A Simple Framework for Contrastive Learning of Visual Representation

Ting Chen, Simon Kornlith, Mohammad Norouzi, Geoffrey Hinton

Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020.

경영과학연구실 김윤석

Introduction

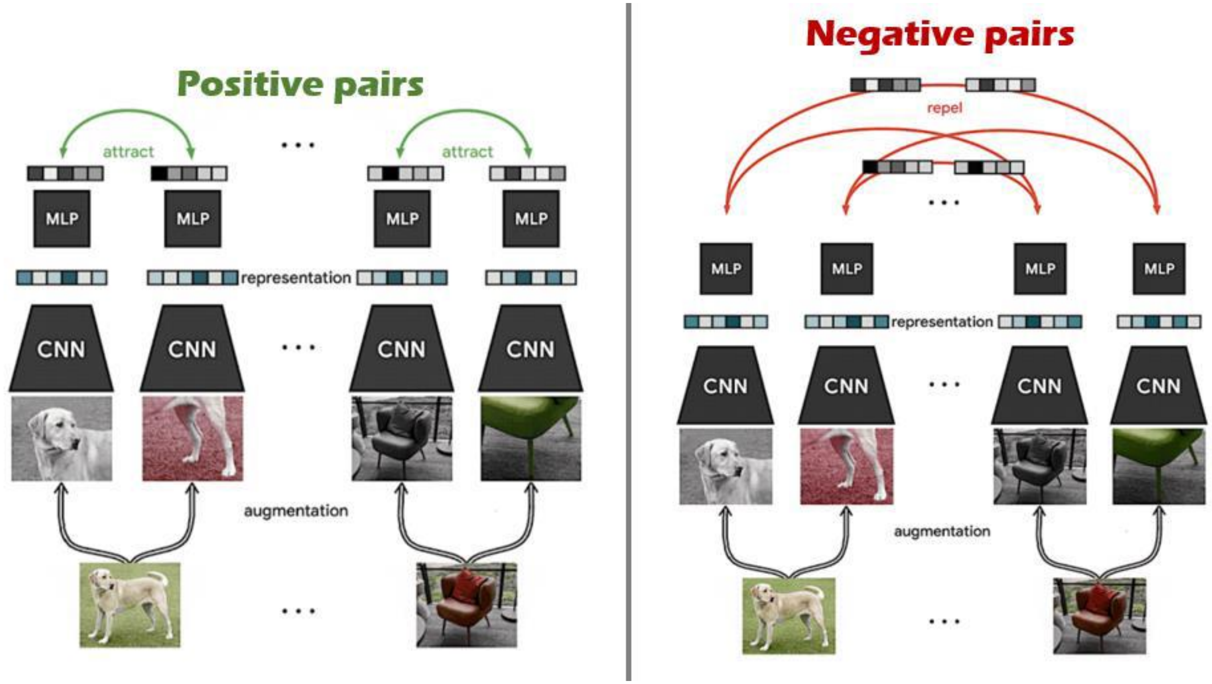
❖Flow of Image Classification Task

Labeled dataset

Unlabeled dataset

Framework of Contrastive Learning

❖Framework



➢ Loss function (Margin Triplet loss)

- y: class, x: image data, θ: neural network parameter, ϵ: margin

$$\mathcal{L}_{\text{cont}}(\mathbf{x}_i, \mathbf{x}_j, \theta) = \mathbb{1}[y_i = y_j] \| f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j) \|_2^2 + \mathbb{1}[y_i \neq y_j] \max(0, \epsilon - \| f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j) \|_2^2)$$

Problem statement & key idea

❖Problem statement

- They want to simplify the recently proposed contrastive self-supervised learning algorithm without requiring special architectures or memory banks.

