****

**学生研究计划(SRP)项目验收**

**项目成员个人结题报告**

|  |  |
| --- | --- |
| **参加项目名称：** | **多媒体数据的情感信息挖掘** |
| **参加项目编号：** | **X201910561102** |
| **参加起止时间：** | **2019 年 3 月至 2020 年 3 月** |
| **指导教师姓名：** | **王伟凝** |
| **指导教师所在学院：** | **电子与信息学院** |
| **学生姓名：** | **成泽森** |
| **学生学号：** | **201730271412** |
| **学生手机号：** | **17306619318** |
| **学生所在学院：** | **电子与信息学院** |
| **学生所学专业：** | **电子科学与技术** |
| **填表日期** | **2020年3月** |

|  |  |
| --- | --- |
| **教务处** | **制** |
| **2020年3月** |

**撰 写 说 明**

**1．本报告的文件名请修改为：学号 （注意文件名不要出现中文）**

**2．本报告要求项目组每位成员单独撰写，共性部分内容可以相似。**

**并提交至系统（**[**http://110.65.10.252/cxxl/Index.aspx**](http://110.65.10.252/cxxl/Index.aspx)**）。**

**3．本“项目成员个人总结报告”建议提纲：（一）文献调研（二）研究方案（三）项目组成员间分工（四）本人主要工作和成效（要细化，并将相关成果如论文首页，专利/软件著作权，竞赛获奖证书等扫描附在本文档后）（五）关于本课题下一步研究工作设想（六）参与科研工作的感想等。**

***以上提纲仅供指导教师参考，各位指导教师可自行制定，但封面须统一用本模板*。**

**4．原则上要求本研究报告不少于2500字，项目成员间报告的共性部分不得多于40%，相关成果可以成员间重复使用。**

## Introduction

近年来，情感计算 ( Affective Computing) 在 AI 的研究中逐渐开始火热起来，情感计算是一个交叉的研究领域，其涵盖了从计算机科学到心理学，从社会学人类学等众多学科，情感计算被分为情绪分析 (Sentiment Analysis) 和情感识别 (Emotion Recognition) 两种[1]，情绪识别更倾向于分析多媒体信息中人所带有的感情信息，而情感识别则更倾向于识别多媒体信息带给阅读多媒体信息的人的情感。这样的区别则带来了应用场景的不同，情绪分析主要应用在文本以及 speech 这种想要获取多媒体信息的语义信息的场景，而且情绪分析本就是属于自然语言处理这个大类之下，而情感识别则用于推荐系统等想要获取情感传递信息的场景，情感识别相比起情绪分析来说，其涉及的领域会相对更广，像人脸表情，自然场景，音频分析，文本分析都是可以研究情感识别的领域。

目前情绪分析相关领域的研究要更多一些，目前社交媒体平台，即时通讯软件，购物平台等文本密集型产品的推出，社会学，心理学，政策学，管理学等领域根据文本挖掘信息的研究越来越多使得情绪分析的研究呼声越来越高，实际上情绪识别更加偏向于挖掘观点，即理解文本中人们在想什么，所以情绪识别往往又是指观点挖掘 (Opinion Mining)[2]，在当前社会无时无刻在进行高密度的文本信息流动，情绪分析作为一种从文本中收集提取人类观点强有力的手段有着很高的研究价值[3]。情感识别虽与情绪识别同属于情感计算，相对来说研究要偏少一点，这主要是因为情感识别这个学习任务更加困难，需要提取更加低级的更加抽象特征。目前情感识别的主要研究分为演讲，对话方面的语音，文本情感识别，例如Mirsamadi S et.al.[4] 通过循环神经网络并加入了本地注意力机制来提取演讲音频特征，然后通过使用多种不同的分类器如 SVM, CNN来对音频特征进行分类，在这一方面当前最新的研究还有将图卷积神经网络 (GCN) 用于对话情感识别的案例[5]。然后还有就是电影，广告，图片甚至还有动图[6]方面的图像，视频情感识别，例如 Y Fan et.al.[7] 通过提出一种全新的卷积神经网络来提取自然场景与人脸场景的多尺度特征而来进行情感识别。目前研究倾向于多模态的情感识别，即将上述两个方面结合起来分析情感，例如 S Albanie et.al. [8] 通过将音频特征与人脸关键点等图像特征融合起来预测演讲视频的情感。

就此项目而言，我们更加专注于情感识别这个情感计算的子领域，目前的情感识别又被分为分类型以及回归型两个子任务，要想完成分类型以及回归型两个子任务，我们需要对应的情感模型，目前被广为认可的有离散情感模型- Ekman 的七种基本情感: 高兴、悲伤、惊讶、恐惧、愤怒、厌恶、中性，其对应的是分类型的任务，还有连续情感模型 – Valence-Arousal 二维情感模型，其对应的是回归型任务，在本项目的策略中，我们需要对多媒体信息的连续情感进行分析所以我们更加关注于 VA 二维连续情感模型。

在介绍完项目的背景之后，我们把聚焦回到项目上，本项目主要是希望能够构建一个能够根据电影中的场景画面以及音频信息来进行广告插入，而想要实现这样一个系统，我们将该系统的流程划分为：场景切割，视频中的多媒体信息的情感分析，广告中的多媒体信息的情感分析，广告插入决策，输出被插入广告的视频，对广告插入效果进行评估。而我负责的部分主要是视频中的多媒体信息的情感分析，以及广告中的多媒体信息的情感分析，所以我需要专注的正是情感计算的情感识别任务，考虑到插入策略要求情感分析能够提供 VA 连续情感，所以我需要专注的是情感识别的回归型任务。

情感识别的回归型任务，其实到目前为止还是没有一个大体上准确率比较高的解决方案，一般此任务的解决方案是通过对 EEG 信号进行分析，例如，或者是针对人脸面部表情来进行分析，显然这与系统构建场景不相符合，我们不可能获取观看视频和广告的人的 EEG 信号，人脸表情也不一定会出现在广告和视频中，我们需要的是一种能够对自然图像直接进行回归获取情感 VA 值的网络，当然除开计算机视觉还有语音分析这一方案，我们的研究主要关注点在于图像信号，语音信号只是作为图像信号分析的补充。

针对以上问题，我采用了两个方案来实现如何从自然图像中直接分析得到 VA 值，第一个方案主要是基于卷积神经网络与卷积自编码器来提取特征，然后通过将这些特征送入全连接层来进行回归；第二个方案主要是通过卷积神经网络来提取特征向量，而后以五秒为单元将特征向量送入循环神经网络中，循环神经网络输出最终的预测值。方案均采用Mediaeval-18 竞赛的电影数据集 Liris-accede[9]作为训练集与测试集。

## Scheme 1 Frame-based Emotion Recognition[10]

### Method

方案一为基于卷积神经网络而直接对图像进行特征提取，能够以单秒为单元来获取情感，然而这样的方法的缺点是只能够分析图像特征，而且不能捕获视频前后帧之间的关联，显然这样的方法会在准确度上有所损失，但是其优点在于需要的信息少，我们采用在 Imagenet[11] 已经预训练的 Resnet50 作为主要的特征提取网络，Imagenet 中包含大量的自然场景图像，通过在 Imagenet 上预训练，能够让网络有着对自然图像分类特征的先验信息，显然这些先验信息是能够对情感识别起到帮助的，这些先验信息能够提供一些具象物体的信息，而具象的物体信息在某些程度上是能够反应情感信息的，例如花会带来高愉悦度，而刀会带来高唤起度，当然更多的可能是不蕴藏着情感信息的，例如香蕉，苹果等等，我们将 Resnet 50 最后的池化操作换为 SPP[11] ,空间金字塔池化 (spatial pyramid pooling) 操作，以求再将卷积获得的信息压缩编码的时候能够能够获得多尺度的信息，而后我们训练一个自编码器，将其编码器部分作为特征信息的补充，整个方案的 pipeline 如下：

