

Apprendre une meilleure cartographie sémantique grâce aux données d’Observation de la Terre et aux données géographiques participatives

Nicolas AUDEBERT^{1,2}, Bertrand LE SAUX¹, Sébastien LEFÈVRE²

¹ONERA, *The French Aerospace Lab*
F-91761 Palaiseau, France

²Univ. Bretagne-Sud, UMR 6074, IRISA
F-56000 Vannes, France

`nicolas.audebert@onera.fr, bertrand.le_saux@onera.fr, sebastien.lefevre@irisa.fr`

Résumé – Ce travail porte sur l’utilisation des données *OpenStreetMap* (OSM) pour la segmentation sémantique d’images de télédétection. Suite aux succès récents obtenus grâce aux réseaux de neurones profonds pour la classification de données multispectrales, hyperspectrales, radar et LiDAR, nous nous intéressons à l’intégration de données géographiques, rarement utilisées dans ces procédés d’apprentissage. En particulier, nous présentons deux architectures permettant d’apprendre simultanément à partir de données OSM et d’images aériennes ou satellites, l’une basée sur le raffinement de cartes et l’autre sur la fusion de données hétérogènes. Nos résultats sur le jeu de données ISPRS Potsdam montrent que l’utilisation des données OSM améliore la qualité des cartes obtenues et accélère la convergence des réseaux lors de la phase d’apprentissage.

Abstract – In this work, we investigate the use of OpenStreetMap (OSM) data for semantic labeling of Earth Observation images. Deep neural networks have been used in the past for remote sensing data classification from various sensors, including multispectral, hyperspectral, SAR and LiDAR data. While OSM has already been used as ground truth data for training such networks, this abundant data source remains rarely exploited as an input information layer. In this paper, we study fusion based architectures and coarse-to-fine segmentation to include the OSM layer into semantic labeling of aerial and satellite images. We illustrate how these methods can be successfully used on the ISPRS Potsdam dataset and show that they significantly improve both the accuracy performance and the convergence speed of the networks in the training phase.

1 Introduction

La cartographie automatisée à partir de données de télédétection par apprentissage profond a fait l’objet de plusieurs études récentes dans la littérature. En particulier, plusieurs approches issues de la vision par ordinateur ont été appliquées avec succès sur des images Rouge-Vert-Bleu (RVB) de télédétection, notamment les *Convolutional Neural Networks* (CNN) pour l’étude de l’occupation des sols [3] et les *Fully Convolutional Networks* (FCN) pour la cartographie de zones urbaines [1], établissant un nouvel état-de-l’art sur ces tâches. En outre, les données de télédétection étant rarement restreintes au simple triplet RVB, plusieurs articles ont étendu ces travaux aux cas multi-capteurs, notamment en fusionnant données optique et LiDAR [1] ou encore données multispectrales et radar [8]. Toutefois, les bases de données géographiques telles que *OpenStreetMap* (OSM) sont encore peu utilisées autrement qu’en vérité terrain là où des annotations expertes ne sont pas disponibles. Pourtant, ces bases de données constituent une source d’information sémantique extrêmement riche, par exemple pour les empreintes au sol des routes et des bâtiments. La question à laquelle nous nous intéressons est donc la suivante : comment peut-on utiliser les données *OpenStreetMap* pour améliorer la cartographie automatisée d’images de télédétection ?

2 Contexte

La classification au niveau pixel d’images aériennes et satellites à partir de CNN a été étudiée de nombreuses reprises depuis les travaux de [11]. Les travaux les plus récents utilisent les modèles qualifiés de FCN [9], obtenant d’excellents résultats sur des images haute résolution en zone urbaine [13], encore améliorés par une régularisation utilisée en post-traitement [15]. Ces mêmes modèles ont été appliqués avec succès à l’extraction d’empreintes de routes et de bâtiments dans des images satellites [10] à des résolutions bien inférieures. Bien que ces travaux se bornent à l’imagerie optique, la fusion de données hétérogènes a également été étudiée par la suite. Notamment, des architectures de réseaux de neurones profonds efficaces utilisant des FCN à double entrée ont été proposées pour la fusion multispectral/SAR [8] et RVB/LiDAR [1].

