人工智能导论课程总结

本课程对现在的人工智能做了较为全面的概览，包含了统计机器学习和深度神经网络的基础内容，同时对现在较为流行的机器学习库sklearn和tensorflow的使用方法进行了讲解。本课程基本覆盖了现在机器学习的基础内容，讲解比较全面，下面将谈谈本课程带来的收获。

本课程的前半段主要讲授了统计机器学习的内容，首先，了解了机器学习的基本概念。比如学习任务一般分为回归和分类两大类，根据训练数据是否带有标记信息，可分为监督学习和无监督学习，如果要进一步划分还有之后提出的半监督学习等。除此之外，也了解了简单的机器学习模型。最为基本的的是线性回归和逻辑回归，线性回归用于回归问题的，可以直接得到数值解，逻辑回归用于解决分类问题。通时还涉及了一些凸优化的基本知识，最小化损失函数的方法最为常见的是梯度下降（gradient decent），随机梯度下降（stochastic gradient decent）和批量梯度下降（batch stochastic gradient decent），还有一些其他的优化方法如拟牛顿法，adam等。如果要优化的目标函数是凸函数，则以上方法能够取得近似的全局最优解，但是在实际操作中目标函数往往不是凸函数，则一般的优化方法只能取得局部最优解而无法保证全局最优解，因此模型的效果会和参数的初始化情况有关。除此之外，还学习了一些较为复杂的模型如支持向量机，决策树，多层感知机，集成学习的模型如随机森林，Adaboost等方法。在此过程中也学习了一些正则化方法（常用的是L1范数和L2范数）改善模型的过拟合，提升模型的泛化能力。

随着计算能力的不断提高，大数据技术的发展，从Alexnet在图像识别领域取得突出效果的开始，深度神经网络模型引起了人们的关注。深度神经网络模型是十分强大的模型，但同时也是难以理解其内部结构的“黑盒模型”。通过近几年的发展，深度学习已经被运用于诸多方面，如自然语言处理，语音识别的，计算机视觉，数据分析处理等。深度学习在这些领域都取得了非常卓越的成果，一般超过了原来深度网络流行之前使用的经典方法，出现了深度学习几乎“一统天下”的局面。最开始的神经网络一般使用S型函数（如sigmod，tanh）函数作为激活函数，但网络容易出现梯度消失的问题，现在在深度模型中一般会使用relu或者leaky relu等等预防梯度消失的激活函数。现在重要的基础网络有循环神经网络、卷积神经网络等。循环神经网络主要用于处理序列数据， 每次会将序列中的一个元素和前一时刻的状态作为输入，因此有一定的记忆能力。由于梯度要在整个序列中传递，简单的RNN会出现梯度爆炸和梯度消失等问题。而lstm通过引入了input gate，output gate， forget gate对梯度进行控制，可以极大改善这个问题，之后提出的GRU与lstm功能类似但是参数更少，节省计算资源同时减少overfit的风险。在自然语言处理领域，基于RNN的模型被大量使用，如实体识别，机器翻译，对话系统等。神经机器翻译已经取得了很好的效果，google的翻译系统已经大规模采用了这样的方式。在语音识别领域，传统的模型是比较复杂和繁琐的，但是利用神经网络可以构建端对端的模型，这样的模型也超过了经典模型的表现。在计算机视觉领域，深度模型被使用的更加广泛，最为重要的便是卷积神经网络，在识别领域经典的网络结构有Alexnet，牛津大学的VGGnet，google的googlenet，微软亚洲研究院的resnet等。在CV的分支目标检测领域，现在主要的方法有faster-rcnn，YOLO（You Only Look Once）和SSD（Single Shot MultiBox Detector），能够在较短的时间内在图片中识别物体并给出物体的位置。除了图像分类和物体检测这样的任务之外，计算机视觉有很多其他的子领域比如人脸识别，行人重识别，人体姿势检测等。初次之外也出现了一些将自然语言处理和计算机视觉进行结合的领域如image caption和video caption，根据图像和视频的内容生成相应的自然语言描述这样的任务，接下来计算机视觉和自然语言处理的相结合的领域和应用将会越来越广泛。

