密级:_____

中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于深度学习的多属性图像分类方法研究

作者姓名：王华阳

指导教师：蒋树强　研究员

培养单位：中国科学院计算技术研究所

学位类别：工学硕士

学科专业：计算机应用技术

研 究 所：中国科学院计算技术研究所

2017 年 5 月

Research on Techniques of Multi-attribute Image Classification Based on Deep Learning

By

Huayang Wang

A Thesis Submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Master of Engineering

Institute of Computing Technology

May, 2017

声　　明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：　　日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：　　导师签名：　　日期：

摘　要

人类所获取的外界信息中有80%是来自视觉的，而且通过视觉获取到的信息是最丰富也是最复杂的。我们人能够很好的看清楚并理解视觉所捕获到的信息，但是如何让计算机看懂并理解图像信息却是一件非常困难的工作。图像分类是让计算机理解世界的基础，也是多媒体技术研究的一个重要方向。而图像分类中的多属性图像分类则可认为是多媒体技术中一个基本而富有挑战性的研究领域。多属性图像分类工作有助于机器从多个层面来更详细的理解图像，从而为计算机理解世界奠定更坚实的基础。

对于多属性图像分类任务，本文研究了如何利用图像多属性标签之间语义关系的嵌入和卷积神经网络模型不同网络层特征的融合来提高模型的分类准确率。文中提出了两种用于图像多属性分类的卷积神经网络模型，分别为：1）局部非对称的多任务卷积神经网络模型（PAMT-CNN），2）融合多层特征的互影响卷积神经网络模型（ME-DAG-CNN）。文中将两种模型分别应用于多属性图像的分类工作中，并在两个数据集上验证所提模型的有效性。

1.与传统的多属性图像分类工作不同的是，局部非对称的多任务卷积神经网络模型在多属性图像特征的提取过程中考虑了图像多属性语义之间的相互影响，以及图像多属性语义相互嵌入对学习图像特征表示的影响。卷积网络低层更多关注的是图像边缘、颜色等共通的特征表示，而高层则关注的是具有类别倾向的区分性特征。在此基础上局部非对称的多任务卷积神经网络模型通过低层网络参数共享实现图像多属性标签语义的相互嵌入，从而提取出共通的低层特征表示。之后通过在相同传统卷积神经网络模型上表现出更好分类性能的图像属性语义在模型中间层对其他图像属性分类任务进行指导，从而提高模型在图像各属性分类任务上的分类正确率。

2.融合多层特征的互影响卷积神经网络模型是在局部非对称的多任务卷积神经网络模型的基础上改进而来的。局部非对称的多任务卷积神经网络模型在设计时考虑了图像多属性标签语义在特征提取过程中的指导作用，但并未考虑融合多层网络特征对模型分类性能的影响。因此，我们在设计融合多层特征的互影响卷积神经网络模型时同时考虑融合多层网络特征和多属性标签语义相互嵌入对模型分类性能的影响。文章通过实验验证了融合多层特征的互影响卷积神经网络模型在提高多属性图像分类任务正确率上的有效性。

关键词：图像分类，深度学习，多属性，语义嵌入

Research on Techniques of
Multi-attribute Image Classification Based on Deep Learning

Wang Huayang (Computer Applied Technology)
Directed By Jiang Shuqiang

80% of the information obtained by human beings comes from human vision, and the information obtained through vision is the most abundant and complex. We are able to see and understand the visual information captured by our vision, but it is a very difficult task to make the computer understand the image information. Image classification is the basis for computer to understand the world, but also an important research direction of multimedia technology. And multi-attribute image classification in image classification can be regarded as a challenging research field in multimedia technology. Multi-attribute image classification can helps the machine to better understand the image, so as to lay a more solid foundation for the computer to understand the world.

For the classification task of multi-attribute images, we study how to use the semantic relation between images and multi-attribute and the fusion of different tasks and multi-scale features to improve the accuracy of image classification. In this paper, we propose two classification models, namely 1) local asymmetric multi-task convolution neural network model (PAMT-CNN), 2) multi-attribute convolution neural network model with mutual influence (ME-DAG -CNN). Both models are used for our multi-attribute image classification tasks.

1. Different from the traditional multi-attribute classification work, our local asymmetric multi-task convolution neural network model considers the relationship between multi-attribute semantics in the process of extracting multi-attribute image features. Through embedding the multi-attribute semantics to achieve a common low-level feature representation. On the basis of this, the semantic label with stronger classification relation is used to guide the task of weak classification, so as to improve the correct rate of the model in the label classification task of the image.

2. The multi-attribute convolution neural network model with mutual influence is improved on the basis of local asymmetric multitask convolution neural network model. The local asymmetric multitasking convolution neural network model is designed to take into account the multi-attribute semantics, but does not take into account the different scales of features. Therefore, we introduce the fusion of multi-scale image features when designing multi-scale convolution neural network models with mutual influence. The fusion network model improves the correct rate of multi-attribute image classification to a certain degree.

Keywords: Image Classification, Deep Learning, Multi-attribute, Semantic embedding

目　录

图目录

表目录

第一章引言

1.1 研究背景和意义

多媒体技术是当前发展最快、最活跃的研究技术，也是计算机科学的重要研究领域之一。人类所获取的外界信息中有 80%是来自视觉的，而且通过视觉获取到的信息是最丰富也是最复杂的。我们人能够很好的看清楚并理解视觉所捕获到的信息，但是如何让计算机看懂并理解图像信息却是一件非常困难的工作。图像分类是让计算机理解世界的基础，也是多媒体技术研究的一个重要方向。而图像分类中的多属性图像分类则可认为是多媒体技术中一个基本而富有挑战性的研究领域。多属性图像分类工作有助于机器从多个层面来更详细、更具体的理解图像，从而为计算机理解世界奠定更坚实的基础。

百度百科对属性的理解如下：属性就是对于一个对象的抽象刻画。因为任何一个具体的事物总是有很多与之相关的性质和对应的关系，我们可以把这些与之对应的性质和关系都叫做该事物的属性。当人来区分一个物体与另一个物体或者说一个事物和另一个事物是否相同时，比较的是一个物体的属性和另一个物体的属性是否相同。由于物体属性的相同或者不同才使得我们所在的客观世界中形成了许许多多种不同的事物类。具有我们限定的相同的物体属性的事物就形成了同一个类，而不具有我们所限定的相同的物体属性的事物就划分到了不同的类。属性是对于一个对象的抽象刻画也是对对象的性质和对象之间的相互关系的统称。属性有本质属性和非本质属性之分，有浅层次属性和深层次属性等不同层次概念的属性之分。所谓本质属性是指决定一个物体之所以成为该物体而不同于其他物体的属性，包括某物体固有的规定性和与其他物体的区别性是我们常说的本质属性的两个最为典型的特点。因此，本质属性一定是物体所特有的属性，而所谓的物体特有属性就不一定是物体的本质属性。但是，我们知道有些物体的特有属性就是从该物体的本质属性上派生而来的。如人能够双脚直立行走的非本质属性就是由人的特有属性派生而来的。而不同层次概念的属性又分为浅层次属性和深层次属性。我们人对物体最初的认识多属于浅层次的属性，浅层次的属性最先反应出的是物体的非本质的特有的属性。只有对物体进行更深层次的研究即反应物体本质属性的研究才称之为深层次概念的属性。

人对事物的理解是从不同层面的多个属性来进行的，这也说明了为什么人能够准确的区分各种各样的物体。为了对物体有更全面的理解，我们需要在传统分类的基础上引出多属性的图像分类问题。多属性的图像分类问题的实现能够使计算机对物体有更全面、更细致的理解，为计算机视觉的其他相关工作或者说多媒体技术的其他相关工作 奠定更加坚实的基础。而传统的多属性的图像分类工作更多的关注的是图像中物体的局部属性

特征，例如：对汽车图像的多属性识别包括汽车车轮数目，车门数目等局部属性，对人脸图像的多属性识别包括头发长短，是否佩戴眼镜等局部属性。对于物体局部属性分类的多属性分类工作不需要考虑各个属性之间的相互关系。从大部物体局部属性来看，他们之间可以认为是相互独立的。而本文要实现的多属性的图像分类工作针对的是图像中物体的全局抽象。我们可以将图像中物体的全局抽象属性认为是介于物体本质属性和非本质属性之间的半本质属性，也可以认为是介于浅层概念属性和深层概念属性之间的中间层概念属性。之所以这么认为，是因为物体的任何一种全局属性虽不能完全限定该物体，但可以极大限度的缩减该物体所在的范围。因此在物体属性的识别过程中对不同属性之间关系的有效利用能够更大程度的提升分类性能。

在 2006 年，Hinton 提出通过神经网络提取图像特征表示的方法，并将该文章发表在 Science 上。该文章提出了两个主要观点：1）包含多个隐含层的深度神经网络具有更好的特征学习能力，通过包含多个隐含层的深度神经网络学习到的特征与原始数据本身表现出的模式更加的接近，并且该特征可以用来更好的实现分类与可视化；2）深度神经网络的训练一直是一个难点，然而通过逐层的无监督训练方法能够有效的克服这一问题。通过研究证明了我们需要去设计更深层的网络结构从而使网络能够学习更高层的抽象特征表示。而更深的网络结构则意味着更多的网络参数和更多的计算复杂度。而近年来随着计算机硬件设备的跳跃式发展，尤其是 GPU 的发展以及大规模的原始的、带标注的训练数据（ImageNet [1]）的出现推动了深度学习更快发展并应用于计算机视觉领域中的图像分类任务上。在 2012 年，Hinton 等[2]采用卷积神经网络 AlexNet 在 ImageNet [1]比赛中取得第一名的成绩。该卷积神经网络在设计上通过 dropout 方法来抑制了模型的过拟合，并在分类任务上 Top5 分类达到 84.7%的分类正确率，比第二名的 Fisher Vector 方法[3]的正确率高出近 10%。

自 2006 年以来，采用深度神经网络的方法来实现图像的分类任务获得了前所未有的关注。深度卷积神经网络在图像分类等多媒体领域引起的革新思潮席卷了整个学术界，带来了让人豁然开朗的新思路，同时也展现出它最振奋人心的魅力。传统的深度神经网络模型在实现图像的单分类问题或是物体的局部多属性分类问题上均可表现出很好的性能。因为无论是图像的但分类问题还是图像中物体的局部多属性分类问题的实现均不需要考虑不同属性之间的相互关系，而面向图像中物体全局属性的分类问题则需要考虑属性间的相互关系从而实现分类性能的提升。因此，在基于全局的多属性图像分类问题上我们要充分考虑图像属性之间的相互关系，将图像属性之间的相互关系体现到深度的卷积神经网络模型中去，从而提高模型的分类性能。

目前的计算机视觉正处在一个被人工智能光环所笼罩下的最好的时代，不论是自动驾驶还是智慧城市等都赋予了计算机视觉技术新的使命与挑战。图像的多属性分类能够实现对图像更全面、更细致的理解，为计算机视觉的其它工作奠定更加坚实的基础。而如何利用图像多属性之间的语义关系以及深度神经网络各网络层之间的交互实现分类性能的提升是我们需要解决的问题。从当前的发展趋势来看，卷积神经网络在计算机视觉

或者说图像分类任务上的应用前景无疑充满了各种可能。

1.2 本文工作

图像分类是让计算机看懂世界的第一步。而图像的多属性分类又是图像分类任务中的一个富有挑战性的工作。我们知道，人对图像的理解除有视觉信息方面以外，更多的是在语义层面来理解图像，这也正是人与机器最大的不同。多属性的图像分类能够让机器在更广的语义层面来理解图像，为计算机视觉的其他相关任务打下更坚实的基础。而近年来，深度学习模型中的卷积神经网络模型在图像分类方面取得了很好的性能。因而一个自然而然的想法便是将卷积神经网络模型应用到图像的多属性分类任务中。而对于多属性图像的分类问题来说，如何对设计卷积神经网络的网络结构并对卷积神经网络结构进行相应改进而使其拥有更高的抽象特征学习能力是利用卷积神经网络来进行图像的多属性分类任务的一个重点。另外，我们认为多属性的图像分类任务的不同属性标签之间一定存在着某种语义关系，而这种标签间的语义关系能够很好的“指导”并实现多任务的卷积神经网络高层特征的抽象性。

本文研究的内容为多属性图像的分类问题。如何设计深度卷积神经网络的网络结构，使得图像在特征提取的过程中能够嵌入图像的多属性语义信息，同时利用这种语义信息来指导图像的多属性分类工作是我们研究的重点。全文工作共分为两部分：提出局部非对称的多任务卷积神经网络模型框架，以及融合多层特征的互影响神经网络模型框架。

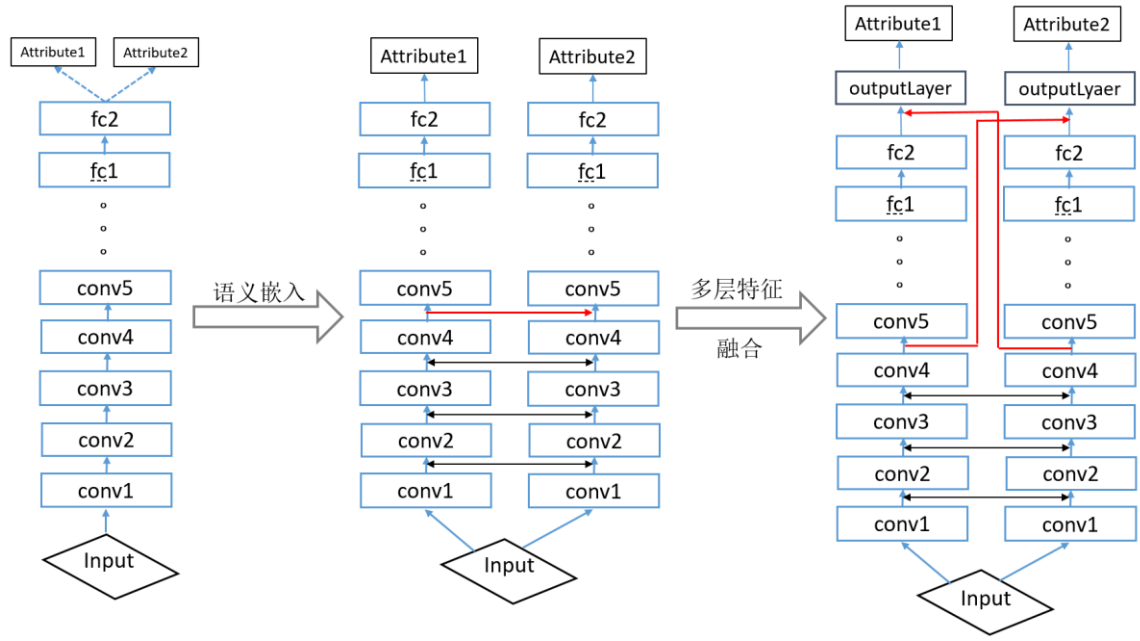


图 1.1本文各部分工作之间的关系。

本文各部分工作之间的关系如图 1.1 所示，下面分别对各部分内容及其创新之处进行简单介绍。

1） 局部非对称的多任务卷积神经网络模型

图像分类是最早采用卷积神经网络模型的几个研究领域之一。卷积神经网络模型在图像分类任务上较传统机器学习方法取得了跨越式的进步。卷积神经网络模型在提取图像的特征表示时，神经网络的卷积层低层的特征表示更倾向于边缘、颜色、纹理等共通的特征表示，而神经网络的卷积层高层的特征表示则更倾向于具有类别倾向的重要区分性的抽象特征描述。在局部非对称的多任务卷积神经网络（Partially Asymmetric Multi-Task Convolutional Neural Network）模型的设计上参考了传统的卷积神经网络结构以及具有图像标签语义嵌入实现的非对称卷积神经网络模型。

该设计的优点在于，在局部非对称的多任务卷积神经网络模型的低层卷积层提取用于分类的图像共通的特征表示的过程中融入了图像多属性的语义信息。这种图像属性语义信息的相互嵌入是的模型底层卷积层提取的图像特征具有了更强的共通性和鲁棒性，也为提取具有类别倾向的抽象性特征打下了坚实的基础。而多语义信息的融入能够更好的帮助实现图像的多属性分类工作。同样，在高层卷积层的非对称性语义嵌入避免了具有类别倾向的重要区分性特征的混淆。文中在两个多属性图像分类数据集上验证了这一改进的有效性。

