An Experimental Comparison of SSL, ImageNet Weight, and Random Initialization as Prior Weights for Deep Learning Models in Medical Image Analysis

Tauqeer Akhtar
M.Tech CSE 2021CSM1021
Dept. of Computer Science and Engineering
IIT Ropar

Dr. Deepti R. Bathula
Assistant Professor
Dept. of Computer Science and Engineering
IIT Ropar

***Abstract*—The use of pre-trained weights from ImageNet during training on a dissimilar dataset has been shown to result in a better solution convergence than using random weights as a prior. However, medical image datasets have unique characteristics that differ significantly from those in ImageNet, thus making pre-training with self-supervised learning a preferable alternative. By leveraging self-supervised training, the model acquires domain-specific knowledge that can enhance its ability to perform downstream tasks such as classification, segmentation, or detection. Additionally, the model's performance can be further improved using deep mutual learning [8] or knowledge distillation [6] techniques. We derive the conclusion that Self Supervised Learning prevents the Model from overfitting on the training data. It helps training with prior weight initialization, but it still is not outperforming ImageNet in terms of accuracy.**

***Index Terms*—Representation Learning, Self Supervised Learning, SimCLR, Pre-training, Medical Image Datasets.**

I. Introduction

Medical Image Datasets are very different than the datasets for which pre-trained weights are present, like Cifar10 or ImageNet. We know that starting from better prior weights helps in faster and more optimal convergence of the Model. Medical Images does not correlate to ImageNet or any available pre-trained weights, and one Medical dataset is entirely different from other Medical datasets. So with SSL the Model can be pre-trained on the same specific dataset on which it has to perform the downstream task. It will help the Model understand the dataset's representation and features before it gets the labels. The Model will have good pre-trained weights, which will help to converge faster and to an optimal solution. After experimenting with different Model configurations and different Medical Image datasets, it is observed that ImageNet weights perform better than Self-Supervised pre-training and Random Initialization performs worst with all the datasets. Self Supervised Learning prevents the Model from overfitting on the training data, whereas ImageNet weights tend to overfit, and validation loss starts to increase after some time. SimCLR was used for Self-Supervised training.

- SimCLR is used to do Self-Supervised Representation Learning, and then the Model is Finetuned for the classification task.
- The results are compared with a random initialized Model and a Model with prior as ImageNet weights.
- We found that the Model with ImageNet weight performs better but is prone to overfitting, whereas SSL tends to prevent overfitting.
- We are planning to use SWaV and then compare the result with the SimCLR as well as other Model configurations.
- We will be using Supervised Contrastive Learning for the Representation Learning for the pre-training in place of Self-Supervised Learning.
- We will use Deep Mutual Learning to improve the performance of the Model further, where two Models can learn from each other.

II. Literature Survey

Self Supervised Learning [3] is a very recent technique in the field of Computer Vision to learn the representation of the data. The pre-training helps achieve very good results with very few labeled data. Some of the applications of Self-Supervised Learning are Representation Learning to help downstream tasks, Colorization of a grayscale image, Context Filling or predicting missing parts of the image, prediction of rotation of an image, etc.

There are several self-supervised learning techniques that have been proposed in recent literature. One such approach, described in [1], is a contrastive learning technique that considers two augmented versions of the same image as positive samples, while all other images in the batch are considered negative samples. The encoder and decoder have identical weights, and negative samples help prevent the model from collapsing to a trivial solution.

Another approach, described in [7], involves using augmented images along with the same class images as positive samples and all other images as negative samples. The model

learns to represent similar and dissimilar images using contrastive loss. Cosine similarity is used, and only the classification layer is trained, with the representation layer frozen.

In [2], the authors demonstrate that even without negative sample pairs, large batches, or momentum encoders, simple Siamese networks can learn meaningful representations if weight copying and gradient flow stopping are used for one of the encoders. Cosine similarity is again used.

In contrast to the use of negative samples in contrastive self-supervised learning, [4] proposes a technique that eliminates their use. The key idea is to use moving average, copying the weight from one encoder to another after some delay. Once the model learns representations, it can be used for fine-tuning.

Finally, there is a momentum-based self-supervised technique described in [5], in which the momentum encoder is updated using delayed update. Keys are sampled from data, and many mini-batches of keys are used, with keys stored in a queue. The new mini-batch replaces the oldest mini-batch keys. The contrastive loss function with cosine similarity is used.

