SimCLR
Simple Contrastive Learning for Visual Representations

“Learn powerful visual features without labels “

https://arxiv.org/pdf/2002.05709

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton — Google Research, 2020

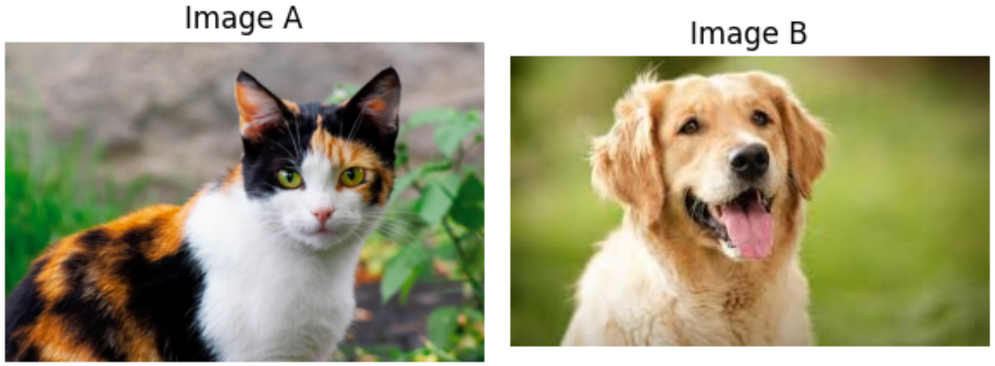
Presented By
Nyein Chan Soe

WHAT is Contrastive Learning

Which things are similar and which are different by showing pairs

Pulls similar items closer together in embedding space

Push dissimilar items apart in embedding space

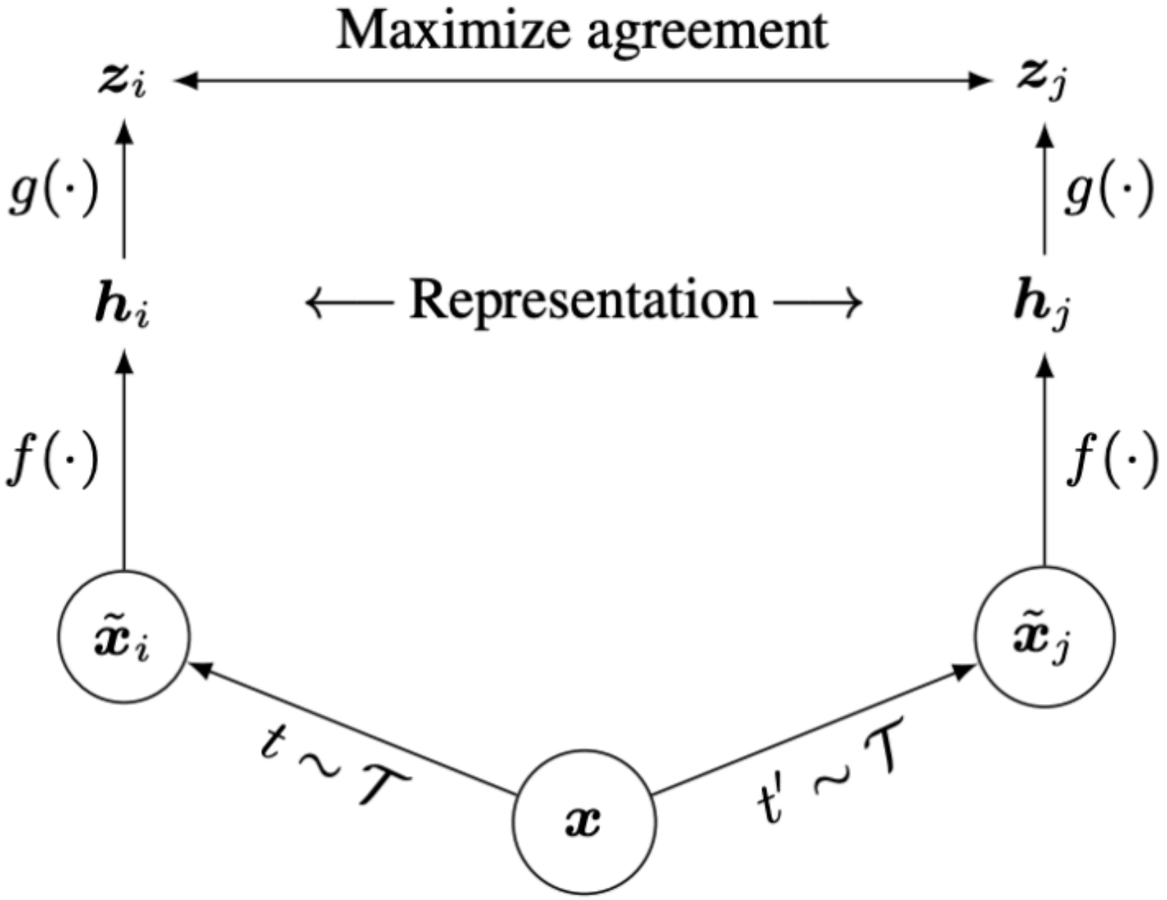


