****

本科毕业设计（论文）

**基于深度学习的鼓转录算法**

|  |  |
| --- | --- |
| **学院**  **专业**  **学生姓名**  **学生学号**  **指导教师**  **提交日期** | **软件学院** |
| **软件工程** |
| **印昊超** |
| **201830680312** |
| **朱金辉** |
| **2022 年 5 月 18 日** |

# 摘 要

音乐转录是音乐信息检索领域的核心任务。抄写乐曲中的鼓音轨是一项明确定义的子任务。鼓声音轨的象征性表现包含了许多关于乐曲的有用信息，如节拍、节奏，以及各种风格和流派的线索。本文评估了多种利用递归与卷积神经网络进行鼓点转录的新方法。本文发现，递归神经网络可以被训练来根据打击乐器的声音的一般性质来识别它们的起始点，并且其效果相比于卷积神经网络更加优秀。本文还使用使用两个著名的数据集来比较和评估递归神经网络的不同架构。将结果与同一数据集上卷积神经网络最先进方法的结果进行比较。此外，使用第二个独立的数据集证明了网络的泛化能力。实验结果表明，所使用的三种不同递归神经网络架构均能够拥有较好的准确率，泛化能力和训练效率，具有很广泛的应用前景，值得引起重视。最后，本文基于已训练好的神经网络开发了一个实时转录模块来进行实时的鼓转录，其能够达到的精度与已训练的最佳的循环神经网络在测试集上达到的精度类似，取得了良好的效果。

关键词：鼓转录，循环神经网络，卷积神经网络，实时转录

# Abstract

Music transcription is a core task in the field of music information retrieval.Copbing the drum tracks in a piece is a well-defined sub-task.The symbolic representation of the drum track contains many useful information about the piece, such as beats, beats, and cues in various styles and genres.This paper evaluates multiple new methods for drum transcription using recurrence with convolutional neural networks.We find that recurrent neural networks can be trained to identify percussion instrument starting points based on the general properties of their sounds, and have better effects compared to convolutional neural networks.We also use the use of two well-known datasets to compare and evaluate the different architectures of recurrent neural networks.The results are compared with those of the state-of-the-art methods for convolutional neural networks on the same dataset.Furthermore, the generalization ability of the network was demonstrated using a second independent dataset.The experimental results show that the three different recurrent neural network architectures used can have good accuracy, generalization ability and training efficiency, which have a wide application prospects and deserve attention.Finally, we developed a real-time transcription module based on the trained neural network to perform real-time drum transcription, with an accuracy similar to that achieved by the best trained recurrent neural network on the test set, with good results.

**Keywords**: Drum transcription；Convolutional Neural Network；Recurrent neural network；Real-time transcription

# 目 录

目录

[摘 要 2](#_Toc17174)

[Abstract 3](#_Toc11914)

[目 录 4](#_Toc28807)

[第一章 绪论 1](#_Toc30135)

[1.1 引言 1](#_Toc15393)

[1.2 研究背景 1](#_Toc162)

[1.3 研究现状 2](#_Toc24273)

[1.4主要工作 3](#_Toc10339)

[1.4 论文结构 4](#_Toc10686)

[第二章 卷积与循环神经网络的基础知识 5](#_Toc11306)

[2.1 卷积神经网络的网络结构 5](#_Toc28507)

[2.1.1卷积层 5](#_Toc27804)

[2.2.2池化层 5](#_Toc28106)

[2.2.3全连接层 6](#_Toc22298)

[2.2 循环神经网络的网络结构 6](#_Toc3471)

[2.2.1循环神经网络隐藏层的基本结构 7](#_Toc10292)

[2.3神经网络的训练模式 8](#_Toc7847)

[2.4本章小结 9](#_Toc9950)

[第三章 基于神经网络的鼓转录算法设计 10](#_Toc20342)

[3.1 特征提取 10](#_Toc17953)

[3.1.1运用短时傅里叶变换的特征提取 11](#_Toc13371)

[3.1.2运用梅尔频谱的特征提取 11](#_Toc18767)

[3.2CNN网络的设计 11](#_Toc10969)

[3.2 RNN网络的设计 12](#_Toc10008)

[3.2.1门控循环单元(GRU) 13](#_Toc1595)

[3.2.2长短时记忆网络(LSTM) 15](#_Toc2531)

[3.2.3双向循环神经网络（Bidirectional RNN) 16](#_Toc1446)

[3.3 峰值选择算法 16](#_Toc93)

[3.4训练过程 17](#_Toc32681)

[3.5代码实现 17](#_Toc18208)

[第四章 实时转录模块与网络传输模块的开发 18](#_Toc9620)

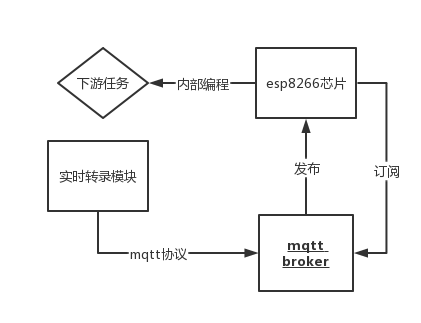
[4.1 实时转录模块的开发 18](#_Toc7385)

[4.1.1实时转录模块的功能 18](#_Toc11163)

[4.1.2实时转录模块的设计 18](#_Toc23639)

[4.2网络传输模块的开发 18](#_Toc17162)

[4.2.1网络传输模块的功能 18](#_Toc13421)

[4.2.2网络传输模块的整体框架 19](#_Toc32575)

[4.2.3网络传输模块的具体实现 19](#_Toc21390)

[第五章 实验结果与分析 20](#_Toc17214)

[5.1实验数据集 20](#_Toc19846)

[5.2结果分析 20](#_Toc1308)

[总结 21](#_Toc29630)

[1. 论文工作总结 21](#_Toc23962)

[2. 工作展望 21](#_Toc23793)

[参考文献 22](#_Toc19866)

[致谢 24](#_Toc12387)

