Augmentation-aware Self-supervised Learning with Conditioned Projector

Marcin Przewięźlikowski[1,2,3] * **Mateusz Pyla**[1,2,3] **Bartosz Zieliński**[1,3]
Bartłomiej Twardowski[3,4,5] **Jacek Tabor**[1] **Marek Śmieja**[1]
[1] Jagiellonian University, Faculty of Mathematics and Computer Science
[2] Jagiellonian University, Doctoral School of Exact and Natural Sciences
[3] IDEAS NCBR [4] Department of Computer Science, Universitat Autònoma de Barcelona
[5] Computer Vision Center, Barcelona

Abstract

Self-supervised methods such as SimCLR and MoCo are able to reach quality on par with supervised approaches, by learning to remain invariant to applied data augmentations. However, this invariance may be harmful to solving downstream tasks that depend on traits affected by augmentations used during pretraining, such as color. We propose to foster sensitivity to such characteristics in the representation space by supplementing the projector network, a common component of self-supervised architectures, with information about augmentations applied to images. In order for the projector to take advantage of this auxiliary conditioning when solving the SSL task, the feature extractor learns to preserve the augmentation information in its representations. Our approach, coined **C**onditional **A**ugmentation-aware **S**elf-**s**upervised **Le**arning (CASSLE), is directly applicable to typical joint-embedding SSL methods regardless of their objective functions. We conduct a series of experiments, which show that CASSLE improves over various SSL methods, reaching state-of-the-art performance in multiple downstream tasks.[2]

1 Introduction

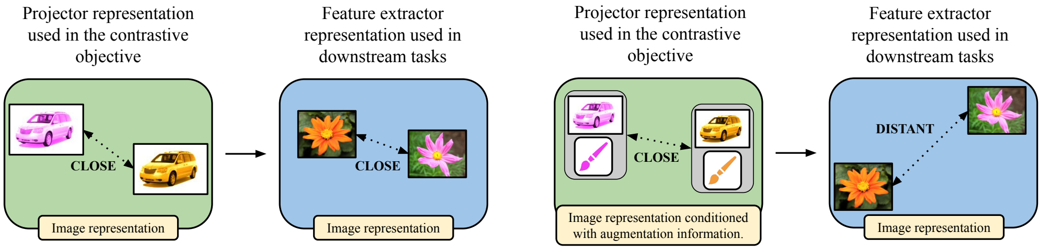
Contrastive methods of learning representations that remain invariant when subjected to various data augmentations [6, 3, 11] have achieved impressive results that have greatly diminished the disparity with representations learned in a supervised way [1]. Nevertheless, contrastive methods may perform poorly when transferred to a downstream task that relies on features affected by augmentation [12]. For example, color jittering can result in a representation space invariant to color shifts, which would be detrimental to the task of flower classification (see Figure 1).

In this work, we propose a new method called **C**onditional **A**ugmentation-aware **S**elf-**s**upervised **Le**arning (CASSLE) that mitigates augmentation invariance of representation without neither major changes in network architecture or modifications to the self-supervised training objective. We propose to use the augmentation information during the SSL training as additional conditioning for the projector network. This encourages the feature extractor network to retain information about augmented image features in its representation. CASSLE can be applied to any joint-embedding SSL method regardless of its objective, provided that it utilizes a projector network [4, 3, 11, 13, 5]. The outcome is a general-purpose, augmentation-aware encoder that can be directly used for any downstream task. CASSLE presents improved results in comparison to other augmentation-aware SSL methods, improving transferability to downstream tasks where invariance of the model representation for specific data changes could be harmful.

[*] Corresponding author: `marcin.przewiezlikowski@doctoral.uj.edu.pl`

[2] We share our codebase at `https://github.com/gmum/CASSLE`.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).


Projector representation used in the contrastive objective
CLOSE
Image representation
Feature extractor representation used in downstream tasks
CLOSE
Image representation
Projector representation used in the contrastive objective
CLOSE
Image representation conditioned with augmentation information.
Feature extractor representation used in downstream tasks
DISTANT
Image representation

Joint-embedding Self-supervised learning (traditional) **Conditional Augmentation-aware Self-supervised Learning** (ours)

Figure 1: Contrastive loss minimization draws the representations of augmented image views closer in the latent space of the projector (green). This may also reduce the distance between their feature extractor representations (blue). Thus, the representation becomes invariant to augmentation-induced perturbations, which may hinder the performance on downstream tasks. In contrast, the self-supervised objective of CASSLE draws together joint representations of images and their augmentations in the projector space. By conditioning the projector with augmentation information, image representations retain more sensitivity to perturbations in the feature extractor space. This proves to be beneficial when solving downstream tasks.

