Transductive self-supervised representation learning for lowering the generalization gap

Andrei-Eusebiu Blahovici*
University of Bucharest, Romania
blahoviciandrei1@gmail.com

Marian-Antonio Bigan*
University of Bucharest, Romania
biganantonio@gmail.com

***Abstract*—Recent studies use deep convolutional neural networks, trained on large datasets, to learn general relevant features which are further used for other image-based tasks or other smaller datasets. Even though these types of pre-trained neural networks show promising results in a lot of computer vision tasks, they also extract the biases in the initial data and are affected by a phenomenon known as domain shift. This happens when a model is evaluated against a novel dataset or task and shows poor performance because of its lack of generalization capabilities. We address this problem by pre-training our encoder in a self supervised manner on data coming from out-of-distribution, evaluating on the ERM and IRM algorithms.**

I. Introduction and related work

a) Relation to transductive transfer learning: Transductive learning assumes that you have access to data from two different distributions: training data accompanied by labels, and testing data that is not annotated. There has been extensive research in this branch, specifically for domain adaptation. We can give information to the model of how the testing distribution looks like using transductive methods as in [1], [2]. In our approach, we expose the encoder in our architecture (see Fig. 1) to out of distribution samples, in an unsupervised manner.

b) Relation to self-supervised representations: Self-supervised learning has been shown repeatedly that it can obtain state-of-the-art results for different tasks. Using self-supervised representations, the authors of BYOL [3] managed to obtain state-of-the-art results without explicitly using negative samples. Meanwhile, we can see important results, using contrastive learning as in SimCLR [4]. In our preliminary experiments, we use just an autoencoder, but we plan to extend it to other contrastive methods as future work.

c) Relation to Generalization: There are several algorithms and setups that addresses the generalization problem, when the input distribution is shifting. We choose Invariant Risk Minimization (IRM [5]) which proposes a training paradigm that is trying to solve a problem that the classical Empirical Risk Minimization algorithm suffers from. That is, machine learning algorithms are able to learn really complex prediction rules, but because of selection biases and confounding factors it does not generalize well in all possible situations. We compare in our experiments the generalization capabilities on out-of-distribution data of IRB and ERM, before and after pre-training our encoder in a transductive way.

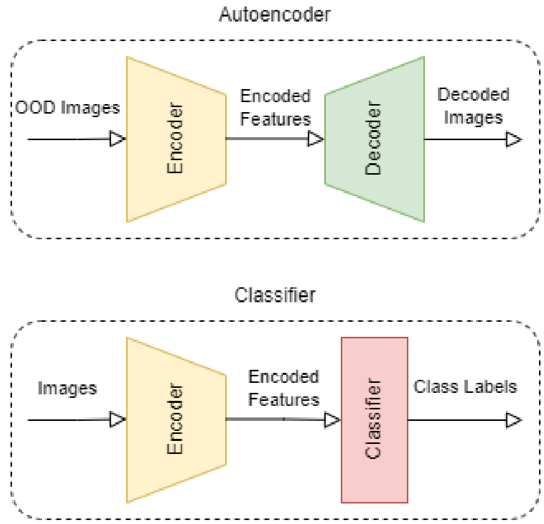

Fig. 1. The above figure shows the architectures of the autoencoder and classifier respectively. We first train the classifier with no pre-trained weights. For the other experiments we pre-train the autoencoder, then copy its weights in the classifier and either fine-tune them or keep them frozen.

II. Our approach

In this paper we aim to study a way of improving these large convolutional neural networks in a transductive manner, using data from test distribution (OOD) in a self-supervised manner. We are going to evaluate classifiers which learned only from training data, with models that have their backbone pre-trained using transductive learning. For the backbone of choice we are going to use an encoder, whilst training the classifier using two different algorithms, ERM and IRM, as shown in Fig. 1.

III. Implementation details and results

a) Dataset and data processing: We use a subset of the fMoW dataset from WILDS [6], whose objective is to stimulate the development of machine learning models that predict the functionality of buildings and land using satellite images. This dataset achieves domain shift by splitting the dataset by time ranges. The training data contains images taken

* Equal contribution.

TABLE I
SIZE OF THE FMOW-WILDS DATASET SUBSAMPLED SPLITS THAT WE USE IN OUR EXPERIMENTS (TRAIN, IN-DISTRIBUTION, OUT-OF-DISTRIBUTION).

train	ood_val	id_val
22,043	3,343	5,807

between 2002-2013, validation has images from 2013-2016 and lastly test in the range 2016-2018.

