Supplementary Material: Understanding the Behaviour of Contrastive Loss

1. Introduction

In this supplementary material, we list some detailed results of our paper including: **(1)** The proof of monotonicity of entropy with respect to temperature coeffecient τ. **(2)** All numerical results of different models trained with contrastive loss are shown in Table 1, and the results of different models trained with hard contrastive loss are shown in Table 2. We train these models using different temperatures ranging from 0.05 to 1.0 on CIFAR10, CIFAR100, SVHN and ImageNet100. **(3)** For the ImageNet100 dataset, we list the 100 labels of ImageNet100 which is shown in Table 3. **(4)**. The illustrations of the local separation on different datasets. In our paper, we have shown the local separation property on CIFAR100 dataset, and have found that the local separation property on all datasets is similar. Figure 1-3 show the local separation on CIFAR10, CIFAR100 and ImageNet100. Fig 4-7 show the local separation property of the hard contrastive loss on the four datasets.

2. Proof in Sec3.2

In Sec3.2 (The role of temperature), we have stated that the entropy $H(r_i)$ increases strictly as the temperature increases. In this part, we prove this statement. Specifically, given a distribution $r_i(s_{i,j})$ as:

$$r_i(s_{i,j}) = \frac{\exp(s_{i,j}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau)}, \quad i \neq j \tag{1}$$

, what we hope to prove is that when other variables including $s_{i,j}$ and $s_{i,k}$ keep invariant, the entropy is $H(r_i)$ is monotonically increasing (Except for the special case when all $s_{i,k}$ is equal, which makes the r_i be a uniform distribution). For simplicity, let:

$$P_l = \exp(s_{i,l}/\tau) \tag{2}$$

We then re-write the entropy H using the above symbol as follows:

$$\begin{aligned} H(r_i) &= -\sum_{j\neq i} r_i(s_{i,j}) \cdot \log(r_i(s_{i,j})) \\ &= -\frac{\sum_{j\neq i} P_j \cdot \log(P_j)}{\sum_{j\neq i} P_j} + \log(\sum_{j\neq i} P_j) \end{aligned} \tag{3}$$

Next, we calculate the gradients of H with respect to P_l for any $l \neq i$ as follows:

$$\frac{\partial H}{\partial P_l} = -\frac{\log P_l}{\sum_{j\neq i} P_j} + \frac{\sum_{j\neq i} P_j \log P_j}{(\sum_{j\neq i} P_j)^2} \tag{4}$$

and the gradient of P_l with respect to $1/\tau$:

$$\frac{\partial P_l}{\partial 1/\tau} = \tau P_l \cdot log(P_l) \tag{5}$$

We have computed the gradient of H with respect to P_l and the gradient of P_l with respect to $1/\tau$. Using the chain rule, we can calculate the gradient of H with respect to $1/\tau$ is as follows:

$$\begin{aligned} \frac{\partial H}{\partial 1/\tau} &= \sum_l \frac{\partial H}{\partial P_l} \cdot \frac{\partial P_l}{\partial 1/\tau} \\ &= \tau \cdot \frac{(\sum_{l\neq i} P_l \cdot \log(P_l))^2 - \sum_{l\neq i} P_l \sum_{l\neq i} P_l \log^2(P_l)}{(\sum_{l\neq i} P_l)^2} \end{aligned} \tag{6}$$

Up to now, we have calculated the gradient of H with respect to $1/\tau$ as the above equation, which only consists of τ and the proposed symbol P_l. Notice that $P_l > 0$, we can apply the Cauchy inequality to the numerator part of the above equation. We have:

$$\begin{aligned} \sum_{l\neq i} P_l \sum_{l\neq i} P_l \log^2(P_l) &= \sum_{l\neq i} \sqrt{P_l}^2 \sum_{l\neq i} (\sqrt{P_l} \cdot \log(P_l))^2 \\ &\geqslant (\sum_{l\neq i} P_l \cdot \log(P_l))^2 \end{aligned} \tag{7}$$

,such that $\partial H/\partial(1/\tau) \leqslant 0$. In another word, the entropy is monotonically increasing as the τ increases. Furthermore, we notice that the equality of the Cauchy inequality is satisfied only if all P_j is equal, which is almost impossible to satisfy in the learning process.

