Apprentissage contrastif multi-modal : Du pré-entrainement auto-supervisé à la classification supervisée.

Paul Berg, Minh-Tan Pham, Nicolas Courty

Apprentissage contrastif multi-modal : Du pré-entrainement auto-supervisé à la classification supervisée

Paul Berg[1] Minh-Tan Pham[1] Nicolas Courty[1]

[1] IRISA, Université Bretagne Sud, UMR 6074, 56000 Vannes, France

paul.berg@univ-ubs.fr

Résumé

Avec l'omniprésence des méthodes d'apprentissage profond pour résoudre des problèmes de vision par ordinateur, le besoin pour des données annotées augmente constamment. Cependant, dans la plupart des cas, le processus d'annotation peut être long et fastidieux. Afin de réduire le besoin pour ces annotations, des méthodes auto-supervisées ont récemment été proposées dans la littérature. Nous nous focalisons ici sur l'apprentissage de représentations pour images multi-modales et comment ces méthodes peuvent être adaptées pour fonctionner pour l'apprentissage multi-modal. Nous proposons une méthode pour l'apprentissage auto-supervisé sur images multi-modales ainsi qu'une méthode pour le finetuning de modèles inspirée par l'apprentissage contrastif.

Mots Clef

Apprentissage contrastif, Classification multimodale, Apprentissage auto-supervisé.

Abstract

With the ubiquity of deep learning methods to solve computer vision tasks, the need for high quality annotated images increases constantly. However, in many cases, the annotation process can be long and tedious. In order to reduce the need for these annotations, so-called self-supervised methods have been proposed in the litterature. In this article, we focus on the problem of representation learning for multi-modal images and how these methods can be adapted for multi-modal learning. We propose a method for self-supervised learning on multi-modal images as well as a finetuning method inspired by contrastive learning.

Keywords

Contrastive learning, Multi-modal classification, Self-supervised learning.

1 Introduction

Pour réduire les besoins en images annotées des méthodes d'apprentissage profond, des méthodes auto-supervisées [6, 7, 4, 5] ont été introduites dans la littérature de vision par ordinateur. Ces méthodes sont bien souvent focalisées sur l'apprentissage de représentations pour images uni-modales. C'est à dire qu'un échantillon est composé uniquement d'une seule modalité. Les images multi-modales qui par opposition sont composées de plusieurs modalités (par exemple, les images optiques et radar en télédétection) sont régulièrement utilisées en télédétection et une les méthodes auto-supervisées doivent donc être adaptées [1]. Plusieurs travaux [11, 13, 8, 2] ont été proposés pour l'apprentissage auto-supervisées sur images multi-modales. Ces travaux ne se focalisent que sur le cas avec deux modalités. Dans cet article, nous proposons une méthode qui peut ingérer un nombre arbitraire de modalités.

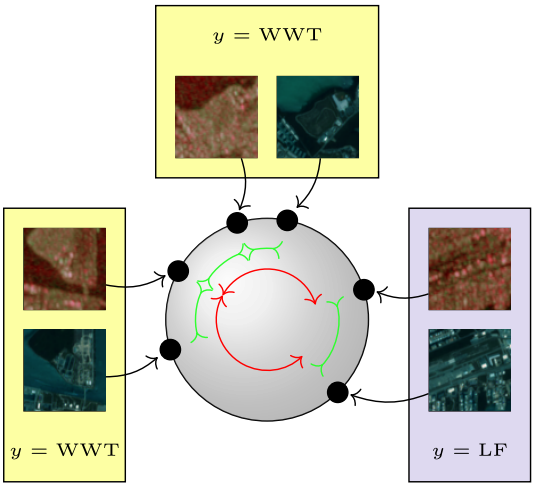

FIGURE 1 – Illustration de l'apprentissage contrastif multi-modal. Les différentes modalités d'une image se rapproche dans l'espace latent et si des labels sont disponibles, les images de même classe se rapproche également dans l'espace latent. Dans ce cas, les représentations des classes *Waste-Water treatment* (WWT) sont rapprochées et celles de la classe *Landfill* (LF) sont éloignées.

