Video Processing and Understanding Lab

Universidad Autónoma de Madrid

SI1/PJI/2019-00414 AISEEME (2020-2022)

Aiding diagnosis by self-supervised deep learning from unlabeled medical imaging

D3

Design of a curriculum-based multi-task self-supervised learning regime

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

CONSEJERÍA DE EDUCACIÓN E INVESTIGACIÓN

Comunidad de Madrid

AUTHORS LIST

Pablo Carballeira López	pablo.carballeira@uam.es
Marcos Escudero Viñolo	marcos.escudero@uam.es
Kirill Sirotkin	kirill.sirotkin@uam.es

HISTORY

Version	**Date**	**Editor**	**Description**
0.1	05/01/2022	Pablo Carballeira López	First draft
1.0	04/08/2022	Pablo Carballeira López	First version

CONTENTS:

1. Introduction

This deliverable describes the work related to tasks T3.1: “Empirical definition and completion of a pretext task curriculum”, T3.2: “Evaluation of the impact of the architecture and training schedule”, and T3.3: “Self-paced multi-task self-supervision”. The aim of tasks T3.1 and T3.2 is to define pretext tasks orderings (curricula) and evaluate the dependencies between the task curriculum and the training framework. Task 3.3 aims to define a learning framework that permits to automatically define a pretext task curriculum for a given target task.

2. Empirical definition and completion of a pretext task curriculum

In scenarios with a lack of training data, it is beneficial to pretrain CNN models on (preferably) domain-similar datasets to obtain a better starting point for training on the downstream task (i.e., the target task, such as skin lesion classification) [1]. Usually, such starting points are obtained by supervised training of a CNN model on a large and widely accepted as representative dataset, such as ImageNet [2]. Importantly, even when the domains of datasets used for pretraining and downstream task training differ significantly, pretraining still yields better performance than using randomly initialized weights [1]. Nonetheless, this limits the model selection to architectures with publicly available ImageNet-pretrained models, implying a computationally expensive and time-consuming training procedure for new model architectures, or architectures designed ad-hoc for specific tasks, such as skin lesion recognition.

In such situations, especially when labeled data in the target domain is scarce or non-existent, a common method to pretrain CNN models is to use Self-Supervised Learning (SSL) - a subset of unsupervised learning methods that leverages automatically generated labels as training objectives. Previous works show the advantages of SSL-pretraining applied for object recognition [3], where SSL-pretrained models outperform models pretrained on ImageNet in a supervised regime, on object recognition tasks or skin lesion assessment [4], where SSL-pretraining makes models more robust to noise. Here, we propose a step further for SSL pretraining by showing that the consecutive use of properly ordered pretext tasks can improve transfer learning results. Curriculum learning strategies [5] propose to order samples during training according to their learning outcomes. Inspired by these techniques, we propose to use a curriculum ordering of pretext tasks.

We have designed a framework to train visual models based on CNN architectures using a sequence of different SSL pretext tasks. A diagram of the implemented training scheme can be found in Figure 1. Curriculum (or anti-curriculum) orderings are defined empirically, based on the performance of individual pretext tasks, after transfer learning to the target task. Specifically, we define as curriculum orderings those coinciding with the increasing individual pretext task transfer learning performances. We group all orderings as follows:

- Curriculum: ordered by increasing performance.
- Anti-curriculum: ordered by decreasing performance.
- Full (anti-)curriculum: (anti-)curriculum comprised by all tasks.

Any other ordering is considered a Mixed-curriculum.

