

BERT 蒸馏在垃圾舆情识别中的探索

原创 凰剑 阿里机器智能 前天



近来 BERT等大规模预训练模型在 NLP 领域各项子任务中取得了不凡的结果，但是模型海量参数，导致上线困难，不能满足生产需求。舆情审核业务中包含大量的垃圾舆情，会耗费大量的人力。本文在垃圾舆情识别任务中尝试 BERT 蒸馏技术，提升 textCNN 分类器性能，利用其小而快的优点，成功落地。

风险样本如下：

风险类型	舆情样本
无效	突然闲下来好无聊啊不知道干啥想吃好吃的打开 饿了么 看了半小时 然后天黑了害我唱会歌儿吧 ???
有效	投诉编号：17349888999投诉对象：饿了么客户关怀投诉问题：服务不到位/态度差,逾期未发货投诉要求：赔偿,道歉,改善服务,作出处罚涉诉金额：16元投诉进度：已回复等了近一个小时，商家在未进行沟通的前提下直接取消订单。原因是动力不足，无法配送。为何让人白等一小时？！

一 传统蒸馏方案

目前，对模型压缩和加速的技术主要分为四种：

- 参数剪枝和共享
- 低秩因子分解
- 转移/紧凑卷积滤波器
- 知识蒸馏

知识蒸馏就是将教师网络的知识迁移到学生网络上，使得学生网络的性能表现如教师网络一般。本文主要集中讲解知识蒸馏的应用。

1 soft label

知识蒸馏最早是 2014 年 Caruana 等人提出方法。通过引入 teacher network（复杂网络，效果好，但预测耗时久）相关的软标签作为总体 loss 的一部分，来引导 student network（简单网络，效果稍差，但预测耗时低）进行学习，来达到知识的迁移目的。这是一个通用而简单的、不同的模型压缩技术。

- 大规模神经网络 (teacher network)得到的类别预测包含了数据结构间的相似性。
- 有了先验的小规模神经网络(student network)只需要很少的新场景数据就能够收敛。
- Softmax函数随着温度变量 (temperature) 的升高分布更均匀。

Loss公式如下：

$$L = \alpha L_{soft} + \beta L_{hard}$$

$$L_{soft} = - \sum_j^N p_j^T \log(q_j^T)$$

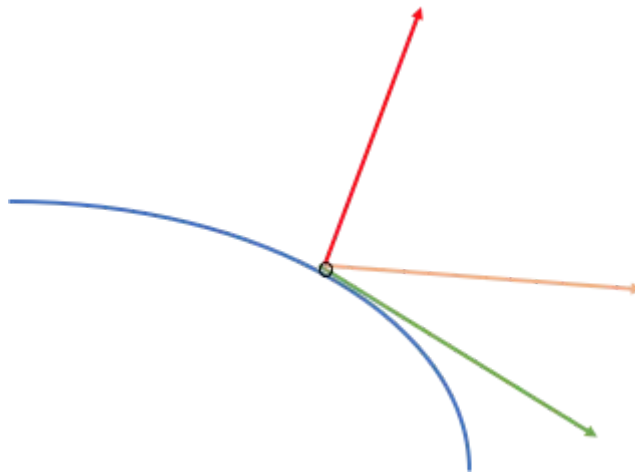
其中，

$$p_i^T = \frac{\exp(v_i/T)}{\sum_k^N \exp(v_k/T)}$$

$$q_i^T = \frac{\exp(z_i/T)}{\sum_k^N \exp(z_k/T)}$$

由此我们可以看出蒸馏有以下优点：

- 学习到大模型的特征表征能力，也能学习到one-hot label中不存在的类别间信息。
- 具有抗噪声能力，如下图，当有噪声时，教师模型的梯度对学生模型梯度有一定的修正性。
- 一定的程度上，加强了模型的泛化性。



红色为噪声数据梯度，黄色为教师模型梯度，绿色为最优梯度

2 using hints

(ICLR 2015) FitNets Romero等人的工作不仅利用教师网络的最后输出logits，还利用了中间隐层参数值，训练学生网络。获得又深又细的FitNets。