Why?

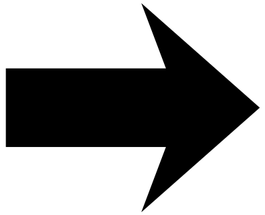
Labels are expensive

SimCLR Framework Blueprint

1. Augmentations
2. Base Encoder
3. Projection Head
4. Contrastive Loss



Augmentations



+

-

Base Encoder

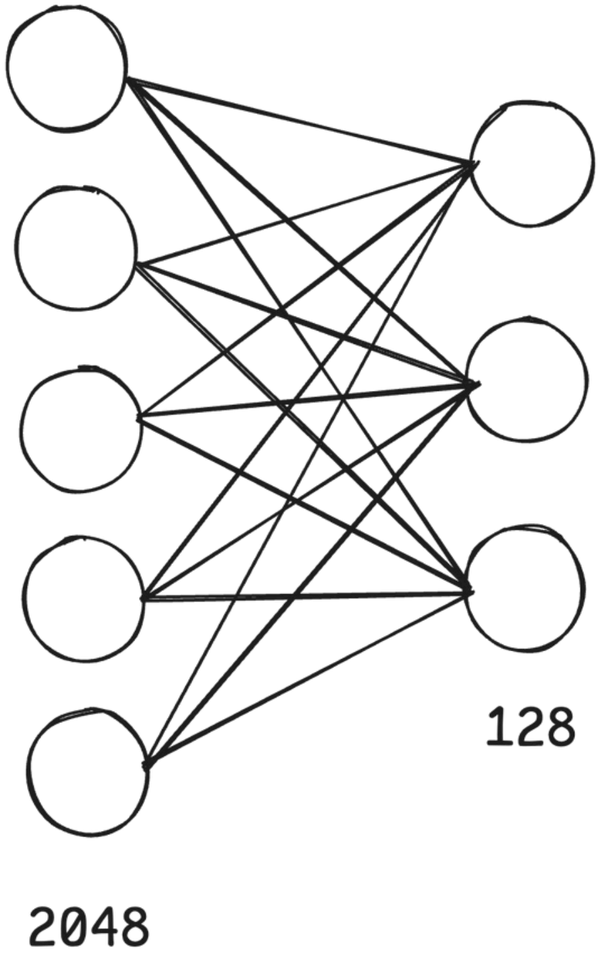
ResNet50

- A neural network *base encoder* $f(\cdot)$ that extracts representation vectors from augmented data examples. Our framework allows various choices of the network architecture without any constraints. We opt for simplicity and adopt the commonly used ResNet (He et al., 2016)

