Spécifications

Introduction

L'explication de décisions des outils d'intelligence artificielle, et notamment des réseaux de neurones profonds, fait partie de la recherche de la communauté scientifique sur l'IA explicable. Tandis qu'il existe à ce jour 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. C'est d'autant plus vrai sur les approches de type extraction des caractéristiques d'entrée à l'aide de cartes de chaleur qui sont majoritairement utilisées.

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 en utilisant des métriques avec et sans 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 classées (pouvant être bruitées ou originales)
- le résultat de leur classification
- une carte d'explications (issue de différentes méthodes, sans que celles-ci ne soient diffusées à l'utilisateur)

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 pour le frontend, JS, PHP ou Python pour le backend, à vous de choisir). Le Mockup d'interface est présente sur la Figure 1 ci-dessous.

Mockup d'interface

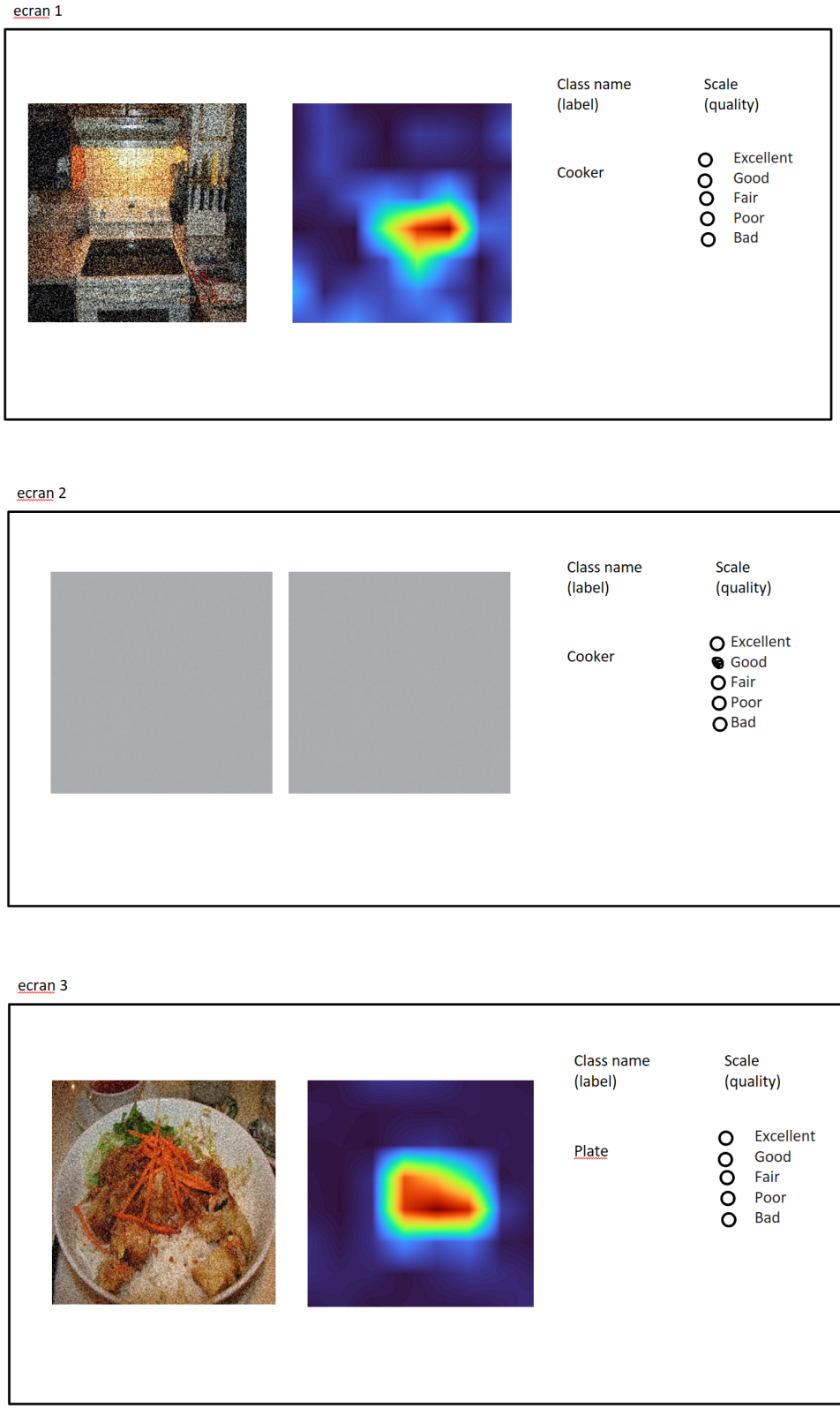

Figure 1: Mockup de l'interface de l'évaluation

Scenario interactif

La séance dynamique de visualisation pour un observateur est composée 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].

Scénario d'évaluation

Un participant réalise un scénario d'évaluation. Il n'est pas encore défini si chaque participant réalise le même, ou un différent. Pour pouvoir s'adapter aux deux cas, la séquence paire d'image/explication ne doit pas être écrite en dure dans l'interface, mais configurable à l'aide d'un fichier CSV. À charge au backend de générer le frontend correspondant (ou au frontend d'adapter son interface en conséquence). Le nom du fichier doit contenir l'identifiant du scénario.

Un fichier de description de scénario devra avoir les informations suivantes (une ligne par page) :

- Identifiant de l'image sélectionnée
- Identifiant de l'altération effectuée
- Identifiant de la méthode d'explication sélectionnée
- Identifiant de la classe prédite
- Identifiant de la classe réelle

ces fichiers pourront être générés automatiquement ou manuellement lorsque le protocole d'évaluation sera défini.

Fichier résultat

Le backend doit sauver un fichier CSV résultat par scénario exécuté. Le nom du fichier doit contenir l'identifiant de l'utilisateur.

Le fichier résultat doit contenir (une ligne par page) :

- toutes les colonnes du fichier de description
- temps de réponse
- score

Pour information, ces fichiers seront utilisés par la suite pour calculer différentes statistiques.

Données utilisées

On peut imaginer l'arborescence de données utilisée de la façon suivante : un dossier image qui contient un dossier par image nommé avec son identifiant. À l'intérieur de chaque dossier d'image, un dossier par altération nommé par son identifiant (nous devons être capable de connaître l'altération effectuée en fonction de l'identifiant). À l'intérieur de ces

dossiers un fichier pour l'image altérée (<identifiant_image>.png) et un fichier pour chaque explication (<identifiant explication>.png).

Le frontend sera capable d'afficher les bonnes images dans son interface en construisant les chemins à partir du fichier de description de scénario.

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)*