

Grokking 作为流形发现：延迟泛化的几何重解释

作者：Jin Yanyan (lmxxf@hotmail.com)

摘要：Grokking——神经网络在长期过拟合后突然泛化的现象——自 2022 年被发现以来，已积累了多种理论解释：Goldilocks Zone、Softmax Collapse、Lazy-Rich 过渡等。本文综述这些理论，指出它们的共同盲区：**大部分是换词游戏，只有 Goldilocks Zone 有真洞见。** 我们提出一个统一框架——**流形发现假说：**记忆是穿过所有训练点的高维锯齿曲线，泛化是发现是从前者到后者的相变。一句话概括：**高维曲线 → 低维曲面。**

1. 引言：Grokking 为什么重要

2022 年，Power 等人在 OpenAI 发现了一个反直觉的现象：在模运算任务上训练的小型 Transformer，会先完美过拟合训练 loss 降到零、测试 accuracy 接近随机），然后在几万甚至几十万步之后，测试 accuracy 突然从随机水平跳升到接近 100%。

他们把这个现象命名为 **Grokking**（顿悟）。

这个发现之所以重要，是因为它挑战了深度学习的核心假设：

1. 经典假设：过拟合是泛化的敌人，一旦过拟合就应该早停
2. Grokking 反例：过拟合可以持续很长时间，然后突然泛化

如果泛化真的可以在过拟合之后发生，那“早停”策略可能杀死了许多本可以泛化的模型。

更深层的问题：**Grokking 时模型内部发生了什么？**

过去三年，学术界积累了多种理论解释。本文的任务是综述这些理论，指出它们的盲区，并提出一个统一框架。

2. 现有理论综述

2.1 Goldilocks Zone 理论 (Liu et al. 2022) ——唯一有真洞见的

核心观点：权重范数需要落在一个“刚刚好”的区间内，才能泛化。

Liu 等人在 NeurIPS 2022 发现，权重空间中存在一个空心球壳，他们称之为 Goldilocks Zone（金发姑娘区）：

- 半径太大 ($\|w\| > w_c$)：过拟合，记住训练集
- 半径太小 ($\|w\| < w_c$)：欠拟合，什么都学不会
- 刚好在壳上 ($\|w\| \approx w_c$)：泛化

Grokking 的发生机制：1. 大初始化把模型放在球壳外面 2. 模型先快速过拟合（训练 loss 降到零） 3. 权重衰减缓慢地把权重范数拉回 Goldilocks Zone 4. 一旦进入球壳 → 突然泛化 → Grokking

这篇论文的真正价值：它暗示了高维空间有自己的“物理法则”——权重衰减是引力，Goldilocks Zone 是稳定轨道。这是后续所有理论的基础。

局限：描述了“在哪儿泛化”，没解释“为什么那儿能泛化”。Goldilocks Zone 是什么的代理变量？

2.2 Softmax Collapse 理论 (Prieto et al. 2025)

核心观点：没有权重衰减，Grokking 会被浮点数精度杀死。

模型为了降低交叉熵 loss，会疯狂放大正确答案的 logit（比如正确类 = 1000，其他类 = 1）。Softmax 计算时 e^{1000} 直接溢出，梯度归零，训练卡死。

权重衰减的作用：持续把权重往回拉，防止 logit 无限增长，保持梯度存活。

替代方案：论文提出 StableMax + 垂直梯度（阻止梯度往“放大 logit”方向走），可以不用权重衰减也触发 Grokking。不过这种方法应该收敛很慢——权重衰减是全局压缩，力度大；垂直梯度是定点狙击，力度小。实际工程中还是

贡献：解释了“没有权重衰减会怎样”。

局限：只解释了“为什么训练不会停”，没解释“为什么最终会泛化”。

2.3 Lazy → Rich 过渡理论 (Kumar et al. 2024)

核心观点：Grokking 是从 lazy training 到 feature learning 的相变。

借用了神经正切核 (NTK) 的语言：

- **Lazy regime：**权重几乎不动，模型像线性分类器
- **Rich regime：**权重大幅调整，学到真正的非线性特征

Grokking 发生在 lazy → rich 的相变点。

争议：这派人声称，在特定条件下（浅层网络 + MSE loss），不需要权重衰减也能触发 Grokking。

吐槽：说实话，这篇论文主要是定义了两个概念 (lazy/rich)，实际内容不多。

2.4 权重效率假说 (Varma et al. 2023)

