2018 CCF BDCI 6th
大数据与计算智能大赛
Big Data & Computational Intelligence Contest

《汽车行业用户观点主题及情感识别赛题》

团队名称：Just a test
团队成员：吴震、范志方
答辩人：范志方

目录

PART 1

团队简介

PART 1 团队简介

队长：吴震

南京大学计算机系自然语言处理研究组博士研究生，导师为戴新宇教授。主要研究方向为情感分析、文本挖掘、舆情分析，曾在AAAI、NLPCC等会议上发表文章。

队员：范志方

南京大学计算机系硕士研究生在读二年级，跟随南京大学自然语言处理研究组戴新宇教授从事NLP的研究。

PART 2

赛题介绍

PART 2 赛题介绍

数据背景：用户对汽车相关内容的讨论或评价。

任务目标：

- 根据用户的评论文本识别出用户讨论的主题和各主题对应的情感极性。
- 主题被分为10类，包括：动力、价格、内饰、配置、安全性、外观、操控、油耗、空间、舒适性。
- 情感分为3类，分别用数字0、1、-1表示中立、正向、负向。

数据集：

- 训练集：12572条句子。
- 测试集：2753条句子。

评论文本	主题	情感极性
2.5豪华，森得空间比q5都大，操控也挺棒	空间	1
	操控	1

评价指标：以<主题，情感>为pair的 F-1 score.

PART 3

任务分析

PART 3 任务分析

任务定位：Aspect-level情感分析。

解决方案：采用pipeline的方式，即先预测主题，再根据主题预测对应的情感极性。

- 主题分类：多标签的文本分类任务；
  - 挑战1：多标签分类
  - 挑战2：文本表示
- 主题情感分类：基于角度的情感分类任务(ABSA)，输入为句子和主题的三分类问题；
  - 挑战1：同一个句子不同的主题可能有不同的情感极性。
  - 挑战2：建模句子和主题的语义关系

PART 4

解决方案

PART 4 解决方案——多模型主题分类

- 模型1： Multi-Label Multi-Attention Model (MLMA)

  - 为每个主题(label)学习一个embedding，用label embedding来做attention，直接学习到和主题相关的词；
  - 每个主题都有一个独立的attention过程，从而学习到各个label对应的句子表示，适合于多标签分类；