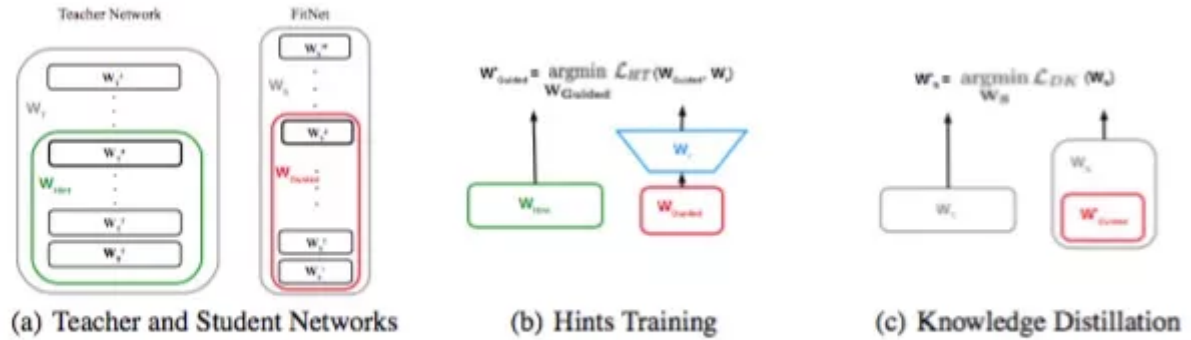


Figure 1: Training a student network using hints.

中间层学习loss如下：

$$\mathcal{L}_{HT}(\mathbf{W}_{\text{Guided}}, \mathbf{W}_{\text{r}}) = \frac{1}{2} \|u_h(\mathbf{x}; \mathbf{W}_{\text{Hint}}) - r(v_g(\mathbf{x}; \mathbf{W}_{\text{Guided}}); \mathbf{W}_{\text{r}})\|^2,$$

作者通过添加中间层loss的方式，通过teacher network 的参数限制student network的解空间的方式，使得参数的最优解更加靠近到teacher network，从而学习到teacher network的高阶表征，减少网络参数的冗余。

3 co-training

(arXiv 2019) Route Constrained Optimization (RCO) Jin和Peng等人的工作受课程学习(curriculum learning)启发，并且知道学生和老之间的gap很大导致蒸馏失败,导致认知偏差，提出路由约束提示学习(Route Constrained Hint Learning)，把学习路径更改为每训练一次teacher network，并把结果输出给student network进行训练。student network可以一步一步地根据这些中间模型慢慢学习，from easy-to-hard。

