MoBYv2AL: Self-supervised Active Learning for Image Classification

Razvan Caramalau[1], Binod Bhatarrai[2], Danail Stoyanov[2], Tae-Kyun Kim[1,3]

[1]Imperial College London, [2]University College London, UK, [3]KAIST, South Korea

Introduction

- **Active Learning** (AL) has recently gained more popularity in the research community. The goal of AL is to sample the **most informative and diverse examples** from a large pool of unlabelled data to query their labels. The existing AL methods can be grouped into two based on the selection criteria.
- The first group is uncertainty-based algorithms that select the challenging and informative examples. Whereas **representative-based** algorithms select the most diverse examples from the data set. To select diverse examples, existing methods first project the images into a feature space followed by applying sampling techniques such as CoreSet. Our work falls in the latter category.
- **Self-supervised learning** (SSL) methods have made tremendous progress in generating discriminative representations of the images. One of the earliest AL works in this direction **CSAL** employed consistency loss between the input image and its geometrically augmented versions along with the objective of downstream tasks. However, this method limits augmentation methods in the primitive form. Similarly, J. Bengar et al. introduced contrastive learning in AL, but the self-supervised method and end-task objective are optimised in multi-stage form. This makes the model sub-optimal, affecting the features' representativeness during selection. Simple random labelling overpasses any AL criteria. Thus, the existing works in this direction show explicit limitations.
- To address the issues of those methods, we introduce contrastive learning as **MoBYv2** (from its SSL predecessor MoBY) in our AL pipeline, **MoBYv2AL**, and jointly train the learner. We choose MoBY SSL because it addresses the computational complexities and shortcomings of other previous methods, such as SimCLR or BYOL.
- **MoBYv2** has two branches: One updates with gradient (query encoder) and another with momentum (key encoder). The parameters of the momentum encoder are updated in slow-moving averages with the query one. Moreover, the memory bank of keys from the momentum encoder keeps long dependencies with several mini-batches. Apart from minimising a contrastive loss, another advantage consists in the asymmetric structure of BYOL that captures distances from mean representation. The AL process of **MoBYv2AL** culminates with the concept-aware selection function, **CoreSet**.

Method

- We tackle the contrastive unsupervised learning approach compared to previous semi-supervised AL techniques that rely on consistency measurement.
- We design the self-supervision framework according to MoBY. This method combines two innovative prior works **MoCo** and **BYOL** on visual transformers. We intuitively explore the contrastive learning strategies from both MoCo and BYOL and align the self-supervision with MoBY.
- From a design perspective, we adopt the asymmetric dual encoders from BYOL as shown in the top (middle) Figure. The top branch culminates with a discriminator to match the outputs from the bottom. Despite this, both branches consist of the same feature extractor architecture followed by an MLP projector for query and key, respectively. Distinctively from MoBY, we tackle convolutional neural networks (CNNs) as feature encoders. Moreover, we reduce the MLP projectors and the query discriminator to a single layer with batch normalisation and ReLU activation.
- We also choose another set of augmentations by alternating strong and weak augmentations, similarly to MoCov2. This change boosted the performance of its predecessor MoCo. We also observed in our experiments that using only strong augmentations can affect the optimisation of the task-aware branch.
- The asymmetric pipeline helps to mimic the contrastive learning principle of BYOL. However, to include the concepts from MoCo, we minimise our objective with the InfoNCE loss. In this case, we will also need to keep the memory bank for the queue of keys. We define the contrastive loss as a sum of InfoNCE from two augmented versions of a query and of a different key:

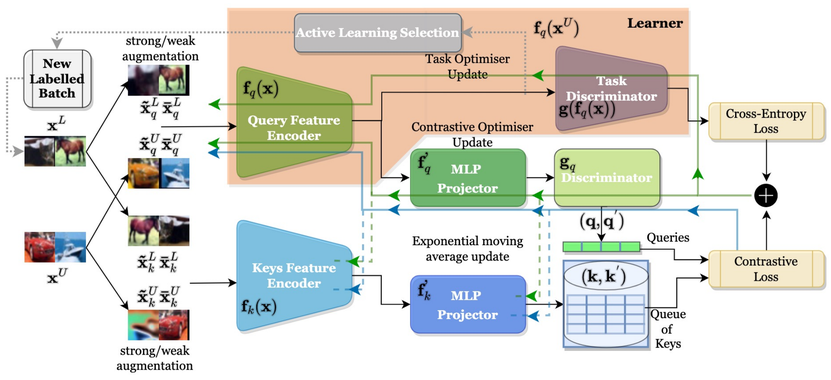
$$\mathcal{L}_{contrastive} = -\log \frac{exp(q \cdot k^{'}/\tau)}{\sum_{i=0}^{m} \exp(q \cdot k_i^{'}/\tau)} - \log \frac{exp(q^{'} \cdot k/\tau)}{\sum_{i=0}^{m} \exp(q^{'} \cdot k_i/\tau))}$$

