基于CNN的辩论激烈程度分类

# 音频特征图提取：语谱图

卷积神经网络（CNN）具有模型参数数量巨大、拟合能力强、可以快速处理输入图像等特点。因此我们将每一段辩论的音频转换成图像（语谱图），然后交给CNN分类器进行训练和推断。为了适应CNN输入图像的尺寸以及尽量减少网络参数个数、提高训练速度，在CNN输入端之前还做了图像的预压缩工作。整个特征提取的过程如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Step** | **操作** | **输入** | **输出** | **备注** |
| **1** | STFT | 30s音频 | 2049x702 | NFFT=4096 |
| **2** | Reshape | STFT语谱图 | 1000x702 | 舍弃频率的信息 |
| **3** | 取对数 | 上一级输出 | 1000x702 | 取进行对数压缩 |
| **4** | 双三次插值：降采样 | 上一级输出 | 138x138 | 调用opencv图像库 |

表1：音频预处理流程表

经过了这样的音频预处理以后，每一段长度为点的音频（）被提取为一幅单通道的图像。

# 卷积神经网络(CNN)构建

为了实现基于语谱图和CNN的分类器，我们在pytorch平台上构建了一个7层卷积神经网络。其具体结构如下图和下表所示：

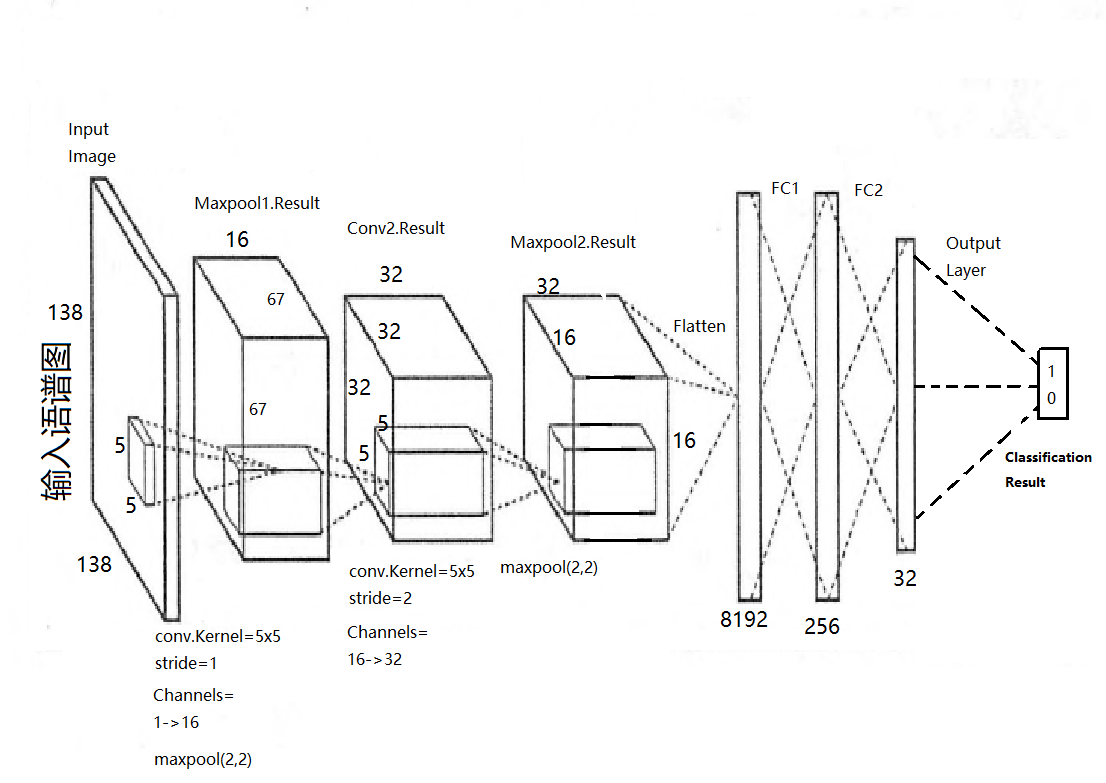


图1：实验中的卷积神经网络结构

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Layer** | **类型** | **输入尺寸** | **输出尺寸** | **备注** |
| **1** | Conv | 1x138x138 | 16x134x134 | 卷积步长=1 |
| **2** | Maxpool | 16x134x134 | 16x67x67 | 下采样2倍 |
| **3** | Conv | 16x67x67 | 32x32x32 | 卷积步长=2 |
| **4** | Maxpool | 32x32x32 | 32x16x16 | 下采样2倍 |
| **5** | FC1 | 8192 | 256 |  |
| **6** | FC2 | 256 | 32 |  |
| **7** | FC3 | 32 | 2 | 输出层 |

表2：卷积神经网络结构表

# CNN训练

为了实现交叉验证，我们将大小为200的训练集随机分割成大小为180的训练集和大小为20的验证集。测试集（大小为100）经过了我们的手工标注，下面的曲线图和表格展示了CNN分别在训练集、测试集和验证集上的Loss和Accuracy随着训练epoch的变化情况：

# CNN分类效果

经过了合理的参数调试，我们最终在pytorch平台上训练出了一个表现优异的模型。训练时，其训练参数如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | **类别** | **参数1** | **参数2** |
| 优化器 | Adam | 学习率 | L2正则化 |
| 损失函数 | 交叉熵 |  |  |
| 迭代次数 |  | 250 |  |
| 批大小 |  | 10 |  |

表4：CNN训练参数

最终正确率截图：