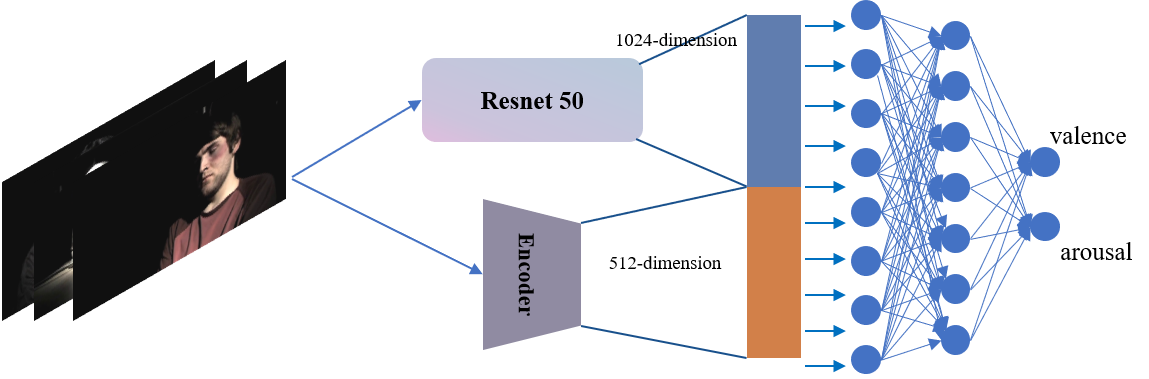


图1 Scheme 1 pipeline

其中 Resnet 50[12] 的结构细节如下，依然保留了原始的残差结构不过选用的是Resnet 二代，即采用 bottleneck 残差结构，这样的残差结构有更多的卷积层而不引入更多的参数，有助于网络提取更多的低级特征：

表1 Resnet 50-spp 网络结构

|  |  |  |
| --- | --- | --- |
| **Layer name** | **Output size** | **50-layer** |
| conv1 | 112x112 | 7x7, 64, stride 2 |
| conv2.x | 56x56 | 3x3, max pool, stride 2 |
| x 3 |
| conv3.x | 28x28 | x 4 |
| conv4.x | 14x14 | x 6 |
| conv5.x | 7x7 | x 3 |
|  | 1x1 | Spatial pyramid pooling,  1024-d fc, softmax |

其中 SPP 模块的细节为:

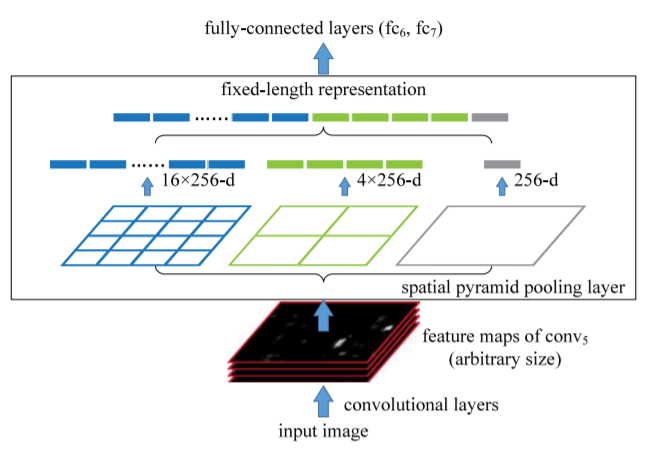


图2 Spatial Pyramid Pooling 模块细节图

自编码器的结构图为：

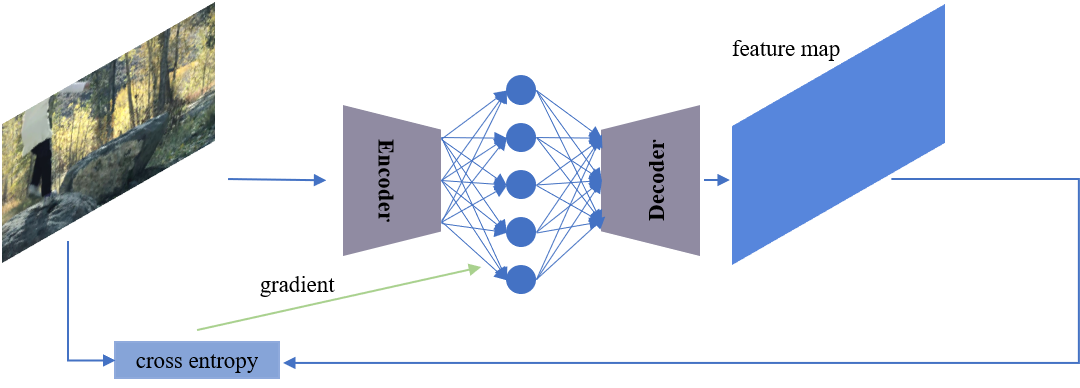


图3 Auto Encoder 结构图

### Experiment

首先对 Liris-accede 数据集进行处理，通过对电影逐帧遍历，每一秒提取一帧，帧与帧之间为等间距的，通过这样处理之后获得了一系列带有 Valence 与 Arousal 标注的时间上连续的图像，在输入神经网络训练之前，我们采取了预处理，通过 z-score standardization 来正则化图像，保留图像的灰度分布特征而不是图像的灰度特征，这样能够使得学习任务的难度下降，问题规模缩小。而后我们将训练集以 0.2 的比例分为训练集与验证集，通过 Adam 优化器对 AE 与 CNN 两个网络进行训练，初始学习率都设置为 0.001，batch size 也都设置为 4， 训练 20 个 epoch 之后选取在验证集上取得最佳结果的模型文件进行测试：

表2 Scheme 1 测试结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Valence | | Arousal | |
| MSE | r | MSE | r |
| Proposed method | 0.0833 | 0.454 | 0.0933 | 0.372 |

## Scheme 2 Sequence-based Emotion Recognition[13][14]

### Method

方案二为通过卷积神经网络来提取以五秒为单元的图像特征，组成一个序列然后将其输入到 RNN 中进行学习得到了处理后的特征，然后将这些特征通过全连接层进行回归，这样的结构能够将很好地与其他特征，例如，音频特征 concatenate 起来，然而单帧图像与音频图像无法建立直接联系所以并不妥，首先展示网络的 pipeline 结构：

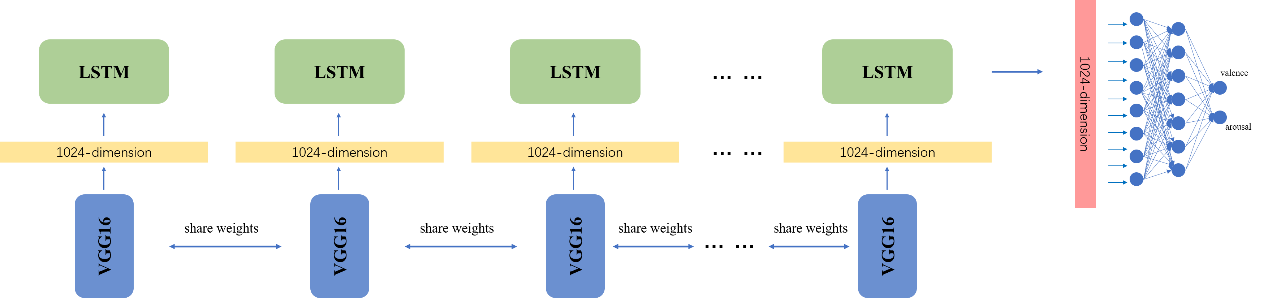


图4 Scheme 2 pipeline

这样的结构比起单纯地直接对单帧进行分析更能够提取 context 信息，即语境的信息，实际上 LSTM 还可以换成 BI-LSTM ，这样可能有更强地提取前后语义信息的能力，能够更高效地获取低级抽象特征。