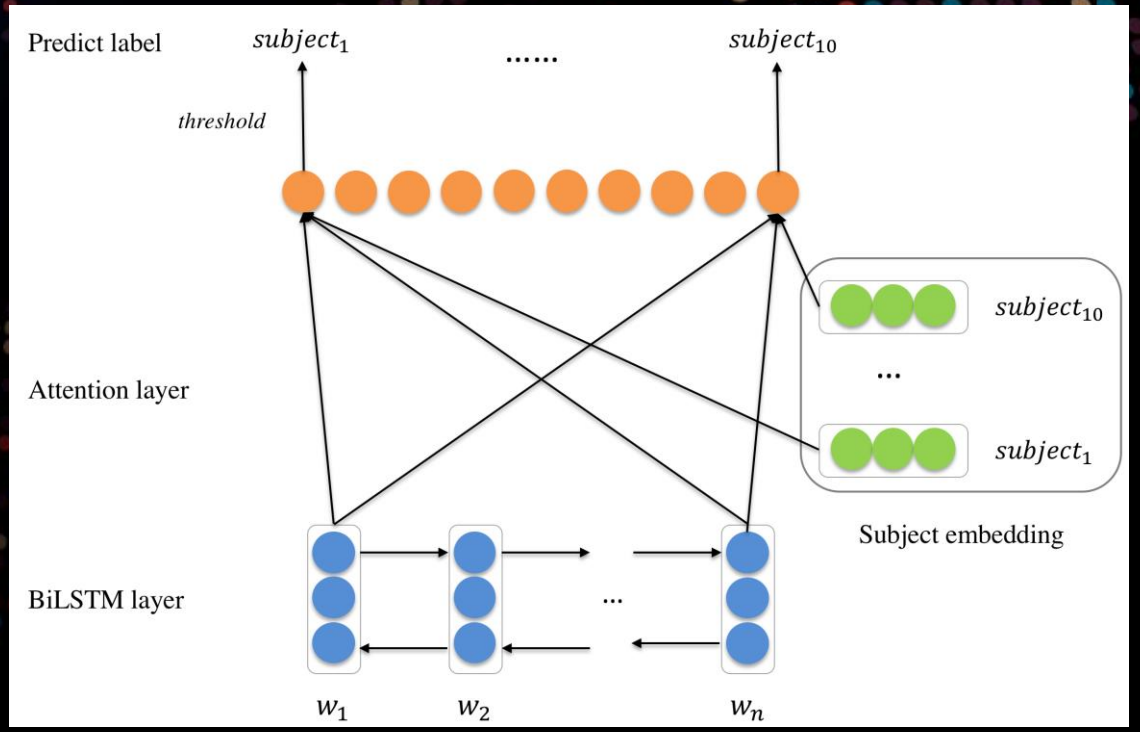


图1：Multi-label Multi-attention Model

- 模型2： 卷积神经网络（CNN）

PART 4 解决方案——多模型情感分类

- 模型1： Attention LSTM[Wang et al., 2016]：单层
- 模型2： HEAT[Cheng et al., 2017]：两层attention
- 模型3： GCAE[Xue and Li, 2018]：门控CNN

