The Impact of Large Receptive Fields on Grad-CAM

Rui Santos[1,2]
rui.m.santos@inesctec.pt
João Pedrosa[1,2]
joao.m.pedrosa@inesctec.pt
Ana Maria Mendonça[1,2]
amendon@fe.up.pt
Aurélio Campilho[1,2]
campilho@fe.up.pt

[1] Faculty of Engineering of the University of Porto (FEUP)
Porto, PT
[2] Institute for Systems and Computer Engineering,
Technology and Science (INESC TEC)
Porto, PT

Abstract

Deep learning models have been widely used in past years for a variety of applications, progressively achieving better results due to an increase in complexity. More complexity leads to a decrease in interpretability, which demands explanations regarding model reasoning. These explanations can be obtained with methods like Grad-CAM that computes the gradients up to the last convolutional layer to form an importance map relative to a specific class. This is followed by an upsampling operation which matches the size of the importance map to the size of the input. However, this step is based on the assumption that the feature spatial organization is maintained throughout the model, which may not be the case. We hypothesize that the spatial organization of the features is not kept during the forward pass for models with large receptive fields, which may render the importance map devoid of any meaning. This also applies to any Grad-CAM variant using the same upsampling step. The obtained results show a significant dispersion of the spatial information, which goes against the implicit assumption of Grad-CAM, and that explainability maps suffer from this dispersion. Altogether, this work addresses a key limitation of Grad-CAM which may go unnoticed for common users, taking one step further in the pursuit for more reliable explainability methods.

1 Introduction

The advent of Deep Learning (DL) has so far enabled the development of countless solutions for an ever-growing number of tasks. Nevertheless, and despite promising performances, the number of DL solutions being implemented in clinical practice is still very limited due to their lack of trustworthiness. This problem arises from the black-box nature of increasingly complex models, with a concomitant decrease in interpretability. To circumvent this issue, several methods have been proposed throughout the years to explain these models and/or their predictions. In this work we will focus on Grad-CAM [5], a common method used for computer vision tasks, with the goal of highlighting areas of an input image contributing to a given prediction. It does so by computing gradients for a given layer of the model, by averaging them for the multiple channels and multiplying them by the activations. The resulting values are summed channel-wise, rectified and scaled and finally upsampled to match the input size and overlap the explanation with the image.

Despite being widely used, in recent years there has been an increased distrust concerning the performance of this method. Different attempts have also been performed to improve Grad-CAM, but some key aspects related to its functioning are almost unexplored, which could help to explain certain flaws associated with this method. One such aspect is the upsampling step, which leads to an implicit assumption in the Grad-CAM method that there is a spatial correspondence between the last feature map and the input. However, due to the large Receptive Fields (RF) of modern architectures, the spatial correspondence assumed by Grad-CAM may not exist since specific features are influenced by very large areas of the input. Previous work has extended this idea to the Effective Receptive Field (ERF) [3] of a model, the area within the RF more likely to influence a given feature, stating that there is a misalignment between the ERF and the implicit RF derived from the upsampling step [6].

In this work, we further investigate this relation between large RFs and the performance of Grad-CAM. Exploring this association will contribute to a better understanding of this method and whether it can be improved or if it should be replaced.

2 Materials and Methods

2.1 Dataset

The dataset used throughout this work was the public version of the VinDr-CXR dataset available on Kaggle [2]. It consists of 15,000 chest x-ray images representing a set of 14 radiographic findings, with respective bounding boxes indicating pathological locations. Each image in the dataset was independently annotated by three radiologists, and labels were extracted based on the majority vote of the participating radiologists. A medical dataset was selected for these analyses since the performance and accuracy of explainability methods acquires an increased importance when applied to this scenario, since DL-based clinical decision support systems need to be highly trustworthy.

