

编号\_\_\_\_\_



长春理工大学

Changchun University of Science and Technology

## 本科生毕业设计

### 基于卷积神经网络的社交图片识别与分类实现

Social Images Recognition and Classification Based on  
Convolutional Neural Networks

学生姓名	刘思源
专业	计算机科学与技术
学号	150511428
指导教师	赵建平
学院	计算机科学技术学院

二〇一九年六月

## 毕业设计（论文）原创承诺书

- 1 . 本人承诺 : 所呈交的毕业设计 ( 论文 ) 《基于卷积神经网络的社交图片识别与分类实现》 , 是认真学习理解学校的《长春理工大学本科毕业设计 ( 论文 ) 工作条例》后 , 在教师的指导下 , 保质保量独立地完成了任务书中规定的内容 , 不弄虚作假 , 不抄袭别人的工作内容。
- 2 . 本人在毕业设计 ( 论文 ) 中引用他人的观点和研究成果 , 均在文中加以注释或以参考文献形式列出 , 对本文的研究工作做出重要贡献的个人和集体均已在文中注明。
- 3 . 在毕业设计 ( 论文 ) 中对侵犯任何方面知识产权的行为 , 由本人承担相应的法律责任。
- 4 . 本人完全了解学校关于保存、使用毕业设计 ( 论文 ) 的规定 , 即 : 按照学校要求提交论文和相关材料的印刷本和电子版本 ; 同意学校保留毕业设计 ( 论文 ) 的复印件和电子版本 , 允许被查阅和借阅 ; 学校可以采用影印、缩印或其他复制手段保存毕业设计 ( 论文 ) , 可以公布其中的全部或部分内容。

以上承诺的法律结果将完全由本人承担 !

作者签名 :           年 月 日

## 摘要

随着“互联网社交文化”的盛行，最初集中在贴吧与论坛等亚文化圈之中的社交图片也渐渐为人们所熟知与认可，这些社交图片在今日被称作“表情包”，并被大量应用于日常交流之中。

它们的素材大多来自于对影视作品或生活中某些人物的有趣瞬间的捕捉，再通过绘图处理或添加配文等方式进行二次创作便可赋予这些社交图片“灵魂”与新的活力。

且区别于“Emoji”和其他图片，社交图片是由网友们自发创作并通过互联网传播开来的，它们可以通过简单的静态内容来承载复杂的信息与内涵，具备单纯文本语言所不具有的感官属性，可以使交流更加立体化，让使用者更加精准和便捷的表达自己的真实状态或内心想法，并使接收者更易代入和理解。

这顺应了现代社会人们的交流需求与趋势，因此社交图片的内容和规模在近些年呈指数式爆炸增长。但这带来的问题是，当面对着一个特定的对话场景，我们往往需要在终端上众多可供选择的内容中翻找出一个或多个最合适以进行使用，但往往这些可使用的内容中夹杂了大量无关或并不适用该场景的干扰选项，如我们日常拍摄的照片或从其他途径保存下来的图片。

虽然我们生来就可以以一种当下任何计算机都无法真正比拟的“神秘”方式快速理解与整合各种抽象内容，但我们的局限也很明显——大多数人完全无法直接依据某种特定的数理规则对大量数据进行不觉枯燥与疲倦的计算处理并得出准确的结果集。且随着社交图片在生活中的使用愈加频繁，对开发一种敏捷高效的方法根据特定场景中人们的主观需求对社交图像的深层语义进行检索的需求变得愈加迫切。

而面对这种依赖大量主观决策的抽象需求，基于深度学习的计算机视觉是我们的不二之选。在此将依据特定的标签规则为数据建立标签库，随后通过一系列的图像处理方法对数据进行归一化操作，最后构建卷积神经网络并使用这些样本进行监督学习，尝试实现基于卷积神经网络的社交图片的识别与分类。

**关键词：计算机视觉 图像处理 深度学习 卷积神经网络 情感识别 情感计算**

## Abstract

With the popularity of ‘Internet Social Culture’ in China, specific images that used to be made, generally with hilarious content, and spread rapidly in subcultures like post bars and other kinds of online forums, have also been gradually accepted and then appreciated by the masses. Furthermore, these ‘Social Function Pictures’ are called ‘memes’ today, and largely used in daily communication.

Most material of them come from capturing amusing moments of some characters in the film and television works or someone in our life, and further processing like secondary creations by fusing with linear hand-painted styles or adding subtitles will add new energy and put the soul into these ‘Social Function Pictures’.

The biggest difference between them and ‘emojis’, and other pictures is, these ‘Social Function Pictures’ are spontaneously produced by net friends, then spread via the Internet without external interventions. They carry exceedingly complicated information and connotation by simple static content, hold the sensory characteristics which text language doesn’t have. They can make our chat more three-dimensional, make users to be more accurately and conveniently when expressing their inner thoughts and real situations, and it is far easier for receivers to understand and feel the empathy.

The appearance of ‘Social Function Pictures’ perfectly meets the demand of the Internet era, so their content and scale have been exponentially increasing in recent years. An accompanying problem is, we usually need to pick out one or more suitable options from a large number of images with a given situation. However, these suitable options are not adjacent for most of the time, there will be a lot of distractions between them, for example images we saved from other apps.

Notwithstanding we can quickly comprehend and integrate extremely abstract content in enigmatic approaches that no single computer can truly compare, our limitations are also very obvious, it is totally impossible for most of us to exert calculation on the mass data by giving an appointed mathematical logic structure and formula, even if it is quite simple and we can assume that we will never feel tired and bored about this, subsequently come out the exact result set.

The more frequently we use ‘Social Function Pictures’ in our daily life, the eagerer are the needs for us to develop an agile and efficient method that can retrieve the deep-seated semantics of them by the subjective wants of people in specific situations.

While we face the abstract needs which depend on a lot of subjective decisions, computer vision based on deep learning is a clever choice for us. Hence, we are going to establish a labelled database according to the specific labelling rules, with subsequent serials of ways of images processing to normalise our data, and finally to construct an appropriate neural network in order to launch surprised training by use those samples. Try to implement social images recognition and classification based on convolutional neural networks.

**Key word:** Computer Vision; Image Processing; Deep Learning; Convolutional Neural Network; Emotion Recognition ; Affective computing

# 目录

摘要.....	I
Abstract .....	II
第 1 章 绪论 .....	1
1.1 研究目的和意义 .....	1
1.2 国内外研究现状 .....	2
1.2.1 前言 .....	2
1.2.2 国外研究现状.....	3
1.2.3 国内研究现状.....	3
1.2.4 国内外文献评析 .....	5
1.3 论文研究内容.....	5
第 2 章 实验设计 .....	6
2.1 模块划分 .....	6
2.2 实验目的 .....	6
第 3 章 研究方法 .....	7
3.1 数据搜集方法与来源 .....	7
3.2 数据预分析与现象总结 .....	7
3.3 情感模型建立方法.....	12
第 4 章 算法设计 .....	14
4.1 数据预处理 .....	14
4.1.1 前言 .....	14
4.1.2 相机位置估计 .....	14
4.1.3 面部平均化 .....	16
4.1.4 图像预处理 .....	17

4.2 深度学习 .....	18
4.2.1 构建标签 .....	18
4.2.2 构建卷积神经网络 .....	20
第 5 章 结果分析 .....	22
5.1 实验结果 .....	22
5.2 结果分析 .....	24
总结与展望 .....	25
致谢 .....	26
参考文献 .....	28

## 第 1 章 绪论

### 1.1 研究目的和意义

社交图片大多依附于出于特定的兴趣爱好而形成的圈子，不同的圈子间流通的社交图片大不相同。比如存在于流量明星偶像效应下的粉丝圈大多使用其“爱豆”的图片并配上文字。而 ACG 圈子中则更喜欢偏向于卡通化风格的社交图片，且内容大部分是出自于圈子内所流行的“梗”，且不同年龄段的人群所使用的社交图片风格也有极大差异（“梗”原指让人记忆颇深且耐人回味的经典桥段，但随着网络与人们的日常生活羁绊的日益加深，其词义也在被不断扩大引申，其可以是抽象的，并且大到某个时间点或某片地理区域与人群的，也可以是具体的，小到某部作品中的情节插曲的）

区别于上述依托兴趣爱好圈的社交图片，网络中也存在着一批具有极强普适性的社交图片，其热度并不依托于商业化的 IP，而是满足了人们摘要中所提到高效表达主观感受的刚需。因此在这里我们将主要选取这类数据进行分析。

