Quartznet: Deep Automatic Speech Recognition With 1d Time-Channel Separable Convolutions

QUARTZNET：具有一维时间通道的深度自动语音识别可分卷积

*塞缪尔·克里曼*†?*斯坦尼斯拉夫·贝利亚耶夫*‡[1]*Boris Ginsburg Jocelyn Huang公司*

*Oleksii Kuchaiev Vitaly Lavrukhin Ryan Leary Jason Li Yang Zhang先生*

NVIDIA，美国

       † ‡

          伊利诺伊大学香槟分校，圣彼得堡大学经济学院

# 摘要

We propose a new end-to-end neural acoustic model for automatic speech recognition. The model is composed of multiple blocks with residual connections between them. Each block consists of one or more modules with 1D time-channel separable convolutional layers, batch normalization, and ReLU layers. It is trained with CTC loss. The proposed network achieves near state-of-the-art accuracy on LibriSpeech and Wall Street Journal, while having fewer parameters than all competing models. We also demonstrate that this model can be effectively fine-tuned on new datasets.

Index Terms— Automatic speech recognition, convolutional networks, time-channel separable convolution, depthwise separable convolution

我们提出了一种新的端到端的自动语音识别神经声学模型。该模型由多个块组成，它们之间有剩余连接。每个块由一个或多个具有1D时间信道可分离卷积层、批标准化和ReLU层的模块组成。它是用CTC损失训练的。该网络在LibriSpeech和《华尔街日报》上达到了近乎最先进的精确度，同时比所有竞争模型的参数都少。我们还证明了该模型可以在新的数据集上进行有效的微调。

*索引项*-自动语音识别，卷积网络，时间通道分离卷积，深度分离卷积

2019.10.22 arxiv, icassp 2020

# 1.    导言

近几年来，端到端（E2E）神经网络（NN）在许多自动语音识别（ASR）任务中取得了新的研究成果。这种模型用一个端到端的训练神经网络代替了传统的多分量ASR系统，直接预测字符序列，大大简化了训练、微调和推理。最新的E2E模型也有很好的精度，但这往往是以计算和内存需求越来越大的模型为代价的。

这项工作的动机是建立一个ASR模型，实现SOTA级精度，同时利用明显较少的参数和较少的计算能力。较小的模型具有多个优点：（1）训练速度更快，（2）部署在计算和内存有限的硬件上更可行，（3）推理吞吐量更高。

我们通过构建一个具有一维时间通道可分离卷积的非常深的神经网络来实现这一目标。这个新的网络在LibriSpeech[1]（见表4）和WSJ[2]（见表7）数据集上达到接近SOTA字错误率（WER），参数少于2000万，而以前的端到端ASR设计通常有超过1亿个参数。我们已经在NeMo工具箱中发布了源代码和预先训练的模型[3].[2]

# 2.    相关工作

在探索紧凑的网络结构和研究神经网络的精度和大小之间的权衡方面已经做了很多工作，如SqueezeNet[4]、ShuffleNet[5]和EfficientNet[6]。我们的方法与MobileNets[7，8]和Xception[9]直接相关，Xception使用深度可分离卷积[10，11]。每个深度可分离卷积模块由两部分组成：深度卷积层和点态卷积层。深度卷积对每个输入通道（输入深度）应用一个滤波器。逐点卷积是1×1卷积，用于创建纵深层输出的线性组合。BatchNorm和ReLU应用于两层的输出。

Hannun等人[12]将类似的方法应用于ASR。他们引入了一种具有时间深度可分离（TDS）卷积的编解码器模型。TDS模型对时频信道（×w×c）格式的数据进行操作，其中是时间步数、输入宽度和信道数。基本TDS块由一个在（T×w）上具有×1卷积的2D卷积块和一个全连接块组成，该全连接块由两个在（w·c）信道上工作的1×1逐点卷积组成，这些信道与层模层交织。相反，在我们的工作中，我们以时间通道格式（×c）操作数据，并且完全解耦卷积的时间和通道部分。TDS块具有×c2+2×（w·c）2参数，而QuartzNet模型具有×c+c2参数，这使得模型尺寸显著减小，同时仍能获得良好的性能。*T型T型w型c级k公司T型k公司k公司*

另一个非常小的ASR模型是由Han等人[13]提出的，该模型使用了多个平行的自我注意流和扩展的、因子化的、尽管不可分离的一维卷积。并行流通过使用不同的扩展率从输入中捕获多个分辨率的语音帧

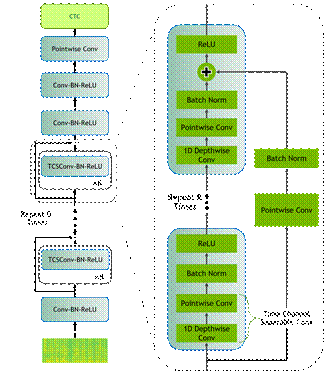


图1。QuartzNet BxR体系结构

并将各个流的结果连接到最终嵌入中。最佳模型有五个流，扩张率为1-2-3-4-5。

# 3.    模型体系结构

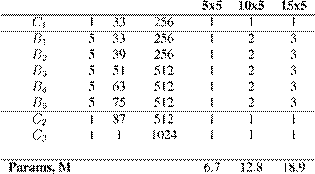
## 3.1.    基本模型

QuartzNet的设计基于Jasper[14]架构，这是一个经过连接主义时间分类（CTC）损失训练的卷积模型[15]。QuartzNet架构的主要创新之处在于，我们将一维卷积替换为一维时间通道可分离卷积，这是一种深度可分离卷积的实现。一维时间信道可分离卷积可分为一维纵深卷积层和逐点卷积层，前者具有单独但跨时间帧在每个信道上操作的核长度，后者独立但跨所有信道在每个时间帧上操作。*K公司K公司*

QuartzNet模型具有以下结构：它们从一维卷积层开始，然后是一系列块。每个块重复多次，并且在块之间具有剩余连接。每个块由相同的基本模块重复组成，包含四个层：1）带通道的纵向卷积层，2）逐点卷积，3）规范化层和4）ReLU。模型的最后一部分由表1中的三部分组成。夸兹涅特体系结构。该模型首先是一个conv层，然后是一个由5组块组成的序列。组中的块是相同的，每个块由R时间通道可分离的K大小卷积模块和C输出通道组成。每个块重复S次。模型的末尾有3个额外的conv层（）。*C级*1 *铋硅铋里K公司库特C级*1 *黑色C级*2*，C级*3*，C级*4



# 块R K C S

*C级*4 1       1       k公司标签k公司1       1       1



卷积层（）。该层具有*C级*2*，C级*3*，C级*4*C级*1

步幅为2，层的扩张为2。*C级*4

表1描述了QuartzNet-5x5、10x5和15x5型号。这些模型有五个独特的模块：-。不同的模型以不同的次数重复这些块，表示为。QuartzNet-5x5（−B2−B3−B4−B5）使每组块重复*B类*1 *B类*5*硅B类*1

1次，夸脱网-10x5（−B1−B2−B2−B5−B5）*B类*1

-重复2次，石英网-15x5（−B1−B1−B1。。。−B5−B5−B5）-重复3次。*B类*1

具有核尺寸、输入通道和输出通道的规则一维卷积层具有×cin×cout权重。时间信道可分离卷积使用×cin+cin×cout权值，分为×cin权值用于深度层和×cout用于点层。*K公司cin公司库特K公司K公司K公司cin公司*

表2。具有不同深度的QuartzNet模型在LibriSpeech上训练了300个时代，贪婪的WER（%）。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 设备清洁 | 开发其他 | 参数，M |
| 5x5个 | 5.39 | 15.69 | 6.7 |
| 10x5个 | 4.14 | 12.33 | 12.8 |
| 15倍 | 3.98 | 11.58 | 18.9 |

深度卷积独立地应用于每个信道，因此它在总权重数中所占的比例相对较小。这允许我们使用更宽的内核，大约是wav2letter[16]或Jasper[14]模型中使用的内核的3倍。我们对四种规格化进行了实验：批规格化[17]、层规格化[18]、实例规格化[19]和组规格化[20]，发现批规格化的模型训练最稳定，给出的WER最好。

## 3.2.    带群的逐点卷积

时间信道可分离卷积块的权值总数为×cin+cin×cout权值。由于它通常比它小几倍，所以大多数权值集中在逐点卷积部分。为了进一步减少参数的数目，我们探索在这一层中使用群卷积。我们还添加了一个组洗牌层来增加跨组交换[5]。*K公司K公司库特*

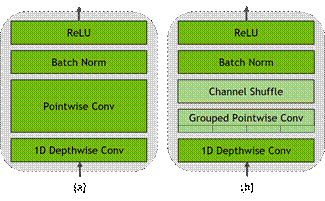


图2。（a） 时间信道可分离一维卷积模块（b）具有分组和混洗的时间信道可分离一维卷积模块

使用分组可以让我们以牺牲一些准确性为代价显著减少权重的数量。表3显示了在LibriSpeech上评估的组大小1、2和4的精度和参数数量之间的权衡。

表3。QuartzNet-15x5，分组卷积，在LibriSpeech上训练300个历元，1%

|  |  |  |  |
| --- | --- | --- | --- |
| #组 | 设备清洁 | 开发其他 | 参数，M |
| 1 | 3.98 | 11.58 | 18.9 |
| 2 | 4.29 | 12.52 | 12.1 |
| 4 | 4.51 | 13.48 | 8.70 |

# 4.    实验

我们评估了QuartzNet在LibriSpeech和WSJ数据集上的性能。此外，我们还进行了一项转移学习实验，展示了如何使用LibriSpeech和Common Voice[21]训练的QuartzNet模型可以在少量音频数据（WSJ数据集）上进行微调，以获得比从头开始训练更好的性能。

## 4.1.    藏书

我们在LibriSpeech数据集上的最佳结果是通过QuartzNet-15x5模型实现的，该模型由15个块组成，每个块有5个卷积模块（见表1）。通过将我们的网络与独立训练的语言模型（即n-gram语言模型和Transformer-XL（T-XL）[22]）相结合，我们得到了与当前SOTA相当的WER。

时间通道可分离卷积模型比规则卷积模型小得多，不易过拟合，因此在训练过程中只使用数据增强和权值衰减进行正则化。我们用SpecAugment[23]、SpecCutout和速度扰动[24]进行了实验。我们在10%的速度扰动下，结合从光谱图中随机切出小矩形的切口[25]，获得了最好的结果。使用NovoGrad优化器[26]和余弦退火学习率策略对模型进行训练。我们还发现，学习率热身有助于稳定早期训练。

表4。LibriSpeech结果，WER（%）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 加强 | 流光溢彩 | 试验  清洁其他 | | 参数，M |
| wav2letter++[27] | 速度扰动 | 康沃尔姆 | 3.26 | 10.47 | 208 |
| 拉斯维加斯[23] | 规格增大 | 循环神经网络 | 2.5 | 5.8 | 360 |
| TDS转换器[12] | 辍学+ | - | 5.36 | 15.64 | 37 |
|  | 标签平滑 | 4克 | 4.21 | 11.87 |  |
|  |  | 康沃尔姆 | 3.28 | 9.84 |  |
| MSSA[13] | 速度扰动 | 4克 | 2.93 | 8.32 | 23 |
|  |  | 4-LSTM型 | 2.20 | 5.82 |  |
| 茉莉花-10x5[14] | 规格增大+ | - | 4.32 | 11.82 | 333 |
|  | 速度扰动 | 6克 | 3.24 | 8.76 |  |
|  |  | T-XL码 | 2.84 | 7.84 |  |
| 夸兹涅特15x5 | 规格断路器+ | - | 3.90 | 11.28 | 19 |
|  | 速度扰动 | 6克 | 2.96 | 8.07 |  |
|  |  | T-XL码 | 2.69 | 7.25 |  |

在一台带有8个特斯拉V100 GPU的DGX1服务器上对400个时代的15x5型号进行培训，每台GPU的批量大小为32。为了减少内存占用和训练时间，我们使用了混合精度训练[28]。我们通过将训练扩展到32个DGX2节点的叠加，将训练时间缩短到4个多小时，这些节点具有更多的历元，并且增加了16K的全局批次（见表5）。[3]

表5。QuartzNet-15x5：关于LibriSpeech的大批量培训，培训时间（小时）和贪婪的WER（%）。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 批量 | 历元 | 时间，h | 开发清洁其他 | 试验  清洁其他 |
| 256 | 400 | 122 | 3.83        11.08 | 3.90      11.28 |
| 16千 | 1500 | 4.3 | 3.71        10.78 | 4.04      11.06 |

## 4.2.    《华尔街日报》

我们在《华尔街日报》数据集的开放词汇任务中训练了一个较小的QuartzNet-5x3模型[2]。我们使用train-si284集进行培训，nov93 dev进行验证，nov92 eval进行测试。使用NovoGrad optimizer（=0.95，=0.5）对QuartzNet-5x3模型（见表6）进行1200个阶段的训练，每个GPU的批大小为32，数据增加（10%速度扰动，SpecCutout）和0.2的退出，其中学习率预热1000步，学习率为0.05，权重衰减为0.001。*β*1 *β*2

我们在推理过程中使用了两种外部语言模型：4gram（波束大小=2048，alpha=3.5，beta=1.5）和TransformerXL（T-XL）。这两种语言模型都是使用《华尔街日报》的官方LM数据构建的。

表6。《华尔街日报》的QuartzNet-5x3。该模型与QuartzNet-15x5具有相同的层，但中间部分仅由五个块组成，每个块重复三次。*C级*1*，C级*2*，C级*3*，C级*4

|  |  |  |  |
| --- | --- | --- | --- |
| 方块 | 右 | K公司 | C级 |
| *C级*4 | 1 | 1 | 克拉贝尔斯克 |

我们使用以下基于标准语音特征训练的端到端模型[4]进行比较：

1） RNN-CTC[30]-CTC模型，5层双向LSTM，每层500个单元；

2） ResCNN LAS[31]：具有深剩余convLSTM编码器和LSTM解码器+标签的拼写模型

平滑的；

3） Wav2Letter++[29]—具有1D卷积层和实例范数的CTC模型。

表7。QuartzNet-5x3，华尔街日报，华盛顿（%）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 流光溢彩 | 93试验 | 92评估 | 参数，M |
| RNN-CTC[30] | 3克 | - | 8.7 | 26.5 |
| ResCNN LAS公司[31] | 3克 | 9.7 | 6.7 | 6.6 |
| Wav2Letter++[29] | 4克 | 9.5 | 5.6 | 17 |
|  | 康沃尔姆 | 7.5 | 4.1 |  |
| 石英网-5x3 | 4克 | 8.1 | 5.8 | 6.4 |
|  | T-XL码 | 7.0 | 4.5 |  |

## 4.3.    迁移学习

由于我们的模型比其他模型要小，所以我们感兴趣的是它能在多大程度上推广到各种来源的数据，特别是当目标语音的数量比训练数据少得多时。我们的设置包括在LibriSpeech[1]和Mozilla的Common Voice[21][5]数据集的组合上训练QuartzNet 15x5，然后在80小时的《华尔街日报》数据集上微调这个训练模型。表8显示了微调前LibriSpeech上实现的WER和微调后WSJ上的结果。

表8。QuartzNet15x5迁移学习。该模型是对librispeechtrain和Mozillas通用语音数据集进行预训练的，并在80小时的WSJ数据集上进行了微调。该模型在LibriSpeech和WSJ上进行了评估，WER（%）。

|  |  |  |  |
| --- | --- | --- | --- |
| 流光溢彩 | LibriSpeech测试clean测试other | 华尔街日报  93试验 | 92评估 |
| - | 4.19            10.98 | 8.97 | 6.37 |
| 4克 | 3.21             8.04 | 5.57 | 3.51 |
| T-XL码 | 2.96             7.53 | 4.82 | 2.99 |

# 5.    结论和未来方向

提出了一种基于一维时间通道可分离卷积层的深度神经网络的端到端语音识别模型。该模型在《华尔街日报》和LibriSpeech上的表现接近最先进水平，同时显著小于所有其他具有类似精确度的端到端系统。这种小巧的模型为移动和嵌入式设备上的语音识别开辟了新的可能性。

这项工作描述了一个基于CTC的模型，但我们正在探索QuartzNet编码器与基于注意的解码器相结合的模型。

# 6.    参考文献

[1] V.Panayotov、G.Chen、D.Povey和S.Khudanpur，“Librispeech:基于公共领域有声图书的asr语料库”，ICASSP，2015年，第5206-5210页。

[2] D.B.Paul和J.M.Baker，“基于华尔街日报的企业社会责任语料库的设计”，演讲与自然语言研讨会论文集。ACL，1992年，第357362页。

[3] O.Kuchaiev、J.Li、H.Nguyen、O.Hrinchuk、R.Leary、，

B、 Ginsburg，S.Kriman，S.Beliaev，V.Lavrukhin，J.Cook，et al.，“Nemo:使用神经模块构建ai应用程序的工具箱，”附件十四：1909.09577, 2019.

[4] F.N.Iandola、M.W.Moskewicz、K.Ashraf、S.Han、W.J.Dally和K.Keutzer，“挤压网：减少50倍参数和1mb模型的Alexnet级精度，”附件十四：1602.07360, 2016.*<*

[5] 张琰，周琰，林琰，孙琰，“Shufflenet:一种非常有效的移动设备卷积神经网络”，CVPR，2018年。

[6] M.Tan和Q.V.Le，“Efficientnet:卷积神经网络模型缩放的再思考”，ICML，2019年。

[7] A.G.Howard、M.Zhu、B.Chen、D.Kalenichenko、W.Wang、T.Weyand、M.Andreetto和H.Adam，

“MOBILENET：用于移动视觉应用的高效卷积神经网络。”附件十四：1704.04861, 2017.

