Machine Learning

Group Project Report

Paper: A Simple Framework for Contrastive Learning of Visual Representations

	Paper Learning	Environment and coding	Report	PPT	Presentation
Wang Guojia 22048145G	√	√	√		
Zhou Yinuo 22043524G	√	√			√
Wu Ruonan 22043492G	√	√		√	

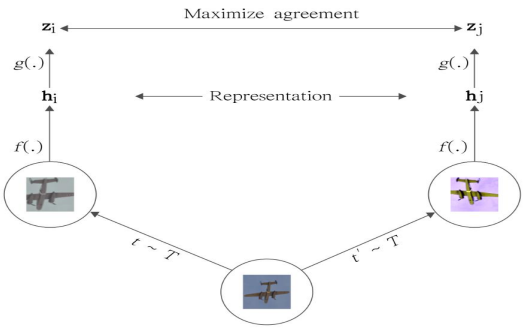
1. Introduction

SimCLR is a simple framework for contrastive learning of visual representations.

The basic idea: the more similar the same kind is, the better the performance is. This paper used cosine similarity to define the similarity function, and generally, the diagonal needs to be removed (The comparison of the same image itself should be removed, or the model deviation will become larger. The similarity of comparison of the same image itself must be the largest, and this will impact the model), and then design the loss function.

2. Methodology

SimCLR maximizes the consistency between different augmented views of the same data example for representation learning through potential spatial loss of contrast, and it consists of four main components.



1) Stochastic data augmentation module

2) Neural network base encoder (base encoder)

3) Projection head for small neural networks

4) Contrast loss function

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

The numerator considers similar classes and the denominator considers different classes.Similar to cross-entropy, this *t* is not described in detail in the original paper and

is a number less than 1. It is understood to be used to amplify the use of differences, which allows the model to converge faster.

3. Experiments,Results and Analysis

3.1Experimental programme

The cifar-10 dataset was used. It contains 10 different categories, each with a reasonable number of observations. Most importantly, it contains a larger set of 60,000 unlabelled images - the majority of the images are used for training.

For this implementation, ResNet-18 is used as the backbone of ConvNet. It takes images of shape (96,96,3) and regular STL-10 dimensions, and outputs a vector representation of size 512. The projection header gram has 2 fully connected layers. Each layer has 512 cells and produces a final 64-dimensional feature representation *z*.

The unlabelled portion of the training + dataset was used - a total of 65,000 images were provided.

After training, a method is used to evaluate the quality of the representation learned by SimCLR. One standard method is to use a linear evaluation protocol.

To reproduce the results of the paper, we replaced the original stlx-10 dataset into the cifar10 dataset for training, and tested three sets of parameters for the experiments and obtained the results.

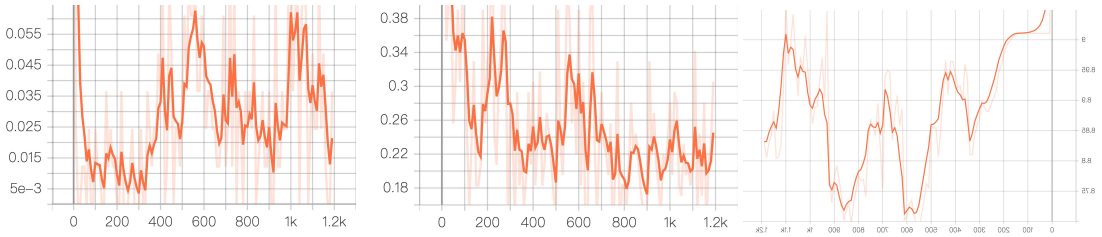
3.2Results and Analysis

NO.	Architecture	BatchSizes	Projection Head dimensionality	Epochs	Top1 %	Top5 %
1	ResNet-18	4096	128	100	27.22	75.06
2	ResNet-18	8192	512	100	53.9	93.90

3	ResNet-50	8192	128	100	64.40	96.72

Experiment NO.1:

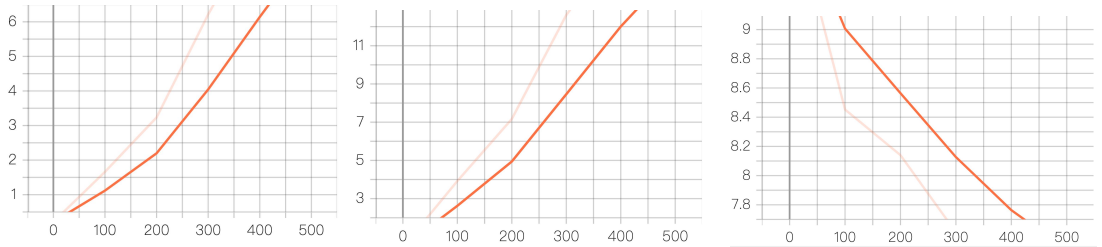
Top1 Top5 loss



As it shown above, the learning rate for this experiment was set too large. This resulted in an increasingly large loss and overfitting.

Experiment NO.2:

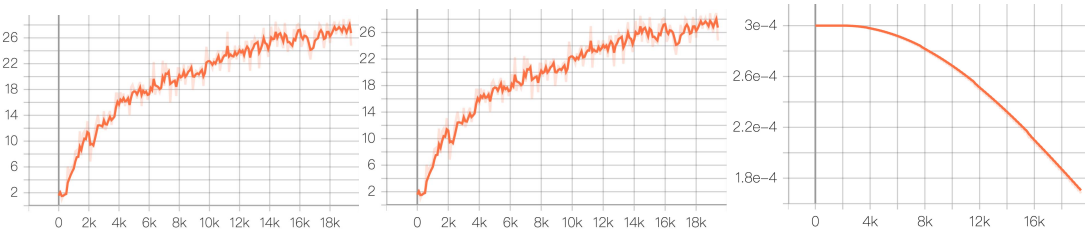
Top1 Top5 loss



Looking at the graphs of the experimental results, you can see that Top5 has a higher accuracy rate than Top1.

Experiment NO.3:

Top1 Top5 loss



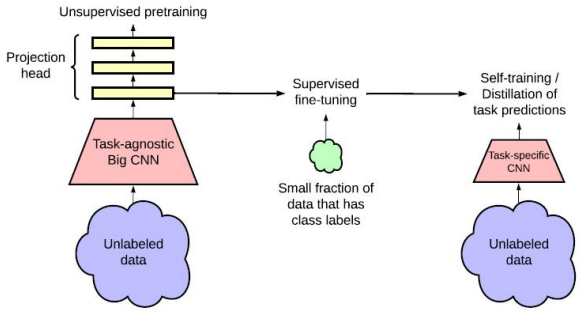
When using smaller training epochs, the performance of larger Batch sizes significantly outperformed that of smaller Batch sizes. The authors found that when larger training epochs were used, the performance of larger Batch sizes became closer to that of smaller Batch sizes.

4. Optimization

1) In this step of the Pre-train using unlabelled datasets, the size of the model is important and using a deep and wide model can help improve performance.

2) After the Pre-train with the unlabelled dataset, it is time to fine-tune with the labelled dataset. This deep and wide model is then distilled into a smaller network.

3) The deeper the projection head, the better the representation can be learned, and the better the fine-tune after the downstream task.



5. Conclusion

This article shows that as long as there are enough machines and large enough batch size, It is sufficient to treat all but the positive samples in each batch as negative samples.

The larger the batch size is (from 256 to 4096), the better the results are. But larger batch size requires better video memory.

The non-linear layer, or projection head, realizes better graph representation. One explanation is that this structure eliminates the effect of data augmentation, which means that features further ahead in the image representation contain more image augmentation information, which is needed for some downstream tasks. Finally, the features need to be normalized, and a suitable temperature value should be adopted, with 0.1 working best in the experiments.