■ 人工智能专题(主持人:谢娟英)

引用格式:高子雄,蒋盛益,欧炎镁,等.基于多任务蒸馏的意图识别和槽位填充[J].陕西师范大学学报(自然科学版),2024,52(3):96-104.[GAO Z X,JIANG S Y,OU Y M,et al. Research on sentence intention recognition and slot filling based on multi-task distillation[J]. Journal of Shaanxi Normal University (Natural Science Edition),2024,52(3):96-104.] **DOI**: 10.15983/j.cnki.jsnu.2024013

基于多任务蒸馏的意图识别和槽位填充

高子雄,蒋盛益*,欧炎镁,禤镇宇
(广东外语外贸大学 信息科学与技术学院/网络空间安全学院,广东 广州 510006)

摘 要:BERT 等预训练模型在很多 NLP 任务上取得了良好的效果,但预训练模型参数规模大,运算量大,对硬件资源要求高,难以部署在小型的配置环境中。模型压缩是解决该问题的关键,知识蒸馏是目前较好的模型压缩方法。基于此,提出基于多任务蒸馏的句意图识别和槽位填充联合模型,该模型将 ALBERT 运用到任务型对话系统中,并利用知识蒸馏策略将 ALBERT 模型知识迁移到 BiLSTM 模型。实验结果表明,基于 ALBERT 的联合模型在 SMP 2019 评测数据集中的句准确率为 77.74%,单独训练的 BiLSTM 模型句准确率为 58.33%,而蒸馏模型的句准确率为 67.22%,在比 BiLSTM 高 8.89%的情况下,推断速度约为 ALBERT 的 18.9 倍。

关键词:意图识别与槽位填充;神经网络;知识蒸馏
中图分类号:O152.1;O413.1 **文献标志码**:A **文章编号**:1672-4291(2024)03-0096-09

Research on sentence intention recognition and slot filling based on multi-task distillation

GAO Zixiong,JIANG Shengyi*,OU Yanmei,XUAN Zhenyu
(School of Information Science and Technology/School of Cyber Security,Guangdong University of Foreign Studies,Guangzhou 510006,Guangdong,China)

Abstract: At present, pre-trained models such as BERT have achieved good results in many NLP tasks, but the pre-trained models are difficult to deploy in small configuration environments because of their large parameter scale, large computation and high requirements on hardware resources. Model compression is the key to solve this problem, and knowledge distillation is currently a better model compression method. A joint model of sentence intent recognition and slot filling based on multi-task distillation is proposed. The model applies ALBERT to task-based dialogue system, and uses the knowledge distillation strategy to migrate the ALBERT model knowledge to the BiLSTM model. Experimental results show that the sentence accuracy rate of the ALBERT based joint model in the SMP 2019 evaluation data set is 77.74%, the sentence accuracy rate of the BiLSTM model trained separately is 58.33%, and the sentence accuracy rate of the distillation model is 67.22%, which is 8.89% higher than the BiLSTM model while offering an inference speed approximately 18.9 times faster than ALBERT.

Keywords: intention recognition and slot filling; neural network; knowledge distillation

收稿日期:2023-10-17
基金项目:国家自然科学基金(61572145)

* **通信作者**:蒋盛益,男,教授,博士生导师,主要从事数据挖掘、自然语言处理等方面的研究。E-mail: jiangshengyi@163.com

任务型对话系统是指以人机对话的形式提供信息或服务的系统。当前，任务型对话系统得到了广泛应用，业务助理系统微软小娜、百度度秘、阿里小蜜等是典型的任务型对话系统。

自然语言理解(natural language understanding，NLU)是任务型对话系统的核心模块，主要作用是对用户输入的句子或者语音识别的结果进行语义解析。在自然语言理解中，一般包含意图识别(intent detection，ID)和槽位填充(slot filling，SF)两个子任务。

本文主要关注基于知识蒸馏的意图识别和槽位填充的研究。目前，意图识别和槽位填充的研究主要基于预训练模型[1-3]。然而，预训练模型由于运算量大且参数较多，在算力较弱的移动设备上难以支持实时性要求较高的智能对话系统。现有研究[4-5]尝试通过知识蒸馏来解决该问题，目前该领域的工作主要针对单一意图识别任务，对意图识别与槽位填充联合模型进行知识蒸馏的相关研究较为缺乏。基于上述背景，本文开展基于多任务蒸馏的意图识别与槽位填充研究，利用联合模型实现意图识别与槽位填充任务间的特征共享，提高模型准确率，并希望通过多任务知识蒸馏解决预训练模型因性能要求难以在常规配置的智能对话系统进行部署的问题。

本文的主要工作如下：1)提出了基于 ALBERT 的汉语意图识别和槽位填充联合模型，并探究不同权重值对联合模型的影响。本文利用 ALBERT 获取文本特征表示，通过 Softmax 函数进行意图分类。同时将 ALBERT 的输出作为 BiLSTM 的输入，并使用 CRF 解码进行槽位填充。2)创新性地尝试在联合模型的基础上进行蒸馏。利用知识蒸馏，将学习能力强、结构复杂的 ALBERT 模型知识迁移到学习能力弱、结构简单的 BiLSTM 模型中。具体做法是通过 ALBERT 联合模型的输出信息对 BiLSTM 模型进行监督学习，提高其预测精度的同时获得较高的推理效率。

1　相关工作

1.1　意图识别与槽位填充

早期，自然语言理解中的意图识别与槽位填充任务采取独立建模的思路。意图识别本质上是文本分类。最早的意图识别是基于规则模版的方法，通过人工分析每个意图下有代表性的例句总结出规则模板，该方法需要耗费大量人力物力。紧接其后的是基于统计机器学习的方法，其效果比基于规则的方法好，但需要大量人工操作设计领域相关的特征，且无法提取到深层特征，效果仍然不理想。

目前，深度学习已经成为主流范式，利用深度学习模型进行意图识别无需特征工程，相较传统模型，具有性能优势。2017 年，Meng 等[6]通过分层 LSTM 进行意图识别，利用单词级 LSTM 获取句子特征表示，句子级 LSTM 提取上下文依存关系。张志昌等[7]提出一种基于独立循环神经网络和词级别注意力融合的用户意图分类方法，有效提高了模型效果。Wang 等[8]提出 CNN-BGRU 模型进行意图分类，利用 CNN 获取深层文本特征，并用 BiGRU 提取上下文语义信息。

早期的槽位填充也是使用基于规则的方法，这种方法缺乏通用性。槽位填充的本质是序列标注问题，随着深度学习的发展，人们开始尝试把深度学习模型如 RNN 应用到序列标注中。Kurata 等[9]使用 LSTM 对槽位填充任务进行建模，利用 LSTM 编码器将整个输入序列编码为固定维度的向量，然后将其作为另一个 LSTM 的输入向量，将整个句子信息融入序列标注过程中，提高了模型性能。张玉帅等[10]利用预训练模型 BERT 和 LSTM 对输入句子进行特征学习，将 BERT 产生的向量表示作为 LSTM 的输入，再利用 softmax 函数和条件随机场(CRF)进行解码。

1.2　联合模型

在实践中发现，意图识别和槽位填充任务很多时候具有较强的相关性。已有研究将两个任务进行联合建模，充分利用意图和槽位中的语义关联。Xu 等[11]利用 CNN 获取底层特征表示，并用 triangular CRF 对意图和槽位的联合条件概率分布进行建模。Goo 等[12]提出了一种通过槽位门控(slot gate)机制来学习意图和槽位向量之间关系的方法，利用意图识别的结果对槽位填充过程进行限制。E 等[13]在此基础上提出了 SF-ID 模型，该方法通过双向交互机制来增强两个任务之间的联系。以往的模型大多依赖于自回归方法，但 Wu 等[14]发现使用自回归方法对整个序列的依赖关系进行建模会导致冗余计算和高延迟，他们提出采用非自回归的方法对意图识别与槽位填充进行联合建模。Chen 等[15]在 2019 年提出基于 Bert[16]的意图识别与槽位填充联合模型，以解决传统 NLU 模型泛化能力差的问题。2020 年，周奇安等[17]在此基础上进行改良，将 BERT 用作编码器，而解码器基于 LSTM 与注意力机制构建。Dao 等[18]进行了越南语的意图识别与槽位填充相关研究，提出了 JointIDSF 模型，该模型是对

Joint BERT+CRF 模型进行改良,通过注意力机制将意图信息融入到槽位填充过程中。2023 年,孟佳娜等[19]利用 BERT 和 RoBERTa 对意图识别和语义槽填充进行联合建模,同时使用长短期记忆网络对历史信息进行语义建模,以解决人机对话系统研究中多轮对话历史信息的意图识别问题。

1.3 知识蒸馏

预训练模型如 BERT 虽然能获得较好的效果,但也存在参数太多(BERT-base 约 110 M 参数)以及预测效率太低的问题。因此,许多学者开始研究模型压缩,知识蒸馏是其中一个研究方向。知识蒸馏是一种基于"教师-学生"的网络训练方法,旨在将结构复杂的教师模型的特征表示传递给参数小、结构简单的学生模型。Urban 等[20]提出在 logit 上监督训练浅层模型去逼近深层模型的效果。Hinton 等[21]采用了 softmax 层内特征匹配的策略,提出蒸馏温度 T,使蒸馏的性能获得提升。Romero 等[22]尝试将教师模型蒸馏到一个更深但比较小的网络中,证明了复杂的深度模型中间层能有效地对学生模型进行监督学习。Liu 等[23]将经过不同初始化训练的多个模型提取为一个模型。Subramanian 等[4]将蒸馏应用到语音增强中,通过模仿多通道输入的软掩码来获得单通道输入的学生模型。Tang 等[24]提出将 BERT 模型知识蒸馏到 BiLSTM 模型。Sun 等[1]采用 PKD-Last 和 PKD-Skip 两种策略有效地利用教师模型隐含层的信息,对 BERT 进行蒸馏。廖胜兰等[2]将知识蒸馏应用到意图分类任务中,尝试将教师模型 BERT 中的知识迁移到学生模型 Text-CNN 和 Text-RCNN; Denisov 等[5]将蒸馏应用到语音和文本两种模态中,构建一种端到端的口语理解模型。郭师光等[3]于 2021 年利用预训练模型 ERNIE 模型[25]知识蒸馏到 FastText 模型。Fukuda 等[26]提出了多教师蒸馏方案,将多个模型的优势整合到单个学生模型中。石佳来等[27]对基于 BERT 的多教师蒸馏方法进行改进,加入了对中间 Transformer 层的知识的提取。

