融合BERT和卷积门控的生成式文本摘要方法

邓维斌[1†], 李云波[1], 张一明[2], 王国胤[1], 朱 坤[1]

(1. 重庆邮电大学 计算智能重庆市重点实验室，重庆 400065；2. 中国人民解放军78111部队，成都 610031)

摘 要: 结合注意力机制的循环神经网络(RNN)模型是目前主流的生成式文本摘要方法,采用基于深度学习的序列到序列框架,但存在并行能力不足或效率低的缺陷,并且在生成摘要的过程中存在准确率低和重复率高的问题. 为解决上述问题,提出一种融合BERT预训练模型和卷积门控单元的生成式摘要方法. 该方法基于改进Transformer模型,在编码器阶段充分利用BERT预先训练的大规模语料,代替RNN提取文本的上下文表征,结合卷积门控单元对编码器输出进行信息筛选,筛选出源文本的关键内容;在解码器阶段,设计3种不同的Transformer,旨在探讨BERT预训练模型和卷积门控单元更为有效的融合方式,以此提升文本摘要生成性能. 实验采用ROUGE值作为评价指标,在LCSTS中文数据集和CNN/Daily Mail英文数据集上与目前主流的生成式摘要方法进行对比的实验,结果表明所提出方法能够提高摘要的准确性和可读性.

关键词: 生成式文本摘要；序列到序列；预训练模型；卷积门控单元；信息筛选；Transformer模型

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0494

引用格式: 邓维斌,李云波,张一明,等. 融合BERT和卷积门控的生成式文本摘要方法[J]. 控制与决策, 2023, 38(1): 152-160.

An abstractive text summarization method combining BERT and convolutional gating unit

DENG Wei-bin[1†], *LI Yun-bo*[1], *ZHANG Yi-ming*[2], *WANG Guo-yin*[1], *ZHU Kun*[1]

(1. Chongqing Key Laboratory of Computational Intelligence，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；2. 78111 Troops of People's Liberation Army of China，Chengdu 610031，China)

Abstract: The recurrent neural network (RNN) model combined with the attention mechanism is the current mainstream abstractive text summarization method, which uses a sequence-to-sequence framework based on deep learning. However, the abstractive summarization model based on the RNN has insufficient parallel ability or performance defects of long-term dependence, and the problem of low accuracy and high repetition rate in the process of generating summary. In order to overcome these problems, an abstractive summarization model method combining the BERT pre-training model and the convolutional gating unit is proposed based on the improved Transformer model. In the encoder stage, it makes full use of the large-scale corpus pre-trained by the BERT to replace the RNN to extract the contextual representation of the text, and then combines the convolutional gating unit to filter the output of the encoder to filter out the source text. In the decoder stage, three different Transformers are designed, for exploring a more effective fusion method of the BERT pre-training model and convolutional gating unit to improve the performance of text summarization. The ROUGE value is used as the evaluation index in the experiments. The experimental results on the LCSTS Chinese dataset and CNN/Daily Mail dataset show that the proposed method improves the accuracy and readability of the abstract.

Keywords: the abstractive summarization；sequence-to-sequence；pre-training model；convolution gating unit；filter；Transformer model

0 引 言

在大数据时代,文本数据呈现爆发式增长,大量且冗杂的文本信息,如新闻、论文和微博等,每天都会推送到人们的视野里,如何快速有效地从中获取关键摘要信息显得异常重要. 将输入的一段长文本压缩精简,提炼出一段短文本的自动文本摘要技术由此诞

收稿日期: 2021-03-25；**录用日期:** 2021-09-22.

基金项目: 国家研发计划项目(2018YFC0832100, 2018YFC0832102)；国家自然科学基金重点项目(61936001)；国家自然科学基金项目(61876027)；重庆市自然科学基金创新群体科学基金项目(cstc2019jcyj-cxttX0002).

责任编委: 胡清华.

†通讯作者. E-mail: dengwb@cqupt.edu.cn.

生. 由自动文摘生成的短文本能够简洁、准确地突出其关键信息,在过滤冗余信息的同时提高了人们阅读的效率.

目前,文本摘要任务依据实现方式可以分为两类:抽取式摘要(extractive summarization)和生成式摘要(abstractive summarization). 抽取式摘要采用特定的评分规则和排序方法,从原文本中抽取若干重要句子组成摘要,一般多用于长文本中,其全部来源于原文,一定程度上确保句法的准确性和信息的完整性. 生成式摘要则通过理解原始文本的上下文语义信息,自动生成语义连贯的简短文本. 与抽取式摘要相比,生成式摘要更加符合人类对于语言认知的习惯. 近几年快速发展的深度神经网络在自然语言处理领域展示出其强大的表征能力,与文本生成任务的结合也是自然语言处理研究的热点.

深度神经网络模型已成功地应用在自然语言处理任务中,并取得优异的成果. 特别地,在文本生成任务中常使用循环神经网络作为编码器和解码器,利用其逐词处理序列的优点,能够有效且准确地理解源文本表达的信息并转换为另一种形式,在生成领域取得了很好的成果. 但是,由于RNN及其变体必须等待上一个神经元输出作为当前神经元输入,难以实现并行化计算,致使在训练和生成阶段的效率较低. Vaswani等[1]提出一种新的序列到序列模型Transformer,因其基于注意力机制而具备较好的并行计算能力,通过在机器翻译任务上的实验表明在减少训练时间的同时能获得更好的翻译结果. 之后,随着Transformer改进方法BERT[2]、XLNet[3]、GPT-2[4]等预训练模型的出现,提升了许多自然语言处理任务的性能水平,如文本分类、命名实体识别和机器阅读理解等,但是预训练模型很少应用或者直接使用在生成式文本摘要.

当前主流的生成式文本摘要大多基于RNN及其变体,不仅存在上述RNN的缺点,还存在生成的摘要准确率低和重复率高的问题. 针对上述问题,同时考虑到Transformer优秀的并行能力和BERT强大的特征提取能力,本文通过改进Transformer的模型架构,提出融合BERT预训练模型的摘要生成模型. 该模型在编码器阶段充分利用预先训练的语言模型,提取出保留着上下文语义的字向量. BERT因为随机选取15%的字符mask,会忽视被mask字符之间可能存在的语义关联,导致可能出现丢失关键短语信息的问题,为了解决这一问题,本文在编码器与解码器之间加入卷积门控单元,使用3个CNN不同大小的卷积窗口进行不同粒度信息提取,负责对编码器的输出进行关键信息筛选. 在解码器阶段使用Transformer及变体对信息进行解码和生成摘要.

本文主要贡献有两个方面:

1) 提出一种融合BERT-base预训练模型和卷积门控单元的模型. 该模型具有利用CNN不同大小卷积核提取不同粒度信息的特点,对经过BERT输出的字向量进行文本关键信息筛选,增强文本关键特征提取的能力,同时减少编码器对数据进行编码上下文表示的过程.

2) 设计3种不同基于Transformer的解码器,提出BERT与卷积门控单元有效的融合方式,有效提升文本摘要生成的质量.

1　相关工作

近年来,深度学习模型表现出来的强大表征能力倍受研究者喜爱,Google Brain[5]团队提出的端到端(Sequence-to-sequence, Seq2Seq)模型结构,成功搭建用于语言处理的神经网络模型,并在各种文本生成任务的处理上表现优秀,开启了自然语言处理(natural language processing, NLP)中端到端网络的火热研究. 该模型包含一个编码器(encoder)和一个解码器(decoder),编码器通过对不同长度文本进行编码,输出固定长度的向量表达,解码器将该向量表达进行解码,重新输出可变长的目标序列. Bahdanau等[6]对Seq2Seq模型进行改进,借鉴人在阅读时注意力会集中在关键词部分这一思维模式,改进Soft-Attention机制,提高了机器翻译任务的效果. 此后,基于注意力机制的Seq2Seq模型在自然语言处理任务中被广泛认可和使用.

文献[7]首次将Seq2Seq模型用于生成式文本摘要,使用带注意力机制的卷积神经网络(convolutional neural network, CNN)为编码器,神经网络语言模型(neural network language models, NNLM)为解码器,并且完成Gigaword数据集和DUC2004数据集的第1次文本摘要任务实验,为后面任务做了基准模型实验. Hu等[8]构建了第1个也是目前最大的一个中文短文本摘要数据集,并在文本摘要模型上完成实验,为国内文本摘要发展又推进一大步. Chopra等[9]在文献[5]的基础上对解码器进行优化,使用循环神经网络代替神经网络语言模型,进一步提升摘要生成的质量. 由于词表大小的限制,摘要生成过程中不可避免地会出现未登录词,针对上述问题,Gu等[10]提出直接从输入序列复制未登陆词到输出序列中的拷贝机制(copy mechanism). See等[11]在指针生成器网络上设计覆盖机制,较好地解决了生成摘要时产生重复词

语的问题. Lin等[12]提出全局编码模型,使用由全局门控网络对编码器RNN的输出结果进行关键信息筛选,进一步提升摘要任务的水平. 吴仁守等[13]在传统编码器对输入文本进行编码后,增加全局自匹配层进行文本自匹配和全局门控单元去除冗余信息.

田珂珂等[14]在transformer基础上进行改进,通过将编码器参数共享到解码器中,使编码器成为解码器的一部分,同时使用门控网络对输入序列进行信息筛选,在摘要评分、训练速度和推理速度方面都得到提升. Peng等[15]使用LSTM和transformer组成双编码器获取更多语义,加上全局门控进一步筛选关键信息. 随着BERT等一系列预训练模型的出现,预训练模型开始在自然语言理解领域取得优秀的成果. Zhang等[16]提出一种基于BERT的自然语言生成模型,在编码过程中充分利用预先训练好的模型,并设计两段式解码器,使其在文本摘要上的效果更进一步. Zhang等[17]提出了一种新的自监督预训练目标GSG (gap sentences generation),以适配transformer-based的encoder-decoder模型在海量文本语料上预训练,结果表明,在多个文本摘要数据集上达到了与人工摘要相媲美的性能. Facebook在ACL2020上提出预训练Seq2Seq的去噪自编码器模型BART[18],BART模型中的文本填充方法让模型学习更多地考虑句子的整体长度,并对输入进行更大范围的转换,从而将BERT中MLM和NSP目标统一起来,加大了模型学习难度. BART在自然语言理解任务上与先进模型不相伯仲,但是在文本生成任务上超越其他模型. 同年,Baidu提出了一个用于语言生成的增强型多流Seq2Seq预训练和微调框架(ERNIE-GEN)[19],其中包括一个填充式生成机制和一个噪声感知生成方法,以减轻预训练和微调的偏差. 此外,ERNIE-GEN还集成了一个新的跨度生成任务来训练该模型生成类似人类书写的文本,进一步提高了下游任务的性能. 通过广泛地实验,ERNIE-GEN在包括摘要生成的NLG任务中达到较好效果.