3. Results

We list detailed experiment results and the chosen ImageNet100 labels as the following tables and figures.

dataset	0.05	0.07	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.0
CIFAR10	76.10	79.75	81.82	83.78	83.27	83.22	82.54	82.97	82.69	82.67	81.97	82.21
CIFAR100	49.80	51.82	52.46	56.05	56.44	55.47	54.17	53.05	50.99	50.08	50.21	48.33
SVHN	88.96	92.55	94.21	95.46	95.47	95.36	94.66	94.47	94.17	93,22	92.66	92.07
ImageNet100	63.91	71.53	74.59	75.41	75.10	72.98	71.10	70.47	69.03	67.91	65.93	65.49

Table 1. All results of different models trained with the ordinary contrastive loss. We test all models on a linear classification task, which freezes all convolutional layers and adds a linear layer on top of the last convolutional layer. We evaluate the above contrastive models on CIFAR10, CIFAR100, SVHN and ImageNet100 respectively.

dataset	0.05	0.07	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.0
CIFAR10	76.22	79.20	80.44	83.28	83.63	84.14	84.31	84.60	84.19	84.60	84.45	84.19
CIFAR100	49.21	50.77	51.91	55.61	56.55	55.66	57.37	57.17	57.54	57.15	56.53	56.77
SVHN	89.07	91.82	93.37	94.58	94.79	94.62	94.93	95.06	95.02	95.03	95.08	95.26
ImageNet100	61.54	68.33	72.10	74.53	74.21	75.04	74.84	74.46	74.70	74.28	74.74	73.78

Table 2. All results of different models trained with the hard contrastive loss. We test all models on a linear classification task, which freezes all convolutional layers and adds a linear layer on top of the last convolutional layer. We evaluate the above hard contrastive models on CIFAR10, CIFAR100, SVHN and ImageNet100 respectively.

n01558993	n01692333	n01729322	n01735189	n01749939	n01773797	n01820546	n01855672	n01978455	n01980166
n01983481	n02009229	n02018207	n02085620	n02086240	n02086910	n02087046	n02089867	n02089973	n02090622
n02091831	n02093428	n02099849	n02100583	n02104029	n02105505	n02106550	n02107142	n02108089	n02109047
n02113799	n02113978	n02114855	n02116738	n02119022	n02123045	n02138441	n02172182	n02231487	n02259212
n02326432	n02396427	n02483362	n02488291	n02701002	n02788148	n02804414	n02859443	n02869837	n02877765
n02974003	n03017168	n03032252	n03062245	n03085013	n03259280	n03379051	n03424325	n03492542	n03494278
n03530642	n03584829	n03594734	n03637318	n03642806	n03764736	n03775546	n03777754	n03785016	n03787032
n03794056	n03837869	n03891251	n03903868	n03930630	n03947888	n04026417	n04067472	n04099969	n04111531
n04127249	n04136333	n04229816	n04238763	n04336792	n04418357	n04429376	n04435653	n04485082	n04493381
n04517823	n04589890	n04592741	n07714571	n07715103	n07753275	n07831146	n07836838	n13037406	n13040303

Table 3. All 100 labels of the ImageNet100 dataset. We take a subset of ImageNet datasets, and list the 100 labels here.

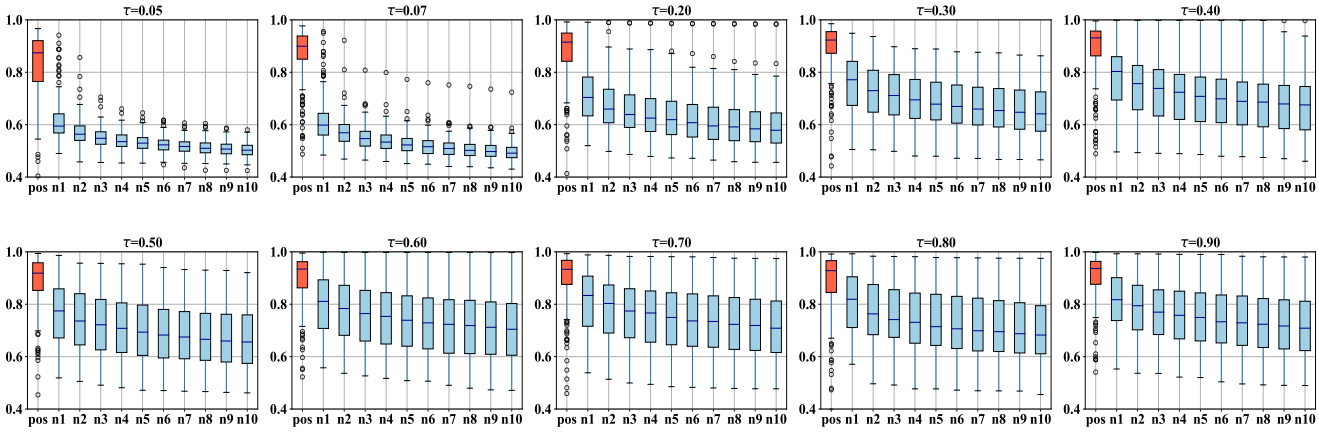

Figure 1. We display the similarity distribution of positive samples marked as 'pos' and the distributions of the top-10 nearest negative samples marked as 'ni' for the i-th nearest neighbour. All models are trained with the ordinary contrastive loss on CIFAR10.

