**《图像基础》**

红绿蓝(RGB)通道：最常用的颜色标准，一幅图由三种颜色叠加而成(0-255)，人眼所能看到的颜色，都可以通过调整红绿蓝三种颜色的比例无限接近

色相、饱和度、亮度(HSV)通道：色相主要控制色彩的变化(0°-360°)，饱和度控制颜色的纯度(0-255)，亮度控制图片的亮度(0-255)。三通道互有影响

图像特征：形状、颜色、纹理

语义鸿沟：相同视觉特性不同语义概念，相同语义概念不同视觉特性

邻域平均法：，邻域平均法去高斯噪声较好，缺点是造成图像模糊，可以设置阈值检测异常灰度值改进。优化算法有：K近邻平均法、梯度导数加权平滑、最大均匀性平滑、小斜面平滑等。

均值滤波：。

高斯滤波：，更好保留轮廓，比均值滤波好

中值滤波法：选取方框内的中位数。非线性，适合滤除脉冲干扰和图像扫描噪声，不适合细节多的图像，去椒盐噪声好。

Sobel算子：和

Scharr算子：和，相比Sobel算子加强了轮廓

Laplacian算子：，不单独用

Canny边缘检测：高斯滤波去噪；计算每个像素点的梯度和方向；非极大值抑制消除散杂响应；双阈值检测；抑制孤立弱边缘。

图像金字塔：高斯金字塔、拉普拉斯金字塔

模板匹配：待匹配的区域在原图中从左到右从上到下进行平方差/相关性/相关系数的计算，选出最接近的区别。模板匹配是最简单的目标检测、计算效率高、可以检测多个相似对象，但只适合静态图与模板匹配。

傅里叶变换作用：高频：变化剧烈的灰度分量，如边界，高通滤波器保留高频分量只保留边界；低频：变化平缓的灰度分量，如大海低通滤波器保留低频分量使图像模糊

**《神经网络》**

静态图和动态图：静态图先定义好网络再计算，速度快；动态图边计算边构建网络，计算速度较慢但便于调试

单层神经网络/感知机：线性分类，凸函数，在前馈型神经网络中，只要有一个隐含层并且神经元足够多，就能近似任何的连续函数或空间分布

多层神经网络：非线性分类，非凸函数，多隐含层可以表示非连续的函数或空间分布、减少每层神经元数量和泛化误差

感受野：每个神经元所能感受到的数据范围，使用卷积提高神经元感受野。如经过一次的卷积感受野就从1提高到了9，因此感受越高理论上越能判断物体的全貌。但在目标检测中，会出现预测的框比感受野大的情况，一种解释是网络可以通过局部预测全貌

特征提取层：对神经网络模块的泛称，特征提取是一个模糊的概念，可抽象理解为网络在卷积和参数更新的过程中，不断把重要的参数权重分配给能区分物体的部分，使要检测的部位通过训练成为一个’感兴趣’区域

上采样：插值法、反池化、反卷积

下采样：缩小图像。当卷积时的步长大于1时也是一种下采样，常用步长为2的卷积代替池化层

**全连接层(FC)**

分别建立每一个输出和每一个输入的线性关系，具有强大的拟合能力，但容易过拟合。全连接层的输入为一维向量，输入长度，输出长度时，模型参数为，1为常数项。通常单纯的叠加全连接层不会进一步提高模型性能，反而容易过拟合

多层感知器(Multi-Layer Perceptron)：由多层全连接层组成的网络

**批归一化层(BN)**

数据归一化操作，加在卷积等网络层后，防止梯度消失有助于收敛。参数：输入通道数，其中一半参数可训练。BN层一般加在激活函数前，因为数据经过运算后的分布会改变，比如sigmoid函数在离零点太远的地方结果相差不大，梯度会很小，导致模型更新缓慢，但如果归一化到零附近会使其趋于线性，因此加上可训练的参数进行一定的弥补。使用的批量越大BN层统计的越精确、归一化的效果越好，如果使用多GPU训练，用SyncBatchNorm可以统计所有卡的批量

其中和为可训练参数，为归一化的逆操作。总体相当于一个可训练的动态归一化，在防止梯度消失和保持更多非线性之间权衡。代表了该网络层的重要程度，如果模型训练后的某个的值很小，则该层在后续的计算中影响程度很小，因此也作为模型剪枝的参考标准

**池化层(Pool)**

下采样的一种，减少参数，防止过拟合，一般用最大池化层。最大池化层：目标偏向纹理轮廓，网络浅层使用最大池化过滤无用信息。平均池化层：目标偏向背景等，网络深层使用平均池化层防止丢失过多信息

**丢弃层(Dropout)**

降低训练集准确率，提高泛化性，提高测试准确率，一般为0.1-0.5，接小层后可取0.2，接大层后可取0.5

**卷积网络(CNN)**

由卷积层组成的网络

卷积层(Conv)：由多个卷积核组成，作用为提取关键信息。提高卷积层中卷积核的数量可提高训练准确率，但测试准确率有瓶颈，后一卷积层的卷积核个数一般大于/等于前一层

卷积核：形状为的卷积核在图片上滑动进行互相关/卷积运算，为奇数一般为3，为前一层维度(通道数)。卷积核计算后，将所有卷积计算结果相加合成一个通道，再加上偏置项，一个卷积核总的参数为。卷积核用于升降维、减少参数量或加在卷积前后缓冲。相同参数下，多用小卷积核比大卷积核有更多感受野

膨胀卷积(Dilated Convolution):卷积核中隔行参数参与计算，相同计算量下膨胀卷积的卷积核更大，感受野更广

因果卷积(Causal Convolution):在一维数据中为了使卷积前后长度不变要在数据前后填充，因果卷积全部填充在前面

**图卷积网络(GCN)**

由图卷积层组成的网络。与CV和NLP的区别为数据格式不固定，数据之间有特殊关系，并非所有的点都有标签，计算损失时也只计算有标签的点(因此经常会是半监督任务)

图卷积层(Graph Conv)：先对邻接矩阵A进行特殊的归一化，然后乘以特征矩阵B0，再乘以权重矩阵C0，最后乘以激活函数relu得到结果result0，此时为一层图卷积。第二层时为：A\*result0\*C1\*relu，直到最后一层时不用加relu。特殊归一化的公式为，目的是归一化的同时保证自身权重的占比更大。图结构经过图卷积后的结构不变，随着图卷积层数的增加，每个点的参数更新就会受到更多点的影响，感受野增大，但通常只需要2、3层时效果就达到最好，过深的层数反而可能导致效果变差(可以理解为感受野增大反而减少了个体的差异)

图的邻接矩阵：由0和1组成的矩阵，表示点与点之间的关系，有关系为1，无关系为0，自己与自己的关系也为1。比如有五个物体，邻接矩阵的尺寸为25\*25

损失函数(Loss)：衡量输出和标签值的误差，每次训练中根据一个批量数据的平均损失更新参数。在网络的训练中损失函数应该下降一个以上数量级

误差反向传播(Back Propagation)：利用链式法则，计算出损失函数对所有参数的梯度，根据梯度更新参数。在特征提取网络中，损失函数(任务)可能有多个，此时参数的更新方向取决于所有损失函数，只有特征提取网络后输出层的参数才取决于对应的损失函数。通常各个损失之间会存在对抗，可在特征提取网络后单独对每个任务进行卷积加强区分

卷积神经网络(Convolutional Neural Network)：由多层卷积层组成的网络

循环神经网络(Recurrent Neural Networks)(1990)：可以处理任意长度的输入序列，可以对序列中的任意位置进行操作；训练过程中容易出现梯度消失或梯度爆炸问题，难以处理长期依赖关系

**长短期记忆网络(LSTM)(1997)**

加强对序列近期位置的记忆，减少对历史久远位置的记忆。内部有输入们、遗忘门、输出门来实现记忆功能

**注意力机制(Transformer)(2017)**

注意力机制在多数情况下的效果比RNN好，并且部分结构能够在GPU上并行计算提高效率。详见后面transformer模型。广泛应用于自然语言处理，少量应用于图片处理。在时间序列预测中效果不好，因为Transformer中会建立每个输入之间对应的关系，输入具有置换不变性，忽略了时序关系，随着输入长度的增加预测性能会下降，虽然Transformer的位置编码对此有一定改善

**视觉注意力机制(Vision Transformer)(2020)**

将Transformer对图片进行特征提取。具体是先用大尺度卷积大幅减少尺寸的同时大幅提高维度，再将尺寸的两个维度展平为一个维度，然后送入到Transformer中

**《激活函数》**

激活函数(Activation)：不加激活函数时无论有多少层隐含层都可以化简为一层线性变换，激活函数应该是简单的非线性函数、连续可导、单调函数

**sigmoid**

函数：torch.nn.Sigmoid()

对数几率函数。将数值压缩在0-1之间。常用于二分类和精确的归一化



**softmax**

函数：torch.nn.Softmax(dim)

将数值压缩在0-1之间，值可以理解为概率分布。、为第i、j个节点的输出值。常用于多分类和不精确的归一化

**tanh**

函数：torch.nn.Tanh()



**relu**

函数：torch.nn.ReLU()



**leaky relu**

函数：torch.nn.LeakyReLu()



**softplus**



**mish**



**silu**

函数：torch.nn.SiLU()

与mish类似，最早在强化学习中提出，具有自稳定性，隐性正则效果



**《损失函数》**

**平均误差(MAE)**

函数：torch.nn.L1Loss()

**平方损失函数(MSE)**

函数：torch.nn.MSELoss()

平方损失函数对w和b求导时要对求导，值可能非常小，导致更新速度缓慢，逻辑回归中可能有多个极小值点

**平均交叉熵损失(Crossentropy)**

函数：torch.nn.BCEWithLogitsLoss()，torch.nn.CrossEntropyLoss

平均交叉熵损失函数对w和b求导时不需要对求导

**IOU(Intersection over Union)**

a为真实框，b为预测框，(,)为左上坐标，(,)为右下坐标。w、h为宽和高。c为极小值防止union为0(特殊情况如边框错误等)，由于计算时+c，也+c

IoU问题：当IoU=0时梯度为零；收敛较慢、回归不够准确

**GIOU(CVPR2019)**

其中G为两框的最小外接框面积

GIoU解决了梯度为零的问题，效果略比IoU好，但当预测框和真实框水平或垂直时，退化成了IoU，依然收敛较慢、回归不够准确

**DIOU**

其中为两框的中心点距离，为最小外接框左上到右下的对角线距离

DIoU收敛速度和准确率都有所提高，但当预测框和真实框接近时退化为IoU

**CIOU**

为极小值防止分母为0，其中当作常数，因此计算损失时关闭自动求导

CIoU中v相对于w和h的梯度：

当w和h在时容易导致梯度爆炸，因此实际中常用1代替。但YOLOV5中由于会对w,h进行筛选从而防止梯度爆炸，因此不需要调整

**正则化**

在损失函数中加入正则化可以让模型的参数尽量变小，但一般不显示出来

L1正则化：函数是线性的(V字)，会让模型中部分不重要的参数接近于0，简化了模型结构，缓解了过拟合，权重系数可为0.0001。模型剪枝时可以根据L1正则化的结果去除不必要的结构。L1正则化一般加在批归一化层的梯度中

L2正则化：函数是曲线的(U字)，会让模型中所有的参数减小但不至于接近于0，减少某些较大参数对模型的影响，防止了过拟合，权重系数可为0.0001。模型剪枝时也可以根据L2正则化的结果去除不必要的结构。L2正则化在函数中有封装

**《CV》**

**数据处理**

[图片]-->[数据增强]-->[归一化]-->[合成批量(batch,size,size,3)]-->[模型]

归一化：除以255、减均值、除以标准差、等比缩放图片和填充灰边(128)，由于图片各通道标准差差不多，可以直接除以255。在有的模型中可以直接缩放图片

数据增强：旋转、仿射变换、翻转、缩放、平移、对比度变换、噪声扰动

马赛克增强(mosaic)(2020)：将四张图片随机进行大小调整、翻转、裁剪等后合在一起，同时标签也做对应的调整。优点：增加了训练的速度；缩小了图像以提高对小目标的检测能力；排除检测目标的背景干扰。但由于增强后的图片与真实场景不一致，因此最后几轮的训练通常不使用马赛克增强

图片分类：通常只需要轻微色彩变换和加噪

目标检测：需要进行马赛克增强

**数据格式：**

YOLO数据格式：(images，labels，train\_images.txt，val\_images.txt，test\_images.txt)，测试集可以不要，images和labels中可以有嵌套文件夹，但.txt中路径要正确，.txt中可以存入相对路径(images/...)或绝对路径。labels中存有‘图片名.txt’，内容为[类别号 Cx Cy w h]相对坐标，以左上角为坐标原点



VOC数据格式：(Annotations、ImageSets/Main、JPEGImages)，Annotations存放每张图片的.xml标注，ImageSets/Main中train.txt、val.txt存放训练和验证的无后缀图片名、JPEGImages中存放所有图片

**任务类型**

**(1)图像分类(Classification)**

输入固定大小的图片，得到图片的类别

[预处理] --> [模型] --> [(batch,dim,n,n)] --> [平均池化(batch,dim,1,1)] --> [(batch,dim)] --> [(batch,类别数)]

**(2)目标检测(Object Detection)**

输入固定大小的图片，得到图片中目标物体的边框

[预处理] --> [模型] --> [(batch,dim,大层长度n,大层长度n)\*大层数3] --> [(batch,小层数3,n,n,内容m)\*3] --> [内容m：(Cx,Cy,w,h,置信度,类别独热编码)] --> [归一化] --> [边框解码] --> [置信度筛选] --> [非极大值抑制]

**目标检测解码：**

YOLO输出的特征图后需要归一化(sigmoid)，然后人为转换成，并赋予其意义：3为小层数、(80,80)为图像的位置、4为物体的边框(cx,cy,w,h)、1为置信度、80为物体的种类

如YOLOV5的解码方式为：

**先验框(anchor)：**固定比例或由大数据kmeans聚类得到，有w和h组成，YOLOV5中每个输出中有3个通道，分别对应3个先验框，使每个预测可用多种先验框匹配。总共有3个输出、9个先验框

**目标检测损失：**

a+b+c=1。输出特征图较大时置信度的损失权重要大些，让网络更专注于识别物体的位置。一般置信度和类别损失用交叉熵损失，边框损失用IoU系列。输出所有位置的置信度都计算损失；但边框和类别只有有标签对应的位置计算损失

边框、置信度、类别损失之间会有一点冲突，在特征提取网络之后可以加入几层卷积网络对各任务分别计算

一个的输出有个预测位置，只有部分位置的预测有用、参与损失计算、有标签矩阵对应，可以利用判断矩阵筛选

在计算置信度损失时，由于大多数地方都不需要预测为0，因此需要提高正样本的权重

**聚焦损失(Focal Loss)：**当训练集中正样本很少时，即使模型全预测为负样本准确率也会比较高，因此需要加大正样本预测错时的损失

**判断矩阵：**

输出的标签矩阵为，判断矩阵为，pred\_matrix[judge\_matrix]取出需要计算损失的位置，label\_matrix[judge\_matrix]取出对应位置的标签

**标签矩阵：**

标签矩阵的形状和网络输出一致，在制作标签矩阵时，需要预测的位置填入对应标签，不需要的位置label\_matrix[judge\_matrix]会自动过滤，填入的标签需要满足一定条件，因为网络输出解码后的范围要包含这个标签的值才能预测

有的网络如YOLOV5扩充了标签中有物体的位置，物体可以由四周最近的两个网格预测

**wh筛选条件：**若解码方式为：

为网络输出，为真是输出。因此标签筛选范围为：

**标签平滑(Label Smoothing):**类别标签一般为0或1，但有时虽然不是目标但和目标很像，可以给个很小的数值，比如目标值给0.99，非目标值0.01

**建议框(proposal)：**根据置信度筛选后还未进行非极大值抑制的框

**非极大值抑制：**将某一种类别的建议框按置信度筛选后，按置信度排序，选出置信度最高的框依次遍历计算与其他框的iou，根据iou阈值去掉重叠的框，记录选出的框，再用置信度次高的框接着遍历剩下的框，直到列表为空

**衡量指标：**

准确率、精确率、召回率、AP

**正负样本：**

一般空标签图片取10%-20%

**目标检测指标：**目标检测一般采用非极大值抑制后的，因此没有TN和准确率

TP：存在真实框，置信度大于阈值，iou大于阈值，类别正确

FN：存在真实框，置信度、iou、类别任意不满足

TN：不存在真实框，置信度小于阈值

FP：不存在真实框，置信度大于阈值

**(3)人脸识别**

用人脸图片提取特征建立人脸特征数据库、目标检测识别人脸、根据人脸关键点提取特征、在人脸特征数据库中比对

与图片库中比对：可以使用np.dot(a=(),b=())，其中a为要比对的人脸特征，b为库中的所有人脸特征，k为特征数量，n为库中所有人脸数

人脸关键点定位：一般用5个或68个关键点定位

**(4)OCR识别**

目标检测识别文字位置、方向分类模型识别文字旋转情况、文字识别模型识别文字

**(5)行人重识别(2022年的热点)**

先用目标检测检测出行人的截图，然后进行相似度匹配

rank1：与目标匹配相似度最高的图片正确的概率

map：所有行人的ap平均值

ap：假设n个匹配结果中第1、3、4、...张图片正确，则ap=，表示第三张图片的位置本应该在第二个位置

**《NLP》**

**基础知识**

词：一个词由一个字符或多个字符组成，每个词对应词表中的一个词索引(token)

词索引(token)：词对应的整数索引，从0开始，有多少词就有多少词索引

词表(vocab)：写入所有模型支持的词，词索引=词的序号-1。一般词索引0是填充符，1是开始符，2是结束符，此外还有特殊符号、单词、汉字、表情等。一个支持中英文的大语言模型(GPT)词表长度可能有5.5万左右

词索引向量(tokens)：由词索引组成的向量(m,)，通常要补齐长度合成批量(batch,n)

词索引编码(tokenizer)：根据词表的对应关系将字符串变为词索引向量，会在前后加上开始、结束标志(1和2)，合成批量时会在短的句子后面填充0

词索引解码：词索引编码的逆过程，将每个词索引变回对应的字符

空白填充(pad\_token\_id)：为了使输入长度一样以合成批量，需要对短的向量填充0

开始标记(bos\_token\_id)：词索引向量的开始标记，通常为1

结束标记(eos\_token\_id)：词索引向量的结束标记，通常为2

编码器(encoder)：主要由词向量编码、位置编码、6个(可变)编码模块组成，编码模块：自注意力层+全连接层。功能为对输入的词索引向量进行运算，只需运算一次

解码器(decoder)：主要由词向量编码、位置编码、6个(可变)解码模块、输出解码组成。解码模块：自注意力层+编解码注意力层+全连接层。使用时会进行多轮运算，将之前生成的结果合并输入到下一轮的预测中。不同任务的解码器输入略有不同，标准transformer模型中每一轮要将编码器的结果输入到解码器的中部。模型预测时是一个字一个字地生成的，但在训练时由于有标签，配合掩码矩阵可以将标签一次性输入

词向量编码(embedding)：编码器和解码器的第一层，编解码器的词向量编码层是共用的。通过特定的算法将词索引向量(batch,n)变为词向量(batch,n,dim)，并且能考虑到词与词之间的距离关系。词索引向量的长度n可以是任意的，但索引值要在词表总数的范围内，词向量长度dim是固定的。词向量编码层的参数是可训练的

词向量解码：解码器的最后一层，可为全连接层。输入词向量(batch,n-1,dim)，词表长度m，则经过全连接层后为(batch,n-1,m)，经过argmax后得到词索引向量(batch,n-1)。如果是语言生成模型，不能直接使用argmax，而是要根据数据的概率分布来选值。如果是在训练，不需要argmax，而是变形为(batch\*(n-1),m)与标签进行交叉熵损失。使用torch.nn.CrossEntropyLoss时标签可以直接传入词索引向量，不用变成独热编码

位置编码(Positional Encoding)：编码器和解码器的第二层，编解码器的位置编码层是共用的。记录词索引的位置信息。有绝对位置编码、相对位置编码等。位置编码层的参数是可训练的

自注意力层(self-Attention)：建立输入的每个词之间的联系。假设词向量长度为m，输入形状为(batch,m,512)。以(m,512)为例，经过查询矩阵(512,n)、键矩阵(512,n)、值矩阵(512,n)得到查询向量(m,n)、键向量(m,n)、值向量(m,n)。每个词的查询向量会与所有词的键向量逐一点乘得到一个注意力得分(m,)，此得分表示每个词对所有词的注意力。然后对得分除以常数和归一化，再与所有词的值向量(m,n)逐一相乘得到(m,n)，在m维度相加得到一个词的输出(n,)。所有词运算后得到(m,n)。通常有多个注意力层并行计算，所有结果合并(m,n\*a)，再降维成原来的(m,512)

