Pune Institute of Computer Technology
Dhankawadi, Pune

A SEMINAR REPORT
ON

SELF SUPERVISED LEARNING IN IMAGES

SUBMITTED BY

Prathamesh Sonawane
Roll No. 31164
Class TE 1

Under the guidance of
Prof. Rujuta Kulkarni

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2021-22

DEPARTMENT OF COMPUTER ENGINEERING
Pune Institute of Computer Technology
Dhankawadi, Pune-43

CERTIFICATE

This is to certify that the Seminar report entitled

"SELF SUPERVISED LEARNING IN IMAGES"

Submitted by
Prathamesh Sonawane   Roll No. 31164

has satisfactorily completed a seminar report under the guidance of
Prof. Rujuta Kulkarni towards the partial fulfillment of third year
Computer Engineering Semester II, Academic Year 2019-20 of
Savitribai Phule Pune University.

Dr. A. R. Deshpande
Internal Guide

Prof. M.S.Takalikar
Head
Department of Computer Engineering

Place: Pune
Date: 29/05/21

ACKNOWLEDGEMENT

Contents

List of Tables

List of Figures

Abstract

Many recent methods self-supervised representation learning train feature extractors by maximizing an estimate of the mutual information (MI) between different views of the data. In this literature survey we discuss and provide evidence to analyze the extent to which maximizing Mutual Information (MI) between vectors has an effect on training feature extractors.2 images are passed through particular feature extractors and 2 independent 2D-vectors for each of the images respectively are created. Following this MI is calculated using a predefined formula between these 2 vectors, using one among a variety of estimators (InfoNCE, InfoMAX, etc). The primary task is to maximize this MI. After noting down initial results they discuss if employing a better MI formula is the only way in which we could get better feature extractors. This paper is to prove that maximization of mutual information, and focusing on improving solely that is not a necessary and/or sufficient condition for beating current SOTA results. It also suggests that while all of this is true we must continue to pursue better formulae to find MI as although it isn't solely responsible for the performance of our experiment but it does play a vital role in it.

Another new approach to self-supervised image representation learning is BYOL (Bootstrap your own latent), it relies on two neural networks, referred to as online and target networks, that interact and learn from each other.

Keywords

1 INTRODUCTION

Self supervised learning these days has become essential, considering the amount of data that is needed to train models it is highly effective to create labels empirically rather than by manual effort. Since in many cases the amount of data may determine the output quality of our predictors, investing millions into creating sub par labelled data sets isn't feasible anymore. The amount of unlabelled data out-numbers labelled data in the ratio 1:1 million, and considering there is no way to label all of the data self-supervised learning comes in handy.

Learning good image representations is a key challenge in computer vision as it allows for efficient training on downstream tasks. Many different training approaches have been proposed to learn such representations, usually relying on visual pretext tasks.

Self supervised learning is a relatively new area of research and requires more than average amount of compute while experimenting because of the sheer scale of things. Training times can vary from 2 days to 2 months and this makes it quite inapproachable to lay persons. But with the resent onset of online cloud computing it has become somewhat accessible and as we get better hardware the training times are being reduced too. The reason they require such high compute is because the weights are always randomly initialized and due to the absence of transfer learning the model takes longer to converge to a minimum.

BYOL introduces a novel method for self supervised learning in image representations. It achieve a higher score than any other state-of-the-art contrastive method.In this there are 2 networks online and target, these two networks work together and help each other learn. Two augmented images are created and passed into each of the networks the aim of this exercise is to get the same vector representation for different versions of an image, in the hope that the network will learning key features along the way of achieving this.

Similarly there are various other techniques used over the years. Jigsaw++[5] divides the images into parts like a jigsaw puzzle and the aim is for the network to predict the relative positions of the patches of the image. Another one is ROT-NET[6] in which the network tries to predict the angle to which the image is rotated. Colorful image colorization[9] aims to convert black and white images to colored ones using U-net style architectures. As we can concur Self supervised learning involves a lot of imagination and coming up with a novel technique which may help the feature extractors understand images in a better way.

2   MOTIVATION

Given a task and enough labels, supervised learning can solve it really well. Good performance usually requires a decent amount of labels, but collecting manual labels is expensive (i.e. ImageNet) and hard to scale up. Considering the amount of unlabelled data is substantially more than human curated labelled datasets, it is kinda wasteful not to use them. However, unsupervised learning is not easy and usually works much less efficiently than supervised learning atleast with the current methods that we are employing.

