中国大学生服务外包
创新创业大赛

inspur 浪潮

【A01】2018 网络零售平台商品分类【浪潮】

--X’团队解决方案

NORTHWESTERN POLYTECHNICAL UNIVERSITY 1938 西北工业大学

问题定义

商品标题　腾讯QQ币148元148QQ币148个直充148Q币148个Q币148个QQB★自动充值

预测目标　本地生活--游戏充值--QQ充值

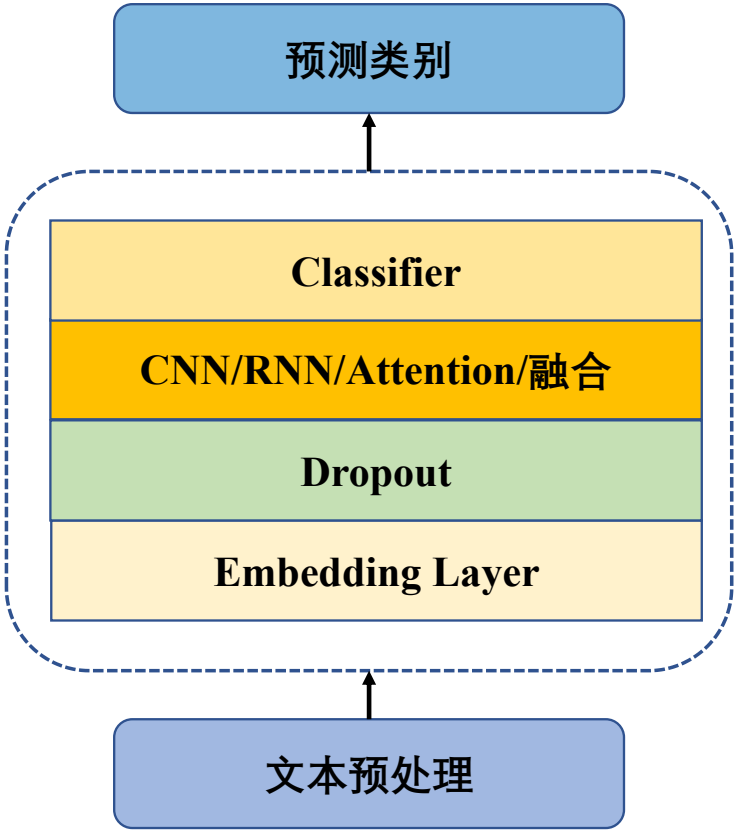
- 1258类商品
- 训练集50万条商品标题
- 测试集450万条商品标题

评估指标：

$$accuracy = \frac{\text{分类正确样本数}}{\text{总的样本数}}$$

基础架构

- ◆文本预处理
- ◆语言模型(Word2vec/Glove/Char)
- ◆分类模型(CNN/RNN/Attention)
- ◆Web可视化交互界面(flask)



数据预处理

腾讯QQ币148元148QQ币148个直充148Q币(148个Q币)148个QQB★自动充值

去除干扰字符 ⇓ 去除停用词

腾讯QQ币148元148QQ币148个直充148Q币148个Q币148个QQB自动充值

⇓ 过滤低频词

腾讯/QQ/币/元/QQ/币/个/直充/Q币/个/Q币/个/QQ/自动/充值

转化为编号 ⇓ 长度不足补0，多余截断

0 0 0 0 0 0 10 301 209 3002 301 209 23 32 390 23 390 23 301 20 21

语言模型

字向量(char vector)

词向量(word vector)

