条件随机场结合深度学习

目录

目录

1.1 概述

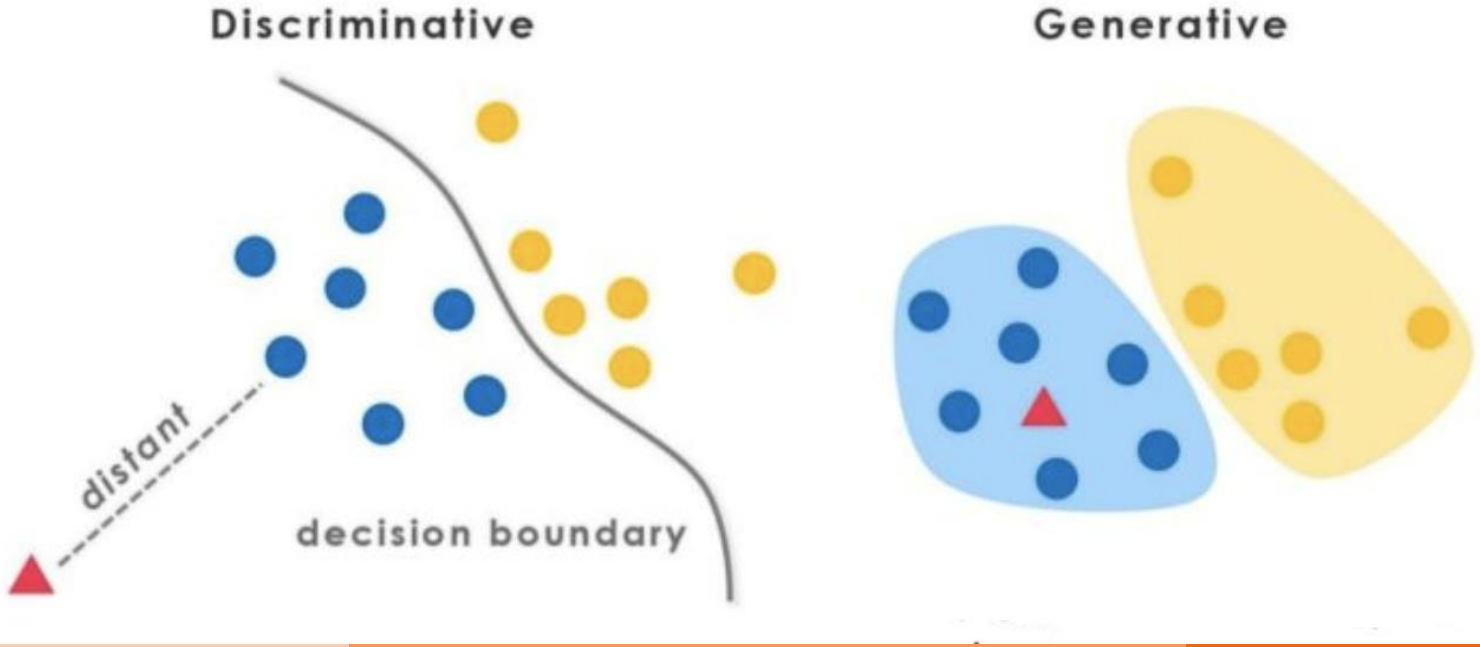
条件随机场模型是Lafferty等人于2001年在最大熵模型和隐马尔可夫模型的基础上提出的一种无向图学习模型,是一种用于标注和切分有序数据的条件概率模型。

CRF最早是针对序列数据分析提出的，现已成功应用于自然语言处理、生物信息学、机器视觉等领域。

1.1 生成模型和判别模型

判别模型：直接将数据的Y（或者label），根据所提供的features，学习，最后画出了一个明显或者比较明显的边界。

生成模型：先从训练样本数据中，学习所有的数据的分布情况，最终确定一个联合分布，来作为所有的输入数据的分布。对于新的样本数据（inference），通过学习到的模型的联合分布 ，再结合新样本给的特征 ，通过条件概率就能出来 。



1.1 生成模型和判别模型

判别方法由数据直接学习决策函数f(X)或者条件概率分布P(Y|X)作为预测的模型，即判别模型。

生成方法由数据学习联合概率分布P(X,Y),然后求出条件概率分布P(Y|X)作为预测的模型，即生成模型：

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

典型的判别模型：K近邻法、感知机、决策树、神经网络、条件随机场等

典型的生成模型：朴素贝叶斯，隐马尔科夫模型

1.1 概率图模型

概率图模型：是一类用图的形式表示随机变量之间条件依赖关系的概率模型，是概率论与图论的结合。

G = (V,E)　　V:顶点/节点，表示随机变量

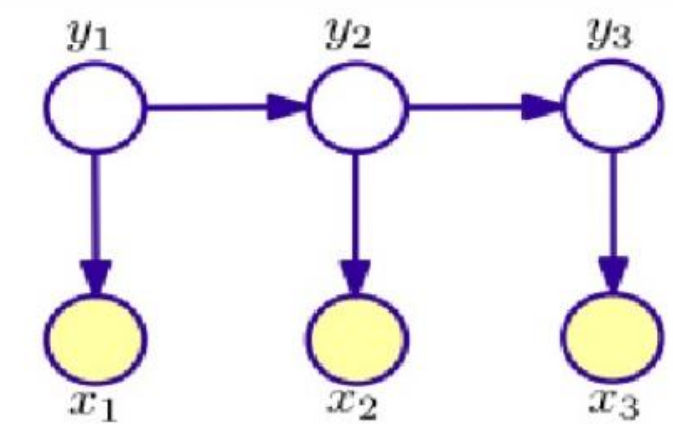
E：边/弧，表示随机变量间的条件依赖关系

根据图中边有无方向，常用的概率图模型分为两类：

有向图：亦称贝叶斯网络或信念网络

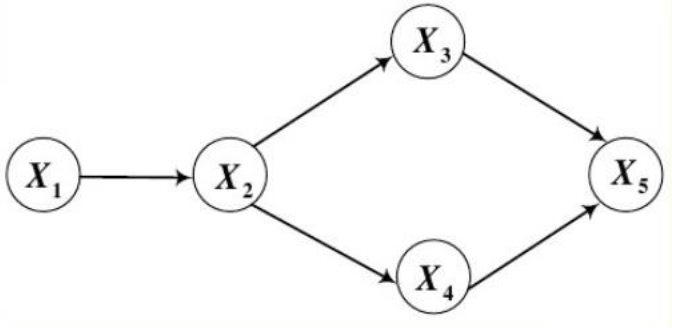
无向图：亦称马尔可夫随机场或马尔可夫网络

1.1 概率有向图模型



有向图的联合概率分布：

$$P(X1,X2,\ldots Xn) = \prod p(Xi|\pi(Xi))$$



图中概率如下：

$$P(X1,X2,\ldots X5) = P(X1)P(X2|X1)P(X3|X2)P(X4|X3)P(X5|X4,X3)$$

1.1 隐马尔科夫随机场HMM

HMM是一个五元组，$\lambda = (Y, X, \pi, A, B)$, 其中Y是状态集合，X是观察值的集合，π是初始状态的概率，A是状态转移概率矩阵，B是输出观察值概率矩阵。

隐马尔科夫模型作了两个基本假设：

1.齐次马尔可夫性假设：即假设隐藏的马尔可夫链在任意时刻t的状态只依赖于其前一时刻的状态，与其他时刻的状态及观测无关，也与时刻t无关。

$$P(i_t \mid i_{t-1}, o_{t-1}, \cdots, i_1, o_1) = P(i_t \mid i_{t-1}), \quad t = 1, 2, \cdots, T$$

