Synthèse bibliographique basée sur l'article : Fully Convolutional Networks for Semantic Segmentation

Lamyae Laaroussi, Oumaima Alehyane
Master 2 SID
Enseignants : Clément Chatelain, Romain Hérault, Rachel Blin.

Abstract
Les réseaux de neurones convolutifs (CNN) sont une forme particulière de réseau neuronal multi-couches. Les auteurs montrent que les CNN entraînés de bout-en-bout et pixel par pixel, entraînent une amélioration de la précision dans plusieurs tâches de segmentation.
Leur but est de créer un réseau entièrement convolutif (FCN) qui prend une entrée de taille arbitraire et produit sa sortie correspondante.

Keywords : CNN, FCN, segmentation sémantique, VGG16

Introduction

Les CNN sont présents dans de multiples applications, comme dans le domaine de la vision par ordinateur et de la reconnaissance d'objets. Les auteurs montrent que les réseaux convolutifs standards pour la reconnaissance d'objets peuvent être réinterprétés comme des FCN qui produisent des cartes de sortie grossières utilisées pour la segmentation sémantique (figure 1), pour prendre des entrées de tailles variées et produire les sorties correspondantes. En effet les CNN type pour la reconnaissance prennent des entrées de taille fixe et produisent des sorties non spatiales, ayant des couches entièrement connectées qui rejettent les informations de type spatial.
On peut ainsi interpréter les couches entièrement connectées comme des convolutions avec des noyaux qui couvrent toute leur région d'entrée.

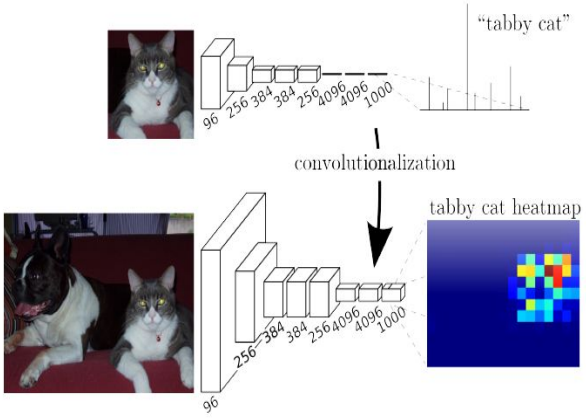

Figure 1 : Transformation des couches entièrement connectées en couches de convolutions qui peuvent générer une carte thermique (heatmap). [1]

Travaux connexes

Les auteurs s'inspirent des succès de réseaux profonds pour la classification d'image et le transfert d'apprentissage. L'idée d'exploiter des réseaux convolutifs avec des entrées de tailles arbitraires est apparu avec Matan *et al*. [2] pour la détection. Des travaux antérieurs ont exploité des réseaux convolutifs dans des cas de prédiction dense pour la segmentation sémantique par Ning *et al.* [3], Farabet *et al.*[4], and Pinheiro and Collobert [5].

Architecture du FCN

Les auteurs ont sélectionné comme base le réseau VGG16 (modèle CNN avec 16 couches dont 13 de convolutions et 3 denses) pour le transformer en un réseau

entièrement convolutif (figure 1). Toutes les couches entièrement connectées du VGG sont converties en couches convolutionnelles pour traiter des entrées de différentes tailles. La couche finale de classification est supprimée et remplacée par des convolutions de noyau 1x1 pour chaque classe. Une couche de déconvolution est ajoutée pour effectuer un sur-échantillonnage en sorties denses en pixels. Cela donne le premier classifieur entièrement convolutif donnant de bons résultats sur les métriques standards (précision par pixel et l'IoU), mais qui a une sortie encore assez grossière. Les réseaux de neurones convolutifs permettent de passer d'une information locale ("où") à une information de plus en plus globale ("quoi"). Introduction des couches dites "skip" (saut) pour préserver certaines informations locales dans les couches de sortie grossières. Ces couches combinent la couche de prédiction finale avec les couches précédentes, qui ont un pas de décalage plus fin et plus d'informations locales (figure 2).

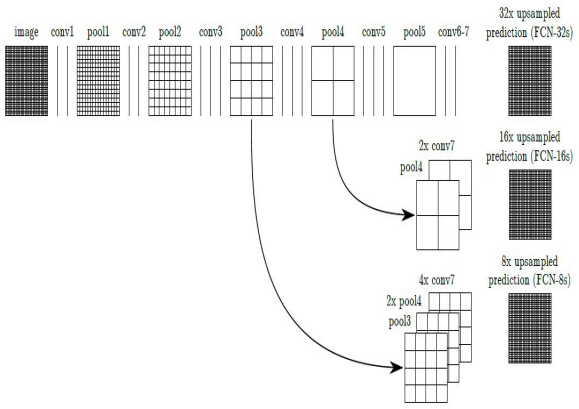

Figure 2: Combinaison d'informations grossières et fines. [1]

