华南师范大学

《人工智能导论》课程项目

课 程 项 目 报 告

项　目　题　目：“猫眼慧识”——基于人脸图像的情绪分析

所　在　学　院：计算机学院

项　目　组　长：李文杰

小　组　成　员：王瀚业 刘乐 江明伦

开　题　时　间：2023 年 3 月 1 日

一、引言

人脸表情识别是近年来计算机视觉领域中备受关注的课题之一，它以分析和识别人类表情为基础，能够利用神经网络提取人类情感信息，在人机交互、远程医疗、安防监控等领域具有重要的应用价值。本项目旨在研究基于图像处理的人脸表情识别方法，通过深度学习等技术实现对人脸表情的自动化识别。本项目使用 FER2013 数据集，采用卷积神经网络(CNN)、VGG 和残差网络(ResNet)等深度学习模型，结合数据预处理、特征提取、模型训练和评估等方法，最终实现了对人脸表情的自动化识别，可以为人脸情感分析领域的部分研究提供可行的技术方案和实现手段。

二、国内外研究现状

人脸情绪识别是近年来受到全球研究关注的热门领域之一。国内的众多研究机构和高校都在该领域积极进行相关研究。较为知名的有中国科学院自动化研究所使用集成学习进行人脸情绪识别，南京大学应用对抗生成网络进行人脸情绪识别等等。

同时，国外的研究机构和企业也在人脸情绪识别领域有所布局。麻省理工学院、卡内基梅隆大学等知名学府在该领域开展了相关研究。微软公司开发了一种情感文本到表情图像的自动转换技术，苹果公司推出了一种基于深度神经网络的人脸表情识别技术等等。这些研究结果和技术应用对未来的人脸情绪识别领域有着重要的推动作用。

三、模型和算法

1、CNN（卷积神经网络）

CNN（Convolutional Neural Networks,卷积神经网络）由一个或多个卷积层、池化层以及全连接层等组成。与其他深度学习结构相比，卷积神经网络在图像等方面能够给出更好的结果。这一模型也可以使用反向传播算法进行训练。相比较其他浅层或深度神经网络，卷积神经网络需要考量的参数更少。

