Classification de Scènes de Nuages de Points 3D par Réseau Convolutionnel Profond Voxelique Multi-échelles

Xavier Roynard[1] Jean-Emmanuel Deschaud[1] François Goulette[1]

[1] Mines ParisTech, PSL Research University, Centre for Robotics
60 boulevard Saint-Michel, 75006 Paris

xavier.roynard@mines-paristech.fr

1 Introduction

Dans cet article nous décrivons un nouveau réseau de neurones convolutionnel (CNN) qui classifie des scènes urbaines et d'intérieur de nuages de points 3D. Ce réseau (décrit en section 3) prend en entrée les voisinages d'un point à plusieurs échelles pour pouvoir interpréter à la fois le contexte éloigné du point et la forme fine de l'objet. On montre en section 4 que notre réseau obtient de meilleurs résultats qu'un réseau mono-échelle, de plus sur le benchmark reduced-8 de Semantic3D [3], ce réseau, classé second, bat l'état de l'art des méthodes de classification par points (celles qui n'utilisent pas d'étape de régularisation).

2 État de l'art

Au cours des trois dernières années, il y a eu un nombre croissant de travaux visant à adapter les méthodes d'apprentissage profond ou à introduire de nouvelles approches "profondes" pour classifier les nuages de points 3D. Parmis les approches développées on peut distinguer quatre philosophies : (a) en projetant le nuage de points sur des images, puis en classifiant chaque image avec des réseaux de segmentation d'images [2], (b) en projetant le nuage dans un grille d'occupation 3D comme dans VoxNet [7] et [5], (c) en travaillant plus sur des graphes du nuage de points via des CRFs comme SegCloud [11] ou plus en faisant des convolutions sur des graphes comme SPGraph [6], (d) en prenant directement le nuage de point en entrée d'un réseau capable de respecter les propriétés ensembliste d'un nuage comme PointNet [8].

Les réseaux convolutionnels comme VoxNet sont performants pour classifier des objets simples [12], mais ne sont pas adaptés à la classification de scènes entières avec des objets à plusieurs échelles.

3 Architecture de notre réseau de classification par points

L'architecture choisie est inspirée de celle décrite dans [10] qui obtient de bons résultats en 2D. Notre réseau suit l'architecture suivante : $Conv(32,3,1,0) \rightarrow Conv(32,3,1,0) \rightarrow MaxPool(2) \rightarrow Conv(64,3,1,0) \rightarrow Conv(64,3,1,0) \rightarrow MaxPool(2) \rightarrow FC(1024) \rightarrow FC(N_c)$. [1] Où N_c est le nombre de classes, et chaque couche $Conv$ et FC est suivie par $BatchNorm \rightarrow PReLU$ et un bloc Squeeze-and-Excitation tiré de [4], excepté la dernière couche FC qui est suivie par une non-linéarité de type $SoftMax$. Ce réseau prend en entrée une grille d'occupation 3D de taille $32 \times 32 \times 32$, où chaque voxel contient 0 (vide) ou 1 (occupé) et a une taille de 10cm $\times$ 10cm $\times$ 10cm.

Pour que notre réseau soit capable de tirer de l'information à la fois du contexte à grande échelle et de la forme de l'objet à petite échelle, nous prenons plusieurs versions de la partie convolutionnelle de ce réseau (sans les couches FC). On donne en entrée de chaque version une grille d'occupation de même taille $32 \times 32 \times 32$, mais avec des tailles de voxels différentes (par exemple 5cm, 10cm et 15cm). Chaque version nous donne en sortie un vecteur de caractéristiques de taille 1024, qu'on concatène ensemble avant de les donner à des couches FC. Voir la figure 1 pour une représentation graphique de notre réseau appelé : MS3_DeepVoxScene (ou abrégé en MS3_DVS).

1. $Conv(n,k,s,p)$ représente une couche convolutionnelle avec n cartes de caractéristiques en sortie, un noyau de taille $k \times k \times k$ un pas de s et un padding de p de chaque coté de la grille.