训练路径如下图：

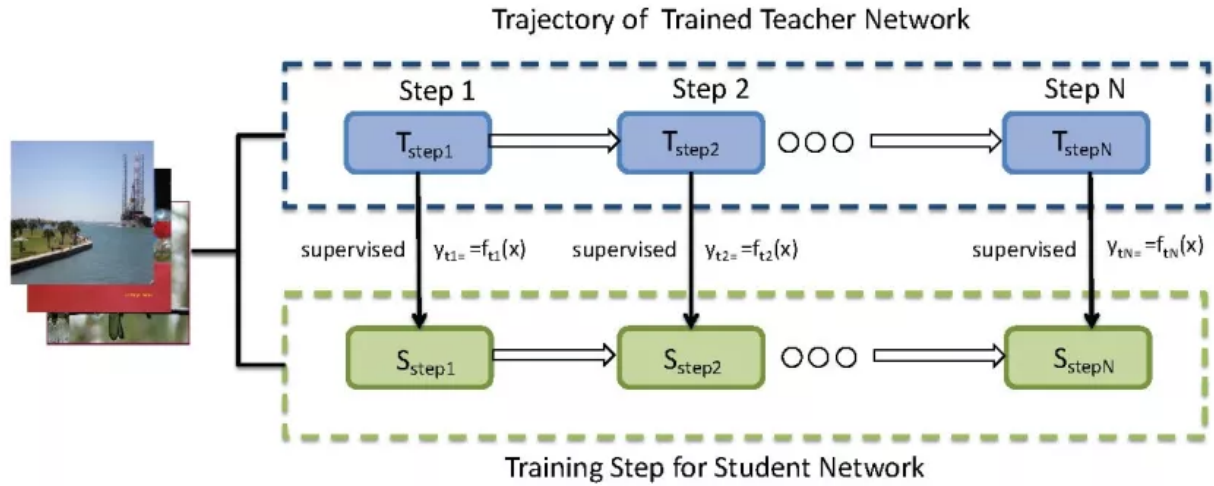


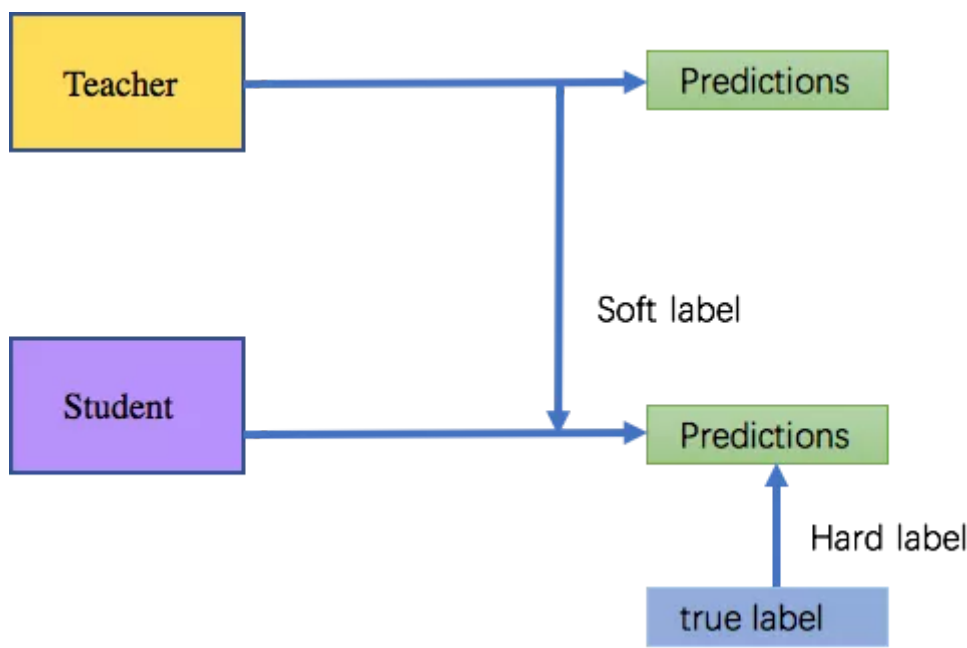
Figure 2: The overall framework of RCO. Previous knowledge transfer method only considers the converged teacher model. While RCO aims to supervise student with intermediate training state of teacher.