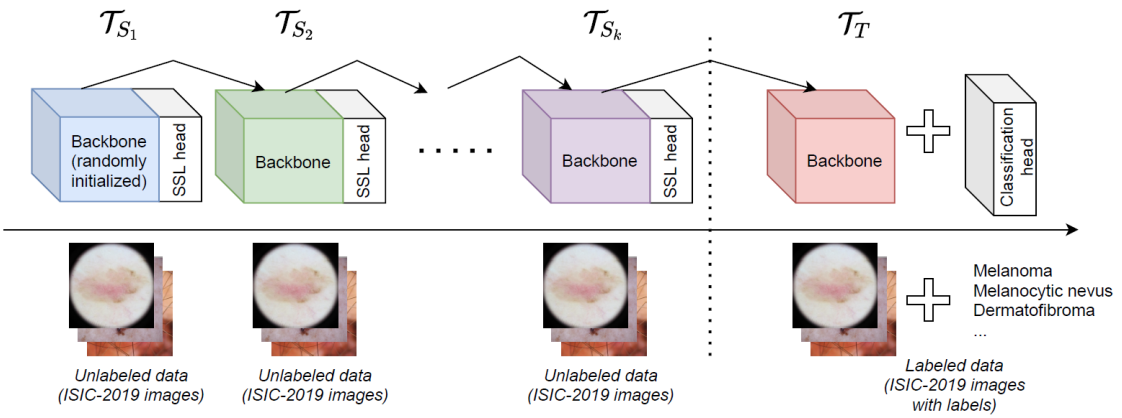

Figure 1 Implemented training scheme using a sequential pretext task curriculum.

The implemented framework is based on the MMSelfSup framework [6], and allows to evaluate the performance of different curriculum SSL training configurations, including number of pretext tasks and variable orderings. The design of experiments on this framework has been based on the hypothesis that, given a downstream task, an advantageous ordering of pretext tasks can be obtained by sorting the tasks according to the increasing order of their individual performances (curriculum orderings).

This training scheme has been described in the following paper [7] that has been submitted to the IEEE Journal BHI, which is currently under review. This training scheme has shown advantages over other state-of-the-art alternatives in the use-case of skin lesion image classification (which will be reported in deliverable D4), and we plan to evaluate its performance on a different medical image domain, namely x-ray lung images (the obtained results will also be reported in deliverable D4).

3. Evaluation of the impact of the architecture and training schedule

3.1. Evaluation of the impact of training schedule

The training process of any CNN architecture requires the configuration of a set of training hyperparameters that have a direct influence on the training pace (how fast the training process converges) and the final performance of the trained model. These hyperparameters include the learning schedule, learning rate or batch size. Generally, the optimization of these hyperparameters requires to repeat training processes with different hyperparameter configurations, increasing the required computational resources (i.e., GPU memory, time, disk space). To address this issue, some approaches attempt to reduce the amount of resources that each configuration uses [8][9] through early stopping, so that a higher number of configurations can be evaluated within a feasible training time and computational resources requirements.

The curricular SSL training scheme depicted in Figure 1 requires the execution of multiple sequential training steps, where the training hyperparameters of each step may influence the subsequent ones, as it sets the starting point for the model weights in the next training step. An exhaustive evaluation of all hyperparameter combinations, for all training steps, would exponentially increase the required training resources beyond a feasible limit.

To verify the influence of hyperparameter configuration in each training step on the performance of the downstream task, we have carried out a series of experiments to estimate the effect of learning-rate selection in each training step of Figure 1. The following process is performed at each training step:

- Uniformly sample *n* initial values of the learning rate (typically 10 or 20) from a pre-defined range, such as [0.01; 0.6] or [0.0001; 0.6] depending on the scenario – models that were already trained before typically benefit from a lower starting learning rate, while models with randomly initialized parameters might require more aggressive learning rates.
- The model is trained with each learning rate on a reduced training dataset and for a limited number of epochs.
- The learning rate that leads to the highest performance/lowest loss value is used to train the model on the full dataset and for the full number of epochs.

The reduction in the dataset size and the number of epochs significantly decreases the computational complexity which allows us to search through a wider range of learning rate values.

The proposed approach is applied sequentially to each step of the curricular SSL training scheme described in Section 2 and shown on Figure 1: T_{S1}, T_{S1},…, T_{Sk} and T_T. Specifically, we train each stage with each of the 20 learning rates sampled from the range [0.0001; 0.6] on 8% of ISIC-2019 training dataset. This approach resulted in the increase in the accuracy on the target task T_{St} for all tested pretraining setups with respect to the default learning rate values defined for ImageNet, including ODC (+10.48%), SwAV (+1%), Relative Location (+1%), Relative Location -> ODC (+3.6%).

