PhD position, Paris, France

Transformers for etiological diagnosis in multi-modal medical data.

Contact : Nicolas Thome, nicolas.thome@sorbonne-universite.fr, Olivier Bernard, olivier.bernard@insa-lyon.fr

Location: Sorbonne Université, Pierre et Marie Curie Campus, 4 Place Jussieu, Paris, Fr

Candidate profile: Master degree in computer science or applied mathematics, Engineering school. Background and experience in machine learning and deep learning. Good technical skills in programming.

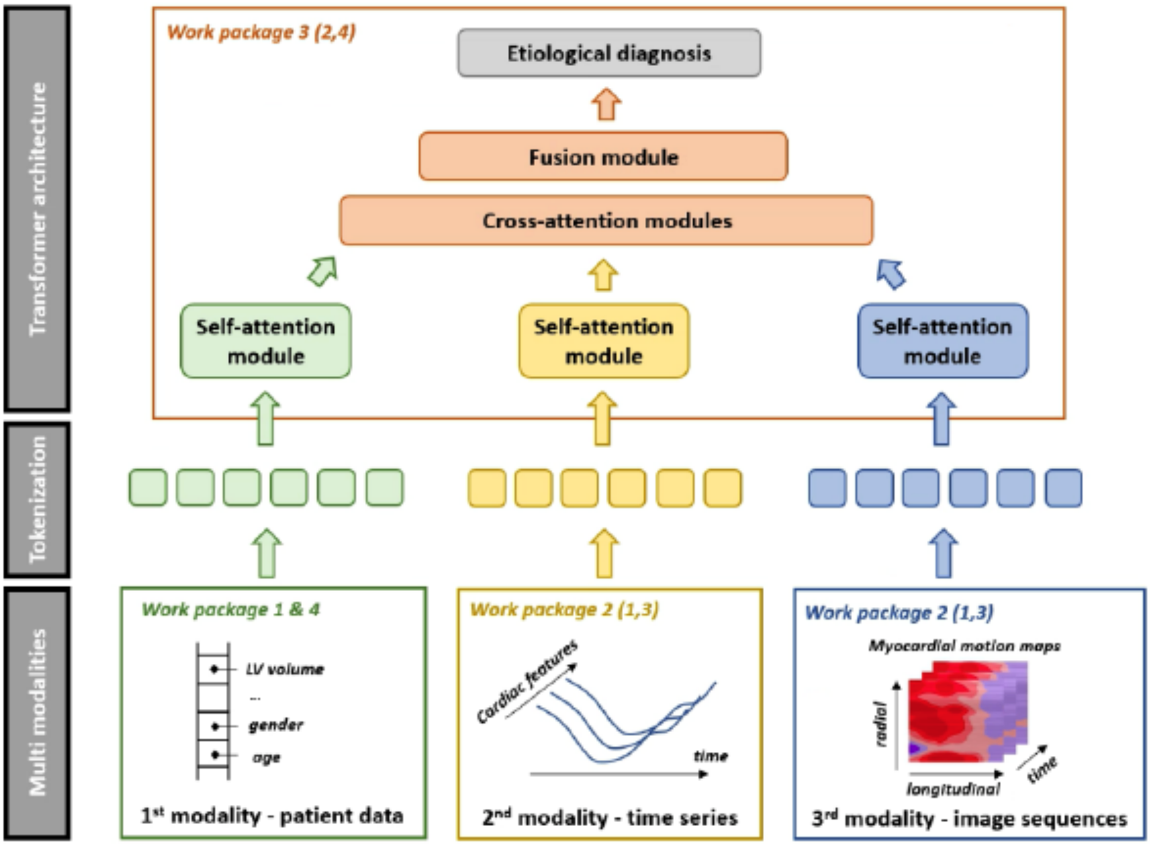
How to apply: please send a cv, motivation letter, grades obtained in master, recommendation letters when possible to nicolas.thome@sorbonne-universite.fr, olivier.bernard@insa-lyon.fr

Start date: October/November 2023

Keywords: etiological diagnosis, deep learning, transformer, multi-modal data

Context

This PhD is founded by the ORCHID ANR project (2023-2027). The main objective in ORCHID is to perform etiological diagnosis, *i.e.* to predict the origin of a given pathology, by combining multi-modal and heterogeneous input data.



As illustrated in the figure above, we are interested in combining diverse sources of information, from raw echography image sequences, times series representing the evolution of cardiac features, and patient data (*e.g.* age, gender, various medical histories).

The objective of this project is to develop rigorous and explainable cardiac disease prediction models based on artificial intelligence (AI). The main challenge is to model complex interactions between high-quality image-based measurements extracted from echocardiograms and relevant patient data to automatically predict etiological diagnosis of cardiac diseases.

Research directions

Regarding methodology, the project will study deep learning models and especially attentional models, *a.k.a.* transformers, for performing the targeted prediction tasks. Transformers have originally been proposed in Natural language processing (NLP) [VAS-17], which use in the field of vision has recently increased significantly with remarkable success [DOS-21]. The use of transformers raises specific challenges in the context of the project that will be explored during the PhD.

Transformers training and architecture

The main goal of this axis is to design efficient transformer architectures able to finely combine the multi-modal inputs of the project and perform accurate diagnosis prediction.

