

基于知识蒸馏与模型集成的事件论元抽取方法

王士浩, 王中卿, 李寿山, 周国栋

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘要: 目前先进的事件论元抽取方法通常使用BERT模型作为编码器, 但BERT巨大的参数量会降低效率, 使模型无法在计算资源有限的设备中运行。提出一种新的事件论元抽取方法, 将事件论元抽取教师模型蒸馏到2个不同的学生模型中, 再对2个学生模型进行集成。构造使用BERT模型和图卷积神经网络的事件论元抽取教师模型, 以及2个分别使用单层卷积神经网络和单层长短期记忆网络的学生模型。先通过均方误差损失函数对学生模型和教师模型的中间层向量进行知识蒸馏, 再对分类层进行知识蒸馏, 使用均方误差损失函数和交叉熵损失函数让学生模型学习教师模型分类层的知识和真实标签的知识。在此基础上, 利用加权平均的方法对2个学生模型进行集成, 从而提升事件论元抽取性能。使用ACE2005英文数据集进行实验, 结果表明, 与学生模型相比, 该方法可使事件论元抽取F1值平均提升5.05个百分点, 推理时间和参数量较教师模型减少90.85%和99.25%。

关键词: 事件论元抽取; 知识蒸馏; 模型集成; 预训练语言模型; 模型压缩

开放科学(资源服务)标志码(OSID):



中文引用格式: 王士浩, 王中卿, 李寿山, 等. 基于知识蒸馏与模型集成的事件论元抽取方法[J]. 计算机工程, 2022, 48(7): 97-103.

英文引用格式: WANG S H, WANG Z Q, LI S S, et al. Event argument extraction method based on knowledge distillation and model ensemble[J]. Computer Engineering, 2022, 48(7): 97-103.

Event Argument Extraction Method Based on Knowledge Distillation and Model Ensemble

WANG Shihao, WANG Zhongqing, LI Shoushan, ZHOU Guodong

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

[Abstract] Existing advanced event argument extraction methods focus on model's performance and ignore model's size and efficiency. These models exist problems of high computation cost and high delay. To address these problems, this paper proposes Event Argument Extraction method via knowledge Distillation and model Ensemble(EAEDE). The event argument extraction teacher model is distilled into two different student models, and then ensemble the student models. Firstly, a event argument extraction teacher model using BERT and graph Convolution Neural Network(CNN) is constructed, and then two student models using Long Short-Term Memory network(LSTM) and CNN respectively are constructed. During the distilling process, the student models learn the middle hidden of teacher model, and then learn the logits of teacher model. The Mean Square Error(MSE) loss function and Cross Entropy(CE) loss function are used to let students learn the knowledge of the teacher's model classification layer and the knowledge of the real label. Finally, the weighted average method is used to ensemble the two student models to get the final model. The experiments using ACE2005 dataset show that this method improves the event argument extraction performance of student models by an average of 5.05 percentage points, while reduces the infer time by 90.85% and reduces the size of model by 99.25%, comparing with the teacher model.

[Key words] event argument extraction; knowledge distillation; model ensemble; pre-trained language model; model compression

DOI: 10.19678/j.issn.1000-3428.0061790

0 概述

事件是信息的一种表现形式, 其定义为特定的人、物在特定时间和特定地点相互作用的客观事实。

事件抽取旨在将非结构化的文本以结构化的形式呈现。在事件抽取过程中, 标志某个事件发生的词称为触发词, 一般为动词或名词, 而事件的参与者称为论元, 一般为实体、时间、数值等。事件抽取包含事

基金项目: 国家自然科学基金(61806137, 61702518); 江苏省高等学校自然科学研究面上项目(18KJB520043)。

作者简介: 王士浩(1997—), 男, 硕士研究生, 主研方向为事件论元抽取; 王中卿(通信作者), 副教授; 李寿山、周国栋, 教授。

收稿日期: 2021-05-31 **修回日期:** 2021-09-16 **E-mail:** shwang10@stu.suda.edu.cn

件检测和事件论元抽取2个子任务,事件检测是识别出文本中的触发词并对事件类型进行分类,事件论元抽取则是识别出事件的论元并对论元角色进行分类。对于例句“在约旦河西岸,一名示威者中弹死亡”,事件检测程序首先会识别出“死亡”是触发词,然后将其事件类型归类为“死亡”。接着事件论元抽取程序识别出“约旦河西岸”“示威者”“弹”是事件的论元,然后将它们的角色类型归类为“地点”“受害者”“工具”。本文主要研究事件论元抽取任务。

