### 智能语音课堂笔记：

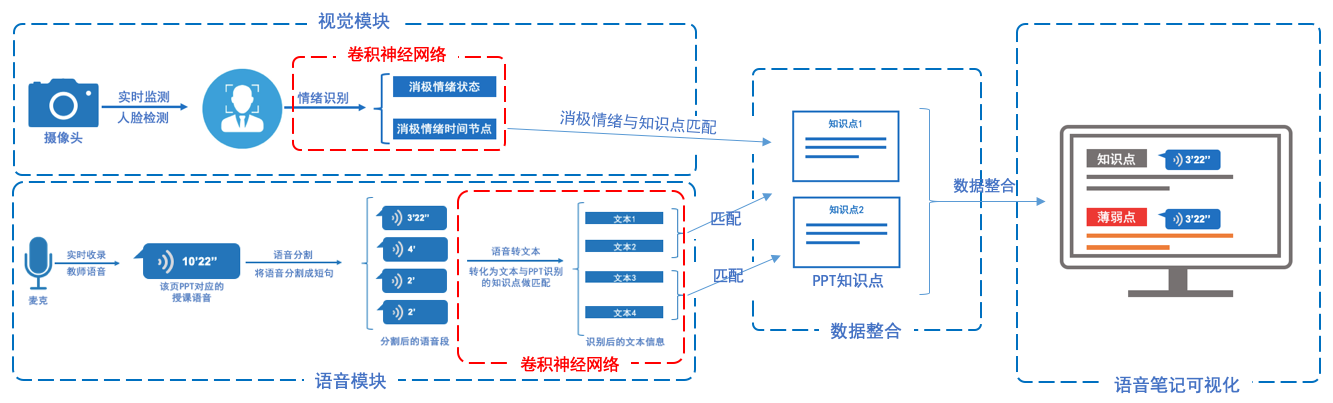
##### 背景：

大学课堂中，知识的传授主要以教师讲解为主，PPT 课件内容多以概括性、辅助教学为主。大学课程专业性强，知识量大，如果学生边听课边记笔记，将会影响听课质量，所以很少学生记录随堂笔记。PPT课件是目前学生复习的主要材料，但课件以概括性主，缺少详细 讲解，这就导致了学生复习时，多为初浅了解，无法深入理解专业知识，甚至是不知所云。

本项目为解决上述问题，提出智能语音解决方案。将华为Atlas 200 DK部署在教室中，利用其麦克风实时获取教师的授课语音，再利用部署在其上的人工智能算法，对语音进行分段和识别，最终实现PPT课件中的知识点和对应教师的讲解语音进行匹配，生成智能语音课堂笔记。学生复习时 可点击 PPT 课件中任意知识点，来听取当时教师对该知识点的详细讲解，提高学生复习效率和质量。 同时，本项目还通过连接在华为 Atlas 200 DK 上的摄像头，实时获取学生上课的情绪状态，将情绪 的时间序列与 PPT 知识点的授课时间序列匹配，为学生提供个性化学习服务。例如，学生在知识点 A 时刻，情绪较为低弱，学习质量下降，智能语音课堂笔记会在学生复习时将该知识点标红，提醒学生对该知识着重复习。

##### 概述：

本项目划分为视觉模块、语音模块、数据整合和语音笔记可视化等四大模块，四个模块相互独立，但数据相互关联。视觉模块包括人脸检测和情绪识别;语音模块包括语音识别;数据整合模块负责将语音模块和视觉模块处理后的结果进行整合，生成语音课堂笔记;语音笔记可视化模块负责将最终的语音课堂笔记可视化，为学生提供服务。项目总体框架如图 1所示。



##### 原理：

本项目中涉及的关键算法包括：人脸检测、情绪识别、语音识别和模型转换。本项目模型主要以 Keras 开发为主，为方便部署在华为Atlas 200 DK，我们利用模型转换算法将训练好的Keras网络模型转为Tensorflow网络模型，再利用华为Mind Studio平台将pb格式的网络模型转换为 Atlas 200 DK 支持的om格式模型。

1.人脸检测

人脸检测从本质上讲是目标检测的一种，也就是将目标(人脸，Target)与干扰(背景， Clutter)区分开来的过程。人脸检测模型是基于Dlib库提供的人脸检测算法进行实现。华为Atlas 200 DK开发板通过调用摄像头拍摄学生上课状态，在拍摄的每帧图像中运用人脸检测算法搜索人脸，找到人脸后返回人脸在图像中的坐标位置。根据返回的人脸坐标位置，对原始图像进行裁剪，裁剪出单独的人脸图像。最终将识别到的人脸图像发送给情绪识别模块进行下一步处理。过程如图 2所示。

