设计题目：视听信息的跨模态匹配

班级：无 55

小组成员：　刘家隆　　2015011116

　　刘可淳　　2015011105

　　石书尧　　2015011103

日期：2017.12.29

一、模型原理与实现

本次实验较为复杂，我们的目的是实现音频信号和视频信号的跨模态匹配。我们所使用的方法是深度学习，此方法在模态匹配等方面已经有了许多成果。经过相关的学习和资料收集，我们有了一些成果。以下是我们创建模型的尝试过程以及最终模型的设计思路。最后还介绍了两个我们实现过但效果一般的网络，我们认为它们还有一定的改进空间。

1. 测试网络的训练

首先我们阅读并理解了所给的框架，并使用测试网络进行了训练。然而网络的 loss 则一直稳定在 0.25，无法降下去，而最终正确率都在 16.67%(5/30)左右，根据 evaluate 的评判标准，这个正确率与完全随机没有区别，说明网络没有起到任何作用。检查 loss 函数后发现，它的两个输入维度不同，网络计算的结果比 target 多了一个数量为 1 的维度，这导致 loss 函数的计算过程与设计不一致，dist 中每个元素与 target 中每个元素都进行了相乘，这样 target 完全没有起到指导训练的的作用。可以证明，此时当且仅当所有输出都为 0.5 左右时，loss 函数的输出达到全局最小值 0.25，这就是此前训练产生随机结果的原因。在 loss 中添加一行降维代码即解决问题。测试网络能达到约 50%~60%的正确率，说明框架工作正常。

2. 最终使用的网络结构

考虑到视频音频的跨模态匹配问题与文本匹配问题有一定的相似程度，如都具有时序性，我们参考了一些 NLP 中的句子建模方面的论文，来构建匹配模型。最终选用的模型主要是参考了 Y Kim 在 2014 年的论文 Convolutional neural networks for Sentence classification[1]，这篇论文主要探讨的是文本匹配问题，Y Kim 采用了交互式处理的思想，首先用一个大型的卷积窗口将两段文本提取成若干个列数为 1 的特征图，之后送入 Max-Pooling 层得到各个特征图的最大值，最后送入全连接+Softmax 层得到匹配结果。

本模型参考了这篇论文的思路，都是通过一个大型的卷积窗口对每一帧的所有特征做卷积，得到若干列数为 1 的特征图。但是与论文不同之处在于，文中的两个模态特征通过的是不同的卷积网络，而我们直接在最开始就把音频特征和视频特征连接起来。虽然这种连接十分简陋，只是单纯的拼接，但是由于卷积核很大，在单个时间点上相当于全连接，因此同一时刻的音频和视频特征已经通过这个卷积网络得到了充分的相互作用。在这之后对 Y Kim 的模型做了一些改进。不同于此论文的思路，由于我们发现测试网络中的 LSTM 层效果很好，而卷积层的输出也是按照时间顺序排列的，每个时间的特征都包含了原有特征的一部分信息，因此将这个结果按照时间分隔输入 LSTM 网络应该能得到较好的结果。首先是将特征矩阵进行降维操作，得到一个帧数*通道数的特征图，然后利用时序性的特点将特征图送入 LSTM 得到一个长度为 128 的特征向量，最后通过全连接层得到匹配分数，其中所有的激活层都采用 Relu 层，模型结构如**图 1.3** 所示。

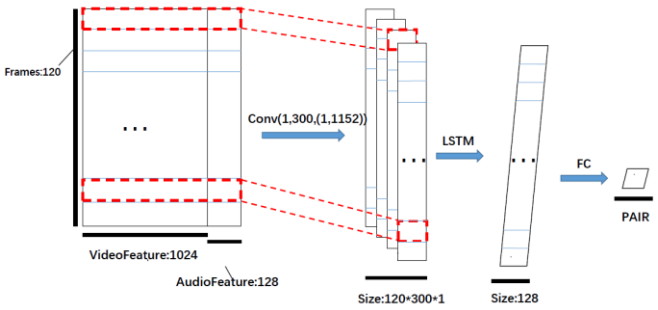
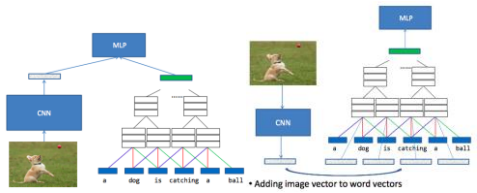


图 1.3 本次试验中设计的网络

2. 一些网络的尝试

（1）参考了 2016 年 ICCV 的论文 Multimodal Convolutional Neural Networks for Matching Image and Sentence[2]，这篇论文使用了 Multimodal CNN 的方法来匹配文本和图片，具体为将文本特征处理成单词、短语、句子层面三种粒度的特征，分别于图片通过 CNN 后得到的特征做匹配，下图展示了单词和短语两种粒度

