On interpretable CNNs representations for visual scene understanding

Context

Convolutional Neural Networks (CNNs) achieve state-of-the-art performance in various computer vision and machine learning tasks, including segmentation, detection and classification [1, 2].

CNNs are usually trained in an end-to-end manner using a lot of annotated data. However, this end-to-end training strategy which might be considered as an important advantage, makes CNNs representations appear as "black-boxes". Indeed, the CNN extracts, from its hidden layers, powerful representations which effectively infer the final decision, yet these representations are difficult to explain or interpret and their meaning is difficult to grasp as well.

In recent years, several works have attempted to understand the "magic" of CNNs towards a better interpretability of the learned representations. For example, in [3] a method is introduced to visualize the filters in a CNN through computation of the gradients of the score of a given unit w.r.t. the input image. This gradient information is used to visualize the intermediate visual patterns encoded in the network layers. Such method is generalized in [4]. Another approach in [5] uses up-convolutions to invert CNN feature maps into images.

A different approach is to study distributions in the feature space for diagnosing CNNs representations, for example by computing adversarial samples [6]. More recently, a method based on an explanatory graph that encodes the semantic hierarchy hidden inside a CNN, has been proposed in [7].

Objectives

Clearly, being able to interpret or to explain the decision of a CNN would be beneficial in many sensitive areas such as security, medical diagnosis or autonomous driving.

In the project, we wish to investigate on this area with a focus on visual scene understanding for autonomous navigation.

1. The first goal of this internship is to study the different approaches for making CNNs representations more interpretable, i.e. provide an overview of the methods and a comparison. This is important as there many different ideas in the literature.
2. The second goal of the internship is an in-depth study of the explanatory graph method proposed in [6], and its application in a visual scene understanding context.

Perspectives

This work could be pursued as a PhD project if the selected candidate provides satisfactory results.

Application and needed skills

Applications

Motivation letter and transcripts are to be sent to Pr D. Sidibé : drodesire.sidibie@univ-evry.fr (https://sites.google.com/view/dsidibe/)

Skills

- Computer vision, machine learning
- Programming, Python, C++
- Familiarity with deep learning framework is a plus

References

[1] LeCun et al. "Deep learning", Nature, 521, 434-444, 2015
[2] Goodfellow et al. "Deep learning", MIT Press, 2016
[3] Zeiler and Fergus, "Visualizing and understanding convolutional networks", in ECCV 2014
[4] Selvaraju et al. "Grad-cam: visual explanation from deep networks via gradient-based loc", in ICCV 2017.
[5] Dosovitskiy and Brox, "Inverting visual representations with convolutional networks", in CVPR 2016
[6] Koh and Liang, "Understanding black-box predictions via influence functions", in ICML 2017
[7] Zhang et al. "Interpreting CNN knowledge via explanatory graph", in AAAI 2018.