- Similarly, we can compute the contrastive loss for the labelled images. In addition, we also minimise the categorical cross-entropy, with the output from the task discriminator. Once computed, we back-propagate both the contrastive and the classification loss. Therefore, the combined loss, adjusted by a scaling factor can be expressed as:

$$\mathcal{L}^{L}_{combined} = \mathcal{L}_{classification} + \lambda_c \mathcal{L}^{L}_{contrastive}$$

- Once trained, the newly set of features enriched with the SSL pipeline of labelled and unlabelled images are passed to the data representativeness AL function CoreSet.

MoBYv2AL Diagram



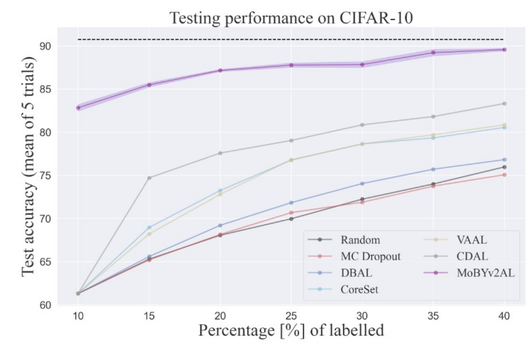
SSL-AL training framework under the proposed MoBYv2AL configuration

The query feature encoder plays two roles: to *map* the features to the task discriminator for classification; to *capture* contrastive visual representation with the asymmetry of the query and key modules. For unlabelled data, the blue lines show the back-propagation of contrastive loss and its exponential moving average (dashed). The green lines also include the cross-entropy loss during training when the annotation is available. Once training ends, the unlabelled samples pass through the learner for AL selection.

Experiments

- **Models.** We use different CNNs for feature encoders. To show that MoBYv2AL is robust to architectural changes, we opt for VGG-16 in the CIFAR-10/100 quantitative experiments and for ResNet-18 in SVHN and FashionMNIST.
- **Training settings**. In terms of training MoBYv2, we optimise the learner and the SSL modules with Stochastic Gradient Descent. We train at every selection stage for 200 epochs, and we keep the batch size at 128. The dictionary size for the keys is set up as in MoBY at 4096. We noticed in our experiments that the contrastive and cross-entropy loss converge together after 200 epochs. The learning rate starts at 0.01, and it follows a schedule for the queue encoder and task discriminator that decreases ten times at 120 and 160 epochs. However, we keep the momentum scheduler update in the key bottom branch (gradual momentum increment from 0.99). In the contrastive loss, for both queues, we fix the temperature parameter to 0.2.

Image Classification – CIFAR10/100



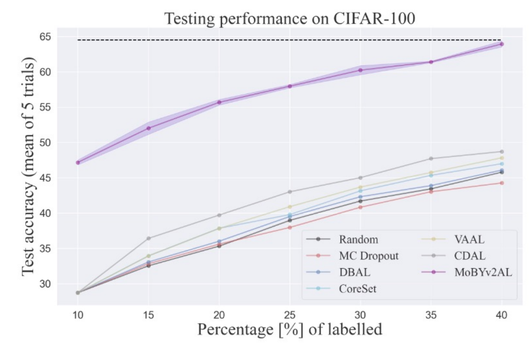


Image Classification – SVHN/FashionMNIST

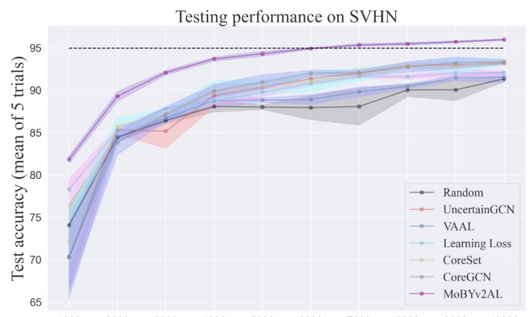
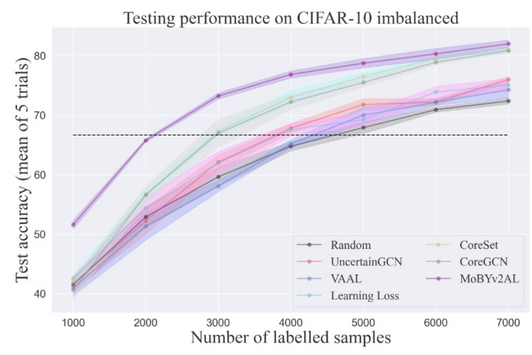
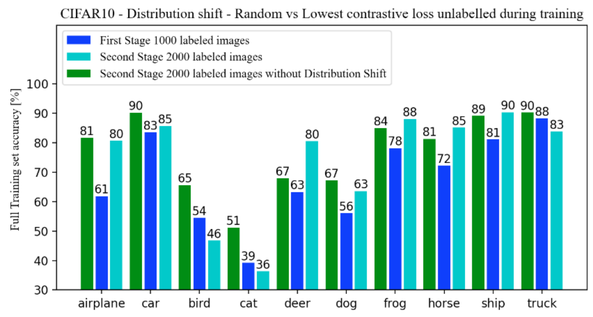