### Experiment

在已经将电影拆成以秒为单位的单帧图片的基础上，我们将连续的五秒，即连续的五帧图像打包，并将五帧图像的最后一张作为要预测 VA 的图像，我们将每一秒图像都处理成了一个将这一秒与前面四秒打包的图像块，对于某一秒图像所在的秒数不足以采样前四秒的情况我们通过填充全黑图像来打包，将最后一秒的 VA 值作为这个图像块的标签，而后我们使用 Adam 优化器，选取初始学习率为 0.001，batch size 为 2，训练 20 个 epoch，从中取出在验证集上测试最好的模型文件进行测试：

表3 Scheme 2 测试结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Valence | | Arousal | |
| MSE | r | MSE | r |
| Proposed method | 0.0789 | 0.473 | 0.0937 | 0.381 |

## Conclusion

本体项目我主要负责关于情感计算的神经网络框架方面的实现，采用了两种方案，第一种方案是基于卷积神经网络和自编码器来提取视觉特征并通过全连接网络来进行回归分析，第二种方案是基于循环神经网络来提取视觉特征并通过全连接层进行回归分析，目前方案只能比较粗粒度的对 VA 进行回归，一方面这是情感分析本身的复杂度决定的，另一方面是方案的实现太过于单薄导致的，之后对于方案的改进主要在于对第二种方案的拓展，对于 CNN-RNN 中 CNN 方面，可以将输入到 RNN 的特征向量进行调整，例如将在人脸数据集 WIDER FACE上预训练的 CNN 提取的视觉特征与在目标检测和语义分割数据集 Pascal voc 上预训练的 CNN 提取的视觉特征以及自然图像数据集 Imagenet 上预训练的 CNN 提取的视觉特征作级联，而后再输入到 RNN 当中，RNN 方面可以引入更新的 RNN 单元，此外，在 CNN-RNN 提取的视觉特征的基础上还可以引入音频特征，将音频特征作为最后全连接层回归的特征信息补充。

### Acknowledgement

首先，非常感谢 srp 项目的指导老师 – 王伟凝老师，通过老师的指导，我掌握了文献阅读，文献检索等方面的知识，以及科研中的常识问题，并且也为我的神经网络相关方面的知识的学习提供了帮助；然后感谢 srp 项目中的师兄师姐，在论文阅读比较困难的时候，能够提供帮助，在方案实现比较困难的时候，也能够提供帮助，十分可靠；然后感谢与我一起参加 srp 项目的队友们，感谢大家能够一起努力为这个项目付出。

## Reference

1. Poria S, Cambria E, Bajpai R, et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37: 98-125.
2. Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends® in Information Retrieval, 2008, 2(1–2): 1-135.
3. Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1253.
4. Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 2227-2231.
5. Ghosal D, Majumder N, Poria S, et al. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation[J]. arXiv preprint arXiv:1908.11540, 2019.
6. Chen W, Rudovic O O, Picard R W. Gifgif+: Collecting emotional animated gifs with clustered multi-task learning[C]//2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2017: 510-517.
7. Fan Y, Lam J C K, Li V O K. Video-based emotion recognition using deeply-supervised neural networks[C]//Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018: 584-588.
8. Albanie S, Nagrani A, Vedaldi A, et al. Emotion recognition in speech using cross-modal transfer in the wild[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 292-301.
9. Baveye Y, Dellandrea E, Chamaret C, et al. Liris-accede: A video database for affective content analysis[J]. IEEE Transactions on Affective Computing, 2015, 6(1): 43-55.
10. Quan K A C, Nguyen V T, Tran M T. Frame-based Evaluation with Deep Features to Predict Emotional Impact of Movies[C]//MediaEval. 2018.
11. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
12. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
13. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
14. Ebrahimi Kahou S, Michalski V, Konda K, et al. Recurrent neural networks for emotion recognition in video[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 2015: 467-474.
15. Fan Y, Lu X, Li D, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 445-450.