人类能够通过各种不同的身体姿态，面部表情，并可以配合发出声音来有意识的或无意识的表达出极其复杂的内容，任意一部分的细微变化就能够引导出截然不同含义。且发出声音的重音位置，语速，语音语调的变化之多更是令人咂舌。这些含义可以是更适合意会的纯粹的情感表露；也可以是能够被直接记录的客观信息，如语言文字，但同时必定会附带很多“额外内容”。我们并不擅长隐藏这些包含了很多我们内心的真实想法与深层情绪，即使这些内容是被相对削弱过的，我们依然能够从中获得很多的信息。

不过，在正常生理状态下，我们大部分情绪的产生与变化都是出于某些外界刺激，并进行连续变化的；但在分析计算机图像时，我们却是脱离因果，直接接触中间某些瞬时过程，这为这项工作添加了很多的不确定性。当我们在对其进行分析时不免会不自觉的尝试补全前因后果，去完善包含这个图片所捕捉的瞬间的全部剧情，这样确实能更好的带入并理解图片中角色的完整心理历程。但很显然，我们在单单浏览一张内

容陌生的图片时，并没有足够的信息去支撑这种补全，我们只能凭借我们单方面的臆想进行“满足个人主观意向的补全”，又由于不同人群在主观上的情感认知与审美喜好是很难通过统一标准进行量化从而进行评估与判断的，每个人都拥有不同的共情能力，三观，人生经历，这就会为结果带来极大的不确定性，因而构建出一个足够精妙的情感模型来在对图片进行分类时尽可能的抵消主观判断上的差异是非常有必要的。

并且近乎每一张图片都在艺术上具备一定的情绪传播能力，无论内容是有生命的，还是无生命的。虽说无生命的图片也能为我们带来极其明显的感官体验，如广袤的宇宙，无垠的沙漠，或者是单纯的色彩与线条的集合。但是出于人类本能的“亲生命性”，我们更喜欢包含着人，动物或植物的图片，因此被用于日常交流的照片大部分也都是围绕于此，我们更能在生命身上找到自身的共鸣，尤其当面对我们的同胞时，我们能够更好的通感，共情。这点同样适用于动漫风格的图片。

所以在此我们并不打算从艺术层面去分析每一张图像所包含的情感，因为这种决定将成几何倍的增加这项任务的工作量与难度。就像上文提到的，使用相对来说更容易代入的内容或许是个明智之选，也就是从能够更加高效传播情感的，包含我们人类，或者其他动物内容的图片入手以展开实验。通过图片中角色的身体姿态，面部表情，并辅之以部分场景内容，判断其包含的“梗”以及所蕴含的情绪，或者所能传达出的情感。从而推断出这张图片可以被应用于何种对话场景之中。

## 1.2 国内外研究现状

### 1.2.1 前言

卷积神经网络是近年来兴起的一种人工神经网络与深度学习理论相结合的模式识别方法，目前已经成为图像分类领域中的研究热点之一。与传统的图像分类方式不同，卷积神经网络不需要针对特定的任务对图像提取具体的手工特征，而是模拟人类的视觉系统对原始图像进行层次化的抽象处理来产生分类结果。该方法采用局部感受野，权值共享和空间采样技术，使得网络的训练参数相比于神经网络大大减少，而且对图像具有一定程度上的平移，旋转和扭曲不变性，已广泛应用于语音识别，人脸识别，手

写体识别，行人检测等应用领域。

### 1.2.2 国外研究现状

来自多伦多大学的 Alex Krizhevsky, Ilya Sutskever 与 Geoffrey E.Hinton(2012), 训练了一个大型的深度卷积神经网络, 将 IMAGENET LSVRC-2010 竞赛中的 120 万幅高分辨率图像分类为 1000 个不同的类别。在测试中分别获得了 37.5% 和 17.0% 的错误率, 这比此前最优秀的结果还要好得多。该神经网络包含 6000 万个参数和 65 万个神经元, 由五个卷积层组成, 其中一些卷积层随后是最大池化层, 还有三个全连接层, 最后是 1000 维的“Softmax”。为了使训练更快, 在其中使用了一种非饱和神经元和一个非常高效的卷积运算的 GPU 实现。且为了减少全连接层中的过拟合, 其采用了一种全新的正则化方法, 称为“dropout”, 并被证明是非常有效的。在 ILSVRC-2012 竞赛中, 其还加入了该模型的一个变种, 获得了 15.3% 的测试错误率与 26.2% 的测试错误率。

来自卡内基梅隆大学的 Zhiding Yu 与微软研究院的 Cha Zhang(2015), 提出了一种基于图像的静态面部表情识别方法, 用于“Wild Challenge”(EmotiW) 挑战中的情绪识别。其将重点放在 SFEW2.0 数据集的分支挑战上。在该数据集中, 其试图将一组静态图像自动分类为 7 种基本情绪。该方法包括一个基于三种最先进的人脸识别技术的检测模块, 以及一个多重深度度卷积神经网络集成的分类模块。每个 CNN 模型都是随机初始化的, 并在 2013 年面部表情识别 (FER) 挑战所提供的更大数据集上进行了预训练。然后又在 SFEW2.0 的训练集上对预先训练的模型进行微调。为了将多个 CNN 模型结合起来, 其提出了两种学习网络响应的组合权重的框架: 通过最小化对数损失与铰链损失。其提出的方法在 FER 数据集上取得了最优秀的结果。SFEW2.0 的验证和测试集也分别达到了 55.96% 和 61.29%, 超过了挑战基线设定的 35.96% 和 39.13%

### 1.2.3 国内研究现状

华南理工大学电子与通信工程专业刘剑聪(2014), 为解决图像美学评价的计算用时问题, 基于已有的工作成果, 将图像美学特征提取过程进行并行化处理, 提升特征提取速度, 并且在阿里云弹性云服务中实现了特征提取, 分类器和评价模型的部署,

构建了一个基于云平台的智能终端图像美感分类与评价系统。针对人物类图像，该学者考虑了人类视觉对人脸的独特感观，提出了新的美学特征。在已有特征的基础上，针对性地设计了人脸区域美学特征，结合全局特征，显著区域特征，采用 SVM 算法建立美感分类器。同时，该学者通过对图像主体区域以及区域分割线的检测，依据常用美学法则，采用基于改进的基于样例的图像修复算法对图像进行主体大小与位置调整，以及采用细缝裁减算法对区域分割线的位置调整，同时对分割线一侧区域进行具有自适应内容保护作用的拉伸操作，实现了对社交图像的构图的优化调整。

哈尔滨工业大学赵思成(2016) 针对图像情感计算中的上述问题进行研究，基于艺术学相关理论，期望提取更具有判别力更容易理解的情感特征；利用社交媒体数据进行以用户为中心的个性化情感预测，探索社交媒体中影响情感感知的因素；对图像情感的分布进行建模，预测一幅图像在多位观察者中所诱发情感的分布情况；研究图像情感在计算机视觉，多媒体技术等领域的应用。该学者指出，通过引入艺术学等相关学科的研究，可以提取出更具有判别力且容易理解的特征，从而提高图像情感识别的准确率；社交媒体中图像情感的感知是个性化的，并且受到时间演变，社交上下文等多种因素的影响，综合考虑这些因素可以显著提高情感预测的性能；从概率分布的角度对图像情感进行建模，是对个性化情感与大众化情感的折中，更符合实际情况，更具有实际意义。

华南理工大学电子与通信工程专业黄杰雄(2018) 提出了一个基于多层次深度卷积神经网络的图像情感分类框架，从图像的主体，颜色，局部信息等多个层次学习图像的情感特征，提升了图像的情感评价性能。而且，该学者还提出了 ResNet-GCN 网络和基于 ResNet-GCN 网络的图像情感分类框架。利用 GCN 结构的大尺寸卷积核提高网络的学习视野，让网络更有效地学习图像的情感特征，提升了框架的图像情感分类性能。此外，该学者为了解决人物社交图像的美学评价研究问题，该学者提取了 78 维人脸美学特征，并结合深度学习方法学习得到的图像特征和已有研究的综合美学特征，利用决策融合提升了图像的美学评价性能。同时，为了更好地将深度学习方法应用到

这一研究中，该学者构建了一个新的人物社交图像美学评价图库。

#### 1.2.4 国内外文献评析

综上所述，国内外研究学者都对卷积神经网络与一些关于图片识别与分类的算法进行了深入的分析与研究。不过，国外学者大多从卷积神经网络的定义和当前使用的领域出发，对卷积神经网络与其他同类型识别方法进行对比，从而提出自己的看法。而国内学者大多基于卷积神经网络的识别方法，并结合具体案例，从而发挥卷积神经网络的作用，并对卷积神经网络在图像识别当中的具体运用提出自己的看法。