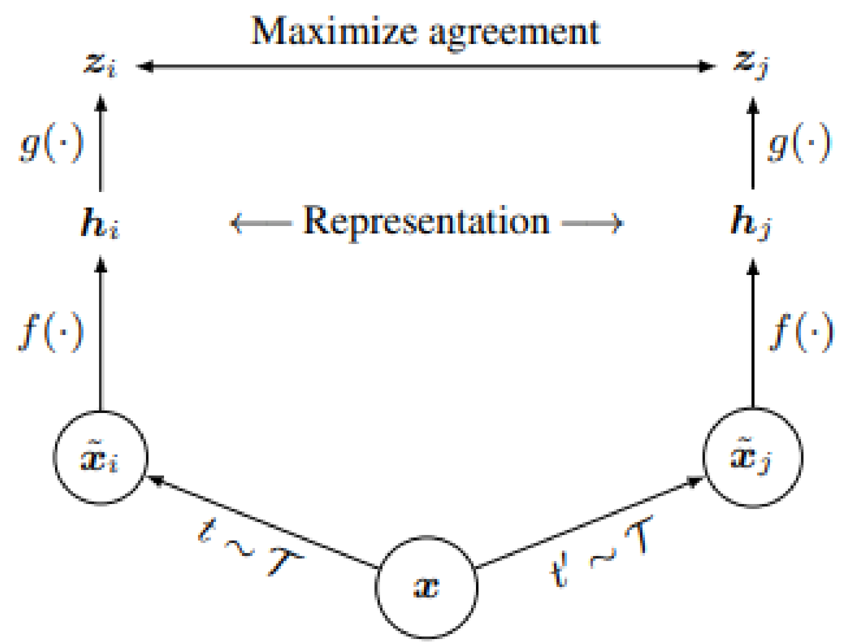
❖Key idea

- Increase the batch size to do a lot of contrast training without memory banks
- Finding the best augmentation combination experimentally

Method

❖ Framework

- $g(\cdot)$: projection head (Multi Layer perceptron)
- $f(\cdot)$: Encoder head (Resnet)
- $t \sim \mathcal{T}, t' \sim \mathcal{T}$: Augmentation function



