南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Understanding the Behaviour of Contrastive Loss

Feng Wang, Huaping Liu †

Beijing National Research Center for Information Science and Technology(BNRist),
Department of Computer Science and Technology, Tsinghua University

wang-f20@mails.tsinghua.edu.cn, hpliu@tsinghua.edu.cn

CVPR 2021

Contrastive learning

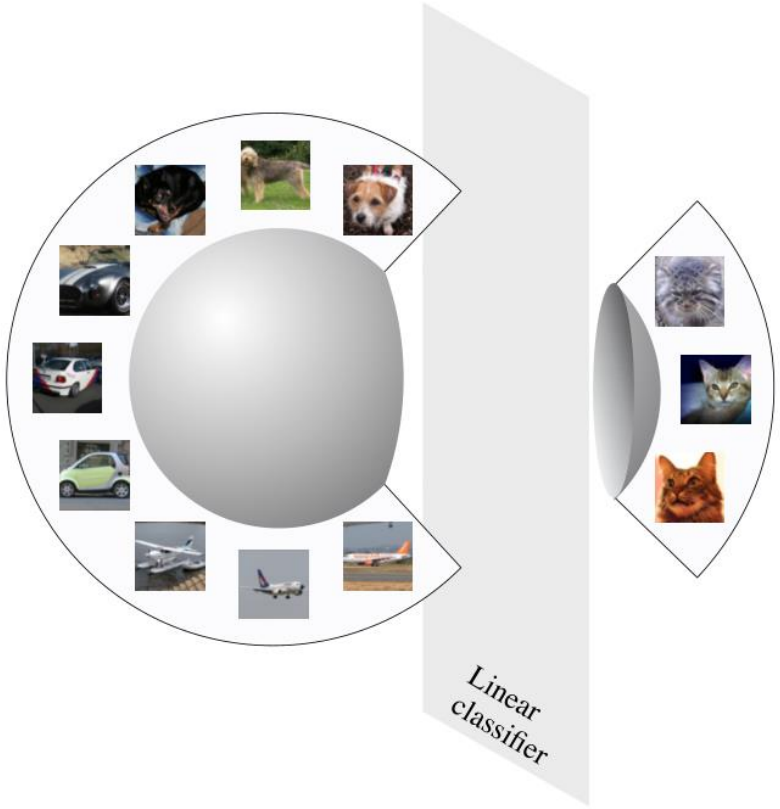


Figure 2: **Hypersphere:** When classes are well-clustered (forming spherical caps), they are linearly separable. The same does not hold for Euclidean spaces.