En général, une fois un encodeur pré-entraîné avec un apprentissage auto-supervisé, celui-ci doit être finetuné. Cela est généralement réalisé en ajoutant un classifieur qui à partir des représentations apprises émet un score de classe pour chaque échantillon. Ce classifieur est souvent finetuné en utilisant l'entropie croisée (CE) qui ne tient pas compte de la nature multi-modale des images. Nous proposons donc un cadre pour finetuner un modèle basé sur l'apprentissage contrastif supervisé [9] qui prend en compte la nature multi-modale des données et améliore la performance finale par rapport au finetuning par CE.

Pré-entraînement	Finetuning						
	S1	S2	NAIP	S1 + S2	S1 + NAIP	S2 + NAIP	S1 + S2 + NAIP
Aucun	47.37%	64.29%	62.03%	65.04%	63.16%	68.42%	65.79%
S1 + S2	53.76%	71.80%	-	71.80%	-	-	-
S1 + NAIP	56.39%	-	70.68%	-	**72.18%**	-	-
S2 + NAIP	-	71.43%	68.42%	-	-	72.18%	-
S1 + S2 + NAIP	**58.65%**	**72.56%**	**72.93%**	**72.18%**	68.80%	**73.31%**	**73.68%**

TABLE 1 – Comparaison des performances de classification sur le jeu de données Meter-ML en fonction du type de modalités utilisées lors du pré-entraînement auto-supervisé et lors du finetuning.

2 Méthodologie

L'apprentissage contrastif [6] est méthode d'apprentissage auto-supervisé qui fonctionne via la création de plusieurs vues pour chaque échantillons via l'utilisation d'augmentations aléatoires. On considère une image $x_i \in \mathbb{R}^{w\times h\times c}$ augmentée au moyen d'augmentations aléatoires $T^A, T^B \sim \mathcal{T}$ et pour laquelle deux vues vont être créées $x_i^A = T^A(x), x_i^b = T^B(x)$. Les images sont encodées et leur représentations sont ensuite projetées sur l'hypersphère [12] par un encodeur $f_\theta(\cdot) : \mathbb{R}^{w\times h\times c} \to \mathbb{S}^{d-1}$, avec $\mathcal{S}^{d-1} = \{x; x \in \mathcal{R}^d, \|x\| = 1\}$. Enfin pour entraîner un modèle, l'objectif contrastif rapproche les représentations des deux vues produites à partir de la même image x et repousse les représentations des autres vues augmentées du jeu de données. L'objectif Information Noise Contrastive Estimation (InfoNCE) utilisé pour une vue x_i^A est défini par :

$$\mathcal{L}_{\text{InfoNCE}}(x_i^A) = -\log\left(\frac{\exp(\langle f_\theta(x_i^A), f_\theta(x_i^B)\rangle/\tau)}{\sum_{j\neq i}\exp(\langle f_\theta(x_i^A), f_\theta(x_j^B)\rangle/\tau)}\right), \tag{1}$$

où τ est un paramètre de température de la log-probabilité résultante.

Dans cet article, nous nous concentrons sur les jeux de données multi-modaux. Cela signifie que pour chaque échantillon il existe plusieurs vues dans différentes modalités. On considère un jeu de données composé de M modalités. Une adaptation de l'objectif contrastif de l'Equation 1 pour le multi-modal est de considérer chaque vue des différentes modalité comme une augmentation de l'échantillon. Avec cette interprétation, l'augmentation correspond à la lecture par un capteur différent, des augmentations peuvent tout de même être ajoutées pour chaque modalités.

Au lieu d'un seul encodeur pour une seule modalité, l'architecture comporte un modèle $f_\theta^m(\cdot)$ par modalité m. Et l'objectif devient de rapprocher les représentations de chacun de ses encodeurs pour les différentes vues d'un même échantillon. En nommant $z_i^m = f_\theta^m(x_i^m)$ la représentation de x_i selon l'encodeur de la modalité m, cet objectif contrastif multi-modal peut être écrit comme ceci :

$$\mathcal{L}_{\text{MM-InfoNCE}}(x_i) = -\sum_{m,n}\log\left(\frac{\exp(\langle z_i^m, z_i^n\rangle/\tau)}{\sum_{j\neq i}\sum_{k=1}^{M}\exp(\langle z_i^m, z_j^k\rangle/\tau)}\right). \tag{2}$$

Lors de l'utilisation des modèles, il est possible de sélectionner uniquement un sous-ensemble des modalités en fusionnant les représentations de ces modalités et en les classifiant.

