arXiv:2305.01638v2 [cs.LG] 2 Nov 2023

S使用多分辨率卷积记忆进行序列建模 y

Jiaxin Shi [1] **Ke Alexander Wang** [1] **Emily B. Fox** [1 2]

摘要

高效捕捉序列数据源中与特定任务相关的长程模式（例如分类和生成建模），是一项根本性挑战。当前主流方法各有取舍：Transformer 等方法需要承担穷举枚举与比较带来的内存负担；循环神经网络等方法需要承担复杂序列依赖带来的计算负担；而卷积网络则需要承担大量或大型滤波器带来的参数负担。我们转而从基于小波的多分辨率分析中汲取灵感，为序列建模定义一种新的构建模块，称为 MULTIRESLAYER。我们模型的关键组件是多分辨率卷积，用于捕捉输入序列中的多尺度趋势。我们的 MULTIRESCONV 可通过共享 滤波器，在扩张因果卷积树上实现。因此，它兼具卷积网络的计算优势与小波分解的严谨理论依据。我们的 MULTIRESLAYER 易于实现，所需参数显著更少，并且长度为 N 的序列最多只需 $\mathcal{O}(N \log N)$ 的内存占用。然而，通过堆叠此类层，我们的模型在使用 CIFAR-10、ListOps 和 PTB-XL 数据集的多个序列分类与自回归密度估计任务上取得了最先进的性能。

1. 引言

序列建模中的一项关键挑战，是对数据中有助于特定任务（如分类、预测或聚类）的长期模式进行概括，或记忆。根据定义，模式是数据中由多个时间步产生的高层结构。然而，模式可能会出现在多个层次上，对应不同的时间尺度。例如，在研究能源消耗时，模式化的变化可能出现在日内、不同日期之间以及季度之间。生理时间序列中也存在类似显著的多尺度趋势，例如糖尿病患者的异常血糖模式，以及心律失常患者的异常心跳。语音音频信号可以用话语、音素和次音素来描述。图像和视频的多尺度结构也已得到充分研究。即使对于无法明确进行多尺度解读的数据来源，多尺度建模方法也能有效捕捉长程模式。

本文提出一种通用且可复用的序列建模构建模块——MULTIRESLAYER——利用多尺度方法记忆过去的数据。我们从多分辨率分析（MRA）的视角看待记忆（Willsky，2002），尤其强调小波分析这一来自信号处理的强大工具，可用于压缩、去噪、特征提取等（Jaw-erth &Sweldens，1994；Akansu 等，2001）。如第2节所述，小波分析可以通过计算高效的方式实现，并可解释为一系列卷积。然而，使用小波只是我们的设计选择，也可以考虑其他多分辨率分析技术，以记忆不同时间尺度上的模式。

受小波启发，MULTIRESLAYER 的核心组件是多分辨率卷积运算（MULTIRESCONV，它保留了多分辨率分析的整体树结构。我们展示了如何在序列的每个时间步，使用 MULTIRESCONV 构建过去信息的记忆，并将其整体实现为一组精心布置的膨胀因果卷积，各层之间共享滤波器。不过，与传统小波分析不同的是，我们学习这些滤波器，并以端到端的方式进行训练。将滤波器固定为预定义的小波滤波器后，MULTIRESCONV便退化为传统的离散小波变换；不过，我们在第5.5节展示了学习滤波器的优势。基础的 MULTIRESCONV 构建模块可以采用深度学习模型中常见的多种方式进行堆叠（例如，跨多个通道堆叠、纵向堆叠为多层等）。我们的模型在使用树结构空洞卷积方面类似于 WaveNet（Oord 等人，2016a），但我们基于原理设计的模型具有独特的跳跃连接结构和滤波器共享模式，因此显著提升了

[1]斯坦福大学 [2]CZ Biohub SF。通信作者：Ji- axin Shi <jiaxins@stanford.edu>。

论文集： 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023。版权所有 © 2023，作者所有。

参数效率和性能（详情见第4 节）。

序列建模领域的机器学习文献正在迅速增长。常见的方法包括循环网络的变体（Hochreiter 和 Schmidhuber，1997年）、自注意力网络（Vaswani 等人，2017年）以及状态空间模型（Gu 等人，2021年）。进一步讨论请参见第4 节。我们的 MULTIRESLAYER 相较于以往的这类工作具有以下关键优势：

- **架构简单：** 我们这一层的核心是简单的空洞卷积和线性变换。
- **训练高效：** 我们的层可以轻松地在执行卷积的硬件加速器上并行化。
- **参数高效：** 我们的层在深度方向堆叠的空洞卷积中复用滤波器。

同样，通过利用多分辨率分析结构，我们从一个有原则且可解释的框架出发，来思考序列建模中的记忆。此外，我们还可以借鉴丰富的多分辨率分析文献，探索模型的推广，例如用于平移不变表示学习的平移不变小波变换（Kingsbury，1998；Se- lesnick 等人，2005），以及扩展到多个输入维度等。

我们的实证评估涵盖序列图像分类和自回归生成建模（CIFAR-10）、语法树推理（ListOps），以及心电图的多标签分类（PTB-XL）。我们还指出，所提出的 MULTIRESCONVs 可轻松应用并扩展至表征学习和长期预测等其他任务。同样，尽管我们专注于序列分析，但所提出的思路也可推广至具有多分辨率结构的其他数据领域，例如图像和视频。探索 MULTIRESLAYER 在这些场景中的应用，是一个令人期待的未来研究方向。

2. 背景：小波分解

与傅里叶变换的频域分析不同，小波提供了时频分析。具体而言，小波是一种具有有限支撑的基函数，并具有多分辨率结构，也就是说，基函数被划分为具有不同分辨率的组——有些聚焦于极短时间尺度上的“局部”函数值，另一些则捕捉较长时间尺度上的更多“全局”结构。接下来，我们将以最简单的小波族——哈尔小波为例，阐释小波多分辨率分析的概念。涵盖所有正交小波的形式化论述见附录A。

假设我们想在时间区间 $[0,1$ 上近似信号 $f(t)$。我们能够得到的最粗略近似是 $\hat{f}^{(0)}(t) \triangleq a_{0,\,0}\phi(t)$，其中 $\phi(t) = \mathbf{1}(0 \le t < 1)$，而 $a_{0,\,0} = \int_0^1 f(t)dt$ 是 f 的平均值。我们使用上标 0 表示这是最低分辨率的

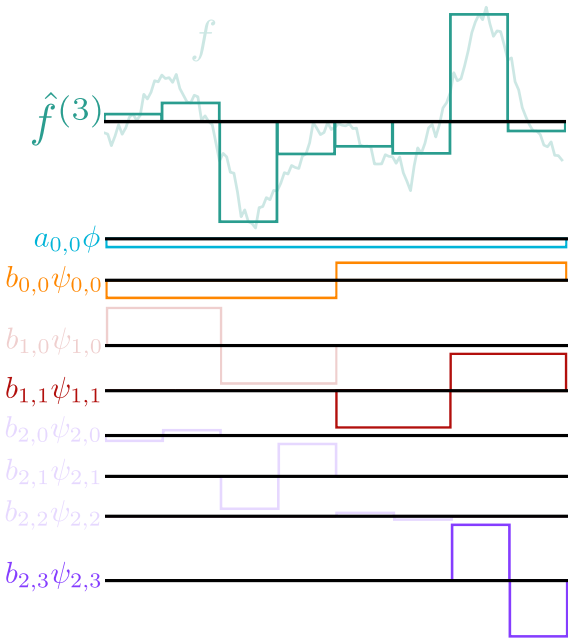

图 *1.* 在标准的多分辨率分析（MRA）中，我们使用 $\hat{f}^{(j)}$近似连续信号 f。这里，我们将 $\hat{f}^{(3)}$ 及其分解过程可视化，展示其如何分解为一系列函数之和，这些函数能够在二叉树划分的输入空间上捕获分辨率逐渐提高的结构，对应于公式（1）。属于同一层级的分量具有相同的颜色。每个分辨率层级中不透明的分量保留了分解过程中的最新信息。仅保留这些分量，相当于采用具有“分辨率渐隐”TREESELECT 的 MULTIRESLAYER。

近似。我们将 f 划分为两个单位区间，并按如下方式近似 f，就能得到更好的近似：f： $f(t) \approx \hat{f}^{(1)}(t) \triangleq a_{1,\,0}\phi(2t) + a_{1,\,1}\phi(2t-1)$，其中$a_{1,\,0} = \int_0^{1/2} f(t)dt$，以及 $a_{1,\,1} = \int_{1/2}^1 f(t)dt$。

我们可以重复这一过程，将区间减半、重新缩放并平移 ϕ，从而得到更精细的近似 $\{\hat{f}^{(j)}\}_{j\in\mathbb{N}_0}$。每个 $\hat{f}^{(j)}$ 都是紧支撑基函数的线性组合，$\{\phi_{j,k}(t) \triangleq 2^{j/2}\phi(2^j t - k)\}_{k\in\mathbb{Z}}$，其分辨率层级由 j 编号：

$$\hat{f}^{(j)}(t) = \sum_{k\in\mathbb{Z}} a_{j,k}\phi_{j,k}(t), \quad \text{where } a_{j,k} = \langle f, \phi_{j,k}\rangle.$$

对于每个层级 $j \in \mathbb{N}_0$，子空间 $V_j \triangleq \text{span}(\{\phi_{j,k}\}_{k\in\mathbb{Z}})$ 包含在长度为$1/2^j$的区间上保持常数的函数。换言之，V_j 中的基函数描述了 f 不大于 $\Delta t \sim 1/2^j$ 的时间尺度的结构。对于足够大的 j，V_j 能够以任意高的精度逼近任意连续时间序列。

可以尝试通过将系数 $\{a_{j,k}\}_{k\in\mathbb{Z}}$ 收集到一个向量中，来概括或表示 f。尽管这些系数共同完整描述了近似值 $\hat{f}^{(j)}$，但每个单独的系数可能过于局部，无法代表 f 中的结构。每个 $a_{j,k}$ 只概括了 f 的值。

在 $1/2^j$ 区间内，而模式可能会出现在更大的区间中。我们需要多个 $a_{j,\ k}$ 来概括这些更大尺度的结构。有没有办法生成一组系数，使每个系数都概括不同尺度上的一种结构?

在互不重叠的分辨率上表示结构。我们确实可以利用多分辨率分析（MRA）中的工具来生成这种表示。在多分辨率分析中，我们反复将 V_j 分解为一个较低分辨率的子空间 V_{j-1} 及其正交补空间 W_{j-1}之和: $V_j = V_{j-1} \oplus W_{j-1}$。由于 V_j 和 V_{j-1} 中的基函数分别描述比 $\Delta t \sim 1/2^j$ 和 $\Delta t \sim 1/2^{j-1}$更粗尺度上的结构，因此，W_{j-1} 中的基函数表示的结构恰好位于 $1/2^j$ 尺度上，并由基系数 $\{b_{j,k}\}_{k\in\mathbb{Z}}$概括。从某个高分辨率层级 J 开始，重复这一过程，我们便得到

$$V_J = V_{J-1} \oplus W_{J-1} = V_0 \oplus W_0 \oplus \ldots \oplus W_{J-2} \oplus W_{J-1}$$

并且，如图 1 所示，

$$f(t) \approx \hat{f}^{(J)}(t) = a_{0,0}\phi(t) + \sum_{j'=0}^{J-1}\sum_{k\in\mathbb{Z}} b_{j',k}\psi_{j',k}(t). \quad (1)$$

基函数 $\{\psi_{j,k}\}$ 称为哈尔小波，而 ϕ称为它们的尺度函数；其函数形式见附录A.1 。系数[1] $\{a_{0,0}\} \cup \{b_{0,k}\}_{k\in\mathbb{Z}} \cup \ldots \cup \{b_{J-1,k}\}_{k\in\mathbb{Z}}$现在概括了 f 在多个分辨率下的结构，范围从 $1/2^0$ 到 $1/2^{J-1}$。

计算表示。我们最初总结以下多分辨率结构的问题 f 归结为计算近似$a_{0、\ 0}$、$\{b_{j,k}\}$的哈尔小波基系数 $\hat{f}^{(J)} \in V_J$。关于如何计算哈尔小波的这些系数，请参见附录 A.1 。一般而言，我们可以使用离散小波变换高效且递归地计算任意小波族的这些系数（DWT；参见附录A.3）。

在附录D中，我们展示了小波变换的表示能力。具体而言，我们考虑一个原始音频波形，它捕获了一段采样率为 16,384 的 1 秒录音。我们使用一棵 10 层小波树，通过总计 2068 个系数来重建音频信号。小波变换能够在这一长序列中“记忆”音频信号的许多重要模式。这种表示能力启发了我们在第ULTIRESLAYER 节概述的3。

3. 多分辨率卷积的序列建模

我们利用离散小波变换的计算结构，为序列构建多分辨率记忆。给定一个序列 $\mathbf{x} \in \mathbb{R}^N$，它表示一个离散采样信号，离散小波变换可以通过以下递归计算来实现，以求得$\mathbf{a}_0$ 和$\mathbf{b}_{0:J-1} \triangleq$（$\mathbf{b}_0, \mathbf{b}_1$，……，$\mathbf{b}_{J-1}$），计算从$\mathbf{a}_J(n) = \mathbf{x}(n)$[2]开始:

$$\mathbf{a}_j(n) \triangleq a_{j,n} = \sum_{k=0}^{K-1} \mathbf{a}_{j+1}(2n+k)\mathbf{h}_0(k),$$
$$\mathbf{b}_j(n) \triangleq b_{j,n} = \sum_{k=0}^{K-1} \mathbf{a}_{j+1}(2n+k)\mathbf{h}_1(k),$$

其中，滤波器 $\mathbf{h}_0$,$\mathbf{h}_1 \in \mathbb{R}^K$ 由小波类别决定。对于哈尔小波，我们有 $\mathbf{h}_0 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ 和 $\mathbf{h}_1 = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$。为使底层计算与滤波器的选择相互独立，我们将这一由卷积和多尺度下采样组成的过程定义为多分辨率卷积操作:

$$\mathbf{a}_0, \mathbf{b}_{0:J-1} = \textsc{MultiresConv}(\mathbf{x}, \mathbf{h}_0, \mathbf{h}_1, J). \quad (2)$$

回顾一下，系数 $\{\mathbf{a}_0(n)\}$ 和 $\{\mathbf{b}j(n)\}$ 构成了$\mathbf{x}$的多分辨率表示。当滤波器取值来自小波时，我们可以通过逆转递归过程，从系数 $\mathbf{a}_0$,$\mathbf{b}_{0:J-1}$ 中完美重建原始序列 $\mathbf{x}$。换句话说，这一过程足以让我们凭借这些系数，完整记忆过去。

然而，我们不将滤波器设为固定值，而是提出将 MULTIRESCONV 用作序列模型的构建模块，并让$\mathbf{h}_0$, $\mathbf{h}_1$ 可学习。学习滤波器使我们能够超越手工设计的小波，同时仍在计算中保留多分辨率结构。尽管我们可能会失去完美重建输入的能力，但我们将在实验中看到，这样做换来了显著的预测性能提升。

3.1. 通用序列建模层

大多数具有竞争力的序列模型都采用将一个序列映射为另一个序列的结构$\mathbf{x} \in \mathbb{R}^N$ 到另一个序列$\mathbf{y} \in \mathbb{R}^N$。每个时间步的输出应当：1）与该步输入相关；2）捕获来自先前时间步的潜在长程依赖。这可以通过穷举并比较来实现（Vaswani 等人，2017年），但其复杂度相对于输入长度呈二次增长。通过维护过去信息的定长记忆，可以避免二次扩展，循环模型便是如此。然而，定长记忆会随着时间推移而丢失信息，因此我们必须确保记忆始终保留最相关的历史信息。

[1] $\{a_{j,k}\}$ 和 $\{b_{j,k}\}$ 在信号处理中称为近似系数和细节系数。请注意，有些资料（例如 Lee 等人 （2019年）称 V_J 为第 0 层，V_{J-1} 为第 1 层，依此类推。

[2] 为处理有限序列，如有必要，我们会在每次递归迭代之前在$\mathbf{a}^{j+1}$ 左侧补零，以确保$\mathbf{a}^{j+1}$ 中的所有元素都不会被排除。

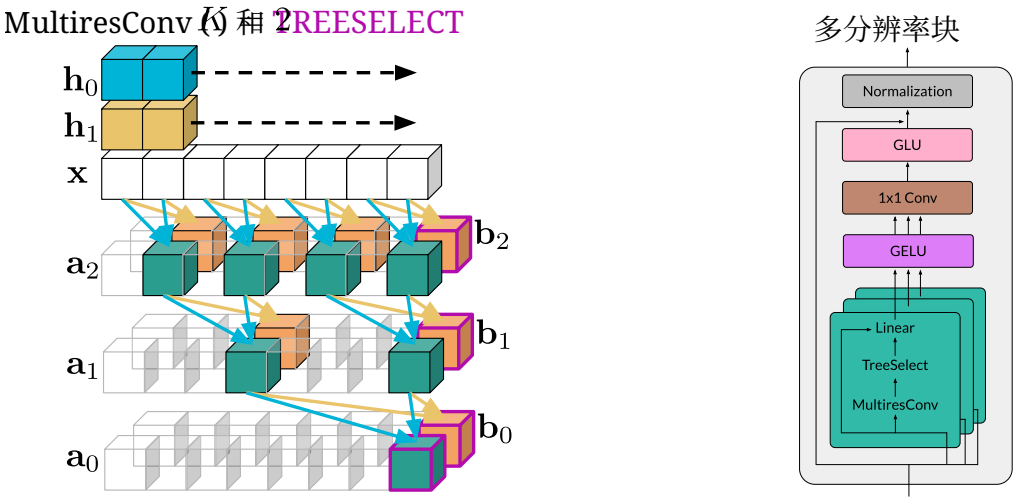

图 2. （左）MULTIRESCONV 由一系列空洞卷积组成，这些卷积在所有层级共享相同的滤波器 $\mathbf{h}_0$,$\mathbf{h}_1$。这里展示了采用“分辨率渐隐”策略的 TREESELECT：在$\mathbf{a}_0$,$\mathbf{b}_{0:J-1}$的每个层级保留最右侧的系数，如洋红色轮廓所示。（右）MULTIRESBLOCK 架构示意图。输入序列的每个通道都在 MULTIRESLAYER 中独立处理。1×1卷积用于混合各通道的信息。我们将多个多分辨率块堆叠起来，构建深层序列模型。

我们将式（MULTIRESCONV 2)中的 MULTIRESCONV 操作用作记忆机制，以捕获多个分辨率下的历史模式。核心思路是在当前时间步 t，对截至该步的序列执行 MULTIRESCONV，$\mathbf{x}(0:t)$，并选取一组相关的表示系数子集（$\mathbf{a}_0^t$，$\mathbf{b}_{0:J-1}^t$），在步骤 t 形成记忆向量$\mathbf{z}_t \in \mathbb{R}^M$：

$$\mathbf{a}_0^t, \mathbf{b}_{0:J-1}^t = \textsc{MultiresConv}(\mathbf{x}(0:t), \mathbf{h}_0, \mathbf{h}_1, J), \quad (3)$$

$$\mathbf{z}_t = \textsc{TreeSelect}((\mathbf{a}_0^t, \mathbf{b}_{0:J-1}^t), M). \quad (4)$$

第 t 步的输出$\mathbf{y}(t)$由该步的输入和记忆通过线性变换生成：

$$\mathbf{y}(t) = \mathbf{w}^\top \begin{bmatrix} \mathbf{z}_t \\ \mathbf{x}(t) \end{bmatrix}.$$

我们对 $t = 0$、1、.....、$N-1$ 重复这一过程，以获得完整的输出序列 $\mathbf{y} \in \mathbb{R}^N$。该架构的示意图见图 2，我们将在第 3.2 节讨论系数选择过程 TREESELECT。

有趣的是，当 J 对所有 t都取相同值时，MULTIRESCONV在各个时间步上的结果可以整体视为多层膨胀因果卷积，与 WaveNet（Oord 等人，2016a）中的卷积类似。因此，这里的所有计算都可以在线性于序列长度的时间内完成，并且能够在实现卷积的现代硬件上大规模并行化。默认情况下，我们设定

$$J = \left\lceil \log_2 \left(\frac{N-1}{K-1} + 1 \right) \right\rceil \quad (5)$$

使得 $\mathbf{a}_0^t$ 只有一个元素$a_{0,\ 0}^t$ 且其感受野覆盖整个 $x(0:t)$，适用于任意 $t \leq N$。

3.2. 记忆机制

理想情况下，我们希望保留整个序列 $\mathbf{a}_0^t$，$\mathbf{b}_{0:J-1}^t$，其中 $t = 0$、.....、N。然而，存储所有这些系数需要占用关于 N 的二次空间。因此，我们需要一种方法，从 MULTIRESCONV 树中选择最能概括历史信息的系数。我们将这一操作称为 TREESELECT。

我们研究了两种策略，用于确定将 MULTIRESCONV 输出的哪些部分纳入记忆：

(1) **随时间均匀：** 传统上，在信号压缩和去噪中使用小波时，会通过阈值处理去除时间局部化的高频分量中大于特定尺度 j 的部分。这里，我们可以通过保留$\mathbf{a}_0^{(t)}$和$\mathbf{b}_0^{(t)}$、.....、$\mathbf{b}_j^{(t)}$ 中低于阈值 j 的部分来实现同样的效果。

(2) **分辨率渐隐：** 另一种策略是随着时间接近 t，逐步提高近似分辨率。这可以通过在每个分辨率层级保留最高索引处的$\mathbf{a}_0^{(t)}$ 和 $\{\mathbf{b}_j^{(t)}\}$ 项来实现。我们在图1和2中对此进行了可视化。这种近似使我们能够保留对重构近期输入值最重要的系数。

尽管我们在第5.5 节进行的消融研究未显示这两种策略之间存在统计学上的显著差异，但我们在所有实验中都采用分辨率渐隐，使模型倾向于着重记忆最近的信息，同时也让实现更简单——我们在附录B中提供了 PyTorch 实现示例。类似地，提高近期历史信息的权重已被证明对状态空间模型至关重要（Gu 等人，2020a；2021年）。

除了上述方法之外，一个显而易见的下一步选择是针对当前任务动态选择相关的多分辨率组件，或许可以通过注意力机制或 L_1 正则化来实现。我们将此留作未来工作。

3.3. 多个通道与混合层

以上内容仅描述了单通道序列之间的映射。为了处理输入 $\mathbf{x} \in \mathbb{R}^{d \times N}$ 具有 d 个通道的情况，我们采用一种简单策略：堆叠 d 个独立层，并可通过 逐通道 因果空洞卷积高效实现。各层的 d 输出序列经过非线性激活函数后，再通过一个 1×1 卷积进行混合。这里我们使用 GELU 激活函数（Hendrycks 和 Gimpel，2016）。值得注意的是，先进行逐通道卷积、再跨通道混合的做法，已在许多序列建模实践中证明是有效的（Wu 等人，2018；Gu 等人，2021）。

3.4. 堆叠为深层模型

我们将序列建模层和混合层封装在一个残差网络块中（He 等人，2016），并采用门控线性激活（Dauphin 等人，2017年）和恒等跳跃连接（见图2）。我们观察到，门控机制能够提升模型性能，这与其他多种序列模型中的观察结果一致（Dauphin 等人，2017年；Oord 等人，2016b；a；Gu 等人，2021年）。随后，我们将多个残差块堆叠成一个深层模型。根据任务的不同，在每个残差块之后应用层归一化（Ba 等人，2016）或批归一化（Ioffe 和 Szegedy，2015）。我们将该模型称为 MULTIRESNET，并在实验中使用它。

3.5. 优化与正则化

我们使用 Adam 优化器，采用默认超参数和解耦权重衰减（Loshchilov 与 Hutter，2018年）。观察到过拟合时，会在 GELU 和门控线性激活函数之后应用 Dropout。

4. 相关工作

近年来，受音频建模（Waibel 等人，1989年；Oord 等人，2016a）和自然语言处理（Collobert 等人，2011年；Kalchbrenner 等人，2016年）等应用的启发，研究人员在构建用于序列数据的卷积神经网络方面开展了大量工作。近期研究旨在通过采用门控单元（Dauphin 等人，2017年）、逐通道卷积（Kaiser 等人，2018年）、输入依赖权重（Shen 等人，2018年；Wu 等人，2018年）、跨深度权重绑定（Bai 等人，2019年）以及自适应滤波器尺寸（Romero 等人，2022a）来提升此类模型的性能。我们的 MULTIRESCONV 融合了这些研究中的许多要素，但在两个方面具有独特之处。首先，我们的方法明确地定义了记忆单元。$\mathbf{z}_t$，这对于建模长程依赖关系非常重要。其次，我们的 MULTIRESCONV 具有更完善的理论基础，因为当滤波器被指定为小波滤波器时，它会退化为标准离散小波变换。

在所有卷积序列模型中，与我们的工作最为接近的是 Oord 等人提出的 WaveNet 架构（2016a）。如图 2 所示，当滤波器尺寸为 2 时，我们针对 a_js 的计算图与 WaveNet 具有相同的连接模式。这两种模型均采用树状扩张因果卷积实现。不过，这两种模型之间有三个主要区别：（1）我们通过 TREESELECT 显式构造记忆，从而形成了与 WaveNet 不同的跳跃连接结构；（2）我们的模型在所有时间尺度上使用相同的滤波器；（3）在不同时间尺度的卷积之间，我们不使用非线性激活函数。因此，我们的架构更简单，且使用的参数远少于 WaveNet。此外，我们建立的小波与树状扩张因果卷积之间的联系，首次为 WaveNet 在原始音频波形建模方面的有效性提供了有原则的解释；原始音频波形是具有多尺度结构的长序列的典型示例。

循环神经网络及其线性变体（例如状态空间模型）也会显式地保留对过去信息的记忆。S4（Gu 等人，2021年）尤其值得关注，因为得益于线性递归，它们可以实现为卷积。然而，这类模型的卷积核与输入序列一样长，而要高效计算这些卷积核，则依赖复杂的参数化和近似技术。尽管近期的一些进展（Gupta 等人，2022年；Gu 等人，2022年；Smith 等人，2022年）解决了这一问题，但初始化这些模型仍需特别处理。许多此类模型采用的共同初始化机制称为 HiPPO（Gu 等人，2020a），其目标是通过投影到正交多项式来记忆历史数据。尽管这与小波（即函数空间中的另一种基）有关，但从表 5中可以看出，我们的方法对初始化并不敏感——标准初始化方案（Glorot 和 Bengio，2010年）与小波初始化的表现同样出色。附录 D 针对音频信号重建任务，提供了一些关于小波基优于 HiPPO 多项式基的经验依据。

5. 实验

我们通过图像、符号序列和生理信号相关的分类与密度估计任务，对模型进行实证测试。除非另有说明，我们对 M 使用标准的 Xavier均匀初始化UL-TIRESCONV 滤波器。所有实验细节均见附录C。PyTorch 代码可在 https://github.com/thjashin/multires-conv.

表 *1.* sCIFAR 数据集上的像素级序列分类性能。粗体表示表现最佳的模型，下划线表示表现第二好的模型。结果取自引用文献或 Hasani 等人(2022年)。

模型	准确率 (%)
注意力：	
Transformer（Trinh 等人，2018年）	62.2
RNN：	
长短期记忆网络 (Hochreiter 和 Schmidhuber，1997)	63.01
r-LSTM (Trinh 等人，2018年)	72.2
UR-GRU(Gu 等人，2020b)	74.4
HiPPO-RNN(Gu 等人，2020a)	61.1
LipschitzRNN (Erichson 等人，2021年)	64.2
状态空间模型：	
S4（Gu 等人，2021年）	91.80
S4D（Gu 等人，2022年）	90.69
S5（Smith 等人，2022年）	90.10
<escape>Liquid-S4（Hasani 等人，2022年） </escape>	<u>92.02</u>
卷积：	
TrellisNet（Bai 等人，2019年）	73.42
CKConv (Romero 等人，2022年b)	63.74
FlexConv (Romero 等人，2022年a)	80.82
MULTIRESNET（我们的）	**93.15**

5.1. 像素级序列图像分类

我们首先考虑图像分类任务，其中图像被视为一维像素序列。模型不得利用图像中的任何二维偏置。因此，模型必须能够捕捉多个不同时间尺度上的模式，包括在原始图像中彼此相近、但在其序列表示中彼此相距较远的像素。

我们在 Sequential CIFAR-10 数据集上评估模型，该数据集长期以来一直被用作评估循环神经网络（RNN）长程依赖建模能力的标准基准测试。我们采用 CIFAR-10 数据集的标准训练集和测试集划分，并从训练集中留出 10% 作为验证集。对于分类任务，我们对输出序列（$\mathbf{y}_{[\text{batchsize},256,N]} \rightarrow \mathbf{y}_{[\text{batchsize},256]}$）在所有时间步上进行均值池化，并将结果输入全连接层以生成类别 logits。我们报告验证准确率最高的模型在测试集上的测试准确率。

结果见表 1。如我们所见，我们的模型在这一基准序列分类任务上取得了最先进的性能（测试准确率最高），优于许多近期实力强劲的竞争模型，包括 Transformer（Vaswani 等人，2017年）、循环神经网络、状态空间模型以及其他卷积模型。尤其是，MULTIRESNET 的表现比此前纯卷积模型高出 10 多个百分点。

令人惊讶的是，我们的模型仅凭一种几乎简单得令人难以置信的架构，就达到了这一全新的性能基准；该架构主要由扩张因果卷积构成。

表*2*。 在 Tay 等人 (2021年) 的长序列ListOps 数据集上预测列表操作结果的性能。粗体表示性能最佳的模型，下划线表示性能第二好的模型。*SGConv 的设计旨在模拟 S4 的全局卷积解释。结果取自引用文献或 Hasani 等人 (2022年)。

模型	准确率（%）
注意力：	
局部注意力（Tay 等人，2021年）	15.82
线性变换（Katharopoulos 等人，2020年）	16.13
Linformer（Wang 等人，2020年）	16.13
Sparse Transformer（Child 等人，2019年）	17.07
Performer (Choromanski 等，2020年)	18.01
Transformer（Vaswani 等人，2017年）	36.37
Sinkhorn Transformer（Tay 等人，2020年）	33.67
FNet（Lee-Thorp 等人，2022年）	35.33
Longformer（Beltagy 等人，2020年）	35.63
BigBird（Zaheer 等人，2020年）	36.05
Nystr¨omformer（Xiong 等人，2021年）	37.15
Luna-256（Ma 等人，2021年）	37.25
Reformer（Kitaev 等人，2020年）	37.27
H-Transformer-1D（Zhu 和 Soricut，2021年）	49.53
状态空间模型：	
S4（Gu 等人，2022年）	59.60
DSS（Gupta 等人，2022年）	57.6
S4D（Gu 等人，2022年）	60.52
S5（Smith 等人，2022年）	<u>62.15</u>
Liquid-S4（Hasani 等人，2022年）	**62.75**
卷积：	
CDIL*（Cheng 等人，2023年）	44.05
SGConv*（Li 等人，2022年）	61.45
MULTIRESNET（我们的方法）	**62.75**

树层级之间共享长度为 2 的滤波器，并通过线性连接将各层相连。具体而言，我们表现最佳的模型使用了 10 个 MULTIRESBLOCK，每个模块都包含一个 MULTIRESCONV层，相当于 10 层权重共享的膨胀因果卷积。总计相当于 100 层膨胀因果卷积，但参数总数仅为 1.4M。相比之下，表现最佳的 S4 模型使用约 7.9M 个参数（Gu 等人，2021年），但性能仍不如我们参数量小 5 倍的模型（并且同样依赖复杂的初始化方法）。

5.2. 符号序列上的层级推理

为了测试模型对层级结构进行推理的能力，我们在 Tay 等人（2021年）提出的长序列ListOps 数据集上进行实验。该数据集由表示多个列表操作组合的序列构成，包括MAX、MEAN、MED（中位数）和SM（求和与取模）。例如，输入可以是

[最大值 1 [最大值 2 3] 5 6 [最小值 7 8]].

本例中的输出为 7。这远短于数据集中最长可达 2048 的示例。我们将预测表述为一个十分类问题。我们使用指定的训练集、验证集和测试集划分

表 3. PTB-XL 数据集上心电图多标签/多类别分类的 AUROC。粗体表示表现最佳的模型，下划线表示表现第二好的模型。其他模型的结果取自 Zhang 等人 (2023年) 和 Strodthoff 等人 (2021年)。

模型（AUROC）	All	Diag	子诊断	超诊断	Form	节律
MULTIRESNET（我们的模型）	**0.938**	0.939	**0.934**	**0.934**	0.897	0.975
Spacetime（Zhang 等人，2023年）	0.936	**0.941**	0.933	0.929	0.883	0.967
S4（Gu 等人，2021年）	**0.938**	0.939	0.929	0.931	0.895	**0.977**
InceptionTime（Ismail Fawaz 等人，2020年）	0.925	0.931	0.930	0.921	**0.899**	0.953
长短期记忆网络（Hochreiter 和 Schmidhuber，1997）	0.907	0.927	0.928	0.927	0.851	0.953
Transformer（Vaswani 等人，2017年）	0.857	0.876	0.882	0.887	0.771	0.831
小波特征（Strodthoff 等人，2021年）	0.849	0.855	0.859	0.874	0.757	0.890

该划分由 Tay 等人（2021年）指定。按照以往工作的做法，我们将序列填充至最大长度（2048）以组成小批次，并在对输出序列进行池化时仅对实际输入取平均值。此实验使用的模型包含 12 个多分辨率块。我们发现，较大的滤波器尺寸（4）对这项任务更有利 [3]。结果如表 2 所示。MULTIRESNET 取得了整体最佳结果，优于所有基于注意力的模型、S4 及其对角变体。只有 Liquid-S4 的表现与之相当，该模型还在状态空间模型中引入了输入之间的内积结构。

MULTIRESNET 是首个使用小卷积核的卷积模型，并在长序列ListOps数据集上实现了可与状态空间模型媲美的性能。尽管 SGConv（Li 等人，2022年）被宣传为卷积神经网络，但该模型实际上旨在模仿 S4 的全局卷积解释。因此，它同样受限于复杂的参数化，并依赖 FFT 来计算大卷积核的卷积。

5.3. 对生理信号进行分类

PTB-XL（Wagner 等人，2020年）是一个公开可用的心电图（ECG）时间序列数据集。该数据集包含来自 18,885 名患者的 21,837 条 12 导联心电图记录，每条记录时长为 10 秒。每条记录至少标注了 SCP-ECG 标准中的 71 种心电图标签之一。数据集划分为六个子集：“全部”、“诊断”、“诊断子类”、“诊断超类”、“形态”和“节律”，每个子集包含完整 71 种标签中的不同子集。“诊断超类”是一项多类别分类任务，而其他子集均为多标签分类任务。

我们使用该数据集的 100Hz 版本，其中每个时间序列包含 1K 个时间步和 12 个通道。我们使用 Strodthoff 等人 （2021年）指定的训练集、验证集和测试集划分。我们将 Sequential CIFAR-10 实验中的架构和学习率迁移了过来，并对 Dropout 在“诊断超类”的验证集上的比率，并在此后对所有子集使用相同的设置。

我们在表3中展示了结果。在 PTB-XL 的全部六个子集中，我们的模型在所有展示的模型中均取得了最佳或次佳 AUROC。与基于固定小波特征训练的神经网络相比（Strodthoff 等人，2021年），我们的模型表现显著更优，这表明将多分辨率计算与端到端学习相结合具有明显优势。

[3]这对应于从 Haar 离散小波变换的树结构切换到更一般的小波（例如 Daubechies 小波）的树结构（Akansu 等人，2001 年）。

表 4。 基于 CIFAR-10 数据集的自回归生成建模。结果以每维比特数（每维比特数）表示（越低越好）。粗体表示性能最佳的模型，下划线表示性能第二佳的模型。结果取自引用文献。

模型	#params	测试每维比特数。
循环神经网络 + *2D* 偏置：		
PixelRNN（Oord 等人，2016c）		3.00
二维卷积：		
PixelCNN (Oord 等人, 2016c)		3.14
门控 PixelCNN (Oord 等人, 2016b)		3.03
PixelCNN++(Salimans 等人,2017年)	53M	2.92
二维卷积 + 注意力：		
PixelSNAIL (Chen 等人, 2018年)	46M	2.85
一维注意力机制：		
Image Trans. (Trinh 等人, 2018年)		2.90
二维注意力：		
稀疏 Transformer（Child 等人，2019年）	59M	**2.80**
状态空间模型 *+U-Net*结构*:*		
S4（Gu 等人，2021年）		2.85
卷积（无二维偏置）：		
MULTIRESNET（我们的方法）	38M	2.84

5.4. 自回归生成建模

在这项实验中，我们不再局限于分类任务，而是评估序列建模层在 CIFAR-10 自回归生成建模中的表现。设置的详细说明见附录C.4。我们使用了 48 个 MULTIRES-LAYER，每个具有 512 个通道，滤波器尺寸为 4。由于该任务的性质需要比前几项实验更多的参数，我们在残差块的 GLU 激活之后添加了一个额外的 1×1 卷积层。

表5。 在长序列ListOps数据集上进行的消融研究结果。我们使用了比第 5.2 节中更小的模型，并报告了三个随机种子的平均性能。

ID	滤波器	初始化	记忆机制	滤波器尺寸	MultiresConv 深度	准确率
1	固定	小波	衰减	2	7	50.15±0.60
2	可训练	小波	衰减	2	7	52.08±0.35
3	可训练	Xavier	衰减	2	7	51.70±0.16
4	可训练	Xavier	衰减	2	11	61.07±0.26
5	可训练	Xavier	均匀	2	7	51.58±0.57
6	可训练	小波	衰减	4	7	59.23±0.13

我们从训练集中留出 2K 个样本用于验证，并报告验证表现最佳时的测试集负对数似然（以每维比特数计）。我们将结果与先前报告的多项结果进行基准比较，见表 4。尽管没有二维偏置且参数显著更少，我们的模型在测试集上仍达到 2.838 每维比特数，大幅胜过基于二维卷积的模型，甚至超过了同时利用卷积和注意力的改进变体（PixelSNAIL）。我们的模型也优于 S4；该模型未报告架构细节，但提到使用了下采样和 U-Net 结构。我们的模型不使用下采样或 U-Net 结构，展现了其内存效率以及捕捉长程依赖的能力。它的表现也胜过采用一维注意力机制的 Image Transformer，只有 Sparse Transformer 优于它；后者还利用了二维偏置，并使用了显著更多的参数。

5.5. 消融研究

我们沿用第 5.2 节中的长序列ListOps数据集，并开展六项实验，以研究模型架构、超参数和初始化方案的影响。每项实验均使用不同的随机种子独立运行三次。结果见表 5。

固定小波滤波器与可训练滤波器。 通过比较实验 1 和实验 2，我们可以看到，训练小波滤波器能够提升性能，这印证了我们将 MULTIRESCONV操作与小波变换解耦的设计选择。

对初始化的敏感性。 接下来，我们考察将滤波器初始化为小波滤波器的效果。从实验 2 到实验 3，我们将初始化方式从小波初始化改为神经网络的标准Xavier初始化（Glorot 和 Bengio，2010年），模型的其他部分均保持不变。我们没有观察到这些方法之间存在统计学上的显著差异。这表明，与需要谨慎初始化的 S4相关方法相比，我们的模型具有优势。

记忆机制。 我们没有发现均匀衰减与分辨率渐隐之间存在统计学上的显著差异，不过分辨率渐隐的实现更简单。

感受野的重要性。 最后，我们展示了，通过增大滤波器尺寸（实验 2 与实验 6 对比）或增加 MULTIRESCONV 的深度（实验 3 与实验 4 对比），可以显著提升该模型的性能。我们认为，这是因为这两种做法都会增大 MULTIRESCONV操作的感受野尺寸，而这对于 ListOps 这类推理任务尤为重要。

6. 结论

我们提出了 MULTIRESLAYER，用于在序列数据源中稳健、高效地记忆长期模式。它借鉴多分辨率分析（MRA）文献中的思想，以小波分解为基础，记忆出现在多个时间尺度上的模式。具体而言，我们的记忆由多分辨率卷积生成，并通过膨胀因果卷积实现；这些卷积使用学习得到的 滤波器，滤波器在树层级之间共享，而各层级通过纯线性运算相连。为生成记忆，可以保留所有多分辨率值，也可以利用小波变换的时间局部化特性，更加重视较近的时间点。

由此得到的 MULTIRESNET 兼具卷积网络的计算优势，其参数数量也远少于竞争模型，同时在多个基准序列建模任务中实现了最先进的性能。这些实验表明，我们的多分辨率记忆结构可迁移到多种任务中，即使某些任务直观上并不适合以多尺度方式处理（例如 ListOps 中的语法树解析）。

借鉴小波文献中的思路，我们构建了一种高效的卷积层，在大幅减少参数数量的同时，并未牺牲模型性能。MULTIRESCONV 的理论基础确保了其具备一种具有强大重建能力的配置（例如，当我们的滤波器等于小波滤波器时）；然而，正如我们所展示的，通过学习滤波器，可以进一步提升预测性能。

从小波框架出发的另一个潜在优势，是能够借助该领域丰富的文献成果，推动未来的建模进展。

尤其是，我们计划探索 多分辨率卷积 在表征学习和长期预测中的应用价值。在表征学习方面，我们可以考虑平移不变小波变换 (Kingsbury，1998；Selesnick 等人，2005)，以构建对输入信号平移具有不变性的表征。例如，即使关键特征彼此发生了位移，我们仍可能希望将心电图信号相似的个体聚类。小波还可以拓展至图像分析，从而在我们的序列建模场景中实现视频分析。

致谢

本研究部分得到了美国空军科学研究办公室拨款 FA9550-21-1-0397、美国海军研究办公室拨款 N00014-22-1-2110、美国国家科学基金会拨款 2205084 以及斯坦福以人为本人工智能研究所（HAI）的资助。EBF 是陈·扎克伯格生物中心旧金山的研究员。KAW 部分获得了斯坦福数据科学项目的支持，担任斯坦福数据科学学者。

参考文献

Akansu, A. N., Haddad,R.A.,and Haddad, P. A. 多分辨率信号分解：变换、子带与小波。Academic press，2001 年。

Ba, J. L., Kiros, J. R., andHinton, G. E. 层归一化。*arXiv* 预印本 *arXiv:1607.06450*，2016 年。

Bai, S., Kolter, J. Z. 和 Koltun, V. 用于序列建模的格状网络。发表于 国际学习表征会议，2019年。

Beltagy, I.、Peters, M. E. 和 Cohan, A. Longformer：长文档 Transformer。*arXiv*预印本 *arXiv:2004.05150*，2020年。

Chen, X.、Mishra, N.、Rohaninejad, M. 和 Abbeel, P. Pix-elSNAIL：一种改进的自回归生成模型。载于国际机器学习会议，第 864–872 页。PMLR，2018年。Cheng, L.、Khalitov, R.、Yu, T.、Zhang, J. 和 Yang, Z. 使用循环膨胀卷积神经网络对长序列数据进行分类。《神经计算》，518：50–59，2023年。

Child, R., Gray, S., Radford, A., 和 Sutskever, I. 使用稀疏 Transformer 生成长序列。*arXiv* 预印本 *arXiv:1904.10509*，2019年。

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mo- hiuddin, A., Kaiser, L. 等。重新思考注意力机制与 Performer。发表于国际学习表征会议，2020年。Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. 从零开始的自然语言处理（几乎）。机器学习研究期刊，12(ARTICLE):2493–2537，2011年。

Daubechies, I. 紧支撑小波的正交归一基。纯粹与应用数学通讯，41(7):909–996，1988年。

Dauphin, Y. N., Fan, A., Auli, M., 和 Grangier, D. 使用门控卷积网络进行语言建模。发表于国际机器学习会议，第 933–941 页。PMLR，2017年。

Erichson, N. B., Azencot, O., Queiruga, A., Hodgkinson, L., 和 Mahoney, M. W. Lipschitz循环神经网络。发表于国际学习表征会议，2021年。

Glorot, X. 和 Bengio, Y. 理解训练深度前馈神经网络的难点。载于国际人工智能与统计学会议，第 249–256 页，2010年。

Gu, A.、Dao, T.、Ermon, S.、Rudra, A. 和 R´e, C. HiPPO：具有最优多项式投影的循环记忆。神经信息处理系统进展，33：1474–1487，2020a。

Gu, A., Gulcehre, C., Paine, T., Hoffman, M. 和 Pascanu, R. 改进循环神经网络的门控机制。载于国际机器学习会议，第 3800–3809 页。PMLR，2020年b。

Gu, A., Goel, K. 和 Re, C. 利用结构化状态空间高效建模长序列。发表于国际学习表征会议，2021年。

Gu, A.、Goel, K.、Gupta, A. 和 R´e, C. 对角状态空间模型的参数化与初始化。载于神经信息处理系统进展，2022年。

Gupta, A.、Gu, A. 和 Berant, J. 对角状态空间与结构化状态空间同样有效。载于 Oh, A. H.、Agarwal, A.、Belgrave, D. 和 Cho, K.（编），神经信息处理系统进展，2022年。

Hasani, R.、Lechner, M.、Wang, T.-H.、Chahine, M.、Amini, A. 和 Rus, D. 液态结构化状态空间模型。*arXiv* 预印本 *arXiv:2209.12951*，2022年。

He, K.、Zhang, X.、Ren, S. 和 Sun, J. 用于图像识别的深度残差学习。在*IEEE*计算机视觉与模式识别会议论文集中，第 770–778 页，2016 年。

Hendrycks, D. 和 Gimpel, K. 高斯误差线性单元（GELUs）。*arXiv* 预印本 *arXiv:1606.08415*，2016 年。

