A Generalized contrastive loss

Inspired by and built upon [16], we propose the following *abstract* form of generalized contrastive loss.

$$\mathcal{L}_{\text{generalized contrastive}} = \mathcal{L}_{\text{alignment}} + \lambda \mathcal{L}_{\text{distribution}} \tag{4}$$

Both terms are defined on hidden representations. $\mathcal{L}_{\text{alignment}}$ encourages representations of augmented views to be consistent, while $\mathcal{L}_{\text{distribution}}$ encourages representations (or a random subset of them) to match a prior distribution (of high entropy).

The standard contrastive loss is a special case of generalized contrastive loss as it can be re-written as Eq. 3. It is worth noting that τ in Eq. 3 appears in two places, one as the scaling of the second term, and the other as the width of Gaussian kernel. They do not necessarily need to be the same, so we could decouple them as follows. The decoupling allows us to study the effects of them separately.

$$\mathcal{L}^{\text{Decoupled NT-Xent}} = -\frac{1}{n}\sum_{i,j}\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j) + \lambda \frac{1}{n}\sum_{i}\log\sum_{k=1}^{2n}\mathbb{1}_{[k\neq i]}\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau) \tag{5}$$

Although the $\mathrm{LogSumExp}$ and the uniform hypersphere prior are widely used, partially due to the popularity of cross entropy loss, here we are interested in knowing *whether or not it is essential to the effectiveness of contrastive loss*. Are there other priors that could also work (e.g. those in Figure A.1)? And how much difference does it make by using other priors?

(a) (b) (c)

Figure A.1: Examples of different prior distribution in 2-D space: (a) uniform hypersphere, (b) uniform hypercube and (c) normal distribution.

One issue of using other priors is we could *not* rely on $\mathrm{LogSumExp}$ for matching the distribution. To this end, we resort to the theory of optimal transport, via Sliced Wasserstein Distance (SWD) [32, 33, 34]. For two sets of equal-sized samples from two 1-D distributions, the optimal transport can be obtained by computing two permutations that order the values of both sets of samples respectively. The 1-D Wasserstein distance can then be computed with ℓ_2 distance between the ordered values. For n-D distributions, we first project the samples to n randomly-generated orthogonal 1-D subspaces, and then compute the sum of 1-D Wasserstein distance across all 1-D subspaces. By adjusting the network weights to minimize the SWD, we are able to reduce the mismatch between the distribution of hidden vectors and a known prior distribution. The detailed algorithm can be found in Algorithm 1.

Algorithm 1 Sliced Wasserstein Distance (SWD) loss.

input: activation vectors $\boldsymbol{H} \in \mathbb{R}^{b\times d}$, a prior distribution (e.g. Gaussian) sampler $\mathcal{S}$
draw prior vectors $\boldsymbol{P} \in \mathbb{R}^{b\times d}$ using $\mathcal{S}$
generate random orthogonal matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d'}$
make projections: $\boldsymbol{H}^{\perp} = \boldsymbol{H}\boldsymbol{W}; \boldsymbol{P}^{\perp} = \boldsymbol{P}\boldsymbol{W}$
initialize SWD loss $\ell = 0$
for $j \in \{1, 2, \cdots, d'\}$ **do**
  $\ell = \ell + \|\text{sort}(\boldsymbol{H}^{\perp}_{:,j}) - \text{sort}(\boldsymbol{P}^{\perp}_{:,j})\|^2$
end for
return $\ell/(dd')$

With SWD loss, we are able to use a wider set of priors, including those in Figure A.1, and potentially more. Table 1 summarizes instantiations of the generalized contrastive loss with different prior distributions and distribution matching loss.

Connection with mutual information. The connection between the standard contrastive loss and mutual information has been shown before [3, 20], where the contrastive loss (a.k.a. InfoNCE loss [3]) is shown to be a lower bound of the mutual information. However, with the generalized contrastive loss, do we still have a similar connection?