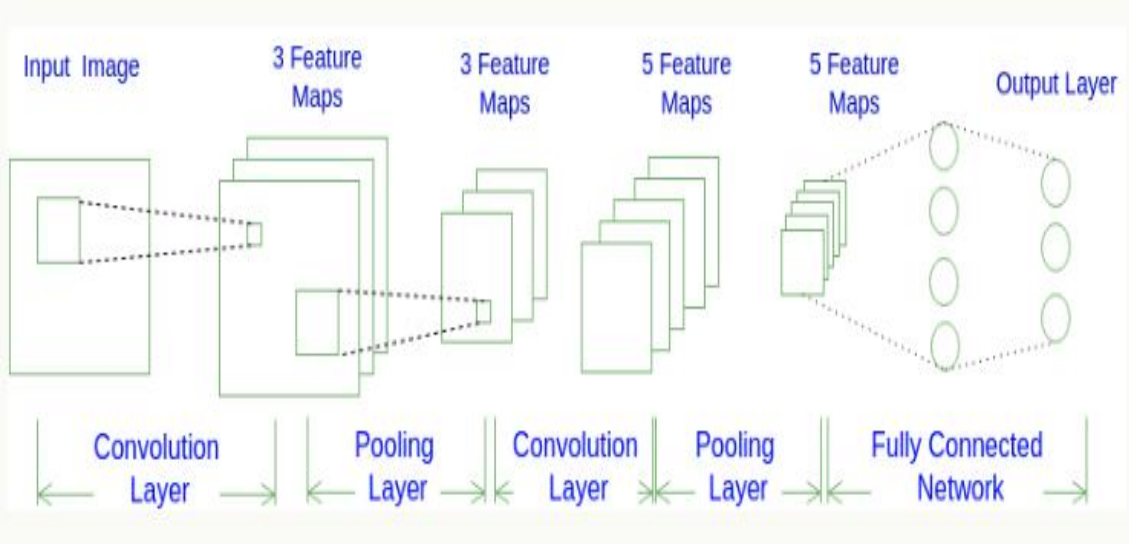


图 1　CNN 结构示意图

这是我们在项目初期使用的一个模型。该模型在算法上存在两个较为典型的问题：

（1）忽略图像的二维特性。

（2）常规神经网络提取的表情特征鲁棒性较差。

于是我们试图寻找一种在提取人脸图像表情特征方面表现性更优的模型。

2、VGG

VGG 是一种深度的卷积神经网络模型，从图像中提取 CNN 特征，VGG 模型是首选算法。VGG 网络的主要工作证明了增加网络深度能够在一定程度上影响网络的最终性能。根据卷积核大小和尺寸的不同可以分为六种配置方式，其中最为人所熟知的配置方式为 VGG16 和 VGG19 两种结构。

VGG 采用连续的小卷积核代替较大卷积核，以获取更大的网络深度。例如，使用 2 个 3×3 卷积核代替 5×5 卷积核。这种方法使得在确保相同感知野的条件下，ＶＧＧ网络具有比一般的ＣＮＮ更大的网络深度，提升了神经网络特征提取及分类的效果。

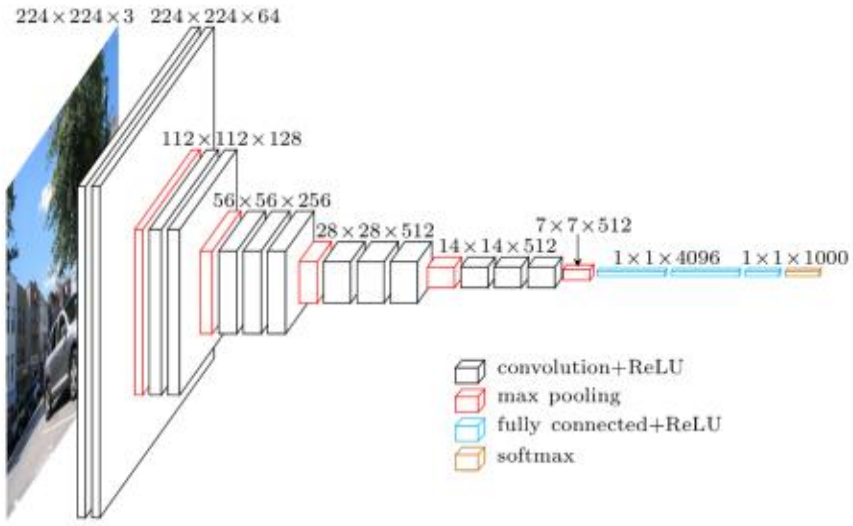


图 2　VGG 结构示意图

在使用 VGG 进行模型训练过程中，也存在一些无法忽视的问题。例如，参数量大，需要更多的存储空间；另外，VGG 使用较多的卷积层，这会导致参数数量的急剧增加，并且由于 VGG 使用的是较小的卷积核，卷积操作的次数更多，也会增加模型的计算量。在训练时发现模型的 GPU 使用率明显上升，导致训练时间更长，这对我们调整模型参数和评估不同参数下模