### 1.3 论文研究内容

通过自建样本数据标签与基于卷积神经网络的计算机视觉，尝试根据图片所包含的情感类型，以及所属于的互联网 IP 类型进行分类，通过立体情感模型并结合以及高级人机交互方式进行单一或组合主观模糊检索，最后根据不同人对不同社交图片所代表内容的强度等级的不同对输出结果进行个性化调整，以提升人们日常在应用社交网络与社交媒体中寻找适用于特定对话情景的社交图片时的效率。

## 第2章 实验设计

### 2.1 模块划分

实线箭头代表模块间的逻辑交互顺序，而虚线箭头代表此处引入了并非来自上一模块的内容。

我们将制定出样本数据的标签规则并由人工处理生成标签样本库，而后搭建出合适的卷积神经网络模型使用已标签数据进行训练，生成权重模型并储存。接下来选取部分并未被学习过的标签样本进行预测，当预测结果达到一定正确率的时候则认为模型可取。最后实现按照需求对从标签样本库中单一或组合检索，返回推荐内容。

而我们的核心难题为如何通过调整标签方法，数据质量，网络结构，学习率，激活函数来提升我们的预测结果的准确率，从而提升推荐内容的准确率。并且将尝试分析在实验进行中遇到的偏向心理学以及社会科学的内容并得出结论。

### 2.2 实验目的

此实验起源于大二时的一个想法，旨在通过技术来节约日常在使用社交软件与社交媒体时的小烦恼。将本科所学计算机视觉内容与即将开始的研究生生活中将接触的机器学习相结合，并涉及到较为感兴趣的认知科学。以此来提升自己对于开展学术研究的能力与经验，加深自己对于上述三个区域的知识的储量。并将此项目代码与样本标签库开源，希望能够为领域做出一些力所能及的微薄贡献。

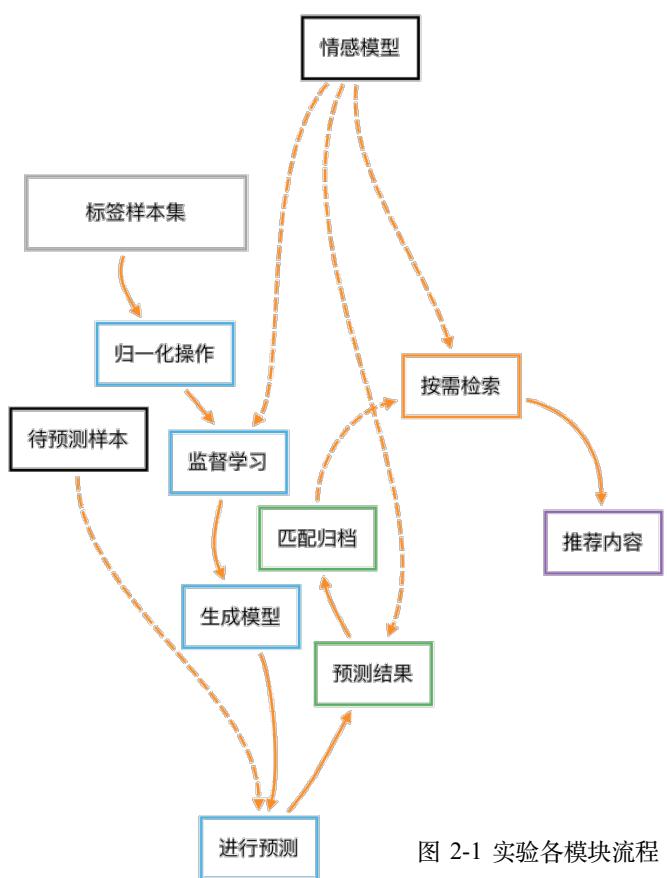


图 2-1 实验各模块流程

## 第3章 研究方法

### 3.1 数据搜集方法与来源

数据是人工智能的食粮，投喂以优质的数据能使其快速茁壮的成长。为了获得足够的优质数据，我向两名挚友寻求了一些必不可少的帮助。在他们的手机与电脑中积累了 100GB 有余的，由众多社交软件缓存在本地的图片，如“腾讯 QQ”，“微信”，“百度贴吧”与“新浪微博”等，并且从未进行过清理。且这其中的大多数图像属于我们所需要的样本，而且相对于直接从图库中或社交媒体上使用网络爬虫遍历下来的样本，这样收集到的样本更具有更优秀的时效性。

这些文件将作为接下来的探索和实验的原材料。事实上，我们过去经常互相分享这些有趣的图片，不单单是为了好玩，我们都认为及时更新我们的“库存”是非常重要的，丝毫不夸张的说这可以帮助我们不被这个飞速运转的社会淘汰。每个人都面临着“万物联网”，闻所未闻的全民的热点话题下一秒就会突然出现在人们的眼前。互联网不再只是年轻人才需要掌握的技术，尽管我们不愿意承认，即便是老年群体也广泛存在着这种现象，如果我们停止使用任何的“社会功能图片”，谈话就会变得非常尴尬奇怪（并且我认为在这种情况下，“emoji”乃至颜文字也应该被算作一种“社交图片”）

### 3.2 数据预分析与现象总结

在开始主要任务之前，我试图通过对这些图像运用简单的大数据分析来挖掘出一些不易察觉的有趣属性与规律。在我这样做的过程中，有一件事情值得注意，那就是会有一定数量的图像没有在它们正确的后缀扩展名中，这会在我们接下来的操作中引起一些因为解码错误而造成的意想不到的错误。为了解决这个问题，我们可以通过二进制打开文件来区分它们的真实格式，读取每个文件的特定序列号，图像文件头中的四组数字可以显示它们的实际编码方法。

在整个数据集中出现的各种格式类型，其序列号及比例如下所示，序列号采用十进制和十六进制形式(同时我们只需关注前四种常见格式，它包含了我们接下来会遇到

的所有可能性)

表 3-1 实验中出现的文件类型比例与类型特征编码

文件类型与占比	十六进制文件头编码	十进制文件头编码
标准 JPEG/JFIF 文件类型(57.92%)	0xFF 0xD8 0xFF 0xE0	[255, 216, 255, 224]
PNG 文件类型(21.75%)	0x89 0x50 0x4E 0x47	[137, 80, 78, 71]
标准 JPEG/EXIF 文件类型(12.06%)	0xFF 0xD8 0xFF 0xE1	[255, 216, 255, 225]
GIF 文件类型(8.21%)	0x47 0x49 0x46 0x38	[71, 73, 70, 56]
三星非标准 JPEG 文件类型(0.20%)	0xFF 0xD8 0xFF 0xDB	[255, 216, 255, 219]
HTML3 文件类型(0.03%)	0x3C 0x21 0x44 0x4F	[60, 33, 68, 79]
其他非标准 JPEG 文件类型(0.03%)	N/A	N/A
其他 16 种无法判别的异常文件类型	N/A	N/A

通过迭代和循环结构，我们可以轻松访问所有的目录与文件，并逐层收集统计信息。随后通过基于“matplotlib”的 python 数据可视化库“seaborn”，绘制了一个具有 686433 个离散点的散点图，图形中的每个点对应一个本地图像，X 轴代表图像的宽度，Y 轴代表图像的高度。而原点(0, 0) 则是我用来初始化点集的一个虚构出来点。

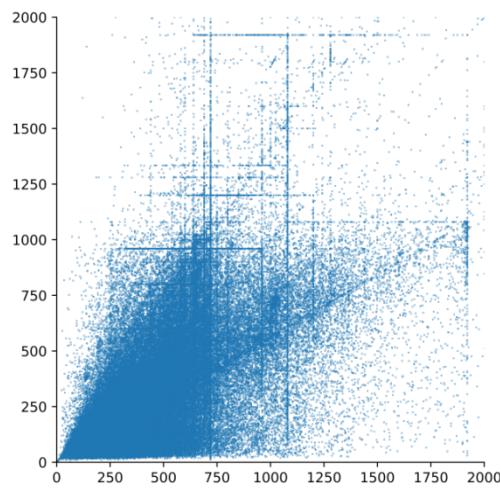


图 3-1 完整数据集的分辨率散点分布图

此外，我们还得到了宽度，高度与宽高比的平均值，分别为 499.864, 461.850 和 1.748。通过生成的散点图，我们能直观的看出两个比较值得注意的现象：

