**视听信息的跨模态匹配**

**——《视听信息系统导论》实验报告**

**一、团队成员及分工情况：**

姓名：陆子阳 学号：2015011135

任务：搭建基于LSTM的预测网络以提取所给视频特征和音频特征在时域上的特性；

撰写实验报告

姓名：林子骞 学号：2014011870

任务：搭建基于Res单元变种的网络；

对三人的最终预测结果合并

姓名：李润桐 学号：2015010960

任务：搭建基于Res单元的网络；

调整接口文件

**二、提交文件清单**

configs 搬运助教代码，配置文件，部分参数没有实际意义

def\_model 基于Res网络的部分模型文件

filelists 仿照助教设计，用txt文件说明测试文件名

LSTM\_MODEL 存放训练LSTM网络所需的文件和

LSTM\_model\_data 存放测试LSTM网络所需的部分参数

report 存放运行结果记录

tools 搬运助教代码

Train 数据集

\*.py 各类代码文件，简要说明如下：

训练Res网络：train\_lrt.py

测试Res网络：\_evaluate\_lrt.py

训练全连接网络：SSSV2AandA2V.py

测试全连接网络：\_test\_lzq.py

训练LSTM网络：参见LSTM\_MODEL文件夹

测试LSTM网络：\_LSTM\_lzy.py

测试网络综合：（即我们给出最终模型）\_simi\_matrix.py

(说明：在测试各网络综合结果时，应先分别运行三个网络对应的测试代码，得到相应文件后才能正确运行)

**三、工作开展与完成情况**

1、原理

助教已经利用Inception V3和VGG网络对视觉、听觉信息进行了特征提取，并得到了两组张量。因此实验的主要任务是训练出一个神经网络，去探索视觉、听觉特征之间的关系，并度量两者间的相似性。

2、采用基于Res单元的网络进行匹配（由李润桐完成）

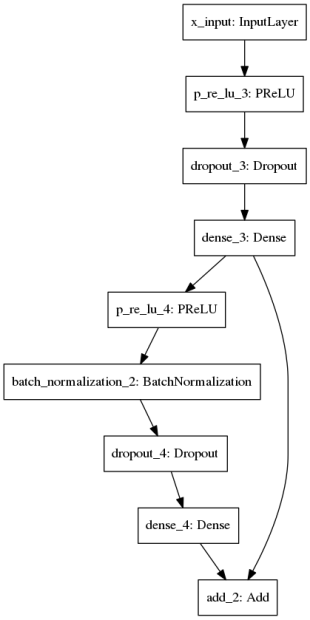
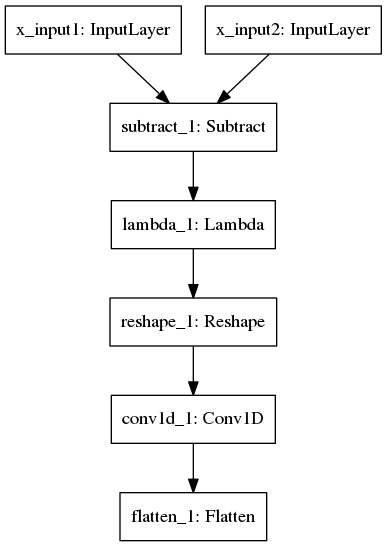
①、数据预处理

对于视觉、听觉特征，这里以视觉为例，向量的大小是120\*1024，样例共计1300个。对于每一个样例，采用每3秒取一次特征的均值，于是每个样例变成了118\*1024大小，相当于手动做了一次conv。而后对其进行规范化（每一个特征0均值、1方差处理）。

②、神经网络的输入与输出

网络输入视频的视觉特征()，输出听觉特征（），即训练一个能由视觉特征计算听觉特征的网络。而后对不同的输出（128维）施加不同的权重，以表征不同特征的重要程度不同，这一步利用另一个conv网络实现。

网络结构分别如下：

具体细节参见代码。

③、相似性判决

这部分把每个样例的120秒拆成了118份，将118份分别取均值后的视觉特征分别输入到网络中，得到输出，再分别与118份真值计算欧式距离的平方（也即mse的优化目标），再乘上conv网络的权重，最后累加。结果越小即越相似。

④、实现

模型定义见dic\_conv.py

训练文件见train\_lrt.py

模型预测与评价见\_evaluate\_lrt.py

⑤、性能评价

时间方面：

本方面测试的开发环境为python3.5+keras2.1.1(tensorflow1.1.0后台)，计算平台为PC端，内存8G，CPU型号为Intel i5-4200M：每次训练（含数据预处理和训练）约耗时8分钟。预测与匹配时间不到一分钟（测试集合大小为60时）。

准确率方面：

经粗略测试，测试集大小为60时，top-5的准确率约80%。运行截图可以参考report文件夹下pic[1-3].png，其使用的测试集参考filenamelists文件夹下的test\_list.txt文件。

