



《计算机视觉相关技术及发展前景》

|  |  |
| --- | --- |
| **专 业** | **计算机科学与技术** |
| **姓 名** | **张晨曦** |
| **日 期** | **2022/8/28** |

**摘 要**

本文由介绍计算机视觉的研究内容切入，随后从图像分类、对象检测、目标跟踪、语义分割、实例分割五个方面详细介绍了计算机视觉领域的主要技术，随后循序渐进地介绍了计算机视觉各阶段所需要学习的知识，并简单介绍了轻量化模块设计、轻量化模型搜索、视觉大模型压缩与知识迁移等前沿研究方向，最终介绍了本专业的本科生就业去向。随后总结全文得出对专业的评价和结论。

**目 录**

[**第1章 专业研究内容介绍** - 1 -](#_Toc112622665)

[1.1 计算机视觉解决的问题 - 1 -](#_Toc112622666)

[1.2 为什么要研究计算机视觉 - 1 -](#_Toc112622667)

[**第2章 专业主要技术介绍** - 3 -](#_Toc112622668)

[2.1 图像分类有关技术介绍 - 3 -](#_Toc112622669)

[2.2 对象检测有关技术介绍 - 3 -](#_Toc112622670)

[2.3 目标跟踪有关技术介绍 - 5 -](#_Toc112622671)

[2.4 语义分割有关技术介绍 - 7 -](#_Toc112622672)

[2.5 实例分割有关技术介绍 - 8 -](#_Toc112622673)

[**第3章 专业学习建议与指导** - 11 -](#_Toc112622674)

[**第4章 目前专业的前沿研究方向** - 12 -](#_Toc112622675)

[4.1 轻量化模块设计 - 12 -](#_Toc112622676)

[4.2 轻量化模型搜索 - 12 -](#_Toc112622677)

[4.3 视觉大模型压缩与知识迁移 - 13 -](#_Toc112622678)

[**第5章 专业本科生就业去向** - 14 -](#_Toc112622679)

[**结 论** - 15 -](#_Toc112622680)

[**参考文献** - 16 -](#_Toc112622681)

**第1章 专业研究内容介绍**

## 1.1 计算机视觉解决的问题

计算机视觉是一门研究如何使[机器](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8/2275865)“看”的科学，更进一步的说，就是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等[机器视觉](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E8%A7%86%E8%A7%89/7414484)，并进一步做图形处理，使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科，计算机视觉研究相关的理论和技术，试图建立能够从图像或者多维数据中获取‘信息’的[人工智能](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/9180" \t "_blank)系统。这里所指的信息指Shannon定义的，可以用来帮助做一个“决定”的信息。因为感知可以看作是从感官信号中提 取信息，所以计算机视觉也可以看作是研究如何使人工系统从图像或多维数据中“感知”的科学。计算机视觉是人工智能的一个重要分支，它要解决的问题就是：**看懂图像里的内容**。