# 绪论

## 引言

音乐转录(Automatic music transcription,AMT)是音乐信息检索领域的核心任务。抄写乐曲中的鼓音轨是一项明确定义的子任务。它是音乐分析领域中的一项十分重要同时也是十分困难的工作, 在音乐检索、音乐可视化、音乐辅助教学等领域具有广阔的应用前景。例如人们学习歌唱或者乐曲演奏时, 可以通过自动标注系统的输出结果对其学习效果进行评价, 从而加快学习的效果和效率。在基于内容的音乐检索技术中, 自动标注技术是实现音频音乐语义级检索的基础, 具有重要地位, 是一项非常有意义的研究工作。

鼓的转录是AMT的一个子任务，它涉及到创建一个由打击乐器（鼓、钹、镣等）演奏的所有音符的象征性表示。源材料通常是，如在AMT中，一个单耳音频来源——要么来自包含多种乐器的复调音频，或一个单独的鼓音轨。由打击乐器演奏的音符的符号表示可以用来获得有节奏的元信息，如节奏、节拍和低拍。

## 研究背景

迄今为止，自动转录研究的一个缺点是缺乏研究和弦打击乐的转录，对此的研究十分有限，且尚未在转录问题的这个特定子集中进行.而现有的研究已经证明，深度学习算法的应用能大大提升鼓转录系统的效率和准确性，尤其是卷积与递归神经网络比目前最先进的算法表现更好。在本工作中，本文介绍了多种利用rnn与cnn自动转录单鼓音轨的新方法。虽然这是从复调音乐转录鼓的第一步，但也存在多个关于独奏鼓轨道转录的应用。在电子音乐制作中，如果需要使用不同的声音进行重新合成，它可以用来转录鼓环。已录制的独奏鼓音轨的转录可以用于录制和摇滚歌曲的制作。如今，使用采样鼓并不罕见，例如，在低成本制作或现代重金属流派。一方面，这是由于录制鼓的复杂性和成本。另一方面，它更容易达到高精度，甚至机器人的声音风格与某些特定类型。与手动编程的鼓轨道相比，自动转录一个简单的低质量的鼓录音可以作为制作歌曲的基础。在这些音频作品中，本文主要关注其中运用的几种典型的鼓种类。这些乐器通常定义了主要的节奏风格，这取决于乐曲的流派。

## 研究现状

大体上，虽然在过去已经提出了许多不同的ADT方法，最新的研究集中在端到端方法计算激活函数的每个鼓仪器。这些方法利用非负矩阵分解(NMF，例如适应NMF等和部分固定NMF等以及网络和双向网络等提取激活函数的音频信号。这种基于激活功能的端到端ADT系统规避了与其他体系结构相关的某些问题。此外，一种被称为分割与分类的方法也被证明取得了相当好的效果：首先是对歌曲进行分段（例如使用开始检测），然后对这些片段进行分类的方法。该方法在分割步骤后会出现信息丢失，即当系统无法检测到一个片段时，该信息就会丢失。这种系统在很大程度上依赖于单个组件的准确性。

由于RNNs，特别是长短期记忆和门控循环单元网络，是被设计用来模拟长期关系的，人们可能会认为基于上述几种算法的RNN系统可以利用鼓轨迹的重复结构并利用这些信息，因此也有不少此方面的研究。此外，CNN的方法也被用于进行鼓转录算法的实现，并且同样被发现有着不错的效果。

从研究方向上看, 当今国内外的鼓转录技术研究大体包括以下几个方面：

1、基于传统机器学习：

独立子空间分析（Independent Subspace Analysis ，ISA),独立分量分析（Independent Component Analysis，ICA),退化分离估计技术（Degenerate Unmixing Estimation Technique ，DUET）[9]非负矩阵分解（non-negative matrix factorization)[8]隐马尔科夫模型[11]

2、基于深度学习

卷积神经网络，循环神经网络[6]。

传统的机器学习方法（如ISA）旨在利用信息论的相关成果来覆盖鼓转录领域的声源分离问题，但ISA算法本身非常少被用于该方面，因此Fitzgerald等人在传统ISA算法基础上做了改进，使其更适用于鼓转录，并更能适应军鼓、底鼓和踩镲或镲片 等有音高的乐器。而结合了深度学习理论的鼓转录算法与传统方法的显著不同是并未采取先验知识空间来建造模型，而是采取了无监督学习建模方式，模仿人类学习音乐的途径来进行鼓转录器的学习[5]

从分类方法上来看，Jason Hockman等人总结了现有鼓转录分类算法共有以下三类[4]：

1. 分割与分类（segement and classify）

分割阶段，首先使用起始检测（onest detection）或从节拍跟踪（beat tracking）中导出的度量网格来将给定的鼓记录划分成不同区域，然后从这些不同的区域中提取相应特征，最后应用分类来确定每个区域中的鼓乐器种类。

1. 匹配与适应（match and adapt）

首先将鼓乐器与预先确定的模板相关联，然后迭代更新模板以反映记录的声谱特征。

1. 分离检测（separate and detect）

识别每个鼓声段的起始点之前，首先尝试在混合的鼓声源中分离声音信号。迄今为止，对鼓转录算法而言最有效的分离检测方法是非负矩阵分解 (NMF)[8]。这种算法将记录划分为多个基函数和相应的时变增益。

## 1.4主要工作

本文基于目前国内外对鼓转录工作的探索和研究，重点探索基于神经网络，运用分割与分离方法的鼓转录算法，并在此基础上对算法进行进一步拓展，旨在开发出一套实时能够获取算法转录结果的同时，还能转发此转录信息到其他软硬件系统，实现实时高精度鼓转录工作的精准落地，同时减轻后续下游任务的开发难度，拓宽鼓转录的研究生态，提升这项技术的实用性。

解决的主要问题为：比较不同结果的卷积神经网络与循环神经网络在鼓转录项目上的效率问题，实时转录模块、网络传输模块的开发问题。

本文最终对比了上述各种算法的效率优劣，并且达到了较高的精度，在此基础上开发了实时转录模块与网络传输模块，能够以较高精度实时转录鼓点音频并且转发到标准esp8266芯片上。

## 论文结构

本文分为五章。其中第一章简述了鼓转录的研究背景和研究现状。第二章详细介绍本文中应用的神经网络的基础结构和相关知识。第三章针对本文中的鼓转录实验具体设计卷积神经网络的网络结构以及训练过程。第四章是实时转录模块与网络传输模块的开发。第五章进行上述模型的实验结果与分析。

