# Vers des Classifieurs Ontologiquement Explicables

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

SysRelC

LISIC

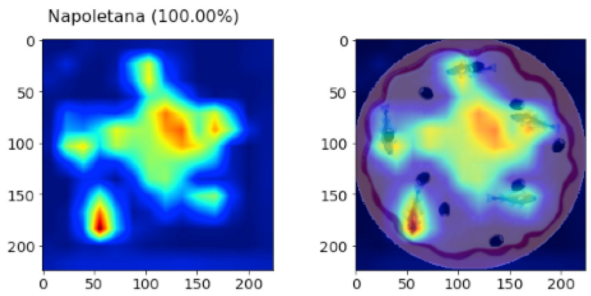
ulco Université Littoral Côte d'Opale

# Contexte : besoin d'explicabilité des IA

- Explicabilité
  - Problème d'opacité des modèles d'IA basés sur l'Apprentissage Profond (AP)
  - Performants **mais** problèmes de confiance des utilisateurs, questions d'éthique, …
  - 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, …

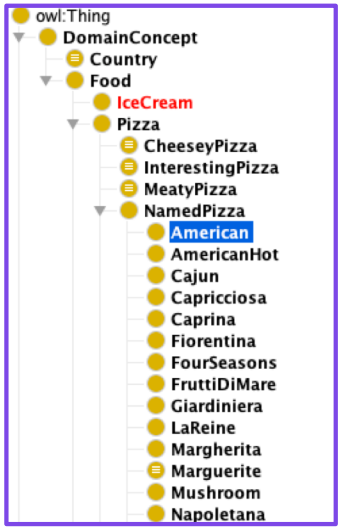



- Problème
  - Inadéquation du niveau d'abstraction des explications fournies

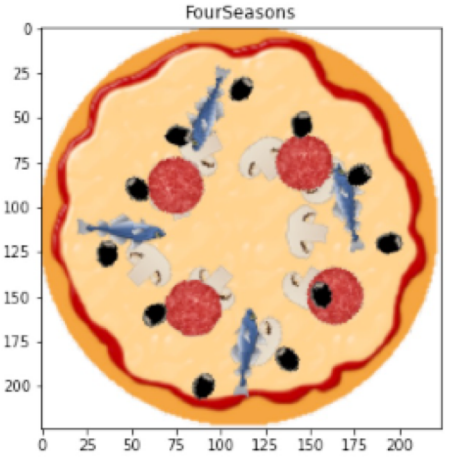
# Illustration

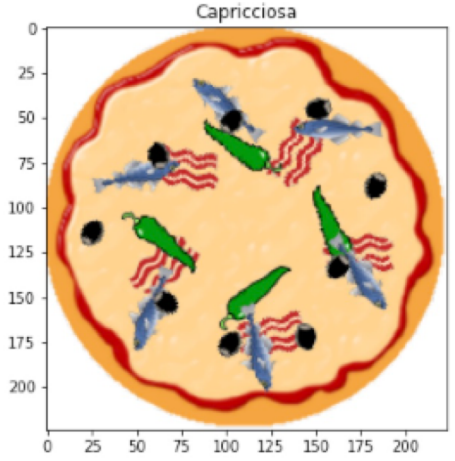
- But : un classifieur d'images
- Classes : ontologie des Pizzas (allégée)
  - ~~22~~ 14 sous-classes de **NamedPizza**
  - ~~36~~ 16 sous-classes de PizzaTopping
  - Exemple de définition :

