Fiche de Lecture 3

Élodie Bouilleteau

Mardi 13 Novembre 2018

1 Référence de l'article

Titre : You Only Look Once : Unified, Real-Time Object Detection

Auteurs : Joseph Redmon, Santosh Divvala, Ross Girshick et Ali Farhadi.

Université : University of Washington

Date de parution : 8 juin 2015

Date de dernière révision : 9 mai 2016

Lien : `https://arxiv.org/abs/1506.02640`

2 Situation des auteurs

Joseph Redmon est un informaticien passionné d'apprentissage automatique, d'analyse de données, ainsi que de conception et de mise en oeuvre de programmes de bas niveau. Il travaille sur la vision par ordinateur. Il possède son propre site web : `https://pjreddie.com/`.

Santosh Divvala est chercheur scientifique chez AI2. Son intérêt principal est la vision par ordinateur, en particulier le problème de compréhension des images.

Ross Girshick est chercheur à Facebook AI Research (FAIR) et travaille sur la vision par ordinateur et l'apprentissage automatique.

Ali Farhadi est professeur associé au département d'informatique et d'ingénierie de l'Université de Washington. Il s'intéresse principalement à la vision par ordinateur, à l'apprentissage automatique, à l'intersection du langage naturel et de la vision, à l'analyse du rôle de la sémantique dans la compréhension visuelle et au raisonnement visuel

3 Introduction

Cette publication traite d'une nouvelle approche de détection d'objet visuel en temps réel. Ils décrivent la détection d'objet comme un problème de régression dans des boîtes englobantes séparées dans l'espace associées à

des probabilités de classe. Un seul réseau de neurones prédit les limites et les probabilités de classe directement à partir d'images complètes.

4 Méthode de détection

Leur méthode de détection se base sur un seul réseau de neurones. Ils unissent les composants de la détection d'objet dans un seul réseau. Le réseau utilise des caractéristiques de l'ensemble de l'image pour prédire des cadre de sélections. Il prédit tous les cadres de sélection de toutes les classes possibles sur une seule image en même temps.

L'algorithme va séparer l'image d'entrée en une grille de dimension SxS. Chaque case de la grille va prédire B boîtes et les scores de confiance pour ses boîtes. Ces scores de confiance reflètent le degré de confiance du modèle sur le fait que la boîte contient un objet et la précision avec laquelle il prédit la boîte. Je ne détaille pas le calcul du score de confidence.

Chaque boîte contient 5 prédictions : x, y, w, h et le score de confiance. Les coordonnées (x, y) représentent le centre de la boîte par rapport aux limites de la cellule de la grille. La largeur et la hauteur sont prédites par rapport à l'image entière. Chaque cellule de grille prédit également C probabilité conditionnelle, P(Classe i | Objet). Ces probabilités sont conditionnées par la cellule de la grille contenant un objet.

5 Désigne du réseau

Ce modèle de détection d'objet est implémenter comme un réseau de neurone de convolution et est évaluer via le jeu de données de détection de PASCAL VOC. Les couches convolutives initiales du réseau extraient les caractéristiques de l'image, tandis que les couches entièrement connectées prédisent les probabilités et les coordonnées de sortie. Le réseau comporte 24 couches convolutives suivies de 2 couches entièrement connectées. Il utilise des couches de réduction 1x1 suivi de couches convolutives de 3x3.

6 Performance

Le modèle YOLO traite les images en temps réel à 45 images par seconde sans traitement par lots sur un Titan X GPU. Une version réduite du réseau, Fast YOLO, traite 155 images par seconde. Par rapport aux autres systèmes

de détection, YOLO fait plus d'erreurs de localisation mais est moins susceptible de prédire de faux positifs. YOLO atteint plus de deux fois la précision moyenne des autres systèmes de détection en temps réel.

7 Conclusion

Cette article explique une méthode en temps réel de détection d'objet en utilisant un réseau de neurone de convolution sur l'image complète.