Larger Batch Size

❖Algorithm

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
  for all k ∈ {1, ..., N} do
    draw two augmentation functions t∼T, t'∼T
    # the first augmentation
    x̃_{2k-1} = t(x_k)
    h_{2k-1} = f(x̃_{2k-1})                    # representation
    z_{2k-1} = g(h_{2k-1})                     # projection
    # the second augmentation
    x̃_{2k} = t'(x_k)
    h_{2k} = f(x̃_{2k})                        # representation
    z_{2k} = g(h_{2k})                         # projection
  end for
  for all i ∈ {1, ..., 2N} and j ∈ {1, ..., 2N} do
    s_{i,j} = z_i^T z_j / (‖z_i‖‖z_j‖)          # pairwise similarity
  end for
  define ℓ(i,j) as ℓ(i,j) = −log exp(s_{i,j}/τ) / Σ_{k=1}^{2N} 1_[k≠i] exp(s_{i,k}/τ)
  L = 1/(2N) Σ_{k=1}^N [ℓ(2k−1, 2k) + ℓ(2k, 2k−1)]
  update networks f and g to minimize L
end for
return encoder network f(·), and throw away g(·)
```

- Do not use memory bank by increasing batch size
- Loss function for positive pair:

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}[\ell(2k-1,2k)+\ell(2k,2k-1)]$$

Experiment

❖ Comparison of results for different experiments

- The authors demonstrate the advantages of SimCLR through an experimental method.

  1. Experiments on data augmentation

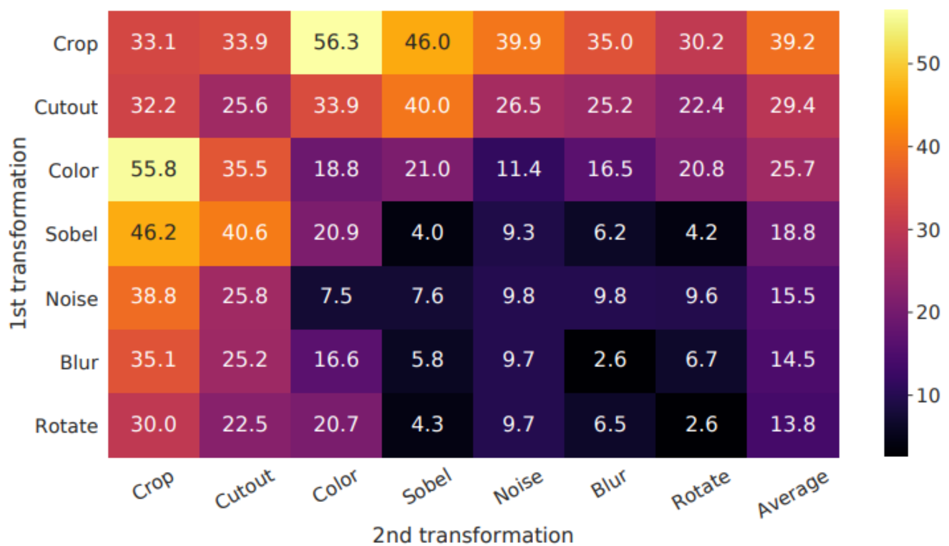
  2. Experiments on projection head configuration

  3. Experiments on batch size

Experiments on data augmentation

❖ Composition of data augmentation operations is crucial for learning good representations

- Comparing the performance of different configurations for two phases of augmentation



Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentation.

Experiments on data augmentation

❖ Contrastive learning needs stronger data augmentation than supervised learning

- Experiments show that unsupervised contrastive learning benefits from stronger (color) data augmentation than supervised learning

Methods	Color distortion strength 1/8	1/4	1/2	1	1 (+Blur)	AutoAug
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
Supervised	77.0	76.7	76.5	75.7	75.4	77.1

Top-1 accuracy of unsupervised contrastive learning and supervised learning using linear evaluation , under varied color distortion strength and other data transformations

Experiments on projection head configuration

❖ A nonlinear projection head improves the representation quality of the layer before it

- Nonlinear projection is better than a linear projection (+3%), and much better than no projection (>10%)

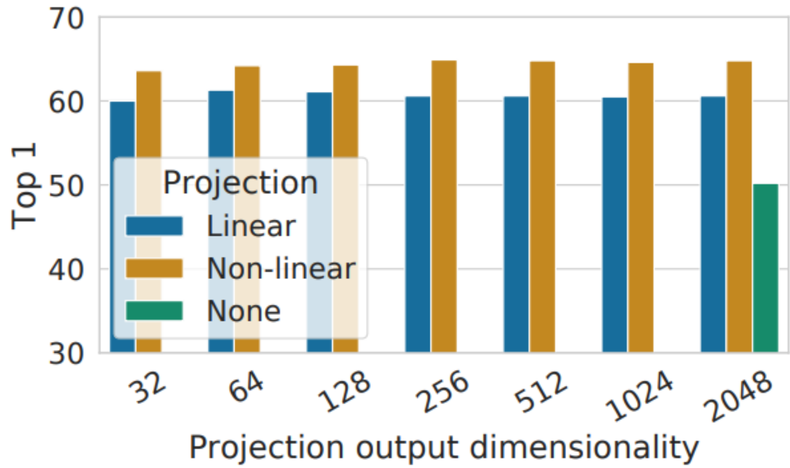
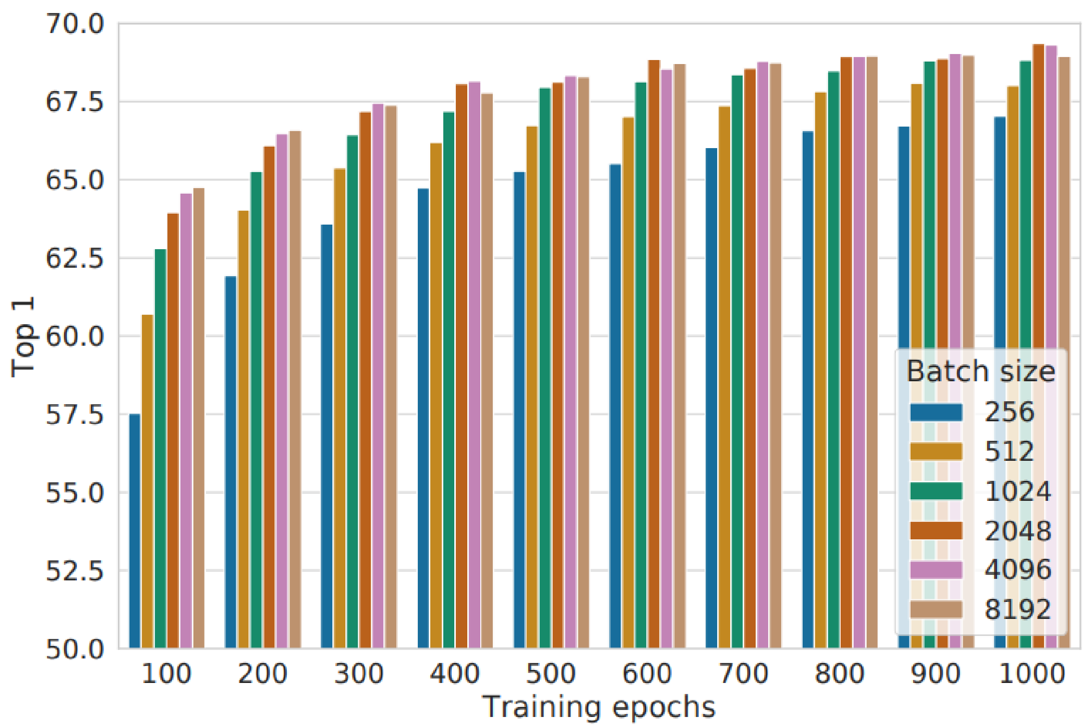


Figure 8. Linear evaluation of representations with different projection heads $g(\cdot)$ and various dimensions of $\boldsymbol{z} = g(\boldsymbol{h})$. The representation $\boldsymbol{h}$ (before projection) is 2048-dimensional here.

Experiments on batch size

❖ Contrastive learning benefits (more) from larger batch sizes

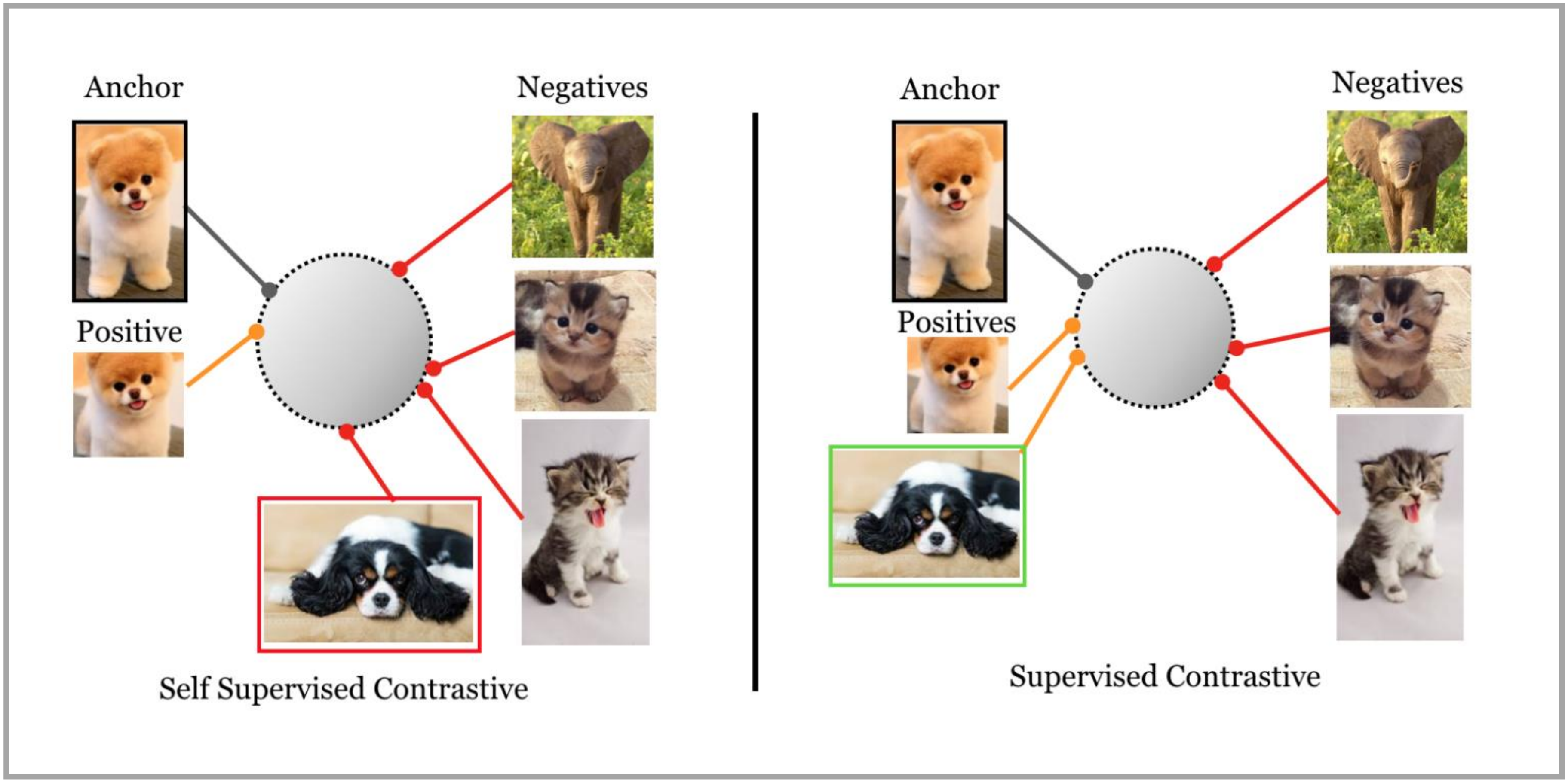
- shows the impact of batch size when models are trained for different numbers of epochs.



Supervised contrastive learning

❖ Issue of SimCLR

- Images of the same class can also be composed of negative pairs



Problem statement & key idea

❖ Problem statement

