Chinese Street View Text: Large-scale Chinese Text Reading with Partially Supervised Learning

论文两个重要的贡献：

1. 现有的文本阅读 benchmark 很难评估更高级的深度学习模型的性能，因此作者提供了大型中文数据集 benchmark
2. 提出了一种部分监督的端到端的训练模型，可以同时定位和识别检测中文文本。与之前以完全监督的方式进行端到端的模型不同，作者的模型是通过部分监督的学习框架进行训练的。通过将大规模的弱注释数据纳入训练过程，可以进一步提高端到端的性能。

Full Annotations：

给出文本的准确位置和文字内容

Weak Annotations:

给出文本的大概区域和文字内容

Text Detection Branch:

Backbone 采用了 ResNet50，并且采用 fpn 将各个层的特征进行融合，生成 feature map F。先将 F 送入检测分支，检测分支会在 F 的每个空间位置进行文本/非文本分类，以计算其属于文本区域的概率。 并预测文本区域的四个顶点之间的偏移量{（Δxm，Δym）| m = 1,2,3,4} 。在训练阶段，检测损失 Ldet 定义为 Ldet = Lloc +λLcls，其中 Lcls 是用于文本/非文本分类的损失，Lloc 是根据位置回归的 SmoothL1 损失计算的，而 λ 是平衡这两个损失的超参数。在测试阶段，检测分支将阈值应用于文本分类的预测概率，并在选定的空间位置上执行 NMS（非最大抑制），以生成四边形文本 proposal。

Perspective RoI Transform:

采用透视 RoI 变换将特征图 F 中的相应区域对齐到小特征图 Fp 中，每个特征图 Fp 都保持固定的高度，且长宽比不变。如果在检测网络中检测出竖直形文字（纵横比大于 1），则将会进行顺时针旋转。（根据论文自己理解的）

Text Recognition Branch:

将变换后的的 feature map 送入 cnn+rnn 网络，提取 feature map 的空间，时间特征，再使用 attention 机制进行每个字符的识别，最后加一个全连接层和 softmax 计算输出字符标签 yt 的概率。（这些都是完全监督的训练过程）

使用 Weak Annotations 数据进行训练，为此论文提出了一个叫 Online Proposal Matching 的模块（OPM）

Online Proposal Matching

OPM 模块旨在定位与关键字注释 yw 相对应的文本区域。 首先利用完全监督模型的检测分支来生成一组文本 proposal。然后，通过透视图 RoI 变换提取每个建议的特征图，并由 CNN-RNN 在文本识别分支中将其编码为顺序特征 Fw。此外，为了计算特征 Fw 和弱标记关键字 yw 之间的相似度，最后将 Fw 和 yw 转换到同一个状态域进行相似度的计算。这是大致的思路，具体细节还不是太了解。

损失函数

总的损失

$$L_{total} = L_{det} + \beta(L_{recog} + L^{w}_{recog})$$

其中 L_{det} 代表检测的损失，L_{recog} 代表完全注释下的识别损失，L^{w}_{recog} 代表弱标注下的识别损失

$$L^{w}_{recog} = \frac{1}{\sum_{i=1}^{N} m(i)} \sum_{i=1}^{N} m(i) l^{w}_{recog}(i), \qquad (4)$$

where $m(i) = 1$ if $d^{w}(i) \leq \tau$, otherwise $m(i) = 0$ and a threshold τ is used to select the matched text proposals. The

$$l^w_{recog}(i) = -\frac{1}{T^w}\sum_{t=1}^{T^w}\log p(\mathbf{y}^w_t|\mathbf{y}^w_{t-1}, \mathbf{h}^w_{t-1}, \mathbf{c}^w_t), \quad (5)$$

where $\mathbf{c}^w_t$ denotes the context vector at time t calculated by attention mechanism. The total loss for the partially super-

Training Pipeline

Stage one:利用完全标注的数据进行先进行训练.

Stage two:训练 **OPM**，利用之前训练好的网络生成一系列的 **proposal**，为了训练 **OPM**，需要产生正负样本。

论文里的思想：利用一张完全标注的图片，随机选择一个文本实例，还有检测网络生成的 **proposal**，分别计算文本实例域每个 **proposal** 的 **IOU**，小于一定阈值的定为负样本，正样本直接就是文本实例（**ground true**）

OPM 的训练也是先用完全标注数据进行训练。

Stage three:利用两种数据放入整个网络进行端到端的训练。

实现细节

将图像填充到 **512×512**。在 **RoI** 变换层中，将扭曲特征图的高度和最大宽度分别设置为 **8** 和 **64**。如果特征图的宽度小于 **64**，则使用零值填充它。否则，我们将使用双线性插值来调整其大小，以将宽度设置为 **64**。