2019-01534 - PhD Position F/M [CORDIS2019-ZENITH] PhD thesis on uncertainty and scalability issues related to the deep learning of a very large number of categories

Type de contrat : CDD de la fonction publique
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Doctorant

A propos du centre ou de la direction fonctionnelle
The Inria Sophia Antipolis - Méditerranée center counts 37 research teams and 9 support departments. The center's staff (about 600 people including 400 Inria employees) is composed of scientists of different nationalities (250 foreigners of 50 nationalities), engineers, technicians and administrators. 1/3 of the staff are civil servants, the others are contractual. The majority of the research teams at the center are located in Sophia Antipolis and Nice in the Alpes-Maritimes. Six teams are based in Montpellier and a team is hosted by the computer science department of the University of Bologna in Italy. The Center is a member of the University and Institution Community (ComUE) "Université Côte d'Azur (UCA)".

Contexte et buts du poste
This PhD thesis will take place within the Inria ZENITH team (https://team.inria.fr/zenith/) which has a strong expertise in data science and an international reputation. It is headed by Patrick Valduriez who has been inducted ACM fellow in 2014. ZENITH's thematic is focused on scientific data and the related challenges: complexity (uncertain, multi-scale data with lots of dimensions), scale (big data, big applications), heterogeneity (in particular, data semantics heterogeneity). The team is working closely with application partners in the biodiversity and cultural heritage domains with a key role in impactful initiatives such as PIGmNT. The PhD will be supervised by Alexis Joly and Antoine Likitais, both permanent researchers with international reputation specializing in machine learning applied to images and sounds respectively.

Mission confiée
In this thesis we will focus on large-scale classification problems typically involving tens to hundreds of thousands of classes, in the case of long-tail distributions (imbalanced classes). These problems are encountered in many contexts: there is a high diversity of objects or concepts to recognize. For instance, biodiversity data is heavily imbalanced, as some species are much more abundant and easier to observe than the vast majority of other species. The same happens for heritage data or wildlife audio/video recordings, where many concepts of interest correspond to only a fraction of the observations.

It has empirically been shown in several studies that deep learning models (convolutional neural network in particular) are more robust to such strong imbalance than other machine algorithms, who perform well on average but wrongly recognize many objects belonging to the long tail of the data distribution. However, the long tail setting remains very challenging today and performance remains extremely heterogeneous from one class to another. We may identify two core avenues for research. First, transfer learning and uncertainty issues are not well understood in the long-tail case. Second, training time and memory/space requirements remain prohibitively expensive at scale. The recruited researcher would investigate both those directions, from a theoretical point of view and would apply his findings in biodiversity and heritage preservation research [2].

Principales activités
Challenges
### Transfer learning and uncertainty

Typically, we distinguish two cases for uncertainty. The first one is "intrinsic ambiguity" and corresponds to the case where an object belongs to several possible classes but with a high confidence. The second one is "model uncertainty" and corresponds to the case where the prediction is globally uncertain because of a lack of training data or a bad generalization of the model. Most current predictive models (such as the widely used multinomial logistic regression) do not make any difference between those two sources for uncertainty [1].

One core duty of the recruited researcher is to build on probabilistic (4) and information theoretic paradigms [5, 6] to disentangle between intrinsic ambiguity and model uncertainty in the long-tail setting.

### Computational load

Training a convolutional neural network from scratch on hundreds of thousands of classes on training data composed of tens of million of images would take up to several months on a single node equipped with four recent GPUs. Moreover, selecting the best performing architecture and optimize the hyper-parameters often requires tens of such networks. Such a cost is prohibitive for most applications. Of course, a first natural route for improvement would be a better engineering of the infrastructures, software frameworks, resource management etc.

However, the second research mission of the project is to circumvent computational load from a more fundamental perspective and focus on transfer learning and hierarchical models. This research direction is inspired from the recent successes of transfer learning for large scale prediction [8] and also the approximate softmax loss function introduced in [2], which allows a significant reduction of the computation costs by introducing some hierarchy, although it still doesn't address the memory cost of the model, which is a critical issue in practice. Improvement will first concern automatic clustering of the target classes [8], and a better understanding of the impact of such hierarchical models in terms of prediction accuracy.


Avantages

- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- Flexible working hours
- Sports facilities

Rémunération

Duration: 36 months
Location: Sophia Antipolis, France
Gross salary per month: 1982€ brut per month (year 1 & 2) and 2085€ brut/month (year 3)