Extraction non supervisée de descripteurs pour des suivis environnementaux aériens

M. Laroze[1,3] R. Dambreville[2,3] C. Friguet[2] S. Lefèvre[2] D. Tuia[4]

[1] Université Rennes 1, IRISA, Rennes, France
[2] Université Bretagne Sud, IRISA, Vannes, France
[3] WIPSEA, Rennes, France
[4] Wageningen University and Research, Wageningen, Pays-Bas.
{mathieu.laroze, romain.dambreville, chloe.friguet, sebastien.lefevre}@irisa.fr, devis.tuia@wur.nl

Résumé

La détection d'objets sur images aériennes est une application spécifique de vision par ordinateur qui nécessite une vérité-terrain pour utiliser des méthodes de classification supervisée, mais celle-ci n'est pas toujours disponible. Afin de réduire le travail d'annotation et de connaissance a priori sur les données, nous évaluons la capacité d'un auto-encodeur convolutionnel à générer des caractéristiques pour la détection d'objets non supervisée.

Mots Clef

Auto-encodeur, détection d'objet, extraction de descripteurs

Abstract

Airborne images require a human pre-processing step to create supervised object detection algorithms. This required ground truth does not always exist when new data are acquired. To alleviate the annotation cost, we evaluate a convolutional autoencoder feature extraction technique for unsupervised object detection.

Keywords

Auto-encoder, object detection, feature extraction

1 Introduction

L'acquisition d'images aériennes pour des services de comptage et de suivi environnemental de zones protégées est un procédé de plus en plus fréquent, qui nécessite néanmoins une extraction manuelle de l'information recherchée : impact humain sur l'environnement, localisation de faune sauvage, suivi de milieux écologique... Les méthodes usuelles, dites supervisées, reposent aujourd'hui sur la nécessité d'une vérité-terrain préalablement établie par des humains mais l'annotation des images est une tâche fastidieuse et coûteuse d'un point de vue humain et économique, et les objets d'intérêt sont généralement peu nombreux lorsque l'on prend en considération la superficie de la zone couverte. Dans notre contexte, nous nous intéressons à l'étape d'initialisation d'un algorithme d'apprentissage automatique, par exemple en apprentissage actif, où la vérité-terrain ne serait pas existante. Une méthode non supervisée est alors nécessaire. Nous étudions ici l'apport d'un auto-encodeur, utilisé plus largement dans la littérature pour l'extraction de descripteurs, ses capacités de classification et pour sa capacité à s'adapter à différents paradigmes d'apprentissage [1]. Après avoir présenté plus particulièrement sa version convolutionnelle, nous évaluons l'apport de cette approche entièrement non supervisée pour de la classification d'objets ou d'imagettes, avec application à la détection d'objets automatique par images aériennes. Cette classification pourrait être utilisée ensuite pour localiser des objets dans une image, à partir des imagettes classifiées.

2 Méthode

Extraction de caractéristiques par un auto-encodeur. Un auto-encodeur (AE) est un réseau de neurones entraîné de manière à copier son entrée à sa sortie [2]. Il est composé d'une première partie encodant son entrée et d'une seconde décodant et reconstruisant celle-ci. Le réseau est ensuite entraîné par une fonction d'erreur basée sur la reconstruction de son entrée et l'erreur quadratique moyenne (MSE-loss) est ici considérée. Les descripteurs obtenus en sortie ont ainsi pu apprendre une représentation des données d'apprentissage. Dans notre cas de détection d'objet où le déséquilibre des classes est en faveur du fond de l'image, un objet ne serait pas correctement reconstruit et serait identifié comme une anomalie par le système. Dans la variante convolutionnelle que nous utilisons ici, l'information spatiale est conservée à l'aide de produits de convolutions permettant de conserver la position de l'objet dans l'image. L'architecture du système est présentée en Figure 1.
Une fonction de *maxPooling* en sortie de l'encodeur permet l'extraction des descripteurs des données d'entrée, après entraînement de bout en bout. Ce vecteur de descripteurs est utilisé ensuite pour une classification non supervisée.

