文章编号：2096-1618(2023)03-0251-07

面向任务型对话机器人的多任务联合模型研究

高作缘， 陶宏才
(西南交通大学计算机与人工智能学院，四川 成都 611756)

摘要：在任务型对话机器人的搭建过程中，一般需要执行多个自然语言处理的子任务。目前传统的训练方式是将每个子任务独立训练后再进行整合，这样忽视了不同子任务之间的关联性，限制了模型的预测能力。现提出一种 Joint-RoBERTa-WWM-of-Theseus 压缩联合模型，一方面通过多任务联合学习训练的方式对意图识别、行业识别和语义槽填充3个子任务进行联合训练，并在多分类的子任务中引入 Focal loss 机制来解决数据分布不平衡的问题；另一方面，模型通过 Theseus 方法进行压缩，在略微损失精度的前提下，大幅提高模型预测速度，提高模型在生产环境中的实时性与实用性。

关 键 词：RoBERTa-WWM 模型；多任务联合学习；Theseus 压缩；Focal loss
中图分类号：TP391.12 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2023.03.001

0 引言

在任务型对话机器人[1]的搭建中，准确地理解用户的意图、判断语句中可能包含的行业分类、提取语句中的关键信息至关重要。因为对话机器人会依据内容，再结合对话的上下文信息来决定下一步的行为走向。

近年来，随着大规模预训练 BERT(bidirectional encoder representation from transformers)模型的提出[2]，自然语言处理的发展迈入了新阶段。BERT 模型采用了双向的 Transformer 作为特征提取器，性能有显著提升，刷新了自然语言处理的多项记录[3]。另外，应用迁移学习[4]后的自然语言处理不再受数据源的限制，轻松解决了目标领域数据样本不足的问题。而 RoBERTa-WWM 模型[5]作为 BERT 模型的改进版本，采用了更大的模型参数量、更多的训练数据和更大的 batch size，还引入了动态掩码、文本编码，比 BERT 模型更好地推广到下游任务。因此，本文将以 RoBERTa-WWM 模型为基础开展研究，实现意图识别和行业识别的子任务。在此基础上，该部分还引入了 Focal loss 机制，解决多分类中数据不平衡的问题，提高模型的稳定性和性能。

语义槽填充子任务本质就是序列标注问题[6]，主要目的就是提取语句中的实体信息，并填充到对应的语义槽中。在序列标注问题中，BiLSTM+CRF 是非常经典的模型[7]。该模型通过双向 LSTM 能更好地捕捉序列中上下文的信息，提高标注的准确性；通过条件随机场(CRF)可以获取全局最优解，避免出现不合理的标注结果。最后，再引入 RoBERTa-WWM 模型来获取语义表示，提高模型的整体性能。综上，将采用 RoBERTa-WWM-BiLSTM-CRF 模型来完成语义槽填充子任务。

自然语言处理常见的任务有文本分类、序列标注、自动文摘等[8]。意图识别和行业识别实际就是文本分类任务，再加上本质为序列标注任务的语义槽填充，模型需处理3个子任务。而在传统的自然语言处理算法中，面对多任务时一般采用不同子任务独立训练，最终以结合的方式来完成模型的整合。而在实际的语言表达中，意图识别、行业识别和语义槽填充3个子任务并非完全孤立[9]，其中一个子任务的预测结果很可能影响其他子任务的预测过程。因此，提出 Joint-RoBERTa-WWM 模型，将意图识别、行业识别和语义槽填充进行联合学习，强化子任务之间的关联性，提高模型的综合预测能力。

研究表明[10]，像 BERT 这样基于 Transformer 的预训练模型，存在参数设置过多、模型过厚重、计算成本过高的问题。因此，本文基于模型的工程性应用考虑，进一步提出了一种基于模块替换[11]的压缩联合模型 Joint-RoBERTa-WWM-of-Theseus，在略微损失预测精度的前提下，大幅加快了预测速度，提高了模型的实时性，更好地为实际工程应用提供服务。

1 相关工作

1.1 任务型机器人

1950年，Turing[12]提出了图灵测试。之后，围绕

收稿日期：2023-02-20
基金项目：国家自然科学基金资助项目(61806170)

人机对话的研究逐渐成为了人机交互[13]领域中的核心研究内容,而对话系统是实现人机对话最直观的表现形式。经过基于规则模板的对话系统、统计对话系统和神经对话系统等 3 个阶段的发展[14],对话系统已经开始向对话机器人演变。在应用场景上,对话机器人可分成 3 类:问答型机器人(QA robot)、闲聊型机器人(chat robot)和任务型机器人(task robot)。

问答型机器人主要为一问一答的形式,机器人在解析用户提出的问题后,需要在知识库中搜索相关的正确答案并将结果返回给用户。其中,每次问答均是独立的,与上下文对话无关。而闲聊型机器人主要以满足用户的情感需求为主,通过有趣、个性化的回复与用户进行互动,较知名产品有微软的小冰。相对于前面两类对话机器人,任务型机器人可以满足更复杂的业务需求,一般指的是机器人为了满足用户的需求目标从而产生多轮对话,通过在对话中不断澄清或调整用户意图完成用户的请求。这就要求机器人能整合上下文信息,根据上一轮对话的内容来决定下一轮对话的子目标。典型的任务型机器人有阿里巴巴的天猫精灵[15]、苹果的 Siri[16]和微软的 Cortana(小娜)[17]。

目前,任务型对话机器人被广泛使用于不同领域的多个场景,如客服行业、医疗行业、生活娱乐场景等。在生活场景中,任务型对话机器人的出现能够帮助人们更方便快捷地工作,提高效率。以 Siri 为例,它可以帮助机主完成打电话、发短信、播放歌曲等任务。在执行任务的过程中,Siri 需要先对机主的语音消息进行识别,再根据识别结果执行意图、领域的预测和语义槽填充 3 个子任务,最后再根据预测结果做出相应的行为来帮助机主完成该次任务。类似此应用场景,本文的模型将用于搭建电商行业的智能客服机器人,因此模型的预测主要包括了意图识别、行业识别和语义槽填充 3 个子任务。意图、行业的预测和语义槽填充的样例如表 1 所示。

表 1 子任务结果样例

语句	意图	行业	语义槽
给我推荐一下黑色的手机	商品求购	手机	颜色:黑色
店里有没有双开门冰箱	商品求购	冰箱	门款式:双开门

1.2 RoBERTa-WWM 模型

BERT 模型的训练过程主要包含掩码语言模型(mask language model, MLM)和下一句预测(next sentence prediction, NSP)两个重要任务。其中,掩码语言模型的原理是随机选取输入序列中 15% 的 Token,在已经选取的 Token 中,以 80% 的概率用标记[MASK]替换掉原始 Token,以 10% 的概率将原始 Token 替换为随机 Token,以剩余 10% 的概率保持原有 Token,这样可以大大提高模型的泛化能力。而 NSP 主要用于判断两个句子之间的关系,对自然语言推理(natural language inference, NLI)这样的下游任务起到至关重要的作用。

对比 BERT 模型,RoBERTa 模型的改进主要体现在:(1) RoBERTa 模型移除了 NSP 任务,采用 Full-Sentences 方式,可以从一篇或多篇文章中连续抽取句子填充到模型的输入序列中,提高了效率。(2) BERT 模型采用的是 Character 级别的字节对编码(byte-pair encoding, BPE),词表大小仅有30 KB;而 RoBERTa 模型采用了 Byte 级别的字节对编码,词表大小50 KB左右,比 BERT 模型词表大近 70%。(3) BERT 模型只在数据预处理期间执行一次掩码,得到一个静态掩码,这样会导致每次训练时 mask 位置都相同,使模型学习的语句模式比较单一;而 RoBERTa 模型采用动态掩码,每次向模型输入一个序列时都会随机地 mask 不同的 Token,可以保证模型逐渐适应不同的掩码策略,学习不同的语言表征。(4) RoBERTa 模型通过采用更大的 batch size、更多的训练数据和训练步骤,较 BERT 模型表现更好。通过以上 4 个方面的改进,RoBERTa 模型在自然语言理解基准测试 RACE、GLUE 和 SQuAD 中达到了 SOTA。

而 RoBERTa-WWM 模型就是在 RoBERTa 模型的基础上,采用全词掩码(whole word masking, WWM)策略。在中文文本中,采用原始策略可能会使一个词语中只有部分字被 mask,而采用 WWM 策略可以使整个词语都被 mask,这样能增强文本的表示效果。

在模型结构上,RoBERTa-WWM 模型继承了 BERT 模型的特点,由 12 层双向 Transformer 组成。初始文本输入,用 $W=\{w_1,w_2,w_3,\cdots,w_n\}$ 表示;模型的输入为该文本字向量、段向量和位置向量的和,用 $E=\{e_1,e_2,e_3,\cdots,e_n\}$ 表示;模型的输出向量用 $T=\{t_1,t_2,t_3,\cdots,t_n\}$ 表示。RoBERTa-WWM 模型结构如图 1 所示。

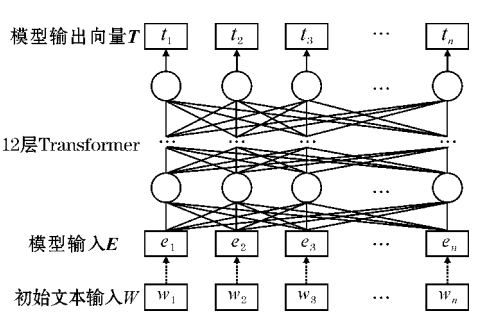

图 1 RoBERTa-WWM 模型结构图

1.3 BiLSTM-CRF 模型

循环神经网络(recurrent neural network, RNN)是一种用来处理序列数据的神经网络[19],它能挖掘数据中的时序信息和语义信息。但是,在实际应用中,RNN 因为单元堆叠导致梯度爆炸或消失较明显。为解决这个问题,Hochreiter 等[20]在 1997 年提出了长短期记忆(long short-term memory,LSTM)网络概念。LSTM 作为 RNN 的一种变体,通过在隐藏层加入记忆单元和门控制器结构,使其具备长期记忆的能力。在 LSTM 中,每个重复的神经元都有三类门,分别为遗忘门(f_t)、输入门(i_t)和输出门(o_t)。LSTM 的单元结构如图 2 所示。

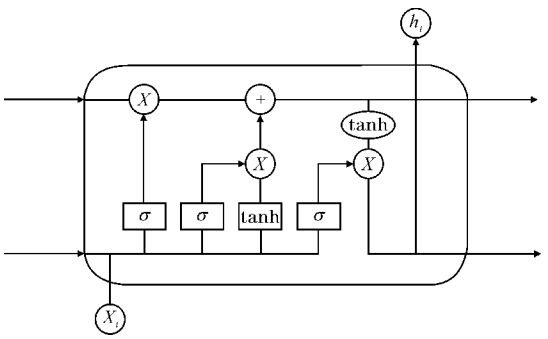

图 2　LSTM 单元结构图

在 LSTM 中,第一步,计算遗忘门,确定要遗忘的信息。遗忘门由 h_{t-1} 和 x_t 线性变换后通过 sigmoid 函数计算输出后并与 C_{t-1} 相乘。遗忘门的计算如下:

$$f_t=\sigma(W_f\times[h_{t-1},x_t]+b_f)$$

第二步,确定要记忆的信息,通过 sigmoid 函数决定需要更新的值 i_t,再通过 tanh 函数创建一个新的候选值向量 $\widetilde{\boldsymbol{C}}_t$,并将其加入到神经元状态中,对神经元状态进行更新得到 C_t。第二步的计算如下:

$$i_t=\sigma(W_i\times[h_{t-1},x_t]+b_i)$$
$$\widetilde{\boldsymbol{C}}_t=\tanh(W_C\times[h_{t-1},x_t]+b_C)$$
$$C_t=f_t\times C_{t-1}+i_t\times\widetilde{\boldsymbol{C}}_t$$

