Ré-entraînement de réseaux de neurones profonds pour la segmentation sémantique par ajout de modalité sans annotation

Jean-Christophe Burnel * Luc Courtrai Sébastien Lefèvre

Université Bretagne Sud, UMR 6074 IRISA
Campus de Tohannic, 56000 Vannes
jean-christophe.burnel@irisa.fr

Résumé

L'entraînement de réseaux de neurones profonds pour la vision par ordinateur nécessite des jeux de données de grande taille, dont l'annotation est particulièrement coûteuse et dépend de la tâche visée. Nous nous intéressons dans cet article à l'ajout d'une nouvelle modalité dans les données, et proposons une méthode tirant parti de cette nouvelle modalité pour offrir des performances accrues sans pour autant nécessiter une nouvelle phase d'annotation. Plus précisément, nous suggérons de réutiliser les données précédemment annotées sur une seule modalité à l'aide d'un pseudo-étiquetage via un auto-entraînement, en générant d'une part des pseudo-labels sur nos données avec les deux modalités, et d'autre part une estimation de la seconde modalité pour les données étiquetées. Nous illustrons l'intérêt de cette méthode pour la segmentation sémantique sur des données RVB-Profondeur, et montrons qu'il peut être plus intéressant d'ajouter une modalité sans annotation plutôt que de n'utiliser qu'une sous-partie annotée d'un jeu de données.

Mots Clef

imagerie multimodale, auto-entraînement, semi-supervisé, segmentation sémantique

Abstract

Training deep neural networks for computer vision requires large datasets. Annotating thse datasets can be expensive depending on the targeted task. In this paper, we focus on the addition of a new modality and propose a method that takes advantage of this new modality to increase our model performances without requiring a new annotation phase. Specifically, we suggest reusing data previously annotated on a single modality by using pseudo-labelling through self-training : on the one hand, by generating pseudo-labels on our data with the two modalities, and on the other hand, by estimating the second modality for the labelled data. We illustrate the interest of this method in the context of semantic segmentation on RGB-Depth data, and show that it may be more interesting to add a modality without annotation rather than to use only an annotated subpart of a dataset.

Keywords

multimodal imagery, self-training, semi-supervised, semantic segmentation

1 Introduction

Les réseaux de neurones profonds et leurs bonnes performances ont permis de faire émerger de plus en plus de projets novateurs reposant sur la vision par ordinateur, et ce dans de nombreux domaines d'application. Contrairement aux jeux généralement utilisés pour tester les algorithmes, comme Imagenet [1], beaucoup de ces projets impliquent des données dont les annotations ne sont pas disponibles en grande quantité, par exemple en télédétection avec l'imagerie multispectrale ou radar. De plus, un projet évoluant dans le temps, les méthodes d'acquisition ou les capteurs peuvent aussi évoluer au fil du projet. Cela peut poser plusieurs problèmes notamment pour la mise en valeur du fastidieux travail d'annotation de ces données. Dans ce contexte, nous nous intéressons aux approches de pseudo-étiquetage et plus précisément à l'auto-entraînement. Nous considérons le scénario où, à l'aide d'un réseau entraîné sur des données RVB, nous cherchons à entraîner un réseau RVB-Profondeur sans données annotées. Nous proposons pour cela de générer des pseudo-labels pour les données non étiquetées, ainsi que d'estimer la profondeur des données annotées.

2 Travaux antérieurs

L'auto-entraînement [6] est une méthode d'apprentissage semi-supervisé où l'on utilise un premier modèle entraîné, que l'on appelle modèle enseignant, pour prédire des labels dont une sous-partie sera ajoutée au jeu d'entraînement. Cette sélection se base sur plusieurs critères, par exemple en utilisant le voisinage dans le cas d'une segmentation grossière ou clairsemée [2] ou en ne sélectionnant que les informations supérieures à un certain seuil [5, 7]. Il existe alors deux façons d'utiliser ces pseudo-labels, soit en ne retenant que la classe avec la probabilité la plus élevée (pseudo-étiquetage dur), soit en retenant l'ensemble

*Ce travail est financé par la région Bretagne et le GIS BreTel via le projet doctoral ALTER, et par le FEAMP via le projet Game of Trawls.

Méthode \ Classes	1	2	3	4	5	6	7	8	9	10	11	12	13	mIOU
Mono supervisé	0.71	*0.38*	0.72	0.54	0.91	0.72	0.57	**0.69**	0.72	0.46	0.57	0.86	*0.74*	51.42
Mono auto-entraîné *	0.69	0.35	**0.75**	*0.68*	0.92	0.72	*0.64*	0.66	0.71	0.44	0.61	*0.89*	0.73	53.90
Mono auto-entraîné 3 rounds *	0.74	0.32	0.68	*0.68*	*0.93*	0.72	**0.66**	0.65	0.71	*0.48*	0.58	*0.89*	**0.75**	*54.95*
Multi supervisé	0.74	0.33	*0.73*	0.67	*0.95*	0.72	0.62	0.66	0.69	0.44	*0.63*	*0.89*	0.73	54.03
Multi auto-entraîné *	*0.75*	0.24	0.62	0.67	*0.95*	**0.75**	0.61	0.67	*0.75*	0.46	**0.64**	0.88	0.73	54.53
Multi auto-entraîné 3 rounds *	***0.77***	**0.43**	0.72	**0.69**	**0.96**	*0.73*	0.62	*0.68*	**0.76**	**0.52**	0.59	**0.90**	**0.75**	**56.41**