型的效果产生了很大影响。

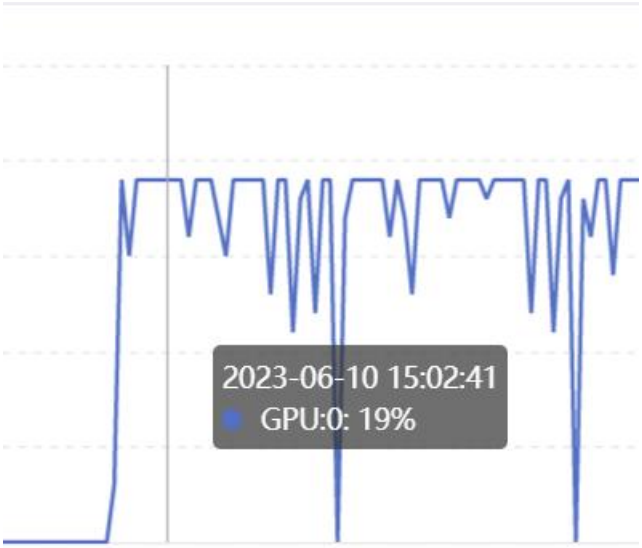


图 3　VGG 模型训练时的 GPU 使用率情况

所以我们决定再次尝试寻找一种在训练效果和训练时间上都具有良好表现的模型。

3、ResNet（残差网络）

ResNet（Residual Network，残差网络）的主要贡献是发现了“退化现象”，并针对退化现象发明了 “直连边/短连接”。使用了残差结构，主分支使用了三个卷积层，第一个是 1×1 的卷积层用来压缩 channel 维度，第二个是 3×3 的卷积层，第三个是 1×1 的卷积层用来还原 channel 维度。同样在捷径分支上有一层 1×1 的卷积层，它的卷积核个数与主分支上的第三层卷积层卷积核个数相同，注意每个卷积层的步距。ResNet 已经被广泛运用于各种特征提取应用中，它的出现解决了网络层数到一定的深度后分类性能和准确率不能提高的问题，深度残差网络与传统卷积神经网络相比，在网络中引入残差模块，该模块的引入有效地缓解了网络模型

训练时反向传播的梯度消失问题，进而解决了深层网络难以训练和性能退化的问题。

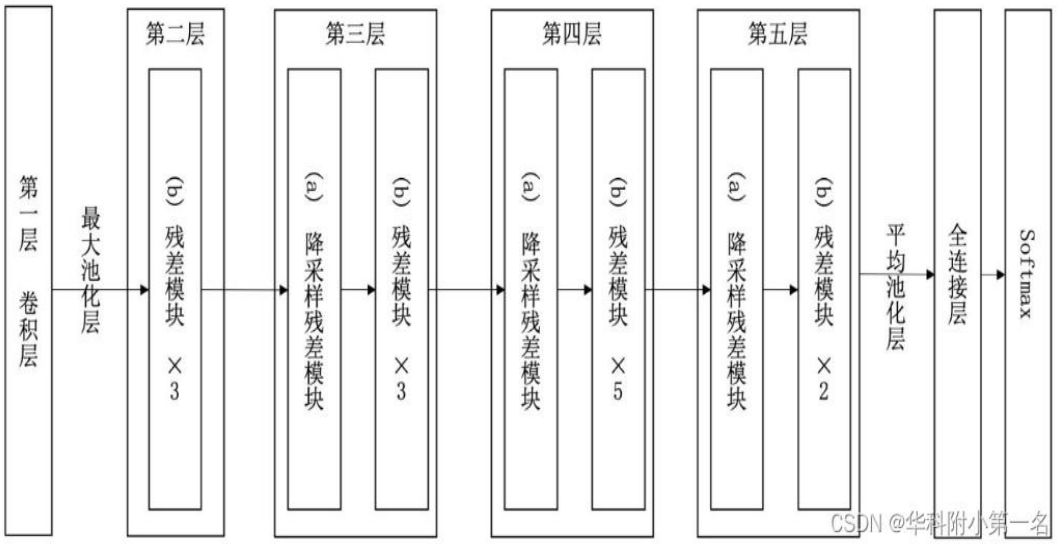


图 4　ResNet 结构示意图

有了前两次模型训练的经验，我们在对 ResNet 模型的训练更加熟练。首先，在模型训练的 Epoch 选取上，我们先是设定一个较大的 Epoch，然后将每一个 Epoch 中模型的训练集准确率和验证集准确率绘制成折线图，以达到直观且清晰观察准确率变化的目的。根据观察，在 Epoch=60 时模型的训练集准确率增长率较小，而验证集准确率基本稳定在 60%，再经过多次测试与对比，确定将 ResNet 设置为 60 个 Epoch。训练集和验证集的准确率变化折线图的绘制，为模型确定了一个合理的 Epoch 参数值，有效防止了欠拟合和过拟合现象的发生。