- ◆Word2Vec
- ◆GloVe
- ◆随机初始化

实验结果

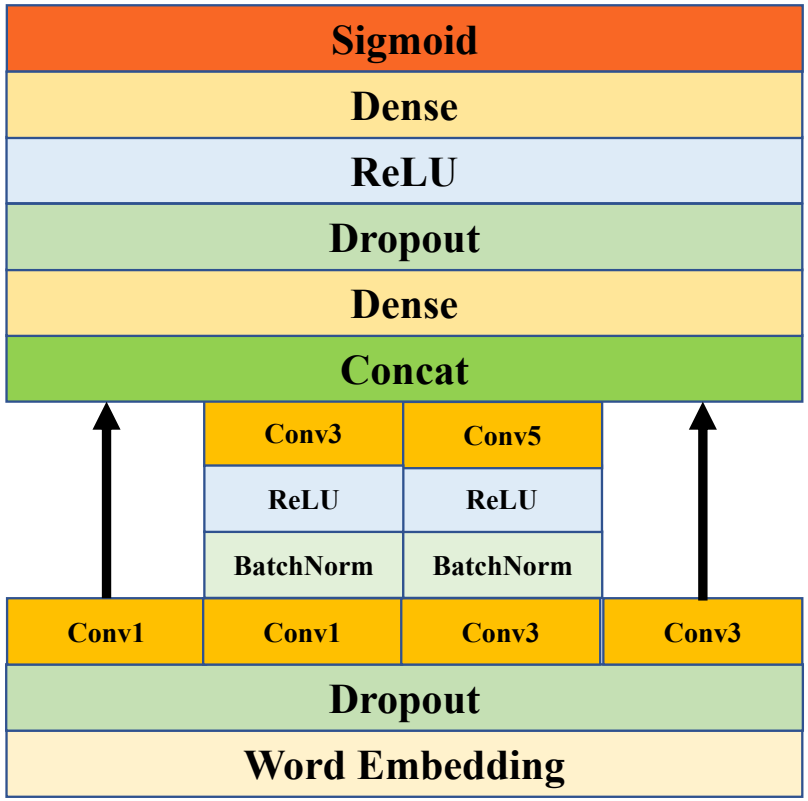
词向量(0.8597)>字向量(0.838)

Word2vec(**0.8615**)>GloVe(0.8595)

GloVe (0.8595 >随机初始化(0.8580)