Pour autant, peu de travaux se sont penchés sur l’intégration des données OSM depuis l’ouverture du site en 2004. Cette source reste principalement utilisée comme vérité terrain pour la détection de routes et de bâtiments [11, 10] dans un contexte d’apprentissage supervisé. Seuls quelques travaux s’appuyant sur les forêts aléatoires ont mis en œuvre l’utilisation des couches OSM comme données d’apprentissages, par exemple pour la prédiction de zones climatologiques locales [4].

3 Méthode

3.1 Segmentation par raffinement

Lorsque les classes d'intérêt de la segmentation sémantique sont déjà présentes dans les données OSM, comme pour les bâtiments ou les routes, il est possible de les utiliser comme des approximations de la vérité terrain afin de n'apprendre qu'un raffinement de celle-ci pour obtenir une carte haute résolution. Ce procédé s'apparente ainsi à l'apprentissage par résidu [7].

Ici, nous utilisons un simple FCN à deux couches afin de convertir les données raster OSM en cartes sémantiques approchant la vérité terrain, et nous appellerons ce modèle OSMNet par la suite. Les données optiques sont traitées par un FCN dérivé du modèle SegNet [2] en suivant l'approche de [1]. SegNet est un réseau de neurones encodeur-décodeur approximant la projection d'une image dans un espace de cartes sémantiques à la même résolution. En utilisant ces deux modèles, nous pouvons alors calculer une carte de prédiction moyenne combinant les deux sources d'entrée. Dans ce cas, si I constitue l'image d'entrée, O la donnée OSM, P_{opt} la fonction de prédiction de SegNet et P_{osm} la fonction de prédiction de OSMNet, la prédiction finale P se calcule par :

$$P(I, O) = \frac{1}{2}(P_{opt}(I) + P_{osm}(O)). \quad (1)$$

Si $P_{osm}(O)$ est une bonne approximation de la vérité terrain VT , alors l'optimisation cherche à minimiser :

$$P_{opt} \propto VT - P_{osm}(O) \ll VT, \quad (2)$$

ce qui devrait avoir un effet similaire à l'apprentissage par résidu [7].

En outre, pour raffiner encore cette prédiction moyenne, nous utilisons la correction résiduelle [1]. Ce réseau de neurones est un FCN à trois couches permettant d'apprendre la correction à appliquer à la prédiction moyenne pour exploiter la complémentarité des données OSM et optiques. En notant C la fonction de prédiction du module de correction résiduelle :

$$P(I, O) = \frac{1}{2}(P_{opt}(I) + P_{osm}(O)) + C(Z_{opt}(I), Z_{osm}(O)), \quad (3)$$

où Z_{opt} et Z_{osm} sont les cartes d'activation finales de SegNet et OSMNet, respectivement.

Dans ce cadre, l'apprentissage par résidu peut être conçu comme la modélisation d'un terme de correction d'erreur. Le procédé complet est illustré dans la figure 1a.

3.2 FCN à double entrée

Les FCN à plusieurs entrées ont été la source de plusieurs études par le passé, notamment dans le cadre du traitement d'images RVB+profondeur (ou 2,5D) [5]. Dans cet article, nous utilisons l'architecture FuseNet [6] pour combiner données optiques et OSM. FuseNet se réapproprie le modèle SegNet en utilisant non pas un, mais deux encodeurs, un pour chaque source. Après chaque succession de 2 ou 3 convolutions, les