目前关于联合模型蒸馏的研究较为缺乏, Chen 等[28]提出一种意图识别和槽位填充联合模型,该方法对模型进行自蒸馏,将最后槽位解码器(final slot decoder)作为初始槽位解码器(initial slot decoder)的软标签。Tu 等[29]提出一种意图识别和槽位填充的联合模型,并利用对比学习和自蒸馏方法有效训练该模型。上述工作采用了自蒸馏的模型压缩方法,而其他相关的蒸馏工作[2-3]主要聚焦于单一意图识别任务。与以上研究不同,本文提出了基于 ALBERT[30]的汉语意图识别与槽位填充联合模型,并尝试将其知识蒸馏到 BiLSTM 模型中,以提高模型的推断效率。

2 研究工作

2.1 基于 ALBERT 的意图识别与槽位填充联合模型构建

2.1.1 预训练模型 ALBERT

BERT 是一种基于微调的多层双向变压器编码器。BERT 利用 Masked LM (MLM) 进行预训练,同时引入 Next Sentence Prediction(NSP) 来捕捉句子级的模式,使模型能够理解句子间的关系,最终生成能融合上下文信息的深层双向语言表征。

ALBERT(a lite BERT)对 BERT 进行了 3 点改良:1)将 embedding 的参数进行因式分解。ALBERT 通过将 one-hot 向量映射到大小为 E 的低维词嵌入空间而不是直接映射到大小为 H 的隐藏空间,然后从低维词嵌入空间映射到隐藏空间,实现将 embedding 的参数进行因式分解,使得词嵌入参数从 $O(V \times H)$降低到 $O(V \times E + E \times H)$,减少模型参数量,其中 V 为词表大小,E 为词向量维度,H 为隐层维度。2)跨层的参数共享。通过共享所有层的参数使模型参数较 BERT 大大减少。3)使用 SOP (sentence order prediction)任务取代 NSP (next sentence prediction)任务进行句间连贯性预测。将难度更小的主题预测和连贯性预测结合的 NSP 任务替换成难度更大的句子连贯性预测的 SOP 任务,让模型学习到更多的信息,提升模型的性能。

2.1.2 基于 ALBERT 的联合模型

ALBERT 较 BERT 参数更少,并且 SOP 任务能让模型学习到更多信息。因此本文使用 ALBERT 替代 BERT 进行意图识别和槽位填充联合建模。模型整体框架如图 1 所示。

进行联合训练时,长度为 Ω 的句子的标记嵌入、段嵌入和位置嵌入拼接起来作为 ALBERT 模型的输入, ALBERT 输出融合了全文语义信息后的各个字符 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{n+2}\}$,其中 $\boldsymbol{x}_1$ 是[CLS]标记的编码表示,代表整句话的句向量,$\boldsymbol{x}_{n+2}$ 是[SEP]标记的编码表示,在单句分类中只作为最后一个标记。

$\boldsymbol{x}_1$ 作为整个文本的语义表示分别输入到领域分类和意图识别任务的输出层中进行分类,得到领域分类任务的预测标签 $\overline{\boldsymbol{y}}^d$ 和意图识别任务的预测

标签 $\overline{\boldsymbol{y}}^i$。

$$\overline{\boldsymbol{y}}^d=\text{softmax}(\boldsymbol{W}^d\boldsymbol{x}_1+\boldsymbol{b}^d),\tag{1}$$

$$\overline{\boldsymbol{y}}^i=\text{softmax}(\boldsymbol{W}^i\boldsymbol{x}_1+\boldsymbol{b}^i)。\tag{2}$$

式(1)中的 $\boldsymbol{W}^d$ 和 $\boldsymbol{b}^d$ 为领域分类任务输出层可训练的权重和偏置，式(2)中的 $\boldsymbol{W}^i$ 和 $\boldsymbol{b}^i$ 为意图识别任务输出层的权重和偏置。

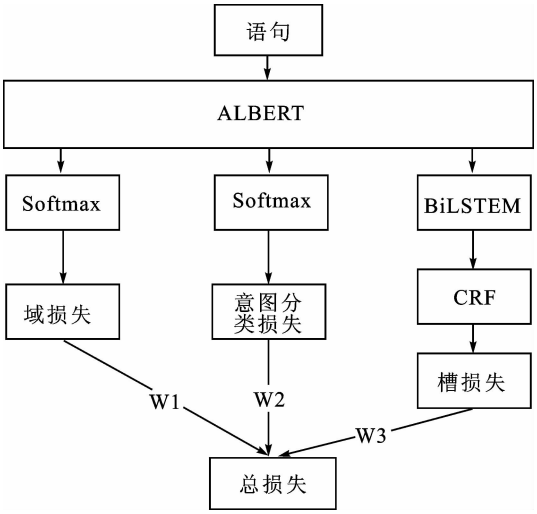

图 1　基于 ALBERT 的意图识别和槽位填充模型

Fig. 1　ALBERT based intent recognition and slot filling model

本文使用交叉熵作为领域分类和意图识别的损失函数，如式(3)所示，式中 m 为样本总数。

$$L_{\text{CE}}=-\sum_{j=1}^{m}y_{(j)}\log\hat{y}_{(j)}。\tag{3}$$

由于存在各个领域、不同意图间数据不平衡的问题。在计算损失时添加一个类别权重 w，类别样本数越少，权重越大，那么在进行反向传播时，求导获得的梯度越大，模型参数变化也越大，进而对模型的影响增大。