2 Method

2.1 Preliminaries

A typical contrastive framework used in self-supervised learning consists of an augmentation function t_ω and two networks: feature extractor f and projector π. Let $\mathbf{v}_1 = t_{\omega_1}(\mathbf{x}), \mathbf{v}_2 = t_{\omega_2}(\mathbf{x})$ be two augmentations of a sample $\mathbf{x} \sim X$ parameterized by $\omega_1, \omega_2 \sim \Omega$. The feature extractor maps them into the embedding space, which is the representation used in downstream tasks. To make the representation invariant to data augmentations, $\mathbf{e}_1 = f(\mathbf{v}_1)$ is forced to be similar to $\mathbf{e}_2 = f(\mathbf{v}_2)$. However, instead of imposing similarity constraints directly on the embedding space of f, a projector π transforms the embeddings into target space where the contrastive loss $\mathcal{L}$ is applied.

Minimizing $\mathcal{L}(\pi(\mathbf{e}_1), \pi(\mathbf{e}_2))$ directly leads to reducing the distance between embeddings $\pi(\mathbf{e}_1)$ and $\pi(\mathbf{e}_2)$. However, $\mathcal{L}$ still indirectly encourages the intermediate network representations (including the output of the feature extractor f) to also conform to the contrastive objective to some extent. As a result, the feature extractor tends to erase the information about augmentation from its output representation. This behavior may however be detrimental for certain downstream tasks (see Figures 1 and 4), which rely on features affected by augmentations. For instance, learning invariance to color jittering through standard contrastive methods may lead to degraded performance on the downstream task of flower recognition, which is not a color-invariant task [9, 12].

2.2 CASSLE

To overcome the above limitations of SSL, we facilitate the feature extractor to encode the information about augmentations in its output representation. In consequence, the obtained representation will be more informative for downstream tasks which depend on features modified by augmentations.

CASSLE achieves this goal by conditioning the projector π on the parameters of augmentations used to perturb the input image. Specifically, we modify π so that apart from embedding $\mathbf{e}$, it also receives augmentation information ω and projects their joint representation into the space where the objective $\mathcal{L}$ is imposed. We do not alter the $\mathcal{L}$ itself; instead, training relies on minimizing the contrastive loss $\mathcal{L}$ between $\pi(\mathbf{e}_1|\omega_1)$ and $\pi(\mathbf{e}_2|\omega_2)$. Thus, π learns to draw $\mathbf{e}_1$ and $\mathbf{e}_2$ together in its representation space *on condition of* ω_1 *and* ω_2. We construct augmentation information ω by concatenating vectors ω^{aug} describing the parameters of each augmentation type [8]. We condition the projector π with ω by processing ω with a 6-layer Multilayer Perceptron and concatenating the output to the image embeddings $\mathbf{e}$ before feeding them to π. We visualize the architecture of CASSLE in Figure 2.

We provide a rationale for why CASSLE preserves information about augmented features in the representation space. Let us examine the impact of including augmentation information vectors ω on the process of solving the contrastive pretext task by CASSLE. Since ω does not carry any information about source images $\mathbf{x}$, its potential usefulness during pretraining could only be explained by providing knowledge of transformations t_ω that had been applied to $\mathbf{x}$ to form views $\mathbf{v}$. Furthermore, for π to act upon the knowledge of ω, feature extractor representation $\mathbf{e} = f(\mathbf{v})$ must preserve information about features of $\mathbf{x}$ modified by t_ω.