1. 图上有数条较为清晰的水平，垂直以及倾斜的直线。

2. 百分之九十以上的点都集中于一个特定的，接近三角形的区域。

通过对这些直线上的点所对应的图像进行挑选，观察后发现数据中所有宽度或高度或宽度与高度成比例的图像大多是屏幕截图，即使不完整，这个截图的尺寸也包含了其电子设备屏幕的长或宽之一。虽然这些截图来自于众多不同尺寸的电子设备，然而大多数产品都是按照特定的行业规则生产制作的，并且是遵循一定的流程开发的。比如说 4:3 是历史最久的比例，它在电视发明之初就已经存在，到现今仍在使用，并且用于许多显示器上。再例如，经过一系列的发展历史，当前主流的屏幕更喜欢 16:9, 16:10 或更高的宽高比。

因此，从遵循一定标准制造出的设备所记录下来的屏幕截图也将遵循类似的规则。我们甚至可以通过分析各种屏幕截图的分辨率规律，预测出市面主流的分辨率规律。然而这并不在我们的研究范围之中。

并且限制于开发商和服务提供商为节约成本和提高效率而制定的媒体传输规则，大多数图像在上传前都会被按照某种规则压缩到相对合适的大小，因此，已经在互联网上多次传播的社交功能图像必然会聚集一个特定的分辨率区域。如图中所示的区域中，图像密度最大的区域是(0, 0) 和(720, 1080) 之间的三角形区域。

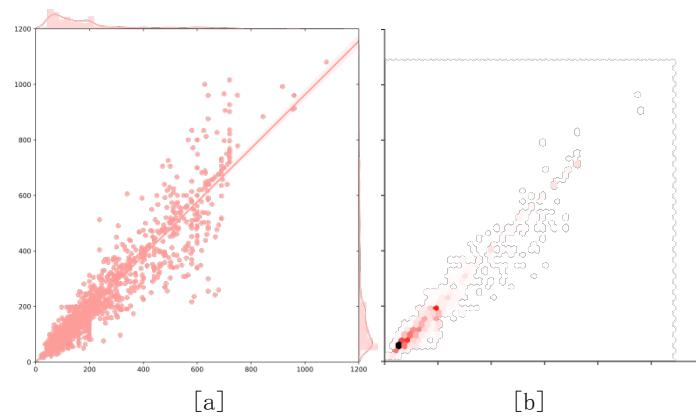


图 3-2 部分“社交图片”样本的分辨率散点图 [a] 与密度蜂窝图 [b]

在对整体数据进行了与分析之后，接下来我们通过人工的主观辨别，从完整的数据集中挑选出了 3037 幅真正的“社交图片”，他们的宽度，高度和宽高比的平均值分别为 193.877, 185.261 和 1.074。并且它们的分辨率大体上遵循线性回归，大多数图片的形状更喜欢以正方形呈现，但高度实际上倾向于略短于宽度。

类似于我们眼睛所呈现的超宽场景，作为记录或模拟现实世界的方式，这类图像更喜欢较高的宽高比，这可以使我们拥有更强的“沉浸感”，如 2:1 或更高的视野，而开阔的视野能让我们感到平静和放松。受此影响，我们生活中的图像会倾向于更宽一些。然而，对于社交图片来说，为什么它们的宽度和高度之间的差异不像我们以前在普通图像上发现的那么明显？(1.073 对比于 1.784)

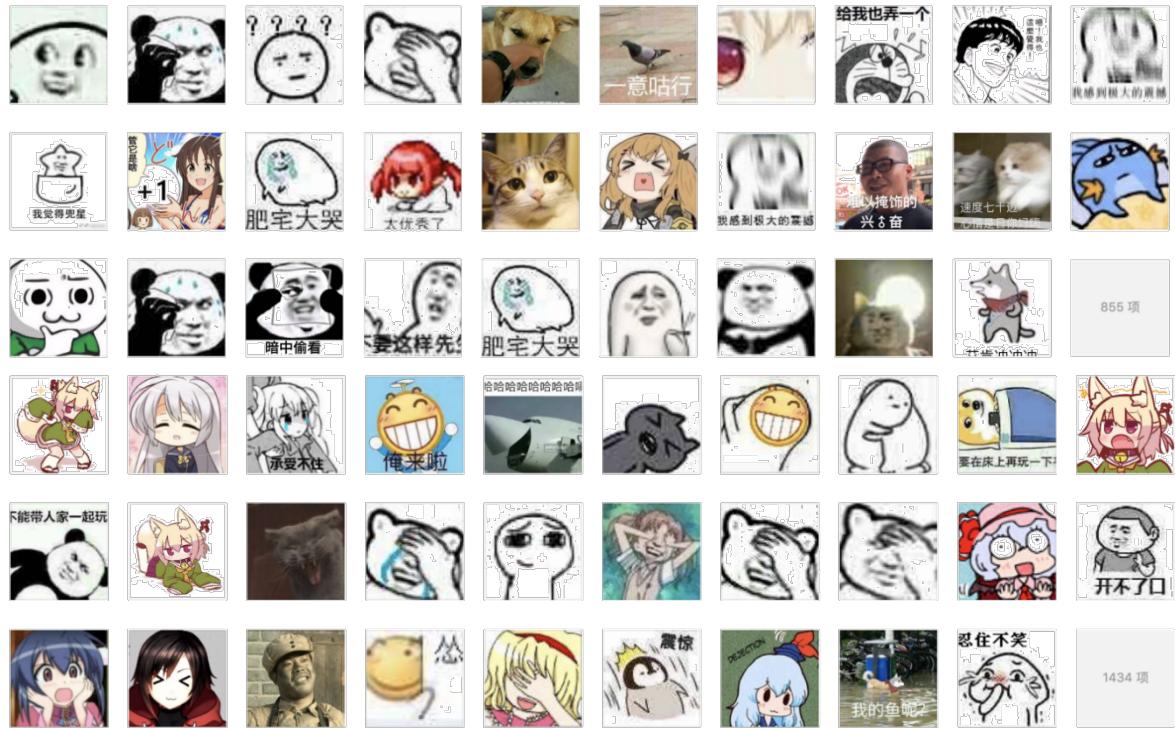


图 3-3 部分完美社交图片样本展示

通过观察和匹配所有的社交图片，我们可以发现几乎每一个例子的主要内容都是围绕一个主角的，尤其是主角的头部，或者说面部特征，而并非整张的图像的所有内容都是必不可少的，有时还配上额外的身体动作或姿势，这也种规则适合大多数肖像图片。例如，下面的例子非常典型地说明了这个想法：



图 3-4 电影“金馆长对金馆长对金馆长”剧照

它拥有一个标准的宽银幕尺寸但唯一需要注意的地方是这个演员的面部表情，由此类推，我们能够很容易找出这些图片的真正内核聚焦在哪里。就仿佛将其用作社交软件的展示头像一样，一张肖像图片的重点就是围绕着主角面部的一个尺寸合适的圆形区域。

但是为什么我们收到的所有正常图像都是矩形而不是圆形？虽然这不是我们的主要任务，但我仍然想为这提供一个理由。

在我们的生活中，制造圆形的东西比制造具有相同功能但呈矩形的东西需要更多的材料成本和更困难的制造技术。而且，我们的祖先在第一次找到方法去分割物体的时候，将兽皮或其他材料沿着简单的线进行划分是更为合理和容易的，因为将事物分为圆需要相对复杂的数学知识，否则很难找到一种行之有效的方法来有效的缩减边角余料等废物的产生，性价比实属不高且浪费较大。因此在大多数情况下，圆形的物品只会是名门贵族之人的特权，或是被用于某些特定的地方，比如车轮。

相反，矩形的东西则会更加便宜，从而被人们很好地使用和接受，且在当时大多数事物都是由普通劳动人民发明与创造的。因此随着人类文明的长期发展，我们的世界实际上是一个矩形的世界，我们的绘画和电影在纸上和屏幕上以矩形呈现，我们居住的建筑物和车辆都是立方体，在我们四四方方的移动设备中，所有的图像也都是矩形的也是比较合理的了。

综上所述，从图片的圆形内核中提取信息的最佳方法是制作一个外接正方形。有时内核也可以是椭圆，那么所生成的图形将是长方形。

### 3.3 情感模型建立方法

关于情绪是否能够被精准量化并记录表示，以及对于是否存在原子情感，即是否存在组成其他复杂情感的基础情感，一直是学术界争论不下的议题，这涉及到了情感计算(Affective computing)与情感受识别(Emotion Recognition)，且不同于文本情感分析，前两者通过者辨识人类的不同情绪，而后者仅辨识词意与语义的情感极性，多与自然语言处理相结合。

