拓扑神经坍缩

Nicholas Enbar-Salo[1],**Jeova Farias Sales Rocha Neto**[1]

[1]鲍登学院，美国缅因州布伦瑞克
nenbarsalo@bowdoin.edu, j.farias@bowdoin.edu

arXiv:2611.28894v1 [q-bio.NC] 21 Nov 2026

摘要

神经坍缩（NC）预测，在训练后期，最后一层类别特征会集中于其均值，且这些均值会形成单纯形等角紧框架（等角紧框架）。我们探究 NC 是否也会在每一层的功能连接组中留下特征：即隐藏层单位激活值之间皮尔逊相关系数构成的图，该图按类别构建，并通过持久图同调进行汇总。使用深层多层感知机，我们观察到两种拓扑现象，它们在终末阶段与最终几层中广为研究的几何神经坍缩同时出现。与 NC1 相反，类别连接组的分半变异性会增加（TNC1）。类别连接组之间的平均成对 H_0 距离 $\bar{D}$ 也会上升（TNC2），但不同类别对之间的增幅并不均匀，这与 NC2 的等距性不同。我们发现，24–56% 的最后一层单元对某些类别而言会变成死亡单元，但对其他类别仍保持活跃，从而增大 $\bar{D}$。校正这些单元的影响后，两种特征仍然偏高。MNIST、Fashion-MNIST、CIFAR-10 和 CIFAR-100 上的这些结果，为拓扑机制可能伴随神经坍缩的几何特征提供了初步证据。

1 引言

神经坍缩（Papyan、Han和Donoho，2020）描述了训练终末阶段（TPT）最后一层特征的几何结构：类内变异性消失（NC1），类别均值构成单纯形等角紧框架（NC2），分类器与这些均值对齐（NC3），且分类器采用最近类别均值规则（NC4）。这一现象意义重大，因为神经坍缩与泛化能力以及网络向新类别迁移的能力有关（Papyan、Han和Donoho，2020；Galanti、Györgyand Hutter，2022）。后续研究还在中间层中发现了神经坍缩，尽管其程度较弱（Parker 等人，2023）。

受生物网络功能性磁共振成像（fMRI）研究中功能连接组的启发，Zhang 等人（2023）提出通过网络的功能连接组来考察网络：即由各单元激活值之间的 Pearson 相关性构成、并通过持久同调加以概括的加权图。本文构建特定类别的逐层连接组，并探究神经坍缩是否留下拓扑痕迹。在此，我们观察到两种与之共同出现的拓扑模式

NC1 和 NC2：**（TNC1）**每个类别连接组在其所属类别内变得更加多变；**（TNC2）**在训练的终末阶段，各类别连接组的均值彼此相距更远。

贡献。(i) 我们定义并观察这两种连接组特征，并追踪它们在不同层和训练检查点上的变化，同时考察 NC1–NC4。(ii) 我们发现，随着训练推进，TNC1 和 TNC2 均会增大，偏离 NC1，并违反 NC2 的等距性质。(iii) 我们报告了一项意外的副产品：在终末阶段，类别选择性死亡单元的数量激增。

2 背景：持久图同调

令 $X_c \in \mathbb{R}^{n_c \times p}$ 表示类别 c 的 n_c 个训练样本在具有 p 个单位的层中的激活值。类别连接组 G_c 是由这 p 个单位构成的完全图，其边权重 $w_{ij} = |$为 r($X_{c,\cdot i}$, $X_{c,\cdot j}$)|，其中 r 表示皮尔逊相关系数。遵循 Songdechakraiwut 和 Chung（2023）的做法，令 G_ϵ 仅保留权重大于 ϵ 的边。随着 ϵ 增大，连通分量的数量 $\beta_0(G_\epsilon)$ 随之增加。0 维持久图由最大生成树（MST）的 $p-1$ 权重（即出生）组成，而 1 维持久图则由其余非树边权重（即死亡）组成。出生表示连通分量的出现，而死亡表示环的消亡。

Songdechakraiwut 和 Wu（2025）证明，对于具有相同节点数的两个类别 $c \neq c'$ 的连接组G_c 和 $G_{c'}$，其图表之间的平方 2-瓦瑟斯坦距离（W_2^2）等于排序后的出生向量 $\mathbf{b}^{(c)}$ 之间的平方欧几里得距离（对于 H_0），或排序后的死亡向量$\mathbf{d}^{(c)}$ 之间的平方欧几里得距离（对于 H_1）。后来，Wu、He 和 Songdechakraiwut（2025）对这一方法进行了改编，用于追踪训练的中间阶段：

$$W_2^2\big(H_0^{(c)}, H_0^{(c')}\big) = \|\mathbf{b}^{(c)} - \mathbf{b}^{(c')}\|_2^2, \tag{1}$$

$$W_2^2\big(H_1^{(c)}, H_1^{(c')}\big) = \|\mathbf{d}^{(c)} - \mathbf{d}^{(c')}\|_2^2. \tag{2}$$

在本文中，我们报告了 TNC2 的逐条目均方距离 MSD $= W_2^2/m$，其中 m 是向量的长度。我们将分析限制在 H_0 和 H_1（连通分量和环）上，遵循 Songdechakraiwut 和 Wu（2025）的做法；他们使用这两个维度来刻画连接组拓扑。为简便起见，我们主要关注H_0，而 H_1 的结果则在讨论中进行总结。

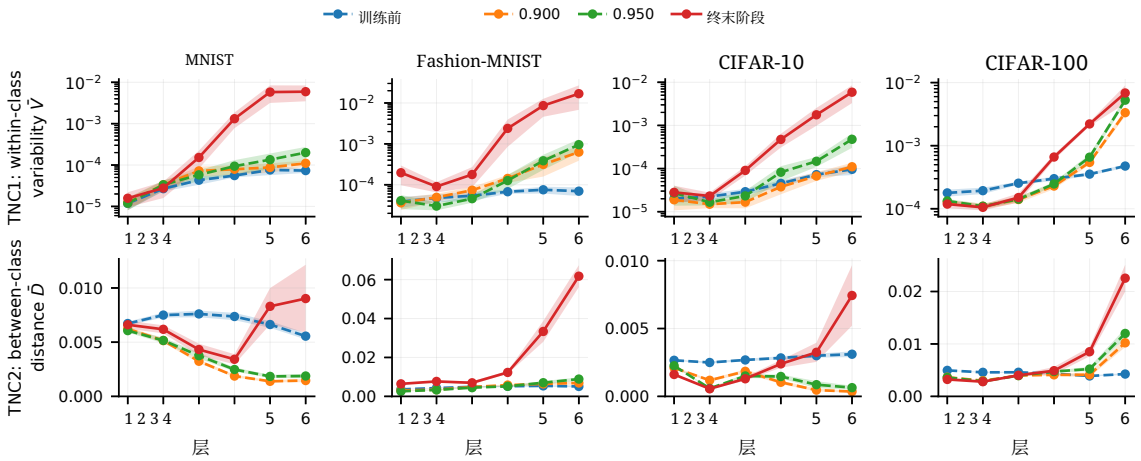


图1：逐层连接组特征（均值 ± 标准差）。上方：分半组内变异性 $\bar{V}$(TNC1)，对数尺度下方：类别连接组之间的 H_0 平均两两 D 距离（TNC2）。

	坍缩		TNC1：类内变异性		TNC2：类间距离	
数据集	NC1	NC2	$\bar{V}$ ($\times 10^{-4}$)	$\bar{V}/\bar{D}_{\text{half}}$	$\bar{D}$	CV*
MNIST	1.25 → .004	.70 → .27	0.7 → 59	.01 → .72	.0056 → .0090	1.64 → 1.15
Fashion-MNIST	2.85 → .002	.84 → .19	0.7 → 169	.01 → .26	.0051 → .0618	1.14 → .79
CIFAR-10	21.3 → .034	1.01 → .19	1.0 → 59	.03 → .63	.0031 → .0074	.98 → .82
CIFAR-100	48.1 → .0013	1.27 → .19	4.7 → 68	.11 → .28	.0043 → .0225	1.01 → .96

表1：最后隐藏层（fc6），对 10 个随机种子取平均；双值条目分别表示初始化 → 终末阶段（TPT）。 *成对距离的变异系数。

3 拓扑连接组特征

TNC1：类内变异性。为了获得类内变异性（NC1）的拓扑类比，我们采用分半重采样，这是一种借鉴功能性磁共振成像文献的方法，用于追踪测量可靠性（Teeuw 等人，2021；Walther 等人，2016）。对于每个随机种子，我们将每个类别随机划分为两半，并根据每一半分别构建一个连接组。令V_c 表示类别 c的两个连接组之间的均方距离，并令 $\bar{V}$ 表示 V_c 在各类别间的均值。随后，我们将其与 D_{half}进行比较，即 不同类别的半规模连接组之间的均方距离均值。比率 $\bar{V}/D_{\text{half}}$ 是 NC1 类内离散度与类间离散度比率（Papyan、Han 和 Donoho，2020）的拓扑对应物。为判断观测到的变异性是否仅由噪声造成，我们还计算 D_{null}，即 C 相同规模的随机、类别无关组之间的均方距离均值，以确定变异性中预期的噪声。由于连接组是样本组的属性，因此需要使用分半变异性。不同于 NC1 的几何特征，TNC1 无法在单个样本上测量。

TNC2：类间距离。$\bar{D}$ 是所有类别对上 MSD($H_0^{(c)}$，$H^{(c')}{}_0$) 的均值。如果类别是可交换的（正如 NC2 中等角紧框架所预期的那样），则成对距离应当近似相等。我们通过成对距离的变异系数（CV）来量化偏离相等的程度：即标准差除以均值。当每一对之间的距离都相等时，CV 等于零；随着距离变得越不均匀，CV 越大。我们的目标是，通过测量成对距离的大小以及成对距离的变化，我们能够观察到 NC2 的拓扑类比。

4 实验

设置。 我们遵循 Parker 等人（2023）的做法，使用均方误差损失训练具有 6 个宽度为 4096 的隐藏层（ReLU）的多层感知机。优化采用动量为 0.9 的随机梯度下降，使用单周期调度（峰值学习率为 0.05）、批量大小为 128、权重衰减10^{-5}，训练 300 轮。与交叉熵相比，我们更倾向于使用均方误差，因为它既能促进神经坍缩，在数学上也更清晰（Han、Papyan 和 Donoho，2022）。每个类别分别使用 500 张（CIFAR-100）、5000 张（MNIST、CIFAR-10）或 6000 张（Fashion-MNIST）训练图像，并对 10 个随机种子的结果取平均。

所有量均在训练集上进行评估，评估时点包括初始化、训练的多个中间阶段以及终末阶段。在相同的检查点，我们计算 NC1= $\text{tr}(\Sigma_W \Sigma_B^{\dagger})/C$、NC2（中心化类别均值的归一化格拉姆矩阵与单纯形等角紧框架的归一化格拉姆矩阵之间的弗罗贝尼乌斯距离）、NC3（归一化分类器矩阵与类别均值矩阵之间的距离）以及 NC4（与最近类别均值规则的不一致程度）。有关推导的进一步说明，请参见 Papyan、Han和 Donoho（2020）以及 Han、Papyan和Donoho（2022）。在这些实验中，我们将隐藏层 1 至 6 分别命名为 fc1 至 fc6。

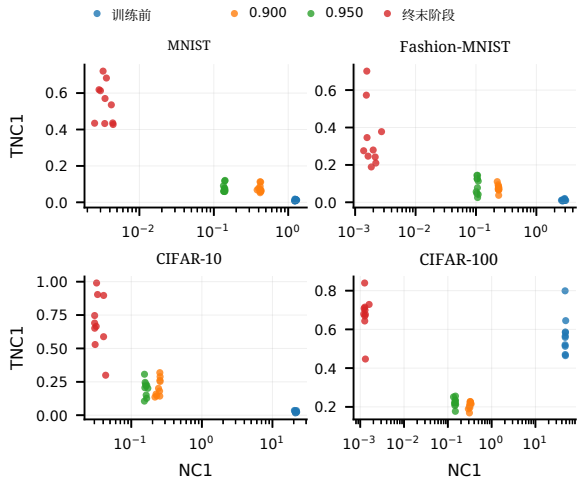


图2：fc6 层的 NC1 与 V/D_{half} (TNC1) 的关系。每个随机种子和检查点对应一个数据点。

神经网络发生坍缩。在最后隐藏层（fc6），NC1 下降了两到三个数量级，NC2 下降了 2.6–5.2×（表1）；NC4 从约 0.9 降至低于10^{-3}。训练开始后，在每一层，NC1 都会随着深度增加和训练持续进行而下降。

TNC1：类内变异性随着 NC1 下降而上升。在 fc6 层，V在终末阶段之前最多增长 14×，随后在终末阶段达到其初始值的 81×、242×、61× 和 14×（图1，上图）。从 fc3 开始，随着深度增加，这种增长愈发显著（CIFAR-100 上从 fc4 开始）。比率 V/D_{half} 从 0.01–0.11 上升至 0.26–0.72（表1，）），而 NC1 则下降。将 NC1 与 V/D_{half} 作图可见二者呈强负相关（斯皮尔曼 $\rho = -0.90$、-0.89、-0.89、-0.21；图2）。这与 NC1 所预期的行为相反；按照 NC1 的预期，类内变异性会持续下降。

TNC2：距离先下降，随后在终末阶段上升。在训练的中间阶段，MNIST 上的 fc6 值从 D 下降 3.8×，CIFAR-10 上则 8.6×，而在 Fashion-MNIST 上略有上升（图1，下方；表1）。在终末阶段，D 在所有数据集上均上升，最终 1.6×、12× 和 2.4× 高于其初始值。在 MNIST 和 CIFAR-10 上，反弹集中于最后两层：在 fc1–fc4，终末阶段的距离仍低于其初始值。在 Fashion-MNIST 上，终末阶段的距离在每一层都超过初始值。初始化后，成对距离的变异系数保持在 0.7 到 1.2 之间，表明成对连接组距离从未如 NC2 所预期的那样趋于均匀。在 CIFAR-100 上，

D 没有出现早期下降：fc6 值在训练中间阶段已经 2.4×上升，最终 5.3×高于其初始值，且终末阶段的反弹仅限于 fc4–fc6。其在 fc6 的变异系数保持在 0.93 到 1.14 之间，同样远未达到均匀状态。

5 类别选择性死亡单元混淆类别连接组距离

尽管类间距离最初会下降，但在终末阶段却大幅上升。在探究这一增长的原因时，我们注意到 fc6 在终末阶段存在大量死亡神经元（对任何样本都从不激活的单位；图3）。此类单位的方差为零，因此其相关性没有定义。对于本研究，我们将这些相关性设为零。我们将单位称为在类别中死亡 c，并用 k_c 表示其数量。先前的研究将这一现象归因于 ReLU，因为 ReLU 会使部分单位的方差坍缩为零（Lu 等，2020 年）。

死亡单元在终末阶段激增，并且具有类别选择性。在fc6中，死亡单元比例随后在终末阶段骤增（图3）：MNIST为 19 → 26%，Fashion-MNIST为 21 → 29%，CIFAR-10为 14 → 24%，CIFAR-100为 20 → 56%。这些单元中的大多数并非对所有类别都处于死亡状态：对于某些类别，该层中有51–79%的单元处于死亡状态，而对于其他类别则处于活跃状态。因此，这些死亡单元具有类别选择性。

死亡单元会放大距离。死亡单元更多的类别拥有更多零出生，因此无论其余单元的行为如何，该类别都与其他类别相距更远。在实证上，死亡单元差距 $|k_c - k_{c'}|$ 与fc6距离之间的斯皮尔曼相关系数（跨类别对计算）从初始化时的0.12–0.65上升至终末阶段的 0.94–0.99。因此，死亡单元会放大类别间距离，并混淆任何纯粹的神经坍缩拓扑特征（图1）。

死亡单元并不能解释全部现象。在此，我们仅保留在少于 20% 的类别中属于死亡状态的神经元，并据此重新计算这两种特征。经过这一修正，拓扑模式依然存在，但有所减弱。fc6 距离在训练中期仍持续下降：在 MNIST 上下降 11×，在 Fashion-MNIST 上下降 1.9×，在 CIFAR-10 上下降 15×（图3；表 2）。在终末阶段，该距离在 MNIST 和 CIFAR-10 上回升至初始值，在 Fashion-MNIST 上则回升至初始值的 5.9×。在每个检查点分别应用 20% 规则，也会得到相同的模式（但此时 Fashion-MNIST 不再表现出早期下降）。修正后神经元集合的类内变异性上升至初始值的 63×、217× 和 47×，V/D_{half} 0.01 0.03 0.34 0.70 1 则从 – 上升至 –；NC 与 V/D_{half} 仍呈负相关（$\rho = -0.86$、-0.88、-0.82）。在 CIFAR-100 上，修正后仅保留 4096 个单位中的 161 个，因此结果的可靠性较低。在该集合上，D 没有早期下降（而是在终末阶段之前上升 1.9×），最终高于其初始值 11×；V上升了 9.4×，但 V/D_{half}仅从 0.57 变为 0.68，且与 NC1 的负相关较弱（$\rho = -0.21$）。总的来说，这些结果表明，除了死亡单元之外，还有其他机制在驱动最终层类间距离于终末阶段上升。

6 讨论与局限性

按标准几何度量来看，神经网络会发生坍缩；然而在终末阶段，拓扑连接组中的类间距离和类内变异性都会上升。

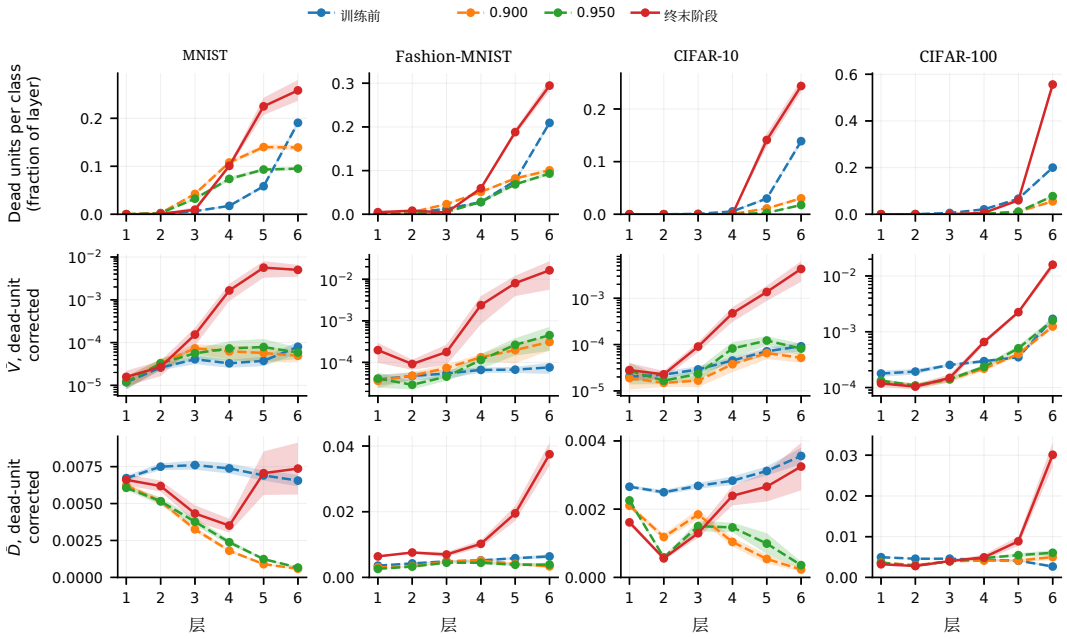


图3：每个类别的死亡单元占该层的比例，以及针对 TNC1 和 TNC2 的死单元校正值（均值 ± 标准差）。

	死亡单元 , fc6		$\bar{D}$, 所有单位		$\bar{D}$ 在校正后的单元集合上 , fc6		
数据集	每个类别：初始化 → 90% →TPT	选择性*	初始化 → TPT	ρ	初始化 → TPT	ρ	单位†
MNIST	781 → 570 → 1058	2446	.0056 → .0090	.97	.0066 → .0074	.95	2577
Fashion-MNIST	857 → 412 → 1207	3230	.0051 → .0618	.99	.0064 → .0375	.98	2232
CIFAR-10	569 → 124 → 998	2082	.0031 → .0074	.94	.0036 → .0033	.57	2624
CIFAR-100	819 → 227 → 2278	4095	.0043 → .0225	.95	.0027 → .0301	.40	161

表 2：fc6 处的死亡单元及死亡单元校正，取 10 个随机种子的均值。死亡单元数为每个类别在 4096 个单元中的计数。 *在 TPT 时，部分但非全部类别中失活的单元。 ρ：TPT 时，死亡单元差距与 H_0 距离在类别对上的斯皮尔曼相关系数。 †校正后的单元集合的大小。

此外，类间距离的增加并非在各对之间均匀分布，这与 NC2 的预测相反。 H_1 的结果类似：fc6 H_1 距离最终比初始值高出 1.9–6.6×，而 H_1 分半比则从最高 0.04 上升至 0.20–0.94（附录A）。

我们认为有两种可能的解释。第一，神经坍缩本身可能是原因，但它对拓扑的影响或许不同于对几何的影响，可能同时增大类间距离和类内变异性。第二，当特征坍缩到非负的类别均值上时，ReLU单元可能会在整个类别中变为静默状态。ReLU特征模型预测类别均值会发生坍缩且彼此重叠很少（Tirer 和 Bruna，2022；Dang 等，2024），这会产生类别选择性失活单元。初始化时就已存在的 ReLU死亡效应（Lu 等，2020）或可解释训练开始前观察到的死亡单元。未来的工作应将这些结果推广到其他激活函数，以探究这些结果究竟源于 ReLU 效应，还是神经坍缩的后果。

局限性。 TNC1 对每个随机种子仅依赖一次随机划分。未来的工作应对多次划分取平均。所有实验均使用均方误差损失和 ReLU多层感知机。我们没有改变样本量、激活函数或权重衰减，而这些因素中的任何一项都可能有助于揭示这些发现背后的机制。最后，结果纯属相关性，因此无法做出因果性断言，认为观察到的拓扑变化是由神经坍缩产生的。

结论与未来方向。 这些结果表明，神经坍缩可能存在一种拓扑对应形式，并与几何神经坍缩同时发生。未来的工作应检验它们能否推广到其他损失函数和激活函数，尤其是 LeakyReLU 和 GELU 这类不会输出精确零值、因而能够避免死亡单元混淆因素的函数。还可以检验直接诱导拓扑坍缩是否能带来与神经坍缩相关的泛化优势。

AI 披露。 我们使用 Claude Opus 5.5 对本文作者撰写的文本进行语法和表述清晰度方面的编辑与润色。所有输出均由作者审核和编辑，作者对内容承担全部责任。此外，我们还使用了 Claude Code 进行调试。作者验证并执行了实验中使用的所有最终代码，所有实验设计决策均由作者作出。

参考文献

Dang, H.; Tran, T.; Nguyen, T.; and Ho, N. 2024. 采用无约束 ReLU 特征模型进行交叉熵类别不平衡学习的神经坍缩。*arXiv* 预印本 *arXiv:2401.02058*。

Galanti, T.; György, A.; and Hutter, M. 2022. 神经坍缩在迁移学习中的作用. 载于国际学习表征会议。Han, X. Y.; Papyan, V.; and Donoho, D. L. 2022. 均方误差损失下的神经坍缩：与中心路径的接近程度及其沿中心路径的动态. 载于国际学习表征会议。Lu, L.; Shin, Y.; Su, Y.; and Karniadakis, G. E. 2020. 死亡 ReLU 与初始化：理论及数值示例。计算物理通讯, 28(5): 1671– 1706. Papyan, V.; Han, X. Y.; and Donoho, D. L. 2020. 深度学习训练终末阶段神经坍缩的普遍性。美国国家科学院院刊, 117(40): 24652–24663. Parker, L.; Onal, E.; Stengel, A.; and Intrater, J. 2023. 分类神经网络中间隐藏层的神经坍缩。*arXiv* 预印本 *arXiv:2308.02760*。Songdechakraiwut, T.; and Chung, M. K. 2023. 脑网络的拓扑学习。应用统计年鉴, 17(1): 403–433. Songdechakraiwut, T.; and Wu, Y. 2025. 神经网络的功能连接组。载于*AAAI* 人工智能会议论文集。Teeuw, J.; Hulshoff Pol, H. E.; Boomsma, D. I.; and Brouwer, R. M. 2021. 静息态功能连接的可靠性建模。神经影像, 231: 117842. Tirer, T.; and Bruna, J. 2022. 用于探索深度神经坍缩的扩展无约束特征模型。载于国际机器学习会议。Walther, A.; Nili, H.; Ejaz, N.; Alink, A.; Kriegeskorte, N.; and Diedrichsen, J. 2016. 多体素模式分析中不相似性度量的可靠性。神经影像, 137: 188–200. Wu, Y.; He, P.; and Songdechakraiwut, T. 2025. 利用动态连接组进行数据高效的神经网络训练。*arXiv* 预印本 *arXiv:2508.06817*。Zhang, B.; Dong, Z.; Zhang, J.; and Lin, H. 2023. 功能网络：深度神经网络可解释性的新框架。神经计算, 519: 94–103。

A H_1结果

本附录用死亡向量$\mathbf{d}^{(c)}$ 替代出生向量，重复正文中的 H_0 分析：$\bar{V}_1$,– – – – $D_{1,\text{half}}$ 和 D_1 定义为 V, D_{half} 和 D，均方位移取死亡向量上的值。

A.1 连接组特征

图 4 和表 3 分别是图1和表1的 H_1 对应版本。至于 H_0，两种特征在 fc6 的终末阶段– –均有所上升：V_1 增长至其初始值的 34–805×倍，D_1 1.9–6.6×,– –；而$V_1/D_{1,\text{ half}}$ 则从不超过 0.04 上升至 0.20–0.94。在 TPT，成对 H_1 距离的变异系数为 1.0–1.9，因此H_1 距离并不比 H_0 距离更接近均匀分布。

A.2 死亡单元校正

图 5 和表 4 在相同的校正后的单元集合上重复了第 5 节中的死亡单元校正。死亡单元差距与 H_1 距离的关联较弱（TPT 为 $\rho = 0.11$–0.71，而 H_0为 0.94–0.99），这表明死亡单元主要增加零出生，而不是改变循环结构。–校正后，D_1 在每个数据集上的最终值仍高于其初始值，且 $V_1/D_{1,\text{ half}}$ 仍然上升。

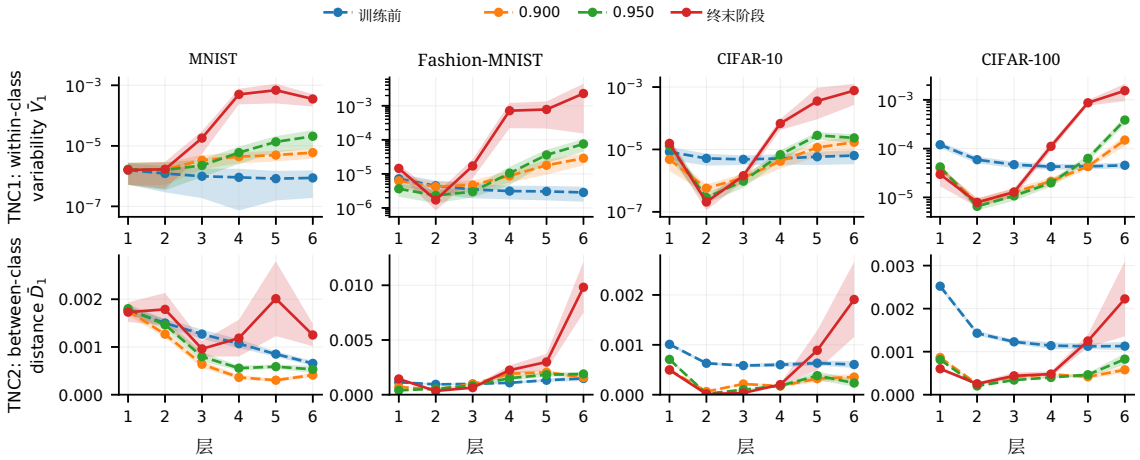


图 4： H_1 连接组特征（均值 ± 标准差）。上图：分半组内变异性 V_1 （TNC1），对数尺度。下图：类别连接组之间的平均成对 H_1 距离 D_1 （TNC2）。参见图1。

	TNC1：类内变异性		TNC2：类间距离	
数据集	$\bar{V}_1$ ($\times 10^{-6}$)	$V_1/\bar{D}_{1,\ \text{half}}$	$\bar{D}_1$	CV*
MNIST	0.9 → 354	.001 → .33	.0007 → .0013	1.74 → 1.19
Fashion-MNIST	2.8 → 2291	.002 →.20	.0015 → .0098	1.17 → 1.02
CIFAR-10	6.4 → 772	.011 → .31	.0006 → .0019	1.07 → 1.08
CIFAR-100	45.5 → 1539	.043 → .94	.0011 → .0022	1.56 → 1.91

表 3： H_1 表1 的对应项：最后隐藏层（fc6），所有单位，取 10 个随机种子的均值；条目为初始化 → 终末阶段（TPT）。 *成对 H_1 距离的变异系数。

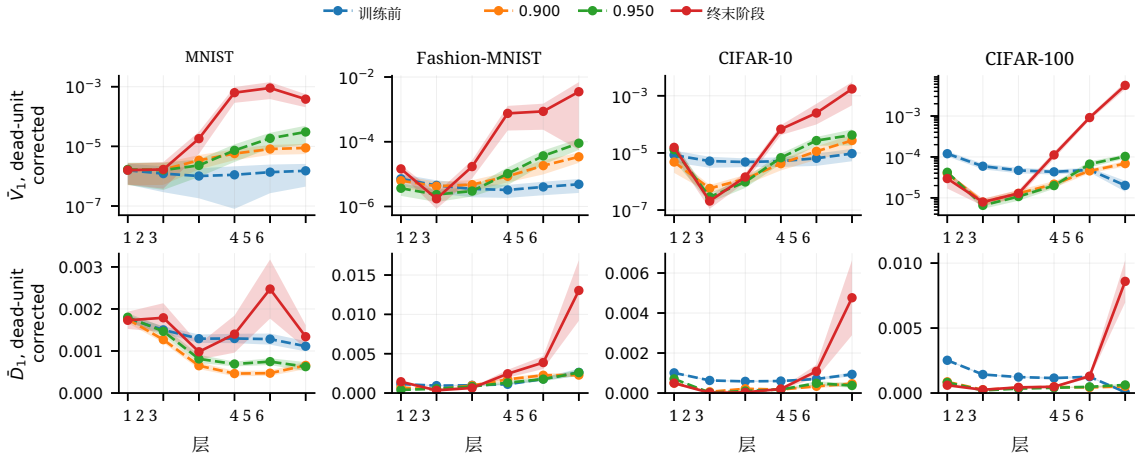


图 5： H_1 死亡单元校正后的特征（均值 ±标准差）。请对比图3的最下面两行。

	$\bar{D}_1$，所有单位		校正后的单元集合，fc6			
数据集	初始 → TPT	ρ	$\bar{D}_1$: 初始 → TPT	ρ	$\bar{V}_1$ ($\times 10^{-6}$)	$V_1/\bar{D}_{1,\ \text{half}}$
MNIST	.0007 → .0013	.45	.0011 → .0013	.28	1.5 → 383	.001 → .29
Fashion-MNIST	.0015 → .0098	.71	.0026 → .0130	.48	4.8 → 3507	.002 → .22
CIFAR-10	.0006 → .0019	.22	.0009 → .0048	.08	9.5 → 1734	.010 → .29
CIFAR-100	.0011 → .0022	.11	.0001 → .0086	.31	20.1 → 5602	.307 → .79

表 4： H_1 表 2 的对应结果：fc6 处的死亡单元校正，取 10 个随机种子的平均值。 ρ：TPT 处死亡单元差距与 H_1 距离之间在类别对上的斯皮尔曼相关系数。校正后的单元集与表 2 中的相同。