Hochreiter, S. 和 Schmidhuber, J. 长短期记忆。神经计算，9(8):1735–1780，1997年。

Ioffe, S. 和 Szegedy, C. 批归一化：通过减少内部协变量偏移加速深度网络训练。载于国际机器学习会议，第 448–456 页。PMLR，2015年。

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A. 和 Petitjean, F. InceptionTime：为时间序列分类寻找 AlexNet。数据挖掘与知识发现，34(6):1936–1962，2020年11月。ISSN 1573-756X。doi: 10.1007/ s10618-020-00710-y。

Jawerth, B. 和 Sweldens, W. 小波多分辨率分析概述。*SIAM*评论，36(3):377–412，1994年。

Kaiser, L., Gomez, A. N. 和 Chollet, F. 用于神经机器翻译的深度可分离卷积。载于国际学习表征会议，2018年。

Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. v. d., Graves, A. 和 Kavukcuoglu, K. 线性时间神经机器翻译。*arXiv* 预印本 *arXiv:1610.10099*，2016 年。

Katharopoulos, A.、Vyas, A.、Pappas, N. 和 Fleuret, F. Transformers 即循环神经网络：具有线性注意力的快速自回归 Transformer。载于国际机器学习会议，第 5156–5165 页。PMLR，2020年。

Kingsbury, N. G. 双树复小波变换：一种实现平移不变性和方向滤波器的新技术。载于*IEEE*数字信号处理研讨会，第86卷，第120–131页。Citeseer，1998年。

Kitaev, N., Kaiser, L., 和 Levskaya, A. Reformer：高效 Transformer。载于 国际学习表征会议，2020年。网址：https://openreview.net/forum?id=rkgNKkHtvB。

Lee, G. R., Gommers, R., Waselewski, F., Wohlfahrt, K., 和 O'Leary, A. PyWavelets：一个用于小波分析的 Python 软件包。《开放源代码软件杂志》，4 (36):1237，2019年4月。ISSN 2475-9066。doi：10.21105/ joss.01237。

Lee-Thorp, J.、Ainslie, J.、Eckstein, I. 和 Ontanon, S. FNet：使用傅里叶变换混合词元。载于北美计算语言学协会分会*2022*年会议论文集：人类语言技术，第4296–4313页，2022年。

Li, Y., Cai, T., Zhang, Y., Chen, D. 和 Dey, D. 是什么使卷积模型在长序列建模方面表现出色？*arXiv* 预印本 *arXiv:2210.09298*，2022年。Loshchilov, I. 和 Hutter, F. SGDR：采用热重启的随机梯度下降。载于国际学习表征会议，2017年。Loshchilov, I. 和 Hutter, F. 解耦权重衰减正则化。载于国际学习表征会议，2018年。

Ma, X., Kong, X., Wang, S., Zhou, C., May, J., Ma, H., 和 Zettlemoyer, L. Luna：线性统一嵌套注意力。神经信息处理系统进展，34: 2441–2453，2021年。

Oord, A. v. d.、Dieleman, S.、Zen, H.、Simonyan, K.、Vinyals, O.、Graves, A.、Kalchbrenner, N.、Senior, A. 和 Kavukcuoglu, K. WaveNet：一种用于原始音频的生成模型。*arXiv* 预印本 *arXiv:1609.03499*，2016a。

Oord, A. v. d.、Kalchbrenner, N.、Espeholt, L.、Vinyals, O.、Graves, A. 等。使用 PixelCNN解码器进行条件图像生成。神经信息处理系统进展，29，2016b。

Oord, A. v. d.、Kalchbrenner, N. 和 Kavukcuoglu, K. 像素循环神经网络。载于国际机器学习会议，第 1747–1756 页。PMLR，2016c。

Romero, D. W., Bruintjes, R., Bekkers, E. J., Tomczak, J. M., Hoogendoorn, M. 和 van Gemert, J. FlexConv：带有可微核大小的连续核卷积。载于国际学习表征会议，2022年a。

Romero, D. W., Kuzina, A., Bekkers, E. J., Tomczak, J. M. 和 Hoogendoorn, M. CKConv：用于序列数据的连续核卷积。载于 国际学习表征会议，2022年b。

Salimans, T., Karpathy, A., Chen, X., 和 Kingma, D. P. PixelCNN++：使用离散化逻辑斯蒂混合似然及其他改进来提升 PixelCNN。在国际学习表征会议，2017年。

Selesnick, I. W., Baraniuk, R. G., and Kingsbury, N. C. 双树复小波变换。*IEEE* 信号处理杂志,22(6):123–151,2005。

Shen, D., Min, M. R., Li, Y., 和 Carin, L. 学习用于文本处理的上下文敏感卷积滤波器。载于*2018*年自然语言处理经验方法会议论文集，第1839–1848页，2018年。

Smith, J. T., Warrington, A., 和 Linderman, S. W. 用于序列建模的简化状态空间层。*arXiv* 预印本 *arXiv:2208.04933*，2022年。

Strang, G. 和 Nguyen, T. 小波和滤波器组。SIAM，1996年。

Strodthoff, N.、Wagner, P.、Schaeffter, T. 和 Samek, W. 用于心电图分析的深度学习：来自 PTB-XL 的基准测试与见解。*IEEE* 生物医学与健康信息学期刊，25(5):1519–1528，2021年5月。ISSN 2168-2208。doi: 10.1109/JBHI.2020.3022989。

Tay, Y.、Bahri, D.、Yang, L.、Metzler, D. 和 Juan, D.-C. 稀疏 Sinkhorn 注意力。载于国际机器学习会议，第 9438–9447 页。PMLR，2020年。

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. 长程竞技场：高效 Transformer 的基准测试。载于国际学习表征会议，2021年。

Trinh, T., Dai, A., Luong, T., and Le, Q. 使用辅助损失学习循环神经网络中的长期依赖关系。发表于国际机器学习会议，第 4965–4974 页。PMLR，2018年。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. 和 Polosukhin, I. 注意力机制就是你所需要的一切。神经信息处理系统进展，第30卷，2017年。

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W. 和 Schaeffter, T. PTB-XL，一个大型公开可用的心电图数据集。《科学数据》，7(1):154，2020年5月。ISSN 2052-4463。doi: 10.1038/s41597-020-0495-6。

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. 和 Lang, K. J. 使用时延神经网络进行音素识别。*IEEE* 声学、语音与信号处理汇刊，37(3)：328–339，1989年。

Wang, S., Li, B. Z., Khabsa, M., Fang, H. 和 Ma, H. Linformer：具有线性复杂度的自注意力。*arXiv* 预印本 *arXiv:2006.04768*，2020年。

Willsky, A. S. 用于信号和图像处理的多分辨率马尔可夫模型。*IEEE* 论文集，90(8)：1396–1458，2002年。

Wu, F., Fan, A., Baevski, A., Dauphin, Y., and Auli, M. 使用轻量级动态卷积减少注意力开销。载于国际学习表征会议，2018年。

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., and Singh, V. Nystr¨omformer：一种基于 Nystr¨om 的自注意力近似算法。载于*AAAI*人工智能会议论文集，第35卷，第14138–14148页，2021年。

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L. 等。Big Bird：用于更长序列的 Transformer。神经信息处理系统进展，33：17283–17297，2020年。

Zhang, M., Saab, K. K., Poli, M., Dao, T., Goel, K., and Re, C. 使用简单的离散状态空间有效地建模时间序列。在第十一届国际学习表征会议，2023年。网址 https://openreview.net/forum?id=2EpjkjzdCAa。

Zhu, Z. 和 Soricut, R. H-transformer-1d：用于序列的快速一维层次化注意力。在计算语言学协会第*59*届年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第3801–3815页，2021年。

A. 背景：小波理论

本节将正式介绍小波理论的背景知识，并在第2节引言的基础上进一步展开。首先，我们将通过多分辨率分析（MRA）框架推导哈尔小波。随后，我们将说明如何利用同一框架推导出多种正交小波。

A.1. 哈尔小波

我们首先介绍哈尔尺度函数的定义。尺度函数也称为母小波。

定义 A.1（哈尔尺度函数）。 哈尔尺度函数定义为

$$\phi(t) = \mathbf{1}(0 \leq t < 1).$$

令 V_0 为由尺度函数的平移张成的空间：

$$V_0 \triangleq \left\{ s_0 \middle| s_0(t) = \sum_{k\in\mathbb{Z}} a_{0,k}\phi(t-k), a_{0,k} \in \mathbb{R} \right\}.$$

它包含在长度为 1 的区间上保持常数的函数。接下来，我们通过将单位区间划分为两半，构造一个包含 V_0 的更大空间 V_1：

$$V_1 \triangleq \left\{ s_1 \middle| s_1(t) = \sum_{k\in\mathbb{Z}} a_{1,k}\sqrt{2}\phi(2t-k), a_{1,k} \in \mathbb{R} \right\}.$$

引入系数 $\sqrt{2}$，使得基函数 $\sqrt{2}\phi(2t-k)$ 经过归一化，即 $\int_{\mathbb{R}}(\sqrt{2}\phi(2t-k))^2\,dt = 1$。这些函数在长度为 $1/2$ 的区间上保持常数。我们可以重复这一过程，为 $j \geq 0$ 构造一系列函数空间：

$$V_j \triangleq \left\{ s_j \middle| s_1(t) = \sum_{k\in\mathbb{Z}} a_{j,k}\phi_{j,k}(t), a_{j,k} \in \mathbb{R} \right\}, \text{ where } \phi_{j,k}(t) \triangleq 2^{j/2}\phi(2^j t - k). \tag{6}$$

