7.3 Supervised Contrastive Learning

2020년 4월 논문

논문 : https://arxiv.org/abs/2004.11362

해당 논문은 Vision카테고리에 가야하나, Contrastive learning(또는 Bi-Encoder)에서 사용할 수 있는 추가 Loss를 제안하여 NLP 카테고리에 추가함

논문이 생각하는 문제

- lack of robustness to noisy label, possibility of poor margins 같은 기법들이 나왔으나 ImageNet같이 큰 데이터셋에는 효과가 없었다
- cross entropy는 인공지능 분야에서 널리 사용되는데 이를 개선한 논문들로 성능이 향상됨 - Label smoothing, self-distillation, Mixup, 등..
- self-supervised learning, triplet loss 같은 Positive:Negative = 1:N, 1:1 사용하면 배치사이즈가 너무 커진다

논문 주장

- classification task에 대해 fine-tuning할때 supervised contrastive learning을 같이 사용하자!

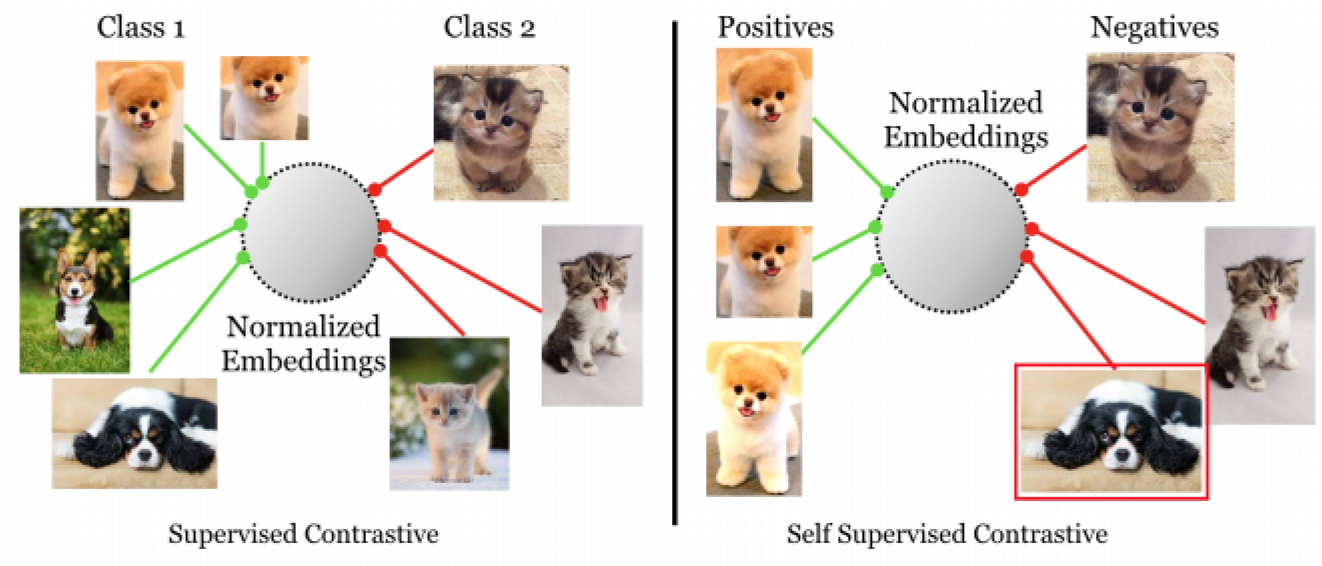

Figure 2: Supervised vs. self-supervised contrastive losses: In the supervised contrastive loss considered in this paper (left), positives from one class are contrasted with negatives from other classes (since labels are provided); images from the same class are mapped to nearby points in a low-dimensional hypersphere. In self-supervised contrastive loss (right), labels are not provided. Hence positives are generated as data augmentations of a given sample (crops, flips, color changes etc.), and negatives are randomly sampled from the mini-batch. This can result in false negatives (shown in bottom right), which may not be mapped correctly, resulting in a worse representation.

타 논문이 많이 사용하는 Self-Supervised Contrastive learning은?

- Positive:Negative = 1:N으로 sample을 설정
    - Positive는 같은 카테고리 이미지 원본 + Data augmentation 기법으로 변형된 이미지
- Positive sample과 Negative sample들간의 거리를 벌리기 위함
- Data augmentation 기법으로 Positive Sample이 불어남 => 1 epoch당 돌려야할 샘플이 너무 많아진다!

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

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

논문이 주장하는 Supervised Contrastive learning은?

- Positive:Negative = N:M으로 sample을 설정하자
- positive에 대한 softmax 총합/positive 개수를 Loss로 사용
- 장점!
    - Positive sample끼리 결집시킬수있음!
    - 전체 샘플내에 다양한 Positive sample 그룹이 있기 때문에 돌려야할 샘플이 적어짐!

