crucial. In this paper, we propose I-GOS, which optimizes for a heatmap so that the classification scores on the masked image would maximally decrease. The main novelty of the approach is to compute descent directions based on the integrated gradients instead of the normal gradient, which avoids local optima and speeds up convergence. Extensive experiments show that the heatmaps produced by our approach are more correlated with the decision of the underlying deep network, in comparison with other state-of-the-art approaches.

1. Introduction

In recent years, there has been a lot of focus on explaining deep neural networks. In the computer vision domain, one of the most important explanation techniques is the heatmap approach [9, 5, 10], which focuses on generating heatmaps that highlight parts of the input image that are most important to the decision of the deep networks on a particular classification target.

Some heatmap approaches achieve good visual qualities for human understanding, such as several one-step backpropagation-based visualizations including Guided Backpropagation (GBP) [7] and the deconvolutional network (DeconvNet) [9]. These approaches utilize the gradient or variants of the gradient and backpropagate them back to the input image, in order to decide which pixels are more relevant to the change of the deep network prediction. However, whether they are actually correlated to the decision-making of the network is not that clear [2]. [2] proves that GBP and DeconvNet are essentially doing (partial) image recovery, and thus generate more human-interpretable visualizations that highlight object boundaries, which do not necessarily represent what the model has truly learned.

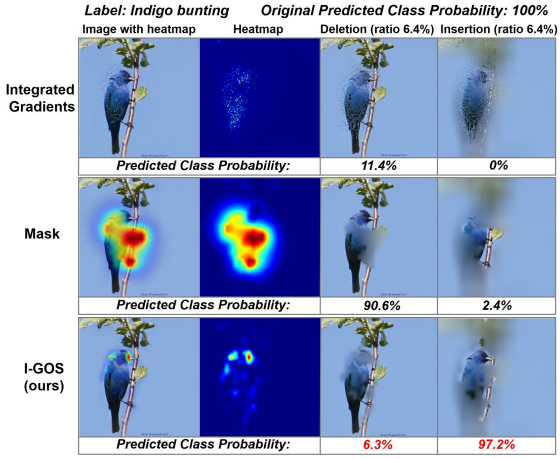

Figure 1. Heatmap visualizations can be verified by testing the CNN on deletion images (column 3), which blur areas highlighted on the heatmap, and insertion images (column 4), which blur areas not highlighted on the heatmap. The first two rows show that Integrated Gradients [8], Mask [1] may fail on these evaluations. Using heatmap generated from our I-GOS, CNN no longer classifies the deletion image to the same category (column 3), and classifies the insertion image correctly with only few pixels revealed (column 4), showing the correlation between the I-GOS heatmap and CNN decision making. For all approaches the same amount of pixels (6.4% in this figure) were blurred/revealed.

If these are the goals of a heatmap, a natural idea would be to directly optimize them. The mask approach [1] generates heatmaps by solving an optimization problem, which aims to find the smallest and smoothest area that maximally decreases the output of a neural network. It can generate very good heatmaps, but usually takes a long time to converge, and sometimes the optimization can be stuck in a bad local optimum due to the strong nonconvexity of the solution space. Another approach called integrated gradients [8] claims that any change in the output can be reflected in

This work is partially supported by the Defense Advanced Research Projects Agency (DARPA) under contract N66001-17-2-4030.

their heatmaps. The basic idea is to explicitly find the image that has the lowest prediction score – a completely grey image, or a highly blurred image usually would not be predicted to any category by a deep network, and then integrate the gradients on the entire line between the grey/blurred image to the original image to generate a heatmap. However, the heatmaps generated by integrated gradients are normally diffuse, thus difficult for human to understand (Fig. 1).

In this paper, we propose a novel visualization approach I-GOS (Integrated-Gradients Optimized Saliency) which utilizes the integrated gradients to improve the mask optimization approach in [1]. The idea is that the direction provided by the integrated gradients may lead better towards the global optimum than the normal gradient which may tend to lead to local optima. Hence, we replace the gradient in mask optimization with the integrated gradients. Due to the high cost of computing the integrated gradients, we employ a line-search based gradient-projection method to maximally utilize each computation of the integrated gradients. I-GOS generates better heatmaps (Fig. 1) and utilizes less computational time than the original mask optimization, as line search is more efficient in finding appropriate step sizes, allowing significantly less iterations to be used.