类别权重的计算方法如式(4)所示，总样本数为 m，领域类别数为 N，属于某领域的样本数为 m_j^d，其中 $\sum_j m_j^d=m$，则某一领域 j 的类别权重为

$$w_j^d=\frac{m}{N+m_j^d},j\in[1,N]。\tag{4}$$

意图的类别数为 M，属于某意图的样本数为 m_g^i，其中 $\sum_g m_g^i=m$，则某一意图 g 的类别权重为

$$w_g^i=\frac{m}{M+m_g^i},g\in[1,M]。\tag{5}$$

最终，领域分类和意图识别的损失函数如下：式(6)为领域分类的损失，式(7)为意图识别的损失。

$$L_{\text{CE}}^d=-\sum_{j=1}^{m}\boldsymbol{y}_{(j)}^d\log\hat{\boldsymbol{y}}_{(j)}^d\boldsymbol{w}_{(j)}^d,\tag{6}$$

$$L_{\text{CE}}^d=-\sum_{j=1}^{m}\boldsymbol{y}_{(j)}^i\log\hat{\boldsymbol{y}}_{(j)}^i\boldsymbol{w}_{(j)}^i。\tag{7}$$

式中：$\boldsymbol{y}$ 为真实标签；$\hat{\boldsymbol{y}}$ 为预测标签；$\boldsymbol{w}^i$ 为领域的类别权重；$\boldsymbol{w}^d$ 为意图的类别权重。

在槽位填充过程中，$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_{n+2}\}$ 作为文本中每个字的向量表示输入到双向 LSTM 网络，将每个时间序列的正向输出 $\overrightarrow{\boldsymbol{h}}_t$ 和反向输出 $\overleftarrow{\boldsymbol{h}}_t$ 进行拼接得到每个时间步的输出 $\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t]$。将 BiLSTM 的输出序列 $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{n+2})$ 输入到概率模型 CRF 中，进一步增强前后标注的约束，最终获得最优的标签序列 $\hat{\boldsymbol{y}}^s=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{n+2})$。

槽位填充任务训练过程将最大化正确标签序列分数在总体分数的比重，其优化目标如下：

$$p(y\mid h;W,b)=\sum_{j=1}^{m}\frac{e^{s(X,y_{(j)}^s)}}{\sum\limits_{\tilde{y}_{(j)}^s\in Y_X}e^{s(X,\tilde{y}_{(j)}^s)}}。\tag{8}$$

式中：m 为样本总数；y^s 为正确的标签序列；$\tilde{y}^s$ 为所有可能的标签序列；$e^{s(X,y^s)}$ 为标签序列为正确序列的分数；$e^{s(X,\tilde{y}^s)}$ 为每种可能序列的分数。

分数包括发射分数和转移分数。发射分数由 BiLSTM 生成，表示一个词选择不同标签的分数，用 E_{k,y_k} 表示。将发射分数输入到 CRF 中，由 CRF 学习到转移矩阵(转移分数)，转移分数表示前一个标签转移到此时标签的分数，用 $T_{y_k,y_{k+1}}$ 表示。其中，k 是单词的位置索引，y_k 是类别的索引。整体的分数为

$$\text{score}(\boldsymbol{X},y)=\sum_{k=1}^{n}T_{y_{k-1},y_k}+\sum_{k=1}^{n}E_{k,y_k}。\tag{9}$$

式(8)取似然对数作为最终槽位填充任务的损失函数

$$L^s=-\sum_{j=1}^{m}\Big[s(\boldsymbol{X},y_{(j)}^s)-\log\Big(\sum_{\tilde{y}_{(j)}^s\in Y_X}e^{s(\boldsymbol{X},\tilde{y}_{(j)}^s)}\Big)\Big]。\tag{10}$$

最终，基于 ALBERT 的意图识别和槽位填充的联合损失函数为

$$L^{\text{total}}=W_dL^d+W_iL^i+W_sL^s。\tag{11}$$

ALBERT 虽然减少了参数的数量，但并不会提高预测效率，因为虽然每层参数共享，但前向传播时还是要对每一层每一个参数进行计算。所以在 ALBERT 模型训练好后，直接获取模型的预测概率，用于知识蒸馏。

2.2　联合模型知识蒸馏策略

2.2.1　基于 BiLSTM 的学生模型

本文采用 BiLSTM 作为学生模型。相比 ALBERT，BiLSTM 是一种浅层神经网络，将 ALBERT 知识蒸馏 BiLSTM，使模型参数进一步

减少,并且以更低的复杂度来获得类似的预测效果。LSTM 虽能很好地捕获较长距离的依赖关系,但无法编码从后向前的信息,而 BiLSTM 能从正向和反向两个方向同时对序列进行建模,更好地捕获双向语义信息。

因此,本文使用 BiLSTM 作为学生模型,更好地捕获双向语义的同时学习到 ALBERT 知识,从而在提高 BiLSTM 模型预测精度的同时,保持较高的推理效率。

2.2.2 基于 ALBERT 的知识蒸馏

本文采用联合蒸馏方式,同时进行意图识别与槽位填充任务的知识蒸馏。知识蒸馏的总体流程如图 2 所示。先将数据集输入到 ALBERT 中进行意图识别与槽位填充联合模型的训练,获取到教师模型(teacher model)。之后将数据集输入 BiLSTM 中进行训练,将 BiLSTM 输出的概率向量与 hard target、教师模型输出的 soft target 相结合计算知识蒸馏损失,使 BiLSTM 的输出分布在保持精确的同时能够尽可能地逼近教师模型,从而学到 ALBERT 的知识。