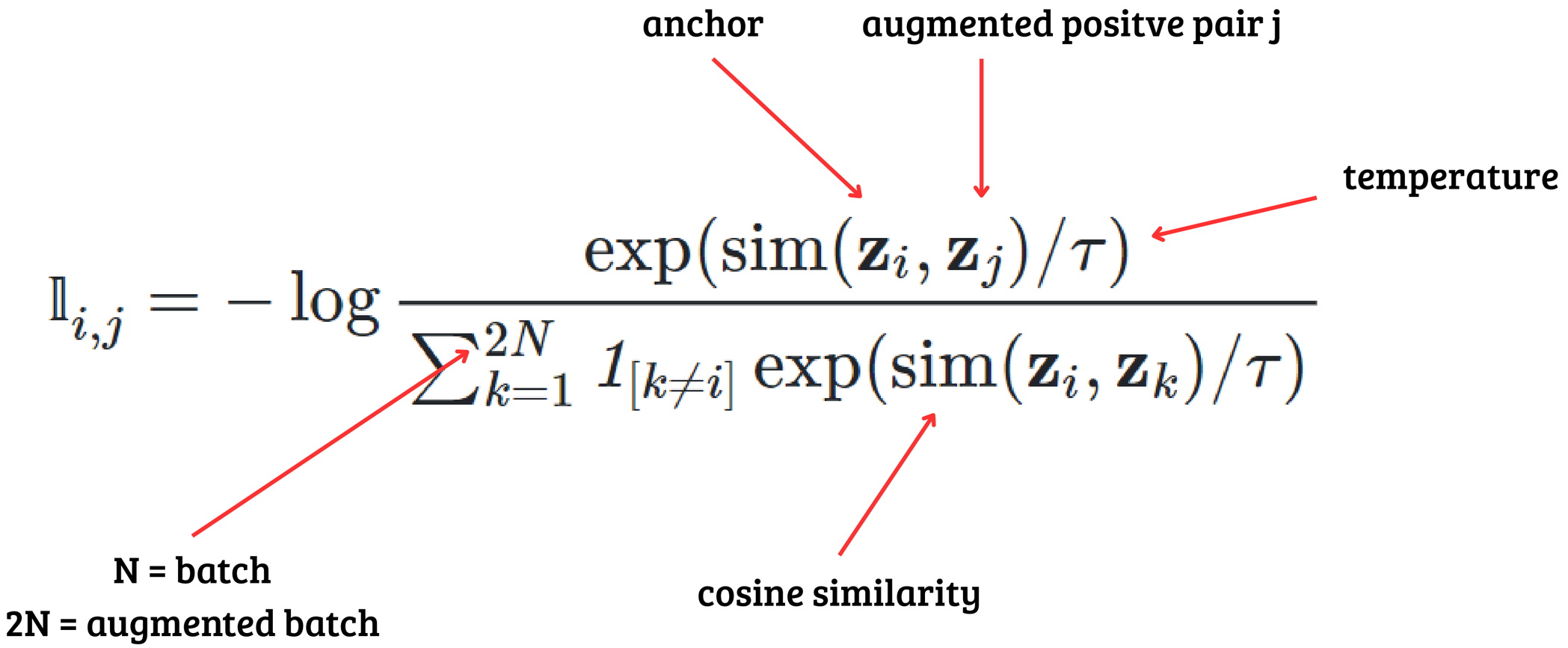
Projection Head

2 Layer MLP

Default setting. Unless otherwise specified, for data augmentation we use random crop and resize (with random flip), color distortions, and Gaussian blur (for details, see Appendix A). 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



Contrastive Loss (NT-Xent)

