Supervised Contrastive learning

reference

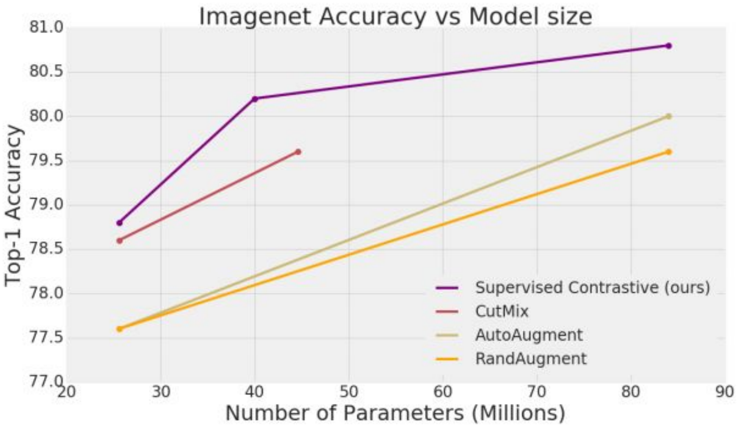
https://amitness.com/2020/03/illustrated-simclr/

https://app.wandb.ai/authors/scl/reports/Improving-Image-Classifiers-with-Supervised-Contrastive-Learning--VmlldzoxMzQwNzE

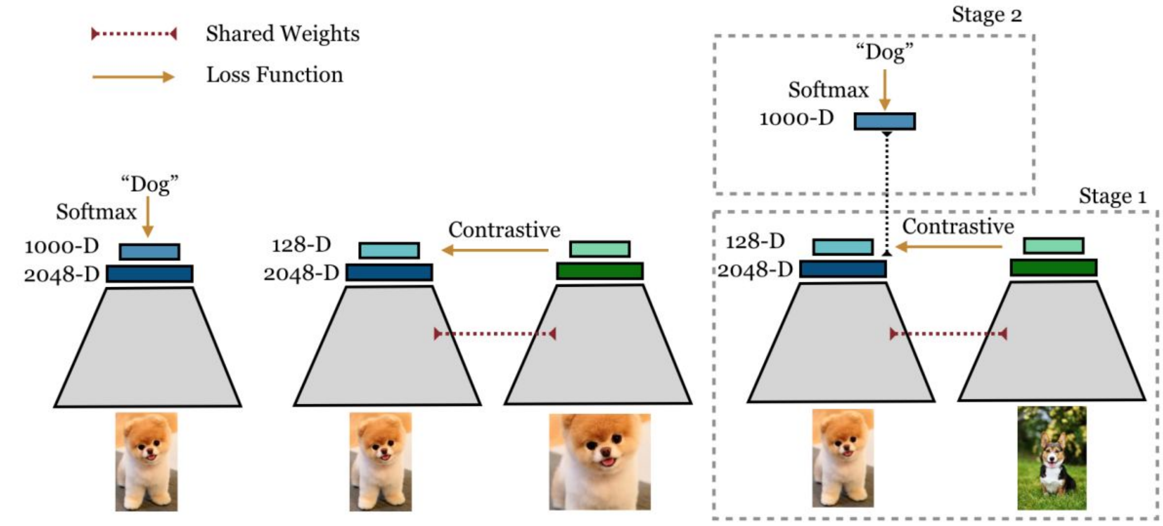
https://www.youtube.com/watch?v=MpdbFLXOOIw

Supervised Contrastive learning

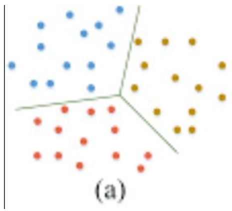
1. cross entropy 단점
   - lack of robustness to noisy labels
   - possibility of poor margin -> reduced generalization
2. proposed
   - label smoothing
   - self-distillation
   - Mixup and related data augmentation strategies
3. propose new loss for supervised training(contrastive loss)



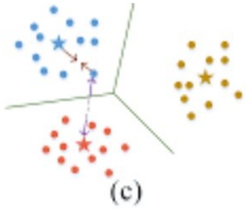
difference with self supervised contrastive learning



(a) Supervised Cross Entropy (b) Self Supervised Contrastive (c) Supervised Contrastive

