Saliency map decoder for Convolutional Neural Networks Interpretability

Environment: QARMA (machine learning) Team at Laboratoire d'Informatique et Systèmes (LIS)
Location: Ecole Centrale Marseille (ECM), Technopôle de Château-Gombert, Marseille
Supervisors: R. Sicre (LIS – ECM), S. Ayache (LIS - Polytech), H. Zhang (LIS-ECM)
Salary: legal minimum
Keywords: computer vision, deep learning, interpretability
Contact: ronan.sicre@lis-lab.fr

In the last decade, machine learning, especially deep learning, received a lot of attention from the research community and industry. Machine learning systems are now used in numerous applications to help people in their daily life.
In the field of computer vision, Deep Neural Networks (DNN) architecture have developed rapidly, not only achieving remarkable performance on basic tasks, such as classification and object detection, but also addressing more diverse and complex problems. With all these improvements however some limitations remain. One main limitation is the interpretability of DNNs, which can help researchers better understand the learning capabilities and bias of their models. In this project, we aim to provide a visualization explanation to interpret the decision of a Convolutional Neural Networks (CNN), for a given input.

Specifically, we follow the works derived from Class Activation Maps (CAM) [1], and gradCAM [2] to generate saliency maps, highlighting the area of an image that contributes the most to the decision of a CNN. CAM revisits the global average pooling layer and sheds light on the localization capabilities of CNNs. Then, methods such as gradCAM [2], gradCAM++ [3], and integratedCAM [4] use gradients to rate the importance of image regions, while ScoreCAM [5] computes a combination of activation maps coming from different layers.

During this internship, we are interested in building a new architecture that learns to build a heat map either from a pre-trained network or when learning to classify images.

- A decoding stream will be added to a common architecture. This stream will take the feature maps of a specific convolutional layer as input, will apply transpose convolution to upsample this map to obtain a saliency image of the same size as the input.
- Several losses will be studied taking inspiration from ScoreCAM [5]. The loss will compare the output logits of the network of the standard image and the image combined with the saliency map. We could later study adversarial loss to improve the saliency map quality.
- Several regularizations will be further evaluated to constrain the saliency map to have low norm, sparse and compact activations, etc.

[1] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In CVPR.
[2] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* .
[3] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)* IEEE.
[4] Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In ICML. PMLR.
[5] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., ... & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.