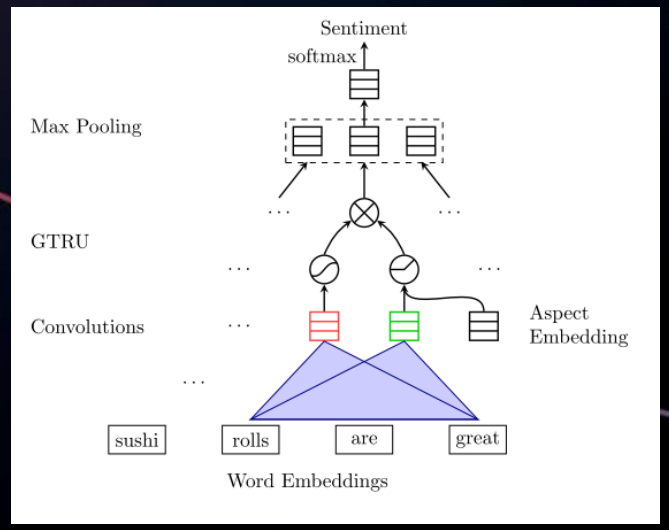


图3：GCAE Model

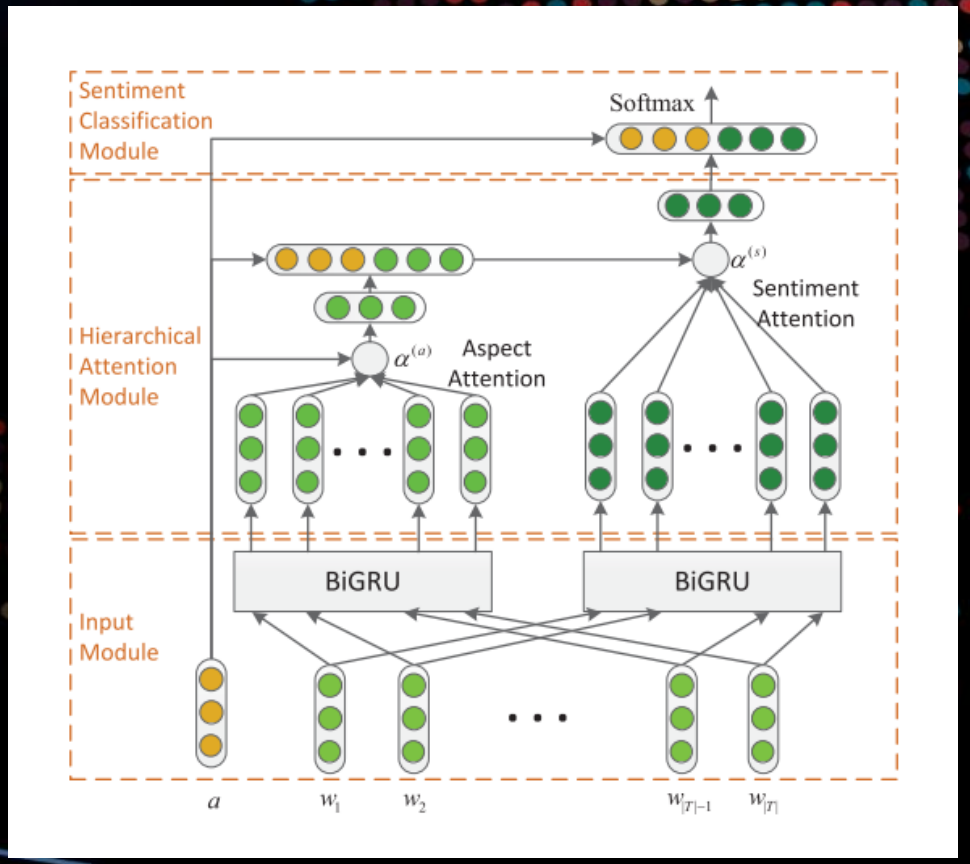


图2：HEAT Model

PART 4 解决方案——引入多种词向量

- Chinese-Word-Vectors [Li et al., 2018]
  (https://github.com/Embedding/Chinese-Word-Vectors)
- Fasttext [Joulin et al., 2016]
  (https://fasttext.cc/docs/en/crawl-vectors.html)
- Tencent AI Lab Embedding [Song et al., 2018]
  (https://ai.tencent.com/ailab/nlp/embedding.html)
- 中文elmo，来自于ELMoForManyLangs [Peters et al., 2018]
  (https://github.com/HIT- SCIR/ELMoForManyLangs)

PART 4 解决方案——引入更深的模型：BERT

- 我们改写了BERT代码，使其能支持多标签文本分类和基于主题的情感分类的问题
- Self-attention可以建模句子和主题之间的语义关系
- 模型很深（12层），有更强的捕捉语义的能力

