矩阵信息论视角下的表征学习

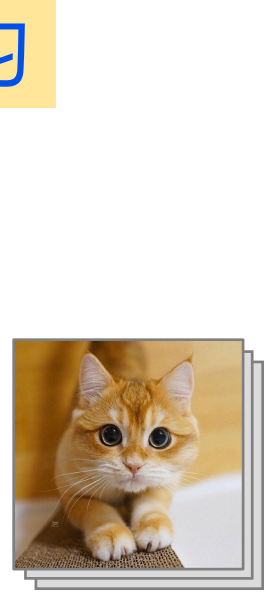
黄维然

上海交通大学

2025.07 JCSDS 杭州

研究背景

监督学习



特征提取网络

特征向量

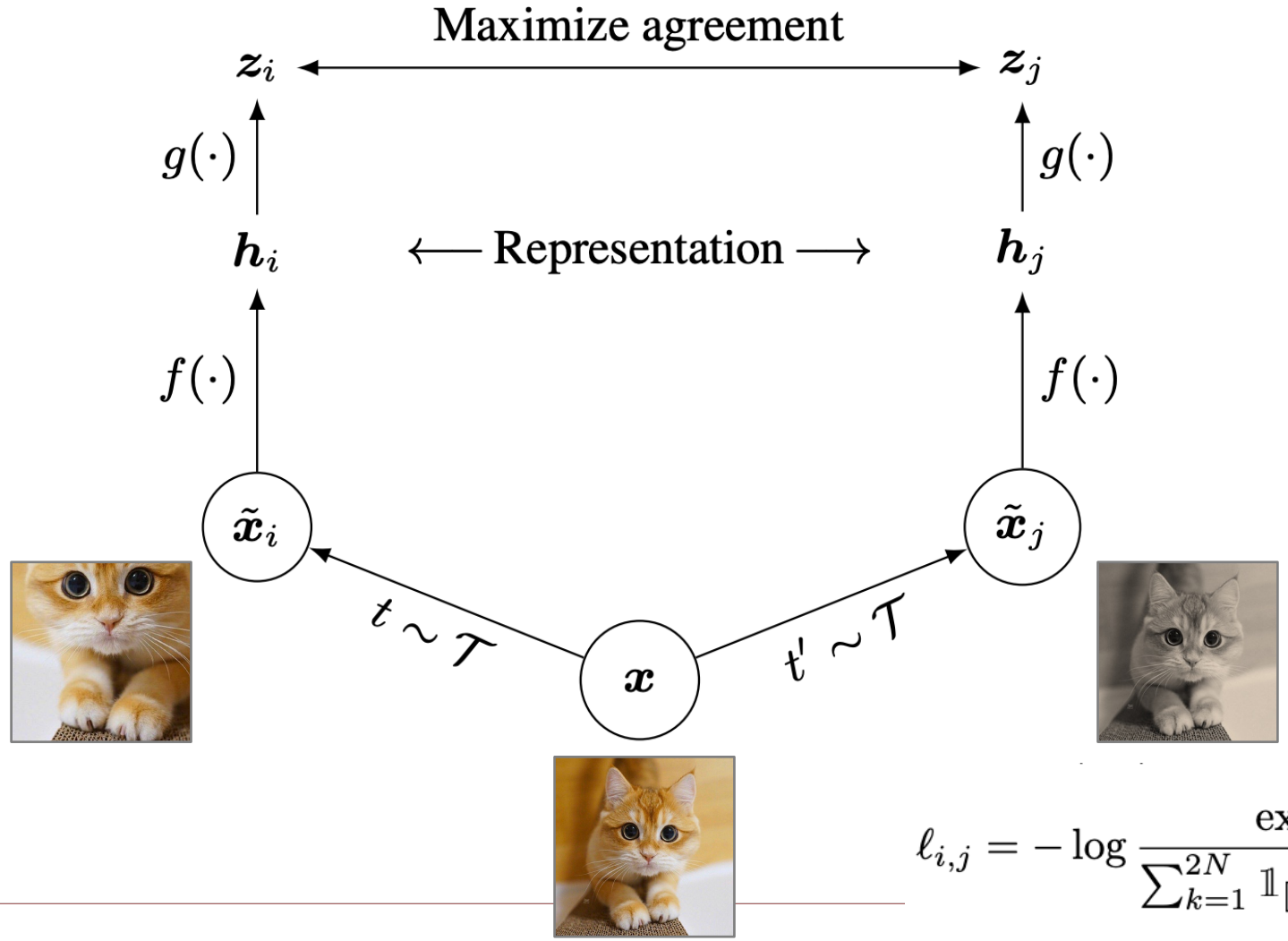
分类网络

“猫”

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log p_i$$

研究背景

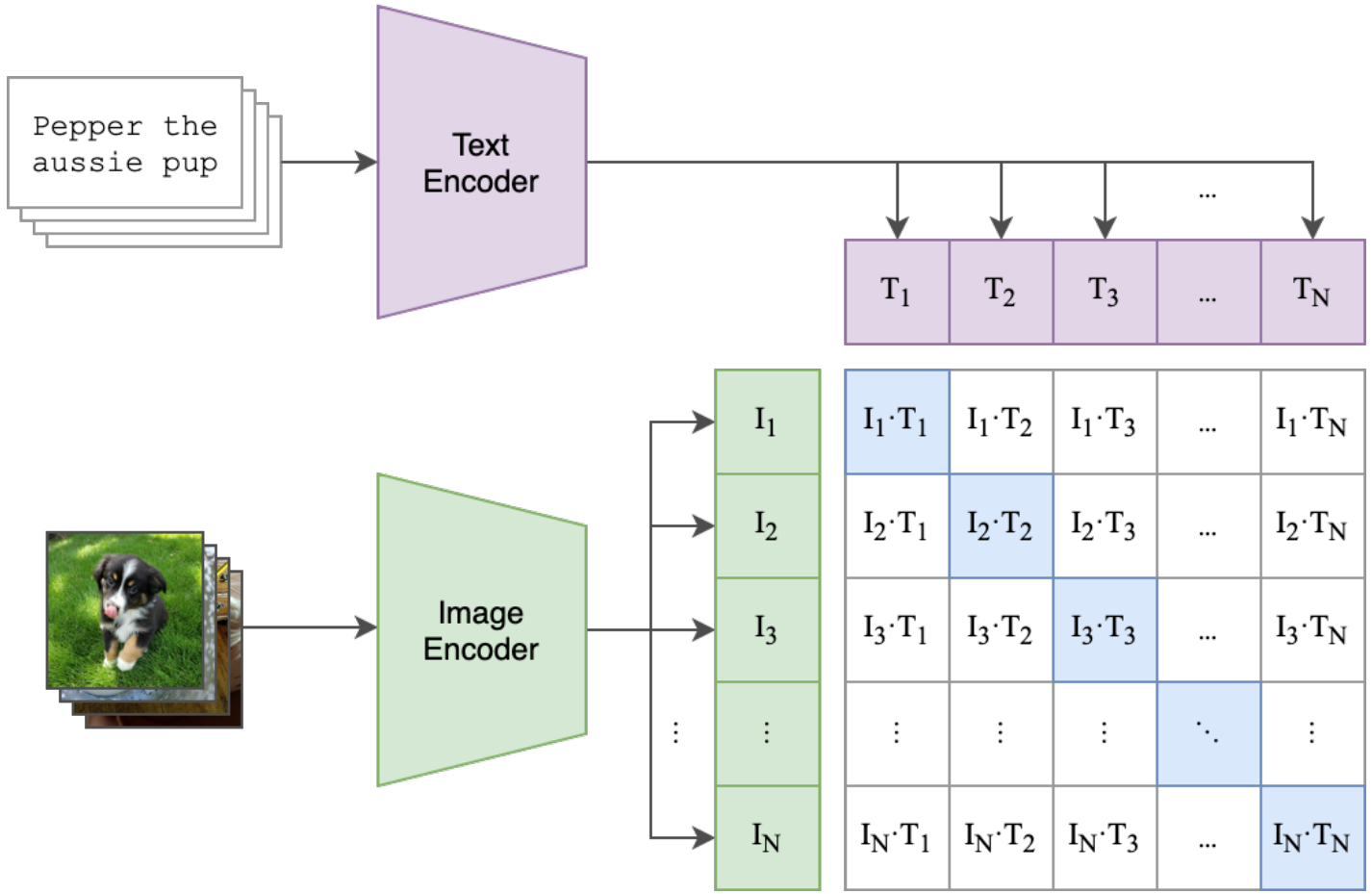
自监督学习



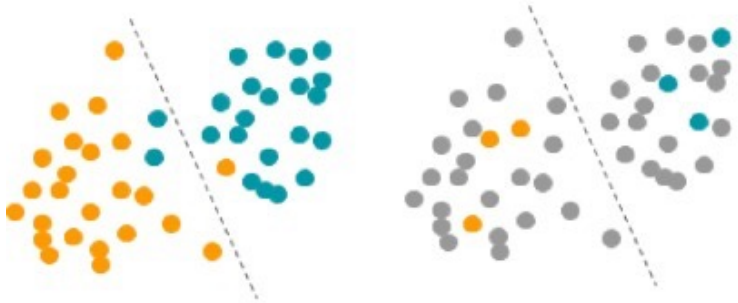
$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

研究背景

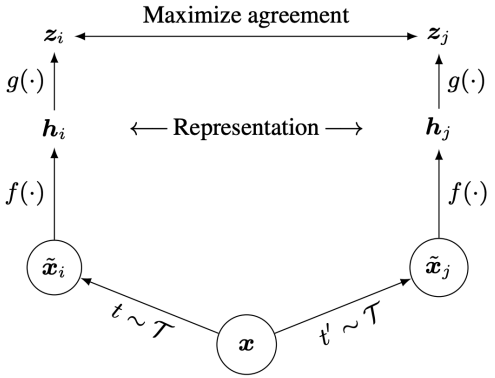
多模态学习


Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1
I_2
I_3
I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

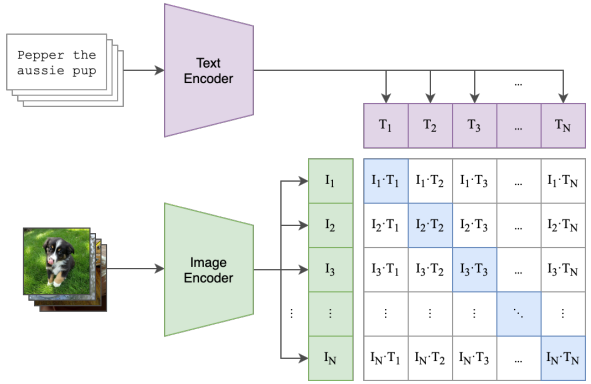
研究背景



(半)监督学习
类别可分性



自监督学习
正样本对齐



多模态学习
跨模态对齐

能否用一个统一的视角理解这些学习过程?

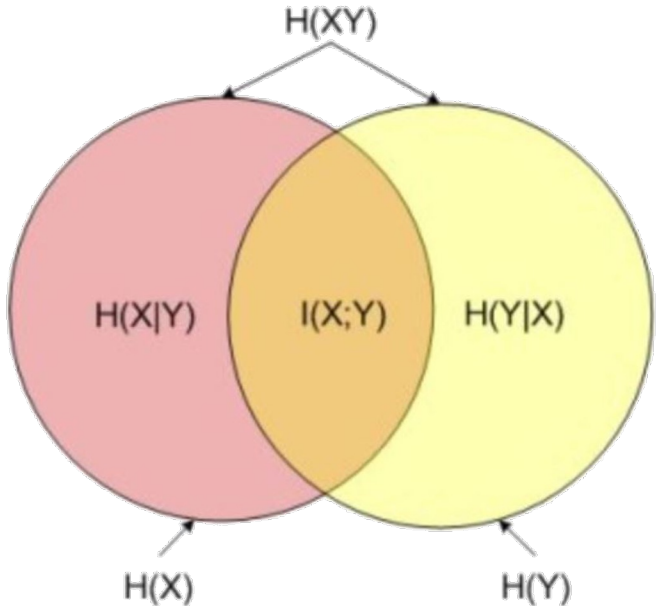
信息熵

信息熵

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i))$$

互信息

$$\mathrm{I}(X;Y) = -\sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$



（相互独立的时候互信息为 0）

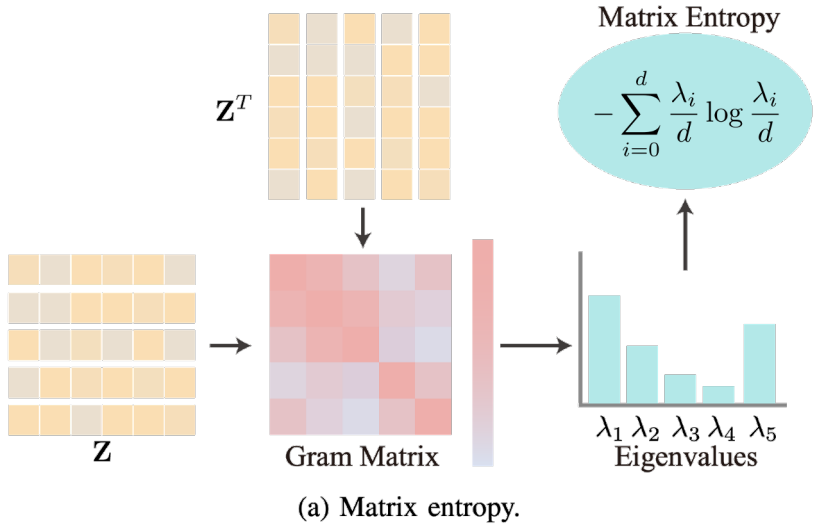
定义在概率分布上，无法用在表征上

矩阵信息熵

矩阵信息熵

$$\mathrm{H}(\mathbf{K}) = -\mathrm{tr}\left(\frac{\mathbf{K}}{d}\log\frac{\mathbf{K}}{d}\right) = -\sum_{i=1}^{d}\frac{\lambda_i}{d}\log\frac{\lambda_i}{d}$$

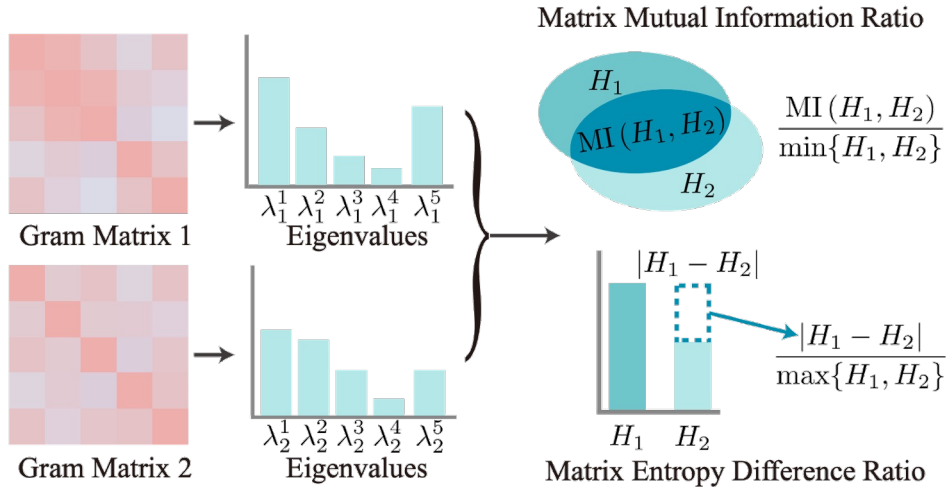
(d 是样本数)



(a) Matrix entropy.

矩阵互信息

$$\mathrm{MI}(\mathbf{K}_1;\mathbf{K}_2) = \mathrm{H}(\mathbf{K}_1) + \mathrm{H}(\mathbf{K}_2) - \mathrm{H}(\mathbf{K}_1 \odot \mathbf{K}_2)$$



Skean, Oscar, et al. "Dime: Maximizing mutual information by a difference of matrix-based entropies." arXiv preprint arXiv:2301.08164 (2023).

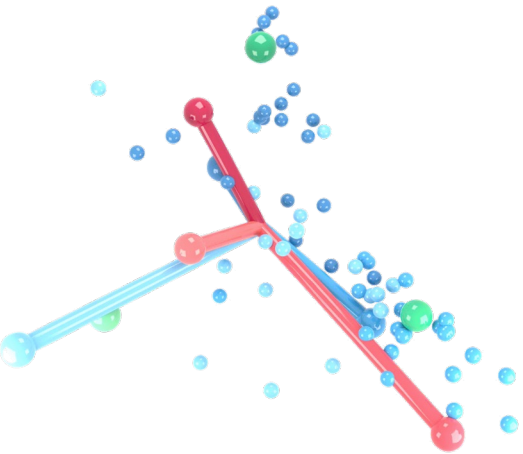
Neural Collapse

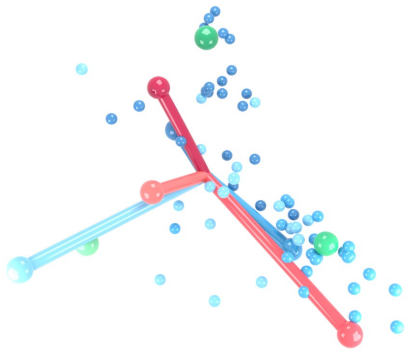
Neural Collapse 是指在深度神经网络训练的终端阶段（在充分大且平衡的数据集上达到零训练误差时），最后一层特征呈现的特殊几何结构。具体来说，它包含以下几个关键特性：

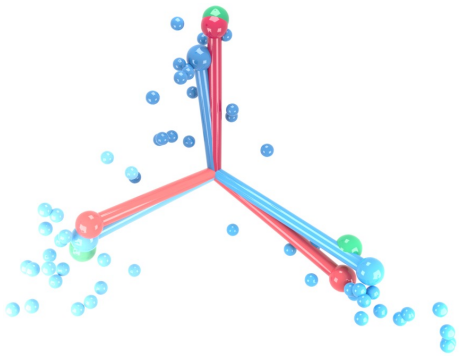
1、随着训练进行，最后一层特征坍缩在类均值

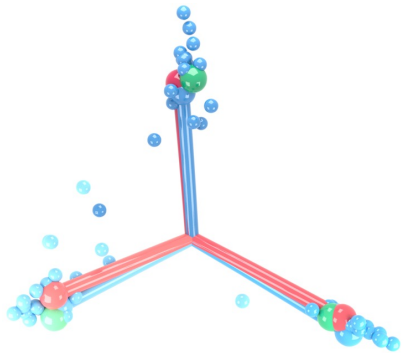
2、类均值收敛到单纯形 ETF 顶点

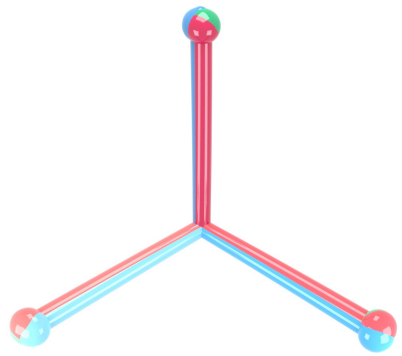
3、线性分类器权重$[W_1, W_2, \ldots, W_C]$与类别均值对齐









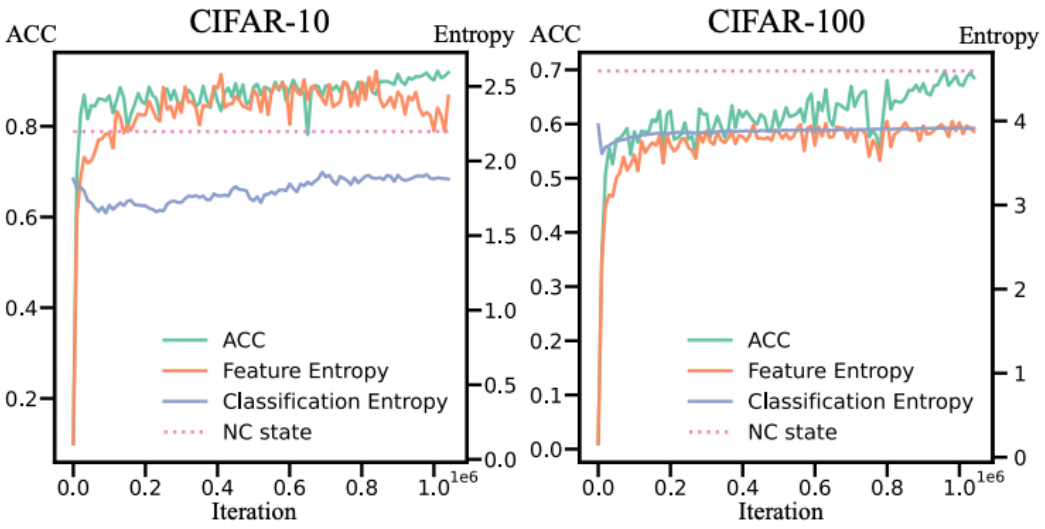


(特征向量归一化)

监督学习中的矩阵熵

当NC发生时，分类器权重，类中心的矩阵信息熵

$$\mathrm{H}\left(\mathbf{W}\mathbf{W}^{T}\right)=\mathrm{H}\left(\mathbf{f}\mathbf{f}^{\mathrm{T}}\right)=\log(C-1)$$



(a) CIFAR-10 (b) CIFAR-100

训练过程中矩阵熵的变化

随机初始化的模型提取的特征具有低秩的结构，信息熵较低。随着训练进行，模型提取的特征逐渐具有的结构信息，熵逐渐增加。

随机初始化的分类头方向各异，构成的Gram矩阵几乎满秩。初始信息熵较高。随着训练进行，分类头像样本特征对齐，矩阵熵逐渐降低。随着训练继续进行，分类头逐渐对齐ETF特征结构。矩阵熵继续升高。

单看特征或者分类器权重的矩阵熵都不是很充分

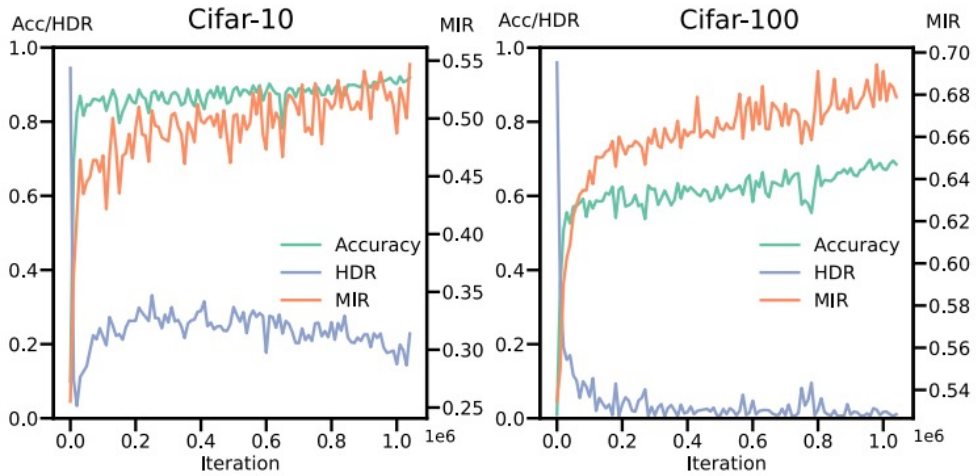
监督学习中的矩阵熵

当NC发生时，我们定义样本表征和分类器权重的矩阵互信息率

$$\mathrm{MIR}\left(\mathbf{f}\mathbf{f}^{\mathrm{T}}, \mathbf{W}\mathbf{W}^{\boldsymbol{T}}\right) = \frac{\mathrm{MI}(\mathbf{f}\mathbf{f}^{\mathrm{T}}, \mathbf{W}\mathbf{W}^{\boldsymbol{T}})}{\min(\mathrm{H}(\mathbf{f}\mathbf{f}^{\mathrm{T}}), \mathrm{H}(\mathbf{W}\mathbf{W}^{\boldsymbol{T}}))} = \frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-1)}$$

当NC发生时，我们定义样本表征和分类器权重的矩阵熵差率

$$\mathrm{HDR}\left(\mathbf{f}\mathbf{f}^{\mathrm{T}}, \mathbf{W}\mathbf{W}^{\boldsymbol{T}}\right) = \frac{\left|\mathrm{H}\left(\mathbf{f}\mathbf{f}^{\mathrm{T}}\right) - \mathrm{H}\left(\mathbf{W}\mathbf{W}^{\boldsymbol{T}}\right)\right|}{\max(\mathrm{H}(\mathbf{f}\mathbf{f}^{\mathrm{T}}), \mathrm{H}(\mathbf{W}\mathbf{W}^{\boldsymbol{T}}))} = 0$$

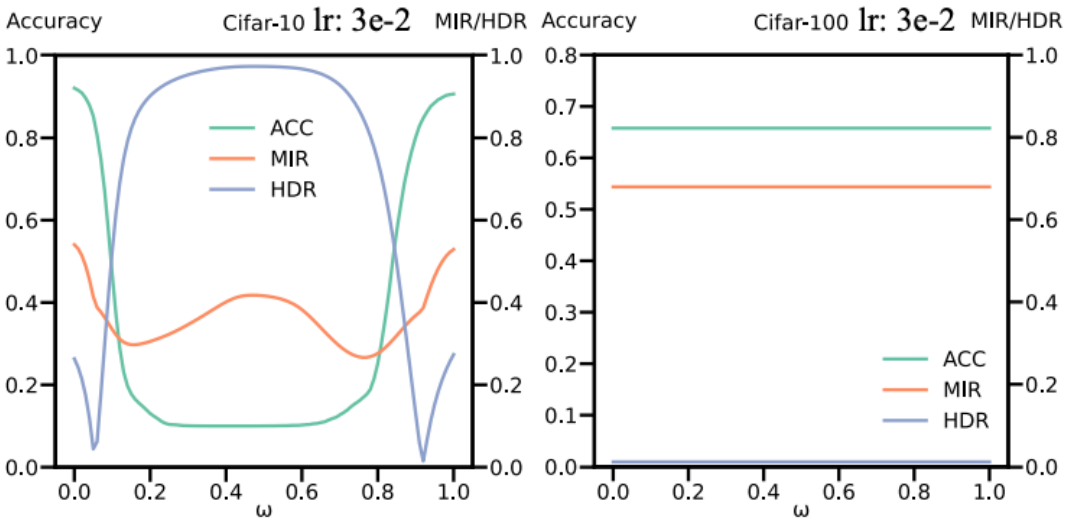


(a) CIFAR-10 (b) CIFAR-100

训练过程中互信息率和矩阵熵差率的变化

随着训练的进行，样本表征和分类器权重的矩阵互信息率逐渐升高直到接近理论值；矩阵熵差率逐渐接近于0。

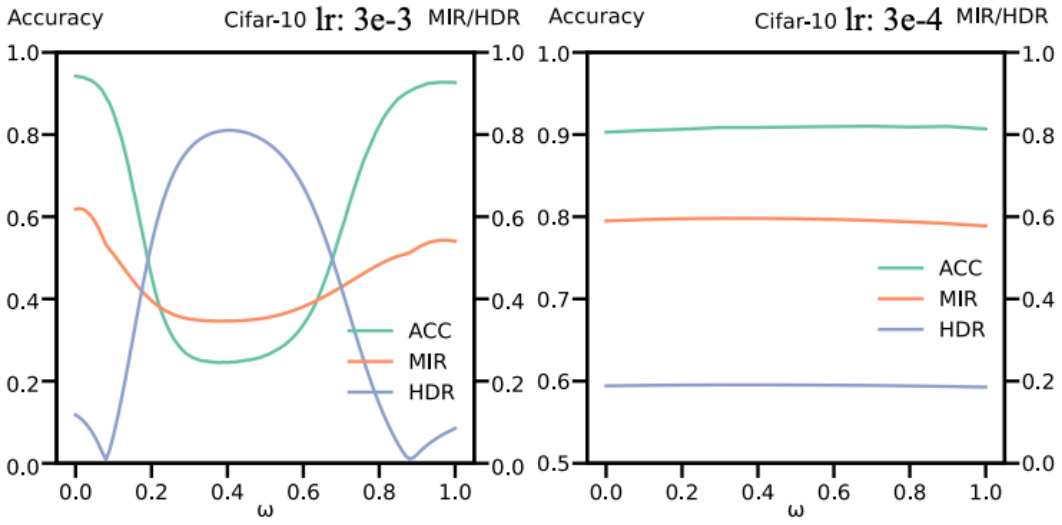
监督学习中的矩阵熵



(a) CIFAR-10 (b) CIFAR-100

Linear Mode Connectivity 分析

$$h = \omega h_1 + (1 - \omega) h_2$$



(a) lr: $3e^{-3}$ (b) lr: $3e^{-4}$

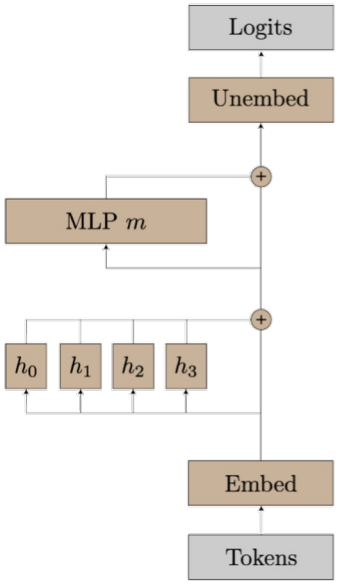
学习率对 LMC 的影响

监督学习中的矩阵熵

Grokking：

模加法运算任务

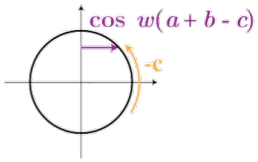
（mod 113）



Computes logits using further trig identities:

$\text{Logit}(c) \propto \cos(w(a+b-c))$

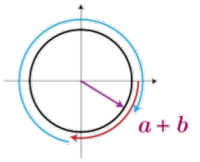
$= \cos(w(a+b))\cos(wc) + \sin(w(a+b))\sin(wc)$



Calculates sine and cosine of $a+b$ using trig identities:

$\sin(w(a+b)) = \sin(wa)\cos(wb) + \cos(wa)\sin(wb)$

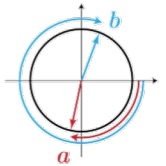
$\cos(w(a+b)) = \cos(wa)\cos(wb) - \sin(wa)\sin(wb)$

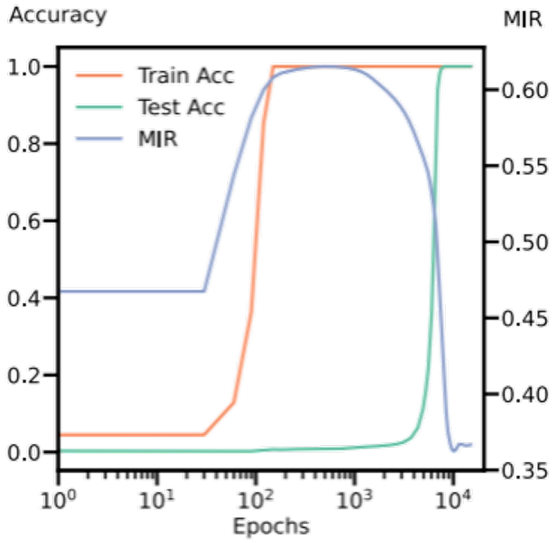


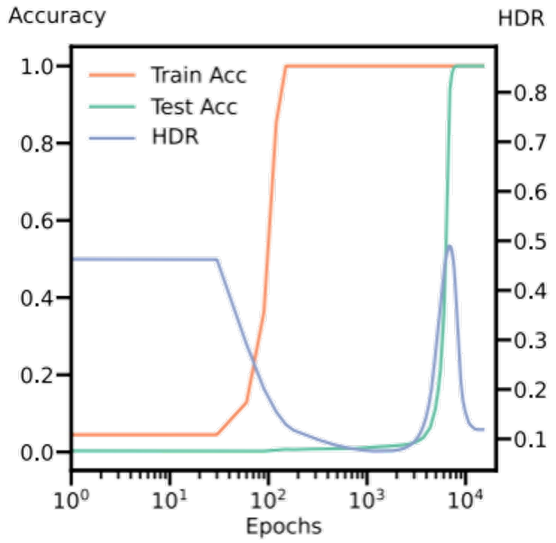
Translates one-hot a, b to Fourier basis:

$a \to \sin(wa), \cos(wa)$

$b \to \sin(wb), \cos(wb)$







监督学习中的矩阵熵

设计为正则项

TABLE II: Error rates (100% - accuracy) on CIFAR-10/100, and STL-10 datasets for state-of-the-art methods in semi-supervised learning. Bold indicates the best performance, and underline indicates the second best.

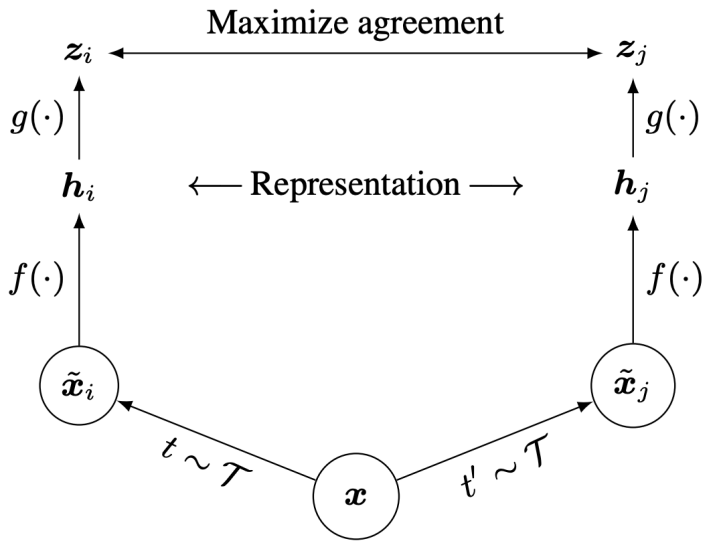
Dataset	CIFAR-10			CIFAR-100		STL-10	
# Label	10	40	250	400	2500	40	1000
Π Model [51]	79.18±1.11	74.34±1.76	46.24±1.29	86.96±0.80	58.80±0.66	74.31±0.85	32.78±0.40
Pseudo Label [52]	80.21± 0.55	74.61±0.26	46.49±2.20	87.45±0.85	57.74±0.28	74.68±0.99	32.64±0.71
VAT [53]	79.81± 1.17	74.66±2.12	41.03±1.79	85.20±1.40	48.84±0.79	74.74±0.38	37.95±1.12
MeanTeacher [54]	76.37± 0.44	70.09±1.60	37.46±3.30	81.11±1.44	45.17±1.06	71.72±1.45	33.90±1.37
MixMatch [32]	65.76± 7.06	36.19±6.48	13.63±0.59	67.59±0.66	39.76±0.48	54.93±0.96	21.70±0.68
ReMixMatch [55]	20.77± 7.48	9.88±1.03	6.30±0.05	42.75±1.05	26.03±0.35	32.12±6.24	6.74±0.17
UDA [56]	34.53± 10.69	10.62±3.75	5.16±0.06	46.39±1.59	27.73±0.21	37.42±8.44	6.64±0.17
FixMatch [25]	24.79± 7.65	7.47±0.28	5.07±0.05	46.42±0.82	28.03±0.16	35.97±4.14	6.25±0.33
Dash [57]	27.28± 14.09	8.93±3.11	5.16±0.23	44.82±0.96	27.15±0.22	34.52±4.30	6.39±0.56
MPL [58]	23.55± 6.01	6.93±0.17	5.76±0.24	46.26±1.84	27.71±0.19	35.76±4.83	6.66±0.00
FlexMatch [26]	13.85± 12.04	4.97±0.06	4.98±0.09	39.94±1.62	26.49±0.20	29.15±4.16	5.77±0.18
FreeMatch [29]	8.07± 4.24	4.90±0.04	4.88±0.18	37.98±0.42	26.47±0.20	15.56±0.55	5.63±0.15
OTMatch [28]	4.89± 0.76	4.72±0.08	4.60±0.15	37.29±0.76	26.04±0.21	**12.10±0.72**	5.60±0.14
SoftMatch [27]	4.91± 0.12	4.82±0.09	**4.04±0.02**	37.10±0.07	26.66±0.25	21.42±3.48	5.73±0.24
FreeMatch + MAX MI (Ours)	4.87± 0.66	4.66± 0.13	4.56± 0.15	**36.41± 1.91**	**25.77± 0.35**	16.61± 1.19	**5.24 ± 0.17**
FreeMatch + MIN HD (Ours)	**4.69± 0.16**	**4.63± 0.25**	4.60± 0.15	37.31± 1.96	25.79± 0.41	14.93 ± 3.28	5.30 ± 0.18

