神经网络的拓扑理解

Tushar Pandey, *[1] Department of Mathematics, Texas A&M University, College Station, TX 77840

在这篇综述论文中，我们考察通常被视为黑箱的神经网络内部结构。最简单、最容易理解的做法是研究二元分类，并尝试理解神经网络所采用的方法。我们回顾不同激活函数的重要性、与之相关的网络架构类型，以及一些经验数据。最后，我们描述针对不同问题和技术选择激活函数的一些可能方案，并以此作结。

arXiv:2301.09742v1 [cs.LG] 23 Jan 2023

I. 引言

深度学习中一个重要的问题是理解黑箱内部（即隐藏层）发生了什么。本文旨在理解数据经过不同层时会发生什么。还有一些其他研究方法，例如在每一层之后考察每个数据点，通常是图像。第二种方法是理解边界流形及其在隐藏层中的变化。尽管这些方法很重要，我们认为更重要的是考察整个数据集经过隐藏层时的变换，并观察数据空间在最终层中的表示。

本文首先考察一些平滑激活函数，其中宽度发挥着重要作用。我们阐释了选择不同架构的神经网络背后的直觉，并指出采用这类方法时可能出现的一些错误以及由此带来的时间复杂度。

在第三节中，我们比较了平滑与非平滑激活函数，并探讨网络宽度和深度的变化。我们试图回答一个常见的问题：在实践中，ReLU 为什么比其他激活函数更好？ [1], [2], [3] 我们从一个已知其拓扑的模拟数据集入手，以了解实际流形中可能发生的变化，并在不同条件下测量这些变化。在了解变化情况后，我们再通过高维真实世界数据进行验证。受计算能力限制，我们没有将部分架构参数从模拟数据调整到维度差异很大的真实世界数据。

回顾这些方法后，我们总结了两种方法的结论，并提供了一些实验，将结果推广到不同的架构，以加深对黑盒的理解。

在论文的大部分内容中，我们将考虑数据分为两个类别的情况。所有定义均列于论文末尾的附录中。

II. 平滑激活函数与拓扑变化

在本节中，我们将探讨激活函数为光滑函数时拓扑的变化。本节内容基于 [4]C 的研究

onsid一段弧 hitect结构 w here there is no hidden l层。 Let'假设 the data forms two li神经网络作为 described in fig 1.

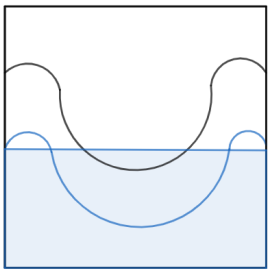


图1. 无隐藏层

* tusharp@tamu.edu

由于没有隐藏层，输出层是一个线性函数，因此神经网络尝试通过一条直线（在更高维度的情况下则是一个超平面）将数据分开来进行分类。在这个例子中，它无法正确分类该数据集，因为不存在一条直线能够将这两个类别完全区分开来。请注意，到目前为止还没有涉及激活函数。
现在考虑只有一个隐藏层的情况。我们观察 $tanh$ 函数，并在图 2 中看到两类数据之间的边界线。

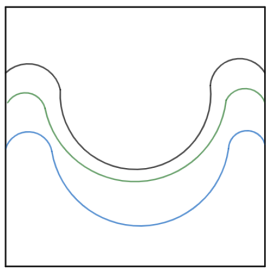


图 2. 一个隐藏层的“tanh”激活函数。绿色线是边界线

这种分离可以在实际数据空间中看到，我们所说的数据空间是指数据最初的表示方式。然而，在网络内部，它并不会试图将边界形状从直线变为曲线，而是改变数据的形状，然后拟合一条线性分离边界。更确切地说，应用激活函数后，数据的环境空间会发生变化，因此神经网络更容易构造线性边界。图 3 展示了一个例子。

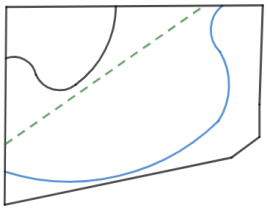


图 3. 数据空间经过变换，最终边界（绿色）是线性的。

对于光滑激活函数，数据空间的这种变化是一个同胚。由此自然会产生一个问题：“对于光滑激活函数，一个隐藏层是否足以改变数据空间（流形），从而通过线性边界分离两个类别？”

回答：不行。事实上，深度不足以判断两个数据集是否可分。网络的深度对应于变换的次数。由于这些函数都是同胚，它们无法改变空间的某些拓扑性质。空间的一种此类性质是同调（或贝蒂数）。回到神经网络：如果一个类别的数据点构成一个圆环，而圆环的孔洞中有一个对应于另一类别的簇，那么对于采用平滑激活的神经网络，$\mathbb{R}^2$ 中的任何同胚都无法区分这两个类别。由于它是一个同胚，β_1 将保持为 1，这意味着没有线性边界能将这两个类别完全分开。

