arXiv:2303.04485v2 [cs.SD] 1 Jun 2023

起音与力度：使用卷积神经网络实现经济实惠的实时钢琴转录

Andres Fernandez图宾根大学¨
和 *IMPRS-IS*图宾根¨，德国
a.fernandez@uni-tuebingen.de

摘要—近年来，复调钢琴转录取得了显著进展，这得益于复杂深度学习方法的应用，以及音符起音、偏移、力度和踏板检测等新子任务的引入。这些进展伴随着所提出模型的复杂度和规模不断提升，而这些模型通常依赖非实时组件和高分辨率数据。本文聚焦于起音和力度检测，并表明一种规模小得多、结构也更简单的卷积方法，即使采用较低的时间分辨率（24 毫秒），仍具有竞争力：我们提出的 ONSETS&VELOCITIES（O&V）模型在 MAESTRO 数据集的起音检测任务上达到最先进的性能（F_1=96.78%），并为起音+力度检测设定了一个良好的全新基准（F_1=94.50%）；该模型仅有 ∼3.1M 个参数，并能在普通通用硬件上保持实时能力。我们提供开源代码以复现研究结果，并提供一个使用预训练模型的实时演示 [1]。

索引词—深度学习、复调钢琴转录

I. 引言

A. 复调钢琴转录

钢琴演奏转录复调钢琴转录 (PPT)这一任务对于音乐分析和再合成等下游任务很有用。考虑一段音频波形 $x(t) \in \mathbb{R}^T$，其时间为 T，对应于一段乐谱 $\mathcal{S}$的钢琴演奏；那么，PPT 的任务就是从 x中恢复 $\mathcal{S}$。这里，$\mathcal{S}$ 是由 N 个音符事件 $\{\mathcal{N}_n$ 组成的集合：$=(k_n,\ v_n,\ \downarrow_n,\ \uparrow_n)\}_{n=1}^{N}$，其中 $k \in \{1,\ldots,K\}$ 指定琴键（通常为 $K = 88$）。值 $v \in [0,1]$表示该事件的强度（也称为琴键力度）。琴键起音（按下）和偏移（松开）的时间戳分别由 $\downarrow$ 和 $\uparrow$, 指定，其中 $0 \leq \downarrow_n < \uparrow_n \leq T\ \ \forall n$。

人们为实现钢琴演奏转录（PPT）的自动化付出了大量努力，通常通过诸如广受欢迎的音乐信息检索评测交流（MIREX）[1]等挑战来体现，并采用了各种技术，例如手工设计特征、频谱图分解、概率模型 [2], [3]，以及近年来的深度学习（DL）[4]。PPT 通常通过在测试集上逐事件比较恢复出的乐谱 $\mathcal{S}$ 与真实标注 $\mathcal{S}$ 来进行评估。MAPS [5], [6], 、SMD [7] 和 MusicNet [8]等数据集的整理工作曾受到标注不精确、训练数据不足、转录结果不切实际和/或录音条件受限等问题的影响，这些问题增加了评估的难度，也阻碍了

作为独立研究人员，在由巴利阿里群岛的*IAM´usica* 项目支持下开展的工作，该项目由*Institut d'EstudisBale`arics*提供支持。[1]设置：
https://github.com/andres-fr/iamusica 训练演示:_
https://github.com/andres-fr/iamusica 演示 _

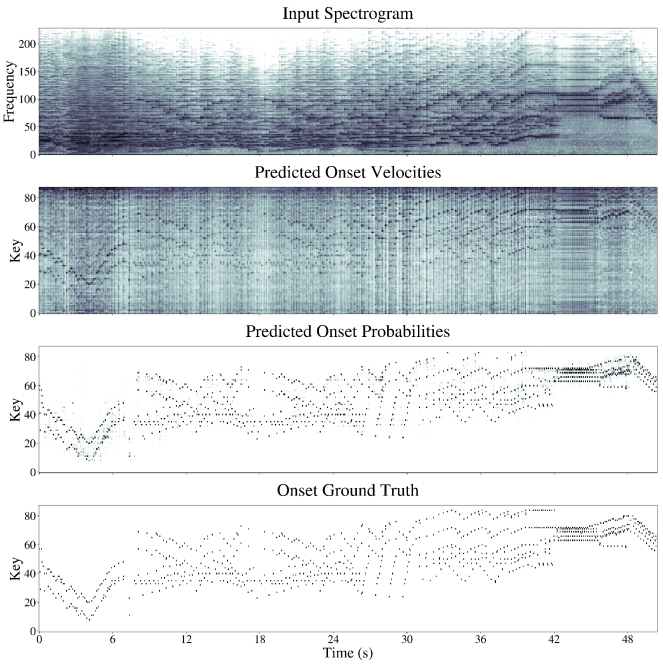


图1：MAESTRO 测试集中的一段炫技乐段的对数梅尔频谱图（X），以及对应的力度（$\hat{}\mathcal{R}_V$）和最后阶段起音（$\hat{}\mathcal{R}_{\downarrow}^{(3)}$）预测结果，还有真实标注钢琴卷帘$\mathbb{1}_{\downarrow_3}$（详情见第II节）。

