Research internship: Mathematical analysis of Transformer-based models

Context. Transformers[1] are a type of neural network designed to efficiently process sequential data, particularly used in natural language processing (NLP). Introduced by Vaswani et al. in 2017, transformers rely primarily on the self-attention mechanism, which enables the model to focus its attention on different parts of an input sequence (such as a sentence) dynamically. The self-attention layer is central to the way transformers work: it allows each word to "linger" on other words in the sequence to understand its context, by calculating the relative importance of each word in the sentence. For example, in the sentence "The cat eats the mouse", the self-attention layer might determine that the word "cat" needs to pay more attention to "eats" and "mouse" to grasp the full meaning. Thanks to this ability to establish links between distant words in a sequence, transformers overcome the limitations of traditional recurrent neural networks (RNNs) for dealing with long-term relationships in text. This makes them extremely powerful for tasks such as machine translation, text summarization and language generation.

We propose to theoretically study toy models of attention layers, in particular for their ability to handle missing data in learning, but also to perform more classical task such as clustering.

Learning with missing data. Various issues can be addressed when learning with missing data: (i) parameter inference from incomplete data. This involves estimating the underlying model for complete data; (ii) prediction from incomplete data in both the training set and the test set; (iii) imputation. The aim here is to provide a complete data set.

We will focus on the last theme of imputation, and in particular on how network architectures including self-attention layers can be relevant in such a context. A survey of the literature on transformer training, will be conducted, highlighting the steps of masking and completion exercises. How are they carried out in practice to help the relevance of such architectures? Next, we will try to propose a simple model based on self-attention layers to complete data. And we will push its theoretical analysis further, for example when incomplete Gaussian data are provided as input to such architectures. This axis of research encompasses theoretical and numerical developments.

[1] corresponding to the 'T' in ChatGPT

PCA. Transformers are renowned for capturing the structure and information contained in data. We will then try to make the link between simple transformer architectures (e.g., with a single layer of attention) and the most classical unsupervised dimension reduction task, principal component analysis (PCA). To do so, one can consider the operation performed by one layer as acting on an input probability distribution (empirical or not). Our aim is to understand the transformation effected by such an architecture on a measure, when the model parameters are trained. As a starter, the input distribution could be considered as Gaussian.

Clustering. Clustering is an unsupervised machine learning technique used to group similar data points together based on their characteristics, without predefined labels. The goal is to identify natural patterns or structures within a dataset, so that data points within a cluster are more similar to each other than to those in other clusters. Common clustering methods include K-means, which partitions data into a specified number of clusters, or model-based clustering, relying on mixture models as priors. Clustering is widely applied in image processing or anomaly detection.

Inspired by the methodology developed in Marion et al. (2024), we will study the abilities of toy architectures based on self-attention layers, to perform clustering.

Practical information.

Supervision. Claire Boyer (LMO, Université Paris-Saclay). Depending on the area explored, discussions with collaborators are envisaged.

Required skills. Two M2 level trainees with high-level technical skills in mathematics/statistics/machine learning. Computer skills will be considered a plus, as each axis contain numerical illustrations. Autonomy and initiative are also desirable qualities. Applicants should send CV, transcripts of the last two years and the name of a referee to claire.boyer@universite-paris-saclay.fr

Location. The internship will take place at LMO (Université Paris-Saclay) in the Probabilities and Statistics team. This is a 6-month internship that can start at the beginning of April. For a PhD continuation, please specify it as soon as possible.

References.

Marion, P., Berthier, R., Biau, G., & Boyer, C. (2024). Attention layers provably solve single-location regression. arXiv preprint arXiv:2410.01537.

Vaswani, A. et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems.