组会汇报

汇报人：王子晨

文献

Inducing Neural Collapse to a Fixed Hierarchy-Aware Frame for Reducing Mistake Severity

Tong Liang, Jim Davis

Ohio State University

Columbus, Ohio 43210

Abstract

There is a recently discovered and intriguing phenomenon called Neural Collapse: at the terminal phase of training a deep neural network for classification, the within-class penultimate feature means and the associated classifier vectors of all flat classes collapse to the vertices of a simplex Equiangular Tight Frame (ETF). Recent work has tried to exploit this phenomenon by fixing the related classifier weights to a pre-computed ETF to induce neural collapse and maximize the separation of the learned features when training with imbalanced data. In this work, we propose to fix the linear classifier of a deep neural network to a Hierarchy-Aware Frame (HAFrame), instead of an ETF, and use a cosine similarity-based auxiliary loss to learn hierarchy-aware penultimate features that collapse to the HAFrame. We demonstrate that our approach reduces the mistake severity of the model's predictions while maintaining its top-1 accuracy on several datasets of varying scales with hierarchies of heights ranging from 3 to 12.

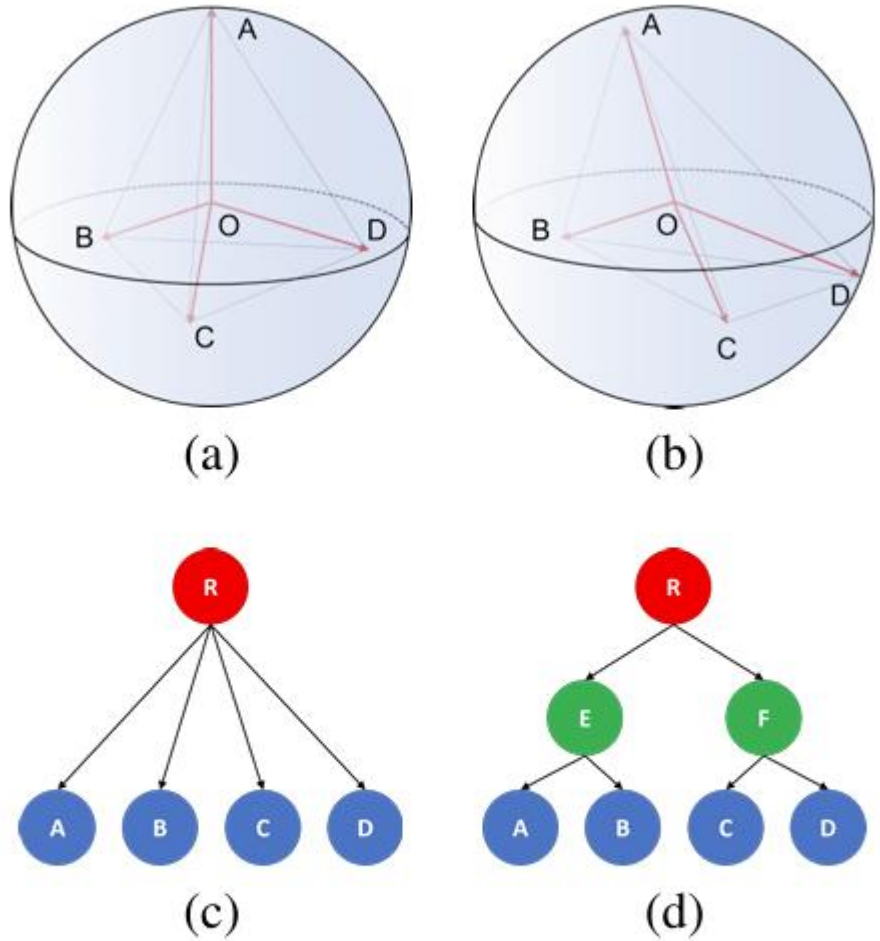

Figure 1. Illustration of a hierarchy-agnostic ETF (a) and a hierarchy-aware HAFrame (b) of four leaf classes and their hierarchies (c) and (d), respectively. All leaf classes in ETF have the same hierarchical distance.

Method

一、计算类别之间的“层级距离”

$$d_{ij} = height(LCA(y_i, y_j)) \tag{1}$$

二、“层级距离”转换成“相似度”

$$S_{ij} = (1 - s_{min}) \cdot e^{-\gamma \cdot \frac{d_{ij}}{d_{max}}} + s_{min} \tag{2}$$

三、分类器权重构造

$$cos\angle(\boldsymbol{w}_i, \boldsymbol{w}_j) = \boldsymbol{w}_i^T \boldsymbol{w}_j = S_{ij}, \forall 1 \leq i \leq j \leq K \tag{3}$$

$$\boldsymbol{S} = \boldsymbol{W}^T \boldsymbol{W} \tag{4}$$

$$\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^T = (\boldsymbol{Q}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{U}^T)(\boldsymbol{U}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{Q}^T) = \boldsymbol{W}^T\boldsymbol{W} \tag{5}$$

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{Q}^T \tag{6}$$

四、额外的转换层

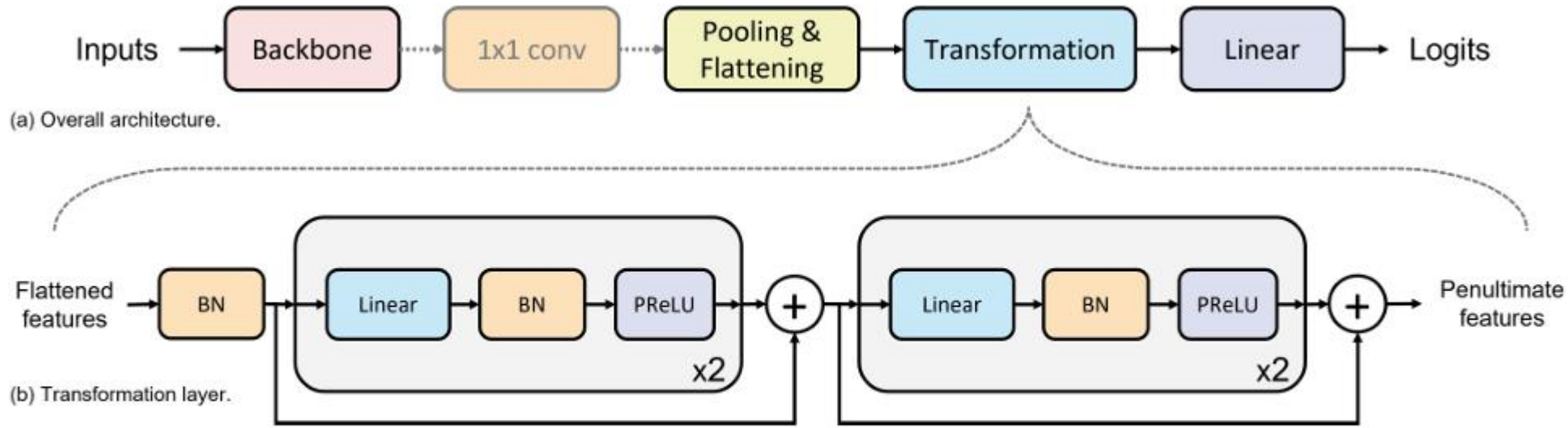

Figure 3. Illustration of the customized network architectures. (a) Top: overall network architecture of our approach, the 1x1 convolutional layer is only used in our type-II models. (b) Bottom: the proposed transformation layer, where BN is 1D batch norm layer.

Dataset	Height	Classes	Train	Val	Test
FGVC-Aircraft	3	100	3,334	3,333	3,333
CIFAR-100	5	100	45,000	5,000	10,000
iNaturalist2019	7	1010	187,385	40,121	40,737
tieredImageNet-H	12	608	425,600	15,200	15,200

Table 1. Statistics of the four datasets used in our experiments.

Model	Method	Top-1 Accuracy ↑	Mistake Severity ↓	HierDist@1 ↓	HierDist@5 ↓	Hierdist@20 ↓
Type-I	cross-entropy	79.18 +/- 0.5511	2.12 +/- 0.0240	0.44 +/- 0.0097	2.10 +/- 0.0033	2.67 +/- 0.0040
	CRM [24]	79.30 +/- 0.5250	2.08 +/- 0.0201	0.43 +/- 0.0091	1.74 +/- 0.0040	2.44 +/- 0.0015
	Flamingo [7]	81.00 +/- 0.5873	2.04 +/- 0.0343	0.39 +/- 0.0072	2.06 +/- 0.0041	2.65 +/- 0.0018
	HAFeature [17]	73.23 +/- 0.6085	2.48 +/- 0.0937	0.66 +/- 0.0152	2.10 +/- 0.0126	2.61 +/- 0.0078
Type-II	cross-entropy	79.58 +/- 0.2727	2.15 +/- 0.0159	0.44 +/- 0.0067	2.11 +/- 0.0055	2.67 +/- 0.0034
	CRM [24]	79.62 +/- 0.2953	2.13 +/- 0.0109	0.43 +/- 0.0058	1.75 +/- 0.0043	2.45 +/- 0.0022
	Flamingo [7]	80.02 +/- 0.7886	2.10 +/- 0.0373	0.42 +/- 0.0168	2.08 +/- 0.0058	2.66 +/- 0.0031
	HAFeature [17]	74.39 +/- 0.7813	2.53 +/- 0.0334	0.65 +/- 0.0226	2.10 +/- 0.0064	2.61 +/- 0.0041
	HAFrame (ours)	80.49 +/- 0.4692	2.02 +/- 0.0381	0.39 +/- 0.0039	1.74 +/- 0.0027	2.45 +/- 0.0024

Table 2. Experiment results on FGVC-Aircraft dataset. The details of type-I and type-II models are included in the training config.

Model	Method	Top-1 Accuracy ↑	Mistake Severity ↓	HierDist@1 ↓	HierDist@5 ↓	Hierdist@20 ↓
Type-I	cross-entropy	77.65 +/- 0.2635	2.34 +/- 0.0271	0.52 +/- 0.0102	2.25 +/- 0.0084	3.19 +/- 0.0045
	CRM [24]	77.63 +/- 0.2800	2.30 +/- 0.0255	0.51 +/- 0.0093	1.11 +/- 0.0077	2.18 +/- 0.0028
	Flamingo [7]	77.91 +/- 0.5733	2.31 +/- 0.0179	0.51 +/- 0.0137	2.07 +/- 0.0198	3.08 +/- 0.0094
	HAFeature [17]	77.49 +/- 0.3391	2.24 +/- 0.0158	0.51 +/- 0.0084	1.43 +/- 0.0108	2.64 +/- 0.0105
Type-II	cross-entropy	76.45 +/- 0.2207	2.43 +/- 0.0235	0.57 +/- 0.0106	2.35 +/- 0.0049	3.30 +/- 0.0030
	CRM [24]	76.48 +/- 0.2278	2.38 +/- 0.0175	0.56 +/- 0.0095	1.15 +/- 0.0074	2.20 +/- 0.0029
	Flamingo [7]	75.19 +/- 0.3188	2.31 +/- 0.0270	0.57 +/- 0.0043	2.42 +/- 0.0161	3.29 +/- 0.0105
	HAFeature [17]	76.44 +/- 0.1560	2.26 +/- 0.0290	0.53 +/- 0.0055	1.71 +/- 0.0130	2.84 +/- 0.0143
	HAFrame (ours)	77.71 +/- 0.2319	2.21 +/- 0.0108	0.49 +/- 0.0066	1.11 +/- 0.0018	2.18 +/- 0.0013

Table 3. Experiment results on CIFAR-100 dataset. The details of type-I and type-II models are included in the training config.

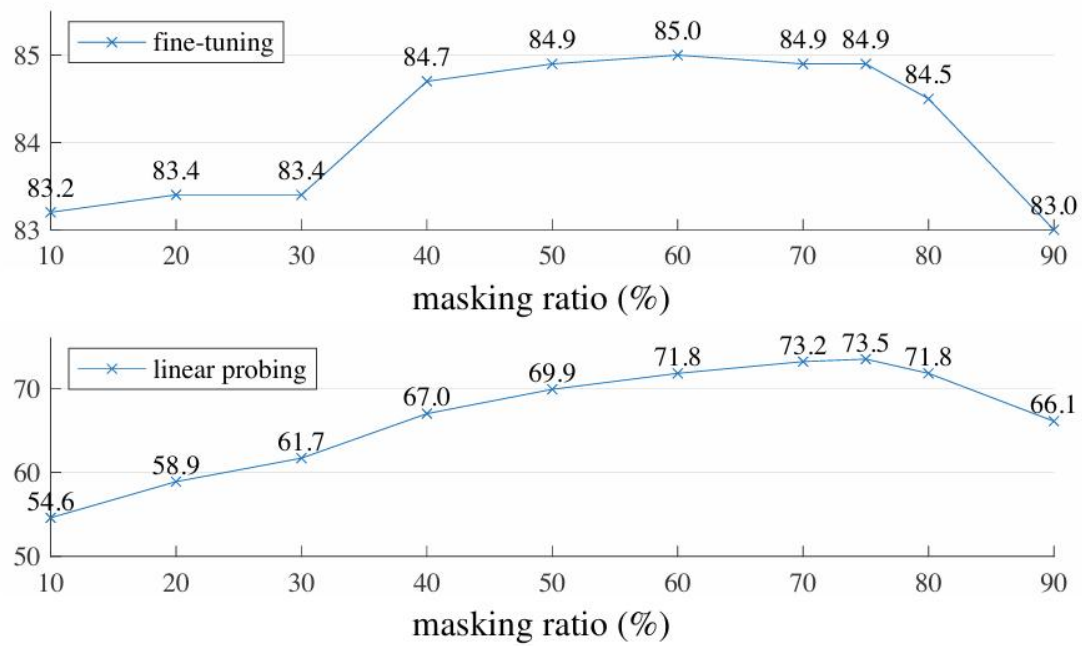

Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

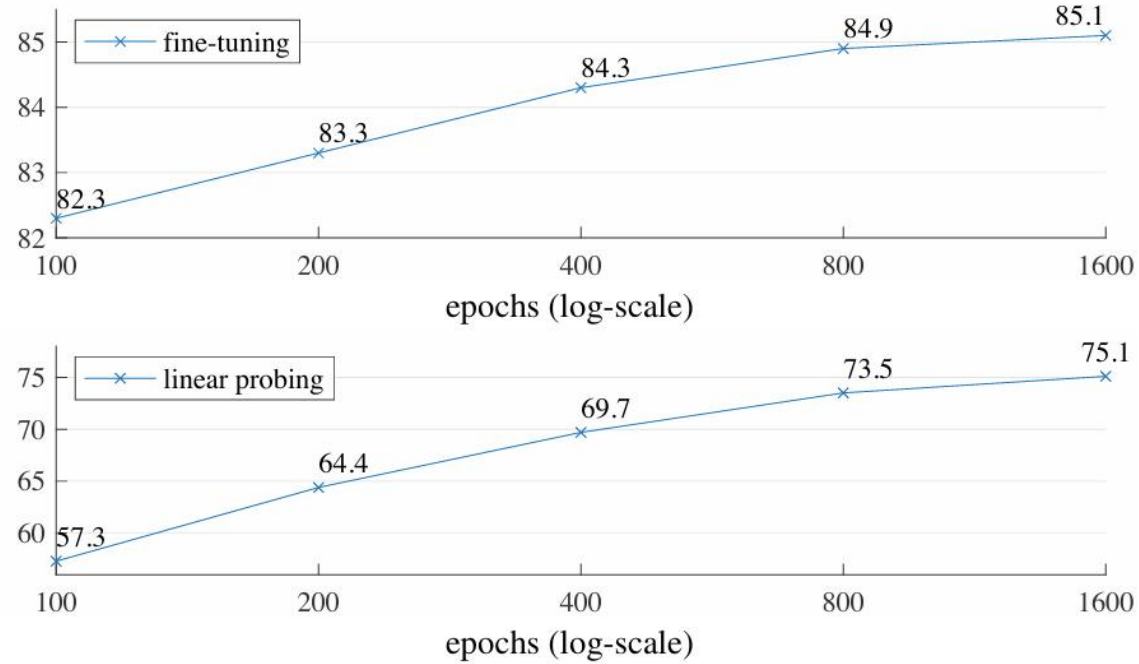

Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.

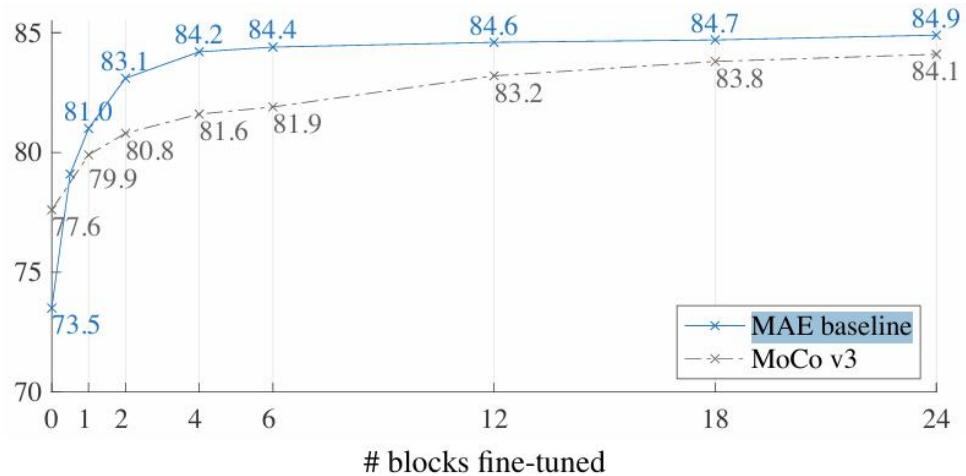

Figure 9. **Partial fine-tuning** results of ViT-L w.r.t. the number of fine-tuned Transformer blocks under the default settings from Table 1. Tuning 0 blocks is linear probing; 24 is full fine-tuning. Our MAE representations are less linearly separable, but are consistently better than MoCo v3 if one or more blocks are tuned.

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

Table 3. **Comparisons with previous results on ImageNet-1K**. The pre-training data is the ImageNet-1K training set (except the tokenizer in BEiT was pre-trained on 250M DALLE data [50]). All self-supervised methods are evaluated by end-to-end fine-tuning. The ViT models are B/16, L/16, H/14 [16]. The best for each column is underlined. All results are on an image size of 224, except for ViT-H with an extra result on 448. Here our MAE reconstructs normalized pixels and is pre-trained for 1600 epochs.

Contribution

- **简单高效的自监督预训练方法：** 提出了掩码自编码器（*MAE*），一种通过随机遮蔽输入图像并重建缺失像素的简单自监督方法。
- **创新的不对称架构设计：** 设计了只在编码器中处理未遮蔽图像块的编码器–解码器架构，从而大幅减少计算量，并通过轻量级解码器完成图像重构。
- **高比例遮蔽策略：** 发现采用高达 *75%* 的遮蔽比例可以形成一个非平凡的自监督任务，既促进了训练效率，也提升了模型性能。
- **优异的扩展性和泛化能力：** 实验结果证明，*MAE* 在仅使用 *ImageNet-1K* 数据的情况下即可训练出高容量、泛化良好的模型，并在多个下游任务（如目标检测、语义分割和图像分类）中超过了传统的监督预训练方法。

Thanks