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




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

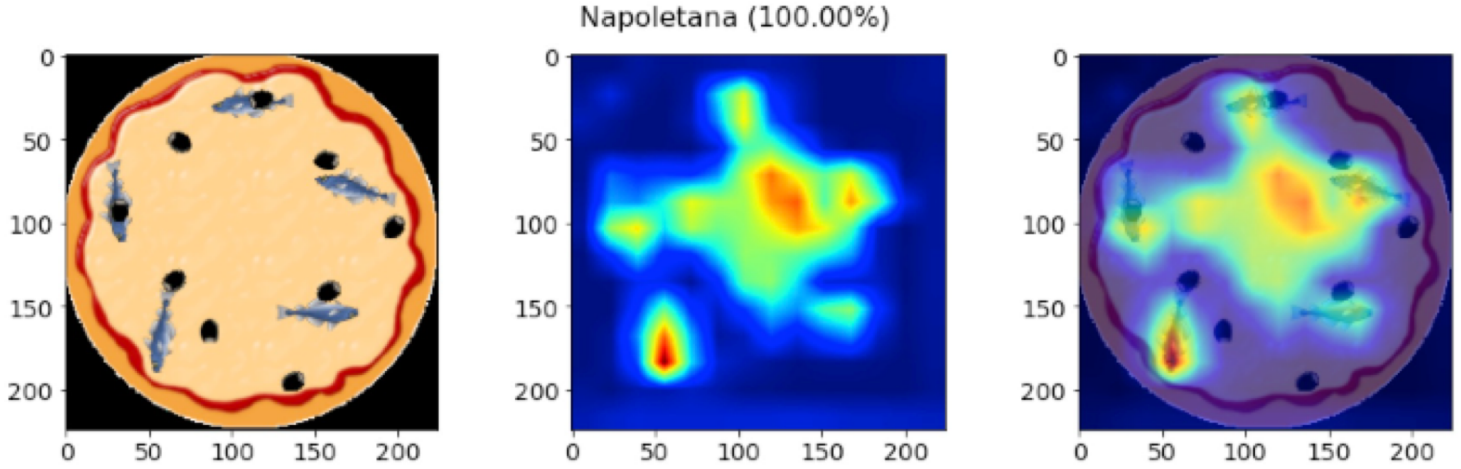


# Problème du niveau d'abstraction

- Modèle de test
  - Spécialisation d'un VGG19 pré-entrainé sur Imagenet
  - Jeu de données simple : 100% de précision

- 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

- Objectif
  - Des IA explicables comblant l'écart sémantique entre les entités manipulées par les algorithmes, et celles permettant d'expliquer les décisions
- 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
- AP, Explicabilité, Sémantique, Ontologies
  - De nombreux travaux introduisent de la sémantique dans les outils pour l'explicabilité des modèles d'AP, mais peu utilisent les ontologies (et l'inférence ontologique)
  - Les approches mêlant AP et ontologies focalisent peu sur l'explicabilité
- Notre approche
  - Mixer AP & Ontologies pour obtenir des IA ontologiquement explicables

# Approche

## Résumé de l'approche (détaillée dans le papier) :

a) Lister l'ensemble C des **classes ontologiques** à identifier

ex. c = Napoletana

b) Lister les ***features ontologiques*** (triplets F) servant à calculer et expliquer une classification à partir des définitions D des classes C

ex. d = (∃ hasTopping . AnchoviesTopping) ⊓ (∃ hasTopping . OliveTopping)
⊓ (∀ hasTopping . (AnchoviesTopping ⊔ OliveTopping))

ex. f = (Napoletana, hasTopping, AnchoviesTopping)

c) Mettre en œuvre une **technique d'AA** permettant de construire l'ensemble FI ⊆ F des ***features ontologiques*** **identifiées** (assertions satisfaites)

ex. Module de **segmentation sémantique**

d) Mettre en œuvre un **raisonnement ontologique** qui utilise D et FI pour calculer CI ⊆ C, l'ensemble des **classes identifiées** pour une donnée