TABLE III: Results for fully supervised learning

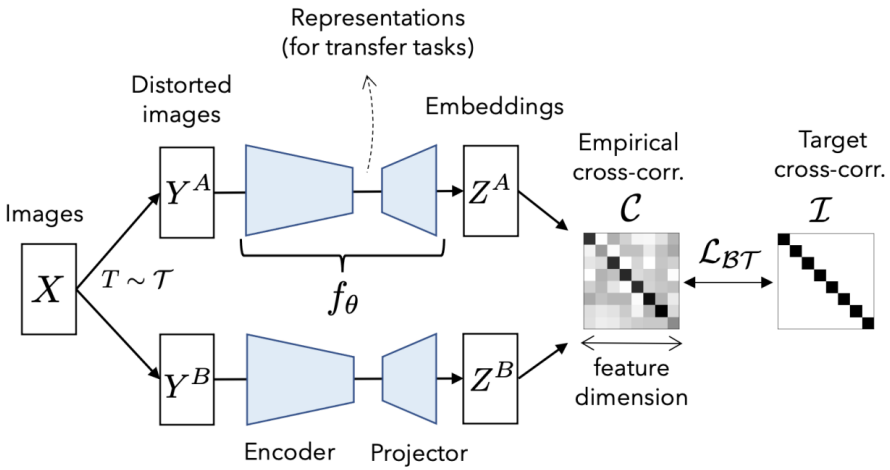
Method	CIFAR-10	CIFAR-100
Fully supervised	95.35	80.77
Ours (MAX MI)	95.52	80.81
Ours (MIN HD)	**95.57**	**80.96**

在标注样本数较少的情况下，对矩阵熵的优化能够更有效的提升模型的性能

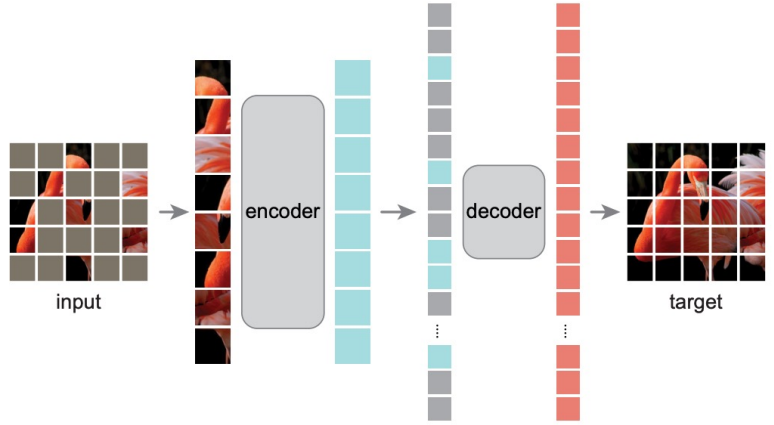
自监督学习中的矩阵熵



SimCLR



Barlow Twins



MAE

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2}\left(\sum_{i=1}^{B}\log\frac{\exp\left((\mathbf{z}_i^{(1)})^\top \mathbf{z}_i^{(2)}\right)}{\sum_{j=1}^{B}\exp\left((\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(2)}\right)} + \sum_{i=1}^{B}\log\frac{\exp\left((\mathbf{z}_i^{(2)})^\top \mathbf{z}_i^{(1)}\right)}{\sum_{j=1}^{B}\exp\left((\mathbf{z}_i^{(2)})^\top \mathbf{z}_j^{(1)}\right)}\right).$$

$$\mathcal{L}_{BT} = \sum_{i=1}^{d}(1-\mathcal{C}_{ii})^2 + \lambda\sum_{i=1}^{d}\sum_{j\neq i}\mathcal{C}_{ij}{}^2$$

重建损失MSE

$$\mathcal{L}_{\text{U-MAE}} = \mathcal{L}_{\text{MAE}} + \gamma\sum_{i\neq j}(\mathbf{z}_i^\top \mathbf{z}_j)^2$$

自监督学习中的矩阵熵

对于双塔模型，每一路都有表征Z_1, Z_2，考虑它们之间的矩阵互信息

$$\mathrm{MI}\left(Z_1^T Z_1; Z_2^T Z_2\right)$$

Theorem 4.3. *1. For the spectral contrastive loss, we have*

$$\mathrm{I}_2(\mathbf{Z}_1^\top \mathbf{Z}_1; \mathbf{Z}_2^\top \mathbf{Z}_2) \geq \log B - 2\log(1 + (2 + \frac{2}{B\lambda})\mathcal{L}_{SC}).$$

2. For the Barlow Twins' loss, we have

$$\mathrm{I}_2(\bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top; \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top) \geq \log d - 2\log(1 + \frac{2}{d\lambda}\mathcal{L}_{BT} + 4\sqrt{d\mathcal{L}_{BT}}).$$

优化损失=
最大化矩阵互信息的下界

自监督学习中的矩阵熵

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

那 MAE 呢？ MAE的损失并没有隐式地优化矩阵熵！

总编码率 TCR

$$\mathrm{TCR}_{\mu}(\mathbf{Z}) = \log\det(\mu\mathbf{I}_d + \mathbf{Z}\mathbf{Z}^{\mathrm{T}})$$

$$\mathrm{TCR}_{\mu}(\mathbf{Z}) = d\log(1+\mu) - \mathrm{KL}\left(\mathbf{I}_d, \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{\mu}}(\boldsymbol{\mu}\mathbf{I}_d + \mathbf{K})\right)$$

TCR是信息熵的下界

$$\mathrm{H}(\mathbf{K}) \geq \frac{1}{d}\mathrm{TCR}_{\mu}(\mathrm{K}) + \mathrm{const}$$

（TCR 的计算只涉及到矩阵的行列式运算，相对于严格计算矩阵熵所需的特征值分解，行列式的计算代价明显更低，也更易于优化。）

自监督学习中的矩阵熵

$$L_{M-MAE} = L_{MAE} - \lambda \mathrm{TCR}_{\mu}(\mathbf{Z})$$

$$\begin{aligned}
\mathcal{L}_{\text{M-MAE}} &= \mathcal{L}_{\text{MAE}} - \lambda \cdot \log\det(\mathbf{I}_d + \frac{1}{\mu}\mathbf{Z}\mathbf{Z}^\top) + \text{Const.} \\
&= \mathcal{L}_{\text{MAE}} - \lambda \cdot \log\det(\mathbf{I}_B + \frac{1}{\mu}\mathbf{Z}^\top\mathbf{Z}) + \text{Const.} \\
&= \mathcal{L}_{\text{MAE}} - \lambda \cdot \mathrm{tr}\log(\mathbf{I}_B + \frac{1}{\mu}\mathbf{Z}^\top\mathbf{Z}) + \text{Const.} \\
&= \mathcal{L}_{\text{MAE}} - \lambda \cdot \mathrm{tr}(\frac{1}{\mu}\mathbf{Z}^\top\mathbf{Z} - \frac{1}{2\mu^2}(\mathbf{Z}^\top\mathbf{Z})^2 + \cdots) \\
&= \mathcal{L}_{\text{U-MAE}} + \text{Higher-order-terms} + \text{Const.}
\end{aligned}$$