A first issue relates to the training strategies. Transformers require huge amount of data to be effective, and are generally pre-trained on generalist datasets, *e.g.* Google JFT dataset for images (300M of images). We will study specific training strategies, including simple transfer, full fine-tuning, or the use of dedicated adapters [CDW-13] amenable to frugal learning.

A second crucial goal is to design efficient transformer architectures able to finely combine the multi-modal inputs of the project and perform accurate diagnosis prediction. We will introduce efficient self-attention modules to model interactions inside each modality, and cross-attention modules for learning advanced fusion operators between modalities. We will propose efficient attention mechanisms able to deal with high-resolution inputs including 3D data. Preliminary solutions proposed by the supervision team for 2D [PET-21] or 3D inputs [THE-23] will be adapted for the echography image sequences of the project.

Multi-modal fusion

The fusion module aims at exploiting the output of the different cross-attention modules in order to perform the final prediction. We will explore different strategies. We will first implement a naïve solution as baseline, e.g. concatenating or averaging the outputs of each cross-attention modules. We will then experiment the option used in [TAN-19], consisting in alternating cross- and self-attention modules full-range contextual interactions before taking the final decision. Finally, we will explore the explicit modeling of fusion operators between the embedded modalities. Beyond simple strategies (averaging, product), we will explore bilinear fusion operators which captures finer-grained correlations. We will study tensor decompositions [BEN-17, BEN-19] for making the approach scalable, and relates them to efficient attention methods.

To make our final pipeline robust to missing data, we will explore the use of imputation techniques [AHJ-16, AWD-18] and adapt these methods for dealing with deep neural networks [TLZ+17]. Beyond these experimental results, we will study theoretical properties of the proposed imputation techniques. For example, [JPS-19] showed that mean imputation techniques provide competitive baseline when dealing with prediction task. We will refine this analysis to the DL predictive models and multi-modal data considered in this project.

Explainability

We will leverage the attention learned from the transformers developed to provide advanced tools to explain the model decision. To give a concrete example, the proposed method will be able to explain a class prediction (e.g. coronary artery disease) by highlighting the most relevant interactions leading to the decision, i.e. the interactions between a given set of patient variables (e.g. age, weight and blood pressure), temporally-localized information from different time series (e.g. important time steps in the left ventricle global longitudinal strain and atrial volume), and spatio-temporally localized information in a given myocardial motion map.

Transformers offer intrinsic explainability features through the use of attention. We will first analyze the learned links between the different modalities to provide an explanation from an applicative point of view, *e.g.* by highlighting the link between image regions and some patient data. Since attention is applied at several places of the network (e.g. different layers and attention heads), the main challenge is to identify the most relevant interactions for explaining the model decision. We will consider two complementary solutions. We will first leverage the idea of selecting peaky distribution attention maps recently proposed in [JKG-21], assuming that maps with few highly activated interactions are important for the final decision. We will then explore more rigorous selection strategies of attention maps by computing their importance with respect to the final decision of the model. To this end, we will adapt popular techniques from deep learning explanation such as GRAD-CAM [SCD-17] or LRP [MBL-19] to quantify each attention map by backpropagation.

A more exploratory study will be conducted to build upon prototype approaches [*CTB-19*], which learns some human-understandable features. We will explore new solutions combining prototypes and global token used in transformers to provide effective and explainable-by-design methods.

References

[AHJ-16] V. Audigier, et al., "A principal component method to impute missing values for mixed data", Adv Data Anal Classif, 10(1):5-26, 2016.

[AWD-18] V. Audigier, et al., "Multiple imputation for multilevel data with continuous and binary variables", Stat Sci, 33(2):160-183, 2018.

[BEN-17] H. Ben-younes. et al.,, “MUTAN: Multimodal Tucker Fusion for Visual Question Answering", ICCV 2017

[BEN-19] H. Ben-younes, et al., "BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection", AAAI 2019.

[CDW-13] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, Yu Qiao, "Vision Transformer Adapter for Dense Predictions". ICLR 2023.

[CTB-19] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C. and Su, J. K. This Looks Like That: Deep Learning for Interpretable Image Recognition. In NeurIPS, 2019.

[DOS-21] A. Dosovitskiy, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", in Proc. ICLR 2021

[JKG-21] T Jaunet, et al., “VisQA: X-raying Vision and Language Reasoning in Transformers”. In IEEE Transactions on Visualization and Computer Graphics (TVCG), 2021.

[JPS-19] On the consistency of supervised learning with missing values. J Josse, N Prost, E Scornet, G Varoquaux, Arxiv 2019.

[MBL-19] G. Montavon et al., "Layer-Wise Relevance Propagation: An Overview", Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019.

[PET-21] O. Petit, et al., "U-Net Transformer: Self and Cross Attention for Medical Image Segmentation", MLMI workshop, MICCAI 2021.

[VAS-17] A. Vaswani, et al., "Attention is all you need", NeurIPS, 30, 2017

[TAN-19] H. Tan, “LXMERT: Learning Cross-Modality Encoder Representations from Transformers“, EMLNP’19

[THE-23] Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation. L. Themyr, C. Rambour, N. Thome, T. Collins, A. Hostettler. WACV 2023.

[SCD-17] R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization". In ICCV, 2021.

[TLZ+17] L. Tran, X. Liu, J. Zhou and R. Jin. Missing Modalities Imputation via Cascaded Residual Autoencoder, CVPR'17.