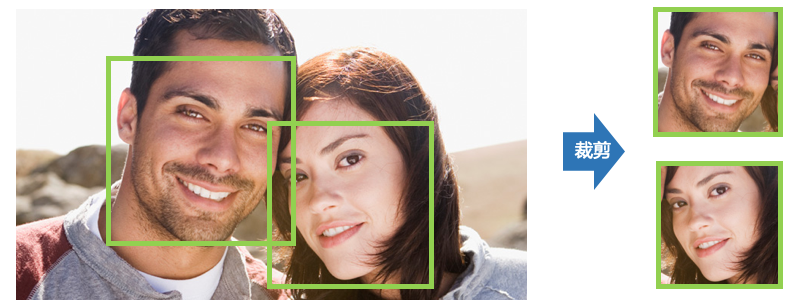


图 2 人脸检测过程示意图

2 情绪识别

情绪识别模型训练的数据集采用Fer2013，该数据集由35886张人脸表情图片组成，其中，训练图片28708张，测试和验证图片各 3589 张，每张图片是由大小固定为48\*48的灰度图像组成，共有7种表情，分别对应数字标签0-6，具体表情对应的标签和中英文如下:0 anger 生气;1 disgust; 2 fear 恐惧;3 happy 高兴;4 sad 伤心;5 surprised 惊讶;6 normal中性。利用Keras深度学习框架设计轻量级情绪识别网络，共包含 12 层。神经网络模型结构如图 3所示。

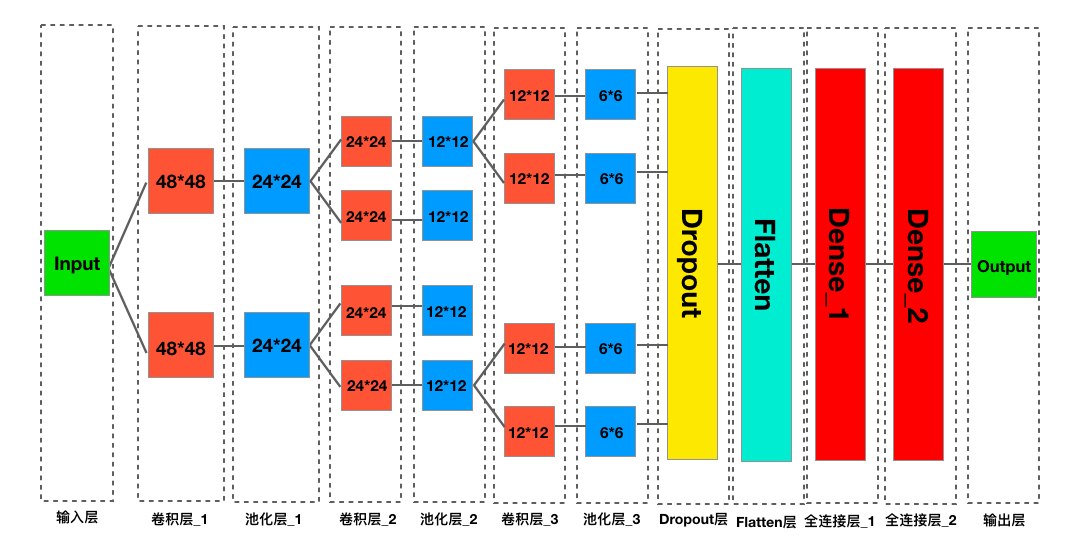


图 3 情绪识别网络结构

3 语音识别

本项目的语音识别模块可以分为声学模型和语言模型。其中声学模型基于Python、Keras构建CNN+CTC神经网络模型。网络结构参考了VGG的深层卷积神经网络，利用多个卷积层和池化层组合，CTC作为损失函数。VGG网络模型在图像识别中表现出色，有着较强的表达能力，相比RNN在鲁棒性上更加出色。在输出端，该模型可以与CTC很好结合在一起。参考图像识别过程，本项目将音频的特征向量设置为神经网络需要的二维频谱图像信号，即语谱图作为输入。输出为识别的拼音序列。

在传统语音识别的声学模型训练过程中，需对每一帧的音频数据进行相应的Label标注，方能有效的训练。在训练数据之前需要做语音对齐的预处理工作，而语音对齐工作本身需要反复迭代，来确保其准确性，这是非常耗时的准备工作。

与传统的声学模型相比，采用CTC作为损失函数的声学模型训练，可以实现端到端的网络训练，不需要预先对齐音频数据，只需要一个输入序列和一个输出序列即可。这种网络结构免去了数据对齐和一一标注的工序，并且CTC可以直接输出序列预测的概率，不需要外部的后处理过程。CTC算法实际上只关注预测输出的序列和真实序列的接近程度，而不会关心预测输出序列中每个结果在时间节点上是否与输入序列对齐。

图 4左为“你好”的音频波形示意图，其中每个红色框代表一帧数据，传统语音识别训练方法需要知道每一帧数据所对应的发音因素，并进行对齐和标注。例如，1-4帧对应“n”的音素；5-7帧对应“i”的音素；8-9帧对应“h”的音素；10-11帧对应“a”的音素；12帧对应“o”的音素（暂且将每个字母作为一个发音因素）。

