用于语音识别中高效唤醒词检测的双重注意力神经转导器

Saumya Y. Sahai、*Jing Liu*、*Thejaswi Muniyappa*、*Kanthashree M. Sathyendra*、*Anastasios Alexandridis*、*Grant P. Strimel*、*Ross McGowan*、*Ariya Rastrow*、*Feng-Ju Chang*、*Athanasios Mouchtaris*、*Siegfried Kunzmann*

Amazon Alexa AI，美国

arXiv:2304.01905v2 [cs.SD] 5 Apr 2023

摘要

我们提出了双注意力神经偏置，这是一种旨在提升唤醒词（WW）识别能力并改善语音识别任务推理时间延迟的架构。该架构利用唤醒词检测，为输入音频帧选择要执行的注意力网络分支，从而动态切换运行时计算路径。通过这种方法，我们有效提升了唤醒词检测准确率，同时降低了以浮点运算次数（FLOPs）衡量的运行时计算成本。我们使用内部去标识化数据集进行演示，结果表明，所提出的双注意力网络可将唤醒词音频帧的计算成本降低 90%，而参数数量仅增加 1%。与基线系统相比，该架构使唤醒词 F1分数相对提升 16%，并使通用稀有词错误率相对降低 3%。

索引词— 语音识别、推理优化、唤醒词检测、注意力、神经偏置、个性化

1. 引言

连接主义时序分类（CTC） [1], 监听-注意-拼写（LAS） [2], 循环神经网络转换器（RNN-T） [3], Transformer转换器 [4–8], 及其变体 ConvRNN-T [9], Conformer [10, 11] 等端到端（E2E）自动语音识别系统日益普及，这得益于它们相比混合HMM-DNN系统的优异性能，也使其成为适用于商业虚拟语音助手的有前景架构。混合模型分别优化声学模型（AM）、发音模型（PM）和语言模型（LM），而端到端系统则联合优化这些模型，直接从输入序列生成词序列。这类完全神经网络端到端方法结构简单，并具备统一的压缩能力，因此是低资源场景的有力候选方案。然而，端到端自动语音识别系统的一项主要局限是：对于配对音频文本训练数据中较少出现的词语，它们难以准确识别，例如客户指定用于称呼虚拟助手的自定义唤醒词（WW）（如 Ziggy、Hey Shaq）、联系人姓名、专有名词及其他稀有命名实体 [12, 13]。为解决这一问题，先前的研究 [8, 14] 提出了基于注意力的神经偏置方法，通过计算编码后的音频表示与个性化目录嵌入之间的相似度，应用偏置适配器机制。基于注意力的神经偏置有望提升个性化实体名称的识别效果；然而，由于该方法需要逐帧处理音频编码，计算复杂度较高，因此产生的运行时延迟使这类基于注意力的偏置网络难以在受硬件限制（例如内存带宽有限、CPU约束）的设备端系统中大规模部署。

*共同贡献。

为解决端侧自动语音识别面临的计算限制，模型压缩是一种常用方法。一般而言，模型压缩技术可分为两类：架构修改和权重表示。前者将复杂架构简化为更精简的替代方案，后者则使用低比特表示来表征权重。我们的工作属于架构修改这一类。该类别还包括 CIFG [15]，它通过合并输入门和遗忘门来简化 LSTM 结构 [16]，从而减少 25% 的参数；简化循环单元 [17]为边缘自动语音识别引入了更高效的循环单元；低秩分解 [18], 、双焦点 [19], 、动态编码器 [20], 、摊销网络 [21, 22], 、Linformer [23], 、Performer[24] 以及时间缩减层 [2, 25, 26]，这些方法旨在降低运行时延迟。第二类则包括量化 [27–29],和稀疏性 [30–32]，它们是采用低比特整数或稀疏表示来表征权重的主要范式。

我们的工作受到双焦点神经转导器的启发 [19], ，该转导器包含两个音频编码器网络，可在运行时动态切换。我们的工作与之相比的一个主要区别在于，采用了计算成本分摊的多头注意力（MHA） [4] 偏置网络，旨在同时增强自定义唤醒词和个性化实体。与不区分句子片段的基础神经偏置 [8, 14] 不同，所提出的双注意力网络仅在句首偏置唤醒词嵌入，而在唤醒词之后的片段偏置专有名称嵌入（例如联系人姓名、设备名称）。