第三步,基于当前时刻的隐藏层状态来决定最终的输出。首先使用 sigmoid 函数决定输出神经元状态的部分 o_t,再使用 tanh 函数处理神经元状态,最后与门控值相乘后即可得到当前时刻的输出 h_t。第三步的计算如下:

$$o_t=\sigma(W_o\times[h_{t-1},x_t]+b_o)$$
$$h_t=o_t\times\tanh(C_t)$$

在此基础上,Fukada[21]提出了双向长短期记忆(bi-directional long short-term memory, BiLSTM)网络概念,更好地捕捉双向的语义依赖。BiLSTM 模型由前向 LSTM 和后向 LSTM 组成,相较于单向 LSTM,它可以获得更加完整的上下文语义信息。但是 BiLSTM 模型的输出没有考虑标签之间的约束和依赖关系,可能会输出无效的序列。如预测的实体开头应该是“B-”而非“I-”,句子的开头应该是“B-”或“O”。为了解决这个问题,在模型中引入条件随机场(condition random field, CRF)模型,为 BiLSTM 模型的输出添加约束关系,保证输出序列的正确性。CRF 由 Lafferty 等[22]于 2001 年提出,结合最大熵模型和隐马尔科夫模型的特点,在序列标注任务中表现突出。在 CRF 中,对于指定的输入序列 $x=(x_1,x_2,\cdots,x_t)$,预测序列 $y=(y_1,y_2,\cdots,y_t)$ 的得分:

$$S(X,Y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i}$$

式中,P 为 BiLSTM 的输出,$\boldsymbol{A}$ 为转移分数矩阵,$A_{i,j}$ 为标签 i 转移到标签 j 的分数。进一步,预测序列 Y 产生的概率:

$$P(Y|X)=\frac{e^{S(X,Y)}}{\sum_{\widetilde{Y}\in Y_X}S(X,\widetilde{Y})}$$

式中,$\widetilde{Y}$ 表示真实的标注序列;Y_X 表示所有可能的标注序列。最后,使用 Viterbi 算法寻找所有 Y 中得分最高的 Y^*:

$$Y^*=\text{argmax}(S(X,\widetilde{Y}))$$

综上,在 BiLSTM 模型后接入条件随机场可以保证最终获取一个有效的预测结果,得到全局最优序列。

2 多任务联合模型及模型压缩

2.1 Joint-RoBERTa-WWM 联合模型

意图识别和行业识别这两个子任务本质就是文本分类问题,本文将使用 RoBERTa-WWM 模型通过对下游任务进行微调来实现。模型的初始输入是文本语句,语句经过分词后形成“[CLS] 语句 [SEP]”的结构。“[CLS]”标签先经过 Encoder 的向量表征,再经过 Pooler 后就能得到句子的向量表征,最后通过 softmax 函数就可以实现文本分类任务,输出句子所属的意图和行业。

在实验过程中发现,数据集中的意图类别分布极其不均匀,导致模型的稳定性较差。因此,在处理意图识别和行业识别这两个多分类子任务时,引入 Focal loss[23]机制,通过改进损失函数来兼顾数据量少的类别。这样,既不影响数据集的原始分布,也能有效提高模型的性能。Focal loss 是交叉熵损失函数(CE loss)的优化版本,简单的二分类交叉熵损失函数如下:

$$CE(p,y)=\begin{cases}-\lg(p) & \text{if } y=1\\ -\lg(1-p) & \text{otherwise}\end{cases}$$

为方便表示,可化简为:

$$CE(p_t)=-\lg(p_t)\quad p_t=\begin{cases}p & \text{if } y=1\\ 1-p & \text{otherwise}\end{cases}$$

针对类别不均匀问题,传统的做法是 α-balanced CE,即在 CE loss 前增加权重系数 α,以此来平衡各类别的分布情况。其中,数据量少的类别 α 越大,而数

