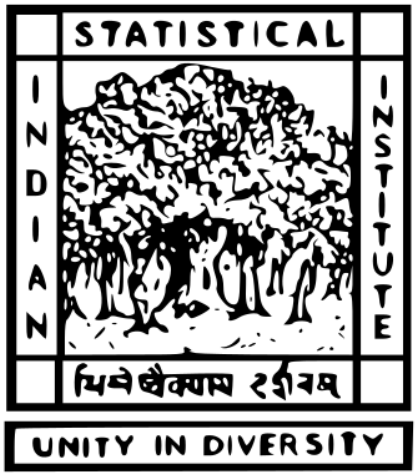
KNEE MRI DIAGNOSIS USING SELF-SUPERVISED LEARNING

By SAYAN RAY

WORD COUNT 7803 TIME SUBMITTED 21-JUL-2022 01:28PM

PAPER ID 88651576



Indian Statistical Institute
Department of Computer Science

KNEE MRI DIAGNOSIS USING SELF-SUPERVISED LEARNING

Prepared By:
Sayan Ray
Roll: CS2038

Supervised By:
Dr. Umapada Pal
Professor
Computer Vision and Pattern Recognition Unit

In the partial fulfillment of the requirements for the degree of MASTER OF TECHNOLOGY In COMPUTER SCIENCE

Kolkata
July, 2022

Abstract

Significant advancements in deep learning have been made in image understanding tasks including object detection, image classification and segmentation. But the effectiveness of image recognition mostly depends on supervised learning, which necessitates a large number of labels that have been manually annotated. But in general , the amount of unlabelled data is present much more than the labelled ones. Also, the pre-trained models from the natural images are not useful on medical images since the intensity distribution is different. Although labelling natural images are easy as because just simple human knowledge is enough. However, the annotation for medical images requires expert knowledge. So, how should we learn representations without labels? We need to get supervision from the data or image itself. We may do this by structuring a supervised learning assignment in a certain way to predict just a portion of information while utilising the remaining data. This is known as self-supervised learning (SSL). SSL algorithms like Jigsaw Puzzle Solving, Rotation Prediction, SimCLR, BYOL, Barlow Twins, MoCo etc. have been proposed over the recent years and have achieved state-of-the-art performance in image recognition tasks. Different variants of these algorithms have also been used for applications in medical image analysis, image segmentation tasks, etc. In this project, we have used BYOLMed2D, a variant of BYOL, for self-supervised learning of representations from Knee MR data. The dataset on which the experiments were performed is MRNet, which is a knee MRI dataset. From the experimental evidence, it is evident that the proposed method succeeds in achieving performance at par with the existing state-of-the-art SSL methods for learning representation on the MRNet dataset.

Keywords: BYOLMed2D, MRNet, CNN, Self-supervised learning

Table of Contents

Abstract **I**

Table of Contents **II**

List of Tables **IV**

List of Figures **V**

1 Introduction **1**

1.1 Introduction 1
1.2 Self-supervised learning: 1
1.3 Motivation 2
1.4 Thesis Outline 3

2 Literature Survey **4**

2.1 Contrastive Learning 4
2.2 SimCLR 5
2.3 MoCo 6
2.3.1 Momentum update 7
2.4 Barlow Twins 7
2.5 BYOL 8
2.5.1 BYOL loss function 9
2.6 Relative Performances 10
2.7 Conclusion 10

3 Robustness Test of SSL on Imbalanced Datasets **11**

3.1 Relative performance gap 13
3.2 CIFAR10-LT 13
3.3 Experiment on CIFAR10-LT 13

3.4 Analysis 14
3.5 Conclusion 14

4 Methodology : BYOLMed2D **15**
4.1 Architecture 15
4.1.1 Residual Networks 15
4.2 Pre-training 16
4.2.1 Maxpooling over frames 17
4.3 Downstream 18
4.3.1 Loss function for downstream task 18
4.4 Conclusion 18

5 Experiments and Results **19**
5.1 Dataset 19
5.2 Augmentations 19
5.3 Result 21
5.3.1 Comparison with other results 21
5.4 Conclusion 21

6 Conclusion **22**

Bibliography **23**

List of Tables

2.1 Relative Performances of SSL algos 10

5.1 Agmentations with probabilities 20

5.2 Comparison to other methods(The result of SimCLR using JIGSAW PUZZLE was collected from the SSLM paper) 21

List of Figures

2.1 Framework for contrastive learning of visual representations 5
2.2 Diagrammatic representation of MoCo Framework 6
2.3 Barlow Twins Framework 8
2.4 BYOL Architecture 9

3.1 Relative performance gap between self-supervised (MoCo v2) representations on long-tailed ImageNet with varying numbers of instances (blue for SSL) and supervised ones (red for SL) n, spanning ID assessment 12
3.2 Relative performance gap between self-supervised (MoCo v2) representations on long-tailed ImageNet with varying numbers of instances (blue for SSL) and supervised ones (red for SL) n, spanning OOD assessment 12
3.3 Confusion Matrix 13

4.3 Dataflow of BYOLMed2D framework 17
4.4 Example of maxpooling over frames 17
4.5 Sigmoid 18

5.1 Image augmentations 20

Chapter 1

Introduction

1.1 Introduction

Medical image analysis is the study of images produced during clinical practice to resolve clinical issues [1]. Learning good representations of image is very much important in computer vision as it helps in efficient training on downstream tasks like classification, segmentation, etc. To learn these representations, a variety of training methods have been developed, most of which rely on visual pretext tasks. Modern contrastive techniques are trained by increasing the distance between representations of augmentations from distinct images, often known as negative pairs, and decreasing the distance between representations of different augmentations of the same image,known as positive pairs. These techniques rely on huge batch sizes and need careful handling of negative pairs. Additionally, the selection of image augmentations has a significant impact on how well they perform.

1.2 Self-supervised learning:

We have focused on self-supervised representation learning, a subtype of unsupervised learning, in our effort. Without the requirement for human supervision, self-supervised learning may be used to extract meaningful feature representations from many forms of data. The two basic components of self-supervised learning are pretext and downstream. The model acquires representations by completing a variety of tasks known as Pretext tasks. Pretext tasks are often created based on the kind of data. The pretext task's goal is to draw from the data explainable and transferable representations that may be applied to the downstream tasks. However, there are very few applications of self-supervised learning techniques in the field of medical image analysis. It should be highlighted that the type of data used in medical image analysis differ greatly from those used in the study of natural images. Additionally, it might be difficult to find a big volume of annotated medical data for end-to-end training of deep neural network. Methods of transfer learning are typically used to address this issue. However, feature specialisation of the higher layers in medical image data may cause problems when utilising models that have been pre-trained on natural image datasets. Additionally, medical professionals are needed for such a delicate operation, making annotations in medical image or video collections difficult to get [2]. All things considered, this qualifies self-supervised learning as a useful technique for use

in medical image analysis.