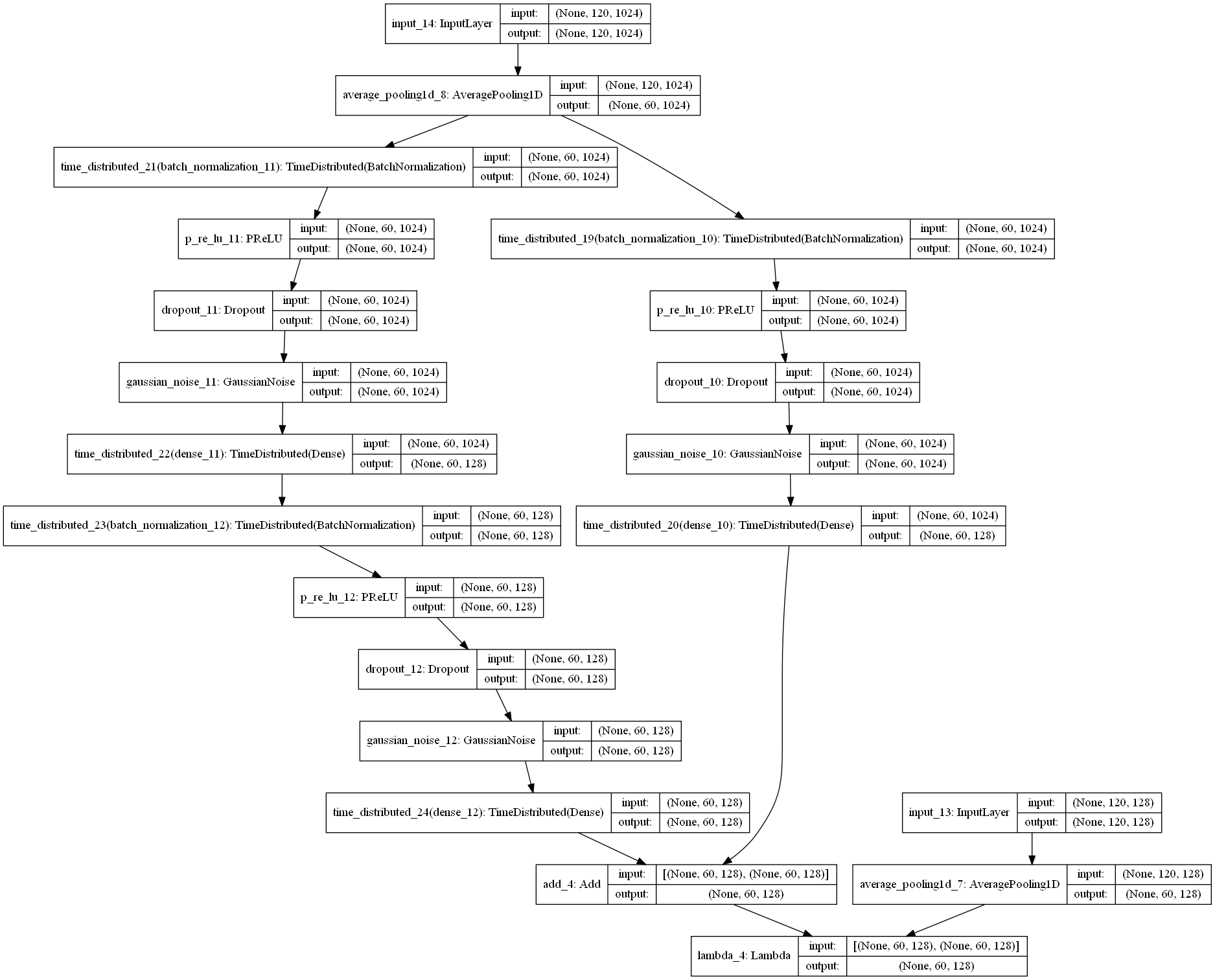
3、采用基于Res变种的网络进行匹配（由林子骞完成）

①、数据预处理

对于视觉、听觉特征，这里以视觉为例，向量的大小是120\*1024，样例共计1300个。对于每一个特征，对其进行规范化（每一个特征0均值、1方差处理）。

②、神经网络的输入与输出

网络输入视频的视觉特征(120x1024)，输出视觉特征预测的听觉特征与实际听觉特征的差值（注，网络第一部对两种特征时间维度pooling使得变为60xdim），所以输出为60\*128网络结构如下：