Image Classification – Imbalanced dataset



Mitigating distribution shift



Ablation studies

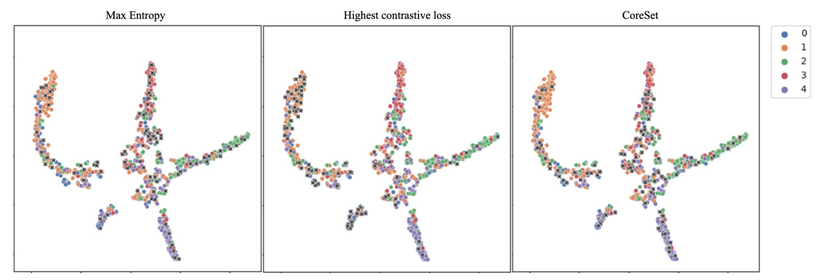
SSL model / No. of labelled	1000	2000	3000
MoCov2	11.62±.9	11.92±.6	12.89±.6
BYOL	12.32±.7	11.72±.4	11.47±.2
MoBY	62.62±.4	72±.5	76.43±.1
MoBYv2AL (Ours)	**63.06±.5**	**76.04±.6**	**80.63±.3**

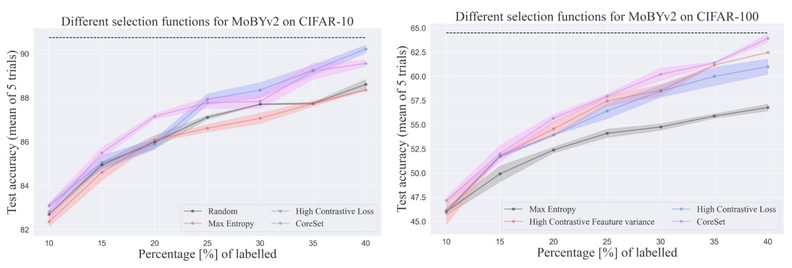
MoBYv2AL / No. of labelled	1000	2000	3000
w/o Discriminator	60.44±.4	72.53±.8	77.89±.3
w/o MLP Projector	58.57±.6	71.96±.5	77.02±.6
w/o Strong Augmentation	47.7±.4	58±.5	64.85±.5
MoBYv2AL (Ours)	**63.06±.5**	**76.04±.6**	**80.63±.3**

MoBYv2AL	1000	2000	3000
Multi-stage semi supervised	34.8±.1	34.96±.2	35.09±.1
Jointly with end-task	63.06±.5	76.04±.6	80.63±.3

SSL method	Supervised	MoCov2	BYOL	DINO	MoBYv2AL
CIFAR-10 Test accuracy	90.08	76.7	77.89	81.2	**88.62**

Active Learning – Selection Function analysis





Conclusion

We have presented an SSL-based AL framework for image classification. The main contributions lie in:

- the task-aware contrastive learning pipeline. MoBYv2 retains the higher visual concepts and aligns them with the downstream task.
- the end-to-end training is efficient and modular, allowing diverse learners and sampling functions.
- quantitative experiments demonstrate the state-of-the-art on four datasets. Our method shows robustness even in simulated class-imbalanced data pools..
- MoBYv2AL tackles the 'cold-start problem' and distribution shift.

Acknowledgements

Sponsored by KAIA grant (22CTAP-C163793-02, MOLIT), NST grant (CRC 21011, MSIT), KOCCA grant (R2022020028, MCST) and the Samsung Display corporation; Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; the Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; and the Royal Academy of Engineering Chair in Emerging Technologies Scheme; and EndoMapper project by Horizon 2020 FET (GA 863146).

Contact Information

Razvan Caramalau

Tel: +447504329733

Email: r.caramalau18@imperial.ac.uk