III. Problem Statement

We know that using pre-trained "ImageNet" weights as a "prior weight" to train a ResNet helps achieve better results and faster convergence than a random initialization with most datasets. Even though "ImageNet" weights are trained on an entirely different type of dataset, it helps in training and provides a better starting weight.

Medical Images are very different than the Images of Imagenet. Medical Images can be XRay Images, Histopathological Images, Skin Patches Images, Tumor Images, etc., which do not correlate with ImageNet as ImageNet has Images of Natural objects. So if we have a method to provide a better starting weight to a specific Medical Image dataset, then we can increase the Model's accuracy for that particular dataset.

A. Our Hypothesis

Our Hypothesis is that we could use Self Supervised Learning before training the Model with the labels. Since the Model could learn the representation and features of the dataset, it will have a good starting point to perform downstream tasks like classification, segmentation, or object detection. Once the Model performs better with the dataset, we could use Deep Mutual Learning to improve the Model further.

B. Why Self Supervised Learning(SSL) [3]?

Medical datasets are very costly to get labeled by experts. Since Self Supervised Learning helps to learn the features and representation of the dataset and doesn't require a label, it is the best way to get the prior weights for any specific dataset. Once the Model learns the dataset's features, it usually helps with the downstream tasks. The self-Supervised Model requires very few labels to perform, similar to Supervised Learning. Hence with few labeled data and a large amount of unlabelled data, we can still get a better-performing Model.

There are many Self Supervised techniques like Simple Siamese Representation Learning(SiamSiam), Momentum Contrast for Unsupervised Visual Representation Learning(MoCo) [5], A Simple Framework for Contrastive Learning of Visual Representations(SimCLR), Unsupervised Learning of Visual Features by Contrasting Cluster Assignments(SwAV), etc. All these techniques are equivalent to some extent and help to learn the representation of the dataset so that we can perform the downstream task with very few labels. The SSL technique used is SimCLR in our experiments.

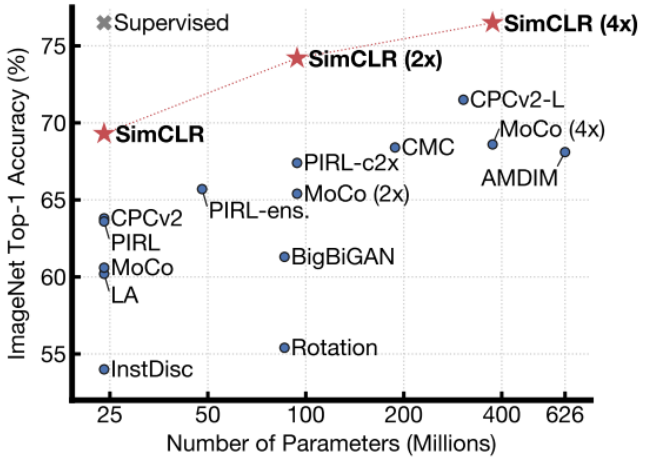

Fig. 1: Accuracy of different SSL technique

1) Why SimCLR [1]?:

A Simple Framework for Contrastive Learning of Visual Representations(SimCLR) is one of the best-performing visual representation learning techniques in Computer Vision tested on the ImageNet dataset. The following figure shows the accuracy obtained by different SSL methods, and we can see that SimCLR outperforms most of the methods. Unlike SimSiam [2], SimCLR never converges to a trivial solution where a model gives the same output no matter the input due to negative samples present in the batch. It is a simple and effective representation learning technique. It is time tested technique.

2) Working of SimCLR: SimCLR consist of four important modules:

1) Data Augmentation module
2) Base Encoder
3) A Projection Head
4) A Contrastive Loss Function

1) Data Augmentation module
   This module makes two augmented versions of the same image using a different set of augmentation. Several augmentations are vertical flip, horizontal flip, center crop, color jitter, gaussian blur, and grayscale. Some