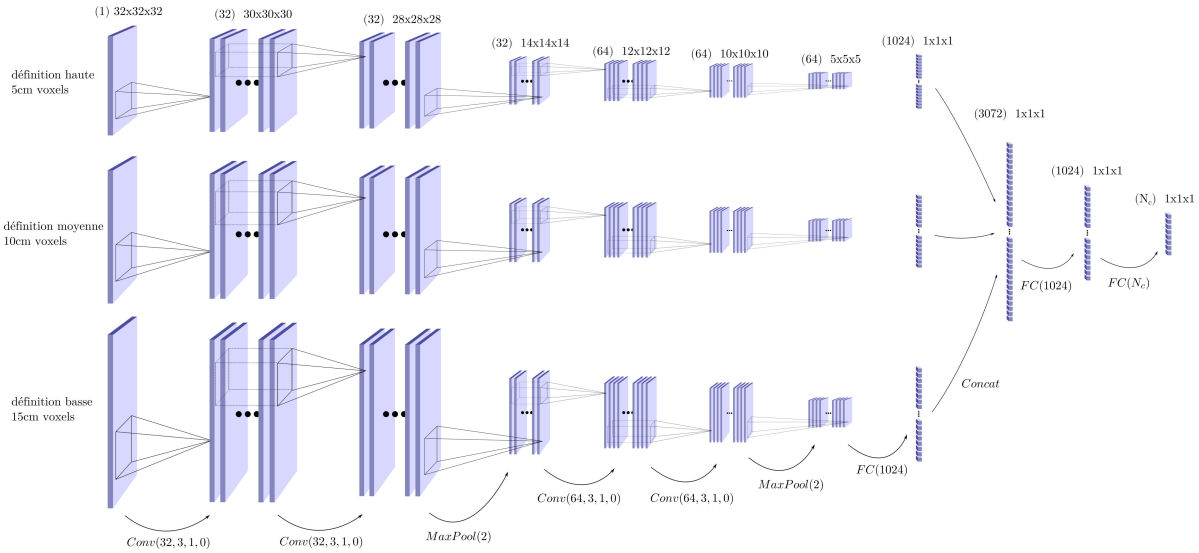

FIGURE 1 – Notre architecture de réseau voxelique multi-échelles : MS3_DeepVoxScene (tous les tenseurs sont représentés en 2D au lieu de 3D pour la simplicité).

4 Résultats expérimentaux

4.1 Comparaison avec l'état de l'art

Sur le jeu de données Semantic3D [3], notre méthode obtient de meilleurs résultats que toutes les méthodes de classification par points (c'est-à-dire sans régularisation). De meilleurs résultats pourrait surement être obtenus en ajoutant par exemple une CRF après la classification. Voir le tableau 1 pour une comparaison avec l'état de l'art (incluant les méthodes avec régularisation) sur Semantic3D.

				IoU par classe							
Rang	**Méthode**	**IoU Moyenné**	**Précision Globale**	terrain artificiel	terrain naturel	végétation haute	végétation basse	bâtiments	paysage rigide	artefacts d'acquisition	voiture
1	SPGraph[6]	**73.2**%	**94.0**%	**97.4**%	**92.6**%	**87.9**%	**44.0**%	**93.2**%	31.0%	**63.5**%	76.2%
2	MS3_DVS(Ours)	65.3%	88.4%	83.0%	67.2%	83.8%	36.7%	92.4%	**31.3**%	50.0%	**78.2**%
3	RF_MSSF	62.7%	90.3%	87.6%	80.3%	81.8%	36.4%	92.2%	24.1%	42.6%	56.6%
4	SegCloud[11]	61.3%	88.1%	83.9%	66.0%	86.0%	40.5%	91.1%	30.9%	27.5%	64.3%
5	SnapNet_[2]	59.1%	88.6%	82.0%	77.3%	79.7%	22.9%	91.1%	18.4%	37.3%	64.4%
9	MS1_DVS(Ours)	57.1%	84.8%	82.7%	53.1%	83.8%	28.7%	89.9%	23.6%	29.8%	65.0%

TABLE 1 – Top-5 sur le benchmark reduced-8 de Semantic3D. MS3_DVS est notre architecture MS3_DeepVoxScene avec pour taille de voxels 5 cm, 10 cm et 15 cm et MS1_DVS est MS1_DeepVoxScene avec taille de voxels 10 cm (ajouté pour comparaison avec une architecture non multi-échelles).

Sur le jeu de données S3DIS [1], on observe une confusion entre les classes *mur* et *tableau* (et plus légèrement avec *pilier*, *colonne*, *fenêtre* et *porte*), ceci s'explique principalement parce que ces classes sont très similaires géométriquement et notre réseau n'utilise pas la couleur. Voir le tableau 2 pour une comparaison avec l'état de l'art sur S3DIS.