随着深度学习技术的发展,研究者通过增加模型复杂度和参数量使模型具有更强的表征能力。在自然语言处理领域,文献[1]提出的BERT模型通过深层的模型结构和大规模数据预训练,在11项自然语言处理任务上取得了较好的效果。此后,越来越多的研究者将BERT模型用于事件论元抽取任务,取得了较大的性能提升。但BERT模型存在以下2个缺点:首先是巨额的参数量会导致模型预测速度变慢、时延增加,使模型无法应用于需要实时处理的场景;其次是模型需要极大的计算资源和内存开销,无法在计算资源有限的设备中运行。大模型无法在计算资源和时间有限的情况下运行,而小模型性能无法达到满意的效果,因此,模型压缩显得尤为重要。

在事件论元抽取领域,先进模型通常使用BERT作为模型编码器。在这些模型中,BERT模型的参数占整体参数的绝大部分。因此,模型压缩的关键在于减少BERT模型的参数。目前,知识蒸馏是常用的模型压缩方法之一。知识蒸馏的主要是让性能较高的教师模型指导性能较差的学生模型训练。教师模型网络结构复杂,表征能力更强,模型的输出(即“软标签”)中包含着标签概率分布知识。通过让学生模型的输出去拟合教师模型的软标签,可使学生模型学到真实标签(即“硬标签”)无法表示的知识。文献[2-4]利用知识蒸馏方法将大尺寸的BERT模型蒸馏到小尺寸的BERT模型中。这种蒸馏方式虽然能够得到较好的结果,但是模型参数量仍然很大。文献[5]将BERT模型蒸馏到单层长短期记忆网络(Long Short-Term Memory network, LSTM)中,大幅降低了模型参数量,但是模型性能仍然需要提升。

本文提出一种基于知识蒸馏与模型集成的事件论元抽取方法(Event Argument Extraction method based on knowledge Distillation and model Ensemble, EAEDE)。将使用BERT的教师模型蒸馏到使用LSTM或单层卷积神经网络(Convolution Neural Network, CNN)编码的学生模型上,以减少模型参数量,加快预测速度。在蒸馏方式上,本文设计一个两段式的蒸馏方式,先将教师模型中间层的信息蒸馏到学生模型中,再对学生模型的分层进行蒸馏,通过将2个学生模型进行集成,进一步提升模型的性能。

1 相关工作

1.1 事件抽取

目前,事件抽取的方法已逐渐从机器学习转换

为深度学习。文献[6]使用卷积神经网络对句子进行编码,并提出了动态多池化技术解决多事件中论元角色重叠的问题。文献[7]将事件检测任务和事件论元抽取任务进行联合学习,并利用触发词与触发词、触发词与论元、论元与论元之间的依赖关系提升分类效果。文献[8]设计了一种依存桥循环神经网络,在对句子编码时融入句子的依存句法特征,在计算当前节点隐层向量时不仅考虑上一个节点输出和记忆细胞中的信息,同时还融入与当前词有依存关系词的信息。文献[9]使用图卷积方法抽取句子的依存句法特征进行联合事件抽取,在事件检测时使用自注意力机制抽取候选触发词的特征。文献[10]使用CRF模型构造一个多任务学习框架,对子任务进行联合训练,同时采用对各事件类型单独训练的方式解决标签重叠问题。自预训练模型BERT推出后,研究者们开始使用BERT进行事件抽取,通过BERT深层次的网络结构和大规模的数据预训练大幅改善了事件抽取效果。文献[11]采用BERT加动态池化层的事件检测模型,同时使用规则的方式生成海量数据并利用对抗训练对生成的数据进行筛选。文献[12]提出层次模块化事件论元抽取模型,对论元角色进行分组并抽取出更抽象的概念,利用论元与抽象概念之间的关系提升论元角色抽取效果。文献[13]通过预测论元的起始位置抽取论元,为每个论元角色标签都设置一个二分类器解决标签重叠问题,同时,还设计一种数据增强方法,解决数据稀缺的问题。文献[14]以阅读理解的方式进行事件抽取,同时利用外部的问答语料库缓解数据稀缺的问题。文献[15]将事件论元抽取任务看作多轮的问答任务,利用历史的抽取信息辅助剩余论元的抽取。