问题：改变宽度又如何？
回答：可以！这样就行。宽度对应于嵌入到更高维度的空间中，在那里可以将孔洞内部的类别抬升起来，再用一个超平面将它们分开。这意味着，如果我们知道能够正确嵌入数据的环境空间的最小维度，就可以使用神经网络对数据集进行完全分类。

定理 II.1. 对于数据空间如上所述，任何宽度为 *1* 或 *2* 的神经网络都无法完全区分这两个数据类别，无论深度如何。

证明。 对于光滑激活函数，隐藏层对应于一个同胚，这意味着 $\beta_1 = 1$ 在所有隐藏层之后。最后一层是线性变换，因此也保持 β_1 。为了用线性边界划分空间，β_1 必须为 0。因此，产生矛盾。 □

请注意，这仅仅是因为激活函数具有平滑性。这意味着，如果激活函数不平滑（例如 ReLU），神经网络就能形成线性边界，将这两个类别区分开来。

对于包含 n 个维度点的数据集，可以将它们嵌入到维度为 2n+2 的环境空间中，使其能够被线性（超平面）边界分开。因此，如果隐藏层的宽度为 $\geq 2n+2$，那么数据就可以被分开。

定理 II.2. *[4]* 当满足以下条件时，输入与网络层的表示之间存在环境同痕：*a) W* 非奇异；*b)* 我们愿意对隐藏层中的神经元重新排序；*c)* 隐藏单元数量大于 *1*。

到目前为止，我们已经找到了一种方法，可以确定一个合适的边界，从而有望在训练数据上达到 100% 的准确率。然而，这并不意味着它就是真正的分界。作者 [4] 还指出，根据一些经验结果，将最后一层从 softmax 改为 KNN 后，准确率会提高。这种方法得出的结论是，对于平滑激活，网络宽度需要足够大，同时还需要足够（但较小）的深度，以免网络陷入局部极小值。

III. 光滑与非光滑激活函数

本节大部分内容遵循 [5] 。数据再次被分为两个类别 $M = M_a \cup M_b$。本节的第一个结论指出，ReLU 的表现优于光滑激活函数（即 tanh）；第二个结论则强调深度，而非浅层网络能够很好地近似大多数函数这一事实。 [5]对真实数据和模拟数据进行了分析，从而得出这些观察结果。拓扑方面的工作通过拓扑数据分析完成，其依据是持久同调这一理念，该理念最初受 [6–8]启发
这项研究的有趣之处在于，它并非关注每个数据点的变化，而是观察数据形状的整体变化，从而获得更深入的认识。研究的重点是考察贝蒂数的变化，以及它在训练过程中受到的影响。



图 4. [5] 贝蒂数的变化：$\beta(red)$：(1，2，0) →(1，2，0) →(2，1，0) →(2，0，0) →(1，0，0) →(1，0，0);$\beta(green)$ :: (2，2，0) →(2，2，0) →(2，1，0) →(2，0，0) →(2，0，0) →(1，0，0)。

如图 4 所示，实际网络会像图中展示的那样改变数据的形状。为了分离各个连通分支，网络必须打破拓扑结构并改变贝蒂数。一般来说，对于二分类问题，目标是降低贝蒂数，使得 $\beta_1(M_a) = 0 = \beta_1(M_b)$ 和 $\beta_2(M_a) = 0 = \beta_2(M_b)$，而 $\beta_0(M_a) = 1 = \beta_0(M_b)$。请注意，最后一个条件并不十分严格。即使 $\beta_0(M_a) > 1$ 而其他 $\beta_j(M_a) = 0$ 对于 $j \geq 1$，分类结果仍然足够好。事实上，如果 $\beta_0 > 1$，则表明可能存在另一个类别（或子类），从而能更深入地了解数据。

这种假设面临的诸多障碍之一是，数据通常并不会以如此理想的形式出现。数据更可能以带有一些噪声的点云形式呈现。不过，这正是持久同调发挥作用的地方。少量噪声的存在不会改变点云数据的有效 β_j 。如果数据确实来自对某个流形结构的采样，那么持久同调几乎可以准确地恢复 β_j 。这篇论文的作者不仅研究了模拟数据中的拓扑变化，也研究了真实世界数据中的拓扑变化，包括图像。

这里将解答一些重要问题：1）为什么 ReLU 在经验上比其他方法表现更好？2）通过这种方法观察到的这些拓扑变化是否稳健？3）即使有逼近定理，为什么深度神经网络的表现仍优于浅层网络？

A. 拓扑

我们试图考察流形的拓扑复杂度，以及数据集的泛化差距，它衡量的是模型在测试集和训练集上的准确率差异。泛化差距定义如下：

$$GG(X_{train}, X_{test}, m) = Acc(X_{train}, m) - Acc(X_{test}, m)$$

其中 m 是模型。
对于分段线性函数，拓扑复杂度的上界可用线性区域来表示。然而，通过训练集确定的线性区域数量在受到微小扰动 [9]时并不稳定。这里关注的不是不同阶段的决策边界，而是数据空间变换。为了理解这些变换，可以观察贝蒂数的变化。此外，无需考察整个流形的贝蒂数，只需考察各个连通分支的贝蒂数变化过程即可。在实践中，通常很难计算点云的同调。

拓扑数据分析（TDA）的标准做法是：i) 去除离群点和噪声；ii) 构造 ϵ–维托里斯-里普斯复形；iii) 在不改变拓扑的情况下简化维托里斯-里普斯复形。根据噪声消除/平滑方式以及维托里斯-里普斯复形中的 ϵ 取值，i) 和 ii) 可能会改变拓扑结构。

B. 设置

1. 假设所讨论的问题是一个二分类问题， $M = M_a \cup M_b$并附加一个假设，$\inf\{||x-y||: x \in M_a, \ y \in M_b\} > 0$。

2. 对于模拟数据集，我们知道其流形，因此通过均匀且密集地选取大量样本来生成点云。所选神经网络采用前馈结构，深度为 l。

3. 网络 函数为 $v: s \circ f_l \circ \cdots \circ f_1: \mathbb{R} \to [0,1]$。s 是评分函数， $s: \mathbb{R}^{n_{l+1}} \to [0,1$]。4. 设 n_j 表示第 j 层的宽度。我们令 $n_1 = d$， $n_{l+1} = p$ 以及 $v_j: f_j \circ \cdots \circ f_1$。

5. 模拟数据在低维空间中并不真实，且具有复杂的拓扑结构。真实世界数据处于高维空间，但就纠缠类别而言，很可能更简单。对于模拟数据，我们关注较大的 β_0、非零的 β_j （针对 $j \geq 1$ ）以及较大的拓扑复杂度。

6. 训练该模型，使其泛化误差接近于零。

剧透预警：当深度足够时，最后一层会将 $v(T_a)$ $v(T_b)$ 映射到 $[0,1]$ 的两端。从模拟数据迁移到真实数据会面临一些挑战。由于拓扑未知，因此难以在每一层计算持久同调。

这项工作使用了三个不同的数据集，如下图 [5] 所示。数据的红色部分是M_b ，绿色部分是 M_a。

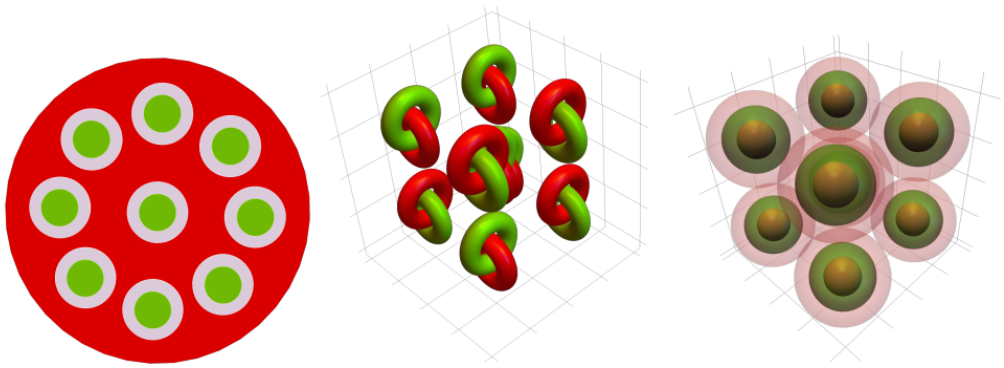


图 5. [5] a) $\beta(M_a) = (9, 0)$, $\beta(M_b) = (1, 9)$, b) $\beta(M_a) = (9, 9, 0)$, $\beta(M_b) = (9, 9, 0)$, c) $\beta(M_a) = (9, 0, 9)$, $\beta(M_b) = (18, 0, 9)$

C. 训练

1. 不同的激活函数：tanh、ReLU、带泄漏的 ReLU
2. 深度不同，范围为 4-10
3. 宽度不同，范围为 6-50
4. 交叉熵分类损失
5. 使用 Adam优化器训练 18000 个训练轮次
6. $\eta = 0.02 - 0.04$ 采用指数衰减 $\eta^{t/d}$, $d = 2500$ 且 t 为训练轮次。
7. 对于瓶颈架构（中间宽度较窄），d $= 4000$, $\eta = 0.5$
8. 分数 $=$ Softmax函数
9. 度量 δ_k 对于 VR_ϵ 复形，是 K近邻上的图测地距离。 $\delta_k(x_i,\ x_j)$ $=$ 它们之间在 K近邻图中的最小边数。它保持连通性并对距离进行归一化。
10. 两个超参数， k 和 ϵ。通过关于 δ_k 在 ϵ处的过滤复形来使用持续同调。首先通过 $\beta_0(VR_{k^*}) = \beta_0(M)$，找到具有 $\epsilon = 1$ 的 k^*。然后通过令 β_1 和 β_2相等来求出 ϵ^*。

D. 结果

1. 所有可能的神经网络架构中， β_0 都明显衰减。tanh 中衰减较慢，带泄漏的 ReLU 中较快，ReLU 中最快。

2. 宽度：
   - 较窄：每层 6 个神经元，拓扑变化更快
   - 瓶颈：其中一个中间层有 3 个神经元，而其他层有 15 个神经元，拓扑在瓶颈处突然变化。
   - 较宽：每层 50 个神经元，拓扑复杂度的降低更平缓。
3. 深度：
   - 对于高度纠缠的类别，需要更深的网络才能达到 100% 的准确率。深度不足会使我们难以达到所需的准确率
   - 最后几层的拓扑复杂度下降得更快。
   - 有时，更深的神经网络会出现非必要的简化。
4. 对于宽度足够的层， $\beta_1(M) = 0$和 $\beta_2(M) = 0$并非必需。浅层网络有时会迫使这种情况发生，但更深的网络则不会，因此能保留更多结构。

E. 图形结果

F. 真实数据集

这些性质已在 MNIST、HTRU2、UCI банкнот 和 UCI 无传感器驱动数据集上得到验证。由于这些数据集的维度较高，持久同调无法高效执行，因此需要在每一层上进行计算。泛化差距有所放宽，幅度为 $2 - 5\%$。网络的宽度和深度保持固定。

MNIST: $\mathbb{R}^{784}$，对于该数据集，拓扑观察基于排名前 50 的主成分进行；HTRU2 数据集:$\mathbb{R}^8$，UCI банкнот: $\mathbb{R}^4$，UCI 驾驶数据集: $\mathbb{R}^{49}$

结果表见附录。观察结果与预期一致：

1. 拓扑 整体复杂度有所降低。网络试图将 β_0 降至 1，并将其他 β_j 降至 $0.$

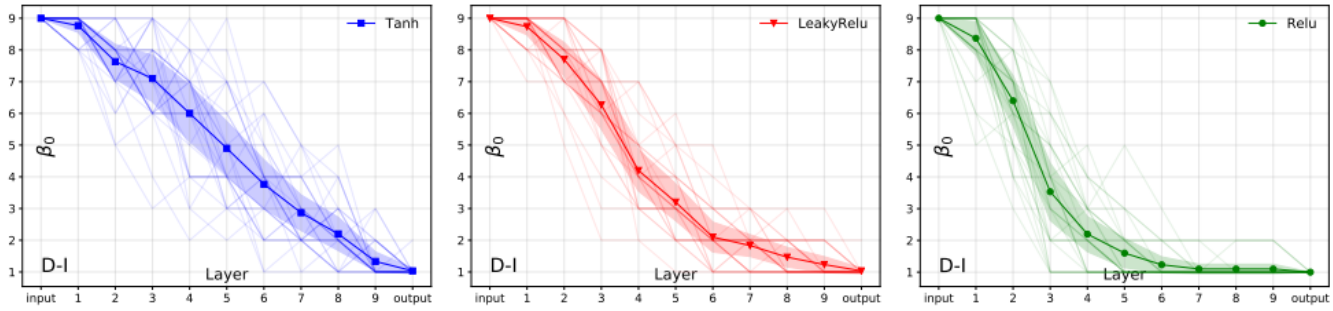


图 6. [5] 数据集 I 中不同激活函数对应的 β_0 平均变化。深色曲线表示平均值，阴影区域表示多次模拟的结果。

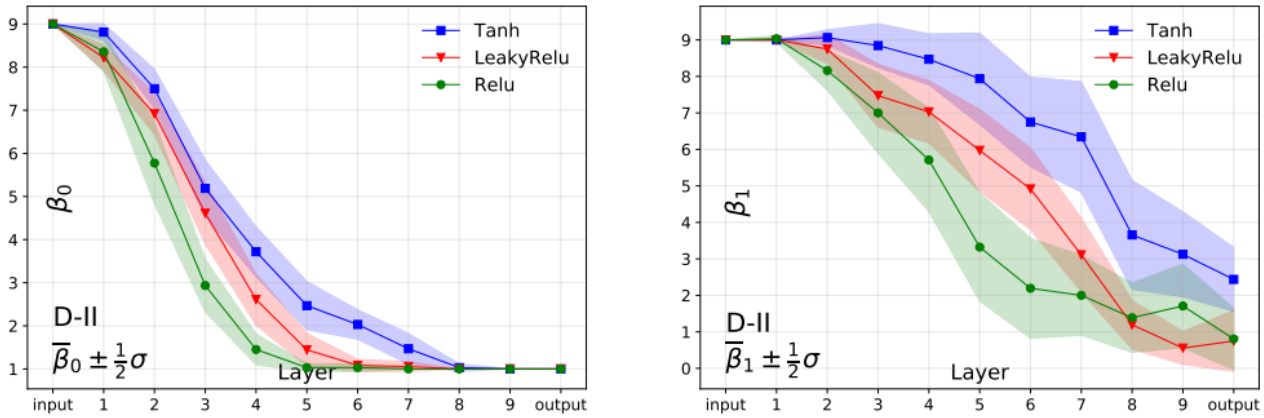


图 7. [5] 数据集 II 中不同激活函数对应的 β_0 和 β_1 变化。

2. 与非光滑激活函数相比，光滑激活函数降低拓扑复杂度的速度更慢，而 ReLU 所实现的简化程度最大。3. ReLU 会对数据空间进行折叠，但并非完全折叠。这可直接从 ReLU 或绝对值函数的定义得出。4. 层数越多，模型越容易训练，因为它可以“逐步花时间一点一点地简化空间”。

IV. 未来工作

这一方向的研究可能性无穷无尽，以下是一些近期研究实验的可行选项：

A. 不同层中的不同激活函数

在多个数据集上训练模型，或使用复合架构训练数据集时，将初始层的激活函数设为光滑函数，并在末端应用 ReLU 或带泄漏的 ReLU，或许能更好地转换数据。其背后的想法是，光滑激活函数能在相同维度中保留结构，而将其嵌入到更高维度中，则可以降低某些复杂性，同时仍保留部分结构。

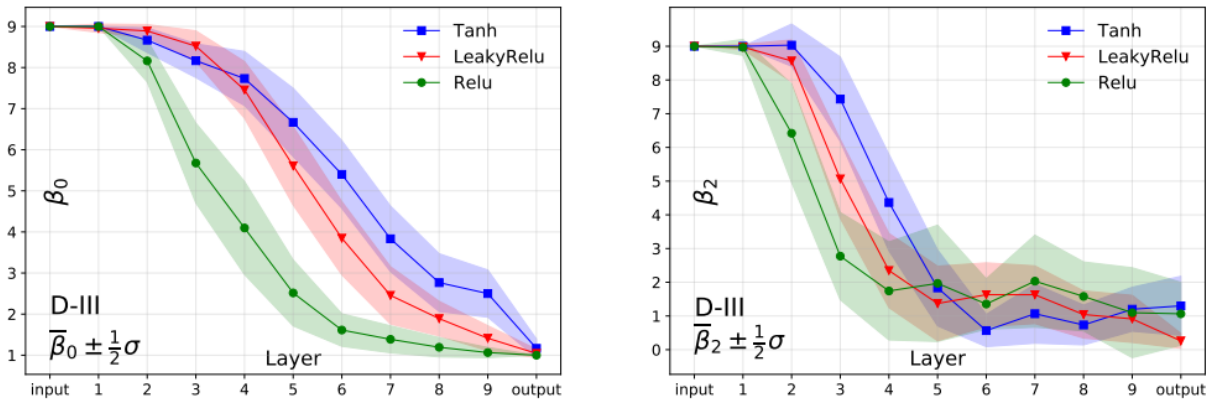


图 8. [5] 数据集 III 中不同激活函数下 β_0 和 β_2 的变化。

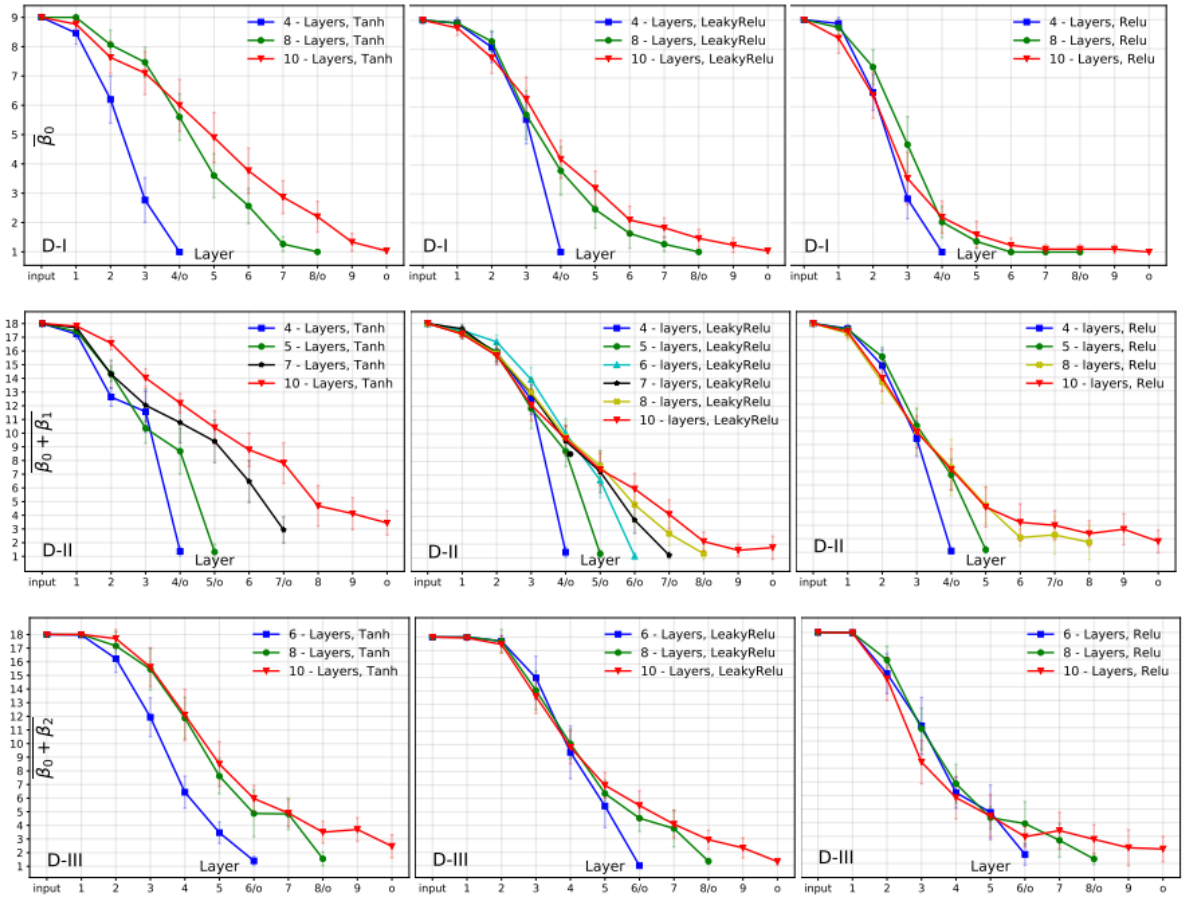


图 9. [5] 三个数据集的拓扑复杂度随网络深度的变化。

B. 时间复杂度与抗噪鲁棒性之间的关系

关于 ReLU 导致的不同线性折叠，有一些结果表明它们在噪声下并不稳定。过度简化与时间之间似乎存在权衡。此外，探究在某些层使用非 ReLU 激活函数的网络架构中，折叠的稳定性也会很有意思。

C. 不同的神经网络架构

应比较同一数据集上的不同架构（例如卷积神经网络、简单前馈网络和残差网络）在拓扑复杂度方面的变化。

V. 参考文献

[1] Vinod Nair 和 Geoffrey E Hinton。修正线性单元改进了受限玻尔兹曼机。载于 *ICML 2010*，第 807–814 页，2010 年。[2] Andrew L. Maas。整流器非线性改进了神经网络声学模型。2013 年。[3] Xavier Glorot、Antoine Bordes 和 Yoshua Bengio。深度稀疏整流神经网络。载于 Geoffrey Gordon、David Dunson 和 Miroslav Dud´ık 编，*Proceedings of the Fourteenth International Conference on Artificial Intelligence andStatistics*，第 15 卷（*Proceedings of Machine Learning Research*），第 315–323 页，美国佛罗里达州劳德代尔堡，2011 年 4 月 11–13 日。机器学习研究论文集。网址：https://proceedings.mlr.press/v15/glorot11a.html。[4] C. Olah。神经网络、流形与拓扑，2014 年。网址：http://colah.github.io/posts/ 2014-03-NN-Manifolds-Topology/。[5] Gregory Naitzat、Andrey Zhitnikov 和 Lek-Heng Lim。深度神经网络的拓扑。 *J. Mach. Learn. Res.*，21(1)，2022 年 6 月。ISSN 1532-4435。[6] G. Carlsson。数据的形状，2012 年。[7] Gunnar Carlsson。点云数据的拓扑模式识别。*Acta Numerica*，23：289–368，2014 年。doi：10.1017/S0962492914000051。[8] Gunnar Carlsson 和 Afra Zomorodian。多维持久性的理论。载于 *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry* 07, 184 193, 2007 ，SCG ’ 页码 – 美国纽约州纽约市，。计算机协会。ISBN 9781595937056。doi:10.1145/1247069.1247105。网址：https://doi.org/10.1145/1247069. 1247105。[9]Liwen Zhang、Gregory Naitzat 和 Lek-Heng Lim。深度神经网络的热带几何。载于 Jennifer Dy 和 Andreas Krause 编，*Proceedings of the 35th International Conference on Machine Learning* 80 *Proceedings* ，第 卷 *of Machine Learning Research* 5824 5832 10 15 2018 80 （），第 – 页。机器学习研究论文集，– 7 月。网址：https://proceedings.mlr.press/v / zhang18i.html。[10] D. Vijay Anand、Zhenyu Meng、Kelin Xia 和 Yuguang Mu。用于渗透调节剂分子聚集和氢键网络分析的加权持久同调。 *Scientific Reports*，10，2019 年。

附录 A：定义

以下列出的定义 为了便于更广泛的读者理解，这些定义比起形式化定义更为直观 .

定义 A.1. 同胚：如果函数 f 连续，且 f 是双射（单射且满射），并且 f^{-1} 连续，则函数 $f: M \to N$ 称为同胚。

定义 A.2. 连通分支：如果空间 M 不能表示为两个分支的不交并，则称其为连通空间。连通分支的数量等于彼此连通的分支的基数。

定义 A.3. 同调：设 M 是一个流形（局部同胚于某个 $n \in \mathbb{N}$ 的 $\mathbb{R}^n$ ）。 0^{th} 同调 $\mathbb{H}_0(M)$ 定义为 M 的连通分支数量。一阶同调 $\mathbb{H}_1(M)$ 与流形 M 中可以嵌入多少个不同的圆（非平凡）有关。根据它们的数量及相互关系，可以为其赋予一个整数系数群。类似地，也可以通过流形中存在多少个球形孔洞来定义 $\mathbb{H}_2(M)$ 。贝蒂数表示同调群的秩。例如 β_0 =秩$(H_0(M))$ β_1 =秩$(H_1(M))$ β_2 =秩$(H_2(M))$

定义 A.4。持久同调：假设环境空间中有一些数据点。从每个点出发，构造一个半径为 ϵ的球。如果两个这样的球相交，就添加一条边。如果三个球两两相交，且三者交集也非空，则绘制一个三角形面。通过这种方式，根据交集逐渐增大每个数据点周围球的半径 ϵ ，便可从数据集构造出一个单纯复形。

在每个半径值下，都会计算数据集的同调。持续时间最长、也就是保持不变的同调，被视为数据集（流形）的同调。

定义 A.5. 拓扑复杂度：对于一个 d 维流形，它是 $\sum_{j=1}^{d} \beta_j$ 。

定义 A.6. 单纯复形：由点、边、面、四面体等组成的集合，其中这些元素的所有面也都属于该复形；此外，若两个元素相交，则它们相交于一个面（对应于该对象）。

定义 A.7. ***Vietoris-Rips*** 复形：这是在持久同调定义中，为不同的 ϵ 半径值构造的单形。

附录 B：示例

同调示例：

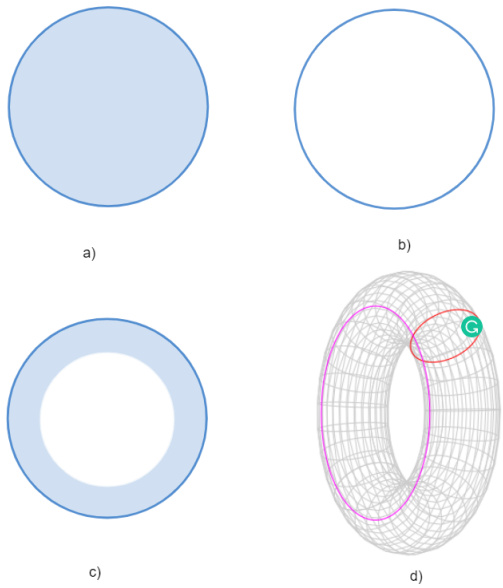


图10. 图像分别为：a) 圆盘 b) 圆 c) 圆环 d) 环面 $\mathbb{H}_0$ 这四个部分中的每一个都是 $\mathbb{Z}$ ，因为只有一个连通分支。

a) $\mathbb{H}_1(D^1) = e$ 这意味着平凡 l. 圆盘内的任何小圆都可以收缩成一个点，并且始终处于圆盘内部。

b) $\mathbb{H}_1(S^1) = \mathbb{Z}$ 因为只有一个圆。

c) $\mathbb{H}_1(Annulus) = \mathbb{Z}$。圆有两种类型：一种绕过孔洞，另一种较小，完全位于蓝色区域内。较小的圆可以收缩成一个点，而另一种则不能。因此，它是 $\mathbb{Z}$ d) $\mathbb{H}_1(\mathbb{T}^2) = \mathbb{Z} \times \mathbb{Z}$。有两个不同的非平凡圆：红色的圆和紫色的圆。它们无法在不离开环面的情况下连续变形为彼此。

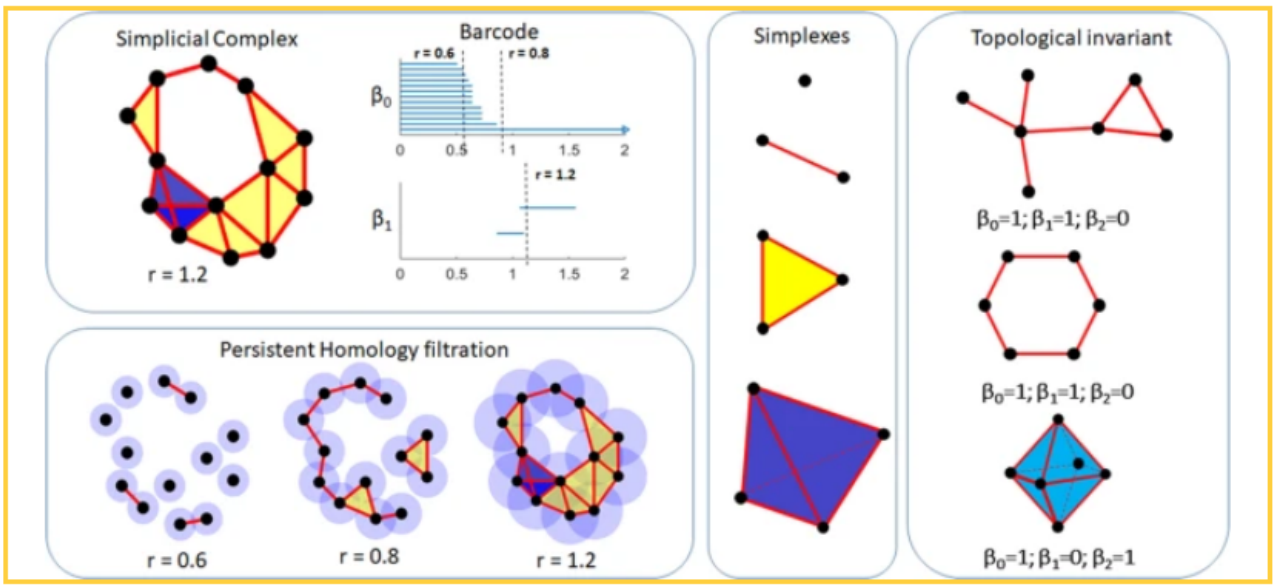


图11. [10] 基于半径和单纯复形同调的持久同调过滤、持久条形码、拓扑不变量

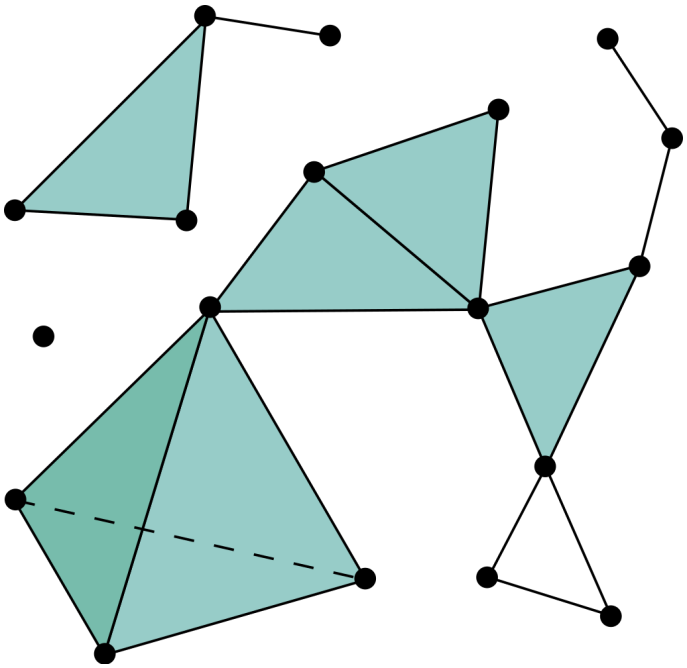


图12. 三维单纯复形