Image and Vision Computing 124 (2022) 104494

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Hyperspherically regularized networks for self-supervision

Aiden Durrant *, Georgios Leontidis

Department of Computing Science & Interdisciplinary Centre for Data and AI, University of Aberdeen, AB24 3UE Aberdeen, United Kingdom

ARTICLE INFO

Article history:
Received 27 March 2022
Accepted 25 May 2022
Available online 30 May 2022

Keywords:
Self-supervised learning
Representation learning
Representation separability
Image classification

ABSTRACT

Bootstrap Your Own Latent (BYOL) introduced an approach to self-supervised learning avoiding the contrastive paradigm and subsequently removing the computational burden of negative sampling associated with such methods. However, we empirically find that the image representations produced under the BYOL's self-distillation paradigm are poorly distributed in representation space compared to contrastive methods. This work empirically demonstrates that feature diversity enforced by contrastive losses is beneficial to image representation uniformity when employed in BYOL, and as such, provides greater inter-class representation separability. Additionally, we explore and advocate the use of regularization methods, specifically the layer-wise minimization of hyperspherical energy (i.e. maximization of entropy) of network weights to encourage representation uniformity. We show that directly optimizing a measure of uniformity alongside the standard loss, or regularizing the networks of the BYOL architecture to minimize the hyperspherical energy of neurons can produce more uniformly distributed and therefore better performing representations for downstream tasks.

1. Introduction

Unsupervised visual representational learning methods [1,2] have recently demonstrated performance on downstream tasks that continues to narrow the gap to supervised pre-training, excelling specifically in classification and segmentation tasks [3–5]. This success is largely contributed to contrastive methods which aim to minimize the distance of the representations pertaining to two *views* of the same image in representations space ('positive pair'), whilst maximizing the distance of views from different images ('negative pair') [6]. This ensures that semantically relevant features encoded by representations of positive pairs are similar, whilst negative pairs are dissimilar. The study of contrastive losses has shown that this repulsion effect between dissimilar views is matching the distribution of features in representational space to a distribution of high entropy [7], in other words, encouraging uniformity of representations in space [8]. This balancing of attraction and repulsion is the mechanism that allows contrastive methods to learn similar semantic features whilst avoiding collapse in representation space.

More recently, alternative approaches aim to explore self-supervised learning avoiding the inherent computational difficulties imposed by contrastive methods reliance on negative samples [3]. One method in particular, Bootstrap Your Own Latent (BYOL) [2] falls under the self-distillation paradigm, where we task an "online network' to predict the image representations produced by a 'target" network in a Siamese fashion, where each network is given a different view of the same image (visually depicted in Fig. 3). Yet this network does away with the negative views (views originating from different images), and the subsequent negative term attributed with contrastive losses. The theoretical understanding of how these networks avoid the seemingly inevitable collapsed equilibria, given no explicit mechanism associated with the negative term of contrastive losses, is still to be investigated [9].

Intrigued by the property of mode collapse and inspired by [8], we empirically demonstrate in this work that BYOL fails to distribute its image representations as uniformly in ℓ_2-normalized unit space (i.e. surface of a unit-hypersphere) compared to its contrastive counterparts. As such, we ask, *can BYOL benefit from mechanisms that introduce feature uniformity found in contrastive methods?* To achieve this, we investigate an alternative to the uniformity constraint posed by [8] and derived from the contrastive loss, aiming to maintain the avoidance of negative sampling advocated in BYOL. We instead propose to utilize minimum hyperspherical energy (MHE) weight regularization [10] to enforce neuron (i.e. kernel) uniformity whilst being independent and therefore robust to smaller batch size (the desirable property of BYOL). We empirically demonstrate how the use of MHE regularization can increase uniformity of representation through the concept of neuron redundancy reduction in the ℓ_2-normalized unit space, and in-turn lead to better learned image representations. Our contributions are summarized as follows: i) we empirically show that BYOL distributes its features poorly in representational space compared to contrastive counter parts and

* Corresponding author.
E-mail address: a.durrant.20@abdn.ac.uk (A. Durrant).

https://doi.org/10.1016/j.imavis.2022.104494

that distribution constraints like those in constrastive losses benefit image representations in BYOL; ii) we propose to hyperspherically regularize the network to improve distribution of neurons and subsequently achieve a greater diversity of representations improving image representation separability and performance on downstream tasks; iii) as a consequence hyperspherically regularized BYOL networks maintain the benefits of avoiding contrastive loss negative terms, resulting in reduced performance drops at smaller batch sizes.

2. Related work

2.1. Unsupervised representational learning

The recent popularity of discriminative unsupervised representational learning, specifically contrastive methods, has sparked keen interest in the theoretical understanding of their underpinnings, emerging from their performance rivaling that of supervised methodologies [1,11]. As to why these methods perform so well has only recently begun to be understood, notably [8] prove that optimizing contrastive loss when under a unit ℓ_2-norm constraint (restricting representational space to a unit hypersphere) is equivalent to optimizing a metric of alignment (distance between positive pairs) and uniformity (all feature vectors should be roughly uniformly distributed on the unit hypersphere). Additionally, [7] extends this work proposing a generic form of the contrastive loss, also identifying the same relations of uniformity to pairwise potential in a Gaussian kernel, to match representations to a prior distribution (of high entropy).