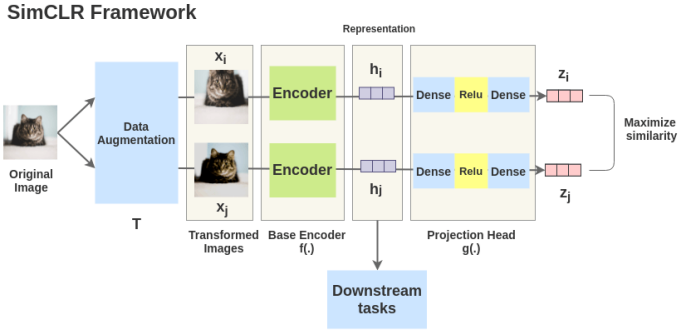

Fig. 2: Architecture of SimCLR

of these augmentations are selected at random with different probabilities and applied to the image to create two versions of the same image. For Medical images, Vertical and Horizontal flips are sometimes dropped. Like in Chest X-ray Images, Vertical or Horizontal flip does not make sense as there is no possibility that we will get a vertically flipped Image in the dataset.

2) Base Encoder
A Base Encoder is the backbone of the SimCLR. It is a neural network that learns the representation. In our experiment, the base encoder is taken as ResNet50. More the depth of the base encoder better the performance of the SimCLR, but it requires a lot more batch size and training. So the optimal base encoder in our case is ResNet50 based on the resources.

3) A projection Layer
It is a Multi-Layer Perceptron(MLP). The output of the ResNet50 is passed through the MLP, and then the loss is calculated and propagated backward. The perceptron layer outputs a vector of size 128.

4) A Contrastive Loss function
In the case of SimCLR, the loss function used is Normalized Temperature-scaled Cross Entropy Loss, also known as NT-Xent Loss. The loss function minimizes the distance between two augmented versions of the same image and maximizes between two different images in the batch. So if the batch size is "N," there are N-1 negative samples. It requires a lot of negative samples, so a large batch size helps simCLR to better learn the features. Following is the expression for the loss function.

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \infty_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

Where i and j are augmented versions of the same image and k are negative samples.

SWaV is another self-supervised technique that uses contrastive learning and clustering. It does this by swapping cluster assignments for positive samples. It is new technique.

We are planning to use SWaV in the future, but simCLR is also a great technique; if the batch size is large and an ample amount of training is done, there is not much difference in performance.

The Model used is ResNet50.

IV. Datasets and Model

The following 5 Medical datasets were used:

a) HAM10000 link
b) Skin Cancer ISIC link
c) Covid19 link
d) Chest X-ray link
e) Indian Diabetic Retinopathy Image Dataset(IDRiD) link

Out of five datasets, HAM10000, Skin Cancer ISIC and Chest X-ray are big, whereas Covid-19 and IDRiD are small. We used both types of datasets in our experiment to check if there would be any marginal difference in results.

1) Why ResNet50?:
ResNet50 has been a standard model in Computer Vision related research. Most researchers use ResNet50 to compare the results from previous work in the same field. It is neither a big model nor a small model, so it fits the sweet spot.

Following are the five trained models and their configurations.

Models	Base Model	SSL	Pretrained-Weights
M1	ResNet50	No	Random
M2	ResNet50	Yes	Random
M3	ResNet50	No	ImageNet
M4	ResNet50	Yes	ImageNet

M1 is directly trained on the dataset without Self Supervised Learning and the weights of ResNet50 in initialized randomly.

M2 is trained on the dataset after Self Supervised Learning on the same dataset. The weights are randomly initialized before Self Supervised training.

M3 is directly trained on the dataset without Self Supervised Learning, but the pre-trained weights taken are ImaneNet weights.

M4 is trained on the dataset after Self Supervised Learning, and the weights of ImageNet is assigned before performing Self Supervised Learning

A. Results and Discussion

We found the following results with all five Medical Image Datasets and four configurations of the Model.

a) HAM10000

Model	Train Accuracy	Train Loss	Test Accuracy	Test Loss
M1	74.75	0.70	62.61	.97
M2	87.04	0.35	70.38	**0.75**
M3	**98.28**	**0.056**	**76.59**	1.14
M4	90.49	0.26	69.6	0.76

b) Skin Cancer ISIC

Model	Train Accuracy	Train Loss	Test Accuracy	Test Loss
M1	89.05	.28	50.03	2.07
M2	81.41	0.50	59.48	**1.03**
M3	**99.2**	**0.024**	**65.52**	1.57
M4	93.82	0.18	62.95	1.03

c) Covid19

Model	Train Accuracy	Train Loss	Test Accuracy	Test Loss
M1	97.3	.068	83.52	0.61
M2	98.09	0.068	86.46	0.24
M3	**98.85**	**0.032**	**95.06**	**0.033**
M4	98.56	.046	83.52	.61

d) Chest X-ray

Model	Train Accuracy	Train Loss	Test Accuracy	Test Loss
M1	98.7	0.03	67.18	1.06
M2	99.59	**0.01**	81.76	0.27
M3	**99.63**	**0.01**	94.79	**0.03**
M4	99.58	**0.01**	**95.59**	0.08

e) Indian Diabetic Retinopathy Image Dataset(IDRiD)

