**基于知识蒸馏的卷积神经网络加速算法的设计与实现**

(三号,黑体，加粗)

**摘 要**（小三，黑体，加粗）

Hihi

（摘要内容部分小四号宋体）（1）一般500 字左右；（2）论文题目为三号黑体字，加粗，可以分成1 或2 行居中打印；（3）论文题目下空一行居中打印“摘要”二字(小三号黑体，加粗)，两字之间空一格（注：“一格”的标准为一个汉字，下同）；（4）“摘要”二字下空一行，打印摘要内容(小四号宋体)，段落按照“首行缩进”2 个字符格式设置；（5）摘要内容后下空一行打印“关键词”三字（小四号黑体，加粗，无间隔），间隔一格，其后为关键词（小四号宋体）。关键词数量为3～5 个，关键词之间间隔一格，不加标点符号。

重点说明本项课题的目的和意义、研究方法、研究（开发）成果、结论，注意突出本人所做工作的部分。摘要中不要出现第一人称，用“本文分析了…”等用语。另外，论文摘要部分至少有一主要段落是对论文的综述，不能写成项目的综述。

**关键词**（小四号黑体，加粗，无间隔） 知识蒸馏 卷积神经网络 关键词3 关键词4（小四号宋体，间隔一格，不加标点符号）

**ENGLISH TITLE(Times New Roman,三号,加粗)**

**ABSTRACT**

**（Times New Roman,小三号,加粗）**

(小四号)The Abstract is a brief description of a thesis or dissertation without notes or comments. It represents concisely the research purpose, content, method, result and conclusion of the thesis or dissertation with emphasis on its innovative findings and perspectives. The Abstract Part consists of both the Chinese abstract and the English abstract. The English abstract should be consistent with the Chinese one in content. The keywords of a thesis or dissertation should be listed below the main body of the abstract, separated by two spaces. The number of the keywords is typically 3 to 5.

以英文摘要为例：论文中的英文一律采用“Times New Roman”字体。论文英文题目首字母采用大写字母（三号，加粗），字可分成1～3 行居中打印。

（1）题目下空一行居中打印“ABSTRACT”（小三号，加粗），再下空一行打印英文摘要内容（小四号），英文摘要与中文摘要相对应；

（2）每段开头留四个字符空格；

（3）摘要内容后下空一行打印“KEY WORDS”（小四号，加粗），间隔两个字符空格，其后关键词（小四号）小写，关键词之间间隔两个字符空格，关键词之间不加标点符号。

**KEY WORDS**（小四号，加粗）key words1 key words2 key words3 key words4（小四号，小写）

**目 录**

**[第一章 引言 1](#_Toc514855714)**

[1.1 课题背景 1](#_Toc514855715)

[1.1.1 卷积神经网络发展 1](#_Toc514855716)

[1.1.2 深度学习模型压缩加速技术 1](#_Toc514855717)

[1.2 课题任务 1](#_Toc514855718)

[1.2.1 课题内容 1](#_Toc514855719)

[1.2.2 本人承担任务 1](#_Toc514855720)

[1.3 论文结构 1](#_Toc514855721)

**[第二章 相关技术介绍(可选) 2](#_Toc514855722)**

[2.1 知识蒸馏 2](#_Toc514855723)

[2.1.1 XXXXX 2](#_Toc514855724)

[2.2 XXXXXXX 2](#_Toc514855725)

[2.3 XXXXXX 2](#_Toc514855726)

[2.4 XXXXXX 2](#_Toc514855727)

[2.n 本章小结 2](#_Toc514855728)

**[第三章 系统的需求分析 3](#_Toc514855729)**

[3.1 系统用户角色分析（可选） 3](#_Toc514855730)

[3.1.1 XXXXX 3](#_Toc514855731)

[3.2系统功能需求分析 3](#_Toc514855732)

[3.3 系统非功能需求分析 3](#_Toc514855733)

[3.4 XXXXXX 3](#_Toc514855734)

[3.n 本章小结 3](#_Toc514855735)

**[第四章 系统的总体设计 4](#_Toc514855736)**

[4.1 系统网络结构设计 4](#_Toc514855737)

[4.1.1 XXXXX 4](#_Toc514855738)

[4.2 系统软件层次架构设计 4](#_Toc514855739)

[4.3 系统功能模块设计 4](#_Toc514855740)

[4.4 系统的数据库设计 5](#_Toc514855741)

[4.5 系统的界面设计 5](#_Toc514855742)

[4.n 本章小结 5](#_Toc514855743)