Lately, alternatives to contrastive methods [3] have been proposed alleviating some of the computational drawbacks associated with contrastive losses, primarily the necessity of large numbers of negative pairs generally requiring increased batch sizes [1] or memory banks [11]. Bootstrap Your Own Latent (BYOL) avoided the use of negative pairs via an 'online' 'target' network approach akin to Mean Teachers [12], where the 'online' network and an additional 'predictor' network aim to predict the representations of a slowly updated 'target' network of the 'online' network. However, it is not clear how these networks avoid collapsed representations, it has been hypothesized Batch Normalization (BN) was the critical mechanism preventing collapse in BYOL [13], yet this hypothesis was refuted, showing batch-independent normalization schemes still achieve comparable performance [9,14].

2.2. Minimal hyperspherical energy and diversity regularization

Many unsupervised representational methods learn their representations under the constraint to lie on the surface of a unit-hypershpere via a ℓ_2-norm constrain leading to desirable traits [15]. As aforementioned [3,8] prove that the negative term (repulsion of negative views) in the contrastive loss is equivalent to the minimization of hyperspherical energy of representations. The minimization of hyperspherical energy, the Thompson Problem [16], is a well studied problem in Physics finding the minimal electrostatic potential energy configuration of electrons. Yet this problem has also found place in providing diversity regularization of neurons [10,17], avoiding undesired representation redundancy. Our work however, investigates whether these regularization methodologies, introducing greater feature diversity and reducing redundancy, can promote more uniformly distributed image representations in BYOL.

3. Uniform distribution of features

We now define the necessary components of our investigation, specifically, explicit uniformity constraints on the representation space derived from the InfoNCE contrastive loss [18] and hyperspherical energy redundancy regularization to enforce representation diversity through neuron uniformity.

3.1. Contrastive learning

We begin by defining the contrastive loss, specifically the InfoNCE loss informally as the softmax cross entropy loss to identify the positive view among the set of unrelated negative views. Formally, we give this in the notation style of [8], in which the popular case of contrastive loss is considered where an encoder $f : \mathbb{R}^n \to \mathcal{S}^{m+1}$ is trained and feature vectors are ℓ_2-normalized onto the unit-hyperspher $\mathcal{S}$ of m dimensions.

$$\mathcal{L}_{contrastive}(f;\tau,M) \triangleq \mathop{\mathbb{E}}_{\substack{(x,y)\sim P_{pos} \\ \{x_i^-\}_{i=1}^M \overset{i.i.d}{\sim} P_{data}}} \left[-log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right] \quad (1)$$

where $P_{data}(\cdot)$ is the distribution of data over $\mathbb{R}^n$, $P_{pos}(\cdot,\cdot)$ is the distribution over positive pairs (augmentations T_1,T_2 of image $X \sim P_{data}$) $\mathbb{R}^n \times \mathbb{R}^n$, $\tau > 0$ is a temperature hyperparameter, and $M \in \mathbb{Z}_+$ a fixed number of negative samples, i.e. $M = 2B - 1$ in [1] where B is the batch size. Additionally, under the assumption of our ℓ_2-norm constraint $f(\cdot) \triangleq f(\cdot)/\|f(\cdot)\|_2$.

3.2. The link to uniformity

It has been shown by the authors of [8] that there is a derivable link to the enforcement of uniformity in contrastive losses. From the loss in Eq.1, it is formally shown in [8] that directly optimizing a metric of *alignment* (encourages positive pair representations to be consistent) and *uniformity* (encourages negative pairs to be dissimilar by uniformity distributing representations) is equivalent when M is sufficiently large.

The *uniformity* loss is given by:

$$\mathcal{L}_{uniformity}(f;t) \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d}{\sim} P_{data}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right], t > 0, \quad (2)$$

This derivation, is of our primary interest, where we ponder if the explicit constraint on uniformity can improve representation diversity and subsequently improve performance of BYOL.

3.3. BYOL and its uniformity on the hypersphere

As previously mentioned, BYOL proposes an alternative to the contrastive paradigm, in which two networks, online (f_θ), and target (f_θ) are each input with a different view of the same image (x,y), with the online network tasked to predict the representations of an identical but temporally aged version of the online network, the target network. The BYOL architecture is depicted in Fig. 1, where the prediction is made via a multi-layer perceptron network $q_\theta(f_\theta(\cdot))$ independent of the target network. For specifics regarding BYOL architecture please refer to 3.6 or the original work [2].