2 融合BERT和卷积门控的文本摘要模型

因transformer有着比RNN更好的并行能力和语义特征提取能力,本文决定采用transformer模型作为生成式摘要模型的基础模型并加以改进. 该模型由3部分组成: BERT编码器、卷积门控单元和解码器. 其中BERT编码器读取输入的文档,并构建其表征;卷积门控单元对编码器的输出进行关键信息筛选,并提供给解码器生成摘要. 为了深入研究BERT编码器和卷积门控单元的有效融合方法,设计3种不同的融合模型:

1) CBC: 编码器使用预训练模型BERT,通过卷积门控单元的筛选进入基础transformer解码器中生成文本.

2) CBC-DA (double attention): 在CBC模型的基础上,将BERT的输出放入解码器第2层多头注意力层中,卷积门控单元筛选的关键信息输入解码器第3层多头注意力层,以加深融合关键信息和全局信息.

3) CBC-RA (residual attention): 在CBC模型的基础上,当前层级编码器解码器的注意力计算过程中加入上层注意力,即将残差网络放到attention矩阵上,具体如图1所示. 下面分别介绍编码器、卷积门控单元和解码器的细节.

2.1 问题形式化描述

文本摘要任务通过给定一段长文本序列$X = \{x_1, x_2, \dots, x_n\}$输入到模型中. 其中:$n$为输入文本长度,经过模型的训练测试最后生成输出想要的短文本摘要序列$Y = \{y_1, y_2, \dots, y_m\}$,$m$为输出文本长度,且要求输入文本长度$n$大于生成摘要长度$m$.

2.2 编 码 器

由于BERT在多种语言理解任务上的出色表现,如分类、实体识别和机器阅读理解,本文采用BERT预训练模型作为编码器. BERT由多个transformer encoder层堆叠起来,每一层都包含一个多头自注意力层(multi-head self-attention layer)和一个前向反馈层(feed forward layer). 在多头自注意力层中,每个头的注意力使用scaled dot-product attention计算注意力,其输入由维度为d的查询向量(query, Q)和键向量(key, K)以及维度为d的值向量(values, V)组成,所有键计算查询的点积,并应用softmax函数获得值的权重. 具体包含3个步骤: 1)每个Query-Key会进行一个点乘的运算过程,同时为防止值过大,除以维度的常数$\sqrt{d_k}$; 2)使用softmax函数将其归一化; 3)乘以Values求得注意力为

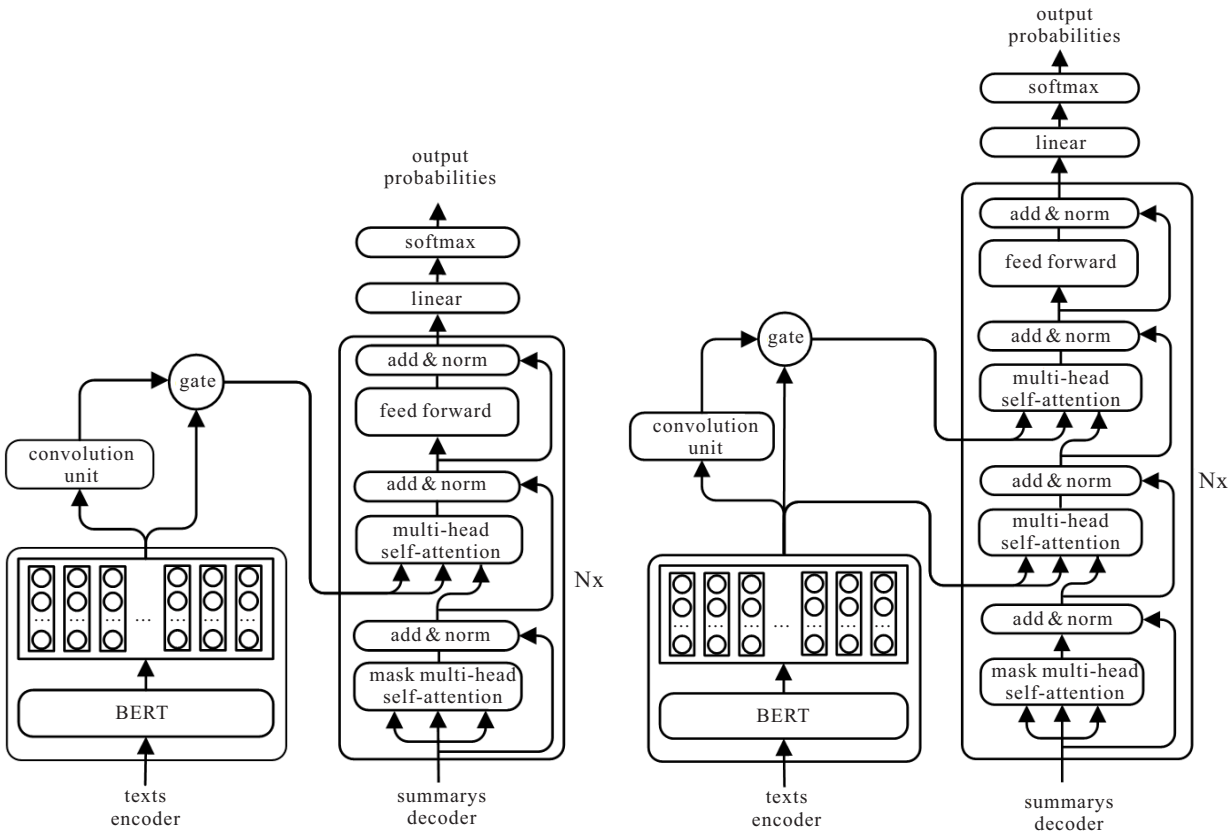
$$\mathrm{attention}(Q, K, V) = \mathrm{softmax}\Big(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\Big)V. \quad (1)$$

