第一讲：导论

基本概念

- 什么是计算机视觉？什么是图像处理？两者的区别是什么？
  1. 计算机视觉： 研究如何让计算机从数字图像中获取深层次理解的技术，将输入的图像转化为对现实世界的理解，并作出相应反馈的技术。
  2. 图像处理：主要有图像增强和图像分析两个方面，包括对图片的去噪，锐化，减轻或者消除图像的退化效果等。
  3. 区别：图像处理是图像到图像的映射，其评价是主观的。计算机视觉是图像到知识之前的映射，其评价是客观的。图像处理是计算机视觉的基础。

- 计算机视觉的难点是什么？
  1. 逆问题：图像是三维世界在二维平面上的投影，而根据投影反求出三维空间信息是难解的问题。

     解决办法：利用先验知识对问题添加约束。
  2. 语义：从图像到语义概念的映射，这个过程是未知、极其复杂的。

     解决办法：数据驱动（模式识别、机器学习、大数据）

- **计算机视觉问题的解决方法**
  - 基于模型的方法（自顶向下）
    - 构造图像的生成模型 `X = G(Y)`
    - 从`X`反算出`Y`. (逆问题): 找到合适的`Y`使得`G(Y)=X`

$$X = G(Y;\theta) \rightarrow Y = argmin_Y E(X, G(Y))$$
$$P(X|Y;\theta) \rightarrow Y = argmax_Y P(Y|X;\theta) = argmax_Y P(Y)(X|Y;\theta)$$

  - 自底向上的方法（数据驱动、手工设计）
    - 手工设计：设计出从X变换到Y的计算过程

$$Y = F(X;\theta) = f_k(f_{k-1}(\dots f_1(x)))$$

    - 数据驱动：用函数`h(x; θ)`近似表示变换`F`，用样本训练函数的参数

      收集样本

$$D = \{X^{(i)}, Y^{(i)}\}_{i=1}^N$$

      学习参数

$$\theta^* = argmin_\theta Loss(\{X, Y\})$$

      推断类别

$$\hat{y} = h(x;\theta^*)$$

第二讲：图像识别-分类器与特征

无参模型：KNN方法

- 基本原理：用欧式距离计算图像向量之间的距离
- 存在的问题：
  - 图像对应位置的像素值可能不具有可比性
  - 像素空间的欧氏距离与语义空间的类别差异之间存在鸿沟
  - 无参方法，需要保存所有训练样本，存储开销较大

有参模型：线性分类器

- Softmax Regression
  - 对任意输入x，k-类线性分类器给出k个分值：

$$\boldsymbol{z} = z_1, z_2, \cdots z_k, z_j = w_j^T x + b_j$$
$$\hat{y} = argmax[\boldsymbol{z}]$$
$$\rho = (\rho_1, \rho_2, \ldots, \rho_k) = softmax(\boldsymbol{z})$$
$$\rho_j = \frac{e^{z_j}}{\sum_{i=1}^{k} e^{z_i}}, \sum_{j=1}^{k} \rho_j = 1$$

-
- Logistic Regression

  Softmax Regression 对应二分类的情况
- 交叉熵损失(Cross Entropy Loss)

  假设p，q是定义在$Y = \{1,2, \ldots, k\}$上的两个离散概率分布

  p，q之间的交叉熵为

$$H(p,q) = -\sum_{y \in Y} p(y) lnq(y)$$

例题

一、Logistic Regression 如下式定义：

$$\rho = g(z; W; b) = \frac{1}{1 + e^{-}(W^T x + b)}$$

采用交叉熵损失$l(w,b;x,y)$训练W，b。

请写出损失函数的表达式，并推导出损失函数对参数W和b的偏导数。

第三讲：图像识别-无监督特征学习

词袋模型

- 什么是视觉单词?
  - 一组重复出现的相似的局部图像块
  - 一个视觉单词在不同图像/不同位置，外观有所不同
  - 一组外观相似的局部图像块的均值可以作为视觉单词
- 视觉词袋模型如何表示图像?
  - 提取特征
  - 从词典中找最近邻的单词
  - 构建词频直方图
- 如何获取视觉词典?
- 词袋模型的特点

卷积运算

- 卷积运算及其性质
  - 可交换：$f * h = h * f$
  - 结合性： $f * g * h = f * g * h$
    - 把大的卷积核分解为多个小卷积核的卷积，通过多次卷 积减少计算量
    - 多次卷积带来多次访问内存操作，增加了一定的开销
  - 对加法的分布律：$f * (g + h) = f * g + f * h$
  - 可分离卷积核
    - 对可分离卷积核，用两个一维卷积可以提高计算效率

- 卷积运算的相关概念：跨度(stride)，填充(Padding),卷积核尺寸
- 理解卷积参数、图像大小与特征图尺寸的关系
- 理解卷积与特征提取的关系

卷积特征

- 图像卷积特征的一般形式（Encoding+Pooling）
- Encoding:如何表示一个图像块？（Kmeans,AutoEncoder,...）
- Deep AutoEncoder

例题

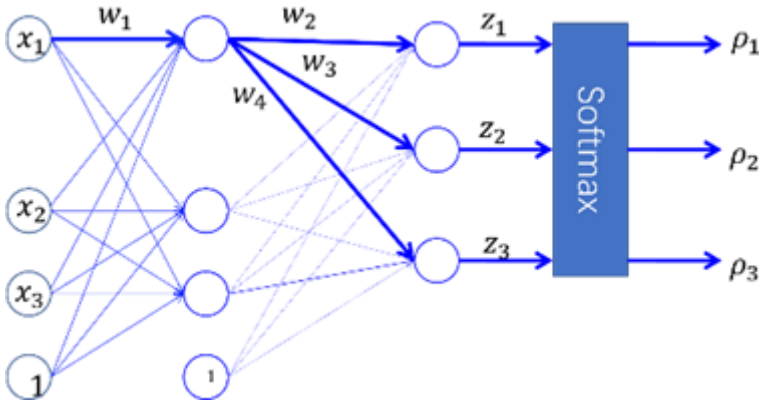
一、使用BoW计算图像特征，假设每一个视觉单词向量的长度为64，词典大小为100，图像中提取到了217个单词，请回答：

（1）采用上述词典描述这幅图像，特征向量的长度是多少？

（2）采用Softmax分类器对图像进行分类，假设类别总数为10，该分类器的参数数目是多少（包含偏置参数）？

（3）x表示图像的特征向量(列向量)，分类器的权重矩阵为W,偏置向量为b，请写出分类器的表达式。

二、下面是一个单隐层神经网络，输出层是一个3-Way Softmax，假设某个样本的真实类别的One-hot向量为`(0，0，1)`，采用交叉熵损失，激活函数使用 sigmoid 函数，请用BP算法，推导出该样本上的损失`l`相对于图中参数`w1`的偏导数。图中`w1，w2，w3，w4`表示对应连接上的权值参数，`z1，z2，z3`表示输出层神经元的净响应。



三、 假设某个pytorch编写的网络模型net如下所示，请回答后面的问题。

```
>>> print(net)
Sequential(
 (0): Conv2d(3, 16, kernel_size=(5, 5), stride=(1, 1) , padding=0)
 (1): ReLU()
 (2): MaxPool2d(kernel_size=3, stride=3, padding=0)
 (3): Conv2d(16, 32, kernel_size=(5, 5), stride=(1, 1) , padding=0)
 (4): ReLU()
 (5): MaxPool2d(kernel_size=3, stride=3, padding=0)
)
```

1） 该网络的输入图像的通道数是多少?

2） 计算第二个卷积层的参数数目（包含bias）。

3） 假设该网络输入图像大小为34×34，请计算上述每一个卷积层和Pooling层输出的特征图的大小。

4） 如果网络输入图像大小为34×34，计算第二个卷积层的乘法运算量。

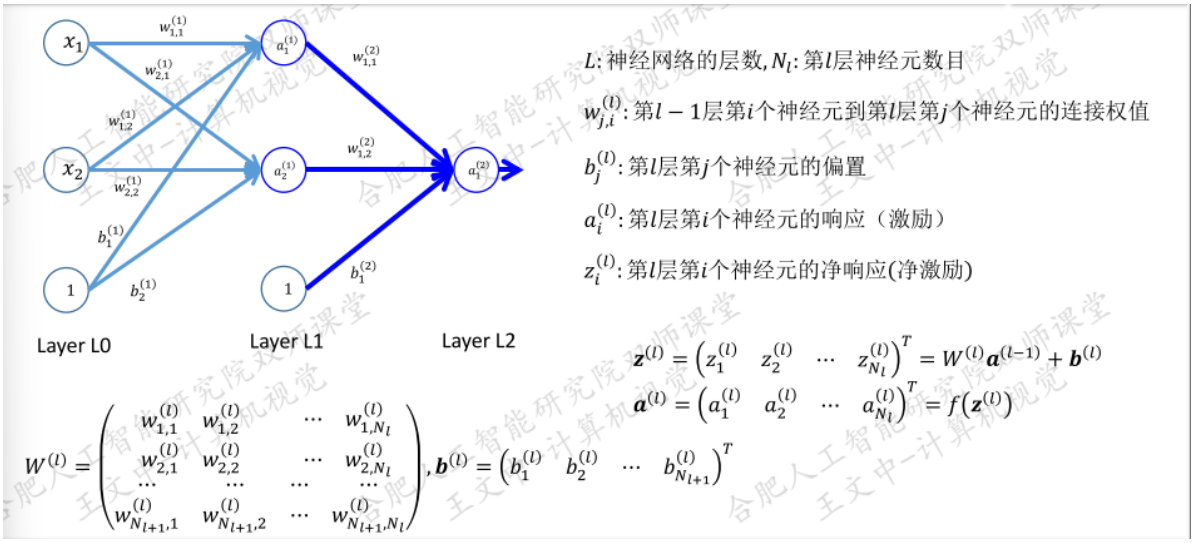
5） 计算第二个卷积层特征图上的神经元在输入图像上的等效感受野的大小。

第四讲：图像分类-端到端学习与卷积神经网络

神经元模型

- 线性聚合，非线性变换

前馈神经网络



`h(x) = ?`

- 多层感知器的结构
- 多层次非线性变换，特征学习，深度学习

BP算法

- BP算法原理
- 对于小规模神经网络，可以手工推导梯度

卷积神经网络

- 卷积网络相对于MLP的优势
- 卷积网络的工作原理（多层特征检测与复合）
- 如何设计一个卷积网络 - 确定结构参数
  - 卷积层(CONV)
    - 每一层卷积核的数目n（确定了该层输出的特征图的通道数目）
    - 每一层卷积核的大小 f
    - 每一层卷积的跨度s
    - 每一层卷积的非线性响应函数（ReLU）
  - Pooling层(POOL)
    - Pooling区域的大小 f
    - Pooling的计算方式（Max,Mean,P-Norm）
    - Pooling的跨度s
  - 全连接层(FC)
    - MLP的层数n
    - 每一层神经元数目 f 与响应函数

- 从低层到高层，特征图的通道数目通常越来越多
- 低层神经元感受野比较小（提取局部特征）、高层神经元的感受野越来越大（提取全局特征）
- Pooling会造成特征定位不准确

例题

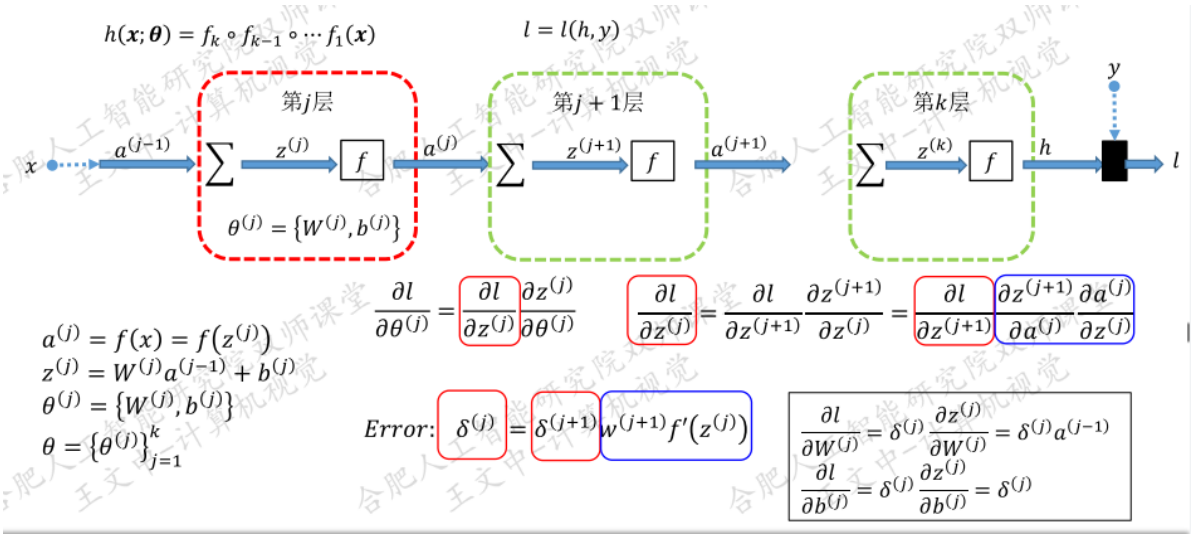
一、神经网络 $h(x;\ \theta)$，在样本 $x,\ y$ 上的损失为 $L(h(x;\ \theta),\ y)$，请写出二分类（交叉熵损失）、多类别分类（交叉熵损失）、回归（平方损失）下损失函数的表达式。

$$\nabla l(\theta)=\frac{\partial l(\theta)}{\partial \theta}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l\left(\theta;\boldsymbol{x}^{(i)},y^{(i)}\right)}{\partial \theta}$$

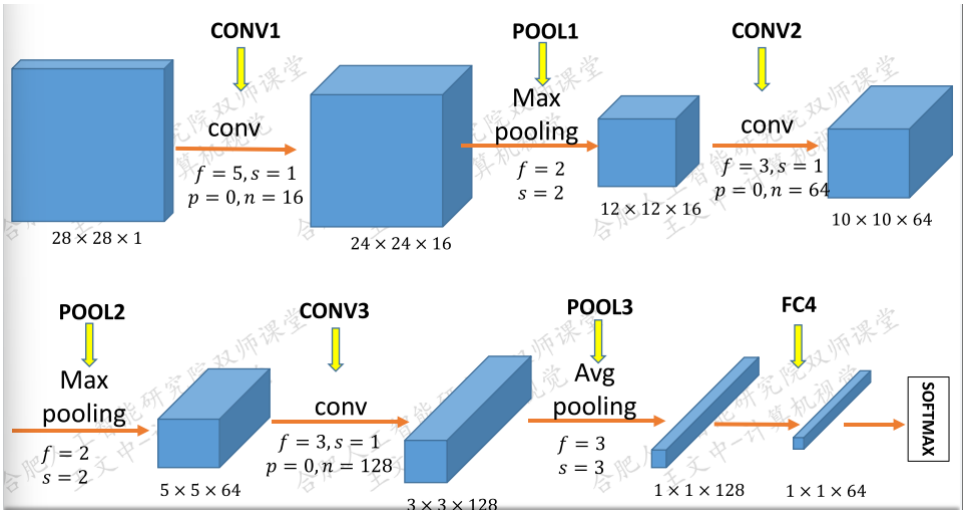
$$\frac{\partial l(\theta;\boldsymbol{x},y)}{\partial \theta}=\frac{\partial l(\theta;\boldsymbol{x},y)}{\partial h(\boldsymbol{x})}\frac{\partial h(\boldsymbol{x};\theta)}{\partial \theta}$$

$$l(\theta;\boldsymbol{x},y)=-y ln\left(h(\boldsymbol{x};\theta)\right)-(1-y)ln\left(1-h(\boldsymbol{x};\theta)\right):\quad \frac{\partial l(\theta;\boldsymbol{x},y)}{\partial h(\boldsymbol{x})}=\frac{h(\boldsymbol{x})-y}{h(\boldsymbol{x})\left(1-h(\boldsymbol{x})\right)}$$

$$l(\theta;\boldsymbol{x},y)=\frac{1}{2}(h(\boldsymbol{x};\theta)-y)^2:\quad \frac{\partial l(\theta;\boldsymbol{x},y)}{\partial h(\boldsymbol{x})}=h(\boldsymbol{x})-y$$



二、结合如下卷积神经网络的结构，填写表格



Layer	Input	Kernel	S/P	#Kernel	Output	#Params	#OPS
CONV1							
POOL1							
CONV2							
POOL2							
CONV3							
POOL3							
FC4							
SOFTMAX							

三、设计用于输入为 32 * 32 * 1 的手写体识别的卷积神经网络，并给出每层的参数个数。

第五讲：PyTorch

Sequential & Module

例题

一、CIFAR10 的卷积、池化以及全连接操作定义如下

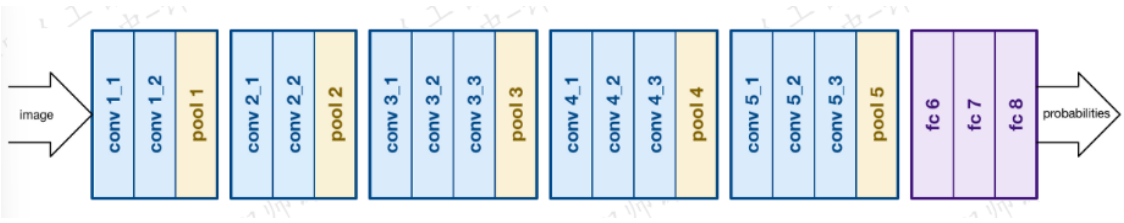
```
self.conv1 = nn.Conv2d(in_channels=3, out_channels=16, kernel_size=3)
self.pool1 = nn.MaxPool2d(kernel_size=2, stride=2)
self.conv2 = nn.Conv2d(in_channels=16, out_channels=32, kernel_size=5)
self.pool2 = nn.MaxPool2d(kernel_size=3, stride=3) self.conv3 =
nn.Conv2d(in_channels=32, out_channels=128, kernel_size=3)
self.fc1 = nn.Linear(128, 128)
self.fc2 = nn.Linear(128, 64)
self.fc3 = nn.Linear(64, 10)
```

请画出卷积神经网络，并给出每层的参数个数。

第六讲：图像分类-现代深层CNN，人脸识别

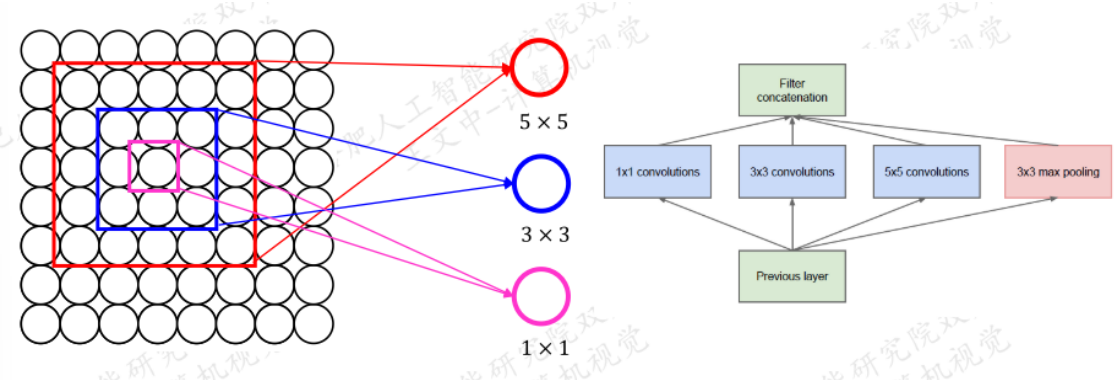
几种典型的DCNN的结构及其设计思想

- VGG



  - 为什么使用两个小的卷积核代替一个大的卷积核?

- GoogLeNet



    - 卷积核的参数数目是多少？给出计算过程
    - 为什么要使用1 × 1的卷积核?

- ResNet

DCNN的训练

- Batch Normalization
- DropOut
- Data Augmentation
- Transfer Learning

人脸识别

- 一般流程：人脸检测 -> CNN -> Result
- Siamese网络
    - 过程
    - 损失函数是什么?
- 人脸比对：人脸特征编码，特征比对
    - TripleLoss
    - Contrastive Loss
    - Pair Classification

例题

一、如下表格

y	**1**	**1**	**0**	**1**	**0**	**1**	**1**
y*	1	1	1	0	0	1	1
TP	1	1	0	0	0	1	1
FP	0	0	1	0	0	0	0

求 tpr 和 fpr 的值

第七讲：从理论到实践

机器学习的基本概念

- 数据、算法、假设空间
  - P(x, y) = P(x)P(y | x)、y = f(x)

$$|err(h) - err(h*)| < \epsilon \Rightarrow h\text{近似正确（}PAC\text{）}$$

  - H 空间太复杂导致过拟合，则可以增加训练样本数量
  - H 较为简单则方差低、偏差较大、容易欠拟合

    H 较为复杂则方差高、偏差较小、容易过拟合

    随着 H 的复杂的的增加，训练误差持续下降，测试误差先降后升
- 过拟合
  - 检查验证集数与训练集数据是否满足独立同分布假设（不满足：更新数据集）
  - 训练样本数量：收集更多样本、样本增广
  - 减少模型复杂度：增加正则化因子、减少可训练网络层数
  - 采用Bagging集成学习（减少方差）
- 欠拟合
  - 增加模型复杂度（比如减少正则化、 增加网络层数）

图像分类实践

- 开发、诊断和调试图像分类模型
- 损失函数不下降
  - 调节学习率
  - 更换优化器
  - 检查数据是否有问题

第八讲：目标检测

目标检测问题的定义

- 目标定位，目标检测
- 目标数量不确定 —> 输出Y的维度不确定

滑动窗方法及其问题

- 需要穷举所有的可能的未知和大小，而绝大多数窗口中都不包含目标，这样会大大增加了计算量和误检率

Viola-Jones人脸检测算法

- Harr小波特征
- AdaBoost分类器

目标包围盒重叠程度评价IOU

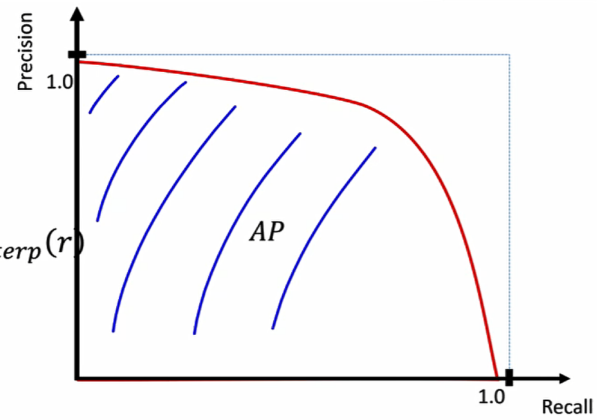
- 0 <= IOD(A, B) <= 1，A, B 之间相交区域大小比并集大小

非极大值抑制NMS

性能评价

- FP, TP, FN, TPR, FPR, AP
  - 某图像中有100个人脸，检测到了80个结果，其中有75个是人脸，则TP, FP, R, P的值分别为?
  - 注意区分R和P
- AP与nms阈值以及得分阈值的关系

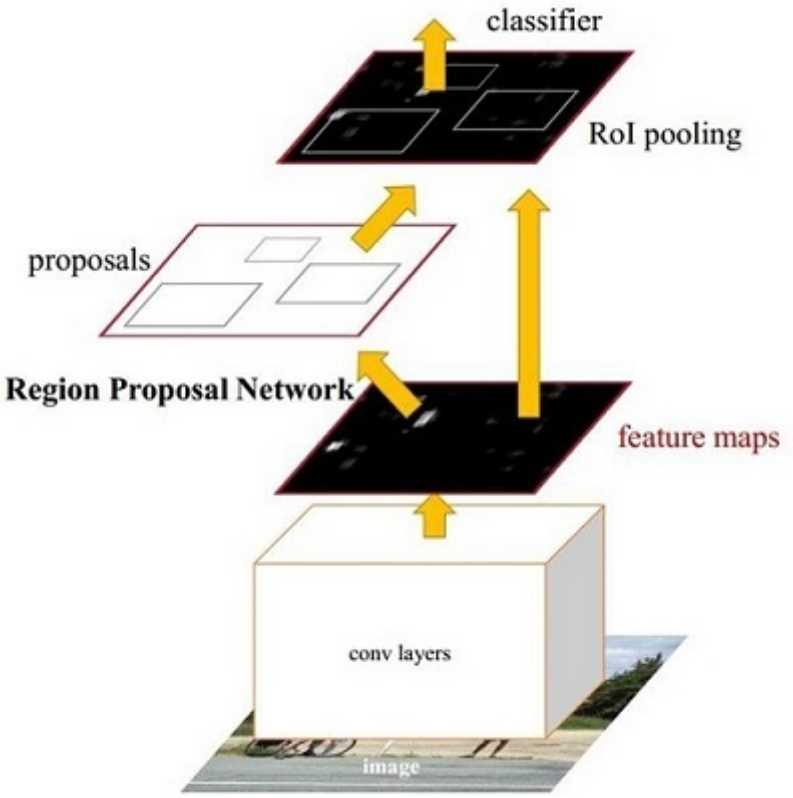
- AP (Average Precision)
  - $AP = \int_0^1 p(r)dr$
  - P-R曲线下方的面积
- PASCAL VOC:
  - $AP = \frac{1}{11}\sum_{r\in\{0,0.1,0.2,\cdots 1.0\}} p_{interp}(r)$
  - $p_{interp}(r) = max_{\tilde{r}\geq r} p(\tilde{r})$



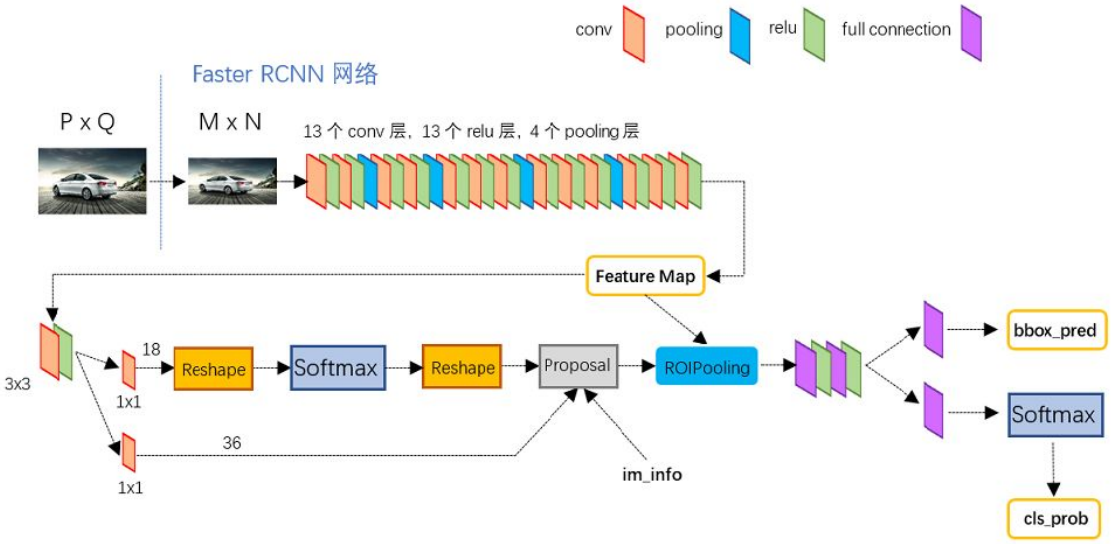
R-CNN

- 二阶段
- Anchor
- RPN
- Box Regression

- Faster R-CNN (https://zhuanlan.zhihu.com/p/31426458)

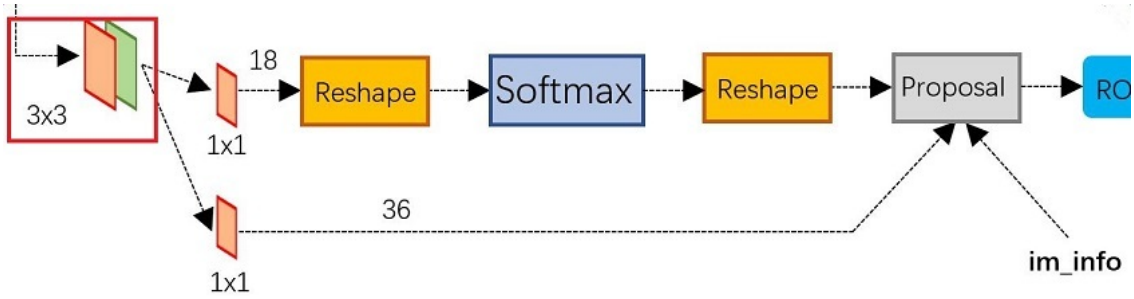


图片大小为 800 * 600，使用VGG-16，详细如下



  - 所有的conv层都是：kernel_size=3，pad=1，stride=1
  - 所有的pooling层都是：kernel_size=2，pad=0，stride=2

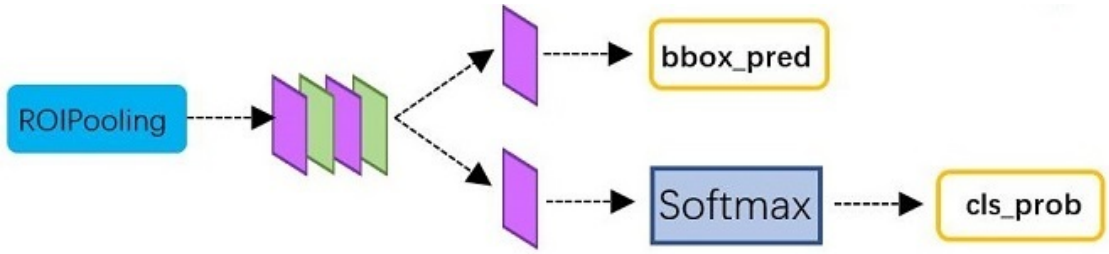
Feature Maps 通过RPN生成检测框，如下为RPN网络结构



  - RPN 输出的 Box 数目必须是固定的

- 每个预测的Box可以计算损失（需要有真值Box）

分类部分网络结构图



- 通过全连接和softmax对proposals进行分类
- 对proposals进行bounding box regression，获取更高精度的rect box

回答下面问题：

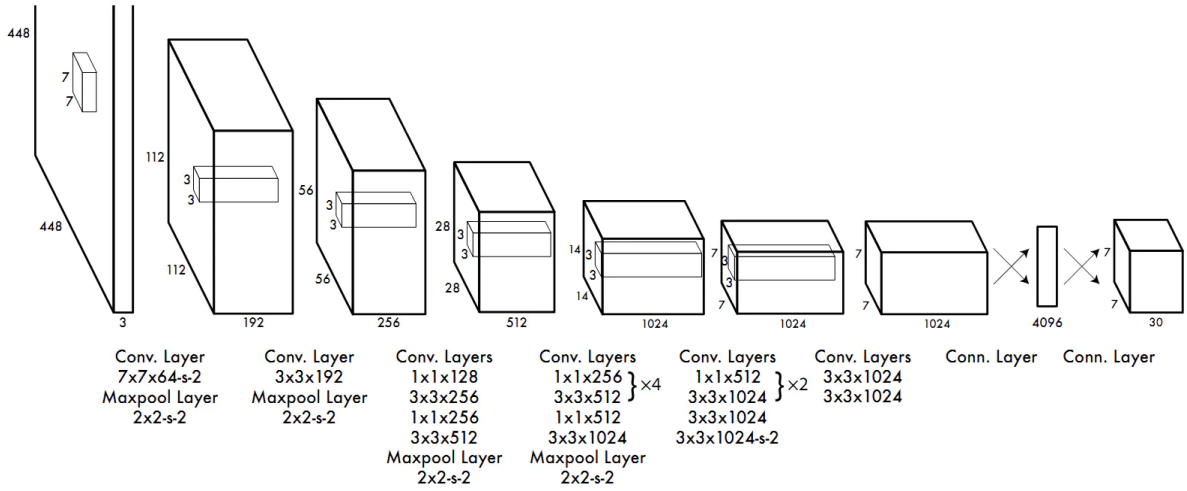
1. 输出的Feature Maps大小和通道数为？
2. 特征图上每个像素点预测 k = 9 个不同的候选区域，则共有多少个候选区域？
3. Anchor Box $a = (x_a, y_a, w_a, h_a)$、预测的Box坐标 (x, y, w, h)、真实的Box坐标 (x^*, y^*, w^*, h^*)。

   请给出损失函数 L 的定义。
4. RoIPooling 获取到的 proposal feature maps 大小为多少？后续如何做Classification？

YOLO

YOLO V4：https://arxiv.org/pdf/2004.10934.pdf

YOLO V1：https://zhuanlan.zhihu.com/p/32525231

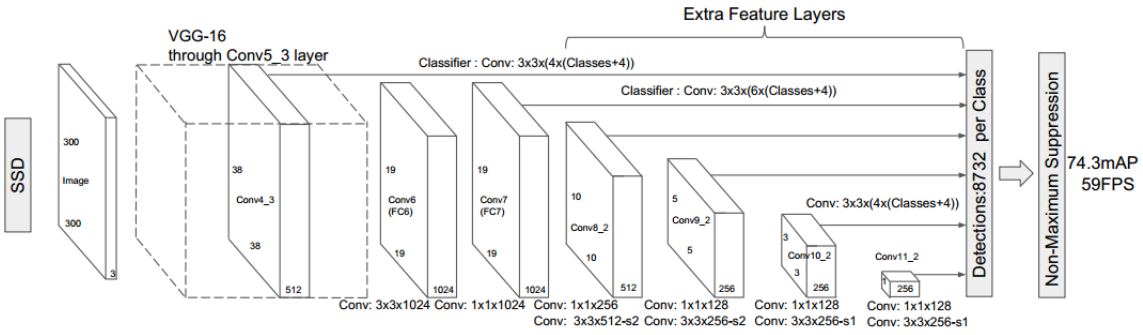


回答下面问题：

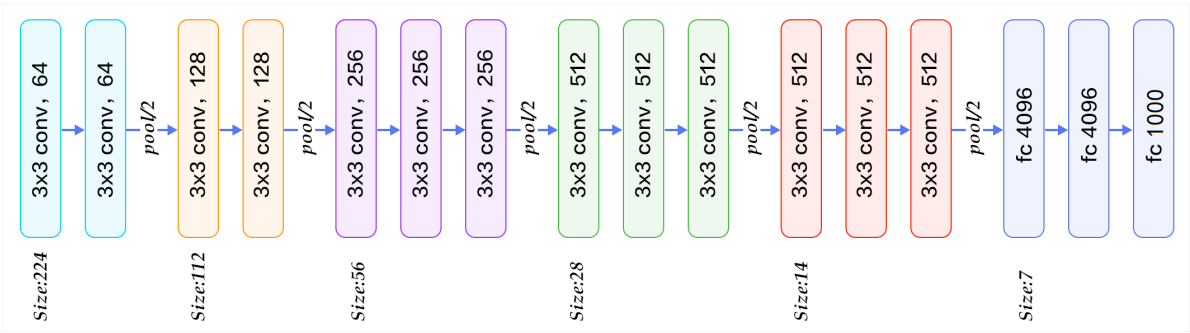
1. 每个网格最多预测两个目标，则最多可以检测出多个目标?
2. 输出 $7*7*30$ 的张量，其具体含义是什么?
3. 小尺度目标检测存在什么问题?

SSD

https://zhuanlan.zhihu.com/p/121117059



VGG-16



请回答下面问题

1. SSD 对 VGG-16做了哪些改动?
2. 用于做预测的 Feature Map 有哪些? Feature Map尺寸与检测目标大小有何关系?
3. 特征图的每一个位置分配 k 个不同长宽比的参考包围盒，对于每一个参考包围盒需要预测出 c 个类别的概率，则特征图上的每一个位置预测多少个值?

第九讲：图像分割

问题定义

传统分割方法

- 阈值分割（像素分类）
  - 大津阈值估计算法（OTSU）

    任取一个阈值 $T \in [0, 255]$，把图像分成两类：C_1, C_2

    最优阈值 T：使分割后两类像素间的方差最大
  - 可变阈值 — 游程平均

$$z_i = \frac{1}{n} \sum_{j=i-n+1}^{i} x_j$$

- 像素聚类
  - SLIC、超像素聚类
- 条件随机场模型（单点势函数与成对势函数的含义）
  - 马尔可夫随机场
    - Gibbs分布
    - 两个势函数

DCNN图像分割

- 如何解决分辨率的问题?
  - 上采样
    - Unpooling
    - 双线性插值
    - 反卷积（转置卷积）

      1. stride 分别为 1、2 的时候下式的值分别为多少

$$\begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} TransposedConv \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} = ?$$

      2. 原图大小为 7 * 7，若卷积的参数为：K = 2，Padding = 0，Stride = 2, 则反卷积的参数为？写出计算过程。

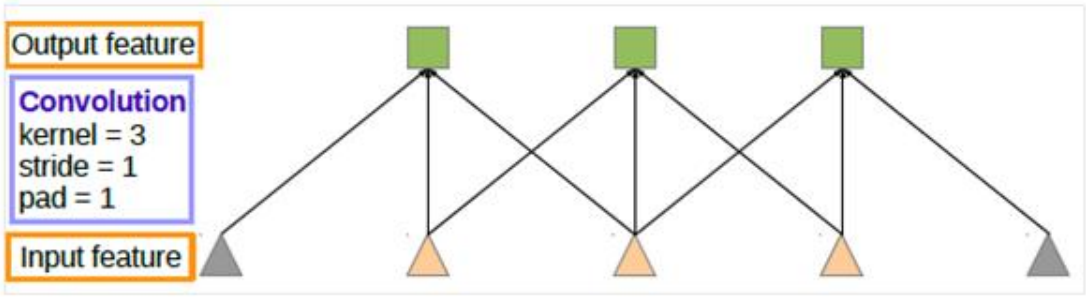
- Encoder-Decoder框架
- UnPooling，Transposed Conv
- Astrous Conv (空洞卷积、膨胀卷积)

  https://www.zhihu.com/question/54149221

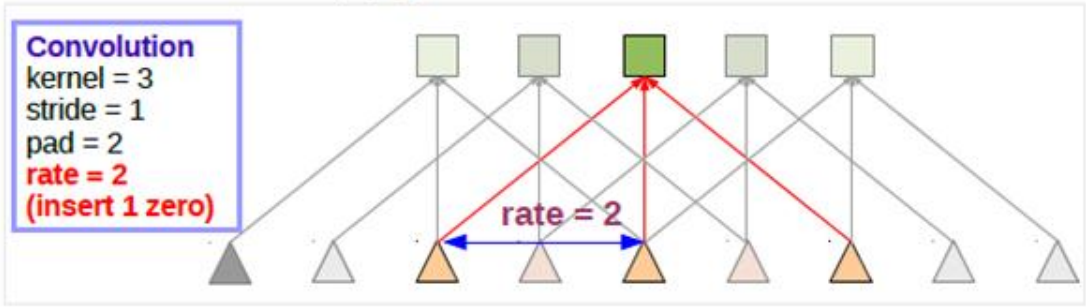
$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k]$$

当r=1时，它就是我们常用的标准卷积

当r>1时即atrous convolution，r表示卷积过程中对输入样例采样的步长



(a) Sparse feature extraction



(b) Dense feature extraction

(a) 是标准卷积，（b）是atrous convolution。

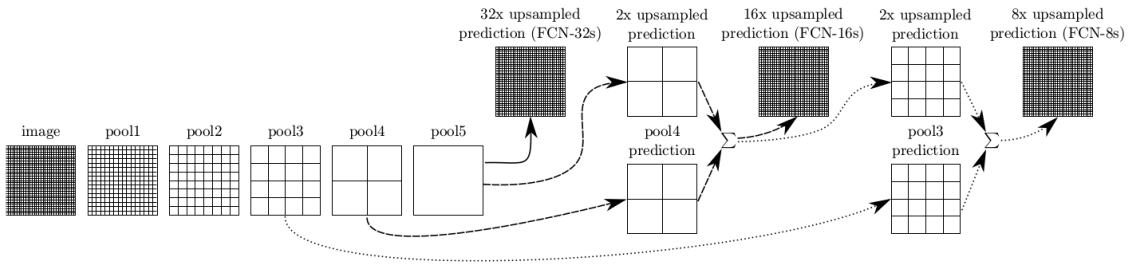
当rate=2时，输入信号被交替采样（sampled alternatively）。首先，pad=2意味着我们在左右两边填充2个0。在rate=2的情况下，我们每2个输入采样输入信号进行卷积。因此我们得到5个输出，这使得输出的特征图变大。

- 实例分割：Region Proposal + FCN

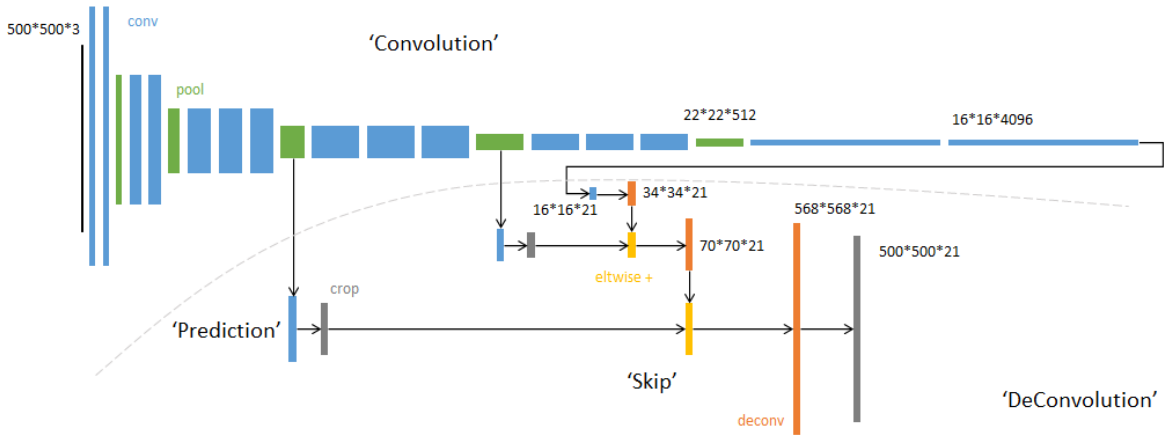
  Image ——> Faster R-CNN ——> FCN

例题

一、FCN



网络结构如下

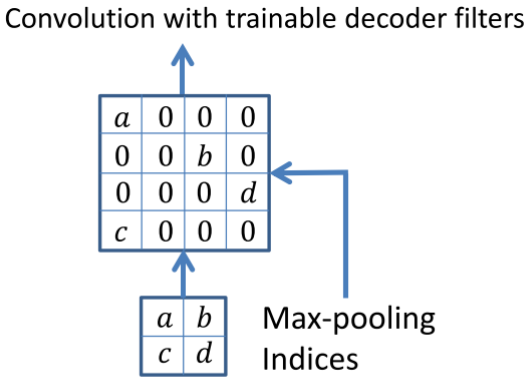


虚线下半部分中，分别从卷积网络的不同阶段，以卷积层（1×1×3）预测深度为21的分类结果。

请回答：

1. 为什么输出的深度为21？
2. 图中做了3次反卷积，每次的步长分别为32、16、8，则卷积核大小分别为？
3. 说明网络结构中跳级结构（Skip）的作用。

二、SegNet(Encoder-Decoder)



SegNet

1. 写出 Encoder-Decoder 的模型结构。
2. 写出上图的 pooling indices。

第十讲：目标跟踪

问题定义

- 目标描述模型（表观模型）
  - 包围盒
- 目标运动模型

  交替进行预测+矫正：

  1. Prediction：$\hat{X}_{t+1} = F(X_t)$
  2. Correction：$X_{t+1} = H(\hat{X}_{t+1}, O_{t+1})$

贝叶斯滤波

- Kalman Filter
  - 预测+矫正 的公式?
- Particle Filter
  - 带权重的粒子

MeanShift Tracking

Correlation Filter

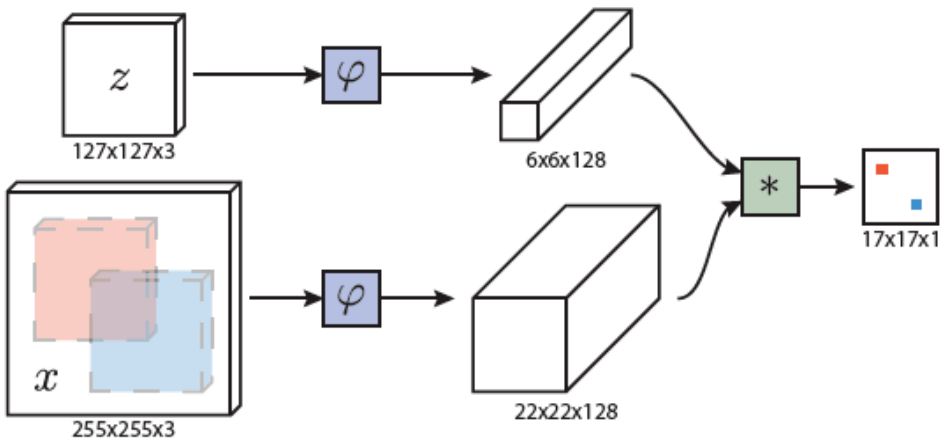
- 设计一个滤波模板，利用该模板与目标候选区域做相关运算，最大输出响应的位置即为当前帧的目标位置。
- MOSSE 以最小化平方和误差为目标函数