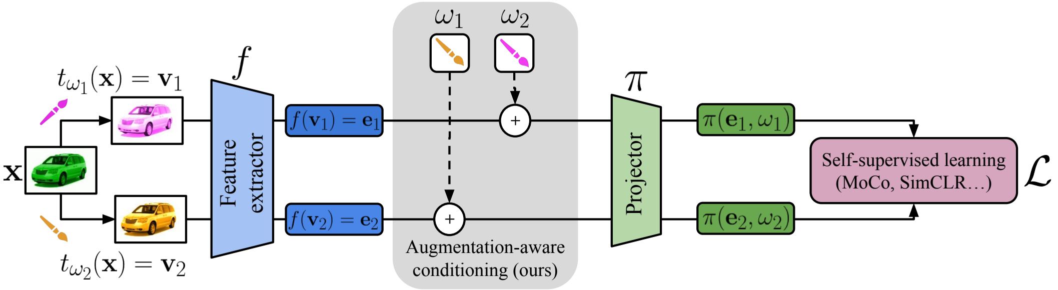

ω_1
ω_2
$t_{\omega_1}(\mathbf{x}) = \mathbf{v}_1$
f
π
$\mathbf{x}$
Feature extractor
$f(\mathbf{v}_1) = \mathbf{e}_1$
$f(\mathbf{v}_2) = \mathbf{e}_2$
Projector
$\pi(\mathbf{e}_1, \omega_1)$
$\pi(\mathbf{e}_2, \omega_2)$
Self-supervised learning (MoCo, SimCLR…)
$\mathcal{L}$
Augmentation-aware conditioning (ours)
$t_{\omega_2}(\mathbf{x}) = \mathbf{v}_2$

Figure 2: Overview of CASSLE. We extend the typical self-supervised learning approaches by incorporating the information of augmentations applied to images into the projector network. In CASSLE, the SSL objective is thus imposed on joint representations of images and the augmentations that had been applied to them. This way, CASSLE enables the feature extractor to be more aware of augmentations than the methods that do not condition the projector network.

To demonstrate that ω is indeed meaningful for solving the contrastive objective by our approach, we measure similarities of augmented image views in the representation space of the CASSLE projector in two scenarios:

1. representations of both images are constructed using true information about augmentations that had been applied to them.
2. representation of one of the images is constructed by supplementing π with falsified augmentation information.

It is evident from Figure 3 that the similarity of representations decreases when false augmentation parameters are supplied to the projector.

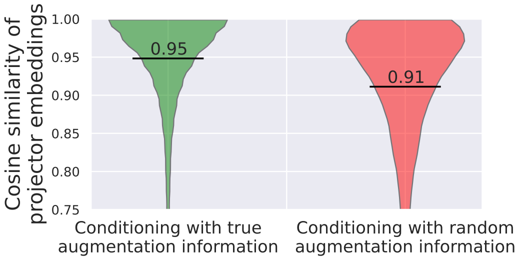


Figure 3: Similarities of CASSLE projector representations when conditioned with augmentation information from either their respective images (green) or randomly sampled (red). Solid lines denote the mean values of similarities. The CASSLE projector relies on correct augmentation information for drawing positive image pairs together.

In CASSLE, *knowledge of information about the applied augmentations (ω) is useful for matching pairs feature extractor representations of augmented image views* ($\mathbf{e}$). This in turn implies that $\mathbf{e}$ indeed preserves information pertaining to data features altered by augmentation.

CASSLE can be applied to a variety of joint-embedding SSL methods, as the only practical modification it makes is changing the projector network to utilize the additional input ω, describing the augmentations. We do not modify any other aspects of the self-supervised approaches, such as objective functions, which is appealing from a practical perspective. Last but not least, the architecture of the feature extractor in CASSLE is not affected by the introduced augmentation conditioning, as we only modify the input to the projector, which is discarded after the pretraining. Just like in vanilla SSL techniques, the feature extractor can be directly used in downstream tasks.

3 Experiments

3.1 Linear evaluation on downstream tasks

We begin the experimental analysis by addressing the most fundamental question – how does CASSLE impact the ability of models to generalize? In order to answer it, we evaluate the performance of pretrained networks on the downstream tasks of classification on a wide array of datasets. We follow the linear evaluation protocol [7, 3, 8], and evaluate multiple self-supervised methods extended with CASSLE, as well as other recently proposed extensions which increase sensitivity to augmentations [8, 2]. We find that CASSLE generally achieves the best downstream results in the vast majority of cases.

3.2 Analysis of representations formed by CASSLE

We also investigate the awareness of augmentation-induced data perturbations in the intermediate and final representations of pretrained networks. As a proxy metric for measuring this, we report the mean InfoNCE loss [10, 3] values under different augmentation types at subsequent stages of ResNet-50 and projectors of MoCo-v2, AugSelf [8] and CASSLE in Figure 4. Representations of CASSLE

Table 1: Linear evaluation on downstream classification and regression tasks. CASSLE consistently improves representations formed by vanilla SSL approaches and performs better or comparably to other techniques of increasing sensitivity to augmentations [12, 8, 2].

