FULLY CONVOLUTIONAL NETWORKS FOR SEMENTIC SEGMENTATION

ANDRIAMAHERISOA Liantsoa
Master 2 : Sciences et Ingénierie des données
Co-accréditation Université de ROUEN et INSA
tolojanahary-liantso.andriamaherisoa@etu.univ-rouen.fr

Résumé

Ce document est un résumé de l'article "Fully Convolutional Networks for Semantic Segmentation" publié en 2016 par Jonathan Long, Evan Shelhamer, et Trevor Darrell.
Dans cet article est présenté le rapport détaillé de leur étude concernant des architectures de réseaux de neurone entièrement convolutifs, en anglais : *"Fully convolutional networks"*, appliqués à la prédiction d'étiquettes ou de classes par pixels d'une image (Segmentation sémantique d'images). Cet article illustre les thèories, les démarches, et les expériences, misent en oeuvre par les chercheurs afin de créer le môdèle FCN à l'aide des réseaux de classifications d'images déjà éxistants et déjà pré-entrainnés *(AlexNet [8], VGGnet[12], et GoogLeNet[13])* tout en améliorant les poids de ces derniers, d'où la méthode de *"backpropagation"*. Cette méthode de Transfert learning est ce que l'on appelle *"fine-tunning"*.

Introduction

A l'époque actuel, les réseaux de neurone à convolutions conduisent à des avancées considérables dans le domaine de la reconnaissance d'images.
Les ConvNets améliorent non seulement la classification d'images entières [8, 12, 13], mais aussi la détecion d'objets, de boîtes englobantes [11, 4, 7] et la prédiction de parties et de points clés [15, 9].
La prochaine étape est d'arriver, pour les tâches de segmentation, à faire une prédiction pixel par pixel.
Dans ce document, les chercheurs mettent le point sur les réseaux entièrement convolutifs (FCN), entrainés de bout en bout, et démontre que ces derniers dépassent largement l'état de l'art pour des tâches de segmenatation sémantique.

Travaux relatifs

Les approches sitées dans ce document s'appuient sur les récents succès du deep nets pour la classification d'images [8, 12, 13] et l'apprentissage par transfert [2, 1]. Le transfert a d'abord été appliqué à de divers tâches de reconnaissance [2, 1], puis sur la détection, et sur deux exemples de segmentation sémantique dans la proposition de modèles de classifieur hybride [4, 6, 5].
La suite est de réarchitecturer et finaliser les réseaux de classification pour prédire directement et de manière dense la segmentation sémantique.

Fully Convolutional Networks

Un FCN est un réseau de neurone convolutif (CNN) sans couches denses.
Les principaux avantages de l'utilisation des réseaux de neurones entièrement convolutifs stipulés dans cet article sont les suivants:

- enlever les couches denses, pour pouvoir travailler avec des images de tailles arbitraires, paru dans le Matan et al. [14], (reconnaissance de chaînes de chiffres).

- dans les ConvNets standards, les couches denses contiennent un très grand nombre de paramètres. Ainsi, éviter les couches denses réduit fortement le nombre de paramètres.

- la rapidité et l'efficacité du modèle grâce à son système de fenêtre coulissante de détection, utilisé dans plusieurs recherches telles que: la segmentation sémantique par Pinheiro and Collobert [10] ou bien dans la réstauration d'images par Eigen et al. [3] (inférence entièrement convolutionnelle).

De classifieurs à FCN denses

Les chercheurs ont délibérement choisis les meilleurs architectures de reconnaissance d'images, notament AlexNet [8], VGGnets [12] et GoogLeNet [13] qui ont eu les meilleurs résultats en ILSVRC14. Et y ont modifié les couches denses. Avant ces modifications, ces derniers 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.
Ainsi, cela les transforme en FCN qui prennent des entrées de différentes tailles et qui retournent en sortie des cartes de caractéristiques.

	FCN-AlexNet	FCN-VGG16	FCN-GoogLeNet[4]
mean IU	39.8	**56.0**	42.5
forward time	50 ms	210 ms	59 ms
conv. layers	8	16	22
parameters	57M	134M	6M
rf size	355	404	907
max stride	32	32	32

Expériences

Concernant les expériences, les auteurs ont utilisé : FCN-VGG16, FCN-AlexNet et FCN-GoogleNet sur les données: PASCAL-voc, NYUD-v2 et SIFLT-FLOW.