多头注意力机制(multi-head attention)对Q、K、V进行h(多头数)个不同线性变换的投影,然后计算h个不同的attention结果,最后通过拼接不同的attention得到multi-head attention,如下所示:

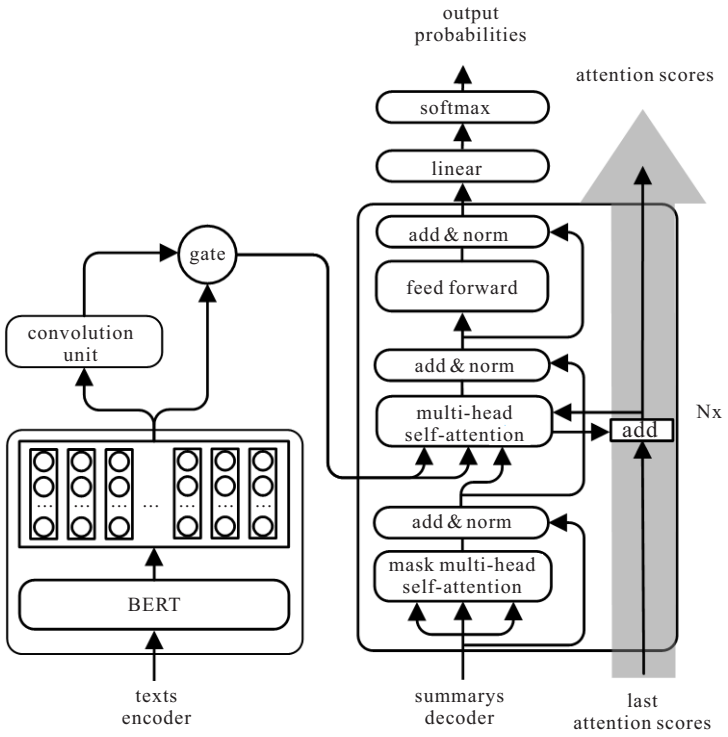
$$X_{\mathrm{enc}} = \mathrm{multihead}(Q, K, V) = \mathrm{concat}(\mathrm{head}_1, \mathrm{head}_2, \dots, \mathrm{head}_h), \quad (2)$$

$$\mathrm{head}_i = \mathrm{attention}(\mathrm{QW}_i^Q, \mathrm{KW}_i^K, \mathrm{VW}_i^V), \quad (3)$$

其中$\mathrm{QW}_i^Q, \mathrm{KW}_i^K, \mathrm{VW}_i^V$为矩阵参数.



(a) CBC　　(b) DBC-DA



(c) CBC-RA

图 1　模型框架

多头注意力层之上存在前向反馈层,其包含两个线性变换和ReLU激活函数,能够有效增加模型的非线性拟合能力,计算为

$$\mathrm{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2. \tag{4}$$

其中: $W_1\in R^{(d\times d_{ff})}$, $W_2\in R^{(d_{ff}\times d)}$ 为线性转换, d_{ff} 为该层的隐藏层大小, b_1 和 b_2 为偏置.

2.3 卷积门控单元

对于生成式模型而言,如何有效地进行关键信息提取是一个核心问题,尤其是对处理长文本、多个句子时问题变得更加突出. 同时输入的序列中包含很多信息,但只有少部分字或词包含整个序列的关键信息. 为实现输入序列的关键信息筛选,使用基于文本上下文全局信息编码的卷积门控单元控制BERT编码器的输出.

卷积门控单元使用CNN对BERT编码器的输出进行卷积,拥有参数共享的卷积核使模型能够提取序列中的局部特征,尤其是n-gram特征. 同时为了进一步加强全局信息,在CNN的输出上加上自注意力机制,如图2所示.

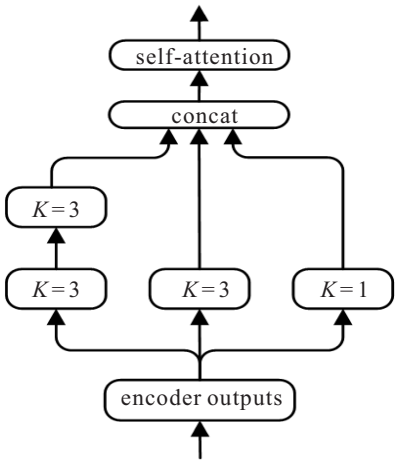


图 2 卷积门控单元

卷积门控单元主要由多个不同卷积核的卷积神经网络加上self-attention机制构成. 在CNN部分通过在一维卷积单元中使用不同尺寸的卷积核提取句子中不同n-gram的信息. 参考文献[17]的工作,根据Inception的设计原则,分别使用核心为1、3、5(用两个3核表示$k=5$的结构避免因结构复杂引起的计算量增大)的kernel,卷积块的细节计算如下:

$$g_i = \mathrm{ReLU}(W[h_{i-k/2},\dots,h_{i+k/2}]+b). \tag{5}$$

其中:ReLU为非线性激活函数,$h_{i-k/2},\dots,h_{i+k/2}$为卷积核窗口的滑动位置,$i$为窗口的中间位置.