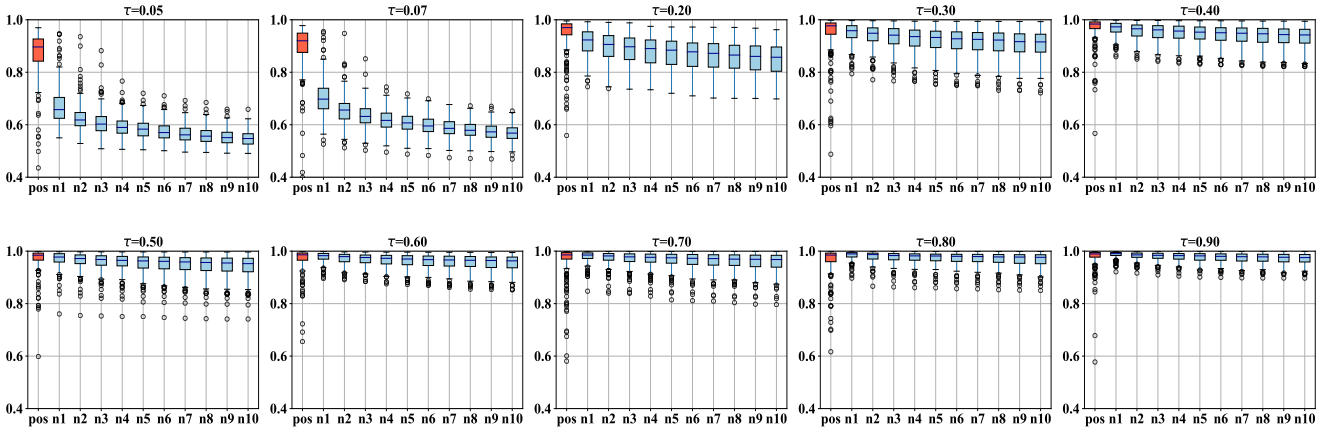

Figure 2. We display the similarity distribution of positive samples marked as 'pos' and the distributions of the top-10 nearest negative samples marked as 'ni' for the i-th nearest neighbour. All models are trained with the ordinary contrastive loss on SVHN.

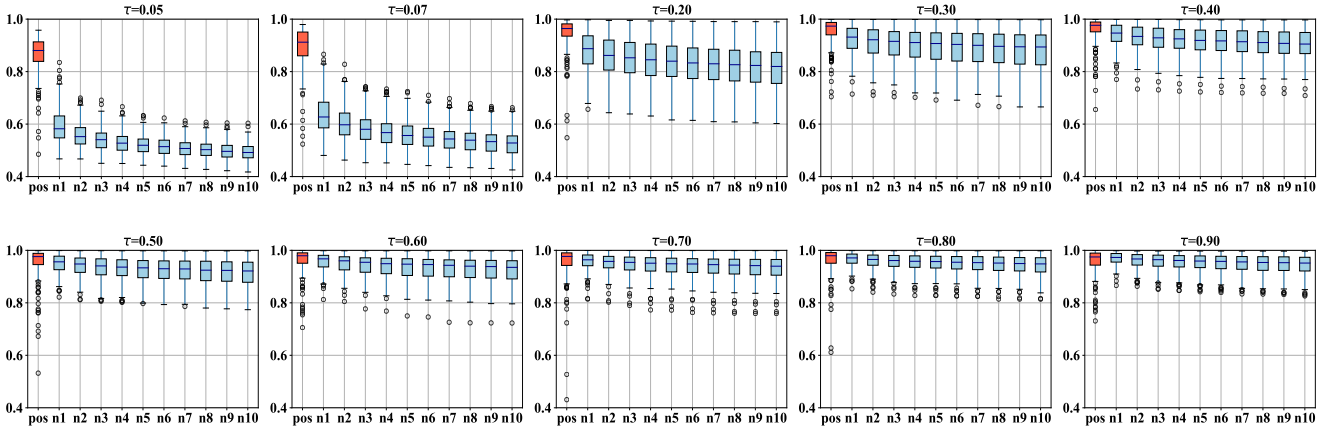

Figure 3. We display the similarity distribution of positive samples marked as 'pos' and the distributions of the top-10 nearest negative samples marked as 'ni' for the i-th nearest neighbour. All models are trained with the ordinary contrastive loss on ImageNet100.

