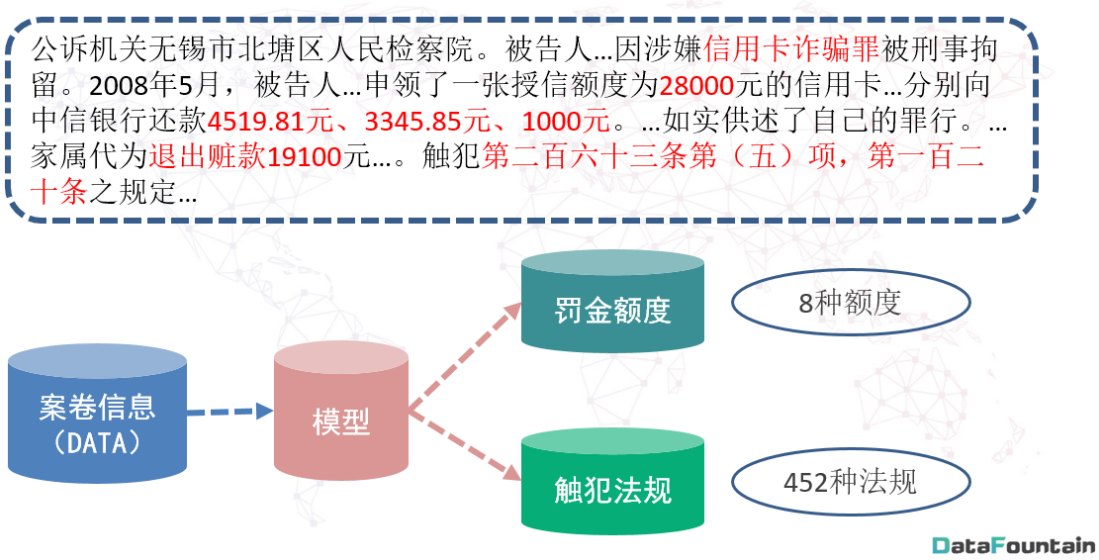
成员介绍

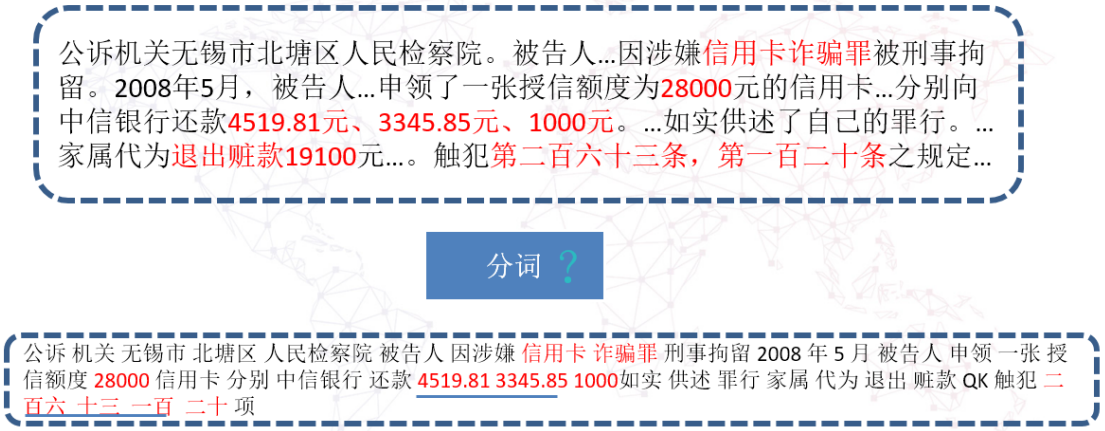
杨亚涛（sysu_Yang）：中山大学数据科与计算机院 17 级博士
刘家豪（kaho）：中山大学数据科与计算机院 15 级硕士
庄业广（Hans）：中山大学数据科与计算机院 17 级硕士
许海城（Seaty）：中山大学数据科与计算机院 18 级硕士
罗志鹏（get_max）:微软

赛题分析

本次比赛的任务是根据犯罪人员的案卷文档，训练模型并预测犯罪人员被罚的罚金额度以及所触犯的法规。罚金额度是有根据金额大小被分为 8 种。法规是有 452 种。根据官方给出信息可知，一个样本只有一种罚金，但可能会触犯多种法规。后面也会根据任务的不同来设计不同的分类器。

数据分析





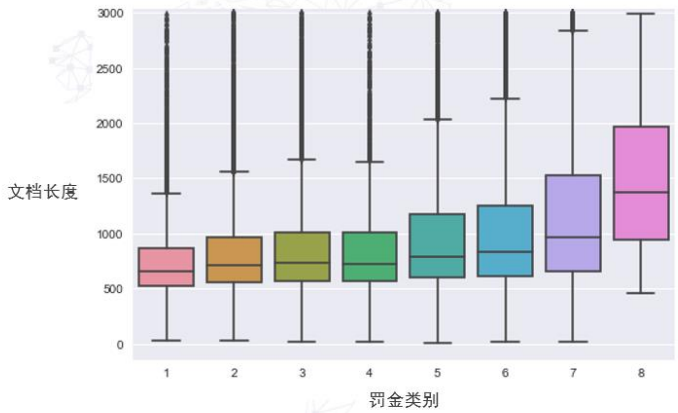
数据预处理

按照中文文本分类的传统思路，原始文本信息是需要分词。如果我们不对文本进行预处理的话会出现两种情况。

第一，4519 可能和 3345 同处在一个罚金范围，但是我们分词后是当作两个词汇来处理的。第二就是二百六十三条可能会被分为二百六和十三。

针对以上问题，我们是需要对数据进行一些预处理。首先我们将文中涉及到的金额（阿拉伯数字，中文数字，两者混合的）先统一转化为阿拉伯数字，然后构建金额映射表将某个区间的金额都映射到一个唯一的标识符。另外我们根据正则表达式将法规条数提取出来使得它分词时候不会被处理。