U-MAE 是 M-MAE 的一个二阶估计

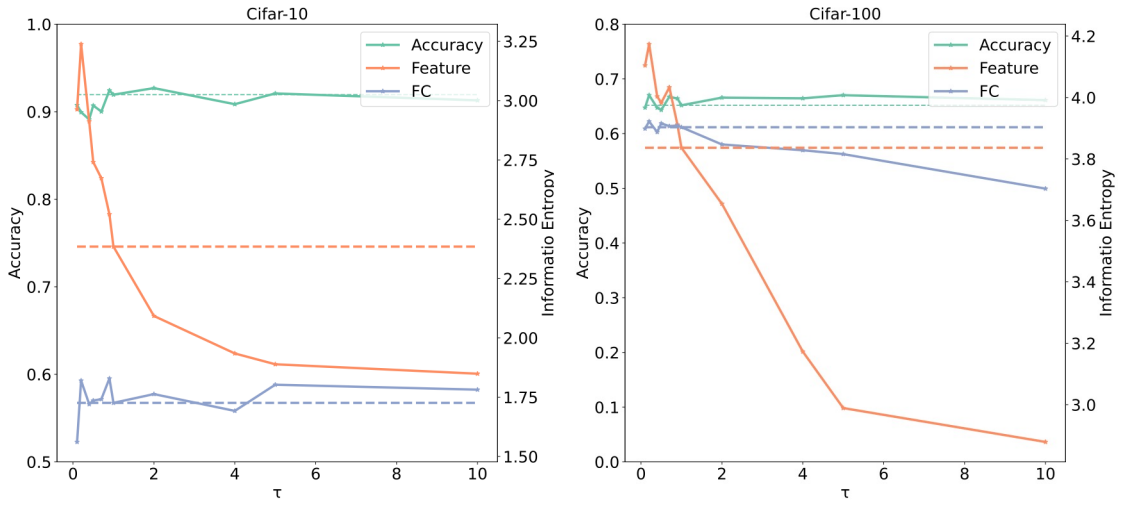
Table 1. Linear evaluation accuracy (%) and fine-tuning accuracy (%) of pretrained models by MAE loss, U-MAE loss, and M-MAE loss with different ViT backbones on ImageNet-1K. The uniformity regularizer TCR loss in the M-MAE loss significantly improves the linear evaluation performance and fine-tuning performance of the MAE loss.

Downstream Task	Method	ViT-Base	ViT-Large
Linear Probing	MAE	55.4	62.2
	U-MAE	58.5	65.8
	M-MAE	**62.4**	**66.0**
Fine-tuning	MAE	82.9	83.3
	U-MAE	83.0	83.2
	M-MAE	**83.1**	**84.3**

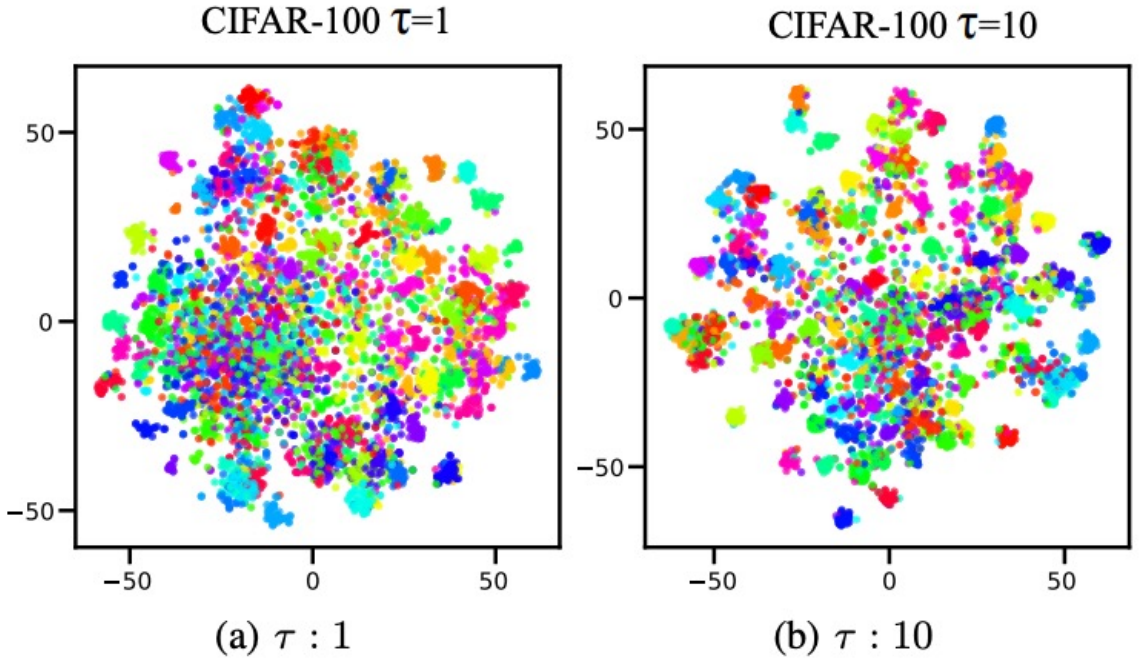
M-MAE: "How mask matters: Towards theoretical understandings of masked autoencoders." NeurIPS 2022.

