# 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,