除此之外，强化学习也越来越受到人们的关注，而且现在的趋势是将强化学习与深度学习相结合，一般称为深度强化学习。AlphaGo，AlphaGo Zero都可以看做使用了深度强化学习的方式，在强化学习的过程中加入了网络对价值进行评估并对行为进行决策。

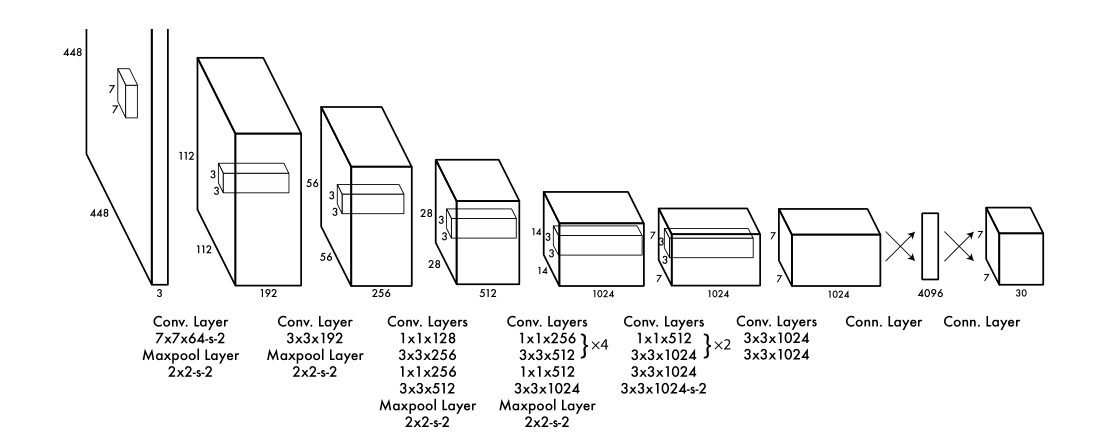
在本学期的课程中，不但学习了一些机器学习的基本理论，同时也动手进行了一些操作实践，通过运行实例代码熟悉了机器学习库sklearn和tensorflow的基本使用方法，并且把这些方法用到了课程作业和自己参与的其他机器学习相关的项目中，通过构建自己的机器学习系统和搭建自己的神经网络进一步加深了对于相关知识的理解。

本学期的大作业是制作一个人工智能小车，在制作小车的过程中小组的各个成员都全程参与，除此之外本人主要负责整个项目的进度和安排，主要负责arduino代码的编写，语音识别和目标检测系统的构建。在我们的作业中，小车主要实现的功能有：小车的随机游走和避障，通过蓝牙连接并控制小车，语音控制小车，利用opencv让小车识别并跟踪预先准备的红色电烙铁，用深度学习的方法进行目标检测追踪特定物体。

arduino板是小车的重要组成部件，组装小车的过程中了解到了许多arduino板的相关知识，并且自己编写了小车的控制代码，掌握了arduino的基本编程，借助超声波发射器，能够得知小车与前方物体的距离，利用此数据通过简单的控制逻辑可以让小车有避障的功能。我们在Arduino板上添加了用于蓝牙控制的装置，使得小车能够在行进中接收蓝牙命令。首先我们通过手机上的蓝牙app连接小车，向小车发送相应的命令，小车根据发送发的命令做出相应的反应，之后我们用电脑蓝牙连接小车，这样可以在电脑上运行复杂的算法并且根据结果跟小车发送相应的命令控制小车的行动。