One SSL algorithm that outperforms cutting-edge contrastive techniques is BYOL, which also avoids the usage of negative pairings. The outputs of a network are iteratively bootstrapped and used as targets for an improved representation. In addition, contrastive methods are less resilient to image augmentations than BYOL; the fact that it does not rely on negative pairs may be one of the main factors contributing to this improvement. In BYOL, direct bootstrapping of the representations has been proposed as an alternative to earlier techniques based on bootstrapping that employed pseudo-labels, cluster indices, or a small number of labels. BYOL employs two neural networks—the online and target networks—that communicate and share knowledge with one another. BYOL trains its online network to predict how further augmentation of the same picture will be represented by the target network starting from an augmented version of the original image. Although this goal allows for collapsed solutions, such as those that produce the same vector for all images, it has been demonstrated that BYOL does not converge to these types of solutions. The assumption of an optimum predictor, however, may be violated by unexpected changes in the target network, in which case BYOL's loss is not necessarily close to the conditional variance. The major objective of BYOL's moving-averaged target network, according to the theory, is to maintain the predictor's near-optimality over training.

1.3 Motivation

The primary goal of the science of medical image analysis is to process and analyse medical images from a variety of modalities in order to get relevant information that helps with accurate diagnosis determinations. Four major tasks derived from the core computer vision tasks and specifically designed for the medical field make up the heavy burden of the analysis of medical pictures. These four tasks are segmentation, registration, detection and localisation, and classification. The methodologies and algorithms used for each of the aforementioned activities help to comprehend and extract valuable information from the medical images. Either manually (engineered) or automatically (learned) from the data, features can be extracted. While statistical machine learning is primarily concerned with manual feature extraction, deep learning is far more interested in automatic feature extraction and is hence recommended.CNNs are a prominent option in the field of medical image analysis and have significantly advanced the many jobs requiring medical image analysis due to their ability to work with image data in its raw forms and performance that can be compared to human performance at faster rates. To capture the underlying distribution in the input data, it is well known that CNNs need to have a vast number of trainable parameters to be estimated, typically in the millions. As a result, improved estimate of these characteristics necessitates the use of a sizable quantity of data. Additionally, human annotation of the input data is required in order to execute supervised training using the gradient descent algorithm.However, there are significant barriers that prevent CNNs from progressing despite the impressive success they have had in the field of medical image processing. It is costly and time-consuming to create a large enough, high-quality human-annotated medical dataset because medical datasets must also be annotated by professional employees, as opposed to natural scene picture data, which may be done so by less experienced staff. Additionally, the annotation process is vulnerable to problems safeguarding patient privacy, especially when dealing with a particular condition. Thus, a significant barrier to machine learning applications in the medical industry

is the lack of labelled data, both in terms of annotation and volume.

In a pretraining-fine-tuning manner, self-supervised learning integrates both supervised and unsupervised learning systems. Self-supervised learning, as it is more precisely known, aims to learn semantically useful features for a specific task by producing supervisory signals from a pool of unlabeled data without the need for human annotation to be used for following tasks where the amount of annotated data is constrained. Self-supervised learning eliminates the necessity for manually labelled data from an unsupervised perspective. On the other hand, with the self-supervised learning technique, the supervised viewpoint is represented in model training with labels generated from the data itself. This is one of the major reasons why self-supervised learning has become a popular option in medical image analysis. Numerous studies have shown that the self-supervised learning method is beneficial for a variety of medical image processing tasks [3, 4].

It's crucial to have a balanced dataset while building a strong training set. When the dataset is severely unbalanced, the majority of available classification algorithms typically do not perform well on minority class cases. Without taking into account the relative distribution of each class, they strive to maximize total accuracy. Real-world data are frequently unbalanced, which is one of the major reasons why machine learning algorithms no longer generalise as well. The imbalance of class is not taken into consideration by conventional learning techniques.Both the dominant class and the minority class receive the same amount of attention. Using traditional learning techniques, it is challenging to create a competent classifier when the imbalance is severe. Missing minority class predictions comes at a larger cost than missing majority class predictions.

Due to the fact that one class is represented by a much higher number of instances than other classes, medical data sets frequently suffer from class imbalance issues. As a result, algorithms frequently become overloaded by huge classes and neglect minor classes. The amount of training datasets may be adjusted through sampling, with under and oversampling being two typical methods. It can be seen that self-supervised learning is more robust to dataset imbalance, hence our motivation for applying SSL to medical image analysis than the other approaches.

1.4 Thesis Outline

The remaining of the thesis is coordinated as follows:

- In Chapter 2, we discuss about the Self-supervised learning algorithms and their relative performances on a particular dataset.
- In Chapter 3, we discuss how the SSL algorithms outperform supervised algorithms when the dataset is imbalanced.
- In Chapter 4, we give the details of the architecture and methodology we have used.
- In Chapter 5, we compare the results obtained by us to other methods.
- In Chapter 6, we draw a conclusion to this project .

Chapter 2

Literature Survey

Self-supervised learning has becoming more common since it may save on the expense of annotating huge datasets. It has the ability to employ the learnt representations for several downstream tasks and use self-defined pseudolabels as supervision. To be more precise, in self-supervised learning for computer vision, natural language processing (NLP), and other fields, contrastive learning [5] has more recently taken the lead. In this chapter , we see how different SSL algorithms work .

2.1 Contrastive Learning

If we have a function f that is represented by any deep neural network, it will output the characteristics of that function as $f(x)$, given an input of x.

Contrastive Learning states that all positive pairs of x_1 and x_2 should have outputs that are similar to each other, and for a negative input x_3, f(x_1) and f(x_2) should both be different to f(x_3).
The positive pair may consist of two augmented views of the same image (for example, a vertically flipped version) or two different crops of the same image, while the negatives may consist of a crop from a different image, a frame from a different video, a different augmented view of a different image, etc. To enforce the similarity between positive pairs and dissimilarity between negative pairs during model training, a loss function is necessary . The set X of N patches, where X is the set of $N-1$ negative samples and 1 positive sample, is used to calculate the loss. The batch's available patches of both the same picture and several other images are randomly selected for the $N-1$ negatives.The loss is called InfoNCE loss [6] , defined by ,

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp\left(q \cdot k^+/\tau\right)}{\exp\left(q \cdot k^+/\tau\right) + \sum_{k^-} \exp\left(q \cdot k^-/\tau\right)} \tag{2.1}$$

Here, q denotes the network prediction, k+ is the correct positive patch, and k- denotes a collection of N-1 negative patches. You should take note that k+, k-, and q are all in representation space. It is comparable to the log-softmax function, as can be shown.With the use of this framework, we can demonstrate verifiable assurances about how well the learned representations perform on the typical classification task, which consists of a subset of the same set of latent classes [7, 8].

2.2 SimCLR

SimCLR [9] framework has following major components:

- Here, three straightforward augmentations are consecutively used: Random colour distortions, random cropping, then resizing to the original size, and random Gaussian blur. It has been seen that the combination of random crop and color distortion is beneficial to get good results.
- For extracting representation vectors from augmented images , a neural network base encoder f is used . Various choices of the network architecture without any constraints is allowed in this framework . For simplicity , Resnet is commonly used.
- For mapping the obtained representations to the space where contrastive loss is applied , a small projection head is used .
- $\{\tilde{x}_k\}$ be a set containing a positive pair of $\tilde{x}_i$ and $\tilde{x}_j$, the contrastive prediction task tries to identify $\tilde{\boldsymbol{x}}_j$ in $\{\tilde{\boldsymbol{x}}_k\}_{k\neq i}$, when a $\tilde{\boldsymbol{x}}_i$ is given , for which a contrastive loss function is defined.