The mutual information between two latent variables U, V can be expressed as

$$I(U;V) = H(U) - H(U|V)$$

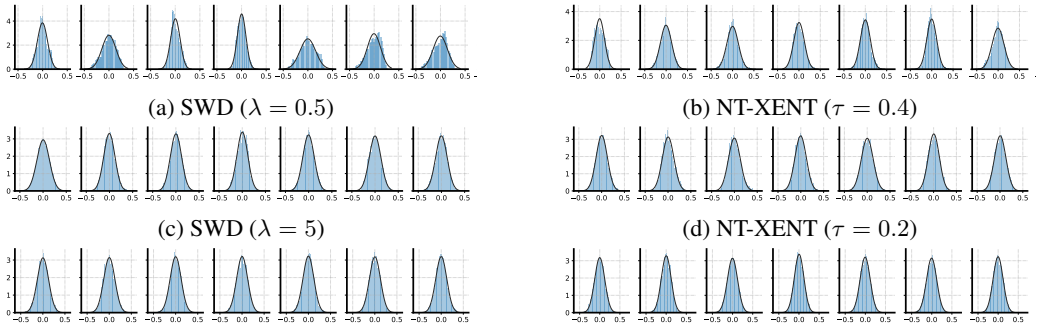
Comparing this factorization of mutual information with generalized contrastive loss, it is not difficult to see that: (1) the alignment term $\mathcal{L}_{\text{alignment}}$ is directly related to $H(U|V)$ which aims to reduce uncertainty of the other views given one view of the example; and (2) the distribution matching term $\mathcal{L}_{\text{distribution}}$ can be considered as a proxy to $H(u)$ for maximizing the entropy in the representation. In particular, for representation in the hypersphere, the entropy is maximized if they are uniformly distributed [16]. It is perhaps worth noting that different from mutual information (Eq. A), the generalized contrastive loss (Eq. 2) allows a tunable weight (λ) between the alignment and distribution matching term, whose relation with the temperature (τ) in standard contrastive loss (Eq. 1) will be discussed later.

Experimental setup. We follow [13, 14] for the use of augmentations and architectures. By default, we use ResNet-50 [35] and a 2-layer projection head [13, 14] after the ResNet's average pooling layer. We set the output ($\boldsymbol{z}$) dimensionality to 64 for CIFAR10 and 128 for ImageNet, since increasing them has little effect on the performance. The batch size and training epoch will be specified for each experiment. We use the linear evaluation protocol, i.e. the accuracy of a trained linear classifier on the learned features is used as a proxy for representation quality.

When comparing the standard contrastive loss (i.e. NT-Xent in Eq. 1) and other instantiations of the generalized contrastive loss (in Table 1), we optimize the hyper-parameters for different losses (for NT-Xent loss, we set $\tau = 0.2$; for decoupled NT-Xent loss, we set $\tau = 1.0, \lambda = 0.1$; for SWD based losses, we set $\lambda = 5$; and since we use mean squared error instead of ℓ_2 distance in alignment loss for losses in Table 1, we find it helpful to scale the loss by 1000 when the hidden vector $\boldsymbol{z}$ is normalized). A batch size of 128 is used for CIFAR-10, and 1024 is used for ImageNet.

B Temperature τ (in standard contrastive loss, Eq. 1) is (within a range) inversely correlated to weighting λ (of distribution matching term in Eq. 2)

To see how well the learned distribution matches the prior distribution (e.g. Gaussian), we randomly project the (high-dimensional) representation vectors into 1-D space and plot the histogram distribution. For prior distribution of Gaussian or uniform in hypersphere, these random projections in 1-D space should be Gaussian like.