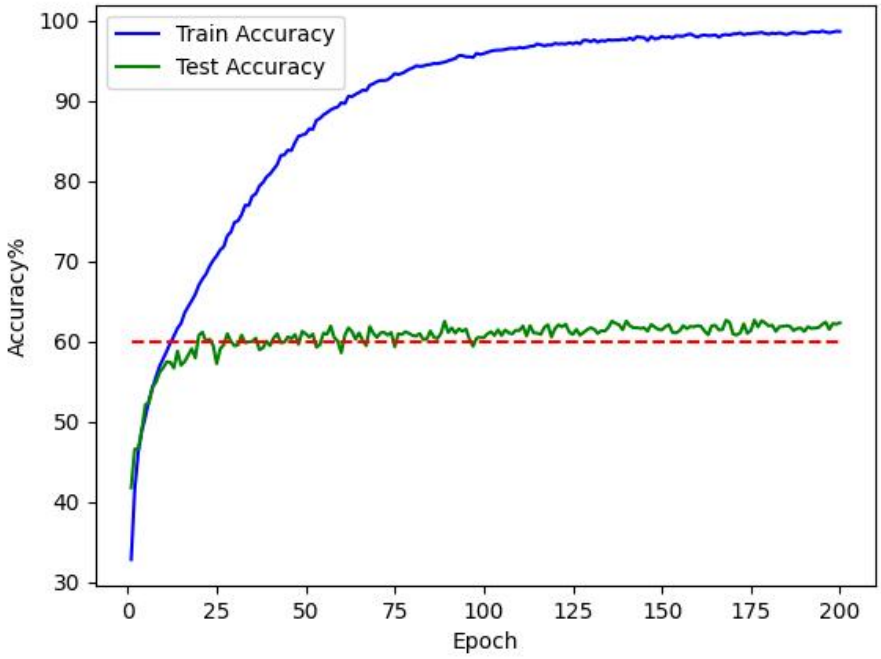


图 5　训练集和验证集在不同 Epoch 下的准确率变化折线图

在数据集方面，我们进行了数据增强，通过使用 transform.compose 函数，对图像进行反转，拉伸，裁剪等等，进一步提高训练效果。

相比于 VGG 等传统的卷积神经网络，ResNet 使用的卷积层数更深，模型参数更少；卷积核的大小相对较大，减少了卷积操作的数量，能够简化网络学习的难度，降低训练时的 GPU 使用率，有效地缩短模型训练的时间。

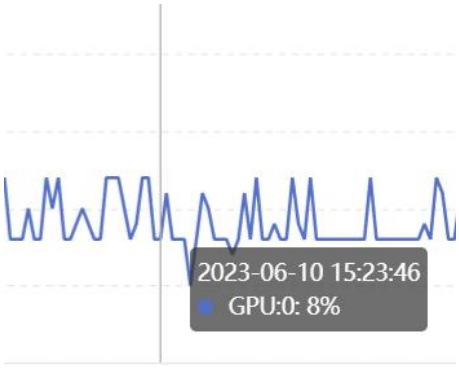


图 6　ResNet 模型训练时的 GPU 使用率情况

四、实验结果分析

1、CNN 模型

（1）部分效果展示

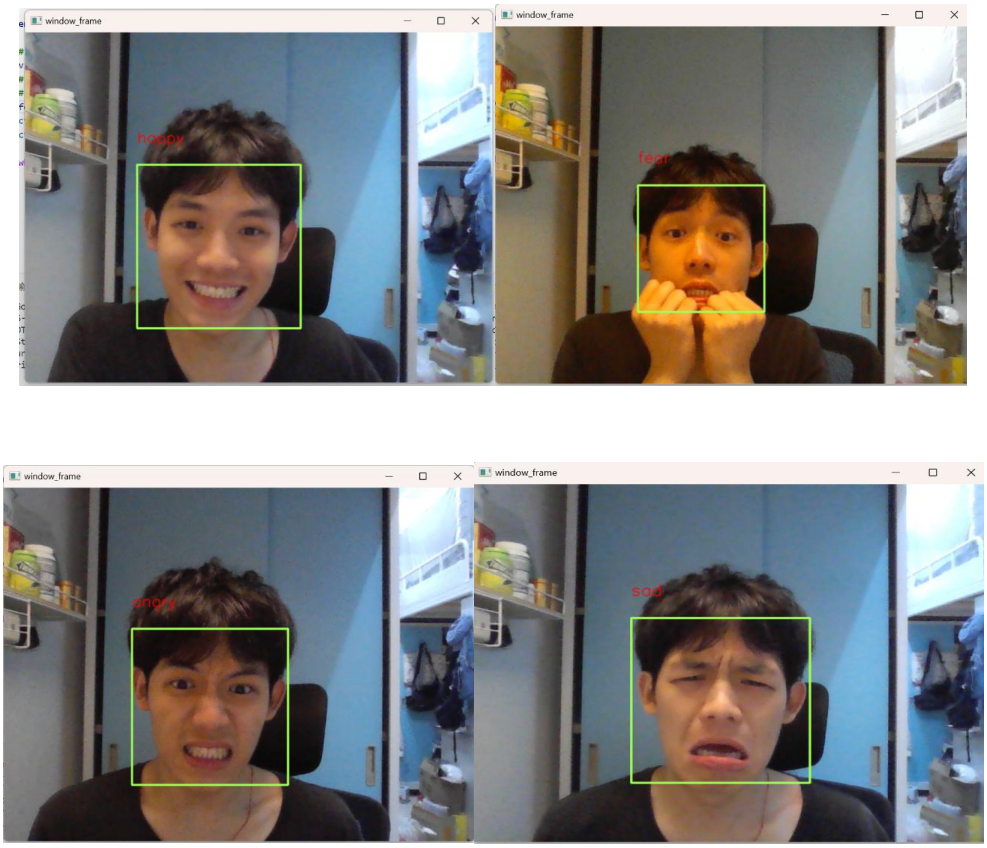


图 7　CNN 模型部分效果展示

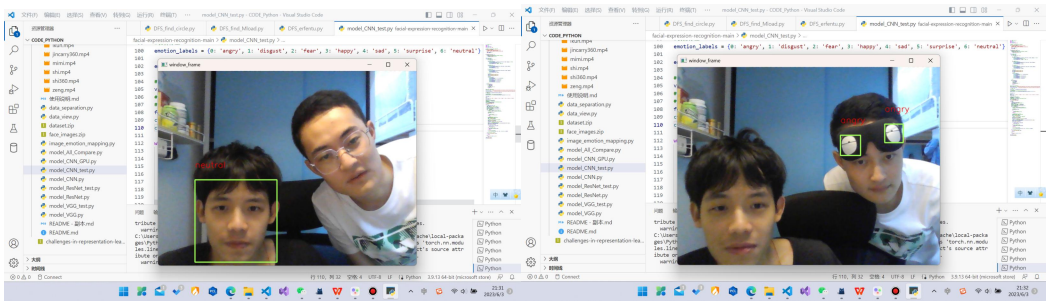


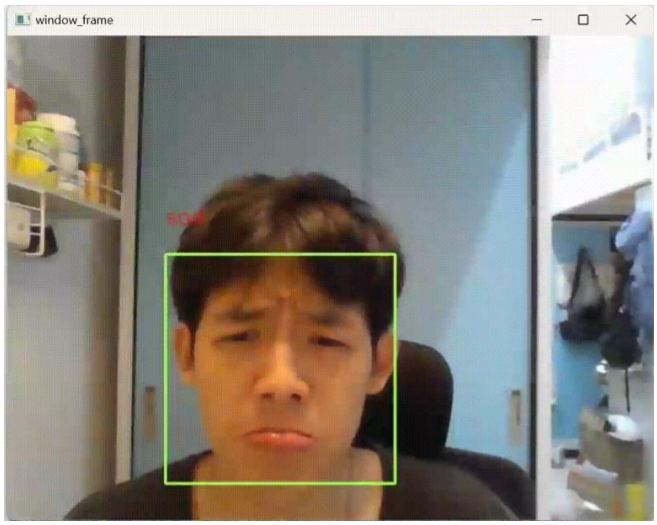
图 8　CNN 模型的缺陷

（2）结果分析

可以看到，CNN 模型已经具有比较优秀的人脸识别能力。但是，在后两幅图中，我们也看到，CNN 模型存在两个问题：忽略图像的二维特性，常规神经网络提取的表情特征鲁棒性较差。这间接导致了训练出的模型在出现多个人脸时有概率识别不出，甚至识别错误（例如将图片中的眼罩错误的识别为人脸）。