3.2. Effect of dataset size on the performance of the pretext task curriculum

In the work reported in [10], we have performed exploratory experiments on: (1) measuring the effect of the training dataset size (number of instances) in the curricular SSL training scheme described in Section 2, and (2) the effect of the relative training dataset size of the SSL pretraining steps with respect to the size of the dataset utilized to fine-tune the model to the target downstream task, i.e., whether using larger unlabeled dataset sizes for the SSL pretraining steps increases the performance in the downstream task. These exploratory experiments are performed using the multi-class ISIC-19 dataset [13].

Table 1 presents a summary of the balanced accuracy results obtained with different dataset sizes for the: i) SSL pretraining steps, and ii) downstream task classifier. Results are obtained for a curricular SSL of: Relative Loc → SwAV → classifier. The results in the diagonal of the table show that the size of the training dataset is relevant for the performance (from an accuracy of 42.19% with a 12.5% size to an accuracy of 71.00% with the whole dataset). However, the size of the dataset in the SSL steps does not influence the results that much (42.9% accuracy using 12.5% of the dataset for SSL pretraining and downstream classifier, to 43.65% accuracy using 100% of the dataset for SSL pretraining and 12.5% for the downstream classifier). More details of these experiments are given in [10].

Table 1 Balanced accuracy results obtained with different dataset sizes for the: i) SSL pretraining steps, and ii) downstream task classifier. Dataset sizes are given as a percentage of the original dataset size (100%). Dataset sizes for the SSL steps are always larger than the downstream classifier.

	SSL pretraining steps				
Downstream classifier		100%	50%	25%	12.5%
	100 %	71.00	-	-	-
	50 %	58.41	59.11	-	-
	25 %	51.45	51.67	49.32	-
	12.5 %	43.65	42.88	43.01	42.19

4. Self-paced multi-task self-supervision

The empirical definition of the curriculum of pretext tasks, described in Section 2, has required a high amount of the design and implementation effort in the project, so this task has been covered with exploratory experiments.

These exploratory experiments have been oriented to analyze:

1) The correlation between downstream task performance obtained with single-SSL pretraining, and the definition of an optimum SSL curriculum.
2) The similarity of features learned by different SSL tasks, to measure their level of complementariness.
3) The possibility of defining a metric that measures the transferability of models trained with different SSL pretext tasks from one image domain to another one.

4.1. Single-SSL pretraining downstream task performance for the definition of a task curriculum order

The fundamental hypothesis we have evaluated in these experiments is based in the following conditions:

- SSL tasks can be ordered by downstream accuracy obtained with single-SSL task pretraining, i.e., training a model in the scheme of Figure 1 with two training steps: a single SSL-task (TS1) and downstream classification (TT), and measuring the model accuracy at the downstream task. We will refer to this as the single-SSL accuracy.

- We want to evaluate whether a curriculum of SSL pretraining tasks, ordered by increasing single-SSL accuracy (defined as curriculum ordering in Section 2) provides a better downstream accuracy than any other ordering.

This hypothesis has been tested for to different medical image domains: dermatoscopic skin lesions in [7], using the ISIC-19 dataset [13], and chest X-Rays in [11], using the SIIM-FISABIO-RSNA COVID-19 Detection dataset [14].

The results in the ISIC-19 dataset show that the optimum combination of pretext tasks is given by two SSL tasks ordered in a curriculum order (ODC → MoCov2 → downstream classification). However, the performance of this scheme outperforms a combination of three pretext tasks in curriculum order (ODC → Rel Loc→ MoCov2 → downstream classification). Accuracy results are given in Table 2. More details are given on [7].

Table 2 Balanced accuracy for SSL curricular training with pretext tasks orered in a curriculum order for the multiclass ISIC-19 dataset.

	Balanced Acc.
ODC → MoCov2 → downstream classification	75.44 %
ODC →Rel Loc→ MoCov2 → downstream classification	73.36 %

The results in the SIIM-FISABIO-RSNA COVID-19 dataset show that the optimum combination of all possible combinations of three pretext tasks is given by a curriculum ordering (MoCov2 → SwAV → Rotation → downstream classification), and this combination outperforms any other option of single-SSL or combination of two SSL tasks (see Table 3). For two SSL-tasks, the second-best option (Relative Location + SwAV: 85.27%) also follows a curriculum order, indicating that arranging the SSL tasks in a curriculum order yields a performance competitive with top-performance combinations. More details are given in [11].

