# End-to-End Radio Traffic Sequence Recognition

# with Deep Recurrent Neural Networks

摘要 - 我们研究序列机器学习技术对原始无线电信号时间序列数据。 通过应用深度循环神经网络，我们学习在不使用专家解调算法的情况下区分几种应用层次特征的包络调制。 我们显示复杂的协议序列可以学习和用于使用这种方法的分类和生成任务。

关键词：机器学习，软件无线电，协议识别，循环神经网络，LSTM，协议学习，交通分类，认知无线电，深度学习

1. 引言

传输分析和深度包检测是确保有线和无线网络中的服务质量（QoS），网络安全性以及适当计费和路由的重要工具。由于这些原因，当今存在系统和算法以区分不同的协议和应用，但是新的方法提供了极大的改进。

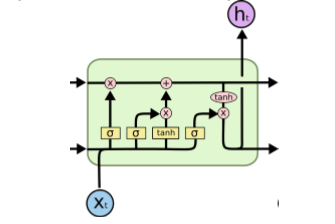
当前技术通常涉及使用许多脆弱的协议解析器，其必须解析组合地大量不同的网络和应用协议，将解析能力限制为其解析器已被手动实现的已知协议，潜在地具有解析器实现漏洞或其他缺陷。在协议解析的基础上，无线信号还需要检测，同步，均衡，符号到位解映射和纠错解码。这些算法中的每一个都增加了开发中的最终解决方案的复杂性，实现成本，漏洞潜力和协议规范。

通过应用机器学习来解释直接携带高级协议的调制无线电信号的任务，我们证明我们可以成功地将这种解映射和解释过程作为机器学习框架内的学习数据映射过程。 这样做，我们形成一个模型，可以学习推广和作出新的不可见的调制和协议的决定。 我们构建了一个不容易出现基于解析器的安全漏洞的模型，并且我们形成一个模型，它不会对开发产生成本和复杂性，而是随着所实现的特定协议的数量而扩展，因为它们是使用一个一般化的模型从数据集导出的。 我们先前已经证明[13]，这种使用深层神经网络学习低层调制的无线电鉴别任务的方法可以是高度有效的，但在这项工作中，我们表明这种潜力还跨越到更高层的业务类型

A．自然语言中的循环网络

循序渐进的神经网络方法时间序列学习不是一个新的东西，他们已经非常成功的近年来在自然语言翻译，自然语言嵌入任务信息检索或映射，自动语音识别领域等应用程序。 在这些中的每一个中，使用诸如长期短期记忆[3]（LSTM）的循环神经网络来编码令牌序列，字符或音素。 基于简单复现单元，LSTM和门控循环单元（GRU）的循环神经网络[5]都被广泛使用，它们的序列学习能力是相当令人印象深刻的，因为在一个简单的任务中可见的呈现自然语言文本、字符到这样的系统[9]。 LSTM基本神经元单元的传递函数和结构如图1所示。

图。 1.基本LSTM单元传递函数图[10]



许多应用程序也成功地使用了基于嵌入，从序列到离散类的映射[4]和许多其他序列相关任务的序列域之间的翻译（例如不同的语言）[7]。 LSTM已经在这一领域被广泛使用，作为一个高度成功的循环网络原语，但并不代表唯一或最少的计算昂贵的选择，因为简单的RNN和GRU都被广泛使用。 在语音和文本建模领域中，用于利用隐马尔可夫模型（HMM）进行序列预测的现有技术方法已经在很大程度上被这种基于RNN的建模方法所取代

B.无线电序列动机的背景

在无线电通信中，无线电发射机和接收机包括多个序列到序列翻译程序[1]。 这些在协议数据比特序列，前向纠错编码比特，随机化和白化比特，成帧比特，以及最终到直接穿过无线电信道的调制和编码符号之间转换。

我们可以通过将数据呈现给适当的机器学习架构来尝试学习这些序列翻译映射，而不是为其中的每一个实现专家算法。 理想地，学习直接从呈现给学习算法的示例数据序列学习消费无线电符号，处理空闲传输模式，数据成帧模式和数据有效载荷模式，而不是依赖于任何数量的专家编码和解码算法描述

II. 监督交通类型学习

在我们的网络中，我们训练一个基于多层LSTM的序列学习器网络在我们调制的无线电信号的一连串切片执行监督分类到11种不同的协议业务类之一。

我们的架构，其中LSTM单元直接操作复杂的基带I / Q信号表示，其中I和Q分量被视为独立和独立的通道，其次是完全连接的层使用线性整流器和softmax激活在最终输出层。

1. 数据集生成

我们生成包括通过无线链路发送的若干不同的公共网络应用协议的数据集。 我们首先捕获对应于感兴趣的网络活动行为的网络业务。 所选应用程序显示在下表中，包括来自多媒体流，典型浏览和文件下载，软件开发和系统管理任务的流量。

•流式传输

◦视频流（通过ABC视频）

◦视频流（通过Youtube）

◦音乐流（通过Spotify）

•实用程序

◦获取

◦ICMP响应测试（Ping）

◦版本控制（git）

•下载/浏览

◦Bit-Torrent

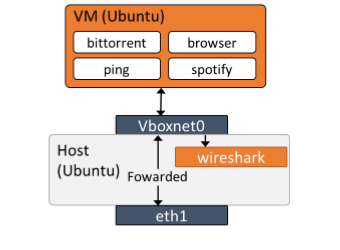
◦Web浏览

◦文件传输协议（FTP）

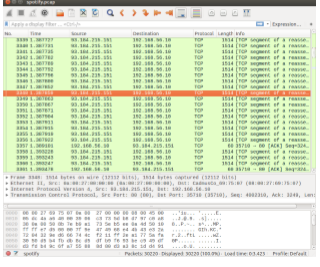
◦HTTP下载

Wireshark和tcpdump用于捕获网络流量并生成后面用于训练和分类的每个网络协议的痕迹。 虽然这些实用程序可以用于目标特定的网络流量（即记录特定的端口/连接/协议），一个更一般的捕获提供了额外的行为数据，将有用的培训和识别，即背景业务和相关业务，如域名称查找也在发生。 这提供了更复杂的异构网络流量看起来像一个更真实的图片，而不是一个设置，可能已明确试图捕获孤立的网络流量只使用一个单一的协议。 这也是一个挑战，因为感兴趣的流量不是在数据集中的任何时候都发生，导致一些时间窗口不包含关于感兴趣的分类任务的信息

图。 2.使用隔离虚拟机进行数据包捕获设置

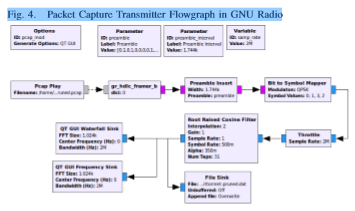


图。 3. Spotify会话的Wireshark数据包捕获



一旦捕获了网络流量，生成数据集的下一步就是通过连续调制调制解调器重放流量，并记录IQ样本数据以形成数据集。我们使用的发射机是使用高级数据链路控制（HDLC）进行成帧的GNU无线电[2]流图，以及用于无任何纠错或随机化的调制的正交相移键控（QPSK）（如图II- A.使用gr-pcap outof-tree模块（OOT）来对每个分组重复具有适当定时信息的分组捕获。消息由HDLC成帧器成帧到恒定速率HDLC比特流中，HDLC成帧器不断地发送空闲如果没有输入数据可用，这使得分类任务很有趣，因为某些东西总是被发送，分类器不能简单地学习功率包络以识别在突发CSMA / CD系统中可能的协议定时。周期性地每1744位允许一个接收器的PHY同步，一个油门块用于施加所需的波特率通过使用该油门选择不同的波特率，PCAP文件中的恒定数据速率在链路中从高到低变化，从而影响空闲和非空闲交通。我们选择大约1MBit / s的比特率，这在所有不同协议记录上平均的链路利用率上提供了合理的中间地带。然后将比特映射到QPSK符号，通过根升余弦滤波器，然后“传输”。在这里，我们只需将IQ符号保存到要用于训练和测试的数据文件

图。 4. GNU收音机中的分组捕获发射机流程图



B.模型数据提取

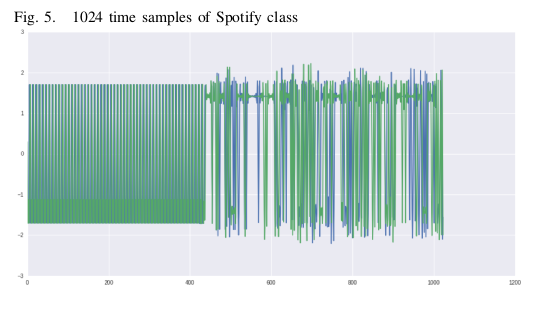
为了在这个大的时间序列上训练模型，我们必须选择如何向RNN模型展示数据。有这里有两个考虑，第一个是如何将序列按照时间划分给序列模型，第二是如何将宏观尺度上的数据划分为训练和测试数据。

在第一种情况下，考虑一个时间序列x（n），这里我们从线性序列建立例子。在这种情况下，我们按照步长M，提取N个大小L的窗口，形成一个三维实例向量。在这种情况下，尺寸按照大小为N\*2\*L的真值张量的形式表现，其中第一尺寸在窗口上方，第二尺寸在I / Q尺寸上，第三尺寸在每个窗口内随时间变化。然后，由原始时间序列中的L +（N-1）\* M个复样本形成每个张量实例。 因为任何任务都不知道最佳切片，所以我们将使用这个符号来指代我们在训练期间测试的输入张量形状。 我们使用python numpy和摄取张量数据执行这个切片到Keras [8]和Theano [16]用于模型训练。

对于我们的监督网络任务分类模型，我们对每个示例使用one-hot目标标签，其中1xK输出值都为零，除非目标索引k是示例类，其中它设置为1.0。 这通常与SoftMAX输出激活层一起使用，以帮助训练类预测，我们用相同方法在这里使用它。

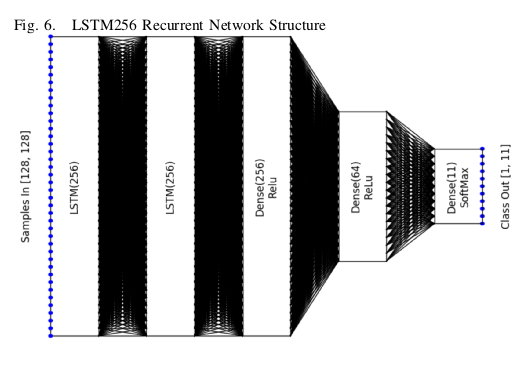
在生成回归模型的情况下，我们使用相同的N个时间步长窗口作为输出的输入张量数据，而使用实值样本的第N + 1个时间步长作为我们的示例目标。

最后，作为预处理步骤，我们考虑是否从我们的样本表示中输入I / Q样本，R /θ样本，仅R或仅θ，其中R，θ表示I / Q样本的极性形式 。 我们这样做是为了考虑捕获同相和正交分量之间的循环关系，当将它们当作实值分离通道时，它们被丢弃。



C、判别模型训练

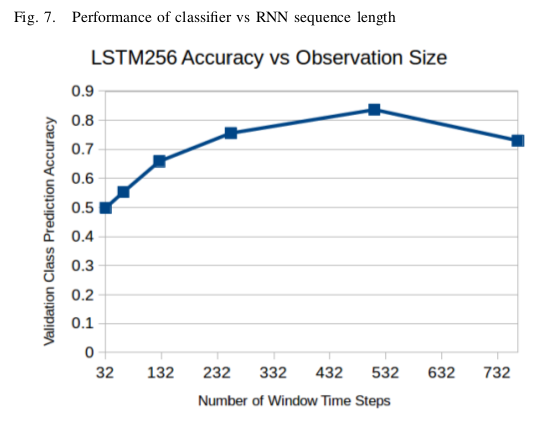
在我们的判别分类器中，我们训练网络以决定哪个traffic由无线网络信号携带。这是一个K类监督学习问题，其寻求选择哪些K个通信类型当前是关注的主要网络通信行为。 我们实现一个CLDNN [11]，或由一系列卷积层，LSTM层，最后全连接层的序列形成的网络，LSTM后面是全连接层。 后者的架构如6所示。



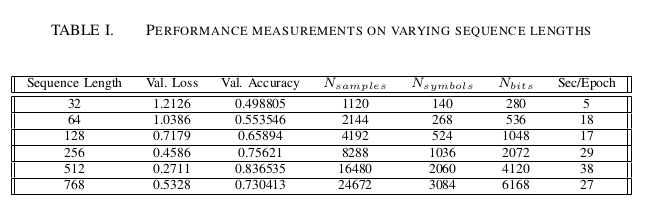
由于这个领域中存在很少的基准数据集，我们在radioml.com上发布了我们的数据集，并且完全描述了用于比较的方法。 我们利用2层LSTM，然后是两个完全连接的层，以使用每层之间0.5的丢失来执行类估计。

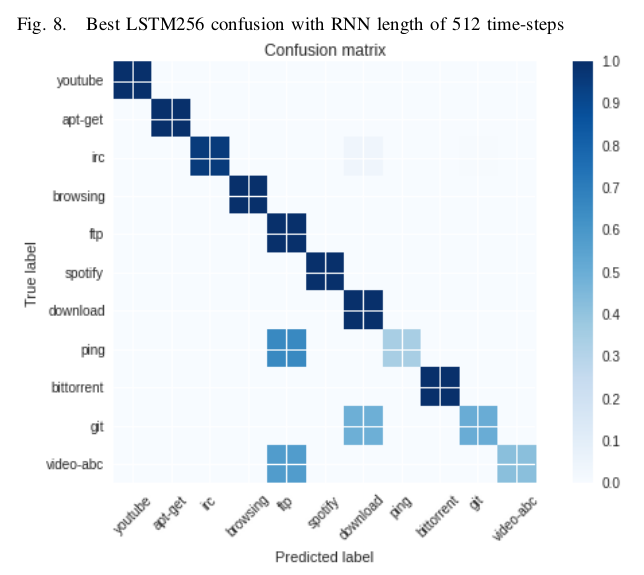
1）重叠的底噪训练：我们从数据集的最简单的情况开始，以确保学习能力，实际上存在于我们提出的模型中。 这里，我们使用具有非常高的信噪比（SNR）的原始调制信号，而没有引入频率或采样定时偏移的影响。 此外，我们的示例每个128个符号对齐，这是我们的初始训练配置的副产品，但对于网络使得任务显着更容易。 最后，在这个训练体系中，我们不重复使用训练和测试集之间的示例，但是我们允许训练和测试集之间的重叠。 也就是说，某些数据窗口可以存在于不同的偏移量。

在这种情况下，我们选择Nx2x128的输入张量形状，其中我们在N个值的范围内搜索以确定用于性能的时间步长的最佳数量，如图7所示

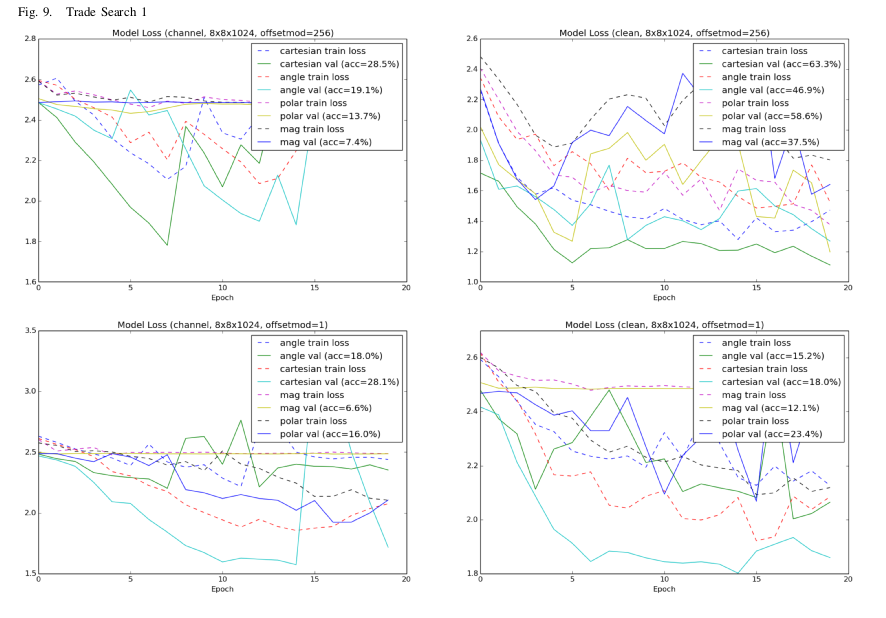


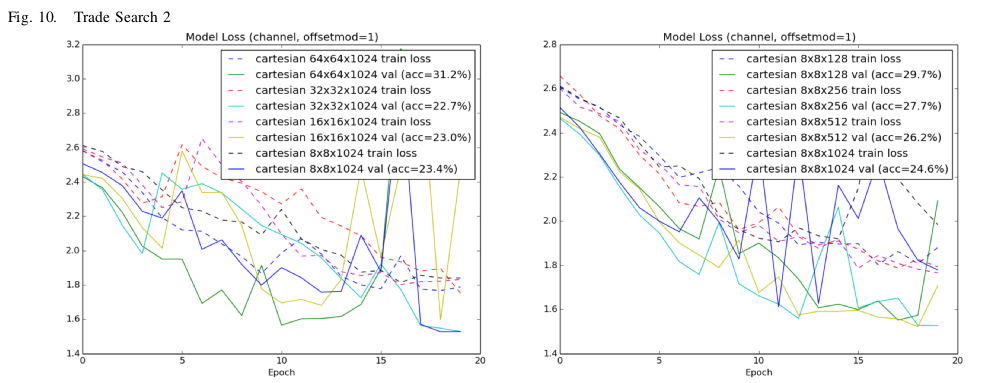
当在LSTM中使用512个时间步长的2x128样本时，我们发现我们获得了最佳精度性能。 不同序列长度评估的细节在表1中详细描述。对于512的序列长度，我们在图8中获得了我们最佳性能分类精度的混淆矩阵，总体精度为84％，具有大多数对角线的精确分类性能，除了几个有点混乱的类。 重要的是注意，一些错误是数据集中固有的，然而，由于数据中的任何给定时间窗口可以具有或不具有表示感兴趣的流量行为的分组，我们在这里查看相当小的时间窗口。



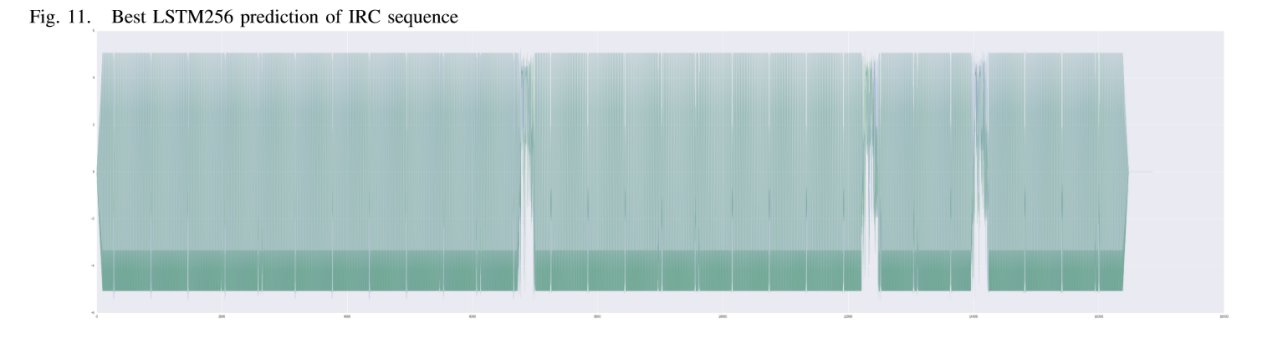


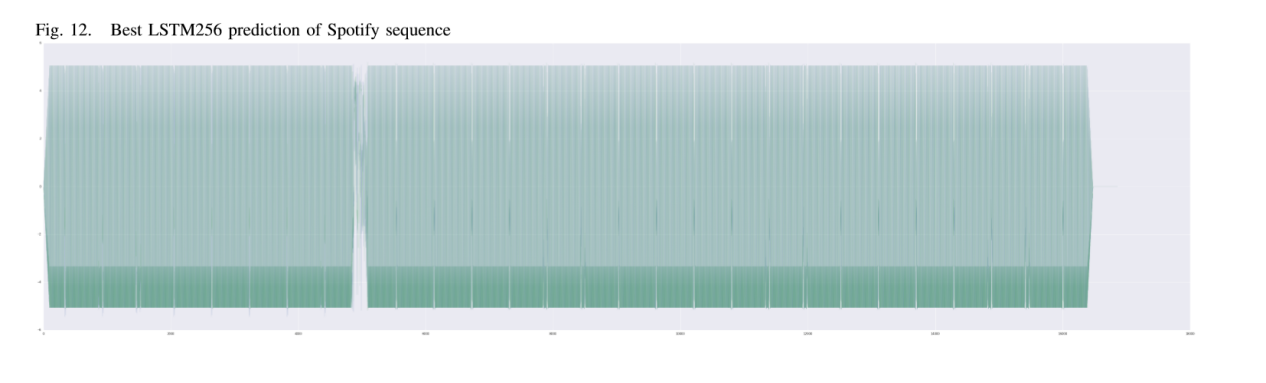
1. 通道训练和无重叠：为了完全区分训练和测试集，我们需要完全消除从每个训练集抽取的例子之间的重叠。在本节中，我们将原始时间序列分为250,000个样本的硬分区，每个样本分配给训练或测试，然后在这些边界内绘制示例用于训练和测试。这确保我们正在学习可概括的序列特征，而不是可以用于回忆一个类的特定窗口示例。此外，我们考虑两种形式的输入信号，一种称为“clean”，其表示没有频率偏移或定时偏移的相同的高SNR信号，以及一种称“channel”，其应用加性高斯白噪声，随机频率偏移和随机定时偏移。后者具有大约20dB的信噪比，仍然相当高，但是相当逼真远低于干净版本。最后，我们放松开始对128个符号对齐偏移的影响，我们考虑“偏移模数”的两个值：1，其中我们可以在任何偏移开始，256，其中开始128符号（256-样本）对齐，考虑这种假设对分类任务的附加效应。这些假设使得任务显得更加困难。  
        由于存在许多架构来评估这个任务，并且对它们进行搜索是一个费力和计算时间密集的任务，我们引入了一种仍然处于早期形式的工具，称为dist hyperas [18]，以帮助在多个不同GPU实例的架构内搜索最佳超参数。  
        在我们的第一次实验中，我们使用8x8x1024输入张量形状评估输入表示、通道效应和步幅对我们的模型的性能。所测试的每个模型的损耗曲线和最终精度如图9所示。在这种情况下，我们使用笛卡尔I / Q输入表示获得了一个通道的最佳性能，当考虑实际信道时，偏移模数似乎没有很大的影响。 （它对干净信号的性能有更大的影响）。一个信道的最佳效果约为28.5％，而没有信道的频率约为63.3％。

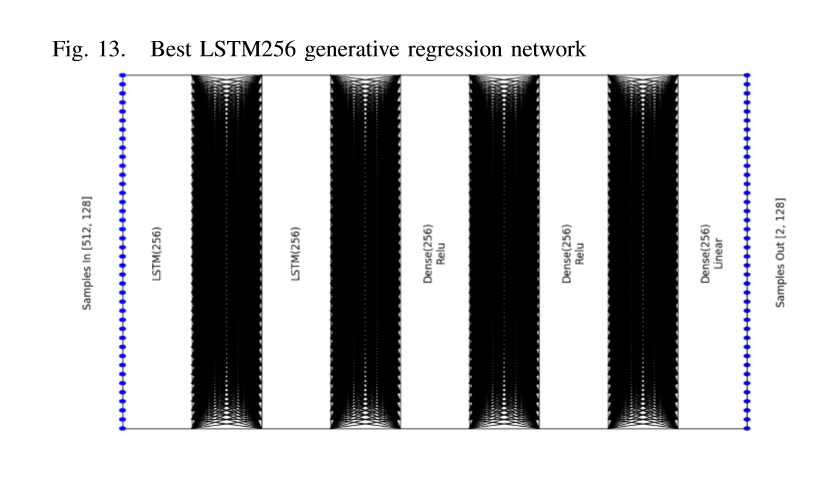
  
     在我们的第二次实验中，我们只考虑笛卡尔I / Q输入和offset\_mod 1.在这种情况下，我们交换序列长度（时间步长数）与所使用的窗口的大小和步幅。结果示于图10中。在这种情况下，我们似乎获得了具有L = 64的窗口大小和N = 1024的序列长度的最佳性能，在现实的信道和采样条件下给出约31.2％的分类精度。我们仍在调查更大的模型和额外的超参数组合，但大型LSTM架构需要大量的内存占用，接近/受到我们的Titan X的限制，并且训练需要大量的计算时间。在未来，我们希望找到更聪明的方式在这些限制中实现。

  
     我们认为，从架构搜索中可以获得一些额外的性能，但也可以从下面描述的改进的基础技术中获益，以帮助应对信道变化。

D. 生成模型训练







我们使用图13所示的简单的第一阶生成模型，其能预测下一个时间步长的窗口。给定N个先前的时间步长窗口作为回归任务。 我们使用线性输出层激活函数的实际输出样本值的平均方差（MSE）来训练网络参数。

在图11中，我们显示出了调制的无线电数据信号，其中前半部分是来自IRC序列的ground truth绝对真实示例，并且使用这里描述的复现神经网络模型从而进行模型预测，生成后半部分的数据。

通过视觉比较预测样本和基准样本的样本，我们可以看到它正确地预测了HDLC空闲模式，等宽成帧模式以及在生成样本数据区域内发生的数据突发的一些相似性。 这对于完全仅基于少数可用训练数据序列的模型训练工作而言是令人印象深刻的，因为它没有用到调制，前导码结构，HDLC协议或顶部应用的专业知识。

在图12中，我们为Spotify音乐流显示出了类似的序列，其中前一半是实例，而后一半是从模型中生成的。 在这种情况下，我们可以看到生成的HDLC空闲模式，等距帧前导码，但没有发生额外的数据突发。

在未来的工作中，我们计划使用生成式对抗网络（GAN）架构[6]方法通过引入一个批评/鉴别器模型来改善我们的生成模型。 这种技术已经在图像域中很好的证明，通过引入一个回馈循环对于一个真实的/可产生的鉴别器，以形成加强学习过程，通过该加强学习过程，两个模型相互改善了更现实的生成输出的结果。

我们认为的对于未来的工作非常适用的两个非常有前途的时间序列生成的方法在[17]和[12]中给出。

III结论

我们在这项工作中表明，循环神经网络模型可以容易地用于高级无线电协议序列识别从预解调的无线电信号数据中，用于标记和生成计算任务

我们已经证明了在理想条件下（高SNR，没有采样率偏移频率）工作得很好的两种任务的基线性能。 然而，引入逼真的通道效应使得任务显的更加困难，并明显降低了模型性能。

在采样率偏移，频率偏移和信道延迟扩展中通过无线信道引入的序列的信道变化使得来自原始数据的学习序列模型困难，但是存在许多可以帮助缓解这个问题的想法，诸如允许注意模型 协调信道效应[15]和在训练期间引入重信道正则化[14]

这些结果对于许多认知和传统无线电应用的序列和协议识别学习具有显着影响。 通过为数据和经验驱动的协议识别学习提供鲁棒的方法，许多未来的无线电分配，QoS，调度和决策算法可以做出关于如何在更大的资源受限的多用户认知无线电联网系统。

致谢

作者感谢弗吉尼亚理工学院和州立大学的布莱德利电气和计算机工程系，休谟中心和DARPA都对这项工作给予了慷ous的支持。 这项研究是由国防高级研究计划署（DARPA）MTO办公室授予HR0011-16-1-0002资助下开发的。 所表达的观点，意见和/或发现的是作者的观点，意见和/或发现，不应被解释为代表国防部或美国政府的官方观点或政策。