二 Bert2TextCNN蒸馏方案

为了提高模型的准确率，并且保障时效性，应对GPU资源紧缺，我们开始构建bert模型蒸馏至textcnn模型的方案。

方案1：离线logit textcnn 蒸馏

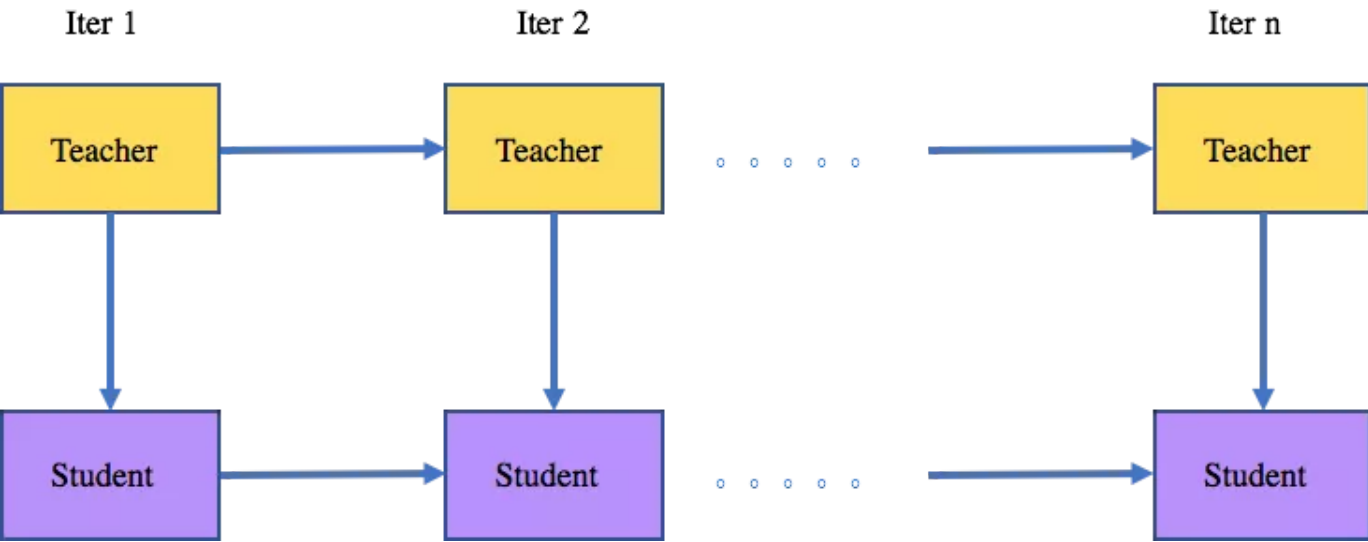
使用的是Caruana的传统方法进行蒸馏。



离线 logit textcnn 蒸馏训练流程

方案2：联合训练 bert textcnn 蒸馏

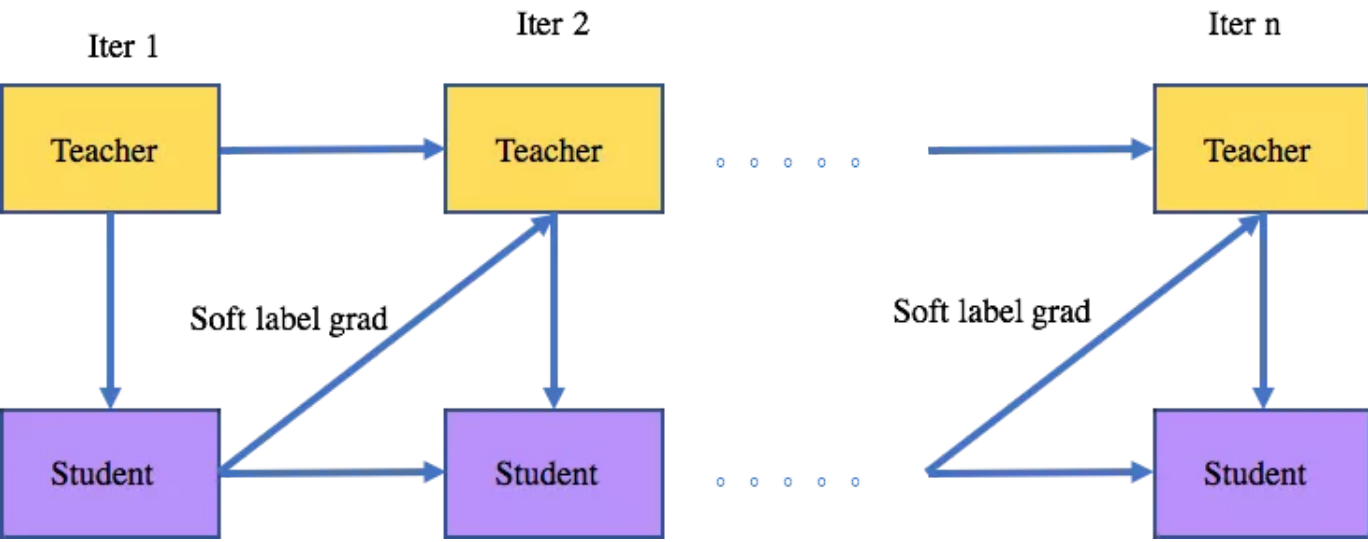
参数隔离：teacher model 训练一次，并把logit传给student。teacher 的参数更新至受到label的影响， student 参数更新受到teacher loigt的soft label loss 和label 的 hard label loss 的影响。



联合训练 bert textcnn 蒸馏参数隔离训练流程

方案3：联合训练 bert textcnn 蒸馏

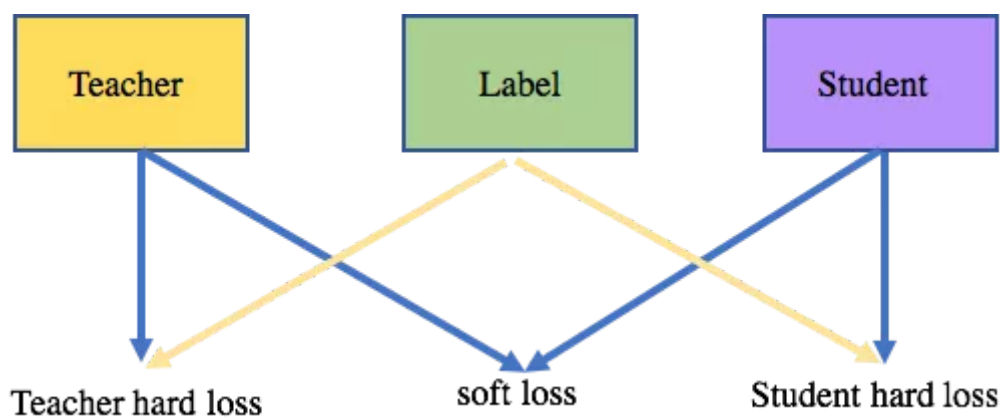
参数不隔离：与方案2类似，主要区别在于前一次迭代的student 的 soft label 的梯度会用于 teacher参数的更新。