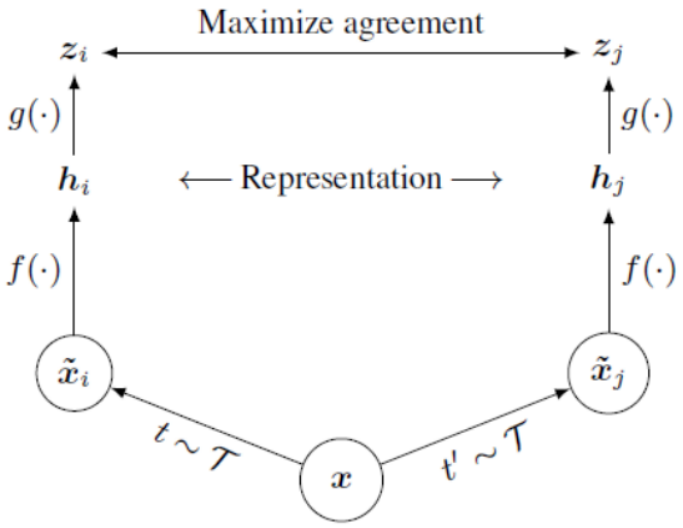


Figure 2.1: Framework for contrastive learning of visual representations

The contrastive prediction task on pairs of augmented examples produced from the minibatch is specified using a minibatch of N examples that was sampled, which creates 2N data points. Given a positive pair , the other 2(N-1) samples are treated as negative samples . The cosine similarity between two vectors u and v is given by $sim(u, v) = u^\top v/\|u\|\|v\|$ .The loss function is therefore defined as follows for a positive pair of instances (i,j):

$$\ell_{i,j} = -\log \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{I}_{[k\neq i]} \exp\left(\operatorname{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right)/\tau\right)} \tag{2.2}$$

where $\mathbb{I}_{[k \neq i]} \in \{0, 1\}$ is the indicator function evaluating to 1 iff $k \neq i$ and τ denotes a temperature parameter.Finally the loss function is defined by,

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)] \tag{2.3}$$

This function is minimized. After training, the projection head g is discarded, and the encoder f and representation it generates are used for subsequent tasks.

2.3 MoCo

Unsupervised visual representation learning utilising methodologies linked to contrastive loss has produced positive results in several research. These are motivated by a variety of factors, but they may also be seen as developing dynamic dictionaries. The "keys" in the dictionary are samples taken from many types of data, including images and patches. An encoder network represents the tokens (keys). Unsupervised learning teaches encoders to execute dictionary searches; a "query" that has been encoded should resemble its corresponding key and differ from other keys. The dictionary is kept as a queue of data samples, with the most recent encoded representations being added to the queue and the older ones being removed. The queue makes it possible to have a big dictionary.Because the dictionary keys we get come from earlier mini-batches, a slowly advancing key encoder is constructed as a momentum-based moving average of the query encoder to preserve consistency. The method that has been talked about here is MoCo [10]

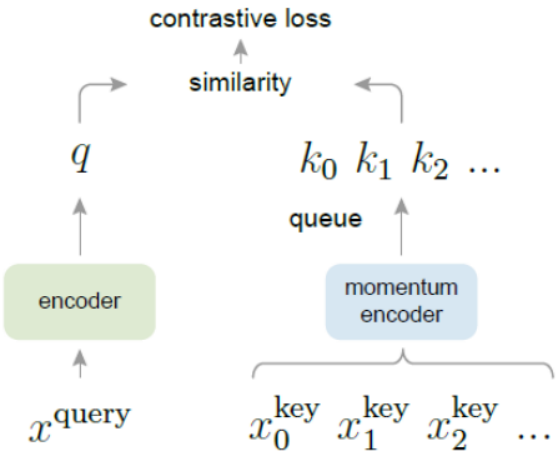


Figure 2.2: Diagrammatic representation of MoCo Framework

Let q be an encoded query and $k_0, k_1, k_2, ...$ are the keys of a dictionary which makes a set of encoded samples. Let's assume that q corresponds to a single key (indicated by k_+) in the dictionary. When q is similar to its positive key k_+ and different to all other keys (negative keys for q), a function called a contrastive loss has a low value.The loss function considered here is InfoNCE ,

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)} \tag{2.4}$$

τ is the temperature hyperparameter
The representation of the query is q = $f_q(x^q)$. where a query sample is x^q and an encoder network is f_q , the input can be images or patches and the network encoders can be similar,partially shared or completely different.

2.3.1 Momentum update

The dictionary may be huge because a queue is being utilised, making it difficult to update the key encoder via back-propagation (the gradient should propagate to all samples in the queue). A simple approach would be to ignore this gradient and replicate the key encoder f_k from the query encoder f_q. However, this technique performs poorly in experiments; the key representations' consistency is reduced by the quickly changing encoder, which is the most likely cause of this failure. To remedy this problem, a momentum upgrade has been suggested .
Let the parameters of f_k be θ_k and those of f_q be θ_q, we make an update to θ_k by:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{2.5}$$

Here $m \in [0, 1)$ is a momentum coefficient. The only variable changed by back-propagation is θ_q. Compared to θ_q, θ_k evolves more smoothly as a result of the momentum update. As a consequence, even if the encoders used to create the keys in the queue differ , the difference among these encoders can be made small. By experiments, it has been seen that a slowly evolving key encoder is particularly crucial when using a queue, as demonstrated by the fact that a relatively big momentum (e.g., $m = 0.999$) performs far better than a lower number (e.g., $m = 0.9$).

2.4 Barlow Twins

The goal of SSL is to develop embeddings that are resistant to distortions in the input sample. However, there is a persistent problem with this strategy, and that is the presence of trivial constant solutions. This may be prevented by careful implementation details, which is what this Barlow Twins [11] aims to do. The cross-correlation matrix between the outputs of two similar networks is readily measured by the suggested objective function. A sample is supplied to these networks in distorted form. Additionally, it aims to keep the matrix as near to the identity matrix as feasible in order to prevent collapse. As a result, embedding vectors of distorted copies of a sample start to resemble one another, reducing redundancy between the components of these vectors. This technique does not require big batches, gradient stopping, asymmetry between the network twins like a predictor network, or a moving average on the weight updates. As a result, it gains from output vectors with extremely high dimensions.