The important distinction to the contrastive paradigm regards the loss function Eq.3 in which no negative samples are draw (i.e. augmentations/views from different source images). Instead, the loss can be seen as an equivalent to the alignment loss derived from Eq.1 by [8] (Appendix A.1), formally the BYOL loss is given as follows,

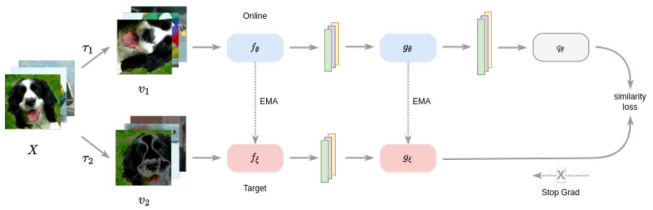


Fig. 1. Visual depiction of BYOL architecture [2].

$$\mathcal{L}_{BYOL}(\theta,\xi) \triangleq \mathop{\mathbb{E}}_{x,y\sim P_{pos}}\left[\|\overline{q}_\theta(f_\theta(x)) - \overline{f}_\xi(y)\|_2^2\right], \tag{3}$$

where $\overline{q}_\theta(f_\theta(x)) \triangleq q_\theta(f_\theta(x))/\|q_\theta(f_\theta(x))\|_2$ and $\overline{f}_\xi(y) \triangleq f_\xi(y)/\|f_\xi(y)\|_2$ are the normalization terms projecting the representation onto the unit-hypersphere.

From Eq. 3 we can observe there exists no term that enforces the separation and therefore diversity of negative views in space. The phenomena associated with the lack of diversity of representations pertaining to different input samples is known as *mode collapse*, where without such term, the network will simple learn trivial and constant representations. The current conjecture as to why BYOL does not exhibit mode collapse lies in the predictor network q_θ [2,14] with the derivation and explanation by [2] given in Appendix A.3.

To explore the distribution of representations and subsequently the effect of the predictor q_θ without explicit diversity constraints, we visualize the uniformity of representations produced by the AlexNet [19] encoder when training on the CIFAR-10 [20] dataset. Fig. 3 depicts the distribution of the validation set features, with Fig. 3c visualizing the distribution under BYOL procedure. We can observe that the distribution of image representations under the BYOL setting, although good and much better than random, is vastly less uniform compared to its contrastive counter-parts.

3.4. Explicit uniformity constraint

This observation leads to the question: *can BYOL benefit from mechanisms that introduce feature uniformity found in contrastive methods?* More specifically, can the uniformity loss, Eq.2, benefit the representations learned under BYOL. This question had been partly explored in [21] and indirectly via exploration of negative samples in [2], yet both of these consider representations produced by the projector and negatives computed from the target network. Given the intuition of BYOL behavior we instead minimize Eq.2 of the online projections f_θ only and independent of the predictor q_θ. The intuition behind this procedure is to enforce uniform distribution of features output by the online network, akin to constrastive, whilst maintaining the properties of the predictor to enforce variation via the maximization of information in the uniformly distributed online network. The combined loss is as follows

$$\mathcal{L}_{BYOL+Uni}(\theta,\xi) = \mathcal{L}_{BYOL}(\theta,\xi) + \lambda_{uni}\cdot\mathcal{L}_{uniformity}(f_\theta;t) \tag{4}$$

where $t = 2$ and λ_{uni} is a hyperparameter controlling the influence of the uniformity metric.

It can be observed from Fig. 3d, the addition of $\mathcal{L}_{uniformity}$ during BYOL training improves the uniformity of representations produced by the encoder, distributing akin to contrastive (Fig. 3b), confirming expected behavior.

Although the addition of the contrastive uniformity term exhibits greater uniformity of representations and subsequent improvement in performance (Table 3), this term relies on the number of negative samples, M, being sufficiently large. This is an undesirable computational necessity which BYOL aimed to remove, therefore introducing the contrastive uniformity metric contradicts the advantage of BYOL, the lack of negative samples. Further analysis of the representations captured and robustness to smaller batch size are examined in later sections.

3.5. MHE regularization

It has been shown by [9,14] that initialization and regularization of weights by batch normalization are fundamental to performant self-supervision. We aim to further extend the power of regularization of self-supervised learning to enforce uniformity of neurons and subsequently the produced representations, whilst avoid the negative sampling constraints imposed by contrastive terms.

We propose the use of hyperspherical regularization [10] alongside batch normalization to explicitly regularize the network to reduce hyperspherical energy of neurons (depicted in Fig. 2) to further improve the diversity of weights in the network and consequently representation uniformity. Fundamentally, such methods aim to reduce undesired representation redundancy occurring through non-uniform distribution of neurons. This choice is particularly motivated the findings in [10], where scenarios of class imbalance where unrepresented classes were shown to be well separated as a result of more uniformly distributed classification neurons. Additionally, [10] argues that the power of neural representations can be characterized by the hyperspherical energy of its neurons (i.e. kernels), and as such a minimal hyperspherical energy configurations can induce better diversity and improve representation separability. The hyperspherical energy for N neurons, in $\mathbb{R}^{(d+1)}$, $\boldsymbol{W}_N = \{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_N \in \mathbb{R}^{(d+1)}\}$ is defined as:

$$\begin{aligned}\boldsymbol{E}_s = \boldsymbol{E}_{s,d}\left(\hat{\boldsymbol{w}}_i|_{i=1}^N\right) &= \sum_{i=1}^N \sum_{j=1,j\neq i}^N r_s\left(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|\right) \\ &= \begin{cases} \sum_{i\neq j}\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-s}, & s>0 \\ \sum_{i\neq j}\log\left(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-1}\right), & s=0 \end{cases}\end{aligned} \tag{5}$$

where $\hat{\boldsymbol{w}}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$ is the i-th neuron weight projected onto $\mathcal{S}^d$, and $r_s(\cdot)$ is a decreasing real valued function, which is chosen to be the Riesz s-kernel, $r_s(z^{-s})$, $s>0$ [10]. We therefore aim to minimize the energy $\boldsymbol{E}_s$ in Eq.5 by manipulating the orientation of the neurons $\boldsymbol{W}_N$ to solve $\min_{\boldsymbol{W}N}\boldsymbol{E}_s$, $s \geq 0$. When $s = 0$, the logarithmic energy minimization problem is undertaken, essentially maximizing the product of Euclidean distance, where in our case this is the angle between neurons.

$$\arg\min_{\boldsymbol{W}_N}\boldsymbol{E}_0 = \arg\min_{\boldsymbol{W}_N}\exp(\boldsymbol{E}_0) = \arg\max_{\boldsymbol{W}_N}\prod_{i\neq j}\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\| \tag{6}$$

As an explicit regularization method, we optimize for the joint objective function:

$$\mathcal{L} = \mathcal{L}_{BYOL}(\theta,\varepsilon) + \lambda_{mhe}\cdot\sum_{j=1}^{L_\theta}\frac{1}{N_j(N_j-1)}\{\boldsymbol{E}_s\}_j \tag{7}$$

where λ_{mhe} is a hyperparameter to control the weighting of our regularization, L_θ the number of layers in the online network f_θ and/or predictor q_θ, and N_j is the number of neurons in layer j. A further variant has also been considered in this work simply extending the hyperspherical energy based on Euclidean distance in Eq.5 to consider geodesic distance on a unit hypersphere. We define this extension in Appendix A.4 For more details and all proofs we refer to [10].

3.5.1. Representation uniformity analysis

Demonstrated in Fig. 3 is the distribution of image representations under MHE regularization following the same visualization methodology

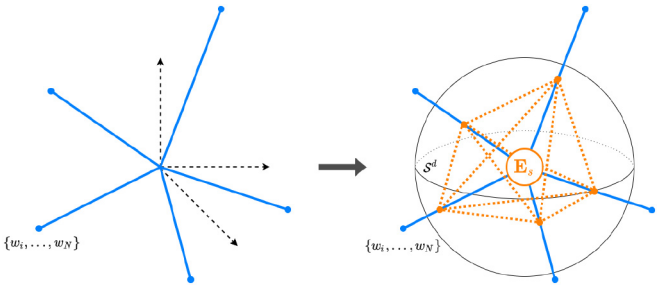
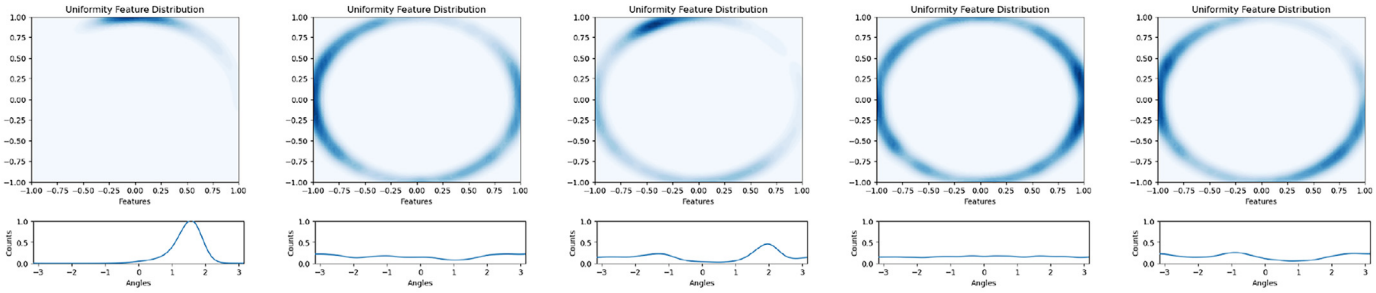


Fig. 2. Visual depiction of the regularization of neurons, $\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_N \in \mathbb{R}^{(d+1)}\}$, to minimum hyperspherical energy, $\boldsymbol{E}_s$, on the unit hypersphere $\mathcal{S}^d$. [10,17].



