Approche Hybride pour des Classifieurs Ontologiquement Explicables

G. Bourguin, A. Lewandowski, M.Bouneffa, A. Ahmad

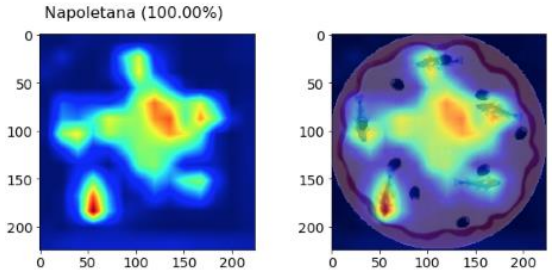
SysRelC

Contexte : besoin d'explicabilité des IA

- Explicabilité
  - Expliquer clairement à un **utilisateur final** le rationnel ayant mené à une décision peut être aussi important que la décision elle-même

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

- Méthodes pour l'explicabilité
  - Explication par mesure d'importance des *features* dans les données d'entrée
  - Outils : Grad-CAM, LIME, …

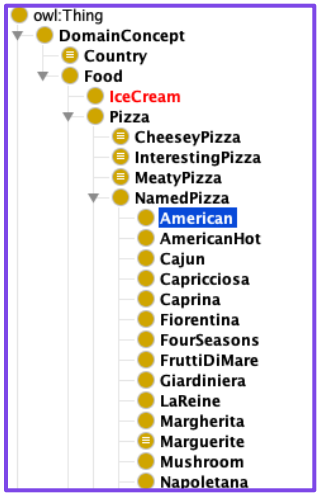


Illustration

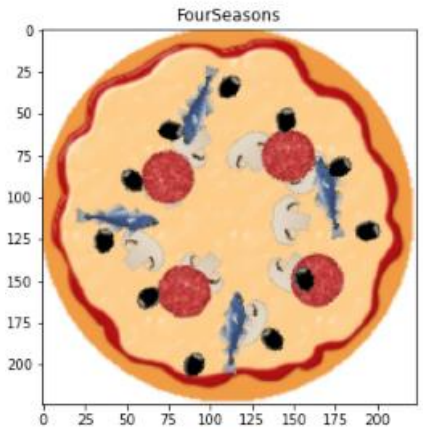
- But : un classifieur d'images
- Classes : ontologie des Pizzas (allégée)
  - ~~22~~ 14 sous-classes de **NamedPizza**
  - ~~36~~ 16 sous-classes de PizzaTopping

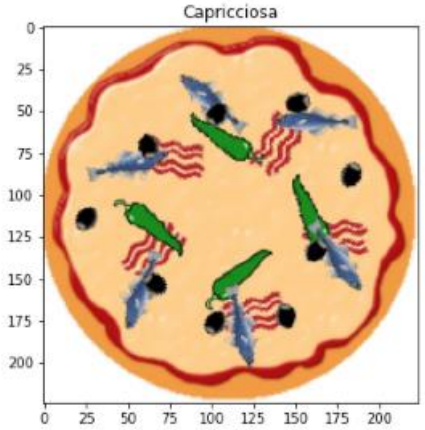
  - Exemple de définition :

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



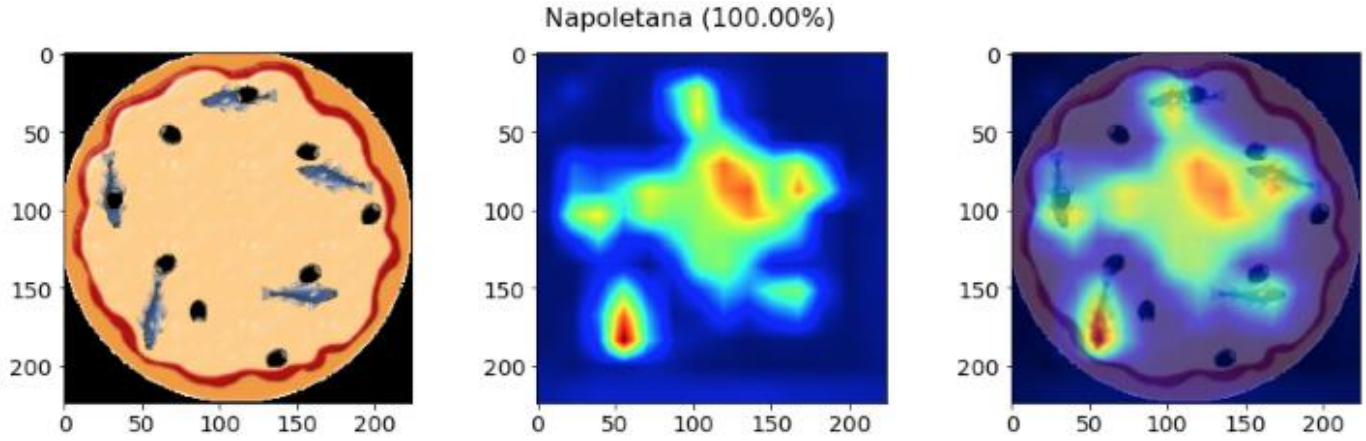
- Création d'un jeu de données contrôlé : images synthétiques