In recent times we invest a lot of time and capital in labelling individual images to build a network around them. A decade ago the solution would have been transfer learning but for transfer learning we need pre-trained weights, pre-trained weights need to be trained in the first place. Now we could say that they could be trained on another data set, but that somewhat defeats the purpose since creating that data set did take time at some point. This is where self supervised learning comes into place. The idea that we don't need labelled data sets in order to build a good feature extractor is really fascinating and that is the reason I decided to get into this field of study for my report.

This doesn't mean that self supervised learning doesn't need labels, many tasks do require labels, but those labels are created by the computer and thus it takes significantly less amount of effort to make those labels. Although all of this is extremely promising, for the time being researchers have been trying to find out the optimum pretext task which could compete with data sets where labels are made manually.

Many ideas have been proposed for self-supervised representation learning on images. A common workflow is to train a model on one or multiple pretext tasks with unlabelled images and then use one intermediate feature layer of this model to feed a classifier on Image-Net classification. The final classification accuracy quantifies how good the learned representation is.

3 PROBLEM DEFINITION AND SCOPE

3.1 Problem Definition

To come up with a technique to extract the meaningful features to increase the efficiency of any particular model architecture that we choose.

3.2 Scope

Self supervised learning is a twist on unsupervised learning where we exploit unlabeled data to obtain labels where there is no explicit annotation or class labels associated with the data we exploit unlabeled data itself to get some kind of labels and induce a supervised learning model on unlabeled data. Specifically we design supervised tasks which are called pretext or auxiliary tasks which can learn meaningful representations through which the model becomes more ready. To then be able to solve a downstream task such as a classification or semantic segmentation or any other supervised learning task a sample task in this context could be to predict a certain part of the input from another part, somewhat like fill in the blanks of of a given input.

You can predict any part of the input from any other part. Between images and videos you could predict the future from the past you could predict the future from the recent past you could try predicting the past from the present. The top from the bottom in case of an image. In a more broader sense you could predict the occluded from the visible in general, you pretend there is a part of the input that you don't know and try to predict that, those are the different tasks or pretext tasks that you can use in self-supervised learning.

4 LITERATURE SURVEY

The Following table shows the literature survey by comparing techniques propose in various references:

Table 1: Literature survey

| No.	Techniques	Dataset	Architecture	parameters	Accuracy
1	Bootstrap Your Own Latent A New Approach to Self-Supervised Learning[1]	imagenet cifar100	online and target networks : Siamese style (Resnets)	250 million	74.3/ 79.6 (Imagenet)
2	On mutual information maximization for representation learning[4]	MNIST CIFAR10	Custom MLP arch.	250 million	85 (MNIST)
3	Self-labelling via simultaneous clustering and representation learning[3]	SVHN CIFAR-10 CIFAR-100 ImageNet	Custom MLP arch.	230 million	77.2 (Imagenet)
4	Unsupervised Visual Representation Learning by Context Prediction[5]	ImageNet	Custom arch.	200 million	54.2 (Imagenet)
5	A critical analysis of self-supervision, or what we can learn from a single image[2]	ImageNet	BIGGAN Rotnet Deep Cluster Alexnet monoGAN	150 million	72.3 (Imagenet)

5 A SURVEY ON PAPERS

5.1 Bootstrap Your Own Latent A New Approach to Self-Supervised Learning[1]

BYOL[1] method helps in learning important representations for a variety of downstream computer vision tasks such as object recognition, object detection and semantic segmentation. Once these representations are learned, they could be used with any standard object classification model like as Resnet, VGGnet, or any semantic segmentation network such as FCN8s, deeplabv3, etc or any other task-specific network and it gets to a better result than training these networks from scratch. This is the major reason behind the popularity of BYOL. The below graph shows that the BYOL representations learned using Imagenet images beats all previous unsupervised learning methods and achieves classification accuracy of 74.1 percent with Resnet50 under linear evaluation protocol.

Another interesting fact is, although a collapsed solution exists for the task curated for BYOL, the model avoids it safely and the actual reason for it is unknown. Collapsed solution means, the model might get away by learning a constant vector for any view of any image and gets to zero loss, but it does not happen