Table 3 Balanced accuracies and AIL scores for the curricular SSL-task pretraining configurations. Sequential orderings for SSL-tasks read left to right. The curriculum column indicates whether a SSL-task combination follows a curriculum ordering. Results in bold refer to the highest score of each block, while results in blue are the highest scores overall.

Curriculum	Pretraining	Validation Acc (%)	AIL (%)
-	**ImageNet**	82.75	**38.16**
-	**Scratch**	**83.69**	37.30
-	**Rel-Loc**	83.62	36.33
-	**MoCo v2**	83.89	37.40
-	**Swav**	83.97	42.82
-	**Rotation**	**84.72**	**45.95**
✓	**MoCo v2 + Rotation**	84.77	**47.92**
	MoCo v2 + Rel-Loc	**85.59**	39.57
✓	**MoCo v2 + SwAV**	83.67	41.79
	Rotation + MoCo v2	76.08	31.87
	Rotation + Rel-Loc	84.33	**44.48**
	Rotation + SwAV	**84.81**	41.46
✓	**Rel-Loc + Rotation**	84.54	**48.63**
✓	**Rel-Loc + MoCo v2**	82.79	39.41
✓	**Rel-Loc + SwAV**	**85.27**	46.26
✓	**SwAV + Rotation**	83.89	**47.16**
	SwAV + Rel-Loc	**84.92**	43.05
	SwAV + MoCo v2	82.37	36.51
	MoCo v2 + Rotation + Rel-Loc	**85.28**	**38.82**
	MoCo v2 + Rotation + SwAV	84.80	30.69
	MoCo v2 + Rel-Loc + Rotation	84.19	38.51
	MoCo v2 + Rel-Loc + SwAV	**85.49**	**46.30**
✓	**MoCo v2 + SwAV + Rotation**	**85.67**	40.19
	MoCo v2 + SwAV + Rel-Loc	83.74	**40.89**

Both results in the ISIC-19 and SIIM-FISABIO-RSNA COVID-19 datasets show evidence of a degree of correlation between single-SSL accuracy and the optimum combination of curricular SSL pretraining tasks. This indicates that single-SSL accuracy can be used as a fundamental of the automatic definition of SSL curriculums. However, this correlation is not perfect, and evidence suggest that single-SSL accuracy is not enough, and should be complemented by other metrics that capture other aspects beyond downstream performance.

4.2. Similarity of features learned by different SSL tasks

Given the conclusions in Section 4.1, we have also explored the use of Central Kernel Alignment (CKA) [12] to measure the similarity of representations, at

different network layers, of features learned by pairs of individual pretext tasks. Such metric on the similarity of features learned by different tasks,