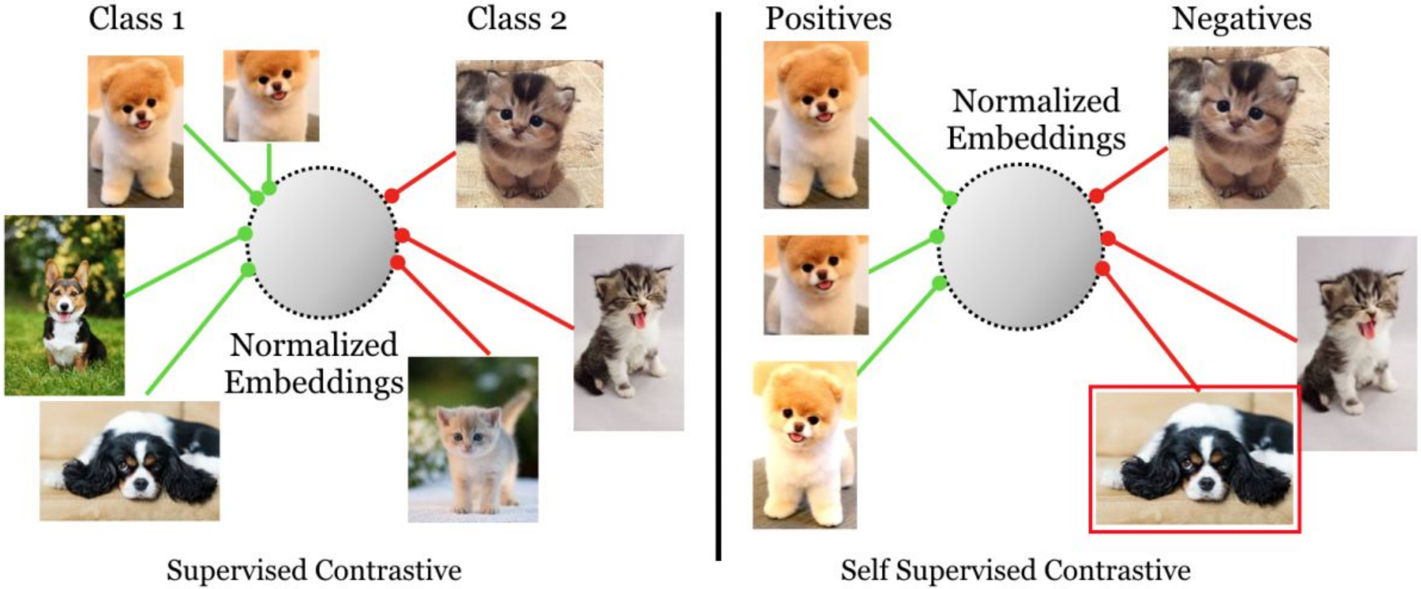


cross entropy



contrastive loss

difference with self supervised contrastive learning

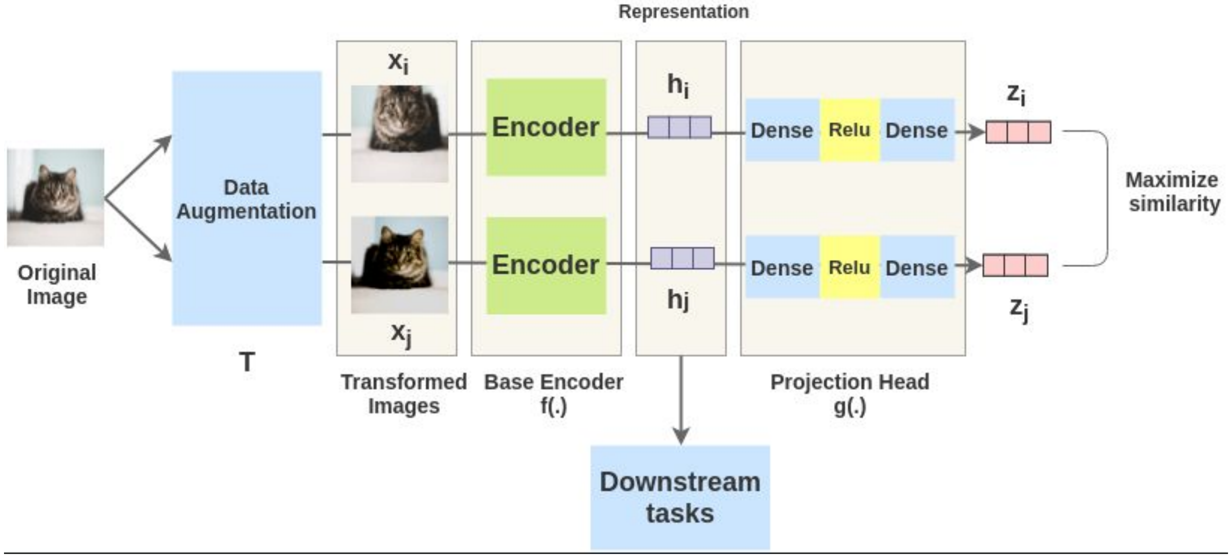


contrastive learning framework

A data augmentation module

An encoder network

A projection network



data augmentation module

generate two randomly augmented images

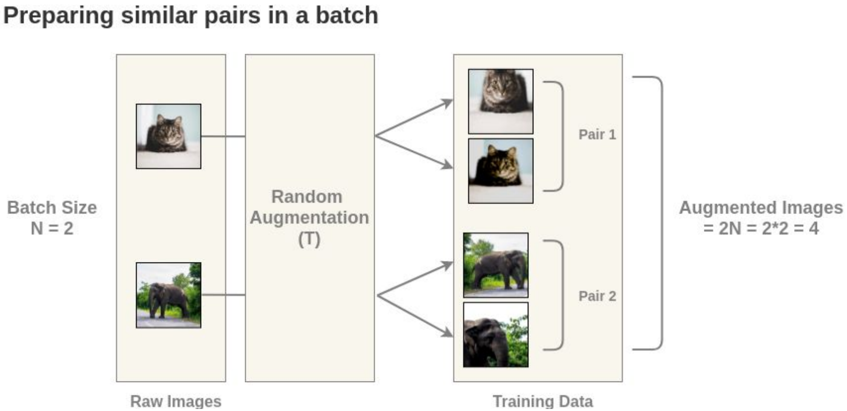
- a random crop to the image and then resizing
- 3 different options

  AutoAugment

  RandAugment