# 卷积与循环神经网络的基础知识

## 卷积神经网络的网络结构

卷积神经网络一般分为输入层、卷积层、池化层、全连接层、输出层等。其中卷积层为卷积神经网络的主要特征

### 2.1.1卷积层

卷积层通常包括多个可学习的卷积核（也称过滤器）。每一个卷积核可以看作一个2D的数字矩阵，而输出特征同样可以看作一个更大的数字矩阵，而卷积的过程事实上是基于这两者之间的，特定的矩阵运算。这一过程可以描述为下述四个步骤：

1、将卷积核叠加在输入矩阵顶部的某个位置。

2、在卷积核中的值与输入矩阵中的相应值之间执行按元素的乘法。

3、对所有运算结果进行求和和，此总和是输出图像中目标像素的输出值。

4、对所有位置重复此操作，并为输出图像中每个目标像素填充步骤三中的输出值。

在卷积层中常常会并列设置多个过滤器，这些过滤器能将输入矩阵处理成格式相同的多个不同输出。将它们堆叠起来便可以得到卷积层的标准输出——通常是一个三维的矩阵。

卷积层还会包括被称为“填充”的操作，这一过程将在输入矩阵周围填充0来让目标输出形成所需要的尺寸。这一步骤不是必需的，如果不进行任何填充，则称为“有效填充”。

卷积层的目的是减小运算量和提取图像特征。它是卷积神经网络最为关键的步骤。

### 2.2.2池化层

池化层的作用是对仍然尺寸较大的输出进行采样。通常在连续的卷积层之间会周期性地插入一个池化层。它的作用是逐渐降低数据体的空间尺寸，这样的话就能减少网络中参数的数量，使得计算资源耗费变少，也能有效控制过拟合。汇聚层使用汇总操作，通常分为MAX-POOLING与MEAN-pooling，对输入数据体的每一个深度切片独立进行操作，改变它的空间尺寸。最常见的形式是汇聚层使用尺寸2x2的滤波器，以步长为2来对每个深度切片进行降采样，将其中75%的激活信息都丢掉。其中MAX操作是从4个数字中取最大值（也就是在深度切片中某个2x2的区域），深度保持不变。而mean操作则是取平均值。经过池化层的采样，原有的数据尺寸得以缩小，计算量进一步降低的同时也避免了过拟合。

### 2.2.3全连接层

全连接层包含多个通常是共享权值的节点，这些节点中的每个都充分地连接上一层的所有输出，并通过矩阵运算与累加输出一个综合了所有图像特征的预测结果。这个计算结果常常需要通过特定的激活函数来输出分类任务所需的概率，如softmax等。

以激活函数softmax为例，这一函数通过将多个线性输出的实数域输入映射到总和为1的，在[0,1]内的有效概率分布空间，能较好地完成分类任务。

其中C为总输出的个数，i表示某个特定的输出，为第i个节点的输出值。

## 循环神经网络的网络结构

循环神经网络通常只包含输入层，隐藏层与输出层。由于输入层与输出层仅为简单的多个并列一维数据，因此本文仅介绍循环神经网络的隐藏层。

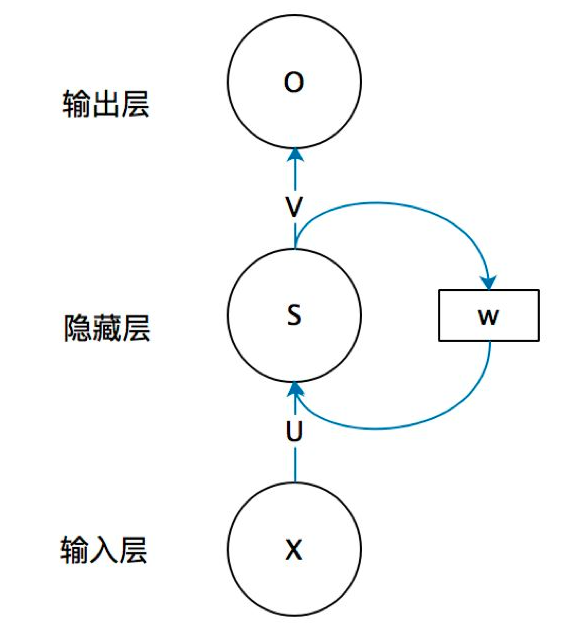


图2-1循环神经网络结构

### 2.2.1循环神经网络隐藏层的基本结构

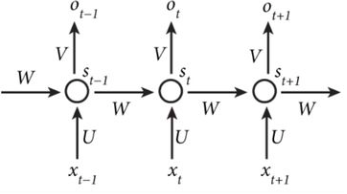


图2-2循环神经网络隐藏层结构

循环神经网络的隐藏层总是由若干个能接受上一层输出、能进行加权运算、能接受新的输入的节点线性组合组成。这样的结构使得循环神经网络能够处理时间顺序的输入。

考察循环神经网络的每个节点，它能够接受来个上个节点的输入，这个输入通常还包含一个可训练的权重W，这个权重决定了上一个节点的输出对此节点输出的影响程度。U,V分别代表着对于输入X与运算结果S的权重，它们的数学关系可以用下式表明：

以此方式链接的多个节点，便组成了一个循环神经网络的隐藏层的网络结构。

## 2.3神经网络的训练模式

神经网络的训练常常以反向传播（Back-propagation）的方式。反向传播是本来只指损失函数对参数的梯度通过网络反向流动的过程，但现在也常被理解成神经网络整个的训练方法，由误差传播、参数更新两个环节循环迭代组成。

其中，误差传播也被称为前向传播，指的是神经网络通过每个层的结构布置，参数设置和网络结构，对输入进行层层运算，并前向推导并最终得出输出的过程。

而参数更新也即反向传播，本质上是求出损失函数L对所有参数梯度，并以此更新参数的过程。这一过程是通过链式法则实现的，以L对某一权重的梯度为例，可以表示如下：

A为某一节点中的权重，B意味着与之链接的另一节点的权重。L意味着损失函数。本公式中，L对A的梯度可以借由L对B的梯度与B对A的梯度来求出，也正因如此每个权重的梯度最终都能由L对最后一层权重的梯度和每一层与前一层的梯度来求出。