Method	C10	C100	Food	MIT	Pets	Flowers	Caltech	Cars	FGVCA	DTD	SUN	CUB	300W
SimCLR [3]													
Vanilla	81.80	61.40	56.59†	61.26†	69.10	81.58†	75.95†	31.20†	38.68†	64.99†	46.37†	28.87†	88.47†
AugSelf [8]†	84.30	63.47	60.76	63.43	**71.86**	**86.59**	79.88	36.56	**42.90**	66.59	48.84	34.46	88.79
AI [2]	83.90	63.10	–	–	69.50	68.30	74.20	–	–	53.70	–	**38.60**	88.00
CASSLE	**85.61**	**64.09**	**61.00**	**63.58**	71.43	85.98	**80.62**	**37.97**	42.26	**67.07**	**49.42**	33.91	**89.05**
MoCo-v2 [6, 4]													
Vanilla	84.60	61.60	59.67	61.64	70.08	82.43	77.25	33.86	41.21	64.47	46.50	32.20	88.77†
AugSelf [8]	85.26	63.90	60.78	63.36	73.46	85.70	78.93	37.35	39.47	66.22	48.52	37.00	89.49†
AI [2]	81.30	64.60	–	–	**74.00**	81.30	78.90	–	–	**68.80**	–	**41.40**	**90.00**
CASSLE	**86.32**	**65.29**	**61.93**	**63.86**	72.86	**86.51**	**79.63**	**38.82**	**42.03**	66.54	**49.25**	36.22	88.93

feature extractor are on average more difficult to match together than those of vanilla MoCo-v2 and AugSelf [8]. This indicates that in CASSLE, the task of augmentation invariance is solved to a larger degree by the projector, and to a smaller degree by the feature extractor, allowing it to be more augmentation-sensitive. As shown in Section 3.1, this sensitivity helps the CASSLE feature extractor achieve similar or better performance than its counterparts when transferred to downstream tasks.

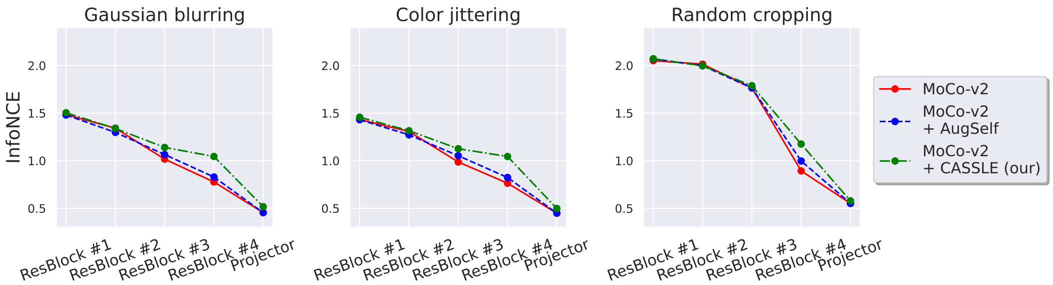


Figure 4: A comparison of InfoNCE loss measured on different kinds of augmentations at subsequent stages of the ResNet-50 and projectors pretrained by vanilla, AugSelf [8] and CASSLE variants of MoCo-v2. Feature extractor representation of CASSLE yields higher InfoNCE values which suggests that it is more susceptible to augmentations.

4 Conclusion

In this paper, we propose a novel method for augmentation-aware self-supervised learning that retains information about data augmentations in the representation space. To accomplish this, we introduce the concept of the conditioned projector, which receives augmentation information while processing the representation vector. Our solution necessitates only small architectural changes and no additional auxiliary loss components. Therefore, the training concentrates on contrastive loss, which enhances overall performance. We show that our solution improves on the downstream performance of vanilla and augmentation-aware SSL techniques. Moreover, it obtains representations more sensitive to augmentations than the baseline methods. Overall, our method offers a straightforward and efficient approach that can be directly applied to a variety of contrastive methods, leading to retaining information about data augmentations in their representation space and improving their quality.

Acknowledgments and Disclosure of Funding

This research has been supported by the flagship project entitled "Artificial Intelligence Computing Center Core Facility" from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University. The research of M. Śmieja was supported by the National Science Centre (Poland), grant no. 2022/45/B/ST6/01117. The research of Bartosz Zieliński is funded in part by National Science Centre (Poland) grant number 2022/47/B/ST6/03397. Bartłomiej Twardowski acknowledges the grant RYC2021-032765-I. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016303.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[2] Ruchika Chavhan, Henry Gouk, Jan Stuehmer, Calum Heggan, Mehrdad Yaghoobi, and Timothy Hospedales. Amortised invariance learning for contrastive self-supervision, 2023.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021.

[6] He K. et al. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[7] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better?, 2019.

[8] Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. *CoRR*, abs/2111.09613, 2021.

[9] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[11] Chen X. and He K. Exploring simple siamese representation learning. In *CVPR*, 2021.

[12] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning, 2020.

[13] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.