(a) SWD ($\lambda = 0.5$) (b) NT-XENT ($\tau = 0.4$)

(c) SWD ($\lambda = 5$) (d) NT-XENT ($\tau = 0.2$)

(e) SWD ($\lambda = 50$) (f) NT-XENT ($\tau = 0.1$)

Figure B.1: Distribution of random orthogonal projection of output vectors on CIFAR-10 test set (each small plot has its own random projection direction). For SWD (uniform hypersphere) loss, distribution becomes more Gaussian as λ increases. For NT-Xent loss, the distribution becomes more Gaussian as τ decreases.

Figure B.1 shows random orthogonal projection of representation from CIFAR-10 test set. We see that both weighting (λ in Eq. 2) and the temperature scaling (τ in Eq. 1) have the effect of controlling distribution matching term, but they have an inverse correlation. In other words, using a higher temperature has similar effect as setting a larger weighting of distribution matching term.

In addition to visualize the representation statistics. We also tune τ and λ separately for the decoupled NT-xent loss (Eq. 5). Figure B.2 shows the linear evaluation of ResNet-18 trained in 200 epochs. We see that the temperature τ and the weighting $\lambda$are inversely correlated for most range. In practice one could simply fix one and tune the other.

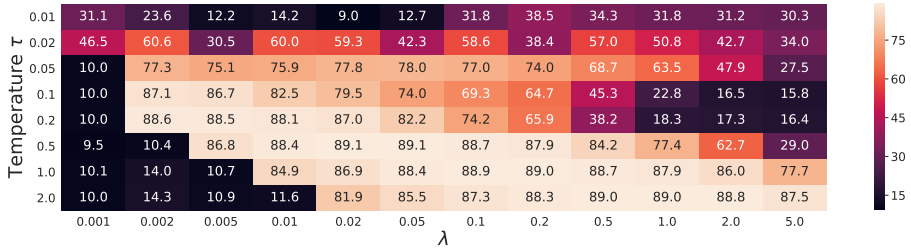
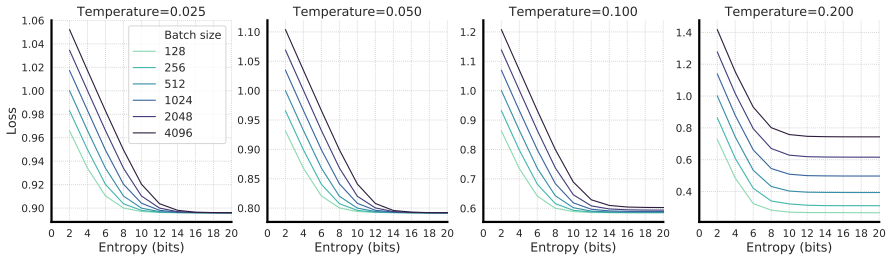
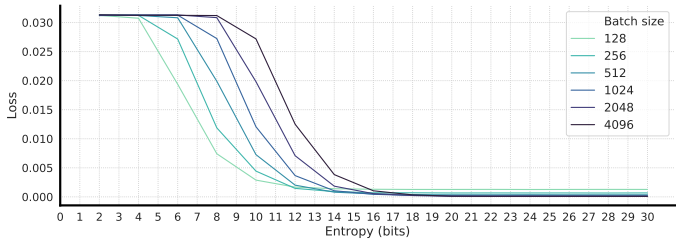


Figure B.2: Linear evaluation of ResNet-18 trained on CIFAR-10 (200 epochs) using decoupled NT-Xent loss (Eq. 5). The temperature τ and the weighting λ are mostly inverse correlated.

C Distribution matching loss, LogSumExp or SWD, saturates with a few bits of entropy



(a) LogSumExp.



(b) SWD (uniform hypersphere).

Figure C.1: Distribution matching loss saturates quickly with a few bits of entropy. The saturation varies slightly across batch sizes.