(a) Random Initalization (b) Contrastive Learning (c) BYOL (d) BYOL + Uni (e) BYOL + MHE Reg

Fig. 3. Learned representations of the CIFAR-10 validation set normalized on the unit hypersphere $\mathcal{S}^1$. The feature distribution is plotted via Gaussian Kernel Density Estimation (KDE) in $\mathbb{R}^2$. The corresponding angles for each (x,y) point in $\mathbb{R}^2$ on the unit hypersphere $\mathcal{S}^1$ is achieved using the von Mises-Fisher KDE [8].

presented in 3.4, we empirically confirm our hypothesis that improving the diversity of weights within the network subsequently results in more diversely distributed representations. Fig. 3e show a significant improvement in representation uniformity compared to the baseline in Fig. 3c. To further confirm these findings, we plot in Fig. 4 the hyperspherical energy of intermediate layer representations of a ResNet-18 encoder during training on the CIFAR-10 dataset between standard BYOL and BYOL with MHE regularization applied. We empirically show that regularizing the neurons via MHE maintains lower hyperspherical energy on its activation/representations throughout the whole network, compared to BYOL baseline. It is motivating to note that the final output layer representations demonstrate immediately lower hyperspherical energy, and increased uniformity by a measure of $G2$ (Appendix A.2), Fig. 5, throughout training by a considerable margin, an important factor for learning good representations applied to downstream tasks. The empirical finding in Fig. 4 support our hypothesis and rationale that increasing diversity of weights leads to representations that are in-turn more uniformly distributed. We report performance benchmarks in 4, and ablations in. 5.

3.6. Implementation details

The implementation follows the procedure presented in [2] with exception to the addition of the regularization loss terms. As to correspond with the BYOL procedure, we employ the same image augmentations as described in [1,2]. Similarly, our experimentation primarily focuses on the use of two different convolutional residual network [22]

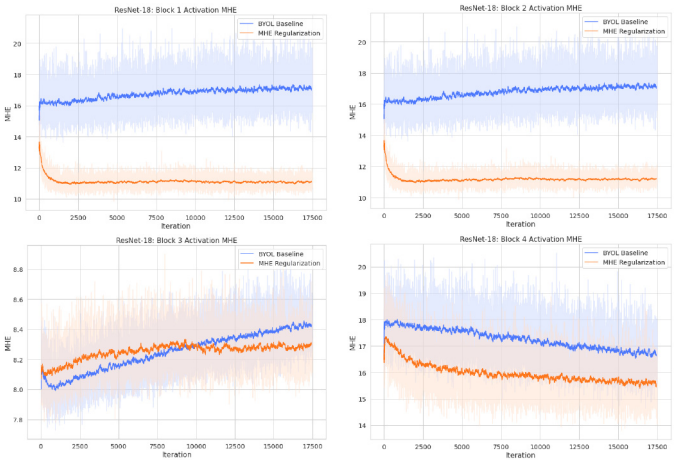
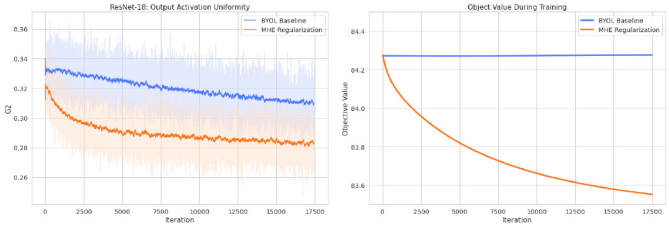


Fig. 4. Hyperspherical energy vs. iteration during training for intermediate representations of the ResNet-18 encoder. We compute the MHE on the output of each ResNet block [22].



(a) G2 (b) MHE Optimization Value

Fig. 5. Dynamics during training of the STL-10 dataset. (a) Uniformity measure $G2$ vs. iteration. (b) MHE regularization objective value vs. iteration.

configurations for our encoders f, ResNet-18 and ResNet-50. Following the procedure described in BYOL [2], the networks replace the standard linear output layer with a Multi-Layer Perceptron (MLP) g, projecting the output of the final average pooling layer to a smaller space. The MLP is a two layer linear network the first outputting in 4096 dimensions, followed by a second outputting to 256 dimensions. The first layer only is followed by batch normalization and Rectified Linear Unit (ReLU) non-linearity. Specifics regarding full augmentation details and optimization settings are given in full in Appendix B.

4. Linear evaluation

To evaluate the quality of representations learned during self-supervised training we employ the standard linear evaluation protocol described in [1,2]. For context, a linear classifier is trained taking as input the representations produced by the encoder f_θ which is pre-trained in a self-supervised manner and then frozen as to not train in a supervised manner. Table 1 reports the top-1 accuracy in % for the

Table 1
ImageNet Linear Classification: encoder trained for 1000 epochs.

Method	Arch.	Batch Size	top-1	k-NN
Supervised	RN50	–	79.3	79.3
SimCLR [1]	RN50	4096	69.1	60.7
MoCov2 [23]	RN50	4096	71.1	61.9
InfoMin [24]	RN50	4096	73.0	65.3
BarlowT [25]	RN50	4096	73.2	66.0
OBoW [26]	RN50	4096	73.8	61.9
SimSiam [14]	RN50	256	71.3	–
BYOL [2]	RN50	4096	74.3	64.8
BYOL* [2]	RN50	4096	74.1	63.7
BYOL-MHE	RN50	4096	**74.4**	**64.9**

We report top-1 accuracy (%) and k-NN accuracy. * = Reproduction, RN50 = ResNet-50.