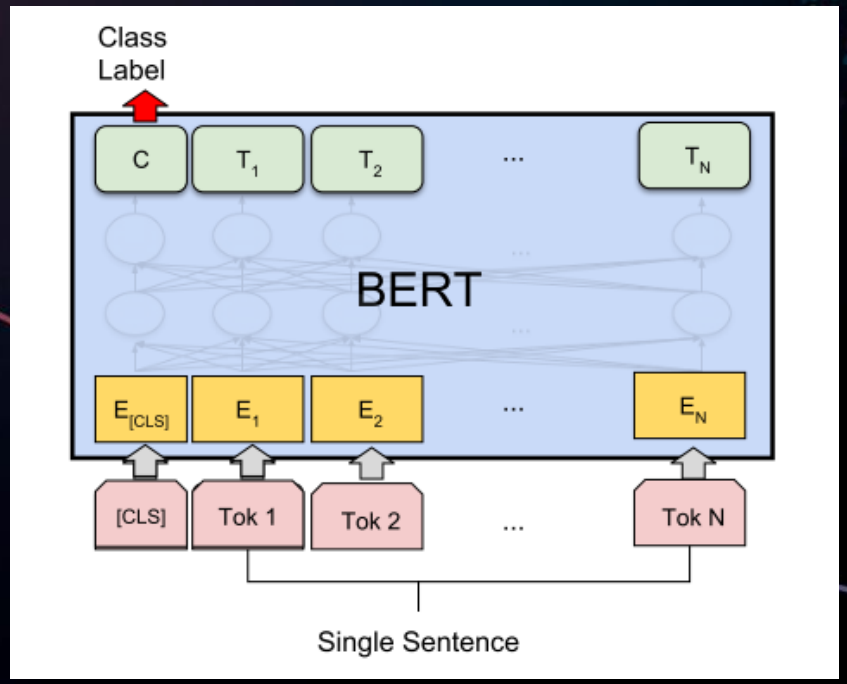


图5：主题分类

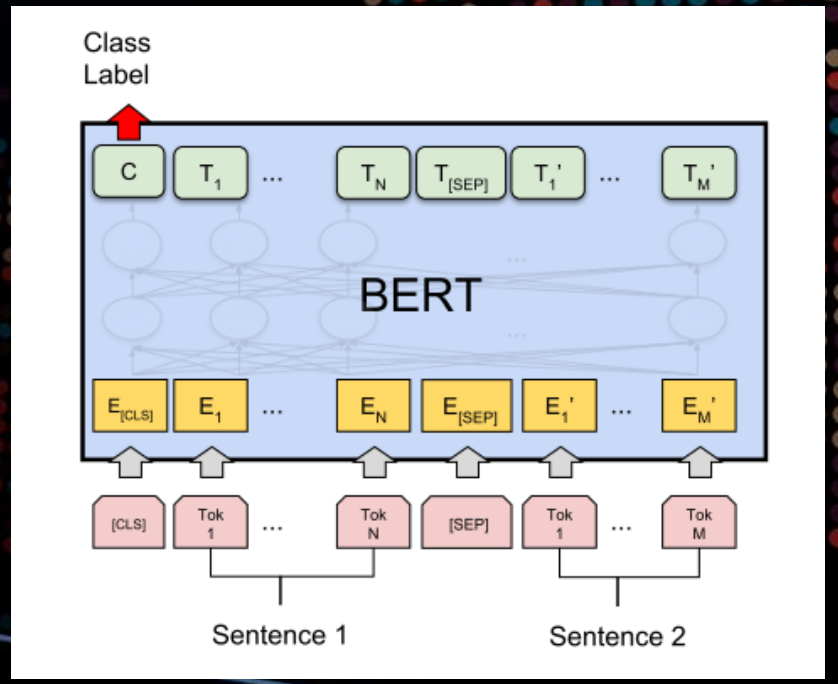