  SimAugment (A simple framework for contrastive learning of visual representations),

  apply random color distortion and Gaussian blurring

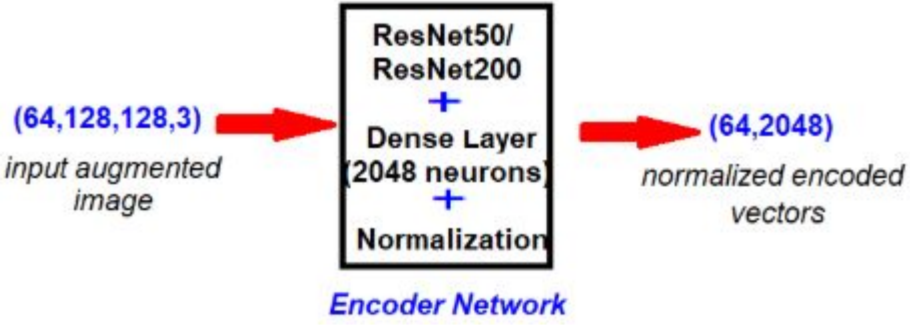


encoder network

maps an augmented image x˜ to a representation vector, r

ResNet-50 and ResNet-200, dim = 2048

normalized to the unit hypersphere



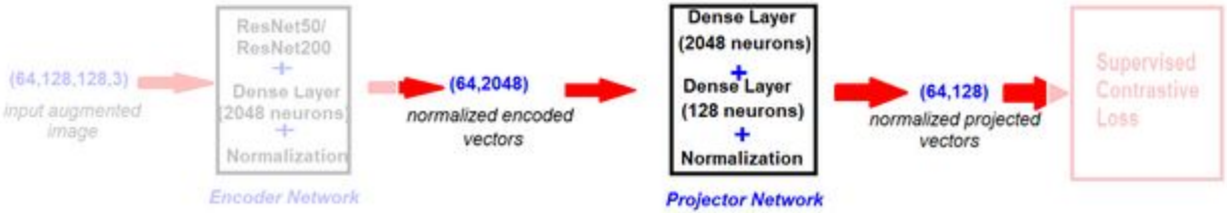
projection network

which maps the normalized representation vector r into a vector z (dim = 128)

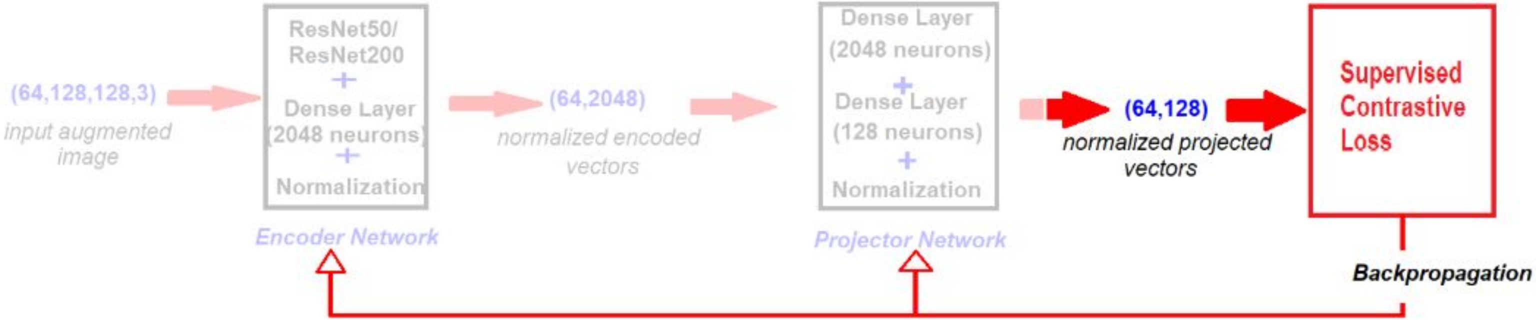
multi-layer perceptron with hidden layer of size 2048