2. 相关工作

2.1. 双焦点RNN-T

神经序列转导器是流式端到端自动语音识别系统 [3]，通常由音频编码器、文本预测器和联合网络组成。编码器的行为类似于声学模型，会为每个输入音频帧 $\mathbf{x}_T = (x_0, \ldots, x_T)$ 生成高级声学表示 $\mathbf{h}_t^{enc}$。文本预测器的行为类似于语言模型，会对先前预测的词片段 $\mathbf{y}_{u-1} = (y_0, \ldots, y_{u-1})$ 进行编码，并输出 $\mathbf{h}_u^{pred}$，其中

$$\mathbf{h}_t^{enc} = \text{AudioEncoder}(\mathbf{x}_t); \quad \mathbf{h}_u^{pred} = \text{TextPredictor}(\mathbf{y}_{u-1}).$$

联合网络融合 $\mathbf{h}_t^{enc}$和 $\mathbf{h}_u^{pred}$，并将它们依次传入全连接层和 softmax 层，从而获得词片段上的输出概率分布。

双焦点RNN-T [19], 是一种特殊类型的神经转导器，由两个音频编码器组成：一个针对包含唤醒词前音频帧和唤醒词音频帧的缓冲的前导音频片段进行训练的小型/快速编码器；以及一个用于处理其余音频的大型/慢速编码器，并利用唤醒词检测在两者之间进行切换（图

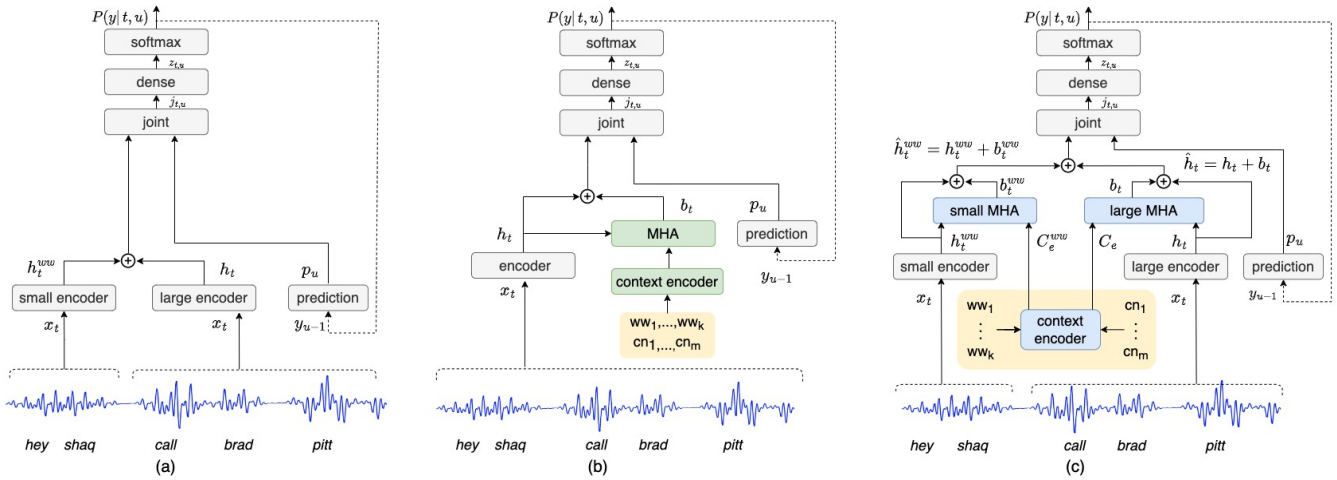

图 1：模型架构：(a) 基线双焦点神经转导器 预训练基础模型 [19]；(b) 基线单注意力神经偏置单注意力基础模型*128* [8, 14]；(c) 所提出的双注意力神经偏置 双注意力-λ，其中 λ 是小型多头注意力的投影大小；使用预训练权重（灰色模块）初始化模型；添加双注意力模块（蓝色模块）；冻结灰色模块，仅训练蓝色模块。