The subset that we use includes data for the most frequent 10 categories in the training data. Also, we subsampled the data even further by taking images taken only from the regions of Europe and Americas. We list the sizes of the training, in-distribution (ID) validation and out-of-distribution (OOD) validation sets in Tab. I.

b) Experimental setup: For these experiments, we decided to use an encoder that maps the images to the feature space which are then being fed to a classifier. The encoder is made up of several blocks, each consisting of two repeating groups of a convolutional layer, an activation function and a batch normalization. The block has a max pooling layer at the end. The classifier is a fully connected layer applied to the flattened output of the encoder. We use cross-entropy as the loss function and Adam with a learning rate of 1e-4 as the optimizer to train all of our models for a varying number of epochs.

c) Autoencoder setup: We use self-supervised learning to pre-train the encoder part of the models. For this reason, we define an autoencoder architecture that includes the encoder and a decoder which uses blocks with a similar structure but with a deconvolutional layer instead of max pooling, for upsampling. The architecture is inspired from U-Net [7] that has its skip connections removed. We train the autoencoder on the OOD validation set for 30 epochs using mean square error as the loss function. For the optimizer we choose Adam with a learning rate of 1e-4 and a L2 regularization of 1e-5.

d) Algorithms setup: We use in parallel two algorithms, ERM and IRM, and two autoencoder block depths while taking four directions to see how we can improve generalisation. The training takes 15 epochs when using ERM and 30 epochs when using IRM.

A. Experiments

We conducted multiple experiments on the presented architecture.

a) ERM and IRM - Base Models: For the first one, we don't use self-supervised learning, but train the encoder and the classifier only on the training set. We use its metrics for reference to the following methods. See columns 1-2 and 5-6 in Fig. 2.

b) Frozen encoder: Our second approach implies using self-supervised learning to pre-train the encoder and fixing all of its parameters. We train only the classifier's parameters. See columns 3-4 and 7-8 in Fig. 2.

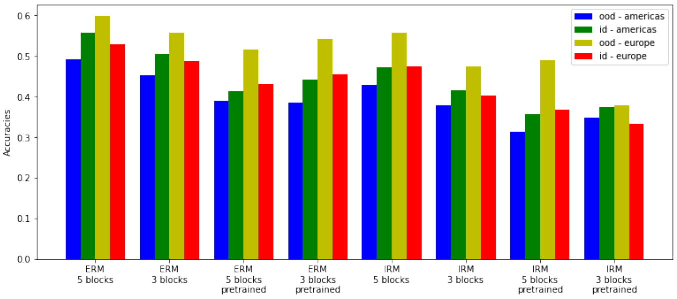

Fig. 2. The bar plot above shows the results we obtained for experiments a) and b). We plotted the accuracies obtained on the out-of-distribution and in-distribution validations sets for both regions, Europe and Americas.

c) Unfrozen encoder: On the third approach, we train the encoder in the same way as on the previous approach, but this time we do not freeze the parameters of the encoder and we fine-tune them while training the classifier from scratch. The results of this approach are similar to those of the first, pre-training does not seem to affect the outcome.

d) Alternate the training iterations: For the fourth direction, we choose to alternate the supervised training of the encoder and the classifier on the training data with the self-supervised learning of the encoder and the decoder. The alternation is done at batch level, each supervised learning batch being followed by a self-supervised learning batch. After several epochs, 5 for ERM and 10 for IRM, we stop using self-supervised learning and train only using supervised learning for the rest of the epochs. For this direction the results do not show improvements, the accuracies being even slightly smaller than those for the first approach.

IV. CONCLUSIONS AND FUTURE WORK

The quantitative results do not show an improvement when pretraining the encoder on ood data, in any of our scenarios. The accuracy is worse on models with a pre-trained backbone than on the ones that have only learned from training data. Interestingly, the IRM algorithm which is developed specifically for generalization tasks seems to perform worse than ERM which is the traditional method of training neural networks.

Our approach can be further improved by switching from an autoencoder to something more complex like an adversarial trained generator. There is already research in this direction, for domain adaptation [8], which shows promising preliminary results. As future work we propose the comparison against our method both on WILDS and on their dataset. Also an other direction would be to take advantage of the contrastive methods for the self-supervised pre-training step.

REFERENCES

[1] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Unsupervised transductive domain adaptation," 2016. [Online]. Available: https://arxiv.org/abs/1602.03534

[2] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 766–785, mar 2021. [Online]. Available: https://doi.org/10.1109/tpami.2019.2945942

[3] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," 2021. [Online]. Available: https://arxiv.org/abs/2103.06695

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[5] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019. [Online]. Available: https://arxiv.org/abs/1907.02893

[6] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "WILDS: A benchmark of in-the-wild distribution shifts," *CoRR*, vol. abs/2012.07421, 2020. [Online]. Available: https://arxiv.org/abs/2012.07421

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[8] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *CoRR*, vol. abs/1702.05464, 2017. [Online]. Available: http://arxiv.org/abs/1702.05464