$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \mathcal{L}_i^{sup} \tag{3}$$

$$\mathcal{L}_i^{sup} = \frac{-1}{2N_{\tilde{\boldsymbol{y}}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \log \frac{\exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_j / \tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_k / \tau\right)} \tag{4}$$

논문이 주장하는 최종 학습 구조

- classification task에 대한 학습 진행
    - BERT embedding (output layer)처럼 Output Layer 추가
- supervised contrastive learning도 같이 진행
    - 별도의 Head를 추가해서 Contrastive Learning 진행

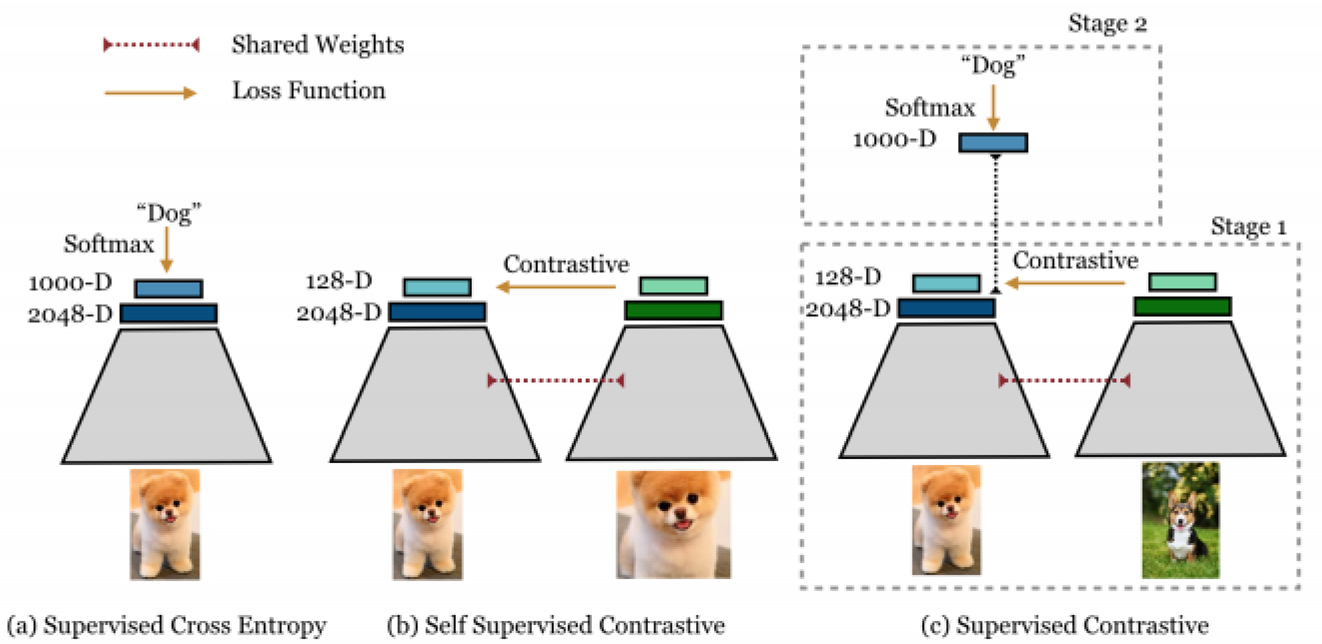

Figure 3: Cross entropy, self-supervised contrastive loss and supervised contrastive loss: The cross entropy loss (left) uses labels and a softmax loss to train a model; the self-supervised contrastive loss (middle) uses a contrastive loss and data augmentations to learn representations about classes; the supervised contrastive loss (right) proposed in this paper has two stages; in the first stage we use labels to choose the images for a contrastive loss. In the second stage, 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.

실험결과

Loss	Architecture	Top-1	Top-5
Cross Entropy (baselines)	AlexNet [27]	56.5	84.6
	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 (our implementation)	ResNet-50	77.0	92.9
	ResNet-200	78.0	93.3
Supervised Contrastive	ResNet-50	**78.8**	**93.9**
	ResNet-200	**80.8**	**95.6**

Table 1: Top-1/Top-5 accuracy results on ImageNet on ResNet-50 and ResNet-200 with AutoAugment [9] being used as the augmentation for Supervised Contrastive learning. Achieving 78.8% on ResNet-50, we outperform all of the top methods whose performance is shown above. Baseline numbers are taken from the referenced papers and we also additionally re-implement cross-entropy ourselves for fair comparison.

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**

Table 2: Training with Supervised Contrastive Loss makes models more robust to corruptions in images, as measured by Mean Corruption Error (mCE) and relative mCE over the ImageNet-C dataset [22] (lower is better).

