SUMMARY

SimCLR v1

*A Simple Framework for Contrastive Learning of Visual Representations*

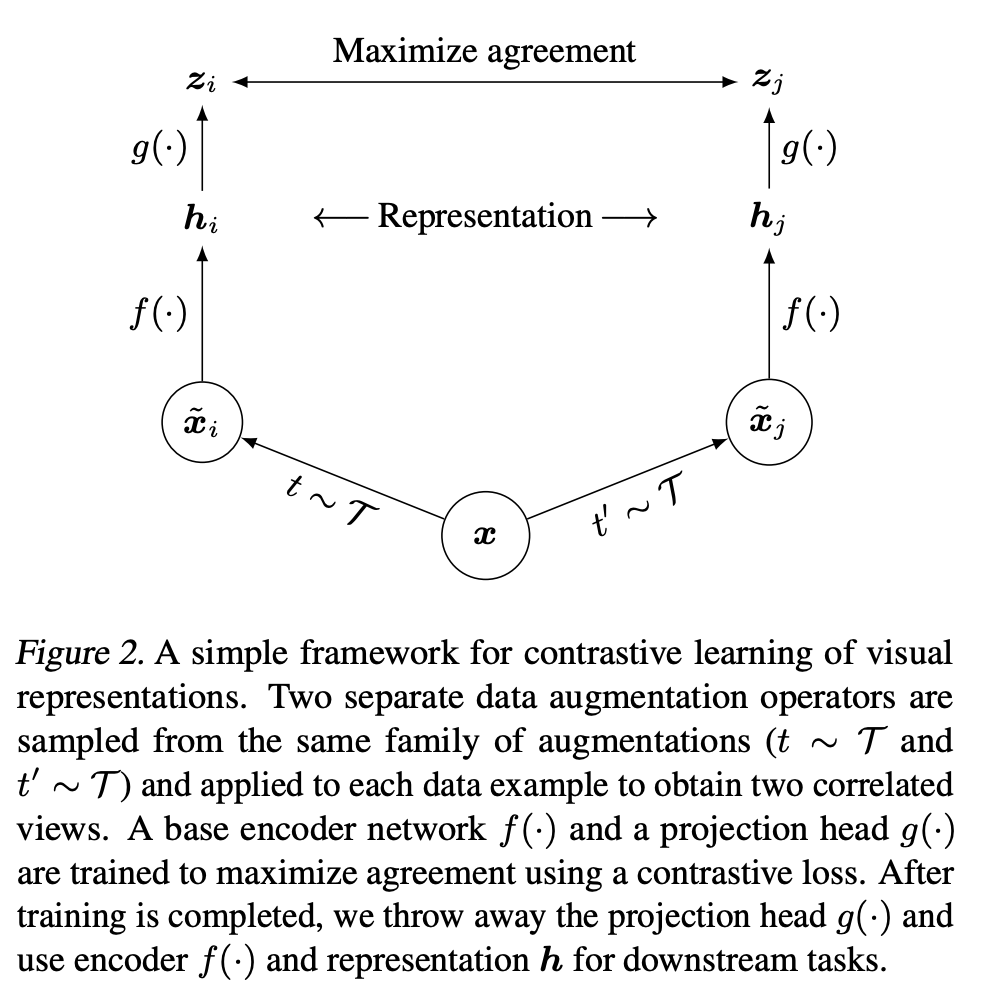
### INTRODUCTION

The paper shows three major findings-

* Composition of data augmentations plays a critical role in defining effective predictive tasks.
* Introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations.
* Contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning.

By combining these, it introduces a simple framework for contrastive learning of visual representations named SimCLR. Not only does SimCLR outperform previous work, but it is also simpler, requiring neither specialized architectures nor a memory bank.

### METHOD



The above figure shows the proposed framework.

