语音识别-食物声音识别

答辩人：李卫、张宇健　　指导老师：黄亮

目 录

CONTENTS

01

团队与选题介绍

The introduction

A company is an association or collection of individuals, whether natural persons, legal persons.

团队与选题介绍

The introduction

队长：铁甲_卡布达(李卫)

研究方向：蟑螂恶霸、鲨鱼辣椒、蜘蛛侦探，蝎子莱莱

队员：铁甲_小宝（张宇健）

研究方向：寻找和平星

团队与选题介绍

The introduction

食物声音识别

INTRODUCE

此次选题以语音识别为背景，要求选手使用提供的语音数据训练模型并完成语音分类的任务。

环境要求：

TensorFlow的版本：2.0 +

keras

sklearn

librosa

团队与选题介绍

The introduction

数据集介绍

INTRODUCE

名称	大小	Link
test_a.zip	1.02GB	http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/531887/test_a.zip
train.zip	3.53GB	http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/531887/train.zip
train_sample.zip	515.64MB	http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/531887/train_sample.zip

数据集中包含20种不同食物的咀嚼声音，赛题任务是给这些声音数据建模，准确分类。

完整的训练集：train文件夹；

部分训练集：train_sample文件夹；

测试集：test文件夹；

预测结果的格式：第一列为语音文件名称，第二列为类别。语音文件顺序无要求。

以预测准确率作为评估标准。

02

思路与研究方法

Research ideas and methods

A company is an association or collection of individuals, whether natural persons, legal persons.

思路与研究方法

Research ideas and methods

声音是如何产生的?

声音以波的形式传播，即声波（Sound Wave）。凭频率（Frequency）、 幅度（Magnitude）、相位（Phase）便构成了波及其叠加的所有，声音的不同音高（Pitch）、音量（Loudness）、音色（Timbre） 也由这些基本“粒子”组合而来。

世上形形色色的声波都可以“降解”到基本波身上，这也是傅里叶变换（Fourier Transform）的基本思想。 不同的声波有不同的频率和幅度（决定音量），人耳也有自己的接受范围。人耳对频率的接受范围大致为 20 Hz至20 kHz， 于是以人为本地将更高频率的声波定义为超声波（Ultrasound Wave）、更低频率的声波定义为次声波（Infrasound Wave）， 虽然其他动物可以听到不同范围的声音；人耳对音量的接受范围已经进化得适应了地球上的常规声音，小到呼吸声、飞虫声， 大到飞机起飞、 火箭发射的声音（已经不是地球默认配置），再往上，人的身心就越来越承受不住了，为了衡量音量的大小，再一次以人为本地将人耳所能听到 的1kHz纯音的音量下限定义为0dB。

思路与研究方法

Research ideas and methods

调用基本库

train_test_split是sklearn中用于划分数据集，将原始数据集划分成测试集和训练集两部分的函数。

分类报告：sklearn.metrics.classification_report()，显示主要的分类指标，返回每个类标签的精确、召回率及F1值

GridSearchCV（自适应模型库），可以省略一些调参过程，直接把最好的呈现出来。

数据预处理（sklearn preprocessing）MinMaxScaler()通过将每个特性缩放到给定范围来转换特性。

```
# 基本库

import pandas as pd
import numpy as np

pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.model_selection import GridSearchCV

from sklearn.preprocessing import MinMaxScaler
```

思路与研究方法

Research ideas and methods

音频相关的库

Python有一些很棒的音频处理库，比如Librosa和PyAudio，还有一些内置的模块用于处理音频的基本处理。

我们将主要使用两个库进行音频采集和回放： 1）Librosa：它通常用于分析音频信号，但更倾向于音乐，它包括用于构建MIR（音乐信息检索）系统的nuts 和 bolts

```
# 加载音频处理库
import os
import matplotlib.pyplot as plt
import librosa
import librosa.display
import glob
import IPython.display as ipd
```