normalize this vector to lie on the unit hypersphere
(inner product to measure distances in the projection space)

The projection network is only used for training the supervised contrastive loss



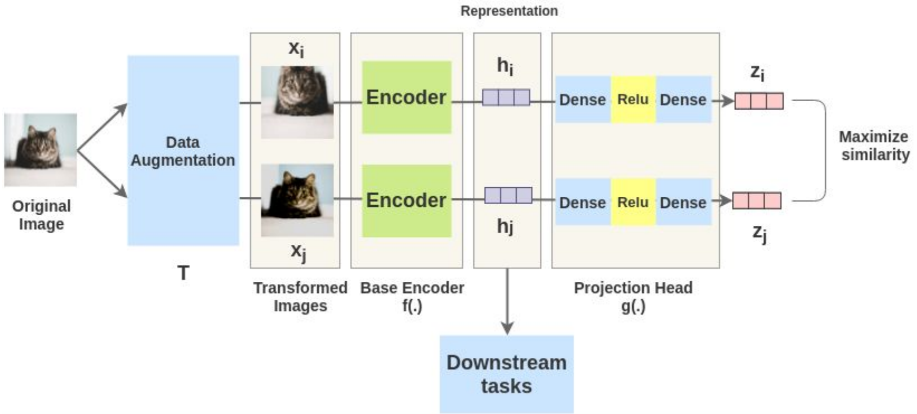
Supervised Contrastive Loss


(64,128,128,3)
input augmented image
ResNet50/ ResNet200
+
Dense Layer (2048 neurons)
+
Normalization
Encoder Network
(64,2048)
normalized encoded vectors
Dense Layer (2048 neurons)
+
Dense Layer (128 neurons)
+
Normalization
Projector Network
(64,128)
normalized projected vectors
Supervised Contrastive Loss
Backpropagation

$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \mathcal{L}_i^{sup}$$

$$\mathcal{L}_i^{sup} = \frac{-1}{2N_{\tilde{\boldsymbol{y}}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \log \frac{\exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_j / \tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_k / \tau\right)}$$

vs Supervised Contrastive Loss



$$\mathcal{L}^{self} = \sum_{i=1}^{2N} \mathcal{L}_i^{self}$$

$$\mathcal{L}_i^{self} = -\log \frac{\exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_{j(i)}/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_k/\tau\right)}$$

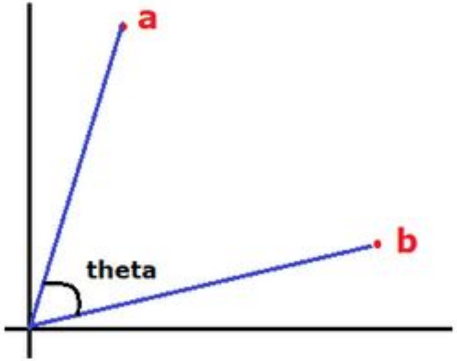
$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \mathcal{L}_i^{sup}$$

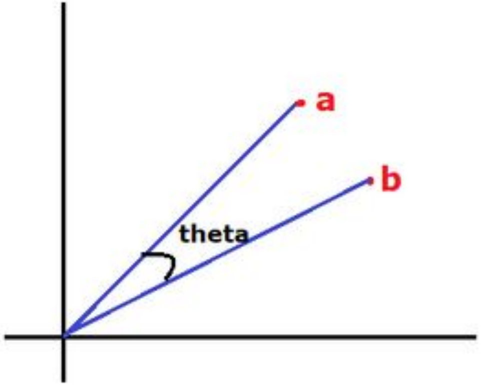
$$\mathcal{L}_i^{sup} = \frac{-1}{2N_{\tilde{\boldsymbol{y}}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \log \frac{\exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_j/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_k/\tau\right)}$$