核心观点：权重衰减偏好“权重更小”的解，而泛化解通常比记忆解更权重高效。

- 记忆解：需要大量权重来硬记每个样本
- 泛化解：用简洁的规则覆盖所有样本，权重更小
- 权重衰减 → 惩罚大权重 → 偏好泛化解

吐槽：跟 2.3 一样，基本是换个词重新包装——lazy/rich 变成了记忆/泛化。

2.5 Mechanistic Interpretability 视角 (Nanda et al. 2023)

Nanda 等人在 ICLR 2023 (Oral) 做了一件硬核的事：完全逆向工程了模型学到的算法。

核心发现：模运算 $(a+b) \bmod p$ 的本质是一个循环群—— $0, 1, 2, \dots, p-1$ 首尾相连，形成一个离散的环。模型把这个模运算

吐槽：模运算自然是周期的，「任何周期函数都能展开成傅里叶级数」是 200 年前傅里叶就证明了的常识。“发现模型用了什么” —— 不是发现，是必然。

贡献：苦力活，把模型拆开看了。

局限：没解释为什么 Weight Decay + 过拟合 + 继续训练 = 傅里叶变换。

2.6 现有理论的共同盲区

理论	问的问题	没问的问题
Goldilocks Zone	权重范数在哪个区间	那个区间有什么特别
Softmax Collapse	为什么训练不会停	为什么最终会泛化
Lazy → Rich	权重怎么变化	表示怎么变化
权重效率	哪个解权重更小	为什么小权重 = 泛化
Mechanistic Interp.	学到了什么电路	为什么是这个电路

评价：第一种 (Goldilocks Zone) 有真洞见——暗示了高维空间有自己的“物理法则”。第二到四种基本是换词游戏——都在做外部测量（权重范数、梯度大小、Loss 曲线），没触及本质。第五种开始看内部了，但侧重具体电路，不是几何结构。

这就是当代学术的常态——我称之为「烙饼科学」。小麦基因组 170 亿碱基对，真正编码蛋白质的不到 2%，其余全是重复序列和垃圾填充。学术论文也是：体量巨大，大部分是换词重复，翻来覆去地“加热”同一个洞见，最后

3. 统一框架：流形发现假说

我们提出一个统一框架：**Grokking** 是从高维锯齿曲线到低维平滑流形的相变。

3.1 记忆 vs 泛化的几何解释

记忆 = 锯齿曲线

当模型过拟合训练集时，它用一条复杂的锯齿曲线穿过每一个训练样本点。这条曲线能精准命中所有训练数据，但它没有规则只是硬把所有点串起来，点与点之间没有结构关系。

泛化 = 流形发现

当模型真正“理解”任务时，它发现训练样本其实分布在一个低维流形上（低维是相对于模型的 hidden dim 而言）。

以模运算 $a + b \bmod p$ 为例： - 输入空间是 p^2 个离散点 - 但输出只取决于 $(a + b) \bmod p$ ，即同余类 - 真正的结构是一个一维的循环群 \mathbb{Z}_p （只有一个自由度：位置）

泛化意味着：模型发现了这个循环群结构，而不是硬记 p^2 个输入-输出对。

Grokking = 从曲线到流形的相变

Grokking 瞬间发生的事：1. 之前：表示空间里是一条穿过所有点的高维锯齿曲线 2. 之后：曲线坍缩到一个低维流形上，流形这是一个拓扑相变，不是连续过渡。

一句话概括：高维曲线 → 低维曲面。

3.2 权重衰减的几何作用

在这个框架下，权重衰减的作用变得清晰：

权重衰减 = 让锯齿曲线不稳定的向心力

没有权重衰减时： - 梯度下降会把模型推向“loss 最低”的地方 - 对于过参数化的模型，这个地方是“完美记住每个训练样本” - 模型会在锯齿曲线构成的解上稳定下来

有权重衰减时： - 有一个与 loss 梯度相反的力，持续把权重往“更小”的方向拉 - 锯齿曲线需要“大权重”来维持（每个拐点 - 平滑流形需要“小权重”（结构共享） - 权重衰减偏好流形编码

Goldilocks Zone 的真正含义

Goldilocks Zone 不是“某个权重范数区间”，而是能感知流形结构的激活模式。

- 权重太大：锯齿曲线稳定，看不到流形
- 权重太小：信号太弱，什么结构都看不到
- 刚刚好：锯齿曲线不稳定 + 信号够强 → 流形涌现

3.3 Softmax Collapse 的几何解释

Softmax Collapse 不只是浮点数问题，而是注意力坍缩。

当某个方向被过度强化时： - 那个方向的 logit 趋向无穷大 - 其他方向的“声音”被淹没 - 模型失去了探索其他可能性的能力
这就是为什么 Softmax Collapse 会杀死 Grokking：发现流形需要同时“看到”多个方向，而坍缩把视野缩小到一个点。