这些函数空间具有日益丰富的结构和表达能力：

$$V_0 \subset V_1 \subset V_2 \subset \cdots \subset V_j.$$

很容易看出，当 j 足够大时，V_j 将具有以任意精度逼近任意连续时间序列 f 的能力。令 $\hat{f}^{(j)}(t) = \sum_{k\in\mathbb{Z}} a_{j,k}\phi_{j,k}(t)$ 表示 f 在 V_j 中的近似值。最佳近似由以下系数给出。

$$a_{j,k} = \int_{\mathbb{R}} f(t)\phi_{j,k}(t)\,dt = \int_{k/2^j}^{(k+1)/2^j} f(t)\,dt. \tag{7}$$

可以用系数向量 f 来表示 $\{a_{j,k}\}_{k\in\mathbb{Z}}$。然而，正如我们在第2节所述，每个单独的系数都过于局部，无法捕捉较大区间内的模式。解决这一问题的一种方法是应用多分辨率分析，生成能够概括 f 在多个不同时间尺度上的模式的表示，从而得到哈尔小波。

小波多分辨率分析包含以下步骤。首先，注意到 $V_{j-1} \subset V_j$，我们将 V_j 分解为 V_{j-1} 与其正交补空间 W_{j-1}之和。重复这一过程，我们得到

$$V_j = V_{j-1} \oplus W_{j-1} = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_{j-1}. \tag{8}$$

由于 V_j 和 V_{j-1} 中的函数分别表示比 $1/2^j$ 和 $1/2^{j-1}$更粗的时间尺度上的结构，因此我们可以预期，W_{j-1} 中的基函数能够唯一表示 $1/2^j$ 时间尺度上的结构。下一步是为 W_{j-1}寻找一组正交归一基。定理 A.3表明，这组基可以由函数ψ经过缩放和平移得到，这些函数统称为哈尔小波。

定义 A.2（哈尔小波）。哈尔小波是一族定义如下的函数：

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \text{ where } \psi(t) \triangleq \phi(2t) - \phi(2t-1). \tag{9}$$

定理 A.3 (哈尔小波作为正交归一基)。设 V_j 定义如 (6) 所示，且 W_{j-1} 为 V_{j-1} 在 V_j中的正交补。那么，$\{\psi_{j,k}\}_{k\in\mathbb{Z}}$ 是 W_j的一个正交归一基。

证明。我们首先注意到 $\psi_{j,\ k} \in V_{j+1}$ 以及 $\langle \psi_{j,\ k},\ \psi_{j',\ k'}\rangle = \delta_{j,\ j'}\delta_{k,\ k'}$。为了证明这一结果，我们需要证明 (1) $\psi_{j,\ k} \in W_j$，即 $\psi_{j,\ k}$ 与 V_j 中的任意函数都正交；以及 (2) W_j 中的任意函数都可以用 $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$ 表示。对于第一个要求，我们有，对于任意 $k' \in \mathbb{Z}$,

$$\langle \psi_{j,k}, \phi_{j,k'}\rangle = \langle \frac{1}{\sqrt{2}}(\phi_{j+1,2k} - \phi_{j+1,2k+1}), \frac{1}{\sqrt{2}}(\phi_{j+1,2k'} + \phi_{j+1,2k'+1})\rangle = \frac{1}{2}(\delta_{k,k'} - \delta_{k,k'}) = 0.$$

因此，$\psi_{j,\ k}$ 与 V_j 中的任意函数正交，这意味着 $\psi_{j,\ k} \in W_j$。为了证明 (2)，我们注意到，对于某个 $s_{j+1} = \sum_{k\in\mathbb{Z}} a_{j+1,\ k}\phi_{j+1,\ k} \in V_{j+1}$，$s_{j+1} \in W_j$ 意味着对于任意 $k' \in \mathbb{Z}$,

$$\begin{aligned}
\langle s_{j+1}, \phi_{j,k'}\rangle &= \sum_{k\in\mathbb{Z}} a_{j+1,k}\langle \phi_{j+1,k}, \phi_{j,k'}\rangle \\
&= \sum_{k\in\mathbb{Z}} a_{j+1,k}\langle \phi_{j+1,k}, \frac{1}{\sqrt{2}}(\phi_{j+1,2k'} + \phi_{j+1,2k'+1})\rangle \\
&= \frac{1}{\sqrt{2}}(a_{j+1,2k'} + a_{j+1,2k'+1}) = 0.
\end{aligned}$$

因此，我们可以将任意 $s_{j+1} \in W_j$ 表示为$a_{j+1,\ 2k+1} = -a_{j+1,\ 2k}$ 的形式

$$s_{j+1} = \sum_{k\in\mathbb{Z}} a_{j+1,2k}(\phi_{j+1,2k} - \phi_{j+1,2k+1}) = \sum_{k\in\mathbb{Z}} \sqrt{2}a_{j+1,2k}\psi_{j,k}.$$

因此，W_j 中的任意函数都可以用 $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$表示。结合（1）和（2）即可证明原命题。 □

现在我们已经得到了小波基，可以重写 f 在 V_j 中的近似，以体现（8）中的分解：

$$f(t) \approx \hat{f}^{(j)}(t) = \sum_{k\in\mathbb{Z}} a_{0,k}\phi(t-k) + \sum_{j'=0}^{j-1}\sum_{k'\in\mathbb{Z}} b_{j',k'}\psi_{j',k'}(t), \tag{10}$$

其中，$\{\phi(t-k)\}_{k\in\mathbb{Z}}$是 V_0的基，$\{\psi_{j',k'}\}_{k'\in\mathbb{Z}}$是 $W_{j'}$的基，对于 $j' = 0$、1、$\cdots$、$j-1$均如此。因此，我们可以使用系数 $\{a_{0,k}\}_{k\in\mathbb{Z}}$和 $\{b_{0,k}\}_{k\in\mathbb{Z}} \cup \cdots \cup \{b_{j-1,k}\}_{k\in\mathbb{Z}}$概括 f的多分辨率结构。

A.2. 推广到所有正交小波

上一节所述的多分辨率分析框架可以从哈尔小波推广到所有正交小波。出发点是，我们可以使用不同于哈尔尺度函数的尺度函数，并满足以下方程：

$$\phi(t) = \sum_{k=0}^{K-1} \mathbf{h}_0(k)\sqrt{2}\phi(2t-k). \tag{11}$$

此条件保证相关函数空间满足 $V_{j-1} \subset V_j$。序列 $\mathbf{h}_0(\cdot) \in \mathbb{R}^K$ 在此称为（低通）滤波器，不同小波的滤波器长度可能各不相同。例如，Haar 小波滤波器是一个二维向量，其中 $\mathbf{h}_0(0) = \mathbf{h}_0(1) = \frac{1}{\sqrt{2}}$。Daubechies-4 小波（Daubechies，1988）滤波器有 4 个元素：

$$\left(\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}\right).$$

类似地，母小波具有以下广义形式：

$$\psi(t) = \sum_{k=0}^{K-1} \mathbf{h}_1(k)\sqrt{2}\phi(2t-k), \tag{12}$$

其中 $\mathbf{h}_1(\cdot)$称为 高通 滤波器。利用 $\phi(t-k)$与 $\psi(t)$之间的正交性，对于任意 k，可以证明（例如参见 Strang 与 Nguyen，1996年）：

$$\mathbf{h}_1(k) = (-1)^k \mathbf{h}_0(K-1-k),$$

其中 N是滤波器 $\mathbf{h}_0$的长度。对于哈尔小波，$\mathbf{h}_1(0) = \mathbf{h}_0(1) = \frac{1}{\sqrt{2}}$，$\mathbf{h}_1(1) = -\mathbf{h}_0(0) = -\frac{1}{\sqrt{2}}$。对于 Daubechies-4 小波，对应的$\mathbf{h}_1$ 滤波器为

$$\left(\frac{1-\sqrt{3}}{4\sqrt{2}}, -\frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, -\frac{1+\sqrt{3}}{4\sqrt{2}}\right).$$

在这项工作中，除消融实验外，我们不固定滤波器（$\mathbf{h}_0$,$\mathbf{h}_1$）的取值，也不强制它们之间存在耦合。相反，我们允许通过梯度下降学习这两个滤波器。我们还发现，使用哈尔小波或 Daubechies 小波初始化滤波器，并不比标准Xavier初始化更有优势。这表明模型能够通过优化发现有效的参数化方式。

A.3. 离散小波变换

我们已经说明，连续时间序列可以用一个系数向量表示，该向量能够捕捉其在多个不同时间尺度上的结构。剩下的挑战是从离散采样信号中确定这些系数$\mathbf{x} \in \mathbb{R}^N$。根据（7），我们知道每个系数 $a_{j,k}$ 概括了时间序列在一个 $1/2^j$ 区间内的平均值。因此，如果我们选择足够大的 J，系数 $\mathbf{a}_J$ 就可以用离散采样的$\mathbf{x}$直接近似得到。我们可以通过一种称为离散小波变换（DWT）的过程，从 10递归地获得小波展开（$\mathbf{a}_J$ 所需的系数。