在第二阶段，尝试了语音控制小车，在此过程中了解了语音处理、语音识别方面的相关知识，并对现在的语音识别系统做了相关的调研，尝试了用现有的语音处理工具（如opensmile，kaldi等）对语音数据提取相关的特征（如mfcc，log Mel特征等）。在传统的非端对端的语音识别系统中一般要先从音频文件中提取特征（常用的是mfcc特征），然后再进一步进行处理，在传统的方法中一般会利用高斯混合模型和隐马尔克夫算法，结合EM算法的思想，经典的语音识别的算法有Baum-Welch算法等。在现在深度学习盛行的情况下，又出现了有些将高斯混合模型与神经网络相结合的方法和端对端的语音识别系统，这些系统往往有更好的语音识别性能。我们将现有的语音识别系统运用在了小车上，语音识别系统运行在PC上，对语音信号进行解析得到相应的文本，根据文本指令对小车发出相应的命令，从而实现语音控制小车。在此阶段，小车还实现了追踪红色的电烙铁的功能。在实现过程中，将手机固定在小车上，通过网络摄像头将小车上的手机拍摄的图像传送到电脑上，利用opencv对得到的视频流进行处理，利用Opencv，可以在得到的帧上进行处理，得到特定的形状和颜色的物体在图像中所处的位置，并根据位置确定物体相对于小车的方位，通过蓝牙命令让小车向物体的方向行驶。这部分工作中熟悉了opencv的基本原理，了解了opencv的使用方法。

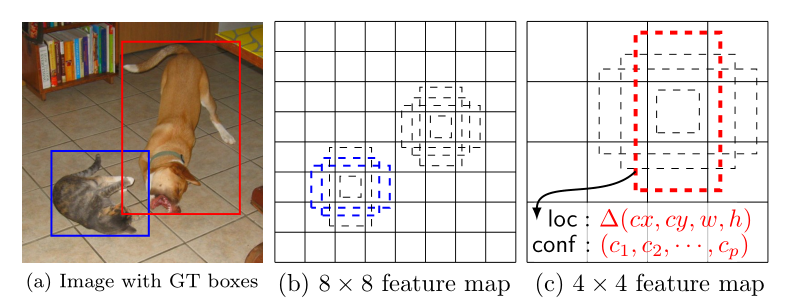
在第三阶段，我们加入了目标检测（object detection）的功能，能够在一个图片中找出某些种类的物体并给出其在图像中所处的位置。现在主要的目标检测的模型有RCNN系列的模型（RCNN，Fast-RCNN，Faster-RCNN，现在主要使用Faster-RCNN），YOLO，SSD。在实现过程中浏览了相关的论文并将模型运用在了小车的系统上， Faster-RCNN ，YOLO和SSD可用于实时的目标检测，计算速度较快。在YOLO的中，整个图片被分成S x S个格子（cell），每个格子（cell）负责中心在该格子的目标的检测，每个格子可以检测B个box，每一个box可以由5个参数进行描述（x，y，w，h，confidence），其中x，y代表box相对于格子的位置，w，h代表了box相对于整张图片的大小，confidence是预测的IoU值（模型产生的目标窗口和原来标记窗口的交叠率），可以代表着个box的定位的置信度。初次之外，每个格子还要负责输出包含某个物体的可能性，比如若此时有C个物体，则此处应该需要长度为C的向量代表包含每种物体的可能性大小，这样，对于每张图片，网络的输出就是一个维度为S x S x (B x 5 + C)的向量，这个向量中就包含了物体的bounding box所处的位置，物体的种类和这个种类上的置信度。网络的结构如下所示：



