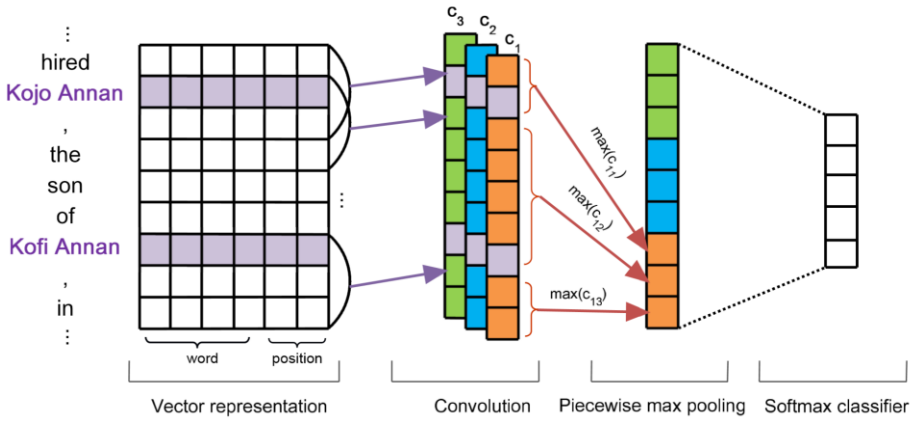
论文阅读笔记：[Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks]

本篇论文是对上一批论文（Relation Classification via Convolutional Deep Neural Network）工作的补充和扩展，核心的模型结构仍然为 CNN，但是针对远程监督的缺点做出了两点针对性的改进： Piecewise Max Pooling 和 Multi-instance Learning

1. Piecewise Max Pooling



首先来看 Piecewise Max Pooling，与 global maxpooling 相比，Piecewise 的输出维度为原来的三倍大小，其原理在于：将整个句子分为三个部分，其中，分割点选取为两个目标词的位置。对每个部分都使用一次 global maxpooling，最后拼接获得的三个特征向量，此举是为了增加获取的特征数量，减少 global maxpooling 信息损失。

2. Multi-instance Learning

Multi-instance Learning 的核心思想是：为了缓解数据集里的错误标记带来的噪声，训练过程中，使用每个 minibatch 中置信度最高的一个样本进行训练，舍弃其他的训练样本。

Algorithm 1 Multi-instance learning

1: Initialize θ. Partition the bags into mini-batches of size b_s.
2: Randomly choose a mini-batch, and feed the bags into the network one by one.
3: Find the j-th instance m_i^j $(1 \leq i \leq b_s)$ in each bag according to Eq. (9).
4: Update θ based on the gradients of m_i^j $(1 \leq i \leq b_s)$ via Adadelta.
5: Repeat steps 2-4 until either convergence or the maximum number of epochs is reached.

根据论文中的算法说明，选取第 i 个 minibatch 选取置信度最大的那个样本 m_i^j 进行训练。

文中置信度计算方法：

$$J(\theta) = \sum_{i=1}^{T} \log p(y_i | m_i^j; \theta) \qquad (8)$$

where j is constrained as follows:

$$j^* = \arg\max_j p(y_i | m_i^j; \theta) \;\; 1 \le j \le q_i \qquad (9)$$

置信度=条件概率

条件概率的计算方法：使用网络正向传播计算，网络的 softmax 输出概率，作为条件概率。