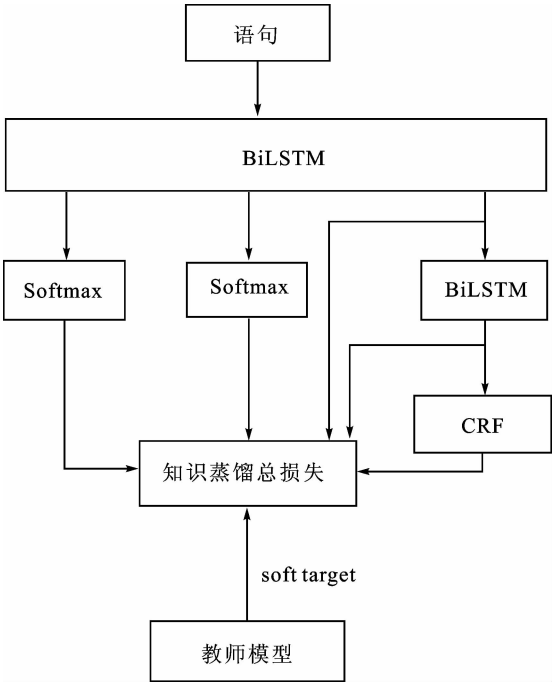

图 2 蒸馏模型

Fig. 2 Distillation model

知识蒸馏的损失函数包含两个部分。其一是 hard target,该部分与教师模型 ALBERT 的损失函数相同。领域分类和意图识别任务使用交叉熵损失函数,同样为了解决类不平衡的问题,在计算损失时添加一个类别权重 w。槽位填充任务使用 CRF 层的损失函数。公式如下:

$$L_{\mathrm{CE}} = -\sum_{j=1}^{m} \boldsymbol{y}_{(j)} \log \hat{\boldsymbol{y}}_{(j)} \boldsymbol{w}_{(j)}, \tag{12}$$

$$L_{\mathrm{CE}}^{s} = -\sum_{j=1}^{m} \left[s(X, y_{(j)}^{s}) - \log\Big(\sum_{\tilde{y}_{(j)}^{s} \in Y_X} e^{s(X, \tilde{y}_{(j)}^{s})} \Big) \right]。 \tag{13}$$

式中:$\boldsymbol{y}$ 为对应任务的真实标签;$\hat{\boldsymbol{y}}$ 为对应任务的预测标签。

其二是 soft target,使用教师模型输出的概率向量与 BiLSTM 输出的概率向量计算均方误差,可以使 BiLSTM 学习到 ALBERT 中更加软化的知识。soft target 包含类别间的信息,这是传统 one-hot label 中没有的。通过最小化损失函数,使学生模型 BiLSTM 的输出分布尽可能地逼近教师模型,从而实现知识蒸馏的目的。公式如下:

$$L_{\mathrm{distil}} = \| \boldsymbol{z}^{(t)} - \boldsymbol{z}^{(s)} \|_2^2。 \tag{14}$$

式中:$\boldsymbol{z}^{(t)}$ 是教师模型对应任务的输出;$z^{(s)}$ 是学生模型对应任务的输出。

最终,每个任务的相应损失函数如下:

$$L = \alpha \cdot L_{\mathrm{CE}} + (1-\alpha) \cdot L_{\mathrm{distil}}。 \tag{15}$$

由于 Softmax 输出分布在正确位置的值会非常大,其他位置很小,对损失函数的影响会比较小,因此直接用教师网络的概率分布对学生网络进行监督学习往往效果有限。相关研究[30] 对 Softmax 函数进行改进,增加参数 T 来调整 Softmax 的输出分布,改进的 Softmax 函数如下:

$$y_i = \frac{\mathrm{e}^{\frac{x_i}{T}}}{\sum_{j=1}^{N} \mathrm{e}^{\frac{x_i}{T}}}, \tag{16}$$

其中 T 数值越大,分布越缓和。当 T 等于 1 时,式(16)相当于原始 Softmax 函数。

具体训练步骤如下:选定合适的 T 值,先在教师模型中训练并预测,然后在相同的 T 值下训练学生模型。当学生模型进行预测时,将值 T 设置为 1,经原始 Softmax 函数获取输出概率。

任务型对话系统中的槽位填充本质上是序列标注,利用 CRF 解码可以获得更好的模型效果。但 CRF 利用相邻标签之间的相关性对标签序列进行全局建模,经 CRF 层解码所得结果丢失了类别间的信息,这增加了从教师模型中提取知识的难度[31]。因此,本文在进行槽位填充任务时,提出使用 BiLSTM 输出的 logits,即 CRF 层的输入对学生模

型进行监督学习，其比 hard label 包含更丰富的信息。在此基础上，为增强学生模型的拟合能力，本文还将 ALBERT 的输出作为学生模型输出的拟合目标。由于 ALBERT 的输出层维度和学生模型输出维度不一样，因此增加了一个额外的线性矩阵进行维度转换。槽位填充的知识蒸馏流程如图 3 所示。