与已有的相关工作不同的是，模型利用了图像不同的属性标签语义之间的相互关系来“指导”网络底层特征描述的学习，并在“指导”的过程中考虑了神经网络各层特征表示的特点。

2） 融合多层特征的互影响卷积神经网络模型

不同网络层特征的融合能够更好的提高神经网络模型分类的正确率。因为图像的分类工作尤其是图像的多属性分类工作中有些类的区分需要对形变鲁棒的高不变性特征来实现（例如区分人和狗），而有些类的区分则更多的需要更强的纹理和轮廓信息来实现（例如区外形相近的两款 SUV 汽车的生产厂家）。因此在设计用于图像多属性分类的神经网络模型的时候需要通过融合网络不同层的不同尺度特征来提高我们分类的正确性。局部非对称的多任务卷积神经网络模型只在模型的底层通过参数共享来实现不同属性标签语义之间的相互指导而并没有考虑多层特征融合对多属性分类问题的影响。

卷积神经网络底层的特征表示更多的关注的是图像的纹理和颜色等共通的特性，而网络高层的特征表示则更多的关注的是图像语义层面的抽象特性。文中通过实验证明了通过融合网络高层特征能够提高模型分类的正确率。因此在局部非对称的多任务卷积神经网络模型的基础上提出融合多层特征的互影响卷积神经网络（Mutual effect DAG Convolutional Neural Network）模型。该网络模型不仅考虑了不同图像属性标签之间的语义关系对分类的影响，同时考虑了融合网络不同网络层特征对于多属性图像分类性能带来的影响。

与已有相关工作不同的是，在融合多层特征的互影响卷积神经网络模型的设计上同时考虑了图像多属性标签之间的语义关系和网络模型多层特征融合对分类性能的影响。文中将两者融合到多属性图像分类的网络模型中，提高了模型的分类准确性。

1.3 论文的组织结构

论文的结构组织如下：第二章介绍了与图像分类相关的国内外研究现状和发展趋势，第三章介绍了基于局部非对称的多任务卷积神经网络模型对多属性分类任务的研究。第四章介绍了融合多层特征的互影响卷积神经网络模型，以及如何基于当前任务选择合适的融合多层特征的神经网络模型。第五章为对本论文的总结。

第二章 国内外研究现状和发展趋势

2.1 传统的图像分类技术

图像分类技术一直被认为是图像理解中的一个核心的研究方向。图像分类是指对于给定的图像的类别进行判断。

在图像分类任务上，传统的图像分类方法包含：基于词袋模型特征（Bag of words）[4]的方法, 基于形变模型（Deformable part Models）[5]的方法，基于纹理特征、形状特征和颜色特征等方法。

上述方法中使用最广泛的是将几种方法组合来实现图像的分类。例如：对图像采样后提取 SIFT[6]，HOG[7]等局部特征，之后基于空间关系进行编码（局部的约束性编码 LLC[8]，Fisher 编码 FV[9]，超向量编码[10]等）和对应的池化（基于空间金字塔的池化[11]，随机池化[12]，最大值池化等）操作，之后基于词袋模型[4]生成最终的标识性特征。

上述方法虽然在其对应的特定分类任务方面表现出了较好的性能，但在分类的准确性上仍然有很大的提升空间。并且，因为传统的图像分类在提取视觉表示的时候采用的是手工设计的特征，这种方法不但需要了解图像数据集的内在模式，费时费力并需要有专业的知识的相关经验。上述缺点导致传统的图像分类技术效率很低，难以实现更广的应用。

2.2 ImageNet 竞赛中的图像分类技术

近年来，在 ImageNet 大规模视觉识别竞赛[13](ILSVRC)中，卷积神经网络模型已经取得了比传统模型更好的性能。并且识别的准确性也在逐步提高。

在 2012 年，Hinton 等[2]采用卷积神经网络 AlexNet 在 ImageNet [1]比赛中取得第一名的成绩。该卷积神经网络在设计上通过 dropout 方法来抑制了模型的过拟合，并在分类任务上 Top5 分类达到 84.7%的分类正确率，比第二名的 Fisher Vector 方法[3]的正确率高出近 10%。[2]中通过端到端（end-to-end）的训练方法形成一个卷积神经网络模型，并采用该模型来进行图像视觉特征提取与分类。AlexNet 网络结构如下如所示：

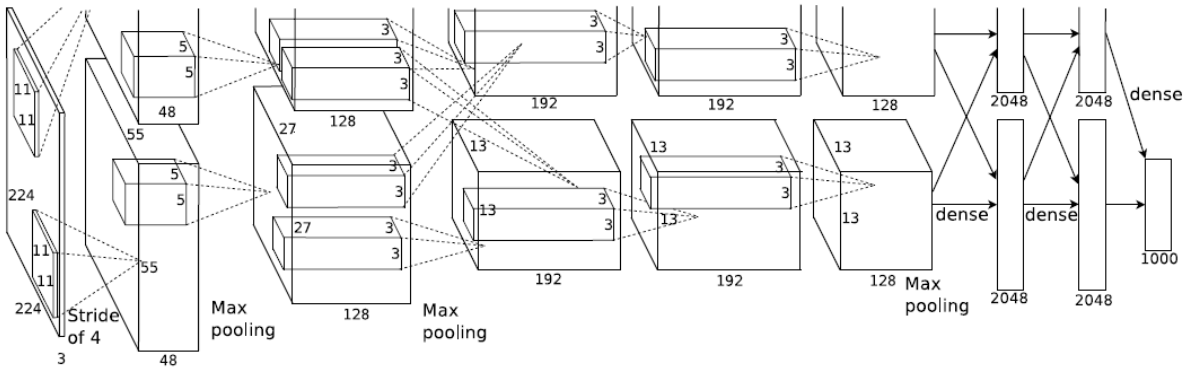
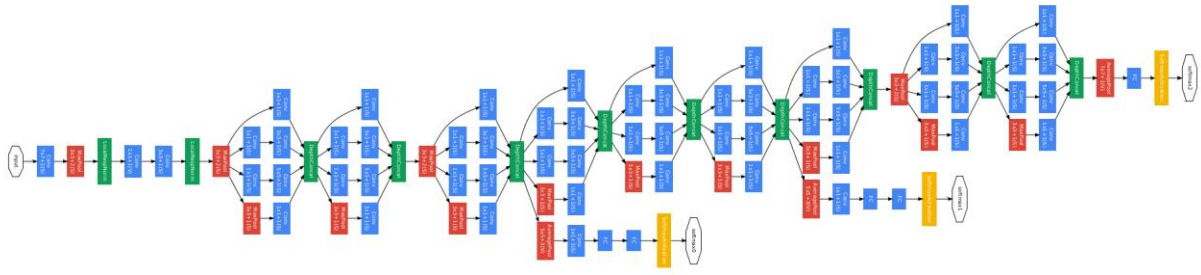


图 2.1 AlexNet 卷积神经网络框架图[2]。

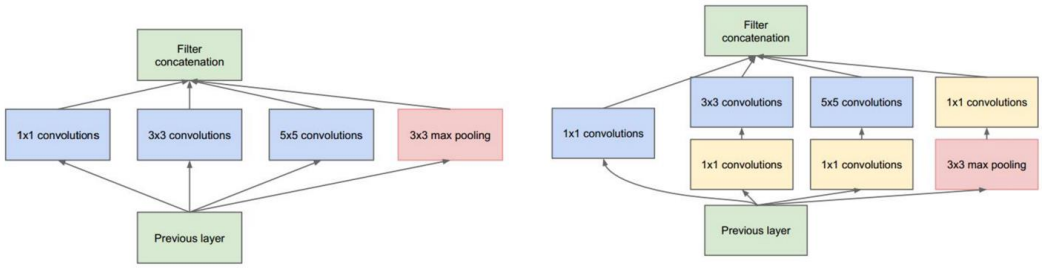
该模型分为上下两层是因为作者采用了两块 GPU 显卡，我们以其中一层模型为例对 AlexNet 模型进行描述。该卷积神经网络模型共包含 8 层，其中有 5 个卷积层和 3 个全连接层。在每一个卷积层中均采用非线性变换 RuLU 来代替传统神经网络中采用大的激活函数 tanh/sigmod，此种替换加速了网络训练完成的速度。同样，在每一个卷积层中采用了局部相应归一化（LRN）对响应进行归一处理。在卷积层之后连接的是降采样的过程。在训练的时候，AlexNet 卷积神经网络采用了 Dropout 技术来防止过拟合的发生。

随着时间的发展，卷积神经网络得到快速的普及和应用。人们更多的关注点开始转移到更新的想法和模型的改进上而不是提升硬件设备和如何获得更大规模的数据集上。一般来说改进网络结构有两种方法：1）增加卷积网络深度，2）增加卷积网络宽度。这两种方法也意味着引入了巨大的训练参数，从而更容易导致过拟合的问题。而解决这个问题的根本方法是将模拟生物神经系统连接，将卷积转化为稀疏连接。

基于上述思考，谷歌在 2014 年提出 GoogLeNet[14]的早期版本——Inception V1，并在 ImageNet2014 的视觉识别竞赛中获得第一名的成绩。而 Inception 卷积神经网络中的结构单元采用的主要思想就是如何用密集的卷积操作来近似最优的局部稀疏结构。Inception V1 整体结构图与基本结构图如下：



（a）



（b） (c)

图 2. 2 GoogLeNet 卷积网络框架图。（a）GoogLeNet 网络结构框图（b）GoogLeNet 网络结构单元的早期版本示意图。（c）GoogLeNet 网络结构单元的改进后版本示意图。

GoogLeNet 基本结构单元的早期设计原因如下：1）采用不同尺寸的卷积核是为了获取不同尺度上的特征信息，并完成信息的融合。2）池化技术能够有效减少参数并防止过拟合，因此在结构单元中引入池化操作。3）卷积核不同尺寸的选择是为了能够在卷积之后得到统一空间中的相同维度特征。但是原始基本结构单元的设计中使用了 5x5 的卷积核，这与涉及初衷减少模型参数，实现局部稀疏是相违背的。因此在接下来的改进中引入 1x1 的卷积核。该卷积核的引入会将模型参数的数量降低 4 倍，同时又不影响最终的特征映射维度。

文章[15]中对简单的加深网络层数会使得训练误差加大这一问题进行了细致的分析，并提出如下假设：如果我们每一层都能够实现将前一层的结果直接映射过来的话，那么虽然加深了网络的层次，我们仍能够保持与前层相同的训练误差而不会增加训练误差。在 2015 年，微软亚洲研究院（MSRA）在此基础上提出深度残差网络（Deep Residual Network）[16]模型，并在 ImageNet2015 视觉竞赛中以 3.57%的错误率取得第一名的成绩。而该成绩也完成了对人类视觉能力的突破。152 层的神经网络模型也比以往任何成功使用的卷积神经网络模型层数高出 5 倍以上。

每一年在 ImageNet 竞赛中取得最佳名词的卷积神经网络结构都将引起用于图像分类的网络结构的变革。大量的工作将围绕当前用于 ImageNet 竞赛的最优网络结构进行改进从而在对应的实验数据集上取得更好的分类性能。

2.3 基于改进网络模型的图像分类技术

卷积神经网络提出之初是以人眼的生理结构作为基础的，这也是卷积神经网络很早的应用是计算机视觉领域的一个重要原因。卷积神经网络的结构的生物学基础奠定了其主要的架构模式——卷积和池化的叠加，并有最后的全连接层提供预测和表示输出。而随着深度学习的不断发展，人们也不断的“反思”这一架构模式，希望从中分析出不同组成部分的作用并改进、替换它们从而能够得到更好的模型结构，从而应用于图像分类任务。

一些工作考虑在原有网络的全连接层上做工作。文章[17]中对 GoogLeNet 网络结构进行更改，将每一个连接层（Concat layers）的输出连接到均值池化层（AveragePool Layers），作为均值池化层的输入；同时将原网络结构中的 Softmax 层替换成包含十棵决策树的深度森林（每棵树的深度固定为 15 层）从而形成一个深度神经决策森林结构(如图 2.3)。更改后的深度神经决策森林结构在 ImageNet 上的分类正确性较 GoogLeNet 网络结构模型提高约 3.7 个百分点。文章[18]则是在传统单分类的卷积神经网络后通过全连接生成 N（N 为分类的任务数）个向量表示，每个向量表示对应一个分类任务。

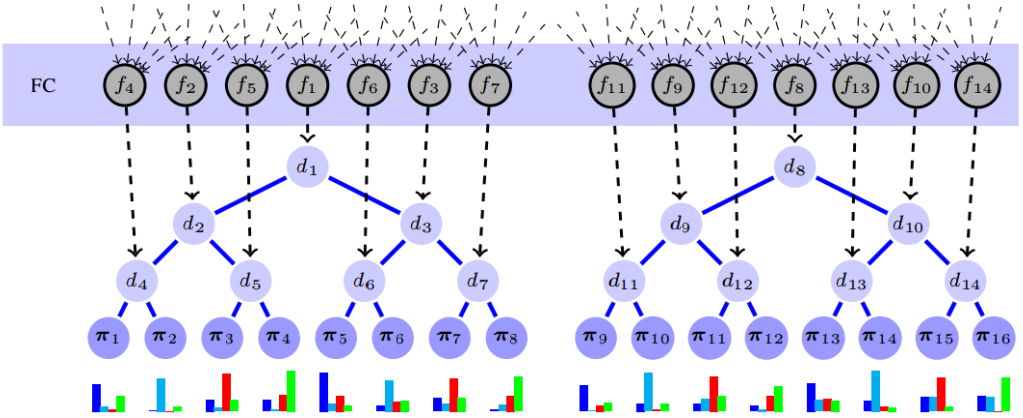


图 2.3深度决策神经森林实现细节图[17]。

在对神经网络结构的深入了解后，一些工作开始考虑替换厚着更改卷积神经网络中的相应单元实现，从而提高卷积神经网络的分类性能。文章[21]中探讨了卷积神经网络中不同构造单元的重要性和其作用，将卷积神经网络中的池化层替换成步幅卷积层（stride convolutional layer）。更改后并通过数学推理与实验来证明了替换后的卷积神经网络在性能上并不会带来任何损失。卷积神经网络提取的表示特征一直被认为是具有空间不变性的，然而这种空间不变性却有很大的局限性。因为，所认为的空间不变性主要是由最大值池化过程带来的。而最大值池化过程却实在一个很小的范围内（2x2 或者 3x3 像素区域）进行的，我们需要非常深的网络层次才能得到所谓的空间不变性。为解决这一问题，文章[22]中提出一种基于自身约束的空间转移模型（self-contained transformation module）。该模型可以添加到网络中的任何需要的部分，从而更加灵活的实现具有不变性特征的提取。文章[23]则提出了双卷积的神经网络结构（如图 2.4）来替换传统神经网络中的卷积过程。通过简单的将原始卷积神经网络中的卷积层替换为文中所提到的双卷积过程就可以在很大性能上提高了卷积神经网络的分类性能。性能提升的原因在于，原始的卷积神经网络中的卷积核（convolutional filters）是相互独立的，他们之间的学习过程可以说没有任何的相关性。双卷积神经网络模型采用的是卷积核组(groups of filters)的形式，保证了组与组之间是通过一定变换得来的，即各组卷积核之间有很强的相关性。这种组与组之间的相关性越强，那么神经网络的分类性能就越好。

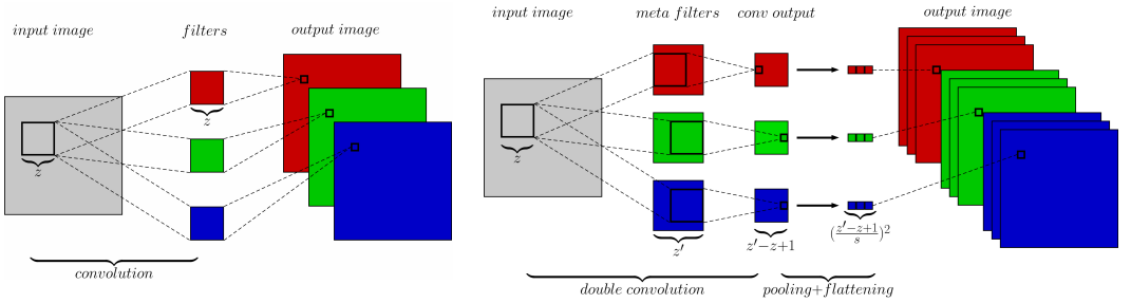


图 2.4传统深度网络模型（左）与双卷积深度网络模型（右）对比图。

最近的关于卷积神经网络结构如何改进的工作中也得出这样一个结论：网络层数的加深会提高卷积神经网络的分类性能，同时如果能够使那些与输入层和输出层更近的层之间包含短连接（shorter connections），那么网络的训练也将更加容易。基于这样的观点，文章[24]提出了 Dense 卷积神经网络模型（网路结构如图 2.5）。在 Dense 卷积神经网络模型中，层与层之间的连接采用了前向的融合。通过这种前向的融合有效的减少了梯度的消失问题，特征表示重用问题，有效的减少了网络模型的参数数目并且加强的特征的传播。