Table 2
ImageNet Linear Classification: ResNet-50 encoder trained for 300 epochs.

Method	Top-1 (%)	Top-5 (%)
SimCLR [1]	67.9	88.5
BYOL [2]	72.5	90.8
BYOL*	71.9	89.2
BYOL-MHE	72.4	89.9

We report top-1 and top-5 Accuracy (%). * = Reproduction.

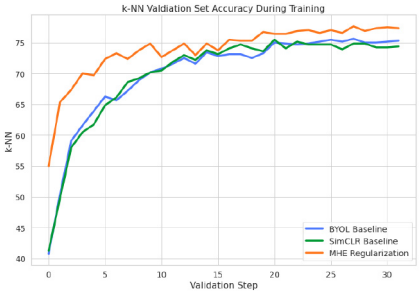


Fig. 6. *k*-Nearest Neighbor accuracy on the STL-10 validation set during training.

ImageNet ILSVRC-2012 test set trained with a standard ResNet-50 for 1000 epochs, we show our reproduction of methods alongside our MHE regularized variant. We empirically set $\lambda_{uni} = 0.125$, and $\lambda_{mhe} = 1.$, choosing the angular variant of MHE regularization with power $s = 2$ applying to all weights in the encoder, projector and predictor, Appendix C.

We report **74.4** top-1 accuracy with the inclusion of MHE regularization, a 0.32% improvement over the standard BYOL baseline, a significant improvement in performance which matches the jump made between other competing methodologies. Additionally, the improvement is maintained at lower epoch counts, and smaller batch sizes, with Table 2 demonstrating 0.5% improvement at 300 epochs and 1024 batch size. Robustness to hyperparameters is further explore in 5, yet these benchmark results are a substantial and clear improvement, further validating the capability of our regularization to provide better representations with minimal overhead. In addition to top-1, we report the weighted nearest neighbor classifier, *k*-NN, as a measure of representation semantic alignment and separability with very little variation due to hyperparameter values. Table 1 reports a large improvement over baseline by 1.0%, and similarly Fig. 6 mirrors such findings when training on STL-10.

In the case of CIFAR-10, CIFAR-100 and STL-10 datasets we see a more substantial improvement, reporting a top-1 accuracy of 94.78%, and 72.56% for CIFAR-10 and CIFAR-100 respectively. This 0.4% improvement in CIFAR-10 is comparable to the improvements found between SimCLR and BYOL, a substantial move towards the supervised baseline of 95.1% reported in [1]. For the explicit uniformity constraint, BYOL + $\mathcal{L}_{uniformity}$, we see on average 0.1% improvement from the MHE regularized variant. This improvement is expected, given the more explicit nature of the uniformity constraint on directly optimizing the representations rather than the implicit MHE regularization, and the observed near uniform distribution depicted in Fig. 3d. Interestingly, we

Table 3
Top-1 (%) Accuracies of Linear Evaluation on CIFAR10 and CIFAR100 datasets with ResNet-50 Encoder trained for 1000 epochs.

Method	CIFAR 10	CIFAR 100	STL-10
SimCLR*	93.81	70.98	82.40
BYOL*	94.46	72.10	82.81
BYOL + $\mathcal{L}_{Uni}$	**94.84**	**72.62**	83.19
BYOL-MHE	94.78	72.56	**83.96**

*= reproduction.

Table 4
Linear Evaluation on CIFAR10 under different regularization configurations, all networks f_θ, g_θ, q_θ, are regularized when selected.

BN	MHE	Accuracy (%)
✘	✘	28.76
✓	✘	90.74
✘	✓	45.72
✓	✓	91.22

find the MHE regularization to be the best performing setting on the STL-10 dataset, with a 1.1% improvement over baseline, and 0.8% improvement over the explicit uniformity constraint. We can conjecture that the large and diverse nature of the semantic classes in STL-10 unlabeled set benefit more from the unique representation neuron effect that enables unrepresented concepts/classes to be uniquely and evenly assigned [10].

5. Ablation and sensitivity analysis

We analyze the behavior of our BYOL constraints exploring the impact of hyperparameter and network configurations. We follow the procedure described in 3.6 and 3.6, training a ResNet-18 encoder for 300 epochs.

5.1. Batch size

One primary advantage BYOL introduced is the robustness to smaller batch sizes, this emerges from the avoidance of negative pairs sampled from within the batch in end-to-end contrastive models. Therefore, with our addition of $\mathcal{L}_{uniformity}$ (Eq.2) being derived from Eq.1, we expect robustness to batch size to degrade. We test the performance under different batch size averaging gradients over N consecutive steps before updating the network parameters, where N is the factor of batch size reduction from the baseline [2]. Fig. 8 shows that the introduction of the explicit uniformity loss reduces robustness to batch size as expected. We see from a baseline of 91.48%, a − 9.06% drop with $\mathcal{L}_{uniformity}(f_\theta)$ compared to BYOL's − 5.74%. This expected result confirms our reasoning to find alternative mechanisms to enforce uniformity of image representations. For MHE regularization, we observe little deviation of performance compared to standard BYOL given the regularization's independence on batch size.