建立统一的 PPT 基准 [9]。MAESTRO 数据集 [9] 的推出解决了其中许多问题：它提供了 ∼200 小时经过精确标注的高质量音频数据，涵盖种类繁多的高难度作品、钢琴家和录音条件，并纳入了评估划分。因此，它很快成为广受欢迎的基准。不过，MAESTRO 中的钢琴都相当相似。为涵盖更普遍的情境并满足日益增长的训练数据需求，[10] 整理了 GiantMIDI 数据集，从 YouTube 收集了超过 1000 小时的钢琴音乐，并使用 DL [11]对其进行标注。

B. 钢琴演奏转录领域的最新进展

一项将深度学习应用于钢琴演奏转录的颇具影响力的工作是 [12]。他们的工作确立了以下主要趋势：

a）频谱图：尽管已有研究尝试将$x(t)$波形用作深度学习的输入 [13], ，时频表征——

例如频谱图 [14, 第19] 章 [11], [15]——总体而言，在钢琴演奏转录和判别式音频任务中仍具有竞争力。

b) 钢琴卷帘监督：考虑一种名为 $\mathcal{S}$ 的替代表示法，即 钢琴卷帘 $\mathcal{R} \in [0,1]^{K\times T}$（见图1），其中条目 $\mathcal{R}(k,t)$ 表示通道 k 在时间 t 的活动状态（未激活时为零）。这种监督方式是通过最小化二元交叉熵损失，来训练模型输出钢琴卷帘 $\mathcal{R}$，使其预测真实标注 $\mathcal{R}$：$\hat{}\hat{}\hat{}l_{BCE}(\mathcal{R},\mathcal{R}) = \langle \mathcal{R}, -log(\mathcal{R})\rangle + \langle (1-\mathcal{R}) -log(1-\mathcal{R})\rangle$。通常，真实标注会被二值化，此时我们使用 $\mathbb{1} \in \{0,1\}^{K\times T}$ 而不是 $\mathcal{R}$。这种方法需要解码预测的钢琴卷帘 $\mathcal{R}$，以获得基于事件的表示 $\mathcal{S} = \hat{}dec_H(\mathcal{R})$，通常使用启发式方法 H，例如将连续的激活帧合并为单个音符。

c) 计算机视觉：通过将频谱图和钢琴卷帘视为图像来处理钢琴演奏转录，可以取得良好效果；卷积神经网络（CNN）等模型只需稍作调整便能胜任。

一个重要的转折点是 ONSETS&FRAMES（O&F） [16],^：它先使用一个子网络预测钢琴卷帘 $\mathcal{R}_{\downarrow}$，其中编码了起音的概率（训练时使用掩码$\mathbb{1}_{\downarrow}$，仅在琴键被按下时激活）；随后使用^另一个子网络预测帧 $\mathcal{R}_{\mathcal{N}}$，并以^$\mathcal{R}_{\downarrow}$ 为条件（训练时使用掩码 $\mathbb{1}_{\mathcal{N}}$，在每个音符的整个持续时间内激活）。之后，O&F 通过多任务^^损失 $l_{BCE}(\mathbb{1}_{\downarrow}, \mathcal{R}_{\downarrow}) + l_{BCE}(\mathbb{1}_{\mathcal{N}}, \mathcal{R}_{\mathcal{N}})$ 进行联合训练。O&F 在所有钢琴转录基准上都取得了显著提升，还引入了一项新颖的子任务——音符力度；该任务由第三个子网络^建模，用于预测力度钢琴卷帘 $\mathcal{R}_V$，并通过掩码 ℓ_2-范数损失 $l_V = \langle \mathbb{1}_{\downarrow}, (\mathcal{R}_V - \hat{}\mathcal{R}_V)^2\rangle$进行训练。由于效果和通用性前所未有，O&F 成为一种流行的基线[17], [18], ，但代价是复杂度有所增加，包括更复杂的解码器启发式方法、更大的模型，以及引入双向循环神经网络层 [19], [20],，这使其实时应用难以实现。

最近， [11] 指出钢琴卷帘存在时间精度方面的问题，并引入可训练的回归模型以提高精确率。随后，他们进一步扩展了模型，加入延音踏板检测功能。在 [21],中，使用现成的 Transformer [22] 设置，以端到端方式直接从频谱图生成 $\mathcal{S}$。除性能出色外，这两个系统的共同点是规模庞大、时间分辨率提高，并以端到端可微解法取代解码器启发式方法，这表明解码是性能瓶颈。