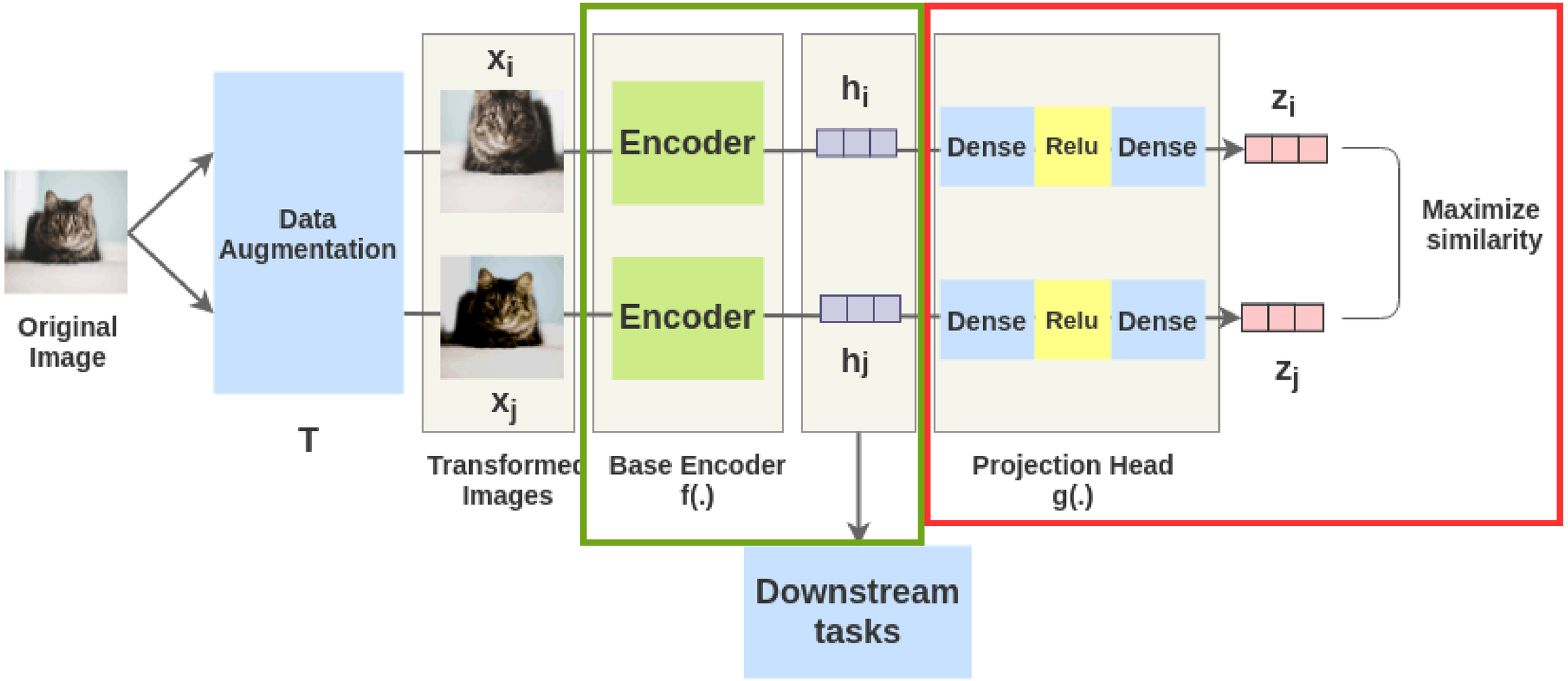


NT-Xent Goal

- Loss = 0 when fraction = 1
- Softmax Competition: $sim(i, j) \gg sim(i, k)$ for all negative pairs where k != j

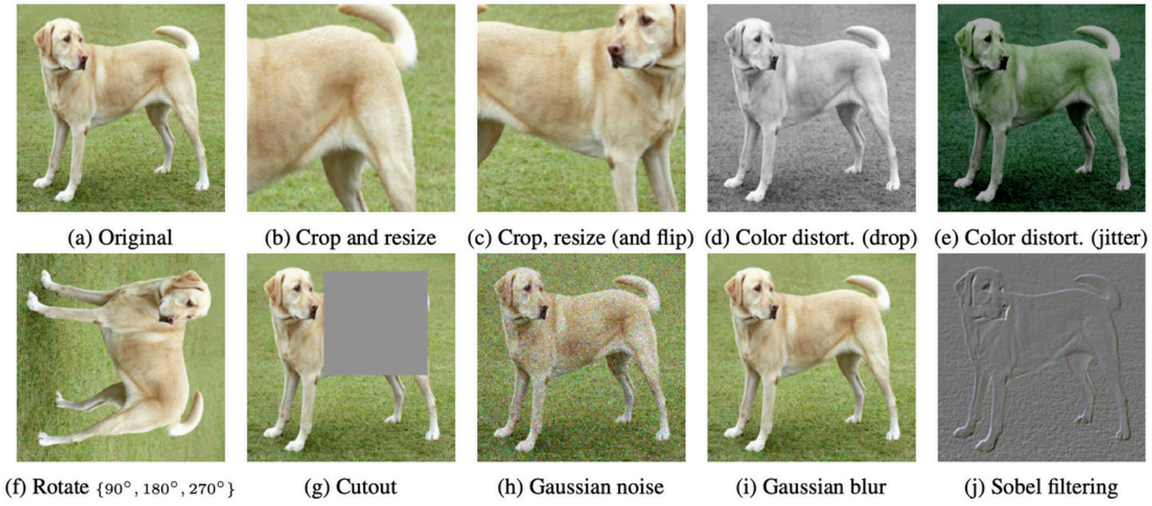
- Temperature is a hyperparameter (T = 0.5)
- Low T (0.1) makes model confident and representation can collapse
- High T (1) gives softer distinction between representations