Ensuite, lors du finetuning, on peut aussi utiliser la nature multi-modale du jeu de données en adaptant la méthode d'apprentissage contrastif supervisée [9]. C'est à dire qu'en plus d'utiliser les différentes modalités comme vues positives dont le modèle essaie d'avoir une représentation similaire, le modèle essaie également d'augmenter la similarité entre les représentations des échantillons qui ont la même classe. Une illustration descriptive de cette architecture est visible dans la Figure 1.

Class	CE [10]	CE (Notre impl.)	**SupCon+CE (Proposé)**
Forest	65±8	63.49±5.90	**75.54±5.01**
Shrubl.	56±11	**58.31±4.59**	52.26±4.42
Grassl.	9±6	**12.60±7.53**	6.18±4.93
Wetland	15±8	14.43±11.55	**20.90±7.11**
Croplan	45±8	52.65±7.94	**59.65±5.07**
Urban	**95±1**	91.14±3.48	89.94±2.63
Barren	39±3	43.91±2.7	**44.09±3.49**
Water	**99±1**	98.63±0.41	98.42±1.25
Global	67±2	68.29±1.66	**72.62±1.31**
Moyenne	53±2	54.39±0.72	**55.87±0.89**

TABLE 2 – Performance sur le jeu de données DFC2020.

3 Expérimentations

Pour évaluer l'impact de cette apprentissage contrastif multi-modal, nous réalisons un pré-entrainement sur le jeu de données multi-modales MeterML [15] avec ensuite différentes modalités utilisées pour le finetuning de la tâche de classification, y compris les images Sentinel-1 (radar), Sentinel-2 (optique) et NAIP (optique). Ces résultats sont visibles sur le Table 1. Nous observons qu'un modèle pré-entraîné sur plusieurs modalités donne les meilleures performances dans le finetuning avec chacune des modalités (S1, S2 ou NAIP) et également avec les combinaisons de deux ou de trois modalités différentes.

Ensuite, pour évaluer notre proposition de l'apprentissage contrastif multimodal supervisé dans le finetuning, nous conduisons les expérimentations sur le jeu de données DFC2020 (IEEE Datafusion Contest 2020) [14] avec des images multi-modales (Sentinel-1 et Sentinel-2). Pour cela, nous comparons la performance du finetuning avec l'entropie croisée (CE) proposé dans [10] par rapport à notre méthode en combinant la CE avec une perte multi-modale contrastive supervisée. Ces résultats sont visible sur le Table 2. Nous observons que notre proposition a amélioré significativement la performance de classification multimodale sur les données DFC2020. Plus des détails sur les expériences, ainsi que les analyses et discussions se trouvent dans notre papier récent [3].

Références

[1] P. Berg, M.-T. Pham, and N. Courty. Self-supervised learning for scene classification in remote sensing : Current state of the art and perspectives. *Remote Sensing*, 14(16) : 3995, 2022.

[2] P. Berg, M.-T. Pham, and N. Courty. Joint multi-modal self-supervised pre-training in remote sensing : Application to methane source classification. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6624–6627. IEEE, 2023.

[3] P. Berg, B. Uzun, M.-T. Pham, and N. Courty. Multimodal supervised contrastive learning in remote sensing downstream tasks. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[4] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33 :9912–9924, 2020.

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] J.-B. e. a. Grill. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33 :21271–21284, 2020.

[8] P. Jain, B. Schoen-Phelan, and R. Ross. Self-supervised learning for invariant representations from multi-spectral and SAR images, 2022. URL `https://arxiv.org/abs/2205.02049`.

[9] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33 :18661–18673, 2020.

[10] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022.

[11] L. Scheibenreif, M. Mommert, and D. Borth. Contrastive self-supervised data fusion for satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3 :705–711, 2022.

[12] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[13] Y. Wang, C. M. Albrecht, and X. X. Zhu. Self-supervised vision transformers for joint SAR-optical representation learning. *arXiv preprint arXiv :2204.05381*, 2022.

[14] N. Yokoya, P. Ghamisi, R. Hänsch, and M. Schmitt. 2020 ieee grss data fusion contest : Global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(1) :154–157, 2020.

[15] B. Zhu, N. Lui, J. Irvin, J. Le, S. Tadwalkar, C. Wang, Z. Ouyang, F. Y. Liu, A. Y. Ng, and R. B. Jackson. Meter-ml : A multi-sensor earth observation benchmark for automated methane source mapping. *arXiv preprint arXiv :2207.11166*, 2022.