思路与研究方法

Research ideas and methods

查看音频数据

```python
# 查看音频数据
def look_data():
    # 音频类别文件夹个数
    print(f'音频文件夹的个数: {len(os.listdir(voice_path))}')

    voice_total = 0
    single_label = {}
    for ind, label_name in enumerate(os.listdir(voice_path)):
        file_path = voice_path + '/' + label_name
        single_num = len(os.listdir(file_path))
        single_label[label_name] = single_num
        voice_total += single_num

    print(f'音频文件总量:  {voice_total}')
    print(f'{"序号":<5}{"类别":<15}{"数量":<10}{"占比"}')
    for ind, (key, value) in enumerate(single_label.items()):
        print(f'{ind:<5}{key:<20}{value:<10}{value / voice_total:.2%}')


look_data()
```

序号	类别	数量	占比
0	aloe	45	4.50%
1	burger	64	6.40%
2	cabbage	48	4.80%
3	candied_fruits	74	7.40%
4	carrots	49	4.90%
5	chips	57	5.70%
6	chocolate	27	2.70%
7	drinks	27	2.70%
8	fries	57	5.70%
9	grapes	61	6.10%
10	gummies	65	6.50%
11	ice-cream	69	6.90%
12	jelly	43	4.30%
13	noodles	33	3.30%
14	pickles	75	7.50%
15	pizza	55	5.50%
16	ribs	47	4.70%
17	salmon	37	3.70%
18	soup	32	3.20%
19	wings	35	3.50%

思路与研究方法

Research ideas and methods

查看音频特征

```
# 播放芦荟的声音
ipd.Audio('./train_sample/aloe/24EJ22XBZ5.wav')

# 播放汉堡的声音
ipd.Audio('./train_sample/burger/0WF1KDZVPZ.wav')
```

使用librosa模块加载音频文件，librosa.load()加载的音频文件，默认采样率（sr）为22050HZ mono。我们可以通过librosa.load(path,sr=44100)来更改采样频率，并查看波形幅度包络图。

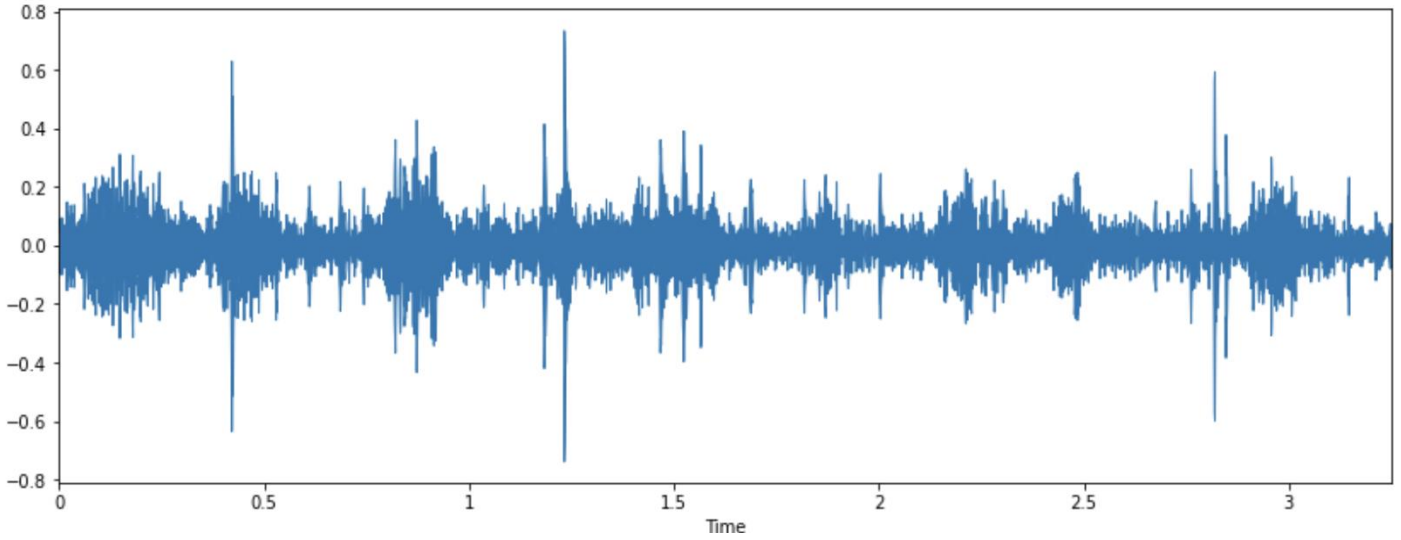
```
data1, sampling_rate1 = librosa.load('./train_sample/aloe/24EJ22XBZ5.wav')
data2, sampling_rate2 = librosa.load('./train_sample/burger/0WF1KDZVPZ.wav')
# 芦荟的波形幅度包络
plt.figure(figsize=(14, 5))
librosa.display.waveplot(data1,sr=sampling_rate1)
```