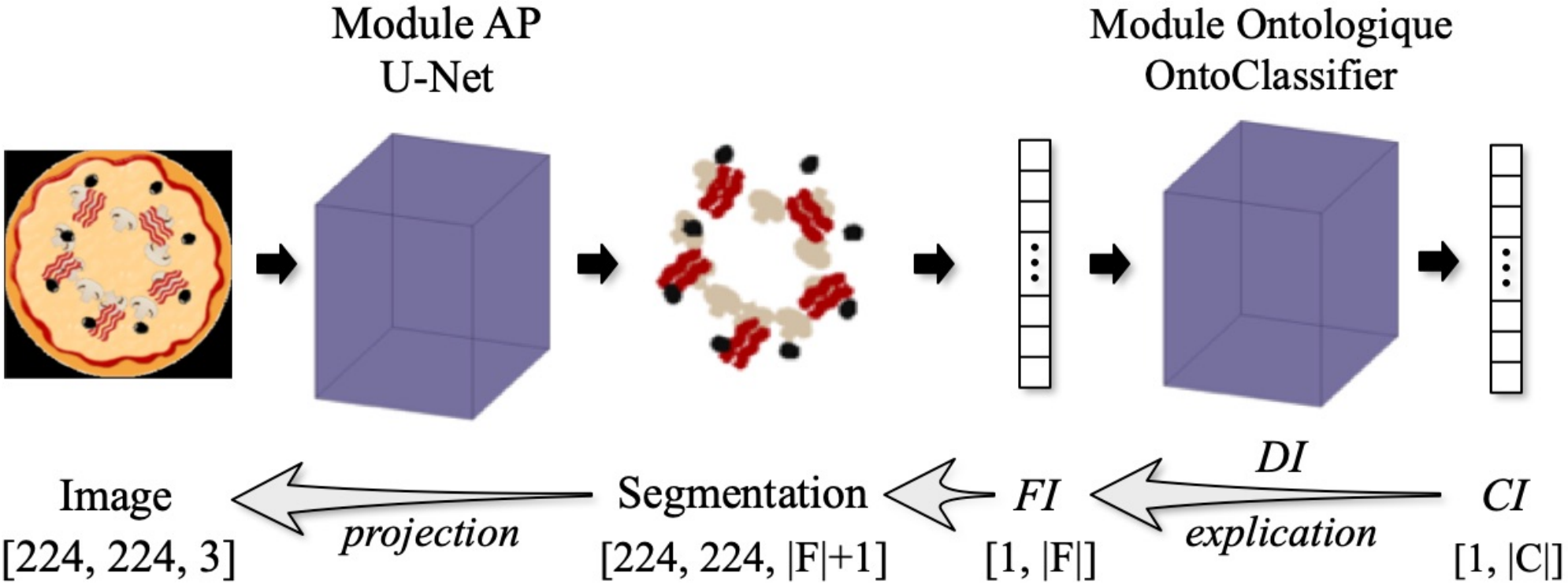
ex. Module **OntoClassifier**

e) Utiliser les axiomes DI ⊆ D tel que DI = { di ≡ ci } et l'ensemble FI pour **expliquer la classification** CI

ex. **Projection des assertions OWL** sur les *features* des données d'entrée

# Classifieur Ontologiquement Explicable

- Pipeline de classification et d'explication




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

# Résultats : classification

- Classification (100% précision)