在CNN之上,通过加入self-attention机制使得模型能进一步关注全局语义信息,self-attention机制使模型在每个时间步都挖掘当前词语与文章中其他词语的关系,同时保证不会产生太多计算复杂度,计算过程如式(1)所示. 其中:Q和V为CNN模块输出的表示矩阵;$K=\mathrm{Watt}\,V$,Watt为一个可学习的矩阵.

将卷积神经网络与自注意力机制组合在一起,即为卷积门控单元g,不仅可以提取整个原文本的n-gram信息,还能挖掘全局相关性. 卷积门控单元门筛选如下:

$$\hat{X}_{\mathrm{enc}} = X_{\mathrm{enc}} \times \mathrm{sigmoid}(g), \tag{6}$$

其中X_{enc}为BERT编码器的输出. 全局门控单元g通过sigmoid函数计算,转化为一个介于0~1之间的向量,在某一维度上的值如果接近0,则表示删除大部分信息,如果接近1,则表示保留大部分信息.

2.4 解 码 器

为探索将BERT编码器和关键短语信息纳入模型的有效方法,使用transformer作为基础解码器的同时,设计两种不同transformer解码器变体.

base-transformer: transformer解码器分析卷积门控单元输出的关键信息生成摘要序列,如图1(a)所示. 首先读取已经生成的摘要序列$y=\{y_0,y_1,\dots,y_{i-1}\}$,$i$表示当前需要生成字号并对序列进行编码,得到新的向量序列$s=\{s_0,s_1,\dots,s_{i-1}\}$;然后解析向量$s$和卷积门控单元的输出,生成当前字$y_i$. 以此类推,最终得到摘要序列.

double-attention: 尽管将关键信息过滤后的输出可以代替原来的编码器输出传入到解码器中,提高模型对关键信息的敏感性,但原始文章中一些有用的全局信息也可能被过滤掉,从而降低生成摘要的质量. 为进一步平衡模型的关键信息敏感性和摘要质量,使用BERT编码器输出和卷积门控单元的输出作为解码器的两个输入并将其融合. 如图1(b)所示,添加一个额外的多头注意力子层,在解码器中依次使用两个块对$\hat{X}_{\mathrm{enc}}$和X_{enc}计算多头注意力,所有子层通过残差进行连接.

residual-attention: 设置解码器层数为12层,与BERT层数相同,因为解码器不同层对编码器的输出不一定计算出相同的注意力权重,为加大模型对关键信息的注意力权重,参考He等[20]的工作,将解码器残差网络放到attention矩阵上,如图1(c)所示. 具体为

$$\mathrm{attention}(Q_n,K_n,V_n) = \mathrm{softmax}(A_n)V_n, \tag{7}$$

$$A_n = \frac{Q_n K_n^{\mathrm{T}}}{\sqrt{d_k}} + A_{n-1}, \tag{8}$$

其中n为第n层transformer decoder. 将残差网络放到attention矩阵上,使得第1层的attention能够直通最后1层,不仅可以融合不同transformer层的注意力,还能有效防止梯度消失的风险.

2.5 训练与推理

对于多分类任务,需要将训练数据中的标签转换为独热编码(one-hot encoding),若样本属于该类标签则概率为1,不属于则为0. 在模型训练阶段常采用交叉熵损失函数鼓励模型预测概率去拟合真实概率,而

拟合独热向量的真实概率不仅无法保证模型的泛化能力造成过拟合,而且全概率和零概率会增加所属类别与其他类别之间的差距,造成模型“过分”相信预测的类别. 标签平滑(label smoothing)策略可以解决上述问题,label smoothing是一种损失函数修正算法,也称为标签平滑归一化,能有效提高分类的准确性.

本文在训练过程中采用标签平滑策略,假定测试误差率为$\varepsilon = 0.1$,one_hot为样本标签转化的独热向量,有

$$\text{one\_hot}_i^* = (1-\varepsilon) \times \text{one\_hot}_i + \frac{(1-\text{one\_hot}_i) \times \varepsilon}{\text{vocab\_size}}. \quad (9)$$

其中:one_hot* 为标签平滑操作后的样本标签,vocab_size为词表大小.

解码器输出向量经过全连接层计算得到向量logit,再经过softmax计算得到模型最终输出:维度为词表大小的向量word_prob,即目标词的预测结果,有

$$\text{word\_prob}_i = \text{softmax}(\text{logit}). \quad (10)$$

最后经过标签平滑后该样本的交叉熵损失为

$$\text{loss} = -\sum_{i-1}^{N} \text{one\_hot}_i^* \times \log(\text{word\_prob}_i). \quad (11)$$

在解码过程中,采用集束搜索(beam search)的方式产生摘要词汇,以较小的代价在相对受限的搜索空间中找出其最优解.

3　实验与结果分析

3.1　实验数据集

实验采用中文数据集LCSTS[8]和英文数据集CNN/Daily Mail.

1) LCSTS数据集. 数据集包含3部分:第1部分包含2 400 591条摘要数据对(短文本、摘要)的主要内容;第2部分是第1部分中随机采样的10 666条摘要数据对;第3部分独立于第1、第2部分,共1 106条摘要数据对. 其中第2、第3部分数据被专家人工评分,评分为1～5,用于衡量短文本与摘要的相关程度. 评分大于等于3分表示短文本与摘要的相关性较强;低于3分表示短文本与摘要相关性较差. 为了更有效地评估摘要模型,选用第1部分数据作为模型训练数据集,第3部分得分为3～5分的数据作为测试集.