据量多的类别 α 越小。α-balanced CE 的表示如下：

$$CE(p_t) = -\alpha_t \lg(p_t)$$

但是，α-balanced CE 只平衡了不同类别对于模型的影响，它无法区分容易样本和困难样本，可能导致容易样本主导梯度而困难样本影响轻微的问题。因此，在模型中引入 Focal loss，以在平衡各类别分布的同时，加强困难样本对 loss 的影响，削弱容易样本的重要性。Focal loss 函数表示：

$$FL(p_t) = -\alpha_t (1-p_t)^{\gamma} \lg(p_t)$$

其中 $(1-p_t)^{\gamma}$ 是调节因子，γ 控制了样本权重的下降程度。

语义槽填充子任务本质上是序列标注问题，本文将使用 RoBERTa-WWM-BiLSTM-CRF 模型来实现该任务。该模型主要分为 3 层：首先，在 RoBERTa-WWM 层，将初始输入的文本语句转换为向量；其次，在 BiLSTM 层中，RoBERTa-WWM 层的向量输出将作为该层的输入，提取上下文信息；最后，在 CRF 层，通过施加约束和标签间的依赖关系保证获取有效的预测结果，获得全局最优序列。

在实际的文本语义中，意图识别、行业识别和语义槽填充 3 个子任务并非是独立的，三者之间存在较强的关联性，其中一个子任务可能对另外两个子任务的预测过程产生一定的影响。因此，提出 Joint-RoBERTa-WWM 联合模型，采用多任务学习（multitask learning）的方式将 3 个子任务进行联合学习，通过最小化 3 个子任务的损失来建立统一的联合损失函数。若用 $L^I(\theta)$、$L^D(\theta)$、$L^S(\theta)$ 分别表示意图识别、行业识别和语义槽填充 3 个子任务的损失函数，则联合损失函数为

$$L(\theta) = L^I(\theta) + L^D(\theta) + L^S(\theta)$$

Joint-RoBERTa-WWM 联合模型的结构如图 3 所示。

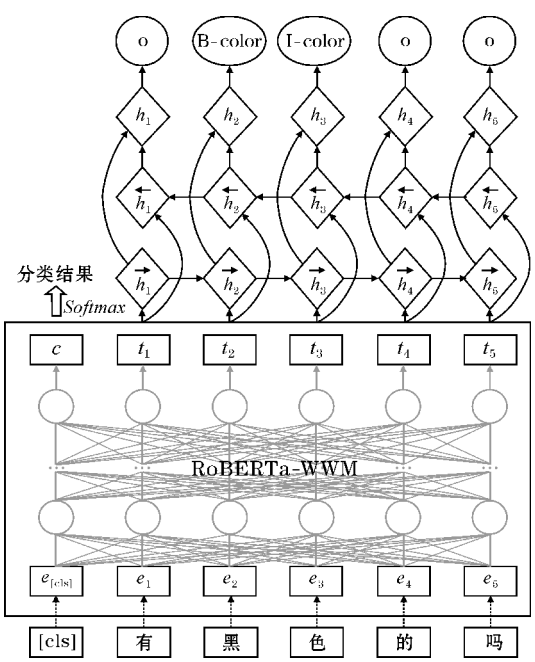

图 3 Joint-RoBERTa-WWM 联合模型结构图

2.2 Joint-RoBERTa-WWM-of-Theseus 压缩联合模型

目前，基于 Transformer 的预训练模型已经占据了自然语言处理领域举足轻重的地位。但有研究表明[24]，这其实是得益于它们“过度参数化”的特点，它们包括了数百万甚至十亿个参数，导致计算成本高并且效率低下，严重阻碍了模型在生产环境中的应用。Xu C 等[11]提出 BERT-of-Theseus，这是一种基于模块替换的模型压缩方法。相比于传统的知识蒸馏[25]，该方法在对初始模型进行压缩后可以保证压缩模型的结构与初始模型仍然相似，使整个压缩过程更加简捷。BERT-of-Theseus 压缩方法可以将原始的 12 层 BERT 教师模型 $P = \{\mathrm{prd}_1, \mathrm{prd}_2, \cdots, \mathrm{prd}_{12}\}$，压缩成一个 6 层的学生模型 $S = \{\mathrm{scc}_1, \mathrm{scc}_2, \cdots, \mathrm{scc}_6\}$，具体可以分为两个阶段：