标准transformer：由编码器和解码器组成。在torch.nn.Transformer函数中为了灵活将词向量编码、位置编码、输出解码部分单独取出。实际中应用中，语言生成模型(GPT)只使用其解码器，多模态文本模型只使用其编码器

翻译任务：使用编码器和解码器，每一轮都要将编码器的结果输入到解码器的中部。解码器的第一轮输入<bos>生成第一个词索引，第二轮输入[<bos>+生成的词索引]得到新的词索引，之后不断重复直到最后一轮生成<eos>停止。生成长度是输入长度-1

语言生成任务(GPT)：只需要解码器，解码器的输入与翻译任务类似，只是输入不是从<bos>开始，而是从[<bos>+输入内容]开始，之后同样将前一轮的输出结果合并后作为下一轮的输入。语言模型的基础训练是输入一句话来预测下一个字，当模型生成能力达标后需要用指令(instruction)数据训练才能达到对话的效果，具体是在输入中加入特殊的提示词(prompt)模版(template)。模型在词向量解码为词索引时，不能直接使用argmax，而是要根据数据的概率分布来选值

自回归(Autoregressive)：解码器将本轮的输入与输出合并，作为下一轮的输入

翻译模型训练：模型预测时需要获取之前的预测结果，所以只能一个字一个字的生成，而训练时由于有标签，可以批量输入。比如输入AB得到C，输入ABC得到D。假设编码器输入=(batch,n)，解码器则输入=(batch,n-1)(不需要最后的结束符)，解码器输出=(batch,n-1)，标签为(n-1)(不需要前面的开始符)。由于批量长度是固定的，因此在代码中为了批量训练还需要配套掩码矩阵

掩码矩阵：掩码矩阵左下边和对角线为0，右上边为torch.-inf，之所以用-inf是因为经过softmax时会得到0

语言模型的generate：会对模型的输出进行一些调整操作

对话模型：对话模型是在语言生成模型的基础上加入了提示词微调，模型的输入需要额外加上提示词模版

单轮对话：一问一答的方式，模型不会考虑之前的对话内容

多轮问答(上下文对话)：模型在回答当前问题时会考虑之前的内容

监督微调(sft)：提示词微调、指令微调，sft数据格式

**数据处理**

[字符串] --> [词索引向量(tokens)] --> [填充0合成批量(长度不定)] --> [词向量(长度固定)] --> [模型]

词向量编码(embedding)也可算作模型的一部分

**sft数据集格式(指令微调)：**

data\_list=[{'instruction':'A','input':'B','output':'C'},...]：sft数据格式。instruction+'\n'+input为真实输入数据，因此input也可以没有，没有input时不用加'\n'；output为输出

prompt='You are a helpful assistant. 你是一个乐于助人的助手。'：提示词

template='[INST] <<SYS>>\n{prompt}\n<</SYS>>\n\n{instruction} [/INST]'：模版

text=template.format(prompt=prompt,instruction=instruction)：在输入中加入提示词模版。instruction为输入文本，除instruction外编码后长度为35

template\_add=' {answer}</s><s>[INST] {instruction} [/INST]'：三个空格

text=text+template\_add.format(answer=answer,instruction=instruction)：多伦对话时要追加历史信息。answer为上一轮的回答，除instruction外编码后长度为10

**任务类型**

**(1)文本分类**

输入一个句子得到句子的类别

[预处理] --> [模型] --> [输出(batch,词数n,长度dim)] --> [(batch,dim,n)] --> [平均池化(batch,dim,1)] --> [(batch,dim)] --> [(batch,类别数)]

衡量指标：准确率、精确率、召回率、F1指标

**(2)序列标注**

对句子中的每一个词做分类

[预处理] --> [模型] --> [输出(batch,词数n,长度dim)] --> [(batch,n,类别数)]

**《TSF》**

**数据处理**

数值可能没有固定的范围，常用z-score：减均值、除以标准差，但如果训练集和测试集数据差别比较大时，以历史数据为标准的处理方法可能不适用于未来数据

z-score：假定数据基本满足高斯分布，

**任务类型**

单变量自标签预测：输入一段序列的前半段，预测的标签是这段序列的后半段。注意有时虽然模型输入的是多个变量，但它们之间没有进行融合，本质上也是单变量。模型学习的是序列的周期变化

单变量异标签预测：输入一段序列，预测的标签是另一段/多段与其相关的序列。模型学习的是各序列之间的关系

多标签预测：输入多段序列，预测的标签是一段/多段与其相关的序列，模型内部要对各输入序列进行融合。模型学习的是各序列之间的关系

《**多模态》**

多模态就是通过特征向量将图片和文本结合起来。比如：输入文本对图片进行操作，输入图片生成文本描述

clip(2021)是多模态发展的经典之作，在clip模型中图片模型和文本模型是分开的，因此可以做多个类型的任务。除此之外也有一些合在一起端到端的模型，但基本原理也是通过特征向量将图片与文本结合

从视觉多模态的角度看，RAM(2023)是图片分类的未来发展方向，GroundingDINO(2023)是目标检测的未来发展方向。即使它们目前的表现还存在不稳定，但在未来或许能集成一种视觉类的通用模型，就像GPT在nlp领域一样，并且通过与GPT结合实现通过语言对图像的精确操控

多模态模型对齐：由于输入数据有多个(如图片和文本)，经过网络后要让它们的特征对齐，这个特征类似标签，让不同的输入数据之间建立关系

对比学习：图片和文本先经过各自的模型得到各自的特征，再利用特征进行相似度匹配。如CLIP(2021)模型

**《模型训练》**

**基础知识**

数据类型：一般训练时用float32，GPU可启用混合精度训练，预测时GPU可用float16

训练集：训练模型的数据

验证集/开发集：每轮训练后测试模型的数据，并以此调整模型和参数

测试集：训练后测试模型性能的数据，不能以此调整模型和参数

预训练模型：加载模型官方在大型数据集上训练好的模型权重，可以大幅加快训练速度

训练速度：通常参数越多的模型训练和推理速度越慢，但除了参数量也与网络结构有关，好的网络结构即使参数很多但训练和推理速度很快，指标也更好。很多模型有官方的预训练权重，可以加快训练速度。模型训练速度比模型验证速度快很多，因为不需要反向求导和更新参数

训练震荡：在训练开始后的一段时间或训练中途，有可能会出现指标异常的情况，越大的模型训练越慢，也越容易出现震荡，之后会慢慢变得平稳。在训练稳定之后继续训练可能出现指标下降的现象，需要保存最佳结果。在加载已有模型接着训练时，学习率、指数移动平均等最好也为保存时的状态，否则可能出现震荡

图片分类训练异常：在多分类的训练过程中，类别样本有的多、有的少，可能会出现只有一个类别指标正常，其他类别指标下降为0的异常现象，一般要等到正常的类别指标稳定之后，其他类别的指标才会依次回归正常。因为模型为了达到损失最小，会优先学习样本多的类别而暂时‘忽略’样本少的类别

轮次(epoch)：当[迭代次数\*批量大小<=训练集]时为一轮，每轮训练前都要打乱训练数据，增加随机性，防止模型对前面的数据遗忘

迭代次数：推理一次为一次迭代。若训练集100个，批量大小为10，则每轮迭代10次

批量大小(batch\_size)：每次迭代放入多少数据（即训练完多少数据更新一次参数）。相同数据时批量越大更新迭代次数越少、收敛越快，在内存允许时一般尽量用最大批量训练，最好是CPU/GPU数2的幂次，但要留有一定的GPU显存，一般在训练+验证完一轮后会略微增加显存占用。对于GNN的数据，可能每组数据都有上万的点，在不大幅度调整学习率的情况下采用批量为1时的效果可能会更好

**训练方法**

混合精度训练：通常采用float32训练，预测用float16。但对于复杂的网络通常有参数不太重要的层，可以用float16精度混合训练

指数移动平均(EMA)：给予近期数据更高权重的平均方法，通常都会使用。在模型每轮训练的每次迭代中，都会计算批量的损失然后反向传播更新参数。但直接使用平均的方法会容易受个别异常批量值的影响，不使用平均则得不到历史信息。EMA的初衷是为了使每次参数的更新更加平滑。EMA指数平均移动公式：

\*其中为当前原始值，加权后的值，为上一次加权后的值，a为权重系数

分布式训练：有模型并行和数据并行，有单机多卡、多机多卡等情况。一般采用数据并行，同一份代码会在不同GPU上运行，先把模型放到不同GPU上，每个GPU对一个batch推理和计算损失（先完成的GPU会一直等待），之后将损失结果平均并进行反向传播，所有GPU上的模型同步更新参数。验证和计算指标也可以设置为分布式。模型保存等代码只需在主设备上运行。分布式训练效果可能会比单卡差，刚开始损失下降可能会很慢

BGD(batch gradient descent)批量梯度下降：每次训练时放入所有样本

SGD(stochastic gradient descent)随机梯度下降：每次训练时放入单个样本，会导致模型对最后使用的样本拟合更好，容易导致参数来回更新，收敛慢

MBGD(mini-batch gradient descent)小批量梯度下降：每次训练时放入部分样本，小批量样本与整体样本之间存在偏差，提高了泛化能力。1，使用前要打乱样本顺序；2，样本数量选取是关键（GPU运算可以考虑使用2的幂次数量）



**学习率**

学习率：图像类adam学习率一般在0.001以内，sgd一般在0.01以内。对于时间序列预测和自然语言处理一般比图像类小一个数量级

模型微调：使用模型官方在大型数据集上训练好的模型权重，可以利用模型在其他数据上学习到的参数，用很小的学习率训练和少量的数据就能得到很好的效果。bert分类模型微调学习率通常在0.00001左右，如果学习率大容易破坏参数

学习率预热：刚开始训练时让学习率从很小开始增加，在几轮后(比如3轮)变为初始学习率。前三轮学习率可以下降为初始学习率的0.1、0.3、0.5

学习率衰减：每轮训练后，一般要调整学习率的大小，使学习率随着训练轮次的增加越来越小。对于复杂的网络，末尾学习率的设置可以不用太小，比如使用adam算法时，初始学习率设置为0.001，末尾学习率设置为0.00005。可以使用余玄下降公式：

其中为训练轮次，n\_为预热训练总轮次，为学习率调整总次数，是初始学习率

Adagrad(adaptive gradient)自适应梯度：在凸函数上有快速单调收敛的效果，但存在学习率过早减小的情况，在非凸函数上效果不好。为梯度，c为防止分母为零的极小常熟，可为

RMSProp(Root Mean Square Prop)均方根道具：在Adagrad基础上加了超参数来衰减历史值，目前为常用优化算法

**指标**

TP：正确判断正类为正类

FN：错误判断正类为负类

TN：正确判断负类为负类

FP：错误判断负类为正类

准确率(accuracy)：T/(T+F)

精确率(precision)：TP/(TP+FP)

召回率(recall)：TP/(TP+FN)

F1\_score：(2\*precision\*recall)/(precision+recall)。其中F1——micro先统计总TP等再计算；F1\_macro计算各类别F1再平均(常用)

排列熵(Permutation Entropy)：度量时间序列复杂性的一种方法，时间序列复杂性越低，序列的可预测性就越强

模型大小：约等于[参数量\*字节数/1024/1024](MB)。float32占4字节

**《模型预测》**

[模型] --> [模型剪枝训练] --> [模型蒸馏训练] --> [模型量化训练] --> [预测代码] --> [gradio可视化] --> [falsk服务部署] --> [gunicorn多进程] --> [docker镜像打包]

模型预热：模型刚开始预测速度会比较慢，之后速度会加快并趋于稳定

模型推理预处理耗时：包括数据解码、图片等尺寸转换、将张量放到gpu上，一般解码和尺寸转换时间略高，将批量张量一起放到gpu上比单个放快很多。但如果是trt模型，批量预测相对单张预测只有在小模型上有一点优势，大模型上差不多，甚至单张预测更快，反而单张预测占显存更少

模型推理：目标检测一般置信度阈值为0.35，非极大值抑制阈值为0.65。一般在内存允许时用大批量推理会快很多。但对于trt模型无论占gpu显存大小，单张预测和批量预测只有在小模型上有一点优势，大模型上差不多甚至单张更快

推理精度：模型推理使用的数据类型一般为float32、float16或int8，一般使用float16几乎不改变精度(归一化到0-1的输出，误差数量级在0.001以内，在图片中不到一个像素)，但量化为int8会有较明显的影响。相同的非极大值抑制阈值下，可能量化前抑制只有一个框，量化后抑制会有两个框，此时需要降低非极大值抑制阈值，但即使只留下一个框，框边的精确度也会下降。使用float16精度时，如果pytorch版本不一样，可能会出现nan的情况

剪枝：使用L1或L2正则化后，批归一化层中不重要结构的参数会很小，根据参数的大小去除卷积层中不必要的通道，一般裁剪后保留的通道数可为30%-80%

蒸馏：需要一个成熟的模型去指导训练的模型

量化**：**将模型浮点数(如float32)替换为整数(int8)

非饱和方式：若浮点数正负绝对值的最大值为a，则将-a至a映射到整数的最大值和最小值，方便计算，但有效值的动态范围会减小

仿射方式：将浮点数最大最小值对应映射到整数的最大最小值

饱和方式：先计算浮点数的阈值T，然后根据浮点数的正负阈值T饱和截断，之后映射到整数的最大最小值

动态离线量化：非饱和方式量化权重，不需要样本数据。

静态离线量化(PTQ)：饱和方式量化权重，非饱和方式量化激活。需使用少量的样本数据进行模型的前向计算，对激活进行数值采样。PTQ在大模型上精度损失比小模型少

量化训练(QAT)：非饱和方式量化权重和激活，并在训练过程中更新权重，需要使用大量有标签样本数据

**《模型项目》**

**ImageNet**

2007年ImageNet项目创立，为ILSVRC比赛提供数据支持

**AlexNet(2012)：**

论文：ImageNet Classifification with Deep Convolutional Neural Networks

2012年ILSVRC比赛冠军，8层(不算pool)