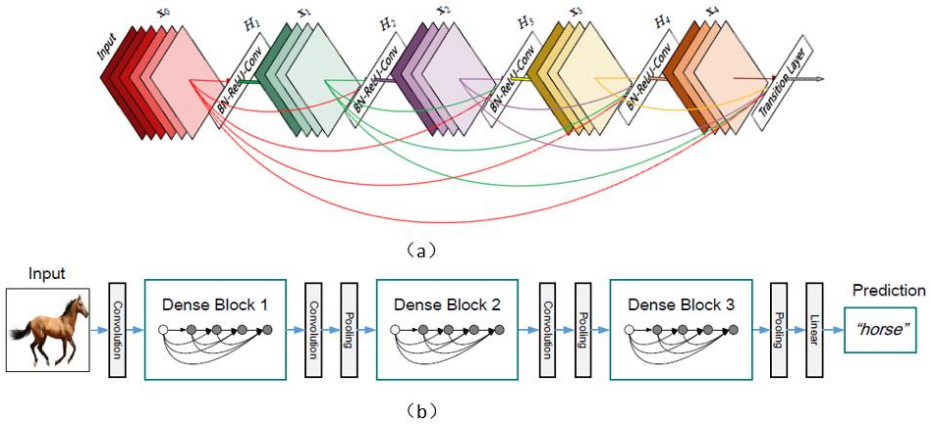


图 2.5 Dense 深度卷积神经网络模型图。（a）为一个五层的密集卷积神经网络的块（dense block），（b）为一个包含三个密集网络块的 Dense 卷积神经网络模型。从图（a）中我们可以看到，在 dense block 中的每一层均采用其前置的所有层输出作为其输入。

虽然传统的神经网络模型在图像的分类工作上取得了跨越式的进步，然而网络却没有考虑网络各层特征融合对图像分类性能的影响。多尺度的图像特征表示一直是计算机视觉领域中的一个经典问题。该问题可以追朔到图像金字塔理论（image pyramids）[25]、尺度空间理论(scale-space theory)[26]和多解决方案模型(multiresolution models)[27][28]等问题。文章[29]探究了融合卷积网络多层特征对于图像分类任务的作用从而提出用于实现多尺度图像分类的 DAG 卷积神经网络模型（如图 2.6）。传统的卷积神经网络模型中提取的特征来自于网络结构最后的输出层，而 DAG 卷积神经网络模型将网络结构中的

低层、中层和高层特征共同应用到最终的分类推理中去。这种改变使得 DAG 卷积神经网络模型相比传统的卷积神经网络模型能够更好的应用到图像的分类问题中去，因为图像的细分类问题相比粗分类问题需要更多的低层的形状、纹理等特征的帮助。所以，网络多层特征的融合能够更好的提高模型的分类性能。文章[29]中还得到这样的结论，融合前层的输出到最后的输出层总能够在一定程度上提高训练中模型的分类性能。

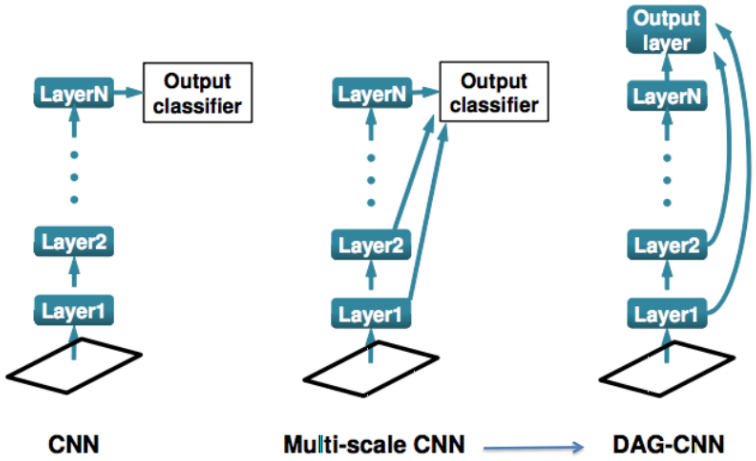


图 2. 6传统深度网络模型和融合多层特征的 DAG 网络模型对比图。传统的卷积神经网络简略图（左）、多尺度的卷积神经网络结构简略图（中）与融合多层特征的 DAG 卷积神经网络结构简略图的对比。

文章[30]中提出了一种新的卷积神经网络方法——深度融合网络（deeply-fused nets）。该创新网络方法的主要核心点在于深度融合（deep fusion）。深度融合就是将中间层的表示结果进行一定程度的融合，然后将融合后的表示作为下层网络的输入。这种方法在深度卷积神经网络模型中可以将不同层的特征表示进行合理的融合从而学习得到多尺度的特征表示，同时卷积神经网络的深层特征表示与浅层特征表示可以实现一个共同学习（jointly learnt）并且能够实现一定的相互指导。深度融合网络将深度基础网络（deep base network）和浅层基础网络（shallow base network）进行融合能够减少传统神经网络的层数，实现训练网络的更加简单化（图 2.7）。

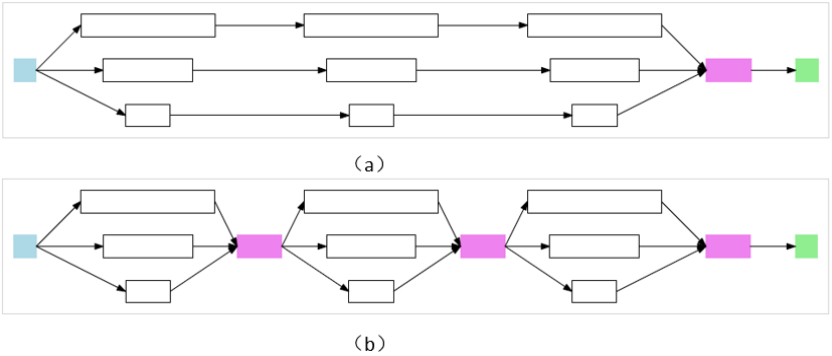


图 2. 7深度神经网络特征融合示意图。图（a）为浅层融合的卷积神经网络，图（b）为深层融合的卷积神经网络。图中的粉色矩形框表示的是融合层（fusion layer）。

梯度消失是训练深度卷积神经网络的一个常见问题。而这一问题会随着卷积神经网络的层数的增加而更加明显。针对这一问题文章[33]中提出随机深度（stochastic depth）

这样一个深度神经网络训练过程。随机深度是指在神经网络的训练过程中采用看似矛盾的训练方法，在训练的时候训练短网络，而在测试的时候使用深网络来测试我们网络的性能。该方法有效解决了梯度消失问题，提高了网络模型的性能同时还减少了网络训练的复杂度。文章[34]中提出一种新的卷积神经网络架构——分形网络（fractal network）。不同于文章[33]中采用的 drop 卷积层的方法，而是在分形网络的基础上提出一种新型的 dropout 机制——drop 路径。在图 2.8 的奇数次迭代（Local）中，文中提到的 drop 方法采用一定的几率来 drop 每一个路径的输入信息，同时保证每个 join 部分至少保留有一个输入。在图 2.8 的偶数次迭代（Global）中，只保留分形卷积神经网络的一列。

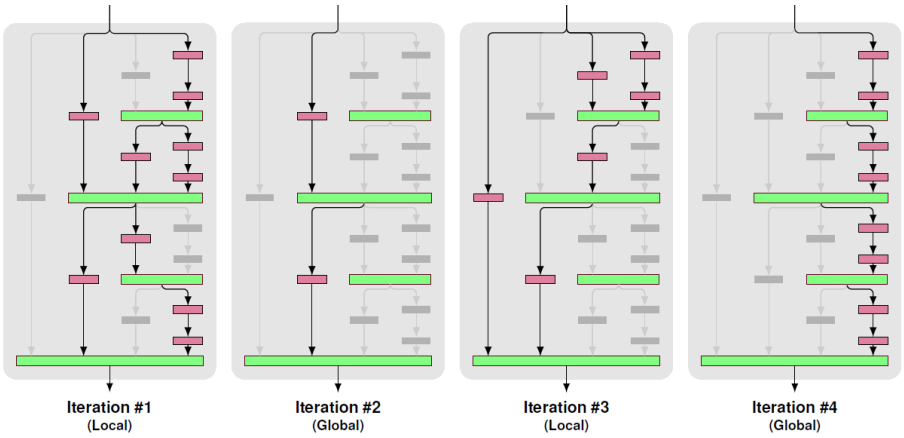


图 2.8分形网络 drop 路径图。

在任务间的相互影响方面，文章[35]提出一种包含“十字绣”单元（“cross-stitch”units）的用于多任务分类的卷积神经网络结构。该“十字绣”单元能够很好的将多任务的各个单独的神经网络进行连接，并实现端到端的训练。通过“十字绣”单元的连接，各个用于单分类任务的卷积神经网络能够实现一定的相互“指导”，实现共通表示的学习，从而提高多分类任务的分类的正确性。

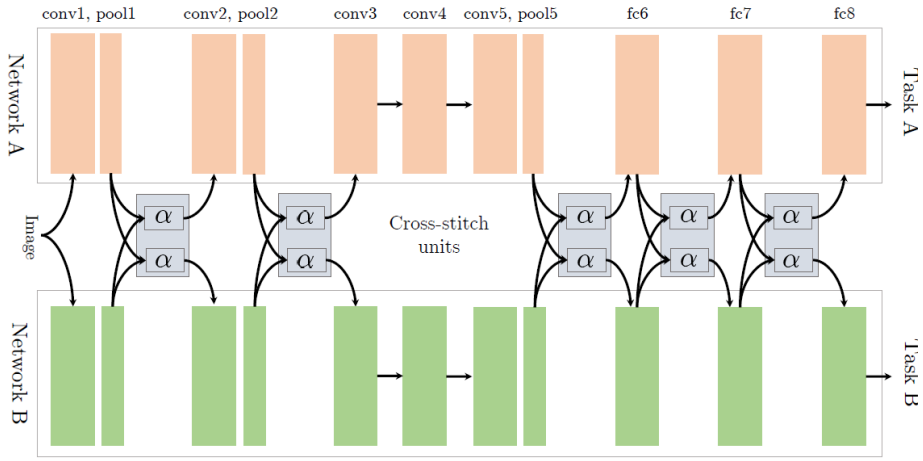


图 2.9包含“十字绣”单元的卷积神经网络框架图。途中采用了五个“十字绣”单元连接了用于实现不同分类任务的两个 AlexNet 卷积神经网络。五个“十字绣”使得两个不同的分类任务在低层和高层的特征提取过程中实现了人物间的相互指导，从而提高了模型的分类正确率。

之前的所有工作都是在传统卷积神经网络这一大的框架下进行相应的改进从而更好的适用于对应的分类任务，而没有任何彻底的颠覆传统神经网络结构的创新。而文章[36]提出了一种全新的用于替代深度卷积神经网络模型的深度 tree 模型——gcForest，采用中文表示就是多粒度级联森林结构模型。同时文章中还提出了一种新的使用多粒度级联结构来做特征表示的决策树级联方法。相比于传统的卷积神经网络模型，多粒度级联森林结构模型在调整模型参数上花费的时间要小很多即该模型更容易训练，同时在实验性能、模型拓展性和模型分析上较深度神经网络模型又具有很强的竞争性。该模型相比于传统的卷积神经网络模型的另一优点在于，传统的神经网络模型的训练需要大规模数据集的支撑，而多粒度级联森林结构模型则可以在较小的数据集上有效运转。

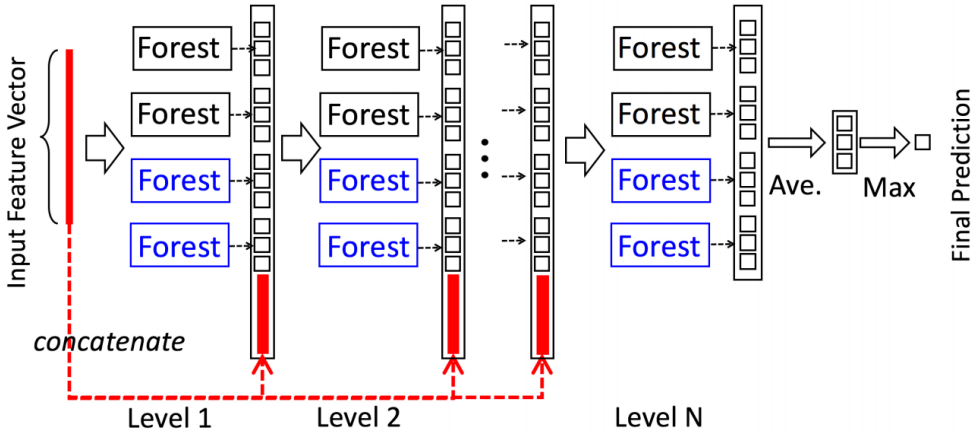


图 2. 10级联的森林结构图[36]。

与已有工作相同的地方是我们也是在传统的卷积神经网络基础上进行相应的研究工作。而与已有分类工作不同的是，我们的图像多属性分类任务的每一个属性标签均对应图像的全局区域，全局的属性标签对应的是更抽象的图像语义。这与传统的多属性分类或者多标签物体识别分类工作是不同的，因此标签间的语义相关性能够更好的进行相互的“指导”。另一点不同的地方在于我们设计图像多属性分类网络结构的时候同时考虑了网络多层特征融合和语义直接的相互“指导”对模型分类性能的影响。我们通过实验验证了二者融合后对模型分类性能提升的帮助性。

2.4 多属性图像分类技术

我们找到的最早的关于属性分类的文章是通过物体属性来描述物体[55]。文章中首次提出将识别的目标由识别物体的标签转变成识别物体对应的属性，从而从不同的角度对物体进行描述。学习属性描述带来了一个新的挑战，每个标签类中的对象可能来自不同的类中，而不像传统的分类工作，每个标签对应的对象来自同一类。这就更大化了相同类内对象的类内差异，为分类工作带来了困难。

之后，属性分类开始引起人们更多的关注。包括室外场景的属性分类，人脸属性分类，衣服属性分类等等。对象的属性分类有助于我们对图像中对象有更好的理解。

在应对户外场景的多属性分类工作时，文章[54]提出一种基于弱监督的同时辨别户外场景和对应属性的方法。文中的弱监督方法包含三个部分：1）组合场景配置，通过一些相对较少的空间层次表示来组合成更多的场景。2）属性关联，为不同词性的属性进行不同的关联，从而为每个名词加形容词的属性对训练一个表示模型。3）联合推理和学习，对于给定的图像，我们选择概率最大的场景和关联属性作为其分类结果。

在应对衣服的多属性分类任务时，一些工作考虑采用多个模型并在模型最后的特征表示进行一定的融合来实现更好的多分类任务性能。文章[19][20]中通过融合多个的独立的卷积神经网络中的特征来实现多任务分类。文章[19]的每一个分类任务采用独立的一个卷积神经网络，之后将图像对应每个任务抽象出来的视觉表示进行一定的融合形成共享矩阵，之后将形成的共享矩阵对应位置赋予不同的权重来形成最终各个分类任务的向量表示（如图 2.11）。文章[20]则是为同一个任务设立不同的卷积神经网络来处理图像的子区域，并通过不同的融合方式获得对应分类标签的特征表示用于提高多属性图像分类任务性能。