为了推导离散小波变换的递归公式，我们从 $s_{j+1} = \sum_m a_{j+1,m}\phi_{j+1,m} \in V_{j+1}$ 开始，并计算其在 V_j 和 W_j中的分解系数：

$$s_{j+1} = \sum_n a_{j,n}\phi_{j,n} + \sum_\ell b_{j,\ell}\psi_{j,\ell}.$$

由于 $\{\phi_{j,n}\}_{n\in\mathbb{Z}}$ 和 $\{\psi_{j,n}\}_{n\in\mathbb{Z}}$的正交性，

$$\begin{aligned}
a_{j,n} &= \langle s_{j+1}, \phi_{j,n}\rangle \\
&= \langle \sum_m a_{j+1,m}\phi_{j+1,m}, \sum_k \mathbf{h}_0(k)\phi_{j+1,2n+k}\rangle \\
&= \sum_k a_{j+1,2n+k}\mathbf{h}_0(k), \text{ and similarly,} \\
b_{j,\ell} &= \langle s_{j+1}, \psi_{j,\ell}\rangle = \sum_k a_{j+1,2\ell+k}\mathbf{h}_1(k).
\end{aligned}$$

DWT 会从 $j = J$ 重复上述计算，直到 0。我们可以看到，这一过程的每一步都是 $\mathbf{a}^{j+1}$ 与滤波器（$\mathbf{h}_0(-)$ 或 $\mathbf{h}_1(-)$）之间的一次卷积，然后应用下采样器（$\mathbf{y}(n) = \mathbf{x}(2n)$）。请注意，尽管卷积是在滤波器 $\mathbf{h}_0$,$\mathbf{h}_1$, 取反之后进行的，但这与神经网络中的一维卷积层完全等价；其卷积核为 $\mathbf{h}_0$ 或 $\mathbf{h}_1$, , 卷积操作为标准卷积再加上滤波器取反。

B. MultiresLayer 实现

我们提供了一个 PyTorch 实现的示例，展示了带有“分辨率渐隐”TreeSelect 策略的多分辨率卷积记忆多分辨率层，具体请参见图 3 中的第 3.2 节。3.2）3。

```
导入 数学来自
torch .nn. 函数式 导入填充，卷积1d

def multires_layer(x, h0, h1, w, depth=None): """ 参数：x：形状为 (batch_size, n_channels, sequence_length) 的输入。h0、h1：形状为 (n_channels, 1, filter_size) 的卷积滤波器。w：TreeSelect 后线性层的权重。形状：(n_ channels, depth + 2)。depth：MultiresConv 的深度，即 J。返回：y：形状为 (batch_size, n_channels, sequence_length) 的输出。""" kernel_size = self.h0.shape[-1 如果 depth 为 None：depth = math.ceil(math.log2((x.shape[-1]- 1) / (kernel_size - 1) + 1)) y = 0. a = x dilation = 1对于 i 从 depth 到 1 递减：a = pad(a, (dilation * (kernel_size - 1), 0), "constant", 0) b = conv1d(a, h1, dilation=dilation, groups= x.shape[1]) a = conv1d(a, h0, dilation=dilation, groups=x.shape[1]) y += w[:, i:i+1] *b dilation *= 2y += w[:, :1] * a y +=w[:, -1:] * x 返回 y
```

图 *3*。 使用“分辨率渐隐”TreeSelect 策略实现 MULTIRESLAYER 的 PyTorch 代码。

C. 实验细节

C.1. 序列 CIFAR-10 分类

输入是长度为 1024 的序列，包含三个对应于 RGB 值的通道。我们将输入重新缩放并居中到 $[-1, 1]$，并使用一个 1×1 卷积层将其编码为 256 个通道。对于此任务，我们使用了 10 个 MULTIRESBLOCK，每个滤波器尺寸为 2、包含 256 个通道，并采用分辨率渐消记忆机制。随后对各时间步进行均值池化，并通过一个线性层生成分类逻辑值。我们使用批量大小为 50，在 250 个训练轮次中训练该网络。我们使用 AdamW 优化器（Loshchilov& Hutter，2018年），采用默认超参数和 0.01 的权重衰减率。我们使用 0.25 的丢弃率、层归一化以及 0.0045 的初始学习率。学习率遵循余弦退火过程（Loshchilov & Hutter，2017年）。

C.2. ListOps

输入是最大长度为 2048 的词元 ID（整数）序列。我们将所有序列填充至这一最大长度，并使用嵌入层将其编码为 128 个通道。我们使用了 10 个多分辨率块，每个块的滤波器尺寸为 4、通道数为 128，并采用分辨率渐消记忆机制。我们仅对实际输入所在的时间步执行均值池化，以生成分类逻辑值。我们使用批量大小为 50，对网络训练了 100 个训练轮次。我们使用 AdamW 优化器，权重衰减率为 0.03。学习率在线性预热 1 个训练轮次后设为 0.003，随后采用余弦退火。我们使用了 0.1 的丢弃率，并在本实验中使用批归一化代替层归一化。

C.3. PTB-XL

输入是包含 1000 个时间步和 12 个通道的时间序列。我们沿用了 Sequential CIFAR-10 实验中的架构和学习率。我们使用了层归一化、0.2 的丢弃率，以及权重衰减率为 0.06 的 AdamW 优化器。网络先经过 5 个训练轮次的线性预热，再以余弦学习率训练 95 个训练轮次。

C.4. 自回归生成建模

CIFAR-10 图像被重塑为长度为 1024 的序列，每个序列包含三个通道。令 $x_i \in \mathbb{R}^3$ 表示序列中第 i 个像素的 RGB 值（经重新缩放并居中至 $[-1, 1]$）。为对 $i = 1$ 的条件分布$p(x_i|x_1, \cdots, x_{i-1})$ 建模，其中 $i = 1, 2, \cdots, 1024$，网络的输入被设为 $(0, x_1, x_2, \cdots, x_{1023})$，目标输出序列则是 $x_1, x_2, \cdots, x_{1024}$ 分布的参数。我们采用了与 Salimans 等人（2017年）相同的离散化逻辑斯蒂分布混合和线性相关结构来建模 RGB 值。我们使用了 48 个 MULTIRESLAYER，滤波器尺寸为 4、通道数为 512，并采用分辨率渐消记忆机制。此外，我们在残差块靠近输出端的位置添加了一个 1×1 卷积层。MULTIRESLAYER 的因果结构确保第 i 个输出仅依赖于前 i 个输入元素，从而保持自回归建模所需的条件依赖模式。网络训练了 250 个训练轮次，批量大小为 64。我们使用了默认超参数的 Adam 优化器。我们采用了 0.1 的丢弃率、层归一化和 0.001 的初始学习率。学习率按照余弦退火过程变化。

D. 附加结果

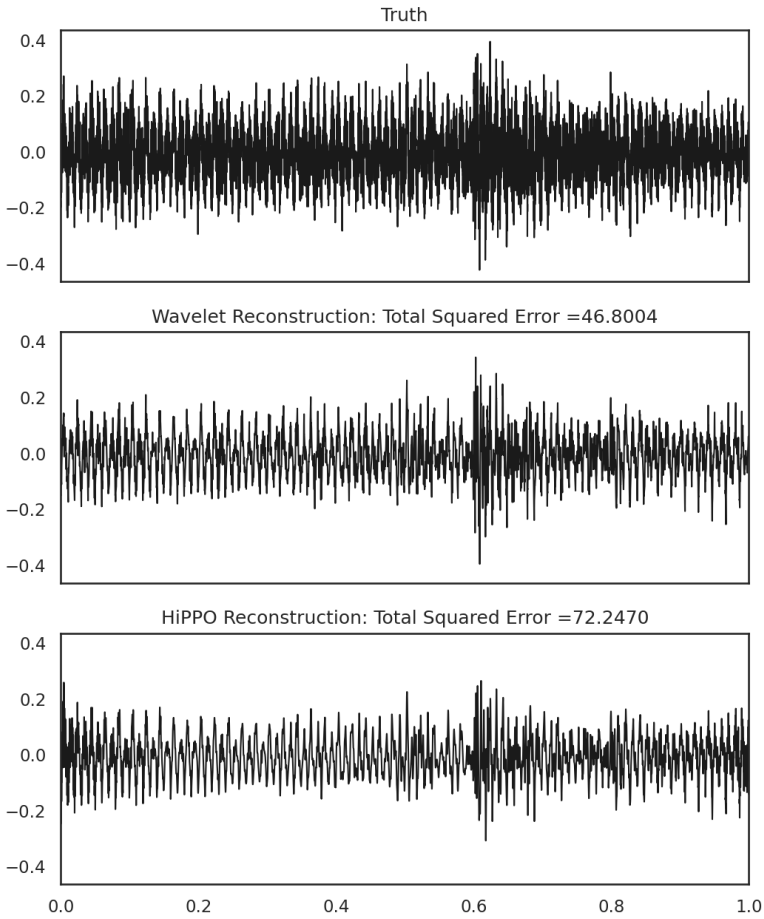

图 *4.* 基于小波和多项式基的音频信号重建。 上： 包含 16,384 个时间点的原始音频波形中的 1 秒片段。 中： 使用含 2068 个系数的 10 级 Daubechies-4 小波分解进行重建。 下： 通过投影到正交多项式进行 HiPPO （ Gu 等人，2020a ）重建，使用 2068 个系数（见第 4 节）。