联合训练 bert textcnn 蒸馏参数不隔离训练流程

方案4：联合训练 bert textcnn loss 相加

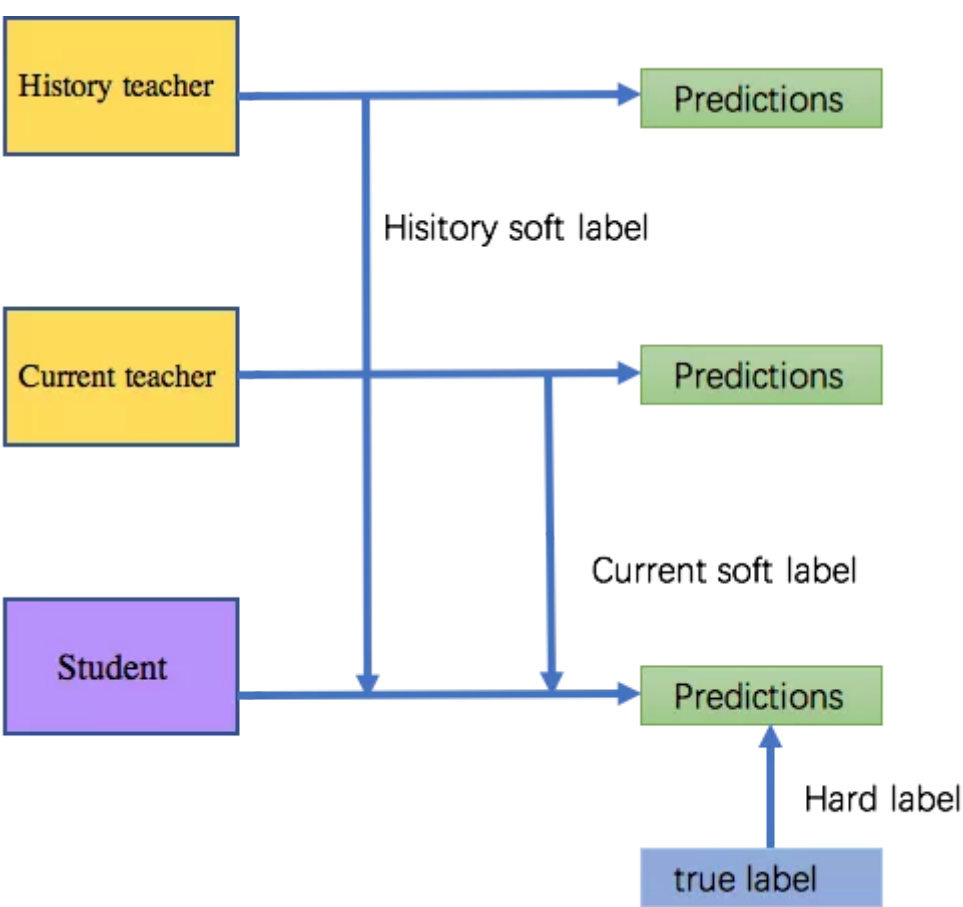
teacher 和student 同时训练，使用muti-task的方式。



联合训练 bert textcnn loss 相加训练流程

方案5: 多teacher

大部分模型，在更新时候需要覆盖线上历史模型的样本，使用线上历史模型作为teacher，让模型学习原有历史模型的知识，保障对原有模型有较高的覆盖。



多 teacher 训练流程

实验结果如下：

训练方式	方案	text-A UC	bert-A UC	准确率	召回
baseline	textcnn	0.758		95%	60%
	bert		0.8057	95%	70%
soft label	离线logit textcnn 蒸馏	0.773		95%	63%
	多teacher 方案	0.767		95%	62%
co-training	联合训练 bert textcnn 蒸馏参数 隔离	0.779	0.795	95%	64%
	联合训练 bert textcnn 蒸馏参数 不隔离	0.777	0.811	95%	65%
	联合训练 bert textcnn loss 相加	0.689	0.8057	95%	52%

从以上的实验，可以发现很有趣的现象。

- 1) 方案2和方案3均使用先训练teacher，再训练student的方式，但是由于梯度返回更新是否隔离的差异，导致方案2低于方案3。是由于方案3中，每次训练一次teacher，在训练一次student，student学习完了的soft loss 会再反馈给teacher，让teacher知道指如何导student是合适的，并且还提升了teacher的性能。
- 2) 方案4采用共同更新的，同时反馈梯度的方式。反而textcnn 的性能迅速下降，虽然bert的性能基本没有衰减，但是bert难以对textcnn每一步的反馈有个正确性的引导。
- 3) 方案5中使用了历史textcnn 的logit，主要是为了用替换线上模型时候，并保持对原有模型有较高的覆盖率，虽然召回下降，但是整体的覆盖率相比于单textcnn 提高了5%的召回率。

Reference

1.Dean, J. (n.d.). Distilling the Knowledge in a Neural Network. 1–9.

2.Romero A , Ballas N , Kahou S E , et al. FitNets: Hints for Thin Deep Nets[J].

3.Jin X , Peng B , Wu Y , et al. Knowledge Distillation via Route Constrained Optimization[J].

欢迎各位技术同路人加入蚂蚁集团大安全机器智能团队，我们专注于面向海量舆情借助大数据技术和自然语言理解技术挖掘存在的金融风险、平台风险，为用户资金安全护航、提高用户在蚂蚁生态下的用户体验。内推直达 lingke.djt@antfin.com，有信必回。

AI 场景体验


机器学习算法: 基于逻辑回归的分类预测

逻辑回归（Logistic regression，简称LR）是一个分类模型，其模型简单和可解释性强，是很多分类算法的基础组件。逻辑回归模型广泛应用于机器学习、大多数医学领域和社会科学等领域。通过本次实验，帮助大家掌握逻辑回归的理论，以及 sklearn 函数调用使用并将其运用到鸢尾花数据集的预测中。

点击“阅读原文”立即体验吧~



关注 机器智能
把握未来可能

 戳我，立即体验。

阅读原文