Internship Master 2 level

Handling missing values based on Deep learning and Attention Mechanism

Keywords: missing values, deep learning, attention mechanism

The issue of missing data has been largely overlooked, especially in the deep learning community. A variety of techniques to handle incomplete training sets was then proposed, (Yoon et al., 2018; Marek et al., 2018; Ivanov et al., 2019; Ghalebikesabi et al., 2021). Many researchers report that the easiest way for handling missing values is to complete all these missing data and thus propose imputation models. While others propose to simply delete rows with missing data (Tlamelo and al. 2021).

We want to investigate different strategies to adapt neural architectures for handling missing values. Here, we focus on attention mechanisms. More precisely, a deep latent variable model can be learned using an attention mechanism.

The attention mechanism is one of the recent advances in deep learning (Yun et al. 2020). An attention layer regularises data distribution to make deep learning models more focused on existing values of data related to the targets of the models, to compensate the lost information due to missing values.

We will focus in particular on the development of a new framework for handling missing values based on attention mechanisms. Many interactions between theory and applications are envisaged.

The studies carried out at LATMOS concerning the atmospheric water cycle on different spatial and temporal scales are based on data from ground-based observation networks (rain gauge networks) or remote sensing data (ground-based radar; satellite constellations). Depending on the datasets considered, missing data may be due to sensor failures (e.g. clogged rain gauges) or to particular meteorological situations that do not allow the measurement of certain characteristics (e.g. undefined polarimetric radar data). The last part will be done in collaboration with LATMOS laboratory.

The internship will take place in the DAVID Lab, University of Versailles (UVSQ campus)-Université Paris Saclay , over a period of 6 months (Laboratoire David, UFR des sciences, 45 avenue des Etats-Unis, 78035 Versailles)

Supervisor team

- Hanane Azzag, LIPN UMR 7030, Université Sorbonne Paris Nord
- Djallel Dilmi, LIPN UMR 7030, Université Sorbonne Paris Nord
- Mustapha Lebbah, DAVID Lab, UVSQ, Université de Paris Saclay
- Cécile Mallet, LATMOS Lab, IPSL, UVSQ, Université de Paris Saclay

Profile

End of engineering degree, M2 in data science, statistics and/or artificial intelligence. Good experience in programming, especially with the PyTorch/deeplearning4j framework.

To apply, simply attach:

- Your current Curriculum Vitae (CV),

- A portfolio of projects, if any,

- Your motivation for the position,

- Your latest university transcripts.

Send it all by email to mlcandidat@gmail.com with "[int-uvsq-23]"

References

- Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. GAIN: missing data imputation using generative adversarial nets. In International Conference on Machine Learning, 2018.
- Marek Smieja, Łukasz Struski, Jacek Tabor, Bartosz Zielinski, and Przemysław Spurek. Processing ´ of missing data by neural networks. In Advances in Neural Information Processing Systems, pp. 2719–2729, 2018.
- Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In International Conference on Learning Representations, 2019.
- Sahra Ghalebikesabi, Rob Cornish, Chris Holmes, and Luke Kelly. Deep generative pattern-set mixture models. In Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, pp. 3727–3735, 2021.
- Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, Sanjiv Kumar. O(n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers. 2020
- Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago & Oteng Tabona. A survey on missing data in machine learning. Journal of Big Data volume 8, Article number: 140 (2021).