比如：

* 图片里的宠物是猫还是狗？
* 图片里的人是老张还是老王？
* 这张照片里，桌子上放了哪些物品？

****

图1：CV的研究内容

## 1.2 为什么要研究计算机视觉

人的大脑皮层， 有差不多 70% 都是在处理视觉信息。视觉是人类获取信息最主要的渠道，没有之一。在网络世界，照片和视频（图像的集合）也正在发生爆炸式的增长。

下图是网络上新增数据的占比趋势图。灰色是结构化数据，蓝色是非结构化数据（大部分都是图像和视频）。可以很明显的发现，图片和视频正在以指数级的速度在增长。

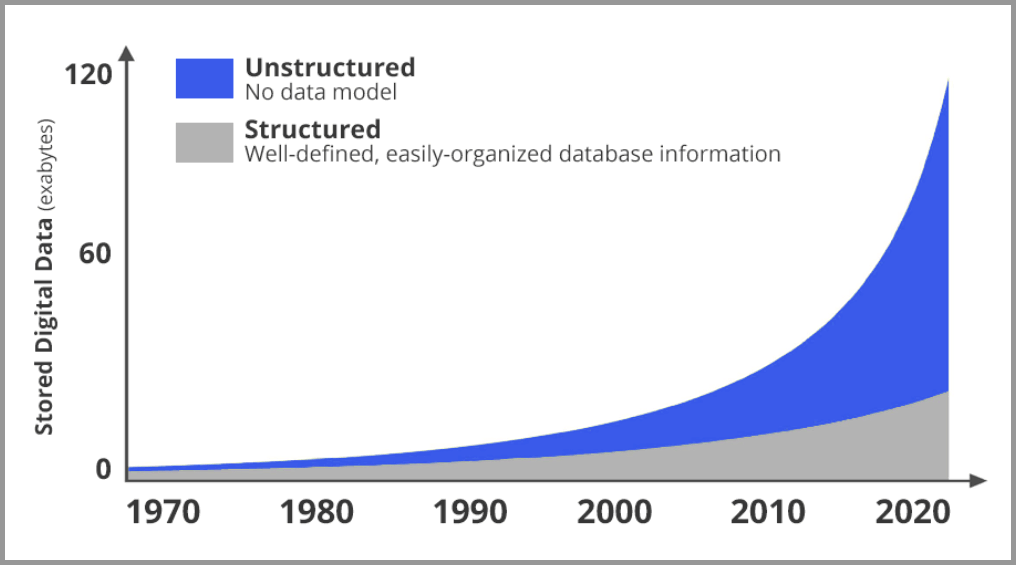


图2：网络上新增数据的占比趋势图

而在计算机视觉出现之前，图像对于计算机来说是黑盒的状态。一张图片对于机器只是一个文件。机器并不知道图片里的内容到底是什么，只知道这张图片是什么尺寸，多少MB，什么格式的。

如果计算机、人工智能想要在现实世界发挥重要作用，就必须**看懂图片**。这就是计算机视觉的主要研究内容。

**第2章 专业主要技术介绍**

## 2.1 图像分类有关技术介绍

给定一组各自被标记为单一类别的图像，我们对一组新的测试图像的类别进行预测，并测量预测的准确性结果，这就是图像分类问题。图像分类问题需要面临以下几个挑战：

视点变化，尺度变化，类内变化，图像变形，图像遮挡，照明条件和背景杂斑。

目前较为流行的图像分类架构是卷积神经网络（CNN）——将图像送入网络，然后网络对图像数据进行分类。卷积神经网络从输入“扫描仪”开始，该输入“扫描仪”也不会一次性解析所有的训练数据。比如输入一个大小为 100\*100 的图像，你也不需要一个有 10,000 个节点的网络层。相反，你只需要创建一个大小为 10 \*10 的扫描输入层，扫描图像的前 10\*10 个像素。然后，扫描仪向右移动一个像素，再扫描下一个 10 \*10 的像素，这就是滑动窗口。

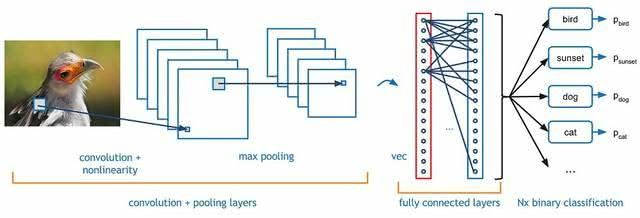
****

图1：滑动窗口

输入数据被送入卷积层，而不是普通层。每个节点只需要处理离自己最近的邻近节点，卷积层也随着扫描的深入而趋于收缩。除了卷积层之外，通常还会有池化层。池化是过滤细节的一种方法，常见的池化技术是最大池化，它用大小为 2\*2 的矩阵传递拥有最多特定属性的像素。

现在，大部分图像分类技术都是在 ImageNet 数据集上训练的， ImageNet 数据集中包含了约 120 万张高分辨率训练图像。测试图像没有初始注释（即没有分割或标签），并且算法必须产生标签来指定图像中存在哪些对象。

## 2.2 对象检测有关技术介绍

识别图像中的对象这一任务，通常会涉及到为各个对象输出边界框和标签。这不同于分类/定位任务——对很多对象进行分类和定位，而不仅仅是对个主体对象进行分类和定位。在对象检测中，你只有 2 个对象分类类别，即对象边界框和非对象边界框。例如，在汽车检测中，你必须使用边界框检测所给定图像中的所有汽车。