DCNN Tracking

- Siamese Tracktor

习题

一、孪生网络（Sinmese Tracktor）提取z和x特征之后，送到相似度函数里计算一下相似度。本文的相似度函数是使用交叉相关。



网络最后的输出，相当于一个判别式方法，用正负样本对来训练网络。搜索图片x中的每一个候选子窗口，其实相当于一个样本，而它的得分，输出的就是它是正/负样本的概率。使用逻辑回归来表示。

1. 写出本文相似度函数。
2. 写出逻辑损失表达式。
3. 写出训练的时候网络的损失函数。

第十一讲：前沿

Generative Adersarial Network

- 一般原理及其训练算法
    - 基本流程
        - 初始化判别器D的参数 θ_d 和生成器G的参数 θ_g 。
        - 从真实样本中采样 m 个样本 { $x^1, x^2, \ldots x^m$ } ，从先验分布噪声中采样 m 个噪声样本 { $z^1, z^2, \ldots, z^m$ } 并通过生成器获取 m 个生成样本 { $\tilde{x}^1, \tilde{x}^2, \ldots, \tilde{x}^m$ } 。固定生成器G，训练判别器D尽可能好地准确判别真实样本和生成样本，尽可能大地区分正确样本和生成的样本。
        - **循环k次更新判别器之后，使用较小的学习率来更新一次生成器的参数**，训练生成器使其尽可能能够减小生成样本与真实样本之间的差距，也相当于尽量使得判别器判别错误。
        - 多次更新迭代之后，最终理想情况是使得判别器判别不出样本来自于生成器的输出还是真实的输出。亦即最终样本判别概率均为0.5。
- GAN的各种应用

Image Captioning

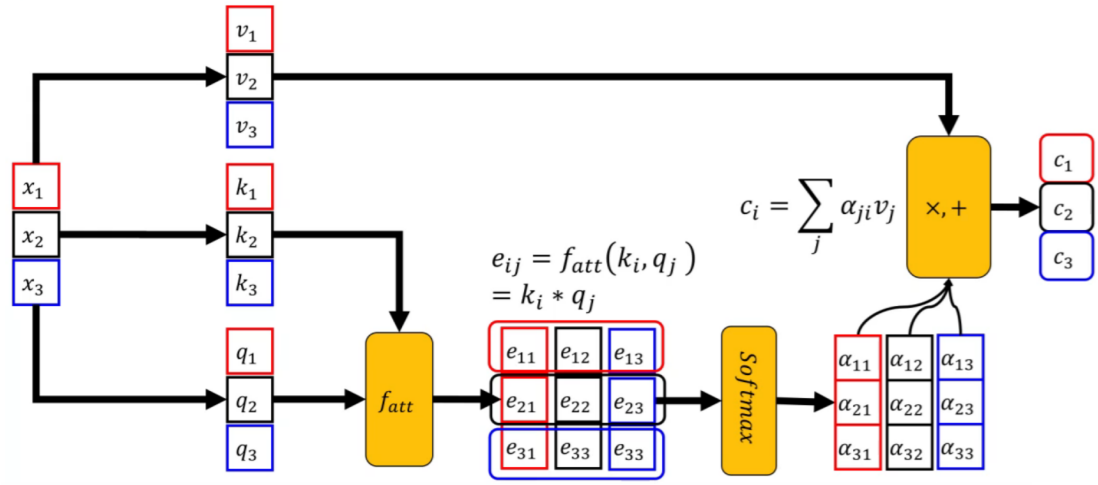
- RNN

Self-Supervised Learning

- 一般原理
- Learning from Image Transformation
- Contrastive Learning

Attention

- Recurrent Attention Model
- Soft Attention
- Self-Attention

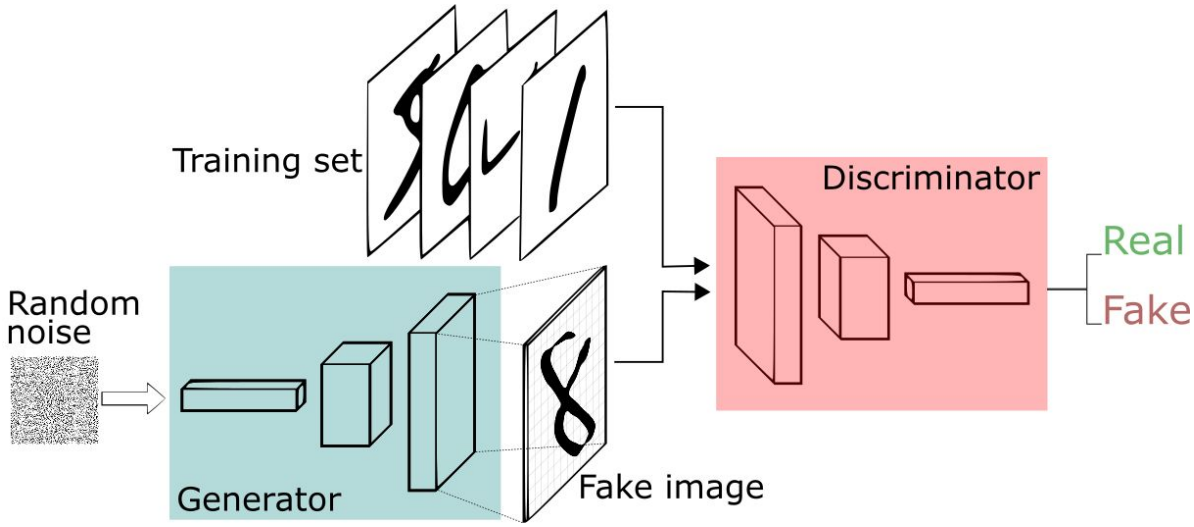


  - Softmax
  - c_i 表示 x_i 对其他 x 观察后生成的，其的生成过程为?
- Transformer
  - Encoder、Decoder

习题

一、生成式对抗网络

我们现在拥有大量的手写数字的数据集，我们希望通过GAN生成一些能够以假乱真的手写字图片。主要由如下两个部分组成：



1. 定义一个模型来作为生成器（Generator），能够输入一个向量，输出手写数字大小的像素图像。
2. 定义一个分类器来作为判别器（Discriminator）用来判别图片是真的还是假的（或者说是来自数据集中的还是生成器中生成的），输入为手写图片，输出为判别图片的标签。

回答下面问题

1. 写出交叉熵损失函数的基本形式。
2. 写出GAN损失函数（Generator & Discriminator）及GAN的训练过程（公式表达）。