The authors of the original paper, conjecture that it might be due to the complex network(Deep Resnet with skip connections) used in the backbone, the model never gets to the straightforward collapsed solution. But in another recent paper SimSiam Chen, Xineli and He, found out it is not the complex network architecture but the "stop-gradient" operation that makes the model to avoid the collapsed representations. "stop-gradient" means that the network never gets to update the weights of the target network directly through gradients and hence never gets to the collapsed solution. They also show that there isn't any need for a momentum target network to avoid collapsed representation but it certainly gives better representations for downstream tasks if used.
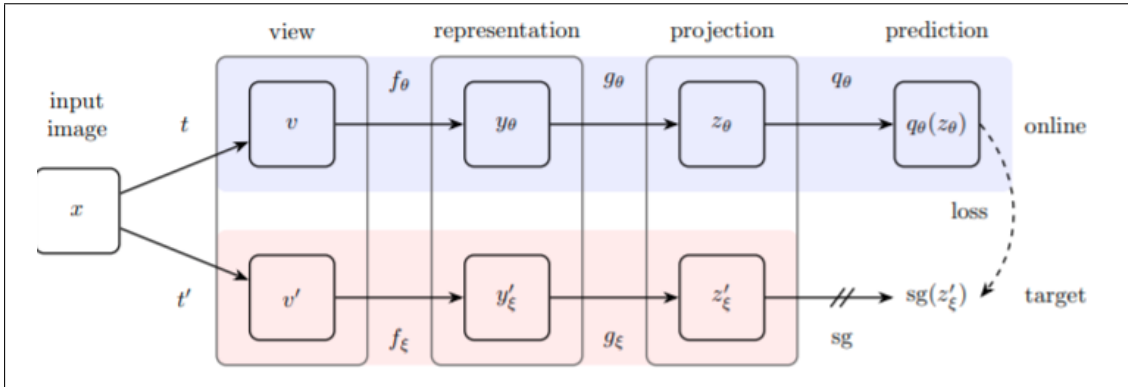
5.2 On mutual information maximization for representation learning

The paper revolves around discussing if maximising Mutual Information (MI) between vectors yields good results. Take an image split it into half, pass

the half images through independent encoders(CNN, MLP layers), get 2 vectors and now calculate the MI between these 2 vectors using a particular estimator (InfoNCE/InfoMAX, etc). The task is to maximise this MI. In the paper we discuss if maximising MI is sufficient. It is said that recently MI has beat many cutting edge scores but is that enough? The authors go on to prove that that alone is not sufficient and that encoder architecture, estimator function, critics etc play an equally important role.

This paper is to prove that maximization mutual information solely is not a necessity condition for the improvement of previous papers. It suggested to use triplet technique in which we use representation of images and we reduce difference between similar images and increasing difference between different images where difference can be called as mutual information. It also suggest to invent better formula for mutual information because current formula although good is not enough.

$$I_{\mathrm{NCE}} = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)}}\right] = \log K - \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\left(1+\sum_{j\neq i}e^{f(x_i,y_j)-f(x_i,y_i)}\right)\right].$$

5.3 Self-labelling via simultaneous clustering and representation learning

In this we try to assign labels to random images by clustering. But we use a better method that traditional K-means. Sinkhorn-Knopp optimization for obtaining pseudo-labels is always better than K-means on all metrics. After this we train the network on a simple downstream task for classification of images.

The method is obtained by maximizing the information between labels and input data indices. We show that this criterion extends standard cross entropy minimization to an optimal transport problem, which we solve efficiently for millions of input images and thousands of labels using a fast variant of the Sinkhorn-Knopp algorithm.

5.4 Unsupervised Visual Representation Learning by Context Prediction

Jigsaw puzzle - From each image 4 random patches are selected. Each patch is then divided into 9 sub-patches in a 3x3 grid. 2 sub patches out of these 9 are selected and our self supervised task is to feed those 2 patches into the model and train it to figure out their relative positions to each other.

Suppose you get two patches A and B from the same image can you tell where B should go relative to a we argue that doing this often requires recognizing object semantics hence given unlabeled images we create a supervised task by randomly sampling one patch then a second patch and then training a deep network to predict the relative positions we find that the features this network extracts can be

used to match patches semantically surprisingly what is learned from this within image tasks captures similarity across images we showed that for Pascal object detection pre-training on unlabeled images gets a significant boost over training on Pascal alone which amounts to more than a third of the benefit from training on a million image-net labels.
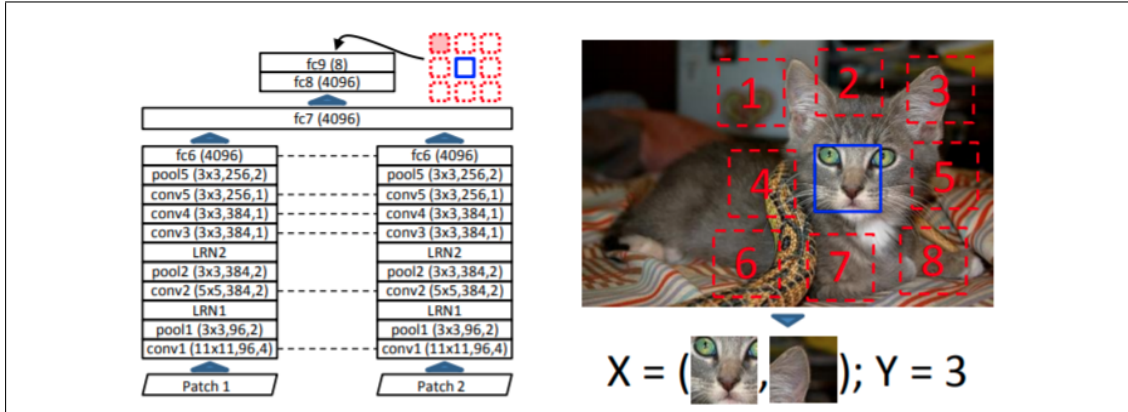
5.5 A critical analysis of self-supervision,or what we can learn from a single image