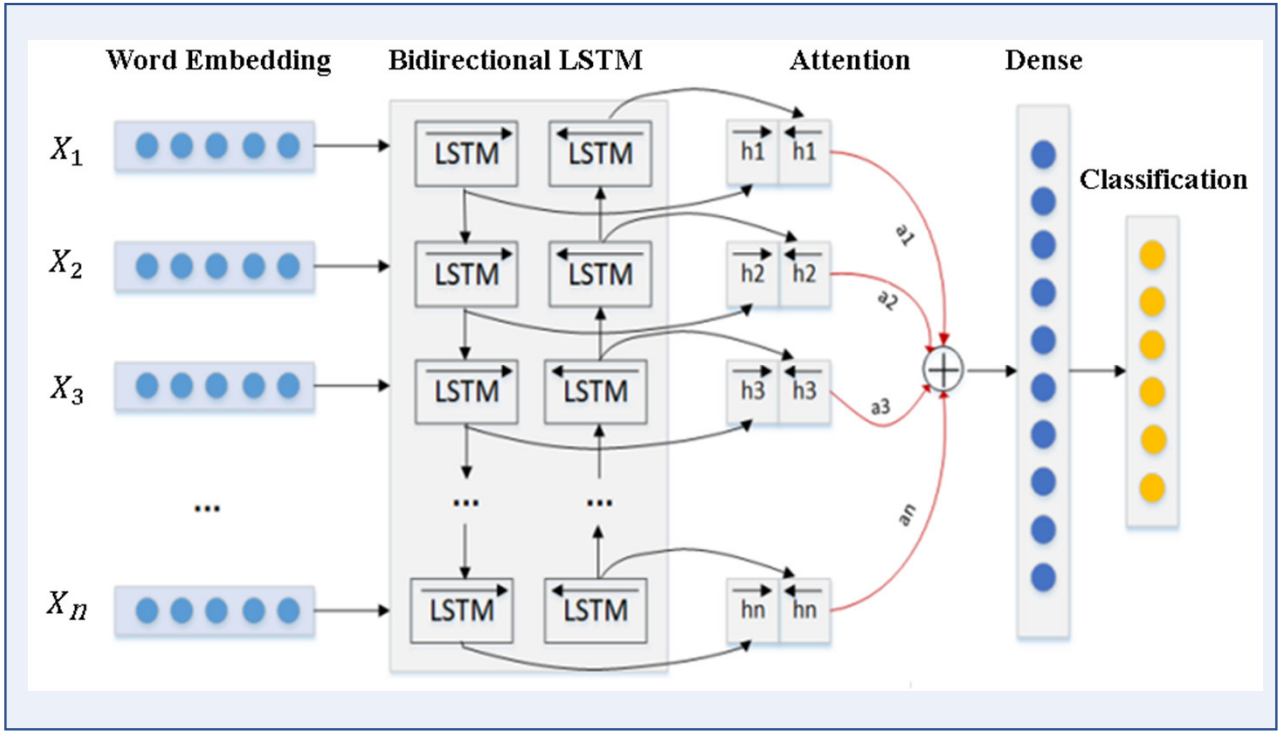
TextCNN

使用多尺度卷积核提取商品标题中不同尺度(N-Gram)的语义信息。



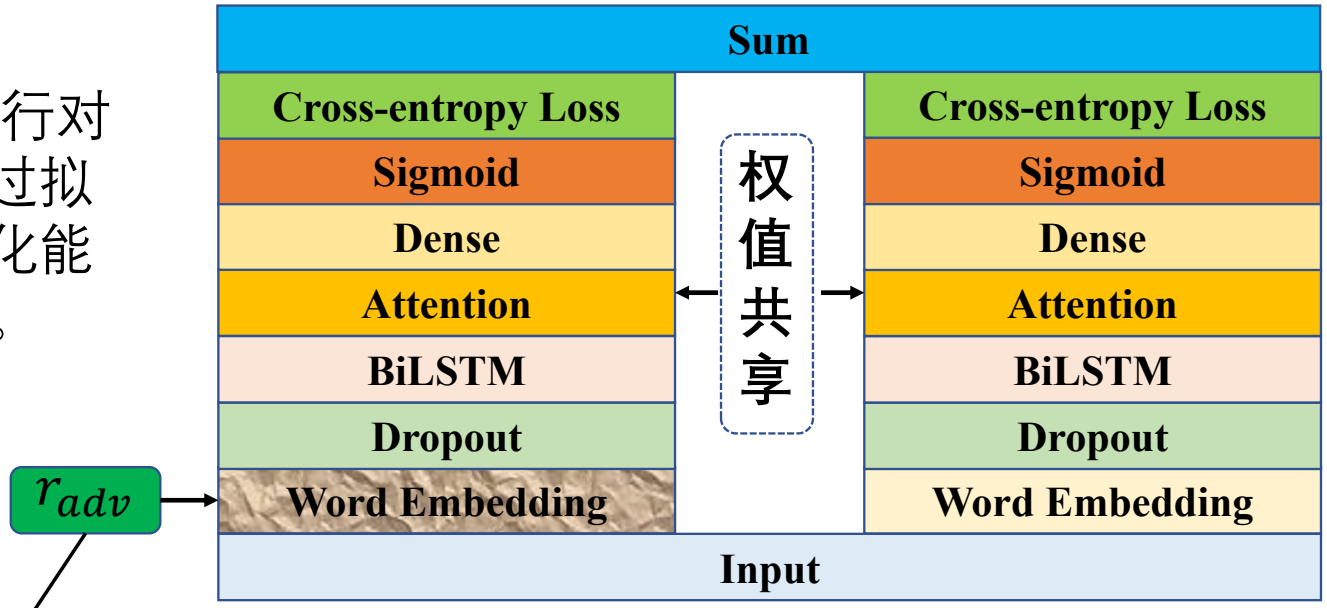
BiLSTM-Attention

利用Bi-LSTM提取文本序列的特征，结合注意力机制赋予特征不同的权重。



Adversarial-BiLSTM-Attention

在Word Embedding 层进行对抗性扰动，缓解模型的过拟合问题，提高模型的泛化能力，起到正则化的作用。



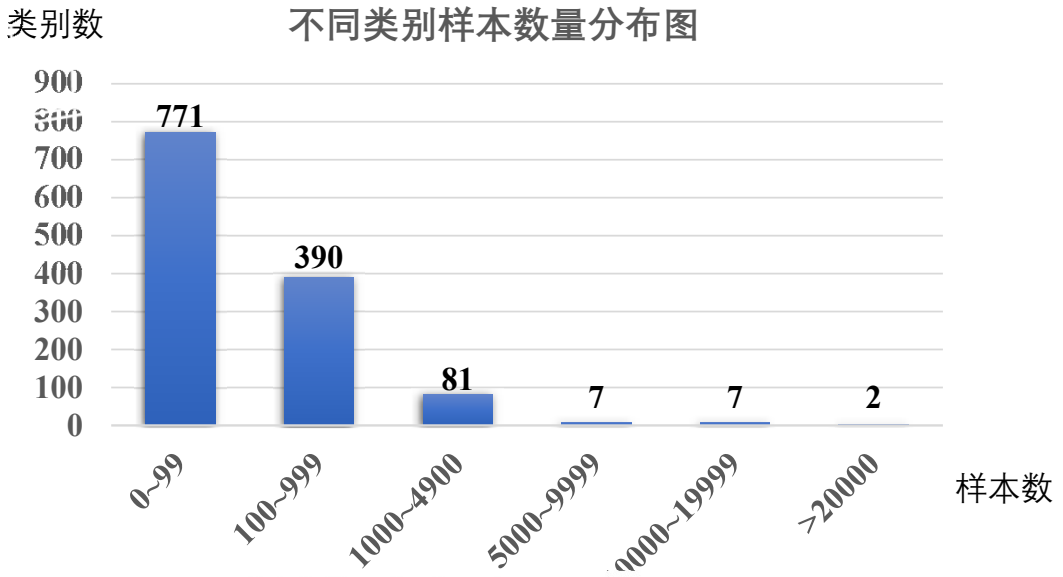
$$r_{adv}=\underset{r,\|r\|\le\varepsilon}{argmin}\ logp(y|x+r;\hat{\theta})\approx-\varepsilon\frac{g}{\|g\|_2},\quad g=\nabla_X logp(y|x;\hat{\theta})$$

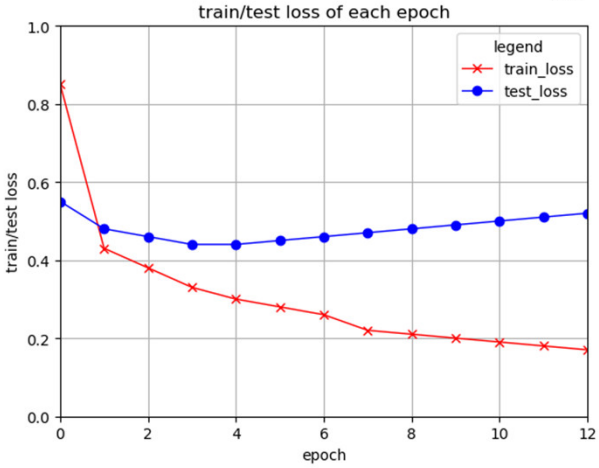
初步实验结果

模型在**50w**数据集上的表现(训练集:测试集=40w:10w)	
Model	**Accuracy**
TextCNN	0.848
BiLSTM	0.860
BiLSTM-Attention (Char Embedding)	0.838
BiLSTM-Attention (Word Embedding)	0.861
Adversarial-BiLSTM-Attention（Char Embedding）	0.844
Adversarial-BiLSTM-Attention（Word Embedding）	**0.871**

问题分析

- ◆过拟合问题
- ◆类别不均衡问题
- ◆标签存在错误(6±1%)