1.2 模型压缩

在现有研究中,针对BERT模型的压缩有知识蒸馏、低秩分解和权值共享2种方式。

1) 知识蒸馏

文献[16]提出知识蒸馏的概念,其使用KL散度衡量教师模型输出与学生模型输出之间的差异,通过最小化KL散度将教师模型中的知识迁移到学生模型中。为充分利用模型中间层的特征,文献[17]通过最大化教师模型和学生模型中间层的互信息进行知识蒸馏。文献[18]通过使学生模型学习教师模型的注意力特征图提高学生模型的性能。文献[2]设计了三重损失函数,计算学生模型和教师模型之间隐层向量的余弦相似度、模型输出的KL散度以及真实标签的交叉熵优化模型,将模型层数减半且保证性能损失较低。文献[3]提出了两段式的蒸馏方式,首先对模型进行通用知识蒸馏,计算学生模型和教师模型Transformer层的隐层向量和注意力的损失值,然后针对具体下游任务进行蒸馏。在神经网络中,深而窄的模型比浅而深的模型具有更强的表征能力但是难于训练,文献[4]采用Bottleneck机制和堆叠多个前馈神经网络的方式训练了一个深而窄的BERT模型,提升了模型的推理效率。在上述工作中,学生模型和教师模型结构相同,只是使用更少的模型层数或是每层的参数。文献[5]使用LSTM构

造了一个学生模型,并让学生模型分类层拟合教师模型分类层,将BERT的知识蒸馏到LSTM中。文献[19]在蒸馏方式上与文献[5]相同,但是使用了对抗生成网络生成大量无标注标签文本扩充数据,使学生模型能够更充分地学习教师模型中的信息。

2) 低秩分解和权值共享

BERT模型的主要参数量来源是词向量和12层的Transformer。文献[20]将词向量从768维减少为128维,再通过一个 128×768 的矩阵将词向量转换成768维,从而使词向量参数减少到原始大小的1/6。为减少Transformer的参数量,该文献将12层的Transformer参数进行共享,即使用一层Transformer反复迭代12次,将Transformer部分参数减少到原

始大小的1/12。虽然这种方法可以大幅减小参数量,但是计算量没有减少,因此,计算速度依然较慢。

2 EAEDE方法

本节详细介绍基于知识蒸馏与模型集成的事件论元抽取方法(EAEDE)。如图1所示,该方法模型包含了2个学生模型和1个教师模型,将学生模型经过蒸馏后进行集成。具体训练流程如图2所示,其中包含4个步骤:1)构建2个学生模型和1个教师模型;2)训练教师模型使其达到最优;3)进行知识蒸馏,使学生模型学习教师模型中蕴含的知识;4)将蒸馏后的2个学生模型进行集成,得到最终的模型。