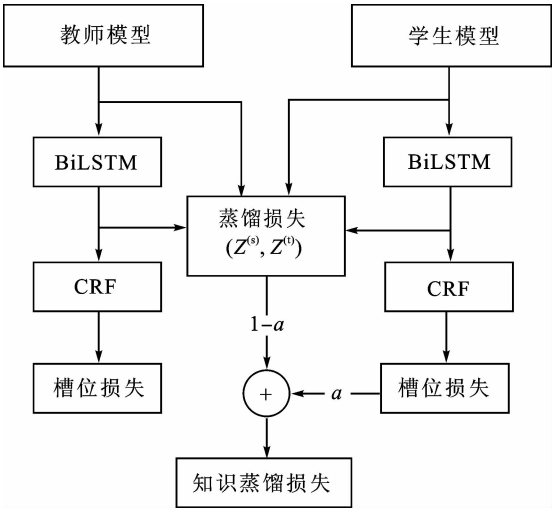

图 3　知识蒸馏损失函数

Fig. 3　Knowledge distillation loss function

除了槽位填充的知识蒸馏损失，领域分类和意图识别都使用教师模型经改进的 Softmax 函数输出的概率向量作为 soft target。最终，对 3 个任务的损失函数加权求和得出总损失函数

$$L_{\text{distill}}^{\text{total}} = W_1 L^d + W_2 L^i + W_3 L^s。\tag{17}$$

3　实验与分析

3.1　数据集

本文使用 SMP 2019 会议的任务型对话系统评测语料作为数据集。其主要包括下面 3 个子任务：领域分类、意图识别和槽位填充。一共涵盖 29 种领域、24 种意图和 60 种语义槽。其中，训练集、验证集和测试集的样本大小分别为 1 694、180、180，部分训练语料如表 1 所示。

3.2　评价指标

本文实验中领域分类和意图识别任务评测指标为准确率，槽位填充任务的评测指标为 F_1 度量值。

本文使用句准确率（sentence accuracy）来衡量模型在解决领域分类、意图识别和槽位填充 3 个问题上的综合能力，即以上 3 项结果全部正确时才算正确。计算公式如下：

句准确率＝3 项任务都取得正确结果的句子数/总句子数。

表 1　部分训练语料

Tab. 1　Part of the training corpus

编号	文本	领域	意图	槽位
1	请帮我打开 uc。	app	LAUNCH	{name:uc}
2	去深圳怎么坐车？	map	ROUTE	{endLoc_city:深圳}
3	北京到成都的汽车时刻表。	bus	QUERY	{Dest:成都,Src:北京}
4	给我放一部最新的电影。	cinemas	QUERY	{}

3.3　实验分析

在实验中先使用 ALBERT 模型对训练语料进行学习，模型的初始参数来源于 github，具体的训练参数如表 2 所示。然后使用 ALBERT 模型作为教师模型，BiLSTM 模型为学生模型，将 ALBERT 知识蒸馏到 BiLSTM 中，BiLSTM 模型的学习率为 1×10^{-3}，其他超参数设置与 ALBERT 模型相同。本文对比了 BERT-base、ALBERT-base、BiLSTM 和 KD BiLSTM（蒸馏模型）在测试集以及验证集上的表现，实验结果如表 3 和表 4 所示。

表 2　ALBERT 模型训练参数

Tab. 2　Training parameter of ALBERT model

实验参数	参数值	参数说明
Learning rate	5×10^{-5}	Adam 优化器学习率
Batch_size	12	每次模型更新的训练样本数
epoch	15	训练迭代轮数
temperature	3.5	蒸馏温度 T
Max_seq_length	32	句子最大长度

表 3 测试集实验结果

Tab. 3 The experimental results of the test set

模型	领域分类精度/%	意图识别精度/%	槽位填充 F_1 值/%	句准确率/%	时间/s
BERT-base	96.11	95.55	81.97	78.33	7.86
ALBERT-base	94.44	96.11	83.04	77.74	7.37
BiLSTM	86.66	91.66	68.12	58.33	0.38
KD BiLSTM(蒸馏模型)	90.56	92.77	73.57	67.22	0.39

表 4 验证集实验结果

Tab. 4 The experimental results of validation set

模型	领域分类精度/%	意图识别精度/%	槽位填充 F_1 值/%	句准确率/%	时间/s
BERT-base	95.55	93.88	84.49	76.66	7.61
ALBERT-base	93.33	93.89	82.35	77.22	7.35
BiLSTM	85.00	88.88	67.92	57.77	0.37
KD BiLSTM(蒸馏模型)	88.89	89.44	73.89	63.33	0.42

从表 3 和表 4 可以看出,ALBERT 相比 BERT 在参数减少的情况下,模型性能相差不大,证明了 ALBERT 模型压缩的有效性,但其推断速度并没有明显变化。蒸馏模型虽然在 3 个任务上的表现不如教师模型 ALBERT,但对比单独训练的学生模型,在测试集的领域分类、意图识别和槽位填充上分别获得了 3.90%、1.11%、5.45%的提升,而且在推断总耗时上,蒸馏模型比教师模型快了 18.9 倍,这体现了知识蒸馏对模型压缩的有效性。

另一点值得注意的是,蒸馏模型相比教师模型在槽位填充任务上的效果相差较大,可能的原因是序列标注任务的复杂度导致其学习效果不佳,但由表 3 可以看出,蒸馏模型在测试集槽位填充上的表现还是比单独训练的学生模型高出 5.45%,这表明知识蒸馏对模型的训练很有帮助。

综上所述,本文提出的 ALBERT 模型取得了不错的效果并证明了其优越性。而将其知识蒸馏到 BiLSTM 模型中,可获得较快的推断速度。此外,虽然蒸馏模型在句准确率上会有所降低,但模型复杂度和参数数量相对于 ALBERT 模型具有优势,对硬件资源要求不高,能部署在小型的配置环境中。