如果使用图像分类和定位图像这样的滑动窗口技术，我们则需要将卷积神经网络应用于图像上的很多不同物体上。由于卷积神经网络会将图像中的每个物体识别为对象或背景，因此我们需要在大量的位置和规模上使用卷积神经网络，但是这需要很大的计算量！

为了解决这一问题，神经网络研究人员建议使用区域（region）这一概念，这样我们就会找到可能包含对象的“斑点”图像区域，这样运行速度就会大大提高。第一种模型是基于区域的卷积神经网络（ R-CNN ），其算法原理如下：

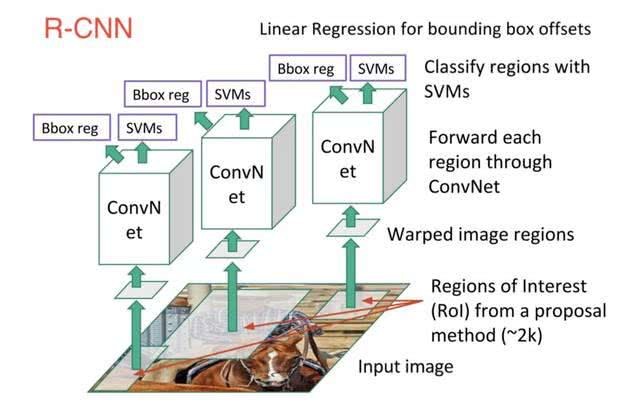


图2：R-CNN原理

实质上，我们将对象检测转换为了一个图像分类问题。但是也存在着问题：训练速度慢，需要大量的磁盘空间，推理速度也很慢。

R-CNN 的第一个升级版本是 Fast R-CNN，通过使用了 2 次增强，大大提了检测速度：

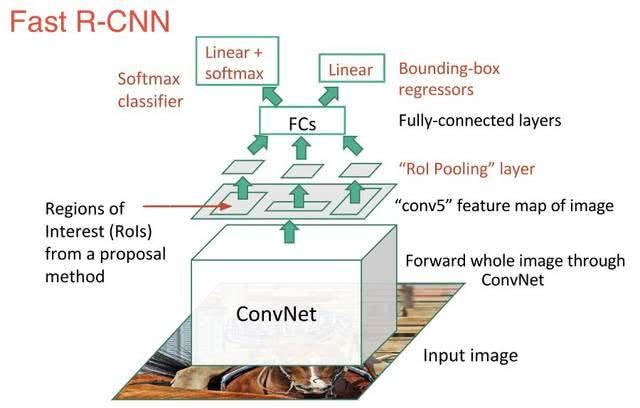


图3：Fast R-CNN原理

Fast R-CNN 的运行速度要比 R-CNN 快的多，因为在一幅图像上它只能训练一个 CNN 。 但是，择性搜索算法生成区域提议仍然要花费大量时间。

Faster R-CNN 是基于深度学习对象检测的一个典型案例。

该算法用一个快速神经网络代替了运算速度很慢的选择性搜索算法：通过插入区域提议网络（RPN），来预测来自特征的建议。RPN 决定查看“哪里”，这样可以减少整个推理过程的计算量。

RPN 快速且高效地扫描每一个位置，来评估在给定的区域内是否需要作进一步处理，其实现方式如下：通过输出 k 个边界框建议，每个边界框建议都有 2 个值——代表每个位置包含目标对象和不包含目标对象的概率。