1(a)）。双焦点架构通过将唤醒词音频帧分流至快速编码器分支，降低了延迟。然而，它适应并识别新词或生僻词的能力有限，尤其是对于用户指定的、交由低容量小型音频编码器处理的唤醒词。这一缺陷会错误拒绝由自定义唤醒词发起的语音查询，或错误地将查询转交给另一个虚拟助手个性，从而降低用户体验。所提出的双注意力架构旨在利用注意力网络缓解这一局限，使其“只专注于”增强所需的唤醒词。

2.2. 神经偏置

为了利用用户的自定义环境和偏好设置，提高对用户向语音助手提出的个性化请求的识别效果， [8] 和 [14]都提出了神经偏置（图 1(b)），其由多头注意力层 [4] 构成，用于衡量音频编码与实体名称嵌入之间的相似度。注意力权重按帧计算，以评估用户预先定义的实体名称与当前音频帧的相关性。神经偏置能有效提升个性化实体名称（例如专有名称，如联系人和设备名称）的识别效果，因为相关性越高的实体名称会获得越高的注意力权重。然而，由于点积注意力的复杂度为二次方， [4], 运行时延迟一直是将神经偏置部署到嵌入式自动语音识别系统中的瓶颈。

3. 双注意力神经偏置

为解决设备端延迟和计算限制问题，并受到双焦点RNN-T和神经偏置（第 2 节）的启发，我们提出了双注意力神经偏置（图 1(c)），使运行时计算路径能够动态切换；具体而言，利用唤醒词检测来选择网络分支，以处理输入音频帧。双注意力的设计理念是引入双焦点“镜头”，使其分别聚焦于话语的不同片段。该设计的显著特点是训练两个替代多头注意力网络（图 1(c) 中以蓝色突出显示的组件）。小型注意力网络 $\mathcal{A}^s$ 与小型音频编码器相结合，针对引入片段进行训练；大型注意力网络 $\mathcal{A}^l$ 则与大型音频编码器配合，处理其余音频。 $\mathcal{A}^s$ 旨在提升用户指定的自定义唤醒词（通常为 10 个左右）的自动语音识别准确率，而 $\mathcal{A}^l$ 则旨在提高对个性化实体名称的识别能力（可扩展至数万个）。 $\mathcal{A}^s$ 的隐藏单元数量少于 $\mathcal{A}^l$，因此帧处理速度更快，但粒度较粗。相比之下， $\mathcal{A}^l$ 容量更大，但需要更多计算。该设计的最后一个组件是上下文编码器，即双向长短期记忆网络编码器，它接收由 SentencePiece分词器 [33]生成的标记化自定义唤醒词/专有名称。该双向长短期记忆网络的最后一个状态用作嵌入 C_e^{ww} 或 C_e。在图 1(c) 中，我们首先对双焦点转导器进行预训练（灰色模块），然后冻结其余预训练权重（灰色模块），仅微调上下文编码器和两个多头注意力模型（蓝色模块） [14]。

3.1. 双注意力偏置网络

小型多头注意力网络 $\mathcal{A}^s$ 经过训练，用于学习前导音频编码与用户启用的唤醒词文本嵌入之间的相关性。由于唤醒词之前的口语词具有天然较低的困惑度，因此该模型较为轻量。相比之下，唤醒词后片段的困惑度较高，需要采用容量更大的多头注意力模型 $\mathcal{A}^l$ 。这种双注意力设计旨在达到与单注意力基线（第 2.2 节）相当的准确率，并减少浮点运算次数，因为该架构重点降低多头注意力推理成本，而这正是主要的运行时瓶颈之一。

3.2. 动态目录掩蔽

与 [8, 14] (图 1(b)) 不同，后者会静态拼接唤醒词和 专有名称 嵌入，而不区分句子片段；我们设计（图 1(c)）的一个显著特点是目录掩码，它会根据标示唤醒词结束位置的帧索引动态确定。在推理时，我们仅通过动态屏蔽 专有名称 词元，在 $\mathcal{A}^s$ 中应用唤醒词嵌入 C_e^{ww}。这样一来，我们便能有效地将偏置候选项从数万个（即 专有名称）缩减到 6 个（即仅自定义唤醒词），从而更好地聚焦，并排除联系人/设备名称等相关性较低的目录，避免它们出现在句首。同理，对于 $\mathcal{A}^l$，我们会丢弃唤醒词目录，仅应用 专有名称嵌入 C_e。这样，我们便能让这两个

