High Resolution Explanation Maps for CNNs using Segmentation Networks

High Resolution Explanation Maps for CNNs using Segmentation Networks

Alessio Mascolini*, Francesco Ponzio*, Enrico Macii*, Elisa Ficarra† and Santa Di Cataldo*
†Università di Modena e Reggio Emilia, Italy
*Politecnico di Torino, Italy
Email: alessio.mascolini@polito.it

***Abstract*—Recent developments have resulted in multiple techniques trying to explain how deep neural networks achieve their predictions. The explainability maps provided by such techniques are useful to understand what the network has learned and increase user confidence in critical applications such as the medical field or autonomous driving. Nonetheless, they typically have very low resolutions, severely limiting their capability of identifying finer details or multiple subjects. In this paper we employ an encoder-decoder architecture with skip connection known as U-Net, originally developed for segmenting medical images, as an image classifier and we show that state of the art explainable techniques applied to U-Net can generate pixel level explanation maps for images of any resolution.**

I. Introduction

The problem of understanding neural network decisions, circumventing their black-box nature, is what we call *interpretability*, or *explainability*. Traditionally, interpretability works focus on visualization techniques revealing the input stimuli that excite individual feature maps at any layer in the model (a.k.a *heatmaps*, *attention maps* or *explanation maps*). Early examples were either very computationally expensive [1] or provided very noisy maps [2]. On the other hand, Class Activation Mapping (CAM), leveraging linear combination of activation maps based on the weights of the last fully connected layer, generates explainability maps that are easier to understand and noise-free, but requires the architecture to end with a Global Average Pooling layer followed by a linear classifier. By employing different techniques to calculate the weights, researchers have recently proposed CAM based techniques which remove this requirement, while still improving the quality of the maps generated and interpretability of deep neural networks (e.g. GradCAM [3], GradCAM++ [4] and ScoreCAM [5]). While the resistance to noise and low computational overhead have made these techniques the de facto standard for explaining CNN classifiers, the low resolution of the heatmaps remains their most significant weakness [6].

In this work we show that by employing an encoder-decoder architecture with skip connections, CAM-based techniques can obtain explainability maps of arbitrary high resolutions. To do so, we test GradCAM, GradCAM++ and ScoreCAM with a U-Net [7] architecture. We also show that the obtained heatmaps allow a better understanding of the learned features.

II. Methods

We employ the 2012-2017 ILSVRC ImageNet Dataset [8], featuring 1000 classes and 1.281.167 training images, 50.000 validation images, 100.000 test images. We obtain heatmaps on the validation set after training a ResNet50 [9], a VGG16 [10] and a U-Net [7] architecture on the training set.

Our aim is to show that the encoder-decoder architecture of U-Net can generate pixel level explanation maps, with extra details that are useful for a better understanding of the network's decision. We specifically compare the results of following CAM-based explainability techniques:
(i) GradCAM [3] "uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept". The gradients of the last convolutional layer are averaged over the spatial dimensions to linearly combine the activation maps. This generates a grayscale map of the areas which contributed to the prediction;
(ii) GradCAM++ [4] improves GradCAM by weighing the activation maps with a different coefficient for each pixel, which significantly improves the ability to identify multiple areas of interest in a single image;
(iii) ScoreCAM [5] generates gradient-free explanation maps by perturbing the input with upsampled activation maps and computing linear combination coefficients from the difference in the logit for the desired class.

All such techniques are not tied to a specific CNN architecture. In our experiments, we used:
(i) ResNet50 [9], a 50 convolutionals layers deep net, followed by a global average pooling layer and a fully connected layer to obtain the logits.
(ii) VGG16 [10], which features higher resolution activation maps and hence it is the default architecture in most of the CAM-based literature.
(iii) U-Net [7], which is a fully convolutional encoder-decoder architecture traditionally used for semantic segmentation, with skip connections between the encoder and decoder block to maintain spatial information. In this work we add global average pooling layer followed by fully connected layer at the end, in order to use U-net as a classifier.

978-1-6654-4214-5/22/$31.00 ©2022 IEEE

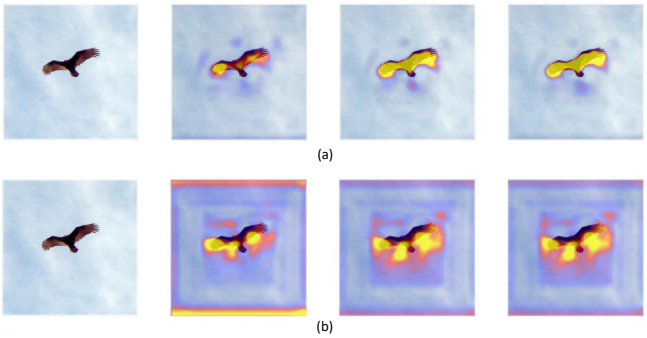

Fig. 1. Explanation maps generated through GradCAM, GradCAM++ and ScoreCAM, respectively using (a) non-ablated and (b) ablated U-Net network.