- They want to train the same class images as positive pairs by using label information.

❖ Key idea

- propose a loss for supervised learning that builds on the contrastive self-supervised literature by leveraging label information

Supervised Contrastive Losses

❖Notation

$\{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1...N}$: Set of N randomly sampled sample/label pairs

$\{\tilde{\boldsymbol{x}}_\ell, \tilde{\boldsymbol{y}}_\ell\}_{\ell=1...2N}$: Set of two random augmentation of $\{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1...N}$

$\tilde{\boldsymbol{x}}_{2k}$ and $\tilde{\boldsymbol{x}}_{2k-1}$ are two random augmentation of $\boldsymbol{x}_k$

$i \in I \equiv \{1...2N\}$: The index of an arbitrary augmented sample

$j(i)$: The the index of the other augmented sample originating from the same source sample

$A(i) \equiv I \setminus \{i\}$

$P(i) \equiv \{p \in A(i) : \tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i\}$

$$\mathcal{L}^{sup}_{out} = \sum_{i \in I} \mathcal{L}^{sup}_{out,i} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \boldsymbol{\cdot} \boldsymbol{z}_p / \tau\right)}{\sum\limits_{a \in A(i)} \exp\left(\boldsymbol{z}_i \boldsymbol{\cdot} \boldsymbol{z}_a / \tau\right)}$$

$$\mathcal{L}^{sup}_{in} = \sum_{i \in I} \mathcal{L}^{sup}_{in,i} = \sum_{i \in I} -\log \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp\left(\boldsymbol{z}_i \boldsymbol{\cdot} \boldsymbol{z}_p / \tau\right)}{\sum\limits_{a \in A(i)} \exp\left(\boldsymbol{z}_i \boldsymbol{\cdot} \boldsymbol{z}_a / \tau\right)} \right\}$$

Comparison of two loss functions