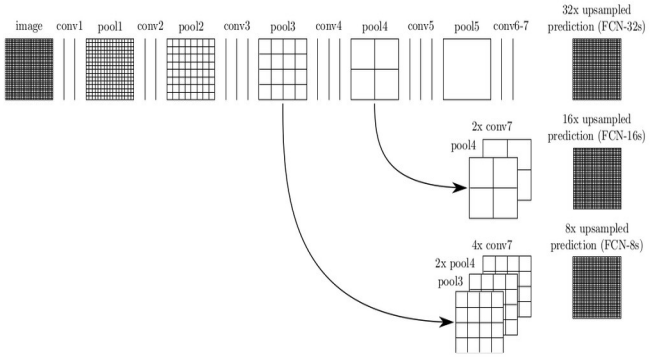
Pour les calculs de préformances, les mesures utilisées sont IOU *"Intersection Over Union"* et *"Pixel eccurancy"*.

Architécture du FCN

Chaque couche de données dans le FCN est de 3-dimensions h x w x d, où h et w sont des dimensions spatials et d, la dimensions de l'entité ou du canal de couleur.

La première couche est l'image, avec la taille de pixel h x w, et d canaux de couleurs. Les emplacements dans les couches supérieurs correspondent aux emplacements dans l'image à laquelle ils sont connectés, qui sont appelés: *"leurs champs recéptifs"*.

Les ConvNets sont contruits sur l'invariance de translation.
Les composantes basiques (Convolution, pooling, et fonction d'activations) agissent en régions d'entrée locale, et dépendent uniquement des coordonnées spatiales relatives.



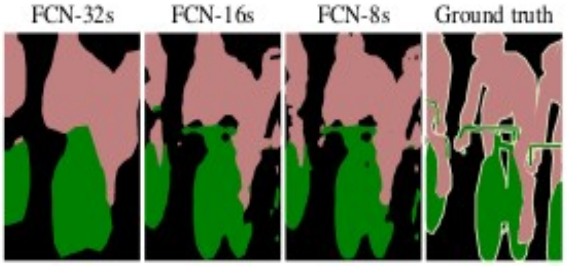
Architecture de segmentation

Les couches pooling (couches de mise en commun) et de prédiction sont représentés comme des grilles qui révèlent une grossièreté spatiale relative, tandis que les couches intermédiaires sont représentées par des lignes verticales.

Première rangée: (FCN-32s)
Suréchantillonnes 32 prédictions en pixels en une seule étape.

Deuxième rangée: (FCN-16s)
Combinaison des prédictions de la couche finale et de la couche pool4, à la foulée 16, permettent de prédire les détails les plus fins, tout en conservant une sémantique de haut niveau d'information.

Troisième rangée: (FCN-8s)
Des prédictions supplémentaires du pool3, à la fouclée 8, fournissent une précision supplémentaires.



Résultats

Les résultats de leurs expériences démontrent que la 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 réduit le temps d'inférence.

	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s-fixed	83.0	59.7	45.4	72.0
FCN-32s	89.1	73.3	59.4	81.4
FCN-16s	90.0	75.7	62.4	83.0
FCN-8s	**90.3**	**75.9**	**62.7**	**83.2**

Conclusion

Il a été présenté dans cet article une approche basée sur les FCN pour la segmentation sémantique d'images. Les résultats ont montré que l'utilisation de cette approche peut dépasser l'état de l'art dans les tâches de segmentation. Ce qui est le fruit d'un démarche logique en utilisant les meilleurs architectures de réseaux déjà existants en les améliorant et en modifiant leurs couches denses. L'avantage de ces réseaux est donc, le non existance des couches denses, ce qui permet de travailler avec des tailles d'entrées variables et conserver les informations spatiales des images. Les résultats obtenues témoignent de l'éfficacité de leurs approches.

Références

[1] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. arXiv preprint arXiv:1412.1283, 2014.

[2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In ICML, 2014.

[3] 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.

[4] 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.

[5] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In ECCV. Springer, 2014.

[6] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In European Conference on Computer Vision (ECCV), 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV, 2014.

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

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

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

[11] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In ICLR, 2014.

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

[13] 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.

[14] O. Matan, C. J. Burges, Y. LeCun, and J. S. Denker. Multidigit recognition using a space displacement neural network.
In NIPS, pages 488–495. Citeseer, 1991.

[15] 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.