C. 钢琴演奏转录的拟议贡献

这些性能方面的先进改进，伴随着模型规模扩大、增加额外组件以及引入新子任务而带来了更高的复杂度 [18]。理解并解构这种复杂性是一个活跃的研究领域：研究者提出了不依赖 O&F 的替代性钢琴演奏转录子任务分解方式，例如非线性去噪与线性解混 [23], 、声源与音符编排[24] 以及 ADSR包络 [25]。此外，还探索了诸如使用可逆神经网络 [26],、重建任务 [27] 和加性注意力 [18]。

在这项工作中，我们追求一个相辅相成的目标：实现实时能力。为此，我们注意到掩码损失 l_V 在起音附近施加了时间局部性，并进一步探讨了若干思路：起音的重要性 [16] 、解码器启发式方法 [18],，以及音符力度与起音天然相关这一观点 [11]。我们提出，用于起音与力度的卷积端到端方法能够提升效率并以可承受的代价实现实时能力，同时不损害性能；此外，高效的解码启发式方法可以取代对高时间分辨率和复杂推理方案的需求。

我们提出 ONSETS&VELOCITIES (O&V)，其特点包括： 1) 在 MAESTRO 上实现了最先进的起音 检测性能，并为起音+力度提供了良好的基线。2) 大幅精简的卷积神经网络（不含循环层）以 24 毫秒分辨率的钢琴卷帘为基础，可在普通硬件上实现经济实用的实时推理。3) 解码机制简单直接，因此无需任何数据增强或扩展即可采用多任务训练方案。

我们在第II节介绍 O&V 方法。第III节通过实验佐证我们的主张。我们还提 [28] 供了基于 PyTorch 的开源实现以及实时演示。第IV节总结全文并提出未来工作方向。

II. 方法

A. 模型

给定一个波形 $x(t) \in \mathbb{R}^T$，采样率为 16kHz，我们使用大小为 2048 的汉宁窗和步长为 δ=384 （即时间分辨率为 Δ_t=24毫秒）计算其短时傅里叶变换（STFT） [14] 。随后，将其映射到 50Hz–8000Hz 范围内的 229 个梅尔频率分箱 [29] ，并取对数，得到输入表示：一个对数梅尔频谱图 $X(f, t') \in \mathbb{R}^{229\times T'}$，其中 $T' = \frac{T}{\delta}$为所得的“紧凑”时域表示（见图1）。我们还计算一阶时间导数 $\dot{X}(f, t') := X(f, t') - X(f, t'-1)$，并将其与 X拼接，构成卷积神经网络的输入。使用相同的Δ_t，我们将 MIDI 标注按时间量化为钢琴卷帘 $\mathcal{R}_V \in [0,1]^{88\times T'}$，其中 $\mathcal{R}_V(k_n, t'_n)$ 在琴键 k_n 于时间 $\Delta_t t'_n \pm \frac{\Delta t}{2}$ 被按下时记录其力度，否则为零。我们进一步将 $\mathcal{R}_V$二值化，得到 $\mathbb{1}_{\downarrow}$。

完整的卷积神经网络如图2所示。我们重点介绍以下设计原则：

a) 不使用循环层：受 [30], [31], 启发，我们采用成熟的卷积神经网络设计：由卷积干和主体构成，后接全连接头，并使用残差瓶颈 [32]。

b) 不进行池化：受 [33],启发，所有残差瓶颈都保持激活形状不变，并通过单个深度卷积层完成从输入形状到输出形状的转换；研究表明，这种方法高效且有效 [35]。请注意，输入域（频谱图）中的卷积具有垂直维度，

而输出域（钢琴卷帘）中的卷积则没有，因为我们假设相邻频率之间存在关联，而相邻琴键之间不存在关联。

c) 多阶段： 受 OpenPose 启发， [36], O&V 采用一系列残差阶段，依次对输出进行细化并生成最终结果。这适用于实时应用，因为无需重新训练即可轻松移除阶段。

d) 时间上下文： O&V 的核心是上下文感知模块 （CAM） [37], ，这是一个残差瓶颈结构，结合了时间扩张卷积和通道注意力 [38]。受时序卷积网络 [39] 和 Inception [40], 启发，我们旨在高效捕捉起音的时间邻域。

e) 模型正则化器： 在输入处以及每个输出之前，O&V 都采用子频谱批归一化 （SBN），即针对每个垂直维度分别使用一个 BN [41], [42]。我们在参数量较大的层之后加入丢弃法 [43], [44]。我们使用带泄漏的 ReLU [45], [46]作为非线性函数。

f) 时间局部性： 设计的输入特征，可直接表示强度变化。

在推理过程中，O&V 会为每个起音阶段生成一个钢琴卷帘（$\hat{\mathcal{R}}_{\downarrow}^{(1)}$，$\hat{\mathcal{R}}_{\downarrow}^{(2)}$，$\hat{\mathcal{R}}_{\downarrow}^{(3)}$），并生成一个力度钢琴卷帘 $\hat{\mathcal{R}}_V$（见图2(d)）。随后，我们提出的解码器 $\mathcal{S}$：