模型	引导音频 片段浮点运算次数	偏置层 参数	引导音频 目录大小	RNN-T			C-T		
				F1R	TRRR	TARR	F1R	TRRR	TARR
预训练基础模型	–	–	–	–	–	–	–	–	–
单注意力基础模型-*128*	$3.3M$	$400K$	300	+14.72%	-14.80%	+28.73 %	+14.96%	-12.84%	+41.96%
单注意力目录掩码	$190K$ (-93.7%)	$400K$(+0.0%)	6	+17.03%	-15.17%	+31.39%	+16.93%	-11.11%	+42.77%
dual-attn-64	$95K$ (-96.9%)	$483K$(+20.7%)	6	+16.42%	-14.07%	+36.45%	+14.14%	-7.98%	+32.42%
双重注意力-*32*	$48K$ (-98.4%)	$441K$(+10.3%)	6	+16.00%	-12.43%	+26.07%	+15.34%	-12.15%	+42.51%
dual-attn-16	$23K$ (-99.2%)	$421K$(+5.1%)	6	+14.34%	-16.63%	+29.01%	+14.38%	-12.15%	+39.23%
dual-attn-8	$12K$ (-99.6%)	$410K$ (+2.5%)	6	+2.38%	-4.21%	+8.52%	+9.30%	-5.90%	+19.07%

表 1：以浮点运算次数（M=10^6、 K=10^3）衡量的计算成本；偏置层中的参数数量；对于图 1(c) 中提出的双注意力-λ，与图 1(a) 所示的基线预训练基础模型 相比，F1分数（记为 F1R）、真拒绝率（TRRR）和真接受率（TARR）的相对变化；单注意力基础模型-*128* 对应图 1(b)； +/- 表示提升/下降。

多头注意力通过屏蔽无关目录，分别偏向各自的目标。如图 1(c) 所示，黄色高亮框中的上下文编码器会离线运行，生成并缓存用户自定义唤醒词 ww_1、.....、ww_k的神经嵌入 C_e^{ww}，以及专有名称 cn_1、.....、cn_m的神经嵌入 C_e。这些缓存的嵌入 C_e^{ww}和 C_e会在运行时推理期间动态屏蔽。我们还在目录中引入了一个特殊的无偏置 词元，如 [14, 34] 所示，以帮助双注意力系统学习何时不进行偏置。

4. 实验

4.1. 数据集

我们使用了 114K 小时的去标识化内部语音助手通用数据集，该数据集从覆盖 20 多个领域（例如音乐、通信和智能家居）的实时流量中随机采样得到，用于预训练基线 RNN-T 和 Conformer-Transducer（C-T）模型[1]。为了训练双注意力网络，我们使用了 290 小时的专有名称数据（包含对命名实体的提及）、按 1:2.5 比例混合的通用数据，以及 3.6K 小时的半监督数据集，其中包含使用教师模型生成的 6 个自定义唤醒词 [35]。为了评估模型，我们使用了一个 75 小时的通用测试集和一个 20 小时的专有名称测试集，二者均由人工转录。为计算唤醒词的真接受率和真拒绝率，我们使用了 25 小时的人工标注数据，其中包含 6 个唤醒词。训练集和测试集均已去标识化，且彼此没有重叠。

4.2. 实验设置

我们使用两种预训练的 ASR 架构 RNN-T 和 C-T 来评估双重注意力神经转换器。

RNN-T 和 C-T 预训练。输入音频特征为 64 维 LFBE 特征，每 10ms 提取一次，窗口大小为 25ms，因此每帧包含 192 维特征。真实标注 token 分别通过 2.5K 和 4K 的词片段分词器 [33, 36] 处理，用于 RNN-T 和 C-T。RNN-T 编码器包含 5 个 LSTM 层，以及一个位于第 3 层、下采样因子为 2 的时间缩减层。引导音频编码器的每个 LSTM 层包含 256 个单元，大型音频编码器的每个 LSTM 层则包含 1120 个单元（图 1(a)）。预测网络包含 2 个 LSTM 层，每层有 1088 个单元。C-T 编码器网络由 2 个卷积层组成，卷积核均为 128 个、大小为 3；第一层和第二层的步幅分别为 2 和 1，后接一个将输入特征投影至 512 维的全连接层。随后，输入被送入 14 个 Conformer 块，这些模块包含层归一化和残差层间连接。每个 Conformer 块包含一个 1024 单元的前馈层、一个具有 8 个 64 维注意力头的 Transformer 层，以及一个卷积核大小为 15 的卷积模块。预测网络包含 2 个 LSTM 层，每层有 736 个单元。为实现流式处理，卷积和注意力计算均基于当前及之前的音频帧。对于 RNN-T 和 C-T，编码器和预测网络的输出均通过一个包含 512 个单元的前馈层进行投影。