改进技巧

◆采用Focal Loss作为训练损失函数用于缓解数据不均衡问题 (0.8)

$$FocaLoss(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)\ ,\ \gamma{=}2, \alpha_t{=}0.25$$

◆ 对50w数据集中标签存在错误的样本进行数据清洗 (0.7)

◆ 将标签根据类别进行分层，训练3个模型分别用于预测不同级别的标签 (1.2)

改进技巧

◆利用爬虫在京东商城上爬取*500w*外部数据，结合迁移学习的思想利用该部分数据预训练语言模型和分类模型。 (0.7+0.5)

◆采用*Self-training*半监督学习方法,挖掘*450w*测试集的信息，并利用该部分信息训练模型。

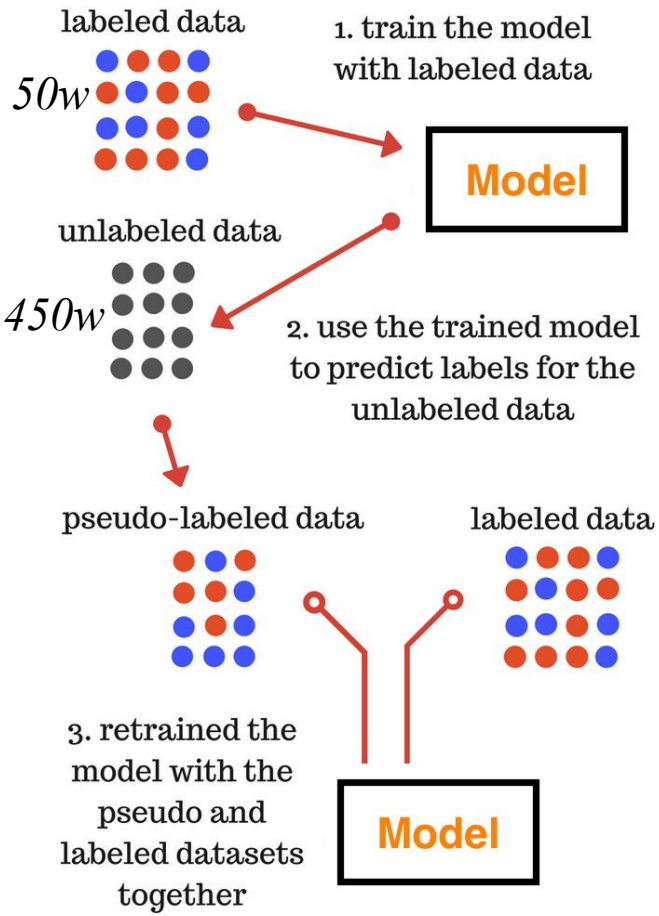
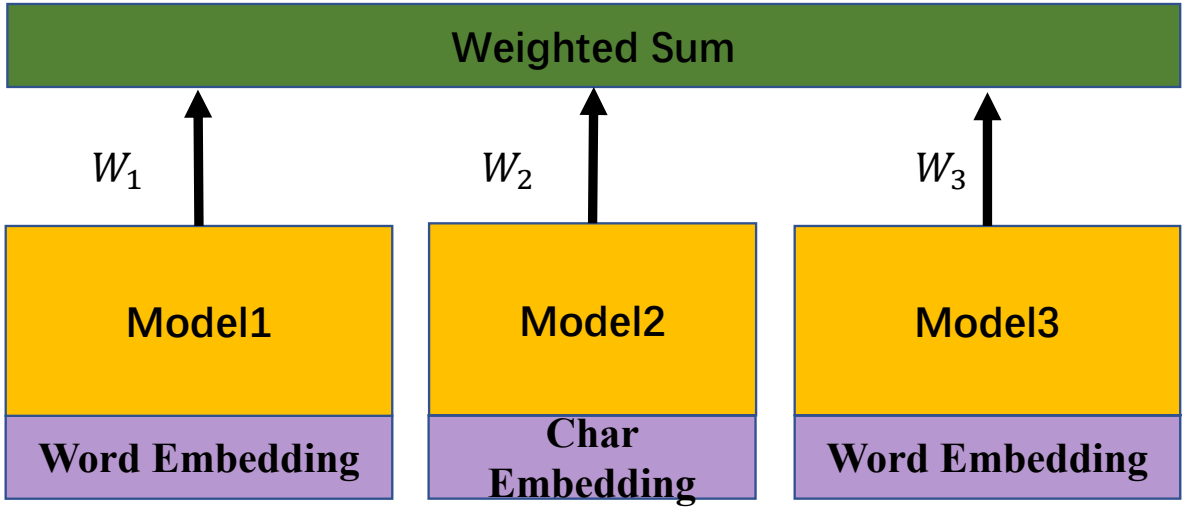