SimCLR Arch

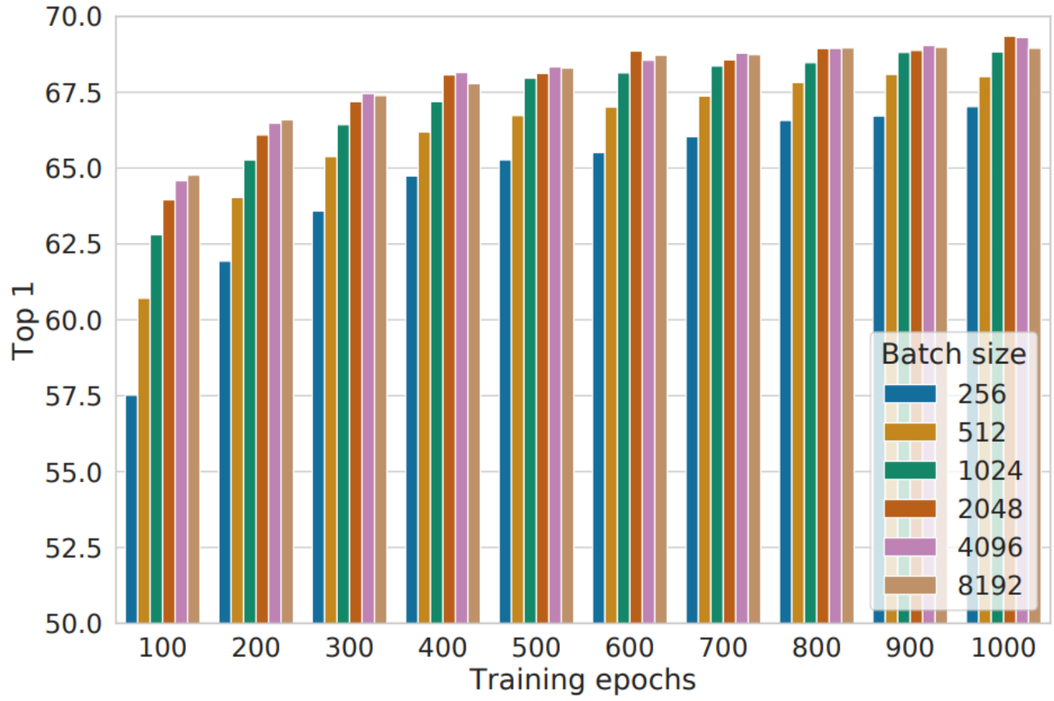

Original Image
Data Augmentation
T
x_i
x_j
Transformed Images
Encoder
Encoder
Base Encoder f(.)
h_i
h_j
Dense
Relu
Dense
Dense
Relu
Dense
Projection Head g(.)
z_i
z_j
Maximize similarity
Downstream tasks

