Visualizing the Resilience of Deep Convolutional Network Interpretations

Bhavan Vasu, Andreas Savakis
Rochester Institute of Technology
Rochester, New York
{bxv7657, andreas.savakis}@rit.edu

Abstract

This paper aims at visualizing the resiliency of deep network interpretations across datasets. We further explore how these interpretations change when network weights are damaged. We utilize Class Activation Maps to obtain heatmaps of deep network interpretations and identify salient local regions. We apply our methods on two remote sensing datasets and demonstrate that representations are resilient across similar datasets. We also demonstrate the benefits of transfer learning for different datasets. We further analyze these interpretations when the network weights are damaged and illustrate that retraining a damaged network is useful in recovering its performance. Our visualization results, based on ResNet50, offer insights in the resiliency of convolutional network architectures.

1. Introduction

This paper aims at visualizing deep convolutional neural network interpretations for aerial imagery [10], [11] and understanding how these interpretations change when network weights are damaged. We focus our investigation on networks for aerial imagery, as these may be prone to damages due to harsh operating conditions and are usually inaccessible for maintenance once deployed. Visualizing changes in the network's interpretation, when the undamaged weights are retrained, allows us to visually assess the resilience of a network.

Visualizing CNN internal representations is a way to better understand the way deep networks interpret images [13], [12], [15]. The work in [9] uncovered salient structures and textures present in network interpretations for aerial imagery, illustrated in Figure 1. This paper is an extension of the work in [9] to include analysis on the resilience of network interpretations to weight damages. Additionally, we demonstrate how recovery of the interpretations is possible by retraining a damaged network. The main contributions of this paper are:

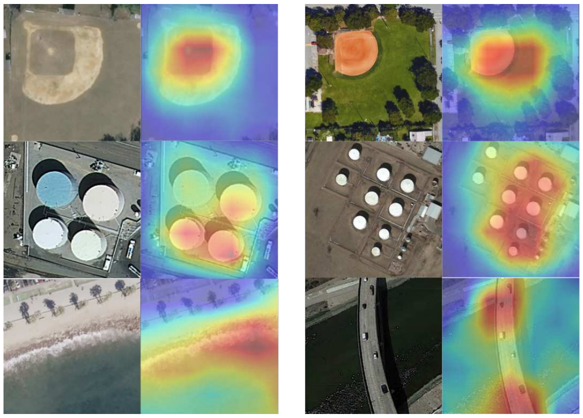

Figure 1. Class activation maps (CAMs) for aerial imagery from UCM (left columns) and AID (right columns) datasets. Images and CAMs are shown for classes: baseball field (top row), storage tanks (middle row), beach (bottom left) and bridge (bottom right).

- Demonstrate the resilience of class activation maps during transfer learning on aerial datasets.
- Visualize the effects of damaged network weights on network interpretations.
- Visualize the effect of retraining the network, as a form of adaptation that helps the network recover its interpretation after substantial damage.

2. Related Work

Zeiler and Fergus, in [13], introduced one of the first visualization techniques for understanding the workings of CNNs. More recently, the Global Average Pooling (GAP) layer has been used to obtain class activation maps as demonstrated in [15]. The work in [1] aims to visualize what the network learns using a DeconvNet [14] to get pixel level predictions for different textures. Bau and Zhou adopted a more flexible approach to classifying a broader set of features by shedding light on how a CNN has remarkable localization abilities [1]. The work in [8] encouraged the use of class activation maps for tasks such as visual