2、VGG 模型

（1）部分效果展示



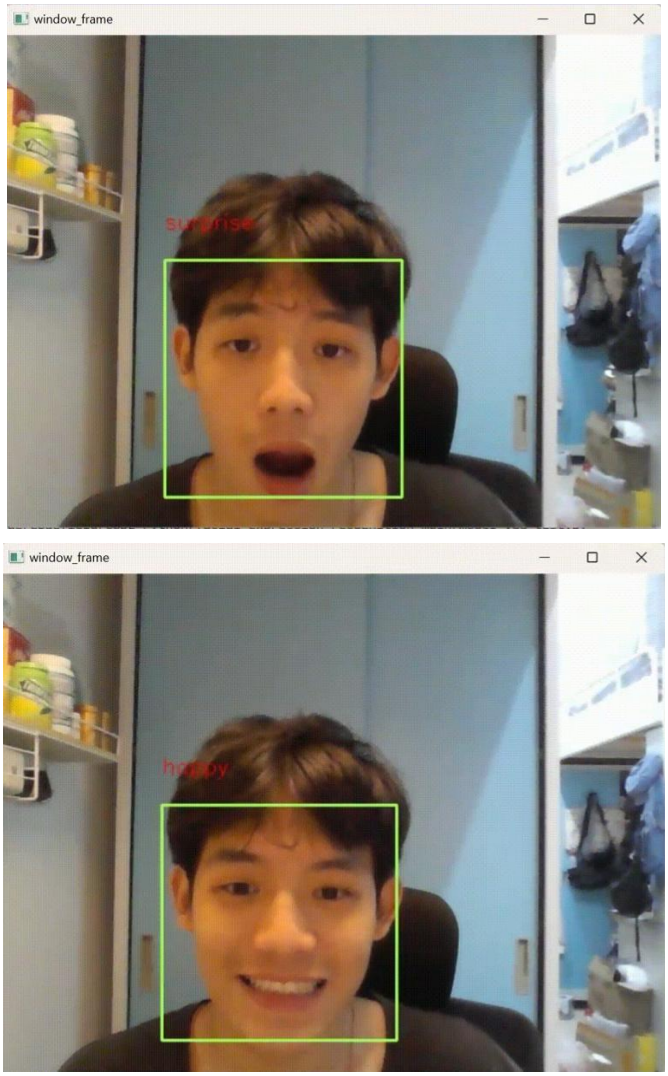


图 9　VGG 捕捉视频中连续变化的表情

（2）结果分析

由于ＶＧＧ网络具有比一般的ＣＮＮ更大的网络深度，提升了神经网络特征提取及分类的效果，所以理论上 VGG 模型的效果将会由于 CNN。而在实际体验中也是如此。在做测试时，我们能够明显的感受到，使用 VGG 训练出的模型识别比 CNN 模型更加灵敏。在连续的表情变化中，VGG 模型往往可以做出更加灵敏快速的反应，同时识别多个人脸并且准确度进一步提升。但是 VGG 模型鲁棒性仍存在不足且网络模型训练时反向传播时会存在梯度消失问题，再加上模型计算量大，所以模型具有进一步优化的空间。

