Polytech network form for PhD Research Grants from the China Scholarship Council

This document describes the PhD subject and supervisor proposed by the French Polytech network of 14 university engineering schools. Please contact the PhD supervisor by email or Skype for further information regarding your application.

Supervisor information

Family name	Pellerin
First name	Denis
Email	denis.pellerin@gipsa-lab.grenoble-inp.fr
Web reference	http://www.gipsa-lab.grenoble-inp.fr/~denis.pellerin/
Lab name	Gipsa-lab
Lab web site	http://www.gipsa-lab.fr/
Polytech name	Polytech' Grenoble
University name	Université Grenoble Alpes
Country	France

PhD information

Title	Deep explanation for Multimedia Indexing and Retrieval.
Main topics regards to CSC list (3 topics at maximum)	I-12. Discernement des modèles et systèmes intelligents Understanding models and intelligent systems

Required skills in science and engineering	Machine learning, mathematics, computer programming

Subject description (two pages maximum)

Deep convolutional neural network and recurrent neural networks have been very successful for a number of tasks, ranging from image classification to speech transcription or language translation. For all of these tasks, one or more neural network are trained by gradient descent according to an objective loss function and to large annotated sets of training data. The gradient descent, associated with a number of regularization techniques, usually leads to models that perform statistically very well on previously unseen test data, thereby showing very good generalization capabilities.

There are however a number of limitations or drawbacks with these approaches, which are linked to the fact that the good performance they provide is only statistical and to the fact that we have little, if any, clue about how the network reach its conclusions and why and where it fails when it does. It has been observed that statistically very good networks makes from time to time mistakes that are really obvious for us. It has also been shown that it is possible to manipulate real images in a way that is imperceptible to us humans but are enough for fooling a state-of-the-art classification network making them predict a visually very different class than the actual image one. Such failures, occurring either by “misfortune” of due to malicious actions, are especially undesirable when the output of the system may have a negative effect on people’s life or well-being.

The objective of the proposed PhD is to study methods for making the inferences made by neural networks as understandable as possible by humans. Traditionally, in machine learning, a trade-off has to be made between prediction accuracy and prediction explanability. This is particularly true in the case of neural networks for which the accuracy is very high while the explanability is very low. The objective here will be to significantly increase the explanability while degrading as little as possible the accuracy. The approaches for achieving this are open. One possibility is to modify the neural network architecture and operation for forcing it to identify meaningful elements in its hidden nodes. Another possibility is to complement it with another neural network that will extract similar information also from its hidden nodes.

The final system should be able to automatically generate “raw” annotations or captions for a given image or for a given video sample with state-of-the-art accuracy. Additionally, the system

should be able to generate a set of visual elements associated to a text jointly explaining the raw predictions, ideally all of it with a single end-to-end training. The main targeted domain is automatic video and image indexing, captioning and retrieval.

References:

Pierre Stock and Moustapha Cisse, ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases, ECCV 2018, http://openaccess.thecvf.com/content_ECCV_2018/papers/Pierre_Stock_ConvNets_and_ImageNet_ECCV_2018_paper.pdf

DARPA Explainable Artificial Intelligence (XAI) program: https://www.darpa.mil/program/explainable-artificial-intelligence