3.4 不同权重对 ALBERT 模型的影响分析

本文探索不同的权重值对教师模型 ALBERT 的影响,联合模型能实现多个任务之间的信息共享,但其损失函数需要考虑到各任务的优化目标,一般做法是对各任务的优化目标进行加权求和,根据子任务的训练情况分配不同的权重值。实验表明,寻找到合适的权重参数能有效提高模型性能,实验结果如表 5 所示。

表 5 不同权重值对 ALBERT 模型的影响

Tab. 5 Influence of different weight values on ALBERT model 单位:%

模型	领域分类精度	意图识别精度	槽位填充 F_1 值	句准确率
$W_i=1,W_d=1,W_s=1$	93.88	95.00	80.61	77.22
$W_i=1,W_d=1,W_s=2$	94.44	96.11	83.04	77.74
$W_i=1,W_d=1,W_s=3$	95.56	95.56	82.17	76.67
$W_i=1,W_d=1,W_s=4$	93.89	95.00	80.61	72.22
$W_i=2,W_d=1,W_s=1$	95.00	95.00	81.66	76.67
$W_i=1,W_d=2,W_s=1$	93.89	97.22	76.25	72.22
$W_i=2,W_d=2,W_s=1$	94.44	96.67	81.20	75.56

从表 5 可以看出在 $W_i=1,W_d=1,W_s=2$ 的情况下,模型能获得最好的效果。之后,随着槽位填充任务权重的增加,领域分类与意图识别的效果受到影响,句准确率下降。另外,在 W_i、W_d 大于 W_s 的时候,模型效果下降,这说明序列标注任务更复杂,模型容易受分类任务影响,需要增加槽位填充任务权重值,使模型在训练过程中能更好地学习到各任务之间的关系。

3.5 探究类别权重对模型的影响

在任务型对话系统数据集中,领域和意图的类别比较多,对于样本比较少的数据集,其往往存在各个领域、不同意图间的数据不平衡的问题。对于样本比例失衡问题,本文通过类别权重对模型进行优化。实验结果如表 6 所示。

表 6　探究类别权重对模型的影响

Tab. 6　The influence of class weight on the model　　单位：%

模型	领域分类精度	意图识别精度	槽位填充 F_1 值	句准确率
ALBERT	94.44	96.11	83.04	77.74
KD BiLSTM	90.56	92.77	73.57	67.22
BERT(no_adjust)	93.33	92.77	82.09	74.44
ALBERT(no_adjust)	93.33	94.44	81.65	75.00
KD BiLSTM(no_adjust)	83.89	91.67	70.64	60.00

通过实验可以看出，与没有添加类别权重的 ALBERT 和 BERT 模型相比，添加了类别权重的 ALBERT 模型在领域分类、意图识别和槽位填充 3 个任务上的效果均获得提升，这表明类别权重的引入能有效解决各类别之间的数据不平衡问题。另外，通过表 6 可以发现，进行类别平衡调整的蒸馏模型比没有经过调整的效果要好，其分别在领域分类、意图识别和槽位填充上获得了 6.67%、1.10%、2.93%的提升，表明了类别不平衡问题的解决对知识蒸馏的结果有较大影响。

4　结语

本文对任务型对话系统进行了研究。目前，虽然预训练模型如 BERT 能在该领域中取得良好效果，但由于其模型复杂、参数量大、推断速度慢等问题，往往难以支持运行环境要求较高的实际业务场景应用。因此，本文提出基于 ALBERT 的意图识别和槽位填充联合模型，大幅度减少预训练模型参数，并且在联合模型的基础上，利用知识蒸馏，将 ALBERT 知识迁移到 BiLSTM 模型，提高了 BiLSTM 模型的泛化能力，并获得较高的推断速度。同时本文还尝试对数据中意图类别不平衡问题进行研究，使用类别权重对模型进行调整。该模型在 SMP 2019 会议评测数据集上进行实验，结果表明，基于 ALBERT 的意图识别与槽位填充联合模型能获得 77.74%的准确率，而蒸馏模型在句准确率为 67.22%的情况下，推断速度约为 ALBERT 的 18.9 倍。此外，未来我们将致力于进一步提升模型的推断速度、减少模型参数数量，同时使其在句准确率上取得更好的性能水平。

参考文献：

[1] SUN S Q, CHENG Y, GAN Z, et al. Patient knowledge distillation for BERT model compression [EB/OL]. [2023-10-17]. http://arxiv.org/abs/1908.09355.pdf.

[2] 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.

[3] 郭师光，崔英花，黄惠燕. 基于 ERNIE 模型的知识蒸馏在智能对话意图识别的应用[J/OL]. 计算机应用，[2023-10-17]. http://kns.cnki.net/kcms/detail/51.1307.TP.20210315.1404.004.html.
GUO S G, CUI Y H, HUANG H Y. Application of knowledge distillation based on the ERNIE model in intelligent dialogue intention recognition[J/OL]. Journal of Computer Applications, [2023-10-17]. http://kns.cnki.net/kcms/detail/51.1307.TP.20210315.1404.004.html.

[4] SUBRAMANIAN A S, CHEN S J, WATANABE S. Student-teacher learning for BLSTM mask-based speech enhancement[EB/OL]. [2023-10-17]. https://arxiv.org/pdf/1803.10013.

[5] DENISOV P, VU N T. Pretrained semantic speech embeddings for end-to-end spoken language understanding via cross-modal teacher-student learning [EB/OL]. [2023-10-17]. http://arxiv.org/abs/2007.01836.