思路与研究方法

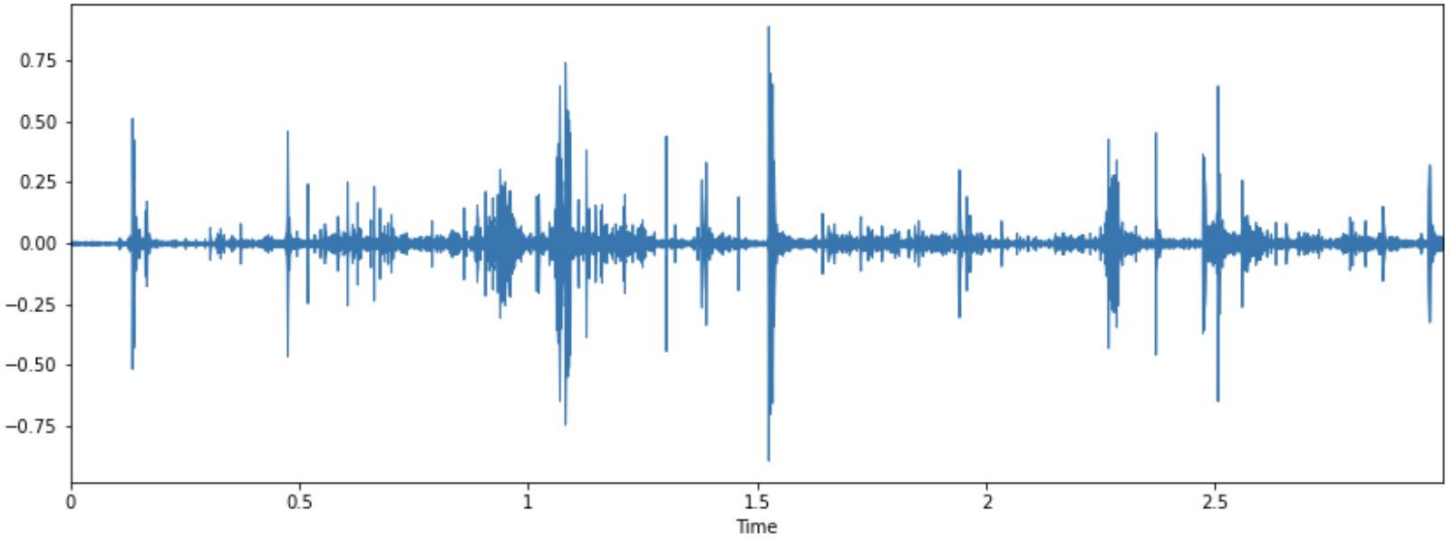
Research ideas and methods

显示包络图

芦荟



汉堡



思路与研究方法

Research ideas and methods

查看声谱图

声谱图（spectrogram）是声音或其他信号的频率随时间变化时的频谱（spectrum）的一种直观表示。声谱图有时也称sonographs,voiceprints,或者voicegrams。当数据以三维图形表示时，可称其为瀑布图（waterfalls）。在二维数组中，第一个轴是频率，第二个轴是时间。我们使用librosa.display.specshow来显示声谱图。

