Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2017

Axe Sophi@Stic :	Environnement|
Titre du sujet :	Deep learning for automatic zooplankton classification
Mention de thèse :	ATSI
HDR Directeur de thèse inscrit à l'ED STIC :	Eric Debreuve

Co-encadrant de thèse éventuel :

Nom :	Irisson
Prénom :	Jean-Olivier
Email :	irisson@normalesup.org
Téléphone :	04 93 76 38 04

Email de contact pour ce sujet :	eric.debreuve@cnrs.fr
Laboratoire d'accueil :	I3S/INRIA

Description du sujet :

In early 2015, the National Data Science Bowl, a Kaggle-hosted [2] competition, asked participants to classify images of plankton (small organisms drifting in the oceans) into taxonomic groups [3]. The top teams relied on deep learning using convolutional neural networks, suddenly popularizing these techniques in the field of biological oceanography. Following this competition, I3S and LOV explored a few solutions to classify the massive amounts of data collected by LOV (20+ million images of organisms, growing by several millions per year now), which started the present project.

The goal of the project is to develop an automatic zooplankton classification method based on potentially several deep convolutional neural networks used in an ensemble learning framework. This branch of statistical learning is accompanied by a good dose of hype, but our collaboration

with environmental biologists, for whom classification is a means to gather structured data rather than an end in itself, imposes to keep our feet on the ground and translates into the following constraints and goals.

(1) Compared to some solutions proposed in the Kaggle competition, which relied on ensembles of over a hundred deep networks, we want to keep the overall complexity down, for the solution to be operational. This will require us to study thoroughly the developed architectures and understand their behavior (instead of just considering the classification procedure as a black box). This will be essential to improve the architectures, both in terms of pure performance and in terms of performance/complexity ratio. Finding ways to reduce the complexity of a deep network is an active field of research. Our objective will be to determine strategies to perform such a task but jointly for a set of networks. Therefore, the proposed project will consist in designing an ensemble of deep networks with optimal performances, and then to study how to preserve most of this optimality while deriving fewer, potentially smaller networks.

(2) Again in the Kaggle competition, the participating teams were experts in statistical learning but had virtually no knowledge of the application field. This led them to make mistakes, like rescaling all images to the same size, as usual for neural networks, while the information about size is of significant importance to discriminate some species. Similarly, they considered all classes/species as equivalent while they can be ordered along a taxonomic tree, hence opening the possibility to first classify into large, easy to discriminate groups and then develop custom, more efficient models to classify into finer taxonomic sub-groups. On the contrary, this project will be conducted in collaboration with specialists of the application domain. It means that we will have the opportunity to study how to tailor network architectures according to some a priori knowledge on the data (such as taxonomy). An additional aspect of this context will be the combination of the features learned by the networks with biologically-based ones (such as size but also time of the day, season...), which is also an active field of research.

(3) Depending on the application, it is clear that classification mistakes do not have the same impact. In a picture labeling task, using a bad classifier would simply be a waste of computing time and money. In medicine however, it would be unconceivable to include a bad classifier in a diagnosis procedure. Even though it might not be as stringent as in this latter case, the reliability of the classifier is also essential in the targeted application since the classification results are to be used in environmental studies with a potentially high societal impact (health of the oceans, climate studies...). Reliability is related to the notion of "adversarial example" that arose three years ago: an image very similar to a correctly classified one that is classified in a wrong category with high confidence. It is possible to make a deep network somewhat more robust to some adversarial examples, typically by including them in the learning set. Nevertheless, we plan to instead try to understand better the underlying phenomena in order to design robust architectures from the ground up.

[1] http://www.lov.obs-vlfr.fr/en/index.html
[2] http://www.kaggle.com/
[3] http://www.kaggle.com/c/datasciencebowl

URL :

http://www.i3s.unice.fr/~debreuve/edstic2017.pdf

English version: