Apprendre sans données, une approche d'apprentissage automatique guidée par simulation en 3D pour l'extraction robuste de texte de cartes nationales d'identité

Edouard Bertrand*, Anaïs Druart*, Axel Thevenot*, Christophe Rodrigues*

* Léonard De Vinci Pôle Universitaire, Research Center, 92 916 Paris La Défense, France

Nous nous intéressons à l'extraction de texte à partir de cartes nationales d'identité (CNI) françaises. Les techniques de reconnaissance optique de caractères (OCR) sont performantes, mais les résultats peuvent être mitigés pour les numérisations à partir de smartphone en raison d'une grande variabilité des angles de vue, de l'éclairage, de la qualité de la caméra... A notre connaissance, il n'existe pas de base de donnée publique d'images de CNI. Dans ce contexte, est-il possible d'utiliser l'apprentissage automatique sans données annotées voire sans données du tout ?

En l'absence de données, nous proposons de créer une simulation qui permet de projeter des documents structurés synthétiques dans un environnement 3D. De cette façon, nous pouvons reproduire et contrôler les différentes difficultés qui seraient rencontrées avec l'image réelle d'un document. Cette solution nous permet de générer des exemples synthétiques d'apprentissage pour entraîner nos modèles d'intelligence artificielle de réseaux de neurones. Les étapes clés de la chaîne de traitements de création de données sont illustrées sur la figure 1.

Nous construisons d'abord une CNI vierge à partir d'un échantillon de Wikipédia, sur laquelle nous rajoutons du texte et une image de profil générée à l'aide du modèle StyleGAN. Cette image est alors projetée dans un environnement 3D fabriqué sur le logiciel Blender. Nous faisons ensuite varier la texture de la table et les paramètres de la simulation (tels que la position de la carte, la distance de la caméra, la distance focale...) afin d'obtenir une variété de rendus synthétiques annotés, pouvant être utilisés pour entraîner des modèles d'intelligence artificielle. Les jeux de données que nous avons ainsi générés sont désormais accessibles au public. [1] Néanmoins, en décidant de simuler entièrement les données, nous nous exposons au

FIG. 1 – *Etapes clés de la chaîne de traitement pour la création de données*

problème de la représentativité des exemples générés. Afin de minimiser ce risque, nous proposons l'utilisation d'une procédure d'apprentissage actif guidée par la lisibilité pour régler

1. https://github.com/ResearchPaper0/Learning-without-real-data

automatiquement les différents paramètres de la simulation et couvrir au mieux les zones les plus réalistes de l'espace de simulation.

Nous décidons de guider notre modèle d'apprentissage actif en fonction de la lisibilité des CNI plutôt que des performances du modèle afin de réduire son coût, en partant du principe que des exemples de cartes à la frontière du lisible pourraient être plus intéressants pour le modèle si nous voulons l'entraîner dans des conditions réalistes. Cette hypothèse nous permet de limiter le processus d'apprentissage actif à la seule sélection des paramètres, sans chercher à obtenir un retour d'information de la part de l'entraînement et de l'évaluation des modèles. Concrètement, notre modèle d'apprentissage actif apprend à prédire pour un ensemble de paramètres de simulation donnés à quel point l'image que ces paramètres permettent de générer serait lisible. Pour déterminer la lisibilité d'une carte, nous mesurons la part de texte (box dice) correctement détecté sur l'image par un modèle oracle auquel on a fourni les coordonnées exactes de la carte. On définit par la suite comme "lisible" toute image dont la box dice est supérieure au seuil manuellement défini de 0,7. On génère ensuite les images tirés des paramètres dont la prédiction de la lisibilité est la plus incertaine. On répète ce processus itérativement en ajoutant à chaque boucle les nouvelles images générées aux données d'apprentissage du modèle d'apprentissage actif pour finalement obtenir un jeu de données d'images de cartes d'identité synthétiques à la frontière de la lisibilité. Cette méthode est schématisé sur la figure 2. Les

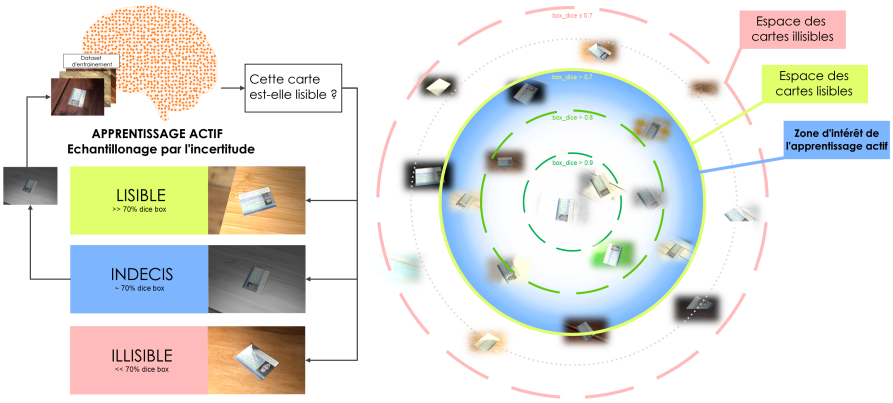

FIG. 2 – *Chaîne de traitement de l'apprentissage actif (gauche) et Espace de lisibilié des CNI (droite)*

principales contributions de ce travail sont les suivantes :

- Construction d'une chaîne de traitement pour créer un jeu de données réaliste d'images synthétiques entièrement annotées afin d'entraîner des modèles d'extraction d'informations dans des documents structurés.
- Création d'un modèle capable de localiser une CNI dans une image, recadrer et redresser la CNI pour une meilleure extraction de texte par un OCR.
- Mise en ligne d'un dataset d'images de CNI françaises synthétiques réalistes entièrement annotées en termes de contenu textuel ainsi que de position d'information permettant à d'autres chercheurs d'entraîner leurs propres modèles.