EECS 6322 Week 8, Paper 1
A Simple Framework for Contrastive Learning of Visual Representations

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper presents a novel framework known as SimCLR for contrastive learning of visual representations. It proposes a few key steps required by the model in order to improve learning in the contrastive learning domain. The key steps include data augmentation, a non linear transformation between the representation and contrastive loss, and an increase in the batch size and training epochs. By combining these steps, the model is able to outperform current state of the art models with significantly fewer labels.

The SimCLR model works by maximizing the agreement of the different representations of the same data. The architecture includes a stochastic data augmentation module that transforms a given input data into two correlated pairs of the same input. For this model, the data augmentation performed include random color distortions, random Gaussian blur, and random cropping followed by resizing back to the original size. The architecture also uses a neural network base encoder that extracts the features and representational vectors from the augmented data. The framework uses a ResNet architecture for the model. The architecture also includes a neural network projection head that maps the representations from where the contrastive loss has been applied, which is followed by a contrastive loss function for the contrastive prediction tasks.

A random sample of N mini batches is input to the system, where given a positive pair, the other $2(N-1)$ augmented examples within a mini batch are treated as negative examples. Hence, the working of the algorithm is as follows. First the data is split into mini batches and for each minibatch an input data is processed. The input is first converted into two correlated pairs by data augmentation, and a pairwise similarity is computed between the two augmented pairs of the same input data. The contrastive loss function then computes the loss between the two pairs and aims to maximize similarity between different representations of the same data, and minimize the similarity between representations of different data.

The architecture does not use a memory bank and instead varies the training batch size N from 256 to 8192. A LARS optimizer is used to stabilize the training process for all batch sizes. The architecture is trained on Cloud TPUs ranging from 32 to 128 cores. The architecture also normalizes the Batch Norm mean and variance over all devices during training. The pre-training for the architecture is done on the ImageNet ILSVRC-2012 dataset and additional experiments on the CIFAR-10 dataset are conducted. A linear evaluation protocol is followed to evaluate the model where a linear classifier is trained on the frozen base network and the test accuracy is used to compute the representational power of the model. The architecture uses a 2-layer MLP projection head and a ResNet-50 base encoder.

A series of data augmentations have been performed and studied by the authors of the paper that include geometric or spatial transformation of data such as resizing , cutout, and cropping. The studies also include appearance transformation such as Gaussian blur, color distortion, and Sobel filtering. These different types of data augmentations were applied to the model and the results showed that no single transformation was sufficient for learning good representation of the data, even though the model fit perfectly to the data. This suggested that a combination of the data augmentations needed to be used in order to increase the learning ability of the model. Through the results, it was also concluded that random color distortion and random cropping were the most important data augmentation methods required to increase the learning for the generalizable features in the model. When compared to supervised learning, it was found that contrastive learning required stronger data augmentations for improving the representational learning processes. Finally, unsupervised learning was also proved to benefit from bigger models as compared to its supervised counterparts.

The projection head in the architecture was tested with three different methods, namely a linear transformation, a non-linear transformation, and no transformation. It was found that the non-linear transformation was best suited for the model. It was also noted that the non-linear activation layer benefitted the representational quality of the layer before it, instead of the layer following the activation layer. Due to the contrastive loss in the layer following the non-linear activation, more information can be maintained in the layer before the activation layer. An l_2 normalization in the NT-Xent loss function used by the model helps the architecture to handle the hard/ semi-hard negatives of the model. In contrast to supervised learning, the contrastive learning also benefits largely by an increase in the batch sizes and the training epochs.

The architecture is compared to the other state of the art results in the semi supervised learning and transfer learning

domain and is able to achieve substantially better results as compared to the previous domain designed architectures. For semi supervised learning, a 1% and 10% sample of the labelled ILSVRC-12 training dataset was taken and for transfer learning, the performance was evaluated across 12 natural image datasets in both fine-tune settings and linear evaluations. The SimCLR framework was able to achieve a better performance with fine tuning in 10 out of 12 datasets with very less number of labelled data as compared to the other models. Thus, it improves considerably over the other domain designed models.

II. Strengths

The paper proposes an architecture that performs considerably well over the other domain designed architectures through simple changes in the architecture such as data augmentation, increasing the batch sizes/ training epochs, and constructing a contrastive loss for the model. It increases the similarity between different representations of the same data and decreases the similarity across different data hence providing it with a simple yet effective model for contrastive learning.

III. Weaknesses

The pre-training in the framework only involves data augmentation which requires the user to select the augmentations and the extent of the augmentations. This might result in augmentations that are not best suited for the model and hence might lead to a poor performance of the model.

IV. Correctness

The claims and empirical methodology of the paper are correct and are supported by their results and comparison with the previous state of the art models. Simple techniques such as data augmentation have been applied to the model and different aspects such as the training epochs and batch sizes have been modified (here increased) to make the model more effective. An in depth comparison has also been provided in the appendix hence supplementing the correctness of the framework.

V. Clarity

The paper is well written in general with a clear formatting and adequate supplementary information such as graphs and comparative tables for the results. The theoretical grounding is also well formatted in the paper, where they are briefly explained along with the concepts, and are later elaborated in the appendix. Overall, the paper is well written and is easy to read.

VI. Relation to prior work

Most approaches for learning visual representation tasks without human supervision are either generative or discriminative in nature. The generative approaches learn to model or generate the pixels in the same input space. On the other hand, the discriminative approaches perform pretext tasks which includes taking the inputs and labels from an unlabelled dataset. The discriminative models then learn the representation by using objective functions similar to those used in supervised learning.

The SimCLR architecture combines the findings of the authors on an unlabelled dataset to perform constative learning. The findings include: data augmentation on unlabeled data lead to better performance, a nonlinear transformation between the contrastive loss and the representation improves the efficiency of the model, including a contrastive cross entropy loss normalizes the embeddings and improve performance, and large batch sizes and training epochs help the model learn better. These findings are then combined with the model which outperforms all the current state of the art models in terms of accuracy and performance.

VII. Reproducibility

The proposed architecture can be reproduced in common deep learning frameworks such as Pytorch. The source code details along with the pseudo code as given in the paper provides adequate information on how to implement the architecture and reproduce major details of the work and the results proposed.

VIII. Additional comments

A study on the applications of this framework on the different real world domains where the model would be best suited as compared to the current models used in the domain would supplement the work proposed by the authors.