similarity

dot product between normalized vector, z

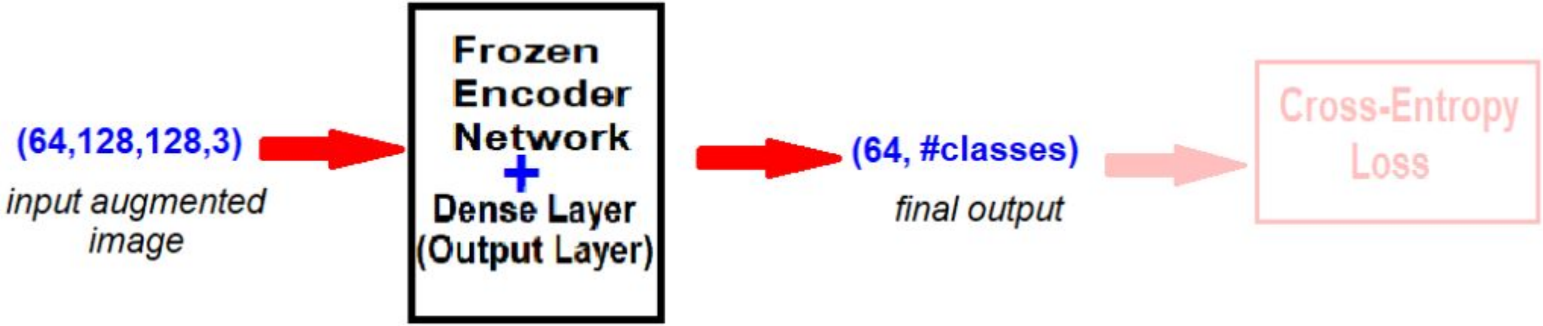
$$\bar{a} \bullet \bar{b} = |\bar{a}||\bar{b}|\cos\theta$$





Downstream Task

Not projection but representation



Supervised Contrastive Loss Gradient Properties

$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \mathcal{L}_i^{sup}$$

$$\mathcal{L}_i^{sup} = \frac{-1}{2N_{\tilde{\boldsymbol{y}}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \log \frac{\exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_j / \tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_k / \tau\right)}$$

$$\frac{\partial \mathcal{L}_i^{sup}}{\partial \boldsymbol{w}_i} = \left.\frac{\partial \mathcal{L}_i^{sup}}{\partial \boldsymbol{w}_i}\right|_{\text{pos}} + \left.\frac{\partial \mathcal{L}_i^{sup}}{\partial \boldsymbol{w}_i}\right|_{\text{neg}}$$

$$\left.\frac{\partial \mathcal{L}_i^{sup}}{\partial \boldsymbol{w}_i}\right|_{\text{pos}} \propto \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \left(\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_j\right) \cdot \boldsymbol{z}_i - \boldsymbol{z}_j\right) \cdot \left(1 - P_{ij}\right)$$

$$\left.\frac{\partial \mathcal{L}_i^{sup}}{\partial \boldsymbol{w}_i}\right|_{\text{neg}} \propto \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \sum_{k=1}^{2N} \mathbb{1}_{k \notin \{i,j\}} \cdot \left(\boldsymbol{z}_k - \left(\boldsymbol{z}_i \bullet \boldsymbol{z}_k\right) \cdot \boldsymbol{z}_i\right) \cdot P_{ik}$$

Supervised Contrastive Loss Gradient Properties

the supervised contrastive loss to focus more on hard positives and negative

$$\|((\boldsymbol{z}_i \bullet \boldsymbol{z}_j) \cdot \boldsymbol{z}_i - \boldsymbol{z}_j)\| \cdot (1 - P_{ij}) = \sqrt{1 - (\boldsymbol{z}_i \bullet \boldsymbol{z}_j)^2} \cdot (1 - P_{ij}) \approx 0$$

$$\|((\boldsymbol{z}_i \bullet \boldsymbol{z}_j) \cdot \boldsymbol{z}_i - \boldsymbol{z}_j)\| \cdot (1 - P_{ij}) = \sqrt{1 - (\boldsymbol{z}_i \bullet \boldsymbol{z}_j)^2} \cdot (1 - P_{ij}) > 0$$

where

$$P_{i\ell} = \frac{\exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_\ell / \tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp\left(\boldsymbol{z}_k \bullet \boldsymbol{z}_\ell / \tau\right)} \quad , \quad i, \ell \in \{1...2N\} \, , \; i \neq \ell$$