2.观测独立性假设：即假设任意时刻的观测只依赖于该时刻的马尔可夫链状态，与其他的观测状态无关。

$$P(o_t \mid i_T, o_T, i_{T-1}, o_{T-1}, \cdots, i_{t+1}, o_{t+1}, i_t, i_{t-1}, o_{t-1}, \cdots, i_1, o_1) = P(o_t \mid i_t)$$

1.1 隐马尔科夫随机场HMM

例 10.1（盒子和球模型） 假设有 4 个盒子，每个盒子里都装有红白两种颜色的球，盒子里的红白球数由表 10.1 列出：

表 10.1 各盒子的红白球数

盒 子	1	2	3	4
红球数	5	3	6	8
白球数	5	7	4	2

按照下面的方法抽球，产生一个球的颜色的观测序列：开始，从 4 个盒子里以等概率随机选取 1 个盒子，从这个盒子里随机抽出 1 个球，记录其颜色后，放回；然后，从当前盒子随机转移到下一个盒子，规则是：如果当前盒子是盒子 1，那么下一盒子一定是盒子 2，如果当前是盒子 2 或 3，那么分别以概率 0.4 和 0.6 转移到左边或右边的盒子，如果当前是盒子 4，那么各以 0.5 的概率停留在盒子 4 或转移到盒子 3；确定转移的盒子后，再从这个盒子里随机抽出 1 个球，记录其颜色，放回；如此下去，重复进行 5 次，得到一个球的颜色的观测序列：

$$O = \{红, 红, 白, 白, 红\}$$

1.1 隐马尔科夫随机场HMM

在这个过程中，观察者只能观测到球的颜色的序列，观测不到球是从哪个盒子取出的，即观测不到盒子的序列.

在这个例子中有两个随机序列，一个是盒子的序列（状态序列），一个是球的颜色的观测序列（观测序列）. 前者是隐藏的，只有后者是可观测的. 这是一个隐马尔可夫模型的例子，根据所给条件，可以明确状态集合、观测集合、序列长度以及模型的三要素.

盒子对应状态，状态的集合是

$$Q=\{\text{盒子1, 盒子2, 盒子3, 盒子4}\},\quad N=4$$

球的颜色对应观测. 观测的集合是

$$V=\{\text{红, 白}\},\quad M=2$$

状态序列和观测序列长度 $T=5$.

初始概率分布为

$$\pi=(0.25, 0.25, 0.25, 0.25)^{\mathrm{T}}$$

1.1 隐马尔科夫随机场HMM

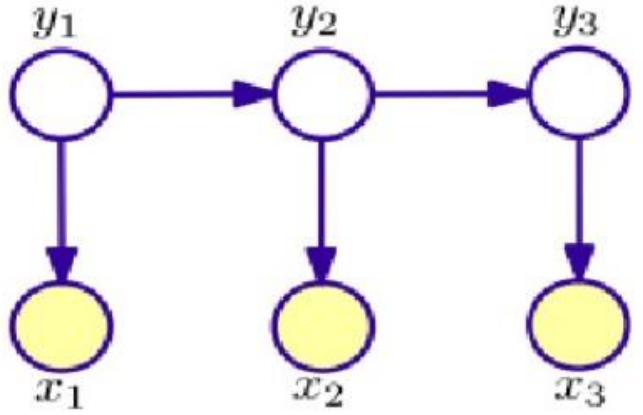
状态转移概率分布为

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

观测概率分布为

$$B=\begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ 0.8 & 0.2 \end{bmatrix}$$

1.1 HMM局限性



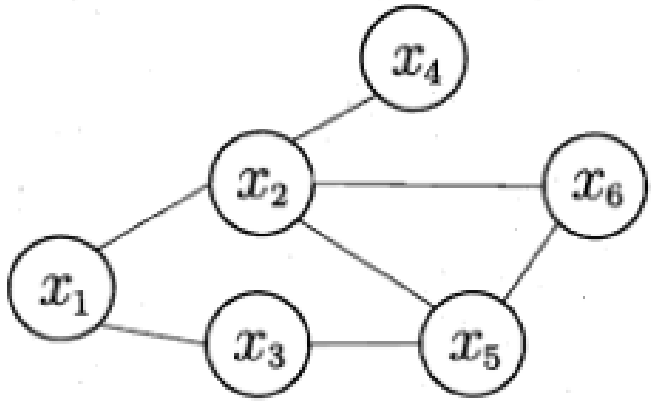
隐马尔科夫模型的**局限性**：

1.模型定义的是联合概率，必须列举所有观察序列的可能值，这对多数领域的来说是比较困难的。

2.基于观察序列中的每个元素都相互条件独立，即在任何时刻观察值仅仅与状态（既要标注的标签）有关。大多数现实世界中的真实观察序列是由多个相互作用的特征和观察序列中较长范围内的元素之间的依赖而成的。

1.1 概率无向图模型

概率无向图模型：设有联合概率分布P(Y),由无向图模型G=(V,E)表示，节点表示随机变量，边表示随机变量之间的依赖关系。如果联合概率分布P(Y)满足成对、局部或全局马尔可夫性，就称此联合概率分布为**概率无向图模型**或**马尔科夫随机场**。



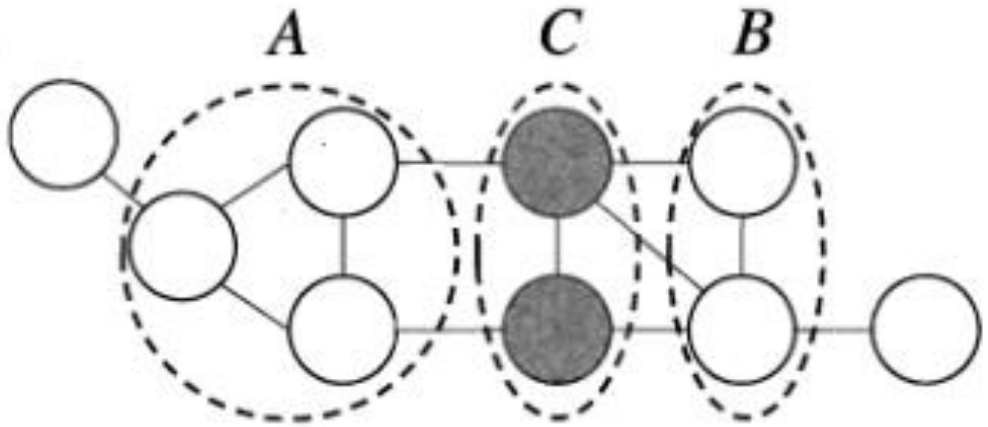
1.1 马尔可夫性

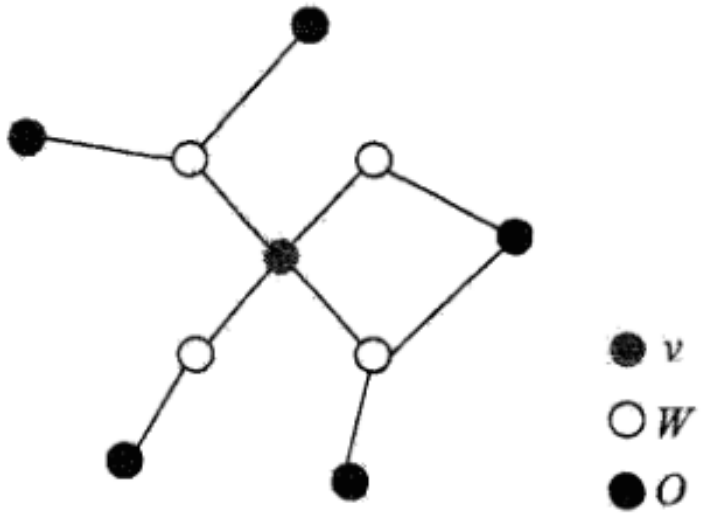
成对的、局部的、全局的马尔可夫性等价

成对马尔可夫性：给定所有其他变量，两个非邻接变量条件独立。

局部马尔可夫性：给定某变量的邻接变量，则该变量条件独立于其他变量。

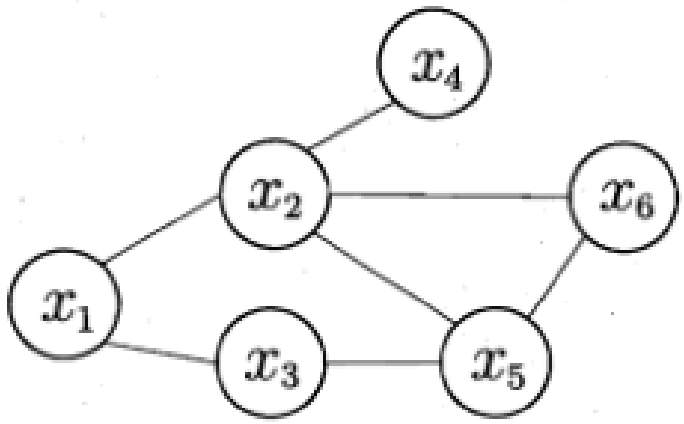
全局马尔可夫性：给定两个变量子集的分离集，则这两个变量子集条件独立。





1.1 团

团与最大团：无向图G中任何两个节点均有边连接节点子集称为团。若C是无向图G的一个**团**，并且不能再加进任何一个G的结点使其成为一个更大的团，则称此C为**最大团**。也就是最大团就是不能被其他团所包含的一个团。



团：

{X2,X4},{X1,X2},{X3,X5},{x1,x3},{x2,x5},

{X2,X6},{X5,X6},{X2,X5,X6}

最大团：

{X2,X4},{X1,X2},{X3,X5},{x1,x3},{X2,X5,X6}

1.1 表达式

将概率无向图模型的联合概率分布P(Y)可写作图中所有最大团C上的函数的乘积形式。

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_C(Y_C)$$

Z：规范化因子

$$Z=\sum_{Y}\prod_{C}\Psi_C(Y_C)$$

势函数要求是严格正的，通常定义为指数函数。

$$\Psi_C(Y_C)=\exp\{-E(Y_C)\}$$

1.1 条件随机场

条件随机场：设X与Y是随机变量，P(Y|X)是在给定X的条件下的条件概率分布，若随机变量Y构成一个由无向图G=(V,E)表示的马尔科夫随机场，即

$$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v)$$

对任意结点v成立，则称条件概率分布P(Y|X)为条件随机场。

条件随机场是给定随机变量X的条件下，随机变量Y的马尔科夫随机场。

1.1 条件随机场示例

词性标注问题：　　“Bob drank coffee at Starbucks”

“Bob（名词）drank(动词）coffee(名词）at(介词）Starbucks

标注序列:(名词，动词，名词，介词，名词)，（名词，动词，动词，介词，名词）...

CRF中的特征函数：

- 句子s（就是我们要标注词性的句子）
- i，用来表示句子s中第i个单词
- l_i，表示要评分的标注序列给第i个单词标注的词性
- l_i-1，表示要评分的标注序列给第i-1个单词标注的词性

输出值：0或1

0表示标注序列不符合这个特征

1表示标注序列符合这个特征

1.1 条件随机场示例

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

当l_i是“副词”并且第i个单词以“ly”结尾时，我们就让f1 = 1，其他情况f1为0。

$$f_1(s, i, l_i, l_{i-1}) = 1$$

如果i=1，l_i=动词，并且句子s是以“？”结尾时，f2=1，其他情况f2=0。

$$f_2(s, i, l_i, l_{i-1}) = 1$$

当l_i-1是介词，l_i是名词时，f3 = 1，其他情况f3=0。

$$f_3(s, i, l_i, l_{i-1}) = 1$$

如果l_i和l_i-1都是介词，那么f4等于1，其他情况f4=(

$$f_4(s, i, l_i, l_{i-1}) = 1$$

1.1 条件随机场示例

对分数进行指数化和标准化得到：

$$p(l|s) = \frac{exp[score(l|s)]}{\sum_{l'} exp[score(l'|s)]} = \frac{exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l'_i,l'_{i-1})]}$$

总结： 为了建一个条件随机场，我们首先要定义一个特征函数集，每个特征函数都以整个句子s，当前位置i，位置i和i-1的标签为输入。然后为每一个特征函数赋予一个权重，然后针对每一个标注序列l，对所有的特征函数加权求和，必要的话，可以把求和的值转化为一个概率值。

1.1 线性链条件随机场

线性链条件随机场： 设X=(X1,X2,…,Xn),Y=(Y1,Y2,…,Yn)均为线性链表示的随机变量序列，若在给定随机变量序列X的条件下，随机变量序列Y的条件概率分布P(Y|X)构成条件随机场，既满足马尔可夫性

$$P(Y_i \mid X, Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n) = P(Y_i \mid X, Y_{i-1}, Y_{i+1})$$

$$i = 1, 2, \cdots, n \quad （在 i = 1 和 n 时只考虑单边）$$

线性链条件随机场的参数化形式：

$$Z(x) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right)$$

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right)$$

1.1 线性链条件随机场

将两个特征函数统一：

$$f_k(y_{i-1},y_i,x,i)=\begin{cases}t_k(y_{i-1},y_i,x,i), & k=1,2,\cdots,K_1\\ s_l(y_i,x,i), & k=K_1+l;\ l=1,2,\cdots,K_2\end{cases}$$

对转移与状态特征在各个位置求和：

$$f_k(y,x)=\sum_{i=1}^{n}f_k(y_{i-1},y_i,x,i),\quad k=1,2,\cdots,K$$

条件随机场可表示为：

$$P(y\mid x)=\frac{1}{Z(x)}\exp\sum_{k=1}^{K}w_k f_k(y,x)$$

$$Z(x)=\sum_{y}\exp\sum_{k=1}^{K}w_k f_k(y,x)$$

1.1 线性链条件随机场

例 11.1 设有一标注问题：输入观测序列为 $X=(X_1,X_2,X_3)$，输出标记序列为 $Y=(Y_1,Y_2,Y_3)$，Y_1,Y_2,Y_3 取值于 $\mathcal{Y}=\{1,2\}$.

假设特征 t_k，s_l 和对应的权值 λ_k，μ_l 如下：

$$t_1=t_1(y_{i-1}=1,y_i=2,x,i),\quad i=2,3,\quad \lambda_1=1$$

$$t_1(y_{i-1},y_i,x,i)=\begin{cases}1, & y_{i-1}=1,y_i=2,x,i,(i=2,3)\\ 0, & \text{其他}\end{cases}$$

这里只注明特征取值为 1 的条件，取值为 0 的条件省略，即

对给定的观测序列 x，求标记序列为 $y=(y_1,y_2,y_3)=(1,2,2)$ 的非规范化条件概率（即没有除以规范化因子的条件概率）.

$$t_2=t_2(y_1=1,y_2=1,x,2)\qquad \lambda_2=0.5$$
$$t_3=t_3(y_2=2,y_3=1,x,3)\qquad \lambda_3=1$$
$$t_4=t_4(y_1=2,y_2=1,x,2),\qquad \lambda_4=1$$
$$t_5=t_5(y_2=2,y_3=2,x,3),\qquad \lambda_5=0.2$$
$$s_1=s_1(y_1=1,x,1),\qquad \mu_1=1$$
$$s_2=s_2(y_i=2,x,i),\quad i=1,2\qquad \mu_2=0.5$$
$$s_3=s_3(y_i=1,x,i),\quad i=2,3\qquad \mu_3=0.8$$
$$s_4=s_4(y_3=2,x,3),\qquad \mu_4=0.5$$

计算：

$$P(y\mid x)\propto \exp\left[\sum_{k=1}^{5}\lambda_k\sum_{i=2}^{3}t_k(y_{i-1},y_i,x,i)+\sum_{k=1}^{4}\mu_k\sum_{i=1}^{3}s_k(y_i,x,i)\right]$$

$$P(y_1=1,y_2=2,y_3=2\mid x)\propto \exp(3.2)$$

1.1 CRF关键问题

关键问题：

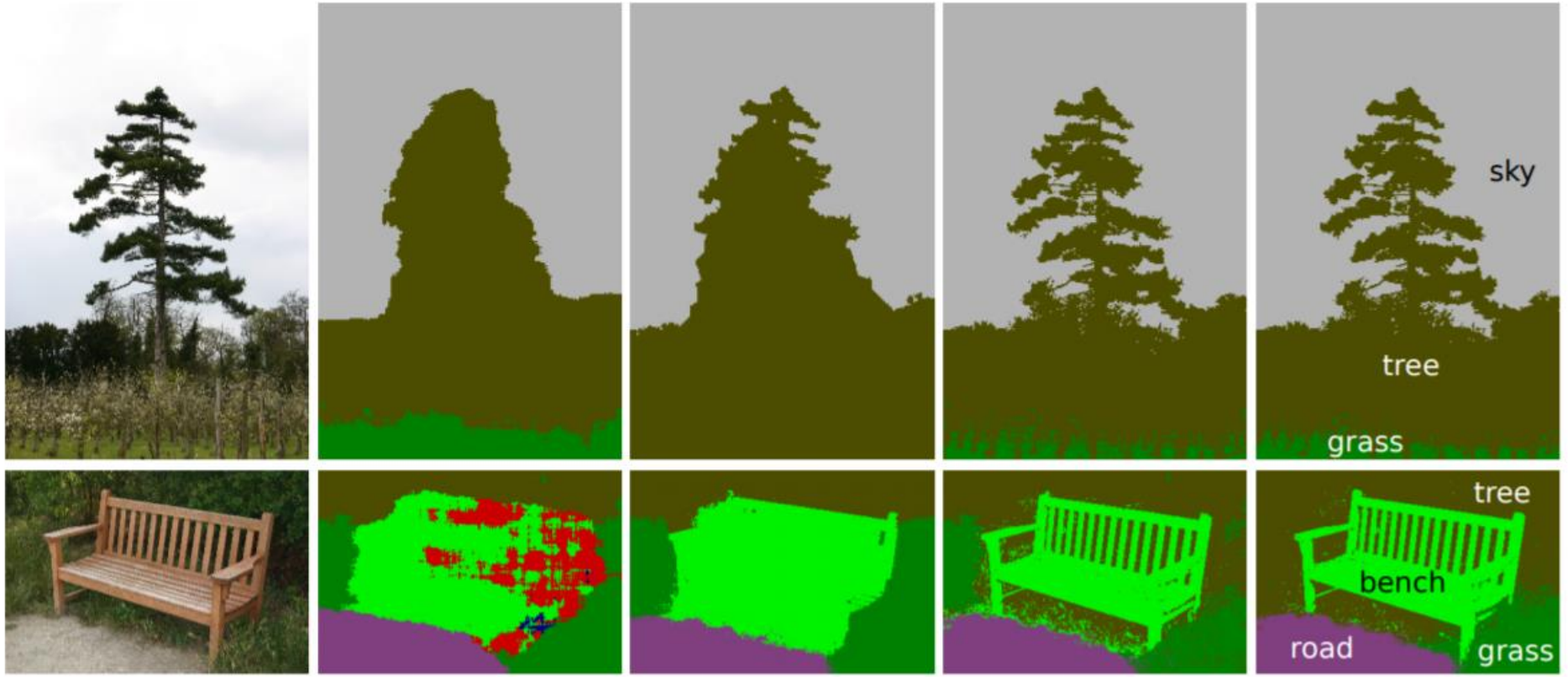
1.特征函数的选择：特征函数的选取直接关系模型的性能

2.参数估计：从已经标注好的训练数据集学习条件随机场模型的参数，即各特征函数权重向量

3.模型推断：在给定条件随机场模型参数下，预测出最可能的状态序列

1.2 全连接CRF模型

论文：Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials
期刊：NIPS(Neural Information Processing Systems)
发表时间：2011



(a) Image (b) Unary classifiers (c) Robust P^n CRF (d) Fully connected CRF, MCMC inference, 36 hrs (e) Fully connected CRF, our approach, 0.2 seconds

1.2 全连接CRF模型

首先定义一个条件随机场，服从于Gibbs分布，联合概率分布可以表达为：

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-\sum_{c \in \mathcal{C}_\mathcal{G}} \phi_c(\mathbf{X}_c|\mathbf{I})),$$

$$E(\mathbf{x}|\mathbf{I}) = \sum_{c \in \mathcal{C}_\mathcal{G}} \phi_c(\mathbf{x}_c|\mathbf{I}).$$

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{L}^N} P(\mathbf{x}|\mathbf{I}).$$

1.2 全连接CRF模型

成对势函数：

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \underbrace{\sum_{m=1}^{K} w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)}_{k(\mathbf{f}_i, \mathbf{f}_j)},$$

$$k(\mathbf{f}_i, \mathbf{f}_j) = w^{(1)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + w^{(2)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}}.$$