Évaluation d'une extraction, comparaison. Nous comparons la capacité d'un AE convolutionnel à classer de ma-

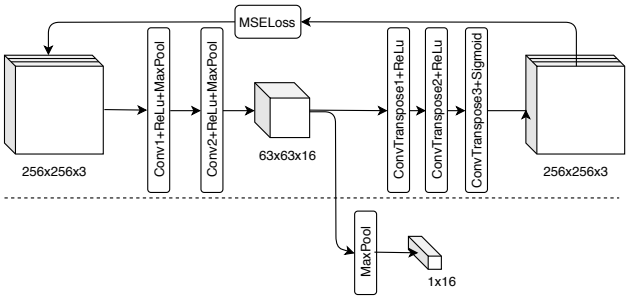

FIGURE 1 – Architecture d'un auto-encodeur. Lors de l'extraction des descripteurs, seul l'encodeur est utilisé.

nière binaire et non supervisée des zones comportant un ou plusieurs objets (zones positives) ou zones vides (négatives). Le classifieur utilisé est un *One-Class SVM* (O-SVM) avec un noyau gaussien. Nous comparons les performances de classification des descripteurs extraits par l'AE avec des descripteurs *hand-craft* comme les HOG [3] et par le réseau convolutionnel profond VGG16 [4], pré-entrainés sur ImageNet, avant la couche *fully connected*.

3 Expériences sur données réelles

Données. La méthode proposée est illustrée à partir de deux jeux de données réduits acquis lors de campagnes de comptage aérien réalisées par le Parc Naturel Régional du Golfe du Morbihan (Bretagne, France). Il s'agit de deux contextes d'activités humaines dans un milieu naturel protégé, où la quantification de cette activité par le dénombrement d'objets d'intérêt (bateaux de plaisance ou pêcheurs à pied respectivement) est importante pour les études environnementales. Deux ensembles de sous-images de 256x256 pixels ont été extraits à partir d'images entières sans recouvrement possible : respectivement 426 négatives et 246 positives (A) ; 1165 négatives et 171 positives (B). Des exemples sont présentés en Figure 2.
Pour ces deux jeux de données, l'AE a été entraîné sur 100 epochs avec une taille de batch de 32.

(A)	γ	Précision	Rappel	F-score
HOG	1	0,41	0,49	0,45
AE	7000	**0,97**	**0,95**	**0,96**
VGG16	0.1	0,43	0,58	0,49
(B)	γ	Précision	Rappel	F-score
HOG	10	0,48	0,80	0,60
AE	10000	**0,49**	**0,83**	**0,62**
VGG16	1	0,48	0,80	0,60

TABLE 1 – Résultats de classification obtenus pour les différents descripteurs sur les deux jeux d'images aériennes.

Résultats & discussion. Les résultats de classification, présentés dans le Tableau 1, sont obtenus par *grid search* sur l'hyperparamètre γ maximisant le F-score.
Nous notons cependant un écart important de la performance de la méthode AE selon les deux jeux de données. Cet écart peut être associé à la différence de résolution des objets. Dans le premier cas, les objets de petite taille permettent à l'AE de capturer le fond et donc de procéder à la détection des objets par détection d'anomalie. Au contraire avec le second cas, on peut attendre de l'auto-encodeur qu'il reconstruise les objets qui sont de la taille de la sous-image, et donc être plus proche des autres descripteurs en termes de résultats. Néanmoins, la dimension de ces derniers est bien moindre en comparaison.
Nous pouvons également remarquer une grande différence de valeur de γ maximisant le F-score entre les descripteurs. Nous pouvons l'interpréter comme un indice sur la capacité de généralisation des descripteurs générés. Dans le cas (A), une valeur aussi grande de γ suppose un sur-apprentissage de l'O-SVM aux descripteurs, contrairement aux descripteurs supposés plus généraux fournis par HOG ou VGG.

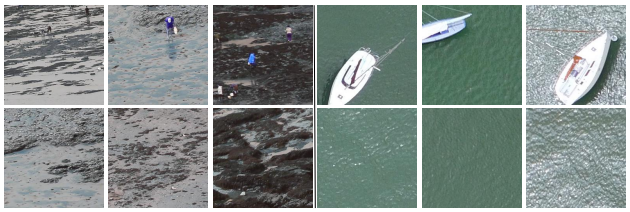

FIGURE 2 – Échantillons de sous-images, négatives (bas) et positives (haut) - pêcheurs à pieds (A) à gauche, bateaux de plaisance (B) à droite.

4 Conclusion et perspectives

Cette étude préliminaire montre que l'auto-encodeur convolutionnel peut être utilisé pour de l'extraction de descripteurs permettent une classification par détection d'anomalie. La conservation des propriétés spatiales des objets dans sa représentation permettra également d'obtenir une localisation plus précise des objets et donc une détection d'objet. Cette approche entièrement non supervisée est compatible avec des modèles d'apprentissage cherchant à optimiser le nombre d'étiquettes disponibles tels que les apprentissages faiblement supervisé ou actif.

Références

[1] G. Dong, G. Liao, H. Liu, and G. Kuang. A review of the autoencoder and its variants : A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE GRSM*, 6(3) :44–68, 2018.

[2] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

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