하이퍼파라미터 분석

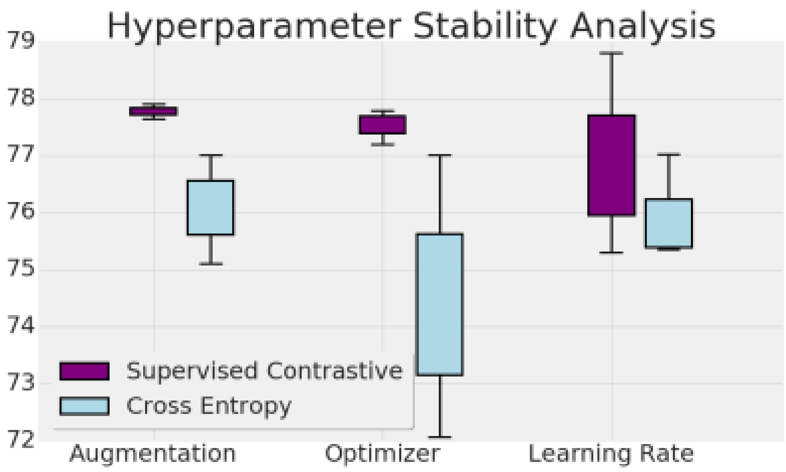

Figure 4: Comparison of top-1 accuracy variability of cross entropy and supervised contrastive loss to changes in hyper-parameters. We compare three augmentations (RandAugment [10], AutoAugment [9] and SimAugment) (left plot); three optimizers (LARS, SGD with Momentum and RMSProp); and 3 learning rates that vary from the optimal rate by a factor of 10 smaller or larger. The supervised contrastive loss is more stable to changes in hyperparameters.

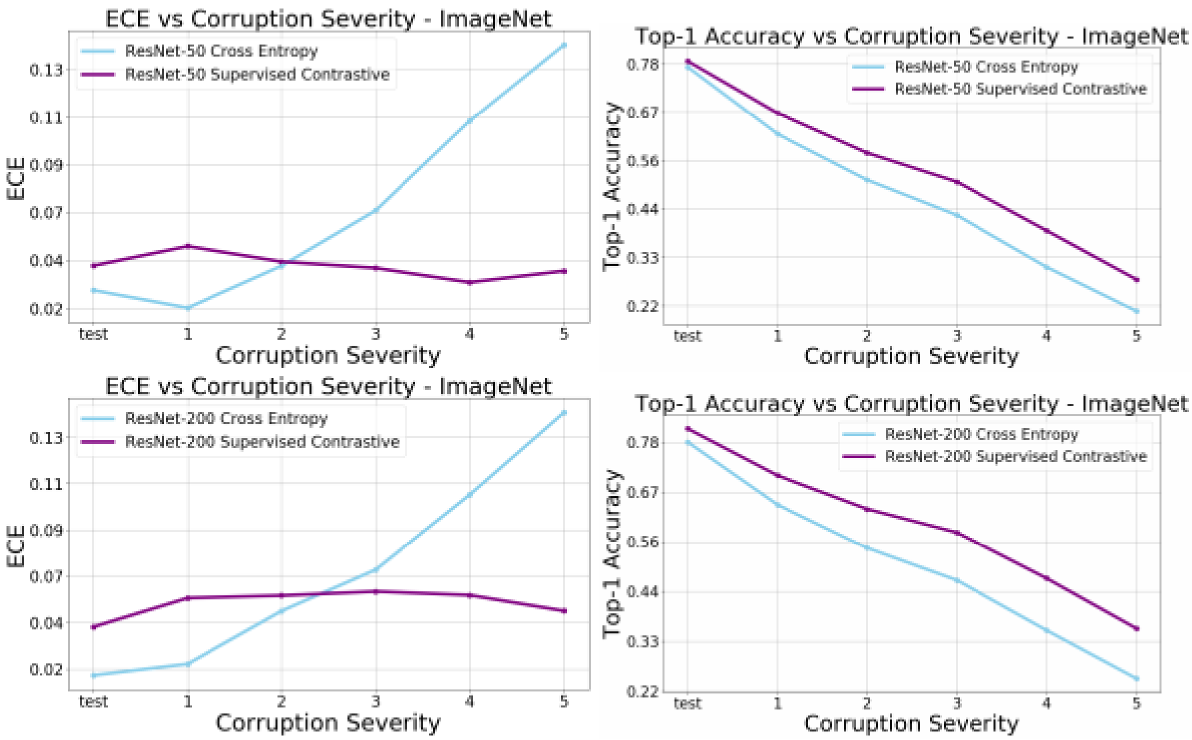

Figure 5: Expected Calibration Error and mean top-1 accuracy at different corruption severities on ImageNet-C, on the ResNet-50 architecture (top) and ResNet-200 architecture (bottom). The contrastive loss maintains a higher accuracy over the range of corruption severities, and does not suffer from increasing calibration error, unlike the cross entropy loss.