基线系统。 图 1(a) 所示的基线 预训练基础模型 是一种双焦点神经转导器 [19]（RNN-T 或 C-T），包含两个音频编码器，但没有神经偏置层 [8, 14]。该模型推理速度快，但对专有名称的准确率较低。图 1(b) 所示的第二个基线 单注意力基础模型-*128*是一种单注意力神经偏置转导器模型，如 [8, 14] 所示，其键和值被投影到 128 维。与预训练基础模型相比，该基线对专有名称 的准确率较高，但运行时推理速度较慢。第三个模型 单注意力目录掩码 与单注意力基础模型-*128*（图 1(b)）相同，不同之处在于，它通过动态目录掩蔽（第 3.2 节）从小型音频编码器中移除了无关目录（例如专有名称）。

双注意力偏置网络的配置。上下文编码器是一个 BiLSTM，每个前向和后向 LSTM 均包含 64 个单元。输入和输出均采用 64 维投影。该上下文编码器从头开始训练，用于生成唤醒词和专有名称的嵌入。随后，这些嵌入被馈送至双注意力层，以对音频编码器的输出进行偏置（图 1(c)）。更具体地说，大型多头注意力网络将大型音频编码器的输出作为查询，将专有名词嵌入作为键和值，并将其投影到 128 维。另一方面，小型多头注意力是一个轻量级模型，它以小型音频编码器的输出作为查询，以唤醒词嵌入作为键和值，并将其投影到大小为 λ 的空间，该模型记作 双注意力-λ。我们对 $\lambda = 64$、32、16 和 8 进行了实验。在后续实验中，小型和大型多头注意力均仅使用 1 个注意力头，因为我们没有观察到使用 2 个或 4 个注意力头能带来准确率提升。

5. 结果

给定模型 A 的词错误率（WER_A）和基线模型 B 的词错误率（WER_B），A 相对于 B 的词错误率相对降低（WERR）计算为 $\text{WERR} = (\text{WER}_B - \text{WER}_A)/\text{WER}_B$。WW 准确率以真接受率（TAR）、真拒绝率（TRR）和 F1分数衡量，其中，TAR 是正确接受的真实正例所占的比例；TRR 是正确拒绝的真实负例所占的比例；F1分数是 TAR 和 TRR 的调和平均值。我们给出相对于基线模型的 F1分数、TAR 和 TRR 的相对提升（分别记为 F1 分数相对提升（F1R）、TARR 和 TRRR）

[1]据我们所知，目前不存在包含多种用户自定义唤醒词的大规模公开数据集。

	RNN-T		C-T	
	通用	专有名称	通用	专有名称
预训练基础模型	–	–	–	–
单注意力基础模型-*128*	+0.2%	+26.73%	+0.2%	+29.08%
单注意力目录掩码	+0.2%	+28.13%	-0.2%	+29.63%
dual-attn-64	+0.2%	+28.13%	0.0%	+29.23%
双重注意力-*32*	+0.8%	+27.24%	0.0%	+29.08%
dual-attn-16	+0.8%	+27.98%	0.0%	+30.42%
dual-attn-8	+0.2%	+28.64%	-0.2%	+29.23%