在认知科学和神经科学两个门类中，用以描述人类感知，并对人类情感进行分类的模型主要有连续(多维)模型和分类(离散)模型两种。

大多数情况下我们可以从一种场景中感受出远不止一种复杂情绪，甚至可以同时存在两种定义和感觉上完全相反的情绪也是合理的。比如当知道了我们所爱的人出于讨好自己做了一件十分愚蠢，令自己十分恼火的事情时，我们会产生一种哭笑不得的状态，这会既包含开心同时也包含愠怒。

且非常容易混淆的一点是，我们使用“社交图片”的目的是为了更加高效便捷向接收者传达我们当前的处境与感受，也就是说将发送者本人代入图片中角色所展示的情绪与所处的状态。

虽说其最主要的部分是围绕着人物主体，但是实际上发挥作用的是整张图片，包括图片的字幕，图片色调，周边景物。其所传递给我们的情绪，并不等同于图片中的角色在拍摄图片的时刻所抱有的情绪，也不等同于图片中的角色面部表情所展现出来的情绪。完全相同内容的图片，调成不同的色调，或配上不同的字库，也能够展现出截然不同的意义。因此在这里我们要脱开这些复杂因素的影响，在对数据进行主观上的情感判断时只围绕着图片的主要部分，并且将色调等会对判定造成影响的部分消除。接下来需要关注的只剩下“社交图片”中的主角的面部特征与身体姿势。

因此我们现在的任务变成了分析使用者在使用这张图片时，想通过图片中角色的

面部特征与身体姿态传播什么样的情绪。在此我们将采用一种只针对于常见“社交图片”进行适配与扩展的混合模型以进行标签，其融合了 Albert Mehrabian 教授于 1996 年提出的三维模型“Pleasure-Arousal-Dominance (PAD) Emotional State Model”，与 Robert Plutchik 教授于 1980 年提出的模型“Plutchik's wheel of emotions”。

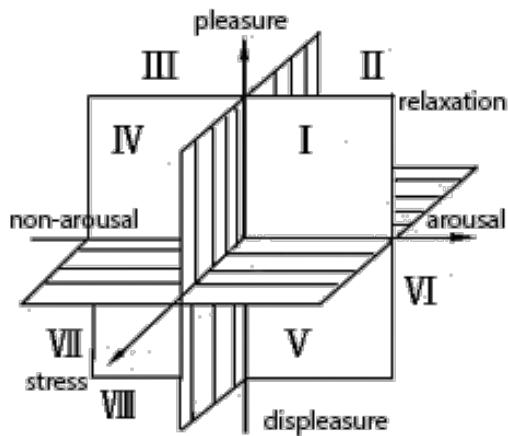


图 3-5 情感模型

X 轴，判断情绪属于 Arousal(唤醒)亦或 non-arousal(未唤醒)：造成某一种或几种情感出现的事件是否会不可避免的造成生理上的一系列变化，即引导情绪的事件是否存在刺激性。比如，当考试作弊被抓到的时，正常人会首先瞬间头脑空白，浑身冒汗甚至双腿发软。从而引导惊慌失措与绝望等情绪，这就是处于唤醒状态；而当我们躺在沙发上听着音乐看着书时，我们则身心愉悦倍感放松，这就是处于未唤醒状态。

Y 轴，判断情绪是否造成压力，即引导情绪的事件是否存在压力源。比如驱动欲，对事情后续发展的担忧(或期待)，从而感到心理上的压力( 并非所有的心理压力都会引导生理压力，即被唤醒，只有心理压力达到一定阈值才会引导出生理压力)

Z 轴，判断该情绪是否会引导愉悦的积极体验，或不愉悦的消极体验。

这三者并不相互独立，但是这三者也不完全相关。即并非有压力就会被唤醒，也并非无压力就属于未唤醒。需要将自己作为事件的经受者，逐个代入理解。

注：此模型经过两位女士与三位男士分别校对勘误，对于常见的“社交图片”的所属类别已经相对准确。

## 第 4 章 算法设计

### 4.1 数据预处理

#### 4.1.1 前言

奇异样本数据的存在会引起训练时间增大，同时也会导致模型无法很好的收敛。而我们的样本来自于互联网上各种平台，包含许多无法预料的不稳定因素，因而在对神经网络进行训练之前对我们做出预分析并贴上标签后的样本进行归一化是十分有必要的。通过前面的分析，我们在此主要使用具有清晰面部结构的“社交图片”进行实验。而在此我们使用 python 的开源库“face\_recognition”，该项目的人脸识别是基于业内领先的开源库 dlib 中的深度学习模型，用美国麻省大学安姆斯特分校制作的“Labeled Faces in the Wild”人脸数据集进行测试有高达 99.38% 的准确率，但对小孩和亚洲人脸的识别准确率尚待提升。为解决此问题，在此我们通过传入参数“--tolerance”来实现调节容错率，将其从默认的 0.6 调节至 0.4，以提升其在识别卡通面部形象与非人类面部形象时的识别率。

#### 4.1.2 相机位置估计

人类的头部可以视作左右对称的结构，所以正常情况下，即使图片中的头部的存在一定程度上的左右偏转，我们也可以通过一定的操作在不存在信息损耗的情况下将其摆正，即通过信息较多的一边补齐信息较少的另一边。

将 X 轴的方向定义为头部直视正前方时的方向。当头向上或向下在 XOZ 平面俯仰时，即使在理想情况下，其面部结构的损耗在一定程度上是不可判断和恢复的，即假设我们拥有当前 2D 图片中目标的精准 3D 模型，从而参照对应点进行坐标转换，我们仍需要对各个区域逐一应用仿射变换乃至进行插值操作。

我们复杂面部肌肉群可以允许我们做出很多并不对称的表情，如不满，蔑视等，并且这些表情的信息正是通过这种不对称来体现的。因此实际情况中，过度的左右转头，即在 XOY 水平转动，也是会影响我们的学习结果的。但单纯的在 YOZ 平面上的

头部旋转则对面部信息丢失的影响不大。

为此我们需要判断一张图片中头部的姿态以明确是否有将其归一化的价值，即判断图片是否其存在过多的头部 YOZ 与 XOY 上的偏转。

在计算机视觉中，物体的姿势指的是其相对于相机的相对取向和位置，可以通过相对于相机移动对象或相对于对象移动相机来更改姿势。对于相继来说 3D 刚体的移动只有两种方式：平移与旋转。而我们可以通过使用 Opencv 中提供的 solvePnP 函数来判断平面图片的姿势，即求出图像的平移向量 translation\_vector 与旋转前后的欧拉角 rotation\_angles，而这个函数的运行总共需要 5 个必须参数：

```
(success, rotation_angles, translation_vector) = cv2.solvePnP(model_points, image_points, camera_matrix, distortion_coefficients, flags=cv2.SOLVEPNP_ITERATIVE)
```

1. image\_points: 此参数为平面图像上的向 3D 模型上投影时的对应点，我们将使用开源库 face\_recognition，并从所识别出的 landmark 点集中选取的我们将使用鼻尖(31 号位点)，下巴(9 号位点)，左眼左角(37 号位点)，右眼右角(46 号位点)，嘴的左角(49 号位点)和嘴的右角(55 号位点)

2. model\_points: 此参数为透视变换时需要的二维特征点所对应的三维点的位置，理想情况下是我们拥有照片中角色的头部 3D 建模。但实际上这并不现实。我们也并不需要通过获取照片中的人的 3D 模型来获取特征点的 3D 位置。只借助一个通用的头部 3D 模型就足够了。在后续归一化操作中，我们将使用以下三维点来进行姿态估计：  
鼻尖: (0.0, 0.0, 0.0); 下巴: (0.0, -330.0, -65.0); 左眼左角: (-225.0, 170.0, -135.0); 右眼右角: (225.0, 170.0, -135.0); 嘴的左角: (-150.0, -150.0, -125.0); 嘴的右角: (150.0, -150.0, -125.0)

3. camera\_matrix 与 distortion\_coefficients: 这两个参数分别为相机内参与相机畸变系数，我们可以通过图像的中心来近似求出光学中心，并以像素的宽度近似求出焦距，并假设不存在畸变。

4. flags: 此参数为 solvePnP 函数所使用的方法，我们选用的是基于 Levenberg-Marquardt 的迭代优化方法。

随后我们将 solvePNP 求到的 translation\_vector 与 rotation\_angles 代入 OpenCV 提供的 projectPoints 方法中，输入需要投影的 3D 点(点集), rotation\_vector, translation\_vector, 相机内参 camera\_matrix, 相机畸变系数 distortion\_coefficients, 从而将输出重投影 2D 点 nose\_end\_point2D。