**OverFeat(2013)：**

论文：OverFeat:Integrated Recognition,Localization and Detection using Convolutional Networks

2013年ILSVRC比赛冠军，OverFeat为AlexNet的改进版，有快速版和精确版两个版本

**VGGNet(2014)：**

论文：VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

2014年ILSVRC比赛分类任务亚军和定位任务冠军，有A-E五种模型

**GoogleNet(2014)：**

论文：Going deeper with convolutions

2014年ILSVRC比赛分类任务冠军和定位任务亚军。GoogleNet有3个分类器，另外两个辅助分类器接4a和4d后，但在比赛中Googlenet只使用了最后的分类器。

GoogleNet提出inception层，用不同卷积核提取网络再合并，利用稀疏性解决参数量大和节约计算资源，加入卷积减少参数量

**ResNet(2015)：**

论文：Deep Residual Learning for Image Recognition



2015年ILSVRC比赛冠军。模型常为50和101层残差网络。加入residual结构，特征融合时采用相加，当输入和输出形状不同时，要用的卷积核对分支升维(不加激活函数)。对于50/101/152层的网络用右边的结构可大幅减少参数

**ResNext(2016)：**

论文：Aggregated Residual Transformations for Deep Neural Networks



2016年ILSVRC比赛冠军。ResNext为ResNet的改进版，相当于inception和residual的结合

**SENet(2017)：**

论文：Squeeze-and-Excitation Networks



2017年ILSVRC比赛冠军

**图像分类**

**Private Detector(2022)：**

github：https://github.com/bumble-tech/private-detector

开源色情分类模型

**Efficientnetv2(2021)：**

<https://github.com/google/automl/tree/master/efficientnetv2>

**INTERN(2023)：**

论文：InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions

github：<https://github.com/OpenGVLab/InternImage>

商汤科技发布的多模态多任务通用大模型，拥有30亿参数，以CNN网络为核心。在相同的主干网络上微调，可以实现不同的任务类型

**目标检测**

**FPN(Feature Pyramid Networks)(2017)：**

论文：Feature Pyramid Networks for Object Detection



**Mask RCNN(2017)：**

论文：Mask R-CNN

github：<https://github.com/matterport/Mask_RCNN>

或https://github.com/facebookresearch/Detectron



在Faster RCNN(2015)上做了改进，速度差不多但性能提高，可用于目标检测、图像分割和人体姿态估计。Mask RCNN最后只取出对应类别的通道做损失计算，不存在类别对抗

**CSPNet(2019)：**

论文：

github：



增强了网络学习能力，使网络轻量化、减少内存的同时保持准确性

**Vision Transformer(2020)：**

论文：An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

github：<https://github.com/google-research/vision_transformer>

2020年goodle提出的将transformer应用于视觉的结构。

**YOLOV4(2020)：**

论文：YOLOv4:Optimal Speed and Accuracy of Object Detection



将Darknet53改为CSPDarknet53。使用特殊池化金字塔SPP(Spatial Pyramid Pooling)和PANet结构。使用CIOU作为损失函数。CSPDarknet使用Mish激活函数，其他依然是Leaky Relu。使用马赛克(Mosaic)增强

SPP结构前后都用卷积降维，池化卷积核大小为5、9、13

**YOLOV5(2020)：**

github：https://github.com/ultralytics/yolov5



P为模型最终预测值，,为预测框的中心点，Y为模型输出，G为网格左上角坐标，stride为缩小的倍数[8,16,32]。YOLOV5的预测值域变为了和[0,4]，是为了将标签网格临近的两个网格也变为标签，[0,4]是为了对应标签筛选条件

各版本YOLOV5的主体网络框架不变，只改变C3层的残差数和通道维度。[第一层输出维度,C3层残差数]：YOLOV5n[16,1]，YOLOV5s[32,1]，YOLOV5m[48,2]，YOLOV5l[64,3]，YOLOV5x[80,4]

SPP层改为SPPF层，输出前后同样用卷积降维，池化卷积核都为5

YOLOV5目前使用CSP层、C3层和SPPF层，使用Silu激活函数，使用CIoU，使用标签平滑。标准输入，输出三层为、、

**YOLOX(2021)：**

论文：YOLOX Exceeding YOLO Series in 2021

github：<https://github.com/Megvii-BaseDetection/YOLOX>

使用Focus和YOLOV4中的CSPnet、SPP进行特征提取

YOLOVX改进解耦合头，原YOLO中只有最后一层的卷积用于得到边框、置信度、类别三种结果，即使前面的的特征提取网络提取很充分，但不用任务之间存在对抗，只用一层卷积难以区分三者的差异。YOLOVX解耦头先将边框与置信度、类别分开卷积，再将置信度与类别分开卷积。YOLOVX不需要先验框(Anchor Free)

SimOTA动态匹配正样本：根据真实框和预测框的重合度计算真实框的k，根据真实框和特征点的预测准确度及包含情况计算Cost代价矩阵，将Cost最低的k个特征点作为真实框的正样本

**YOLOV7(2022)：**

论文：YOLOv7:Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors

github：[https://github.com/WongKinYiu/yolov7](https://github.com/Megvii-BaseDetection/YOLOX)



提出了elan、elan-h层作为YOLOV5的C3层改进版，提出了加速推理的repconv结构

repconv：其思想源于repVGG，在训练时加入特殊的残差结构辅助训练，但在预测时可以等效为一个的卷积，在网络复杂度下降的同时预测性能却没有下降

官方版本训练：与YOLOV5基本一致，只是.yaml文件读取略有区别，小模型使用train.py训练，大模型(W,E,D)使用train\_aux.py训练

**YOLOV8(2023)：**

github：<https://github.com/ultralytics/ultralytics>

根据YOLOV5的C3和提出了YOLOv7的elan提出了C2f结构

**insightface(2021)：**

github：<https://github.com/deepinsight/insightface>

github：<https://github.com/TWK2022/insightface>

github：<https://github.com/TWK2022/FaceRecognition>

insightface是2021年的人脸检测模型。模型输出(xmin,yxmin,xmax,ymax)

**图片分割**

**SAM(2023)：**

论文：Segment Anything

github：<https://github.com/facebookresearch/segment-anything>

示例网址：<https://segment-anything.com/>

2023年facebook发布的图像分割模型(pytorch)，能对图片进行分割，可用于抠图等。有vit\_b(358M)、vit\_l(1.16G)、vit\_h(2.39G)三个型号，要单独下载模型