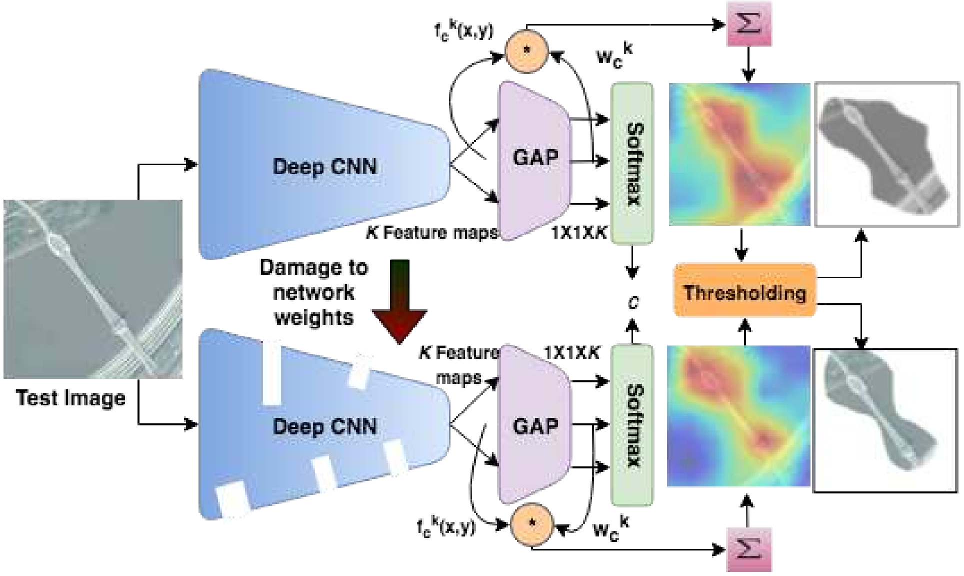

Figure 2. Block diagram for class activation map extraction for intact CNN (top path) and CNN with damaged weights (bottom path).

questioning answer and image captioning. Recent work in [5] was able to link spatial information importance to classification accuracy by generating randomized occlusion binary maps.

A different line of research deals with understanding how errors in the network weights affect classification performance. The effects of various modifications in the network weights were analyzed in [2], [4]. In [2] quantization and Principal Component Analysis (PCA) were used to drop network weights. The work in [7] established a framework for estimating resiliency of DCNN's by understanding the relationship between bit error rate and classification error. Recent work in [6] examines the classification accuracy for various amounts of weight damage in deep convolutional architectures. In this paper, we visualize the effects of weight damage on class activation maps for network interpretation.

3. Methodology

3.1. Aerial-CAM

Class activation maps (CAMs) are extracted for aerial images [9] using the GAP layer [15]. The CAMs are obtained using the intact network and the network with various degrees of weight damage, as illustrated by Figure 2. After the CAMs are obtained, the Highest Activation Region (HAR) is found to localize the image region that most contributes to the network activation.

A CAM is the image region that most influences the classification decision of the network. The GAP layer is a weighted sum of the feature maps from the last convolutional layer and is used to generate CAMs. For a given feature map, let $f^k(x, y)$ represent the activation of unit k in the last convolutional layer at the (x, y) location in the test image and w^k is the corresponding GAP layer output. The CAM, $F(x, y)$ for an aerial scene belonging to class c is given by

$$F_c(x, y) = \sum_k w_c^k f_c^k(x, y). \tag{1}$$

Our work investigates how these interpretations are affected by transfer learning to gain insight into the CAM's transferability across tasks. We also explore the effects of weight damage on network interpretations.

3.2. CAM Resilience to Weight Damage

In this work, weight damages are spread randomly using the framework shown in Figure 2. We introduce weight damage at four levels $D1$, $D2$, $D3$ and $D4$, for the popular ResNet50 architecture [3]. These selected nodes are all present in the first convolution layer of each network block. If a layer contains N trainable weights, we drop $D\%$ of N weights.

Dropping network weights is forcing the selected weights to zero, simulating the stuck at zero damage. The amount of damage $D\%$ is swept from 5% to 95% of the total number of weights in a given layer across all filters between the convolutional layers l and $l + 1$. The skip connections in ResNet50 are not altered.

Once damaged, the network weights are permanently disabled in the test and retrain phase. Therefore, only the undamaged weights are updated during retraining, while gradient flow to the damaged weights is disabled. Retraining can help reinforce existing network connections and adapt to new interpretations with the limited number of weights remaining in the damaged network.