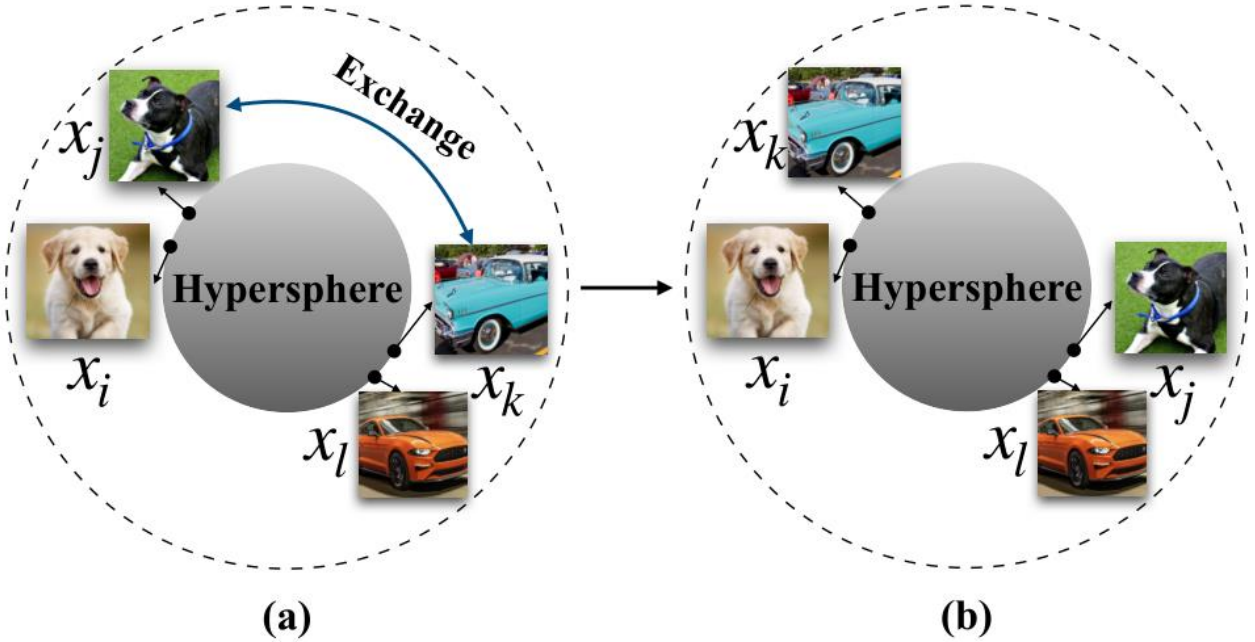
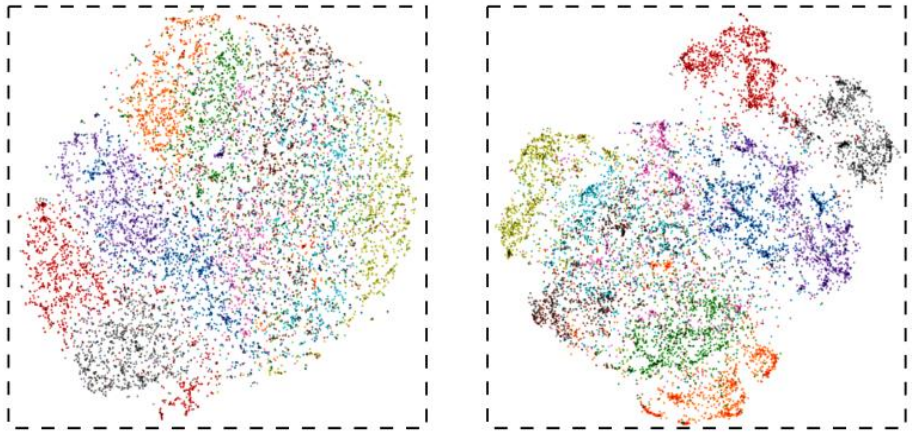


Figure 1. We display two embedding distributions with four instances on a hypersphere. From the figure, we observe that exchanging x_j and x_k, as well as their corresponding augmentations, will not change the value of contrastive loss. However, the embedding distribution of (a) is much more useful for downstream tasks because it captures the semantical relations between instances.

Hardness-aware Property

$$\mathcal{L}(x_i) = -\log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau)+\exp(s_{i,i}/\tau)}\right] \quad (1)$$

$$\mathcal{L}_{simple}(x_i) = -s_{i,i} + \lambda\sum_{i\neq j} s_{i,j} \quad (3)$$



$\tau = 0.07$ $\tau = 0.2$

Figure 2. T-SNE [29] visualization of the embedding distribution. The two models are trained on CIFAR10. The temperature is set to 0.07 and 0.2 respectively. Small temperature tends to generate more uniform distribution and be less tolerant to similar samples.

Gradients Analysis

$$\mathcal{L}(x_i) = -\log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau)+\exp(s_{i,i}/\tau)}\right] \quad (1)$$

$$P_{i,j} = \frac{\exp(s_{i,j}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau)+\exp(s_{i,i}/\tau)} \quad (2)$$

$$\frac{\partial\mathcal{L}(x_i)}{\partial s_{i,i}} = -\frac{1}{\tau}\sum_{k\neq i}P_{i,k}, \quad \frac{\partial\mathcal{L}(x_i)}{\partial s_{i,j}} = \frac{1}{\tau}P_{i,j} \quad (4)$$

- Gradients $\propto exp(s_{i,j}/\tau)$
- $(\sum_{k\neq i}|\frac{\partial L(x_i)}{\partial s_{i,k}}|)/|\frac{\partial L(x_i)}{\partial s_{i,i}}| = 1$