2.2 Models and training

We used three different architectures from torchvision.models for the conducted experiments, namely EfficientNet-b0, DenseNet121 and ResNet50. Every architecture contained 14 output neurons (one for each class in the dataset) and was trained on five different splits, starting with pre-trained weights on ImageNet. Each split consisted of a train, validation and a test sets, with the validation set being used to implement an early stopping strategy with a patience of 10. Every batch of images was submitted to data augmentation transforms, including rotations, cropping and changes in brightness, contrast, saturation and hue. The input size was 224 and a learning rate of 10^{-3} and a binary cross-entropy loss function were used. The batch size varied between 32 and 64 depending on memory usage for each model (64 for EfficientNet-b0 and ResNet50, 32 for DenseNet121).

2.3 Receptive field computation

The RF for each model was computed using an approach based on the work of Araujo *et al.* [1]. Through Equation 1, where r, k, s and l stand for the RF, kernel size, stride and layer index, respectively, it is possible to iteratively compute the RF. Nevertheless, this formula has its limitations (for instance, when dealing with skip connections), so the results were carefully assessed considering the properties of each architecture.

$$r_{l-1} = r_l * s_l + k_l - s_l \tag{1}$$

2.4 Explainability metrics

Two distinct metrics were used to evaluate the maps given by Grad-CAM, the Intersection over Union (IoU) and the Hit Rate (HR), following a similar procedure to [4]. However, in this work, the IoU was computed between the bounding boxes and the maps binarized according to the area of the bounding boxes. Regarding the HR, this metric is computed by checking whether the highest value in the explainability map is located inside the corresponding bounding box.

The explainability maps and metrics were computed not only for the final layer, but also for intermediate layers. It is worth noting that for these intermediate layers in which the feature map resolution was higher, the extracted explainability maps were first downsampled to the size of the last feature map and only then upsampled to match the input size, thus discarding the impact of different feature map resolutions on the results.

2.5 Effective receptive field computation

The ERFs were computed using backpropagations from specific feature maps up to the input, similarly to past approaches [3, 6]. To perform these backpropagations and compute the ERFs, a gradient signal of 1 at the center coordinates across all channels of the feature maps, and 0 otherwise, was used. Equation 2 describes this process, where A_{ijk} are the center coordinates of a given feature map, C is the number of channels of the input, I_{xyc}, and N is the number of images in the test set.

$$ERF = \frac{1}{N}\sum_{n}\left|\frac{1}{C}\sum_{c}\frac{\partial A_{ijk}}{\partial I_{xyc}}\right| \tag{2}$$

3 Results

3.1 Receptive fields

We first computed the RFs for each model as shown in Table 1. Surprisingly, our results show that the RF for the EfficientNet-b0 model is infinite. This is caused by squeeze-and-excitation blocks present on the EfficientNet-b0 architecture, which condense the channel information into single values that are then processed and used as weights to recalibrate the feature maps. On the other hand, the DenseNet121 architecture has a theoretical RF that ranges from 1047 to 2071. This occurs due to the skip connections and concatenation operations that are present in the model, which leave certain channels unaltered and thus with smaller RFs. Regarding ResNet50, it presents the smallest RF from the three models but still large when compared to typical input sizes.

Table 1: Receptive fields for the different models.

EfficientNet-b0	DenseNet121	ResNet50
Infinite	1047-2071	427

3.2 Explainability evaluation

The next step consisted in producing explanations for each model using Grad-CAM and computing metrics to evaluate the overlap with the pathology masks. We performed this not only for the last layer, but also for intermediate layers where the RF is smaller, as shown in Table 2. Interestingly, we can see a significant improvement for intermediate layers in terms of IoU and HR regardless of feature map resolution, since we downsample attributions from intermediate layers before the upsampling, even though these features are *a priori* less meaningful. We also show some examples for the Aortic enlargement class in Figure 1 where the explanations for intermediate layers are more accurately targeting the aorta.

Table 2: Mean IoU and HR values for different layers and models.