2. Model Formulation

Mask Optimization: For one-step backpropagation-based approaches [7, 9], there exits an issue that they only reflect infinitesimal changes of the prediction of a deep network. In the highly nonlinear function estimated by the deep network, such infinitesimal changes are not necessarily reflective of changes large enough to alter the decision of the deep network. What we would expect is that the heatmaps indicate the areas that would really change the classification result significantly. In [1], a perturbation based approach is proposed which introduces a mask M as the heatmap to perturb the input I_0. M is optimized by solving the following objective function:

$$\begin{aligned} &\underset{M}{\operatorname{argmin}}\ F_c(I_0, M) = f_c\big(\Phi(I_0, M)\big) + g(M), \\ &\text{where}\ \ \Phi(I_0, M) = I_0 \odot M + \tilde{I}_0 \odot (\mathbf{1} - M), \\ &\qquad g(M) = \lambda_1 ||\mathbf{1} - M||_1 + \lambda_2 \text{TV}(M), \quad \mathbf{0} \le M \le \mathbf{1} \end{aligned} \tag{1}$$

In (1), $f_c(I)$ represents the prediction output of a black-box deep network f on class c from an image I; M is a matrix which has the same shape as the input image I_0 and whose elements are all in $[0, 1]$; $\tilde{I}_0$ is a baseline image with the same shape as I_0, which should have a low score on the class c, $f_c\big(\tilde{I}_0\big) \approx \min_I f_c(I)$, and in practice either a constant image, random noise, or a highly blurred version of I_0. This optimization seeks to find a deletion mask that significantly decreases the output score, i.e., $f_c\big(I_0 \odot M + \tilde{I}_0 \odot (\mathbf{1} - M)\big) \ll f_c(I_0)$ under the regularization of $g(M)$. $g(M)$ contains two terms, with the first term on the magnitude of M, and the second term a total-variation (TV) norm to make M more piecewise-smooth.

Although this approach of optimizing a mask performs significantly better than the gradient method, there exist inevitable drawbacks when using a traditional first-order algorithm to solve it. First, it is slow, usually taking hundreds of iterations to obtain the heatmap for each image. Second, since the model f_c is highly nonlinear in most cases, optimizing (1) may only achieve a local optimum, with no guarantee that it indicates the right direction for a significant change related to the output class (Fig. 1 and Fig. 3).

Integrated Gradients: Note that the problem of finding the mask is not a conventional non-convex optimization problem. For $F_c(I_0, M) = f_c(I_0, M) + g(M)$, we (approximately) know the global minimum (or, at least a reasonably small value) of $f_c(I_0, M)$ in a baseline image $\tilde{I}_0$, which corresponds to $M = \mathbf{0}$. The integrated gradients approach [8] considers the straight-line path from the baseline $\tilde{I}_0$ to the input I_0. Instead of evaluating the gradient at the provided input I_0 only, the integrated gradients would be obtained by accumulating all the gradients along the path. [8] proved that it satisfies an axiom called completeness that the integrated gradients for all pixels add up to the difference between the output of f_c at the input I_0 and the baseline $\tilde{I}_0$, if f_c is differentiable almost everywhere. In practice, the integral in integrated gradients is approximated via a summation. We sum the gradients at points occurring at sufficiently small intervals along the straight-line path from the input M to a baseline $\tilde{M} = \mathbf{0}$: $\nabla^{IG} f_c(M) = \frac{1}{S} \sum_{s=1}^{S} \frac{\partial f_c\left(\Phi\left(I_0, \frac{s}{S} M\right)\right)}{\partial M}$, where S is a constant, usually 20. Integrated gradients have some nice theoretical properties and perform better than the gradient-based approaches. However, the heatmap generated by the integrated gradients is still diffuse (Fig. 1 and Fig. 3).

I-GOS: We believe the above two approaches can be combined for a better heatmap approach. The integrated gradient naturally provides a better direction than the gradient in that it points more directly to the global optimum of a part of the objective function. One can view the convex constraint function $g(M)$ as equivalent to the Lagrangian of a constrained optimization approach with constraints $\|\mathbf{1} - M\|_1 \le B_1$ and $TV(M) \le B_2$, B_1 and B_2 being positive constants, hence consider the optimization problem (1) to be a constrained minimization problem on $f_c(\Phi(I_0, M))$. In this case, we know the unconstrained solution in $M = \mathbf{0}$ is outside the constraint region. We speculate that an optimization algorithm may be better than gradient descent if it directly attempts to move to the unconstrained optimum.

To illustrate this, Fig. 2 shows a 2D optimization with a starting point A, a local optimum C, and a baseline B. The area within the black dashed line is the constraint region which is decided by the constraint function $g(M)$ and the boundary of M. A first-order algorithm will follow the gradient descent direction (the purple line) to the local optimum C; while the integrated gradients computed along

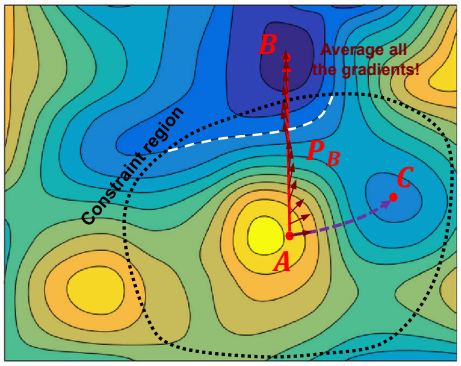

Figure 2. (Best viewed in color) Suppose we are optimizing in a region with a start point A, a local optimum C, and a baseline B which is the unconstrained global optimum; the area within the black dashed line is the constraint region which is decided by the constraint terms $g(M)$ and the bound constraints $\mathbf{0} \leq M \leq \mathbf{1}$, we may find a better solution by always moving towards B rather than following the gradient and end up at C.

the path P_B from A to the baseline B may enable the optimization to reach an area better than C within the constraint region. We can see that the integrated gradients with an appropriate baseline have a global view of the space and may generate a better descent direction. In practice, the baseline does not need to be the global optimum. A good baseline near the global optimum could still improve over the local optimum achieved by gradient descent.

Hence, we utilize the integrated gradients to substitute the gradient of the partial objective $f_c(M)$ in optimization (1), and introduce a new visualization method called Integrated-Gradient Optimized Saliency (I-GOS). For the regularization terms $g(M)$ in optimization (1), we still compute the partial (sub)gradient with respect to M: $\nabla g(M) = \lambda_1 \cdot \frac{\partial ||\mathbf{1}-M||_1}{\partial M} + \lambda_2 \cdot \frac{\partial \mathrm{TV}(M)}{\partial M}$.

The total (sub)gradient of the optimization for M at each step is the combination of the integrated gradients for the $f_c(M)$ and the gradients of the regularization terms $g(M)$: $TG(M) = \nabla^{IG} f_c(M) + \nabla g(M)$. Note that this is no longer a conventional optimization problem, since it contains 2 different types of gradients. The integrated gradients are utilized to indicate a direction for the partial objective $f_c(M)$; the gradients of the $g(M)$ are used to regularize this direction and prevent it to be diffuse.

Due to the high cost of computing the integrated gradients, we employ a line-search based gradient-projection method to maximally utilize each computation of the integrated gradients. Line search is more efficient in finding appropriate step sizes, allowing significantly less iterations to be used. In order to avoid adversarial examples, we add different random noise n_s to I_0 at each point along the straight-line path when computing the integrated gradients; and we set the resolution of the mask M be smaller than the shape of the input I_0, and perturb I_0 with $\Phi(I_0, \mathrm{up}(M))$, where $\mathrm{up}(M)$ upsamples M to the original resolution of I_0.

3. Experiments

We follow [3] to adopt *deletion* and *insertion* as metrics to evaluate the performance of the heatmaps generated by different approaches. The intuition behind the *deletion* metric is that the removal of the pixels most relevant to a class will cause the original class score dropping sharply. The intuition behind the *insertion* metric is that only keeping the most relevant pixels will retain the original score as much as possible, which can eliminate the disturbing from the adversarial attacks. The insertion metric would not score adversarial examples highly, since to achieve a good insertion score, the deep model needs to make a confident, consistent prediction using a small part of the image.

We utilize the pretrained VGG19 network [6] from the PyTorch model zoo to test $5,000$ randomly selected images from the validation set of ImageNet [4]. Table 1 shows the comparative evaluations of I-GOS with other state-of-the-art approaches in terms of both *deletion* and *insertion* metrics. Fig. 3 shows some comparison examples between different approaches on 224×224 heatmaps. From Table 1 we observe that our proposed approach I-GOS performs better than Excitation BP [10] and Mask [1] in both deletion and insertion scores for heatmaps with all different resolutions. RISE [3] and Integrated Gradients can only generate 224×224 heatmaps. GradCam [5] can only generate 14×14 heatmap on VGG19. And our approach also beats RISE, Integrated Gradient, and GradCam in both deletion and insertion scores on heatmaps with the same resolutions. Although Integrated Gradients has some good properties theoretically, it gets the worst insertion score among all the approaches, which indicates that it indeed contains lots of diffused pixels uncorrelated with the classification, as in the *Cucumber* and *Oboe* examples in Fig. 3. Excitation BP is a one-step backpropagation-based approach that is better than other one-step backpropagation-based approaches, and during the experiments we find that sometimes it just fires on the border and corner of the image instead of the contents, or on irrelevant parts of the image as argued in [2]. Thus, it performs the worst in the deletion task. RISE also suffers on the deletion score maybe because of the randomness on the masks it generates.