$$dec_{\sigma\ \rho\ \mu}\ \hat{\mathcal{R}}_{\downarrow}^{(3)}\ \hat{\mathcal{R}}_V = \ldots(\hat{\ },\hat{\ })$$ 遵循一种简单的启发式方法：先进行时间高斯平滑（平滑），方差为 σ^2，再执行非极大值抑制（*nms*）、阈值处理 ρ 和移位 μ，$\hat{\ }$从而得到包含音符起音与力度的预测乐谱 $\mathcal{S}$：

$$\begin{aligned}\hat{\downarrow} &:= \{(k,t') : nms(smooth_\sigma(\hat{\mathcal{R}}_{\downarrow}^{(3)}))_{(k,t')} \geq \rho\} \\ \hat{\mathcal{S}} &:= \{(k_n,\ \hat{\mathcal{R}}_V(k_n,t'_n),\ \Delta_t t'_n + \mu) : (k_n,t'_n) \in \hat{\downarrow}_\rho\}\end{aligned} \quad (1)$$

nms 操作会将任何严格小于（k，$t+1$）或（k，$t-1$）的（k，t）项置零。在所得位置读取音符事件，并统一偏移常数 μ。在本研究中，我们使用通过在 MAESTRO 验证集划分的一个子集上对训练好的卷积神经网络进行交叉验证而得到的数值 $\sigma=1$，$\mu=-0.01s$，$\rho=0.74$（请注意，这与用于评估的测试集划分不同）。虽然最优的 ρ会在训练过程中波动，但我们发现 $\sigma=1$，$\mu=-0.01s$较为稳定。

B. 训练

我们通过最小化以下多任务损失，训练卷积神经网络联合预测起音概率和力度：

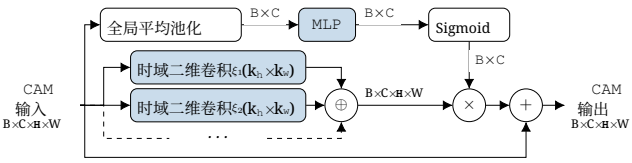
$$l_{\downarrow V}(\mathbb{1}_{\downarrow_3},(\hat{\mathcal{R}}_{\downarrow}^{(1)},\hat{\mathcal{R}}_{\downarrow}^{(2)},\hat{\mathcal{R}}_{\downarrow}^{(3)}),\mathcal{R}_{V_3},\hat{\mathcal{R}}_V) :=$$

$$\sum_i^3 l_{BCE}(\mathbb{1}_{\downarrow_3},\hat{\mathcal{R}}_{\downarrow}^{(i)}) + \lambda \cdot l_{V'}(\mathcal{R}_{V_3},\hat{\mathcal{R}}_V), \text{ where}$$

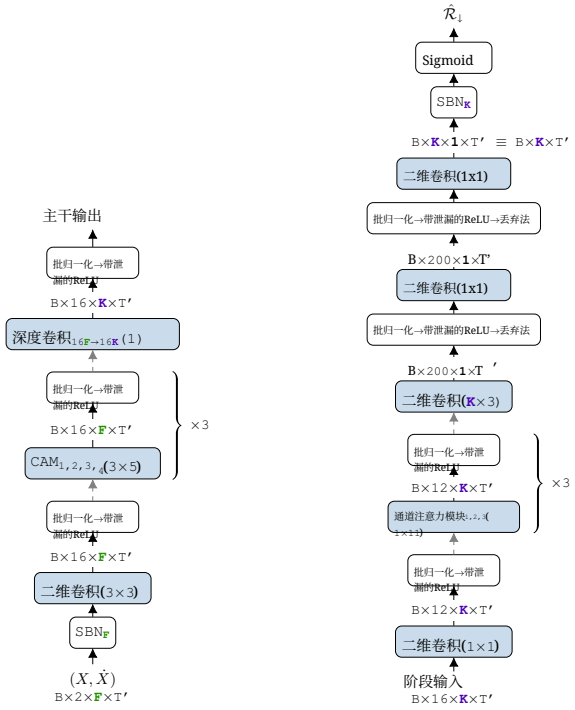
$$l_{V'}(\mathcal{R}_{V_3},\hat{\mathcal{R}}_V) :=$$

$$\langle \mathbb{1}_{\downarrow_3},\ (\mathcal{R}_{V_3}\cdot -log(\hat{\mathcal{R}}_V))((1-\mathcal{R}_{V_3})\cdot -log(1-\hat{\mathcal{R}}_V))\rangle$$

$\mathbb{1}_{\downarrow_3}$ 和 $\mathcal{R}_{V_3}$ 卷帘是对 $\mathbb{1}_{\downarrow}$ 和 $\mathcal{R}_V$的一种直接改动：在（k，t）处的每个激活帧也会扩展为 $t+1$ 和 $t+2$（即音符起音覆盖 3 帧，而非 1 帧）。这一简单扩展对于达到目标性能至关重要，并且与我们的解码器结合后，使得