[6] MENG L, HUANG M L. Dialogue intent classification with long short-term memory networks[M]// Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2018: 42-50.

[7] 张志昌，张珍文，张治满. 基于 IndRNN-Attention 的用户意图分类[J]. 计算机研究与发展，2019，56(7)：1517-1524.
ZHANG Z C, ZHANG Z W, ZHANG Z M. User intention classification based on IndRNN-Attention[J]. Journal of Computer Research and Development, 2019, 56(7): 1517-1524.

[8] WANG Y F, HUANG J W, HE T T, et al. Dialogue intent classification with character-CNN-BGRU networks [J]. Multimedia Tools and Applications, 2020, 79(7): 4553-4572.

[9] KURATA G, XIANG B, ZHOU B W, et al. Leveraging sentence-level information with encoder LSTM for semantic slot filling[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1601.01530.

[10] 张玉帅，赵欢，李博. 基于 BERT 和 BiLSTM 的语义槽填充[J]. 计算机科学，2020，48(1)：247-252.
ZHANG Y S, ZHAO H, LI B. Semantic slot filling based on BERT and BiLSTM[J]. Computer Science, 2020, 48(1): 247-252.

[11] XU P Y,SARIKAYA R. Convolutional neural network based triangular CRF for joint intent detection and slot filling[C]// 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc:IEEE,2013:78-83.

[12] GOO C W,GAO G,HSU Y K,et al. Slot-gated modeling for joint slot filling and intent prediction[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,Volume 2 (Short Papers). New Orleans:Association for Computational Linguistics,2018:753-757.

[13] E H H,NIU P Q,CHEN Z F,et al. A novel Bi-directional interrelated model for joint intent detection and slot filling[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1907.00390.

[14] WU D,DING L,LU F,et al. SlotRefine:a fast non-autoregressive model for joint intent detection and slot filling[EB/OL]. [2023-10-17]. http://arxiv.org/abs/2010.02693.

[15] CHEN Q,ZHUO Z,WANG W. BERT for joint intent classification and slot filling[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1902.10909.

[16] DEVLIN J, CHANG M W, LEE K, et al. BERT:pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, Seatle: NAACL, 2019, 4171 4186.

[17] 周奇安,李舟军. 基于 BERT 的任务导向对话系统自然语言理解的改进模型与调优方法[J]. 中文信息学报,2020,34(5):82-90.
ZHOU Q A,LI Z J. Improved model and tuning method of natural language understanding in task-oriented dialogue system based on BERT[J]. Journal of Chinese Information Processing,2020,34(5):82-90.

[18] DAO M H,TRUONG T H,NGUYEN D Q. Intent detection and slot filling for Vietnamese[EB/OL]. [2023-10-17]. http://arxiv.org/abs/2104.02021.

[19] 孟佳娜,单明,孙世昶,等. 融入历史信息的多轮对话意图识别[J]. 大连民族大学学报,2023,25(3):244-249.
MENG J N,SHAN M,SUN S C,et al. Multi-round dialogue intention recognition with historical information[J]. Journal of Dalian Minzu University,2023,25(3):244-249.

[20] URBAN G,GERAS K J,KAHOU S E,et al. Do deep convolutional nets really need to be deep and convolutional?[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1603.05691.

[21] HINTON G,VINYALS O,DEAN J. Distilling the knowledge in a neural network[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1503.02531.

[22] ROMERO A,BALLAS N,KAHOU S E,et al. fitNets:hints for thin deep nets[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1412.6550.

[23] LIU Y J, CHE W X, ZHAO H P, et al. Distilling knowledge for search-based structured prediction[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1805.11224.pdf.

[24] TANG R,LU Y,LIU L Q,et al. Distilling task-specific knowledge from BERT into simple neural networks[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1903.12136.pdf.

[25] SUN Y,WANG S H,LI Y K,et al. ERNIE:enhanced representation through knowledge integration[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1904.09223.pdf.

[26] FUKUDA T,KURATA G. Generalized knowledge distillation from an ensemble of specialized teachers leveraging unsupervised neural clustering[C]// ICASSP 2021 - 2021 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). Toronto:IEEE,2021:6868-6872.

[27] 石佳来,郭卫斌. 一种针对 BERT 模型的多教师蒸馏方案[J/OL]. 华东理工大学学报(自然科学版). [2023-10-17]. https://doi.org/10.14135/j.cnki.1006-3080.20230118001.
SHI J L, GUO W B. A multi teacher distillation scheme for BERT model[J/OL]. Journal of East China University of Science and Technology. [2023-10-17]. https://doi.org/10.14135/j.cnki.1006-3080.20230118001.

[28] CHEN L S,ZHOU P L,ZOU Y X. Joint multiple intent detection and slot filling via self-distillation[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore:IEEE,2022:7612-7616.

[29] TU N A,UYEN H T T,PHUONG T M,et al. Joint multiple intent detection and slot filling with supervised contrastive learning and self-distillation[M]// Frontiers in Artificial Intelligence and Applications. Amsterdam:IOS Press,2023.

[30] LAN Z Z,CHEN M D,GOODMAN S,et al. ALBERT:a lite BERT for self-supervised learning of language representations[EB/OL]. [2023-10-17]. http://arxiv.org/abs/1909.11942.pdf.

[31] WANG X Y,JIANG Y,BACH N,et al. Structure-level knowledge distillation for multilingual sequence labeling[EB/OL]. [2023-10-17]. http://arxiv.org/abs/2004.03846.pdf.

〔责任编辑 宋轶文〕