统计分析



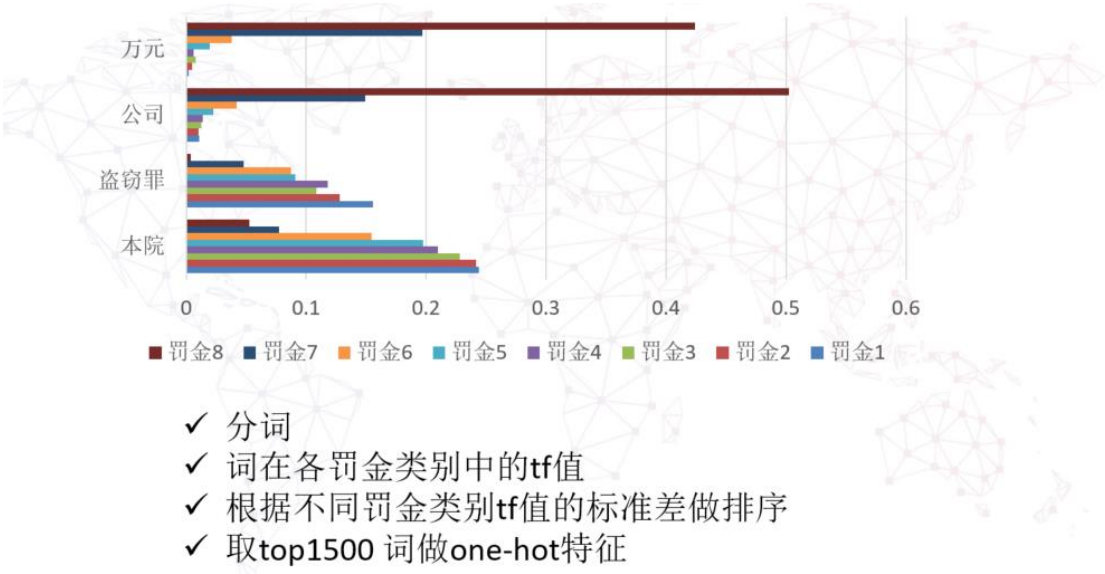
- ✓ 统计各罚金下文档长度
- ✓ 罚金越重，长度越长

此外，我们还做了一些对数据的统计分析。统计各罚金类别下文档的长度分布，发现罚金越重，长度会越长。

- ✓ 统计出现在金额前七个词汇
- ✓ 词汇分为四大类
- ✓ 做每个类别下金额的统计特征

std	sum	mean	max	min	ptp

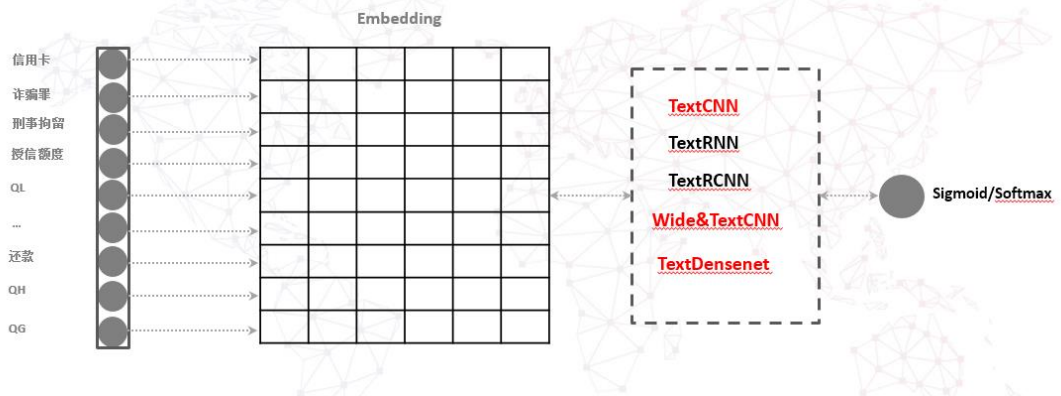
另外，为了关注文档中金额含义的不同，我们统计出现金额的前 7 个词汇。词汇被分为四大类。罚金类，金融类，汇总类，刑事类等等。从而来统计不同类别的金额的一些统计特征，将金额统计进行细分化。



为了提取出对文本类别判断比较有用的词汇，我们分词后会统计不同罚金类别下词汇所出现的频率，根据不同类别下出现频率的标准差来表示词汇对于不同类别的区分度，然后依照标准差来取 top1500 个词汇做 onehot。用于后面的模型。

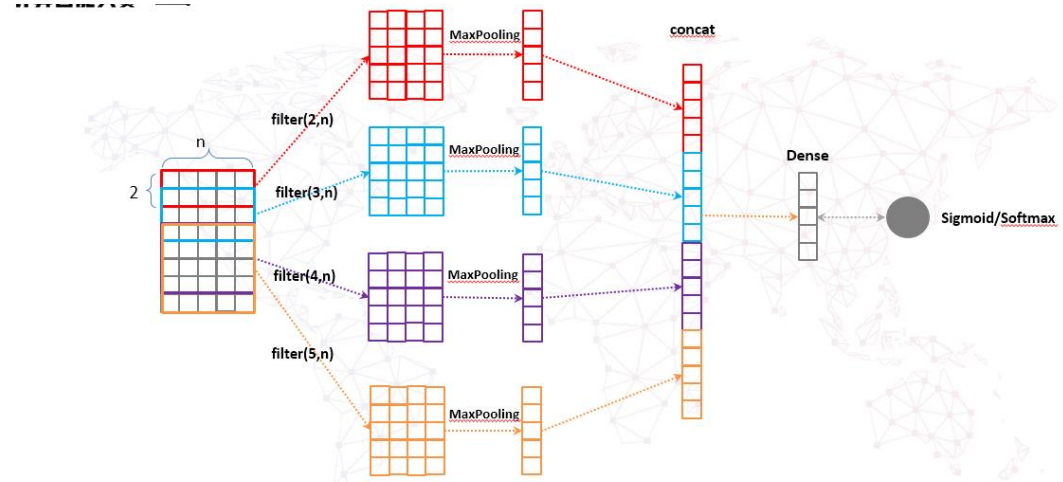
模型设计及分析

框架结构



首先，介绍下模型的框架结构。一系列分词后的词汇会代表案卷的原始文本信息，我们通过 embedding 将词汇转换为一个向量来表述，整体的 embedding 矩阵将会表述整个案卷的信息，之后会使用常见的 TextCNN/RNN 以及我们本次比赛中做的一些改进的模型。最后根据任务的不同使用不同的分类器激活函数。

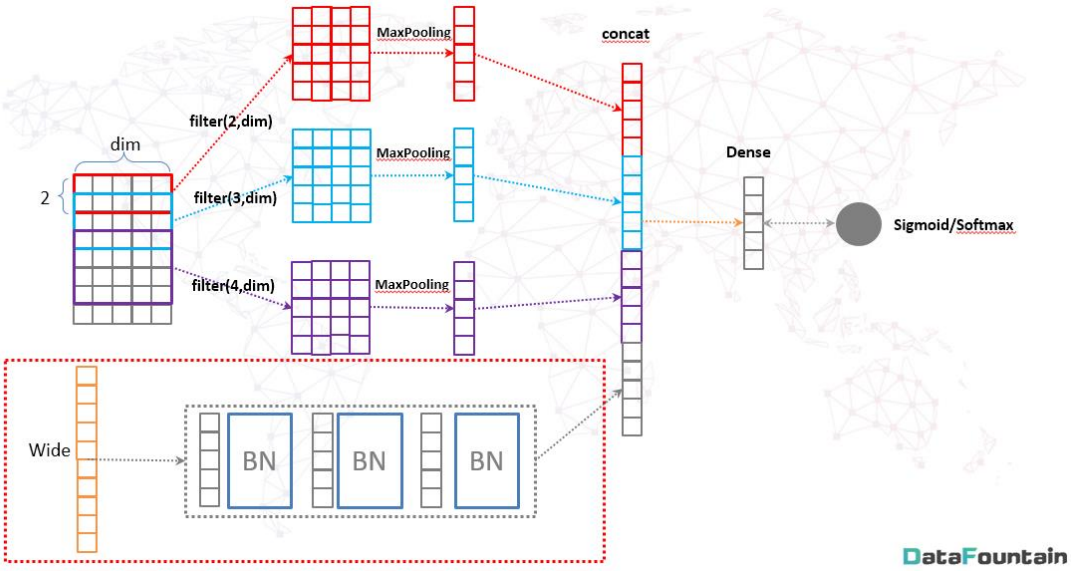
TextCNN



- Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
- Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [J]. arXiv preprint arXiv:1510.03820, 2015.

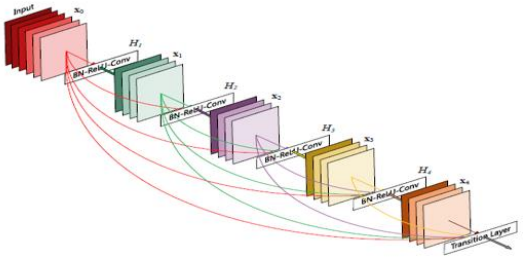
TextCNN 模型，该模型是文本分类中常用的模型，通用的步骤是设定卷积的 kernel 为 2*n 来对相邻的两个词汇进行卷积，多个卷积核可以产生多个 featureMap，最后会通过 MaxPooling 来降低模型复杂度以及提取 featureMap 的关键信息。后面也有相关论文指出可以定义不同的 kernel 来同时构造多种 featureMap 最后再 concat 各自 kernel 下学习的 feature 来学习会带来更好的效果。

Wide&TextCNN



接下来是我们对 TextCNN 做的一些改进，在数据分析部分我们统计了一系列特征，在此被称为 Wide 部分。红色虚框部分是我们的人工统计特征神经网络结构，我们将利用三层全链接来学习 Wide 部分的内容，在每个全连接后加 BatchNormalize 来防止梯度消失问题。Wide&CNN 可以实现人工特征与神经网络自动特征的结合，提高了 CNN 的效果。