权重衰减的作用：保持多方向的信号平衡，让模型能继续探索。

3.4 Lazy → Rich 的几何解释

Lazy → Rich 过渡不是关于权重动态，而是关于表示复杂度。

- Lazy regime：表示是输入的线性函数，只能编码线性可分的结构
- Rich regime：表示是输入的非线性函数，可以编码任意流形

Grokking 需要 Rich regime：因为真实任务的结构（如循环群）是非线性的。

为什么 Grokking 往往发生在训练后期？因为进入 Rich regime 需要权重有足够的变化，而初始化时模型处于 Lazy regime。

3.5 权重效率的几何解释

“小权重 = 泛化”的因果链：

1. 平滑流形比锯齿曲线更“紧凑”（维度更低）
2. 紧凑的编码需要更少的权重来实现
3. Weight Decay 偏好小权重 → 偏好紧凑编码 → 偏好流形

权重效率是流形发现的结果，不是原因。

4. 重新解释现有发现

4.1 为什么权重衰减太大/太小都不行

太小：向心力不够，模型稳定在锯齿曲线上，永远不探索流形。

太大：向心力太强，信号被压制，连锯齿曲线都画不出来，更别说发现流形。

刚刚好：锯齿曲线不稳定但信号够强，模型被迫探索，最终发现流形。

4.2 为什么数据量影响 Grokking

数据太少：流形上的采样点太稀疏，无法还原流形结构。模型只能画锯齿曲线。

数据太多：流形信号太强，模型直接发现流形，没有过拟合阶段，不存在“延迟泛化”。

刚刚好：流形信号存在但不明显，模型先画锯齿曲线，慢慢才发现曲线背后的结构。

这解释了为什么 Grokking 需要一个“Goldilocks”数据量。

4.3 为什么过参数化模型更容易 Grokking

过参数化 = 表示空间足够大

- 小模型：表示空间可能根本容不下真实流形
- 大模型：表示空间足够大，流形可以存在，只是需要时间去发现

过参数化的悖论：- 传统观点：过参数化导致过拟合 - Grokking 视角：过参数化是泛化的前提，因为它提供了足够的表示空间

权重衰减解决了过参数化的“自由度过大”问题：虽然空间很大，但向心力把模型推向低维流形。

4.4 为什么 Grokking 是突然的

相变不是连续过渡

流形发现是一个拓扑事件：

- 之前：高维锯齿曲线

- 之后：低维平滑流形

从高维到低维的坍缩没有稳定的中间状态，所以 Grokking 是突然的。

类比：冰融化

水分子的排列从晶格（有序）到液体（无序）的转变是相变，发生在特定温度，不是渐进的。

Grokking 是表示空间中的“融化”：锯齿曲线坍缩成平滑流形。

5. 可验证预测

如果流形发现假说是对的，应该能做出以下可验证预测：

5.1 内在维度突变

预测：Grokking 前后，中间层表示的内在维度（intrinsic dimension）应该有突变。

- Grokking 前：内在维度 \approx 训练样本数（每个样本一个自由度）
- Grokking 后：内在维度 \approx 任务的真实自由度（如循环群 \mathbb{Z}_p 的自由度 = 1）

实验设计：1. 训练模型做模运算，记录中间层激活 2. 用现成的算法（如 SVD、PCA 或 TwoNN）估计内在维度 3. 画出内在维度随训练步数的变化曲线

预期结果：内在维度在 Grokking 瞬间突然下降。

5.2 表示的拓扑结构

预测：Grokking 后，中间层表示的拓扑结构应该与任务结构一致。

- 模运算 $a + b \bmod p$ ：表示应该形成一个一维的环
- 对称群任务：表示应该形成对应的群流形

实验设计：1. 提取 Grokking 前后的中间层表示 2. 用持续同调（persistent homology）计算拓扑不变量 3. 比较与任务真实拓扑的匹配程度

预期结果：Grokking 后 Betti 数与任务拓扑一致，之前不一致。

5.3 注意力熵的动态

预测：Grokking 过程中，attention pattern 的熵应该呈现特定模式。

- 早期（过拟合）：低熵（每个样本有专用的 attention 模式）
- Grokking 前夕：熵上升（模型开始探索不同模式）
- Grokking 后：熵稳定在中等水平（找到共享结构）