[8] M.Sandler、A.Howard、M.Zhu、A.Zhmoginov和L.Chen，“Mobilenetv2：反向残差和线性瓶颈”，CVPR，2018年。

[9] F.Chollet，“例外：深度可分离卷积的深度学习”，CVPR，2017年。

[10] L Sifre和S.Mallat，“用于纹理分类的刚性运动散射”附件十四：1403.1687, 2014.

[11] V.Vanhoucke，“学习规模视觉表征”，ICLR，2014年。

[12] A.Hannun、A.Lee、Q.Xu和R.Collobert，

“具有时间深度可分离卷积的序列到序列语音识别”，过程。Interspeech 2019，2019，第3785-3789页。

[13] K.J.Han，R.Prieto，K.Wu，和T.Ma，“基于扩展一维卷积的多流自我注意的语音识别技术现状，”附件十四：1910.00716, 2019.

[14] J.Li、V.Lavrukhin、B.Ginsburg、R.Leary、O.Kuchaiev、J.M.Cohen、H.Nguyen和R.T.Gadde，“Jasper:端到端卷积神经声学模型，”附件十四：1904.03288, 2019.

[15] A.Graves、S.Fernandez、F.Gomez和J.Schmidhuber，“连接主义时间分类：用递归神经网络标记未分段序列数据”，ICML，2006年。

[16] V.Liptchinsky，G.Synnaeve和R.Collobert，“基于字母的语音识别与门控网络，”附件十四：1712.09444, 2017.

[17] S.Ioffe和C.Szegedy，“批量标准化：通过减少内部协变量变化加速深层网络训练，”附件十四：1502.03167, 2015.

[18] J.Ba、J.R.Kiros和G.E.Hinton，“层标准化”附件十四：1607.06450, 2016.

[19] D.Ulyanov、A.Vedaldi和V.Lempitsky，“实例规范化：快速样式化的缺失成分，”附件十四：1607.08022, 2016.

[20] 吴彦宏，“群体规范化”，《计算机科学讲稿》，2018年第319页。

[21]Mozilla，“普通声音”https://voice。mozilla.org/en, 2019.

[22]Z.Dai、Z.Yang、Y.Yang、J.G.Carbonell、Q.V.Le和R.Salakhutdinov，“变形金刚xl：超出固定长度语境的注意力语言模型，”附件十四：1901.02860, 2019.

[23]D.Park，W.Chan，Y.Zhang，C.Chiu，B.Zoph，E.D.Cubuk和Q.V.Le，“Specaugment:一种用于自动语音识别的简单数据增强方法，”附件十四：1904.08779, 2019.

[24]T.Ko、V.Peddinti、D.Povey和S.Khudanpur，“语音识别的音频增强”，Interspeech，2015年。

[25]T.Devries和G.W.Taylor，“带截断的卷积神经网络的改进正则化，”附件十四：1708.04552, 2017.

[26]B.Ginsburg，P.Castonguay，O.Hrinchuk，O.Kuchaiev，V.Lavrukhin，R.Leary，J.Li，H.Nguyen，和Cohen J.M.，“深层网络训练的分层自适应矩随机梯度法，”附件十四：1905.11286, 2019.

[27]N.Zeghidour、Q.Xu、V.Liptchinsky、N.Usunier、G.Synnaeve和R.Collobert，“完全卷积语音识别，”附件十四：1812.06864, 2018.

[28]P.Micikevicius，S.Narang，J.Alben等人，《混合精度训练》附件十四：1710.03740, 2017.

[29]N.Zeghidour、N.Usunier、G.Synnave、R.Collobert和E.Dupoux，“从原始波形进行端到端语音识别，”附件十四：1806.07098, 2018.

[30]A.Graves和N.Jaitly，“使用递归神经网络实现端到端语音识别”，ICML，2014年。

[31]J.Chorowski和N.Jaitly，“在序列到序列模型中实现更好的解码和语言模型集成，”附件十四：1612.02695, 2016.

[[1]](" \l "_ftnref1" \o ")这项工作是在S.Kriman和S.Beliaev在NVIDIA时进行的

[[2]](" \l "_ftnref2" \o ")https://github.com/NVIDIA/NeMo

[[3]](" \l "_ftnref3" \o ")训练时间更长（3000学时），测试清洁率提高到3.87%，其他测试清洁率提高到10.61%。

[[4]](" \l "_ftnref4" \o ")注意，具有可训练前端和convLM的wav2letter++[29]具有更好的WER：对于nov93测试为6.8%，对于nov92-dev为3.5%。在这里，我们只考虑具有标准mel滤波器组前端的模型。

[[5]](" \l "_ftnref5" \o ")我们用的是经过验证的普通语音，ver。en 1087h 2019-06-12。