Object Detection

Introduction

L'intelligence artificielle est un domaine à l'intersection, principalement, des mathématiques et de l'informatique. Cependant, les récentes découvertes de ces 10 à 20 dernières années dans l'IA a permis d'étendre cette intersection à de multiples disciplines. La vision par ordinateur (en anglais : **Computer Vision**), est une branche de l'intelligence artificielle à la croisée de la physique optique, la géométrie et la statistique. Ce domaine consiste à reproduire et à automatiser les tâches que le système visuel humain peut effectuer à partir d'images et/ou de vidéos numériques. La détection d'objet (en anglais : **Object Detection**) est l'une des méthodes utilisée en vision par ordinateur. Cette méthode consiste à localiser avec précision une ou plusieurs instances de classes sur une image.

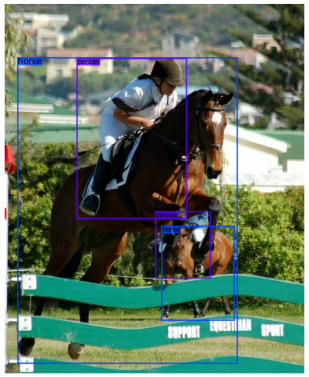

FIGURE 1 – Exemple d'une image annotée (Dataset : VOC2007)

Modèles/Méthodes

Depuis l'un des premiers réseaux de neurones convolutifs (CNN) connus et utilisés avec succès, **LeNet** en 1998, les CNN ont connu de nombreuses évolutions. La création d'ensembles de données massifs comme ImageNet, l'organisation de compétitions liées à ces ensembles de données, et l'utilisation généralisée des unités de traitement graphique (GPU) ont eu un impact direct sur la recherche et le développement d'outils dans ce domaine. Ainsi, durant la dernière décennie, les technologies en détection d'objets se sont développées sur une approche similaire, mais avec toutefois des différences notables, notamment avec des architectures telles que :

- **R-CNN** (**R**egion Based **C**onvolutional **N**eural **N**etworks, 2013)
- **Fast R-CNN** (Avril 2015)
- **Faster R-CNN** (Juin 2015)
- **YOLO** (**Y**ou **O**nly **L**ook **O**nce, Juin 2015)
- **SSD** (**S**ingle **S**hot MultiBox **D**etector, Décembre 2015)
- **ViT** (**Vi**sion **T**ransformer, 2020)

Outils/Métriques

IoU (Intersection over Union)

IoU, (en français : "Intersection sur Union"), est une métrique évaluant l'intersection sur l'union de deux cadres englobants (boîtes englobantes).

$$\text{IoU}(A, B) = \begin{cases} 0 & \text{si } A \cup B = \emptyset \\ \frac{A \cap B}{A \cup B} & \text{sinon} \end{cases}$$

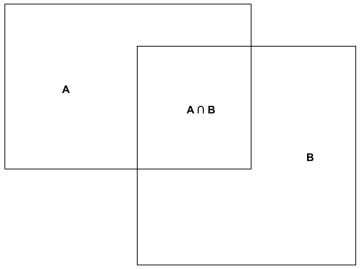

FIGURE 2 – Intersection over Union

Dans le contexte de la détection d'objet, A réprésente une prédiction réalisée par un modèle et B un rectangle cartérisant la présence réelle d'un objet. IoU permet alors de mesurer la prédiction effectuée par le modèle par rapport au réel.

Algorithme Intersection sur l'union
Fonction intersection_over_union(box_p, box_t) :
On considère que box_p et box_t soient de la forme [x1, y1, x2, y2]
x1 = max(box_p[0], box_t[0])
y1 = max(box_p[1], box_t[1])
x2 = min(box_p[2], box_t[2])
y2 = min(box_p[3], box_t[3])
intersection = (x2 - x1) × (y2 - y1)
area1 = abs((box_p[2] - box_p[0]) × (box_p[3] - box_p[1]))
area2 = abs((box_t[2] - box_t[0]) × (box_t[3] - box_t[1]))
Retourner intersection / (area1 + area2 - intersection)

À noter que les cadres englobants peuvent également avoir un angle rotationnel par rapport au centre de l'image. Ce qui peut compliquer le calcul par la forme de l'intersection qui n'est ni carré ni un rectangle.