随后计算原 2D 点 p1 和重投影 2D 点 p2 的向量 p12 与 x 轴的成角就是头部下低或上抬的俯仰角。当头部上抬或下低的程度小于正负 15 度(即 tangent 值小于 $\pm(2-\sqrt{3})$ )时则认为存在归一化价值。

而 p1 于 p2 的距离 p2p1\_length 则可以代表其沿着 x 轴左右偏转的程度，将 p1 与 p2 在 3D 模型中所对应的点 np.array([(0.0, 0.0, float(length))]) 与(0.0 , 0.0 , 0.0)的距离 length 与 2D 投影后 p1 与 p2 的距离 p2p1\_length 做对比即可求出水平偏转系数 horizontal\_coefficient = p2p1\_length/length, 当其小于 0.2 的时候我们则认为当前样本存在归一化价值。

#### 4.1.3 面部平均化

对于存在归一化价值的样本，为了尽可能的消除头部左右偏转对图片带来的影响，我选择通过图片已有的部分头部结构来补全信息损耗较大的另一部分，即用两张互为水平翻转副本的图片生成一个平均面部。

首先我们要将面部对齐，即通过 68 个 face\_recognition 识别出的关键点的，分别照出在图片中最大与最小的 X 值与 Y 值，四个常函数构成一个矩形区域，随后以矩形中心为正方形中心，并以矩形最长边乘以(1+10 %)为正方形边长构建正方形，并判断是否超越了图像大小，若超越了图像大小则对超出部分以 0 值填充，随后使两张图片的正方形区域完全重叠对齐并去掉交集以外的区域。

接下来我们求出 68 对关键点间连线中点的坐标，组成新一组的 68 个点，并标以蓝色，我们再将原本的限制区域的顶点与各边中点标记为额外的 8 个关键点，并标记为如图所示的绿色。

随后的工作是对这 76 个点进行应用 Delaunay 三角剖分，得到如图 4-7 所示的三角网格，随后我们遍历三角网格中的每个三角形，按照其三个顶点的序号分别去两张原图中寻找所对应的原三角形区域，并向当前网格中的三角形区域应用仿射变换从而得到两张可以近乎重叠但是存在细微差别的图片，而这些差别就是我们需要应用另一张水平翻转的图片进行补充的内容。

最后我们可以直接将仿射到同一个三角网格上的两个图片完全对齐，求出每一个像素对于两张图片的平均值并应用于最终的结果图片上，所得到的结果就是进行左右翻转补齐后，消除了 YOZ 和 XOY 平面上头部偏转的面部图片。

#### 4.1.4 图像预处理

首先我们按照当前三角剖分所得网格外围正方形的四个顶点对应用“面部平均化”后的图片进行剪裁，对于越界的区域进行 0 值填充。

随后，因之前针对 3037 张标准“社交图片”所求出的图片的高度与宽度的平均值分别为 193.877, 185.261，且平均宽高比为 1.074，因此我们决定将标签样本的尺寸统一至边长为  $3 \times 2^6$  的正方形，即将全部图片调整尺寸至 192\*192。(至于为何采用  $a \times 2^b$  这种规则来限定图像分辨率不做论述)

而将图片由正方形调整成正方形是一个十分令人愉悦的过程，在此我们可以简单地运用 OpenCV 所提供的 resize()方法。

随后我们将当前图片转换为矩阵数组并对其进行归一化线性变换，即求出矩阵内的全部元素的最大值与最小值，并将矩阵内全部的元素减去一个最小值后除以最大值与最小值的差，即矩阵内全部元素的极值，从而将样本的特征值转换到同一量纲下并把数据映射到 [0,1] 区间段内，且线性变换的性质决定了改变后的数据不会“失效”，反而能够使计算梯度下降进行求解时更快的收敛。

## 4.2 深度学习

### 4.2.1 构建标签

因国内暂无开源的样本标签库适用于本论文，且此设计并非商业用途，所以我们将自费雇佣人员，或找同学与朋友帮忙，按照需求针对手中的“社交图片”进行人工分类。

为了尽可能的减少因主观认知差异带来的误差，我们对标签规则进行了细致的制定与解释，并对所有参与标签工作的人员在展开工作前进行了培训与测试，并由开发者对标签样本进行了检查，以确保结果正确。

表 4-1 所采用的标签规则

位名称	位功能	位内容
一号位 (异议较小)	判断主观上是否为社交图片	1 位, 布尔型, [0 , 1] 0: 不属于社交图片 1: 属于社交图片
二号位 (无异议)	判断社交图片的基础风格类别	1 位, 整型, [0 , 3] 0: 无法判断 1: 真实世界风格内容 2: ACG(动漫动画游戏)类 3: 改编创作类
三号位 (无异议)	判断社交图片的基础内容类别	1 位, 整数值, [0 , 3] 0: 组合情况 1: 人类/卡通人类 2: 动物/卡通动物 3: 非生物/卡通非生物

---

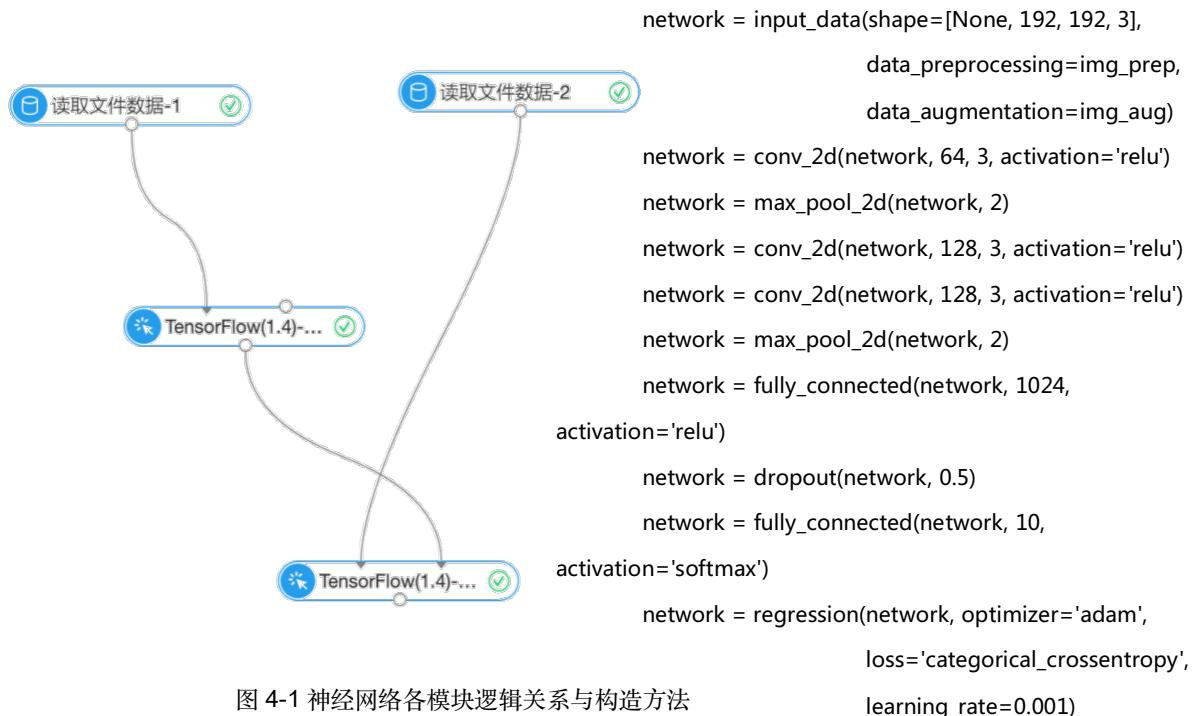
四号位 (异议较大)	针对较为流行与常见的社交图片, 按照其内容属于的 IP 进行分类	3 位, 字符串, [0 , 999] 每一个序号代表一个不同的类型
(若该社交图片存 在面部结构)	对图片所表达的情绪进行粗略与细 致分类	5 位, 字符串, 由三部分组成, 分别为:
五号位(异议较大)		象限[1 , 8] + 主类[01 , 99] + 子类[01 , 99] (除象限外的其他两部分 为非必须内容, 若无法判 断则用 0 占位)
文件序号位	对不同文件进行编号以从根本进行 区分	7 位, 字符串 由三部分组成, 分别为: 月份[01 , 12] + 日期[01 , 31] + 文件编号[000 , 999]

---

标签采用直接针对样本文件重命名的方式，在操作时按照路径读取样本时也顺带拆解样本标签并进行学习，如果这个文件不是“社交图片”，直接就以“一号位+文件序号位”命名，不继续考虑其他属性(此便捷命名规则只受用于非社交图片情况)

