A Simple Framework for Contrastive Learning of Visual Representations

Seri Lee
Computer Science and Engineering
Seoul National University
Seoul, Republic of Korea
sally20921@snu.ac.kr

***Abstract*—This paper presents SimCLR: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1)composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 85.8% top-5 accuracy, outperforming AlexNet with 100x fewer labels.**

I. Introduction

Learning effective visual representations without human supervision is a long-standing problem [1]. Most mainstream approaches fall into one of two classes: generative or discriminative. Generative approaches learn to generate or otherwise model pixels in the input space. However, pixel-level generation is computationally expensive and may not be necessary for representation learning. Discriminative approaches learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset. Many such approaches have relied on heuristics to design pretext tasks, which could limit the generality of the learned representations. Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-art results.

In this work, we introduce a simple framework for contrastive learning of visual representations, which we call SimCLR. Not only does SimCLR outperform previous work, but it is also simpler, requiring neither specialized architectures nor a memory bank.

In order to understand what enables good contrastive representation learning, we systematically study the major components of our framework and show that:

- Composition of multiple data augmentation operations is crucial in defining contrastive prediction tasks that yield effective representations. In addition, unsupervised contrastive learning benefits from stronger data augmentation than supervised learning.
- Introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations.
- Representation learning with contrastive cross entropy loss benefits from normalized embeddings and an appropriately adjusted temperature parameter.
- Contrastive learning benefits from large batch sizes and longer training compared to its supervised counterpart. Like supervised learning, contrastive learning benefits from deeper and wider networks.

We combine these findings to achieve a new state-of-the-art in self-supervised and semi-supervised learning on ImageNet ILSVRC-2012. Under the linear evaluation protocol, SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art. When fine-tuned with only 1% of the ImageNet labels, SimCLR achieves 85.8% top-5 accuracy, a relative improvement of 10%. When fine-tuned on natural image classification datasets, SimCLR performs on par with or better than a strong supervised baseline on 10 out of 12 datasets.

II. Method

A. The Contrastive Learning Framework

Inspired by recent contrastive learning algorithms, simCLR learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. This framework comprises the following four major components.

- A stochastic data augmentation module that transforms any given data example randomly resulting in two correlated views of the same example, denoted $\tilde{x}_i$ and $\tilde{x}_j$, which we consider as a positive pair. In this work, we sequentially apply three simple augmentations: random cropping followed by resize back to the original size, random color distortions, and random Gaussian blur. The combination of random crop and color distortion is crucial to achieve a good performance.

- A neural network base encoder $f(\cdot)$ that extracts representation vectors from augmented data examples. Our framework allows various choice of the network architecture without any constraints. We opt for simplicity and adopt the commonly used ResNet to obtain $h_i = f(\tilde{x}_i) = ResNet(\tilde{x}_i)$ where $h_i \in \mathbb{R}^d$ is the output after the average pooling layer.
- A small neural network projection head $g(\cdot)$ that maps representations to the space where contrastive loss is applied. We use a MLP with one hidden layer to obtain $z_i = g(h_i) = W^{(2)}\sigma(W^{(1)}h_i)$ where σ is a ReLU nonlinearity. We find it beneficial to define the contrastive loss on z_i's rather than h_i's.
- A contrastive loss function defined for a contrastive prediction task. Given a set $\{\tilde{x}_k\}$ including a positive pair of examples $\tilde{x}_i$ and $\tilde{x}_j$, the contrastive prediction task aims to identify $\tilde{x}_j$ in $\{\tilde{x}_k\}_{k\neq i}$ for a given $\tilde{x}_i$.