$$\mu(x_i, x_j) = [x_i \neq x_j].$$

1.2 CRF图像分割

Deeplab分割

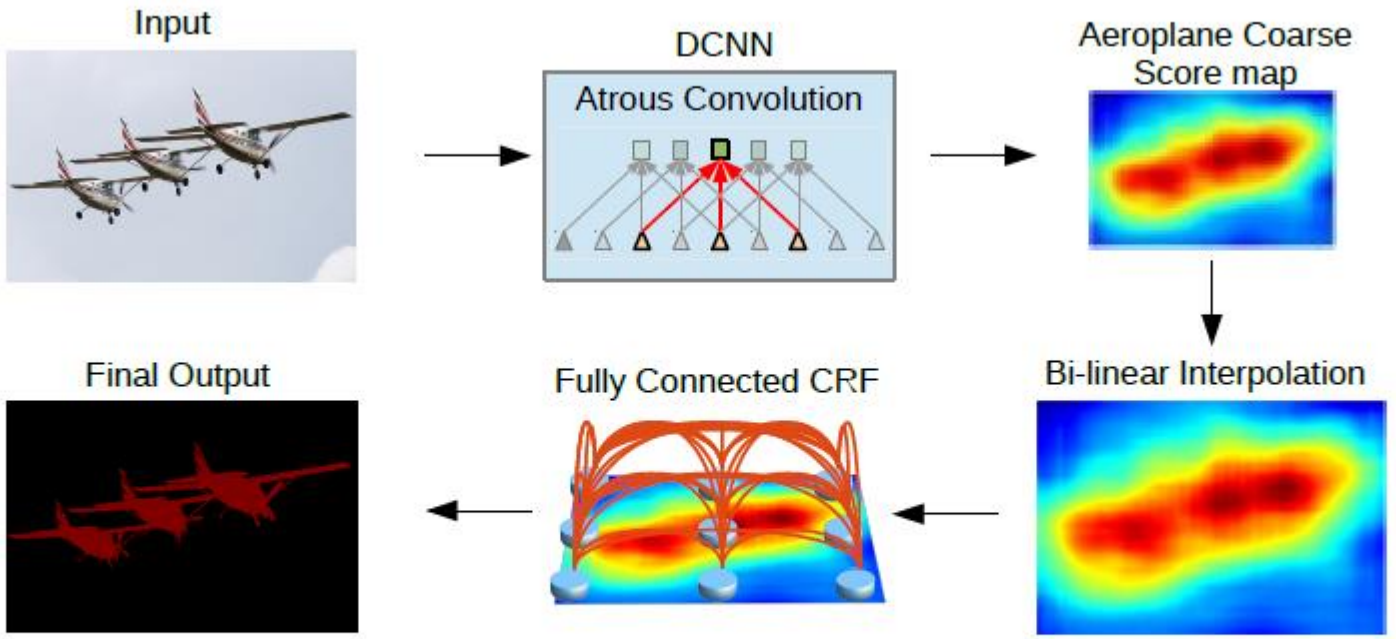


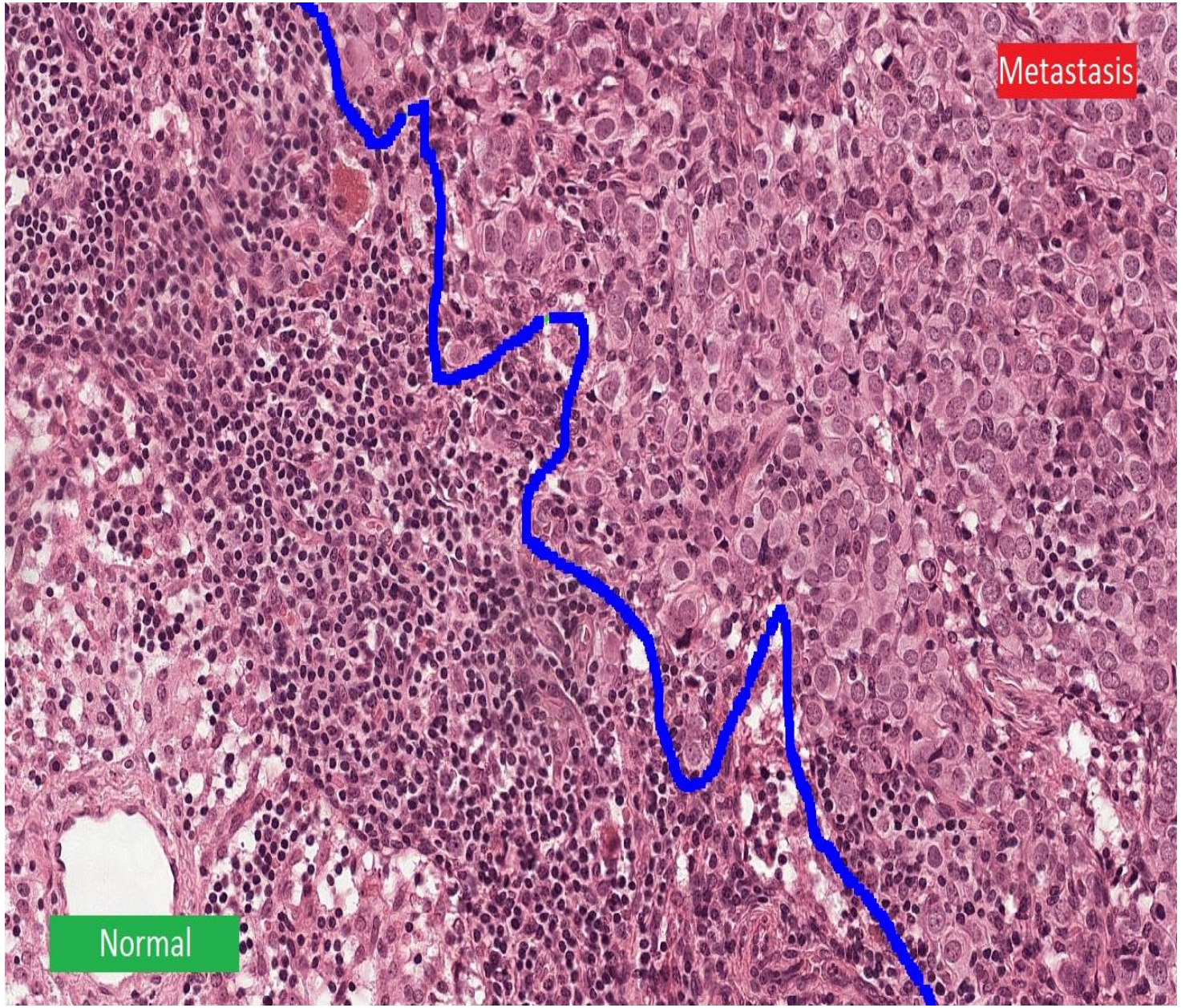
Fig. 1: Model Illustration. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries.

目录

2 Camelyon16

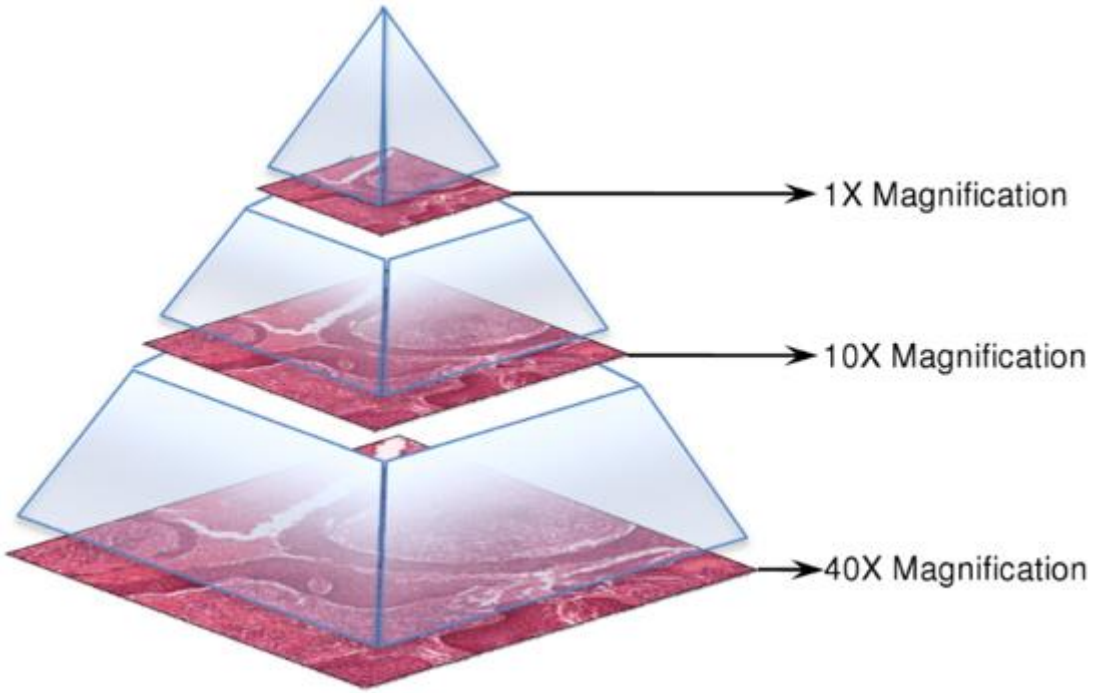
Camelyon16:对乳腺癌在淋巴结中的转移进行病理切片的分类与定位。

主要任务为对测试集中的120张淋巴结病理切片进行判断是否发生了癌变(classification)，同时需要对发生癌变的位置区域精准定位(segmentation)。