表 2：相对于预训练基础模型的词错误率相对降低（WERR）； +/- 表示词错误率有所改善 / 恶化。

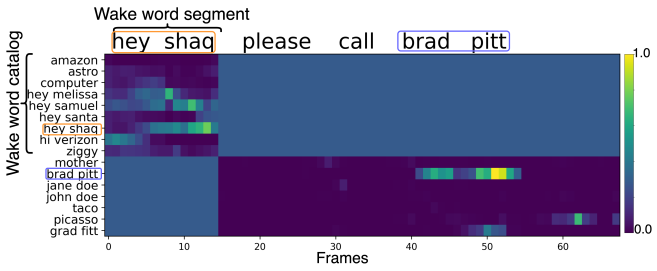

图 2：*dual-attn-64* 模型对不同目录实体的注意力权重可视化（颜色越亮表示权重越大）。每帧代表 60 毫秒。

。WERR、F1R、TARR 和 TRRR 的值越高，性能越好。负值表示性能下降。为衡量计算成本，我们报告处理引导音频所需的每帧 MHA 层浮点运算次数（FLOPs）总数 [19, 21] 。

5.1. 唤醒词准确率

随着模型学会偏向目标 WW，TAR 自然会提高。由于模型倾向于接受更多查询，TRR 可能会下降。从表 1 可以看出，与各自的双焦点基线预训练基础模型 相比，所提出的 双重注意力-λ 模型将 WW TARR 提高了最多 36%（RNN-T）和 42%（C-T）。当我们将投影维度 λ 从 64 降低到 16 时，RNN-T 的 TARR 仍提高了 29%，C-T 则提高了 39 %。当 λ=8时，RNN-T 和 C-T 的 TARR 分别提高了 8.5% 和 19%。另一方面，我们观察到 RNN-T 和 C-T 的 TR-RR 分别最多回归 14% 和 12%。事实上，对于 $\lambda = 64$,32,8，所提出的双重注意力模型 双重注意力-λ 在 TRRR 方面略微优于图 1(b) 中的基线 单注意力基线-*128* ；对于 C-T 的所有取值也同样如此。综合考虑 TAR 和 TRR，双重注意力-λ 使 RNN-T 和 C-T 的 F1 分数均最多提高了 17%。

5.2. 自动语音识别准确率

表 2 展示了所提出的双重注意力-λ （图 1(c)）相对于双焦点基线预训练基础模型 （图 1(a)）的词错误率相对降低（WERR）。在专有名称 测试集上，得益于目录屏蔽（详见第 3.2 节）以及经过专门训练、能够“聚焦于”更相关目录的多头注意力，双重注意力架构使 RNN-T 的词错误率最多降低 28.6%（而单注意力基础模型-*128*为 26.7%），使 C-T 的词错误率最多降低 30.4%（而单注意力基础模型-*128*为 29.1%）。当我们将 WW 多头注意力网络的 λ 从 64 减至 8 时，在专有名称 或通用 测试集上，相较于单注意力基础模型-*128* ，准确率没有下降。这表明双重注意力-λ 在降低计算成本的同时，不会损害 RNN-T 和 C-T 的自动语音识别准确率。

5.3. 计算成本与注意力可视化

表 1 展示了以浮点运算次数（FLOPs）衡量的计算成本。以神经偏置 [8, 14] 单注意力基础模型-*128* 为基线（图 1(b)），参数仅略有增加（从 $400K$ 增加到 $483K$，即参数总量的 1%），FLOPs（目录大小为 300 时）便从 $3.3M$（单注意力基础模型-*128*）降至 $95K$（*dual-attn-64*）。当我们将 λ 从 64 减少到 8 时，FLOPs 进一步从 95K 降至 12K。然而，当 $\lambda = 8$时，我们观察到唤醒词准确率的提升有所减少：RNN-T 和 C-T 的 F1分数相对提升分别为 1% 和 6%。值得注意的是，在引导音频片段中将偏置候选项从 300 个缩减至 6 个时，目录屏蔽在降低 FLOPs（从 $3.3M$ 降至 $190K$）方面发挥了重要作用。图 2 展示了应用目录屏蔽后小型和大型多头注意力层的注意力热图。小型多头注意力层对真正的唤醒词*hey shaq* 显示出较高的取值，同时屏蔽掉相关性较低的专有名称目录；而大型多头注意力层则偏向目标专有名称*brad pitt* ，同时屏蔽掉唤醒词目录。

6. 结论

我们提出了一种双重注意力神经转导网络，其设计灵感来自双焦点RNN-T和基于注意力的神经偏置。该架构利用唤醒词检测动态选择偏置分支，从而有效提升了专有名称和自定义唤醒词的自动语音识别准确率，同时降低了运行时浮点运算量，并减少了注意力网络的运行时延迟。

