AUGMENTED GRAD-CAM: HEAT-MAPS SUPER RESOLUTION THROUGH AUGMENTATION

Pietro Morbidelli[⋆] *Diego Carrera*[†] *Beatrice Rossi*[†] *Pasqualina Fragneto*[†] *Giacomo Boracchi*[⋆]

[⋆] Politecnico di Milano, DEIB, Italy
[†] STMicroelectronics, Agrate Brianza, Italy

ABSTRACT

We present *Augmented Grad-CAM*, a general framework to provide a high-resolution visual explanation of CNN outputs. Our idea is to take advantage of image augmentation to aggregate multiple low-resolution heat-maps – in our experiments Grad-CAMs – computed from augmented copies of the same input image. We generate the high-resolution heat-map through *super-resolution*, and we formulate a general optimization problem based on Total Variation regularization. This problem is entirely solved on the GPU at inference time, together with image augmentation. Augmented Grad-CAM outperforms Grad-CAM in weakly supervised localization on Imagenet dataset, and provides more detailed heat-maps. Moreover, Augmented Grad-CAM turns to be particularly useful in monitoring the production of silicon wafers, where CNNs are employed to classify defective patterns on the wafer surface to detect harmful faults in the production line.

Index Terms— Visual interpretation, Superresolution, Class Activation Mapping, Image Augmentation

1. INTRODUCTION

Several works have recently addressed the problem of highlighting which pixels in the input image have mostly influenced the output of a CNN [1, 2, 3, 4]. These *interpretability* tools, of which Grad-CAM [2] is the most prominent example, generate heat-maps that shade light on why a CNN made a specific decision and are very useful to support a decision-making process (e.g. in industrial monitoring). Meaningful heat-maps have to be well localized around the object of the queried category (*class-discriminative*) and should capture fine-grained details (*high-resolution*). Unfortunately, heat-maps are computed from the activations at the last convolutional layer, which contains very informative features but has a low spatial extent. As a matter of fact, bilinear/bicubic interpolation is used to severely upsample (e.g. 32x) and then superimpose a heat-map to the input image, resulting in necessarily poor localization.

Our idea is to increase heat-maps resolution through image augmentation. While image augmentation is widely used for improving CNN training, and for increasing robustness of predictions at test time [5], none of the existing *visual explanation tools* [1, 2, 3, 4] leverages augmentation to perform super-resolution. Nevertheless, all the responses that the network generates to the multiple augmented versions of the same input image are very informative for reconstructing an high-resolution heat-map. Here we show that, by modeling the interplay between image augmentation and heat-map computation, it is possible to exploit this information and achieve better high-resolution heat-maps than interpolation.

(a) Grad-CAM. (b) Augmented Grad-CAM.

Fig. 1: Example of heat-maps of the *mastiff* class computed using VGG16. Augmented Grad-CAM provides a better localized heat-map than the traditional Grad-CAM and better captures fine-grained details like the dog's legs and ears. Moreover Augmented Grad-CAM does not highlight areas outside the dog, as opposed to Grad-CAM.

Our contribution is *Augmented Grad-CAM*, a general framework to perform heat-map super-resolution by taking advantage of the information shared in multiple low-resolution heat-maps computed from the same input under different – but known – transformations. In particular, we model heat-maps computed by Grad-CAM as the result of an unknown downsampling operator, which we invert to recover a high-resolution heat-map, as in multiframe super-resolution [6]. We formulate heat-map super-resolution as an optimization problem where Anistropic Total Variation regularization is used to preserve the edges in the recovered high-resolution heat-map. Our framework is flexible and can be applied to heat-maps computed form different visualization tools, not necessarily Grad-CAM. Moreover, the whole

super-resolution procedure can be efficiently performed on the GPU at inference time. Our TensorFlow implementation of Augmented Grad-CAM has been publicly released[1]. Our experiments demonstrate that Augmented Grad-CAM yields more detailed heat-maps than its original counterpart Grad-CAM (see for instance Figure 1), and that it outperforms Grad-CAM in the primary test for quantitatively assessing interpretability, namely the Imagenet weakly supervised localization challenge [7]. We also show that Augmented Grad-CAM is particularly useful in wafer monitoring, a crucial problem for semiconductor manufacturing where CNNs are employed [8, 9] to classify spatial arrangements (patterns) of defects detected on the surface of silicon wafer. This is a very important problem since certain patterns might indicate specific problems in the production machinery that have to be promptly fixed. In this monitoring context, interpretability techniques and high-resolution maps can help to figure out ambiguous patterns and support the decision-making process of an operator.

