PhD Position - University of Orléans

Learning code embeddings : application to Computer Science Education

Contact : Guillaume Cleuziou

guillaume.cleuziou@univ-orleans.fr

Supervisors : Guillaume Cleuziou and Matthieu Exbrayat

We offer a PhD position at the LIFO (Laboratoire d'Informatique Fondamentale d'Orléans), University of Orléans, France. It will start in October 2022 for 3 years.

Related to the research domain of **Artificial Intelligence**, the PhD will include the following research fields : Machine Learning, Deep Learning, Representation Learning, Educational Data Mining.

PhD position requirements

- The PhD position is fully funded by the University of Orléans with a monthly gross salary of 1,975€.
- It will be conducted at LIFO, University of Orléans, France.
- It will start in October 2022.
- The deadline to apply is June 20th, 2022.
- Interviews with pre-selected candidates will take place in the end of June 2022.

Required technical skills

- Proficiency (speaking and writing) in French or in English.
- Strong skills in programming languages such as Java and Python.
- Experience in machine learning, data mining and deep learning. Interest in Educational data analysis is appreciated.

Required documents to apply

The complete application consists of the documents below, which must be sent as a single PDF file to **Guillaume Cleuziou** (guillaume.cleuziou@univ-orleans.fr) LIFO, University of Orléans.

- CV
- One-page cover letter (clearly indicating available starting date as well as relevent qualifications, experience and motivation)
- University certificates and transcripts (both B.Sc and M.Sc degrees marks)
- Contact details of up to three referees
- Possibly an English language certificate and a list of publications. **All documents should be in English or in French.**

PhD Subject Description

Improving the pedagogical efficiency of programming training platforms is a fast-growing problem that requires the construction of fine-grained and exploitable representations of learners' programs. In this PhD thesis, we are interested in learning representations (or embeddings) of programs for pedagogical purposes.

Two main strategies for learning program embeddings have been proposed so far: approaches based on the observation of program execution results (Wang et al., 2018) and those based on the syntactic analysis of programs (Alon, 2019). In this thesis, we will consider an original approach at the intersection of these two strategies based on a representation of programs via an abstract execution sequence and thus aiming to jointly take advantage of both functional and syntactic descriptions of programs (Cleuziou&Flouvat, 2021).

In order to carry out this work, it will be necessary to draw inspiration from models developed for text mining purpose and to study their adaptability for computer programs. Given the specificities of this type of data (restricted vocabulary, importance of 'words' order, etc.), it will be interesting to consider either simple (e.g. word2vec), recurrent (e.g. LSTM, GRU), convolutive or Transformer-like (e.g. BERT) neural models.

The fundamental part of the thesis will be backed up by applicative concerns on educational data, aiming at the development of 'Augmented Pedagogy' environments for teachers. The aim will be to identify support tasks on which the teacher could be assisted (e.g. detection of learner 'drop-outs', suggestion of feedbacks, etc.) and to implement them in a Research & Development process integrated with the digital tools used by the institution's training courses.

Bibliography

Alon, U., M. Zilberstein, O. Levy, and E. Yahav (2019). code2vec : Learning distributed representations of code. Proceedings of the ACM on Programming Languages 3(POPL), 1–29.

Cleuziou, G. and F. Flouvat (2021). Learning student program embeddings using abstract execution traces, In International Conference on Educational Data Mining (EDM'2021).

Wang, K., R. Singh, and Z. Su (2018). Dynamic neural program embeddings for program repair. In International Conference on Learning Representations (ICLR’2018).