cartes d'activations neuronales du deuxième encodeur sont sommées aux activations du premier. Cela permet au réseau d'apprendre une représentation conjointe des données exploitant les deux modalités. Un unique décodeur transforme ensuite cette représentation en sur-échantillonnant et en effectuant la classification dans l'espace des classes sémantiques pour chaque pixel. Comme détaillé dans la figure 1b, une branche principale apprend la représentation conjointe tandis que la branche auxiliaire n'apprend que les activations liées aux données OSM. En notant P la fonction de prédiction de FuseNet, I l'image d'entrée, O la donnée OSM, $E_i^{\{opt,osm\}}$ les cartes d'activations après le $i^{\text{ème}}$ bloc de l'encodeur, $B_i^{\{opt,osm\}}$ les fonctions représentées par le $i^{\text{ème}}$ bloc de convolutions et D la fonction du décodeur :

$$P(I, O) = D(E_5^{opt}(I, O)) \quad (4)$$

et

$$E_{i+1}^{opt}(I, O) = B_i^{opt}(E_i^{opt}(I, O)) + B_i^{osm}(E_i^{osm}(O)). \quad (5)$$

4 Expériences

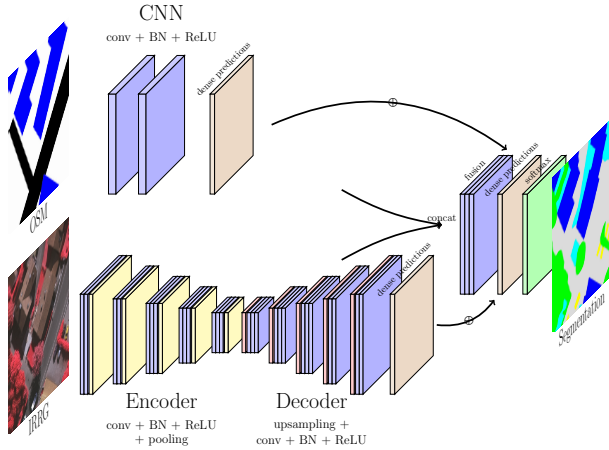
4.1 Jeu de données

Le jeu de données *ISPRS Potsdam Semantic Labeling* [12] est constitué de 38 images aériennes acquises sur la ville de Potsdam (Allemagne) à une résolution de 5cm/pixel. Les images sont composées de 4 bandes spectrales (IR, R, V et B) et contiennent 6000 × 6000 pixels. Une vérité terrain dense est fournie pour 24 tuiles contenant les classes de routes, bâtiments, végétation basse, arbres, véhicules et une classe de rejet (cf. Figure 2). Les tuiles étant géo-référencées, nous générons les rasters OSM associés contenant les empreintes de routes, bâtiments, des zones de végétation et de l'eau en utilisant Maperitive¹. Les résultats sont obtenus par validation croisée sur 3 divisions du jeu de données.

