自媒体文章推荐

目标

鉴别优质的自媒体文章以便于进行后序的推荐等任务

难点

影响人阅读体验的因素繁多

- 多种形式的载体, e.g. 文字, 视频, 图片,
- 多样的写作风格和写作习惯
- 迥异的文段布局
- 巨大的内容跨度

思路

模仿人阅读的基本流程, 从文章的组织布局 (直观感受), 写作特征 (浅层观感), 语义逻辑 (深层内容) 三个方面分别建模再汇总, 将问题建模为有监督二分类问题

要点

组织布局网络

文章切割成文本块, 图片块等等, 人工定义每块的特征描述

- 每块的长度, 宽度, 离标题的距离, etc
- 变长, 序列, 主要参考了手机等移动设备上的阅读习惯

使用GRU进行整体布局的建模,
使用 1D-CNN 进行局部布局的建模

- GRU 模拟人从头到尾浏览文章的顺序性

最终输出是 GRU 和 1D-CNN 的输出连接

写作特征网络

人工设计底层特征

- 文本
  - 文本长度, 段落总数, etc
- 图片, 视频
  - 清晰度, 图中文字占比, etc
- 综合
  - 图文占比, etc

self-attention 构造组合特征

- embedding转化类别和数值特征
- attention 得到组合特征
- 多个attention得到多个组合特征
- 多层attention得到多个层级的组合特征
- 输出为最终的组合特征的连接

语义逻辑网络

句子层级

- 精调BERT
- 得到各个句子的表示

文档层级

- 双向transformer
- 利用各个句子的表示得到文档的表示

网络综合

子网络输出各自过一个全连接, 将输出结果拼接,
拼接结果过全连接后 sigmoid

训练

有监督, 二分类

一些讨论

- 语义逻辑子网络作用显著
- 文本长度和图片中文字总数发挥了重要作用
- 果然文章长了大家就不看