Supervised Contrastive Learning

0. Abstract

- a novel training methology that consistently outperforms cross entropy on supervised learning tasks

- modify the batch contrastive loss

- leverage key ingredients such as large batch sizes and normalized embeddings

1. Introduction

- cross-entropy loss

- 1) KL-divergence between two discrete distributions: the label distribution and the empirical distribution of the logits

- 2) lack robustness to noisy labels

- 3) possibility of poor margin

- 4) reduced generalized performance

- propose a new loss for supervised training which completely does away with a reference distribution:

- simply impose that normalized embeddings from the same class are closer together than embeddings from different classes

- inspired by the family of contrastive objective functions

- contrastive loss
- consists of two “opposing force”
- 1) ‘positives’ pulls the anchor closer in representation space to other points
- 2) ‘negatives’ pushes the anchor farther away from other points
- consider many positives per anchor in addition to many negatives

- **Figure 2: Supervised vs self-supervised contrastive losses** In the supervised contrastive loss, positives from one class are contrasted with negatives from other classes. In self-supervised contrastive loss, labels are not provided. Hence positives are generated as data augmentations of a given sample and negatives are randomly sampled from the mini-batch. This can result in false negatives, which may not be mapped correctly, resulting in a worse representation.

-

Supervised contrastive

class 1

class 2

dog 1

cat 1

dog 2

cat 2

normalized embeddings

dog 3

Cat 3

self-supervised contrastive

positives

negatives

cat 1

cat 2

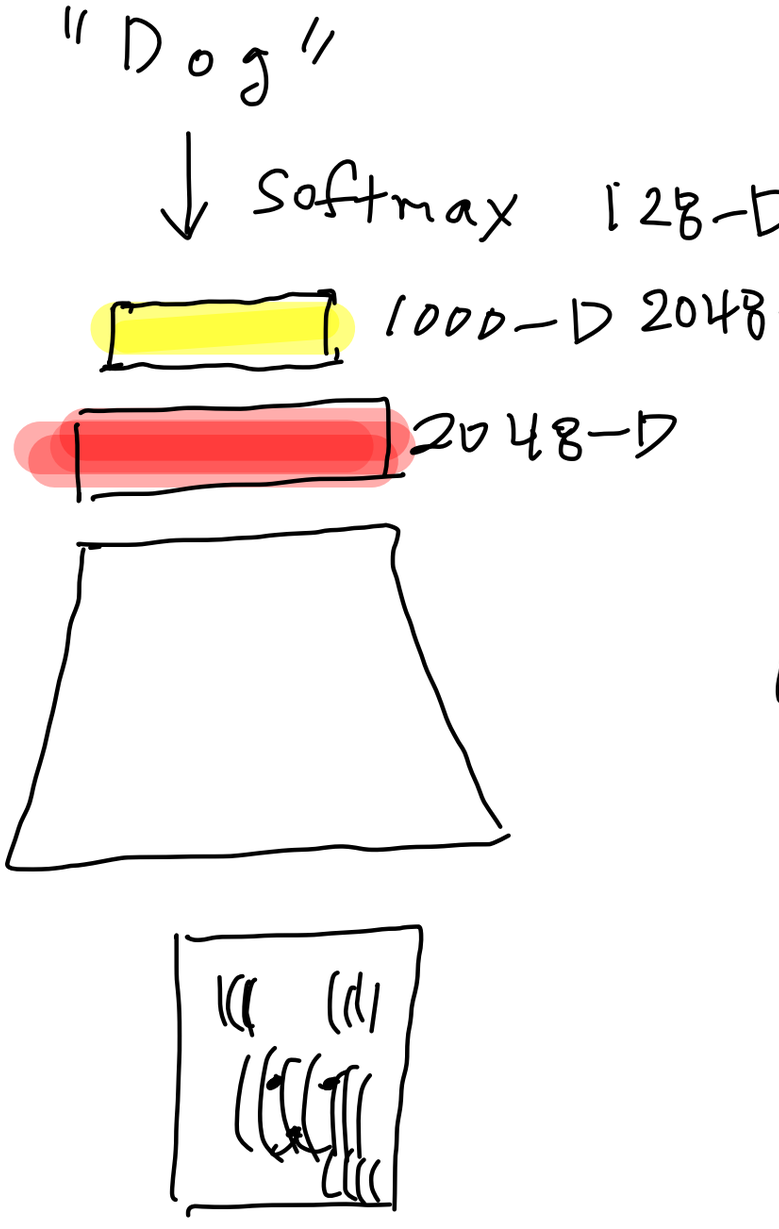
normalized embeddings

dog 1

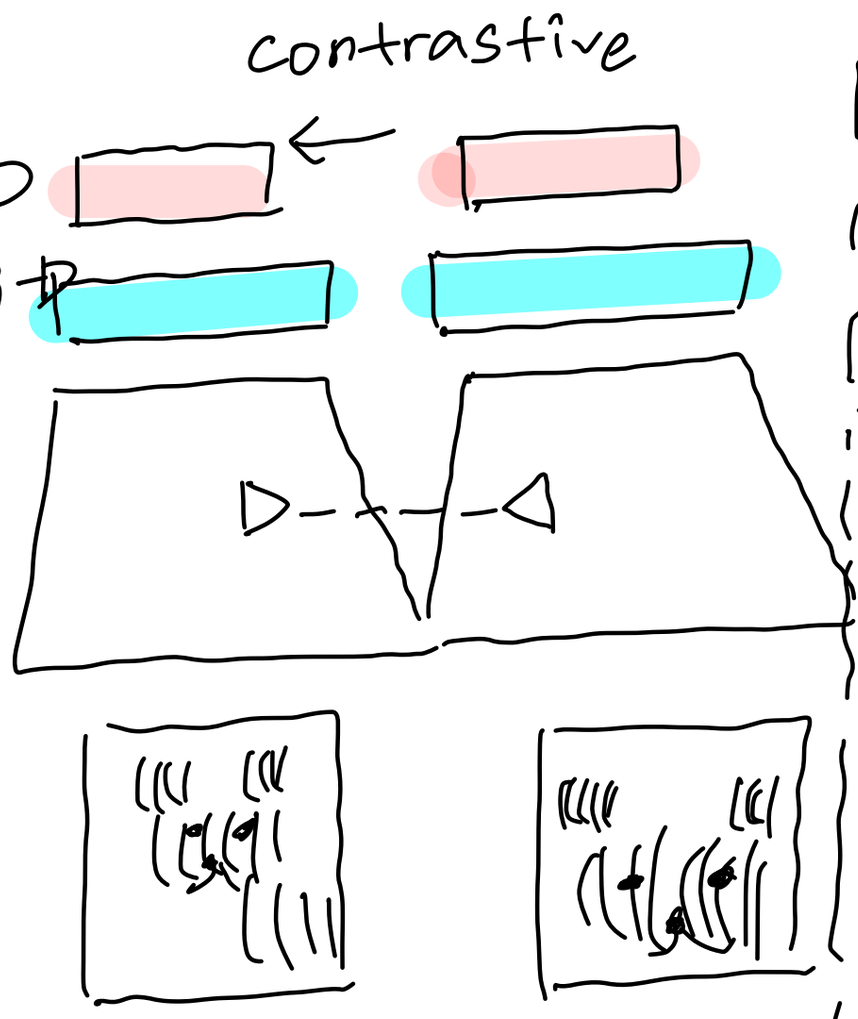
- **Figure 3: Cross entropy, self-supervised contrastive loss and supervised contrastive loss** The cross entropy loss uses labels and a softmax loss to train a model; the self-supervised contrastive loss uses a contrastive loss and data augmentations to learn representations about classes; the supervised contrastive loss has two stages; we use labels to choose the images for a contrastive loss. Then we freeze the learned representations and then learn a classifier on a linear layer using a softmax loss. Thus combining the benefits of using labels and contrastive losses.