7. 参考文献

[1] A Graves、S Fern´andez、F Gomez 和 J Schmidhuber，“连接主义时序分类：使用循环神经网络标注未分段序列数据”，载于第*23*届国际机器学习会议论文集，2006年，第369–376页。[2] W Chan、N Jaitly、Q Le 和 O Vinyals，“聆听、关注与拼写：用于大词汇量会话语音识别的神经网络”，载于*2016*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2016年，第4960–4964页。[3] A Graves，“使用循环神经网络进行序列转导”，国际机器学习会议表示学习研讨会，2012年。[4] A Vaswani、N Shazeer、N Parmar、J Uszkoreit、L Jones、A N Gomez、L Kaiser 和 I Polosukhin，“注意力即一切”，神经信息处理系统进展，第30卷，2017年。[5] Z Tian、J Yi、J Tao、Y Bai 和 Z Wen，“用于端到端语音识别的自注意力转换器”，*Interspeech*，2019年。[6] C Yeh、J Mahadeokar、K Kalgaonkar、Y Wang、D Le、M Jain、K Schubert、C Fuegen 和 M L Seltzer，“Transformer转换器：采用自注意力的端到端语音识别”，*arXiv*预印本 *arXiv:1910.12977*，2019年。[7] Q Zhang、H Lu、H Sak、A Tripathi、E McDermott、S Koo 和 S Kumar，“Transformer转换器：采用Transformer编码器和RNN-T损失的可流式语音识别模型”，载于*ICASSP*，2020年。[8] F Chang、J Liu、M Radfar、A Mouchtaris、M Omologo、A Ras- trow 和 S Kunzmann，“用于语音识别的上下文感知Transformer转换器”，载于2021年IEEE自动

语音识别与理解研讨会（*ASRU*）。IEEE，2021年，第503–510页。

[9] M Radfar、R Barnwal、R V Swaminathan、F Chang、G P Strimel、N Susanj 和 A Mouchtaris，“ConvRNN-T：用于流式语音识别的卷积增强型循环神经网络转导器”，*arXiv* 预印本 *arXiv:2209.14868*，2022年。

[10] A Gulati、J Qin、C Chiu、N Parmar、Y Zhang、J Yu、W Han、S Wang、Z Zhang 等，“Conformer：用于语音识别的卷积增强型 Transformer”，*arXiv* 预印本 *arXiv:2005.08100*，2020年。

[11] B Li、A Gulati、J Yu、T N Sainath、C Chiu、A Narayanan、S Chang、R Pang、Y He 等，“用于流式自动语音识别的更优、更快速的端到端模型”，载于 *ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第5634–5638页。

[12] T N Sainath、R Prabhavalkar、S Kumar、S Lee、A Kannan、D Rybach、V Schogol、P Nguyen、B Li 等人，《不需要词典吗？评估发音词典在端到端模型中的价值》，载于*2018*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018年，第 5859–5863 页。

[13]A Gourav、J Liu、S Kalmane、G Tiwari、D Filimonov、A Rastrow 等人，《端到端语音识别系统的个性化策略》，载于*ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第 7348–7352 页。

[14] K M Sathyendra、T Muniyappa、F Chang、J Liu、J Su、G P Strimel、A Mouchtaris 和S Kunzmann，“神经转导器中的个性化语音识别上下文适配器”，载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第8537–8541页。

[15] K Greff、R K Srivastava、J Koutn´ık、B R Steunebrink 和 J Schmidhuber，“LSTM：搜索空间奥德赛”，*IEEE*神经网络与学习系统汇刊，第28卷，第10期，第2222–2232页，2016年。

[16] S Hochreiter 和 J Schmidhuber，“长短期记忆”，神经计算，第9卷，第8期，第1735–1780页，1997年。

[17] TLei、Y Zhang、S I Wang、H Dai 和 Y Artzi，“用于高度并行化循环的简单循环单元”，*arXiv* 预印本 *arXiv:1709.02755*，2017年。

[18] D Povey、G Cheng、Y Wang、K Li、H Xu、M Yarmohammadi 和 S Khudanpur，“用于深度神经网络的半正交低秩矩阵分解”，载于 *Interspeech*，2018年，第 3743–3747 页。