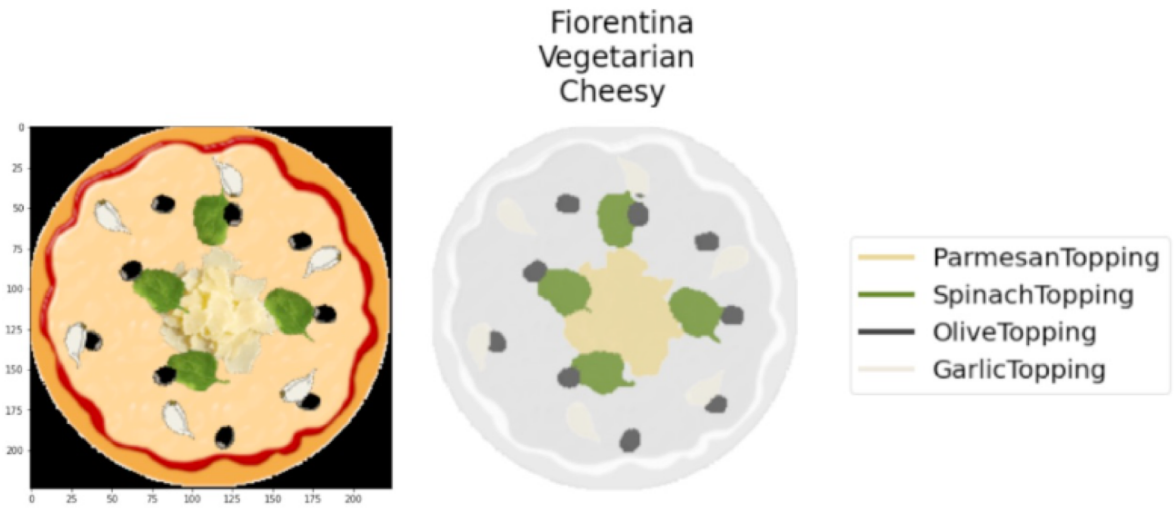
  Expressions « simples » :

  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**)