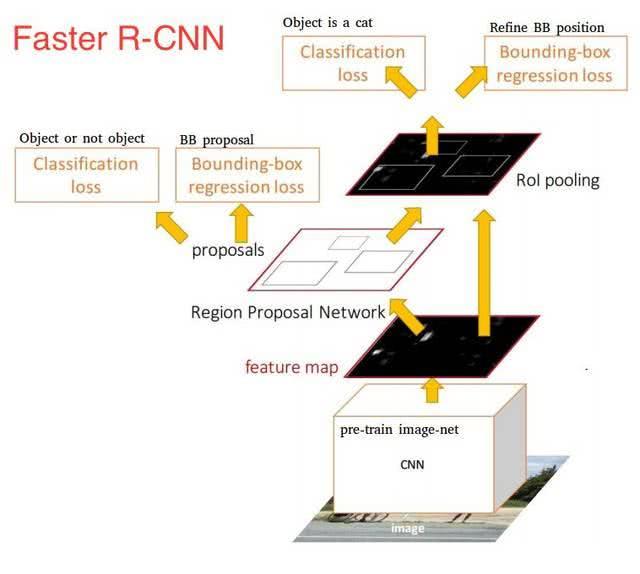


图4：Faster R-CNN原理

一旦我们有了区域建议，就直接将它们送入 Fast R-CNN 。并且，我们还添加了一个池化层、一些全连接层、一个 softmax 分类层以及一个边界框回归器。

总之，Faster R-CNN 的速度和准确度更高。值得注意的是，虽然以后的模型在提高检测速度方面做了很多工作，但很少有模型能够大幅度的超越 Faster R-CNN 。换句话说， Faster R-CNN 可能不是最简单或最快速的目标检测方法，但仍然是性能最好的方法之一。

近年来，主要的目标检测算法已经转向更快、更高效的检测系统。这种趋势在 You Only Look Once（YOLO），Single Shot MultiBox Detector（SSD）和基于区域的全卷积网络（ R-FCN ）算法中尤为明显，这三种算法转向在整个图像上共享计算。因此，这三种算法和上述的3种造价较高的R-CNN 技术有所不同。

## 2.3 目标跟踪有关技术介绍

目标跟踪，是指在特定场景跟踪某一个或多个特定感兴趣对象的过程。传统的应用就是视频和真实世界的交互，在检测到初始对象之后进行观察。现在，目标跟踪在无人驾驶领域也很重要，例如 Uber 和特斯拉等公司的无人驾驶。

根据观察模型，目标跟踪算法可分成 2 类：生成算法和判别算法。

为了通过检测实现跟踪，我们检测所有帧的候选对象，并使用深度学习从候选对象中识别想要的对象。有两种可以使用的基本网络模型：堆叠自动编码器（ SAE ）和卷积神经网络（ CNN ）。

目前，最流行的使用 SAE 进行目标跟踪的网络是 Deep Learning Tracker（DLT），它使用了离线预训练和在线微调。其过程如下：

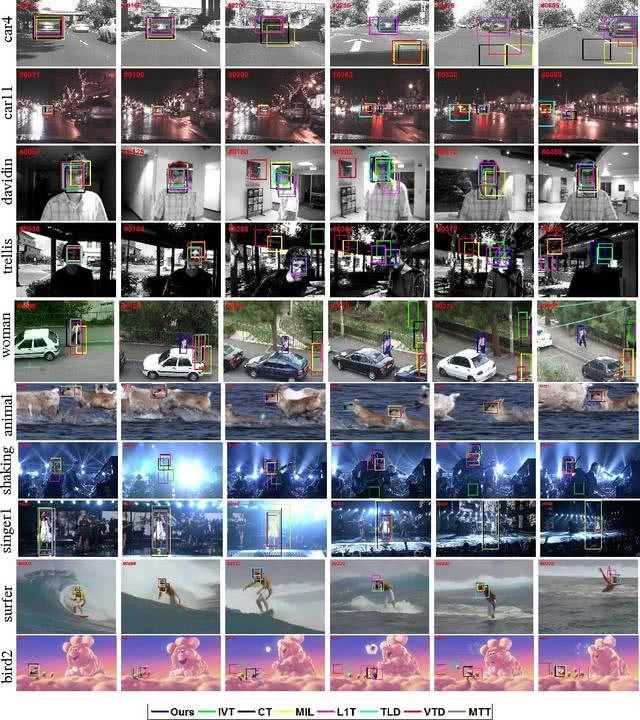


图1：DLT过程

鉴于 CNN 在图像分类和目标检测方面的优势，它已成为计算机视觉和视觉跟踪的主流深度模型。一般来说，大规模的卷积神经网络既可以作为分类器和跟踪器来训练。具有代表性的基于卷积神经网络的跟踪算法有全卷积网络跟踪器（FCNT）和多域卷积神经网络（MD Net）。

