# 项目报告

## 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。

所以就此项目而言，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，就计算机视觉方面一般此任务的解决方案是通过对 EEG 信号进行分析，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Mediaeval18 电影数据集作为训练集。

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

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 已经预训练的 Resnet50 作为主要的特征提取网络然后将最后的池化操作换为 SPP[3] ,空间金字塔池化 (spatial pyramid pooling) 操作，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：

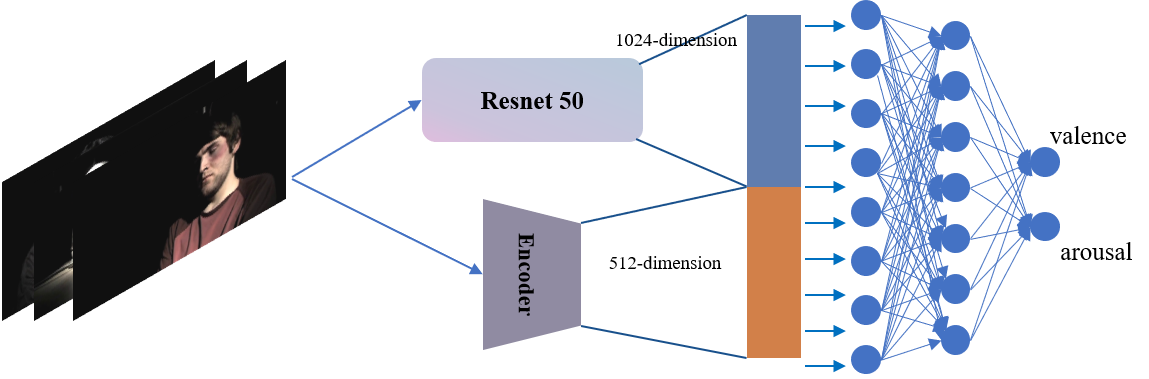


图1 Scheme 1 pipeline

其中 Resnet 50[4] 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:

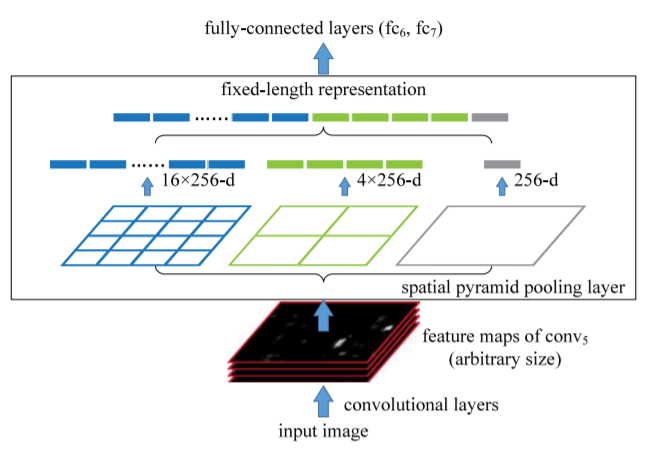


图2 Spatial Pyramid Pooling 模块细节图

自编码器的结构图为：

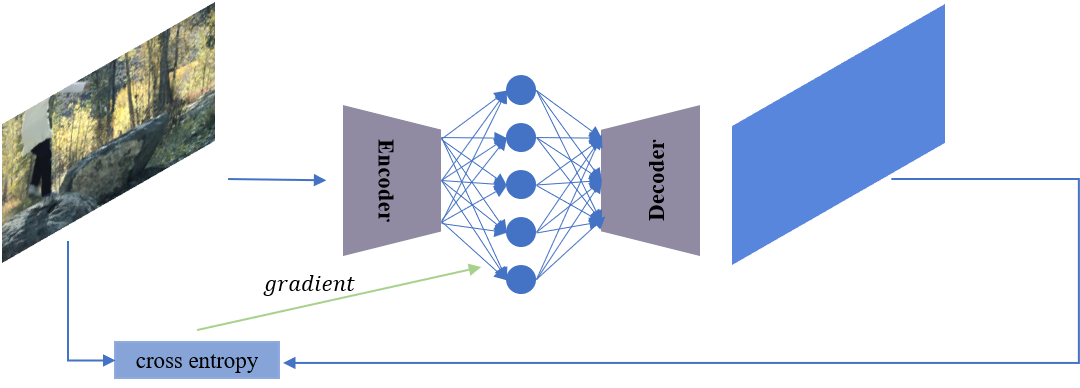


图3 Auto Encoder 结构图

## Scheme 2 Sequence-based Emotion Recognition[5][6]

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：

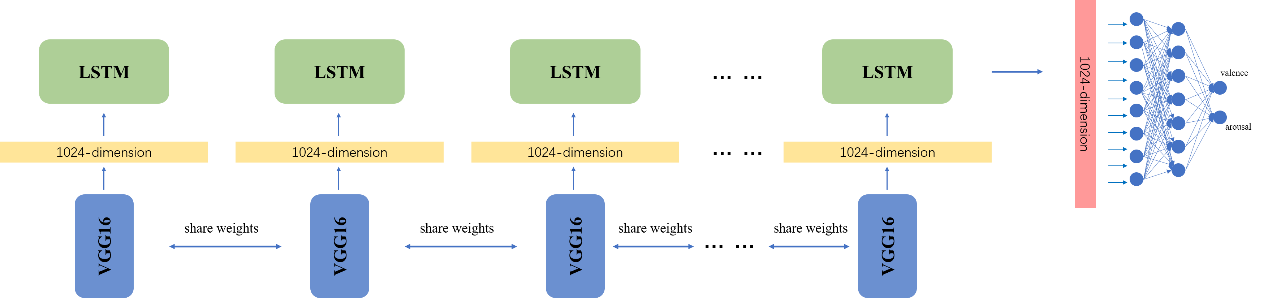


图4 Scheme 2 pipeline

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。

## 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. 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.
3. 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.
4. 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.
5. 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.
6. 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.