Towards Considering Explicit Sensitivity to Augmentation in Visual Instance Discrimination Tasks

Alexandre Devillers, Mathieu Lefort

Towards Considering Explicit Sensitivity to Augmentation in Visual Instance Discrimination Tasks

A. Devillers[1], M. Lefort[1]

[1] University Lyon 1 Claude Bernard, LIRIS

alexandre.devillers@liris.cnrs.fr
mathieu.lefort@liris.cnrs.fr

Résumé

Les méthodes récentes d'apprentissage autosupervisé de représentations visuelles sont basées sur des tâches de discrimination d'instance visant à apprendre des représentations non triviales insensibles à un ensemble d'augmentations soigneusement choisi. Les performances de ces méthodes se rapprochent rapidement des approches supervisées, et les surpassent même dans certains cas, et ce, sans supervision experte. Cet article donnera un aperçu général de ces méthodes et discutera d'une direction de recherche visant à inclure de la sensibilité à certaines augmentations.

Mots-clés

Apprentissage de représentation visuelle, apprentissage profond autosupervisé, discrimination d'instance.

Abstract

Recent methods of self-supervised visual representation learning are based on instance discrimination tasks aiming at learning non-trivial representations insensitive to a carefully chosen set of augmentations. These methods are closing the gap with the supervised approaches, even outperforming them in some cases, while having the advantage of not requiring expert supervision. This paper will overview some of these methods, and discuss a research direction aiming at including sensitivity to some augmentations.

Keywords

Visual representation learning, deep self-supervised learning, instance discrimination.

1 Introduction

Learning pertinent visual representations is a crucial and challenging problem to achieve good performance on downstream tasks while allowing for better data efficiency. Learning such representations in a self-supervised manner, *i.e.* without human supervision, allows the use of plentiful raw data, opening the application of deep learning to domains suffering from a lack of annotations. Yet, self-supervised learning requires finding a supervisory signal obtainable from the data.

Recent successful approaches in visual representation learning are based on instance discrimination tasks, and more precisely on building augmentations-invariant embeddings [1, 2, 3, 5, 6]. These methods are closing the gap with the supervised approaches, even being competitive in some cases as few-shot learning. Still, learning such an invariant property encourages embeddings to be insensitive to augmentations, *i.e.* to leave the information modifiable by the transformations. This way, the set of possible transformations has to be carefully selected. For instance, if one requires color information to be in the representations, then the set of transformations should avoid color manipulation. Sec. 2 will briefly overview some of these recent methods relying on an invariance task, while in Sec. 3 we will identify why using only insensitivity may be sub-optimal, and show, based on recent work and our preliminary results, how explicit sensitivity can be beneficial to representations. Finally, Sec. 4 will stand for the perspectives and future works we aim for.

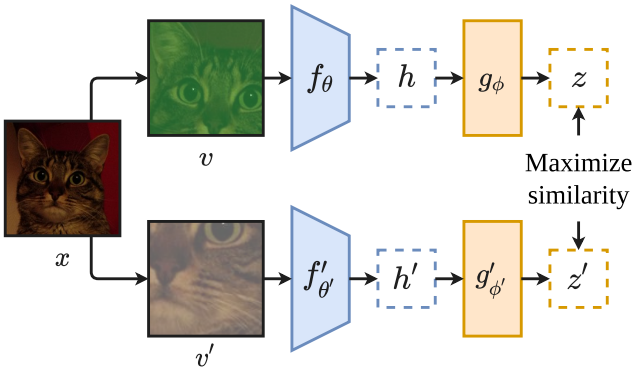

FIGURE 1 – Common base shared by recent methods. The input image x is augmented in two ways to give the views v and v'. These views are then passed through an encoder to give respectively the representations h and h', which are then passed in a projection head that outputs respectively the embeddings z and z' on which the loss is applied.