(i)第一阶段是模块替换训练。将每个教师模块 prd_i 替换为相应的学生模块 scc_i。若第 i 个模块的输出向量表示为 y_i，则教师模型第 $i+1$ 个模块的前向计算输出：

$$y_{i+1} = \mathrm{prd}_i(y_i)$$

对于第 $i+1$ 个模块，通过伯努利分布采样一个随机变量 r_{i+1}，采样概率为 p，如下：

$$r_{i+1} \sim \mathrm{Bernoulli}(p)$$

则第 $i+1$ 个模块在学生模型中的最终输出为

$$y_{i+1} = r_{i+1} * \mathrm{scc}_i(y_i) + (1-r_{i+1}) * \mathrm{prd}_i(y_i)$$

其中 $*$ 表示按元素计算的乘法。

第一阶段的替换流程如图 4 所示。

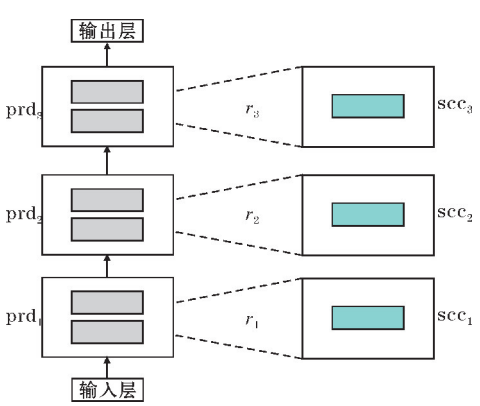

图 4 第一阶段替换流程图

(ii)第二阶段是学生模型 S 自身的微调，让所有的学生模块都参与到训练中，最后组合成学生模型 S：

$$S = \{\mathrm{scc}_1, \mathrm{scc}_2, \cdots, \mathrm{scc}_6\}$$

$$y_{i+1} = \mathrm{scc}_i(y_i)$$

第二阶段的训练流程如图 5 所示。

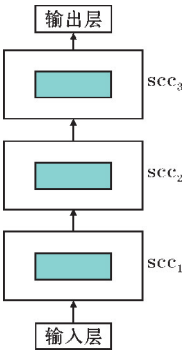

图 5　第二阶段训练流程图

经过以上两个阶段，每个prd_i模块都压缩成更小的scc_i模块，这样教师模型 P 被压缩成一个更小的学生模型 S。借助 Theseus 压缩的思想，Joint-RoBERTa-WWM 模型经过相同的方法进行压缩后，便构成了本文所提出的 Joint-RoBERTa-WWM-of-Theseus 模型，大幅提高了模型的预测速度，使模型能更好地服务于生产环境。

3　实验

3.1　实验环境

实验环境如下：Windows10 操作系统，Ryzen 5 5600X@3.70 GHz CPU，NVIDIA GeForce RTX 3070 显卡，16 GB内存。另外，实验中采用 Python 编程语言和 Tensorflow 深度学习框架实现模型的搭建。

3.2　实验数据集

以某大赛提供的真实对话数据为基础，添加了通过 Scrapy 框架爬取的某电商平台 4 个品类共 9865 条商品数据，抽取、标注了 3075 条文本语料作为实验数据集。在数据集中，共有 4 个行业分类、14 个意图分类和 22 个槽位。数据分布不平衡的问题在多分类任务，尤其是在意图识别子任务中，表现得尤其明显，图 6 展示了不同意图在数据集中的分布对比。由图可以看出，不同的类别之间数据量相差较大。因此，在模型的调优过程中解决数据分布不平衡的问题是十分必要的。