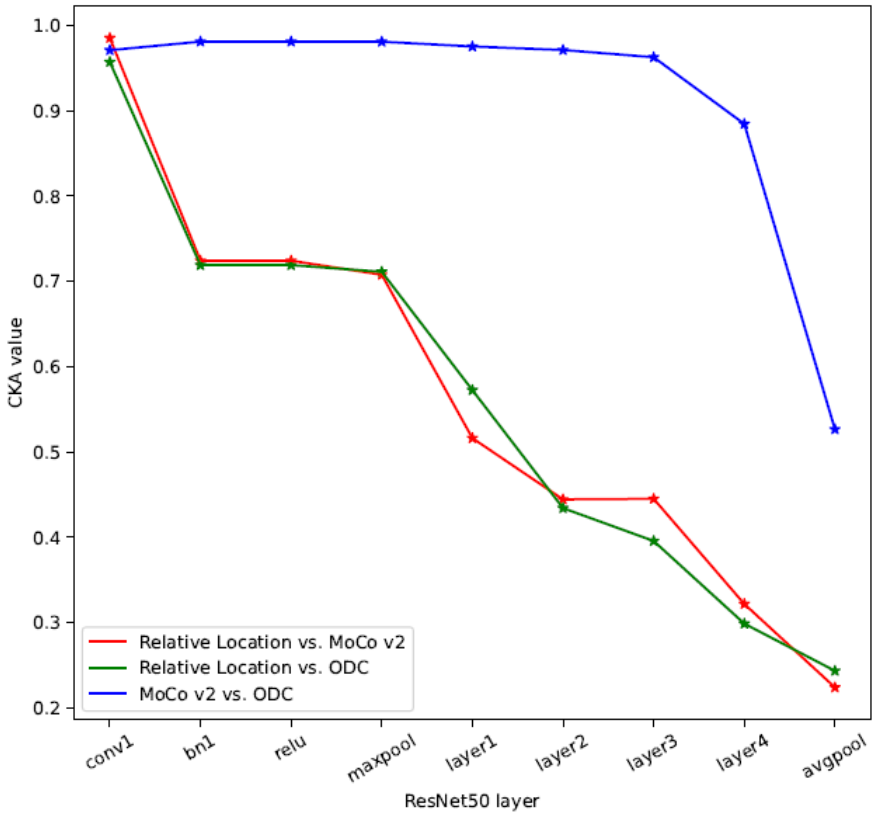

Figure 2 Similarities of feature maps learned by pretext tasks per layer. The similarity of feature maps is measured with Central Kernel Alignment (CKA): values of CKA closer to 1 indicate strong similarity of feature maps, while values closer to 0 indicate that they are dissimilar.

Results for the ISIC-19 dataset are shown in Figure 2. In these results, we compare the CKA between pairs of pretext tasks (Relative Location, ODC and MoCo v2). ODC and MoCo v2 produce very similar representations (the CKA values are close to one for all layers except the final average pooling layer), while the representations of Relative Location are very dissimilar from those of ODC and MoCo v2. Our hypothesis is that, given that the training directions of ODC and MoCo v2 features are aligned, i.e, they both follow similar training paths, intermediate Relative Location features in the full curriculum (ODC →Rel Loc→ MoCov2 → downstream classification, see Table 2) somehow shift the feature space, decreasing the performance with respect to ODC → MoCov2 → downstream classification. These results also show of the multi-path training idea, i.e., the fact that two models reaching similar downstream accuracy (ODC and Relative Location) are relying on dissimilar representations.

This discussion sheds light on the limitations of using single-SSL accuracy to define the full curriculum ordering (where all pretext tasks are used). While is

better than any anti-curriculum and mixed curriculum orderings using all three tasks, it is not always the best option. In this regard, as a direction for the future work, we suggest incorporating the CKA analysis (alongside with individual pretext tasks accuracy) to consider the relationships between the learned representations of the individual tasks when establishing the curriculum ordering.

4.3. Evaluation of the transferability of SSL-trained models to a different image domain:

As another experiment to evaluate the possibility of defining a pretext task curriculum automatically, we have performed a series of exploratory experiments on the transferability of features learned by SSL pretext tasks to a downstream task, or datasets of the same or different image modality of the one used for training.

To this aim, we start by a set of models trained used different pretext tasks, but all learned using the ImageNet dataset and the same architecture (ResNet50). The used models are available at this link. Then, we forward-pass images from different datasets [16][17][18][19][20][21][22][23] on these models and compare the empirical distributions of the features of a given dataset with those obtained by feeding the model using ImageNet [24] images. The comparison is performed in terms of the Fréchet Inception Distance (FID) [15]. As a reference, we randomly divided ImageNet into two disjoint sets and obtain the FID between these two sets using the same process. We then normalized FID between datasets by this reference to yield normalized FIDs.

Results of this process can be observed arranged on a per dataset and a per tasks basis in Figure 3 and Figure 4 respectively. Initial results suggest that: (1) Features learned using BYOL [25] yield very close results to supervised learning when transferred to the target task (image classification on ImageNet) and (2) The FID distances between ImageNet and the other dataset are higher as the datasets are more different both in terms of content and modality, shaping roughly three sets, (3) the performance of the features transferred to the target task is proportional (to some extent) to the FID distance of the learned features of the pretext task with respect to those of the target task. We are currently validating these observations by analysing additional target tasks.