而CTC算法引入了Blank（该帧没有预测值）和Spike（尖峰）概念。每一个预测的分类结果对应一整音频中的一个Spike，其他不是Spike的位置则被认为是Blank。对于一段音频，CTC处理后的输出结果是一系列Spike的序列，而并不关心每一个因素持续了多长时间。如图 4右所示，进过CTC预测的序列结果在时间上可能稍晚于真实的发音时间节点，但顺序与真实顺序相同，其他位置则全部标记为Blank。

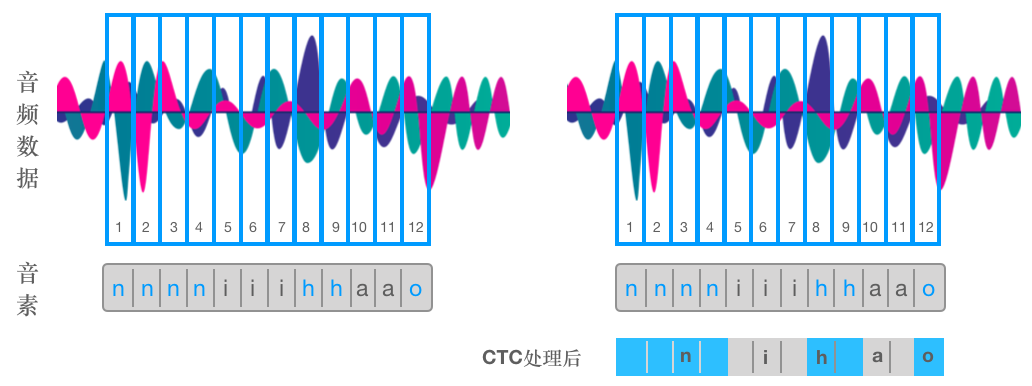


图 4 音频数据对齐示意图

本项目的语言模型是基于统计的语言模型（基于概率图的最大熵隐马尔可夫模型），来实现从拼音到文本的解码过程。而拼音转汉字的过程时动态规划的问题，与寻找最短路径的算法基本相似。我们可以将汉语输入看成是一个通信问题，每一个拼音可以对应多个汉字，而每个汉字一次只读一个音，把每个拼音对应的字从左到右连接起来，就成为了一张有向图。如图 5所示，是输入的拼音序列，是第一个音对应的候选字，以次类推。整个网络就变成在有向图中寻找从起点开始，到终点概率最大的路径问题，可以使用各种求最短路径算法来实现，这里我们使用维特比算法来进行拼音到汉字的解码。

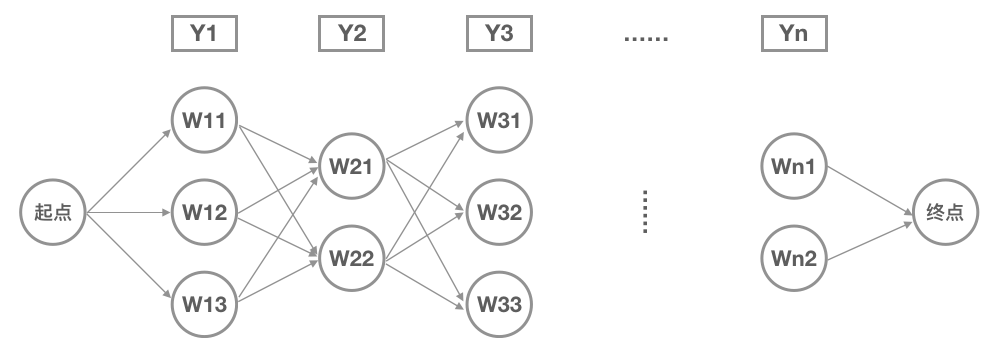


图 5 拼音转文字过程

维特比算法是先计算第一步的概率，然后将概率按大小排序，剔除概率较低的路径，然后再进行第二步，再剔除概率较低的路径，以此类推。通过设置阈值来剔除概率较低的路径。例如，设置每一步的阈值为0.001，每一步的概率便与比较，小于阈值则剔除，其中n为当前的步数。反复执行，直到路径终点，我们便可以获得概率最大的一个句子。

##### 技术

* 华为基于Ascend 200 DK芯片的MindSpore DDK（Device Developer Kit）。
* 基于opencv的人脸检测以及图像预处理，基于CNN+HMM的语音识别。
* 基于tensorflow的ASR 模型和情绪识别模型，使用Matrix提供的模型管家接口将其转换为Ascend 200DK芯片支持的模型。

硬件

* Atlas 200 Developer Kit开发者板
* 外接萤石摄像头
* 板载麦克风

优化点

智能语音课堂笔记的语音识别目前使用的是CNN+CTC+HMM，之后可以采用RNN(LSTM)进行特征提取，同时对于数据集，由于实验采用的数据集只用200小时进行训练，之后的优化可以考虑采用更多的数据集进行重新训练，对于结果展示方面，我们采用的是网站展示，需求者可以根据自己的的相应需求做出调整。

效果展示

见文件夹视频 “”