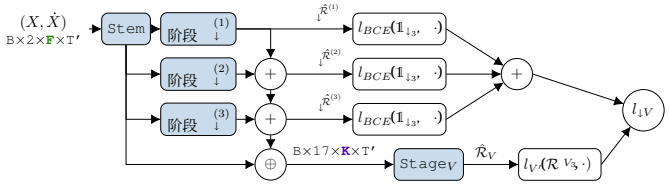


（a）基于 [37] 的**通道注意力模块**$\xi_1,\ \xi_2\ldots(k_h \times k_w)$ 示意图。中间分支将（⊕）个时间扩张卷积（TDConv）进行拼接，每个卷积的时间扩张率为 ξ。每个 TDConv 的输出通道数为$\mathbf{H}$ 除以所拼接的 TDConv 层数，并调整填充以保持形状不变。顶部支路作为通道注意力机制（×），其中多层感知机（MLP）充当瓶颈结构（我们使用了包含 ReLU激活和 8 个隐藏维度的 2 层结构）。底部支路为残差连接（+）。



（b）**主干**示意图，它是一个由通道注意力模块驱动的残差卷积神经网络。SBN 包含 **F** 个频带（每个垂直维度对应一个）。×3 花括号表示 3 个顺序块。请注意，通过深度逐通道卷积 **F→K** 实现了单步转换；该卷积为 DepthConv(1)，时间卷积核宽度为 1。

(c) 起音**阶段**$_{\downarrow}$示意图。这是对主干的修改，后接卷积，其作用类似于沿时间维度移动的多层感知机 T'。我们在多层感知机中加入丢弃法。力度**阶段**$_V$ 与阶段$_{\downarrow}$类似，但仅包含 {×1} 个 CAM模块，且输入通道数为 17，而非 16。



（d）我们提出的包含3个阶段的 ONSETS&VELOCITIES 完整架构及训练损失示意图。每个阶段$_{\downarrow}{}^{(i)}$ 都会生成一个起音钢琴卷帘$\hat{\ }\mathcal{R}_{\downarrow}^{(i)}$，后续阶段通过残差连接细化输出（+ 保持形状不变）。速度阶段$_V$ 利用最终起音阶段提供的起音位置相关信息，并对所有部分进行联合训练。参见图1），其中给出了X、$\hat{\ }\mathcal{R}_{\downarrow}^{(3)}$和$\hat{\ }\mathcal{R}_V$的定性示例。第II-B节介绍了损失 $l_{\downarrow V}$。

图2：我们提出的卷积神经网络。四阶张量的维度为批次×通道×高度×宽度。设计原理、接口和损失函数见第II节。

表 I：按规格（仅起音+力度的参数数量、架构和功能）和性能（精确率、召回率、F_1-分数及 MAESTRO 版本）比较表现最佳的模型。

模型	起音+力度 # 参数	架构	偏移/踏板?	起音 (%)			起音+力度 (%)			MAESTRO 版本
				P	R	F_1	P	R	F_1	
O&F [16]	10M	双向循环神经网络	√ /	98.27	92.61	95.32	-	-	-	v1
回归 [11]	12M	双向循环神经网络	√ / √	98.17	95.35	96.72	-	-	-	v2
Transformer [21]	–	Transformer	√ /	-	-	96.13	-	-	-	v3
O&V（我们的）	**3.13M**	CNN	/	98.58	95.07	**96.78**	96.25	92.86	94.50	v3

从而无需采用 [11]中讨论的复杂解码方案。掩码损失 $l_{V'}$ 是前文所述 l_V的交叉熵变体，在 [16] 和[11], 中提出，旨在仅在起音附近预测正确的力度。

所有模型权重均采用高斯-何分布初始化 [47], ，偏差初始化为 0；CAM 通道注意力模块的偏差（位于 Sigmoid 之前）则初始化为 1，以促进信号流动。我们使用 Adam优化器，并采用解耦权重衰减 [48], [49] ，衰减值为 3×10^{-4}。训练时随机抽取 5 秒片段组成批次（批次大小为 40， ∼14 个批次为一个周期），共训练 ∼70K 个批次。学习率先在 500 个批次内从 0 逐步升至 0.008，随后采用带热重启的余弦退火 [50], ，周期为 1000 个批次，并在每个周期后衰减 97.5%。批归一化/子批归一化的动量为 95%，丢弃率为 15%，带泄漏的 ReLU 斜率为 0.1。在 l'_V 中，我们使用 $\lambda=10$。鉴于 $\mathbb{1}_{3\downarrow}$较为稀疏，我们将 $l_{BCE}(\mathbb{1}_{3\downarrow}, \cdot)$ 中的正项权重设为负项的 8 倍。使用 NVIDIA 2080Ti 图形处理器时，训练速度为每小时 1800 个批次。