Some Emperical Findings From Authors

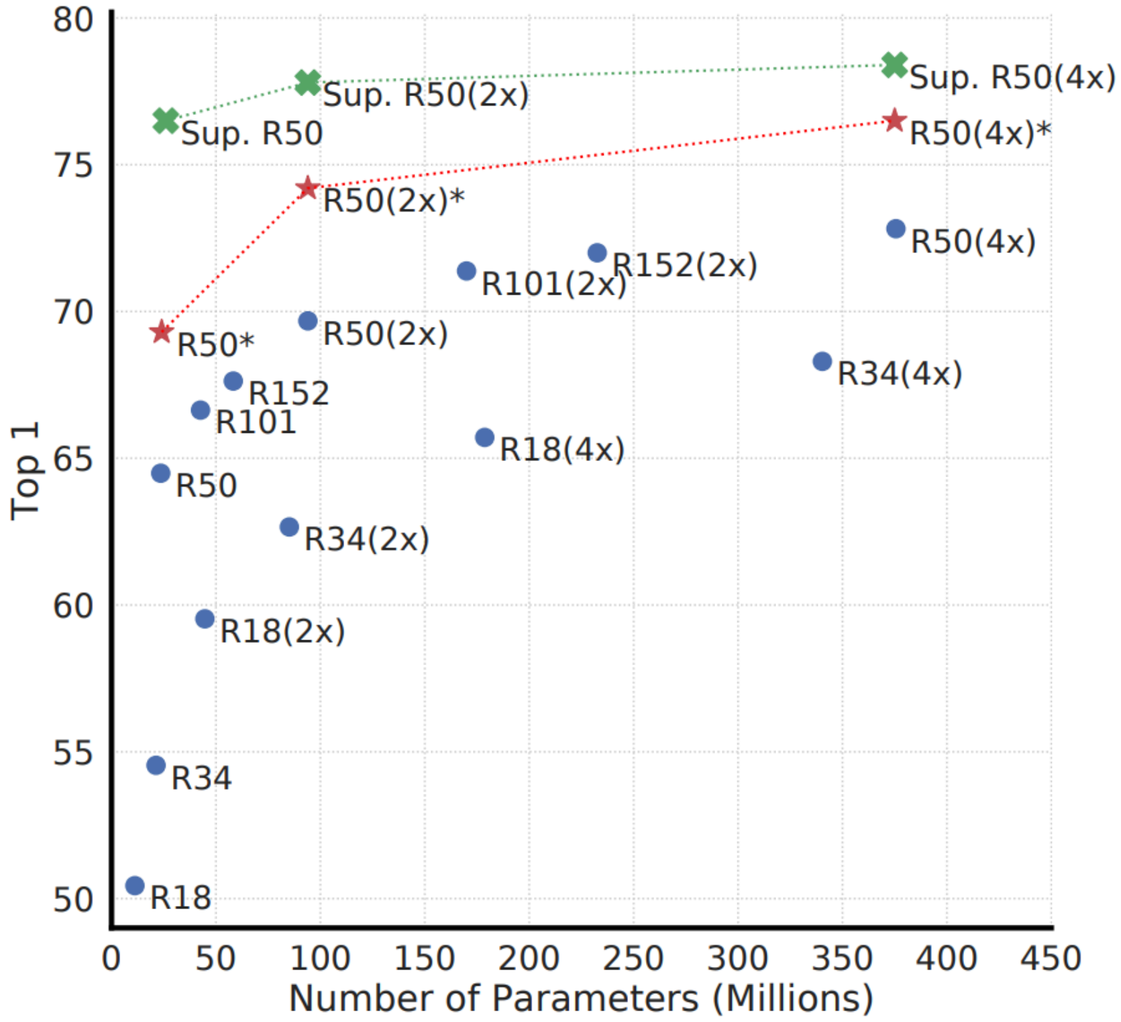
1. Strength of Augmentations Matters alot **(composition of crop + color jitter + blur)**
2. Projection Head improves representation quality **(encoder)**
3. Large batch size better learning **(batch size of 4096 for 100 epochs)**



igure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal arameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our iodels* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)