求出梯度后，通常在几个不同的策略间选择一个来进行梯度更新，如：RMSPROP，ADAM，BGD，SGD等。这些方式在整体上都将令损失函数的预期逐渐下降。

## 2.4本章小结

在本章，本文介绍了两种神经网络的基本网络结构和核心计算过程，包括卷积神经网络与循环神经网络。本文还讨论了这两种神经网络共同的参数更新方法——反向传播，并指出导数的链式传播法则是参数更新成为可能的关键。在下一章本文将会介绍这样的两种神经网络是如何应用到本文的鼓转录任务中，且是如何进行训练的。

# 基于神经网络的鼓转录算法设计

在本节中，本文将更仔细地探讨神经网络模型的特性。在前面提到的神经网络类别中，本文研究了四种不同的类型：双向递归网络(BRNN)、长短时记忆网络(LSTM)、门控循环单元(GRU)、和卷积网络(CNN)

## 特征提取

特征提取是鼓转录任务最首要的任务。在进行神经网络训搭建与训练之前，本文必须设法让任务输入能够以规范化的数据形式转化，并能够恰好适应神经网络的需要，以便于将它们输入神经网络并完成训练任务。而故转录任务中，本文的输入为许多格式为.WAV的音频文件与这些音频文件对应的标注文件，其中标注文件以.TXT形式保存，其中的每一行都表示某个特定的时间点，一个对应鼓事件的发生。本实验包含两个不同的数据源，其中一个包含6种不同的鼓种类所形成的鼓音频，而另一个仅包含4种类型的鼓种类。为了便于说明，本节以第一个数据源为例进行介绍。

为了从音频文件提取特征信息，常用的方法包含诸如傅里叶变换，快速傅里叶变换，短时傅里叶变换等算法，此外的一些方法也都建立在傅里叶变换的基础上。傅里叶变换是数字信号处理中的基本操作，广泛应用于表述及分析离散时域信号领域，它的功能是将复杂的频谱种的频域信息提取来用于进一步的分析与计算。

然而，原始的傅里叶变换由于运算量过大，并不适用于包括本实验的大多数实验中。而快速傅里叶变换能够大大减少运算量，却在变换中丢失了时域信息。因此，能够适应本实验的算法是短时傅里叶变换。这种方法能够结合快速傅里叶变换中运算效率高的优点，同时能够保留原频谱中关于时域的信息。本实验中运用了这种方法进行特征提取。

此外，本实验还运用了另一种优秀的算法来提取特征：梅尔倒谱系数。这种算法在短时傅里叶变换的基础上叠加了一种被称为梅尔滤波器的滤波器，它通过在原频率上叠加基于人耳的声学认知特征推导的非线性变换将原频率映射到更适应人耳听觉规律的频率域中，这种提取方式可以更好地处理如鼓等乐器的音频，适应这类转录任务，并已被证实取得了更好的效果。本实验同样运用了这种方法来进行特征提取。

接下来详细介绍此两种不同的特征提取在本实验中的实际运用：

### 3.1.1运用短时傅里叶变换的特征提取

作为输入，使用具有44.1kHz采样率的16位分辨率的单声道音频文件。音频被标准化。此后，使用2048的窗口大小和1024的滑动大小来进行短时傅里叶变换。然后，使用鼓包含的方式来进行帧的标注，将每一帧的开始与结束时间转化为秒为单位，并在标注文件中匹配每一个鼓时间与对应帧数，并与前面得到的标记结果拼接在一起。最后，将每一帧的后十帧的特征数据剪切起来，使得cnn也能够获得包含每一帧前后背景的信息。

这两种方式得到的处理结果进行切分，得到对应的训练集特征，训练集标注；以及测试集特征，测试集标注用于训练。

### 3.1.2运用梅尔频谱的特征提取

使用具有44.1kHz采样率的16位分辨率的单声道音频文件。音频被标准化。使用2048个样本的窗口大小计算出一个对数功率谱图。然后将频率轴转换为对数尺度，每个八度使用12个三角形滤波器，频率范围从20到20,000Hz。总共有84个频率箱。

梅尔频谱中的梅尔转换过程可以用下列公式描述：

其中f为原频率。

## 3.2CNN网络的设计

首先对CNN网络的总体架构进行总结：输入数据包含mel频谱图数据。随后的层是具有ReLU激活层和最大池化层的卷积层的交替堆栈。它最后以一个完全连接的ReLU单元层和一个包含sigmoid的输出层结束，具体结构如图所示：

卷积神经网络不仅成功地应用于图像处理，而且还成功地应用于许多其他机器学习任务。卷积层由两层32 3x3滤波器组成，再加上一个3x3最大池化层和一个随机层（drop out layer)（λ=0.3）。

为了克服CNN无法记忆，无法获取上下文信息的弊端，本文为CNN增加了一个被称为光谱上下文的结构。这个结构能够令CNN的输入中的每一帧都包含其后10帧上下文的特征信息用于训练。虽然普通CNN没有任何记忆，但光谱上下文允许CNN在训练和分类过程中访问周围的信息。然而，10帧的背景并不足以在节奏模式中找到重复的结构。因此，CNN只能依赖于声学特征，即信号的光谱特征。然而，通过先进的训练方法，如批处理标准化，以及cnn训练更加方便快捷的优势，可以期待CNN能取得较好的结果。