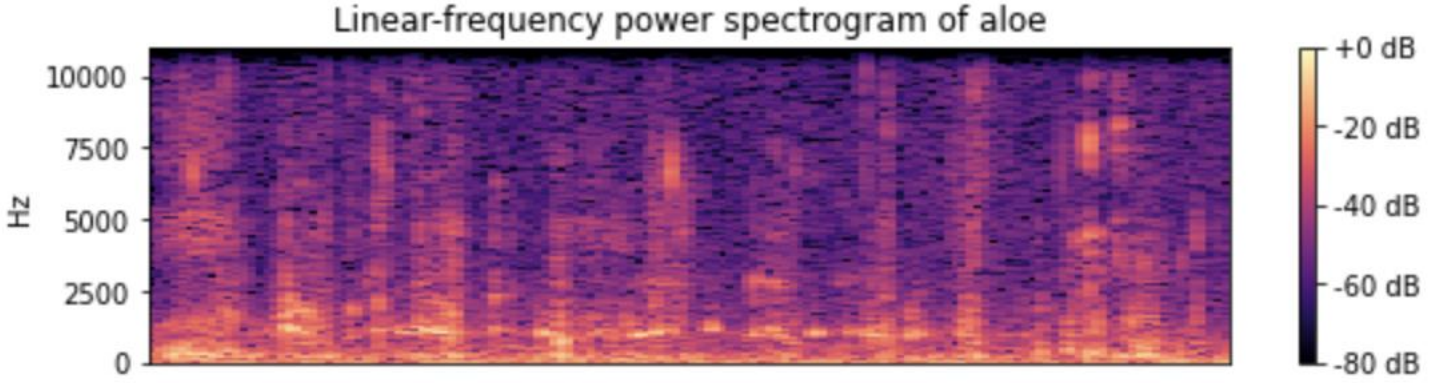
```
# 芦荟的声谱图
plt.figure(figsize=(20, 10))
D = librosa.amplitude_to_db(np.abs(librosa.stft(data1)), ref=np.max)
plt.subplot(4, 2, 1)
librosa.display.specshow(D, y_axis='linear')
plt.colorbar(format='%+2.0f dB')
plt.title('Linear-frequency power spectrogram of aloe')
# 汉堡的声图谱
plt.figure(figsize=(20, 10))
D = librosa.amplitude_to_db(np.abs(librosa.stft(data2)), ref=np.max)
plt.subplot(4, 2, 1)
librosa.display.specshow(D, y_axis='linear')
plt.colorbar(format='%+2.0f dB')
plt.title('Linear-frequency power spectrogram of burger')
```

思路与研究方法

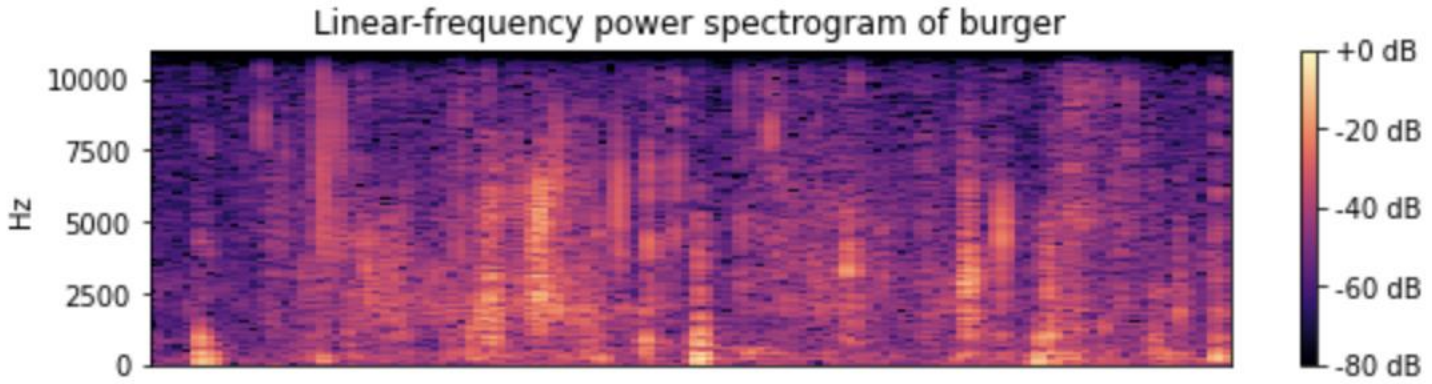
Research ideas and methods

显示声谱图

芦荟



汉堡



思路与研究方法

Research ideas and methods

MFCC特征提取

当拿到上述这些音频数据之后，接下来就需要进行特征提取（过滤掉背景噪音等不需要的信息）从而筛选出我们所需要的信息。

人的耳朵在接收信号的时候，不同的频率会引起耳蜗不同部位的震动。耳蜗就像一个频谱仪，自动在做特征提取并进行语音信号的处理。在语音识别领域中MFCC（Mel Frequency Cepstral Coefficents）特征提取是最常用的方法，也是本次音频分类任务中涉及到的特征提取方法。

具体来说，MFCC特征提取的步骤如下：

1.对语音信号进行分帧处理
2.用周期图(periodogram)法来进行功率谱(power spectrum)估计
3.对功率谱用Mel滤波器组进行滤波，计算每个滤波器里的能量
4.对每个滤波器的能量取log
5.进行离散余弦变换（DCT）变换
6.保留DCT的第2-13个系数，去掉其它

```
feature = []
label = []
# 建立类别标签，不同类别对应不同的数字。
label_dict = {'aloe': 0, 'burger': 1, 'cabbage': 2,'candied_fruits':3, 'carrots': 4, 'chips':5,
              'chocolate': 6, 'drinks': 7, 'fries': 8, 'grapes': 9, 'gummies': 10, 'ice-cream':11,
              'jelly': 12, 'noodles': 13, 'pickles': 14, 'pizza': 15, 'ribs': 16, 'salmon':17,
              'soup': 18, 'wings': 19}
label_dict_inv = {v:k for k,v in label_dict.items()}
```

思路与研究方法

Research ideas and methods

建立提取音频特征的函数并输出

```
def extract_features(parent_dir, sub_dirs, max_file=10, file_ext="*.wav"):
    c = 0
    label, feature = [], []
    for sub_dir in sub_dirs:
        for fn in tqdm(glob.glob(os.path.join(parent_dir, sub_dir, file_ext))[:max_file]):  # 遍历数据集的所有文件

            # segment_log_specgrams, segment_labels = [], []
            # sound_clip,sr = librosa.load(fn)
            # print(fn)
            label_name = fn.split('/')[-2]
            label.extend([label_dict[label_name]])
            X, sample_rate = librosa.load(fn, res_type='kaiser_fast')
            mels = np.mean(librosa.feature.melspectrogram(y=X, sr=sample_rate).T,
                           axis=0)  # 计算梅尔频谱(mel spectrogram),并把它作为特征
            feature.extend([mels])

    return [feature, label]
```

思路与研究方法

Research ideas and methods

```
# 自己更改目录
parent_dir = './train_sample/'
save_dir = "./"
folds = sub_dirs = np.array(['aloe','burger','cabbage','candied_fruits',
                             'carrots','chips','chocolate','drinks','fries',
                            'grapes','gummies','ice-cream','jelly','noodles','pickles',
                            'pizza','ribs','salmon','soup','wings'])

# 获取特征feature以及类别的label
temp = extract_features(parent_dir,sub_dirs,max_file=100)

temp = np.array(temp)
data = temp.transpose()

# 获取特征
X = np.vstack(data[:, 0])

# 获取标签
Y = np.array(data[:, 1])
print('X的特征尺寸是：',X.shape)
print('Y的特征尺寸是：',Y.shape)

# 在Keras库中：to_categorical就是将类别向量转换为二进制（只有0和1）的矩阵类型表示
Y = to_categorical(Y)

# 最终数据
print(X.shape)
print(Y.shape)
```