随后我们将每 1000 个样本进行预处理后压缩成一个无后缀文件并命名为“data\_batch\_”+上数字编号。而我们共有 5 个标签样本集，“data\_batch\_1”, “data\_batch\_2”, “data\_batch\_3”, “data\_batch\_4”, “data\_batch\_5”用于训练，有 1 个标签样本集 “test\_batch” 用于测试并得到正确率。

### 4.2.2 构建卷积神经网络



我们的深度学习工作除小型化网络的本地测试外，都使用了阿里云 PAI 机器学习平台与阿里云 OSS 对象储存服务进行搭建，平台采用了适用于混合型工作负载的 Tesla-P100 集群，NVIDIA 的 Pascal 架构使其能为高性能计算和超大规模工作负载提供卓越性能。并且其为传统深度学习提供了从数据处理，模型训练，服务部署到预测的一站式服务，支持 CPU/GPU 混合调度，具有高效的资源复用；后者则提供了海量，安全，低成本，高可靠的云存储服务，并可以使用使用 RESTful API 可以在互联网任何位置存储和访问。

出于时间原因我们暂无法将样本的全部标签都使用上，首先我们将针对识别 10 种表达最常用的情感的“社交图片”进行训练与测试。这些类型在样本总体中的占比一般大于  $1E-2$ 。

用于预测与训练的网络结构相同，共包含了 5 个加权的层，3 个为卷积层，2 个为全连接层。其中一个全连接层用以输出一个 2 维的“softmax”来表达对于 10 个类别的预测。这 10 个类别是我们在进行人工分类时出现频率最高的。我们的网络采用取最大值

的多目标逻辑回归，最大值池化层设置在一号与三号卷积层后，非线性激活函数“ReLu”被应用在每个卷积层和全连接层。因采用“sigmoid”时的程序计算量要远大于使用“ReLu”函数时的，且反向传播求误差梯度时还需求导并涉及除法。而采用 Relu 激活函数时只需要判断输入是否大于 0，因此使用“ReLu”函数时的模型收敛速度会提升很多。且“ReLu”会使一部分神经元的输出为 0，这样就造成了网络的稀疏性，并且减少了参数的相互依存关系，缓解了过拟合问题的发生。且对于深层网络，在“sigmoid”接近饱和区时的数值变换太过缓慢，这种情况会造成信息丢失，从而在误差反向传播时，更容易出现梯度消失的情况，且在网络层数多的时候尤其明显。与之相反，“ReLu”的 gradient 大多数情况下是常数，十分有助于解决深层网络的收敛问题。进一步而言，“ReLu”函数相比“sigmoid”更符合生物神经元的特征。

过拟合是机器学习容易出现的通病，若出现过了拟合，所计算出的模型大概率会是无效的。因而为了解决过拟合问题，一般会采用“ensemble”方法，即训练多个模型做组合。而此时，时间成本的开销就成为一个大问题，不仅训练工作费时，测试工作也十分费时。而“dropout”的出现则可以很好的解决这个问题，其更像一种添加噪声的方法，“dropout”会迫使一个神经单元与随机挑选出来的其他神经单元共同工作，减弱乃至消除了神经元节点间的联合适应性，增强了泛化能力。

随后我们需要利用损失函数来检验算法模型的优劣，同时利用损失函数来提升算法模型。这个提升的过程就叫做优化(Optimizer)。而我们选用的是“adam(adaptive moment estimation)”方法，即适应性矩估计方法是对 RMSProp 优化器的更新，利用梯度的一阶矩估计和二阶矩估计动态调整每个参数的学习率。每一次迭代学习率都有一个明确的范围，使得参数变化相对平稳。并且配合“交叉熵损失函数(categorical\_crossentropy)”以用来评估当前训练得到的概率分布与真实分布的差异情况。同时选用 0.001 作为该神经网络的学习率。

## 第 5 章 结果分析

### 5.1 实验结果

我们并没有过多的纠结于神经网络结构与参数对于测试结果的影响并尝试对其进行校调，因这只是一个较为粗浅的试验。

实验共使用了 5000 张已标签的“社交图片”进行训练，其中每种类别分别有  
 ①162(3.24 %), ②131(2.62 %), ③99(1.98 %), ④98(1.96 %), ⑤60(1.2 %), ⑥134(2.68 %),  
 ⑦79(1.58 %), ⑧125(2.5 %), ⑨120(2.4 %), ⑩136(2.72 %)个样本。

而测试使用的标签样本集中包含了①17(1.7 %), ②16(1.6 %), ③21(2.1 %),  
 ④22(2.2 %), ⑤18(1.8 %), ⑥28(2.8 %), ⑦19(1.9 %), ⑧20(2.0 %), ⑨35(3.5 %),  
 ⑩20(2.0 %)个样本。

各个类型的预结果测准确率分别为①15(88.24 %), ②13(81.25 %), ③14(66.67 %),  
 ④15(68.18 %), ⑤14(77.78 %), ⑥23(82.14 %), ⑦14(73.68 %), ⑧18(90.00 %),  
 ⑨29(82.86 %), ⑩17(85.00 %), 总计 172/216, 即 79.63 %, 表格如下。

表 5-1 实验结果中各类型在训练集与测试集中所占的数量与比例

样本类型	训练集中各类型样	测试集中各类型样	预测结果正确/测
	本数/样本总量	本数/样本总量	试集中各类型样
			本总数
①Disdainful (对...)	162/5000(3.24 %)	17/1000(1.7 %)	15/17(88.24 %)
轻蔑的;鄙夷的;满			
脸不屑的			
②Mischievous 调	131/5000(2.62 %)	16/1000(1.6 %)	13/16(81.25 %)
皮的;淘气的;滑稽			
的;幽默的;诙谐的;			
表情行为夸张的;			

③Bored (对...)感到厌倦(烦)的;不耐烦的;拒绝的;	134/5000(2.68 %)	21/1000(2.1 %)	14/21(66.67 %)
④Puzzled (对...)困惑(迷惑)的;	98/5000(1.96 %)	22/1000(2.2 %)	15/22(68.18 %)
⑤Overbearing 霸道的;专横的;盛气凌人的;	60/5000(1.2 %)	18/1000(1.8 %)	14/18(77.78 %)
⑥Gruesome (面部表情)阴森的;可怕的;残忍的;(丧心病狂的)	134/5000(2.68 %)	28/1000(2.8 %)	23/28(82.14 %)
⑦Embarrassed 窘迫的;尴尬的;(尬笑的)	79/5000(1.58 %)	19/1000(1.9 %)	14/19(73.68 %)
⑧Shocked 震惊的;惊愕的;(因...)而身心受到打击的;	125/5000(2.5 %)	20/1000(2.0 %)	18/20(90.00 %)
⑨Sad 悲伤的;伤心的;哭泣的;	120/5000(2.4 %)	35/1000(3.5 %)	29/35(82.86 %)
⑩Angry 生气的;发怒的;冲动的;暴躁的;	136/5000(2.72 %)	20/1000(2.0 %)	17/20(85.00 %)
总计	1182/5000(23.64 %)	216/1000(21.60 %)	172/216(79.63 %)

## 5.2 结果分析

即便是我们人类，也会不可避免的出现误判，比如我们会想不起来向你走来并且朝你招手的人是谁，我们也会误以为前面的某个人是自己的熟人而上去打招呼。因此归根结底深度学习是一门仿生学，而其模仿的生物，也就是我们自身，都无法做到的事情也不应该去苛求采用不完全算法的冰冷的机器。而从当前结果可以看出，对于包含变体越多的类别，其判断时的准确率就越低。且虽暂时没有可进行对比的参照案例，但总体的预测准确率还是较为可观的。

而我们当前面对的一部分实际问题就是，对监督学习所需样本进行标签的人工成本过高，因此我们应该专注于如何尽可能的提升数据的使用效率以及采取何种方法更加符合现实世界情况的根据已有样本扩充样本数量。除此之外，回归到神经网络本身，小到节点权重的初始化赋值方式，神经网络的学习率，所应用的激活函数与损失函数，大到神经网络结构与样本的学习方式，每一部分对于最后模型的影响，以及各个部分对于最后模型的影响都值得我们细致的探究。

这个结果并非建立在可以让神经网络真正意义上的理解人类感情上，但这至少可以证明通过深度学习并之配以合理的情感模型依然可以满足很多应用场景，例如我们所尝试的通过主观情感对“社交图片”进行识别与分类，将训练好的神经网络模型并配之以其他智能算法或与模糊理论相结合，就可以在一定程度上做到快速匹配当前需求以提供最合适的“社交图片”，即最优解。

