Vers des Classifieurs Ontologiquement Explicables

G. Bourguin, A. Lewandowski, M. Bouneffa, A. Ahmad
SysReIC – LISIC - ULCO

Problématique

(In)Explicabilité des IA utilisant l'Apprentissage Profond (AP)

Besoin d'entrouvrir les boîtes noires : mouvement **XAI** (eXplainable AI)

« **Expliquer à un utilisateur final** le rationnel ayant mené à une décision **peut être aussi important que la prédiction** »

[L. A. Hendricks et al., Generating visual explanations. In ECCV, 2016.]

Domaine d'illustration

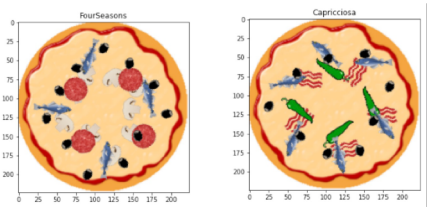
Une petite Pizza ?

Ontologie des Pizzas (*allégée*) - Université de Manchester -

- ~~22~~ 14 sous-classes de NamedPizza
- ~~36~~ 16 sous-classes de PizzaTopping

Napoletana ≡ Pizza ⊓ (∃ hasTopping . AnchoviesTopping)
⊓ (∃ hasTopping . OliveTopping)
⊓ (∀ hasTopping . (AnchoviesTopping ⊔ OliveTopping))

But : **classifier** des images (synthétiques) de pizzas + **expliquer**

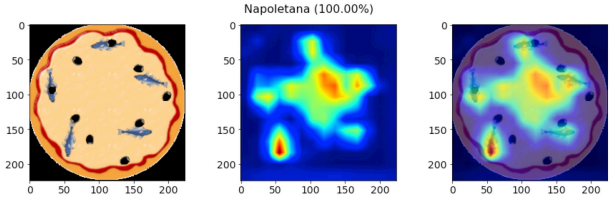


Explicabilité : état de l'art

Pondération des *features* d'entrée

Méthodes *post-hoc*, agnostiques
Grad-CAM, LIME, SHAP, …

Exemple avec Grad-CAM :
Classifieur « classique » VGG19 pré-entrainé (Imagenet)



Napoletana = anchois + vide (!)

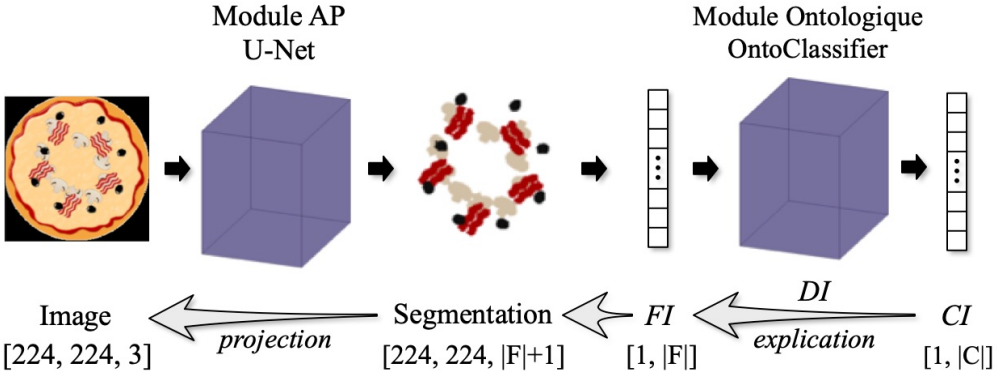
Ne correspond pas à la définition des experts

Proposition : marier AP & Ontologies pour des IA ontologiquement explicables

Fournir des **explications au niveau d'abstraction des experts** du domaine

Les **ontologies** capturent les **connaissances** du domaine. **L'inférence** ontologique **peut être expliquée**.

Réalisation : module d'**apprentissage profond** (***segmentation sémantique***) + module **ontologique** (***OntoClassifier***).



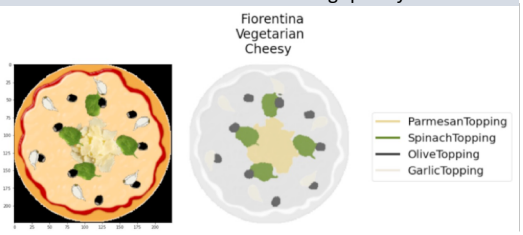
L'**OntoClassifier** *est* ***généré automatiquement*** *à partir de l'ontologie sous la forme d'un* ***graphe de tenseurs***.

Directement intégré au pipeline.

Classification de 100 images :
Hermit reasoner : ~ 130s
OntoClassifier : ~ 1,6s

Résultat : un Classifieur Ontologiquement Explicable

Classification multi-label ontologique + *features*



Focus sur les classes identifiées + explications