C. 评估

遵循 O&F [16],REGRESSION [11]和 TRANSFORMER [21],的相同评估流程，并采用 [51]实现于 mir_eval 库中的标准评估指标 [52], ，我们报告预测起音的精确率（P）、召回率（R）和 F_1 分数；若预测起音与真实标注的时间差在 50 毫秒以内，则视为正确。遵循 O&F [16, 3 的起音+力度评估 1],还增加了一项约束：预测力度与归一化到 0 至 1 范围内的真实标注之间的差值也必须在 0.1 以内。

MAESTRO 数据集仍在积极扩充和整理中，目前已有 3 个版本。我们在表 I 中报告了各版本的相应结果。需要指出的是，版本 2 和 3 几乎相同，不过跨版本比较的结果应视为近似值。

III. 实验与讨论

我们使用 MAESTRO v3 的训练划分训练了 O&V，未进行任何扩充或增强，并在起音检测方面取得了当前最先进的性能（见表 I）。下面我们讨论其中的一些影响：

a) 时间分辨率：我们的结果似乎与提高时间分辨率的必要性相矛盾； [11]采用了 8 毫秒的分辨率，而我们的结果表明，24 毫秒的钢琴卷帘表示配合式（1）中给出的解码器，仍具有竞争力。

b) 降低内存占用：**表 I 报告了仅负责起音检测和帧检测的各组件的模型参数**（TRANSFORMER 总共有 ∼54M 个参数，但它采用联合方式转录所有内容，因此无法进行公平比较）。O&V 的表现优于最佳替代方案 REGRESSION，且参数量少 ∼4 倍。

c）经济实惠的实时推理：O&F 和回归中使用的双向循环层不适合实时处理。TRANSFORMER 在英特尔至强 CPU（1 个核心）上转录一个 120 秒的文件耗时 ∼380s，在 Tesla-T4 GPU 上耗时 ∼20s（包括偏移），以上数据均来自运行官方 Colab 实现[2]。O&V 在 8 核英特尔 i7-11800H CPU 上处理同一文件耗时不到 2 秒。即使考虑核心数量，O&V 的速度也比 TRANSFORMER 快约一个数量级。

d）概念简洁性：从本质上说，O&V 通过将前馈卷积神经网络应用于基于钢琴卷帘的判别任务，并辅以简单的解码启发式方法，重拾了 [12] 所体现的简洁性。其架构允许在无需重新训练的情况下移除起音阶段，从而以很少的额外复杂性，在运行时间和性能之间实现灵活权衡。

e) 延迟：我们提出的 O&V 组件的感受野为：主干：60 帧（1.44 秒），阶段$_{\downarrow}{}^{(i)}$：99 帧（∼2.38 秒），阶段$_{\text{V}}$：35 帧（0.84 秒）。理论上，这会导致超过 9 秒的延迟，远非响应迅速的系统。我们注意到，非正式测试表明，可以截短延迟而不会显著影响结果（我们在一次现场研讨会中使用了 4 秒的延迟），因此鼓励实际应用。我们还指出，我们的重点是寻找一种推理成本可承受且性能有竞争力的卷积神经网络，并未针对低感受野进行优化；通过对架构稍作调整（例如减少连续阶段的数量或减少 CAM），或许就能实现低感受野。

IV. 结论与未来工作

我们介绍了 ONSETS&VELOCITIES（O&V），这是一种卷积架构，可在性能不打折的情况下，借助配置适中的通用硬件实现实时能力，同时显著缩减模型规模并降低时间分辨率。O&V 在钢琴演奏转录（PPT）的音符起音 检测中达到了最先进的性能，并为起音+力度 检测建立了良好的基准。未来的工作可以包括缩短感受野、扩展至偏移 和踏板 检测、在不同乐器上进行训练和评估，以及通过消融研究分析设计选择（例如阶段数量和时间分辨率）。

[2]https://github.com/magenta/mt3

致谢

´A.F. 感谢 Jes´us Monge Alvarez 和 Christian J. Steinmetz 提供宝贵的反馈，感谢巴利阿里群岛研究所 ` 通过研究资助 389062 INV-23/2021 支持本项工作，并感谢国际马克斯·普朗克智能系统研究生院（IMPRS-IS）提供进一步支持。