Connections to Triplet Loss

Contrastive learning is closely related to the triplet loss

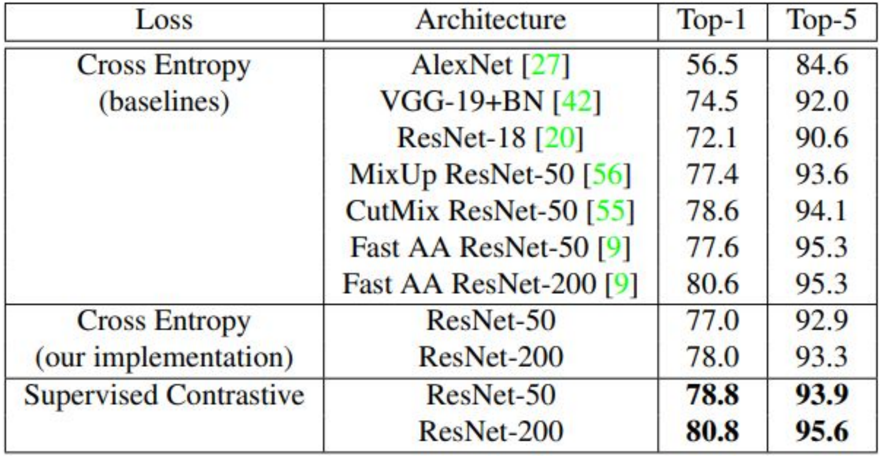
$$\begin{aligned}
\mathcal{L}_{con} &= -\log \frac{\exp\left(\boldsymbol{z}_a \bullet \boldsymbol{z}_p / \tau\right)}{\exp\left(\boldsymbol{z}_a \bullet \boldsymbol{z}_p / \tau\right) + \exp\left(\boldsymbol{z}_a \bullet \boldsymbol{z}_n / \tau\right)} \\
&= \log\left(1 + \exp\left(\left(\boldsymbol{z}_a \bullet \boldsymbol{z}_n - \boldsymbol{z}_a \bullet \boldsymbol{z}_p\right) / \tau\right)\right) \\
&\approx \exp\left(\left(\boldsymbol{z}_a \bullet \boldsymbol{z}_n - \boldsymbol{z}_a \bullet \boldsymbol{z}_p\right) / \tau\right) \quad \text{(Taylor expansion of log)} \\
&\approx 1 + \frac{1}{\tau} \cdot \left(\boldsymbol{z}_a \bullet \boldsymbol{z}_n - \boldsymbol{z}_a \bullet \boldsymbol{z}_p\right) \\
&= 1 - \frac{1}{2\tau} \cdot \left(\left\|\boldsymbol{z}_a - \boldsymbol{z}_n\right\|^2 - \left\|\boldsymbol{z}_a - \boldsymbol{z}_p\right\|^2\right) \\
&\propto \left\|\boldsymbol{z}_a - \boldsymbol{z}_p\right\|^2 - \left\|\boldsymbol{z}_a - \boldsymbol{z}_n\right\|^2 + 2\tau
\end{aligned}$$

$$d(X,Y)^2 = \langle X - Y, X - Y \rangle = \langle X, X \rangle + \langle Y, Y \rangle - 2\langle X, Y \rangle = 2(1 - \langle X, Y \rangle)$$

Training detail

700 epochs during the pretraining stage. (350 epochs only dropped the top-1 accuracy by a small amount)
training step is about 50% slower than cross-entropy
batch sizes of up to 8192
a temperature of $\tau = 0.07$