的匹配模型。在我们自己的模型中，将视频和音频按照相同的结构处理成三种粒度的特征图，进行匹配。但由于网络规模过大，参数过多，训练时无法平衡内存占用过大和舍弃参数导致表达不足的问题，导致最后并没有成功跑通这个模型。该模型的结构图如**图 1.1** 所示。



（a）　　（b）

图 1.1　Multimodal CNN

（2）参考了 2014 年 ACL 的论文 A Convolutional Neural Network for Modelling Sentences[3]，这篇文章的主要思路是先在底层组合邻近的词语信息，然后利用动态 k-max Pooling 的方法，使得句子中相离较远的词语也有交互行为，提取出重要的语义信息，具体结构如下图所示。我们参考了这个论文中动态池化的思路，先将视频和音频的特征串接起来，然后通过卷积层来组合相邻帧之间的信息，接着利用 k-max pooling 组合不同帧的信息，最后通过全连接层得到匹配分数。但在训练过程中，出现了 Loss 一直不降的问题，由于时间比较紧张来不及仔细思考这其中的原因，所以只能放弃这个模型。该模型原理图如**图 1.2** 所示。

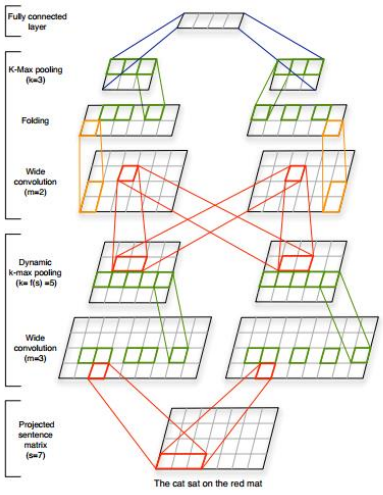


图 1.2　k-max Pooling

二、模型性能分析

我们进行了较长时间的参数调试，这主要包括了学习率参数调试和网络参数调试。

1. 优化器的改进

首先针对 SGD 方法 loss 下降速度过慢的问题，我们改用了 Adam 方法，两者 loss 函数收敛情况**图 2.1** 所示：

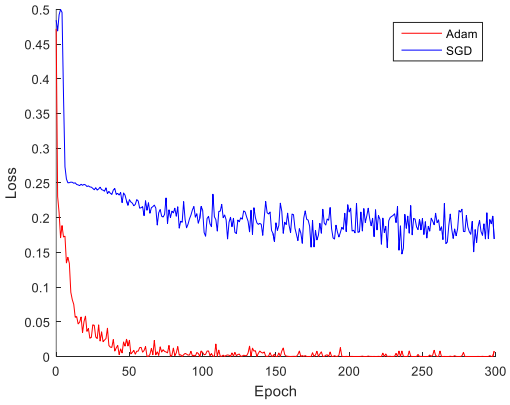


图 2.1 SGD 与 Adam 算法 loss

此外，所给代码中虽然提及了学习率降低这一方法，但有参数未定义，无法正常工作。我们实现了这一部分，学习率 lr 与训练轮数 epoch 的关系为：

$$\mathrm{lr} = \mathrm{lr}(0) * lr\_decay^{epoch/10}$$

通过调整参数 lr_decay 可以调整学习率下降速率。

2. 改变不同的卷积核通道数

在我们的模型中，最重要的参数就是卷积核的通道数。如果这个参数过小可能会导致卷积层丢失太多信息，而如果过大则可能会因为训练数据不足效果降低。

最终正确率结果如**图 2.2**（lr = 0.001, lr_decay = 0.9）：

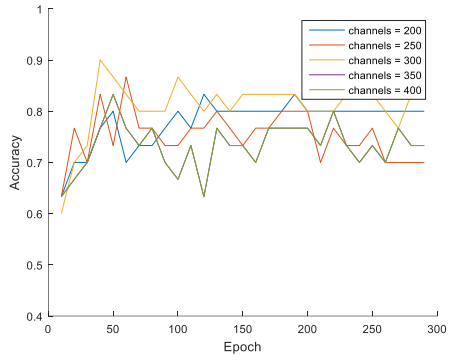


图 2.2 不同数目卷积核训练正确率

我们大致可以判定，取通道数为 300 效果较好。

3. 改变不同的学习率

最终正确率结果如**图 2.3**（channels = 300, lr_decay = 0.9）：

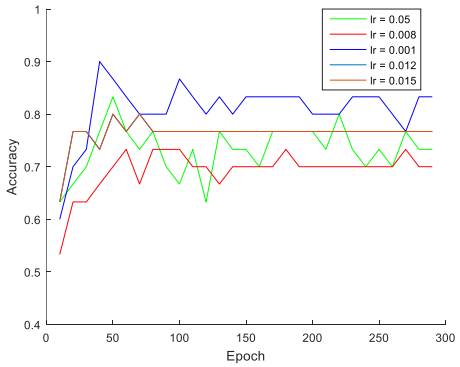


图 2.3 不同学习率训练正确率

可见取 lr = 0.001 效果较好。

4. 改变不同的学习率降低速率

最终正确结果如**图 2.4**（channels = 300, lr = 0.001）：

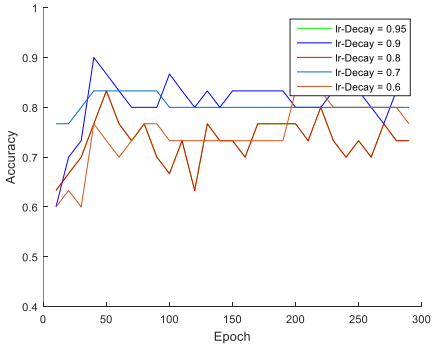


图 2.4 不同学习率下降速度训练正确率

可见取 lr_decay = 0.9 效果较好。

5. 使用更多测试集进行验证