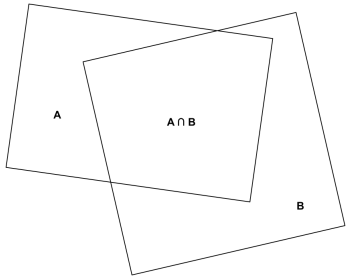

FIGURE 3 – Rotated Intersection over Union

Generalized Intersection over Union (GIoU)

GIoU (en français : **l'Intersection sur l'Union Généralisée**), une métrique et une fonction de coût développées par une équipe de Stanford, présente une corrélation significative avec l'IoU. Cette caractéristique permet de simplifier le processus d'apprentissage en fournissant une estimation de la précision du modèle sans nécessiter le calcul précis de cette dernière. Pour plus de détails, https://giou.stanford.edu/.

Non Max Suppression (NMS)

NMS est une méthode utilisée pour la détection d'objets. Tous les modèles vu auparavant peuvent prédire en fonction de leurs caractéristiques des centaines de cadres englobants sur l'image. D'où l'intérêt d'utiliser cette méthode dont la principale fonction est de nettoyer l'image des cadres redondants et non pertinents :

1. Conserver les cadres englobants dont la valeur de fiabilité est supérieure ou égale à un seuil de confiance (valeur entre 0 et 1 de la qualité de la prédiction selon le modèle). Tri décroissant des cadres englobants selon leur valeur de fiabilité.
2. Sélectionner du cadre englobant prédit avec la meilleure fiabilité et le considérer comme le meilleur.
3. Parcourir le reste du tri, supprimer ceux dont la valeur IoU avec le "meilleur" cadre est égale ou supérieur à un seuil définit. Alternative, supprimer ceux dont la valeur IoU avec le "meilleur" cadre est égale ou supérieur à un seuil définit et avoir la même classe.
4. Conserver le meilleur cadre et répéter le processus jusqu'à que la liste de tri de prédictions soit vide.

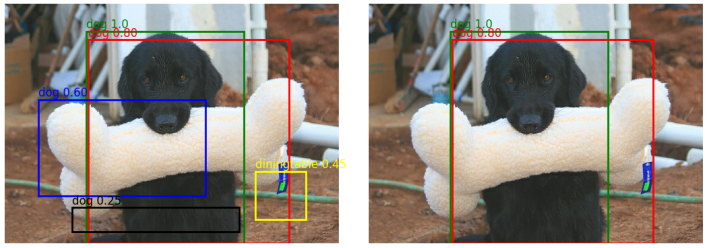

FIGURE 4 – Avant NMS - Après NMS (Version non alternative)

La troisième étape du processus de Non Max Suppression dépend du problème et représente un défi majeur dans le domaine de la détection d'objets. Il s'agit de parvenir à éliminer les prédictions redondantes du modèle tout en étant capable de détecter avec précision des objets qui sont proches, voire superposés. Cette problématique peut être résolue en optant par la segmentation. En effet, nous pouvons voir la segmentation comme une version de YOLO extrême, car la subdivision de l'image est maximale. Une cellule correspond à un pixel, et cette dernière a pour nouveau rôle de classifier au lieu d'être responsable de la prédiction d'un objet. Cette classification par cellule permet de prédire les contours réelles des objets, contrairement à la détection d'objets.

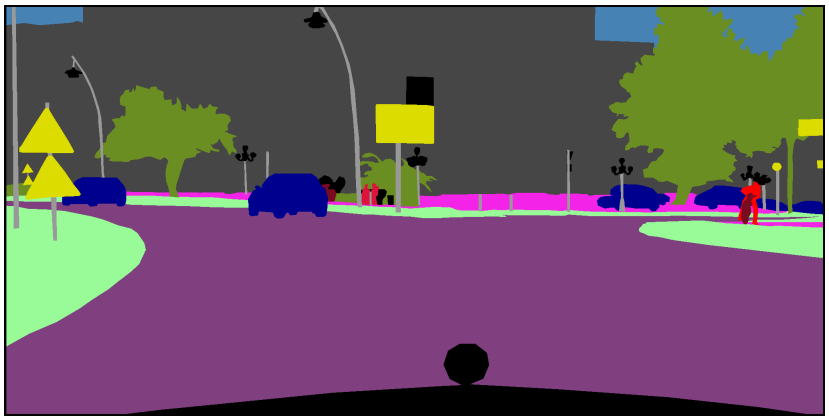

FIGURE 5 – Exemple d'une image segmentée par instance (Dataset : Cityscapes)