跨模态对齐的矩阵熵

1. 同类样本矩阵熵越低，说明样本特征越集中。
2. 对于整个数据集，在监督学习中，降低矩阵熵也会提升类内特征聚类效果

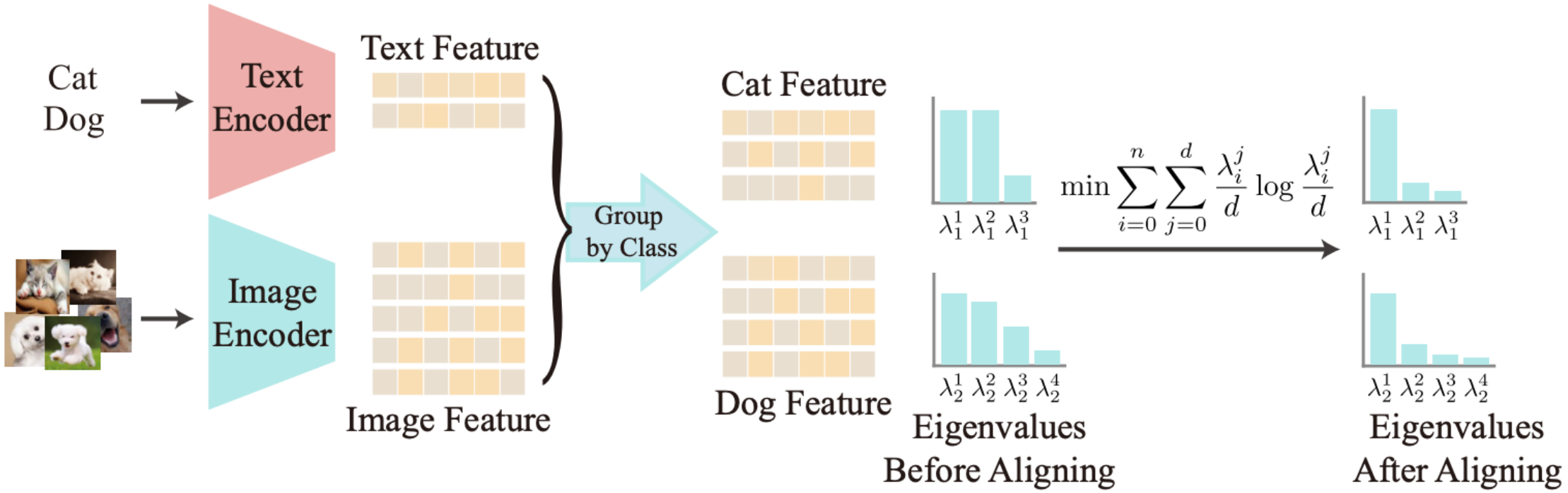


温度系数对特征的矩阵熵的影响



不同温度的特征可视化

跨模态对齐的矩阵熵



跨模态特征对齐损失

跨模态对齐的矩阵熵

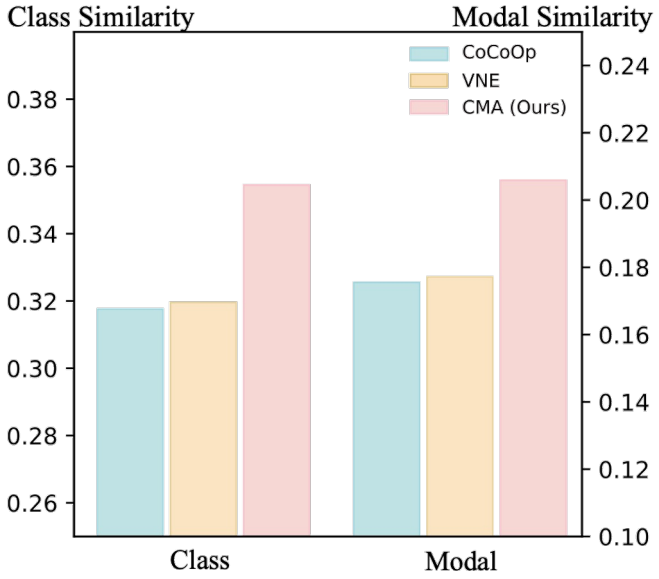
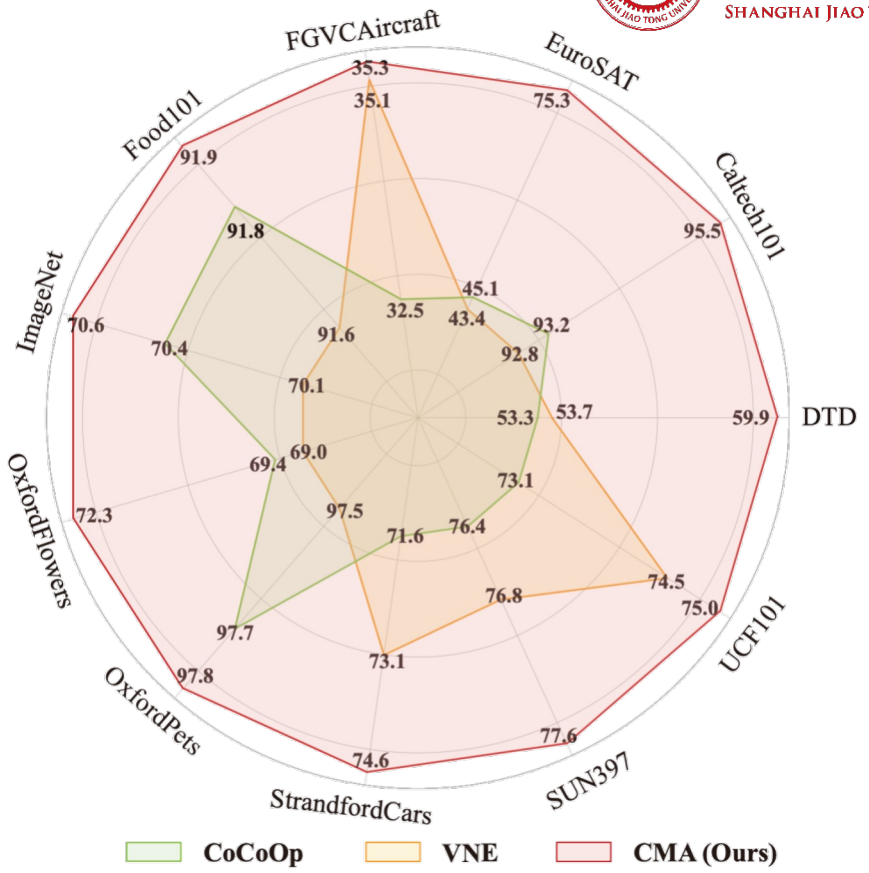


TABLE I: Performance of novel classes under different λ. When $\lambda = 0$, it indicates the performance of vanilla CoCoOp.

Dataset	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Caltech101 [42]	93.2	92.5	92.8	93.2	93.0	93.2	**95.5**	93.3	93.8	92.5	88.5
DTD [45]	52.3	56.3	56.3	54.5	59.8	58.9	57.6	54.0	52.2	**59.9**	56.5
EuroSAT [46]	45.1	47.2	47.8	44.2	55.9	53.1	59.6	59.6	66.1	**75.3**	62.8
FGVCAircraft [47]	32.5	33.7	32.9	34.0	33.5	**35.3**	33.4	30.7	30.2	31.3	23.5
Food101 [49]	91.8	91.4	91.8	91.6	91.6	91.6	**91.9**	91.8	91.6	87.8	80.4
ImageNet [39]	70.4	**70.6**	70.4	70.1	70.3	70.0	69.7	69.3	68.7	61.1	46.7
OxfordFlowers [43]	69.4	71.2	72.3	68.9	61.7	**71.9**	70.2	69.1	70.5	65.5	58.4
OxfordPets [48]	97.7	**97.8**	97.7	97.6	97.4	**97.8**	**97.8**	96.7	97.7	97.3	84.8
StanfordCars [40]	71.6	72.9	74.3	73.1	**74.6**	**74.6**	73.3	73.5	74.0	66.7	57.5
SUN397 [44]	76.4	76.8	77.1	76.4	77.2	77.3	77.5	77.0	**77.6**	72.1	61.5
UCF101 [41]	73.1	72.8	72.6	72.9	74.5	73.4	**75.0**	74.8	69.9	72.0	58.2



在数据集类别数较少的的情况下，对矩阵熵的优化能够更有效的提升模型的对齐性能

总结

- 在监督学习/半监督学习中，最优样本表征与分类头权重应呈现低熵差高互信息的关系。
- 在自监督学习中，可以用表征矩阵熵来统一理解三种不同的学习范式。
- 在多模对齐中，可以通过约束类内样本表征的跨模态矩阵熵来对齐不同模态的信息。

涉及论文：

[1] Unveiling the dynamics of information interplay in supervised learning. ICML 2024.
[2] Information Flow in Self-Supervised Learning. ICML 2024.
[3] Exploring Information-Theoretic Metrics Associated with Neural Collapse in Supervised Training. Arxiv 2025.
[4] Diff-eRank: A Novel Rank-Based Metric for Evaluating Large Language Models. NeurIPS 2024.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

谢谢观看

饮水思源 爱国荣校