参考文献

[1] J. Downie、D. Byrd 和 T. Crawford，“ISMIR 十年：对挑战与机遇的思考。”ISMIR 论文集，第 13–18 页，2009 年 1 月。[2] E. Benetos、S. Dixon、D. Giannoulis、H. Kirchhoff 和 A. Klapuri，“自动音乐转录：挑战与未来方向”，《智能信息系统杂志》，第 41 卷，2013 年 12 月。[3] J. Driedger，“使用音频分解技术处理音乐信号”，博士论文，FAU Erlangen-N¨urnberg，2016 年。[4] Y. LeCun、Y. Bengio 和 G. Hinton，“深度学习”，《自然》，第 521 卷，第 436–44 页，2015 年 5 月。[5] V. Emiya、N. Bertin、B. David 和 R. Badeau，“MAPS——用于多音高估计和自动音乐转录的钢琴数据库（研究报告）”，2010 年。[6] A. Ycart 和 E. Benetos，“A-MAPS：加入节奏和调性标注的增强版 MAPS 数据集”，第 19届 ISMIR 会议，2018 年 9 月。[7] M. M¨uller、V. Konz、W. Bogler 和 V. Arifi-M¨uller，“萨尔兰音乐数据（SMD）”，第12届 ISMIR 的最新成果与演示环节，2011 年。[8] J. Thickstun、Z. Harchaoui 和 S. M. Kakade，“从零开始学习音乐特征”，载于 ICLR，2017 年。[9] C. Hawthorne、A. Stasyuk、A. Roberts、I. Simon、C.-Z. A. Huang、S. Dieleman、E. Elsen、J. Engel 和 D. Eck，“借助 MAESTRO 数据集实现因式分解钢琴音乐建模与生成”，载于《国际学习表征会议》，2019 年。[10] Q. Kong、B. Li、J. Chen 和 Y. Wang，“GiantMIDI-Piano：大型古典钢琴音乐 MIDI 数据集。”ISMIR 汇刊，第 87–98 页，2022 年。[11] Q. Kong、B. Li、X. Song、Y. Wan 和 Y. Wang，“通过回归起始与结束时间实现带踏板的高分辨率钢琴转录”，IEEE/ACM 音频、语音与语言处理汇刊，第 29 卷，第 3707–3717 页，2021 年 10 月。[12] R. Kelz、M. Dorfer、F. Korzeniowski、S. B¨ock、A. Arzt 和 G. Widmer，“简单逐帧方法在钢琴转录中的潜力”，载于 ISMIR 论文集，2016 年 8 月，第 475–481 页。[13] N. Zeghidour、O. Teboul、F. de Chaumont Quitry 和 M. Tagliasacchi，“LEAF：用于音频分类的可学习前端”，载于《国际学习表征会议》，2021 年。[14] R. Bracewell，《傅里叶变换及其应用》，第 2 版。东京：麦格劳-希尔讲谈社有限公司，1978 年。[15] Q. Kong、Y. Cao、T. Iqbal、Y. Wang、W. Wang 和 M. D. Plumbley，“PANNs：用于音频模式识别的大规模预训练音频神经网络”，IEEE/ACM《音频、语音与语言处理汇刊》，第 28 卷，第 2880–2894 页，2020 年。[16] C. Hawthorne、E. Elsen、J. Song、A. Roberts、I. Simon、C. Raffel、J. H. Engel、S. Oore 和 D. Eck，“Onsets and Frames：双目标钢琴转录。”载于 ISMIR，2018 年，第 50–57 页。[17] J. Kim 和 J. Bello，“通过对抗学习改进 Onsets and Frames 音乐转录”，载于 ISMIR 论文集，2019 年，第 670–677 页。[18] K. Cheuk、Y. Luo、E. Benetos 和 D. Herremans，“使用加性注意力重新审视 Onsets and Frames 模型”，载于国际神经网络联合会议（IJCNN）论文集，2021 年。[19] S. Hochreiter 和 J. Schmidhuber，“长短期记忆”，《神经计算》，第 9 卷，第 1735–80 页，1997 年 12 月。[20] M. Schuster 和 K. Paliwal，“双向循环神经网络”，IEEE 信号处理汇刊，第 45 卷，第 11 期，第 2673–2681 页，1997 年。[21] C. Hawthorne、I. Simon、R. Swavely、E. Manilow 和 J. H. Engel，“使用 Transformer 进行序列到序列钢琴转录”，载于 ISMIR 论文集，2021 年 11 月，第 246–253 页。[22] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. u. Kaiser 和 I. Polosukhin，“注意力就是你所需要的一切”，载于《神经信息处理系统进展》，第 30 卷，2017 年。[23] R. Kelz 和 G. Widmer，“非线性去噪，线性解混”，载于《我（仍然）不敢相信它没有更好！》NeurIPS 2021 研讨会，2021 年。[24] L. S. Mart´ak、R. Kelz 和 G. Widmer，“在复调钢琴转录系统中平衡偏差与性能”，《信号处理前沿》，第 2 卷，2022 年。[25] R. Kelz、S. B¨ock 和 G. Widmer，“深度复调 ADSR 钢琴音符转录”，ICASSP，第 246–250 页，2019 年。[26] R. Kelz 和 G. Widmer，“利用可逆神经网络实现可解释的复调转录”，载于 ISMIR 论文集，2019 年。