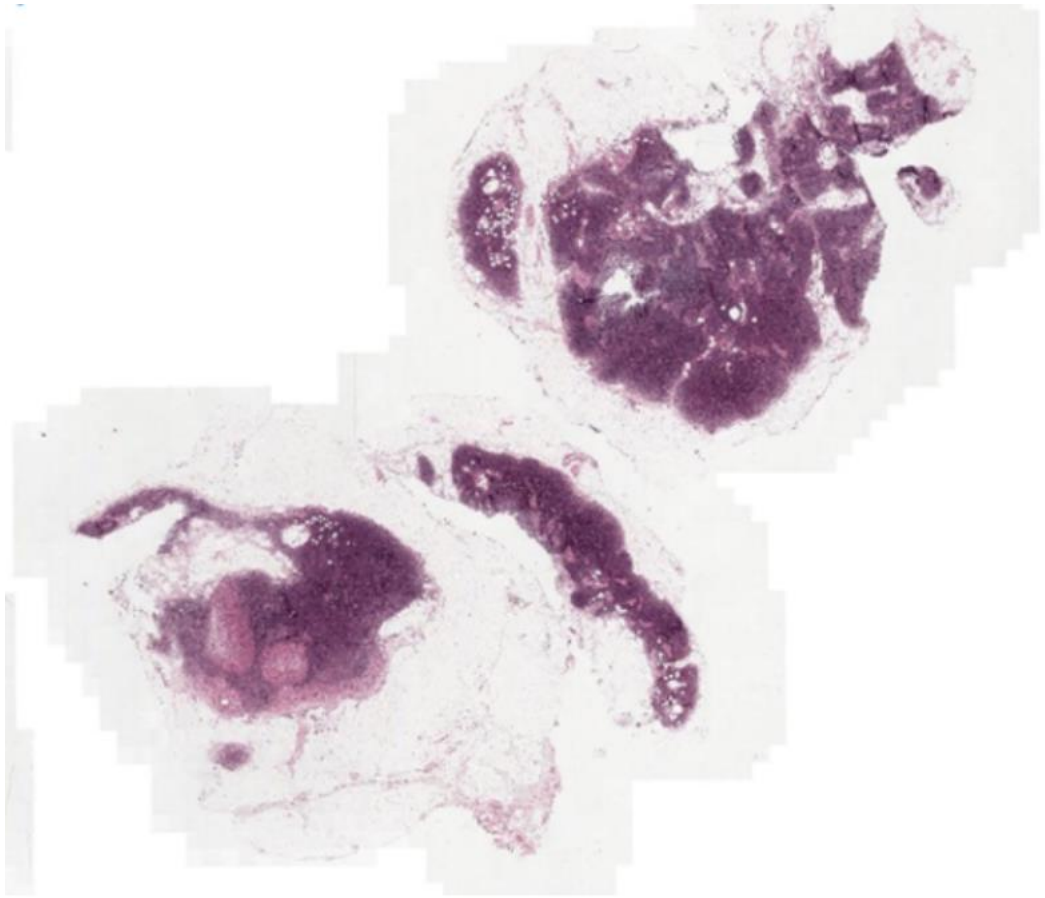
2 数据集

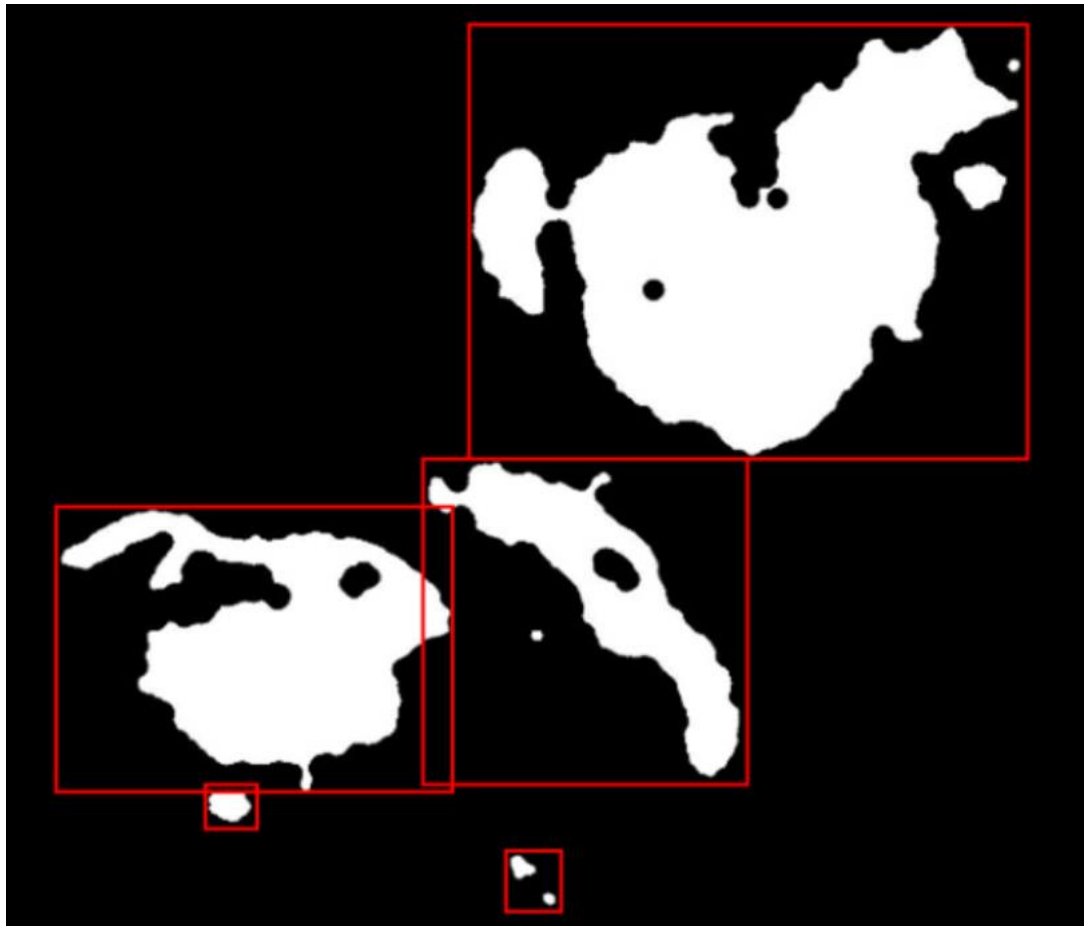
病理数据：160张含有癌症组织的切片，240张正常组织的切片



在最大分辨率40x的图像大小为300000x150000,一个样本的所占硬盘空间大小大概是5~6G。

2 预处理





2 NCRF

论文：Cancer Metastasis Detection With Neural Conditional Random Field
期刊：MIDL(Medical Imaging with Deep Learning)
发表时间：2018.4