TABLE 1 – Comparaison des différentes méthodes testées (nos contributions sont marquées avec un astérisque). Chaque colonne correspond à une classe. On compare l'exactitude par classe et l'intersection sur l'union globale. En **gras** on montre le meilleur résultat et en *italique* le second meilleur. Pour chaque métrique, plus la valeur est haute, meilleur est le résultat.

des probabilités (pseudo-étiquetage souple). On entraîne ensuite un second réseau, que l'on appelle modèle étudiant, avec les données étiquetées et les données non étiquetées. L'auto-entraînement est une méthode itérative (composée de *rounds*), et à la fin de chaque round, le réseau étudiant devient le réseau enseignant du prochain round. Cependant, même si l'on peut changer de domaine, on reste sur les mêmes modalités pour pouvoir combiner les deux jeux de données.

3 Ajout de modalité sans annotation

On propose ici de réutiliser les données étiquetées dans l'entraînement en prédisant la seconde modalité. Ainsi on ajoute un modèle enseignant qui sera chargé de générer une pseudo-modalité. Les architectures mono-modales et multi-modales sont basées sur Adapnet et SSMA [4] qui ont l'avantage de reposer sur des blocs simples. On utilise le même type de réseau pour la segmentation et l'estimation de profondeur. La sélection de pseudo-labels s'effectue suivant une méthode proche de CBST [7], où l'on retiendra les prédictions des 3 classes les plus prédites (on se place donc dans le cas d'un pseudo étiquetage souple) et l'on mettra toutes les autres à 0. Cela nous permet d'avoir des pseudo-labels souples clairsemés, profitant d'un stockage plus efficient, mais également de réduire le bruit des classes les moins probables. Dans notre fonction de coût, on minimise l'entropie croisée et la divergence K-L entre les deux distributions préalablement lissées.

Nous évaluons notre méthode à l'aide du jeu de données NYU Depth V2 [3]. Celui-ci propose à la fois des données annotées (1449 images) et non annotées (vidéos). Pour créer une base de données non annotées, on vient sélectionner une image sur ~ 160 (on prend l'image où la profondeur est la plus synchronisée possible avec l'image couleur), on retient donc un total de 3085 images non étiquetées. Dans un second temps, on utilise un réseau préentrainé pour détecter les images trop proches de celles du jeu de validation et les retirer. Les résultats obtenus avec notre méthode sont donnés dans la Table 1. On compare ici les résultats de nos réseaux RVB et RVB-Pronfondeur entrainés sur le jeu de données annotées avec un auto-entraînement sans utiliser la nouvelle modalité d'une part, et en RVB-Profondeur d'autre part. Le modèle RVB supervisé est celui qui servira de premier modèle enseignant, tandis que le modèle RVB-Profondeur supervisé est donné à titre de comparaison. On notera que le modèle RVB après 3 rounds d'auto-entraînement est meilleur que le modèle RVB-Profondeur supervisé, montrant l'importance des données pour une tâche comme la segmentation sémantique. On remarque également que l'ajout de l'information de profondeur (réelle et générée) permet d'obtenir de bien meilleures performances

4 Conclusion

On montre ici qu'il est possible d'ajouter une modalité sans annotation, et de profiter de cette modalité pour accroître les performances d'un modèle. Des tests plus poussés et une étude ablative précise permettront de savoir si ces résultats dépendent ou non de l'initialisation, et de quantifier les apports de chaque partie indépendamment.

5 Remerciements

Ce travail est financé par la région Bretagne et le GIS BreTel via le projet doctoral ALTER, et par le FEAMP via le projet GAME OF TRAWLS.

Références

[1] Jia DENG et al. "Imagenet : A large-scale hierarchical image database". In : *CVPR*. 2009, p. 248-255.

[2] Inmaculada DÓPIDO et al. "Semisupervised self-learning for hyperspectral image classification". In : *TGRS* 51.7 (2013), p. 4032-4044.

[3] Nathan SILBERMAN et al. "Indoor Segmentation and Support Inference from RGBD Images". In : *ECCV*. 2012.

[4] Abhinav VALADA, Rohit MOHAN et Wolfram BURGARD. "Self-Supervised Model Adaptation for Multimodal Semantic Segmentation". In : *IJCV* 128.5 (2020), p. 1239-1285.

[5] Qizhe XIE et al. "Self-Training With Noisy Student Improves ImageNet Classification". In : *CVPR*. 2020, p. 10684-10695.

[6] David YAROWSKY. "Unsupervised word sense disambiguation rivaling supervised methods". In : *ACL*. 1995, p. 189-196.

[7] Yang ZOU et al. "Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training". In : *ECCV*. 2018.