- ▷- - ◁ shared weights

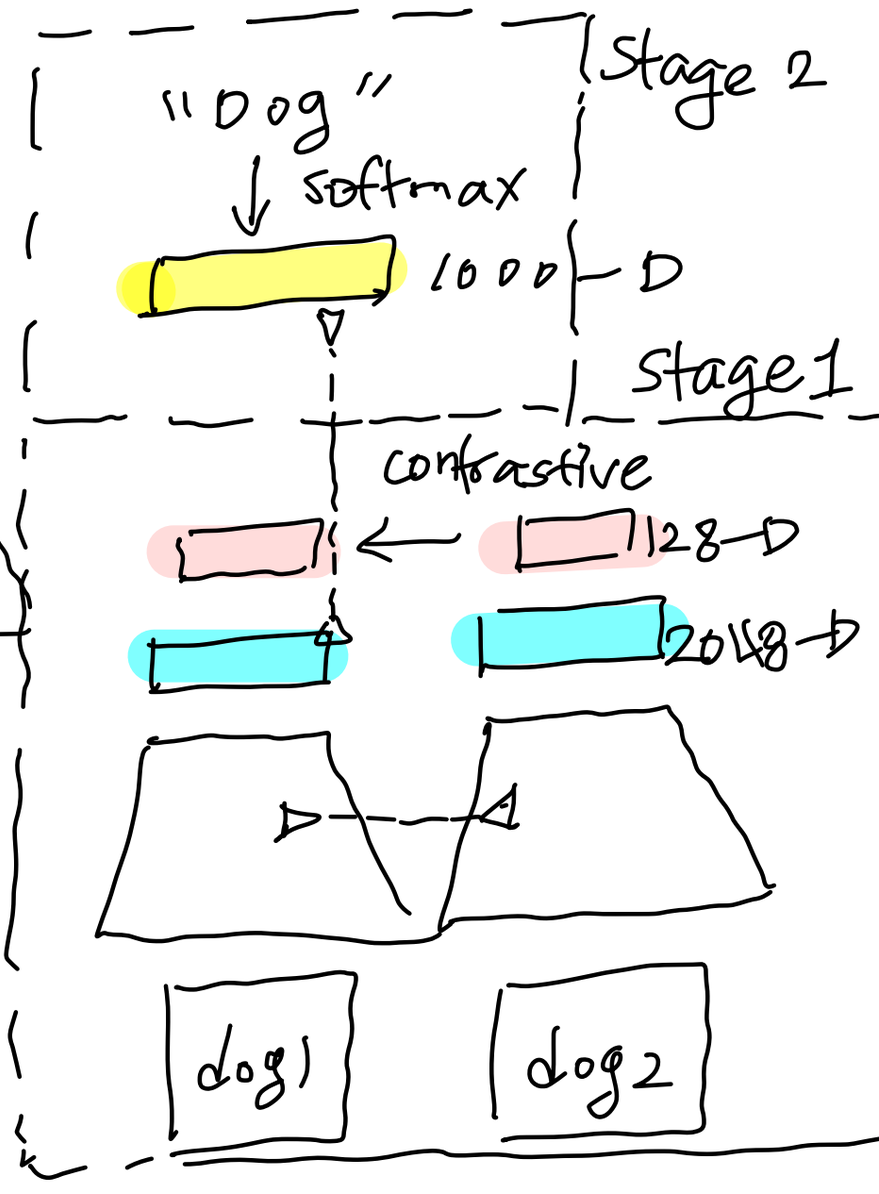
(a) supervised
cross entropy



(b) self-supervised
contrastive



(c) supervised
contrastive



→ loss function

3. Method

- how can we modify self-supervised representation learning to be suitable for fully supervised learning?

3.1 Representation Learning Framework

- 1) A data augmentation module $A(\cdot)$
- 2) An encoder network $E(\cdot)$
- 3) A projection network $P(\cdot)$

3.2.1 Self-supervised Contrastive Loss

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

$$L_i^{self} = -\log \frac{\exp(z_i \cdot z_{j(i)} / \tau)}{\sum_{k=1}^{2N} 1_{i \neq k} \cdot \exp(z_i \cdot z_k / \tau)}$$

$i \in \{1, \cdots, 2N\}$ index of arbitrary augmented image

$j(i)$ index of the other augmented image originating from the same source image

index i is called the anchor

index $j(i)$ is called the positive

the other $2(N-1)$ indices ($k=1\cdots 2N, k\notin\{i,j\}$) are called the negatives

$z_k = P(E(\tilde{x}_k))$

$1_B \in \{0,1\}$ indicator function that returns 1 Iff B evaluates as True

$\tau > 0$ scalar temperature parameter

$z_i \cdot z_{j(i)}$ computes the inner (dot) product between normalized vectors $z_i, z_{j(i)}$ in 128-dimensional space

* there is 1 positive pair and $2N-2$ negative pairs

The encoder learns to map similar views to neighboring representations

3.2.2 Supervised Contrastive Loss

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

$$L_i^{sup} = \frac{-1}{2N\tilde{y}_i - 1} \sum_{j=1}^{2N} 1_{i \neq j} \cdot 1_{\tilde{y}_i = \tilde{y}_j} \cdot \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2N} 1_{i \neq k} \cdot \exp(z_i \cdot z_k / \tau)}$$

$N\tilde{y}_i$ total number of images in the mini-batch that have the same label, $\tilde{y}_i$, as the anchor i.