5.2. MHE regularization and batch normalization

Following our intuition and empirical findings that MHE regularization encourages representation uniformity, we further investigate the effect of regularization components. We first explore the uniformity of CIFAR-10 validation set representations as done in Fig. 2. We can see the representations in $\mathcal{S}^1$ plotted in Fig. 7 and their corresponding linear evaluation results in Table 4. The results empirically show how without batch normalization the network fails to learn whilst poorly distributing representations in space, resulting in collapsed representations coinciding with [9]. Confirming our previous empirical results, BYOL with MHE

Table 5
Linear Evaluation (Accuracy %) on CIFAR10 given different network configurations of MHE regularization, '✓' denotes that MHE regularization has been applied to that sub-network.

Layer								
Encoder $f_{\theta,\xi}$	–	✓						
Projector $g_{\theta,\xi}$	–	✓	–	✓				
Predictor q_θ	–	✓	–	✓	–	✓	–	✓
MHE (a2)	90.74†	**91.64**	91.36	91.46	91.38	91.10	90.96	91.22
BYOL + $\mathcal{L}_{Uni}$	91.48							

The encoder is ResNet-18 trained for 300 epochs. †= BYOL (repro).

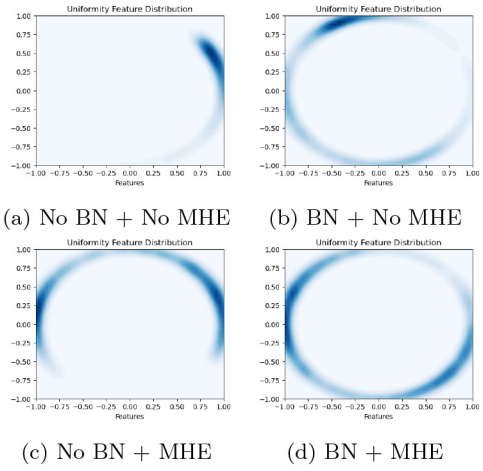


Fig. 7. Uniformity of representations on $\mathcal{S}^1$ under different regularization configurations plotted with Gaussian KDE.

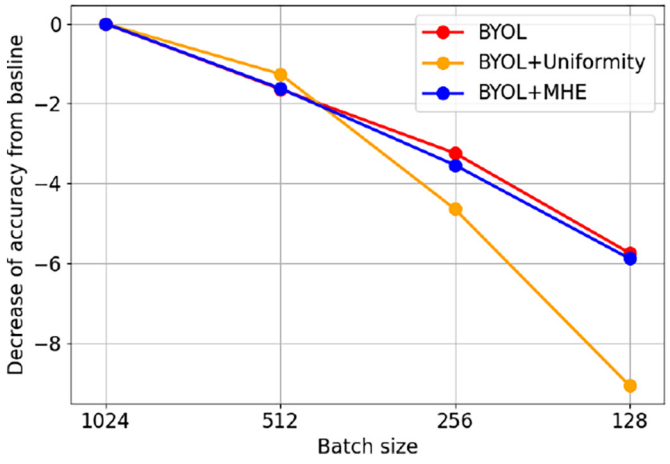


Fig. 8. Reduction in CIFAR-10 linear evaluation top-1% when decreasing batch size for the proposed variants and repro of BYOL.

regularized alone produces image representations that are distributed far more uniform than batch norm, Fig. 7c. However, the linear evaluation performance suffers compared to batch normalization, although our regularization has a similar effect to batch norm in avoiding collapse of representations, albeit with less impact. We conclude from these finding that regularization is a key component to avoid mode collapse in self-distillation methods, where batch normalization is not a fundamental necessity rather the diversity of neurons and reduction in redundancy provided by MHE is enough to encourage variation in learned representations. This is a promising finding which warrants further investigation in future work.

5.3. MHE regularization parameterization

To investigate how varying hyperparameters for the MHE regularization affects performance, we report results for network configurations in Table 5. Additionally, the weight of the regularization λ_{mhe} and powers s are given in Appendix D.

We report in Table 5 the linear evaluation performance under varying configurations of MHE regularization to individual sub-networks. We show that across all configurations we see an increase in performance, showing that the improved weight diversity and subsequent representation diversity improves the quality of representations learned. Additionally, referring to our previous notion that it is not preferable to directly enforce uniformity at the predictor of the BYOL architecture based on the intuition of BYOL's behavior [2], we do not see any degradation in performance when MHE is applied at the predictor level. We conjecture that the improved diversity of features helps assist the online network in capturing more varied representations.