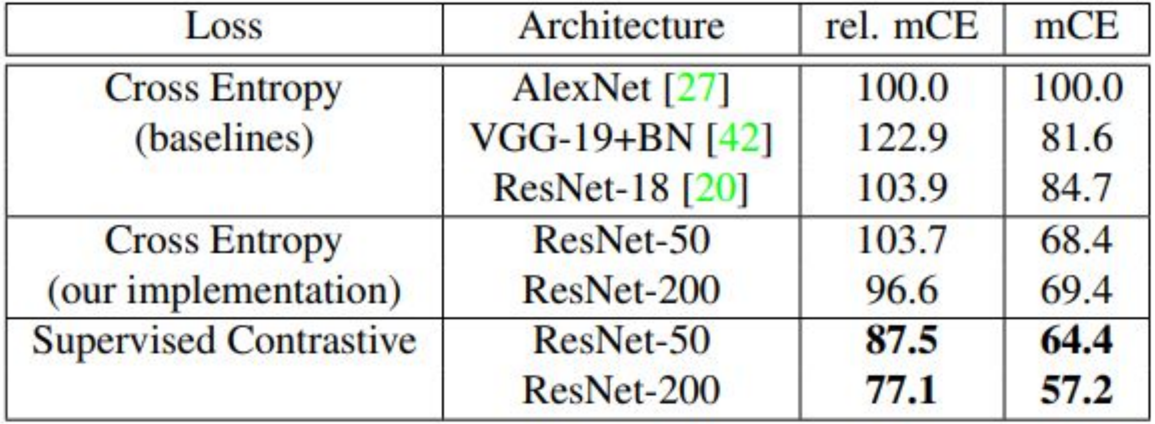
ImageNet Classification Accuracy



Loss	Architecture	Top-1	Top-5
Cross Entropy	AlexNet [27]	56.5	84.6
(baselines)	VGG-19+BN [42]	74.5	92.0
	ResNet-18 [20]	72.1	90.6
	MixUp ResNet-50 [56]	77.4	93.6
	CutMix ResNet-50 [55]	78.6	94.1
	Fast AA ResNet-50 [9]	77.6	95.3
	Fast AA ResNet-200 [9]	80.6	95.3
Cross Entropy	ResNet-50	77.0	92.9
(our implementation)	ResNet-200	78.0	93.3
Supervised Contrastive	ResNet-50	**78.8**	**93.9**
	ResNet-200	**80.8**	**95.6**

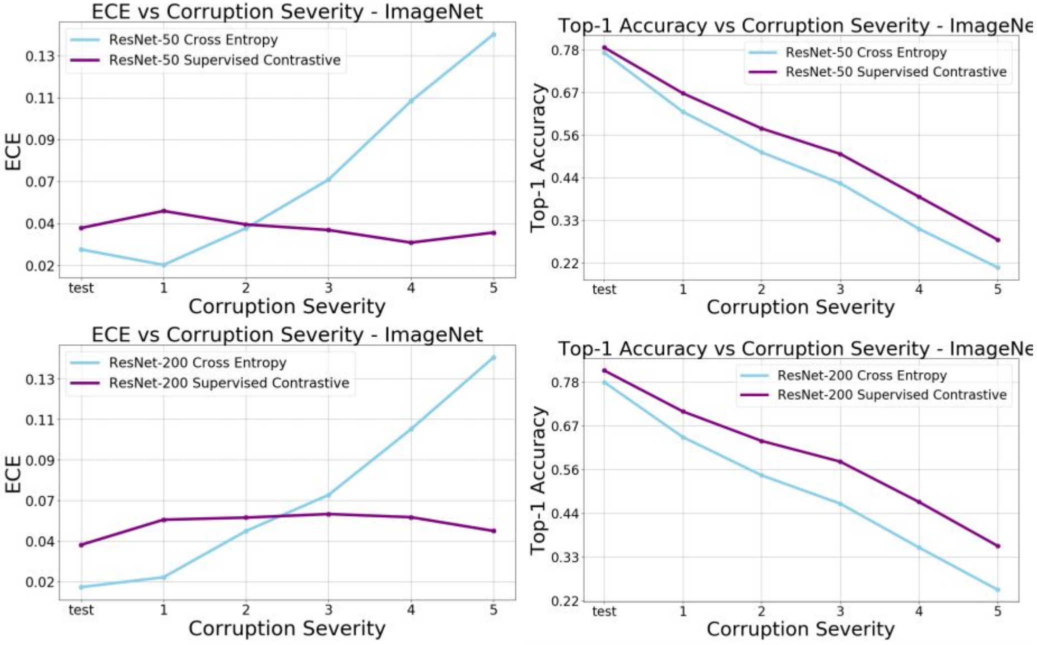
Robustness to Image Corruptions and Calibration

Training with Supervised Contrastive Loss makes models more robust to corruptions in images



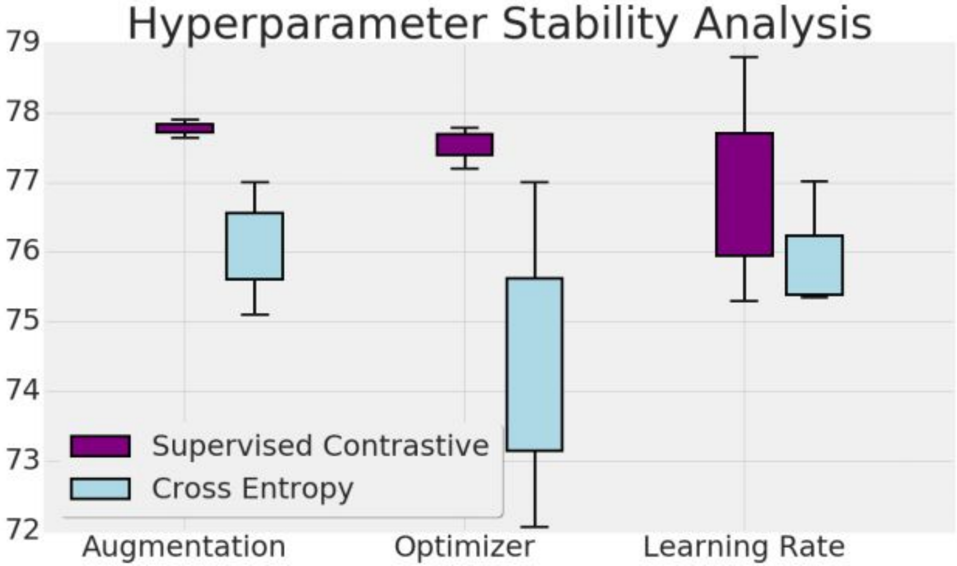
Loss	Architecture	rel. mCE	mCE
Cross Entropy (baselines)	AlexNet [27]	100.0	100.0
	VGG-19+BN [42]	122.9	81.6
	ResNet-18 [20]	103.9	84.7
Cross Entropy (our implementation)	ResNet-50	103.7	68.4
	ResNet-200	96.6	69.4
Supervised Contrastive	ResNet-50	**87.5**	**64.4**
	ResNet-200	**77.1**	**57.2**