该方法不仅分析单个小图片，也将图片四周邻近的网格一并输入进行肿瘤细胞分析。相邻切片之间的空间相关性通过条件随机场进行建模。

2 NCRF

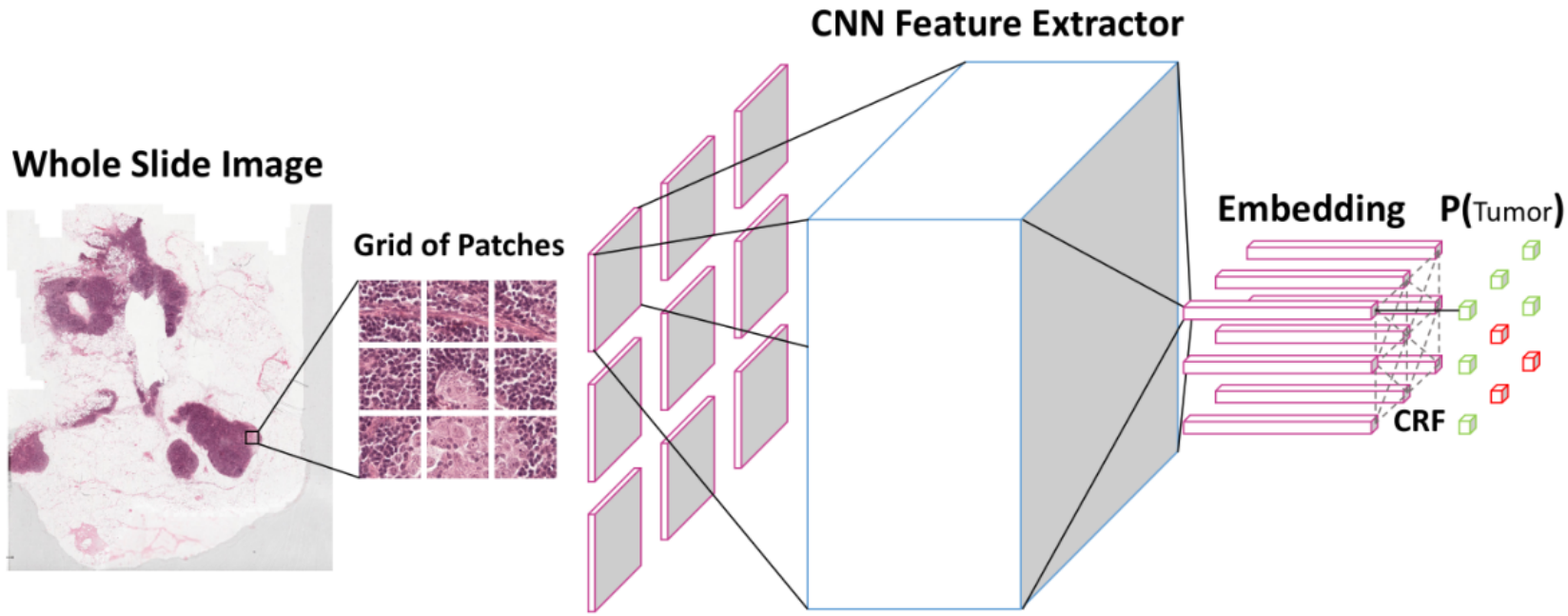


Figure 1: The architecture of NCRF model.

2 NCRF

Gibbs分布：如果无向图模型能够表示成一系列在G的最大团（们）上的非负函数乘积的形式，这个无向图模型的概率分布P(X)就称为Gibbs分布。

定义一个满足Gibbs分布的条件概率P(Y|x),

$$P(\mathbf{Y}=\boldsymbol{y}\mid\boldsymbol{x})=\frac{1}{Z(\mathbf{x})}\exp(-E(\boldsymbol{y},\boldsymbol{x}))$$

$$y^{*}=argmax_{y}P(Y|x)$$

3 NCRF

能量函数E(x):

$$E(\mathbf{y}, \boldsymbol{x}) = \sum_i \psi_u(y_i) + \sum_{i<j} \psi_p(y_i, y_j)$$

$$\psi_p(y_i, y_j) = \mathbb{I}(y_i = y_j) \cdot w_{i,j} \left(1 - \frac{x_i \cdot x_j}{\|x_i\| \, \|x_j\|}\right)$$

2 NCRF

平均场论(简称MFT，也叫作**自洽场理论**)是对大且复杂的随机模型的一种简化。未简化前的模型通常包含巨大数目的含相互作用的小个体。平均场理论则做了这样的近似：对某个独立的小个体，所有其他个体对它产生的作用可以用一个平均的量给出。

相对熵，又称为**KL散度**。**KL散度**是两个概率分布P和Q差别的非对称性的度量。KL散度是用来度量使用基于Q的编码来编码来自P的样本平均所需的额外的位元数。 典型情况下，P表示数据的真实分布，Q表示数据的理论分布、模型分布、或P的近似分布。

$$D_{\mathrm{KL}}(P\|Q) = -\sum_{i} P(i)\ln\frac{Q(i)}{P(i)}.$$

2 NCRF

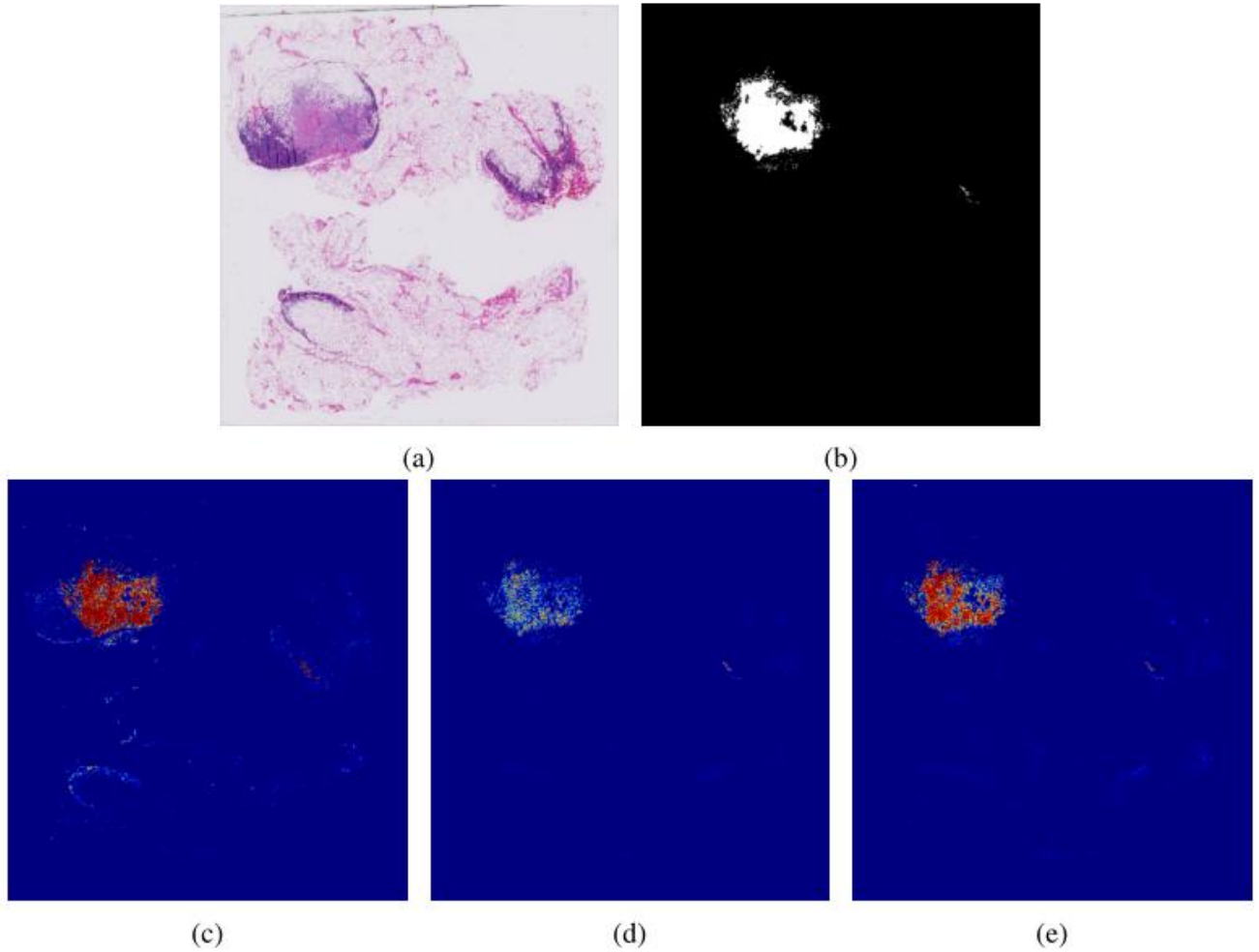
平均场近似推理：

$$\log Q_i(y_i) = \mathbb{E}_{-Q_i}\left[\log \tilde{P}(\mathbf{Y})\right] + \text{const}$$