Figure. *Self-training of semi-supervised learning*

模型融合

- ◆加权融合（1.0）
- ◆Stacking融合



改进结果

模型的分类准确率(训练集:40w+半监督+爬虫数据验证集:10w)	
model	**accuracy**
Multi-Head-Attention	0.9073
BilSTM	0.9156
[0.42BiLSTM+0.58Attention](加权融合)	0.9194
[0.67BiLSTM+0.09Attention+0.24BiLSTMAttention](加权融合)	0.9201

最优模型

采用model1、model2以及model3进行模型融合，最终取在该模型上取得了92. %的最高分类准确率。

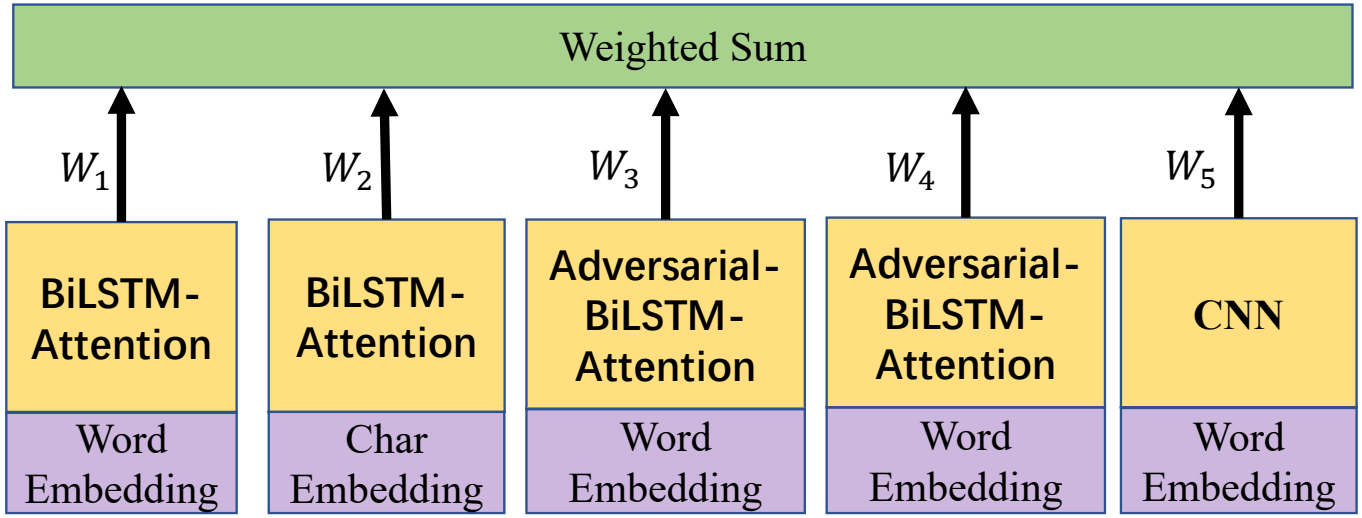


Figure. 加权融合

改进与思考

模型压缩技术：

(模型加速)：

- ◆矩阵分解
- ◆网络剪枝
- ◆知识蒸馏

反思

(面向应用场景)：

- ◆采用增量学习方法(Incremental Learning)