Studying Tools For AI Explainability On Convolutional Networks

C. Hurter[1], N. Couellan[2]

[1] ENAC, DATAVIZ, Université de Toulouse, F-31055 Toulouse, France.
[2] ENAC, OPTIM, Université de Toulouse, F-31055 Toulouse, France.

Context

Artificial Intelligence is used in a wide range of applications in modern systems, using an increasingly important volume of data of several types (continuous, discrete). The display of power in some methods (Deep Neural Networks, Random Forest...) allows us to solve problems that were unaccessible before, but it also comes at the cost of a bigger opacity. In some domains (medical, aeronautics...) where diagnosis is the center of the equation, this is unacceptable. Indeed, understanding **why** an AI-based system took a decision that led to an issue is more important than knowing **what** decision it took in those fields. Opening those commonly named "black-box" models to extract some meaningful - and more importantly, human interpretable - visualization is our final objective for this PhD.

Objectives of PhD

This PhD thesis on algorithm transparency is structured around 3 main approaches: descriptive, predictive and prescriptive.

- **descriptive**: consists in retracing an output back to the input in a clear way. Answers the question "How did my system predict this?"
- **predictive**: consists in learning to anticipate the results of my system. Widely used for predictive maintenance of planes, hence ensuring their total flight time is as long as possible.
- **prescriptive**: consists in fixing currently-established systems on-the-fly by analyzing them and detecting bugs in predictions. This is without a doubt the most complex one to achieve.

Overfitting issue

Overfitting is a well-known issue in modern machine learning architectures. An "overfitting" model means it is too specialized on training data, and didn't learn to generalize. It is generally confirmed manually by viewing the accuracy and validation curves per epoch - accuracy continues to increase while validation accuracy stays equal or starts to decrease. We believe that we can find a relation of the weights in a model between consecutive epochs to confirm overfitting with another technique.

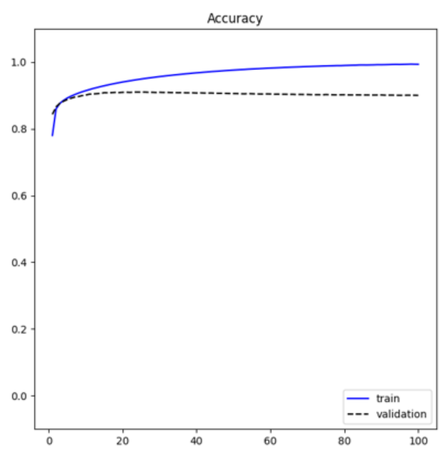

Figure 1: Usual way to assess overfitting is by comparing the curves.

Convnet filters

Most modern architectures for convnets use convolutional layers at the start, then end with densely-connected layers at the bottom. They are made like that because it is believed that the convolutional layers act as **parameters extractors** of the input image and project it in a new space composed of new axis learnt by the axis during training. These axis, called feature maps in the convolutional layers, can be visualized to try and understand what kind of feature the system learnt to recognize.

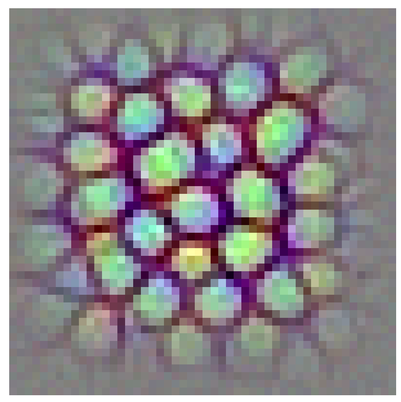

Figure 2: Convnet filter from a VGGNet trained on ImageNet.

Grad-CAM

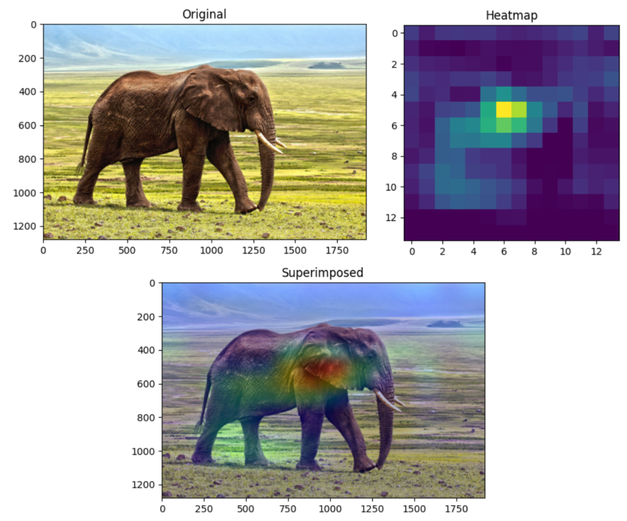

Figure 3: Grad-CAM on an african elephant.

It is becoming more important to understand how the systems we use to help our decision work in depth. Multiple gradient-based techniques to visualize the effect of a network over an input exist; Vanilla backpropagation [1], Guided backpropagation [2], Gradient o Input [3][4], SmoothGrad [5], Integrated Gradients [6], Score-CAM [7] and many others. Among them, Grad-CAM [8][9] is probably one of the most famous ones for assessing the features from the input images the system learned.

References

[1] Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013)
[2] Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net." arXiv preprint arXiv:1412.6806 (2014)
[3] Shrikumar, Avanti, et al. "Not just a black box: Learning important features through propagating activation differences." arXiv preprint arXiv:1605.01713 (2016).
[4] Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." International Conference on Machine Learning. PMLR, 2017.
[5] Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." arXiv preprint arXiv:1706.03825 (2017).
[6] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International Conference on Machine Learning. PMLR, 2017.
[7] Wang, Haofan, et al. "Score-CAM: Score-weighted visual explanations for convolutional neural networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.
[8] Selvaraju, Ramprasaath R., et al. "Grad-CAM: Why did you say that?." arXiv preprint arXiv:1611.07450 (2016).
[9] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.