2. RELATED WORKS

Several tools have been recently proposed to provide a visual explanation of CNN decisions and Class Activation Mapping (CAM) [1] is one of the first solutions. CAM needs to modify a trained CNN by removing all the fully-connected layers at the top and introducing a Global Average Pooling (GAP) layer followed by a single fully connected layer. This simplified architecture allows CAM to compute a heat-map for each output class, by summing the activations of the last convolutional layer rescaled by the weights of the newly introduced fully-connected layer associated to the selected class. Modifying the network architecture results in a few drawbacks: need for retraining and lower performance due to the simple layer on top of the CNN. Grad-CAM [2] is a generalization of CAM that does not require to change the CNN architecture, thus can be also applied to networks not necessarily designed for classification, but also for other tasks like captioning or visual question answering. There are a few recent extensions of Grad-CAM. Grad-CAM++ [3] computes higher-order derivatives to increase localization accuracy of Grad-CAMs, in particular in presence of multiple occurrence of the same objects in the image. Smooth Grad-CAM++ [4] adopts Smoothgrad [10] to smooth the gradients computed during Grad-CAM++, and provide visually sharpen heat-maps. Smoothgrad adopts multiple noisy versions of the same input image, which can be seen as a simple form of image augmentation. Except from noise addition in Smooth Grad-CAM++, interpretability tools based on heat-maps ignore augmentation, and none of the existing solutions address heat-maps computation as a super-resolution problem.

[1]Our implementation is available at `https://github.com/diegocarrera89/AugmentedGradCAM`

3. GRAD-CAM AND PROBLEM FORMULATION

Whilst our super-resolution framework can be used on any heat-map, thus not necessarily Grad-CAM [2] or its extensions, here we focus on Grad-CAM being the best known. Grad-CAM computes a heat-map $\mathbf{g} \in \mathbb{R}^{n\times m}$ that highlights which regions of the input image $\mathbf{x} \in \mathbb{R}^{N\times M}$ have mostly influenced the classifier score in favor of the class c (uppercase letters indicate sizes that are larger than lowercase one, i.e. $N > n$ and $M > m$). Let y_c denote the score of the class c and $\mathbf{a}^k \in \mathbb{R}^{n\times m}, k = 1, \ldots, K$, the activation maps corresponding to the k-th filter of the last convolutional layer. The Grad-CAM for the class c is defined as a weighted average of $\mathbf{a}^k, k = 1, \ldots, K$, followed by a ReLU activation:

$$\mathbf{g}_c = \text{ReLU}\big(\sum_k \alpha_c^k \mathbf{a}^k\big), \tag{1}$$

where the *importance weights* $\{\alpha_c^k\}$ are defined as the average derivatives of y_c with respect to each pixel (i, j) in the activation $\mathbf{a}^k$:

$$\alpha_c^k = \frac{1}{nm} \sum_i \sum_j \frac{\partial y_c}{\partial \mathbf{a}^k(i, j)}\,. \tag{2}$$

The ReLU in (1) indicates that only positive contributions are relevant to compute the heat-map.

The heat-map $\mathbf{g}_c$ has typically low-resolution and it is upsampled to the size of $\mathbf{x}$ via bilinear/bicubic interpolation. This resizing allows to superimpose the Grad-CAM to the input image and interpret network decisions, as in Figure 1. Another option to increase the resolution of the Grad-CAM is to select activations at previous convolutional layers of the network, as these exhibit a higher spatial resolution. However, the high-level semantic features are mainly in the last layer [2], and this option would yield heat-maps that are less class-discriminative.

The problem we address here is to increase the resolution of Grad-CAM in an input image. Our idea is to leverage image augmentation, which is often employed to improve classification performance of a CNN.