输出：

(1000, 128)

(1000, 20)

当进行过音频数据的分析以及特征提取后，我们就要开始进行模型建立了，选择的是CNN模型

03

技术与实践

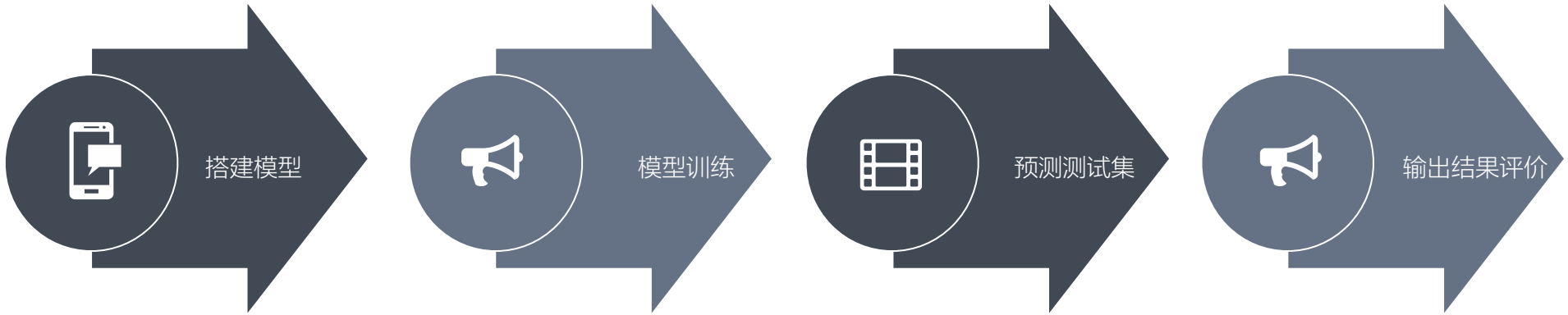
Technology and practice

A company is an association or collection of individuals, whether natural persons, legal persons.

技术与实践

Technology and practice

卷积神经网络模型的建立搭建和使用



技术与实践

Technology and practice

卷积神经网络搭建

1)输入层：用于数据的输入

2)卷积层：使用卷积核进行特征提取和特征映射------>可以多次重复使用

3)激励层：由于卷积也是一种线性运算，因此需要增加非线性映射(也就是激活函数)

4)池化层：进行下采样，对特征图稀疏处理，减少数据运算量----->可以多次重复使用

5)Flatten操作：将二维的向量，拉直为一维的向量，从而可以放入下一层的神经网络中

6)全连接层：通常在CNN的尾部进行重新拟合，减少特征信息的损失----->DNN网络

```python
from keras.models import Sequential

model = Sequential()

input_dim = (16, 8, 1)

model.add(Conv2D(64, (3, 3), padding="same", activation="tanh", input_shape=input_dim))# 卷积层
model.add(MaxPool2D(pool_size=(2, 2)))# 最大池化
model.add(Conv2D(128, (3, 3), padding="same", activation="tanh")) #卷积层
model.add(MaxPool2D(pool_size=(2, 2)))# 最大池化层
model.add(Dropout(0.1))
model.add(Flatten())# 展开
model.add(Dense(1024, activation="tanh"))
model.add(Dense(20, activation="softmax"))# 输出层：20个units输出20个类的概率

# 编译模型，设置损失函数，优化方法以及评价标准
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])# 交叉熵损失函数
```

技术与实践

Technology and practice

89	violethao1	Avionics	37.45	2021-04-13
90	铁甲_卡布达	浙江工业大学	37.40	2021-05-30
90	铁甲_小宝	浙江工业大学	37.40	2021-05-30
92 ↓2	豌豆子333333	广西民族大学	37.35	2021-05-09
93 ↓2	熊猫吃鸡翅吖	广东商学院	37.20	2021-04-14