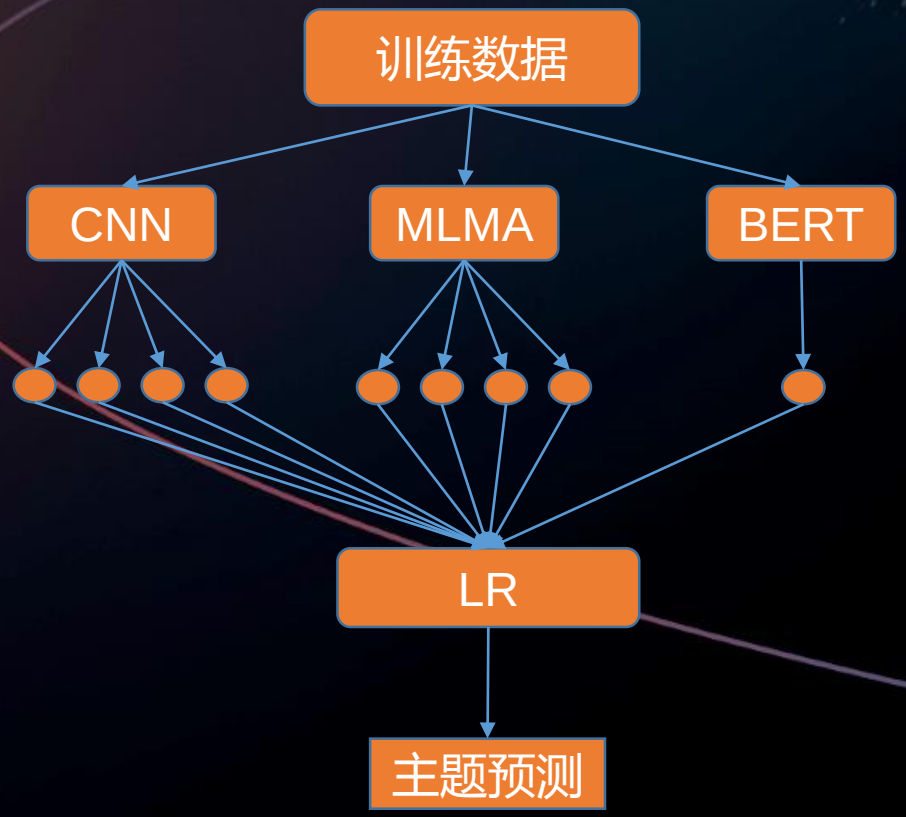
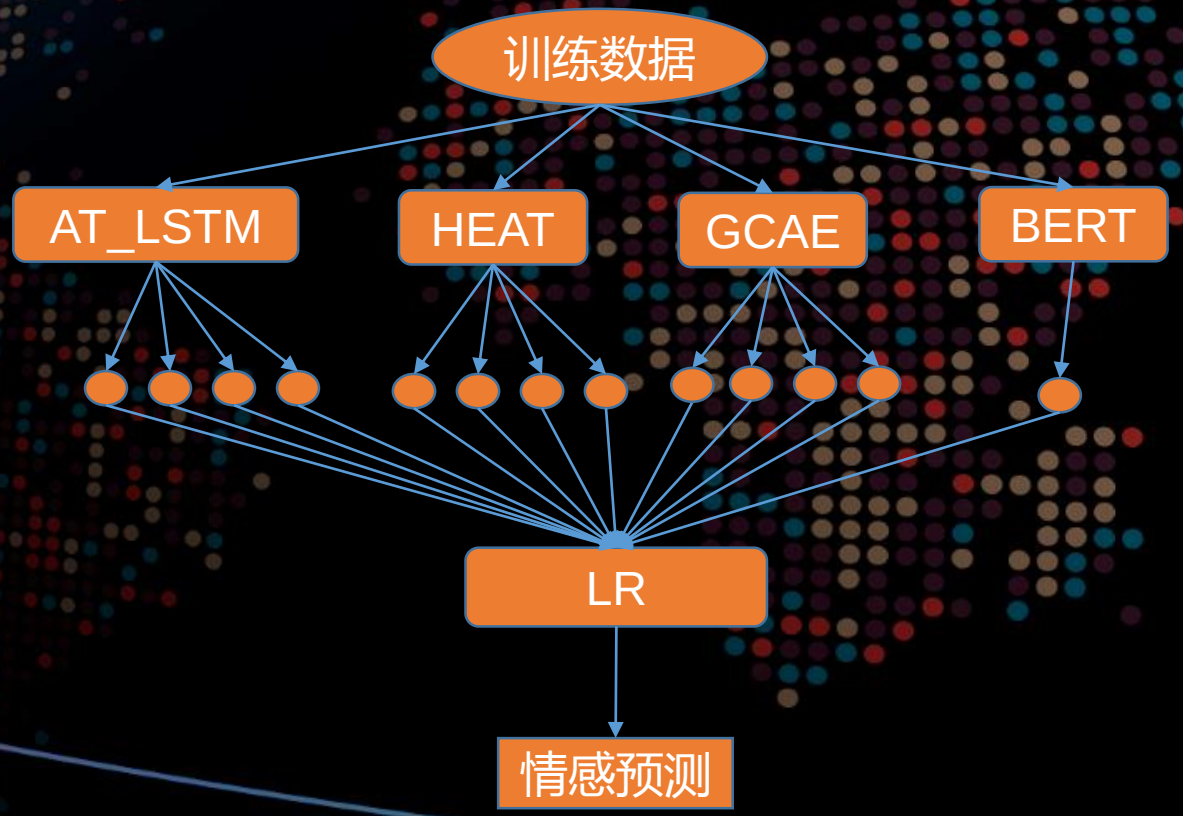