[27] K. Cheuk、Y. Luo、E. Benetos 和 D. Herremans，“频谱图重建对自动音乐转录的影响：提高转录准确率的另一种方法”，载于 2020 年第 25 届国际模式识别会议（ICPR），2021 年，第 9091–9098 页。[28] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga、A. Desmaison、A. Kopf、E. Yang、Z. DeVito、M. Raison、A. Tejani、S. Chilamkurthy、B. Steiner、L. Fang、J. Bai 和 S. Chintala，“PyTorch：一种命令式风格的高性能深度学习库”，载于《神经信息处理系统进展》，2019 年，第 8024–8035 页。[29] S. S. Stevens 和 J. E. Volkmann，“音高与频率的关系：修订后的音阶”，《美国心理学杂志》，第 53 卷，第 329 页，1940 年。[30] I. Radosavovic、R. P. Kosaraju、R. Girshick、K. He 和 P. Dollar，“设计网络设计空间”，载于 CVPR 论文集，2020 年 6 月。[31] Z. Liu、H. Mao、C.-Y. Wu、C. Feichtenhofer、T. Darrell 和 S. Xie，“面向 2020 年代的卷积网络”，CVPR 论文集，2022 年。[32] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于 CVPR 论文集，2016 年 6 月，第 770–778 页。[33] J. T. Springenberg、A. Dosovitskiy、T. Brox 和 M. A. Riedmiller，“追求简洁：全卷积网络”，载于 ICLR，2015 年。[34] F. Chollet，“Xception：采用深度可分离卷积的深度学习”，载于 CVPR 论文集，2017 年，第 1800–1807 页。[35] A. G. Howard、M. Zhu、B. Chen、D. Kalenichenko、W. Wang、T. Weyand、M. Andreetto 和 H. Adam，“MobileNets：面向移动视觉应用的高效卷积神经网络”，CoRR，2017 年。[36] Z. Cao、G. Hidalgo Martinez、T. Simon、S. Wei 和 Y. A. Sheikh，“OpenPose：利用部件亲和场进行实时多人二维姿态估计”，IEEE《模式分析与机器智能汇刊》，2019 年。[37] J. Zhang、Z. Chen 和 D. Tao，“通过渐进式上下文细化进行人体关键点检测”，载于 ICCV 研讨会：COCO关键点检测挑战赛赛道，2019 年 10 月。[38] J. Hu、L. Shen 和 G. Sun，“压缩与激励网络”，载于 CVPR 论文集，2018 年，第 7132–7141 页。[39] S. Bai、J. Z. Kolter 和 V. Koltun，“用于序列建模的通用卷积网络与循环网络的实证评估。”CoRR，2018 年。[40] C. Szegedy、W. Liu、Y. Jia、P. Sermanet、S. Reed、D. Anguelov、D. Erhan、V. Vanhoucke 和 A. Rabinovich，“通过卷积更进一步”，载于 CVPR 论文集，2015 年，第 1–9 页。[41] S. Ioffe 和 C. Szegedy，“批量归一化：通过减少内部协变量偏移加速深度网络训练”，载于 ICML 论文集，2015 年，第 448–456 页。[42] S. Chang、H. Park、J. Cho、H. Park、S. Yun 和 K. Hwang，“用于神经音频数据处理的子频谱归一化”，ICASSP，2021 年。[43] N. Srivastava、G. Hinton、A. Krizhevsky、I. Sutskever 和 R. Salakhut-dinov，“丢弃法：防止神经网络过拟合的简易方法”，《机器学习研究杂志》，第 15 卷，第 56 期，第 1929–1958 页，2014 年。[44] X. Li、S. Chen、X. Hu 和 J. Yang，“通过方差偏移理解 Dropout 与批量归一化之间的不协调”，载于 CVPR 论文集，2019 年，第 2682–2690 页。[45] V. Nair 和 G. E. Hinton，“整流线性单元提升受限玻尔兹曼机的性能”，载于 ICML 论文集，2010 年，第 807–814 页。[46] A. L. Maas、A. Y. Hannun 和 A. Y. Ng，“整流器非线性函数可改善神经网络声学模型”，载于 ICML 论文集，2013 年。[47] K. He、X. Zhang、S. Ren 和 J. Sun，“深入研究整流器：在 ImageNet 分类上超越人类水平的表现”，载于 ICCV 论文集，2015 年，第 1026–1034 页。[48] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于 ICLR 论文集，2015 年。[49] I. Loshchilov 和 F. Hutter，“解耦权重衰减正则化”，载于 ICLR 论文集，2019 年。[50] I. Loshchilov 和 F. Hutter，“SGDR：采用热重启的随机梯度下降”，载于 ICLR 论文集，2017 年。[51] M. Bay、A. F. Ehmann 和 J. S. Downie，“多基频估计与跟踪系统的评估”，载于 ISMIR 论文集，2009 年，第 315–320 页。[52] C. Raffel、B. McFee、E. J. Humphrey、J. Salamon、O. Nieto、D. Liang 和 D. P. W. Ellis，“mir_eval：常用音乐信息检索指标的透明实现”，载于 ISMIR 论文集，2014 年，第 367–372 页。