成功

初次搭建神经网络模型成功，也得到了相应的预测测试集，小有成就

01 LOREM

比赛

正确率达到了37.40%，登上了排行榜的第90位

02 LOREM

03 LOREM

不足

此次模型的正确率远远不够，没有超过50%，也处于排行榜末端

04 LOREM

改进

经过对已有代码的分析，我们的模型只有两层隐层，存在欠拟合的问题。

技术与实践

Technology and practice

改进模型

初步建立的隐层是两层的卷积神经网络模型，通过训练集训练以后，去测试，得到的结果不如人意。正确率只有37%左右。经过分析，我们小组认为可能是存在欠拟合或过拟合的问题，导致结果不好。仔细思考，隐层只有两层，过拟合的概率其实并不大，更有可能是欠拟合。这样一来，改进模型的目标和方向就明显了，我们需要尝试增加隐层数目，且要避免过拟合的问题发生

```
model = Sequential()

# 输入的大小
input_dim = (16, 8, 1)

model.add(Conv2D(64, (3, 3), padding = "same", activation = "tanh", input_shape = input_dim))# 卷积层
model.add(MaxPool2D(pool_size=(2, 2)))# 最大池化
model.add(Conv2D(128, (3, 3), padding = "same", activation = "tanh")) #卷积层
model.add(MaxPool2D(pool_size=(2, 2))) # 最大池化层
model.add(Dropout(0.1))
model.add(Flatten()) # 展开
model.add(Dense(1024, activation = "tanh"))
model.add(Dense(20, activation = "softmax")) # 输出层：20个units输出20个类的概率
```

技术与实践

Technology and practice

改进模型的表现：

将改进模型预测测试集后得到的结果上传比赛平台，可以看到无论是在正确率方面还是在排名方面都有了一个显著的提升。正确率提升到72.80%，排名上升到第40名。

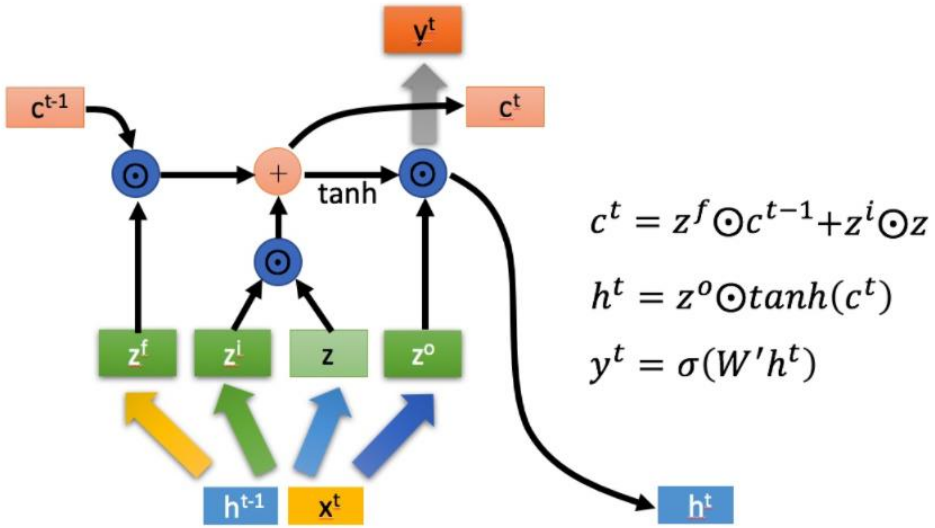
27	菜鸟01#	北方信息工程学院	86.90
28	jia9611	北京航空航天大学	85.20
29	yan837764072	湖南科大	82.80
30	颜士力	湖南科技大学	82.55
31	我是nppc	湖南科大	82.45
32	提莫不好玩	天池	82.35
33	Ger文	广州大学	81.95
34	xa7rfcxyiuvjs	吉林大学	80.15
35	forelike	东莞理工	79.10
36	1217508999679176	绿动资本	77.75
37	yhssmxirdufdq	内蒙古农大	77.30
38	游客bbkktrpr6ntji	咸宁学院	75.15
39	guilinpang	北京交通大学	73.75
40	铁甲_小宝	浙江工业大学	72.80