此前测试的数据只有测试集中的 30 组，因此我们还在训练集中额外划分出 30 组进行了训练（lr = 0.001, lr_decay = 0.9, channels = 300），结果如**图 2.5**：

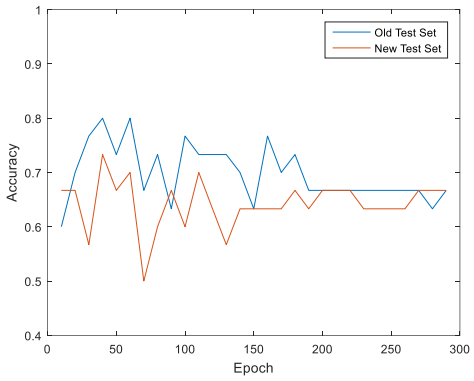


图 2.5 新旧两测试集测试结果

可以看出对于新划分出来的训练集，此网络正确率较低，这可能是因为划分的训练集的前 30 个特征有共同的特性。

综上，此神经网络在 30 个测试集中基本上能保持 77%左右的准确率，而参数调整合适之后正确率能稳定在 83%左右，但在其他测试集中的表现则不能完全保证。

三、实验总结

最终我们选出了 lr= 0.001, lr_decay = 0.9, channels = 300 条件下，某次训练第 100 个 Epoch 时的网络，它在测试集中的表现为正确率 86.7%。这一数据比起随机排序或者一些简单的算法实现要好了很多，这说明我们已经取得了一定的成果。

在本次实验过程中我们从不熟悉 python 语言、从未实现过神经网络、完全不了解 pytorch 框架开始，通过不断的学习、调试、debug，成功实现第一个神经网络，并不断提高它的正确率，我们每个人都在其中起到了作用，也都获得了很多收获。本次实验也让我们对于深度学习和视觉听觉信息模式匹配有了一些初步的了解。这对于我们今后的学习都大有帮助。

但是由于我们对这一领域仍然不甚熟悉，本次实验我们仍有很多不足之处：

1. 正确率仍然不够高。这可能包含了多方面的原因，既可能是神经网络的问题，也可能是特征本身的提取存在问题。我们的神经网络还是比较简单，因此我们相信一定存在更好的方法解决本次实验的问题。

2. 测试集数量太少。我们在调参过程中只使用了 30 组测试集数据检查效果，这不能保证我们的网络在其他数据集上的表现。

3. 对于查找相关论文得到的网络模型，我们虽然进行了尝试修改，但结果并不理想。这可能是因为看起来类似的问题其实并不完全相同，也可能是我们对论文中的想法理解并不透彻，结果很多结构复杂的网络实际表现很不好，另外十分遗憾的是由于时间复杂度或空间复杂度较高，有的网络结构不能正常运行。

4. 没有尝试进行音频和视频特征提取的工作。所给的音频特征和视频特征并非我们自己提取，这样我们对于它的特性本身并不了解，如果我们对特征提取更熟悉一点，

四、小组分工

刘可淳：文献搜集、模型设计讨论、模型代码编写、参数调试讨论
刘家隆：文献阅读、模型设计讨论、模型代码编写、参数调试
石书尧：文献阅读、模型设计讨论、参数调试、报告撰写

五、文件清单

report.pdf	课程设计报告
VA_METRIC_FINAL.pth	训练好的模型文件
models.py	模型定义文件
evaluate.py	evaluate 测试接口

六、参考文献

[1] Y Kim.Convolutional neural networks for sentence classification [D]. 《Eprint Arxiv》, 2014.
[2] L Ma,Z Lu,L Shang,H Li. Multimodal Convolutional Neural Networks for Matching Image and Sentence [D].IEEE,2015.
[3] N Kalchbrenner,E Grefenstette,P Blunsom. Convolutional Neural Network for Modelling Sentences [D]. 《Eprint Arxiv》,2014.