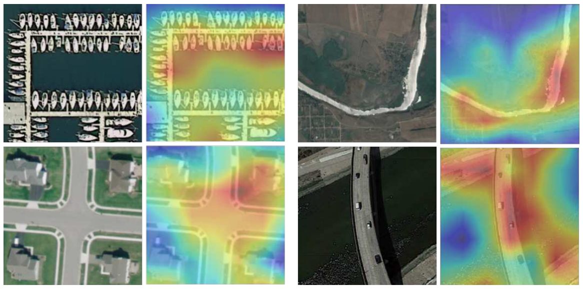

Figure 3. Visualizing CAM resilience across aerial datasets, when training one one dataset (UCM/AID) and testing on another (AID/UCM) without retraining. Results shown when training on UCM and testing on AID (left columns) or vice versa (right columns), and for shared classes among datasets: harbor, river (top row), or for new previously unseen classes: intersection, bridge (bottom row).

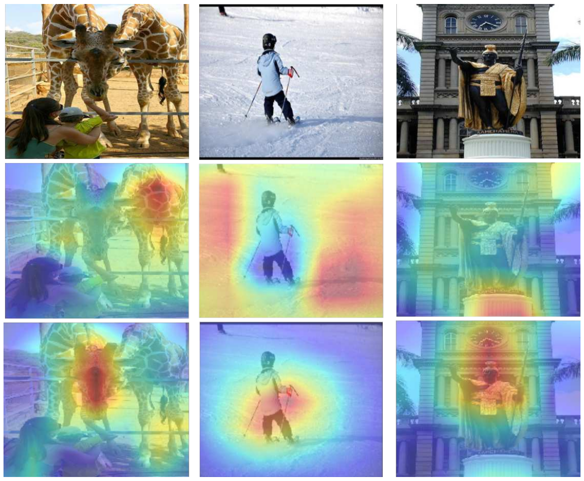

Figure 4. Visualizing CAM failure and recovery after retraining to illustrate the benefit of transfer learning from aerial imagery to MS-COCO. Images (top row), CAMs without retraining (middle row) and CAMs after transfer learning (bottom row).

4. Results

4.1. CAMs For Transfer Learning

We begin our experiments by illustrating the resilience of deep network interpretations across datasets. In these experiments, we obtain CAM visualization results for the ResNet50 architecture [9]. The resilience of CAMs across aerial datasets, UCM and AID, is illustrated in Figure 3. Good CAM heatmaps and localizations are obtained even for classes that were not used during training. However, useful representations are not maintained when the datasets have significant differences, for example when training with AID aerial data and testing on MS-COCO. As illustrated in Figure 4, transfer learning plays a crucial role in improving the CAMs and the ability of the network to localize useful image regions.

	AID	**UCM**
Training Accuracy	96.3	94.2
Validation Accuracy	92.4	91.7

Table 1. Percent accuracy of ResNet50 on aerial datasets.

4.2. CAMs With Weight Damage

Table 1 shows accuracy results for ResNet50, without any damage, when trained separately on the AID and UCM datasets to classify 14 classes. When introducing damage to network weights, we experiment with the layer location and amount of damage. Four layer locations were chosen and the amount of damage at these layers was independently varied from 5% to 95% in increments of 5%.

Figure 5 shows results of CAMs for ResNet50 with 50% damaged weights and retraining. In Figure 5(a)(b) we observe the network is able to classify test images from class airport and baseball field correctly even after substantial (75%) damage to the network weights. We also observe a drop in the highest activation region (HAR), which is later recovered after retraining, as shown in the last two columns. In Figure 5(c)(d) the network misclassifies an image from class storage tanks and resort as airport. The reasoning behind the misclassification can be explained by the class activation map after damage (highlighted in red). The activation is focused on regions that resemble an airport/airplanes. Retraining helps the network regain its lost interpretations as seen in the last two columns of Figure 5.

5. Conclusion

This paper explores the resilience of deep network interpretations during transfer learning and when network weights are damaged. We illustrate that network representations are resilient across aerial datasets even without retraining, but transfer learning greatly improves representations when the test domain is significantly different, as is the case with MS-COCO. We consider the commonly used ResNet50 network in our investigation of resiliency under weight damage. Visualizing how class activation maps behave when weights are damaged, provided insight into the network's decision making process under failure conditions. Retraining to overcome the effects of damage, gave us activation maps illustrating how network decision making improves.