**[第五章 系统主要功能模块的详细设计与实现 6](#_Toc514855744)**

[5.1 XXXXX模块的详细设计与实现 6](#_Toc514855745)

[5.1.1 XXXXX 6](#_Toc514855746)

[5.2 XXX模块的详细设计与实现 6](#_Toc514855747)

[5.3 XX模块的详细设计与实现 6](#_Toc514855748)

[5.4 XXXXX模块的详细设计与实现 6](#_Toc514855749)

[5.n 本章小结 6](#_Toc514855750)

**[第6章 系统测试 7](#_Toc514855751)**

[6.1 系统测试方法介绍 7](#_Toc514855752)

[6.2 系统的功能测试 7](#_Toc514855753)

[6.3 系统的性能测试 7](#_Toc514855754)

[6.n 本章小结 7](#_Toc514855755)

**[第N章 结束语 8](#_Toc514855756)**

[N.1 论文工作总结 8](#_Toc514855757)

更新后注意删除页码及省略号,保持原有格式!!!!

[N.2 问题和展望 8](#_Toc514855758)

**[参考文献](#_Toc514855759)**

**[致 谢](#_Toc514855760)**

注意：“目录” 两字之间间隔两格，两字居中打印（三号黑体，加粗），下空两行依次为章、节、小节及其开始页码。目录内容的“第 X 章”用小4号黑体，其余使用小4号宋体, 目录的行间距为20磅。

# 引言

（一级标题：黑体三号，加粗，序号和标题间间隔一格）

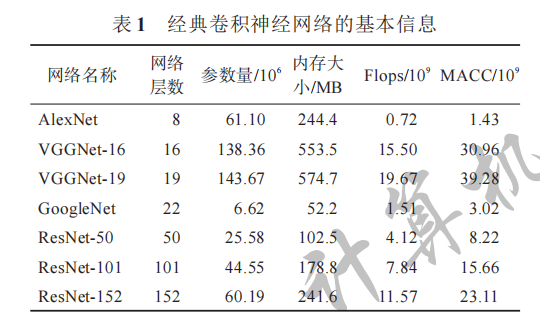
## 1.1 课题背景

二级标题：黑体四号，加粗，序号和标题间间隔一格

### 卷积神经网络发展（未写）

近年来，卷积神经网络（Convolutional neural network, CNNs）在计算机视觉、自然语言处理、语音识别等领域取得了突飞猛进的发展,其强大的特征学习能力引起了国内外专家学者广泛关注。然而，深度卷积神经网络普遍规模庞大、计算复杂，在实时要求高和资源受限环境下的性能很受限制。随着现实生活中移动端设备广泛普及，越来越多深度学习模型向边缘设备部署。而边缘设备的内存容量、计算资源十分有限。如何在不影响深度学习模型性能的情况下进行模型加速压缩迅速成为研究热点.国内外学者相继提出许多方案，探索实现缩小模型、加快推理计算速度的深度学习模型压缩与加速方法。

（表：经典卷积网络参数量表）



三级及以下标题：黑体小四号，加粗，首行缩进2字符，序号和标题间间隔一格

### 深度学习模型压缩与加速

随着深度学习模型不断发展，模型结构、计算复杂度不断上升，计算开销、内存存储需求日益增大。在图像处理领域，He提出Resnet网络缓解深层网络梯度消失问题，广泛用于图片特征提取，其Resnet-50参数量26M，FLOPs高达3.9亿，内存消耗约为102MB。bert模型用于诸多文本分析下游任务的预训练，其bert-base模型参数量达110M。在模型性能提高的另一面，网络参数量不断增大，带来了更多的内存消耗，高浮点型运算量意味着计算时间的增加。在存储计算资源有限设备模型性能受限，如移动手机、智能手环等终端设备。现实生活中移动端设备广泛普及，越来越多深度学习模型向边缘设备部署。如何在不影响深度学习模型性能的情况下进行模型加速压缩迅速成为研究热点.国内外学者相继提出许多方案，探索实现缩小模型，加快推理计算速度的深度学习模型压缩与加速方法。