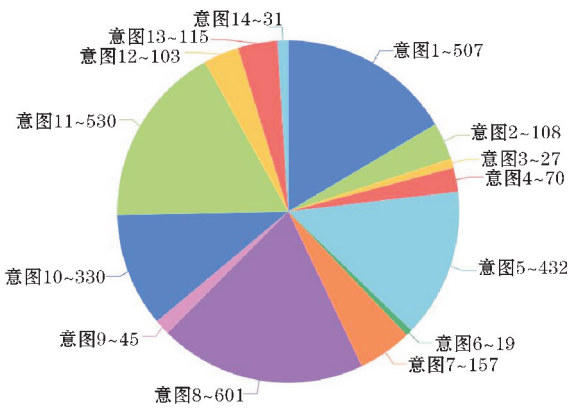

图 6　各类意图分布情况

3.3　实验参数与评价指标

实验中采用 Adam 优化器；设置的最大文本长度是 128；学习率为 0.00002；RoBERTa-WWM 层数为 12 层，隐藏层大小为 768；训练时加入早停机制，并设置 Dropout 比例为 0.5 来避免过拟合；为使损失函数下降更稳定，设置 Warmpu 比例为0.1；在进行 Theseus 压缩时，模块替换概率设置为0.5。实验中主要的超参数如表 2 所示。

表 2　超参数设置

参数名称	参数值
优化算法	Adam
最大文本长度	128
初始化学习率	0.00002
RoBERTa-WWM 隐藏层大小	768
LSTM 隐藏层大小	128
Dropout 比例	0.5
Warmup 比例	0.1
模块替换概率	0.5

在实验中，采取 F_1 值作为模型的评价指标，同时兼顾准确率和召回率。若用 F_1^I、F_1^D、F_1^S 分别表示意图识别、行业识别和语义槽填充 3 个子任务的 F_1 值，则模型整体的 F_1：

$$F_1=\frac{F_1^I+F_1^D+F_1^S}{3}$$

3.4　实验结果与分析

实验将 Joint-RoBERTa-WWM 模型和 Joint-BERT 模型进行比较，并对比了 Joint-RoBERTa-WWM 模型在使用 CE loss、α-balanced CE、Focal loss 等不同的损失函数时的表现情况，证明了 Joint-RoBERTa-WWM 模型使用 Focal loss 解决数据不平衡问题的优势。这些模型的对比实验结果如表 3 所示。

表 3　模型对比实验结果

模型	F_1
Joint-BERT	96.38
Joint-RoBERTa-WWM+CE loss	96.60
Joint-RoBERTa-WWM+α-balanced CE	96.80
Joint-RoBERTa-WWM+Focal loss(本文模型)	97.06

另外，实验将 Joint-RoBERTa-WWM-of-Theseus 模型、Joint-RoBERTa-WWM 模型进行对比，证明经过

Theseus 方法压缩,可以使模型在略微损失精度的前提下,大幅提高预测速度,帮助其为生产环境提供性能良好的实时预测服务。为了模拟真实生产环境下的模型运行情况,实验使用 Flask 框架分别将两个模型接口化部署后,测试 900 条不同文本的实时预测接口请求平均时长并将其作为评估标准,对比两个模型的预测速度,对比结果如表 4 所示。经过实验验证发现,通过 Theseus 方法压缩后的联合模型预测速度可以提高至压缩前的2.33倍,为模型在实际生产环境中的顺利使用奠定了基础。

表 4 模型压缩前后预测速度对比

模型	F_1	预测速度
Joint-RoBERTa-WWM	97.06	1.00X
Joint-RoBERTa-WWM-of-Theseus	96.79	2.33X

4 结束语

在任务型对话机器人的应用场景下,提出了 Joint-RoBERTa-WWM-of-Theseus 压缩联合模型。该模型充分考虑不同子任务之间的相互影响,将意图识别、行业识别和语义槽填充 3 个子任务进行联合学习训练;其次,在多分类子任务中引入了 Focal loss 机制,通过损失函数的优化来解决数据分布不平衡问题;另外,采用 Theseus 方法将模型进行压缩,使模型以很小的精度损失为代价换取了更快的预测速度,大幅提高了其在生产环境下的服务能力。实验表明,Joint-RoBERTa-WWM-of-Theseus 压缩联合模型为任务型对话机器人的搭建提供了良好的算法基础。

参考文献:

[1] 于丹,闫晓宇,王艳秋,等. 任务型对话机器人的设计及其应用[J]. 软件工程,2021,24(2):55-59.

[2] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv, 2018,1810:4805.

