Internship Proposal: Transfer Learning for Multi-Output Regression with Few-Shot Data

***Abstract*—This internship focuses on developing a transfer learning approach for multi-output prediction, particularly in scenarios with small datasets (few-shot learning). The objective is to improve predictive performances by incorporating similarity information between tasks, enabling the model to leverage shared features across related predictions.**

***Index Terms*—Deep Learning, Few-Shot Learning, Transfer Learning, Multivariate Regression**

A. Context

- **Supervisors & projects members:** Philippe Caillou, Florian Yger, Olivier Allais, Julia Mink, Cyriaque Rousselot
- **Laboratory** TAU (Inria), LISN Université Paris Saclay, INRAE, INSA Rouen Normandie

This internship will take place at LISN Laboratory at Université Paris Saclay. This internship is part of the HORAPEST project (Fig. 1), which evaluates pesticide exposure and its impact on children health in France. The focus is now on estimating airborne concentrations for each phytosanitary molecule with high accuracy. The application involves predicting molecular concentrations in the air using campaign data of air quality in France. We have access to few data points of sensors, temporal correlations and common characteristics between molecules of interest. While some molecular concentrations are predicted with high accuracy, others remain challenging to predict. By embedding the relationships among molecules into the model, the goal is to predict more effectively the pesticide exposure of patients.

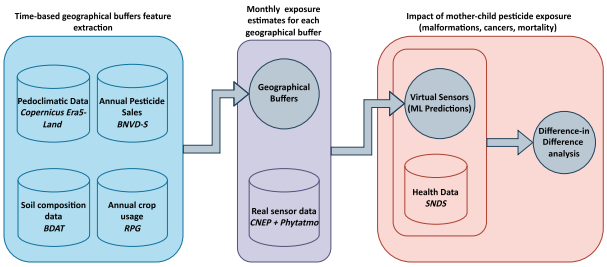

Fig. 1: HORAPEST project methodology

B. Related Works

Predicting multiple outputs with limited data is a significant challenge in machine learning. Multitask Learning (MTL) has been a foundational approach, where models are trained on multiple related tasks simultaneously to improve generalization performance [1].

In scenarios with scarce data, Few-Shot Learning techniques have been developed to enable models to learn from a limited number of examples [2]. Transfer Learning has been widely used to address data scarcity by transferring knowledge from related tasks or domains with abundant data to those with limited data [3]. In the context of regression tasks, transfer learning can help improve predictions by leveraging patterns learned from similar tasks. In the domain of environmental exposure assessment, predictive models have been employed to estimate pollutant concentrations using limited sensor data [4]. Integrating molecular similarity and temporal correlations has the potential to improve the accuracy of these models, particularly when predicting concentrations of less-studied substances.

C. Tasks and Responsibilities

1) Literature Review
   - Explore state-of-the-art methods in multitask learning, transfer learning, and few-shot learning.
   - Understand existing toy datasets and synthetic data generation techniques for analyzing the predictors.
2) Model Development
   - Develop predictors leveraging multitask learning frameworks.
   - Compare independent predictors against joint predictors trained with multi-task setups.
   - Assess uncertainty in transferring predictions from one substance to another.
   - Integrate techniques such as LoRA [5] to adapt predictors to specific substance.
3) Integration and Evaluation
   - Test the method on real-world datasets.
   - Incorporate the improved predictor into the project's pipeline.
4) Potential Outcome
   - Contribution to scientific literature through conference publications.
   - Opportunities for long-term collaboration.

a) *Skills Required:*

- Knowledge of machine learning, deep learning, Python and Pytorch.
- Ability to process and analyze datasets, including generated or synthetic data.
- Excellent algorithmic skills.
- Autonomy and curiosity.

D. Internship Calendar

Months 1–2

- Develop familiarity on multitask learning and transfer learning literature, oriented toward few-shot situations.
- Develop familiarity with our lab computing resources.
- Build and test initial predictors.

Months 3–4 Depending on first months outputs:

- Implement techniques to exploit covariance matrix geometry.
- Begin integrating similarity information to enhance predictor performance.
- Begin leveraging LoRA or related approaches for task adaptation.

Months 5–6

- Finalize and evaluate the integrated predictor on real-world datasets.
- Prepare results for conference submissions.
- Collaborate on potential integration into the production pipeline.

E. Benefits of the Internship

- Work on a real-world problem with tangible applications.
- Gain hands-on experience in advanced machine learning techniques in a specialized team.
- Opportunity to publish and collaborate on research.

F. Application Process

Interested candidates are encouraged to apply with a CV and a brief statement of motivation sent to cyriaque.rousselot(at)inria.fr with the mail object **[Internship application]**. Applications will be reviewed on a rolling basis with interviews until the position is filled.

References

[1] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.

[2] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] K. de Hoogh *et al.*, "Development of West-European PM2. 5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data," *Environmental research*, vol. 151, pp. 1–10, 2016.

[5] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2106.09685