4. AUGMENTED GRAD-CAM

To improve the resolution of heat-maps provided by Grad-CAM we take advantage of image augmentation performed at test time and, given an input image $\mathbf{x}$ and a class c [2], we generate L augmented copies of $\mathbf{x}$ as

$$\mathbf{x}_l = \mathcal{A}_l(\mathbf{x}), \;\; l = 1 \ldots, L, \tag{3}$$

where the augmentation operators $\mathcal{A}_l : \mathbb{R}^{N\times M} \to \mathbb{R}^{N\times M}$ include random rotations and translations of $\mathbf{x}$.

[2]for sake of simplicity hereinafter we remove the subscript c, being often the network prediction.

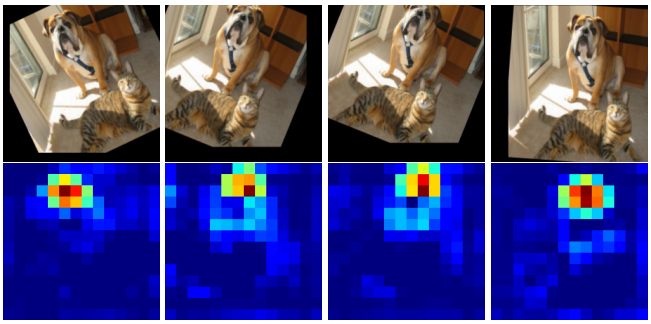


Fig. 2: Example of Augmented Grad-CAMS (top row), each one generated from different augmented version of the same image (bottom row). The Augmented Grad-CAMs contain different information that can be exploited to achieve a more detailed high-resolution heat-maps, see Figure 1.

Our intuition is that the set of Grad-CAMs $\{\mathbf{g}_l\}$ computed from augmented images $\{\mathbf{x}_l\}$ is much more informative than the Grad-CAM computed from the single input image $\mathbf{x}$, as can be observed in Figure 2. In fact, the trained CNN is not perfectly invariant to roto-translations, and Grad-CAMs computed from augmented versions of the same input image do not correspond to roto-translations of the Grad-CAM of the original image. Therefore, each augmented Grad-CAM $\{\mathbf{g}_l\}$ might bring useful information to a super-resolution algorithm, resulting in superior high-resolution heat-maps. Super-resolution allows us to exploit the whole semantic information in the network coming from activations at the last convolutional layer, and at the same time leverage information spread through multiple maps provided by image augmentation.

The modeling assumption underlying Augmented Grad-CAM is that each $\mathbf{g}_l$ is obtained from downsampling an augmented version of an unknown high-resolution heat-map $\mathbf{h} \in \mathbb{R}^{N\times M}$, which is the one we want to reconstruct:

$$\mathbf{g}_l \approx \mathcal{D} \circ \mathcal{A}_l(\mathbf{h}), \tag{4}$$

where $\mathcal{D}\colon \mathbb{R}^{N\times M} \to \mathbb{R}^{n\times m}$ is the downsampling operator and $\circ$ denotes the function composition. Moreover, since the downsampling and the augmentation operators, at least in case of augmentation made of roto-translations, are linear, (4) can be written as $\mathbf{g}_l \approx DA_l\mathbf{h}$, where D and A_l are the matrix representing transformations $\mathcal{D}$ and $\mathcal{A}_l$ respectively.

We reconstruct $\mathbf{h}$ by solving this optimization problem:

$$\min_{\mathbf{h}} \frac{1}{2}\sum_{l=1}^{L} \|\mathcal{D}A_l h - \mathbf{g}_l\|_2^2 + \lambda \mathrm{TV}_{\ell_1}(\mathbf{h}) + \frac{\mu}{2}\|\mathbf{h}\|_2^2, \tag{5}$$

where TV_{ℓ_1} denotes the *anisotropic total-variation* regularization, defined as the sum of the absolute values of the horizontal and vertical derivative of $\mathbf{h}$, i.e.

$$\mathrm{TV}_{\ell_1}(\mathbf{h}) = \sum_{i,j} |\partial_x \mathbf{h}(i,j)| + |\partial_y \mathbf{h}(i,j)|.$$

Table 1: Top-1 and Top-1 localization errors achieved over the ILSVRC 2015 validation set.