[3] 李法来,金震,熊婷,等. 基于中文 Bert 模型智能机器人的实现方法和系统[P]. 中国:CN113553405A,2021-10-26.

[4] Karl Weiss, Taghi M Khoshgoftaar, DingDing Wang. A survey of transfer learning[J]. Journal of Big Data,2016,3(1):1-40.

[5] Yinhan Liu, Myle Ott, Naman Goyal, et al. Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. CoRR,2019.

[6] Xuezhe Ma, Eduard H. Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. CoRR,2016.

[7] 柏兵,侯霞,石松. 基于 CRF 和 BI-LSTM 的命名实体识别方法[J]. 北京信息科技大学学报(自然科学版),2018,33(6):27-33.

[8] 赵京胜,宋梦雪,高祥. 自然语言处理发展及应用综述[J]. 信息技术与信息化,2019(7):142-145.

[9] Qian Chen, Zhu Zhuo, Wen Wang. BERT for Joint Intent Classification and Slot Filling[J]. CoRR,2019.

[10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[J]. CoRR,2019.

[11] Xu C, Zhou W, Ge T, et al. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing[J]. arXiv preprint arXiv,2020,2002:2925.

[12] Turing A M. Computing machinery and intelligence[J]. Mind,1950,59(236):433-460.

[13] 俞凯,陈露,陈博,等. 任务型人机对话系统中的认知技术——概念、进展及其未来[J]. 计算机学报,2015,38(12):2333-2348.

[14] 陈龙,孙泽健. 面向任务的对话系统现状研究[J]. 电子技术与软件工程,2017(23):172-173.

[15] 天猫精灵鲍娟:天猫精灵用 AI 连接家庭全场景智慧营销[J]. 国际品牌观察,2021(20):47-48.

[16] Aron J. How innovative is Apple's new voice assistant, Siri? [J]. New Scientist,2011,212(2836):24.

[17] Hoy Matthew B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants[J]. Medical reference services quarterly,2018,37(1):81-88.

[18] Yiming Cui, Wanxiang Che, Ting Liu, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. CoRR,2019.

[19] Schmidhuber J. Deep Learning in Neural Networks: An Overview[J]. Neural Networks,2015, 61:85-117.

[20] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[21] Toshiaki Fukada, Mike Schuster, Yoshinori

Sagisaka. Phoneme boundary estimation using bidirectional recurrent neural networks and its applications[J]. Systems and Computers in Japan, 1999,30(4):20-30.

[22] Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. International Conference on Machine Learning. San Francisco, USA, 2001:282-289.

[23] Lin Tsung-Yi, Goyal Priya, Girshick Ross, et al. Focal Loss for Dense Object Detection[C]. Proceedings of the IEEE international conference on computer vision. 2017:2980-2988.

[24] Geoffrey E Hinton, Oriol Vinyals, Jeffrey Dean. Distilling the Knowledge in a Neural Network [J]. CoRR, 2015.

[25] Nakkiran P, Kaplun G, Bansal Y, et al. 2020 Deep double descent: where bigger models and more data hurt Int. Conf. Learning Representations[J]. Journal of Statistical Mechanics: Theory and Experiment, 2021(12):124003.

Research on Multi-task Jointing Model for Task Chat Robot

GAO Zuoyuan, TAO Hongcai

(School of Computing & Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: In the process of building a task-oriented chatbot, it is generally necessary to execute several subtasks of Natural Language Processing. And the traditional training method is to integrate each subtask after training independently, which will ignore the relevance between different subtasks and limit the predictive power of the model. This paper proposes a compressed jointed model, i. e., Joint-RoBERTa-WWM-of-Theseus. On the one hand, intention classification, domain classification and semantic slot filling are jointly trained through multi-task joint learning and training. And the focal loss mechanism is introduced to the multi-class classification subtask to solve the problem of data distribution imbalance. On the other hand, the model is compressed by means of Theseus compression method, which greatly improves the prediction speed of the model and improves the applicability and the real-time in the production environment with a slight loss of accuracy.

Keywords: RoBERTa-WWM model; multi-task joint learning; Theseus compression; Focal loss