It is mentioned that initial layer features can be trained by just training on one image and its augmentations. We do not need a huge data set. They noted that even if we train on a single image(and its augmentations) we can get 2/3rd the accuracy meaning big data sets only provide small increments in self supervised learning and goes on to prove that augmentations are very important. in the end they discovered that one image is enough to train the first 3 layers and produces similar results but for deeper layer where more intricate features are recognised we need bigger data sets. Our main conclusion is that these methods succeed perfectly in capturing the simplest image statistics, but that for deeper layers a gap exist with strong supervision which is compensated only in limited manner by using large data sets. This novel finding motivates a renewed focus on the role of augmentations in self-supervised learning and critical rethinking of how to better leverage the available data

6 Some highly cited Self-supervised techniques

6.1 Rotation[6]

In this they propose to produce 4 copies of a single image by rotating it by 0°, 90°, 180°, 270° and let a single network predict the rotation which was applied—a 4-class classification task. Intuitively, a good model should learn to recognize canonical orientations of objects in natural images.

# Rotations	Rotations	CIFAR-10 Classification Accuracy
4	0°, 90°, 180°, 270°	**89.06**
8	0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°	88.51
2	0°, 180°	87.46
2	90°, 270°	85.52

6.2 Exemplar[10]

In this technique, every individual image corresponds to its own class, and multiple examples of it are generated by heavy random data augmentation such as translation, scaling, rotation, and contrast and color shifts. We use data augmentation mechanism from. proposes to use the triplet loss [40, 18] in order to scale this pretext task to a large number of images (hence, classes) present in the ImageNet dataset. The triplet loss avoids explicit class labels and, instead, encourages examples of the same image to have representations that are close in the Euclidean space while also being far from the representations of different images. Example representations are given by a 1000-dimensional logits layer.
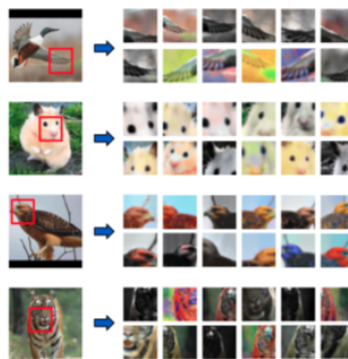

Figure 1: Examples of random transformations applied to the patches extracted from the unlabeled set of STL-10

7   CONCLUSION

In conclusion we can see that many techniques have been employed in order to make networks learn features of an image. Self-supervised learning isn't restricted to only images, similarly people are conducting research for videos, audio, 3D images, text and reinforcement learning too. The ultimate aim to find a perfect task that includes all the characteristics in the surrounding, and get appropriate outputs as we, as humans do in out daily life, and progress is being made in that direction everyday.

References

[1] Jean-Bastien Grill ,1 Florian Strub,1 Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya1, Carl Doersch1, Bernardo Avila Pires1, Zhaohan Daniel Guo1, Mohammad Gheshlaghi Azar1, Bilal Piot1, Koray Kavukcuoglu1, Rémi Munos1, Michal Valko1. Bootstrap Your Own Latent A New Approach to Self-Supervised Learning. In NeurIPS, 2020.

[2] Yuki M. Asano, Christian Rupprecht, Andrea Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. International Conference on Learning Representations (ICLR) 2020

[3] Yuki Markus Asano, Christian Rupprecht, Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. International Conference on Learning Representations (ICLR) 2020.

[4] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, Mario Lucic. On Mutual Information Maximization for Representation Learning. ICLR 2020.

[5] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In International Conference on Computer Vision (ICCV), 2015.

[6] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In International Conference on Learning Representations (ICLR), 2018.

[7] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In European Conference on Computer Vision (ECCV), 2016.

[8] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), 2014.

[9] Richard Zhang, Phillip Isola, Alexei A. Efros. Colorful Image Colorization. arxiv.com

[10] Eric Arazo, Noel E. O'Connor, Kevin McGuinness. Improving Unsupervised Learning With Exemplar CNNS. IMVIP, 2019

attach your review and visit log here......

temporary plagarism