FCNT 充分分析并利用了 VGG 模型中的特征映射，这是一种预先训练好的 ImageNet 数据集，并有如下效果：

因此， FCNT 设计了特征选择网络，在 VGG 网络的卷积 4-3 和卷积 5-3 层上选择最相关的特征映射。 然后为避免噪音的过拟合，FCNT还为这两个层的选择特征映射单独设计了两个额外的通道（即SNet和GNet）： GNet捕获对象的类别信息；SNet将该对象从具有相似外观的背景中区分出来。

这两个网络的运作流程如下：都使用第一帧中给定的边界框进行初始化，以获取对象的映射。而对于新的帧，对其进行剪切并传输最后一帧中的感兴趣区域，该感兴趣区域是以目标对象为中心。最后，通过 SNet 和 GNet ，分类器得到两个预测热映射，而跟踪器根据是否存在干扰信息，来决定使用哪张热映射生成的跟踪结果。 FCNT 的图如下所示。

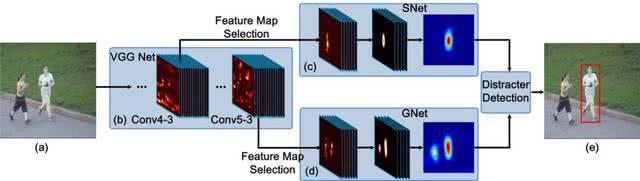


图2：FCNT示意图

与 FCNT 的思路不同， MD Net 使用视频的所有序列来跟踪对象的移动。上述网络使用不相关的图像数据来减少跟踪数据的训练需求，并且这种想法与跟踪有一些偏差。该视频中的一个类的对象可以是另一个视频中的背景，因此， MD Net 提出了“多域”这一概念，它能够在每个域中独立的区分对象和背景，而一个域表示一组包含相同类型对象的视频。

如下图所示，MD Net 可分为两个部分，即 K 个特定目标分支层和共享层：每个分支包含一个具有 softmax 损失的二进制分类层，用于区分每个域中的对象和背景；共享层与所有域共享，以保证通用表示。

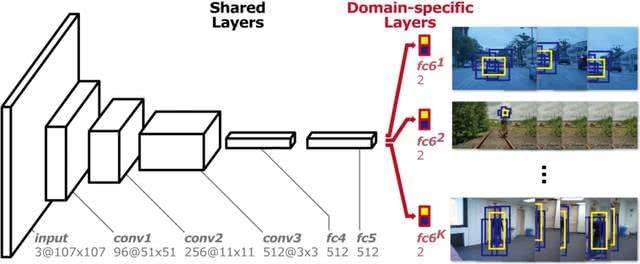


图3：MD Net示意图

## 2.4 语义分割有关技术介绍

计算机视觉的核心是分割，它将整个图像分成一个个像素组，然后对其进行标记和分类。特别地，语义分割试图在语义上理解图像中每个像素的角色（比如，识别它是汽车、摩托车还是其他的类别）。如上图所示，除了识别人、道路、汽车、树木等之外，我们还必须确定每个物体的边界。因此，与分类不同，我们需要用模型对密集的像素进行预测。

与其他计算机视觉任务一样，卷积神经网络在分割任务上取得了巨大成功。最流行的原始方法之一是通过滑动窗口进行块分类，利用每个像素周围的图像块，对每个像素分别进行分类。但是其计算效率非常低，因为我们不能在重叠块之间重用共享特征。