Problème du niveau d'abstraction

- Modèle de test
  - Spécialisation d'un VGG19 pré-entrainé sur Imagenet
- Explication d'une classification par Grad-CAM



  - C'est une **Napolitaine** car elle contient des **anchois** et du **vide**…
  - (les olives sont ignorées)
  - **Ne correspond pas à la définition des experts du domaine**

Positionnement

- Injecter de la connaissance : Ontologies
  - Les ontologies capturent les connaissances liées aux domaines d'expertises des utilisateurs, et permettent aux algorithmes de les manipuler
  - L'inférence ontologique est un processus déductif qui peut être expliqué

- Approche
  - Hybrider AP & ontologies pour obtenir des IA ontologiquement explicables
  - (cf. article…)

Classifieur Ontologiquement Explicable

- Pipeline de classification et d'explication

Image
[224, 224, 3]

- OntoClassifier :
  - Généré automatiquement à partir des éléments de l'ontologie (C, D, F)
  - Implémenté sous forme de tenseurs (Tensorflow 2, PyTorch)
    => rapide, directement ajouté en sortie du module d'AP

Résultats : classification

- Classification

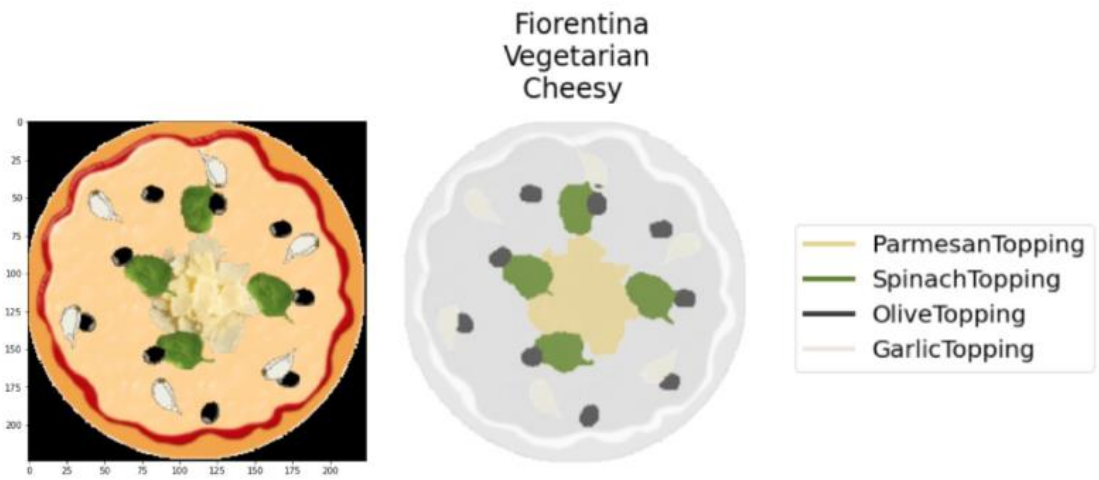
Expressions « simples »

Fiorentina ≡ Pizza ⊓ (∃ hasTopping.GarlicTopping) ⊓ (∃ hasTopping.OliveTopping) ⊓
(∃ hasTopping.ParmesanTopping) ⊓ (∃ hasTopping.SpinachTopping) ⊓
(∀ hasTopping.(GarlicTopping ⊔ OliveTopping ⊔ ParmesanTopping ⊔ SpinachTopping))

