Fiche de Lecture 5

Élodie Bouilleteau

Mercredi 15 Novembre 2018

1 Référence de l'article

Titre :Bounding Box Embedding for Single Shot Person Instance Segmentation

Auteurs : Jacob Richeimer et Jonathan Mitchell.

Date de parution : 20 juillet 2018

Sujets : Computer Vision and Pattern Recognition.

Lien vers le document : `https://arxiv.org/abs/1807.07674`

2 Situation des auteurs

Jacob Richeimer est le directeur de l'entreprise nommé OCTI, INC. L'entreprise OCTI INC développe la technologie de vision par ordinateur et d'apprentissage automatique de l'application de messagerie vidéo en réalité augmentée.

Notamment : estimation de la pose humaine 3D mobile en temps réel, segmentation d'instances sémantiques et reconnaissance des actions squelettiques. Langues : Python. Outils : Keras, Tensorflow.

Jonathan Mitchell est ingénieur dans cette entreprise et se concentre sur la vision par ordinateur, l'apprentissage en profondeur (deep learning), et en particulier la détection de pose humaine et la segmentation d'instances.

3 Introduction

Cette publication présente une nouvelle approche pour la tâche de segmentation d'instance d'une personne en utilisant un modèle en single-shot (en un coup). Le modèle proposer emploie un réseau de neurone de convolution qui est entraîné pour prédire aussi bien les masques de segmentation par classe (ici les personnes) que les boîtes englobantes des instances d'objet (de personnes) auxquelles chaque pixel appartient. Les auteurs de cette article cherche à associer un pixel à l'instance de l'objet (de la personne) auquel il appartient.

4 information

4.1 Deux approches de détection

1. L'approche "Top-down" consiste à d'abord localiser les instances de l'objet puis à obtenir le masques de pixel pour chaque instance détecter.
2. l'approche "Bottom-up" consiste à d'abord déterminer la classe d'objet de chaque pixel puis à les grouper en une seule instance d'objet.

Dans cet article, les auteurs adoptent l'approchent "Bottom-up" et propose une méthode simple qui ne requiert qu'un minimum de calculs en plus des actuelle approche de l'état de l'art sur la segmentation sémantique catégorique.

L'approche "Bottom-up" demande, après la segmentation sémantique, d'ajouter des étapes supplémentaires de regroupement de pixels en instances.

5 Méthode de détection

Ils ont développés une approche "single-shot" de segmentation d'instance de personne. Pour une image donnée, cela consiste d'abord à classifier chaque pixels comme appartenant à une personne ou au fond de l'image, puis à regrouper les pixels qualifiés comme personne dans une instance de personne.

5.1 Segmentation sémantique de personne

Ils utilise un réseau de neurone de convolution standard pour faire la segmentation. Ils prédissent pour chaque emplacement de pixel, la probabilité qu'il appartient à une instance de personne.

5.2 Proposition de boîte de détection

Dans le but de prédire l'instance de personne à laquelle appartient chaque pixel, chaque emplacement de pixel est associé à une "proposition" ou à une "ancre". Une "ancre" est une boîte englobante qui est centré sur ce pixel et à une largeur w et une hauteur h.

Pour chaque pixel, le réseau prédit les décalages (dx, dy, dw, dh) entre sa boîte d'ancrage et la boîte englobante de l'instance à laquelle il appartient.

5.3 Regroupement de pixels en instance

Dans cet article, les auteurs ont choisi de faire correspondre les coordonnées des pixels pour qu'ils soient compris dans les coordonnées du cadre de sélection de l'instance à laquelle appartient chaque pixel.

Cette méthode n'est pas parfaite, en effet, il est possible que plusieurs instances se chevauchent et que les boîtes englobantes soient presque identiques.

Le point positif est que cette méthode est facile à implémenter à la suite des architecture de segmentation sémantique déjà existante.

La méthode regroupement des pixels à 2 étapes qui se suivent :

1. La sélection de boîte globale est la première étape. Les auteurs traitent la valeur de la probabilité attribuée à chaque emplacement de pixel comme la valeur de confiance associée au cadre de sélection prévu pour cet emplacement. Ils recueillent toutes les boîtes englobantes prédites qui correspondent aux pics locaux de la carte de segmentation sémantique et ont un indice de confiance supérieur à un seuil (t = 0.6). La suppression non-maximal est ensuite appliquée aux boîtes englobantes collectées pour obtenir les détection globales de la boîte englobante Bg pour l'image donnée.
2. L'assignation de pixel à une instance est la seconde étape. Sp est l'ensemble des pixels trouvé comme personne. Chacun de ses pixels a besoin d'être assigner à une des instances globales des boîtes englobantes Bg de l'étape précédente. Pour chaque location de pixel xi dans Sp, ils prennent la boîte englobante correspondante bi, et effectue l'intersection sur l'union entre bi et chacune des boîte englobante globale Bg. La localisation du pixel est ensuite assigné à la boîte englobante Bg. Si toutes les boîtes de Bg se chevauchent avec bi avec un score IoU inférieur à un seuil tiou, alors la localisation du pixel xi est supprimé de Sp. Ce pixel est supposé être un résultat faux positif de la segmentation sémantique et n'est attribué à aucune des instances.

6 Jeu de données

Ils ont utilisé le jeu de données COCO pour l'apprentissage et l'évaluation. Ils ont réaliser l'apprentissage seulement avec les images d'apprentissage contenant des annotation de personnes, soit 64 115 images.

7 Architecture du modèle

Les auteurs utilisent le réseau de base ResNet-50, qu'ils ont choisi pour son équilibre riche en fonctionnalités et sa consommation de mémoire, auquel ils y attachent le module Atrous Spatial Pyramid Pooling et les couches de décodeurs DeepLabv3 +.

La seule divergence par rapport à DeepLabv3 + réside dans le fait qu'en plus de la couche de convolution finale 1x1 avec un filtre par classe (dans leur cas, il n'existe qu'une seule classe) au-dessus des cartes de caractéristiques de sortie du décodeur, ils disposent d'une couche de convolution supplémentaire 1x1 avec quatre filtres pour prédire les décalages de la boîte englobante dense.

8 Conclusion

Cette article présente une méthode unique pour la segmentation d'instances d'objets et montré son efficacité lors de la segmentation d'instances de personnes.