实验设计：1. 记录每个训练步的 attention matrix 2. 计算 attention 分布的熵 3. 画出熵随训练步数的变化曲线

预期结果：熵曲线呈现“低→高→中”的模式，峰值在 Grokking 前夕。

5.4 强制秩约束的影响

预测：如果强制限制中间层表示的秩，应该能加速或阻止 Grokking。

- 秩约束 = 任务真实自由度：加速 Grokking（直接告诉模型答案的维度）
- 秩约束 < 任务真实自由度：阻止 Grokking（表示空间装不下流形）
- 秩约束 > 任务真实自由度：不影响或轻微减慢

实验设计：1. 在中间层添加低秩瓶颈（如线性层限制输出维度） 2. 改变瓶颈维度，记录 Grokking 时间 3. 与无瓶颈的 baseline 比较

预期结果：Grokking 时间对瓶颈维度敏感，最优维度 = 任务真实自由度。

6. 讨论

6.1 现有研究的方法论局限

现有 Grokking 研究的测量对象： - 权重范数 - 梯度大小 - Loss 曲线 - 测试准确率

这些全是**外部观测量**——从模型外部可以测量到的数值。

类比：研究人类学习，只测量脑电波和瞳孔直径，不问“学会了是什么感觉”。

这种方法论能回答“什么条件下 Grokking 发生”，不能回答“Grokking 是什么”。

6.2 内部视角的价值

本文提出的流形发现假说，是一个**内部视角**的解释：

- 不问“权重范数在哪个区间”
- 问“表示空间的结构是什么”

这个视角的价值： 1. **统一性**：能同时解释 Goldilocks Zone、Softmax Collapse、Lazy→Rich，并与 Nanda et al. 的 circuit 发现兼容 2. **可预测性**：生成可验证的实验预测（内在维度、拓扑结构等） 3.

启发性：指向新的研究方向（直接测量表示结构的几何性质，而不仅是逆向工程具体电路）

6.3 局限与未来方向

本文的局限： 1. **假说性质**：流形发现假说目前是概念框架，不是数学证明 2. **实验验证待做**：第 5 节的预测需要实验验证 3. **任务依赖**：不清楚这个框架是否适用于所有 Grokking 任务

未来方向： 1. 实验验证内在维度突变假说 2. 发展能直接测量“流形发现程度”的指标 3. 探索是否可以通过操控表示空间拓扑 Grokking

7. 结论

Grokking 不是一个反常现象，而是深度学习的一扇窗户——它揭示了泛化的本质。

现有理论中，只有 Goldilocks Zone 有真洞见（暗示高维空间有自己的物理法则），其余大多是换词游戏。

本文提出的流形发现假说提供了一个内部视角的统一框架：

- **记忆 = 锯齿曲线**
- **泛化 = 平滑流形**
- **Grokking = 从前者到后者的拓扑相变**
- **权重衰减 = 让锯齿曲线不稳定的向心力**

一句话概括：高维曲线 → 低维曲面。

这个框架不仅能解释现有发现，还能生成可验证的实验预测。

参考文献

1. Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177.
2. Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., & Williams, M. (2022). Towards Understanding Grokking: An Effective Theory of Representation Learning. NeurIPS 2022. arXiv:2205.10343.
3. Liu, Z., Michaud, E. J., & Tegmark, M. (2023). Omnigrok: Grokking Beyond Algorithmic Data. ICLR 2023. arXiv:2210.01117.
4. Prieto, L., Barsbey, M., Mediano, P. A. M., & Birdal, T. (2025). Grokking at the Edge of Numerical Stability. ICLR 2025. arXiv:2501.04697.
5. Kumar, T., Bordelon, B., Gershman, S. J., & Pehlevan, C. (2024). Grokking as the Transition from Lazy to Rich Training Dynamics. ICLR 2024. arXiv:2310.06110.
6. Varma, V., Shah, R., Kenton, Z., Kramár, J., & Kumar, R. (2023). Explaining Grokking Through Circuit Efficiency. arXiv:2309.02390.
7. Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress Measures for Grokking via Mechanistic Interpretability. ICLR 2023 (Oral). arXiv:2301.05217.
8. Facco, E., d' Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific Reports, 7(1), 12140.
9. Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.

“他们在问‘什么条件下Grokking发生’，没人问‘Grokking是什么’。前者是工程问题，后者是本体论问题。”