DenseNet



◆ **省参数**

在 ImageNet 分类数据集上达到同样的准确率，DenseNet 所需的参数量不到 ResNet 的一半

对于工业界而言，小模型可以显著地节省带宽，降低存储开销。

◆ **省计算**

达到与 ResNet 相当的精度，DenseNet 所需的计算量也只有 ResNet 的一半左右

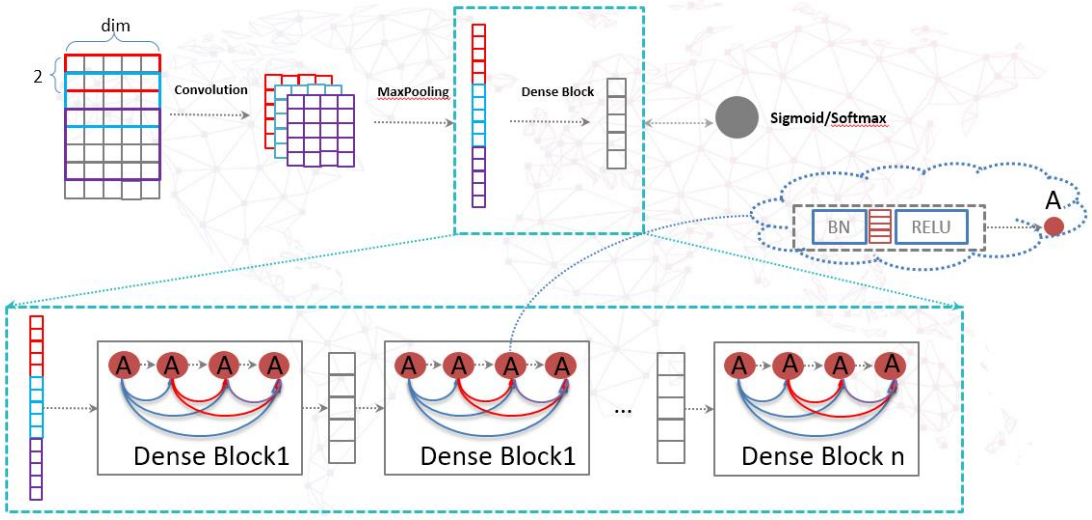
◆ **抗过拟合**

DenseNet 具有非常好的抗过拟合性能，尤其适合于训练数据相对匮乏的应用

□ Huang G, Liu Z, Weinberger K Q, et al. Densely connected convolutional networks[J]. CVPR 2017 Best Paper

另外，我们本次比赛也对 textCNN 做了其他结构的改进。在之前参加的一些比赛或现在研究论文中发现，TextCNN 中卷积的层数最多达到 2，再加深没有多大的效果。可能是因为文本数据和图像数据的差别还是很大的。今年 CVPR 的 Best Paper 提出了 DenseNet 的结构，该结构使得每一卷积层都会利用到前面所有层所学习到的 feature，使得达到很高准确率的同时节省模型参数，节省计算时间，以及有很好的抗过拟合的能力。另外无论是图像中的 ResNet 还是 DenseNet 都会证明出 当你的网路结构很深的情况下，当你的输入层信息越接近于输出层就会得到越好的效果。基于此，我们在想是否可以对 TextCNN 做深度网络测试。