**Semantic-SAM(2023)：**

github：<https://github.com/UX-Decoder/Semantic-SAM>

2023年的图像分割模型(pytorch)，在SAM模型的基础上进行了改进，提升了分割性能，并且可以更灵活的选择不同的分割程度(比如只分割出窗户还是分割出整辆车)

**SEEM(2023)：**

论文：Segment Everything Everywhere All at Once

github：<https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once>

示例网址：https://huggingface.co/spaces/xdecoder/SEEM

2023年的图像分割和识别模型(pytorch)，在对图片进行分割的同时，还会对分割的物体进行识别。分割性能相比SAM有一定的提升

**时间序列预测**

**LSTM(1997)：**

论文：LONG SHORT-TERM MEMORY

LSTM规避了标准RNN中梯度爆炸和梯度消失的问题，学习速度更快

****

其中为函数。LSTM中丢弃层不宜太高，一般为0.2；LSTM输出维度太低对结果有明显影响，可为64/128...；LSTM后加卷积可小幅度提高网络性能，但卷积不宜过多

**Informer(2020)：**

论文：Informer Beyond Efficient Transformer for Long Sequence Time Series Forecasting

github：<https://github.com/zhouhaoyi/Informer2020>

2020年提出的模型，目的是在transformer 的基础上改进，降低时间复杂、提高计算效率和提高预测能力。它提出一种新的结构：ProbeSparse自注意机制，这是一种稀疏的注意力机制，只对部分的输入使用注意力机制，并证明可以在保持性能的同时大幅减少计算量

**Autoformer(2021)：**

论文：Autoformer Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting

github：<https://github.com/thuml/Autoformer>

2021年提出的模型，目的是在transformer 的基础上改进，提高对长期时间序列预测的效果。transformer 有强大的学习能力，但在长期时间序列预测上效果不佳。而Autoformer将序列分解成两部分：周期趋势部分和偏差部分，周期趋势部分代表序列主要的趋势，偏差部分代表噪声和季节性等偏离部分

**SCINet(2021)：**

论文：Appendix: Time Series is a Special Sequence: Forecasting with Sample Convolution and Interaction

github：<https://github.com/cure-lab/SCINet>

论文在7个流行的数据集上测试，大多数情况下优于现有的时间序列预测（TSF）方法。论文证明TCN网络中的因果卷积并非必要。SCINet可以降低时间序列的排序熵，理论上更容易预测

**FEDformer(2022)：**

论文：FEDformer Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting

github：<https://github.com/MAZiqing/FEDformer>

**NLinear(2022)：**

论文：Are transformers effective for time series forecasting?

github：<https://github.com/cure-lab/LTSF-Linear>

文章证明用一些简单的全连接层其实就比几乎所有基于的transformer的模型好

NLinear主体就是全连接层，只是在全连接前/后要减去/加上序列的最后一个值，起到归一化的效果，让模型更快收敛

**自然语言处理**

**Transformer(2017)：**

论文：Attention is All You Need

github：<https://github.com/huggingface/transformers>

使用文档：<https://blog.csdn.net/zhaohongfei_358/article/details/126019181>

内部机制：<https://zhuanlan.zhihu.com/p/109484539>



当前NLP模型基本都以transformer结构展开，更多理论详见《NLP》部分

**ERNIE 3.0(2022)：**

github：<https://github.com/PaddlePaddle/PaddleNLP/tree/develop/model_zoo>

百度提出的超大语言模型，在多个数据集上均优于以往模型

**NLLB(2022)：**

github：<https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/modeling>

2022年facebook发布的语言翻译模型，能够在200种语言之间翻译

**LLaMA(2022)：**

2022年facebook发布的大型语言模型

**ChatGLM-6B(2023)：**

github：<https://github.com/THUDM/ChatGLM-6B>

2023年清华大学发布的中英双语大型对话模型

**Chinese-Vicuna(2023)：**

github：<https://github.com/Facico/Chinese-Vicuna>

2022年Meta发布LLaMa模型，有7B、13B、33B三个型号型号。此后中文团队训练出中文模型补丁，中文模型只是部分权重，要与原模型合并才能使用。由于LLaMa的使用需要申请，因此官方不提供直接合并好的中文模型

**Chinese-LLaMA-Plus/Chinese-Alpaca-Plus(2023)：**

github：<https://github.com/ymcui/Chinese-LLaMA-Alpaca>

2022年Meta发布LLaMa模型，有7B、13B、33B三个型号。此后中文团队训练出中文模型补丁，中文模型只是部分权重，要与原模型合并才能使用。Chinese-LLaMA-Plus是基础模型补丁，要与原LLaMA合并才能使用；Chinese-Alpaca-Plus是额外精调的模型补丁，要与原LLaMA、Chinese-LLaMA-Plus合并才能使用。由于LLaMa的使用需要申请，因此官方不提供直接合并好的中文模型

**Chinese-LLaMA-Alpaca-2(2023)：**

github：<https://github.com/ymcui/Chinese-LLaMA-Alpaca-2>

2023年Meta发布LLaMa2模型，有1.3B、7B、13B三个型号。此后中文团队训练出中文模型补丁，中文模型只是部分权重，要与原模型合并才能使用。由于LLaMa2完全开源，因此可以直接从官方下载合并好的中文模型

**\*vllm：**

vllm是llama等模型的加速库，可以合理利用cpu等资源实现加速

**\*llama.cpp：LLaMA的c++部署**

github：<https://github.com/ggerganov/llama.cpp>

**多模态**

**CLIP(2021)：**

论文：Learning Transferable Visual Models From Natural Language Supervision

github：<https://github.com/openai/CLIP>

2021年openai发布的基于对比学习的多模态模型(pytorch)，用于将文字描述和图片匹配。由一个图片编码模型和一个文本编码模型组成，一张图片经过图片模型得到的特征向量和这张图片的描述经过文本模型得到的特征向量会相近，通过计算余弦相似度可以通过文本找图片、通过图片找文本(官方文本模型只支持英文)

CLIP模型训练：在互联网上用4亿个图片文本对进行训练。图片送入图片模型，文本送入文本模型，将两者输出的特征向量用交叉熵计算损失，也叫对比学习

**\*粤港澳数字经济研究院太乙模型(2022)：**

国内配套的中文文本模型：<https://huggingface.co/IDEA-CCNL/Taiyi-CLIP-Roberta-large-326M-Chinese>

中文文本模型训练：用1.23亿个图片文本训练了一个中文文本模型，图片模型不变，中文文本模型以图片模型的输出特征作为标签，进行交叉熵损失计算

**\*阿里达摩院Chinese-CLIP(2022)：**

github：<https://github.com/OFA-Sys/Chinese-CLIP>

用2亿个图片文本对重新训练了clip模型，训练方法与原clip一致

**DALL-E(2021)：**

2021年openai发布的文本生成图片模型

**DALL-E2(2022)：**

github：<https://github.com/lucidrains/DALLE2-pytorch>

2021年openai发布的文本生成图片综合模型，由文本编码模型BART、图片编/解码模型VQGAN、CLIP模型组成。输入文本经过文本编码模型得到特征向量，然后经过图片解码器得到多张图片，再用CLIP模型找到最合适的图片