3、ResNet 模型

（1）部分效果展示

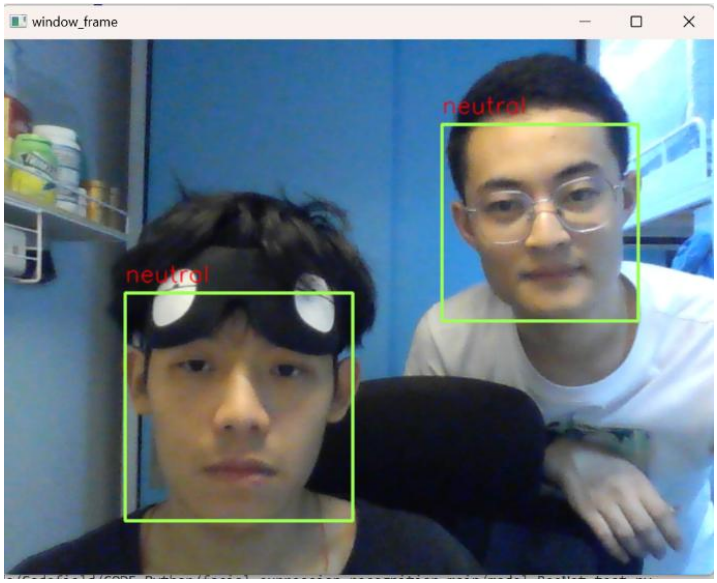


图 10　ResNet 在防漏识别、错识别方面的优化效果

图 11　ResNet 模型捕捉电影片段演员微表情变化

（2）结果分析

对于我们最终训练完成的 ResNet 模型，缓解了网络模型训练时反向传播的梯度消失问题，鲁棒性较好，而且在训练时间相对较短的同时还能保持较好的性能，进而解决了深层网络难以训练和性能退化的问题，从上图中可以看出，相对于 CNN 模型，该模型在防止漏识别、错识别方面表现较好；在微表情以及头部动作幅度和表情变化幅度较大的场景下也能保持较好的捕捉和识别能力。

五、结论

通过本项目的实验研究，我们使用了三种不同的深度学习模型（CNN、

VGG 和 ResNet）来实现基于人脸图像的情绪分析。实验结果表明，三种模型在人脸识别和情绪分析方面都取得了良好的效果。

首先，CNN 模型在人脸识别方面表现出较好的能力，但存在忽略图像二维特性和提取表情特征鲁棒性较差的问题。其次，VGG 模型相较于 CNN 模型，在图像特征提取和分类方面取得了进一步的提升。它具有更大的网络深度，能够更好地提取和表示人脸表情特征。最后，我们采用了 ResNet 模型，通过引入残差结构解决了深层网络训练和性能退化的问题。实验结果显示，ResNet 模型在人脸识别和情绪分析方面具有较高的准确率和灵敏度，对多人脸的捕捉和准确识别表现出优异的性能。

综上所述，基于我们的实验结果，我们得出以下结论：

1、CNN 模型、VGG 模型和 ResNet 模型都能够有效实现基于人脸图像的情绪分析。

2、在三种模型中，ResNet 模型在人脸识别和情绪分析方面表现出最佳的性能。

3、深度学习模型的选择对于人脸情绪分析的准确性和灵敏度具有重要影响，需要根据具体需求选择合适的模型。

本研究为人脸情绪分析领域提供了初学者可行的技术方案和实现手段，为初学者相关研究和应用提供了有益的参考和借鉴。

六、参考文献

[1] M. M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in Proceedings of the 8th International Conference on Automatic Face and Gesture Recognition, 2008, pp. 1-6.

[2]Z. Zhong, L. Zheng, and S. Zhang, "Recovering 3D face shape and texture using deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5163-5171.

[3]M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," Journal of Multimedia, vol. 1, no. 6, pp. 22-35, 2006.

[4] S. Wu, Y. Zheng, X. Zhang, and X. Xue, "A face recognition method based on deep learning with single sample per person," Multimedia Tools and Applications, vol. 77, no.