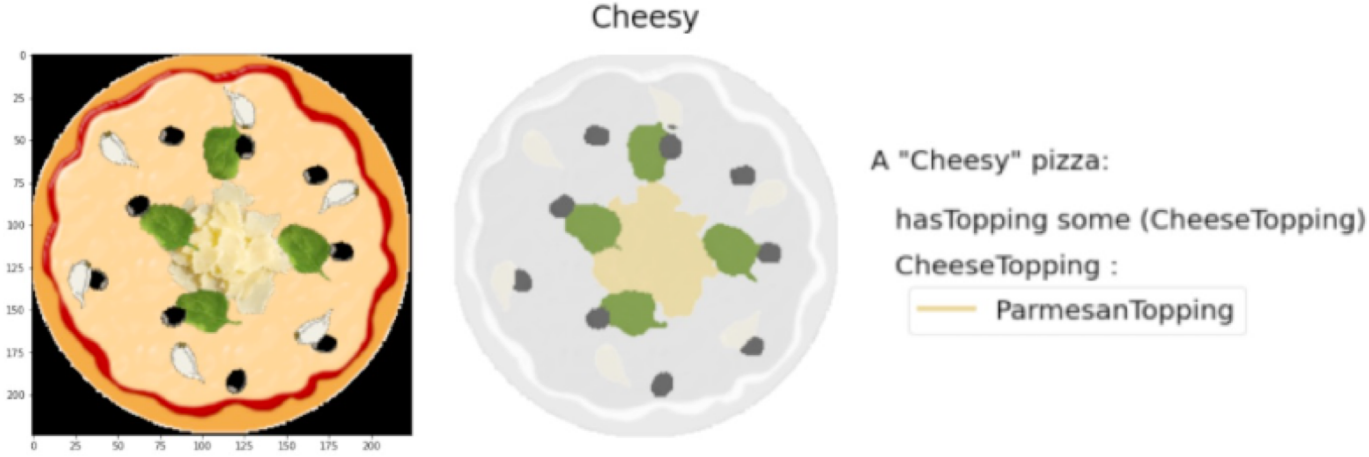



- Niveau d'abstraction du masque en adéquation avec l'ontologie

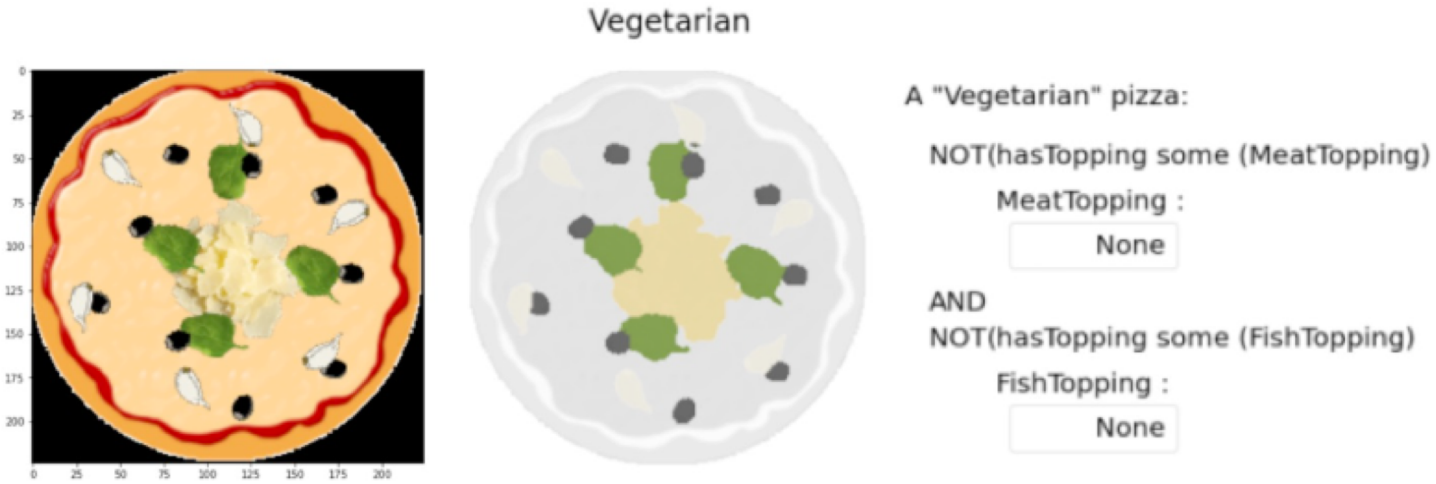
# Résultats : explications

- Projection des définitions ontologiques sur les données d'entrée

Explications pour la classification en CheesyPizza :




Explications pour la classification en VegetarianPizza :



# Discussion

- Les explications s'arrêtent au niveau d'abstraction des *features ontologiques*
  - *Pourquoi c'est une olive ?*
  - Possibilité de raffiner en détaillant l'ontologie & le module d'AP (?)
  - Objectif : «  Expliquer clairement à un **utilisateur final** … »
  - ➢ Pose la question du plus bas niveau d'abstraction utile à l'utilisateur final

- Travail préparatoire plus complexe que pour un « simple » classifieur
  - Création d'une ontologie, technique d'AP plus complexe

- Possibilité de modifier l'ensemble des classes (et définitions) en sortie du classifieur
  - Par simple nouvelle génération de l'OntoClassifier
  - Tant que les *features ontologiques* ne changent pas, pas de ré-entraînement

- Possibilité d'envisager différents points de vue

  Point de vue 1: $\text{VegetarianPizza} \equiv \neg (\exists\ \text{hasTopping} . \text{FishTopping}) \sqcup \neg (\exists\ \text{hasTopping} . \text{MeatTopping})$

  Point de vue 2: $\text{VegetarianPizza} \equiv \forall\ \text{hasTopping} . \text{VegetarianTopping}$

# Conclusion

- Vers des classifieurs ontologiquement explicables

  - Approche mise en œuvre dans de nouveaux projets
    - Jeux de données « réelles » (analyse d'impact écologique, …)
    - Classification d'images à grain fin

  - Demande à être développée plus avant
    - Explications par projection des définitions en OWL ?
    - Interface Homme Machine de manipulation des explications
      - Expérimentations avec des utilisateurs finaux
    - Autres domaines (NLP, …)

# Vers des Classifieurs Ontologiquement Explicables

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

SysRelC

LISIC

ulco Université Littoral Côte d'Opale