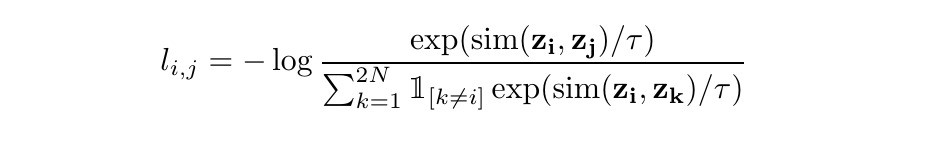
This framework comprises the following components:

* A stochastic data augmentation module, where we sequentially apply three simple augmentations: random cropping followed by resize back to the original size, random color distortions, and random Gaussian blur.
* A neural network base encoder f(·) that extracts representation vectors from augmented data examples. Resnet is used in the paper.
* A small neural network projection head g(·) that maps representations to the space where contrastive loss is applied. Here, they propose a MLP with one hidden layer to obtain

*zi = g(hi) = W(2)σ(W(1)hi)*

where σ is a ReLU nonlinearity.

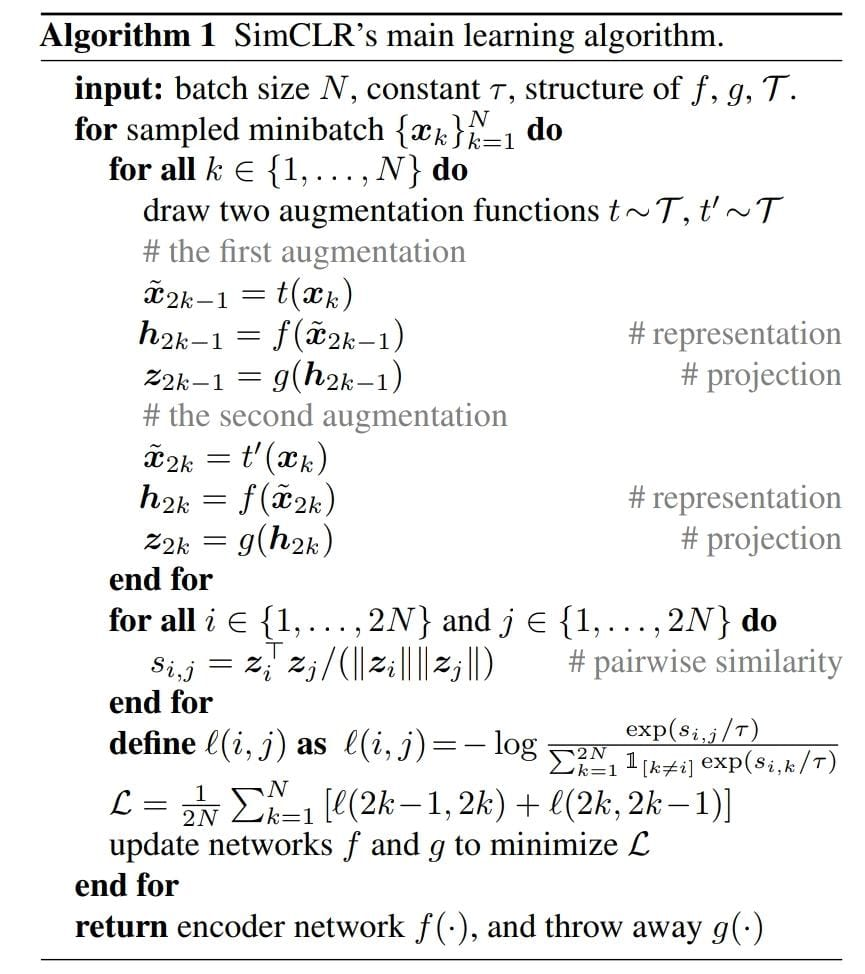
* A contrastive loss function.



Where sim(zi, zj) refers to the cosine similarity.

Negative examples are not sampled explicitly. Instead, given a positive pair,we treat the other 2(N−1) augmented examples within a mini batch as negative examples.