The Role of temperature

The **relative penalty** on negative sample xj:

$$r_i(s_{i,j}) = |\frac{\partial L(x_i)}{\partial s_{i,j}}| / |\frac{\partial L(x_i)}{\partial s_{i,i}}|$$

$$r_i(s_{i,j}) = \frac{\exp(s_{i,j}/\tau)}{\sum_{k \neq i} \exp(s_{i,k}/\tau)}, \quad i \neq j$$

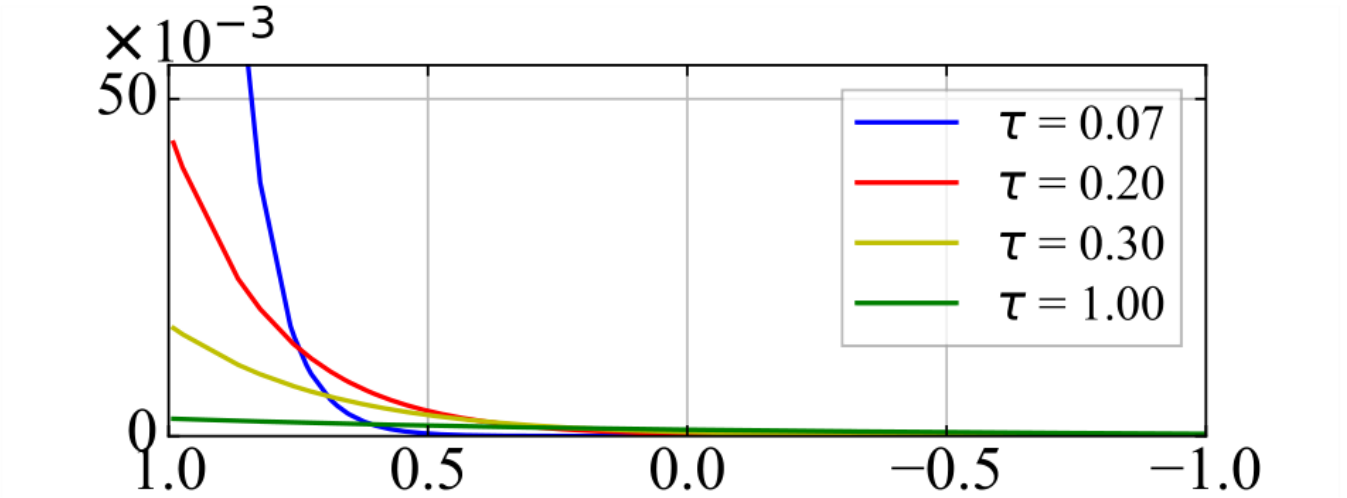


Figure 3. The gradient ratio $r_{i,j}$ with respect to different $s_{i,j}$. We sample the $s_{i,j}$ from a uniform distribution in $[-1, 1]$. As we can see, with lower temperature, the contrastive loss tends to punish more on the hard negative samples.

The Role of temperature

$$
\begin{aligned}
&\lim_{\tau \to 0^+} -\log \left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k \neq i} \exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)} \right] \\
&= \lim_{\tau \to 0^+} +\log \left[1 + \sum_{k \neq i} \exp((s_{i,k} - s_{i,i})/\tau) \right] \\
&= \lim_{\tau \to 0^+} +\log \left[1 + \sum_{s_{i,k} \geqslant s_{i,i}}^{k} \exp((s_{i,k} - s_{i,i})/\tau) \right] \\
&= \lim_{\tau \to 0^+} \frac{1}{\tau} max[s_{max} - s_{i,i}, 0]
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
&\lim_{\tau \to +\infty} -\log \left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k \neq i} \exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)} \right] \\
&= \lim_{\tau \to +\infty} -\frac{1}{\tau} s_{i,i} + \log \sum_{k} \exp(s_{i,k}/\tau) \\
&= \lim_{\tau \to +\infty} -\frac{1}{\tau} s_{i,i} + \frac{1}{N} \sum_{k} \exp(s_{i,k}/\tau) - 1 + \log N \\
&= \lim_{\tau \to +\infty} -\frac{N-1}{N\tau} s_{i,i} + \frac{1}{N\tau} \sum_{k \neq i} s_{i,k} + \log N
\end{aligned}
\tag{7}
$$