（图片摘自原论文）

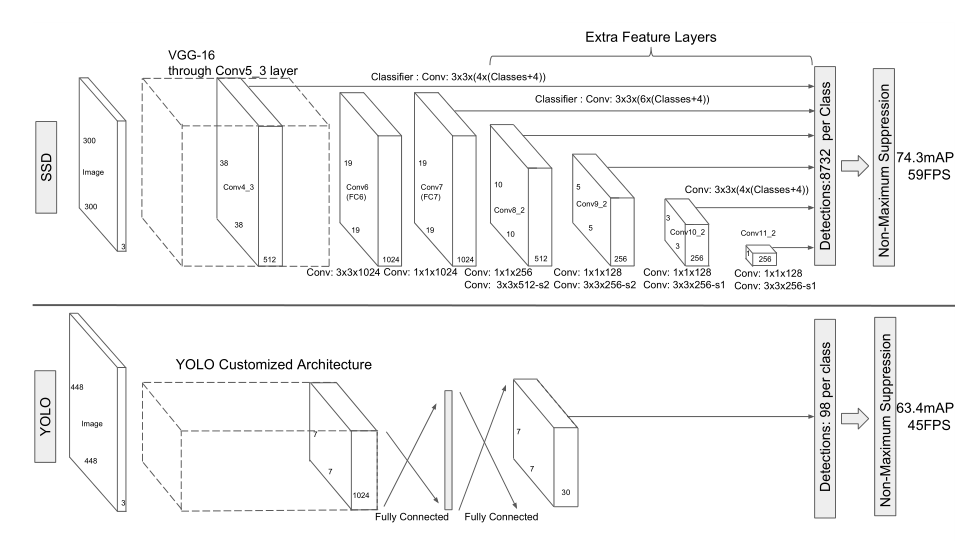
论文中网络共有24层卷积，其中前20层来自于在一个图像分类任务上通过预训练中得到的VGGnet的卷积层，之后加上4层卷积与2层全连接网络之后再一起进行训练进行微调。网络中多次使用了1x1的卷积核减少参数，网络可以同时得到物体的种类位置和bounding box，是一种比较有效率的模型，YOLO在之后还有进一步的改进，在此不再进一步说明。

在小车的具体实现中使用了SSD的目标检测模型。SSD模型与在YOLO在网络结构上有所不同，前几层卷积上两者是类似的，都是通过VGGnet在分类问题上进行预训练进行初始化，但之后YOLO加了若干卷积层和全连接层并将结果在全连接层进行输出，而 SSD没有全连接层，而是添加了若干卷积层，每个卷积层得到的feature map上用特定的filter产生网络的输出，每个filter数出一个偏移值或者置信度。使每个位置对应k个default box，要输出的是每个default boxes 对于不同物体类别的偏移量(shape offsets，由4个参数描述) 以及其置信度(c1,c2,…,cp)，如下图：



（图片摘自原论文）

因此在大小为m x n的feature map上，若物体种类为c个，则需要(c+4)k 个filter，产生(c+4)kmn个输出，不同的feature map上default box尺度不同，因此可以用于不同尺度的物体检测。两者的结构对比如下：



（图片摘自原论文）

我们用了预训练好的模型进行目标检测，利用小车上手机传回的图像，在图像上运用模型得到相应物体的种类和位置，利用这些信息进行判断，可以使得小车行驶到特定的物体附近。

综上，在本课程中，打下了一定的机器学习基础，熟悉了机器学习方面相关工具的使用，也培养了查阅资料和文献的能力，在组队完成大作业的过程中加强了团队协作能力，在多方面都有所提高。

参考文献：

[1] Joseph Redmon , Santosh Divvala , Ross Girshick , Ali Farhadi：You Only Look Once: Unified, Real-Time Object Detection. *In Computer Vision and Pattern Recognition (CVPR)*， 2016

[2] [W Liu](http://xueshu.baidu.com/s?wd=author%3A%28Wei%20Liu%29%20UNC%20Chapel%20Hill&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)， [D Anguelov](http://xueshu.baidu.com/s?wd=author%3A%28Dragomir%20Anguelov%29%20Zoox%20Inc&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)， [D Erhan](http://xueshu.baidu.com/s?wd=author%3A%28Dumitru%20Erhan%29%20Google%20Inc&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)， [C Szegedy](http://xueshu.baidu.com/s?wd=author%3A%28Christian%20Szegedy%29%20Google%20Inc&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson)， [S Reed](http://xueshu.baidu.com/s?wd=author%3A%28Scott%20Reed%29%20University%20of%20Michigan&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson):　SSD: Single Shot MultiBox Detector．*In European Conference on Computer Vision（ECCV），*2016

樊树霖

2015201919