For every image in a batch X sampled from a dataset, Barlow Twins generates two distorted views. It makes use of a distribution of data augmentations $\mathcal{T}$ to obtain those . The two batches of distorted views, Y^A and Y^B, are then sent as inputs to a function. f_θ which is generally a deep network that can be trained and generates batches of the embeddings Z^A and Z^B, respectively. To make notations simpler, it is assumed that Z^A and Z^B have a mean output of 0 for the batch.This framework has an unique loss function defined by ,

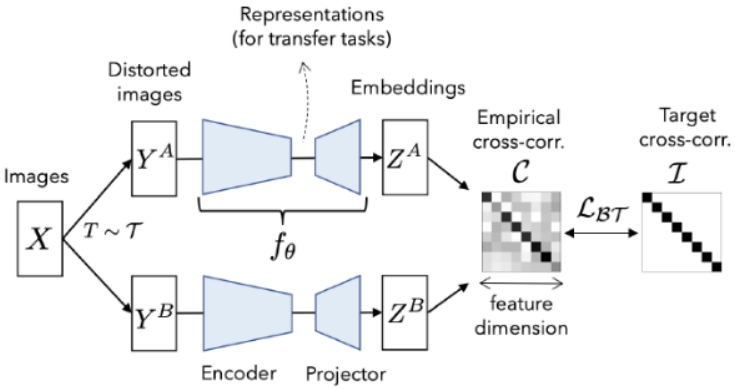


Figure 2.3: Barlow Twins Framework

$$\mathcal{L}_{\mathcal{BT}} = \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \underbrace{\lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}} \tag{2.6}$$

λ is a positive constant and $\mathcal{C}$ is the cross-correlation matrix computed between the outputs of the two identical networks along the batch dimension , where the elements of the matrix is given by ,

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2} \sqrt{\sum_b \left(z_{b,j}^B\right)^2}}$$

.

where b denotes batch samples and i,j denote the vector dimension of the networks outputs. $\mathcal{C}$ is a square matrix having entries between -1 (perfect anti-correlation) and 1(perfect correlation) and size equal to the dimensionality of the network's output. The embedding becomes invariant to the applied distortions when the invariance term attempts to equate the diagonal components of the cross-correlation matrix to 1. In order to decorrelate the various vector components of the embedding, the redundancy reduction term of the goal attempts to equate the off-diagonal members of the cross-correlation matrix to 0. The redundancy between the output units is decreased by this decorrelation.

2.5 BYOL

The objective of BYOL [12] is to learn a representation y_θ that can be used for downstream tasks. The online and target networks are the two neural networks that BYOL utilises to learn. The online network is composed of three stages: an encoder f_θ, a projector g_θ, and a predictor q_θ. It is specified by a set of weights. The target network utilises a different set of weights while having the same architecture as the online network. To train the online network, the target network supplies regression goals, and its parameters ξ are an exponential moving average of the online parameters θ. An image x sampled uniformly

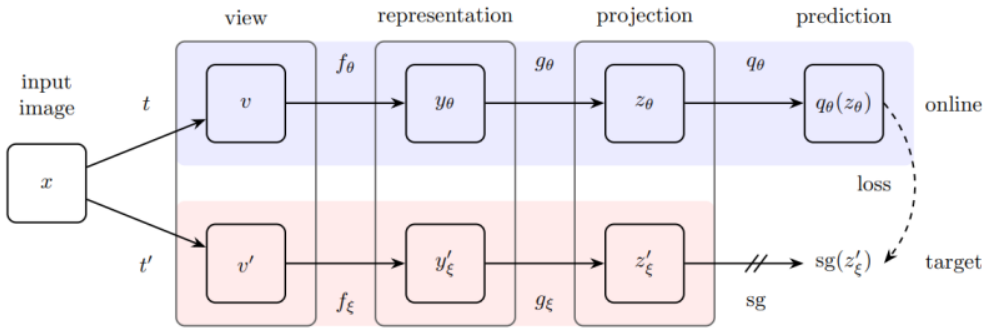


Figure 2.4: BYOL Architecture

from the set of images $\mathcal{D}$, where $\mathcal{T}$ and $\mathcal{T}'$ are two distributions of image augmentations, BYOL produces two augmented views v and v' from x by applying respective image aumentations $t \sim \mathcal{T}$ and $t' \sim \mathcal{T}'$. A representation $y_\theta = f_\theta(v)$ and a projection $z_\theta = g_\theta(y)$ is produced by the online network from the initial augmented view v . Similarly , from the second augmented view v' , the target network gives output $y'_\xi = f_\xi(v')$ and the target projection $z'_\xi = g_\xi(y')$. Then we get an output $q_\theta(z_\theta)$ predicted by z'_ξ and l_2 normalized to $\overline{q_\theta}(z_\theta)$ and $\overline{z'_\xi}$ respectively .The architecture between the online and target pipelines is asymmetrical since this predictor is only applied to the online branch.

2.5.1 BYOL loss function

The mean squared error between the target projections and the normalized predictions serves as the loss function,

$$\mathcal{L}_{\theta,\xi} = \left\|\overline{q_\theta}(z_\theta) - \bar{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\left\langle q_\theta(z_\theta), z'_\xi \right\rangle}{\|q_\theta(z_\theta)\|_2 \cdot \left\|z'_\xi\right\|_2} \tag{2.7}$$

In order to calculate $\tilde{\mathcal{L}_{\theta,\xi}}$, the target network and the online network are each fed with v and v' separately, we symmetrize the loss $\mathcal{L}_{\theta,\xi}$. With regard to just, but not, we conduct a stochastic optimization step at each training step to reduce $\mathcal{L}_{\theta,\xi}^{BYOL} = \mathcal{L}_{\theta,\xi} + \tilde{\mathcal{L}_{\theta,\xi}}$, as shown by the stop-gradient(sg) in Figure 3.4 .
More specifically, given a target decay rate of $\tau \in [0, 1]$, we update the model as follows after each training step:

$$\theta \leftarrow \text{optimizer}\left(\theta, \nabla_\theta \mathcal{L}_{\theta,\xi}^{\text{BYOL}}, \eta\right) \tag{2.8}$$

$$\xi \leftarrow \tau\xi + (1 - \tau)\theta \tag{2.9}$$

where optimizer and η is optimizer and learning rate respectively .

2.6 Relative Performances

The top-1 and top-5 accuracies obtained on the ImageNet [13] validation set are given in the following table after training a linear classifier on ImageNet on top of fixed representations of a ResNet-50 pretrained with the methods mentioned.

Table 2.1: Relative Performances of SSL algos

Method	Top-1	Top-5
Supervised	76.5	
MoCo	71.1	90.1
SIMCLR	69.3	89.0
BYOL	74.3	91.6
BARLOW TWINS	73.2	91.0

2.7 Conclusion

While other methods (SimCLR, MoCo) use both positive and negative pairs and the contrastive loss is then minimized for training, BYOL doesn't use negative pairs. Also, addition of a predictor in the online network helps BYOL to avoid collapsed solutions. These might be the reasons behind BYOL's improved performance over the other methods which can be seen by their performance on the ImageNet dataset.

Chapter 3

Robustness Test of SSL on Imbalanced Datasets

Class imbalance is a frequent issue that has been thoroughly researched in classical machine learning, but there is relatively little systematic study accessible in the context of deep learning [14]. Because SSL doesn't require labels for learning, it is a useful method for learning common visual representations. Although the behaviour of SSL is not well understood, large unlabelled datasets in the wild frequently contain long-tailed label distributions. Extensive investigations have shown that commercially available self-supervised representations are already more resistant to class imbalance than supervised representations. With SSL, there is a substantially lower performance disparity between balanced and unbalanced pre-training, across sample sizes, for both in-domain and, especially, out-of-domain assessment than there is with supervised learning.