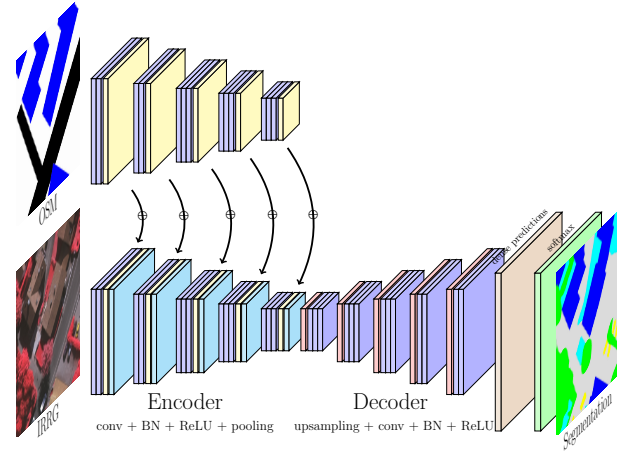
4.2 Cadre expérimental

Lors de l'apprentissage sur le jeu de données ISPRS Potsdam, nous extrayons aléatoirement des images de 128 × 128 pixels dans le jeu d'apprentissage, en appliquant optionnellement des symétries afin d'augmenter le nombre d'exemples. Le réseau est entraîné par descente de gradient en traitant 10 images en parallèle, comme suggéré dans [1]. Les taux d'apprentissage sont initialisés à 0,005 pour l'encodeur et 0,01 pour le décodeur, puis sont divisés par 2 tous les 30 000 itérations. L'encodeur pour les données RVB est initialisé à partir des poids de VGG-16 [14] pré-entraînés sur ImageNet, les autres poids étant initialisés aléatoirement. Le taux d'apprentissage pour l'encodeur est fixé à la moitié du taux d'apprentissage pour le décodeur (soit 0,005). À l'évaluation, chaque tuile est traitée via une fenêtre glissante de taille 128 × 128 avec un pas de 64. Les prédictions surnuméraires sur un même pixel

1. <http://maperitive.net/>



(a) Fusion RVB/OSM par correction résiduelle [1].



(b) FuseNet [6] appliqué aux données optiques et OSM.

FIGURE 1 – Réseaux de neurones profonds pour le traitement simultané de données optiques et *OpenStreetMap*.



FIGURE 2 – ISPRS Potsdam : données RVB et VT.

sont moyennées pour lisser la cartographie finale. L'apprentissage jusqu'à convergence (150 000 itérations) prend environ 20 heures sur une carte graphique NVIDIA K20c, tandis que l'évaluation sur le jeu de test prend moins de 30 minutes.

4.3 Résultats

Les résultats obtenus sur les données de validation de l'ISPRS Potsdam sont détaillés dans le tableau 1. En suivant les recommandations des auteurs du jeu de données, nous indiquons le pourcentage global de pixels correctement classés et les scores F1 pour chaque classe, calculés sur une version alternative de la vérité terrain dans laquelle les bordures ont été érodées de 3 pixels.

$$F1_i = 2 \frac{precision_i \times rappel_i}{precision_i + rappel_i}, \quad (6)$$

$$rappel_i = \frac{tp_i}{C_i}, \quad precision_i = \frac{tp_i}{P_i}, \quad (7)$$

où tp_i est le nombre de vrais positifs de la classe i , C_i le nombre de pixels appartenant à la classe i et P_i le nombre de pixels associés à la classe i par le modèle.

Comme attendu, l'inclusion de données OSM améliore les performances de classification du modèle, notamment pour les routes et les bâtiments qui bénéficient de la présence des empreintes au sol dans les couches OSM. En effet, cette infor-

mation additionnelle permet de discriminer certaines ambiguïtés où un modèle purement optique aurait des difficultés, par exemple pour distinguer un parking au sol et sur un toit, à l'apparence très similaire.

Par ailleurs, l'intégration des données OSM dans l'apprentissage permet d'accélérer la convergence du modèle. En effet, sur le même jeu de données, le modèle SegNet appris par raffinement depuis OSM nécessite 25 % d'itérations en moins que le SegNet RVB classique pour converger à la même performance de classification finale. En outre, le minimum local obtenu est meilleur, avec une fonction de coût à 0,39 contre 0,45 dans le cas précédent, ce qui laisse présager une meilleure capacité de généralisation. Enfin, l'inclusion des données OSM rend la sortie du réseau visuellement plus cohérente et mieux structurée spatialement, comme illustré dans la figure 3.

5 Conclusion

Nous avons présenté dans cet article comment intégrer de l'information géographique dans une méthode d'apprentissage profond pour la cartographie d'images de télédétection. Nous avons proposé deux méthodes, l'une par raffinement de cartes et la seconde par fusion de données hétérogènes, permettant de combiner imagerie optique et données *OpenStreetMap*. Nous avons validé ces deux méthodes sur le jeu de données ISPRS Potsdam sur lequel nous avons amélioré la performance de classification du modèle de 2,5 % par rapport à un FCN classique. Nos travaux montrent qu'il est possible d'utiliser efficacement de l'information géographique non-visuelle pour la cartographie automatisée dans le cadre des réseaux de neurones convolutifs profonds. Ce faisant, nous espérons pouvoir bénéficier de cartes sémantiques de meilleure qualité exploitant toutes les sources d'information à notre disposition, aussi bien acquises par capteurs aériens et satellites que collectées par des plateformes de cartographie collaboratives.