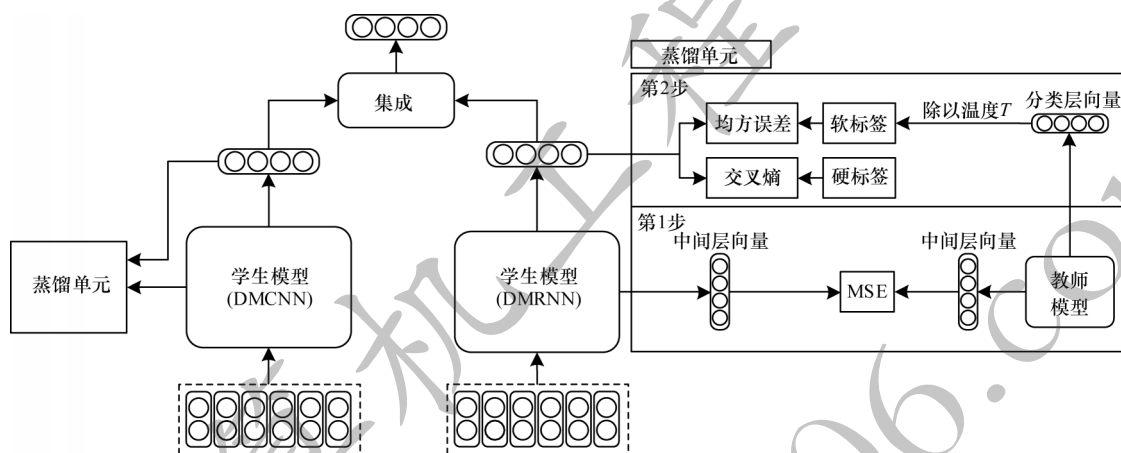


图1 EAEDE模型架构

Fig.1 Structure of EAEDE model

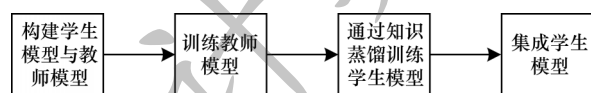


图2 模型训练流程

Fig.2 Procedure of model training

2.1 学生模型

学生模型架构如图3所示,其中包含输入层、句级特征抽取层、词级特征抽取层、分类层4个部分。2个学生模型分别采用CNN和LSTM对句子进行编码。

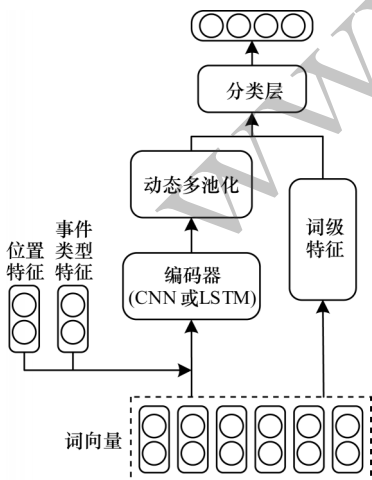


图3 学生模型架构

Fig.3 Structure of student model

2.1.1 输入层

给定一个句子 $S=\{w_1, w_2, \dots, w_t, \dots, w_a, \dots, w_n\}$,其中:下标 t 表示触发词位置;下标 a 表示候选论元位置;下标 n 表示句子长度。模型的输入由词向量、位置特征、事件类型特征三部分组成:词向量通过使用GloVe模型^[21]在维基百科语料上训练得到,获得词向量;位置特征是当前词与候选论元和触发词的相对位置,每个相对位置都由一个随机向量表示;事件类型特征与位置特征相似,每种事件类型由一个随机向量表示。将词向量、位置特征和事件类型特征进行拼接,得到输入 $X=\{x_1, x_2, \dots, x_t, \dots, x_a, \dots, x_n\}$ 。

2.1.2 句级特征抽取层

本文中的2个学生模型分别使用CNN和LSTM作为编码器获得模型句级特征,使模型之间具有一定的差异性。句子通过输入层处理后放入编码器中,得到隐层向量,如式(1)所示:

$$\{h_1, h_2, \dots, h_n\} = \text{Encoder}(x_1, x_2, \dots, x_n) \quad (1)$$

然后,使用动态多池化方法对隐层向量进行融合,得到候选论元 w_a 的句级特征,如式(2)~式(5)所示:

$$h_1 = \max(h_1, h_2, \dots, h_t) \quad (2)$$

$$h_m = \max(h_{t+1}, h_{t+2}, \dots, h_a) \quad (3)$$

$$h_r = \max(h_{a+1}, h_{a+2}, \dots, h_n) \quad (4)$$

$$H_s = [h_l, h_m, h_r] \quad (5)$$

其中: t 表示触发词位置; a 表示候选论元位置; $[]$ 表示拼接。

2.1.3 词级特征抽取层

词级特征是事件论元抽取的重要特征。给定一个句子,句子中每个词的词向量为 $\{c_1, c_2, \dots, c_t, \dots, c_a, \dots, c_n\}$ 。本文将候选论元、触发词以及两者上下词的词向量拼接起来作为词级特征,具体如式(6)所示:

$$H_L = [c_{t-1}, c_t, c_{t+1}, c_{a-1}, c_a, c_{a+1}] \quad (6)$$

2.1.4 分类层

在得到句级特征 H_s 与词级特征 H_L 后,将两者拼接起来得到最终候选论元 w_a 的最终特征表示 H 。然后,使用全连接层和 Softmax 函数对候选论元进行分类,得到每个类别的概率。

2.2 教师模型

教师模型采用与学生模型相同的架构,但教师模型使用了 BERT 模型和图卷积来提高分类效果,具体见图 4。本小节主要介绍教师模型的句级特征抽取层及教师模型的训练流程。