Systematically examining the representation quality of SSL methods under class imbalance has revealed that pre-trained SSL representations are already more resistant to dataset imbalance than pre-trained representations. By using a linear probe on in-domain (ID) data and fine-tuning on out-of-domain (OOD) data, the representation quality has been assessed.

Under several configurations, such as dataset sizes and imbalance ratios, and with both in-domain and out-of-domain assessments, it has been found that the balance-imbalance gap for SSL is significantly lower than that for SL (hence superior) (see Figure 3.1 and Figure 3.2 for more details). Although SSL does not require labels and is therefore more easily applicable to bigger datasets than SL, this resilience is true even when the number of samples for the two algorithms is equal.

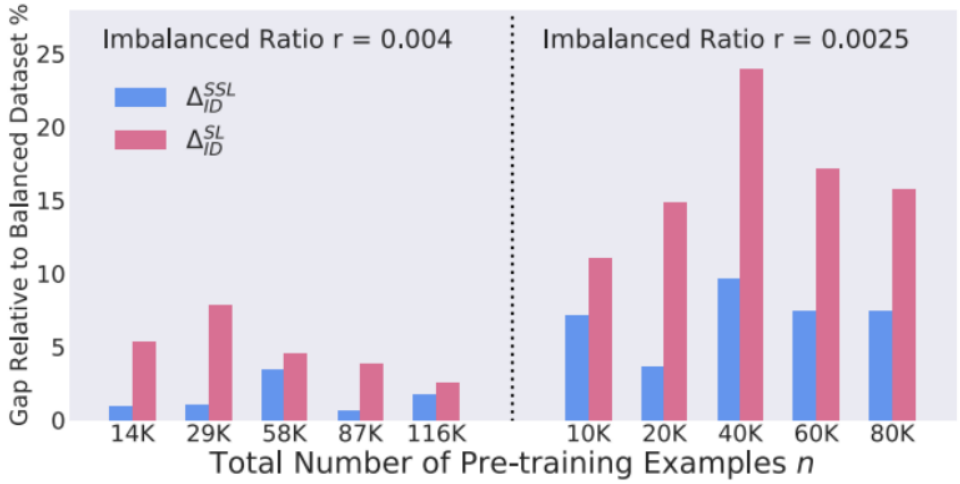


Figure 3.1: Relative performance gap between self-supervised (MoCo v2) representations on long-tailed ImageNet with varying numbers of instances (blue for SSL) and supervised ones (red for SL) n, spanning ID assessment

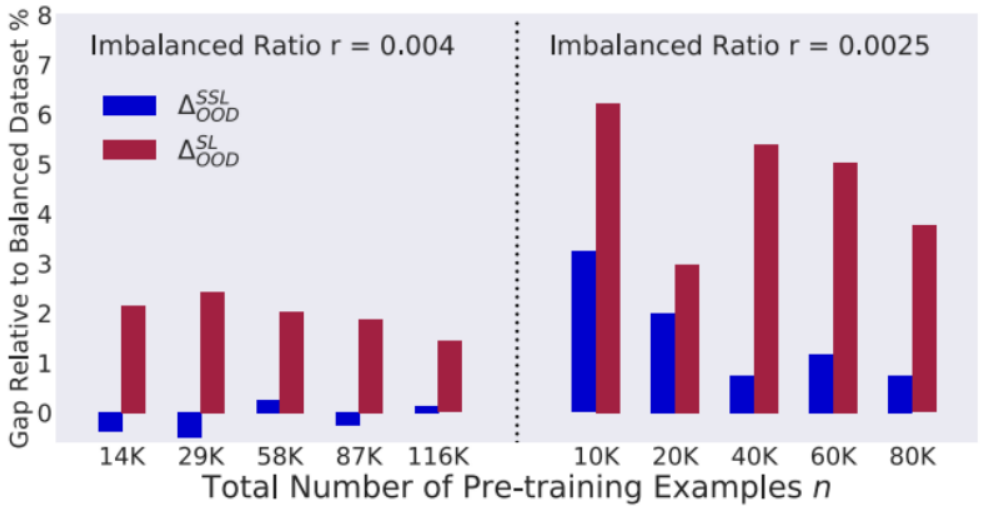


Figure 3.2: Relative performance gap between self-supervised (MoCo v2) representations on long-tailed ImageNet with varying numbers of instances (blue for SSL) and supervised ones (red for SL) n, spanning OOD assessment

3.1 Relative performance gap

$A^{SSL}(n,r)$ and $A^{SL}(n,r)$ indicate the accuracy of the generated supervised and self-supervised representations, respectively when there are n datapoints present in the dataset and the imbalance ratio is r . When r=1 , the dataset is balanced. By extensive experiments [15], it can be seen that ,

$$\Delta^{\mathrm{SSL}}(n,r) = \frac{A^{\mathrm{SSL}}(n,1) - A^{\mathrm{SSL}}(n,r)}{A^{\mathrm{SSL}}(n,1)} \ll \Delta^{\mathrm{SL}}(n,r) = \frac{A^{\mathrm{SL}}(n,1) - A^{\mathrm{SL}}(n,r)}{A^{\mathrm{SL}}(n,1)}$$

where Δ denotes the relative performance gap .

3.2 CIFAR10-LT

The imbalanced version of CIFAR10 is called CIFAR10-LT .While the CIFAR10 dataset has 10 classes with 6000 32*32 images in each class , with the imbalance factor 0.01 , the classes of CIFAR10-LT now consists with 5000,2997,1796,1077,645,387,232, 139,83,50 samples respectively.

3.3 Experiment on CIFAR10-LT

Using this imbalanced dataset for pre-training with SimCLR framework for 200 epochs using Adam optimizer and learning rate 0.001, we achieved 83% accuracy in the downstream classification using linear evaluation on CIFAR10. In linear evaluation , a linear classifier was trained on the features obtained by passing the images through the pre-trained encoder .The resulting confusion matrix was ,

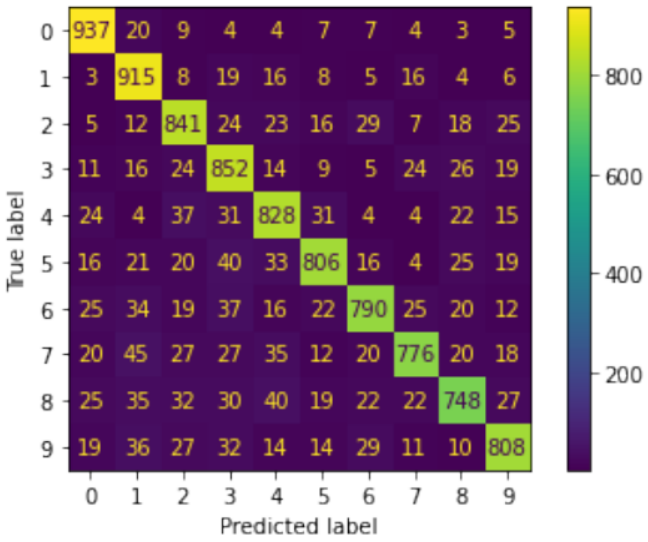


Figure 3.3: Confusion Matrix

3.4 Analysis

The classes which are rare in the imbalanced dataset contain only few datapoints, so it becomes hard to learn proper features and hence it also becomes hard to classify those classes.Seeking assistance from the characteristics learnt by the frequent classes is one method to solve this issue. However, because classification tasks are supervised, the model learns features that are primarily effective for categorising the common classes and also tends to disregard characteristics that can transfer to the rare classes and other downstream tasks. Jamal et al. urge the model to learn properties that may be transferred from frequent classes to rare classes as a result of this. The models in SSL, on the other hand, are able to acquire deeper features that capture the fundamental structures of the inputs, characteristics that are helpful for identifying the frequent classes, as well as features that are transferable to rare classes.