Here we study the saturation of distribution matching loss (based on LogSumExp or SWD), without presence of the alignment term. To do so, we create square images with k binary channels (instead of RGB channels), and all pixels at different locations of a 32×32 image share the same value, this allows us to use the same architecture as one for CIFAR-10 (i.e. ResNet-18 and 2-layer projection head with output dimensionality of 64). We note that this experiment can also be conducted on images of 1×1 size with other architecture. It is not difficult to see the entropy of this dataset is k bits. A mini-batch of data points (without augmentations) are first encoded via the network, and then the distribution matching loss is defined on the network's outputs. The network is trained for 400 epochs, and longer training epochs makes little difference. Figure C.1 shows that distribution loss

saturates quickly with a few bits of entropy in the dataset (same or less bits in representations), and both temperature and batch sizes have effects on the saturation behavior.

D Linear evaluation of generalized contrastive losses on CIFAR-10 and ImageNet

Table D.1, D.2 and D.3 show linear evaluation performance of ResNet-50 trained with different losses (numerical results of Figure 1). Similar to [13, 14], a square root learning rate is used. In addition, results of different batch sizes are also compared, and we find the differences are small with reasonable sizes (e.g. 128 for CIFAR-10 and 1024 for ImageNet).

Table D.1: Linear evaluation accuracy (top-1) of ResNet-50 trained with different losses on CIFAR-10.

Loss	Epoch / Batch size	100	200	400	800
NT-Xent	128	87.4	91.0	93.0	93.9
	256	88.0	91.3	93.0	93.6
	512	87.9	91.3	92.9	93.7
	1024	88.2	91.2	92.7	93.3
Decoupled NT-Xent	128	87.8	91.0	93.0	94.0
	256	87.7	91.1	92.8	93.6
	512	87.5	91.3	92.7	93.6
	1024	87.5	91.0	92.6	93.7
SWD (normal)	128	86.3	90.5	92.8	93.8
	256	86.2	90.8	93.1	94.1
	512	85.0	90.7	92.9	94.1
	1024	83.3	89.9	93.0	93.9
SWD (uniform hypercube)	128	85.1	90.1	92.6	93.4
	256	84.6	89.9	92.9	93.8
	512	83.1	89.8	92.8	93.8
	1024	81.3	88.3	92.2	93.6
SWD (uniform hypersphere)	128	87.0	90.9	92.9	93.8
	256	87.1	90.9	92.5	93.7
	512	86.6	90.8	92.9	93.4
	1024	86.0	90.3	92.5	93.2

E Extra results on CIFAR-10 with random bits added

Figure E.1 shows linear evaluation on CIFAR-10 with different random bits added trained with a wider range of batch sizes. It is worth noting that the bits (in the x-axis) are calculated based on the total size of uniform integer distribution. However, this is an overestimation of actual bits as due to collision in generated integers.

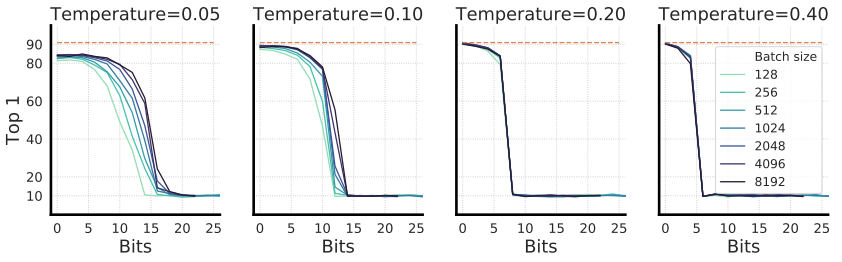
We observe that the linear evaluation accuracy decreases quickly with a few bits of the extra channel competing feature added. And this detrimental effect on the representation quality cannot be avoided by different contrastive loss functions, batch sizes, or memory mechanism in momentum contrast [10]. Although a smaller temperature (τ) or larger weighting (λ) slightly mitigate the degeneration effect, its baseline performance when no extra bits are added is also worse. With less than 15 bits of competing features added, the representation quality degenerates to the level where RGB channels are completely ignored.

Table D.2: Linear evaluation accuracy (top-1) of ResNet-50 trained with different losses on ImageNet (with 2-layer projection head).

