AVIS DE SOUTENANCE DE THÈSE

Monsieur Hamed BENAZHA est autorisé à présenter ses travaux en vue de l'obtention du diplôme national de DOCTORAT délivré par l'école CENTRALE de MARSEILLE

Le vendredi 21 novembre 2025 à 14h00

Lieu : Salle 020, batiment la Jetée, 38 Rue Frédéric Joliot Curie, 13013 Marseille

Titre : **Améliorer l'interprétabilité des réseaux de neurones par l'apprentissage de représentations structurées**

Ecole doctorale : ED 184 - Mathématiques et Informatique de Marseille

Spécialité : Informatique

Composition du jury :

M. Thierry ARTIERES	Centrale Méditerranée	Directeur de thèse
M. Stéphane AYACHE	Aix-Marseille Université	Co-directeur de thèse
Mme Séverine DUBUISSON	Aix-Marseille Université	Présidente
M. Bertrand DELEZOIDE	Amanda	Rapporteur
M. Vincent GUIGUE	AgroParisTech	Rapporteur
M. Guillaume RABUSSEAU	Université de Montréal	Examinateur

Résumé (FR)

L'apprentissage profond a révolutionné l'intelligence artificielle, permettant des avancées dans de nombreux domaines d'application. Cependant, cette performance s'accompagne souvent d'un manque de transparence. Les modèles neuronaux, en raison de leur complexité, sont fréquemment perçus comme des boîtes noires, ce qui constitue un frein majeur à leur compréhension et à leur adoption dans des secteurs sensibles où la confiance et l'explicabilité sont primordiales. Cette thèse aborde ce défi central de l'interprétabilité en se concentrant sur l'apprentissage de représentations internes structurées. Notre hypothèse est qu'en contraignant la manière dont les modèles encodent l'information, il est possible de concevoir des architectures intrinsèquement plus transparentes. Nos contributions s'articulent autour de deux axes principaux : l'apprentissage de représentations discrètes pour les modèles séquentiels, et l'apprentissage de représentations démêlées pour les données complexes. La première partie de cette thèse vise à remplacer les représentations continues, traditionnellement utilisées dans les réseaux de neurones, par des représentations discrètes ou quantifiées. En forçant les états internes d'un modèle à appartenir à un ensemble fini de symboles, nous transformons son fonctionnement pour le rapprocher de celui des systèmes logiques et symboliques. Nous explorons cette idée dans le cadre des réseaux de neurones récurrents, en développant des méthodes pour discrétiser leurs représentations latentes, que ce soit pendant ou après la phase d'apprentissage. Cette approche permet d'établir une équivalence formelle entre un RNN et un automate fini. Une telle structure symbolique permet de transformer une boîte noire en boîte blanche. Cela ouvre la voie à de nouvelles stratégies d'explicabilité, en facilitant l'analyse du comportement du modèle de manière exhaustive. En outre, nous montrons qu'un avantage complémentaire majeur de cette discrétisation est une accélération de la vitesse d'inférence. Les

opérations matricielles complexes sont remplacées par de simples lectures de table, ce qui rend ces modèles particulièrement adaptés aux environnements contraints en ressources ou en latence. La seconde partie de la thèse s'intéresse à l'apprentissage de représentations démêlées, dans lesquelles chaque facteur de variation des données est encodé de manière indépendante. Dans le cadre des modèles génératifs de type auto-encodeur variationnel (VAE), nous introduisons et formalisons le concept d'hallucination. Ce phénomène, bien qu'étudié dans le domaine de la génération de texte, est ici adapté à la génération d'images pour désigner la création d'attributs irréalistes ou incohérents lors de manipulations de l'espace latent. Nous proposons une métrique pour quantifier ce phénomène et montrons empiriquement que, bien qu'ils puissent être liés, le démêlage et l'hallucination sont deux propriétés distinctes d'une représentation. Enfin, nous réexaminons le paradigme dominant du démêlage, fondé sur l'idée que chaque facteur génératif devrait être encodé dans une seule dimension latente. Constatant les limites de cette approche pour des facteurs complexes, nous proposons un cadre de démêlage vectoriel, dans lequel chaque facteur génératif est représenté par un groupe de composantes latentes. Nous adaptons à ce nouveau cadre les métriques de l'état de l'art, et démontrons sa capacité à mieux démêler sur des données réelles.

Mots-clés : Interprétabilité, Apprentissage de représentations, Discrétisation, Représentations démêlées

Abstract (EN)

Deep learning has revolutionized artificial intelligence, enabling advances in many application domains. However, this performance often comes at the cost of transparency. Due to their complexity, neural models are frequently perceived as black boxes, which is a major obstacle to their understanding and adoption in sensitive sectors where trust and explainability are crucial. This thesis addresses the central challenge of interpretability by focusing on learning structured internal representations. Our hypothesis is that by constraining how models encode information, it is possible to design architectures that are intrinsically more transparent. Our contributions are organized around two main directions: learning discrete representations for sequential models, and learning disentangled representations for complex data. The first part of this thesis aims to replace the continuous representations traditionally used in neural networks with discrete or quantized representations. By forcing a model's internal states to belong to a finite set of symbols, we transform its functioning to resemble that of logical and symbolic systems. We explore this idea in the context of recurrent neural networks (RNNs), developing methods to discretize their latent representations, either during or after training. This approach allows us to establish a formal equivalence between an RNN and a finite-state automaton. Such a symbolic structure turns a black box into a white box. This opens the way for new explainability strategies by enabling exhaustive analysis of the model's behavior. Furthermore, we show that an additional major benefit of discretization is a speed-up in inference. Complex matrix operations are replaced by simple table lookups, making these models particularly suited to resource or latency constrained environments. The second part of the thesis focuses on learning disentangled representations, where each factor of variation in the data is encoded independently. In the context of generative models based on variational autoencoders (VAEs), we introduce and formalize the concept of hallucination. Although this phenomenon has been studied in the field of text generation, we adapt it here to image generation to refer to the creation of unrealistic or inconsistent attributes during latent space manipulations. We propose a metric to quantify this phenomenon and show empirically that, while related, disentanglement and hallucination are two distinct properties of a representation. Finally, we revisit the dominant disentanglement paradigm, which assumes that each generative factor should be captured by a single latent dimension, and propose a different approach. Noting the limitations of this approach for complex factors, we propose a vector-wise disentanglement framework, in which each generative factor is represented by a group of latent components. We adapt state-of-the-art metrics to this new setting and demonstrate its improved disentangling capability on real-world data.

Keywords: Interpretability, Representation learning, Discretization, Disentangled representations