深度学习实践初稿

恶意评论识别

周文杰
黄志鹏
徐铖

Background

- 如今随着社交网络和社交媒体的高速发展，对网络上发表文字的监管显得尤为重要，一些欧洲国家新颁发的规定要求在72小时内强制删除不合法的内容，这对于人工智能处理恶意评论识别提出了巨大的挑战。
- 过去的一些传统模型在处理这类问题时通常对可能带有歧视意义的少数群体名词给予过高的权重，比如错将“I am a gay woman”判别为恶意评论，这是由于这类词通常和辱骂性文字一同出现。

Background

- 我们参加的kaggle竞赛

https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview/description

期望在之前模型的基础上使用新的深度学习模型解决这个问题，这类词包括至少出现500次的transgender, Muslim等性别种族专有名词。

Dataset Overview

- 这项任务的数据集来自已关闭的civil comment平台上将近200万的评论，每一条由人工标注恶意指数（0.0 ~ 1.0），大于等于0.5认为该评论为恶意评论。
- 最终分数由评分的损失函数和由带敏感词的权重矩阵综合计算得出，具体如下图
- 我们目前对数据集只做了一些简单的处理，包括去掉所有的特殊字符，以及把所有字符转化为小写

Generalized Mean of Bias AUCs

To combine the per-identity Bias AUCs into one overall measure, we calculate their generalized mean as defined below:

$$M_p(m_s) = \left(\frac{1}{N} \sum_{s=1}^{N} m_s^p \right)^{\frac{1}{p}}$$

where:

M_p = the pth power-mean function
m_s = the bias metric m calulated for subgroup s
N = number of identity subgroups

For this competition, we use a p value of -5 to encourage competitors to improve the model for the identity subgroups with the lowest model performance.

Final Metric

We combine the overall AUC with the generalized mean of the Bias AUCs to calculate the final model score:

$$score = w_0 AUC_{overall} + \sum_{a=1}^{A} w_a M_p(m_{s,a})$$

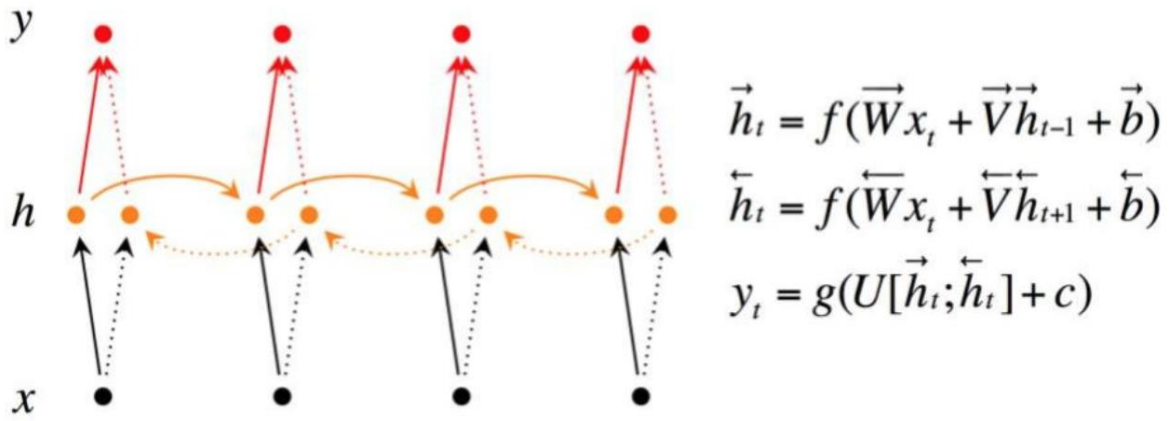
where:

A = number of submetrics (3)
$m_{s,a}$ = bias metric for identity subgroup s using submetric a
w_a = a weighting for the relative importance of each submetric; all four w values set to 0.25

Baseline Model

- Wording embedding我们采用Glove拼接FastText的方式，以此使语义信息表示的更丰富。
- 模型主体我们先尝试最基本的BiLSTM，LSTM在RNN基础上增加了遗忘和记忆的功能，提炼出句子重要部分和关键词间的前后联系。
- Baseline model在测试集上的bias AUC值为0.934

Bidirectional LSTM，由两个LSTMs上下叠加在 一起组成。输出由这两个LSTMs的隐藏层的状态决定。



Improvement Work

- 我们尝试增加多任务学习，增加一个language model层在encoder上。
- 如果只是利用last hidden或者使用mean/max pooling的结果进行预测，训练目标对模型的约束有限，不能有效的指导模型学到复杂的表征。
- 为了使模型更充分地使用training data，我们引入language model。具体地，使用前向LSTM的hidden state $\overrightarrow{h_t}$ 预测下一个词，使用后向LSTM的hidden state $\overleftarrow{h_t}$预测下一个词