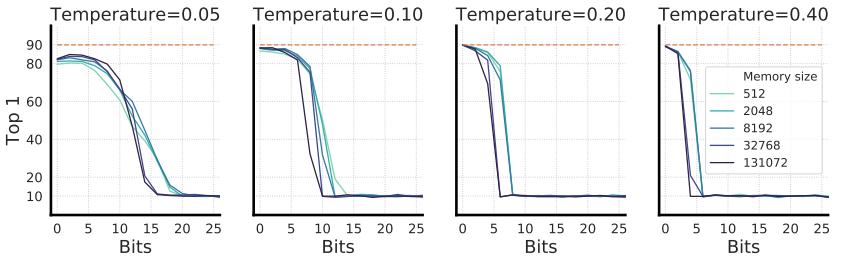
Loss	Epoch / Batch size	100	200	400	800
NT-Xent	512	65.4	67.3	68.7	69.3
	1024	65.6	67.6	68.8	69.8
	2048	65.3	67.6	69.0	70.1
Decoupled NT-Xent	512	65.8	67.6	68.9	69.5
	1024	66.0	67.9	69.0	70.1
	2048	65.8	67.9	69.3	70.2
SWD (normal)	512	64.9	66.8	68.0	69.0
	1024	65.0	67.1	68.2	69.3
	2048	65.0	66.9	68.4	69.7
SWD (uniform hypercube)	512	64.3	66.4	67.8	68.7
	1024	64.2	66.5	67.9	68.9
	2048	63.9	66.6	67.9	69.0
SWD (uniform hypersphere)	512	65.6	67.7	69.0	70.0
	1024	65.8	67.9	69.0	69.6
	2048	65.6	67.8	69.2	69.8

Table D.3: Linear evaluation accuracy (top-1) of ResNet-50 trained with different losses on ImageNet (with 3-layer projection head).

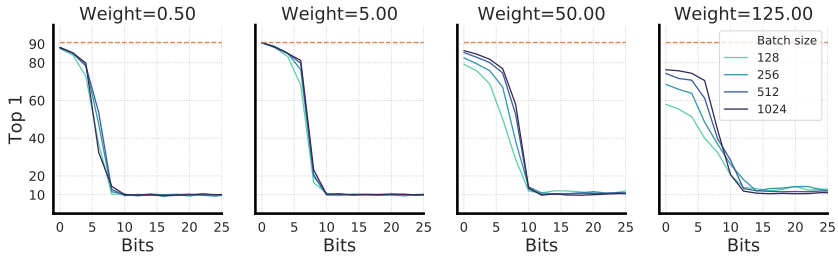
Loss	Epoch / Batch size	100	200	400	800
NT-Xent	512	66.6	68.4	70.0	71.0
	1024	66.8	68.9	70.1	70.9
	2048	66.8	69.1	70.4	71.3
Decoupled NT-Xent	512	66.8	68.4	69.6	70.6
	1024	66.6	68.9	69.9	70.8
	2048	66.6	69.0	70.1	70.8
SWD (normal)	512	66.5	68.4	69.8	70.8
	1024	66.6	68.8	70.1	71.1
	2048	66.7	69.1	70.2	71.1
SWD (uniform hypercube)	512	66.1	68.3	69.7	70.7
	1024	66.3	68.5	70.0	71.3
	2048	65.8	68.2	70.1	71.1
SWD (uniform hypersphere)	512	66.5	68.3	69.5	70.5
	1024	66.6	68.6	69.8	70.8
	2048	66.5	68.7	70.2	70.9



(a) Standard NT-Xent



(b) NT-Xent with Momentum Contrast (MoCo)



(c) SWD (uniform hypersphere)

Figure E.1: Linear evaluation accuracy on CIFAR-10 of ResNet-18 (400 epochs) when different random bits are added. Different contrastive losses and batch sizes are compared.