The paper proceeds to provide the algorithm of implementation of the above methods.



The paper further proves aforementioned claims with experimentation results showing that data augmentation is largely beneficial for self-supervised learning and large batch sizes help the proposed framework in learning.

SimCLR v2

*Big Self-Supervised Models are Strong Semi-Supervised Learners*

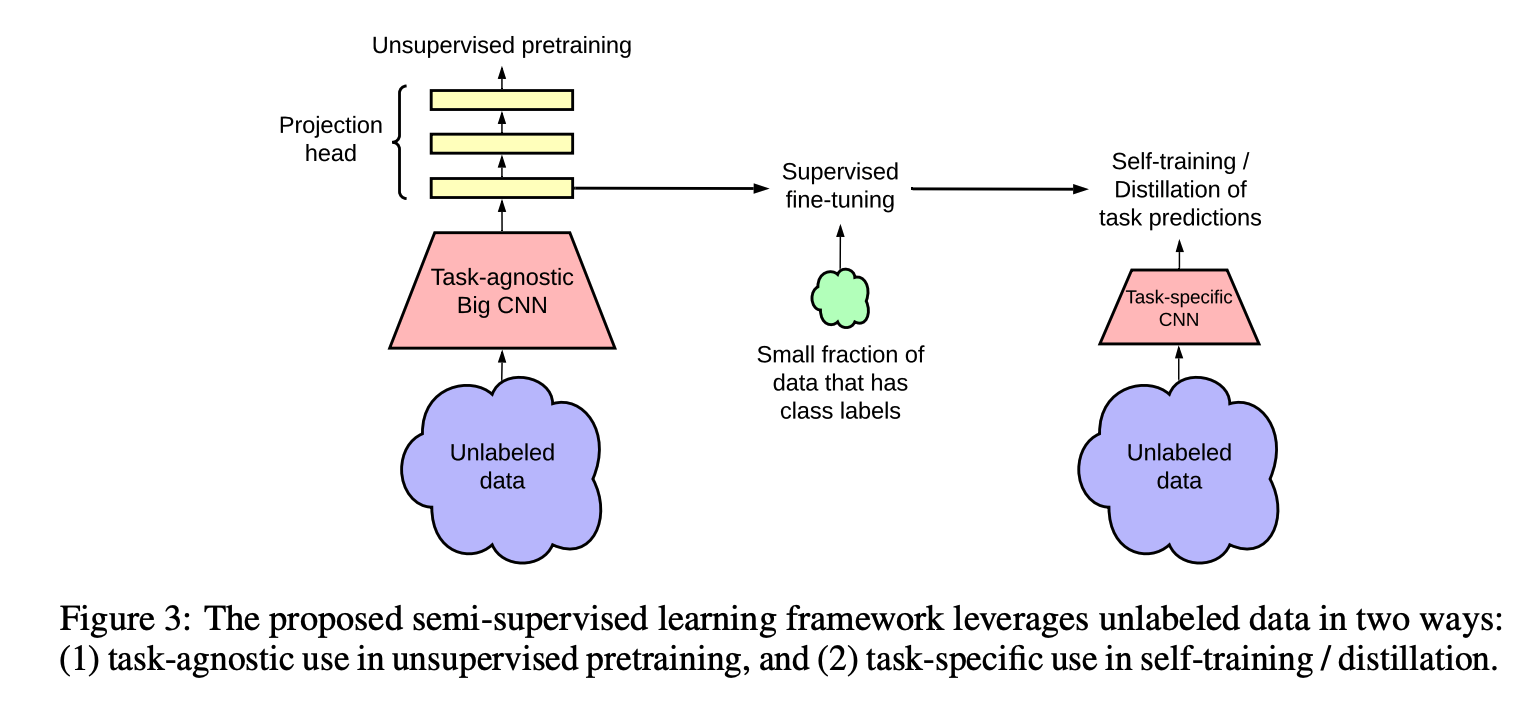
### INTRODUCTION

This paper introduces another paradigm for learning with a few labels, adding unsupervised task-agnostic pretraining and supervised fine-tuning as modifications. One of the main objectives is also to prove that larger models lead to better training in a semi-supervised setting.