	Single	Avg	Max	Augmented
Top-1 Error	0.5908	0.5964	0.5817	**0.5725**
Top-5 Error	0.4974	0.5025	0.4859	**0.4747**

This regularization is typically employed in super-resolution algorithms [6] as it preserves the edges in the reconstructed image. We also adopt an ℓ^2 penalization term, as we experience it improves the stability of the minimization problem. Problem (5) is convex and nonsmooth due to the TV_{ℓ_1} term. We solve (5) by means of subgradient descent, where we set the step size using the Adam optmizer [11] and use a TensorFlow implementation to exploit GPU computing parallelism.

5. EXPERIMENTS

Our experiments are meant to demonstrate that: *i)* Augmented Grad-CAM improves the quality of heat-maps used for visual explanation of CNNs and *ii)* simple aggregation strategies are not sufficient to produce good high-resolution heat-maps. We consider the following solutions:

Single Grad-CAM: the Grad-CAM [2] computed from the input image, which does not involve any augmentation.

Max / Avg Grad-CAM: two baseline aggregation solutions where $\{\mathbf{g}_l\}$ are first upsampled by bilinear interpolation, and then registered to the original image $\mathbf{x}$ by the inverse operator $\mathcal{A}_l^{-1}$. Finally, the registered heat-maps are aggregated by pixel-wise maximum / average.

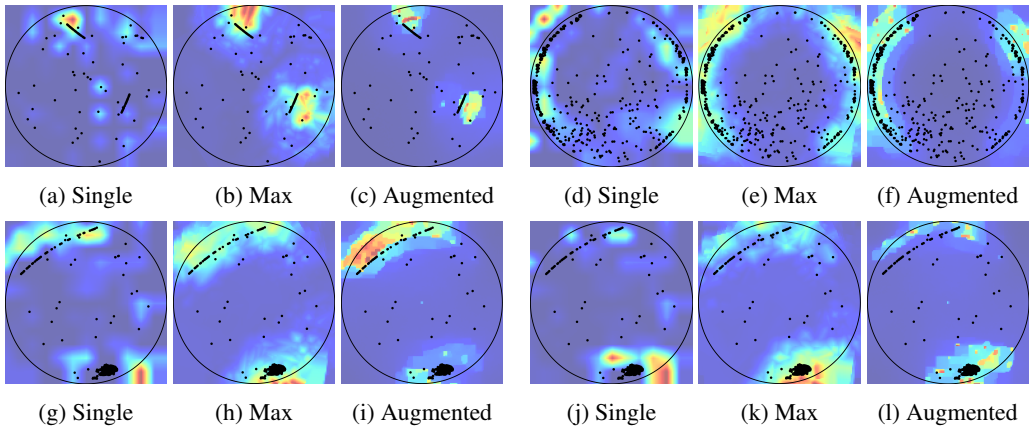
Augmented Grad-CAM: the proposed solution, where $\{\mathbf{g}_l\}$ are aggregated by solving (5). The parameters λ and μ in (5) are manually selected to achieve good quality high-resolution heat-maps in a small subset of 5 images.

We perform two experiments: at first we consider weakly supervised localization to quantitatively assess interpretability performance. Then, we show that our solution yields high-resolution heat-maps for wafer monitoring that are qualitative better than the alternatives.

In all our experiments, we generate $L = 99$ augmented input images $\mathbf{x}_l$ plus the original image to compute 100 low-resolution Grad-CAMs $\{\mathbf{g}_l\}$. The augmentation consists in roto-translating the input images, where the choice of the roto-translation parameters depends on the considered application, and is described in what follows.