This loss has important properties well suited for supervised learning:

1) "Generalization to an arbitrary number of positives"
For any anchor, all positives in a mini-batch contribute to the numerator.
The loss encourages the encoder to give closely aligned representations to all entries from the same class.

2) "Contrastive power increases with more negatives"
The ability to discriminate between signal and noise (negatives) is improved by adding more examples of negatives.
Add large numbers of negatives to the denominator!

"By using many positives and negatives", we are able to better model both intra-class and inter-class variability.

3.2.3 Supervised Contrastive Loss Gradient Properties

$$\frac{\partial L_i^{sup}}{\partial W_i} = \frac{\partial L_i^{sup}}{\partial W_i}\bigg|_{pos} + \frac{\partial L_i^{sup}}{\partial W_i}\bigg|_{neg}$$

$$\frac{\partial L_i^{sup}}{\partial W_i}\bigg|_{pos} \propto \sum_{j=1}^{2N} 1_{i \neq j} 1_{\tilde{y}_i = \tilde{y}_j} \left((z_i \cdot z_j) \cdot z_i - z_j\right) \cdot (1 - P_{ij})$$

$$\frac{\partial L_i^{sup}}{\partial W_i}\bigg|_{neg} \propto \sum_{j=1}^{2N} 1_{i \neq j} 1_{\tilde{y}_i = \tilde{y}_j} \sum_{k=1}^{2N} 1_{k \notin \{i,j\}} \cdot \left[z_k - (z_i \cdot z_k) \cdot z_i\right) \cdot P_{ik}$$

W : projection network output immediately prior to normalization (i.e. $z = W / \|W\|$)

where

$$p_{il} = \frac{\exp(z_i \cdot z_l / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \exp(z_k \cdot z_l / \tau)} \qquad i, l \in \{1, \cdots, 2N\} \quad i \neq l$$

The gradient has a structure that naturally causes learning to focus more on hard positives and negatives (ones against which continuing to contrast the anchor greatly benefits the encoder)

The addition of a normalization layer at the end of the projection network allows the gradient to have this structure

Easy positives and negatives have small gradient contributions while hard positives and negatives have large ones.

for easy positives, $z_i \cdot z_j \approx 1$ and thus P_{ij} is large

$$\| ((z_i \cdot z_j) \cdot z_i - z_j) \| \cdot (1 - P_{ij}) = \sqrt{1-(z_i \cdot z_j)^2} \cdot (1 - P_{ij}) \approx 0$$

for a hard positive, $z_i \cdot z_j \approx 0$ and P_{ij} is moderate

$$\| ((z_i \cdot z_j) \cdot z_i - z_j) \| \cdot (1 - P_{ij}) = \sqrt{1-(z_i \cdot z_j)^2} \cdot (1 - P_{ij}) > 0$$

the general $((z_i \cdot z_\ell) \cdot z_i - z_\ell)$ structure plays a key role in ensuring the gradients appears only if a normalization layer is added to the end of the projection network

4. Experiments

4.1 ImageNet Classification Accuracy

Evaluate by measuring classification accuracy on ImageNet

After training the embedding network with supervised contrastive loss on ImageNet, we replace the projection head of the network with a new randomly initialized linear dense (fully connected) layer.

This linear layer is trained with standard cross entropy while the parameters of the embedding network are kept unchanged.

<Result> Table 1.

Loss	Architecture	Top-1
Cross-entropy	AlexNet	56.5
supervised contrastive	ResNet-50	78.8
	ResNet-200	80.8

4.2 Robustness to Image Corruptions and Calibration

Deep neural networks often lack robustness to out of distribution data or natural corruptions.

Compare the supervised contrastive models to cross entropy using mean Corruption Error (mCE) and relative mean Corruption Error (rel. mCE)

<Result> Table 2

Loss	architecture	rel.	mCE
cross entropy	AlexNet	100.0	100.0
supervised contrastive	ResNet-50	87.5	64.4
	ResNet-200	77.1	57.2

4.5 Training Details

The supervised contrastive loss needs an additional step of training a final linear classifier to compute top-1 accuracy.
This is not needed if the purpose is to representations for transfer learning tasks or retrieval.

Batch size : 8192
Small number of steps suffice to train the dense layers on top of the frozen embedding network
Temperature: 0.07
Smaller temperature benefit training more than higher ones
AutoAugment
optimizer: LARS for pre-training and RMSProp for training the dense layer on the top of the frozen network