模型训练：训练时用的是文本编码模型和图片编码模型，对两者的输出计算交叉熵。推理时图片编码模型换成的图片解码器

**DDPM(2022)：**

github：<https://github.com/hojonathanho/diffusion>

**Imagen(2022)：**

github：<https://github.com/lucidrains/imagen-pytorch>

2022年google发布的文本生成图片模型(pytorch)(1G)。由生成模型DDPM级联组成

**Stable Diffusion(2022)：**

github：<https://github.com/Stability-AI/StableDiffusion>

提示词网站CIVIT：<https://civitai.com/?tag=base+model>

各平台翻译API：<https://bobtranslate.com/service/translate/volcengine.html>

由clip文本模型+unet去噪生成模型+vae模型+各类风格peft模型组成。以2.1版本为例官方提供了512\*512和768\*768两种模型，每种模型都有剪枝、ema+剪枝两个版本，并且提供了.ckpt和.safetensors两种模型，safetensors是一种更安全的格式

扩散(diffusion)：图片的加噪或去噪

**RAM(2023)：**

github：<https://github.com/xinyu1205/recognize-anything>

2023年的图像识别模型(pytorch)——高级图片分类模型。在clip的基础上进行了改进，输入一张图片后，可以识别图片中常见的所有物体的名称，能够应用到图片的数据标注中

**GroundingDINO(2023)：**

github：<https://github.com/IDEA-Research/GroundingDINO>

2023年的文字提示目标检测模型(pytorch)——高级目标检测模型。输入一张图后可以通过文本选择性的检测图片中的物体，比如：“左边的人”；还可以结合文生图模型做到“将左边的猫换成狗”

**3D**

**PointNet(2016)**

PointNet是斯垣福大学在2016年提出的一种点云分类/分割深度学习框架。刚提出的原始模型有些冗余的地方，后来人们提出了改进版PointNet++

**NeuralRecon(2021)**

github：https://github.com/zju3dv/NeuralRecon

NeuralRecon是商汤2021年提出的单摄像头三维重建模型。三维重建坐标合成阶段会非常消耗性能，但NeuralRecon没有很复杂的结构速度很快，还能取得不错的效果

模型的输入是一个物体不同视角的9张图片

模型要先根据摄像头中的参数将图片中的像素对应到重建后的结果上

**GrabCAD**

网址：<https://grabcad.com/engineers>

GrabCAD 是全球工程师的CAD 文件共享平台，有400万个可免费下载的CAD文件。但文件格式有近10种：stl、3d object、dxf、c4d、step、stp...

**OpenUSD**

github：<https://github.com/PixarAnimationStudios/OpenUSD>

通用场景描述USD是一个开源的软件，可以满足3D场景需求

**三维重建**

传统方法：深度估计、扫描点云、坐标融合

神经网络方法：通过神经网络端到端的预测

**原理：**

[图片坐标系]-->[拍摄坐标系]-->[重建坐标系]

图片坐标系：拍摄时物体会通过小孔映射到摄像机内部形成图片，二维坐标系

拍摄坐标系：拍摄的时候以相机为中心建立的三维坐标系

重建坐标系：自定义的三维坐标系，需要先构建一个空白的立方体(volume)，再不断填充立方体的体积元素(voxel)。立方体越大、体积元素越小则重建模型越细致，但耗性能

[图片坐标系]-->[拍摄坐标系]：经过一个变换矩阵，需要摄像机的内参

[拍摄坐标系]-->[重建坐标系]：经过一个变换矩阵，由旋转矩阵和平移矩阵组成，需要摄像机的外参。旋转矩阵：仿射变换，形状为3\*3。位置矩阵：形状为3\*1，代表在x、y、z上的平移量

相机标定：从图片坐标系到重建坐标系需要摄像机的内外参数，反之如果知道部分点在图片坐标系和重建坐标系的位置就可以求出摄像机的内外参数

流程：

[关键帧选择]-->[输入局部的图]-->[片段重建]-->[全局合成]

关键帧选择：比如要重建一个物体，只需要几个方位的图片

输入局部的图：如果要对一个房间重建，需要先重建房间中的各个物体

片段重建：根据输入的图在重建坐标系的空白立方体中生成相应的体积元素

全局合成：不断地片段重建直到最终填满空白立方体

**点云数据**

最终形式：x、y、z、其他信息

可视化软件：cloudcompare

**特点：**

点云数据为非结构化数据，输入点的数量是不定的、顺序是无顺序的，一个物体可由上万个点来表示。模型输入需要具有置换不变性

点云数据可以通过激光雷达扫描得到，根据扫描的视角不同，点会有近密远疏的特点，需要进行点云补全，点云补全可以通过点云配准，也可以通过GNN网络生成

点云配准：将不同视角下扫描得到的点云进行对齐，合成一组点云数据

数据集格式举例：

**|数据集：**

|类别1：

|类别2：

|A物体的点云数据

|B物体的点云数据

|...

**存储格式：**

.las、.laz：LiDAR数据的工业标准格式，二进制文件格式

.pcd：PCL库官方指定格式

.obj：从几何学上定义的3D模型文件格式，文本文件

.pcap：Velodyne公司激光雷达的默认采集数据的二进制文件格式

.ply：斯坦福大学设计开发的多边形文件格式

.pts：简便的点云格式，文本格式。只包含点坐标信息，按x、y、z顺序存储

.xyz、.txt：文本格式

**其他项目**

**PaddleSlim()：飞桨模型量化**

github：https://github.com/PaddlePaddle/PaddleSlim

**FastDeploy()：飞桨模型部署开发套件**

github：<https://github.com/PaddlePaddle/FastDeploy>

**PaddleNLP：飞桨自然语言处库**

github：<https://github.com/PaddlePaddle/PaddleNLP>

文本分类：<https://github.com/PaddlePaddle/PaddleNLP/tree/develop/applications/text_classification>

**timm：集成了很多pytorch的图像模型**

github：<https://github.com/rwightman/pytorch-image-models>

**transformers：集成了很多pytorch的自然语言处理模型**

github：<https://github.com/huggingface/transformers>

**Pycorrector：中文文本纠错库**

github：<https://github.com/shibing624/pycorrector>

默认模型Kenlm(2.8G)，库中提供了其他模型，推荐MacBert4csc(420M)。

**Fiftyone：目标检测数据集可视化工具**

官方教程：https://docs.voxel51.com/index.html

可以选择本地数据(图片分类、目标检测)并后以浏览器的方式查看，会自动去除空标注的图片和找不到图片的标注，每次加载数据后会留下记录到本地，下次就可以直接查看。需要pycocotools库，最好新建一个单独的环境

**doccano：文本分类标注工具**

github：<https://github.com/doccano/doccano>

**其他github作者：**

CV：<https://github.com/bubbliiiing>

NLP：<https://github.com/shibing624>

封神榜大模型：<https://github.com/IDEA-CCNL/Fengshenbang-LM>