Explicit Hard Negative Sampling

$$\mathcal{L}_{\text{hard}}(x_i) = -\log \frac{\exp(s_{i,i}/\tau)}{\sum_{s_{i,k} \geqslant s_{\alpha}^{(i)}} \exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)} \tag{9}$$

数据集	Contrastive Loss (Eq1)	Simple Loss (Eq2) + Hard
CIFAR-10	79.75	84.84
CIFAR-100	51.82	55.71
ImageNet-100	71.53	74.31
SVHN	92.55	94.99

Uniformity-Tolerance Dilemma

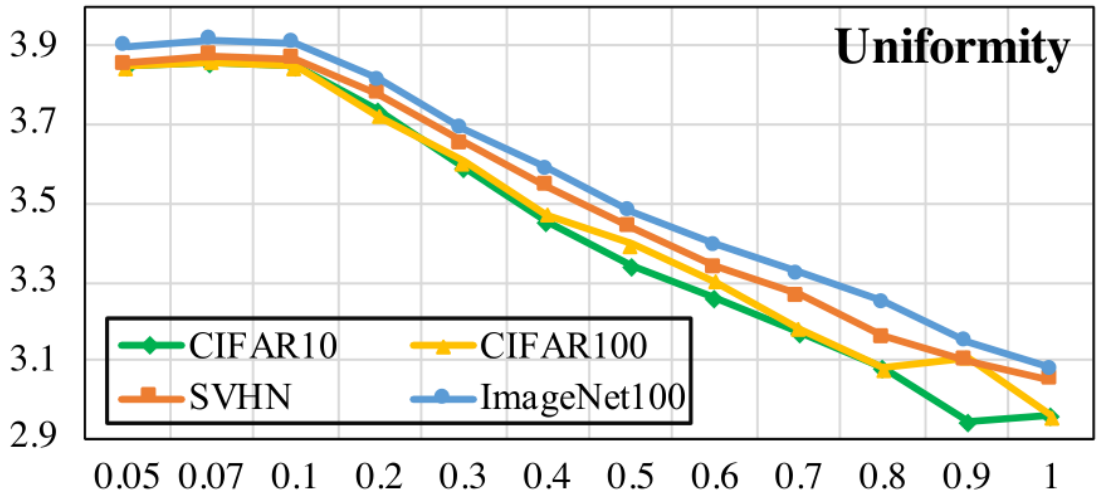


Figure 4. Uniformity of embedding distribution trained with different temperature on CIFAR10, CIFAR100 and SVHN. The x axis represents different temperature, and y axis represents $-\mathcal{L}_{\text{uniformity}}$. Large value means the distribution is more uniform.