## RNN网络的设计

本次实验运用了三种不同的RNN：门控循环单元（GRU）、长短时记忆网络（LSTM）、双向RNN，并对比了它们实际的准确率和训练效率，性能。

神经网络整体结构如图所示：

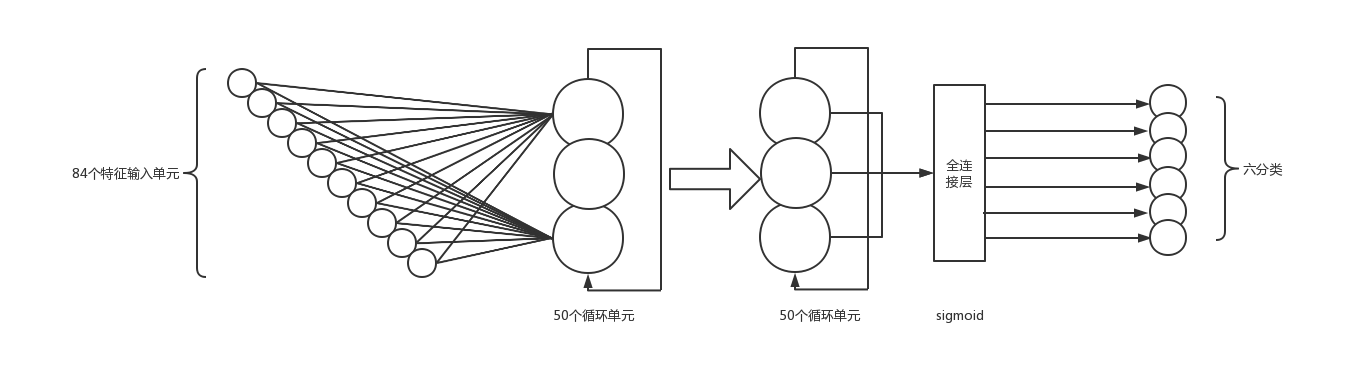


图3-1 RNN网络的网络结构

需要指出的是，RNN模型所输出的概率并非直接用于鼓事件的识别，而仅作为激活函数，用于后续的峰值选择算法的输入。全连接层选择sigmoid函数作为激活函数的原因是因为每个鼓点事件都相互独立，因此并没有采用softamx函数。该神经网络对于每种鼓种类而言，都只输出单一鼓种类的事件概率，形成了一个对应的激活函数。在此之上运用峰值选择算法来识别那些可能的起始点。

这些激活函数将会与已经过标注的鼓事件经处理产生的目标函数进行对比，生成方式是在每个发生鼓事件的起始点将函数值设置为1，而在其他所有时刻置0。激活函数与该函数的损失用交叉熵的方式计算，并成为此神经网络的激活函数。

sigmoid函数定义如下：

交叉熵的计算公式定义如下：

接下来对几种不同的RNN结构进行介绍。

### 3.2.1门控循环单元(GRU)

门控循环单元具有84个特征的输入节点，并通过两层相互连接的循环层对输入进行处理。输入层与循环层之间的连接、循环连接以及循环层与输出层之间的连接都被密集地实现了（每个节点都连接到所有其他节点）。结果表明在输出层使用sigmoid函数取得了更好的效果。

门控循环单元的特殊性体现在每一个循环单元中：

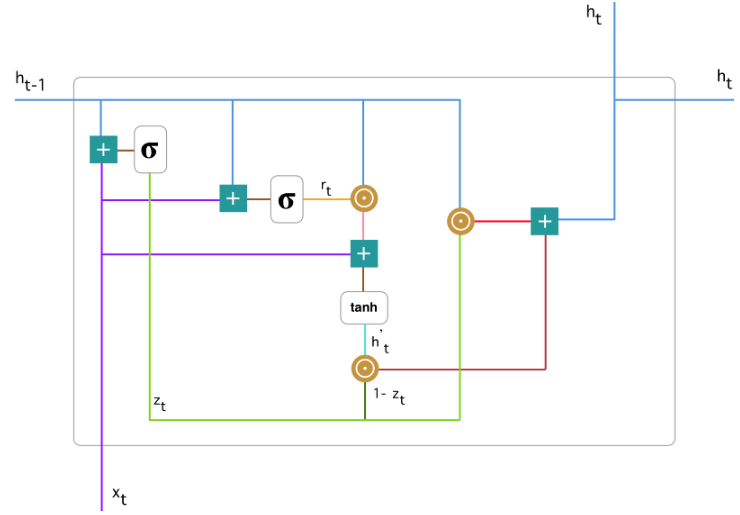


图3-2门控循环单元的内部结构

其中，为前一个时间布保存的记忆信息，为本单元接受的输入。GRU 有两个有两个门，即一个重置门（reset gate）和一个更新门（update gate）。从直观上来说，重置门决定了如何将新的输入信息与前面的记忆相结合，更新门定义了前面记忆保存到当前时间步的量。如果本文将重置门设置为 1，更新门设置为 0，那么本文将再次获得标准 RNN 模型。Gru模型的主要运算公式给出如下：

其中，z表示更新门。更新门帮助模型决定到底要将多少过去的信息传递到未来，或到底前一时间步和当前时间步的信息有多少是需要继续传递的。这一点十分重要，由此模型能决定从过去复制所有的信息以减少梯度消失的风险。

r表示重置门。本质上来说，重置门主要决定了到底有多少过去的信息需要遗忘。

意味着本单元的记忆内容，而则意味着在经过门的选择之后传递到下一单元的记忆内容。

门控循环单元的意义在于，它能够自主通过门来对记忆内容的传递来进行选择——可以选择重要的内容而充分保留，也可以选择不太重要的内容紧接着抛弃。它们能够保存长期序列中的信息，且不会随时间而清除或因为与预测不相关而移除。因此，这也大大缓解了RNN可能会出现的梯度消失问题。

### 3.2.2长短时记忆网络(LSTM)

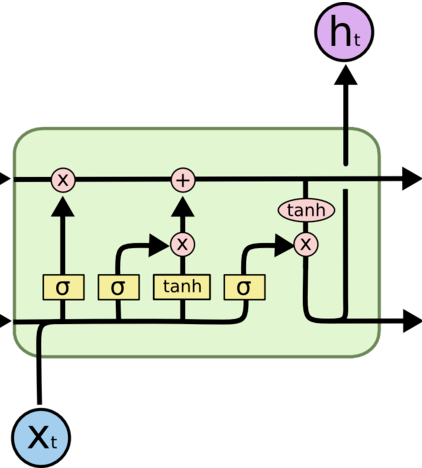


图3-3长短时记忆网络单元内部结构

Lstm的结构与Gru非常类似，但是更加复杂。Lstm的每一个单元包含两个传递自上一单元的信息：记忆细胞C（即图中从上至下的第一条水平线）与单元输出h（即图中从上至下的第二条水平线）。记忆细胞是一段能够很方便地进行传递的信息量，它几乎可以不进行什么运算就直接传递到下一层，这赋予了Lstm长时记忆的能力：较为初始的记忆细胞量能够几乎不损耗地传递到位于队列末尾的单元。除此之外，x表示该单元的输入。