TABLE 1 – Résultats sur le jeu de données ISPRS Potsdam (score F1 par classe et pourcentage global de pixels bien classés).

Méthode	routes	bâtiments	vég. basse	arbres	véhicules	Global
SegNet RVB	93,0 %	92,9 %	85,0 %	85,1 %	95,1 %	89,7 %
RC RVB+OSM	93,9 %	92,8 %	85,1 %	85,2 %	95,8 %	90,6 %
FuseNet	95,3 %	95,9 %	86,3 %	85,1 %	96,8 %	92,3 %

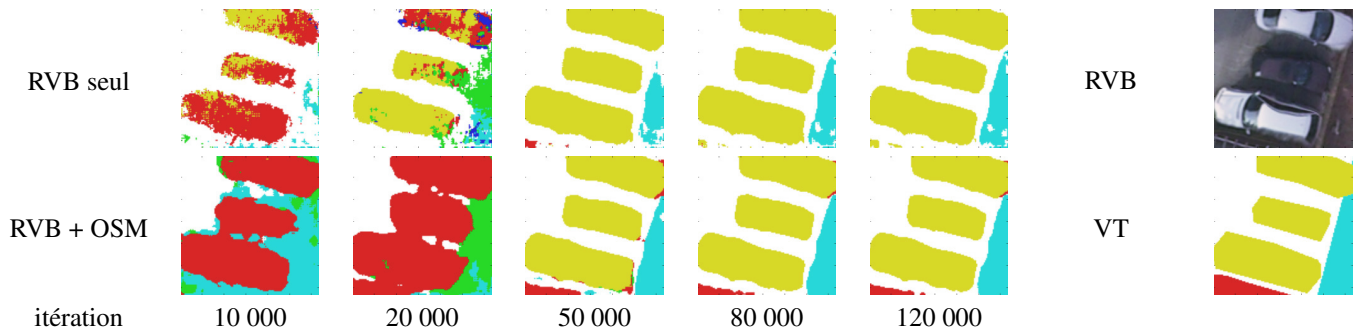


FIGURE 3 – Évolution des prédictions de SegNet RVB et RBV+OSM. L'ajout de OSM rend les prédictions visuellement plus structurées, dès le début de l'apprentissage.

Légende : blanc : routes, bleu : bâtiments, cyan : végétation basse, vert : arbres, jaune : véhicules, rouge : autre

Références

- [1] N. Audebert, B. Le Saux, and S. Lefèvre. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks. In *Computer Vision – ACCV 2016*, pages 180–196. Springer, Cham, Nov. 2016.
- [2] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet : A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv :1511.00561 [cs]*, Nov. 2015.
- [3] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva. Land use classification in remote sensing images by convolutional neural networks. *arXiv :1508.00092 [cs]*, Aug. 2015.
- [4] O. Danylo, L. See, B. Bechtel, D. Schepaschenko, and S. Fritz. Contributing to WUDAPT : A Local Climate Zone Classification of Two Cities in Ukraine. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(5) :1841–1853, May 2016.
- [5] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, Sept. 2015.
- [6] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. FuseNet : Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision – ACCV 2016*, pages 213–228. Springer, Cham, Nov. 2016.
- [7] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.
- [8] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu. FusioNet : A Two-Stream Convolutional Neural Network for Urban Scene Classification using PolSAR and Hyperspectral Data. In *2017 Joint Urban Remote Sensing Event (JURSE)*, Mar. 2017.
- [9] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015.
- [10] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Fully convolutional neural networks for remote sensing image classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5071–5074, July 2016.
- [11] V. Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.
- [12] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf. The ISPRS benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1 :3, 2012.
- [13] J. Sherrah. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. *arXiv :1606.02585 [cs]*, June 2016.
- [14] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv :1409.1556 [cs]*, Sept. 2014.
- [15] M. Volpi and D. Tuia. Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2) :881–893, 2017.