解决方案就是加州大学伯克利分校提出的全卷积网络（ FCN ），它提出了端到端的卷积神经网络体系结构，在没有任何全连接层的情况下进行密集预测。

这种方法允许针对任何尺寸的图像生成分割映射，并且比块分类算法快得多，几乎后续所有的语义分割算法都采用了这种范式。

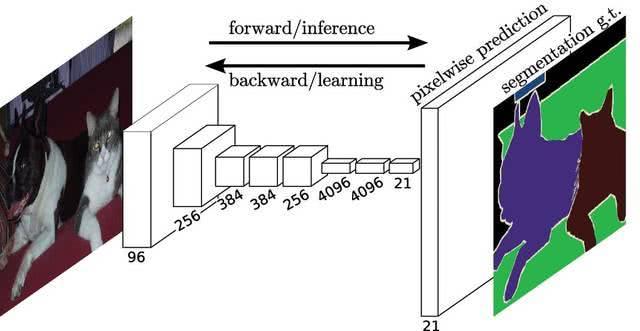


图1：FCN示意图

但是，这也仍然存在一个问题：在原始图像分辨率上进行卷积运算非常昂贵。为了解决这个问题， FCN 在网络内部使用了下采样和上采样：下采样层被称为条纹卷积（ striped convolution ）；而上采样层被称为反卷积（ transposed convolution ）。

尽管采用了上采样和下采样层，但由于池化期间的信息丢失， FCN 会生成比较粗糙的分割映射。SegNet 是一种比 FCN （使用最大池化和编码解码框架）更高效的内存架构。在 SegNet 解码技术中，从更高分辨率的特征映射中引入了 shortcut/skip connections ，以改善上采样和下采样后的粗糙分割映射。

目前的语义分割研究都依赖于完全卷积网络，如空洞卷积 ( Dilated Convolutions ），DeepLab 和 RefineNet 。

## 2.5 实例分割有关技术介绍

除了语义分割之外，实例分割将不同类型的实例进行分类，比如用 5 种不同颜色来标记 5 辆汽车。分类任务通常来说就是识别出包含单个对象的图像是什么，但在分割实例时，我们需要执行更复杂的任务。我们会看到多个重叠物体和不同背景的复杂景象，我们不仅需要将这些不同的对象进行分类，而且还要确定对象的边界、差异和彼此之间的关系。

举个例子，Facebook AI 使用了 Mask R-CNN 架构对实例分割问题进行了探索。



图1：Mask R-CNN 架构示意图

就像 Fast R-CNN 和 Faster R-CNN 一样， Mask R-CNN 的底层是鉴于 Faster R-CNN 在物体检测方面效果很好，我们是否可以将其扩展到像素级分割？

Mask R-CNN 通过向 Faster R-CNN 添加一个分支来进行像素级分割，该分支输出一个二进制掩码，该掩码表示给定像素是否为目标对象的一部分：该分支是基于卷积神经网络特征映射的全卷积网络。将给定的卷积神经网络特征映射作为输入，输出为一个矩阵，其中像素属于该对象的所有位置用 1 表示，其他位置则用 0 表示，这就是二进制掩码。

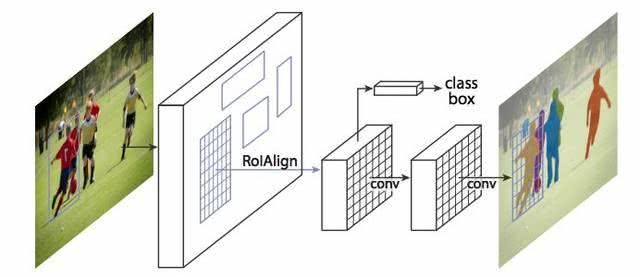


图2：二进制掩码

另外，当在原始 Faster R-CNN 架构上运行且没有做任何修改时，感兴趣池化区域（ RoIPool ） 选择的特征映射区域或原始图像的区域稍微错开。由于图像分割具有像素级特性，这与边界框不同，自然会导致结果不准确。 Mas R-CNN 通过调整 RoIPool 来解决这个问题，使用感兴趣区域对齐（ Roialign ）方法使其变的更精确。本质上， RoIlign 使用双线性插值来避免舍入误差，这会导致检测和分割不准确。

一旦生成这些掩码， Mask R-CNN将RoIAlign与来自Faster R-CNN的分类和边界框相结合，以便进行精确的分割：



图三：Mask R-CNN的分割示意图

**第3章 专业学习建议与指导**

计算机视觉专业的主干学科仍为计算机科学与技术，所以学习的过程中不可避免的会涉及编程等实战环节。

计算机视觉入门要从编程语言（如python、octave等）开始，在了解其基础用法后不要过多地纠结于语言的偏难怪细节，而应将细节问题的发现和解决留到实现具体项目的过程中。

