



## 基于动态路由序列生成模型的多标签文本分类方法

王敏蕊, 高 曙\*, 袁自勇, 袁 蕾

(武汉理工大学 计算机科学与技术学院, 武汉 430063)

(\* 通信作者电子邮箱 gshu418@163.com)

**摘 要:** 现实世界中, 多标签文本比单标签文本具有更广泛的应用场景, 但其输出空间的庞大给分类任务带来了更多的挑战。将多标签文本分类问题看作标签序列生成问题, 把序列生成模型(SGM)应用于多标签文本分类领域, 并针对该模型的顺序结构容易产生累积误差等不足, 构建了基于动态路由(DR)的序列生成模型(DR-SGM)。该模型基于Encoder-Decoder模式; Encoder层中使用双向长短期记忆(Bi-LSTM)神经网络+Attention进行语义信息编码; Decoder层设计了一种基于动态路由的解码器结构, 该结构在隐含层后添加了动态路由聚合层, 利用路由参数的全局共享减弱了累积误差产生的影响。同时, 动态路由能捕获文本中部分-部分、部分-整体的位置信息, 并且通过优化动态路由算法进一步提高了语义聚合效果。将DR-SGM应用于多标签文本分类, 实验结果表明, 在RCV1-V2、AAPD和Slashdot数据集上, 多标签文本分类效果得到了有效的提升。

**关键词:** 多标签文本分类; 序列生成模型; 胶囊网络; 动态路由; 双向长短期记忆神经网络

**中图分类号:** TP391 **文献标志码:** A

### Sequence generation model with dynamic routing for multi-label text classification

WANG Minrui, GAO Shu\*, YUAN Ziyong, YUAN Lei

(School of Computer Science and Technology, Wuhan University of Technology, Wuhan Hubei 430063, China)

**Abstract:** In the real world, multi-label text has a wider application scenario than single-label text. At the same time, due to its huge output space, it brings a lot of challenges to the classification task. The multi-label text classification problem was regarded as label sequence generation problem, and the Sequence Generation Model (SGM) was applied to the multi-label text classification field. Aiming at the problems such as that the sequence structure of the model is easy to produce the cumulative error, an SGM based on Dynamic Routing (DR-SGM) was proposed. The model was based on Encoder-Decoder mode. In the Encoder layer, Bi-directional Long Short-Term Memory (Bi-LSTM) neural network+Attention was used to encode the semantic information. In the Decoder layer, a decoder structure with the dynamic routing aggregation layer was designed which reduces the influence of the cumulative error added behind the hidden layer. At the same time, the part-part and part-glob position information in the text was captured by dynamic routing. And by optimizing the dynamic routing algorithm, the semantic clustering effect was further improved. DR-SGM was applied to the classification of multi-label texts. The experimental results show that DR-SGM improves multi-label text classification results on the RCV1-V2, AAPD and Slashdot datasets.

**Key words:** multi-label text classification; Sequence Generation Model (SGM); capsule network; Dynamic Routing (DR); Bi-directional Long Short-Term Memory (Bi-LSTM) neural network

## 0 引言

文本分类是自然语言处理领域的重要问题之一, 传统的监督学习方法大多假设数据样本是单标签形式的, 即一个样本对应一个类别标签, 但现实生活中, 往往并不如此理想, 一个数据样本通常会表达极其复杂的多重语义。与单标签不同, 多标签样本给一个样本标注多个标签, 从而更加准确、有效地表达单标签所不能表达的复杂语义关系。多标签文本在日常生活中十分常见, 例如: 一条新闻可能同时包含“华为”“5G”“通信技术”等多个主题, 一条微博可能同时标注“明星”

“综艺”“搞笑”等多个标签, 因此, 研究多标签文本分类对挖掘具有丰富语义的现实世界文本对象具有重要的意义。

多标签的传统分类方法包括二值分类(Binary Relevance, BR)方法、分类器链(Classifier Chain, CC)等。BR方法不考虑标签之间的相关性, 但由于其简单而应用广泛。CC方法考虑每一个标签与其他所有标签之间的关系, 将多标签学习问题转化为一组有序的二分类问题, 其中, 每个二分类器的输入都要基于之前分类器的预测结果。传统多标签分类方法中文本特征的提取往往需要人工干预, 容易带来噪声, 同时又非常耗费人力。近年来, 深度学习方法在单标签文本分类任务上取

收稿日期: 2019-11-28; 修回日期: 2020-02-10; 录用日期: 2020-02-14。 基金项目: 国家自然科学基金资助项目(51679180)。

作者简介: 王敏蕊(1995—), 女, 江西南昌人, 硕士研究生, 主要研究方向: 自然语言处理; 高曙(1967—), 女, 湖北武汉人, 教授, 博士, 主要研究方向: 智能计算与语义识别、数据分析与应用; 袁自勇(1995—), 男, 安徽亳州人, 硕士研究生, 主要研究方向: 自然语言处理; 袁蕾(1997—), 女, 安徽滁州人, 硕士研究生, 主要研究方向: 自然语言处理。



