Exploring Simple Siamese Representation Learning

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

***Abstract*—Siamese networks have become a common structure in various recent models for unsupervised visual representation learning. These models maximize the similarity between two augmentations of one image, subject to cerrtain conditions for avoiding collapsing solutions. In this paper, we report surprising empirical results that simple Siamese networks can learn meaningful representations even using none of the following: (1) negative sample pairs, (2) large batches, (3) momentum encoders. Our experiments show that collapsing solutions do exist for the loss and structure, but a stop-gradient operation plays an essential role in preventing collapsing. We provide a hypothesis on the implication of stop-gradient, and further show proof-of-concept experiments verifying it. Our "SimSiam" method achieves competitive results on ImageNet and downstream tasks. We hope this simple baseline will motivate people to rethink the roles of Siamese architectures for unsupervised representation learning.**

I. Introduction

Recently there has been steady progress in self-supervised representation learning, with encouraging results on multiple visual tasks [**?**]. Despite various original motivations, these methods generally involve certain forms of Siamese networks. Siamese networks are weight-sharing neural networks applied on two or more inputs. They are natural tool s for comparing (including but not limited to contrasting) entities. Recent methods define the inputs as two augmentations of one image, and maximize the similarity subject to different conditions.

An undesired trivial solution to Siamese network is all outputs "collapsing" to a constant. There have been several general strategies for preventing Siamese networks from collapsing. Contrastive learning repulses different images (negative pairs) while attracting the same image's two views (positive pairs). The negative pairs preclude constant outputs from the solution space. Clustering is another way of avoiding constant output, and SwAV incorporates online clustering into Siamese networks. Beyond contrastive learning and clustering, BYOL relies only on positive pairs but it does not collapse in case a momentum encoder is used.

In this paper, we report that simple Siamese networks can work surprisingly well with none of the above strategies for preventing collapsing. Our model directly maximizes the similarity of one image's two views, using neither negative pairs nor a momentum encoder. It works with typical batch sizes and does not rely on large-batch training.

Thanks to the conceptual simplicity, SimSiam can serve as a hub that relates several existing methods. In a nutshell, our method can be thought of as "BYOL without the momentum encoder". Unlike BYOL but like SimCLR and SwAV, our method directly shares the weights between the two branches, so it can be thought of as "SimCLR without negative pairs" and "SwAV without online clustering". Interestingly, SimSiam is related to each nethod by removing one of its core components. Even so, SimSiam does not cause collapsing and can perform competitively.

We empirically show that collapsing solutions do exist, but a stop-gradient operation is critical to prevent such solutions. The importance of stop-gradient suggests that thereshould be a different underlying optimization problem that is being solved. We hypothesize taht there are implicitly two sets of variables, and SimSiam behaves like alternating between optimzing each set. We provide proof-of -concept experiments to verify this hypothesis.

Our simple baseline suggests that the Siamese architectures can be essential reason for the common success of the related methods. Siamese network can naturally introduce inductive biases for modeling invariance, as by definition "invariance " means that two observations of the same concept should produce the same outputs. Analogous to convolutions, which is a successful inductive bias via weight-sharing for modeling tanslation-invariance, the weight-sharing Siamese networks can model invariance w.r.t more complicated transformations (e.g., augmentations). We hope that our exploration will motivate people to rethink the fundamental roles of Siamese architectures for unsupervised representation learning.

II. Related Works

Siamese networks Siamese networks are general models for comparing entities. Their applications include signature and face verification, tracking, one-shot learning, and others. In conventional use cases, the inputs to Siamese networks are from different images, and the comparability is determined by supervision.

Contrastive learning The core idea of contrastive learning is to attract the positive sample pairs and repulse the negative sample pairs. This methodology has been recently popularized for unsupervised representation learning. Simple and effective instantiations of contrastive learning have been developed using Siamese networks.

In practice, contrastive learning methods benefit from a large number of negative samples. These samples can be amintained in a memory bank. In a Siamese network, MoCo maintains a queue of negative samples and turns one branch

into a momentum encoder to improve consistency of the queue. SimCLR directly uses negative samples coexisting in the current batch, and it requires a large batch to work well.

Clustering Another category of methods for unsupervised representation learning are based on clustering. They alternate between clustering representations and learning to predict the cluster assignment. SwAV incorporates clustering into a Siamese network, by computing the assignments from one view and predicting it from another view. SwAV perfroms online clustering under a balanced partition constraintfor each batch, which is solved by the Sinkhorn-Knopp transform.

While clustering-based methods do not define negative exemplars, the cluster centers can play as negative prototypes. Like contrastive learning, require either a memory bank, large batches, or a queue to provide enough samples for clustering.

BYOL BYOL directly predicts the output of one view from another view. It is a Siamese network in which one branch is a momentum encoder. It is hypothesized in that the momentum encoder is important for BYOL to avoid collapsing, and it reports failure results if removing the momentum encoder. Our empirical study challenges the necessity of the momentum encoder for preventing collapsing. We discover that the stop-gradient operation is critical. This discovery can be obscured with the usage of a momentum encoder, which is always accompanied with stop-gradient. While the moving-average behavior may improve accuracy with an appropriate momentum coefficient, our experiments show that it is not directly related to preventing collapsing.