[19] J Macoskey、G P Strimel 和 A Rastrow，“双焦点神经自动语音识别：利用关键词检出进行推理优化”，载于*ICASSP 2021-2021 IEEE International Conference on Acous- tics, Speechand Signal Processing (ICASSP)*。IEEE，2021年，第 5999–6003 页。

[20] Y Shi, V Nagaraja, C Wu, J Mahadeokar, D Le, R Prabhavalkar, A Xiao, C Yeh, J Chan, C Fuegen, O Kalinli, 和 M L Seltzer，《动态编码器转导器：一种灵活的准确率与延迟权衡方案》，2021年。

[21] Y Xie, J Macoskey, M Radfar, F Chang, B King, A Rastrow, A Mouchtaris, 和 G P Strimel，《用于流式自动语音识别的计算成本摊销型 Transformer》，*arXiv* 预印本 *arXiv:2207.02393*，2022年。[22] J Macoskey, G P Strimel, J Su, 和 A Rastrow，《用于低延迟语音识别的摊销神经网络》，*Interspeech*，2021年。[23] S Wang, B Z Li, M Khabsa, H Fang, 和 H Ma，《Linformer：具有线性复杂度的自注意力》，*arXiv* 预印本 *arXiv:2006.04768*，2020年。[24] K Choromanski, V Likhosherstov, D Dohan, X Song, A Gane, T Sarlos, P Hawkins, J Davis, A Mohiuddin 等，《重新思考 Performer 中的注意力机制》，*arXiv* 预印本 *arXiv:2009 14794*，2020年。[25] H Soltau, H Liao, 和 H Sak，《神经语音识别器：用于大词汇量语音识别的声学到词级 LSTM 模型》，*arXiv* 预印本 *arXiv:1610.09975*，2016年。[26] Y He, T N Sainath, R Prabhavalkar, I McGraw, R Alvarez, D Zhao, D Rybach, A Kannan 等，《面向移动设备的流式端到端语音识别》，载于*ICASSP 2019-2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019年，第6381–6385页。[27] I Hubara, M Courbariaux, D Soudry, R El-Yaniv, 和 Y Bengio，《量化神经网络：以低精度权重和激活值训练神经网络》，机器学习研究期刊，第18卷，6898，2017年，1，6869，页码 –。[28] J Choi, S Venkataramani, V V Srinivasan, K Gopalakrishnan, Z Wang, 和 P Chuang，《准确高效的2位量化神经网络》，机器学习与系统会议论文集，第1卷，第348–359页，2019年。[29] H D Nguyen, A Alexandridis, 和 A Mouchtaris，《采用绝对余弦正则化的量化感知训练在自动语音识别中的应用》，载于*Interspeech*，2020年，第3366–3370页。[30] M Zhu 和 S Gupta，《剪枝还是不剪枝：探索剪枝对模型压缩的效果》，*arXiv* 预印本 *arXiv:1710.01878*，2017年。[31] V Niculae 和 M Blondel，《稀疏结构化神经注意力的正则化框架》，神经信息处理系统进展，第30卷，2017年。[32]A Roy, M Saffar, A Vaswani, 和 D Grangier，《利用路由 Transformer 实现高效的基于内容的稀疏注意力》，计算语言学协会汇刊，第9卷，第53–68页，2021年。[33] T Kudo，《子词正则化：利用多个子词候选项改进神经网络翻译模型》，*arXiv* 预印本 *arXiv:1804.10959*，2018年。[34] G Pundak, T N Sainath, R Prabhavalkar, A Kannan, 和 D Zhao，《深层上下文：端到端上下文语音识别》，载于*2018*年*IEEE*口语技术研讨会（*SLT*）。IEEE，2018年，第418–425页。[35]J Liu, R V Swaminathan, S H K Parthasarathi, C Lyu, A Mouchtaris, 和 S Kunzmann，《利用大规模教师-学生训练构建端侧声学模型》，柏林，海德堡，2021年，第413–424页，Springer-Verlag。[36] R Sennrich, B Haddow, 和 A Birch，《利用子词单元实现生僻词的神经机器翻译》，*arXiv* 预印本 *arXiv:1508.07909*，2015年。