Fully Convolutional Networks for Semantic Segmentation

Nasreddine Menacer

Master Sciences et Ingénierie des données Co-accréditation Université de Rouen et INSA

Résumé

Ce document est un résumé de l'article intitulé «Fully Convolutional Networks for Semantic Segmentation » publié en 2016 par Jonathan Long, Evan Shelhamer, et Trevor Darrell, cet article présente un rapport d'étude qui détaille l'architecture des réseaux entièrement convolutionels et leurs applications notamment de prédiction de classe par pixel (segmentation sémantique d'images). Et comment des réseaux de classification contemporains (Alex Net, VGG Net...) sont utilisés pour des taches similaires.

Introduction

De nos jours les réseaux à convolutions conduisent à des avancées considérables en termes de reconnaissance d'images. Les ConvNet améliorent non seulement la classification des images entières [1, 2, 3], mais progressent également dans les tâches locales comme la détection d'objets, de boîtes englobantes et la prédiction de parties et de points-clés [4, 5]. Le but aujourd'hui et d'arriver pour des taches de segmentation à faire une prédiction pixel par pixel, Dans Ce document les chercheures mettent le point sur les réseaux entièrement convolutionels (FCN), entrainé de bout en bout, et démontre que ces Derniers dépasse largement l'état de l'art dans les taches de segmentation sémantique.

Travaux Connexes

L'approche présenté dans l'article s'appuie sur les succès récents des réseaux profonds pour la classification des images [1, 2, 3] et le « transfert learning » [6, 7], Le calcul convolutionnel a également été exploité dans ces réseaux multicouches. Plusieurs travaux récents ont utilisé des ConvNets pour résoudre des problèmes de prédiction, y compris de segmentation sémantique par Ning et al. [8], Farabet et al [9], et Pinheiro et Collobert [10].

Fully Convolutional Networks

Un FCN est un réseau de neurone convolutionnel (CNN) sans couches denses. Cette caractéristique apporte de multiples avantages. D'abord, enlever les couches denses, permet de travailler avec des images de tailles variables. Deuxièmement, dans les ConvNets standard, les couches denses contiennent un très grand nombre de paramètres. Ainsi, éviter les couches denses réduites fortement le nombre de paramètres.

Vers les réseaux de classification aux réseaux segmentation

Les réseaux de reconnaissance d'image typiques, notamment LeNet [13], AlexNet [1], prennent des entrées de taille fixe et produisent des sorties non spatiales. Les couches entièrement connectées de ces réseaux ont des dimensions fixes et rejettent les coordonnées spatiales. Cependant, ces couches peuvent également être considérées comme des convolutions avec des filtres qui couvrent l'ensemble de leurs régions d'entrée. Cela les transforme en réseaux entièrement convolutionels qui prennent des entrées de différents tailles et donnent en sortie des cartes de caractéristiques. Pour construire ces FCN les auteurs vont utiliser des classifieurs typique comme VGG16, AlexNet, et Google Net et effectuer des modifications sur les couches denses. Les expériences et les résultats de tous les FCN générées sont les suivants.

Expériences

Pour leurs expériences les auteurs vont utiliser : FCN-VGG16, FCN-AlexNet et FCN-GoogleNet. Les bases de test : PASCAL-Voc, Nyud-V2 et SIFT-Flow.
Et pour mesurer les performances les métriques utilisées sont IOU « Intersection Over Union » et « Pixel accuracy ».

Résultats

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [12]	47.9	-	-
SDS [17]	52.6	51.6	~ 50 s
FCN-8s	**62.7**	**62.2**	~ **175 ms**

On remarque que les FCN-8s donne une amélioration relative de 20% par rapport à l'état de l'art sur les ensembles de tests PASCAL VOC 2011 et 2012 et réduisent le temps d'inférence.

	pixel acc.	mean acc.	mean IU	f.w. IU
Gupta *et al.* [15]	60.3	-	28.6	47.0
FCN-32s RGB	60.0	42.2	29.2	43.9
FCN-32s RGBD	61.5	42.4	30.5	45.5
FCN-32s HHA	57.1	35.2	24.2	40.4
FCN-32s RGB-HHA	64.3	44.9	32.8	48.0
FCN-16s RGB-HHA	**65.4**	**46.1**	**34.0**	**49.5**

Sur le tableau précédent on remarque que les résultats des réseaux entrainés sur des bases de données d'images en utilisant quatre canaux (un canal de profondeur en plus des canaux RGB), ne sont pas mauvais en regardant le « pixel accuracy » qui s'élève à 65,4 %

Conclusion

Dans cet article, une nouvelle approche basée sur les FCN pour la segmentation sémantique d'images a été présentée. Les résultats ont montré que l'utilisation de cette approche peut dépasser l'état de l'art dans les taches de segmentation et l'avantage de ces réseaux c'est qu'on ne trouve pas des couches denses, ce qui présente de nombreux avantages en termes de réduction du nombre de paramètres, et qui permet de travailler avec des tailles d'entrée variables et de conserver les informations spatiales des images.

Référence

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014.

[4] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Partbased r-cnns for fine-grained category detection. In Computer Vision–ECCV 2014, pages 834–849. Springer, 2014.

[5] J. Long, N. Zhang, and T. Darrell. Do convnets learn correspondence? In NIPS, 2014.

[6] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. In Computer Vision (ICCV), 2013 IEEE International Conference on, pages 633–640. IEEE, 2013.

[7] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In Computer Vision–ECCV 2014, pages 818–833. Springer, 2014.

[8] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. CoRR, abs/1406.2984, 2014.

[9] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013.

[10] P. H. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In ICML, 2014.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition, 2014.

[12] Y. Ganin and V. Lempitsky. N4-fields: Neural network nearest neighbor fields for image transforms. In ACCV, 2014.

[13] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to hand-written zip code recognition. In Neural Computation, 1989.