在了解python语言的基础语法后可以学习数据分析所涉及到的库，如Numpy、Pandas等，并进行一些数据分析的简单实战。

在了解数据分析有关库的使用后可以学习机器学习、深度学习等理论基础，让自己不仅知道使用方法更清楚使用原理，并可以通过Kaggle等平台不断通过项目实践来加深自己的掌握。

随后可以去学习一些传统图像处理算法库，如OpenCV，并通过项目进行OpenCV的实践。

在完善了理论基础也了解了传统算法后可以学习一些深度学习框架（如pytorch、TensorFlow）的使用，了解一些图像算法模型，随后通过项目实践深度学习框架。

在完成上述步骤后可以阅读一些引用量高的CV论文进行进阶。

**第4章 目前专业的前沿研究方向**

在很多实际场景中，由于存储、计算资源的限制，大模型难以直接部署或者无法满足实时需求。因此，轻量级的视觉预训练模型研究变得越来越重要，且具有很强的实际应用价值。尽管目前有一些工作在探讨轻量级模型，但是这些方法大多是针对特定任务、特定结构设计的，在设计和训练过程中没有考虑到模型的通用性，存在跨数据域、跨任务的泛化局限性。

由于提高轻量化预训练模型通用性的核心在于如何在资源受限（参数量，时延等）的情况下强化模型的学习能力，使其能够更好地在大规模数据中学习通用特征，因此，以下三个方向是比较前沿的研究方向。

## 4.1 轻量化模块设计

轻量、低延时的模块是组成轻量级模型的重要部分。在卷积神经网络中，具有代表性的轻量级模块有MobileNet的反向残差模块（Inverted Residual Block）以及 ShuffleNet 的通道随机交叉单元（Shuffle Unit）。在视觉 Transformer 结构中，由于图像块之间注意力的计算没有很好地考虑相对位置编码信息，因此研究人员们设计了即插即用的轻量级二维图像相对位置编码方法 iRPE [1]，它不需要修改任何的训练超参数，就能提高模型的性能。此外，针对视觉 Transformer 参数冗余的问题，研究员们设计了权重多路复用（Weight Multiplexing）模块 [2]。如图1所示，该方法通过多层权重复用减少模型参数的冗余性，并且引入不共享的线性变换，提高参数的多样性。

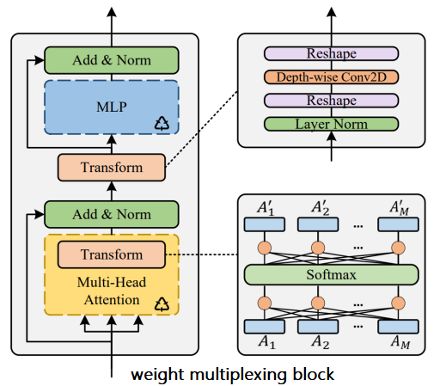


图1：Transformer 中的权重多路复用模块

## 4.2 轻量化模型搜索

网络结构搜索（Neural Architecture Search）可以从模型设计空间中自动找到更加轻量、性能更加优异的模型结构 [3]。在卷积神经网络中，代表性工作有 NASNet 和 EfficientNet 等。在视觉 Transformer 结构搜索中，针对视觉模型中的通道宽度、网络深度以及 head 数量等多个维度，研究员们先后提出了 AutoFormer [4] 和 S3 [5]，实现了视觉模型的动态可伸缩训练与结构搜索。在同样模型精度的情况下，搜索得到的新模型具有更小的参数量和计算量。值得注意的是，在 S3 中，研究人员们利用 E-T Error [5]以及权重共享超网来指导、改进搜索空间，在得到更高效的模型结构的同时也分析了搜索空间的演进过程，如图2所示。与此同时，模型结构搜索的过程为轻量化模型的设计提供了有效的设计经验和参考。