2 Existing Methods

Recent methods are mostly siamese networks with an instance discrimination task [1, 2, 3, 5, 6]. They aim to build a latent space where two augmentations of the same image

have similar embeddings, see Fig. 1, while using various tricks to avoid simple solution collapse. Thus, the resulting embeddings are insensitive to the possible augmentations, requiring the set of used augmentations to be carefully chosen, as specified in the previous section. Current state-of-the-art methods are almost all using the same set of transformations, which have been constructed experimentally by testing multiple combinations [2]. These transformations are strong enough to make the views very different at the pixel level (i.e. crop, color jitter, etc.), while preserving the original semantic information of the source image, thus forcing the representations to encode this shared semantic to be similar.

On top of this, the embeddings on which the invariance loss is applied are not directly the representations, but a non-linear projection of the representations, as this has shown to improve performance. One hypothesis is that it could allow some augmentation-related information to be present in the representations, as this projection could filter it before the loss. This projection can be seen as an invariance-head, taking the form of a multi-layer perceptron, which is only used by the invariance task during representation learning.

3 Research Hypothesis

The importance of the selected augmentations, and the final projection applied to the representations, shows us that we can separate augmentations into two groups : the ones for which the representations benefit from insensitivity (crop, color jitter, etc.), and the ones for which sensitivity is beneficial (rotation, vertical flip, etc.), more details can be found in [4]. Moreover, this separation experimentally seems the same for all recent methods performing augmentation-invariance tasks. While a large number of recent methods have simply ignored the augmentations requiring sensitivity [1, 2, 3, 5, 6], only one, to the best of our knowledge, has tried to add explicit sensitivity to these augmentations while keeping an invariance task [4].

This last method consists of a framework that proposes to add an extra head with a second task that predicts the transformation, i.e. rotation, that has led to a given view based on its representation. It can be seen as image invariance, as two images augmented in the same way are classified similarly, whereas recent methods perform augmentation invariance. Consequently, it forces the model to encode augmentation-related information into the representations, making them sensitive to these augmentations. Therefore, the transformations used for this task have to be the ones for which sensitivity has shown to be beneficial, such as rotation or vertical flip. This addition has demonstrated to improve existing methods of the state of the art such as [2, 3, 6].

On our side, we have also explored the same idea of using an extra head with a second task to add an explicit sensitivity to some augmentations. We have started experimenting with a task of non-trivial equivariance, aiming at building a latent space in which displacements caused by augmentations in the image space are significant. Therefore, we differ from recent methods that rely on invariance, where such displacements are null. We also differ from [4] as we do not have image invariance and as we structure the latent space so that the displacements are predictable from the augmentations parameters. Altogether, our addition guarantees that some augmentation-related information is present and structured in the representations. Preliminary results using SimCLR [2] as a baseline, on which we have added our extra head and task, have shown to lower the error from around 9% to 7.5% on CIFAR10 using the usual linear classification evaluation.

4 Discussion and Perspectives

We believe explicit augmentation sensitivity has been under-explored in recent visual representation learning methods. We also believe that some discarded augmentations could benefit a non-invariant task. Following recent work and our preliminary results, we suggest that such explicit augmentation sensitivity may be a good research direction and could lead to more state-of-the-art improvements in visual representation learning. Our future work will aim at scaling our experiments to more baselines such as BYOL [5] and more datasets such as ImageNet. We also plan to perform few-shot evaluations while experimenting with the set of augmentations for which explicit sensitivity is significantly beneficial.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011013160 made by GENCI.

Références

[1] A. Bardes, J. Ponce, and Y. LeCun. Vicreg : Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv :2105.04906*, 2021.

[2] 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.

[3] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[4] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, and M. Soljačić. Equivariant Contrastive Learning. *arXiv preprint arXiv :2111.00899*, 2021.

[5] J-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, and M. G. Azar. Bootstrap your own latent : A new approach to self-supervised learning. *arXiv preprint arXiv :2006.07733*, 2020.

[6] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins : Self-supervised learning via redundancy reduction. *arXiv preprint arXiv :2103.03230*, 2021.