Ils divisent d'abord le pas de décalage de sortie en 2 et fusionnent les informations d'une couche de 16 pixels de pas de décalage. Ils combinent les prédictions de pool4 avec les prédictions sur-échantillonnées calculées en plus de conv7 et les somment. Ensuite, les prédictions de pas de décalage 16 sont sur-échantillonnées à la taille de l'image. Le réseau résultant est appelé FCN-16.
Le même processus est effectué pour une couche de pas de 8 pixels, combinant les couches de saut et les prédictions de conv7, qu'ils appellent FCN-8. Le réseau d'origine sans couches de saut est appelé FCN-32.
Les FCN-16 et FCN-8 sont également appris de bout en bout et initialisés avec le dernier réseau plus grossier (FCN-32 et FCN-16 respectivement).
FCN32: on obtient une segmentation plus globale après le sur-échantillonnage (déconvolution pour élargir l'image) avec stride 32 (pas de décalage).
FCN8 : on obtient une segmentation plus précise (plus affinée) après le sur-échantillonnage (déconvolution pour élargir l'image) avec stride 8 (pas de décalage) car ça prend en compte des voisinages plus proches.
⇒ moins de sémantique globale
=> plus de résolution

Expériences

Les auteurs évaluent leurs modèles sur les jeux de données PASCAL VOC, NYUDv2, SIFT Flow, et PASCAL-Context.

Résultats

Le FCN-8 donne une amélioration relative de 30% sur les ensembles de test de PASCAL VOC 2011 et 2012 avec une inférence et un apprentissage plus rapides par rapport à SDS [6], et à R-CNN [7] (Table 1).

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [5]	47.9	-	-
SDS [14]	52.6	51.6	~ 50 s
FCN-8s	67.5	67.2	~ 100 ms

Table 1: Résultats du FCN-8 avec les données de PASCAL VOC 2011 et 2012 [1].

Le FCN-8 donne une amélioration des performances sur les ensembles de test de SIFT Flow par rapport à d'autres réseaux convolutionnels (Table 2).

	pixel acc.	mean acc.	mean IU	f.w. IU	geom. acc.
Liu *et al.* [57]	76.7	-	-	-	-
Tighe *et al.* [58] transfer	-	-	-	-	90.8
Tighe *et al.* [59] SVM	75.6	41.1	-	-	-
Tighe *et al.* [59] SVM+MRF	78.6	39.2	-	-	-
Farabet *et al.* [12] natural	72.3	50.8	-	-	-
Farabet *et al.* [12] balanced	78.5	29.6	-	-	-
Pinheiro *et al.* [13]	77.7	29.8	-	-	-
FCN-8s	85.9	53.9	41.2	77.2	94.6

Table 2: Résultats obtenus pour les données de SIFT Flow [1].

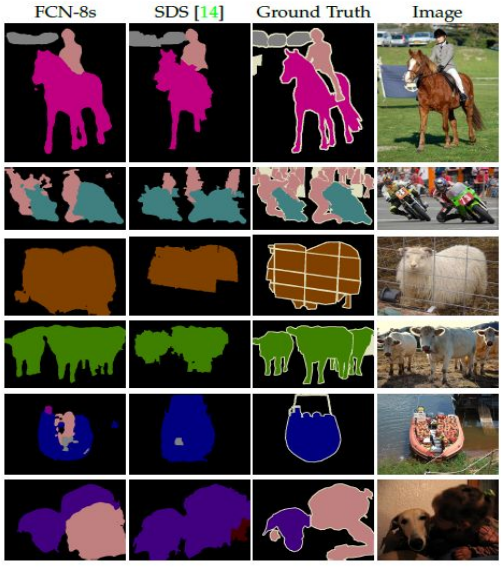

Figure 3: Les réseaux entièrement convolutifs améliorent les performances sur PASCAL[1].

La colonne de gauche montre une sortie plus précise du réseau FCN-8. La seconde montre la sortie de la méthode SDS du Hariharan et al. [6].

En première ligne on remarque de fines structures récupérées (cheval et cavalier). En deuxième ligne on remarque une bonne capacité de séparer les objets en interaction étroite.
Les cinquième et sixième lignes indiquent les cas de défaillance: Le réseau voit les gilets de sauvetage dans un bateau comme des personnes et confond les cheveux humains avec le chien.

Analyses

Avantages:

- Segmentation sémantique est possible avec entrées de tailles variables.
- Prédiction bout en bout et pixel par pixel.

Inconvénient:

- Donne de meilleures performances qu'avec beaucoup de données.
- Nécessite de travailler avec un GPU puissant.

Conclusion

La conversion des réseaux convolutifs classiques en réseaux entièrement convolutifs est une approche relativement simple mais puissante.

Les réseaux entièrement convolutifs sont une classe riche de modèles. Ils permettent la segmentation sémantique en apprenant les images de tailles arbitraires de bout en bout et pixel par pixel.

Références :

[1]: Evan Shelhamer, Jonathan Long, Trevor Darrell, ”Fully Convolutional Networks for Semantic Segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[2]: O. Matan, C. J. Burges, Y. LeCun, and J. S. Denker, “Multi-digit recognition using a space displacement neural network,” in NIPS, 1991, pp. 488–495.

[3]: F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E.Barbano, “Toward automatic phenotyping of developing embryos from videos,” Image Processing, IEEE Transactions on, vol. 14, no. 9,pp. 1360–1371, 2005.

[4]: C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” *PAMI*, 2013.

[5]: P. H. Pinheiro and R. Collobert, “Recurrent convolutional neural networks for scene labeling,” in ICML, 2014.

[6]: B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, “Simultaneous ´detection and segmentation,” in *ECCV*, 2014.

[7]:R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Region-based convolutional networks for accurate object detection and segmentation,” PAMI, 2015.