Algorithm 1 Mean-field inference algorithm

compute $\psi_u(y_i)$ for all i and $\psi_p(y_i, y_j)$ for all i, j
$\log \tilde{P}(\mathbf{Y}) \leftarrow -\left[\sum_i \psi_u(y_i) + \sum_{i<j} \psi_p(y_i, y_j)\right]$
initialize $Q_i(y_i) \leftarrow \exp(-\psi_u(y_i))$ for all i
normalize $Q_i(y_i)$ for all i
for T iterations **do**
  $\log Q_i(y_i) \leftarrow \mathbb{E}_{-Q_i}\left[\log \tilde{P}(\mathbf{Y})\right]$ for all i
  normalize $Q_i(y_i)$ for all i
end for

2 NCRF



a.原始wsi影像

b. Ground truth

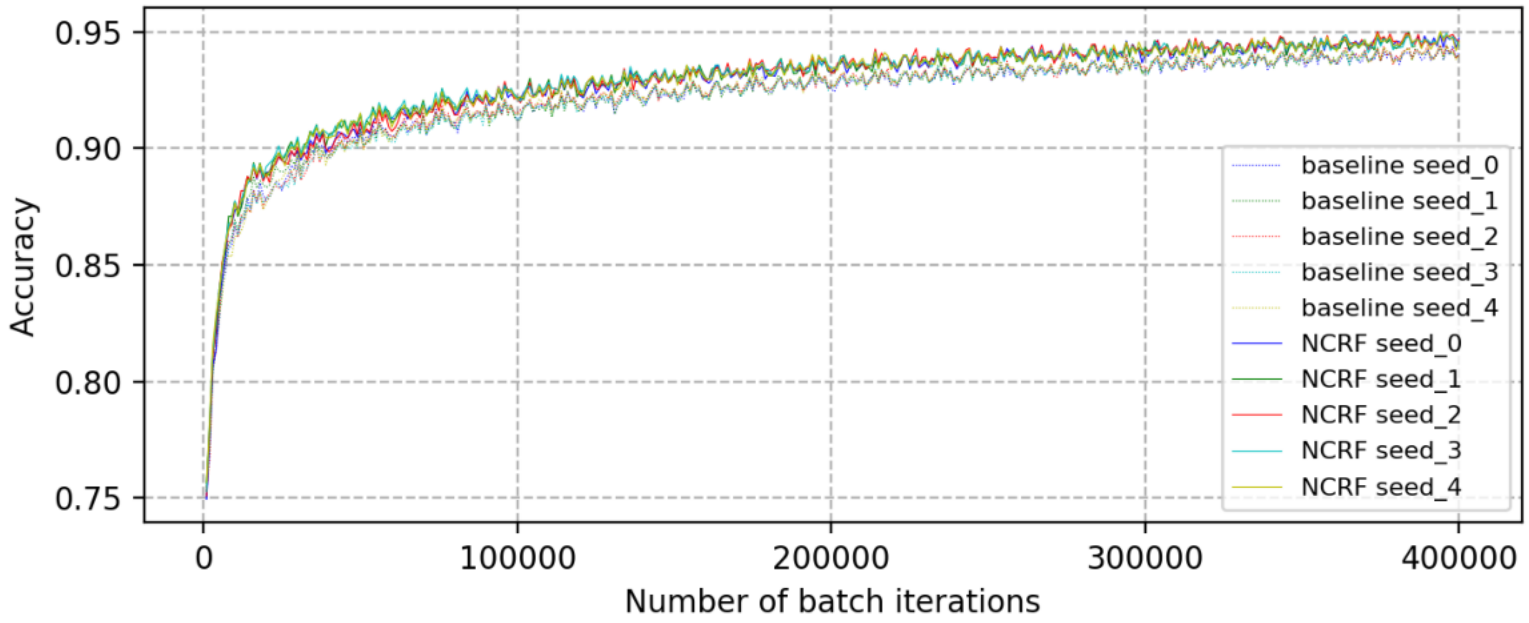
c.标准方法

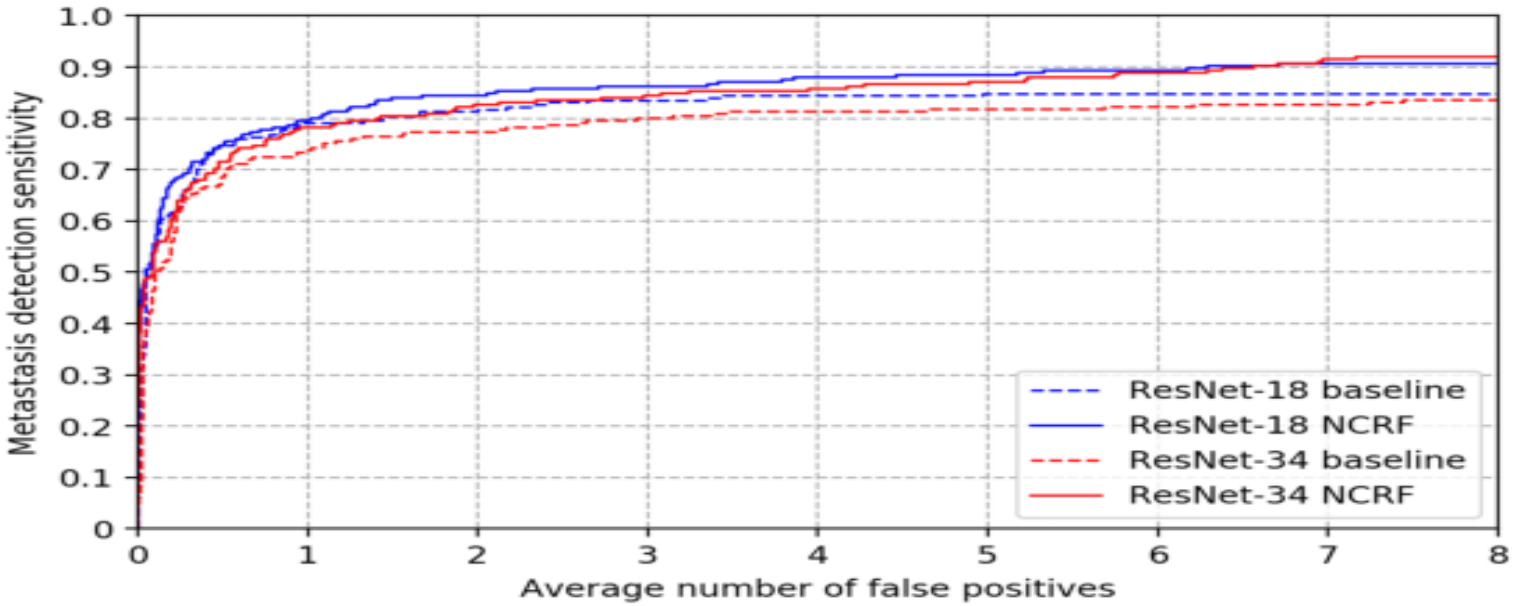
d.标准方法+hard negative mining

e.NCRF+hard negative mining

2 NCRF

AUC曲线：





FROC曲线
FROC指标达到0.8069

感谢聆听