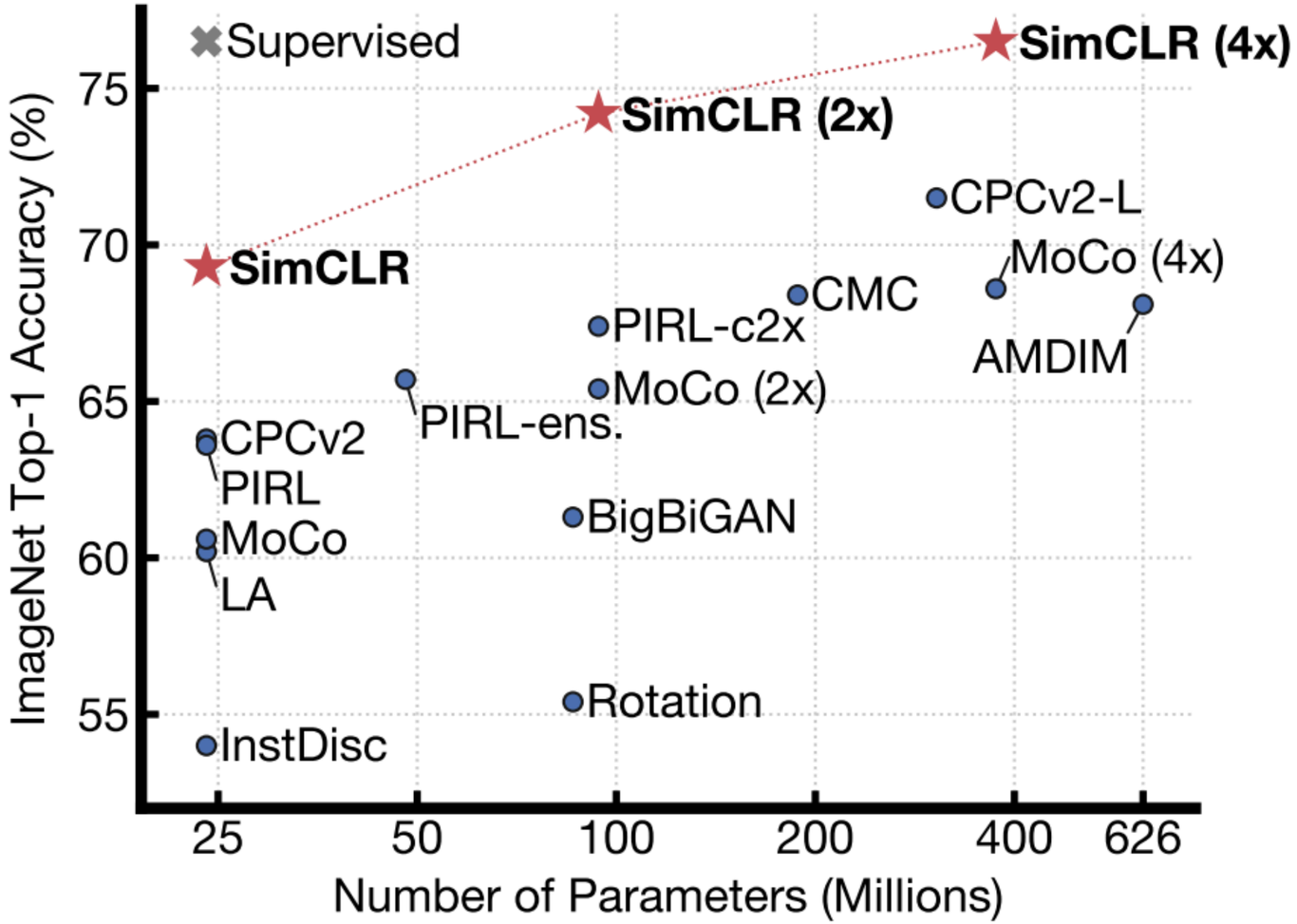
Linear Evaluations



Bigger models consistently achieve better accuracy, with wider models (4x) outperforming deeper models (R152). At the largest scale, SimCLR nearly matches supervised learning

Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs

SimCLR vs The Competitions



SimCLR dominates across all model sizes. At ~96m parameters, SimCLR achieves 74% - better than MoCo with 4× more parameters! This is the first self-supervised method to match supervised learning.

Conclusions

- **Self-supervision through augmentations**

- **Matches supervised learning performance**

  Achieves 76.5% Top-1 accuracy on ImageNet with just a linear classifier.

- **Scalability is key**

  Larger and wider models consistently improve performance, with the contrastive learning framework effectively leveraging additional model capacity

- **Practical impact**

  Enables pre-training on unlabeled data then fine-tuning on small labeled datasets, making deep learning accessible when labels are expensive or scarce

Thank you !