Contrat doctoral – ED Galilée

Titre du sujet : Interpretability for Multiple Instance Learning for Medical Imaging

- **Unité de recherche** : LIPN
- **Discipline** : Informatique
- **Direction de thèse** : Céline Rouveirol
  Co-encadrement : Thomas Papastergiou
- **Contact** : rouveirol@lipn.univ-paris13.fr; papastergiou@lipn.univ-paris13.fr
- **Domaine de recherche** : Informatique, Apprentissage Automatique, Exlainabilité
- **Mots clés** :Machine Learning, Epxlainable Machine Learning, Multiple Instance Learning

Introduction

Multiple Instance Learning

Multiple Instance Learning (MIL) is a weakly supervised Machine Learning (ML) paradigm [1]. In contrary to the classical Machine Learning paradigm, where each object is represented by a single feature vector and a label is associated to each object, in Multiple Instance Learning each object is represented by a set of feature vectors (called *instances*) and labels are associated to this set of features (called *bag*). This representation offers a more complete description of an object, since different feature vectors can capture different aspects of the object (e.g. instances could refer to paragraphs and bags to documents, instances could refer to part of an image and bag to the entire image, or instances could refer to molecular substructures and bag to compounds, etc.) but it comes with a drawback: labels are provided only for the bags and not for the individual feature vectors. Thus, the bag labels are determined based on the unknown labels of the instances, and it is not obvious which instance is supporting which class. This fact implies an underlying assumption on how the bag labels are determined.

In the original definition of the MIL paradigm, Dietterich et al. [2] formulated the Standard Multiple Instance Learning assumption, which initially was referring to a binary classification problem with two classes: a positive and a negative class. The bag's label is recognized as positive if *at least* one positive instance is present in the bag and negative otherwise. Even that this assumption might seems very restrictive, it captures though the nature of many real life problems. For example, in the histopathology images classification problem, an image must be characterized as cancerous even if only a small part of it is cancerous.

However, the restriction of standard MIL assumption to binary classification problems, implies that only one concept is contained in the data. Besides the standard MIL assumption, Weidman et al. [3] defined a hierarchy of assumptions defining the relationship between the instances' and bags' labels. Moreover, individual algorithms have been proposed that extend the one-concept framework to the multiclass classification problem [4], [5].

Interpretability in MIL

In MIL, the label of a bag is determined by a few instances, (e.g. in the standard MIL assumption only one positive instance can characterize a bag as positive). Thus, in order to interpret the inference process of a MIL classifier, the key-instances, which make a bag positive, must be recognized [6]. Furthermore, an interpretation method should be able to answer to two questions: (1) Which are the key instances for a bag and (2) What outcome does each key instance support (i.e. to the prediction of which class each instance contributes) [7]. In the case of the standard MIL assumption (i.e. in the binary classification setting) identifying the key-instances answers to both questions at once, since there is only one positive class. This is not the case, though, for the multiclass classification task, where several positive classes need to be defined. The explanation of the

classification task, should not to be confused with the instance selection procedures (e.g. [8], [9]) where the non-discriminatory instances are recognized and removed, for improving classification performance, and no information is given on the classification of which class these instances contribute and how much.

Several interpretation methods have been proposed for the Multiple Instance Learning setting, like methods that produce instance-level predictions and are local inherently interpretable like [10], [11], attention based methods like in [12], Graph Neural Network based methods like in [13] or model agnostic MIL interpretation methods like in [7].

Research directions

In this PhD, we will propose new approaches and algorithms for the interpretability of Multiple Instance Learning classifiers. In contrast to the domain of interpreting Machine Learning and Deep Learning models the field of MIL interpretability is an underexplored field and gives many research opportunities. For example, the methodology in [7] is based on the identification of the contribution of the instances in the classification outcome, as aforementioned. In contrary, in [14] pixel-level contributions are calculated without paying attention to the contribution of each instance. Furthermore, in the same work, well known classical machine learning explicability methods, like GradCAM, Layer-wise Relevance Propagation (LRP), Information Bottleneck Attribution etc., have been adapted to the Deep MIL paradigm.

The purpose of this PhD will be to go a step further and to propose interpretability methods for the MIL paradigm that will account for both the contribution of each instance as well as of each feature (i.e. pixel in the case of imagery) in the prediction of a model. To give an intuitive example in the case of histopathology images, where each slide is divided to different patches, cancerous instances can co-exist with non-cancerous instances, and furthermore cancerous cells can coexist with non-cancerous cells in the same cancerous instance. The challenge here, is to propose an adequate interpretability methodology for accounting both for instance- and feature level explanations. A further challenge, is the management of potential contradictions between instance- and feature-wise explanations. Another challenge of interpretability methods that account for both instance- and feature-wise explanations is the complexity of the interactions as we move from the classical MIL assumption and the two-class problem formulation to alternative MIL assumptions in the frame of multiclass classification problems. Finaly in this PhD, the previous work of the co-supervisors will be exploited. More specifically the previous work of C. Rouveirol on symbolic explanations of ML models [15] as well as the work of T. Papastergiou on MIL algorithms [5], [16] and instance selection techniques [8], [9] can be exploited to explore, between other approaches, symbolic MIL explanations.

In terms of applications we will focus primarily, on the field of medical imaging, and more specifically on histopathology images, without excluding other application areas if necessary. The domain of histopathology offers different types of tasks: e.g. the binary task of classifying malignant and benign tumors, or the more challenging multiclass classification tasks of predicting the type of cancer or related properties as the Gleason score in prostate cancer, like in [17]. A methodology of interpretability of MIL models accounting for instance- and feature-wise explanations in the multiclass classification task, would be a very useful tool for analyzing MIL inferences in histopathological images and in consequence to enhance the trustworthiness of MIL models in the medical domain.