2) CNN/Daily Mail数据集. 大型英文文本摘要数据集,其中包含28万训练数据对,1.3万验证数据对和1.1万测试数据对. 每条文档为较长训练文本,平均包含766个词、29.74个句子;对应摘要平均包含53个词、3.72个句子.

3.2　实验评价标准

采用Lin[21]提出的ROUGE指标进行评估. ROUGE指标主要基于n元词召回率,通过对比模型生成摘要和多个专家提前书写的标准参考摘要之间相同的基本n-gram、词序列或词对数目,达到评价摘要质量的目的. 评价指标分数越高表明模型生成摘要质量越高. ROUGE评价指标细分为ROUGE-N和ROUGE-L等,本文使用ROUGE-1和ROUGE-2分别测试生成的摘要与标准之间1-gram和2-gram的重叠程度,使用ROUGE-L测试最长公共子序列的重叠程度.

3.3　实验参数

实验融合BERT预训练模型,使用BERT字表,大小为21 128字,字向量维度为768,并将解码器层数和多头注意力数分别设置为与编码器相同的12层和12头. 在训练过程中,设置BERT输入输出只有一个句子,以[CLS]作为开始符、[SEP]作为结束符. 编码器与解码器均采用AdamW优化函数,其中学习率r设置为5e-5,$\beta_1 = 0.9$, $\beta_2 = 0.999$,并设置学习计划采用预热(Warmup)与衰减(Decay)策略. 为进行正则化,将dropout和标签平滑均设置为0.1,针对LCSTS数据集,批大小设置为16,输入字符最大长度为130,输出文章字符最大长度为30;针对CNN/Daily Mail,批处理大小为4,输入文章字符最大长度500,输出字符最大长度100. 实验使用单张RTX-2080Ti(GPU)进行训练,在测试阶段使用束大小为3的束搜索.

3.4　对比结果与分析

本文模型将对比以下使用LCSTS中文数据集的基准模型,并从相关文献中直接抽取实验结果:

1) RNN和RNN-context[8]: 基于RNN的文本摘要模型首次使用LCSTS短文本作为数据集进行训练测试,其中RNN-context模型在RNN的基础上加入注意力机制.

2) CopyNet[10]: 使用RNN作为编码器,解码器采用生成模式和拷贝模式.

3) S2S + superAE[22]: 在训练模型阶段利用autoencoder监督Seq2Seq模型的学习,以提高encoder性能.

4) CGU[12]: 基于全局编码的Seq2Seq模型,通过设置全局门控单元对编码器的输出信息进行筛选.

5) transformer: 基于注意力机制的Seq2Seq模型,且具有较好的并行能力. 本文复现基于transformer的文本摘要生成方法,并在数据集进行实验.

6) GDE[15]: 基于全局门控双编码器的生成式文本摘要方法,通过使用双编码器获取更多的全局语义信息由全局门控单元筛选,进一步过滤出关键信息,进而使模型能够生成更准确的文本摘要.

7) BERTabs: 将文本通过预训练模型BERT编码表征,输入到transformer解码器中生成摘要. BERT因其强大的表征能力可以很好地输出上下文语义信息,与transformer模型同样是本文的基础模型.

表1展示了各生成式摘要模型在LCSTS数据集上的实验结果.

表1 自动文摘在LCSTS上的实验结果

model	ROUGE-1	ROUGE-2	ROUGE-L
RNN	21.5	8.9	18.6
RNN-context	29.9	17.4	27.2
CopyNet	34.4	21.6	31.3
S2S+superAE	39.2	26.0	36.2
CGU	39.4	26.9	36.5
transformer	38.2	24.8	36.3
GDE	40.2	25.0	36.5
BERTabs	41.1	27.1	39.1
CBC	41.6	27.6	39.3
CBC-DA	41.8	28.2	39.6
CBC-RA	41.9	28.1	39.3

由表1可见,对于CGU模型,其编码器、解码器均使用双向LSTM,同时在编码器与解码器中间加入卷积门控单元对信息进行筛选. 在中文数据集上与CGU对比,CBC在ROUGE-1、ROUGE-2和ROUGE-L上分别提高2.2、0.7、2.8个百分点,CBC-DA提高2.4、1.3、3.1个百分点,CBC-RA提高2.5、1.2、2.8个百分点,表明CBC三个模型使用预训练模型BERT虽未直接训练LCSTS数据集,但在大规模预训练的基础上,通过卷积门控单元能够很好地表示上下文特征和关键信息筛选.

为了验证本文模型在英文数据集上的效果,选择在CNN/Daily Mail英文数据集上与相关学者的实验结果进行对比,具体见表2. 对比模型包括基于RNN的CGU[17]模型、transformer模型、基于BERT的One-Stage[22]和BERT-ext+abs[23]模型. 几种对比模型的实验结果直接来自于所引用文献.

表2 自动文摘在CNN/Daily Mail上的实验结果

model	ROUGE-1	ROUGE-2	ROUGE-L
CGU	39.53	17.28	36.38
transformer	39.5	16.06	36.63
One-Stage	39.5	17.87	36.65
BERT-ext + abs	40.14	17.87	37.83
BERTabs	39.2	18.02	39.23
CBC	40.06	18.24	39.33
CBC-DA	40.57	18.49	39.77
CBC-RA	40.33	18.34	39.76

1) One-Stage: 一种基于BERT-base和复制机制的生成式摘要.

2) BERT-ext+abs: 一种基于BERT的新提取器架构和新的训练信号,旨在全面优化摘要的ROUGE值.

