

Self-Attention Enhanced CNNs and Collaborative Curriculum Learning for Distantly Supervised Relation Extraction

Yuyun Huang[†] Jinhua Du^{‡§}

[†]University College Dublin, Ireland

[‡]Investments AI, AIG (American International Group, Inc.)

[§]ADAPT Centre, School of Computing, Dublin City University, Ireland

{yuyun.huang}@ucd.ie

{jinhua.du}@aig.com

IJCAI 2019

这篇paper研究的task也是DS，然后也是bag-level 的relation prediction，使用的也是NYT+Freebase的dataset。

之前讲过MIL，有用bag中probability最高的，也有用attention soft结合起来的。

这篇paper把两个都拿过来了，当成两个student network。

然后用curriculum learning的方法来学习（其实我感觉更像是共同学习多一点？）

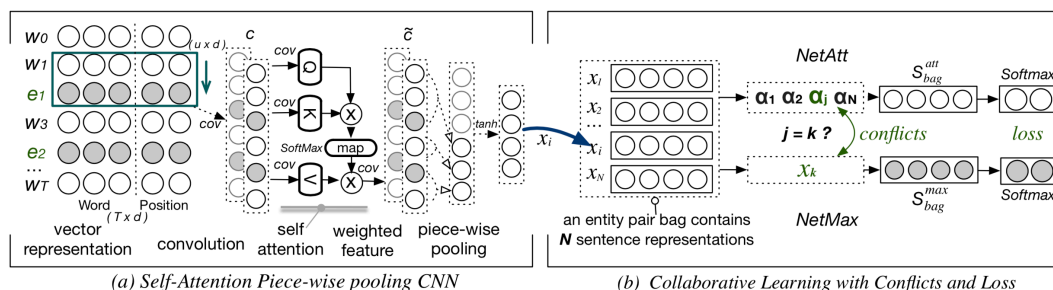
首先模型的输入是bag的一个一个的sentence，那还是比较常规的embedding模式。

然后呢，就过CNN。

然后呢，论文说的一个创新点就是在CNN之后再做一个self-attention。

当然了，我觉得它说的原因就是扯淡。。

然后最后还是一个PCNN那种形式的max-pooling



反正最后就是得到bag中每个sentence的representation。

之后怎么预测呢？

大白话就是：两个student网络分别预测，然后比较他们预测的概率最高的那个

sentence是不是同一个，一样的话就认为是easy的样本，加进去算loss；如果不一样就认为是difficult，把它mask掉，不参与loss的计算。注意两个网络的loss是分别各自计算的，由hyperparameter加权求和得到最终总loss。然后为了让两个网络互相regularize，还加了一个两个模型预测最高概率的index的交叉熵。这个什么意思呢？如果两个网络预测同一个，那么交叉熵为0，loss为0。如果不是同一个，那么交叉熵就可以衡量两个的距离啦。（似乎不是对称的哦？）

Algorithm 1 Update using conflicts

inputs: mini batch size m ; j_{att}^i and k_{max}^i ; two students: p_{att} (*NetAtt*), p_{max} (*NetMax*); bag representation: S_i^{att} , S_i^{max}

for i in $\{1, 2, 3, \dots, m\}$ **do**

if $j_{att}^i = k_{max}^i$ **then**

 # no conflicts $v_i^c = 1$

 add S_i^{att} to set $\{S^{att}\}$

 add S_i^{max} to set $\{S^{max}\}$

end

end

update p_{att} with $\{S^{att}\}$ & p_{max} with $\{S^{max}\}$

$$j(S_i; \theta) = \eta \log p(r_i | S_{bag-i}^{att}; \theta) + (1 - \eta) \log p(r_i | S_{bag-i}^{max}; \theta)$$

$$\min_{\theta, v} E(\theta, v, \lambda) = \frac{1}{m} \sum_{i=1}^m v_i l_i + L_{jk}$$

实际上，这些easy的样本，应该就可以看成是noise比较小的instance啦！

最后当然效果不错啦！

个人感觉，这个两个student network用他们的conflict来筛选鉴定出比较不noisy的training data参与训练，想法还是挺不错的！