TextDenseNet

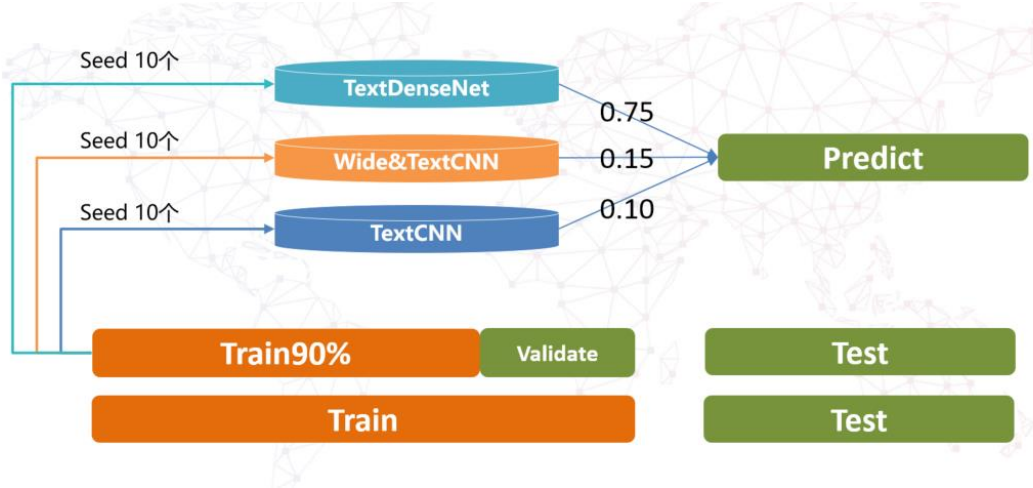


通过实验我们发现对卷积进行 DenseNet 形式的加深是不会对效果起到很好的作用，并且训练的时间会越来越慢。后面我们调整 Concat 后的全链接结构时会发现对结果有很大的影响。我们认为传统的 concat 后接全链接是利用信息不充分的。后面我们就转向不对卷积进行加深而是对 Pooling 后的 feature 进行加深。我们设定了 A 结构，其由 BN，全链接，RELU 组成 ，使得 feature 激活的同时能尽量避免梯度消失，然后我们每个 A 结构都会利用到前面所有 A 结构的信息使得深度学习的同时能够使得每一层的信息距离输出层都很近。在这里，我们每个 Dense Block 是由四个 A 结构组成。在罚金中四个 DenseBlock 是最好，法规中是 1 个到两个。

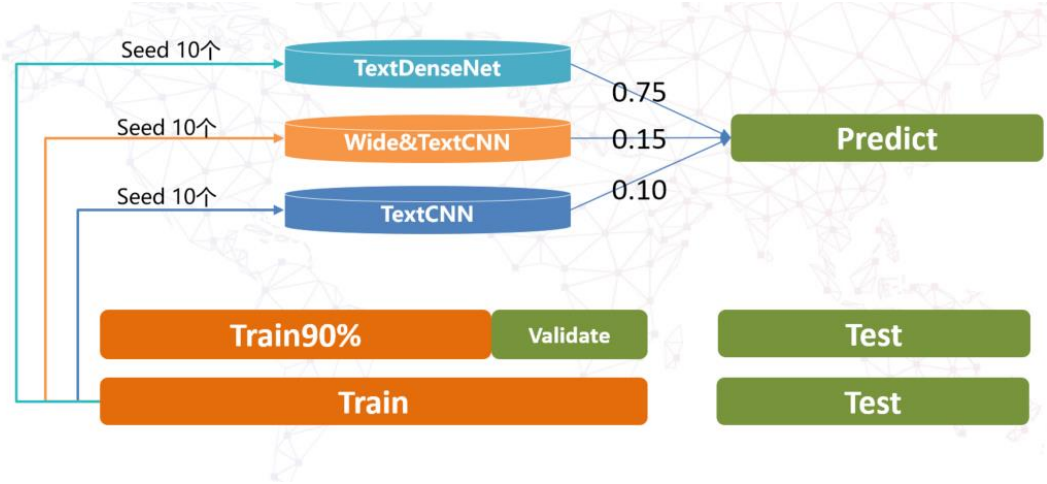
效果对比

另外，我们对各类模型效果做了以下对比。传统模型 TextCNN 是效果最差的。Wide&TextCNN 是要比传统的 TextCNN 要好。TextDenseNet 是要比两种模型结构都好很多。并且单 TextDenseNet 模型的效果会达到排行榜的第二。另外我们也做了其他模型的实验如 TextRNN，RCNN 等等。由于自己硬件条件限制只有一个 1080，太耗时，并且初步试验效果不佳。就没有使用。

模型融合



首先，我们先将线下训练集按 9：1 比例划分训练集和验证集，利用验证集来选择最优参数以及最优轮数，最后不同模型按照 10 个不同的 seed 训练出的 10 个模型。最后三个模型按照 0.75，0.15.0.10 的比例进行融合。



通过验证集已经获得了最优参数和最优轮数，然后我们就扩展到全部数据集上进行同样的模型训练，然后以同样的系数进行融合。效果是要单用 90%的数据要好。