Résumé de thèse – Candidature au Prix de thèse GdR ISIS 2025

Ma thèse s'inscrit dans le domaine de l'explicabilité des réseaux de neurones profonds (XAI), avec pour objectif de mieux comprendre leurs mécanismes internes et d'aligner leurs stratégies de prise de décision sur celles des humains.

I) Vers une Attribution de l'Importance Fiable des Modèles de Vision

Dans un premier temps, mes travaux ont porté sur l'amélioration des méthodes d'attribution [1,2,3], en développant des estimateurs rigoureux issus de l'Analyse de Sensibilité Globale pour quantifier de manière robuste l'importance des caractéristiques exploitées par les modèles. Ces avancées ont abouti à la création de cartes de saillance fiables, offrant une première ouverture sur la manière dont ces modèles perçoivent les images.

Dans une démarche expérimentale, j'ai encadré un projet de psychophysique à grande échelle [4], visant à évaluer l'impact des explications sur la capacité des observateurs humains à prédire les décisions des modèles. Nos résultats ont révélé une limitation fondamentale des approches existantes : les méthodes d'attribution indiquent ***où*** un modèle regarde, mais échouent à expliquer ***ce qu'il perçoit réellement*** dans ces régions.

II) Dépasser l'Attribution : Vers une Explication Conceptuelle

Face à cette problématique, j'ai introduit trois contributions majeures :

- CRAFT [5], une méthode permettant d'extraire automatiquement les concepts fondamentaux utilisés par les réseaux de neurones.
- Un outil de visualisation spécialisée [7,8] pour rendre ces concepts accessibles.
- Une unification théorique des approches d'extraction de concepts [6], accompagnée d'une méthodologie rigoureuse pour évaluer leur importance dans la prise de décision des modèles.

Ces travaux ont permis de révéler les stratégies sous-jacentes des architectures de vision artificielle et de démontrer la complémentarité entre différentes méthodes d'explicabilité. Cette synergie a conduit à la conception de LENS [11], un projet illustrant les stratégies variées employées par ResNet-50 pour chacune des 1000 classes d'ImageNet.

III) Alignement des Modèles avec la Perception Humaine

Enfin, mes recherches se sont orientées vers l'alignement des réseaux de neurones profonds avec la perception humaine [9,10]. À l'aide du jeu de données ClickMe (500 000 annotations psychophysiques), j'ai introduit un benchmark pour comparer les cartes d'importance des modèles avec celles des humains. Nos résultats ont mis en évidence un problème préoccupant : les modèles récents s'éloignent progressivement des stratégies humaines. Pour y remédier, j'ai proposé un schéma d'apprentissage guidé par l'explicabilité [9], permettant de contraindre les modèles à utiliser des indices pertinents du point de vue humain, conciliant ainsi performance et alignement cognitif.

Conclusion

En résumé, mes recherches doctorales ont poursuivi deux objectifs distincts. D'une part, j'ai développé des outils pour mieux comprendre où et quels *indices visuels* sont utilisés par les modèles neuronaux. D'autre part, j'ai proposé des techniques visant à rendre ces modèles plus interprétables en les alignant davantage avec les stratégies humaines. Je pense que mes travaux en intelligence artificielle explicable et en alignement des modèles s'intègrent pleinement aux thématiques du GdR ISIS, notamment en ce qui concerne la compréhension et l'interprétabilité des modèles d'apprentissage profond appliqués au traitement de l'image et de la vision par ordinateur.

[1] Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. T.Fel, R.Cadène, M.Chalvidal, M.Cord, D.Vigouroux, T.Serre. Conference on Neural Information Processing Systems, **NeurIPS 2021**

[2] Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis. T.Fel, M.Ducoffe, D.Vigouroux, R.Cadène, M.Capelle, C.Nicodème, T.Serre. Conference on Computer Vision and Pattern Recognition, **CVPR 2023**

[3] Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure. P.Novello, T.Fel, D.Vigouroux. Conference on Neural Information Processing Systems, **NeurIPS 2022**

[4] What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. T.Fel, J.Colin, R.Cadène, T.Serre. Conference on Neural Information Processing Systems, **NeurIPS 2022**

[5] CRAFT: Concept Recursive Activation FacTorization. T Fel, A Picard, L Bethune, T Boissin, D Vigouroux, J Colin, R Cadène, T Serre. Conference on Computer Vision and Pattern Recognition, **CVPR 2023**

[6] A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation. T.Fel, V Boutin, M Moayeri, R Cadène, L Bethune, M Chalvidal, T Serre. Conference on Neural Information Processing Systems, **NeurIPS 2023 (Spotlight)**

[7] Unlocking Feature Visualization for Deeper Networks with Magnitude Constrained Optimization. T.Fel, T Boissin, V Boutin, A Picard, P Novello, J Colin, D Linsley, T Rousseau, R Cadène, L Gardes, T.Serre. Conference on Neural Information Processing Systems, **NeurIPS 2023**

[8] Feature Accentuation. C Hamblin, T Fel, S Saha, T Konkle, G.A Alvarez. Under review

[9] Harmonizing the object recognition strategies of deep neural networks with humans. T.Fel, I.Felipe, D.Linsley, T.Serre. Conference on Neural Information Processing Systems, **NeurIPS 2022**

[10] Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. D Linsley, IF Rodriguez, T Fel, M Arcaro, S Sharma, M Livingstone, T Serre. Conference on Neural Information Processing Systems, **NeurIPS 2023**

[11] https://serre-lab.github.io/Lens/