4.2 Étude de l'intérêt des multi-échelles

Pour évaluer nos choix d'architecture nous avons testé cette tâche de classification par point avec un des premiers réseaux convolutionnels 3D : VoxNet [7]. Ceci nous permet à la fois de valider les choix pour l'architecture générique MS1_DeepVoxScene et de valider l'intérêt d'un réseau multi-échelles. Voir le tableau 3 pour une comparaison des 3 architectures. Pour une comparaison par classe entre MS1_DeepVoxScene et MS3_DeepVoxScene sur le jeu de données Paris-Lille-3D voir le tableau 4. On observe que l'utilisation de réseaux multi-échelles améliore les résultats sur la plupart des

Methode	Mean IoU	Mean Accuracy	Per class IoU (in %)												
			plafond	sol	mur	pilier	colonne	fenêtre	porte	chèse	table	étagère	canapé	tableau	autre
PointNet [8]	41.09%	48.98%	88.80	97.33	69.80	**0.05**	3.92	46.26	10.76	52.61	58.93	40.28	5.85	**26.38**	33.22
MS3_DVS(Ours)	46.32%	57.93%	79.03	88.07	53.55	0.00	**20.47**	29.01	37.29	68.84	63.72	47.44	**61.62**	16.50	36.64
SEGCloud [11]	48.92%	57.35%	90.06	96.05	69.86	0.00	18.37	38.35	23.12	75.89	70.40	58.42	40.88	12.96	41.60
SPG [6]	**54.67%**	**61.75%**	**91.49**	**97.89**	**75.89**	0.00	14.25	**51.34**	**52.29**	**86.35**	**77.40**	**65.49**	40.38	7.23	**50.67**

TABLE 2 – Résultats sur le 5ème fichier du jeu de données S3DIS [1].

classes, en particulier les *bâtiments*, *barrières* et *piétons* (surtout en rappel), tandis que la classe *voiture* perd beaucoup en précision.

Dataset	MS3_DVS	MS1_DVS	VoxNet [7]
Paris-Lille-3D	**89.29%**	88.23%	86.59%
Semantic3D	**79.36%**	74.05%	71.66%
S3DIS	**73.08%**	69.36%	66.28%

TABLE 3 – Comparaison des F1-scores de MS3_DVS, MS1_DVS et VoxNet [7] moyennés par classe. Pour chaque jeu de données le F1-score est moyenné sur les différents fichiers de validation.

Classe	Précision		Rappel	
	MS3_DVS	MS1_DVS	MS3_DVS	MS1_DVS
sol	**97.74%**	97.08%	**98.70%**	98.28%
bâtiments	**85.50%**	84.28%	**95.27%**	90.65%
poteaux	**93.30%**	92.27%	92.69%	**94.16%**
potelets	98.60%	**98.61%**	93.93%	**94.16%**
poubelles	**95.31%**	93.52%	79.60%	**80.91%**
barrières	**85.70%**	81.56%	**77.08%**	73.85%
piétons	**98.53%**	93.62%	**95.42%**	92.89%
voitures	93.51%	**96.41%**	**98.38%**	97.71%
naturel	**89.51%**	88.23%	**92.52%**	91.53%

TABLE 4 – Precision et rappel par classe moyenné sur les 4 fichiers de validaiton du jeu de données Paris-Lille-3D [9].

Références

[1] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[2] A. Boulch, B. L. Saux, and N. Audebert. Unstructured point cloud semantic labeling using deep segmentation networks. In *Eurographics Workshop on 3D Object Retrieval*, volume 2, page 1, 2017.

[3] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys. Semantic3d.net : A new large-scale point cloud cassification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-1-W1, pages 91–98, 2017.

[4] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation Networks. *ArXiv e-prints*, Sept. 2017.

[5] J. Huang and S. You. Point cloud labeling using 3d convolutional neural network. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2670–2675. IEEE, 2016.

[6] L. Landrieu and M. Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. *arXiv preprint arXiv :1711.09869*, Nov. 2017.

[7] D. Maturana and S. Scherer. Voxnet : A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 922–928. IEEE, 2015.

[8] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet : Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv :1612.00593*, 2016.

[9] X. Roynard, J.-E. Deschaud, and F. Goulette. Paris-Lille-3D : a large and high-quality ground truth urban point cloud dataset for automatic segmentation and classification. *ArXiv e-prints*, Nov. 2017.

[10] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv e-prints*, Sept. 2014.

[11] L. P. Tchapmi, C. B. Choy, I. Armeni, J. Gwak, and S. Savarese. Segcloud : Semantic segmentation of 3d point clouds. *arXiv preprint arXiv :1710.07563*, 2017.

[12] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets : A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015.