Breaking the proposition into three steps: (1) unsupervised or self-supervised pretraining, (2) supervised fine-tuning, and (3) distillation using unlabeled data, the paper includes the following-

* Empirical results suggesting that for semi-supervised learning (via the task-agnostic use of unlabeled data), the fewer the labels, the more it is possible to benefit from a bigger model.
* Showcasing that although big models are important for learning general (visual) representations, the extra capacity may not be necessary when a specific target task is concerned. Therefore, with the task-specific use of unlabeled data, the predictive performance of the model can be further improved and transferred into a smaller network.
* Demonstration of the importance of the nonlinear transformation (a.k.a. projection head) after convolutional layers used in SimCLR for semi-supervised learning. A deeper projection head not only improves the representation quality measured by linear evaluation, but also improves semi-supervised performance when fine-tuning from a middle layer of the projection head.

### METHOD



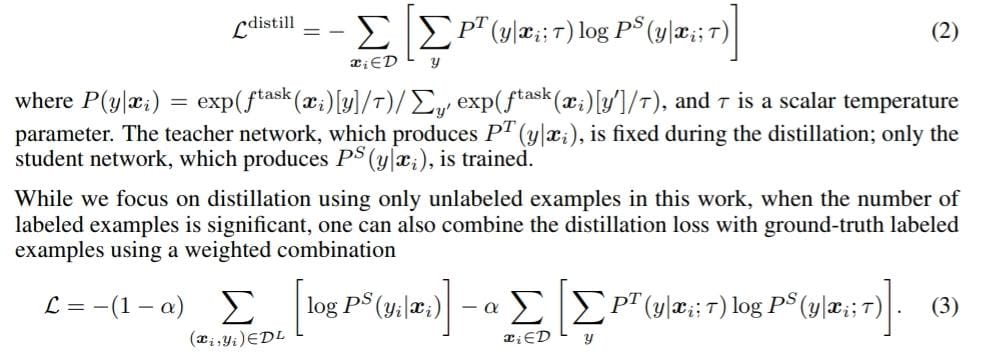
The first time the unlabeled data is used, it is in a task-agnostic way, for learning general (visual) representations via unsupervised pretraining. The general representations are then adapted for a specific task via supervised fine-tuning. The second time the unlabeled data is used, it is in a task-specific way, for further improving predictive performance and obtaining a compact mode.

Improving on the contrastive loss of SimCLR(1), the paper proposes the following modifications-

* To fully leverage the power of general pretraining, it includes an exploration of larger ResNet models.
* An increase in the capacity of the non-linear network g(·) (a.k.a. projection head), by making it deeper. Furthermore, instead of throwing away g(·) entirely after pretraining as in SimCLR [1], they propose a fine-tune from a middle layer.
* Incorporation of the memory mechanism from MoCo.

#### Self-training / knowledge distillation via unlabeled example

Use of the fine-tuned network as a teacher to impute labels for training a student network. Specifically, we minimize the following distillation loss where no real labels are used:



#### Distillation

For distillation, only unlabeled examples are used, unless otherwise specified. We consider two types of distillation: self-distillation where the student has the same model architecture as the teacher (excluding projection head), and big-to-small distillation where the student is a much smaller network.

Distillation with unlabeled examples improves fine-tuned models in two ways, as shown in Figure 6: (1) when the student model has a smaller architecture than the teacher model, it improves the model efficiency by transferring task-specific knowledge to a student model, (2) even when the student model has the same architecture as the teacher model (excluding the projection head after ResNet encoder), self-distillation can still meaningfully improve the semi-supervised learning performance.