5.1. Weakly Supervised Localization

The goal of weakly supervised localization is estimate a bounding box of the principal object of the image on top of its label. The term "weakly" refers to the fact that no bounding box annotations are provided for training. We address this problem following the approach in [2]: given the predicted class, we compute a heat-map $\mathbf{h}$ using each of the considered methods. Each bounding box is computed by



(a) Single (b) Max (c) Augmented (d) Single (e) Max (f) Augmented

(g) Single (h) Max (i) Augmented (j) Single (k) Max (l) Augmented

Fig. 3: Heat-maps generated by the considered methods superimposed to WDMs reporting the following patterns: two *scratches* (a,b,c), a *ring* (d,e,f), a *scratch* and a *fingerprint* simultaneously (g,h,i and j,k,i respectively). Heat-maps in the second row are computed for the same WDM considering the *scratch* class (g,h,i) and the *fingerprint* class (j,k,i).

thresholding the heat-map $\mathbf{h}$ at 15% of its maximum value, and by drawing a bounding box around the largest connected component of the resulting binary mask. As a backbone for the classification network we adopt a pre-trained VGG-16 [5] as in [2]. The augmentation is performed using rotations of angles uniformly sampled in $[-30°, 30°]$ followed by a translations in the range $[-10, 10]$ pixel along each direction. Localization performance are computed over the validation set of the ILSVRC 2015 localization challenge [12] using the evaluation tool provided. As for classification, we consider both the top-1 and top-5 localization errors, reported in Table 1. The proposed method outperforms the others showing superior localization performance both in terms of top-1 and top-5 errors. Although we were not able to reproduce the localization performance provided in [2], we employ the same implementation (publicly available in the released package) of the Grad-CAM in all the approaches.

5.2. Augmented Grad-CAMs for Wafer Monitoring

Here we show that Augmented Grad-CAM is an effective interpretability tool for monitoring silicon wafer production. In particular, we analyze *Wafer Defect Maps (WDMs)*, namely the output of inspection machines that report the location of each defect detected on the wafer surface, and train a CNN made of 6 convolutional layers followed by two fully connected layers to classify defect patterns in 13 classes. We generate the input image by binning the WDM, obtaining a grayscale image of 128×128 pixels, while the resolution of the computed Grad-CAM is 12×12. The network is trained over the ST dataset described in [8] which comprises 29,746 WDMs acquired in the production site of STMicroelectronics of Agrate Brianza, Italy. In our experiments we generate $L = 99$ augmented copies of the original input image $\mathbf{x}$ by random rotations of angles uniformly sampled in $[-360°, 360°]$.

Figure 3 depicts some interesting cases we analyzed. Single Grad-CAM fails in identifying multiple occurrences of the same class within the same WDM (e.g. the two *scratches* in 3.a), while Max and Augmented Grad-CAMs provide more class-discriminative responses (3.b and 3.c). As expected, our super-resolution provides better localized results than Max aggregation. We also observe that Single Grad-CAM is poorly localized (3.d), as few regions belonging to the *ring* pattern are highlighted, whereas Max and Augmented Grad-CAMs cover larger areas of the pattern. The second row of Figure 3 reports a WDM containing two different patterns (*scratch* and *fingerprint*). For the *scratch* class, Augmented Grad-CAM (3.i) reports the highest activation values for regions covering the target pattern, while Single Grad-CAM (3.g) mostly highlights regions corresponding to the *fingerprint* class. Similarly, in the Augmented Grad-CAM for the *fingerprint* class the regions corresponding to the *scratch* class are less active than in other methods (3.l).

6. CONCLUSIONS

We have presented Augmented Grad-CAM, an efficient tool for providing high-resolution heat-maps for explaining CNN outputs. Grad-CAM performs super-resolution from multiple heat-maps obtained from augmentation and this allows to improve the localization performance w.r.t. Grad-CAM of a substantial margin, and to achieve better visual interpretation of WDM in the considered industrial monitoring problem. Despite our experiments have been focused on Grad-CAM only, the super-resolution framework underlying Augmented Grad-CAM is general and can account for other techniques to compute low-resolution heat-maps. Ongoing works concern extending Augmented Grad-CAM to sub-manifold sparse convolutional networks [13], which would require much higher upsampling factors [8] and raise alignment issues due to the sparse activations over network layers.

7. REFERENCES

[1] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[4] Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint arXiv:1908.01224*, 2019.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Roberto di Bella, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, and Giacomo Boracchi, "Wafer defect map classification using sparse convolutional networks," in *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, 2019.

[9] Takeshi Nakazawa and Deepak V Kulkarni, "Wafer map defect pattern classification and image retrieval using convolutional neural network," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 2, pp. 309–314, 2018.

[10] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[11] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[13] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.