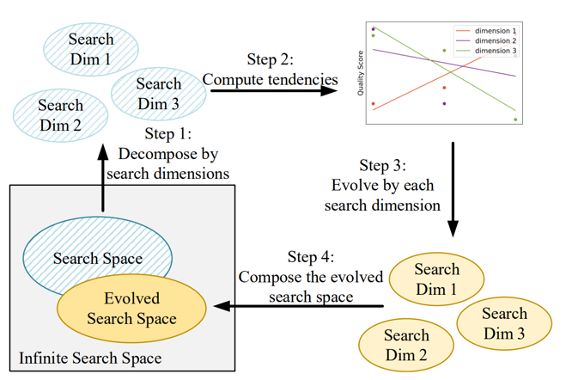


图2：轻量级模型搜索空间进化过程

## 4.3 视觉大模型压缩与知识迁移

轻量级预训练模型的另一难题在于，由于模型容量有限，难以直接学习大规模数据中包含的丰富信息和知识。为了解决这一问题，研究人员们提出了快速预训练蒸馏方案，将大模型的知识迁移到轻量化的小模型中 [6]。如图3所示，和传统的单阶段知识蒸馏不同，快速预训练蒸馏分为两个阶段：1）压缩并保存大模型训练过程中使用的数据增广信息和预测信息；2）加载并恢复大模型的预测信息和数据增广后，利用大模型作为教师，通过预训练蒸馏指导轻量化学生模型的学习和训练。不同于剪枝和量化，该方法在权重共享的基础上使用了上文中提到的权重复用[2]，通过引入轻量级权重变换和蒸馏，成功压缩视觉预训练大模型，得到了通用性更强的轻量级模型。在不牺牲性能的情况下，该方法可以将原有大模型压缩数十倍。

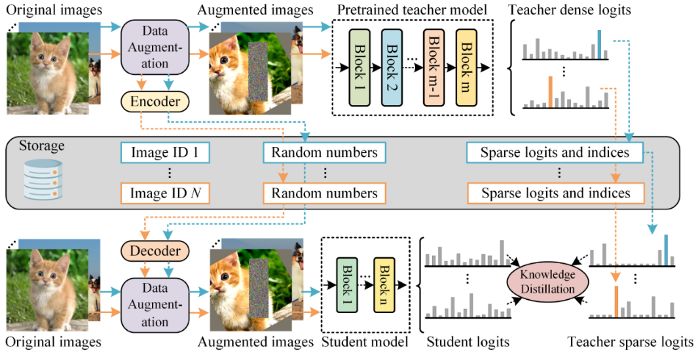


图3：快速预训练知识蒸馏

**第5章 专业本科生就业去向**

计算机视觉专业本科生的就业去向大致可分为：

1) 搜索方向：百度、谷歌、微软、yahoo等(包括智能搜索、语音搜索、图片搜索、视频搜索等都是未来的热门方向)。

2) 医学图像处理：医疗设备、医疗器械很多都会涉及到图像处理和成像，大型的公司有西门子、GE、飞利浦等。

3) 计算机视觉和模式识别方向：前面说过的指纹识别、人脸识别、虹膜识别等。还有一个大的方向是车牌识别。目前鉴于视频监控是一个热点问题，做跟踪和识别也不错。

4) 还有一些图像处理方面的人才需求的公司如威盛、松下、索尼、三星等。

当然，由于计算机视觉行业竞争较为激烈且岗位要求的知识储备量一般较大，所以大部分本科同学选择了升学来提升自己。

**结 论**

计算机视觉专业以让计算机看懂图像内容为目的，借助于程序设计，深度学习等知识和方式来实现，且需要大量的知识储备才能充分理解和使用。计算机视觉专业目前仍有轻量化模块设计等前沿研究方向等待我们去探索，还有十足的发展潜力。就业去向一般为高新技术相关公司。

综上所述，计算机视觉专业十分值得学习与深入！

**参考文献**

[1] Rethinking and Improving Relative Position Encoding for Vision Transformer, ICCV 2021.

[2] MiniViT: Compressing Vision Transfor[mers](https://www.zhihu.com/search?q=mers&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2597507628%7D) with Weight Multiplexing, CVPR 2022.

[3] Cyclic Differentiable Architecture Search, TPAMI 2022.

[4] AutoFormer: Searching Transformers for Visual Recognition, ICCV 2021.

[5] Searching the Search Space of Vision Transformer, NeurIPS 2021.

[6] TinyViT: Fast Pretraining Distillation for Small Vision Transformers, ECCV 2022.