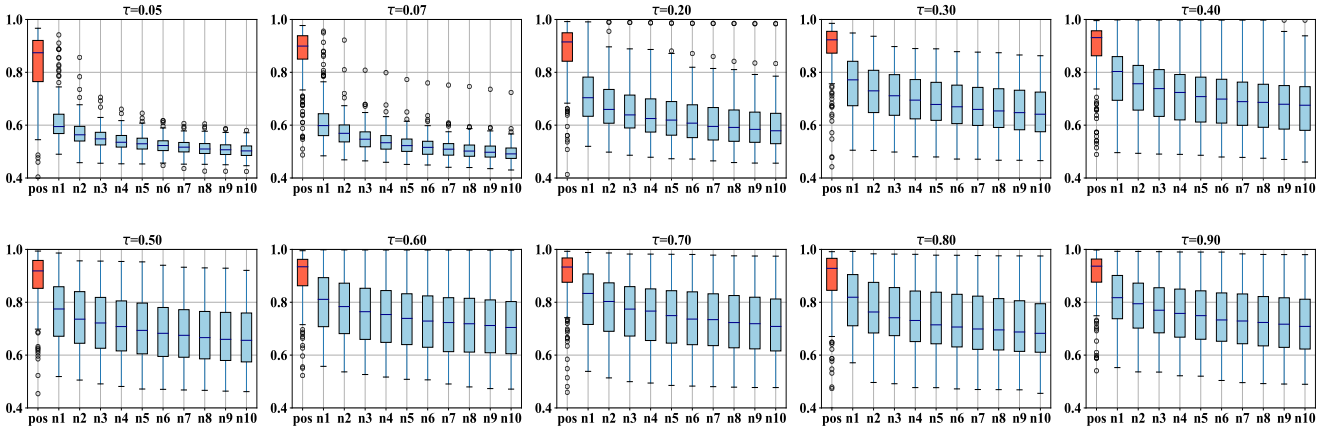

Figure 4. We display the similarity distribution of positive samples marked as 'pos' and the distributions of the top-10 nearest negative samples marked as 'ni' for the i-th nearest neighbour. All models are trained with the **hard** contrastive loss on CIFAR10.

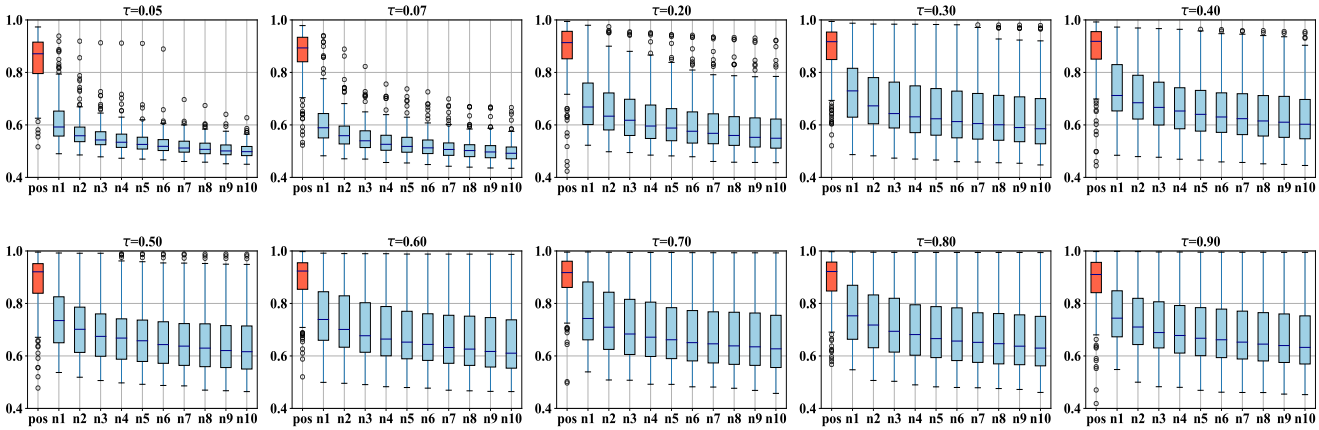

Figure 5. We display the similarity distribution of positive samples marked as 'pos' and the distributions of the top-10 nearest negative samples marked as 'ni' for the i-th nearest neighbour. All models are trained with the **hard** contrastive loss on CIFAR100.