We randomly sample of minibatch of N examples and define the contrastive prediction task on pairs of augmented examples derived from the minibatch, resulting in $2N$ data points. We do not sample negative examples explicitly. Instead, given a positive pair, we treat the other $2(N-1)$ augmented examples within a minibatch as negative examples. Let $sim(u, v) = u^\top v / \|u\| \|v\|$ denote the cosine similarity between two vectors u and v. Then the loss function for a positive pair of examples (i, j) is defined as

$$l_{i,j} = -\log \frac{exp(sim((z_i, z_j)/\tau))}{\sum_{k=1}^{2N} \mathbb{1}_{k\neq i} exp(sim((z_i, z_k)/\tau))} \quad (1)$$

where $\mathbb{1}_{k\neq i} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$ and τ denotes a temperature parameter. The final loss is computed across all positive pairs, both (i, j) and (j, i), in a mini-batch. This loss has been used in previous work; for convenience, we term it *NT-Xent*(the normalized temperature-scaled cross entropy loss).

Algorithm 1 summarizes the proposed method.

Algorithm 1: SimCLR's main learning algorithm

Input: batch size N, constant τ, structure of f,g,T
for *sampled minibatch* $\{x_k\}_{k=1}^N$ **do**
  forall $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $\tau \sim T$, $\tau' \sim T$;
    `/* the first augmentation */`
    $\tilde{x}_{2k-1} = t(x_k)$;
    $h_{2k-1} = f(\tilde{x}_{2k-1})$;
    ; `/* representation */`
    $z_{2k-1} = g(h_{2k-1})$;
    ; `/* projection */`
    `/* the second augmentation */`
    $\tilde{x}_{2k} = t(x_k)$;
    $h_{2k} = f(\tilde{x}_{2k})$;
    ; `/* representation */`
    $z_{2k} = g(h_{2k})$;
    ; `/* projection */`
  end
  forall $i \in \{1, \ldots, 2N\}$ *and* $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = z_i^\top z_j / (\|z_i\| \|z_j\|)$;
    ; `/* pairwise similarity */`
  end
  $L = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)]$;
  update networks f and g to minimize L
end
return *encoder network* $f(\cdot)$*, and throw away* $g(\cdot)$

B. Training with Large Batch Size

We do not train the model with a memory bank. Instead, we vary the training batch size N from 256 to 8192. To stabilize training, we use the LARS optimizer for all batch sizes. We train our model with Cloud TPUs, using 32 to 128 cores depending on the batch size.

Global BN Standard ResNets use batch normalization. In distributed training with data parallelism, the BN mean and variance are typically aggregated locally per device. In our contrastive learning, as positive pairs are computed in the same device, the model can exploit the local information leakage to improve prediction accuracy without improving representations. We address this issue by aggregating BN mean and variance over all devices during the training. Other approaches include shuffling data examples, or replacing BN with layer norm.

C. Evaluation Protocol

Here we lay out the protocol for our empirical studies, which aim to understand different design choices in our framework.

Dataset and Metrics Most of our study for unsupervised pretraining (learning encoder network f without labels) is done using the ImageNet ILSVRC-2012 dataset. Some additional pretraining experiments on CIFAR-10 can be found in Appendix B.9. We also test the pretrained results on a wide range of datasets for transfer learning. To evaluate the learned representations, we follow the widely used linear protocol, where a linear classifier is trained on top of the frozen base network, and test accuracy is used as a proxy for representation quality. Beyond linear evaluation, we also compare against state-of-the-art on semi-supervised and transfer learning.

Default setting Unless otherwise specified, for data augmentation we use random crop and resize (with random flip), color distortions, and Gaussian blur. We use ResNet-50 as the base encoder network, and a 2-layer MLP projection head to project the representation to a 128-dimensional latent space. As the loss, we use NT-Xent, optimized using LARS with linear learning rate scaling and weight decay of 10^{-6}.

We train at batch size 4096 for 100 epochs. Furthermore, we use linear warmup for the first 10 epochs, and decay the learning rate with the cosine decay schedule without restarts.

D. Data Augmentation for Contrastive Representation Learning

Data augmentation defines predictive tasks

REFERENCES

[1] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2), 2020.