Robustness to Image Corruptions and Calibration

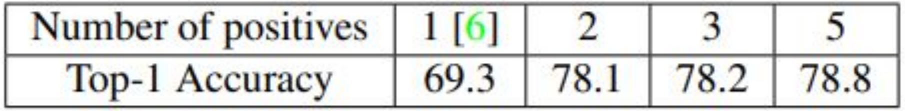


Hyperparameter Stability

maybe due to the smoother geometry of the hypersphere compared to labels which are the endpoints of the n-dimensional simplex (as cross-entropy requires)



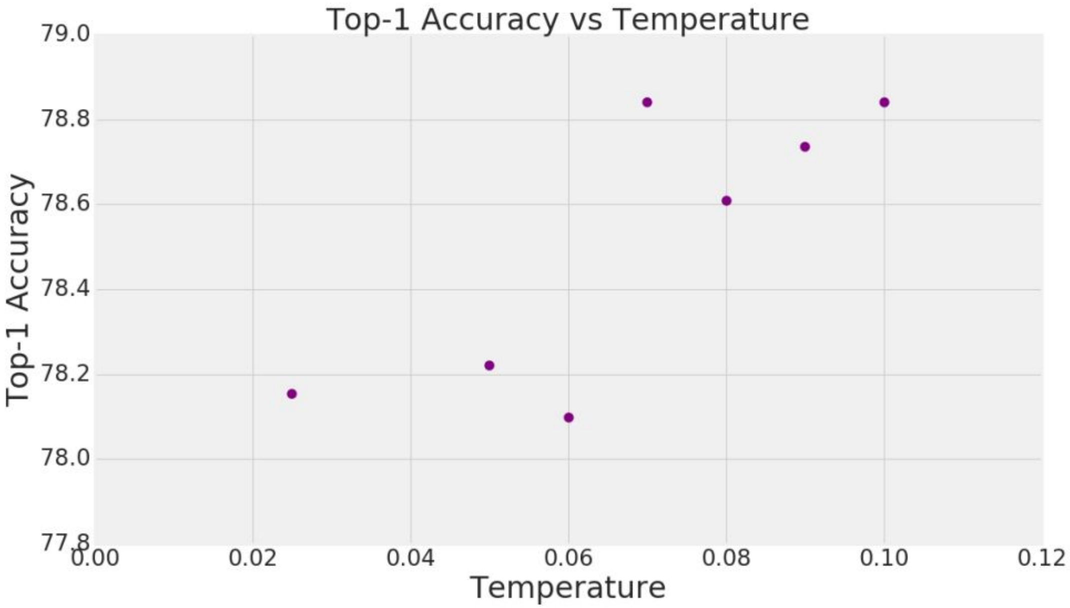
Effect of Number of Positives



Number of positives	1 [6]	2	3	5
Top-1 Accuracy	69.3	78.1	78.2	78.8

Effect of Temperature in Loss Function

temperature is important



Introduction - Meta Learning

Learning to Learn

적은 수의 sample만으로 학습할 수 없을까?

크게 3가지 방식으로 분류

- metric 기반의 representation을 학습하는 방식
- model기반의 external/internal memory를 통한 recurrent network를 학습하는 방식
- optimization 기반의 fast learning을 위한 model의 hyperparameter를 최적화하는 방식

few shot learning/Meta learning

Not class but example specific -> overfit -> much data

Is it possible to learn with few data like person

few shot learning/Meta learning

Multi task learning

배경(잔디)가 개의 특징이 아님을 확인 가능

매번 다른 task를 학습함으로 인해 specific feature들에 대한 학습은 서로 상쇄되고 class에 대한 general featrue들이 학습이 되게 된다.