## 总结与展望

人类始终希望机器可以做得更多，更好。从我们能够使用斜面，杠杆，滑轮等较为原始的省力机械，到开创了以机器代替手工劳动的第一次工业革命，再到我们现代的工业自动化。而我们依旧不满足于此，我们希望计算机能够按照人类的方式来思考，来处理事务，或提供解决方案，并且运用计算机独特的优势，将其做得更加的完美，而2016年3月，谷歌旗下的人工智能机器人“AlphaGo”首次击败人类围棋选手。

但是其归根结底也只是采用了更为高级的算法。若要计算机可以真正的按照人类的思维思考，其是否首先要具有情感，具有欲望，具备自我意识，乃至自由意志呢？这仍是个充满争议且具有浓厚科幻色彩的论题。但对于这些虚无缥缈的感觉究竟该以何种模型被模拟与计算，以及人工智能是否可以真正意义上理解，而不去是如同找规律一般的识别呢。这仍需要通过计算科学，生命科学，神经科学，脑科学，心理学，伦理学以及应用物理，量子物理等众多领域的几代，乃至十几代科学家们的不懈钻研与探索来实现。

而在此之前，我们是否可以找到一种方法使神经网络能够像我们一样通过概念进行学习，并能够通过挖掘数据间的联系进行类比学习，亦或能否以一种更加尖端的仿生学方式来模拟思维的形成过程呢？

通过本科阶段的毕业设计工作，我对深度学习相关内容有了更加深刻的认识，知晓了人工智能这门学科的发展历程与现状。这也让我产生了很多需要探索与解决问题。为什么这种拓扑结构的网络能拥有这种特性，为什么对于学习率的改变可以使预测结果朝着这种方向改变，我还想知道我们为什么选择这样的损失函数。我发觉这应该就是学习与探索的乐趣，我也意识到了自己当前所欠缺的大量知识，更是明确了自己所感兴趣的领域与接下来需加倍努力的方向。本科并非我求学路上的终点，我希望在接下来的求学路上，亦或学术生涯中，能够砥砺前行，为自己所专注的事业做出一定的贡献。

## 致谢

机缘巧合，来到了长春理工大学，我感激这四年里遇到的每一个人，我珍惜这四年间经历过的每一件事。

首先感谢我的父亲刘守峰，母亲王辉，我最爱的人。他们给予了我世上父母所能够给儿子的全部的爱与关怀。纵使面对着我在成长道路上不断暴露出的罄竹难书的顽劣，纵使一次又一次的令他们伤心与难过，也终选择宽容，不厌其烦的教育，引导与守护着我。在面临人生选择，在人生道路上蹒跚前行之时，正视，重视我的每一个任性的，不成熟的决定，让我有勇气与能力走得更高更远。

感谢赵建平教授在学术上所给予我的深刻指导。感谢王睿老师，从立钢老师对我论文初稿数次耐心细致的审阅与校对，以及感谢计算机科学技术学院的每一位授课的老师，他们一丝不苟的治学态度与各具特色的授课方式都将成为我记忆中无法抹去的感动，他们的每一次拿起的粉笔都刻下了我生命中华丽的注解。

感谢张昕老师，在项目组内工作学习与提升自己的那段时光回忆起来格外的幸福与充实。感谢蒋振刚老师，在我留学申请时为我纠察文书中的逻辑错误与语言疏忽，并多次为我出具推荐信。与二人的数次畅谈更是让我对职业与学术生涯有了更明晰的认知与规划，在他们身上我感受到了他们对学生的尊重与关心，以及为师的一颗炽热的心。

感谢周俞彤，在 16 年夏天出现，作为我人生中结识到的第一位优秀的女生，她的自信与上进氤氲出了独特的人格魅力。数年来作为我的榜样与至交，给予我关怀与鼓励，让我明白自律给我自由，努力与奋斗让我远离空想与焦虑。让我内心变得温暖，让我的思想与行为更加成熟。

感谢韩春鑫，谢组鹏，我大学收获的最宝贵的友谊，学业上，事业上，生活上。在我最为迷茫困顿之时出现，打消了我大二下正在准备实施的休学支教的计划，将我招入项目团队一员，让我明白了一个有趣的点子是如何变成现实并创造价值的，也让我

体会到了实现自我价值的满足与骄傲，并且有机会与途径接触到了全国各地许多优秀的人，明白了自己想成为一个什么样的人，过上什么样的生活。

感谢冯啸迪学长，在项目组内耐心细致的将其所学知识毫无保留的传授予我，答疑解惑，这也让我掌握并熟悉了互联网方向的相关内容与技术，对大型软件工程业务需求有较全面的了解，为日后的许多工作打下了坚实的基础。

感谢高国栋，王双娇，周思哲，姚鉴城，作为我大学期间关系最为密切的伙伴，是他们陪我熬过了我四年里三千小时的课程，一千小时的自习，数千顿早饭午饭晚饭。他们为这平淡无奇的日常生活添上了艳丽的色彩。他们四年里为我提供了无数的帮助，甚至包括提供这篇文章中所使用的原数据以及完成项目中繁琐冗杂的标签工作，都有他们的身影。他们掌握着我无数的黑历史，也见证了我的努力，成长，与改变。

感谢刘冬，佟皓东，阮文佳，张文龙，曾泽宇，我一起生活了的室友们，能够为了他人收起自己的锐气和小性子，相互包容，理解。每一次猜拳决定谁下楼取饭，每一次考试前夜的挑灯夜读，所有鸡毛蒜皮，嬉嬉闹闹的事情都是幸福。

感谢导员宋子祥，班长高酉权，学习委员黄新龙，以及带班杨鑫学长多年来的辛勤付出。感谢与我一起奋斗过的同学们，我们见证了彼此的青春，也共同经历了生活的辛酸苦辣。以及感谢我在网宿科技实习期间导师苏江水，组长庄贤荣，人事部 HR 刘霜，以及朱向清，饶丽萍，李凯，陈成等众多同事的照顾。

最后感谢刘一帆同学，在最后一学期以奇妙的方式降临，与我产生了复杂深刻的情谊。也感谢挚友王博，恩师原卓，以及李宗蔚，肖文超，徐士尧，陶雨欢，以及许许多多曾在我生命中驻留过，给予了我极大帮助与深刻影响的人。并向在我完成毕业设计过程中，创作了我所阅读到的各领域精彩文献的工程师与科学家们致敬。并且感谢所有开发阿里云 PAI 机器学习平台的前辈们。

年复一年，春绿冬藏，花开花落，人来人往。太多面孔与瞬间浮现在脑海，每一个人都带来回忆，每一件事都值得思考。毕业在即，又要踏上另一段旅程，希望人生的新阶段在充满挑战的同时，也常有温暖相伴。最后，愿师长安康，同窗如意。

## 参考文献

- [1] Yu Z . Image based Static Facial Expression Recognition with Multiple Deep Network Learning[C]// Acm on International Conference on Multimodal Interaction. ACM, 2015.
- [2] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.
- [3] Cowen A S , Keltner D . Self-report captures 27 distinct categories of emotion bridged by continuous gradients[J]. Proceedings of the National Academy of Sciences, 2017:201702247.
- [4] David Eberly, Perspective Mappings[J]. Geometric Tools, Redmond WA 98052 , 2011.
- [5] Plutchik R . A psychoevolutionary theory of emotion.[J]. Emotion Theory Research & Experience, 2000, 21(4-5):529-553.
- [6] Pei S C , Lin C N . Image normalization for pattern recognition[J]. Image and Vision Computing, 1995, 13(10):711-723.
- [7] Zhang S J , Cao X B , Zhang F , et al. Monocular vision-based iterative pose estimation algorithm from corresponding feature points[J]. Science in China Series F (Information Science), 2010, 53(8):1682-1696.
- [8] Kazemi V , Sullivan J . One Millisecond Face Alignment with an Ensemble of Regression Trees[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014.
- [9] Liu N , Wang K , Jin X , et al. Visual affective classification by combining visual and text features[J]. PLoS ONE, 2017, 12(8):e0183018..
- [10] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]// International Conference on Neural Information Processing Systems. 2014.
- [11] Ahmad K . Affective Computing and Sentiment Analysis[J]. IEEE Intelligent Systems, 2016, 31(2):102-107.
- [12] 刘剑聪. 社交图像美学分类与优化算法研究[D]. 2014.
- [13] 赵思成. 图像情感感知的计算与应用研究[D]. 2016.
- [14] 黄杰雄. 社交图像的情感和美学评价研究[D]. 2018.
- [15] 刘海龙, 李宝安, 吕学强, et al. 基于深度卷积神经网络的图像检索算法研究[J]. 计算机应用研究, 2017(12):302-305.