6. Acknowledgement

This research was supported in part by a grant from the AFOSR Dynamic Data Driven Applications Systems (DDDAS) program.

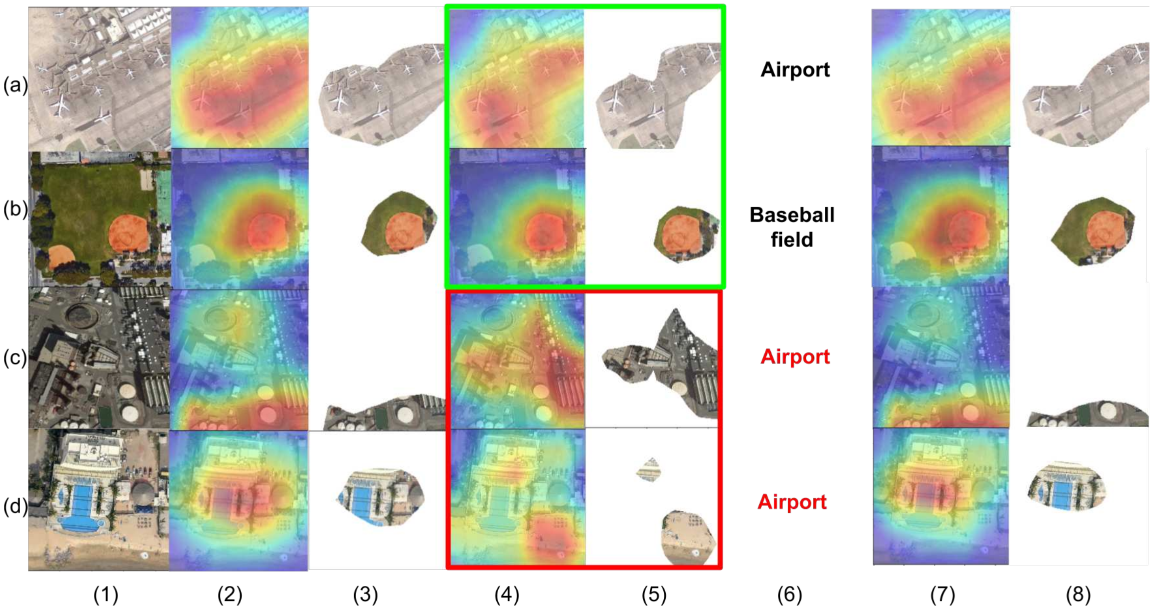

Figure 5. Results with class activation maps for ResNet50 on AID dataset for classes (a) airport (b) baseball field (c) storage tanks and (d) resort. Columns show (1) test image, (2) CAM without damage, (4) CAM after weight damage, (7) CAM after retraining; (3) HAR without damage, (5) HAR after weight damage, (8) HAR after retraining; (6) predicted label after weight damage, where red indicates incorrect prediction.

References

[1] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

[2] M. Carvalho, M. Cord, S. E. F. de Avila, N. Thome, and E. Valle. Deep neural networks under stress. *CoRR*, abs/1605.03498, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.

[5] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[6] F. U. Rahman, B. Vasu, and A. Savakis. Resilience and self-healing of deep convolutional object detectors. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3443–3447. IEEE, 2018.

[7] B. Reagen, U. Gupta, L. Pentecost, P. Whatmough, S. K. Lee, N. Mulholland, D. Brooks, and G.-Y. Wei. Ares: A framework for quantifying the resilience of deep neural networks. In *Proceedings of the 55th Annual Design Automation Conference*, DAC '18, pages 17:1–17:6, New York, NY, USA, 2018. ACM.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization., in ICCV, 2016.

[9] B. Vasu, F. U. Rahman, and A. Savakis. Aerial-cam: Salient structures and textures in network class activation maps of aerial imagery. In *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5, June 2018.

[10] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, July 2017.

[11] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279. ACM, 2010.

[12] J. Yosinski, J. Clune, A. M. Nguyen, T. J. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.

[13] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[14] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2528–2535, June 2010.

[15] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.