3.5 Conclusion

We can conclude that, when the dataset is imbalanced, SSL indeed achieves good results compared to SL(Resnet18 achieves 93.02% accuracy on CIFAR10) as we can see from our experiment of SimCLR on CIFAR10-LT . The most probable reason for it being the nature of supervised learning to learn features from frequent classes and not from rare classes while SSL learns richer features from frequent data that are transferable to rare data.

Chapter 4

Methodology : BYOLMed2D

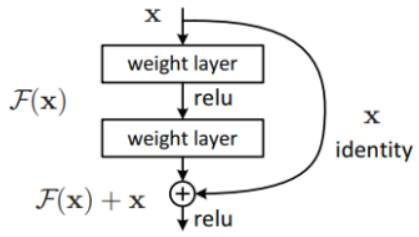
In this chapter, we discuss the architecture and loss functions used in our project. In the first section 4.1, we describe the architecture use as the backbone encoder in our proposed algorithm. In Sec. 4.2, we describe the model configuration used in pre-training stage of the proposed self-supervised learning framework. Finally, in Sec. 4.3, we give a brief description of the downstream task in our work.

4.1 Architecture

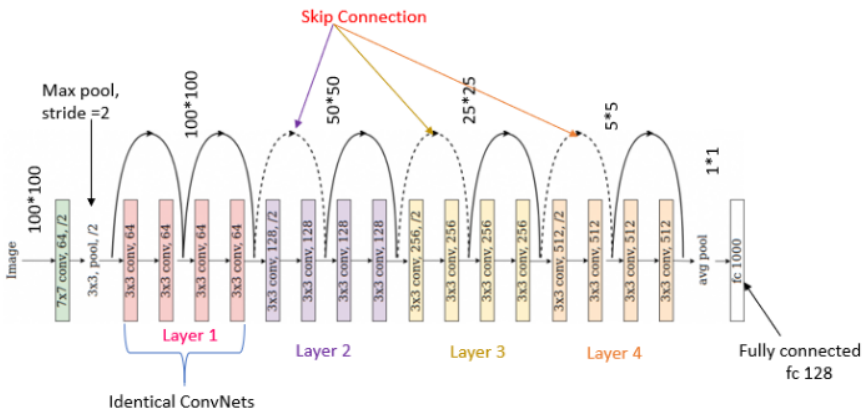
4.1.1 Residual Networks

Deep residual networks [16],[17] are the only ones we use for our research because of their excellent performance and simplicity. There have been reports of practical challenges while training neural networks to carry out tasks when the input/output sequences' temporal dependencies cover extensive time periods [18]. ResNet is a type of model that employs 2D convolution . Resnet18 is a 18 layers deep network . Let's assume that the underlying mapping $\mathcal{G}(x)$ of a residual block can be fitted by a few stacked layers , with the inputs to the first of these layers being x. It is similar to hypothesizing that multiple nonlinear layers can asymptotically approach the residual functions, i.e., $\mathcal{G}(x)-x$, if one believes that many nonlinear layers can asymptotically approximate the complex functions . Therefore, we allow these layers to approach a residual function $\mathcal{F}(x) := \mathcal{G}(x) - x$ rather than expecting stacked layers to roughly approximate $\mathcal{H}(x)$. Thus, the original function is thus $\mathcal{F}(x)+x$. Dimensions of x and $\mathcal{F}$ should be same.

In Resnet18 we have 18 residual blocks stacked on each other. In every block , the input is



(a) Residual Block used in ResNet architecture



(a) ResNet18 Architecture

passed through a convolution or weight layer . The output of this is then batch normalized to be fed to the RELU [19] activation function and then again through a weight layer . After the last block , there is a fully connected layer. No. of parameters present in Resnet18 is 11.14 million.

4.2 Pre-training

In order to extract crucial characteristics from training data and discover significant patterns, network depth is crucial for solving complicated image analysis issues utilising deep learning. However, because of the gradients, adding more neural layers can be computationally expensive and difficult. Resnet solves this problem well and that's why BYOLMed2D method uses Resnet18 as the backbone. In the pre-training part, we consider only 16 frames from each video. The average number of frames in each video in the MRNet dataset is approximately 30. We randomly choose 16 frames from each video. This allows the model to discard any temporal bias in the features. Furthermore, the change in features over consecutive frames in medical videos are low. Thus, the variation in features in these type of data are harder to learn, than fast changing features, in self-supervised learning. interleaving more than 1 frames between two sampled frames, on an average, we intend to change the slow changing features to fast changing ones. This makes the slow varying features in the data more contributing. Consequently, this helps in learning a more generalizable feature pool from the data without any ground truth annotations.

The randomly sampled frames are passed through Resnet18. However, the each sample in the MRNet dataset is 3D in nature, that is, each sample is in the form $Frames \times 3 \times Height \times Width$. To extract the encoded representations, we reshape the data obtained from the dataloader in the form $BatchSize \times 16 \times 3 \times Height \times Width$ to the shape $(BatchSize \times 16) \times 3 \times Height \times Width$ before passing through ResNet18. While θ denote the set of weights for the online network , ξ is the exponential moving average(**EMA**) of the online parameters θ and work as the parameters for the target network. After obtaining the encoded representations from the ResNet18 encoder, it is reshaped to $BatchSize \times 16 \times 3 \times Height \times Width$. Then the feature of those frames are maxpooled over the frames to obtain a single feature vector of shape $BatchSize \times 3 \times Height \times Width$.

To feed the representations to the projector, we take a global average pooling of the representations obtained after maxpooling over the frames. This feature vector is then given as input to the projector and predictor in BYOL framework. The loss function used was stated in the Sec. 2.5.1.

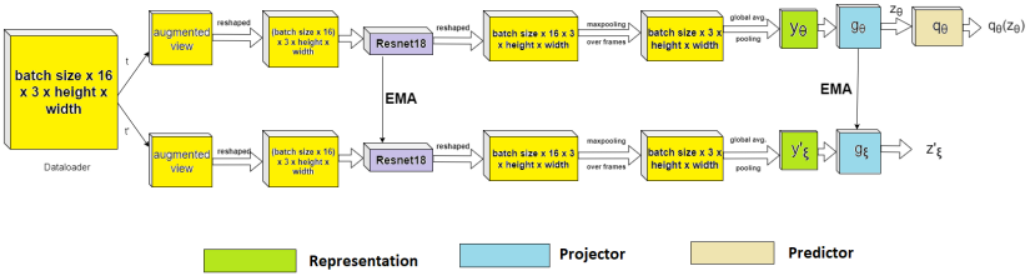


Figure 4.3: Dataflow of BYOLMed2D framework

4.2.1 Maxpooling over frames

A pooling layer's main goal is to gather features from maps produced by convolving a filter over an image. Formally, its purpose is to gradually shrink the representation's spatial dimension in order to minimise the number of parameters and computations required by the network. Max pooling is the most typical type of pooling. It is preferred over AvgPool because, while individual qualities in MaxPool are highlighted regardless of location, average features are highlighted in AvgPool.