Lstm的运算公式给出如下：

### 3.2.3双向循环神经网络（Bidirectional RNN)

本工作中使用的双向RNN的体系结构由前向层的50个节点和后向层的50个节点组成。前向层和后向层都直接连接到输入层。双向rnn通常比单向rnn产生更好的结果，因为它们也可以使用未来框架的上下文进行分类。在这项工作中，他们意味着结合向前和向后RNN的优势。

## 峰值选择算法

峰值选择算法对于基于RNN的鼓转录算法而言是必需的。初始特征经过神经网络后，将会输出对应每个鼓种类在时间线上的逐帧发生概率，如果将此视为一个关于时间n的函数F(x)，峰值选择算法则将上述函数进行进一步处理，从中选取出最可能发生鼓事件的每一个时间点。

本次实验采用的峰值选择算法是基于Richard Vogl, Matthias Dorfer, Peter Knees等人使用的选择算法开发的，该算法以下式给出：

满足上述三个式子的时间点n被视为一个最可能的鼓起始点。简单地说，一个峰值必须是某个窗口的最大值，并且高于另一个窗口的平均值加上某个阈值。此外，必须有一个至少组合宽度的距离，这个组合宽度也被称为容忍距离。当考虑到10ms是听到两个不同事件的阈值，选择窗口的参数是为了在开发数据集上获得良好的结果[1]

## 3.4训练过程

本工作中rnn必须解决的任务是一个六方二元分类问题。当提供输入光谱图时，RNN必须通过预测输出神经元的激活功能来识别六种鼓类别的起始点。训练算法必须调整网络的权值和偏差，以实现这种功能。此外，在rnn的循环层和输出层之间的辍学[30]被用于训练。该训练算法的目标是使损失函数最小化。损失函数测量网络在再现目标函数时产生的误差。使用六个输出神经元和目标函数的值的二值交叉熵的平均值作为训练的损失函数。

使用的初始学习率为0.02，更新方法选用Adam，每一批次的样本量选取20，迭代次数选择50，最后选择在测试集准确率最高的权重。

## 3.5代码实现

本次实验全程使用基于python语言的tensorflow框架进行开发，

# 实时转录模块与网络传输模块的开发

## 实时转录模块的开发

### 4.1.1实时转录模块的需求分析

鼓转录的一个重要应用就是实时转录，在现代许多音频能以更加方便的形式进行传播，因此包括流行乐在内的许多含鼓音乐音频能够及时获取，并且催生出了即时转录的需求。即时转录功能可以成为鼓转录许多下游任务的起点，如音乐机器人等。即时转录功能还能够满足现场音乐的转录需求，便于音乐研究人员在第一时间获取鼓的时间信息。因此鼓的实时转录是一项十分有价值，且应用前景十分广阔的功能。

鼓转录功能要求可以通过pc端的录音设备实时读取外界的音频输入，并且能狗通过这些音频自动转录出其中含鼓的可能起始点的时间信息，并将其转化为图表或文字的形式进行输出，要求能同时识别现有的多种鼓乐器，并且能识别同一时间点上的多种不同起始点，确保每一个可能的鼓点都能被识别并且转录。

整体的转录流程大体如下：

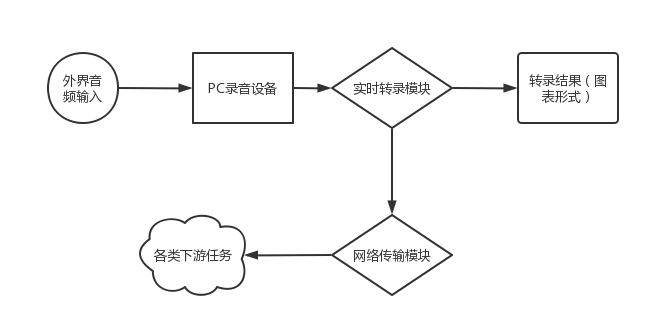


图4.1 实时转录模块的转录流程

此外对该系统开发的可行性进行分析如下：

1. 技术可行性：本系统可以通过pyaudio库来便捷地调用PC端接口来获取外界录音，同时利用wave库的相关函数来对数据进行进一步的加工处理，同时利用tensorflow模块来对上一节已保存的神经网络进行调用，实时分析转录结果，最终使用matplotlib库的图表输出功能来进行输出。
2. 经济可行性：本系统并不需要额外的经济开销进行支持，完全使用开源的python库和一台个人pc就可以完成开发和测试，因此本模块的经济可行性是十分充分的。

综上所述，对实施转录模块的功能需求总结如下，要求本模块具有：

1. 接收外界音频
2. 实时对音频数据化处理形成转录结果
3. 图形化输出

等功能，而本模块同时具有技术可行性和经济可行性，因此可以成为本阶段的开发目标。

### 4.1.2实时转录模块的设计

实时转录模块通过python的pyaudiio库进行开发。

一、监听者模块

定义监听者类listener，其包含一个run方法来调用pyaudio中wave模块的open方法，开启音频流并读入实时音频。设置参数Chunk = 1024，Format = pyaudio.paInt16 channels= 2 Rate = 44100

监听者模块的主要功能对应上一节中的需求（1），能够实现开启音频流读取外界输入，转化为数据流并进行存储便于下一步处理的功能。

二、音频处理模块

定义队列Q用于存放解析的帧信息，定义回调函数来对读取的音频信息进行预处理，转化为标准的梅尔频谱形式。

开启一个while循环来实时监听外界音频输入，并调用tensorflow库读取保存的模型信息进行预测，将结果规范化后调用matplotlib的接口来进行图形化输出。

音频处理模块的主要功能对应上一节中的需求（2）和（3）主要对数据进行处理，同时转化为图形界面输出。

### 4.1.3实时转录模块的实现

本模块的开发使用windows10操作系统，python3.6作为解释器，使用音频处理库pyaoudio、wave，数据处理库numpy，神经网络框架tensorflow进行开发，

