Internship: Wheat grain classification from Hyper-Spectral Images for agro-ecology

Key words: CNN - classification - large number of channels - real data - agro-ecology

Context:

In order to reduce the use of synthetic pesticides, or even eliminate it altogether, the transition to agroecology involves a spatio-temporal diversification of agricultural populations. Among the possible strategies, **mixing multiple varieties of the same species** is a technique that is increasingly being used by farmers. Its benefits in controling epidemies of fungal pathogens are well-documented in wheat, but the heterogeneity can also exacerbate certain forms of resource competition. To understand the underlying mechanisms, we are particularly interested in **comparing the yields of numerous wheat varieties grown in monoculture and in various mixtures**. However, at the time of harvest, we currently only have information about the total yield per mixture. The **information about the proportion of grains harvested by variety within each mixture is missing**, a proportion that is not necessarily the same as at the time of sowing. Such changes in proportions between sowing and harvest would inform us about the ability of a particular variety to dominate its neighbors or, conversely, to promote complementarity. This is interesting not only from a fundamental perspective in terms of the evolution of populations in a fluctuating environment but also from an applied perspective for a range of agricultural and agri-food stakeholders (breeders, cooperatives, sorters, millers, etc.).

The PerfoMix Project:

As part of the PerfoMix project (coordinated by T. Flutre), a field trial was conducted on the agricultural plots of the UMR GQE (INRAE Ile-de-France) on the "plateau de Saclay". Eight varieties of bread wheat were cultivated in monoculture and in mixtures composed of two to four varieties. The field trial was carried out in 2019-2020 and was repeated in 2020-2021.

The grains from each plot were analyzed at the UMR AGAP (INRAE Montpellier) using **hyperspectral imaging** (HSI). This technique provides a 2D image with, for each pixel, the reflectance level at several hundred wavelengths in the visible and near-infrared, known as a "spectrum". **Concretely, instead of the three RGB channels of a classical image, the HS images have more than 200 channels.**

Data availability:

All the images have already been collected and are stored on a server at the AGAP unit in Montpellier. We have images for about 150 wheat grains per variety with known orientation (wheat grains are asymmetrical, with one side having a furrow and not the other). We have images for about 1600 wheat grains of a single variety (known label) for training, and this for all the K=8 classes (varieties) per year. We also have unlabelled data for 40k grains in the first year and 70k grains in the second year (could be used for further self-supervised learning).

Preliminary results:

From grains harvested in monoculture plots, machine learning models (such as PLS, SVM, RF, kNN) were trained to recognize each variety based on spectra averaged over all pixels per grain (it is easy to segment the grain from the background with conventional techniques). Then, mixtures of the same grains were created in silico to evaluate prediction accuracy by cross-validation. The prediction results are promising (M2 internship by A. Belny from AgroParisTech in 2021). There is also some **litterature** on the subject, with which the internship student should get familiar: [JMAVM21, LWAW18, NZF+19].

Internship goal:

The goal is to use the three dimensions of the data, not just the spectrum averaged over all pixels, to predict the variety of a given grain. This should be doable especially with deep learning (CNNs).

A first **sanity check will be to train a furrow detection model**. This classification task can be easily done by visual inspection (in other words, it can be done by humans with 100%

accuracy). Images have been specifically collected to study this.

Then, the goal will be to **train a CNN to recognize the variety of a grain from its HS image**, regardless of the grain orientation (the side it fell on). This may be done by building a CNN from scratch or using foundation models (pre-trained or not), such as VGG-16 or more recent models.

References

[JMAVM21] Shima Javanmardi, Seyed-Hassan Miraei Ashtiani, Fons J. Verbeek, and Alex Martynenko. Computer-vision classification of corn seed varieties using deep convolutional neural network. *Journal of Stored Products Research*, 92:101800, May 2021.

[LWAW18] Wen-Xuan Liao, Xuan-Yu Wang, Dong An, and Yao-Guang Wei. Hyperspectral Imaging Technology and Transfer Learning Utilized in Identification Haploid Maize Seeds, May 2018. arXiv:1805.11784 [cs].

[NZF+19] Pengcheng Nie, Jinnuo Zhang, Xuping Feng, Chenliang Yu, and Yong He. Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning. *Sensors and Actuators B: Chemical*, 296:126630, October 2019.

Learned Skills:

- learn about agro-ecology, esp. about varietal mixtures
- understand, develop and deploy CNNs
- developped pyTorch practice
- using a job scheduler (slurm) on a computer cluster (a GPU farm)

Expected abilities:

- good python skills
- good basic theoretical knowledge of CNNs, and about ML in general
- some knowledge of pyTorch
- interest for agro-ecological problems is a plus
- already knowing about slurm is a plus

Duration: The preferred duration would be of at least 4 months.

Labs: (Co-supervision between 2 labs)
LISN, Université Paris-Saclay
IDEEV, Université Paris-Saclay

Teams:
A & O (Algorithmes et Optimisation), INRIA team: TAU
DEAP (Diversité, Évolution et Adaptation des Populations) inside gQE (génétique Quantitative et Évolution)

Advisors:

François Landes (francois.landes@inria.fr), Maitre de Conférences (CS)
Timothée Flutre (timothee.flutre@inrae.fr), chargé de Recherche INRAe
Also in collaboration with Martin Ecarnot martin.ecarnot@inrae.fr, Ingéneur d'études INRAe

Head of Lab: Sophie Rosset (LISN)

Location: Gif-sur-Yvette (Plateau de Saclay), Batiment 660 (Digiteo building)
Gif-sur-Yvette (Plateau de Saclay), 12 route 128 (IDEEV building)

Website: http://lptms.u-psud.fr/francois-landes/internships-phd/