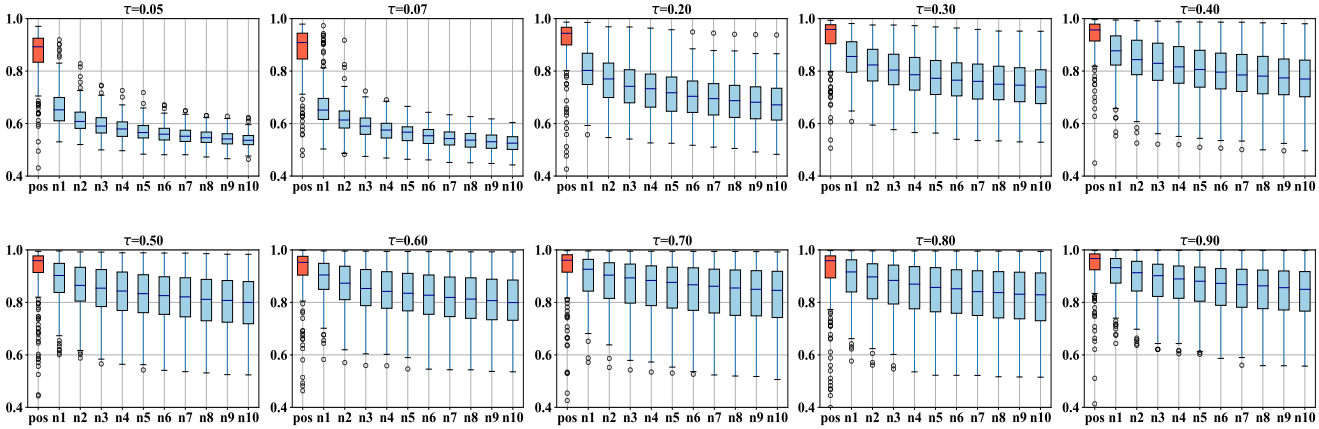

Figure 6. We display the similarity distribution of positive samples marked as 'pos' and the distributions of the top-10 nearest negative samples marked as 'ni' for the i-th nearest neighbour. All models are trained with the **hard** contrastive loss on SVHN.

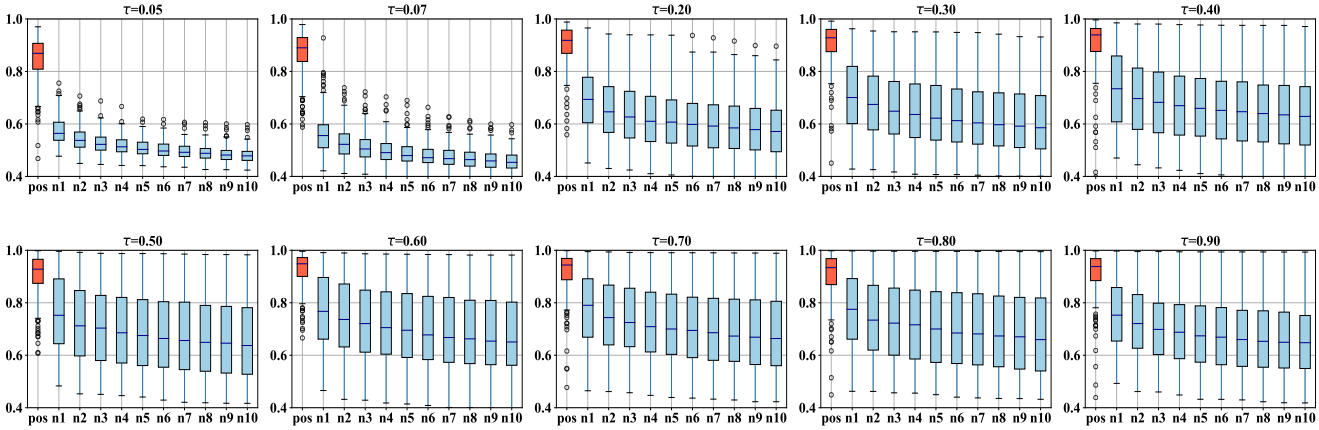

Figure 7. We display the similarity distribution of positive samples marked as 'pos' and the distributions of the top-10 nearest negative samples marked as 'ni' for the i-th nearest neighbour. All models are trained with the **hard** contrastive loss on ImageNet100.