Spécifications

Introduction

L'explication de décisions des outils d'intelligence artificielle et notamment des réseaux de neurones profonds fait partie de recherche de la communauté scientifique sur l'IA explicable. Tandis que ce jour, il existe une grande variété d'approches en explication des décisions des réseaux profonds (DNNs) pour la classification des images, des vidéos et d'autre information, l'évaluation de ces outils reste un problème ouvert.

Les départements "Image et Son" (TAD - Traitement et analyse de données) et "Systèmes et Données" (BKB - Bench to Knowledge and Beyond) ont proposé des protocoles d'évaluation des outils d'explication avec les métriques sans référence et avec référence. Ces protocoles ne supposent pas d'implication directe de l'humain dans la boucle d'évaluation. Une implication a priori, indirecte, est prévue dans le protocole avec référence [1] par comparaison de la carte d'explication automatique fourni par un outil d'explication avec la carte de la saillance visuelle humaine construite à partir des points de fixations du regard ou des cliques souris obtenus lors d'expérience avec un groupe de sujets.

Objectif de PFE

Dans le cadre du PFE, il s'agit de développer un outil pour une expérience participative d'évaluation des cartes d'explication par des sujets humains.

Spécifications fonctionnelles

Nous envisageons le protocole suivant :

- les images classifiées
- le résultat de leur classification
- les cartes d'explications

seront affichés sur l'écran.

Également, la composante d'évaluation avec des boutons radio sera affichée selon l'échelle de Likert pour le scoring participatif des cartes. Le score attribué à chaque carte par chaque sujet humain sera enregistré de façon anonyme dans un fichier .csv.

Spécifications techniques

L'interface de scoring sera implantée comme un service web. (HTML5 + JS ou Python à vous de choisir). Le Mockup d'interface est présente sur la Figure 1 ci-dessous.

Mockup d'interface

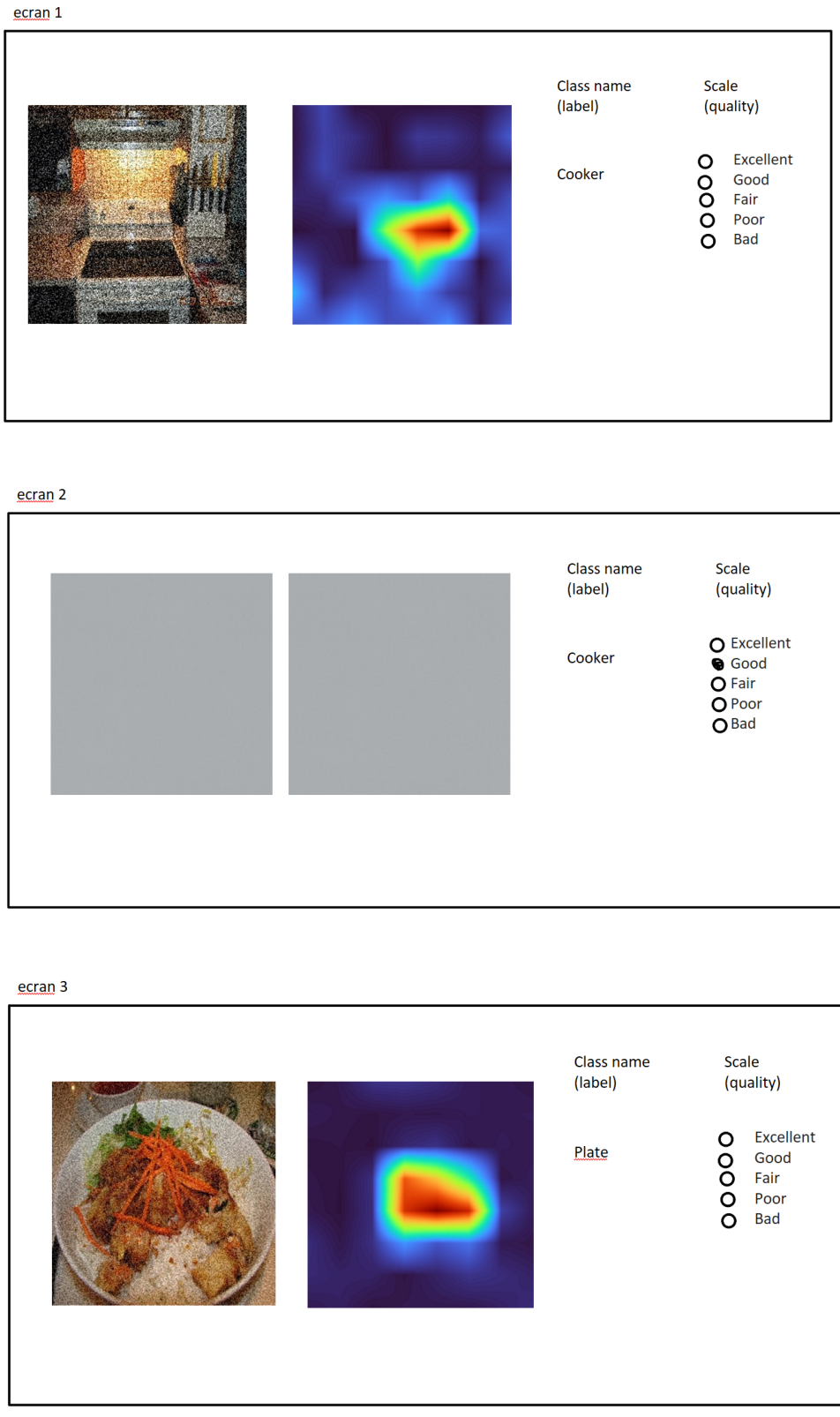

Figure 1: Mockup de l'interface de l'évaluation

Scenario interactif

La séance dynamique de visualisation pour un observateur et compose de deux types d'écrans d'interfaces :

1) L'affichage des images et des cartes
2) L'affichage de la même fenêtre graphique, mais avec des images grises.

Cf. Figure 1.

Ce deuxième type d'écrans remplacera l'écran de premier type dès que l'utilisateur a donné son avis en cliquant sur un des boutons de l'échelle Likert. Le temps d'affichage sera à régler à fin d'éviter au maximum la fatigue visuelle du sujet. Pour cela, nous allons utiliser la norme ITU-R Rec. BT.500-11 et le protocole du [2].

Références bibliographiques

[1] *A. Zhukov, J. Benois-Pineau, R. Giot: Evaluation of FEM and MLFEM AI-explainers in Image Classification tasks with reference-based and no-reference metrics- arXiv preprint arXiv:2212.01222, 2022 - arxiv.org*

[2] *A. Montoya Obeso, J. Benois-Pineau, MS. García-Vázquez, AA. Ramírez-Acosta: Visual vs internal attention mechanisms in deep neural networks for image classification and object detection. Pattern Recognit. 123: 108411 (2022)*