6. Conclusion

We empirically show that uniformity constrains like those in contrastive losses can be beneficial in BYOL and self-distillation methods in general where negative samples are negated. To maintain the computation benefits proposed by BYOL we investigate the use of regularization methods that minimize the hyperspherical energy between network neurons. We show that this type of redundancy regularization implicitly improves distribution uniformity representations learned by the encoder, leading to improved results in all experimentation over the baseline whilst remaining robust to changes in batch size, with minimal additional computational. Empirical exploration demonstrates the degree in which MHE regularization impacts the uniformity of representations during training throughout the encoder network, validating our intuition that more diverse neurons result in more diverse representations.

Performance gains from our regularization are significant given no architectural change, nor augmentation change, common in alternative approaches. We believe further performance improvements can be made with tuning of hyperparameters. Yet how the avoidance of fully collapsed equilibria in the presence of MHE regularization identified in this work is still yet to be understood, as is how the maximization of kernel diversity improves activation diversity. However, from this work we have identified the importance of regularization in self-supervision and its effect on learned image representations in space.

6.1. Future work

This works empirically identifies unexpected training behavior of the self-supervised, self-distilled method BYOL, and as such expanding this exploration to alternative methods is a natural continuation. In addition, the further analysis of regularization in self-supervision as a whole is an importance next step to understand the training dynamics. Furthermore, the identified phenomena shows such regularization impacting uniformity may be enough to solely avoid mode collapse currently prevented by the predictor network [14], establishing the hypothesis for future investigations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk). Access granted through the project: ec173 - Next gen self-supervised learning systems for vision tasks, supported by the Engineering and Physical Sciences Research Council (EPSRC).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imavis.2022.104494.

References

[1] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, International Conference on Machine Learning, PMLR 2020, pp. 1597–1607.

[2] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Pires, Z. Guo, M. Azar, et al., Bootstrap your own latent: A new approach to self-supervised learning, Neural Information Processing Systems, 2020.

[3] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, 34th Conference on Neural Information Processing Systems, NeurIPS'20, vol. 33, Curran Associates, Inc 2020, pp. 9912–9924.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging Properties in Self-Supervised Vision Transformers, IEEE/CVF International Conference on Computer Vision, 2021 9650–9660.

[5] Y. Zhang, Q. Liang, K. Zou, Z. Li, W. Sun, Y. Wang, Self-supervised part segmentation via motion imitation, Image Vis. Comput. 104393 (2022).

[6] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, IEEE 2005, pp. 539–546.

[7] T. Chen, L. Li, Intriguing Properties of Contrastive Losses, Advances in Neural Information Processing Systems, 34, 2021.

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

[9] P.H. Richemond, J.-B. Grill, F. Altché, C. Tallec, F. Strub, A. Brock, S. Smith, S. De, R. Pascanu, B. Piot, et al., Byol Works Even Without Batch Statistics, arXiv Preprint, 2020 (arXiv:2010.10241).

[10] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, L. Song, Learning towards minimum hyperspherical energy, Adv. Neural Inf. Proces. Syst. 31 (2018) 6222–6233.

[11] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 9729–9738.

[12] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, Advances in Neural Information Processing Systems 2017, pp. 1195–1204.

[13] A. Fetterman, J. Albrecht, Understanding Self-Supervised and Contrastive Learning with Bootstrap Your Own Latent (byol), https://www.untitled-ai.com/understanding-self-supervised-contrastive-learning.html 2020.

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

[15] J. Xu, G. Durrett, Spherical latent spaces for stable variational autoencoders, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 2018, pp. 4503–4513.

[16] J.J. Thomson, Xxiv. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure, Lond. Edinb. Dublin Philos. Mag. J. Sci. 7 (39) (1904) 237–265.

[17] R. Lin, W. Liu, Z. Liu, C. Feng, Z. Yu, J.M. Rehg, L. Xiong, L. Song, Regularizing neural networks via minimizing hyperspherical energy, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 6917–6927.

[18] A.V.D. Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv Preprint, 2018 (arXiv:1807.03748).

[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.

[20] A. Krizhevsky, G. Hinton, et al, Learning Multiple Layers of Features from Tiny Images, University of Toronto, 2009.

[21] H. Shi, D. Luo, S. Tang, J. Wang, Y. Zhuang, Run Away from your Teacher: Understanding Byol by a Novel Self-Supervised Approach, arXiv Preprint, 2020 (arXiv:2011.10944).

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 770–778.

[23] X. Chen, H. Fan, R. Girshick, K. He, Improved Baselines with Momentum Contrastive Learning, arXiv Preprint, 2020 (arXiv:2003.04297).

[24] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, P. Isola, What Makes for Good Views for Contrastive Learning, Advances in Neural Information Processing Systems, 33, 2020 6827–6839.

[25] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow Twins: Self-Supervised Learning Via Redundancy Reduction, In *International Conference on Machine Learning*, PMLR, 2021 12310–12320.

[26] S. Gidaris, A. Bursuc, G. Puy, N. Komodakis, M. Cord, P. Pérez, Online Bag-of-Visual-Words Generation for Unsupervised Representation Learning, InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021 6830–6840.