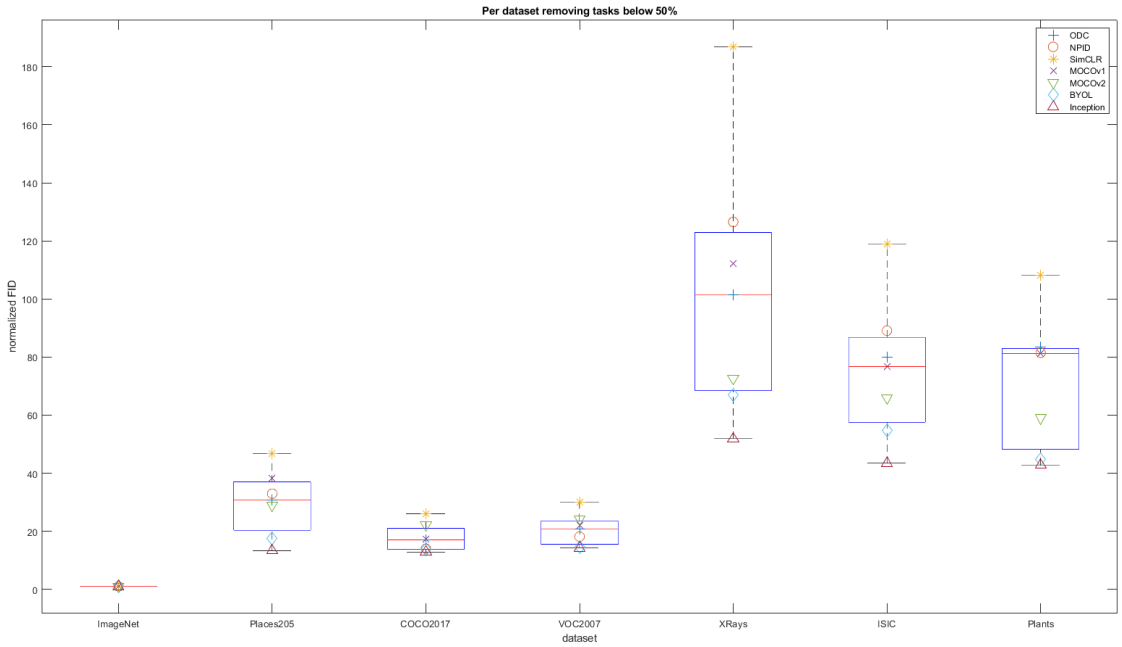

Figure 3 FIDs between features extracted for pairs of datasets (ImageNet vs X) by forwarding images of these datasets to models learned by different pretext-tasks. FIDs are normalized by dividing them by the ImageNet vs ImageNet pair. Results for Rotation task are removed for the shake of visualization.