## 4.2网络传输模块的开发

### 4.2.1网络传输模块需求分析

网络传输功能是实时鼓转录应当实现的重要功能。将实时转录出的结果通过网络分发到各个软硬件系统中去，可以极大地丰富实时转录模块的应用场景，提升实时转录模块的实用价值，并且便于下游任务对实施转录的结果进行进一步利用。因此网络传输模块是本系统实施转录模块后的关键一环。

网络传输模块是一个能够将实时转录得到的结果，通过mqtt转发到其余软、硬件系统，以实现转录结果功能拓展，复用的模块，能够被用于如智能音乐机器人等项目的开发，具有较好的应用前景。

本文的网络传输模块要求能够实现将实时转录模块的结果通过mqtt分发到标准esp8266芯片，并使其操控两个舵机的转动，能够实现不同的转录结果带动不同的舵机转动，体现出系统的分类功能。

网络传输模块的开发可行性分析如下：

1. 技术可行性

本模块应用嵌入式开发语言micropython，配合thonny进行程序的编写，录入。使用python进一步拓写上一部分中的实施转录模块，并将转录结果分发至公用broker上，最终转发至硬件系统esp8266芯片。

1. 经济可行性

本模块的开发，设计，测试环节需要一片esp8266芯片，一台pc，两个舵机即可完成。因此本模块具有充分的经济可行性。

### 4.2.2网络传输模块的设计

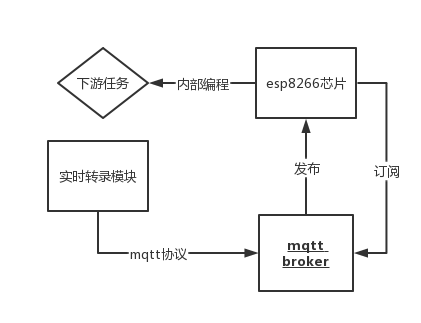
网络传输模块的整体流程如下图所

图4.2 网络传输模块的整体框架

网络传输模块需要通过mqtt架构来实现信息在软、硬件间的传递。具体流程为：实时转录模块通过mqtt协议，成为一名mqttbroker的客户端（client），并向其定期上传实时转录结果。Broker作为信息中转站，无差别地向所有客户端发布特定主题的信息包，并最终以唯一主题名称标识由已编程的esp8266芯片接收，并通过内部编程来进一步实现信息处理，来完成诸如机器人循鼓点转动，闪光，转圈等的下游任务。

在本文中，网络传输模块将三分类的实时转录结果传输至esp8266芯片，并控制两台舵机的转动。

1. PC端传输模块

PC端传输模块主要用于配合实施转录模块来进行信息传输，调用paho库来进行数据的转发。设置公用broker（后称broker），设置专用topic，定义不同的分类结果对应信息，并通过条件判断来实现分类结果到信息的传递。

1. esp8266芯片内部模块

本模块用于实现网络传输模块的下游任务：舵机实时转动。通过mqtt协议接受broker传输的特定topic，对应不同的信息进行条件判断，完成不同舵机对应转动的任务。

### 4.2.3网络传输模块的实现

网络传输模块主要由语言micropython提供支持，使用thonny进行开发，使用网络公共broker进行信息转运，使用标准esp8266芯片作为客户端。

# 实验结果与分析

## 5.1实验数据集

本次实验采用了两个不同的数据集来进行训练，以及对神经网络的精度进行评估。

其中包括来自ENST-Drums以及来自IDMT-SMT-Drums database的原音频和标注文件。

ENST-Drums数据库是一个大型而多样的鼓自动转录和处理的研究数据库。这个数据库记录了三个专攻不同音乐类型的专业鼓手。每个鼓手录制的音频材料的总持续时间约为75分钟。每个鼓手都演奏自己的鼓序列，在每个序列中，使用棒、杆、刷子或木槌来增加鼓声音的多样性。鼓包本身是各种各样的，包含一个小的，便携的种类，有两个鼓和两个钹，适合爵士和拉丁音乐；到一个更大的摇滚鼓，其中包含4个钹和5个钹。

这个由三个专业鼓手表演的视听数据库记录在8个音频频道和2个视频频道。鼓序列已完全注释，并将在很大程度上为研究目的自由分配。其内容的多样性应服务于涉及鼓的音频信号处理的各个领域的研究，从单鼓事件分类到复杂的多模态鼓音轨转录和复音音乐的提取。

IDMT-SMT-Drums数据库是一个用于自动鼓转录和源分离的中型数据库。该数据集由608个WAV文件组成(44.1kHz，Mono，16位)。大约持续时间为2：10小时。这些音频文件均为44100采样率，前一部分包含六种类别的鼓种，而后一部分仅包含三种类别的鼓种。

|  |  |
| --- | --- |
| 所用算法 | 最佳的测试集准确率 |
| BRNN | 91.4% |
| LSTM | 98.2% |
| GRU | 95.8% |
| CNN | 82% |

图5.1 在IDMT-SMT-Drums database上的准确率

|  |  |
| --- | --- |
| 所用算法 | 最佳的测试集准确率 |
| BRNN | 73..1% |
| LSTM | 93.2% |
| GRU | 92..7% |
| CNN | 70% |

图5.2 在ENST-Drums上的准确率

## 5.2实验步骤

在本次实验中为了评估所提出的系统，我们对测试子集的音频文件进行了预处理，如第3.1节所述。随后，该音频文件的频谱图被输入到RNN的输入层中。在第一个数据集ENST-Drums上，输出层的六个神经元为六种鼓点提供激活函数，然后通过拾峰算法识别相关的峰值，这些峰值最终将会被转换为不同鼓的起始点。在第二个数据集IDMT-SMT-Drums database上鼓的种类和输出层神经元的个数则为3。然后使用10ms的窗口进行峰值筛选，识别真起始点和假的起始点。

对于所有四种架构，都进行了上述两个不同的实验。首先，使用ENST-Drums数据集来对数据进行预处理，然后将其随机切分成训练集和测试即，并提取训练集用四种不同的神经网络进行训练。然后，使用测试集对每种神经网络进行测试。第二次重新搭建新的神经网络结构，主要是适应不同的鼓种类，然后利用IDMT-SMT-Drums database重新进行特征提取，分割与训练，并在最后进行测试。这样做是为了评估训练后的模型的泛化能力，以及模型是否与训练数据集过拟合。值得注意的是，在这两种数据集上我们都使用帧包含的方式进行标注，这样做的目的是方便提取每一帧的起始点，这使得它更容易识别所有注释的起始点。另一方面，在音频中有一些起始点其实不应该被转录，如果神经网络错误地将这样的点进行转录，则这个结果将会被视为错误的。