技术与实践
Technology and practice

进一步提升:

我们希望能够按照之前的经验，通过调节某个参数来提升模型的表现。例如：改变卷积层数、调整卷积核、池化层的大小调整、以及对激活函数的选择。但得到的结果并没有变得更好。

经过网络上的搜索，发现了他人对于此类语音识别所使用的模型LSTM。LSTM模型(长短期记忆神经网络模型)，可以有效解决长序列训练过程中的梯度消失和梯度爆炸问题。



技术与实践

Technology and practice

LSTM主要阶段：

LSTM内部主要有三个阶段：
1. 忘记阶段。这个阶段主要是对上一个节点传进来的输入进行选择性忘记。“忘记不重要的，记住重要的”。

2. 选择记忆阶段。这个阶段将这个阶段的输入有选择性地进行“记忆”。哪些重要则着重记录下来，哪些不重要，则少记一些。

将上面两步得到的结果相加，即可得到传输给下一个状态的。
3. 输出阶段。这个阶段将决定哪些将会被当成当前状态的输出。

抱着尝试一下的态度，我们使用了这个模型。可以看到有提升，但是提升并不大，进步了一点点的。

27	菜鸟01#	北方信息工程学院	86.90
28	jia9611	北京航空航天大学	85.20
29	yan837764072	湖南科大	82.80
30	颜士力	湖南科技大学	82.55
31	我是nppc	湖南科大	82.45
32	提莫不好玩	天池	82.35
33	铁甲_小宝	浙江工业大学	81.95
34	xa7rfcxyiuvjs	吉林大学	80.15

04

总结与展望

Summary and prospect

A company is an association or collection of individuals, whether natural persons, legal persons.

总结与展望

Summary and prospect

题目分析

学习了有关声波方面的知识，以及声谱图等

构建模型

搭建出了一个含两层隐层的神经网络模型

改进模型

这个过程是最消磨耐心的，也是机器学习必不可少的

选择题目

选择的是语音识别方面的赛题，以正确率作为比赛标准。

选择模型

选择使用·的·是·卷积·神经·网络模型

预测结果

得到的预测结果也能排的上榜

总结与展望

Summary and prospect

此次机器学习课程设计，我们选择使用的是卷积神经网络模型，需求驱动学习，这段时间也学习了大量和这方面有关的知识。当然，图中也有想过放弃，尤其是调参的环节，繁琐复杂，不仅如此，还需要花费大量时间去运行，一点一点消磨耐心，真真切切体会到了调参的重要性，好在最后都坚持了下来。除此之外，合适模型的选择也是非常重要的，用适宜的模型去解决适宜的问题，具体问题具体分析。在机器学习这条路上，走得越远，发现自己不会的东西也就越多，更需要静下心来去努力学习。非常遗憾的是，比赛平台的要求严格、我们的疏忽大意，导致了在实名验证时的失败，使得我们这几个星期以来的成果付诸东流，可以发现有很多队伍是和我们犯一样错误的，如果当时能够验证好，说不定还有机会拿下奖品。这是遗憾，也是教训，提醒我们在今后的人生路上更要一丝不苟，注重细节。

未来：我们也在github上找到了有关声音模型的开源项目，等这段忙碌的时间过去，可以去进一步深入学习的。网上资源很多，剩下的就是需要自己坚持了。

最后，对于黄老师这几个星期以来的指导、督促和鼓励深表感谢!

恳请老师 批评指正

A company is an association or collection of individuals, whether natural persons, legal persons, or a mixture of both. A company is an association or collection of individuals,