图6：情感分类

PART 4 解决方案——多模型融合的集成学习

- Stacking：第二层模型采用Logistic Regression。





PART 4 解决方案——改进路线

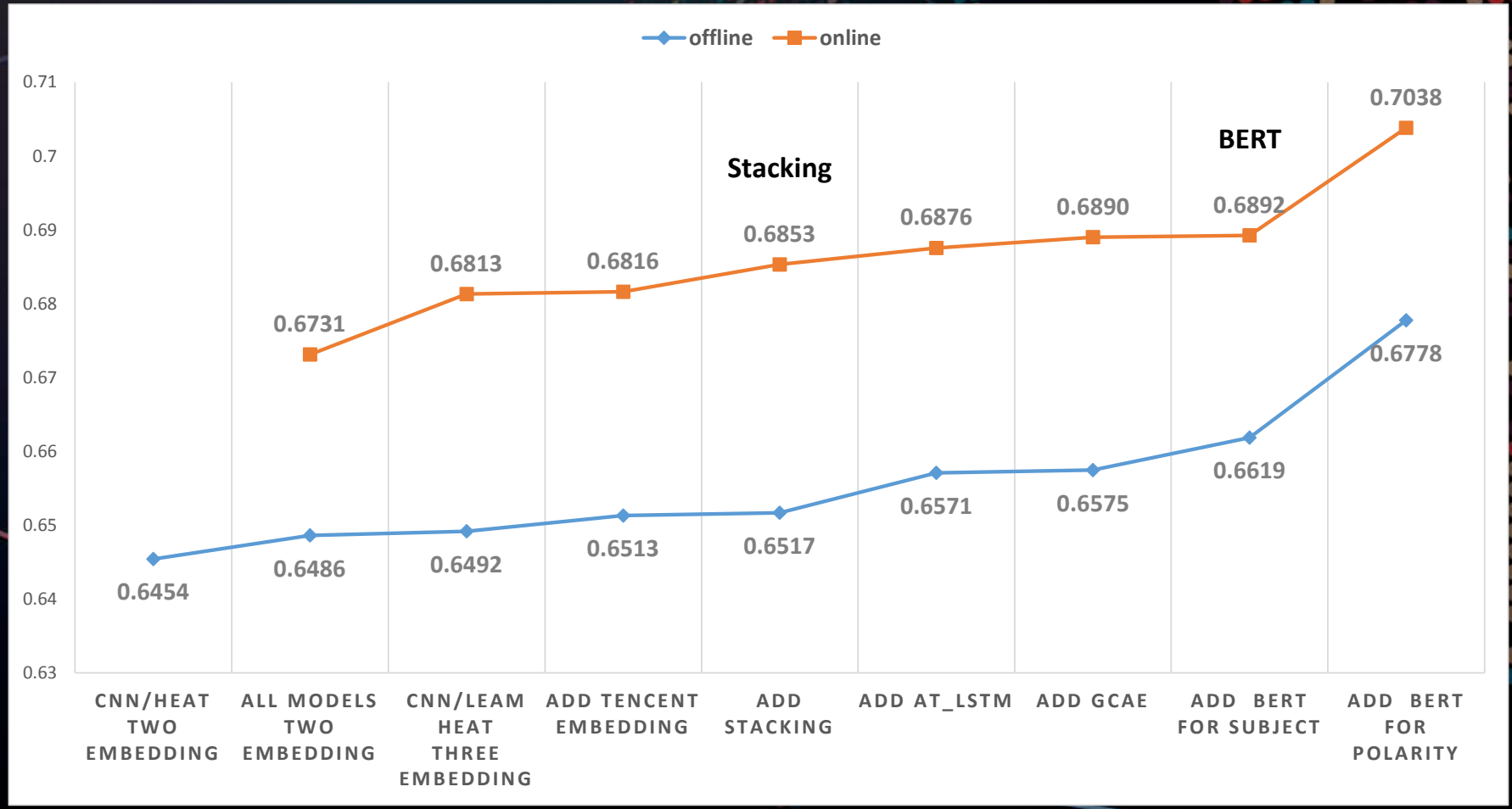


图7：改进路线

PART 5 系统亮点总结

- 深度学习：BERT模型有12层比其他模型要深很多，能够建模复杂语义
- 集成学习：模型结构的差异性+输入特征的差异性（和而不同）
- 迁移学习：将基于大规模语料的BERT在当前数据集上fine-tune
- 任务定制化：
  - 设计了Multi-Label Multi-Attention模型
  - 首次将BERT引入到多标签分类和基于角度的情感分类任务
- 完全数据驱动：无人工规则，完全基于数据驱动，不需要人工干预

PART 6

未来改进

PART 6 未来改进

- 利用标注好的情感词信息。
- 抓取汽车语料对BERT做无监督的领域迁移。
- BERT的压缩和裁剪，使其适用于工业界的需求。
- Multi-task learning。

Reference

1. Cheng, J., Zhao, S., Zhang, J., King, I., Zhang, X., and Wang, H. (2017). Aspect-level Sentiment Classification with HEAT (HiErarchical ATtention) Network. Proceedings of the 26th ACM International Conference on Information and Knowledge Management, pages 97–106.
2. Kenton, M.-w. C., Kristina, L., and Devlin, J. (2017). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
3. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification.
4. Wang, Y., Huang, M., Zhao, L., and Zhu, X. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 606–615.
5. Xue, W. and Li, T. (2018). Aspect Based Sentiment Analysis with Gated Convolutional Networks.

THANKS

2018 CCF BDCI 大数据与计算智能大赛
2018.12.01

开源地址：
https://github.com/yilirin/BDCI_Car_2018