Expressions complexes utilisant l'héritage

CheesyPizza ≡ ∃ hasTopping . **CheeseTopping**

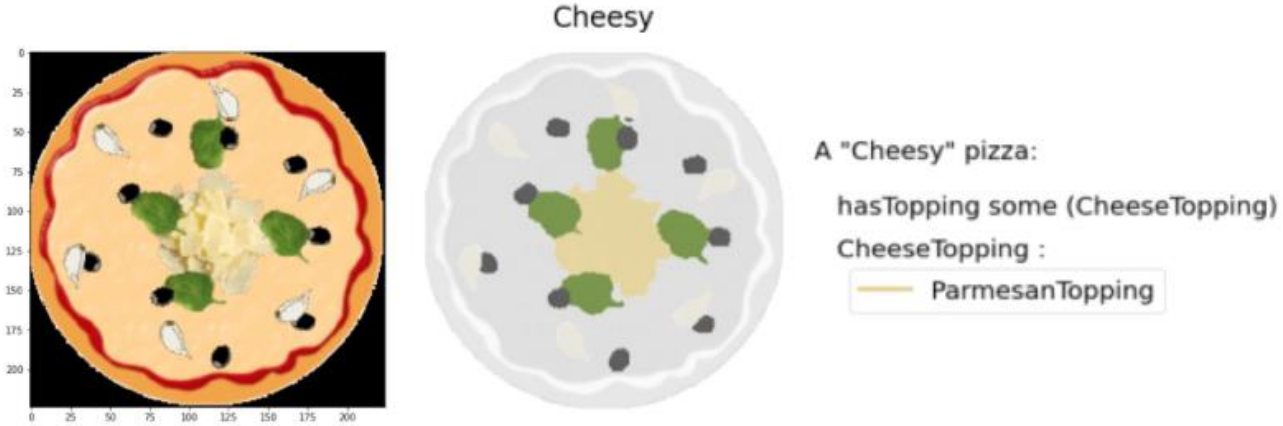
VegetarianPizza ≡ ¬ (∃ hasTopping . **FishTopping**) ⊓ ¬ (∃ hasTopping . **MeatTopping**)



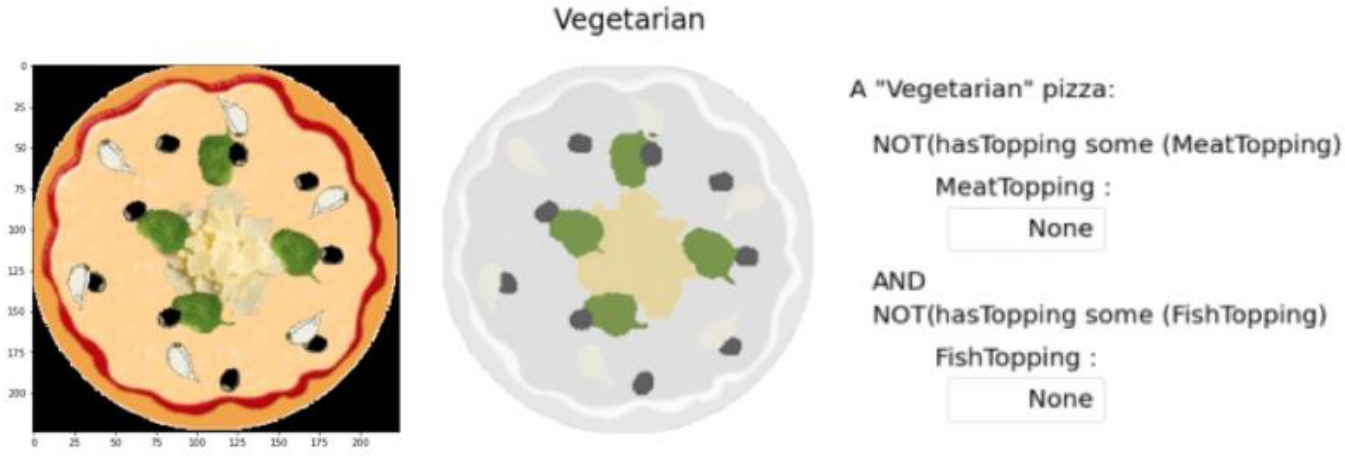
Résultats : explications

- Projection des définitions ontologiques sur les données d'entrée

Explications pour la classification en CheesyPizza :



Explications pour la classification en VegetarianPizza :



Discussion

- Formalisme des explications
  - Projection des définitions OWL ?!?
  - IHM de manipulation, co-construction d'une explication…
- Les explications s'arrêtent au niveau d'abstraction des *features ontologiques*
  - *Pourquoi c'est une olive ?*
  - Possibilité de raffiner en détaillant l'ontologie + module d'AP plus bas niveau (?)
  - Objectif : « Expliquer clairement à un utilisateur final … »
  - ➢ Pose la question du plus bas niveau d'abstraction utile à l'utilisateur final
  - ➢ Tâche utilisateur -> besoins émergents …
- Application (Thèse 2022) :
  - Problème écologique de suivi et d'observation des trajectoires de la faune volante marine
  - Images fournies par des dispositifs volants, expertise des ornithologues

Approche Hybride pour des Classifieurs Ontologiquement Explicables

G. Bourguin, A. Lewandowski, M.Bouneffa, A. Ahmad

SysReIC