具体细节参见代码。

③、相似性判决

网络输出即为相似度矩阵，越小说明匹配效果越好。且最终对所得排名矩阵行列进行复合操作，提取其列信息以优化。

④、实现

模型函数 ResFunction.py

训练文件见SSSV2AabdA2V.py

模型预测与评价见\_test\_lzq.py   
⑤、性能评估

训练结果：采用60个样本进行测试，平均top5准确率在80%左右。

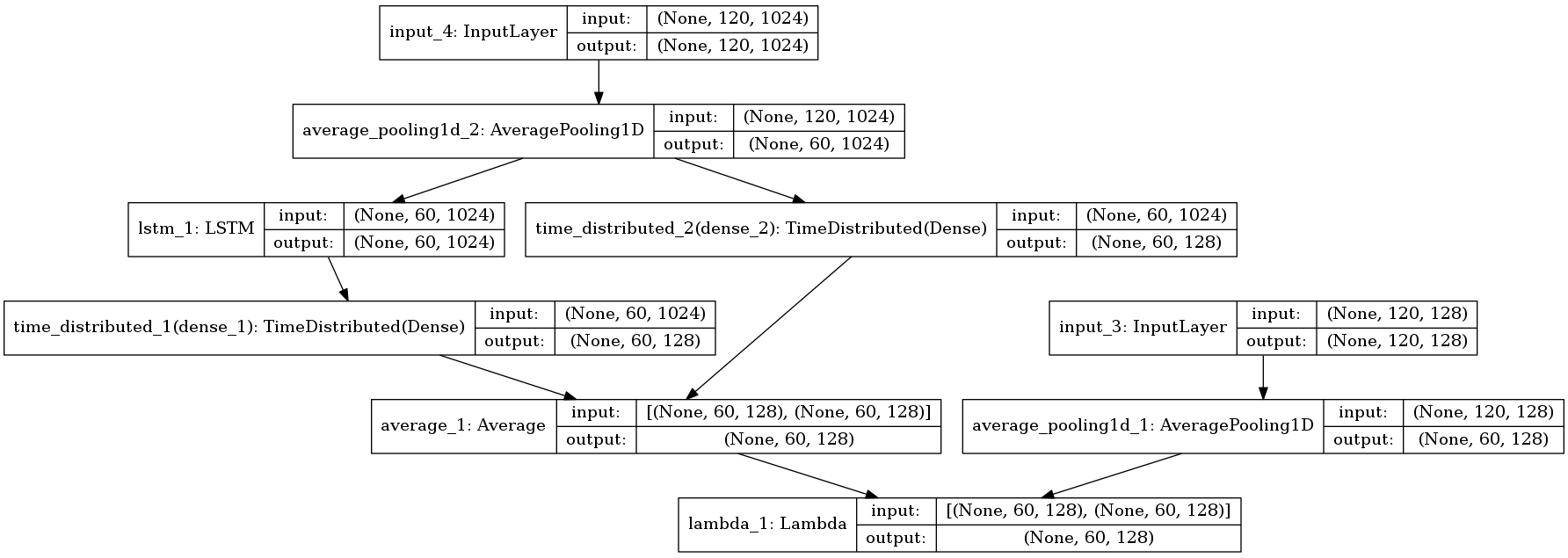
4、采用LSTM网络进行匹配（由陆子阳完成）

①、数据预处理与输入输出接口

数据预处理部分和神经网络输入输出接口与3中网络相同。

②、网络结构

搭建基于LSTM的预测网络以提取视频特征和音频特征在时域上的特性。网络结构如下图所示：



具体细节参见代码。

③、相似性判决

网络输出即为相似度矩阵，越小说明匹配效果越好。

④、实现

LSTMFunction.py 函数文件，编写需调用的函数

LSTMSSS.py 训练主程序

\_LSTM\_lzy.py 测试程序

⑤、性能评估

由于搭建的网络较为复杂，加上LSTM层本身训练耗时较长，因此在服务器上每次训练约耗时40分钟，在本地上对60个样本进行预测输出相似度矩阵约耗时8分钟。

在准确率方面，对60个样本进行预测，top5准确率稳定在75%左右。运行结果及说明参考report文件夹下的LSTM.docx

5、对三个网络的输出结果综合

对三个模型得到的排名矩阵，将行列信息分别复合以优化，然后直接进行求和，最后再复合行列信息优化，作为最终模型的输出。

需要先分别运行：

\_evaluate\_lrt.py 获得lrt\_mtx.npy

\_test\_lzq.py 获得 lzq.npy

\_LSTM\_lzy.py 获得 LSTM\_lzq.npy

最后运行 \_simi\_matrix.py 会打印并保存最终输出矩阵answer.npy

注：值越低相似度越高

**四、工作总结**

在本次实验中，每个组员均完成了各自网络的搭建，因此我们认为在这次实验中，我们的收获是较为全面而巨大的，主要体现在更为深入地了解了机器学习的实际应用，实际对于神经网络进行了参数调整，对于这个过程有了一个初步的认识，且对于python这门语言有了更进一步的了解。

同时，这次实验中我们还存在一些不足，主要体现在：①、没有设计出完整的结合时间信息和空间信息的网络，而是分别训练后对结果矩阵进行合并处理；②、网络的正确率方面仍有提升的空间；③、提取时间信息的LSTM网络时间开销太大，没有找到较好的降低时间开销的方法。