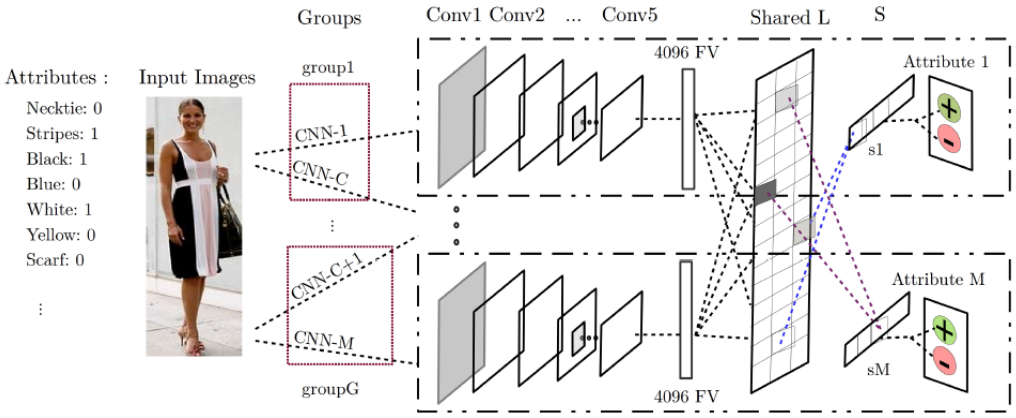


图 2. 11多任务卷积神经网络模型框架图。

图像的多属性分类工作中，有些属性信息是语义层面的而不是视觉层面的。文章[31]中提出非对称的多任务卷积神经网络模型（如图 2.12）将用户意图嵌入到图像的语义提取中去从而实现图像的分类工作。通过非对称的多任务卷积神经网络模型提取的图像的特征表示不仅包含了图像的语义信息，还包含的用户的意图信息。因此，该特征能够更好的应用于以用户为中心的各种任务中去。而文章[32]提出了一种新的用于图像多标签分类的神经网络模型——空间正则化的卷积神经网络模型（spatial regularizetion network）。该模型在仅采用图像层面的监督信息的条件下去探索图像的不同标签之间的语义和空间关系（如图 2.13）。

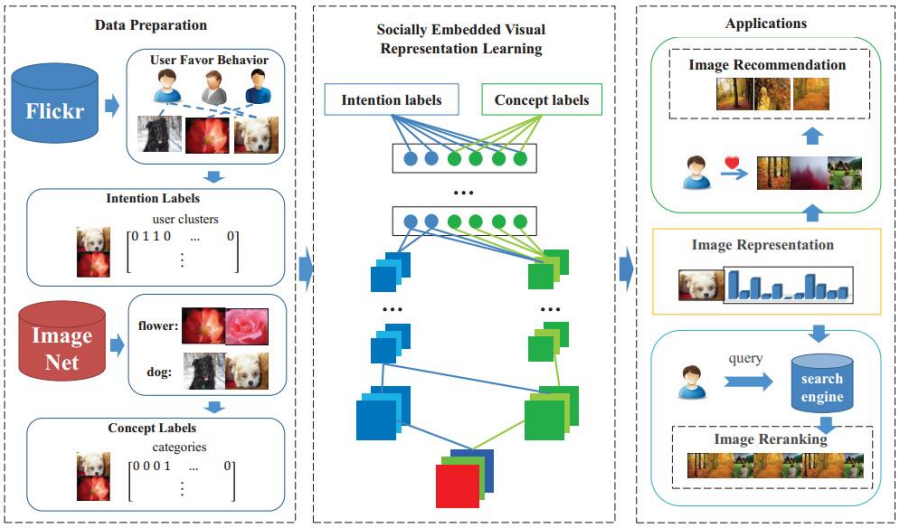


图 2. 12非对称的多任务卷积神经网络模型框架图。

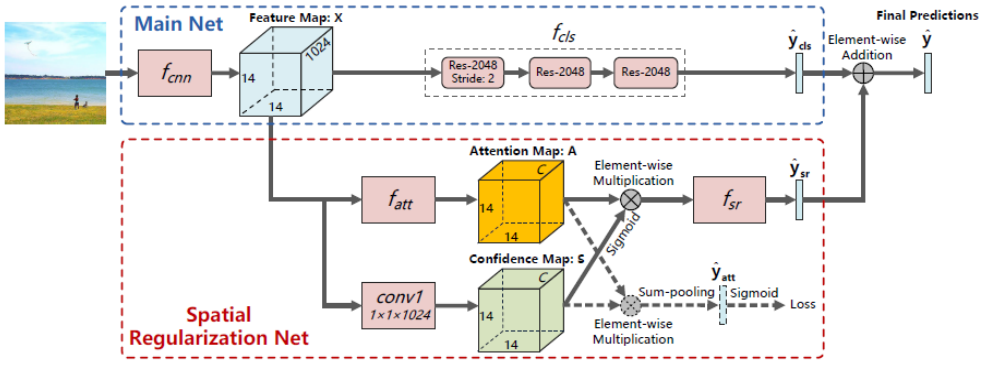


图 2. 13空间正则化的深度神经网络模型。图 2.9 上部分的 Main Net 是在沿用的 Resnet 的基础上重新训练新的分类器。下部分是采用 attention 机制来获取图像标签之间的空间和语义信息的空间正则化网络模块。

在属性分类工作之后，人们也更多的考虑对象属性分类反作用于图像分类来提高模型的分类正确性的办法。文章[56]提出一种基于属性分类的人脸验证方法。通过训练的二分类器来判别具有或者不具有某种特定的属性，从而实现对人脸的验证工作。文章[57]中同样采用基于对象的属性分类来进行 Zero-Shot 的物体类别分类的工作。

我们通过对象的属性分类能够看出，属性分类能够帮助我们更好的获取图像的关键信息并更好的理解图像。而我们知道，属性是对于一个对象的抽象刻画也是对对象的性质和对象之间的相互关系的统称。属性有本质属性和非本质属性之分，有浅层次属性和深层次属性等不同层次概念的属性之分。因此，属性不仅包含对象局部特征的浅层次属性，同样，属性也应该包含图像全局特征的深层次属性。当人来区分一个物体与另一个物体或者说一个事物和另一个事物是否相同时，比较的是一个物体的属性和另一个物体的属性是否相同。这种比较更多的是基于物体的深层次属性而来的。本文中提到的多属性的图像分类是基于对象全局的深层次属性而言的。全局的深层次属性具有浅层次属性不具有的更强的相关性。这种相关性是浅层次属性之间所不具备的。因此，在实现物体

深层次属性分类工作的同时，我们可以充分利用属性之间的这种相关性来提高我们模型的分类性能。同时，更多深层次属性的辨别，有助于我们更好的对当前对象进行分析和理解。

第三章 局部非对称的多任务卷积神经网络模型

3.1 概述

对于计算机视觉来说，其主要研究内容为如何对于给定的图像或者图像序列集进行分，从而得到对于图像尽可能详细的、正确的、不同角度的正确的描述[37]。图像理解在研究内容上与计算机视觉有一定的交叉，图像理解更注重在分析图像的基础上去理解图像本身所包含的内容和含义。我们知道，图像理解是深度学习模型较早的一个应用领域，同时也是被最广泛研究的一个领域。随着大规模标记的图像数据集（如 ImageNet）的发展，深度卷积神经网络模型在大规模数据集的处理中显示出了不可替代的优越性能。卷积神经网络模型被广泛的应用于图像理解任务中[38]。

近年来，以后很多关于图像分类任务的研究。但大多数根据图像内容进行单任务（单标签）的或者多属性的分类工作。而对于多任务图像的分类问题来说，之前的方法更多的是采用同一个网络提取相同的表示特征采用不同的分类器来进行分类或者采用不同的神经网络模型提取不同的表示特征并对特征进行融合，对融合之后的特征采用不同的分类器来进行分类。

由文章[39]可知，卷积神经网络在训练过程中底层特征更多的是纹理和色彩特征即是一些共通的表示，而越高层的特征表示能力、抽象能力越强。对于图像的多属性分类（对同一幅图像标记多个标签）来说， 底层特征应该是共通的或者说应该共享某些参数使得学习到的底层特征更加的鲁棒，而高层特征抽象能力更强，分类模型应该独立来学习而避免因为融合或者一起学习而造成的模型抽象能力削弱。

根据上文的分析，如何对设计应对多任务分类的卷积神经网络结构并对卷积神经网络结构进行更改而使其拥有更高的特征学习能力是模型研究的关键。

3.2 局部非对称的多任务卷积神经网络模型架构

局部非对称的多任务卷积神经网络（PAMT-CNN）架构如图 3.1 所示。从图中可以看到对应相同分类任务的两个卷积网络，而网络之间有一定的交互（权值共享）。其中“Conv”，“Pool”和“Fc”分别表示的是卷积层，池化层（Pooling Layer）和全连接层。而“Conv5_2”则表示为任务二中的第五个卷积层，“3*3@56”则表示的是该层的卷积核大小为 3 乘以 3，同时该层有 56 个不同的该尺寸的卷积核。

网络结构如此设计的原因在于：1）不同的任务对应于输入图像不同层面的语义信息学习，因此我们需要两条从原始输入数据到标记的监督信息的学习路径。为了实现对不同层面语义的学习，PAMT-CNN 模型中上层虚框对应的路径用于实现任务一对应的图像

属性语义的学习，而下层虚框对应的路径则用于实现任务二对应的图像属性语义的学习。2）不同的任务是对于同一幅输入图像而言的，因此两者之间有一定的相互关联，而低层特征更多的是一些通用的表示。因此模型在低层进行了一定的权重的共享[40]，共同来学习基于输入图像的更加鲁棒的低层特征表示。3）就像前面讨论的一样，网络高层的特征表示更加的抽象[39]。我们将更非常规的，意识中更需要接收高层指导的任务放在下层卷积路径中。然后在“Conv5”中接收上下两条路径的语义信息指导从而形成非对称的多任务的卷积神经网络结构。

而模型在语义嵌入时选择的是卷积神经网络的低层卷积层。这样做的原因在于文章[39]中指出，卷积神经网络不同的卷积层关注着图像不同的特征描述：卷积层二更多关注的是图像的边缘、颜色和梯度特征，卷积层三更多关注的是图像的复杂的不变性和纹理的特征，卷积层四关注的是具有类别倾向的重要区分性特征而卷积层五则对应着重要位置变化的抽象表特征表示。因为我们认为，同一幅图像的第二卷积层的边缘、颜色和梯度特征，第三卷积层所关注的复杂的不变性和纹理特征应该是一致的。我们通过语义的相互嵌入，使得低层的特征能够实现更共同的通用特征表示的低层特征的学习，从而为更高层抽象特征的提取打下基础，并在一定程度上弥补了网络底层特征与高层特征之间的语义鸿沟。同时我们将任务一与任务二的第四个卷积层连接后作为任务而第五个卷积层的输入，目的是为实现一种单向的指导。我们默认为任务一对应的属性有更强的可分性，通俗的说就是在同样网络结构的情况下，分类的正确率能够更高。因为第四个卷积层对应的图像特征更多的是具有类别倾向的，这样的指导能够实现任务二分类性能的更大程度上的提升。

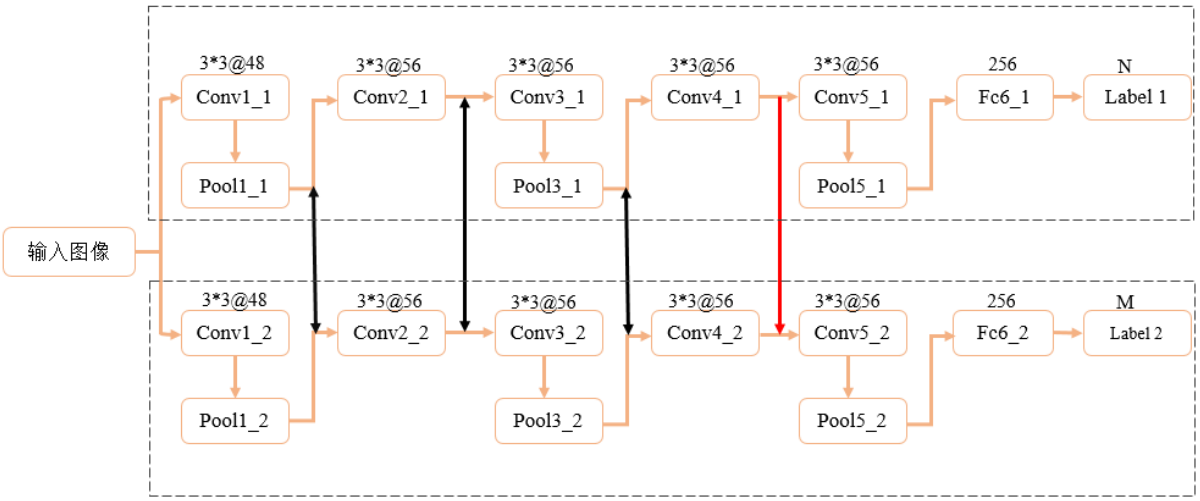


图 3. 1 局部非对称的多任务卷积神经网络结构图。顶端的虚线框卷积用于实现对于输入图像任务一的分类，而底端的虚线框卷积用于实现对同一幅输入图像的任务二进行分类。在两层之间的黑色实线双向箭头表示两个网络间低层的交互，用于实现特征学习过程中的相互指导。红色向下箭头表示任务一对于任务二高层抽象特征学习的指导。

现在我们更详细的介绍我们的非对称的多任务卷积神经网络结构。在分类任务一中，

有五个卷积层外加两个全连接层。在任务一的五个卷积层中的第一个，第三个和第五个卷积层的后面都会有一个池化层与之相连。从图 3.1 中我们可以看到每个卷积层的卷积核大小与个数。而池化层中，池化的核（Kernel）的尺寸为 3 乘以 3，同时池化层的步长同样为 3 个像素点。

在前馈阶段（forward propagation），从第$(l-1)^{th}$层向第l^{th}层的传递函数可以形式化成如下两种不容的形式：

$$x_l^1 = \sigma(W_{l-1}^{1,1}x_{l-1}^1 + b_{l-1}^{1,1} + W_{l-1}^{1,2}x_{l-1}^1 + b_{l-1}^{1,2}), 1 < l \le 4 \quad \text{（式 3.1）}$$

其中，$\sigma(*)$表示激活函数，当$x > 0$时时$\sigma(x) = x$，反之则为 0。而$W_{l-1}^{1,i}(i = 1,2)$则表示的是从第i^{th}条路径的第$(l-1)$层向第 1 条路径上的第l层传递的权重，b_{l-1}^{i}则是对应的偏置。

$$x_l^2 = \sigma(W_{l-1}^1 x_{l-1}^1 + b_{l-1}^1), 4 < l \le 7 \quad \text{（式 3.2）}$$

其中W_{l-1}^1表示的是从第$(l-1)$层向第l层传递的权重，b_{l-1}^1则是对应的偏置。最终的输出层由$\tilde{d}$表示，我们定义$\tilde{d}$的数学表示形式如下：

$$\tilde{d} = soft\max(\sigma(W_7^1 x_7^1 + b_7^1)) \quad \text{（式 3.3）}$$

其中$soft\max(*)$为$soft\ \max$函数。

与上层网络卷积路径相同，我们在下层卷积路径上采用相同的设置。但是，我们在连接上有一点不同，除了最后的两个连接层以外，在下层卷积路径上的所有卷积层的输入均来自当前路径的前一层与上层路径上对应的前一层的输出。我们形式化上述描述如下：

$$x_l^2 = \sigma(W_{l-1}^{2,1}x_{l-1}^2 + b_{l-1}^{2,1} + W_{l-1}^{2,2}x_{l-1}^2 + b_{l-1}^{2,2}), 1 < l \le 5 \quad \text{（式 3.4）}$$

最终的下层输出层由$\tilde{o}$表示，我们定义$\tilde{o}$的数学表示形式如下：

$$\tilde{o} = soft\max(\sigma(W_7^2 x_7^2 + b_7^2)) \quad \text{（式 3.5）}$$

3.3 局部非对称的多任务卷积神经网络模型训练

对于给定的输入图像，我们期望对应的输出$\tilde{d}_i$和$\tilde{o}_i$能够与目标分类向量d_i和o_i能够更加的接近。此处的接近是指欧式距离数值更小。d_i和o_i为对应的 N 维和 M 维一维向量表示，其中 N 和 M 分别对应在图像数据集中的不同类别数。第一层卷积神经网络中

损失函数定义为：

$$L_1 = \frac{1}{2}\sum_{i=1}^{N}(d_i - \tilde{d}_i)^2 + \frac{\lambda}{2}\sum_{I=1}^{3}(\left\|W_1^{11}\right\|_F^2 + \left\|W_1^{12}\right\|_F^2) + \frac{\mu}{2}\sum_{I=4}^{7}\left\|W_I^1\right\|_F^2 \quad （式 3.6）$$

其中 $\tilde{d}_i$ 是对于输入图像 I_i 从局部非对称卷积神经网络模型的第一个路径上的输出；λ 和 μ 是平衡损失的因子并用来正则化之前的过拟合影响。

而对于第二层的卷积路径来说，我们设定其损失函数为 L_2，该损失函数可以形式化表示成如下形式：

$$L_2 = \frac{1}{2}\sum_{i=1}^{N}(o_i - \tilde{o}_i)^2 + \frac{\lambda}{2}\sum_{I=1}^{4}(\left\|W_1^{21}\right\|_F^2 + \left\|W_1^{22}\right\|_F^2) + \frac{\mu}{2}\sum_{I=5}^{7}\left\|W_I^2\right\|_F^2 \quad （式 3.7）$$

在训练的过程中，我们从数据集中选择一批原始图像作为输入。采用 L_1 去更新上层卷积神经网络的权重参数 W_1，采用 L_2 去更新下层卷积神经网络的权重参数 W_2。我们重复上述操作知道损失误差收敛。对于任意给定的图像 I_i，我们可以通过上述训练的模型来同时预测其对应的不同的标签。

3.4 实验评测

3.4.1 数据集

为验证我们网络结构设计的有效性，我们选择了两个数据集：Food 数据集和 CompCars 数据集。我们从选择的两个数据集中筛选出满足我们要求的部分子集作为我们的实验数据集。

我们选择的 Food 数据集的子集包含 24690 幅图像，每一幅图像分别被标记上菜品名称和对应的餐馆名称作为我们多任务分类的多标签。我们保证最终选定的子集中对于给定的菜品标签，其中包含不少于 145 幅图像。同样我们去除了包含少于等于三种菜品的餐馆，来保证每个餐馆标签下包含的菜品种类大于三种。我们的子数据集中包含 100 个餐馆标签和 100 个对应的菜品标签。我们从中选择 18626 幅图像用于训练，1122 幅图像用于验证，4942 幅图像用于测试。因为该子集中的图像数目分布不是很均匀，因此我们在图 3.2 中给出了数据集中对应菜品和餐馆中菜品图像数目的分布图。出于空间的考虑，我们选用阿拉伯数字来表示对应的菜品和餐馆。图 3.3 给出了我们当前 Food 数据集[41]子集的一些样例。

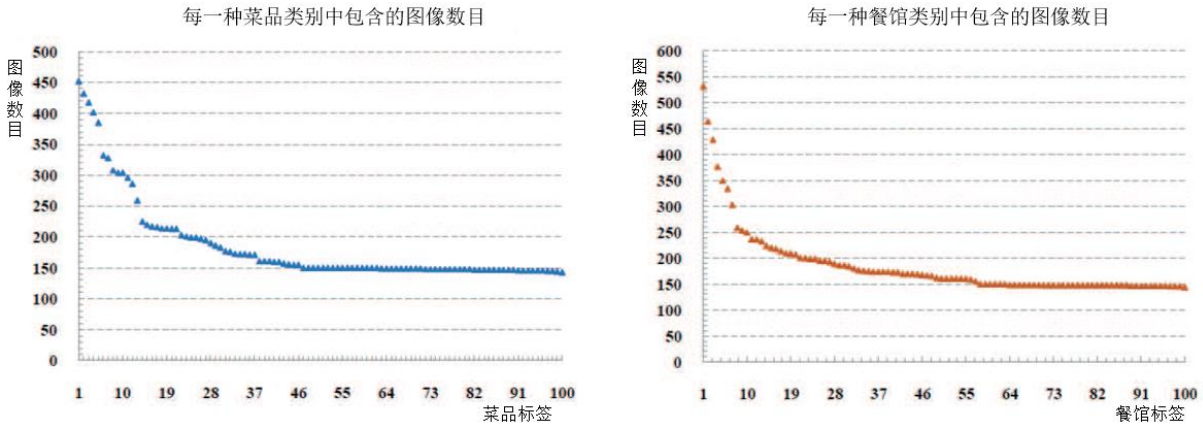


图 3.2 Food 数据集中图像数目分布图。为了空间的考虑，我们采用阿拉伯数字来表示对应的餐馆和菜品名称。

我们选择的 CompCars 数据集[42]的子集包含 25000 幅汽车图像，每一幅图像分别被标记上汽车厂商名称和对应的汽车类型作为我们多任务分类的多标签。我们最终选定的子集中包含 12 种汽车类型标签以及 100 种汽车厂商标签。我们保证最终选定的子集中对于给定的每一种标签，其中包含不少于 200 幅图像。我们从中选择 15000 幅图像用于训练，2000 幅用于验证，8000 幅用于测试。图 3.4 给出了我们当前 CompCars 数据集子集的一些样例。

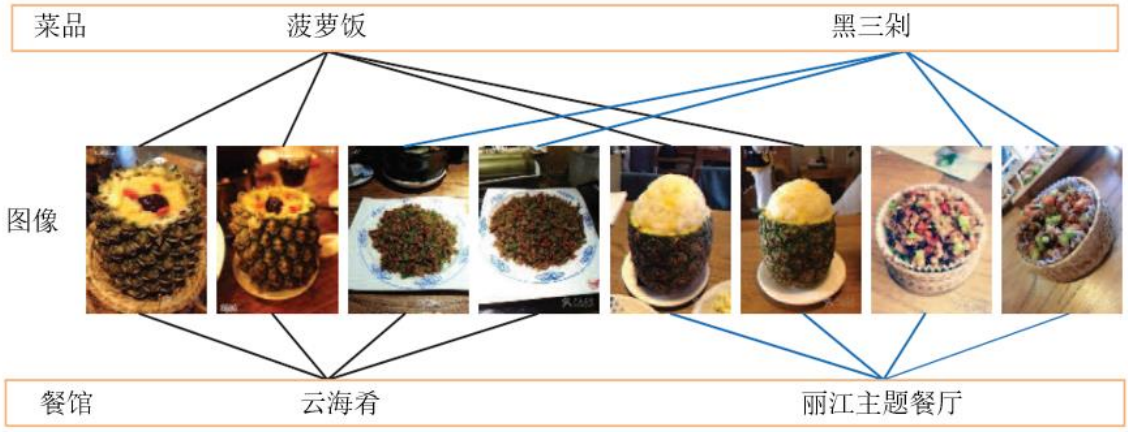


图 3.3 Food 数据集子集举例。样例图像中我们列出了两种菜品：菠萝饭和黑三剁。样例中包含两个餐馆：云海肴和丽江主题餐厅。其中每个餐馆均包含上述两种菜品，每一种菜品中包含的菜品图像分别来自不同的两个餐馆。

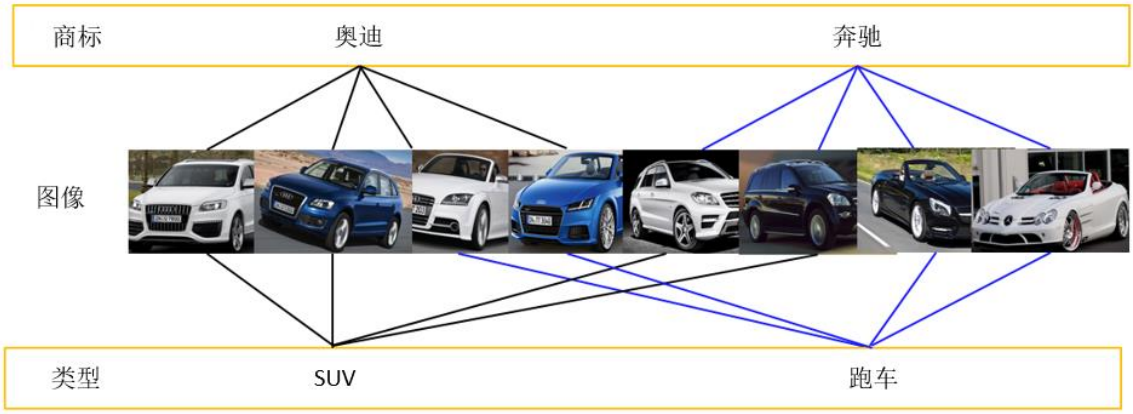


图 3.4 CompCars 数据集子集举例。样例图像中我们列出了两种汽车厂商：奥迪和奔驰。样例中包含两种汽车类型：SUV 和跑车。其中每个厂商均包含上述两种类型的汽车，同时每一种类型的骑车中包含的汽车图像分别来自不同的两个厂商。

3.4.2 实验参数设定

在实验参数设置方面，我们选择了和 AlexNet[2]相似的设置。每一幅输入图像均被缩放成 256x256 后分别从上下左右四个角和中心裁剪出五个 224x224 尺寸的区域图像。最后两个全连接层的 dropout 值设置成 0.5。每一层的学习率起始值设置成 0.01，每一次学习率降低时除以 10，直到错误率不再下降为止。实验是在单个内存为 3G 的 GTX780Ti 显卡上完成的。

3.4.3 图像的多属性分类评测

为了验证我们模型的有效性，我们选择了六种比较的 baseline：

1）单网络模型（CNN-ST）：分别将多任务分类中的每个任务单独训练一个卷积神经网络模型并单独去验证每一个模型的性能；

2）非对称的多任务卷积神经网络模型[31](amtCNN)：多任务中的两个分类任务采用相同的神经网络结构，并且任务二的每一层的输入来自两个任务的前一层的输出[43]。

3）反转非对称的多任务卷积神经网络模型(amtCNN-Inv)：多任务中的两个分类任务仍然采用相同的卷积神经网络结构，但是任务二的每层卷积输出将作为任务一对应的上层的输入，即任务一的每一层将收到任务二对应的前一层的指导。

4）局部对称的多任务卷积神经网络模型(PAMT-CNN-4D)：多任务中的两个分类任务仍然采用相同的卷积神经网络结构，但是两个网络的卷积层输入均来自两个任务对应的前一层神经网络输出。

5）局部非对称的多任务卷积神经网络模型 1（PAMT-CNN-4D-2S）：在 PAMT-CNN-4D 的基础上，任务二的前两个全连接层的输入由原本对应的前层更改为将任务一和任务二对应的前层网络输出连接后作为对应的输入。

6）局部非对称的多任务卷积神经网络模型 2（PAMT-CNN-4D-3S）：在 PAMT-CNN-4D-2S 的基础上，将任务二最后一个全连接层的输入由原来对应的任务二前层改为任务一和任务二对应的前层网络输出连接后后作为对应的输入。

表 3.1 不同方法在 Food 数据集上的分类准确率。

方法	在菜品分类上的准确性	在餐馆分类上的准确性	平均准确性
CNN-ST	70.01%	56.16%	63.085%
amtCNN	70.65%	58.89%	64.77%
amtCNN-Inv	73.72%	56.21%	64.965%
PAMT-CNN-4D	73.25%	61.41%	67.33%
PAMT-CNN-4D-2S	72.78%	63.48%	68.13%
PAMT-CNN-4D-3S	70.92%	63.19%	67.055%
PAMT-CNN	**74.87%**	**64.75%**	**69.81%**

表 3.2 不同方法在 CompCars 子数据集上的分类准确率。

方法	在类型上的准确性	在厂商分类上的准确性	平均准确性
CNN-ST	67.06%	35.67%	51.365%
amtCNN	NA	NA	NA
amtCNN-Inv	NA	NA	NA
PAMT-CNN-4D	64.72%	42.68%	53.7%
PAMT-CNN-4D-2S	67.51%	44.59%	56.05%
PAMT-CNN-4D-3S	67.32%	43.74%	55.53%
PAMT-CNN	**68.50%**	**46.84%**	**57.67%**

在两个数据集上的分类比较结果分别展示在表 3.1 和表 3.2。从两个表中的实验结果中我们可以得到以下结论：

1）在不分低层特征和高层特征学习中所表现出的与原始数据的关联性而直接进行单向的指导训练（amtCNN）对最终的分类结果或有提升或有下降，并没有起到决定性的作用。从表 3.1 中我们可知，当多任务中两个任务的分类性能相差更小，或者说区分度相当（在菜品分类上的准确性较在餐馆分类的准确性高仅 14 个百分点）的情况下，amtCNN 能够对分类结果有较小的提升，而当多任务中的两个任务分类性能相差较大，或者说区分度相差较大（在类型上分类的准确性高于在厂商分类的准确性近 32 个百分点）的情况下甚至起到的反作用。因此，我们在多任务网络的训练过程中要区分来对待高层特征和低层特征。

2）在仅考虑低层特征（PAMT-CNN-4D）并实现两个任务的相互“指导”时，我们的网络结构在两个实验数据集上的分类性能均得到了显著的提升（平均提升约 3.3 个百分点）。我们认为该结论证明了我们之前的观点，神经网络所提取的图

像知识表示的低层特征更多的是通用的、鲁棒性的特征。而多任务分类中是对同一幅图像对应的不同标签进行预测，低层网络之间的交互学习能够使得学习到的特征更加鲁棒。也就是实现了学习过程中的相互指导，从而使低层特征更好的为高层抽象特征的提取打基础。也正是如此才使得该网络模型在分类性能上取得了更好的结果。

3）通过对比 PAMT-CNN-4D-2S、PAMT-CNN-4D-3S 和 PAMT-CNN 我们发现，当我们进行高层抽象特征之间的过多“指导”时，我们的分类性能不升反而进行了一定的降低。因此我们也能够得出这样的结论：高层抽象特征之间的相互指导或者学习，会影响当前特征的抽象性。因此，我们在高层特征中的低层特征网络上对分类能力相对较差的任务进行一定的指导会有利于分类的整体效果。

为了更好的分析局部非对称的多任务卷积神经网络模型对分类的影响，我们选择了 PAMT-CNN-4D-2S、PAMT-CNN-4D-3S 和 PAMT-CNN 模型在 Food 子数据集中多任务分类中的每一类的正确性进行了比较。比较结果展示在图 3.4 和图 3.5 中。

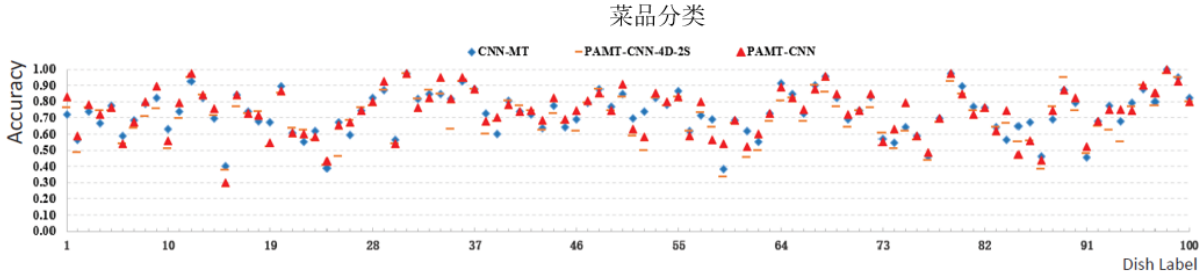


图 3. 5 不同方法在 Food 数据集菜名标签分类正确率比较。不同的局部非对称的多任务卷积神经网络模型在 Food 子数据集的菜品分类中的每一类的分类正确性。为空间的考虑，我们选用阿拉伯数字来代替每一类对应的菜品名称。

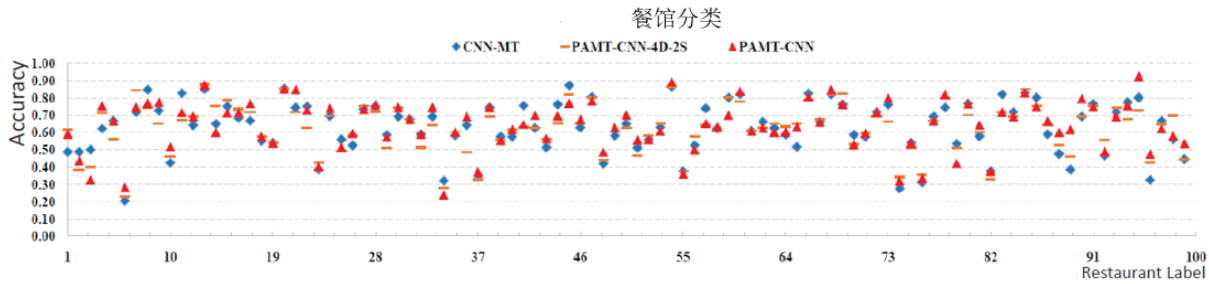


图 3. 6 不同方法在 Food 数据集餐馆标签分类正确率比较。不同的局部非对称的多任务卷积神经网络模型在 Food 子数据集的餐馆分类中的每一类的分类正确性。为空间的考虑，我们选用阿拉伯数字来代替每一类对应的餐馆名称。

3.5 小结

在多任务分类中，一幅图像会对应多个属性标签。文中提出了一种局部非对称的多任务卷积神经网络模型用属性任务图像分类。局部非对称的多任务卷积神经网络模型在

设计上考虑到多属性语义标签的相互嵌入，通过语义的嵌入能够在一定成都上弥补网络底层和网络高层语义鸿沟，实现低层共同的特征表示更加的鲁棒。低层特征的鲁棒性也为高层抽象特征的学习打下了坚实的基础。在神经网络的第五层的输入是两个任务第四层的输出进行融合，也就是在一定程度上采用了任务一来指导任务二的学习。通过采用的局部的非对称的“指导”，从而在一定程度上提升了多任务分类模型分类的准确性。

最后通过实验证实了，卷积神经网络低层参数的共享对分类结果的提升有较明显的作用。同时也验证了文章[39]所说，卷积神经网络在训练过程中底层特征更多的是纹理和色彩特征即是一些共通的表示。两个卷积神经网络低层参数的共享使得学习到的这个低层特征具有更共通的表示。同时文章也通过实验验证了高层特征的抽象性。抽象性则与分类语义更加的接近。因此，高层的过多“指导”会影响图像多属性分类任务的最终分类性能。

第四章 融合多层特征的互影响卷积神经网络模型

4.1 概述

在上一章中提出了应用于图像多属性分类的局部非对称的多任务卷积神经网络模型。同时，我们也在上一章中通过局部非对称的多任务卷积神经网络模型验证了文章[39]中的观点，卷积神经网络的低层关注的更多的是颜色、边缘、梯度、具有复杂不变性和纹理等共通性的特征表示，而卷积的高层关注的更多的是具有类别倾向的重要区分性特征表示。我们将多个应用于单任务分类的深度卷积神经网络模型在网络低层进行一定参数共享，从而使学习到的不同任务间低层共通性的特征表示更加鲁棒。更加鲁棒的低层共通性特征表示的实现也在一定程度上为高层特征的抽象奠定了基础。我们在局部非对称的多任务卷积神经网络模型的高层卷积层采用的是非对称的单向“指导”，其目的是为了在不影响区分性更强的分类任务的分类准确率的前提下提高其他分类任务的分类正确性。

传统的深度卷积神经网络包括局部非对称的多任务卷积神经网络均采用模型最后一层的抽象特征作为最终的分类特征，而没有考虑到多层特征融合对分类性能带来的影响。由文章[29]可知，采用多层特征融合的卷积神经网络可以被看作是一个有向无环的卷积神经网络模型（DAG-CNNs），而该模型可以有效的应用于图像的分类工作中去，并提高模型的分类性能。通过图 3.4 和图 3.5 我们可以看到，抽象层在不同尺度上的相互“指导”使得每一类的分类正确性有一定的浮动，这也正说明了模型不同网络层特征的融合对最终分类性能一定的影响。

在本章我们逐一分析了不同网络层特征的融合对于传统卷积神经网络模型分类性能的影响。之后我们在局部非对称的多任务卷积神经网络模型的基础上加入多层特征融合，并借鉴上章中提到的多任务间的相互“指导”机制，提出融合多层特征的互影响卷积神经网络模型。最后，我们在上述两个数据集上通过实验证明了该方法在提高模型分类性能上的有效性。

4.2 融合多层特征的神经网络模型选择

文章[29]中指出，图像的分类任务需要融合卷积神经网络的多层特征形成最终的特征表示，从而更好的提高模型的分类性能。当应对粗分类任务时，我们需要提取出对每一类物体形变具有鲁棒性的高不变性的特征表示。而这一工作由传统的拥有一个输出层的卷积神经网络就能够很好的胜任。高层的抽象特征表现出很强的抽象性与不变性。但应对细分类问题时，我们不但需要高层的极度抽象的特征表示，我们还需要低层或者说

中层的一些形状、纹理等低层次共通性的特征表示的。不同卷积层的图像特征的有机结合才能更大限度的提升模型的分类性能。

也正是基于这样的考虑，文章[29]引出用于多尺度图像识别的 DAG-CNNs 模型。利用 DAG-CNNs 模型从不同的卷积层提取不同尺寸的特征，并验证了该模型能够在一定程度上较好的提高模型的分类性能。多尺度特征的提取是不需要其他额外的计算量的，因为他们是在网络进行前馈的过程中计算好的。而最终的 DAG-CNNs 模型是一个前向的有向无环图的结构。DAG-CNNs 网络模型与传统的 CNN 模型的对比图如下。

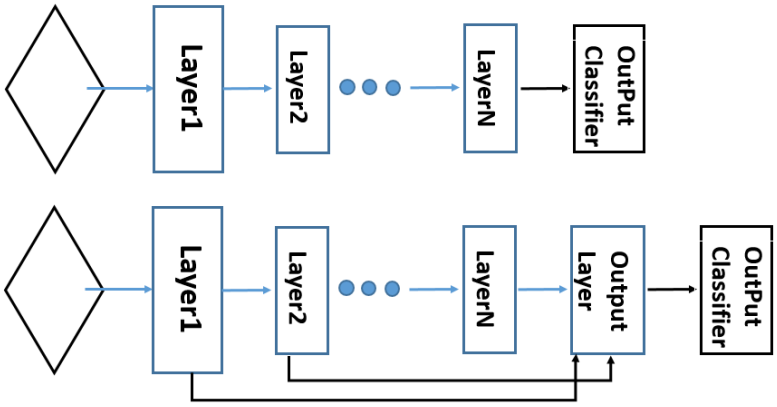


图 4.1传统深度网络模型与多层特征融合的 DAG 网络模型对比图。

然而，对于多层特征融合的方法来说，有一个很显而易见的困难问题是特征维度：融合后的特征在特征维度上很容易达到很难控制的程度。这也使得我们采用了多层特征融合的卷积神经网络模型难以训练。因此，我们在网络的多层特征融合的引入上加入最大值池化这一步骤。从这一角度来看，DAG-CNN 卷积神经网络模型和之前的通过空间池化（spatial pooling）来得到多尺度特征就很相似了。类似的工作也包括：BOW（bag of words）模型[44]，空间金字塔理论（spatial pyramids）[45]，多尺度模板[46]和随机池化模型[47]。

该模型仍然可以通过端到端（end-to-end）的训练方法来训练，仍然是一个前馈网络模型，但是在整体的结构上看来，已经不再是传统的线性的或者说是链式的结构了，而是一个有向无环图的结构。因为很多卷积神经网络工具包[48][49]的支持，该网络架构模型可以很容易的被应用到当前的卷积神经网络框架中来。 DAG 结构的卷积神经网络最早是在递归神经网络模型[50][51]的上下文中被探讨。同样，在最近的相关研究中，卷积神经网络结构中经常采用“skip”连接来关联不同的卷积层[52][53]。通过融合卷积神经网络不同层特征，可以在一定程度上提高网络的分类性能。

当所有的卷积层输出均连接到最后的输出层会影响最终的分类性能，造成分类准确率的降低。造成这一现象的原因是因为过拟合引起的，即添加所有层到最终的输出层仍可以提高最终的分类性能，但是因为过拟合会损失最终测试时的性能。因此，要将 DAG-CNNs 模型应用到我们的数据集上就需要考虑这样一个问题：将哪些层连接到最终的输出层会提高我们最终的分类性能而不会造成过拟合这样的问题。

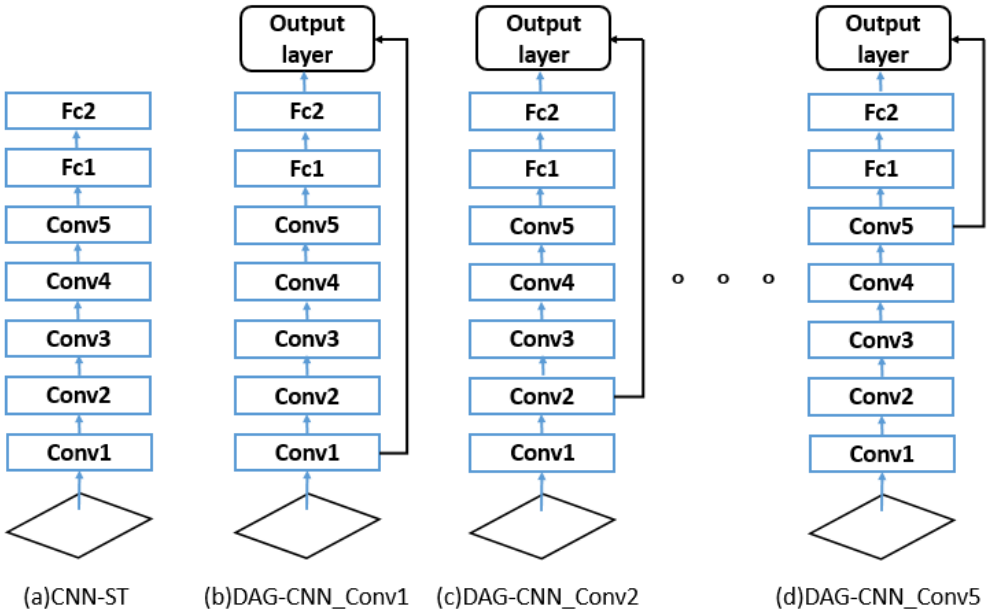


图 4.2融合不同层特征的 DAG 网络模型结构示意图。（a）单卷积神经网络 CNN-ST 结构示意图，（b）由第一层卷积层引出的多特征融合的 DAG 模型 DAG-CNN_Conv1，（c）由第二层卷积层引出的多特征融合的 DAG 模型 DAG-CNN_Conv2，（d）由第五层卷积层引出的多特征融合的 DAG 模型 DAG-CNN_Conv5。

我们采用上一章提到的 CNN-ST 模型作为基准，并在该模型的基础上引入多层特征融合。我们在图 4.2 中给出我们各多层特征融合的 DAG 网络模型的结构示意图。我们仅考虑由某一层引出的多特征融合问题而不考虑由多个网络层特征进行融合的问题（我们只会将中间某一层的输出连接到最后的输出层形成融合后的特征）。不同结构在不同数据集上对应的实验结果我们展示在表 4.1 和表 4.2 中。

表 4.1 融合不同层特征的 DAG 模型在 Food 数据集上的分类性能对比。“Conv5”表明仅由第五个卷积层的输出连接到最终的输出层形成最终的特征表示并用来实现分类的多层特征融合的 DAG 卷积神经网络模型。

方法	在菜品分类上的准确性	在餐馆分类上的准确性
CNN-ST	70.01%	54.16%
DAG-CNN_Conv5	**71.81%**	54.90%
DAG-CNN_Conv4	68.70%	**55.22%**
DAG-CNN_Conv3	58.44%	40.27%
DAG-CNN_Conv2	47.35%	32.65%
DAG-CNN_Conv1		

表 4.2 融合不同层特征的 DAG 模型在 CompCars 子数据集上的分类性能对比。“Conv5”表明仅由第五个卷积层的输出连接到最终的输出层形成最终的特征表示并用来实现分类的多层特征融合的 DAG 卷积神经网络模型。

方法	在产商分类上的准确性	在类型分类上的准确性
CNN-ST	35.67%	67.06%

DAG-CNN_Conv5	35.67%	**69.04%**
DAG-CNN_Conv4	**38.36%**	65.94%
DAG-CNN_Conv3	30.98%	60.55%
DAG-CNN_Conv2	24.77%	51.02%
DAG-CNN_Conv1		

通过上表中的实验结果我们能够看出，DAG-CNNs 网络模型在 Food 子数据集和 CompCars 子数据集的多分类任务上均能实现分类结果准确性的提高。然而我们依然可以看出：1）将卷积神经网络的第一个卷积层的输出连接到最终的输出层会导致网络训练不收敛，我们并没能很好的分析出不收敛的原因。2）卷积神经网络的第二层以及第三层的输出分别连接到最终的输出层时，分类效果都有大幅度的降低。卷积层次越低，导致的分类效果越差。没有实现文章[33]中所验证的结果，添加卷积层输出到最终的输出层总是会提高最终的分类性能。3）从表中的两个多分类任务的实验性能中我们还能发现一个有趣的问题，当一个任务的可分性越强（Food 子数据集中的菜品分类任务和 CompCars 子数据集的餐馆分类任务）时，高层（卷积层 5）的输出在提高分类性能上的作用越明显。同样，当一个任务的可分性相对较弱（Food 子数据集中的餐馆分类任务和 CompCars 子数据集的厂商分类任务）的时候，较高层(卷积层 4)的输出对最终分类结果在性能上的提高越明显。4）实验结果没有体现出添加卷积层总会提高最终的分类效果的另一个原因可能是因为我们模型的简单性，在引入 DAG 之后导致模型参数的过多。在模型参数过多而数据量不够的情况下是无法有效实现网络模型的训练从而到损失函数不收敛和分类性能达不到预期的结果。

4.3 融合多层特征的互影响卷积神经网络模型

在本小节，我们将会更详细的介绍融合多层特征的互影响卷积神经网络模型（ME-DAG-CNN）。融合多层特征的互影响卷积神经网络模型是在上一章介绍的局部非对称的多任务卷积神经网络模型的基础上引入多层特征融合而构成的。虽然上一章提出的局部非对称的多任务卷积神经网络模型较传统的单独训练的神经网络模型在图像的多属性分类任务上有较大的性能提升，同时也验证了我们之前的假设，在卷积网络低层实现图像属性标签语义的相互嵌入能够使得网络低层特征在提取过程中融入高层抽象语义特征，也能够使得任务间低层共通性表示的特征更加鲁棒。这种语义的嵌入和更共通的特征表示提升了模型的分类性能。

然而我们发现，局部非对称的多任务卷积神经网络在最终的分类时仍采用的是最后一层抽象的特征表示，而没有考虑多层特征融合对分类性能的影响。多层特征融合是提高卷积神经网络模型分类正确率的另一个有效的办法。因为，图像的分类工作更好的实现其本身就是需要融合不同网络层特征来共同实现。例如：区分人和狗，我们更多需要的是对形变鲁棒的高不变性特征，而区分形状相似的两款汽车的生产厂商则需要更强的

纹理和轮廓等信息。

由文章[29]可知，融合了网络多层特征的卷积神经网络可以被看作是一个有向无环图（DAG-CNNs），而该模型可以有效的应用于图像的分类工作中去，并得到较好的性能提升。而通过上一章的图 3.5 和图 3.6 我们可以看到，抽象层在不同网络层上的相互“指导”使得每一类的分类正确性有一定的浮动，这也正说明了融合不同网络层特征对最终分类性能一定的影响。

融合多层特征的互影响卷积神经网络模型（ME-DAG-CNN）的网络整体框架图我们可以参考图 4.3。该结构可以被看作是对应的 DAG 前向反馈网络。同时，该结构支持端到端的模型训练。

4.3.1 融合多层特征的互影响卷积神经网络模型架构

融合多层特征的互影响卷积神经网络模型是在上一章的局部非对称的多任务卷积神经网络模型的基础上改进而来的。通过图 4.3 局部非对称的多任务卷积神经网络模型和图 4.4 融合多层特征的互影响的卷积神经网络模型对比可以发现，融合多层特征的互影响的卷积神经网络模型是在局部非对称的卷积神经网络模型基础上取消了高层卷积层的单向语义嵌入而引入了网络多层特征的融合“指导”。

在网络多层特征融合部分，我们选择了第五卷积层和网络最后的全连接层特征进行融合。选择第四卷积层的原因是因为我们在前面一节融合多层特征的神经网络模型选择部分通过实验验证了卷积层高层（第四层和第五层）特征和网络最高层的全连接层融合后能在分类性能上得到最大的提升。因此，在融合多特征的互影响卷积神经网络模型中也选择将卷积层的高层（第四层卷积层）的输出同最高层的全连接层特征进行融合。而为什么在融合多层特征的互影响的卷积神经网络模型的特征融合选择部分选择的是交互的方式，是因为我们认为图像的多属性中的某种属性能够在一定程度上帮助提升图像另一种属性的分类性能。例如：我们的 Food 数据集中，有汉堡的图像。汉堡的图像对应有两个属性的标签——汉堡（Food 名）和 KFC（餐馆名）。但当我们看到汉堡图像时，我们脑海中想到的汉堡对应的餐馆中一定会有 KFC（餐馆），而当我们想到 KFC（餐馆）时一定会想到餐馆中的菜品（Food）。所以在融合多层特征的互影响卷积神经网络模型的设计中将引入的多层特征进行了交互的融合。

由图 4.4（融合多层特征的互影响卷积神经网络模型框架图）我们能够看出，该模型结构中包含卷积层（Conv），ReLu 层，池化层（Pool）和全连接层（Fc）。“3x3,Conv1,48”表示的是该层的卷积核尺寸是三乘以三，该层为一个卷积层，包含的卷积核的个数为 48。该模型最终的全连接处接连一个对应有 K 个输出的 Softmax 函数用于预测最终的判别结果。在模型的多层网络特征融合部分，我们将卷积层之后对应的 ReLU 层作为输入，连接上一个均值池化层之后通过全连接层形成一个 1x256 维的输出向量。该输出将与对应的全连接层进行加和操作之后作为最终 Softmax 层的输入，从而得到最终的预测结果。

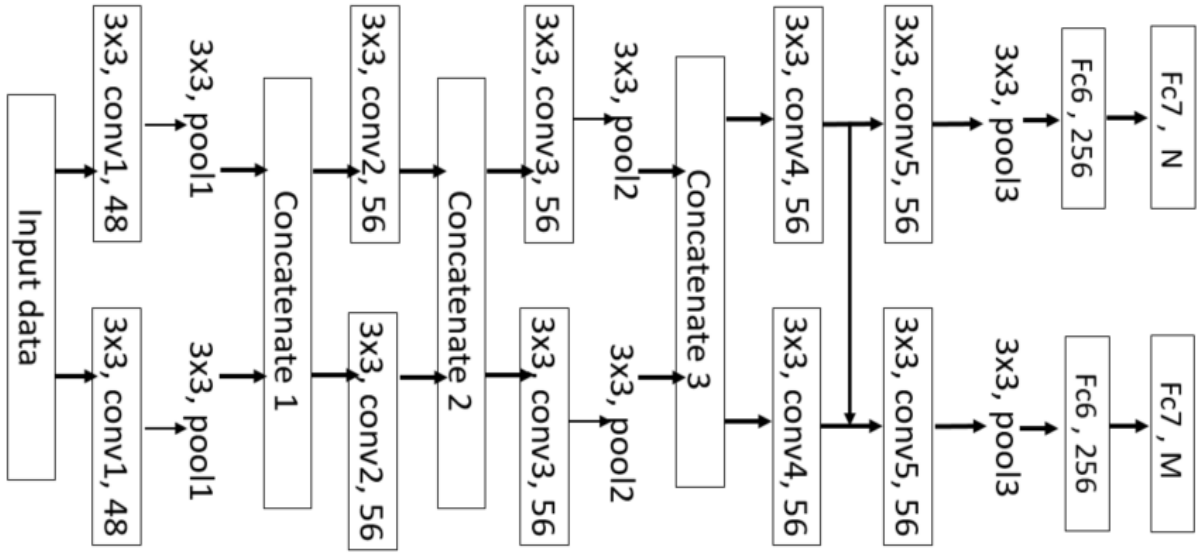
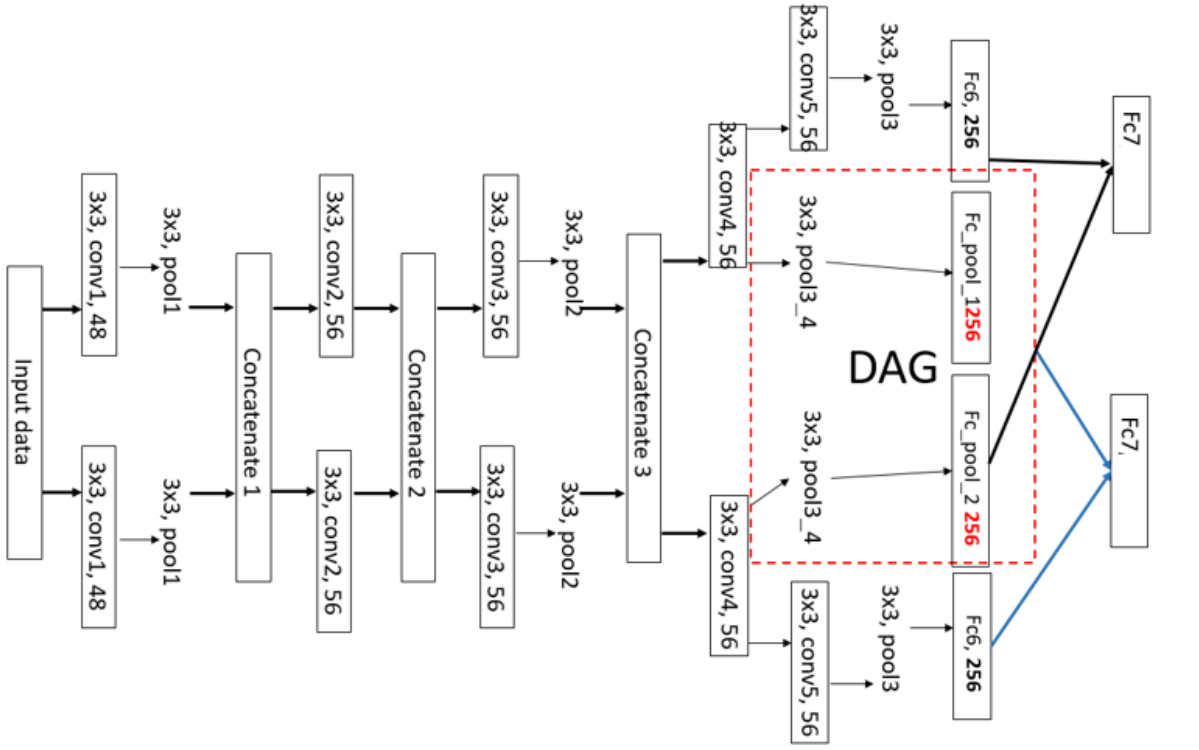


图 4. 3局部非对称的多任务卷积神经网络模型框架图。其中“conv”，“pool”和“Fc”分别代表着卷积层，池化层（Pooling Layer）和全连接层。每一个卷积层之后都会连接一个 ReLU 层用于实现网络的稀疏性。



（b）

图 4. 4融合多层特征的互影响卷积神经网络模型结构图。其中“conv”，“pool”和“Fc”分别代表着卷积层，池化层（Pooling Layer）和全连接层。每一个卷积层之后都会连接一个 ReLU 层用于实现网络的稀疏性。

在模型的训练过程中，我们设定 $w_1, \ldots, w_k$ 为融合多层特征的互影响卷积神经网络

模型的参数。训练数据集是$\left(x^{\mathrm{i}}, y_1^i, y_2^i\right)$，其中$\left(x^{\mathrm{i}}\right)$是第 i 幅输入图像而$\left(y_1^i\right)$则是其是该图像对应的第一个任务的标签，$\left(y_2^i\right)$则是其是该图像对应的第二个任务的标签。我们设定函数f_1 (*)是任务一对应的模型而f_2 (*)是任务二对应的模型。我们的目的是为了去优化以下两个问题：

$$\arg\min_{w_1,\ldots w_k} \frac{1}{n}\sum_{i=1}^{n} L(f_1(x^i, w_1, \ldots, w_k), y_1^i) \tag{式4.1}$$

$$\arg\min_{w_1,\ldots w_k} \frac{1}{n}\sum_{i=1}^{n} L(f_2(x^i, w_1, \ldots, w_k), y_2^i) \tag{式4.2}$$

之后我们将采用随机梯度下降算法来优化上面对应的两个目标函数。该模型最终的整体求导过程可以在分别求导后采用链式法则递归地进行计算最终的导数值的方法来进行有效的计算。对于融合多层特征的互影响卷积神经网络模型与传统卷积神经网络模型在求导过程中不同的地方就在于引出多层特征的卷积层和多层特征融合部分的输出层（Output Layer）。融合多层特征的互影响卷积神经网络模型虽然于传统的卷积神经网络存在不同，但亦然可以采用式 4.3 和式 4.4 并通过链式法则来递归的求解。

$$\frac{\partial \mathrm{z}}{\partial \alpha} = \sum_{i=1}^{2} \frac{\partial z}{\partial \beta_i} \frac{\partial \beta_i}{\alpha} \tag{式4.3}$$

$$\frac{\partial \mathrm{z}}{\partial \alpha} = \frac{\partial \mathrm{z}}{\partial \beta} \frac{\partial \beta}{\partial \alpha} = \frac{\partial z}{\partial \beta} \sum_{j=1}^{2} \frac{\partial \beta}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial \alpha} \tag{式4.4}$$

对于引出多层特征的卷积层处采用式 4.3 来计算，对于多层特征融合后的输出层则采用式 4.4 来计算。式 4.3 中的α为多层特征融合引入部分的输入，而β_{i}则是多层特征引出部分的第 i 个分支，参数 z 对应的是融合多层特征的互影响卷积神经网络特征融合后的 softmax 输出。式 4.4 中采用与式 4.3 中相同的参数设置，而式 4.4 中的参数α_{j}则是多层特征融合部分的一个输入分支。

4.3.2 实验评测

本章中我们所提出的融合多层特征的互影响卷积神经网络模型是在局部非对称的卷积神经网络模型的基础上引入多层特征融合而来的。为了验证我们模型的有效性，我们选择了两种比较的 baseline，并在上章中提到的 Food 数据集和 CompCars 子数据集上验证我们模型的有效性：

1）单神经网络模型（CNN-ST）：分别将图像多属性分类中的每个属性分类任务单独训练一个卷积神经网络模型并单独去验证每一个模型的性能。模型在网络结构上采用了五个卷积层和两个全连接层的七层网络结构，在参数设置上采用与前面章节相同的参数设置。

2）多尺度 DAG 网络模型（Multi-scale DAG-CNN）：该模型是在 CNN-ST 的基础上引入多尺度特征融合的有向无环图的网络结构。我们在上一节介绍了多尺度 DAG 网络模型，并在两个数据集上验证了该模型的性能。我们选择改模型在两个数据集上最好的实验性能作为对比。

3）局部非对称的多任务卷积神经网路模型（PAMT-CNN）：该模型的实验设置和实验对比结果沿用第三章中的实验设置和实验性能的各种描述。

表 4.3 不同方法在 Food 数据集上的分类准确率。

方法	在菜品分类上的准确性	在餐馆分类上的准确性	平均准确性
CNN-ST	70.01%	54.16%	63.085%
DAG-CNN	71.81%	55.22%	63.515%
PAMT-CNN	74.87%	64.75%	69.81%
ME-DAG-CNN	**74.96%**	**65.42%**	**70.19%**

表 4.4 不同方法在 CompCars 子数据集上的分类准确率。

方法	在类型上的准确性	在厂商分类上的准确性	平均准确性
CNN-ST	67.06%	35.67%	51.365%
DAG-CNN	69.04%	38.36%	53.70%
PAMT-CNN	68.50%	46.84%	57.67%
ME-DAG-CNN	**69.78%**	**47.42%**	**58.6%**

不同模型在两个数据集上的分类结果分别展示在表 4.3 和表 4.4 中。从两个表中的实验结果中我们可以得到以下结论：

1）DAG-CNN 在两个数据集的两个分类任务上较 CNN-ST 在分类性能上均有一定程度的提高。这也说明了 DAG 的引入（即多尺度的引入）对我们的细分类分类任务在某种程度上有一定的帮助。这种帮助就是细分类任务需要更多的低层的共通的特征表示用于帮助最终的分类，体现在性能上就是两个数据集的不同分类任务的分类正确性均有所提高。

2）通过 PAMT-CNN 和 DAG-CNN 的对比我们可以发现，在两个数据集子集的分类任务的平均正确性均有较大幅度的提高(在 Food 数据集子集上的分类平均正确性提高了近 6.3 个百分点，在 CompCars 数据集子集上的分类平均正确性上提高了仅 4 个百分点)，而在 CompCars 的类型分类任务上的正确性却没有采用 DAG-CNN 模型的分类正确性高（降低了约 0.5 个百分点）。这一现象的原因我们认为是，在两个分类任务在正确性上相差较多的时候，相互知道学习的共通的低层表示会更多的向分类性能较低的任务靠近，从而大幅度提升分类性能较

差任务的准确性，从而在一定程度上拉低了原分类性能较好的分类任务的分类性能。但不论如何，PAMT-CNN 的分类性能均高于单网络 CNN 所能达到的分类准确性。

3）融合多层特征的互影响卷积神经网络模型（ME-DAG-CNN）在两个数据集子集的分类任务上均表现出了当前最好的分类性能。我们认为其中原因在于具有相互影响的DAG卷积神经网络模型在网络低层采用了任务间的相互学习即参数的共享，这种学习会使得学习到的共通的低层特征表示更具有共通性。这种共同性的特征表示对于细分类任务来说至关重要。同时，在网络的高层上引入了多层特征融合的概念，将网络低层学习到的共通性的特征描述直接应用于最终的特征表示从而使得最终学习到的特征表示不但具有很强的抽象性，还具有一定的尺度性，从而提高了模型最终的分类性能。

除了以上结论外，我们很难不想到这样一个问题，对于深度的卷积神经网络来说，不同的餐馆怎么表示，不同的菜品菜品怎么表示，不同的汽车厂商怎么表示，不同的汽车类型又怎么表示呢？对于不同的深度模型来说，同样的汽车类型，其表示又有哪些不同呢?对于以上问题，我们选择汽车类型举例，通过可视化的方法对其进行解释。

图 4.5 中的每一列对应的是同一种汽车类型在不同深度网络中的可视化表示，分别为 MPV 商务车、SUV 越野车、SEDAN 小轿车、HATCHBACK 掀背车和 MINIBUS 小型公共汽车。图 4.5 的第一行表示的是从五种类型的汽车中分别选择的示例图像，第二行是单网络模型对应的五种汽车类型的可视化结果，第三行是 DAG 网络模型对应的五种汽车类型的可视化结果，第四行是局部非对称的多任务卷积神经网络模型对应的五种汽车类型可视化结果，第五行融合多层特征的互影响卷积神经网络模型对应的五种汽车类型的可视化结果，所有可视化结果对应的是一种汽车类型的可视化结果而不是某一幅图像的可视化结果。从第二行单一神经网络的可视化结果中我们能隐约看到一丝汽车类型的影子。从视觉方面来看，不同类型之间的可分性相对较弱。从第三行是 DAG 网络的可视化结果。DAG 网络在特征表示层进行了不同尺度特征的融合，从而在视觉表示上我们很难看出对应的汽车类型的影子。但是在可分性上较单网络模型有所提高。而第四行的局部非对称的多任务卷积神经网络模型和第五行融合多层特征的互影响卷积神经网络模型对应的汽车类型的可视化结果与单网络和 DAG 网络的可视化结果在视觉上相比明显有更多的色彩交互部分，即不同色彩的分布更加均衡，色彩之间的重叠更多了。我们认为这种色彩分布的均衡是由不同语义的相互嵌入而引起的。同样我们也可以认为，语义的嵌入会使得图像对应的特征表示更加的抽象，更加的接近人对图像的理解。因此，模型在分类性能上较单网络和 DAG 网络模型有所提高。

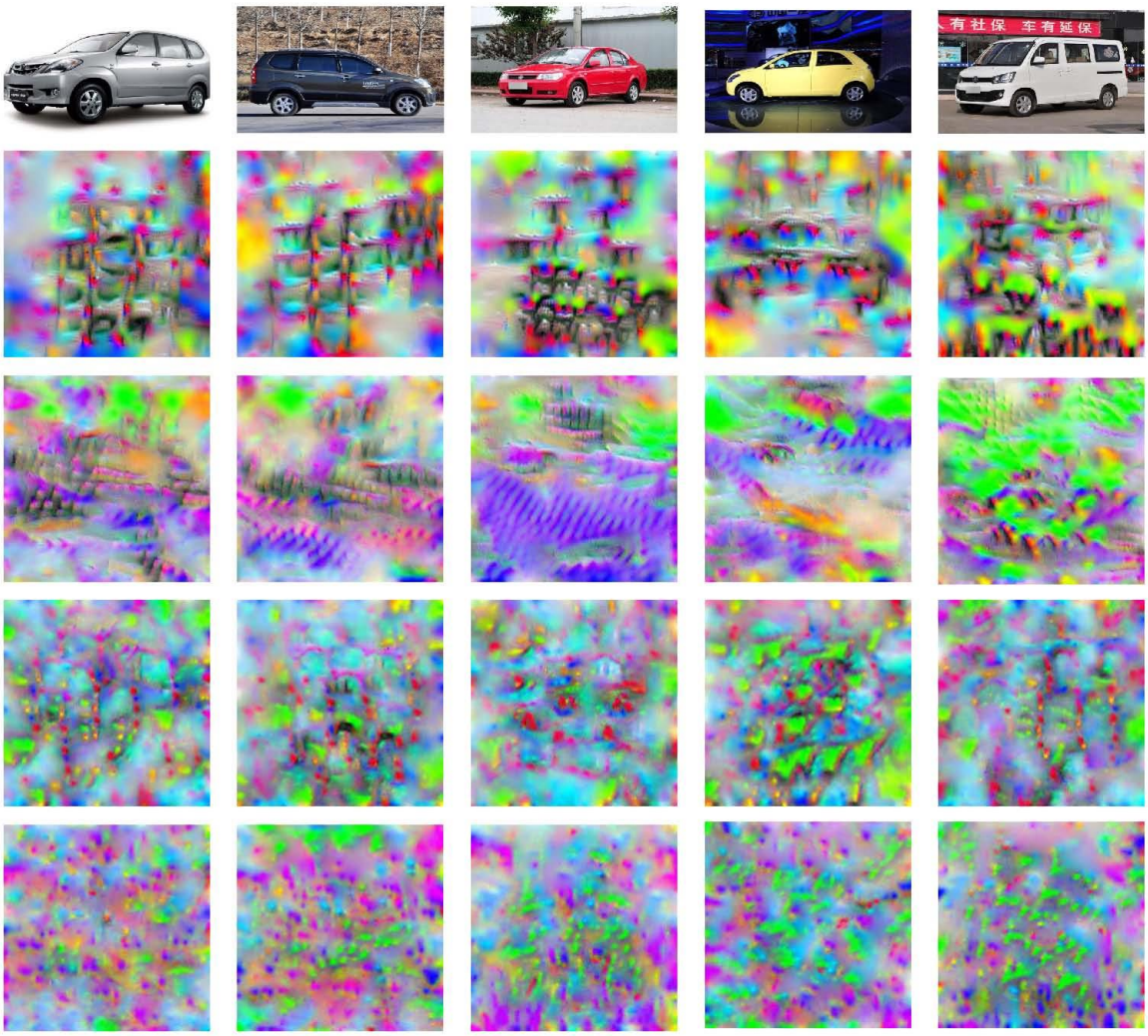


图 4.5不同网络模型的汽车类型可视化举例。

4.4 小结

在本章中，我们介绍了融合多层特征的互影响神经网络模型，并通过实验验证了在我们当前任务中融合多层特征的互影响神经网络模型对分类性能的影响。同时，我们在融合多层特征的卷积神经网络模型和局部非对称的多任务卷积神经网络模型的基础上引入了融合多层特征的互影响卷积神经网络模型。我们通过实验验证了新引入的融合多层特征的互影响卷积神经网络模型在多分类任务上的有效性。