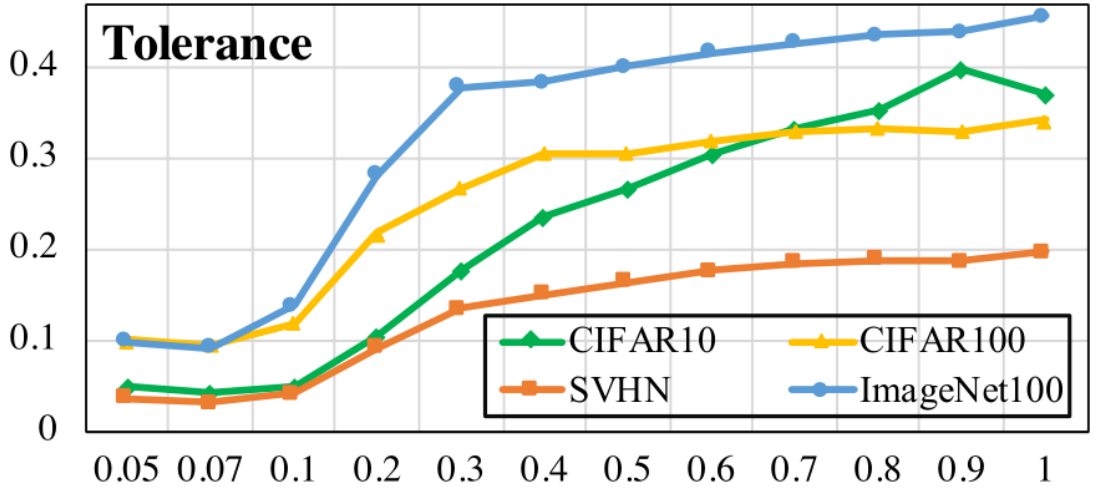


Figure 5. Measurement of tolerance on models trained on CIFAR10, CIFAR100 and SVHN. The x axis represents different temperatures, and y axis represents the tolerance to samples with the same category. Large value means the model is more tolerant to semantically consistent samples.

$$\mathcal{L}_{\text{uniformity}}(f;t) = \log \mathop{\mathbb{E}}_{x,y\sim p_{data}} \left[e^{-t||f(x)-f(y)||_2^2} \right] \quad (10)$$

$$T = \mathop{\mathbb{E}}_{x,y\sim p_{data}} \left[(f(x)^T f(y)) \cdot I_{l(x)=l(y)} \right] \quad (11)$$

Uniformity-Tolerance Dilemma

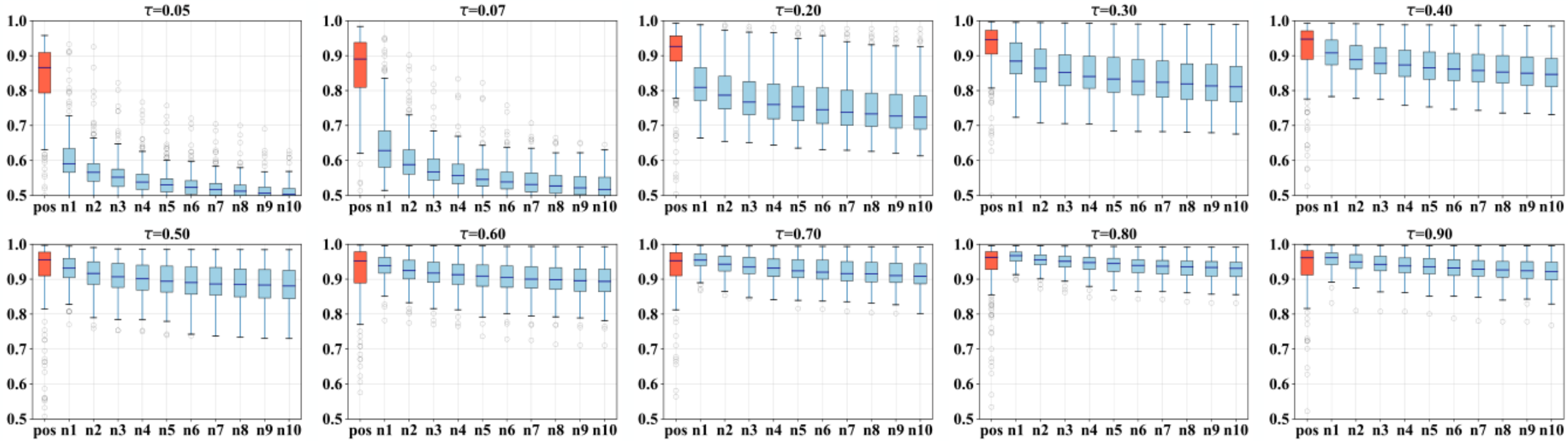


Figure 8. We display the similarity distribution of positive samples and the top-10 nearest negative samples that are marked as 'pos' and 'ni' for the i-th nearest neighbour. All models are trained on CIFAR100. For models trained on other datasets, they present the same pattern with the above figure, and we display them in the supplementary material.

Uniformity-Tolerance Dilemma

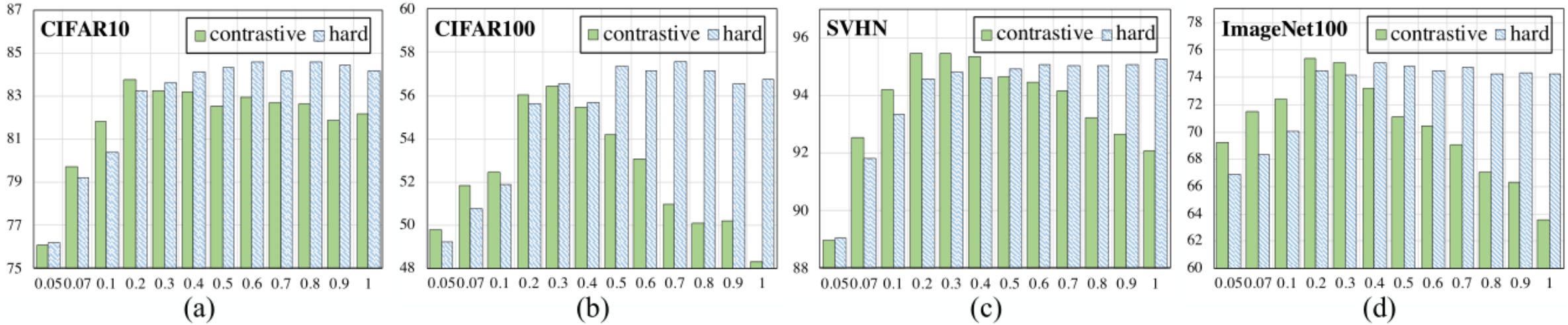


Figure 9. Performance comparison of models trained with different temperatures. For CIFAR10, CIFAR100 and SVHN, the backbone network is ResNet-18, and for ImageNet, the backbone network is ResNet-50. After the pretraining stage, we freeze all convolutional layers and add a linear layer. We report 1-crop top-1 accuracy for all models.

Experiments

Dataset	Result	Contrastive				Simple	HardContrastive				HardSimple
		0.07	0.3	0.7	1.0		0.07	0.3	0.7	1.0	
CIFAR10	accuracy	79.75	83.27	82.69	82.21	74.83	79.2	83.63	84.19	84.19	84.84
	uniformity	3.86	3.60	3.17	2.96	1.68	3.88	3.89	3.87	3.86	3.85
	tolerance	0.04	0.178	0.333	0.372	0.61	0.034	0.0267	0.030	0.030	0.030
CIFAR100	accuracy	51.82	56.44	50.99	48.33	39.31	50.77	56.55	57.54	56.77	55.71
	uniformity	3.86	3.60	3.18	2.96	2.12	3.87	3.88	3.87	3.86	3.86
	tolerance	0.10	0.269	0.331	0.343	0.39	0.088	0.124	0.158	0.172	0.174
SVHN	accuracy	92.55	95.47	94.17	92.07	70.83	91.82	94.79	95.02	95.26	94.99
	uniformity	3.88	3.65	3.27	3.05	1.50	3.89	3.91	3.90	3.88	3.85
	tolerance	0.032	0.137	0.186	0.197	0.074	0.025	0.021	0.021	0.023	0.026
ImageNet100	accuracy	71.53	75.10	69.03	63.57	48.09	68.33	74.21	74.70	74.28	74.31
	uniformity	3.917	3.693	3.323	3.08	1.742	3.929	3.932	3.927	3.923	3.917
	tolerance	0.093	0.380	0.427	0.456	0.528	0.067	0.096	0.121	0.134	0.157

Table 1. We report the accuracy of linear classification on CIFAR10, CIFAR100 and SVHN, including models trained with the ordinary contrastive loss, simple contrastive loss, hard contrastive loss and hard simple contrastive loss. For models trained on ordinary contrastive loss and hard contrastive loss, we select several representative temperatures. More results are shown in the supplementary material.

Uniformity-Tolerance Dilemma

The dilemma is caused by the inherent defect of unsupervised contrastive loss that it pushes all different instances ignoring their semantical relations!

An ideal model should be both locally clustered and globally uniform.

Conclusion

- 1. Contrastive loss is a **hardness-aware** loss function.
- 2. The **hardness-aware property** is significant to the success of the contrastive loss.
- 3. The **temperature** plays a key role in controlling the **local separation** and **global uniformity** of the embedding distributions.

THANKS