Model	Layer	IoU	HR
EfficientNet-b0	features[-1]	0.21 ± 0.05	0.35 ± 0.08
	features[-2]	0.20 ± 0.04	0.35 ± 0.06
	features[-3]	**0.23** ± **0.03**	**0.44** ± **0.04**
	features[-4]	0.13 ± 0.02	0.32 ± 0.03
DenseNet121	denseblock4	0.18 ± 0.02	0.21 ± 0.04
	transition3	**0.27** ± **0.02**	**0.51** ± **0.05**
	denseblock3	0.26 ± 0.02	0.48 ± 0.05
	transition2	0.14 ± 0.01	0.30 ± 0.03
ResNet50	layer4	0.14 ± 0.02	0.30 ± 0.03
	layer3	**0.20** ± **0.03**	**0.43** ± **0.04**
	layer2	0.07 ± 0.01	0.17 ± 0.04

3.3 Effective receptive fields

To complement this analysis, we computed and compared the ERFs for the last and best intermediate layers (Figure 2). As expected, the ERF is much more concentrated in intermediate layers which can lead to better explainability maps. There is, however, an apparent trade-off between the ERF concentration and feature importance. Therefore, going back one or two blocks can result in a significant improvement in the faithfulness of explanations, but going too far back is no longer beneficial.

4 Conclusion

In this work, we studied a possible limitation of the Grad-CAM method which may also apply to other CAM-based methods. As shown here, RFs of common architectures can be much larger than typical input sizes, meaning that any part of the image may influence a given feature. This can lead to the obvious problem that features and gradients for the last layer, and consequently attributions, do not have any spatial correspondence relative to the input. We complemented the RF measures with the computation of the ERF and a link was found between its concentration and the performance of Grad-CAM. More importantly, we demonstrated this phenomenon while excluding the effect of different feature map resolutions. Ultimately, these results can lead to a better understanding of why this method can produce inaccurate explanations and pave the way for its improvement/replacement.

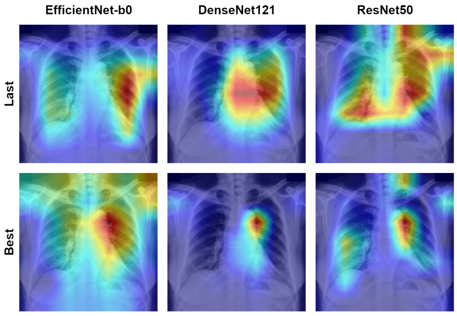

Figure 1: Explanations from the last layer and the best one in terms of IoU and HR. The image used corresponds to an Aortic enlargement case.

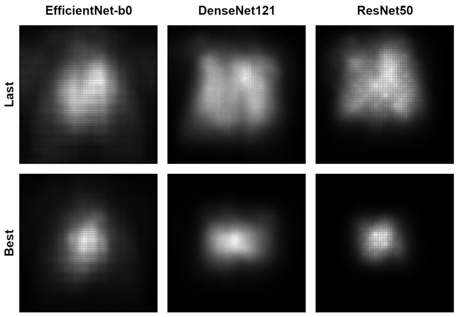

Figure 2: ERFs from the last and best layers.

Acknowledgements

This work was funded by the ERDF - European Regional Development Fund, through the Programa Operacional Regional do Norte (NORTE 2020) and by National Funds through the FCT - Portuguese Foundation for Science and Technology, I.P. within the scope of the CMU Portugal Program (NORTE-01-0247-FEDER-045905) and LA/P/0063/2020. João Pedrosa is supported by Fundação para a Ciência e Tecnologia (FCT) through application reference 2022.06138.CEECIND.

References

[1] André Araujo et al. Computing Receptive Fields of Convolutional Neural Networks. *Distill*, 2019.

[2] DungNB et al. VinBigData Chest X-ray Abnormalities Detection. *Kaggle*, 2020.

[3] Wenjie Luo et al. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[4] A. Saporta et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat Mach Intell*, 4:867–878, 2022. ISSN 0167-8655.

[5] R. R. Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[6] Pengfei Xia et al. On the receptive field misalignment in CAM-based visual explanations. *Pattern Recognition Letters*, 152:275–282, 2021. ISSN 0167-8655.