The purpose of using maxpooling in our framework is to extract the most contributing feature over all the randomly selected frames, which is invariant under transformations. As the two views or augmented versions, v and v' of the MR video x are differently augmented, it is intuitive that the latent feature vectors encoded by the backbone will be mapped to different locations in the latent space. The purpose of BYOL is to bring the two differently augmented MR clip closer to each other in the latent space. By extracting the features which have maximum response over each unit spatial dimension in the encoded representations, we enforce the model to learn the temporally varying representations which are invariant under transformations. This allows us to incorporate the capability of learning temporally varying information in addition to the spatial information in the model. Consequently, this allows the model to learn generalized anatomy-aware representations from the MR videos.

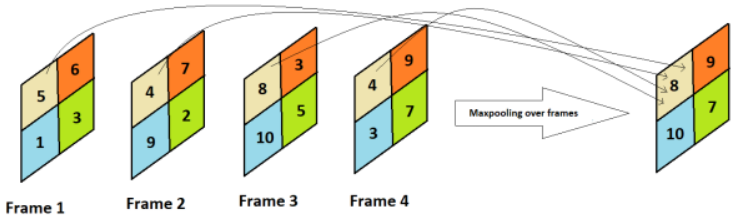


Figure 4.4: Example of maxpooling over frames

4.3 Downstream

In the downstream task, the objective is to classify MRNet [20] videos for detecting ACL tear injury. Hence, the downstream task is a binary classification task. But the number of samples with occurence of ACL tear is far less than the number of samples which do not have ACL tear injury. To deal with the imbalance in the data, we use oversampling technique [21] to balance the dataset. By randomly selecting samples with occurrence of ACL tear, we make the number of positive samples equal to the number of samples not containing ACL tear injury. Here also 16 frames from each video were used. For the optimization of the binary cross entropy loss, we used a batch size of 1. The binary cross entropy loss was optimized with Adam optimizer for 30 epochs, with an initial LR 0.0001 and multistep LR decay of 0.1 at 18th and 24th epoch.

4.3.1 Loss function for downstream task

A Sigmoid layer and the BCELoss are combined into a single class in this loss. By integrating the operations into one layer, we are able to take use of the log-sum-exp method for numerical stability, making this version more stable numerically than one that uses a simple Sigmoid followed by a BCELoss. The unreduced loss can be described as:

$$l(x, y) = -[y.log(\sigma(x) + (1 - y).log(1 - \sigma(x))] \tag{4.1}$$

where

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

is the sigmoid function.

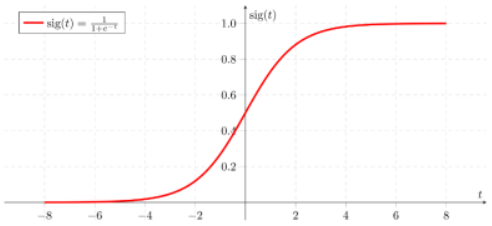


Figure 4.5: Sigmoid

4.4 Conclusion

In the methodology , we have used a framework that is similar to BYOL as BYOL gives better result than the other self-supervised learning algorithms when performed on a particular dataset as seen earlier .

Chapter 5

Experiments and Results

In this chapter we talk about the dataset(s) used for pretraining and downstream task , the augmentations used by the model , results obtained by the model and its comparisons to other methods which also detect ACL tear injury.

5.1 Dataset

We utilise the MRNet dataset [20] as our reference dataset in the downstream task. There are 1370 knee MR video clips altogether in the MRNet collection. The training set consists of 1130 MR video clips, whereas the tuning or validation set consists of 120 MR video clips. Only 208 videos out of the 1,130 training examples—include an ACL injury. The dataset we are utilising for this investigation is obviously quite imbalanced. Hence we have the chance to investigate how self-supervised learning methods perform on datasets with imbalance.

5.2 Augmentations

The collection of augmentations used by BYOL during self-supervised training is listed below:

- Random cropping: An area uniformly sampled between 8% and 100% of the original picture and an aspect ratio logarithmically sampled between 3/4 and 4/3 are used to choose a random patch of the image. Using bicubic interpolation, this patch is then scaled to the desired size of 224 x 224.
- optional left-right flip
- Color jittering: A uniformly random offset that is applied to all the pixels of the same picture shifts the brightness, contrast, saturation, and hue of the image. For every patch, a different random sequence for these shifts is chosen.
- color dropping: an optional conversion to grayscale.
- Gaussian blurring: for a 224×224 image, a square Gaussian kernel of size 23×23 is used, with a standard deviation uniformly sampled over [0.1, 2.0]

- solarization: an optional color transformation for pixels with values in $[0, 1]$.

The augmentations from the sets $\mathcal{T}$ and $\mathcal{T}'$ are combinations of the aforementioned image augmentations performed with varying probabilities in the sequence given above.

Table 5.1: Agmentations with probabilities

Parameter	$\mathcal{T}$	$\mathcal{T}'$
Random crop probability	1.0	1.0
Flip probability	0.5	0.5
Color jittering probability	0.8	0.8
Brightness adjustment max intensity	0.4	0.4
Contrast adjustment max intensity	0.4	0.4
Saturation adjustment max intensity	0.2	0.2
Hue adjustment max intensity	0.1	0.1
Color dropping probability	0.2	0.2
Gaussian blurring probability	1.0	0.1
Solarization probability	0.0	0.2

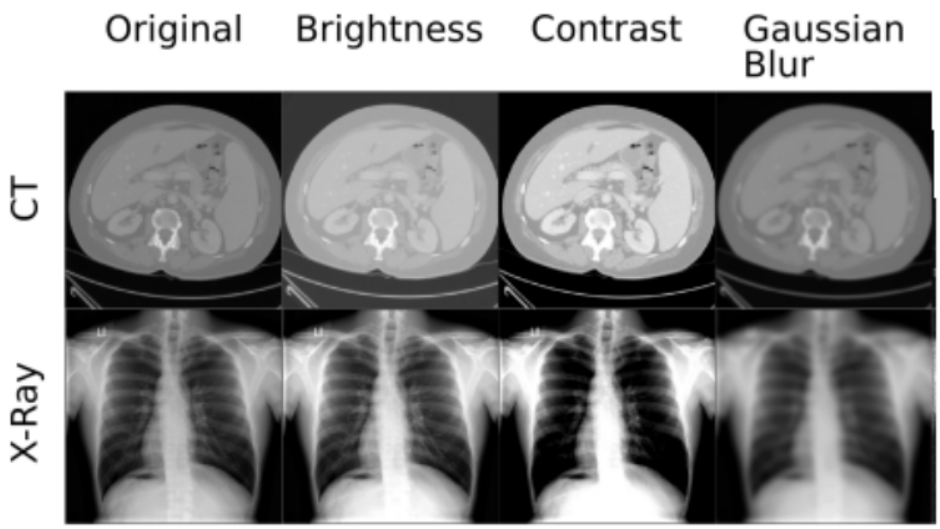


Figure 5.1: Image augmentations

5.3 Result

The results obtained by this model were **73.3%** accuracy and **0.801** AUC score for ACL tear detection on saggital plane.