Model	Train Accuracy	Train Loss	Test Accuracy	Test Loss
M1	57.81	0.98	34.09	2.223
M2	53.17	1.173	37.03	**1.52**
M3	**63.27**	**0.88**	**38.54**	1.90
M4	53.54	1.11	35.09	1.67

We observe that the performance can be arranged in descending order as M3, M4, M2, and M1. Self Supervised Learning is helping the training as it is outperforming random initialization in all the cases, but it is not outperforming the ImageNet weights initialization in most of the cases except in the Chest-Xray dataset. We can also see that Self Supervised Learning benefits more with larger datasets than with smaller datasets.

Following are the train and test graphs for ISIC and HAM10000

a) ISIC

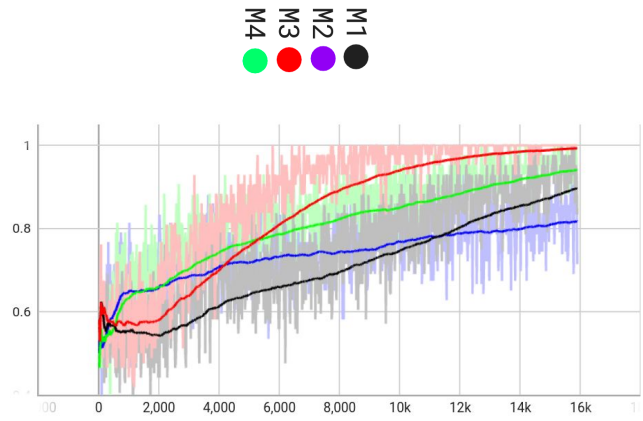

Fig. 3: Train Accuracy

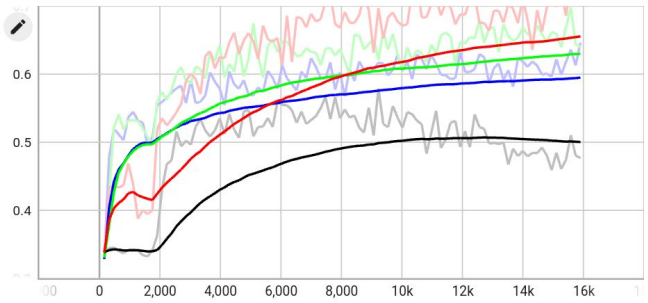

Fig. 4: Test Accuracy

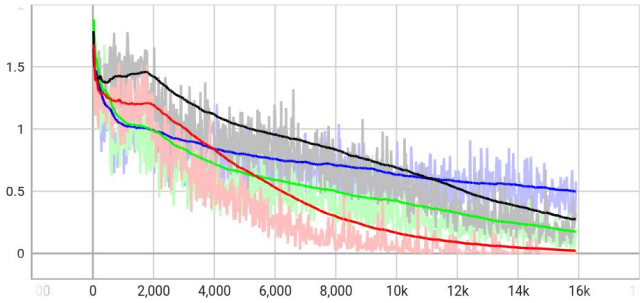

Fig. 5: Train Loss

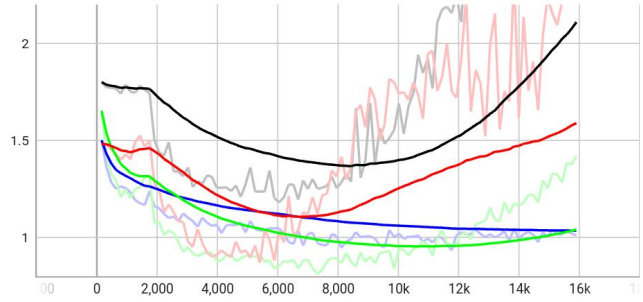

Fig. 6: Test Loss

b) HAM10000

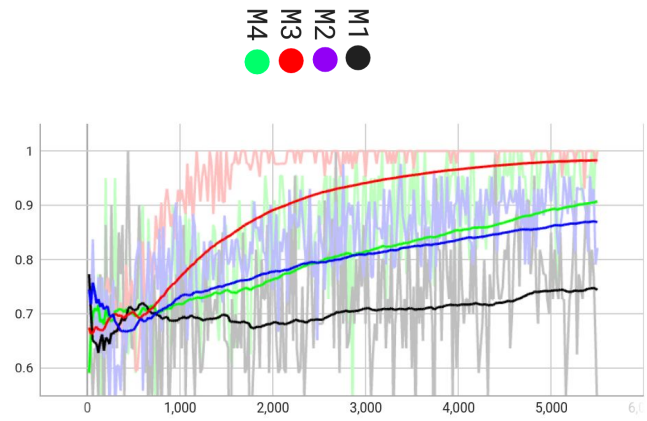

Fig. 7: Train Accuracy

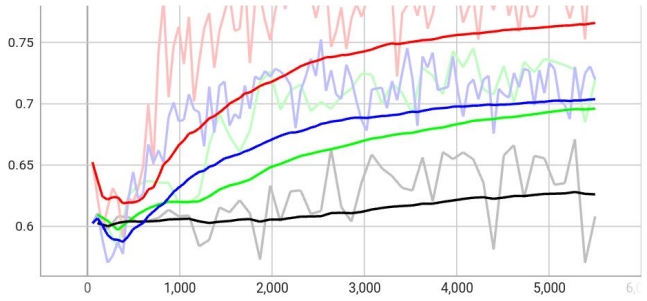

Fig. 8: Test Accuracy

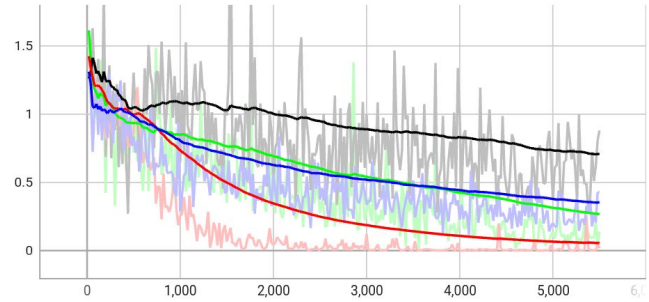

Fig. 9: Train Loss

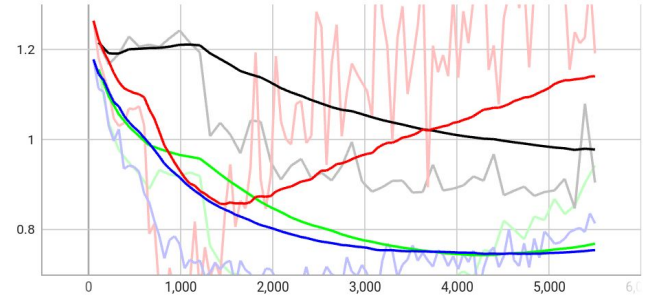

Fig. 10: Test Loss

We observe from the above graph that the validation loss is increasing in random initialization and ImageNet initialization where Self Supervised Learning is not applied. Even though we have high validation accuracy, Model confidence is low as test loss is high and increases with more training, which means the Model will not perform better in real-world scenarios.

V. Conclusion and Future Plans

We derive the conclusion that Self Supervised Learning prevents the Model from overfitting on the training data. It helps training with prior weight initialization, but it still is not outperforming ImageNet in terms of accuracy. We also find that performing Self Supervised Learning with prior as ImageNet certainly performs better when SSL is performed with random initialization.

Since we are unable to outperform ImageNet with Self Supervised Learning, We will try to implement "Supervised Contrastive Learning", which uses the label while learning the representation. Hence we expect it to provide a better result than ImageNet. Our next step will be to use Mutual Learning to improve the Model's performance further. We will perform more experiments with Model M4(SSL with ImageNet as prior).

We will try to use SWaV instead of SimCLR and see if it helps.

REFERENCES

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[3] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances and challenges. *CoRR*, abs/2110.09327, 2021.

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020.

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[7] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020.

[8] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.