We also compare the running time for I-GOS with those for Mask, RISE, GradCam, and Integrated Gradients on VGG19. For each approach, we only use one Nvidia 1080Ti GPU to test $5,000$ images. For I-GOS, the maximal iteration is 15; for Mask, the maximal iteration is 500; for RISE, the number of random input samples is $4,000$. For I-GOS, it takes about 5 seconds to generate the heatmap for each image; for Mask, it takes about 14 seconds; and for RISE, it takes over 30 seconds. To the best of our knowledge, I-GOS is the fastest one among these perturbation-based methods. The average running times for the backpropagation-based methods (GradCam and Integrated Gradients) are all less

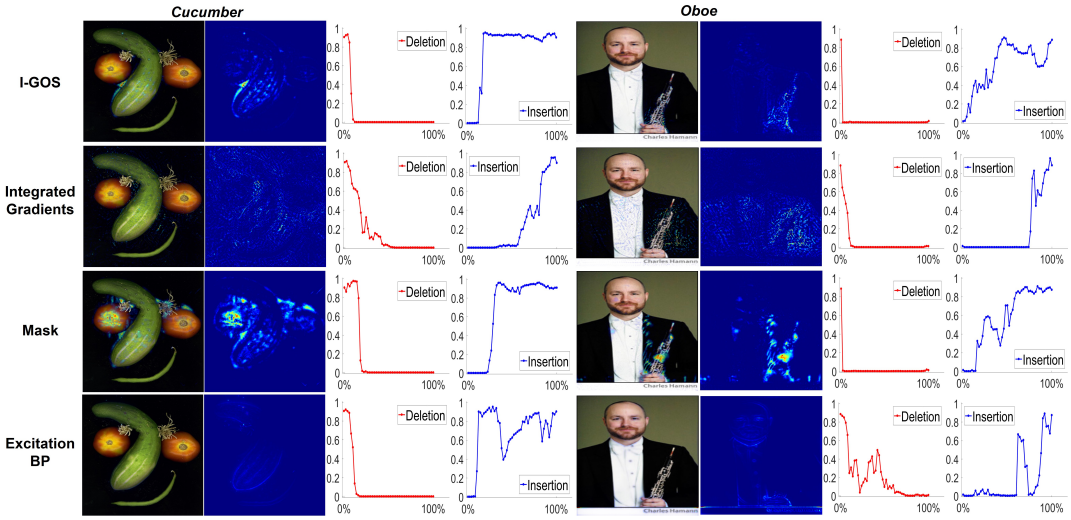

Figure 3. A comparison among different approaches with heatmaps of 224×224 resolution. The red plot illustrates how the CNN predicted probability drops with more areas masked, and the blue plot illustrates how the prediction increases with more areas revealed. The x axis for the red/blue plot represents the percentage of pixels masked/revealed; the y axis for the red/blue plot represents the predicted class probability. One can see with I-GOS the red curve drops earlier and the blue plot increases earlier, leading to less area under the deletion curve (deletion metric) and more area under the insertion curve (insertion metric). (Best viewed in color)

Table 1. Evaluation in terms of deletion (lower is better) and insertion (higher is better) scores on ImageNet dataset using the VGG19 model. GradCam can only generate 14×14 heatmaps for VGG19; RISE and Integrated Gradients can only generate 224×224 heatmaps.

	224×224		112×112		28×28		14× 14	
	Deletion	Insertion	Deletion	Insertion	Deletion	Insertion	Deletion	Insertion
Excitation BP [10]	0.2037	0.4728	0.2053	0.4966	0.2202	0.5256	0.2328	0.5452
Mask [1]	0.0482	0.4158	0.0728	0.4377	0.1056	0.5335	0.1753	0.5647
GradCam [5]	– –	– –	– –	– –	– –	– –	0.1527	0.5938
RISE [3]	0.1082	0.5139	– –	– –	– –	– –	– –	– –
Integrated Gradients [8]	0.0663	0.2551	– –	– –	– –	– –	– –	– –
I-GOS (ours)	**0.0336**	**0.5246**	**0.0609**	**0.5153**	**0.0899**	**0.5701**	**0.1213**	**0.6387**

than 1 second. However, our approach achieve much better performance than these approaches, especially with higher resolutions.

4. Conclusion

In this paper, we propose a novel visualization approach I-GOS, which utilizes integrated gradients to optimize for a heatmap. We show that the integrated gradients provides a better direction than the gradient when a good baseline is known for part of the objective of the optimization. The heatmaps generated by the proposed approach are human-understandable and more correlated to the decision-making of the model. Extensive experiments show that I-GOS advances state-of-the-art deletion and insertion scores on all heatmap resolutions.

References

[1] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pages 3449–3457, 2017. 1, 2, 3, 4

[2] W. Nie, Y. Zhang, and A. Patel. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. *ArXiv e-prints*, May 2018. 1, 3

[3] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. *ArXiv e-prints*, June 2018. 3, 4

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 3

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 1, 3, 4

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. 3

[7] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR Workshop*, 2015. 1, 2

[8] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 1, 2, 4

[9] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1, 2

[10] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, pages 543–559, 2016. 1, 3, 4