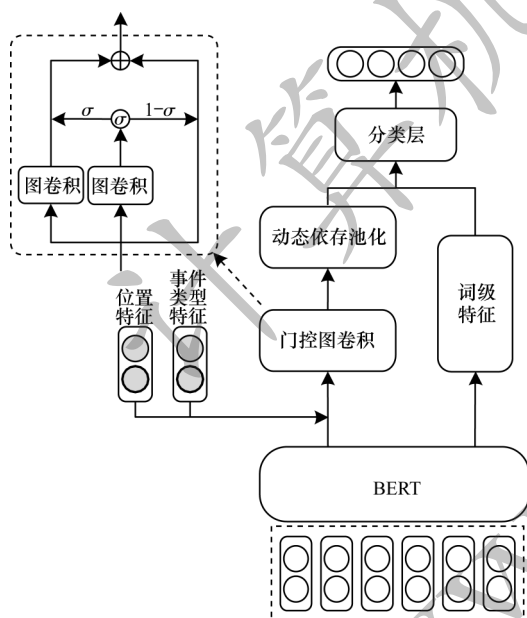


图4 教师模型架构

Fig.4 Structure of teacher model

2.2.1 句级特征抽取层

教师模型的编码器包含 BERT 和门控图卷积 2 个部分。给定一个句子 $S=\{w_1, w_2, \dots, w_n\}$ 。首先,使用 BERT 对句子进行编码,得到句子的隐层向量 X 。然后,将句子隐层向量与位置特征向量、事件类型特征向量拼接作为图卷积的输入 X 。位置特征向量和事件类型特征向量的初始化方式与学生模型相同。

图卷积的邻接矩阵根据依存句法树进行构造。使用 stanfordCoreNLP 工具(<https://stanfordnlp.github.io/CoreNLP/>)进行依存句法分析。本文将句子中每一个词作为一个节点,根据依存关系构造该句子的邻

接矩阵 A 。

参考文献[22]的工作,本文首先对每一个节点创建一个自环 $A_{ii}=1, i \in [1, n]$ 。如果节点 i 和 j 之间有依存关系,则向邻接矩阵中添加边 $A_{ij}=A_{ji}=1$ 。在构建邻接矩阵后,对隐层向量 X 进行两次图卷积,得到句子的依存句法特征。然后,对其中一个依存句法特征使用 Sigmoid 函数,将其作为门控单元,记为 σ 。接着,将依存句法特征与 σ 进行点乘,将输入的隐层向量 X 与 $(1-\sigma)$ 进行点乘,最后将两者相加融合,具体如式(7)、式(8)所示:

$$H = X \otimes (1 - \sigma) + \text{GraphConv}_1(X) \otimes \sigma \quad (7)$$

$$\sigma = \text{Sigmoid}(\text{GraphConv}_2(X)) \quad (8)$$

经过图卷积后,得到隐层向量 $H_s = \{h_1, h_2, \dots, h_n\}$,本文将候选论元、触发词以及和候选论元有直接依存关系的词进行最大池化,得到候选论元的句级特征,如式(9)所示:

$$H_s = \max \left(\{h_a, h_t\} \cup \{h_i | A_{i,a} = 1, i \in [1, n]\} \right) \quad (9)$$

其中: t 表示触发词的位置; a 表示候选论元的位置; A 为邻接矩阵。

2.2.2 教师模型训练

本文使用交叉熵(Cross Entropy, CE)损失函数计算教师模型的输出与真实标签之间的损失值,如式(10)所示:

$$L_{CE} = - \sum_{i=1}^K y_i \ln(\hat{y}_i) \quad (10)$$

其中: K 表示标签类别数; y 表示真实标签; \hat{y} 表示教师模型的预测值。

2.3 知识蒸馏

对学生模型的蒸馏共分为 2 个阶段:第一阶段对模型的中间层进行蒸馏;第二阶段对模型的分层进行蒸馏。

在第一阶段,选取教师模型 Gate-GCN 层的输出与学生模型 Encoder 层的输出进行蒸馏。教师模型 Gate-GCN 层的输出包含了由 BERT 获得全局的语义信息和由 GCN 模型获得的依存句法信息。2 个学生模型使用了 CNN 和 LSTM 对句子进行编码,但两者都无法有效地获取长句子的全局语义信息,更无法获取句子的依存句法信息。通过知识蒸馏,学生模型的 Encoder 层会拟合教师模型 Gate-GCN 层的输出,使其输出的向量包含全局的语义信息和依存句法信息。由于两者输出的向量维度不同,因此学生模型 Encoder 层的输出需要进行线性变换匹配教师模型的维度,然后使用均方误差(Mean Square Error, MSE)损失函数进行蒸馏,如式(11)所示:

$$L_{MSE} = \frac{1}{K} \sum_{i=1}^K (h_i^t - h_i^s)^2 \quad (11)$$

其中: h^t 表示教师模型的隐层向量; h^s 表示学生模型的隐层向量。

在第二阶段,学生模型的分层会学习教师模

型分类层的概率分布知识,即“软标签”。软标签中包含了比真实标签(硬标签)更丰富的信息。以“在约旦河西岸,一名示威者中弹死亡”中的论元“示威者”为例,硬标签只会告诉模型“示威者”是“受害者”,而软标签则会使模型学习到“示威者”很可能是“受害者”,较低可能是“攻击者”,绝不可能是“地点”。教师模型输出的软标签往往过于陡峭,某一类别的概率很大,而其他类别概率很小。为了放大小概率类别的信息,使软标签更加平滑,需要在使用Softmax函数对概率归一化时设置温度系数。温度越大,软标签越平滑,各标签概率趋向于相同;温度越小,软标签越陡峭,各标签概率相差越大。具体如式(12)所示:

$$\hat{y}_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \tag{12}$$

其中: z 表示模型的输出; T 是温度参数; \hat{y} 表示归一化后的概率。本阶段将软标签和预测值的均方误差损失函数与硬标签和预测值的交叉熵损失函数相加作为最终的目标函数,具体如式(13)所示:

$$L = \frac{1}{K} \sum_{i=1}^K (\hat{y}_i^t - \hat{y}_i^s)^2 - \sum_{i=1}^K y_i \ln(\hat{y}_i^s) \tag{13}$$

其中: \hat{y}^t 表示教师模型输出的软标签; \hat{y}^s 表示学生模型的预测结果; y 表示真实标签; K 表示标签类别个数。

2.4 模型集成

本文采用加权平均的方法对蒸馏后的学生模型进行集成。模型集成能够降低单个学生模型的偏差对整体的影响,提升模型的泛化能力。为了更好的效果,被集成的模型之间应当保持一定差异性。本文分别使用CNN和LSTM构造2个不同的学生模型,使模型关注于不同的特征。CNN由于卷积核大小的限制,使模型倾向于局部的特征,而LSTM具有记忆细胞,可以记住较远词的信息,使模型倾向于全局的特征。模型集成方法的计算公式如式(14)所示:

$$\hat{y}^E = a \cdot \hat{y}^{s_{cnn}} + (1 - a) \cdot \hat{y}^{s_{lstm}} \tag{14}$$

其中: \hat{y}^E 表示模型集成后的预测结果; $\hat{y}^{s_{cnn}}$ 和 $\hat{y}^{s_{lstm}}$ 分别表示2个蒸馏后学生模型的预测结果; a 表示权重系数。

3 实验

3.1 数据集与评价标准

本文使用ACE2005英文数据集进行实验。该数据集包含了新闻、广播对话、微博等6个领域的数据,共599篇文档,其中定义了33种事件类型和35种论元角色。本文采取与文献[6-7]相同的划分方式,随机选取40篇新闻领域文档作为测试集,在剩下的文档中选取30篇作为测试集,选取529篇作

为训练集。本文使用准确率、召回率和F1值作为评价标准,当一个论元的所属的事件类型、位置和论元角色分类都正确时,则该论元分类正确。

3.2 模型参数

教师模型和学生模型的超参数设置分别如表1和表2所示。在训练过程中,模型通过Adam优化器更新参数。学生模型和教师模型均使用Dropout机制,防止在训练过程中模型出现过拟合。

表1 教师模型超参数设置

Table 1 Hyperparameter setting of teacher model

超参数名称	参数值
学习率	2×10^{-5}
批处理大小	8
BERT向量维度	768
图卷积向量维度	918
事件类型特征向量维度	50
位置特征向量维度	50
丢弃率	0.5

表2 学生模型超参数设置

Table 2 Hyperparameter setting of student model

超参数名称	参数值
中间层蒸馏学习率	3×10^{-4}
分类层蒸馏学习率	1×10^{-3}
批处理大小	50
词向量维度	100
CNN向量维度	300
卷积核大小	3
BiLSTM向量维度	300
事件类型特征向量维度	50
位置特征向量维度	50
丢弃率	0.5
温度	2
权重系数a	0.5

3.3 实验结果

本文针对事件论元抽取任务,评测模型论元角色分类的结果。论元角色分类所用的触发词由DMCNN^[6]模型抽取获得。为了验证知识蒸馏与模型集成的事件论元抽取方法的有效性,本文选取以下基准方法进行对比:

- 1) DMCNN: 本文方法的学生模型之一,由CHEN等^[6]提出,使用CNN对句子进行编码。
- 2) DMRNN: 本文方法的学生模型之一,使用RNN对句子进行编码。
- 3) Teacher Model: 本文2.2小节介绍的教师模型。
- 4) DMCNN(distill): 经过教师模型蒸馏后的DMCNN模型。
- 5) DMRNN(distill): 经过教师模型蒸馏后的DMRNN模型。

6)EAEDE:本文提出的方法,由2个蒸馏后的学生模型集成后所得。

表3列出了基准模型与EAEDE在事件论元抽取上的实验结果。可以看出:Teacher Model的F1值达到60.3%,分别比DMCNN与DMRNN高出6.8个百分点和6.3个百分点,三者之间差距较大;学生模型经过蒸馏后,论元抽取效果显著提升,DMCNN(distill)在F1值上比DMCNN高出2.9个百分点,DMRNN(distill)在F1值上比DMRNN高出2.6个百分点,这表明在事件论元抽取任务中,知识蒸馏可以提升浅层模型的表征能力,获得更好的效果;此外,模型集成后EAEDE的F1值分别比DMCNN(distill)和DMRNN(distill)高出2.4个百分点和2.2个百分点,比DMCNN高出5.3个百分点,比DMRNN高出4.8个百分点,这说明模型集成可以有效弥补单个模型性能不足的缺点,得到一个更全面的模型。

表3 事件论元抽取实验结果对比

Table 3 Experiment results comparison of event argument extraction

对比方法	准确率	召回率	F1值
DMCNN	62.2	46.9	53.5
DMRNN	60.2	48.9	54.0
Teacher Model	57.4	63.5	60.3
DMCNN(distill)	58.0	54.9	56.4
DMRNN(distill)	55.8	57.4	56.6
EAEDE	59.4	58.3	58.8

为进一步分析本文方法的有效性,选取以下模型压缩方法进行对比:

1)Distill-Logits:方法架构与本文方法相同,都是将教师模型蒸馏到2个学生模型上,然后进行集成。不同之处在于,该方法只对分类层进行蒸馏。

2)Distill-Encoder:方法架构与本文提出的方法相同,但该方法只对学生模型的Encoder层进行蒸馏,然后使用真实标签对模型分类层进行训练。

3)Teacher-TinyBert:本文2.2小节提出的教师模型,蒸馏方式使用的是JIAO等^[3]提出的TinyBert。本文中使用的模型尺寸是6层768维。

4)Teacher-MobileBert:文献[4]所提出的BERT蒸馏方式,本文使用的模型尺寸是24层512维。

Teacher-TinyBert的模型压缩思路是减少BERT的参数量,而EAEDE的模型压缩思路是将大模型知识蒸馏到多个小模型上,然后对小模型进行集成。从表4可以看出:在论元分类F1值上,本文提出的EAEDE比Teacher-TinyBert高出2.1个百分点,比Teacher-MobileBert高出1.7个百分点,充分证明了本文模型压缩思路的有效性;同时,EAEDE的性能比Distill-Logits高出1.9个百分点,比Distill-Encoder高出2.2个百分点。Distill-Logits和Distill-Encoder都只对模型的一个部分进行了蒸馏,而EAEDE对两个部分都进行了蒸馏,这表明对教师模型各层分别进

行蒸馏能够使学生模型更有效地学习教师模型的知识,提升学生模型的分类性能。

表4 不同模型压缩方法性能对比

Table 4 Performance comparison of different model

compression methods			%
对比方法	准确率	召回率	F1值
Distill-Logits	61.5	52.9	56.9
Distill-Encoder	59.6	53.8	56.6
Teacher-TinyBert	54.7	58.8	56.7
Teacher-MobileBert	58.2	56.0	57.1
EAEDE	59.4	58.3	58.8

3.4 模型效率分析

本文在相同的硬件条件下对模型预测速度进行测试,GPU型号为Tesla K40m,CPU型号为E5-2680 v4。表5列出了EAEDE与基准方法在参数量和单条数据推理时间上的对比结果。可以看出:EAEDE在参数量上比Teacher Model减少99.25%,比Teacher-TinyBert减少98.81%,比Teacher-MobileBert减少96.91%;EAEDE仅使用了单层的CNN和RNN编码器,参数量极小;Teacher Model和Teacher-TinyBert中的BERT分别含有12层和6层的Transformer,隐层向量维度为768,通过计算可得,仅仅单层的Transformer参数量就达到了 7×10^6 ,是EAEDE的8.1倍;在单条数据推理时间上,EAEDE比Teacher Model减少90.85%,比Teacher-TinyBert减少85.01%,比Teacher-MobileBert减少81.54%,这充分说明了本文方法在有限的硬件资源和时间下能够获得较好的效果。