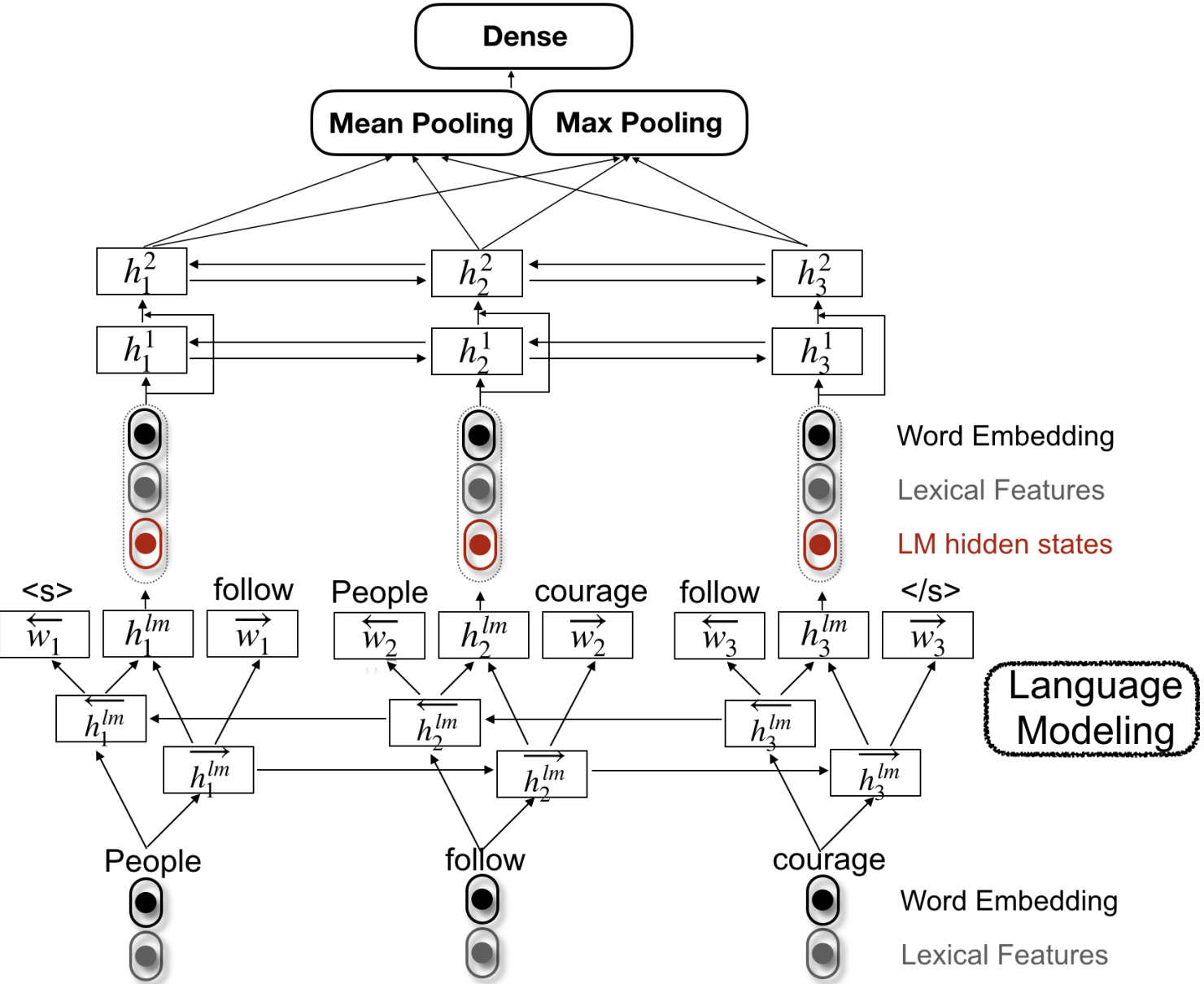
Improvement Work

- 损失函数为主任务的对数似然损失与language model的对数似然损失的加权和:

$$\overrightarrow{E} = -\sum_{t=1}^{T-1} log(P(w_{t+1}|\overrightarrow{m_t}))$$

$$\overleftarrow{E} = -\sum_{t=2}^{T} log(P(w_{t-1}|\overleftarrow{m_t}))$$

$$\widetilde{E} = E + \gamma(\overrightarrow{E} + \overleftarrow{E})$$


Dense
Mean Pooling
Max Pooling
Word Embedding
Lexical Features
LM hidden states
<s>
follow
People
courage
follow
</s>
Language Modeling
People
follow
courage
Word Embedding
Lexical Features

Improvement Work

- 最后我们可能尝试在输入中加入NER、POS等特征，识别敏感词汇，对数据集中包含性别、种族等专有名词的subgroup进行识别，并对subgroup进行不一样的处理，如对subgroup加一个新的隐含层对其敏感词汇进行判断评分进一步让总体模型达到更好效果，实现竞赛的预期。