融合多层特征的互影响卷积神经网络模型是将多层特征融合的问题引入到卷积神经网络模型中来。融合多层特征的互影响的引入实现了不同网络层特征表示的融合，从而能够更好的应用于图像的多属性分类任务提高模型的分类性能。我们验证了在数据集图像数目较少、卷积神经网络层数较少（即模型较简单）的情况下，低层卷积的输出对最终的分类性能是有害的。因此我们不应该将低层卷积的输出特征直接应用于最终的特征表示。同样，也并不是越高卷积层的输出对分类正确性的提升效果越明显，我们需要通

过实验来找到适合当前任务最好的网络结构。

之后我们在用于图像分类工作的融合多层特征的神经网络模型（DAG-CNN）和局部非对称的多任务卷积神经网络模型(PAMT-CNN)的基础上，我们提出了一种新的用于图像多出行分类任务的卷积神经网络架构——融合多层特征的互影响卷积神经网络模型。该模型沿用了上述两种模型的优点，通过网路参数的共享来学习对于当前多个分类任务更共通的特征表示。共通的特征表示能够更好的服务于高层抽象特征的提取任务。在最终的抽象特征表示学习的时候引用了网络多层特征的融合。文章[33]解释了融合网络多层特征的网络模型提高图像分类性能的具体原因。通过 DAG-CNN 模型在当前两个数据集子集的分类任务上的实验，我们选择了第四层卷积层输出的尺度特征连接高层的抽象特征形成我们最终的用于分类的特征表示。实验结果显示，该模型结构提取的特征表示在两个数据集的多属性分类任务上均表现出了更好的性能，从而验证了我们模型的有效性。

第五章 结束语

多媒体技术是当前发展最快、最活跃的研究技术，也是计算机科学的重要研究领域之一。人类所获取的外界信息中有80%是来自视觉的，而且通过视觉获取到的信息是最丰富也是最复杂的。我们人能够很好的看清楚并理解视觉所捕获到的信息，但是如何让计算机看懂并理解图像信息却是一件非常困难的工作。图像分类是让计算机理解世界的基础，也是多媒体技术研究的一个重要方向。而图像分类中的多属性图像分类则可认为是多媒体技术中一个基本而富有挑战性的研究领域。多属性图像分类工作有助于机器从多个层面来更详细、更具体的理解图像，从而为计算机理解世界奠定更坚实的基础。

本文的工作是研究基于深度学习的多属性图像分类方法。与已有的大多数工作不同是，我们更细的分析了不同图像属性标签间的相互关联性以及不同网络层特征交互对于图像分类问题性能的影响。之后我们利用不同图像属性标签之间的互相嵌入与不同网络层特征的有机结合来提高模型在不同分类任务上的准确性。

为验证之前假设的内容，图像高层抽象属性标签之间的语义嵌入能够提升模型的分类性能，我们提出了局部非对称的多任务卷积神经网络模型。不同属性分类任务模型之间的参数共享实现了在提取图像特征过程中的语义嵌入，使得提取的特征中不仅包含图像的视觉信息，还包含了不同属性对应的语义信息。同时，我们在考虑语义嵌入的过程中考虑了语义嵌入对图像特征表示学习的影响，仅在不具有类别倾向的重要区分性的卷积层采用语义的相互嵌入。而具有类别倾向的重要区分性的卷积层，我们选择单项的语义嵌入。我们通过模型在不同数据集上的实验，验证了局部非对称的多任务卷积神经网络模型的分类有效性。

多层特征的融合能够在一定程度上提升深度卷积神经网络模型的分类性能。基于这样的结论，我们在局部非对称的多任务卷积神经网络模型的基础上引入多特征融合机制，提出了融合多层特征的互影响卷积神经网络模型。我们逐一验证了融合不同网络层特征的神经网络模型在不同数据集上的分类性能。实验结果证明了，高层次特征的融合能够在一定程度上提高模型的分类性能，而低网络层特征的融合则会对模型的分类性能造成一定的损失。我们认为造成这一现象的原因是因为高层特征具有一定的类别特性，而低层特征更多关注的是图像共通的特征表示。在单模型上多层特征融合实验为我们融合多特征的互影响卷积神经网络模型中融合特征的选择奠定了基础，也揭示了我们从网络高层卷积层进行多层特征融合的原因。而多尺度特征在不同属性任务之间的交互也是为了在一定程度上实现不同语义的相互嵌入。之后我们通过在不同数据集上的实验，验证了融合多层特征的互影响卷积神经网络模型在多属性分类任务中分类性能的有效性。

我们在模型设计中考虑了不同属性标签语义嵌入对图像分类性能的影响，也考虑了

融合多层特征在提升模型分类性能上的有效性。虽然如此，但我们的模型仍然存在一定的缺陷和不足：1）我们只是在小规模的数据集上验证了我们模型的有效性，而我们的模型能否在大规模数据集上取得较好的分类性能还有待验证，2）我们的模型要求数据集中的所有图像数据同时具有多个属性标签而不能有标签的缺失，这也使得模型在鲁棒性和通用性方面存在一定的不足。

在文中我们并没有通过很好的可视化结果去分析语义嵌入和融合多层特征对模型分类性能的影响，未来我们希望通过可视化的结果来分析图像的不同属性语义的相互嵌入对深度卷积神经网络模型特征表示学习的影响，从而实现更合理的语义的相互嵌入。目前我们也仅在小规模数据集上验证了我们模型的有效性，而未来我们希望能够在大规模的多属性分类数据集上验证我们模型的有效性，同时提高我们模型涉及的鲁棒性和通用性。最后，我们希望能够提出更通用的用于图像多属性分类的模型框架。

参考文献

[1] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 248-255.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[3] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification[J]. Computer Vision–ECCV 2010, 2010: 143-156.

[4] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-8.

[5] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.

[6] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.

[8] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 3360-3367.

[9] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification[J]. Computer Vision–ECCV 2010, 2010: 143-156.

[10] Zhou X, Yu K, Zhang T, et al. Image classification using super-vector coding of local image descriptors[J]. Computer Vision–ECCV 2010, 2010: 141-154.

[11] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European Conference on Computer Vision. Springer International Publishing, 2014: 346-361.

[12] Zeiler M D, Fergus R. Stochastic pooling for regularization of deep convolutional neural networks[J]. arXiv preprint arXiv:1301.3557, 2013.

[13] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[14] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[15] Srivastava R K, Greff K, Schmidhuber J. Highway networks[J]. arXiv preprint arXiv:1505.00387, 2015.

[16] 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.

[17] Kontschieder P, Fiterau M, Criminisi A, et al. Deep neural decision forests[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1467-1475.

[18] Xia Y, Feng J, Zhang B. Vehicle Logo Recognition and attributes prediction by multi-task learning with CNN[C]//Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on. IEEE, 2016: 668-672.

[19] Abdulnabi A H, Wang G, Lu J, et al. Multi-task CNN model for attribute prediction[J]. IEEE Transactions on Multimedia, 2015, 17(11): 1949-1959.

[20] Zhu J, Liao S, Yi D, et al. Multi-label cnn based pedestrian attribute learning for soft biometrics[C]//Biometrics (ICB), 2015 International Conference on. IEEE, 2015: 535-540.

[21] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net[J]. arXiv preprint arXiv:1412.6806, 2014.

[22] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in Neural Information Processing Systems. 2015: 2017-2025.

[23] Zhai S, Cheng Y, Zhang Z M. Doubly convolutional neural networks[C]//Advances In Neural Information Processing Systems. 2016: 1082-1090.

[24] Huang G, Liu Z, Weinberger K Q, et al. Densely connected convolutional networks[J]. arXiv preprint arXiv:1608.06993, 2016.

[25] Burt P, Adelson E. The Laplacian pyramid as a compact image code[J]. IEEE

Transactions on communications, 1983, 31(4): 532-540.

[26] Lindeberg T. Scale-space theory in computer vision[M]. Springer Science & Business Media, 2013.

[27] Mallat S. A Wavelet Tour of Signal Processing, (Wavelet Analysis & Its Applications)[J]. 1999.

[28] Brown I J. A Wavelet Tour of Signal Processing: The Sparse Way[J]. Investigacion Operacional. 2008, 29(3): 277-278.

[29] Yang S, Ramanan D. Multi-scale recognition with DAG-CNNs[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1215-1223.

[30] Wang J, Wei Z, Zhang T, et al. Deeply-fused nets[J]. arXiv preprint arXiv:1605.07716, 2016.

[31] Liu S, Cui P, Zhu W, et al. Learning socially embedded visual representation from scratch[C]//Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 109-118.

[32] Zhu F, Li H, Ouyang W, et al. Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification[J]. arXiv preprint arXiv:1702.05891, 2017.

[33] Chai S, Raghavan A, Zhang D, et al. Low Precision Neural Networks using Subband Decomposition[J]. arXiv preprint arXiv:1703.08595, 2017.

[34] Larsson G, Maire M, Shakhnarovich G. Fractalnet: Ultra-deep neural networks without residuals[J]. arXiv preprint arXiv:1605.07648, 2016.

[35] Misra I, Shrivastava A, Gupta A, et al. Cross-stitch networks for multi-task learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3994-4003.

[36] Zhou Z H, Feng J. Deep forest: Towards an alternative to deep neural networks[J]. arXiv preprint arXiv:1702.08835, 2017.

[37] David A, Jean P. Computer vision: a modern approach[J]. Prentice Hall, 2002: 654-659.

[38] Wang X. Deep Learning in Object Recognition, Detection, and Segmentation[J]. Foundations and Trends® in Signal Processing, 2016, 8(4): 217-382.

[39] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer International

Publishing, 2014: 818-833.

[40] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[41] Xu R, Herranz L, Jiang S, et al. Geolocalized modeling for dish recognition[J]. IEEE transactions on multimedia, 2015, 17(8): 1187-1199.

[42] Yang L, Luo P, Change Loy C, et al. A large-scale car dataset for fine-grained categorization and verification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3973-3981.

[43] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends® in Machine Learning, 2009, 2(1): 1-127.

[44] Sivic J, Zisserman A. Video google: A text retrieval approach to object matching in videos[C]//iccv. 2003, 2(1470): 1470-1477.

[45] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]//Computer vision and pattern recognition, 2006 IEEE computer society conference on. IEEE, 2006, 2: 2169-2178.

[46] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.

[47] Gong Y, Wang L, Guo R, et al. Multi-scale orderless pooling of deep convolutional activation features[C]//European conference on computer vision. Springer International Publishing, 2014: 392-407.

[48] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[49] Vedaldi A, Lenc K. Matconvnet: Convolutional neural networks for matlab[C]//Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 689-692.

[50] Baldi P, Pollastri G. The principled design of large-scale recursive neural network architectures--dag-rnns and the protein structure prediction problem[J]. Journal of Machine Learning Research, 2003, 4(Sep): 575-602.

[51] Graves A, Schmidhuber J. Offline handwriting recognition with multidimensional recurrent neural networks[C]//Advances in neural information processing systems.

2009: 545-552.

[52] Raiko T, Valpola H, LeCun Y. Deep Learning Made Easier by Linear Transformations in Perceptrons[C]//AISTATS. 2012, 22: 924-932.

[53] Sermanet P, Kavukcuoglu K, Chintala S, et al. Pedestrian detection with unsupervised multi-stage feature learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3626-3633.

[54] Wang S, Joo J, Wang Y, et al. Weakly supervised learning for attribute localization in outdoor scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3111-3118.

[55] Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 1778-1785.

[56] Kumar N, Berg A C, Belhumeur P N, et al. Attribute and simile classifiers for face verification[C]//Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009: 365-372.

[57] Lampert C H, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 453-465.

致 谢

转眼间，研究生学习生涯已如同手中紧握的沙子，无声无息的流失。回首过往，三年的研究生生活让我感触良多，收获良多。借此，向一路上指导我，关心我和帮助我的恩师、亲友表示诚挚的感谢。

首先在这里感谢我的导师蒋树强研究员。在过去三年的学习时间里，感谢蒋老师对我研究学习的指导，对我工作的帮助和生活上的建议。蒋老师严谨的治学态度，精益求精的工作作风和诲人不倦的高尚师德都深深的感染和激励着我。蒋老师的严以待己、宽以待人的崇高风尚和平易近人的人格魅力让我终身受益。此外，从毕业论文的选题到答辩，蒋老师的耐心监督和指导了我论文各部分的完成，所以我想首先感谢蒋老师，感谢蒋老师对我的栽培，感谢蒋老师对我的耐心与理解。

感谢实验室的闵巍庆博士，我硕士期间的很多研究工作也都与闵巍庆博士的指导是分不开的，特别是实验的一些完成细节以及英文论文的修改上，闵巍庆博士帮助了我很多。在每一次组会后，闵巍庆博士都会指导我如何更详细的去分析问题并帮助我提出切实可行的解决思路。所以在这里要感谢闵巍庆博士。

感谢实验室的师兄、师弟们，主要有宋新航师兄、徐瑞邯师兄、吕雄师兄、黎向阳师兄、朱耀辉师兄、周佳齐师兄、李雪同学、乔雷先同学、朱永清师弟、孙健师弟，梅舒欢师弟等。实验室的很多工作是在我们大家共同的努力下完成的。感谢大家在我每一次遇到问题时提供的帮助与建议，感谢大家三年的陪伴。希望大家在今后的生活、工作和学习上都能一帆风顺，同时也衷心希望我们实验室的发展能越来越好。

此外，我要感谢我的父母和朋友，谢谢你们和我一同分担我在生活、学习中的喜怒哀乐。

作者简介

姓名：王华阳　　性别：男　　出生年月：1990.10.21　　籍贯：山东威海

2014.9 – 2017.7	中科院计算所	计算机应用技术专业	硕士生
2009.9 – 2013.7	山东科技大学	计算机科学与技术专业	本科生

【攻读硕士学位期间发表的论文】

[1]**Wang H,** Min W, Li X, et al. Where and What to Eat: Simultaneous Restaurant and Dish Recognition from Food Image[M], Advances in Multimedia Information Processing - PCM 2016. Springer International Publishing, 2016. (已录用)

[2]Min W, Jiang S, Sang J, **Wang H** et al. Being a Super Cook: Joint Food Attributes and Multi-Modal Content Modeling for Recipe Retrieval and Exploration[J]. IEEE Transactions on Multimedia, 2016, PP(99):1-1. (已录用)

[3]Jiang S, Min W, Li X, **Wang H** et al. Dual Track Multimodal Automatic Learning through Human-Robot Interaction. International Joint Conference on Artificial Intelligence, IJCAI 2017. (已录用)

[4]**Wang H,** Min W, Jiang S. Interaction-based Multimodal Knowledge Graph Construction and Search. International Conference on Intelligent Robots and Systems, IROS 2017. (已投稿)