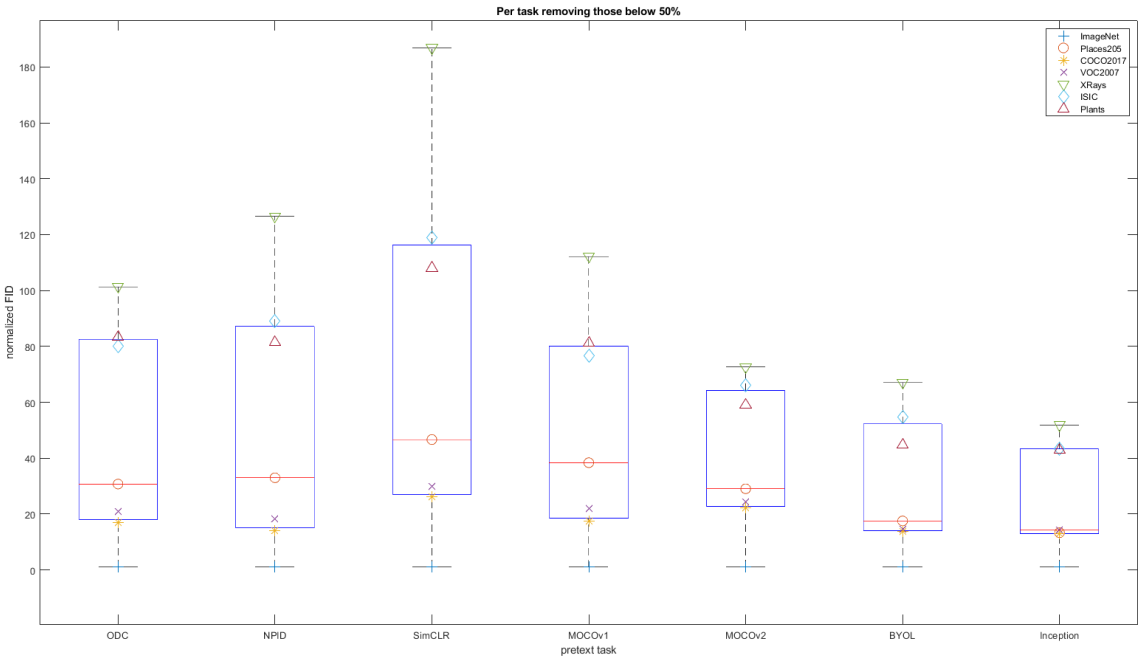

Figure 4 FIDs between features extracted for pairs of datasets (ImageNet vs X) by forwarding images of these datasets to models learned by different pretext-tasks. FIDs are normalized by dividing them by the ImageNet vs ImageNet pair. Results for Rotation task are removed for the shake of visualization. Pretext tasks are arranged according to their transferred performance (higher accuracy to the right). The last task, Inception, stands for the comparison between the features of a supervised Inception model trained on ImageNet.

References

[1] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in Advances in Neural Information Processing Systems, vol. 27, 2014.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012

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

[4] F. Navarro, C. Watanabe, S. Shit, A. Sekuboyina, J. C. Peeken, S. E. Combs, and B. H. Menze, “Evaluating the robustness of self-supervised learning in medical imaging,” arXiv preprint arXiv:2105.06986, 2021.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.

[6] MMSelfSup: https://github.com/open-mmlab/mmselfsup

[7] Kirill Sirotkin, Marcos Escudero-Viñolo, Pablo Carballeira, Juan Carlos San Miguel, "Improved skin lesion recognition by a Self-Supervised Curricular Deep Learning approach", submitted to IEEE Journal of Biomedical and Health Informatics, (under review). Preprint available at: https://arxiv.org/abs/2112.12086

[8] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research, 18(1):6765–6816, 2017. 2

[9] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. Proceedings of Machine Learning and Systems, 2:230–246, 2020.

[10] Análisis de la influencia del volumen de datos en el rendimiento de técnicas de aprendizaje autosupervisado para clasificación de imagen dermatoscópica, Álvaro Rojo Torío, (advisor: Pablo Carballeira López), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2022.

[11] Self-Supervised Curricular Learning For Chest X-Ray Image Classification, Iván de Andrés Tamé, (advisor: Pablo Carballeira López), Trabajo Fin de Máster (Master Thesis), Máster Universitario en Deep Learning for Audio and Video Signal Processing, Univ. Autónoma de Madrid, Jul. 2022.

[12] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, “Similarity of neural network representations revisited,” in International Conference on Machine Learning. PMLR, 2019, pp. 3519–3529.

[13] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin lesion classification using ensembles of multi-resolution efficientnets with meta data,” MethodsX, vol. 7, p. 100864, 2020.

[14] P. Lakhani, et al., The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. 2021.

[15] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Proc. NIPS (2017)

[16] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[17] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. Ieee, 2009.

[18] Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew

Zisserman. "The PASCAL visual object classes challenge 2007 (VOC2007) results." (2007).
[19] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.
[20] Zhou, Bolei, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. "Places: A 10 million image database for scene recognition." IEEE transactions on pattern analysis and machine intelligence 40, no. 6 (2017): 1452-1464.
[21] Sharma, Saroj Raj, Dataset of diseased plant leaf images and corresponding labels. GitHub - spMohanty/PlantVillage-Dataset: Dataset of diseased plant leaf images and corresponding labels. 2018.
[22] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J. & Soyer, P. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data 8, 34 (2021). https://doi.org/10.1038/s41597-021-00815-z
[23] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097-2106. 2017.
[24] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. IEEE, 2009.
[25] Grill, Jean-Bastien, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch et al. "Bootstrap your own latent: A new approach to self-supervised learning." arXiv preprint arXiv:2006.07733 (2020).