❖ They determine a better loss function through experimentation.

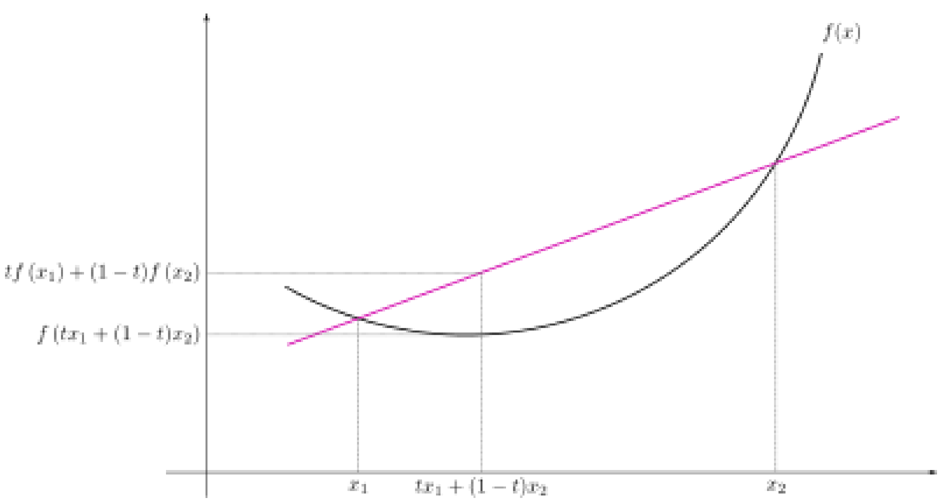
- In the experiment, $\mathcal{L}_{out}^{sup}$ shows better performance.

Loss	Top-1
$\mathcal{L}_{out}^{sup}$	78.7%
$\mathcal{L}_{in}^{sup}$	67.4%

- Also, Jensen's inequality shows that $\mathcal{L}_{out}^{sup}$ is the upper limit of $\mathcal{L}_{in}^{sup}$.

➢ Jensen's inequality

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$



Performance Comparison

❖ Comparison of multiple datasets

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers	Mean
SimCLR-50 [3]	**88.20**	**97.70**	**85.90**	**75.90**	**63.50**	91.30	**88.10**	84.10	73.20	89.20	92.10	**97.00**	**84.81**
Xent-50	87.38	96.50	84.93	74.70	63.15	89.57	80.80	**85.36**	**76.86**	92.35	**92.34**	96.93	**84.67**
SupCon-50	87.23	97.42	84.27	75.15	58.04	**91.69**	84.09	85.17	74.60	**93.47**	91.04	96.0	**84.27**
Xent-200	**89.36**	97.96	86.49	**76.50**	**64.36**	90.01	84.22	**86.27**	**76.76**	**93.48**	93.84	**97.20**	**85.77**
SupCon-200	88.62	**98.28**	**87.28**	76.26	60.46	**91.78**	**88.68**	85.18	74.26	93.12	**94.91**	96.97	**85.67**

❖ Comparison to Imagenet

Dataset	SimCLR[3]	Cross-Entropy	Max-Margin [32]	SupCon
CIFAR10	93.6	95.0	92.4	**96.0**
CIFAR100	70.7	75.3	70.5	**76.5**
ImageNet	70.2	78.2	78.0	**78.7**

Conclusion

❖ The performance of contrastive learning was improved by increasing the batch size.

❖ Through experiments, they proposed an effective augmentation combination for contrastive learning.

❖ The performance of contrastive learning was improved by using label information.