The choice of architecture has a direct impact on the obtained explanation maps. Since CAM methods linearly combine the activations in the last convolutional layer, they can only provide information as fine as the spatial resolution of such layer. For visualization purpose, the activation maps are typically upscaled to the original image size using bilinear interpolation, which typically implies issues identifying multiple subjects and fine details and localization errors. While VGG16 can typically obtain higher resolution than ResNet, this resolution is still limited to 16x16 pixels. On the other hand, being U-Net a segmentation network, the output layer is of the same size as the input layer, which allows more detailed and accurate explanation maps.

III. Results

The rationale of this study is to evaluate whether skip connections between encoder and decoder have a measurable effect on the localization ability of the network. To do so, we remove skip connections from a U-Net architecture, in a process known as *ablation*, and we compare the localization ability to an identical non-ablated network. To this date there is still no generally agreed upon metric to evaluate the quality of explanation maps in a quantitative way [11]. In our study, we consider the combination of network and CAM technique which provides the broadest explanation map (ResNet50 with GradCAM++). In this map, the low energy pixels are reasonably the ones with the lowest relevance possible to the classification task. Then, we obtain a *localization score* s as follows. We calculate a binary reference mask M by thresholding the explanation map of ResNet50 and GradCAM++. Then, we compute the portion of the energy of A, the explanation map to be evaluated, that is located within the reference mask M:

$$s = \frac{\sum_i \sum_j A_{ij} \cdot M_{ij}}{\sum_i \sum_j A_{ij}}$$

The higher s, the higher the localization ability of A.

From Table I we can see localization scores almost 10% higher for the non-ablated network (U-Net with skip connections) compared to the ablated counterpart, for every CAM method employed. The reduced localization ability of the ablated network is also evident from the examples in Fig. 1. To further demonstrate the high localization capability of our solution, Fig.2 shows more representative examples from the ImageNet Validation Set, respectively generated using GradCAM++ with ResNet50, VGG16 (the de-facto default architecture for high-resolution explainability maps) and U-Net.

TABLE I
Localization scores of U-Net and ablated counterpart.

	GradCAM	GradCAM++	ScoreCAM
U-Net	53.55%	72.47%	71.52%
Ablated U-Net	47.81%	63.04%	64.36%

In standard semantic segmentation tasks the spatial relationship between input and output of the U-Net is maintained through a specific supervised loss term pushing the output towards the ground truth segmentation mask, which is clearly absent in a classification tasks. Through the ablation study and the obtained heatmaps, we show that the skip connections between encoder and decoder preserve such spatial relationship without additional loss terms or regularizations. While VGG16 does a fairly good job of localizing the subject in the images, the higher resolution map generated by U-Net provides finer information about which features are decisive for the classification task. For example, in the last column of Fig. 2 it allows us to infer that the network has learned to identify the sticker on the apples rather than the apples themselves. This is information that is impossible to obtain with lower resolution maps, and can be very useful in identifying wrong predictions and unexpected behaviours in CNNs.

IV. Conclusions

CAM-based techniques with encoder-decoder architectures generate high resolution explanation maps, where skip connections preserve the spatial relationship with the original image. This provides better insights on how CNNs operate and what they focus on when learning.

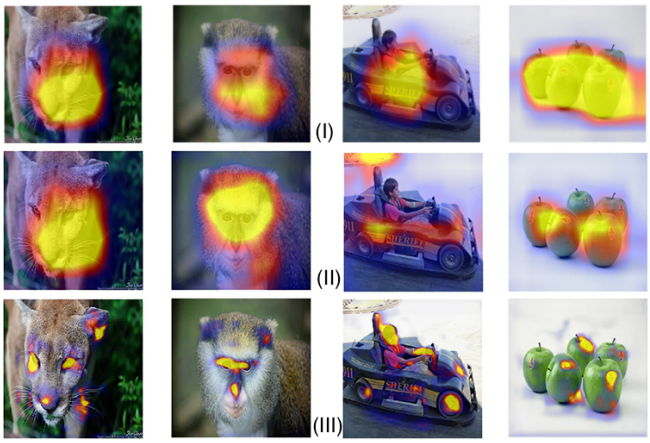

Fig. 2. GradCAM++ explanation maps: (I) ResNet50 , (III) VGG16, (III) U-Net.

References

[1] Matthew D Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*. 2013. arXiv: 1311.2901 [cs.CV].

[2] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. arXiv: 1312.6034 [cs.CV].

[3] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[4] Aditya Chattopadhay et al. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Mar. 2018). DOI: 10.1109/wacv.2018.00097. URL: http://dx.doi.org/10.1109/WACV.2018.00097.

[5] Haofan Wang et al. *Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks*. 2020. arXiv: 1910.01279 [cs.CV].

[6] Ioannis Kakogeorgiou and Konstantinos Karantzalos. "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing". In: *International Journal of Applied Earth Observation and Geoinformation* 103 (Dec. 2021), p. 102520. ISSN: 0303-2434. DOI: 10.1016/j.jag.2021.102520. URL: http://dx.doi.org/10.1016/j.jag.2021.102520.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[8] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[9] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[10] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[11] Forough Poursabzi-Sangdeh et al. *Manipulating and Measuring Model Interpretability*. 2021. arXiv: 1802.07810 [cs.AI].