5.3.1 Comparison with other results

There are some other methods which detect ACL tear injury like MRNet(supervised) [22] achieving 86.7% accuracy and 0.915 AUC score , SSLM [23] achieving 80% accuracy and 0.893 AUC score , SIMCLR using JIGSAW PUZZLE [9] achieving 62.5% accuracy and 0.691 AUC score (this result was collected from the SSLM paper) , MoCo v2 using JIGSAW PUZZLE [24] achieving 45.8% accuracy and 0.389 AUC score , S.Manna et. al [25] achieving 76.62% accuracy and 0.848 AUC score ..

Table 5.2: Comparison to other methods(The result of SimCLR using JIGSAW PUZZLE was collected from the SSLM paper)

Method	Accuracy	AUC score
MRNet(supervised) [22]	86.7%	0.915
SSLM [23]	80%	0.893
SIMCLR using JIGSAW PUZZLE* [9]	62.5%	0.691
MoCo v2 using JIGSAW PUZZLE [24]	45.8%	0.389
S.Manna et. al [25]	76.62%	0.848
BYOLMed2D	73.3%	0.801

5.4 Conclusion

Comparing with other methods , we can see that our method achieves average result as while our method performs better than few methods, there are few which give better result than BYOLMed2D as well.

Chapter 6

Conclusion

This study aims to investigate self-supervised learning algorithms' potential in medical image analysis .Although BYOLMed2D achieves good results still it continues to rely on pre-existing augmentation sets that are unique to vision-related applications. It is essential to generate equally appropriate augmentations for each of the other modalities (such as audio, video, text, etc.) in order to generalise BYOLMed2D to them. Such augmentations could need a lot of knowledge and work to design. To extend the method to other modalities, it would be crucial to automate the search for these augmentations. We look to improve the accuracy in downstream classification task(ACL tear injury detection) by experimenting with different hyperparameters .

Bibliography

[1] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, pp. 1–13, 2018.

[2] K. Ohri and M. Kumar, "Review on self-supervised image recognition using deep neural networks," *Knowledge-Based Systems*, vol. 224, p. 107090, 2021.

[3] 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.

[4] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *arXiv preprint arXiv:2109.08685*, 2021.

[5] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[6] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[7] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, "A theoretical analysis of contrastive unsupervised representation learning," in *International Conference on Machine Learning*, pp. 5628–5637, PMLR, 2019.

[8] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie, "When does contrastive visual representation learning work?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14755–14764, June 2022.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[10] 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*, pp. 9729–9738, 2020.

[11] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, pp. 12310–12320, PMLR, 2021.

[12] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[14] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol. 106, pp. 249–259, 2018.

[15] H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma, "Self-supervised learning is more robust to dataset imbalance," *arXiv preprint arXiv:2110.05025*, 2021.

[16] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

[17] X. Du, Y. Li, Y. Cui, R. Qian, J. Li, and I. Bello, "Revisiting 3d resnets for video recognition," *arXiv preprint arXiv:2109.01696*, 2021.

[18] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[19] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[20] D. Azcona, K. McGuinness, and A. F. Smeaton, "A comparative study of existing and new deep learning methods for detecting knee injuries using the mrnet dataset," in *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pp. 149–155, IEEE, 2020.

[21] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," *Int. J. Recent Trends Eng. Res*, vol. 3, no. 4, pp. 444–449, 2017.

[22] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. N. Patel, K. W. Yeom, K. Shpanskaya, *et al.*, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet," *PLoS medicine*, vol. 15, no. 11, p. e1002699, 2018.

[23] S. Manna, S. Bhattacharya, and U. Pal, "Sslm: Self-supervised learning for medical diagnosis from mr video," *arXiv preprint arXiv:2104.10481*, 2021.

[24] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[25] S. Manna, S. Bhattacharya, and U. Pal, "Self-supervised representation learning for detection of acl tear injury in knee mr videos," *Pattern Recognition Letters*, vol. 154, pp. 37–43, 2022.

KNEE MRI DIAGNOSIS USING SELF-SUPERVISED LEARNING

ORIGINALITY REPORT

9%
SIMILARITY INDEX

PRIMARY SOURCES

1	arxiv.org Internet	144 words — 2%
2	researchers.lille.inria.fr Internet	84 words — 1%
3	towardsdatascience.com Internet	78 words — 1%
4	Lidan Wu, Daoming Zong, Shiliang Sun, Jing Zhao. "A Sequential Contrastive Learning Framework for Robust Dysarthric Speech Recognition", ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021 Crossref	35 words — < 1%
5	www.arxiv-vanity.com Internet	33 words — < 1%
6	arxiv-export-lb.library.cornell.edu Internet	32 words — < 1%
7	deepai.org Internet	31 words — < 1%
8	Hao Geng, Fan Yang, Xuan Zeng, Bei Yu. "When Wafer Failure Pattern Classification Meets Few-shot Learning and Self-Supervised Learning", 2021 IEEE/ACM	29 words — < 1%

International Conference On Computer Aided Design (ICCAD), 2021
Crossref

9	keerc.snu.ac.kr Internet	26 words —	< 1%
10	export.arxiv.org Internet	25 words —	< 1%
11	Saeed Shurrab, Rehab Duwairi. "Self-supervised learning methods and applications in medical imaging analysis: a survey", PeerJ Computer Science, 2022 Crossref	22 words —	< 1%
12	Xiaohong Zhang, Liqing Jiang, Dongxu Yang, Jinyan Yan, Xinhong Lu. "Urine Sediment Recognition Method Based on Multi-View Deep Residual Learning in Microscopic Image", Journal of Medical Systems, 2019 Crossref	21 words —	< 1%
13	Md Mahfuzur Rahman Siddiquee, Andriy Myronenko. "Chapter 15 Redundancy Reduction in Semantic Segmentation of 3D Brain Tumor MRIs", Springer Science and Business Media LLC, 2022 Crossref	20 words —	< 1%
14	Abebech Jenber Belay, Ayodeji Olalekan Salau, Minale Ashagrie, Melaku Bitew Haile. "Development of a chickpea disease detection and classification model using deep learning", Informatics in Medicine Unlocked, 2022 Crossref	19 words —	< 1%
15	ir.ahduni.edu.in Internet	17 words —	< 1%

16	library.isical.ac.in:8080 Internet	17 words — < 1%
17	libraries.io Internet	16 words — < 1%
18	"Computer Vision – ECCV 2020", Springer Science and Business Media LLC, 2020 Crossref	15 words — < 1%
19	proceedings.mlr.press Internet	14 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 14 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 14 WORDS