由表2中的ROUGE值可见,所提出的CBC模型对比CGU和transformer模型有所提升,如对于transformer模型,CBC在ROUGE值上分别提升0.46、2.18和2.7,与相同条件下的基础BERT优化模型BERT-ext+abs有所不如. 但是,本文在CBC基础上优化解码器的两个模型CBC-DA和CBC-RA,对比BERT-ext+abs有不同程度的提升,CBC-DA较结果分别提高0.44、0.62和1.94,CBC-RA分别提高0.19、0.47和1.93.

3.5 消融分析

为分析每个组件的重要性,在中文数据集LCSTS上对模型进行消融研究. 本文使用三级消融模型进行实验: 一级为基础模型BERTabs,直接使用BERT作为编码器,不对信息进行筛选和加强全局信息;二级为在基础模型BERTabs上加入卷积门控机制的优化编码器模型CBC;三级为在CBC模型上进行双层multi-head self-attention的融合模型CBC-DA或者加入上层残差网络的模型CBC-RA两种优化解码器模型. 消融实验结果如表1最后4行所示.

CBC模型与基础BERTabs模型相比,在ROUGE三个指标上分别提升0.5、0.5、0.2个百分点,表明在解码器不变的情况下,加入卷积门控单元能够有效地指导模型获取更多的关键信息. 同时,对比其他基于RNN的模型可知,融合BERT预训练模型和卷积门控单元的生成式摘要方法在语义分析上弥补了RNN编码器提取特征不足的缺点,更能有效提取全局关键信息.

由最后两行实验结果可以看出,模型CBC-DA比CBC取得了更高的准确率,表明将BERT输出放入解码器中进行解码,再与关键信息一同解码能够进一步抓取全局信息;模型CBC-RA通过加入上层残差网络的方式加大对关键信息的注意力权重,在ROUGE三个指标上比CBC效果更好.

综上,通过消融分析表明,每个模块都是模型所必需的,且在评价指标上均有提升.

3.6 案例分析

为了进一步直观地评估CBC模型的摘要生成能力,下面对比基于transformer模型和本文基准模型BERTabs在实际案例中生成摘要的性能,从LCSTS数据集中抽取3个案例文本进行分析展示,如表3所示.

表 3　LCSTS数据集部分生成式摘要示例

		文本
例1	源文本	雅虎发布2014年第4季度财报,并推出了免税方式剥离其持有的阿里巴巴集团15%股权的计划,打算将这一价值约400亿美元的宝贵投资分配给股东. 截止发稿前,雅虎股价上涨了大约7%,至51.45美元.
	参考摘要	雅虎宣布剥离阿里巴巴股份
	transformer	雅虎股价400亿美元投资阿里巴巴股价400亿美元分配股权
	BERTabs	雅虎剥离阿里巴巴15%股权股价上涨7%
	CBC	雅虎推出免税方式剥离阿里15%股权计划
	CBC-DA	雅虎推出免税方式剥离阿里巴巴15%股权计划
	CBC-RA	雅虎剥离阿里巴巴15%股权价大涨7%
例2	源文本	李克强总理18日在美药典公司餐厅与1家进驻自贸区的中外企业家座谈,请他们给自贸区各项改革“打分”. 他对10位参会企业家说:“希望我们在留有饭菜余香中进行的座谈会,不仅friendly(友好),而且frankly(坦率),有什么问题直来直去讲出来.”
	参考摘要	李克强邀10企业给上海自贸区打分
	transformer	李克强邀企业家给自贸区打分
	BERTabs	李克强请进驻自贸区中外企业打分
	CBC	李克强邀10企业给上海自贸区打分
	CBC-DA	李克强邀10家企业家给自贸区打分
	CBC-RA	李克强邀10企业给自贸区打分
例3	源文本	2014年,51信用卡管家跟宜信等P2P公司合作,推出线上信贷产品“瞬时贷”,其是一种纯在线操作的信贷模式. 51信用卡管家创始人孙海涛说,51目前每天放贷1 000万,预计2015年,自营产品加上瞬时贷,放贷额度将远超30亿.
	参考摘要	51信用卡管家,预计2015年放贷额度远超30亿
	tansformer	51信用卡管家自营销售楼瞬时代来自营业
	BERTabs	51信用卡管家孙海涛:[UNK]瞬时贷[UNK]每天放贷1 000万
	CBC	51信用卡管家创始人孙海涛: 2015年放贷额度将超30亿
	CBC-DA	51信用卡管家: 2015年放贷额度将远超30亿
	CBC-RA	51信用卡管家孙海涛: 2015年放贷额度将超30亿

表3中3个案例的实验结果分析如下:

1) 在例1中,transformer模型在语义理解上表达错误,原文“雅虎宣布剥离阿里巴巴股份”,表达含义是“剥离股份”,而并非“投资阿里巴巴股价”;BERTabs和本文提出的3个CBC模型与参考摘要表达含义完全一致,成功表明模型使用BERT后提升了语义理解准确性.

2) 在例2中,transformer模型错误地将“企业”生成“企业家”,并遗漏“10企业”和“上海”两个关键信息;BERTabs模型则抓错重点,原文表达的是“给上海自贸区打分”,而非“自贸区中外企业打分”,错误生成摘要,这是不合理的;CBC和CBC-RA模型很好地抓住了文章主旨,也注意到“上海”或者“10企业”两个关键信息,表明卷积门控单元的融入能够有效提取关键信息短语,不足在于CBC-DA模型仍错误将“企业”生成“企业家”.

