AVIS DE SOUTENANCE DE THÈSE

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

Le lundi 23 septembre 2024 à 10h00

Lieu : Salle Amphi Sciences Naturelles, - 3 place Victor Hugo, 13331 Marseille cedex 3

Titre : **Apprentissage des représentations discriminantes pour interpréter les modèles de reconnaissance d'image**

Ecole doctorale : ED 184 - Mathématiques et Informatique de Marseille

Spécialité :

Composition du jury :

M. Stéphane AYACHE	Aix Marseille Université	Directeur de thèse
M. Ronan SICRE	École Centrale Mediterranée	Co-directeur de thèse
M. Frederic JURIE	Université de Caen Normandie	Rapporteur
M. Giorgos TOLIAS	Czech Technical University in Prague	Rapporteur
M. Frederic PRECIOSO	Université Côte d'Azur	Examinateur
Mme Diane LARLUS	Naver Labs Europe	Examinatrice

Résumé (FR)

Les capacités de vision par ordinateur se sont améliorées au cours de la dernière décennie, une meilleure utilisation du matériel permettant aux ordinateurs de traiter davantage d'images plus rapidement, entraînant l'avènement de l'apprentissage profond. De plus, au cours de cette période, des architectures de modèles telles que les réseaux de neurones convolutifs et les transformers ont été introduites, permettant aux applications de vision par ordinateur de réaliser des tâches plus complexes. En particulier, les modèles de reconnaissance d'images sont désormais capables d'identifier et de reconnaître des éléments sur une image, même dans des conditions difficiles. Ces facteurs ont contribué à l'introduction de ces modèles dans la société. Avec la diffusion des technologies de l' apprentissage profond au sein de la société, une nouvelle exigence a émergé pour ces méthodologies. Puisqu 'elles interagissent désormais et affectent directement les vies humaines, il est impératif de comprendre leur fonctionnement et de fournir des explications. Pour répondre à ces questions, un nouveau domaine de recherche a vu le jour: l'interprétabilité et l'IA explicable. Dans cette thèse, notre objectif est de comprendre et de développer des modèles d'interprétabilité pour les modèles de reconnaissance d'images de pointe. Nous présentons et expliquons brièvement certains des modèles de reconnaissance d'images les plus performants et pertinents pour les Réseaux de Neurones Convolutifs et les Transformers. Ensuite, nous examinons les approches actuelles en matière d' interprétabilité conçues pour fournir des explications, ainsi que leurs protocoles d' évaluation. Nous faisons des observations sur ces méthodes et protocoles d'évaluation, mettant en évidence les difficultés rencontrées et suggérant des

idées pour surmonter leurs limitations. Notre première contribution, s'appuie sur le raisonnement des Cartes d'Activation de Classe. En particulier, cette proposition optimise le coefficient de pondération requis pour calculer une carte de saillance, générant une représentation qui maximise la probabilité spécifique à la classe. Cette carte de saillance offre les meilleurs résultats selon les mesures d'interprétabilité, et met en évidence que le contexte est pertinent pour décrire une prédiction. De plus, une nouvelle métrique pour compléter l'évaluation de l'interprétabilité est dévoilée, remédiant aux lacunes de cette procédure. Notre deuxième contribution, est un ajout aux modèles actuels de reconnaissance d'images, améliorant les mesures d'interprétabilité. Inspiré par des modèles novateurs performants tels que les Transformers, nous construisons un flux qui calcule l'interaction d'une représentation de classe abstraite avec les caractéristiques profondes des réseaux neuronaux convolutionnels. Cette représentation est finalement utilisée pour effectuer la classification. Notre flux affiche des améliorations lors de l'évaluation quantitative, tout en préservant les performances de reconnaissance à travers différents modèles. Enfin, notre dernière contribution présente un nouveau paradigme d'entraînement pour les réseaux neuronaux profonds. De plus, ce paradigme débruite les informations de gradient des modèles profonds dans l'espace d'entrée. La représentation de rétropropagation guidée de l'image d'entrée est utilisée pour régulariser les modèles lors de leur phase d' entraînement. En conséquence, nos modèles entraînés affichent des améliorations pour l' évaluation interprétable. Nous appliquons notre paradigme à de petites architectures dans un cadre contraint, ouvrant la voie au développement futur dans des ensembles de données à grande échelle, ainsi qu'avec des modèles plus complexes. Mots clés: Apprentissage Profond, reconaissance d'image, interpretabilité, explicabilité.

Mots-clés : vision par l'ordinateur, apprentissage profond, reconnaissance d'image, interprétabilité, apprentissage non supervisé

Abstract (EN)

Computer Vision capabilities improved in the past decade, a better utilization of hardware enabled computers to process more images faster, ensuing the dawn of deep learning. Moreover, over this timespan, model architectures such as convolutional neural networks and transformers have been introduced, enabling computer vision applications to conduct more complex tasks. In particular, image recognition models are now capable of identifying and recognizing elements on an image, even on challenging conditions. These factors have contributed towards the introduction of these models into society. With the permeation of deep learning technologies within society, a new requirement emerged for these methodologies. Since they are now interacting and affecting human lives directly, it is mandatory to understand their functioning and provide explanations. To address these questions a new research field has emerged: interpretability and explainable AI. In this thesis, our goal is to understand and further develop interpretability models for state-of-the-art image recognition models. We introduce and briefly explain some of the most relevant high performance image recognition models for both Convolutional Neural Networks and Transformers. Then, current interpretability approaches designed to provide explanations, as well as their evaluation protocols. We make observations upon these methods and evaluation protocols, highlighting difficulties upon them and suggesting ideas to address their limitations. In the following chapters we present our contributions. Our first contribution, builds upon the reasoning of Class Activation Mappings. In particular, this proposal optimizes the weighting coefficient required to compute a saliency map, generating a representation that maximizes class specific probability. This saliency map performs the best across interpretability metrics on multiple datasets. Plus, it highlights that context is relevant towards describing a prediction. Additionally, a novel metric to complement interpretability evaluation is unveiled, addressing shortcomings in this procedure. Our second contribution, is an addition to current image recognition models, enhancing interpretability measurements. Inspired novel high performing models such as Transformers, we construct a stream that computes the

interaction of an abstract class representation, with deep features of convolutional neural networks. This representation is ultimately used to perform classification. Our approach displays improvements on quantitative evaluation, as well as preserves recognition performance across different models. Lastly, our final contribution presents a novel training paradigm for deep neural networks. Moreover, this paradigm denoises the gradient information of deep models in the input space. The guided backpropagation representation of the input image is used to regularize models during their training phase. As a result, our trained models display improvements for interpretable evaluation. We apply our method to small architectures in a constrained setting, paving the way for future development in large scale datasets, as well as with more complex models. Keywords: Deep Learning, image recognition, interpretability, explainability.

Keywords: computer vision, deep learning, image recognition, interpretability, unsupervised learning