通过实验表明，许多深度学习网络中一定比例参数是冗余部分。早在20世纪80年代末期，Lecun提出optimal brain damage 算法利用损失函数计算参数的二阶导数，以此衡量网络中参数的重要程度，发现删除近半数参数后深度学习网络依然可与原来的网络性能一样好，证明了深度学习模型压缩的可行性。同时网络中的冗余结构容易引发模型过拟合、模型泛化性下降等问题。由此可见，进行深度学习模型压缩加速相关研究十分必要。常见深度学习模型压缩加速方法包括网络剪枝、参数量化、知识蒸馏、设计轻量型架构等。网络剪枝将训练后模型中冗余结构进行剪裁以此压缩模型大小。参数量化是指较低位宽表示典型的 32 位浮点网络参数,网络参数包括权重、激活值、梯度和误差等等。目前主流方案是将模型参数从float32类型压缩到int8类型，既节省内存又提高计算速度。知识蒸馏利用原始模型学习出来的知识作为先验，将先验知识迁移到小规模的神经网络中训练更小但性能仍较好的小型神经网络。知识蒸馏取得了一定的成效，它适用于几乎所有的网络架构，并可以与其他策略（如网络剪枝和量化）相结合，以进一步改进网络设计。

正文：采用宋体小四号字，段落按照“首行缩进”2 字符格式设置，1.2倍行距。

理工科专业要求15000 汉字以上；经管文法类专业要求20000 汉字；英语、日语专业要求7000词以上。（不含图表、程序和计算数字）。

（3）图：每一图应有简短确切的图名，图中标注可采用中文或英文。图名的英文字体为五号Times New Roman ，中文字体为五号楷体。引用图应在图名的右上角标明文献来源。图中坐标上标注的符号和缩略词必须与正文中一致。

图序号一律采用阿拉伯数字分章依序编排，如：图 3－2 为第三章第二图。如果图中含有几个不同部分，应将分图序号标注在分图的左上角，并在图名下列出各分图图名。

绘图必须工整、清晰、规范。示意图应能清楚反映图示内容；照片应在右下角标明放大比例；实验结果曲线图应制成方框图。

（4）表：表序号一律采用阿拉伯数字分章依序编排，如：表 5－4 为第五章第四表。每张表格应有简短确切的标题，表标题及序号置于表的正上方，英文字体为五号Times New Roman，中文字体为五号楷体。表内必须按规定的符号标注单位。

（5)公式：公式序号一律采用阿拉伯数字分章依序编排，如：式(2-13)、式(4-5)，其标注应于该公式所在行的最右侧；公式书写方式应在文中相应位置另起一行居中横排，对于较长的公式只可在符号处（+、-、\*、/、 ≤、≥等）转行。

## 1.2 课题任务

### 1.2.1 课题内容

### 1.2.2 本人承担任务

## 1.3 论文结构

# 相关技术

## 2.1 知识蒸馏

知识蒸馏是指从预训练好的教师网络提取知识，引导较小参数量的学生网络学习的方法。知识蒸馏将原神经网络中蕴含的知识传递到目标网络。通常，原神经网络规模较大，经过一系列训练得到教师模型。用来指导参数量较小的学生网络学习训练提高学生模型的性能。相同的学习样本经过预训练的教师模型为学生网络提供有监督的中间层、输出层信息指导，帮助学生模型更快更高效的学习样本特征。

随着模型压缩加速的需求不断扩大，许多基于知识蒸馏的深度学习模型压缩工作相继开展并取得一定成效。2015年，hinton首次提出知识蒸馏概念，确立了教师-学生框架，通过计算师生模型输出分布的差异做为损失信息，令学生模型学习教师模型预测的概率分布。于此同时，庞大教师模型内部同样蕴含着指导学生模型训练的知识。Romero等人在Fitnet中引入中间提示层指导学生模型的训练，旨在使学生网络学习教师网络的中间表示层中隐含的知识。Heo等，利用教师模型的决策边界指导学生模型学习。在Zagoruyko等人将注意力机制引入知识蒸馏，该方法让学生网络模仿教师网络中间层的注意力特征图，有效的提升了学生网络的分类性能。除了直接利用教师网络的输出信息或者中间层信息外，其他知识蒸馏算法对传递的知识作了全新的定义。FSP[ ]矩阵提取不同层之间的二阶统计量，学生网络学习教师网络模型中层与层之间的映射关系。Tian等人将对比学习的机制引入到基于关系的知识蒸馏方案中。尽管关于如何提炼和传播知识的蒸馏方法不断提出，但如何将师生模型之间的中间层特征联系起来仍然是一个问题。

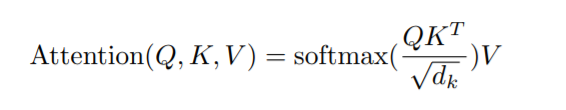
知识蒸馏中师生网络结构不同，在提取教师模型中间层信息、监督学生模型学习时产生中间层匹配问题。之前许多工作致力于在师生模型同一层次上进行知识传递。然而忽略了教师模型层其它中间层对学生网络训练提供指导信息。另一方面，师生网络存在层数不同的情况，如何合理划分师生模型中间层对映关系仍是具有挑战性的问题。为了将教师模型的知识更充分地向学生模型转移，本文对师生网络不同层次间的相关性进行量化计算。

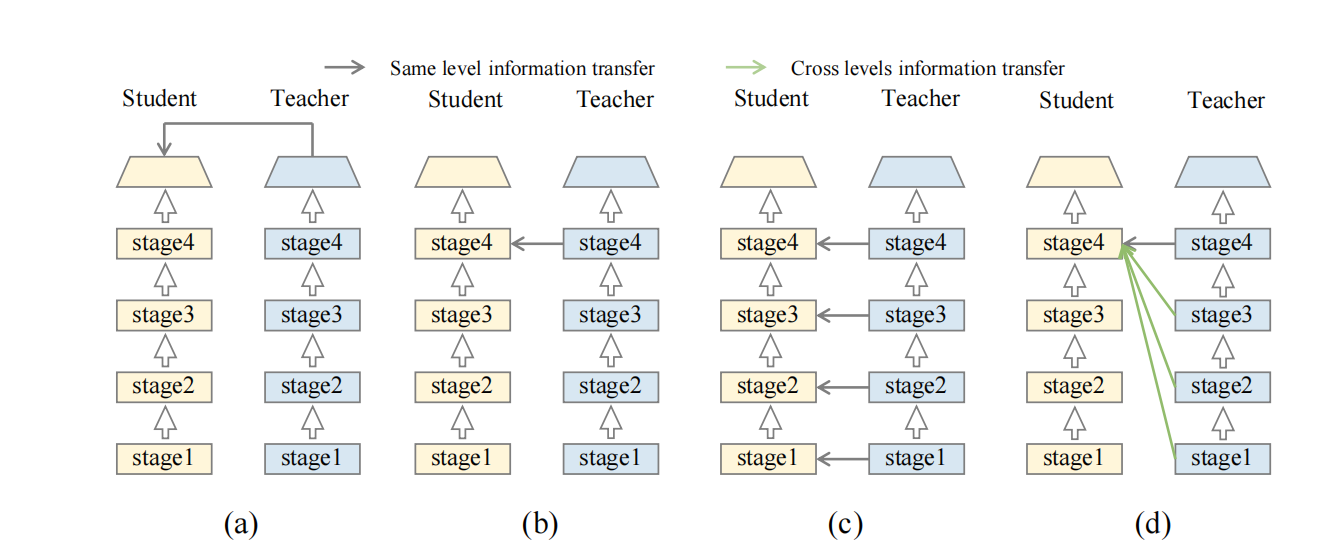
## 2.2 注意力机制

注意力机制首先在人类视觉领域被提出。认知科学认为，当处理信息能力有限时，人类往往将注意力集中于部分更重要的信息，而忽略其它可见信息。注意力机制的主要工作原理是从复杂信息中选择对当前工作目标更有价值的部分，削减次要信息对目标工作的影响因子。注意力机制广泛用于深度学习领域。在计算机视觉任务中，建立了注意机制来进行视觉信息处理，如局部图像特征提取、显著性检测、滑动窗口处理方法等。

与此同时，Ashish Vaswani等人将注意力机制应用于机器翻译等各类自然语言处理任务中，计算每个词语和所在序列所有词语的相关性，有效降低了序列过长对循环神经网络造成的影响，进一步激发引发了学者对注意力机制的关注。SCAN在跨模态检索任务中，利用注意力机制匹配文本和图像。词语信息可由不同权重的图像区域信息加权描述，在此过程中实现了文本图片双模态的特征融合，图像信息通过注意力机制传递到文本信息中。这种知识传递的方式同样适用于旨在通过教师模型引导学生模型学习训练的知识蒸馏。

简单的注意力函数可以表示为以下公式，其中Q,K,V作为输入向量，query表示教师网络提取的中间层特征向量，key表示学生网络提取的中间层向量，可以通过点积、拼接、使用感知机等多种方法计算二者相关性系数。在缩小师生特征图距离的过程中实现知识传递。





图引用

引用：

## 2.3 卷积神经网络

近年来，伴随着深度学习不断发展，在计算机视觉、自然语言处理等领域中涌现着各种神经网络模型，尤其是卷积神经网络在图像处理领域取得很大成功。通常，卷积神经网络包括卷积层、归一化层、池化层、全连接层等。其中提取样本特征的模块主要为卷积层。人们受到生物视觉处理过程的启发，提出卷积概念，在输入图像上多个卷积核滑动进行卷积运算得到特征图。设置不同的卷积核大小可以改变局部感受野大小，改变滑动步长可以控制输出特征图的尺寸，同时设置卷积核个数改变特征图通道数。这样卷积层可以产生一组平行的特征图（feature map），实现了特征提取。卷积层的产生增强了模型局部感知性。以下是卷积操作示意图。加图

早在20世纪末期，运用核心卷积操作的卷积神经网络诞生，经典卷积网络LeNet进行手写数字识别，为深度卷积神经网络的发展奠定了基础。2012年，卷积神经网络AlexNet模型添加激活函数、归一化层，缓解梯度消失问题，并在ImageNet比赛中获得分类任务冠军。AlexNet网络中部分卷积核过大（如11\*11），Simonyan 和 Zisserman提出了 VGGNet ，改用3 × 3 小卷积核的卷积层,增加了网络深度,提升了网络性能。卷积神经网络深度不断增加，梯度消失问题越来越明显，He 等人提出残差卷积神经网络，通过将输入和卷积结果叠加使不同层的特征向下传递，缓解梯度消失问题。为进一步优化模型，学者相继提出各种Resnet变体网络，例如宽残差网络（WRN）、金字塔残差网络（PyramidalNet）等。随着模型精度的提高，卷积神经网络层数越来越多、越来越宽，考虑到内存占用、计算消耗等问题，开始了轻量化卷积神经网络的探索。Iandola 等 人(2016) 提出了 SqueezeNet,其核心模块为挤压卷积层和扩展卷积层两部分。使用1\*1小卷积核减少参数量。

与此同时，许多规模较大的卷积神经网络模型在图像分类等领域已突破人类识别水平，新的问题是如何将这种参数量较大的模型压缩投入智能手机、无人驾驶汽车等资源紧张型设备使用。本文将结合之前的模型压缩加速方法，针对图像分类任务提供模型压缩方案。

# 

# 第N章 结束语

## N.1 论文工作总结

本文分析了XXX，论述了XXX，讨论了XXX，给出了XXX，提出了XXX，设计了XXX，编码实现了XXX，组织实施了XXX，等

本文取得的主要成果有：（注意：是“成果”，不是你做的工作）

1. XXX
2. XXX

本人在该项目中的主要工作有：

1. XXX
2. XXX

## N.2 问题和展望

# 参考文献

在学位论文中引用参考文献时，应在引出处的右上方用方括号标注阿拉伯数字编排的序号；参考文献的排列按照文中引用出现的顺序列出，参考文献全部列于正文之后，每篇论文至少查阅10篇以上文献资料。

“参考文献”四个字 黑体三号，加粗，居中。

参考文献内容中文字体采用宋体五号、英文字体采用Times New Roman（五号）。其排列参考格式为：

（1）专著中的文献

[序号] 作者.专著名称.版本(第１版不加标注).出版者.出版年:参考页码.

（2）期刊中的文献

[序号] 作者.文献名称.期刊名称.卷号（期号）.年,月:页码范围.

（3） 论文集

[序号] 作者.论文题目.见(英文用 In).主编.论文集名.出版地.出版年:页码范围.

（4）学位论文

[序号] 作者.题目.［学位论文］.(英文用［Dissertation］) 保存地点. 保存单位:年份.

（5） 专利

[序号] 专利申请者.题目.国别.专利文献种类.专利号.批准日期.

（6）技术标准

[序号] 起草责任者.标准代号.标准顺序号－发布年.标准名称.出版地.出版者.出版年度.

注：文献中的作者数量低于三位时全部列出；超过三位时只列前三位，其后加“等”字即可；作者姓名之间用逗号分开；中外人名一律采用姓在前，名在后的著录法。

# 致 谢

“致谢”两个字之间间隔两格，三号黑体，加粗，居中。致谢主要是对导师和给予指导或协助完成毕业设计（论文）工作的组织和个人表示感谢。文字要简洁、实事求是，切忌浮夸。

# 附 录

附录一般作为学位论文主体的补充，主要包括：正文内过于冗长的公式推导；供读者阅读方便所需要的辅助性的数学工具或重复性数据图表；由于过分冗长而不宜放置在正文中的计算机程序清单；本专业内具有参考价值的资料；论文使用的缩写说明等。附录编于正文后，其页码与正文连续编排。

“附录”两个字之间间隔两格，三号黑体，加粗，居中。

论文的附录依序编排为附录1，附录2……。附录中的图表公式另编排序号，与正文分开。

# 外 文 资 料

外文资料应主要选自近期外文期刊,内容应与毕业设计(论文)内容紧密相关，外文译文不少于5000汉字，外文资料的原件可以为电子版打印稿也可以是原件影印件。