3) 在例3中,transformer和BERTabs都没能成功生成与参考摘要相似的文本,而CBC三个模型的摘要接近于参考摘要. 在基础模型都没能生成相似摘要的情况下,融合模型的生成摘要不仅保留了关键信息,还与参考摘要高度重合.

综合上述3个实际案例,所提出的3个CBC模型在短文本数据集中能够很好地抓住全文关键信息,准确地表达原文含义,生成的摘要与原文摘要不仅更加相似,并且在语义上也更加相近.

4　结　论

本文提出了一种融合BERT和卷积门控单元的生成式文本摘要方法,通过使用大规模语料训练的预训练模型BERT作为编码器获取更多上下文语义信息,由卷积门控单元进一步筛选关键信息,排除冗余信息,同时设计3种不同融合方式的模型探讨如何生成更准确的文本摘要. 在LCSTS数据集上的实验表明,所提出模型在ROUGE评价指标上比当前主流中文生成式模型效果好,在CNN/Daily Mail英文数据集的实验表明,所提出模型优于基于基础BERT模型的生成式摘要. 下一步研究将考虑在现有工作的基础上,加强对长文挡语义结构、上下文语义关系、关键句子信息获取等方面的研究,构建长文本、多标题的文本摘要生成的模型和方法.

参考文献(References)

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. The 31st Conference on Neural Information Processing Systems. Long Beach: MIT Press, 2017: 6000-6010.

[2] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. Proceeding of the 2019 Conference of the North American Chapter of the Association for

Computational Linguistices. Piscataway: IEEE, 2019: 4171-4186.

[3] Yang Z L, Dai Z H, Yang Y M, et al. XLNet: Generalized autoregressive pretraining for language understanding[J]. CoRR, 2019: 1906.08237.

[4] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI Blog, 2019,1(8): 9-33.

[5] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. CoRR, 2014: 1409.3215.

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translation[C]. Proceedings of the Conference on Learning Processings. Piacataway: IEEE, 2015(1): 1-15.

[7] Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, 2015: 379-389.

[8] Hu B T, Chen Q C, Zhu F Z. LCSTS: A large scale Chinese short text summarization dataset[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, 2015: 1967-1972.

[9] Chopra S, Auli M, Rush A M. Abstractive sentence summarization with attentive recurrent neural networks[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, 2016: 93-98.

[10] Gu J T, Lu Z D, Li H, et al. Incorporating copying mechanism in sequence-to-sequence learning[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016: 1631-1640.

[11] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017: 1073-1083.

[12] Lin J Y, Sun X, Ma S M, et al. Global encoding for abstractive summarization[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018: 163-169.

[13] 吴仁守, 王红玲, 王中卿, 等. 全局自匹配机制的短文本摘要生成方法[J]. 软件学报, 2019, 30(9): 2705-2717.
(Wu R S, Wang H L, Wang Z Q, et al. Short text summary generation with global self-matching mechanism[J]. Journal of Software, 2019, 30(9): 2705-2717.)

[14] 田珂珂, 周瑞莹, 董浩业, 等. 基于编码器共享和门控网络的生成式文本摘要方法[J]. 北京大学学报: 自然科学版, 2020, 56(1): 61-67.
(Tian K K, Zhou R Y, Dong H Y, et al. An abstractive summarization method based on encoder-sharing and gated network[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2020, 56(1): 61-67.)

[15] Peng L, Liu Q, Lv L B, et al. An abstractive summarization method based on global gated dual encoder[C]. Natural Language Processing and Chinese Computing. Berlin, 2020: 355-365.

[16] Zhang H Y, Cai J J, Xu J J, et al. Pretraining-based natural language generation for text summarization[C]. Proceedings of the 23rd Conference on Computational Natural Language Learning. Hong Kong, 2019: 789-797.

[17] Zhang J Q, Zhao Y, Saleh M, et al. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization[C]. ICML. New York, 2019: 1-54.

[18] Lewis M, Liu Y H, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2020: 7871-7880.

[19] Xiao D L, Zhang H, Li Y K, et al. ERNIE-GEN: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation[C]. Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, 2020: 3997-4003.

[20] He R N, Ravula A, Kanagal B, et al. RealFormer: Transformer likes residual attention[J/OL]. 2020, arXiv: 2012.11747.

[21] Lin C. Rouge: A package for automatic evaluation of summaries[C]. Proceedings of the ACL Workshop: Text Summarization Braches Out. Barcelona, 2004: 74-81.

[22] Ma S M, Sun X, Lin J Y, et al. Autoencoder as assistant supervisor: Improving text representation for Chinese social media text summarization[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018: 725-731.

[23] Bae S, Kim T, Kim J, et al. Summary level training of sentence rewriting for abstractive summarization[J]. CoRR, 2019: 1909.08752.

作者简介

邓维斌(1978−), 男, 教授, 博士, 从事智能信息处理、确定性决策等研究, E-mail: dengwb@cqupt.edu.cn;

李云波(1996−), 男, 硕士生, 从事自然语言处理的研究, E-mail: lybde_email@163.com;

张一明(1982−), 男, 工程师, 硕士, 从事智能信息处理的研究, E-mail: 110054527@qq.com;

王国胤(1970−), 男, 教授, 博士生导师, 从事智能信息处理、粒计算等研究, E-mail: wanggy@cqupt.edu.cn;

朱坤(1997−), 男, 硕士生, 从事文本分类的研究, E-mail: 1209562838@qq.com.

(责任编辑：郑晓蕾)