表5 模型参数量、预测速度与F1值对比

Table 5 Comparison of model parameter quantity, prediction speed and F1 value

对比方法	参数量/ 10^6	单条数据推理时间/ms	F1值/%
Teacher Model	114.81	47.0	60.3
Teacher-TinyBert	72.29	28.7	56.6
Teacher-MobileBert	27.84	23.3	57.1
EAEDE	0.86	4.3	58.8

4 结束语

本文针对现有事件论元抽取模型参数量大、时延高等问题,提出一种基于知识蒸馏和模型集成的事件论元抽取方法。设计两阶段的蒸馏方式,将含有BERT的教师模型蒸馏到一个使用单层CNN或LSTM编码器的学生模型中,大幅减少模型参数量,并通过模型集成进一步提升性能。实验结果表明,本文方法能够快速完成事件论元抽取任务并取得较优性能。下一步将通过加入词级特征、词向量等进行知识蒸馏,提升学生模型的事件论元抽取性能。

参考文献

- [1] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021-05-10]. <https://arxiv.org/abs/1810.04805>.
- [2] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[EB/OL]. [2021-05-10]. <https://arxiv.org/abs/1910.01108>.
- [3] JIAO X Q, YIN Y C, SHANG L F, et al. TinyBERT: distilling BERT for natural language understanding[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2020: 4163-4174.
- [4] SUN Z Q, YU H K, SONG X D, et al. MobileBERT: a compact task-agnostic BERT for resource-limited devices [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2020: 2158-2170.
- [5] TANG R, LU Y, LIU L Q, et al. Distilling task-specific knowledge from BERT into simple neural networks[EB/OL]. [2021-05-10]. <https://arxiv.org/abs/1903.12136>.
- [6] CHEN Y B, XU L H, LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2015: 409-419.
- [7] NGUYEN T H, CHO K, GRISHMAN R. Joint event extraction via recurrent neural networks[C]//Proceedings of NAACL-HLT 2016. Stroudsburg, USA: Association for Computational Linguistics, 2016: 300-309.
- [8] SHA L, QIAN F, CHANG B, et al. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI, 2018: 5916-5923.
- [9] LIU X, LUO Z C, HUANG H Y. Jointly multiple events extraction via attention-based graph information aggregation [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 1247-1256.
- [10] 贺瑞芳,段绍杨. 基于多任务学习的中文事件抽取联合模型[J]. 软件学报, 2019, 30(4): 1015-1030.
HE R F, DUAN S Y. Joint Chinese event extraction based multi-task learning[J]. Journal of Software, 2019, 30(4): 1015-1030. (in Chinese)
- [11] WANG X, HAN X, LIU Z, et al. Adversarial training for weakly supervised event detection [C]//Proceedings of NAACL-HLT 2019. Stroudsburg, USA: Association for Computational Linguistics, 2019: 998-1008.
- [12] WANG X, WANG Z, HAN X, et al. HMEAE: hierarchical modular event argument extraction[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2019: 5781-5787.
- [13] YANG S, FENG D W, QIAO L B, et al. Exploring pre-trained language models for event extraction and generation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2019: 5284-5294.
- [14] LIU J, CHEN Y, LIU K, et al. Event extraction as machine reading comprehension [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2020: 1641-1651.
- [15] LI F Y, PENG W H, CHEN Y G, et al. Event extraction as multi-turn question answering [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2020: 829-838.
- [16] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. [2021-05-10]. <https://arxiv.org/abs/1503.02531>.
- [17] AHN S, HU S X, DAMIANOU A, et al. Variational information distillation for knowledge transfer [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 9155-9163.
- [18] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer[EB/OL]. [2021-05-10]. <https://arxiv.org/abs/1612.03928>.
- [19] 廖胜兰, 吉建民, 俞畅, 等. 基于BERT模型与知识蒸馏的意图分类方法[J]. 计算机工程, 2021, 47(5): 73-79.
LIAO S L, JI J M, YU C, et al. Intention classification method based on BERT model and knowledge distillation [J]. Computer Engineering, 2021, 47(5): 73-79. (in Chinese)
- [20] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. [2021-05-10]. <https://arxiv.org/abs/1909.11942>.
- [21] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2014: 1532-1543.
- [22] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2021-05-10]. <https://arxiv.org/abs/1609.02907>.

编辑 金胡考