ENST-Drums数据集比IDMT-SMT-Drums数据集包含更有表现力和更快的鼓点风格，这就导致前者的转录难度更大，导致转录结果也不如后者理想，效果有所偏差。这种情况反映在最先进的算法和所提出的方法的转录性能上。

## 5.2结果分析

本次在IDMT-SMT-Drums database上的实验结果中采用的三种RNN算法均取得了好于上下文帧数10的CNN算法的准确率，且其中LSTM算法的准确率最高。可以看出,LSTM在与其相似的GRU算法的对比中取得了极其微弱的优势，因此在本数据集上更长的依赖关系对结果产生的影响十分微弱，相比之下GRU算法带来的算力节省和较高效率不可忽视，综合来看GRU更适合被应用于进行鼓转录算法的神经网络训练。而CNN由于缺乏捕捉时间序列上下文之间的关系，因此在准确率上与RNN算法产生了明显的差距。

除此之外，BRNN在结果上也取得了较为不错的准确率，同时观察训练历史可以得知，BRNN的准确率提升是最为平稳的，此外BRNN的效率也优于GRU，这表明双向rnn的短维持信息能够产生较多的额外用于分类的信息。

此外，观察这两个不同数据集上此四种算法的表表现可以发现，LSTM更复杂的记忆功能所体现出的高性能在风格更加突出的ENST-Drums数据集上表现的更不明显，这说明影响ENST-Drums精度的负面因素收到长时记忆的改善非常微弱，这是因为它风格的多变稀释了长时记忆所带来的额外信息量。

# 总结

## 论文工作总结

本文介绍了三种利用rnn的单鼓轨道转录方法。他们的转录性能比使用CNN方法的最先进的方法的结果更好。RNN方法似乎能够很好地进行泛化，因为合理的高转录结果是在另一个的、独立的、更不同的数据集上产生的。本实验还开发了一个用于实时转录的模块，它能够使用已保存的神经网络进行实时的鼓事件预测并进行转录，以图形化页面显示结果，结果表明这部分模块的准确率相当高，达到了测试时的平均水平。

## 工作展望

使用神经网络进行鼓转录的进一步工作包括使用不同的信息提取和分析思路，除去本文使用的分割与分类方法外，还有包括匹配与适应（match and adapt）与分离检测（separate and detect）等的方法可以用于鼓音频的分析。此外，还有更多的神经网络方面的成果可以用于鼓转录工作，如近几年兴起的注意力机制等。鼓转录的神经网络算法势必可以扩展到更大的数据集，在更广的范围内进行验证，适应和提升。

# 

# 参考文献

1. Vogl, R., Dorfer, M., & Knees, P. (2016, August). Recurrent Neural Networks for Drum Transcription. In ISMIR (pp. 730-736).
2. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.
3. Ngiam J, Chen Z, Chia D, et al. Tiled convolutional neural networks[C], Advances in Neural Information Processing Systems. 2010: 1279-1287.

1. [EB/OL].http://www.cajcd.edu.cn/pub/wml.txt/980810-2.html'1998-08-16/1998-10-04.
2. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C], Advances in neural information processing systems. 2012: 1097-1105.
3. Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[M], Computer Vision–ECCV 2014. Springer International Publishing, 2014: 818-833.
4. Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional networks[C],Proc. CVPR, 2010: 2528-2535.
5. Marius Miron, Matthew EP Davies, and Fabien Gouyon. An open-source drum transcription system for pure data and max msp. In Proc. 38th IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP), 2013.
6. Derry FitzGerald, Bob Lawlor, and Eugene Coyle. Drum transcription in the presence of pitched instruments using prior subspace analysis. In Proc. Irish Signals & Systems Conf, 2003.
7. Olivier Gillet and Gael Richard. Automatic transcription of drum loops. In Proc. 29th IEEE Intl Conf on Acoustics, Speech, and Signal Processing (ICASSP), 2004
8. Olivier Gillet and Gael Richard. Supervised and unsupervised sequence modelling for drum transcription. In Proc. 8th Intl Conf on Music Information Retrieval (ISMIR), 2007.
9. Olivier Gillet and Gael Richard. Transcription and separation of drum signals from polyphonic music. IEEE Transactions on Audio, Speech, and Language Processing, 16(3):529–540, 2008.
10. Marius Miron, Matthew EP Davies, and Fabien Gouyon. Improving the real-time performance of a causal audio drum transcription system. In Proc. 10th Sound and Music Computing Conf (SMC), 2013.
11. Arnaud Moreau and Arthur Flexer. Drum transcription in polyphonic music using non-negative matrix factorisation. In Proc. 8th Intl Conf on Music Information Retrieval (ISMIR), 2007.
12. Jouni Paulus and Anssi Klapuri. Drum sound detection in polyphonic music with hidden markov models.EURASIP Journal on Audio, Speech, and Music Processing, 2009.
13. Richard Vogl, Matthias Dorfer, and Peter Knees. Drum transcription from polyphonic music with recurrent neural networks. In Proc. 42nd IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP), 2017
14. Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G Okuno. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression. IEEE Transactions on Audio, Speech, and Language Processing, 15(1):333–345, 2007.

# 致谢

在实验、写作的过程中，我确实遇到了数不胜数的苦难，一度灰心丧气。如果没有朱金辉老师，及戴培助教两位良师益友的鼓励和指导，想必也就看不到这篇论文，因此我要郑重地感谢他们，是他们的悉心付出，成就了我的这篇成果。

我更要感谢在生活中，耐心陪伴我的她——周雨晴女士，是她的温柔和智慧给了我勇气和力量，在这个过程中我得以重新审视自己的专业素养，在灰暗的日子里找到阳光！

感谢她忍受我的坏脾气，我的抱怨，我们之间的爱和支持会永远延续。

感谢我的父母和兄弟姐妹一直以来的鼓励。

这篇论文是献给上述所有人的。