Research methodology and approach

The research will be performed in two stages, as described below.

In the first phase (M1-M18), after a rigorous state-of-the art exploration, the different open source datasets (i.e. histopathological imaging for different types of cancer and different tasks) will be acquired, categorized and the different types of tasks will be identified. Next, we will propose, implement and evaluate interpretability methods for accounting both instance- and feature-wise explanations for the classical instance

learning assumption for the binary classification problem. In this frame we will apply our methodologies to malignant-benign histopathogical images classification task.

In the second phase (M18-M36), we will tackle the problem of proposing, implementing and evaluating novel methodologies for the multi-class classification task in the frame of alternative MIL assumptions. In this frame we will apply our methodologies in the interpretation multi-class classification problems e.g. the prediction of the Gleason score from histopathology images for the prostate cancer. Finally, a 3-5 months period is assigned to the writing of the dissertation.

In the frame of this PhD, we can have a potential collaboration with the Institut Curie (https://curie.fr/), which is specialized in cancer research and treatment, for acquiring histopathology data of breast cancer.

Bibliography

[1] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013, doi: 10.1016/j.artint.2013.06.003.

[2] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1, pp. 31–71, Jan. 1997, doi: 10.1016/S0004-3702(96)00034-3.

[3] N. Weidmann, E. Frank, and B. Pfahringer, "A Two-Level Learning Method for Generalized Multi-instance Problems," in *Machine Learning: ECML 2003*, vol. 2837, N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski, Eds., in Lecture Notes in Computer Science, vol. 2837. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 468–479. doi: 10.1007/978-3-540-39857-8_42.

[4] T. Papastergiou, J. Azé, S. Bringay, M. Louet, P. Poncelet, and L. Gavara, "Multiple Instance Learning Based on Mol2vec Molecular Substructure Embeddings for Discovery of NDM-1 Inhibitors," in *Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022)*, F. Fdez-Riverola, M. Rocha, M. S. Mohamad, S. Caraiman, and A. B. Gil-González, Eds., in Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2023, pp. 55–66. doi: 10.1007/978-3-031-17024-9_6.

[5] T. Papastergiou, E. I. Zacharaki, and V. Megalooikonomou, "Tensor Decomposition for Multiple-Instance Classification of High-Order Medical Data," *Complexity*, vol. 2018, pp. 1–13, Dec. 2018, doi: 10.1155/2018/8651930.

[6] Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou, "Key Instance Detection in Multi-Instance Learning," in *Proceedings of the Asian Conference on Machine Learning*, Steven C. H. Hoi and Wray Buntine, Eds., PMLR, Nov. 2012, pp. 253–268. [Online]. Available: https://proceedings.mlr.press/v25/liu12b.html

[7] J. Early, C. Evers, and S. Ramchurn, "Model Agnostic Interpretability for Multiple Instance Learning," 2022, *arXiv*. doi: 10.48550/ARXIV.2201.11701.

[8] T. Papastergiou, E. I. Zacharaki, and V. Megalooikonomou, "TensMIL2: Improved Multiple Instance Classification Through Tensor Decomposition and Instance Selection," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain: IEEE, Sep. 2019, pp. 1–5. doi: 10.23919/EUSIPCO.2019.8902500.

[9] E. Branikas, T. Papastergiou, E. I. Zacharaki, and V. Megalooikonomou, "Instance Selection Techniques for Multiple Instance Classification," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, PATRAS, Greece: IEEE, Jul. 2019, pp. 1–7. doi: 10.1109/IISA.2019.8900679.

[10] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018, doi: 10.1016/j.patcog.2017.08.026.

[11] S. A. Javed, D. Juyal, H. Padigela, A. Taylor-Weiner, L. Yu, and A. Prakash, "Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 20689–20702. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/82764461a05e933cc2fd9d312e107d12-Paper-Conference.pdf

[12] M. Ilse, J. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., Proceedings of Machine Learning Research: PMLR, 2018, pp. 2127--2136. [Online]. Available: https://proceedings.mlr.press/v80/ilse18a.html

[13] M. Tu, J. Huang, X. He, and B. Zhou, "Multiple instance learning with graph neural networks," 2019, *arXiv*. doi: 10.48550/ARXIV.1906.04881.

[14] A. Sadafi *et al.*, "Pixel-Level Explanation of Multiple Instance Learning Models in Biomedical Single Cell Images," in *Information Processing in Medical Imaging*, vol. 13939, A. Frangi, M. De Bruijne, D. Wassermann, and N. Navab,

Eds., in Lecture Notes in Computer Science, vol. 13939. , Cham: Springer Nature Switzerland, 2023, pp. 170–182. doi: 10.1007/978-3-031-34048-2_14.

[15] V. Ventos *et al.*, "Construction and Elicitation of a Black Box Model in the Game of Bridge," in *Advances in Knowledge Discovery and Management*, vol. 1110, R. Jaziri, A. Martin, A. Cornuéjols, E. Cuvelier, and F. Guillet, Eds., in Studies in Computational Intelligence, vol. 1110. , Cham: Springer Nature Switzerland, 2024, pp. 29–53. doi: 10.1007/978-3-031-40403-0_2.

[16] T. Papastergiou *et al.*, "Discovering NDM-1 inhibitors using molecular substructure embeddings representations," *J. Integr. Bioinforma.*, vol. 20, no. 2, p. 20220050, Jul. 2023, doi: 10.1515/jib-2022-0050.

[17] M. Ren, M. Huang, Y. Zhang, Z. Zhang, and M. Ren, "Enhanced hierarchical attention mechanism for mixed MIL in automatic Gleason grading and scoring," *Sci. Rep.*, vol. 15, no. 1, p. 15980, May 2025, doi: 10.1038/s41598-025-00048-9.