得了非常好的成绩<sup>[1-3]</sup>,但国内外基于深度学习的多标签文本分类模型尚处于研究阶段,针对现有深度学习模型挖掘标签相关性效果差问题,有学者提出将多标签文本分类问题看作标签序列生成,并取得了较好效果<sup>[4-6]</sup>。对于多标签文本分类,每个样本对应的标签集都可以看作一个标签序列,为文本进行多标签标注可以看成标签序列生成,而循环神经网络(Recurrent Neural Network, RNN)及其变体已应用于各种序列建模任务中。文献[5]中首次将多标签文本分类看作序列生成任务,序列生成模型(Sequence Generation Model, SGM)中Decoder使用RNN的变体长短期记忆(Long Short-Term Memory, LSTM)神经网络,基于已经预测的标签产生下一个标签,这种顺序结构由于考虑了标签之间的相关关系,从而获得了更好的多标签文本分类效果。但是由于其序列性,容易造成累计误差。针对以上问题,本文受到胶囊网络中的动态路由(Dynamic Routing, DR)思想启发,将序列生成模型和动态路由方法结合,增加动态路由聚合层,克服序列生成中的累积误差缺陷,并将其应用于多标签文本分类。本文的主要工作如下:

1) 将序列生成模型与胶囊网络中的动态路由思想结合,设计了一种基于动态路由的解码器结构。这种解码器结构能减弱序列生成模型中累积误差的影响,其中,优化的动态路由算法能提升语义聚合效果。

2) 利用所提出的解码器结构,构建了基于动态路由的序列生成模型(SGM based on DR, DR-SGM),并将DR-SGM应用于多标签文本分类。该模型能通过其顺序结构捕捉标签相关性,从而提升多标签分类效果。

3) 将本文模型在三个多标签文本数据集进行测试,实验结果表明,本文模型性能优于7个基准模型。

## 1 相关研究

多标签文本分类任务一直是自然语言处理领域一个十分重要却又富有挑战性的任务。多年来,国内外学者在多标签文本分类领域投入了大量研究。多标签文本分类,顾名思义,即是对具有多个标签的文本样本进行标签预测,它相对于单标签文本分类更加复杂。现有的多标签文本分类方法可划分为传统方法和深度学习方法,综述如下:

按照解决策略准则,传统机器学习方法中将多标签分类分为问题转化和算法适应两类。问题转化方法指将多标签问题转化为一个或一组单标签问题,从而运用已有的单标签算法解决,如标签幂集(Label Powerset, LP)<sup>[7]</sup>、分类器链<sup>[8]</sup>等。算法适应方法指通过改进现有单标签算法以完成多标签学习任务。例如:Osojnik等<sup>[9]</sup>设计了一种基于流式多目标回归器iSOUP-Tree的多标签分类方法;李兆玉等<sup>[10]</sup>为每个训练样本的近邻集合计算其近邻密度和近邻权重,提出了一种基于引力模型的多标签分类算法;刘慧婷等<sup>[11]</sup>设计了基于去噪自编码器和矩阵分解的联合嵌入多标签分类算法Deep AE-MF。

基于深度学习模型的多标签文本分类模型尚处研究阶段,并没有很完整的体系分类,但已经有学者取得了一些成果:Baker等<sup>[12]</sup>设计了一种基于卷积神经网络(Convolution

Neural Network, CNN)架构的标签共现的多标签文本分类方法;Kurata等<sup>[13]</sup>提出了一种新颖的基于标签共现神经网络初始化方法;Shimura等<sup>[14]</sup>提出一种针对短文本多标签文本的分层卷积神经网络结构,该方法利用类别之间的层次关系解决短文本数据稀疏问题;Yang等<sup>[15]</sup>提出了一种可以“重新考虑”预测的标签的深度学习框架;宋攀等<sup>[16]</sup>提出了一种基于神经网络探究标签依赖关系的算法执行多标签分类任务;Liu等<sup>[17]</sup>针对极端多标签文本分类中巨大的标签空间引发的数据稀疏性和可扩展性,考虑标签共现问题,提出了专为一种多标签学习设计的新的卷积神经网络模型;He等<sup>[18]</sup>将标签关联、缺失标签和特征选择联合起来,提出一种新的多标签分类学习框架;Banerjee等<sup>[19]</sup>将多标签文档按层次划分,制定了一种新的基于迁移学习的分类策略HTrans。序列生成思想应用于多标签文本分类已有部分成果:Chen等<sup>[4]</sup>提出通过将CNN与RNN组合以捕捉全局和局部文本语义,并通过RNN输出标签序列;Yang等<sup>[5]</sup>首次提出将序列生成思想应用于多标签文本分类;Qin等<sup>[6]</sup>延续序列生成思想,构建新的训练目标,以便RNN能发现最佳标签顺序。

综上所述,深度学习方法被越来越多地应用于多标签文本分类领域,序列生成模型是多标签文本分类中一次成功的尝试,但其标签序列生成过程中容易产生累积误差,严重影响时间靠后的标签生成,从而降低标签标注准确率。本文主要针对这个不足展开研究工作。

## 2 基于动态路由的RNN序列生成模型

多标签文本指一个实例被多个标签标注的文本,多标签文本分类问题的目标是为每个未分类文本样本标注合适的类别标签。形式化地描述为:

假设文本样本空间 $X = \{x_1, x_2, \dots, x_m\}$ ,对应包含 $n$ 个类别的标签空间 $Y = \{y_1, y_2, \dots, y_n\}$ ,现有多标签文本训练集 $D = \{(x_i, y_i) | 1 \leq i \leq k\}$ ,多标签分类任务的目的就是利用训练集 $D$ 学习到一个分类器 $C: X \rightarrow 2^Y$ 。对于每一个样本 $x_i$ ,都有一个标签集合 $Y_i$ 与之关联<sup>[20]</sup>。

为更好探究标签之间的相关性,本文构建了一种基于动态路由的RNN序列生成模型(DR-DGM),以取得更好的多标签文本分类效果。

### 2.1 面向多标签文本分类的序列生成模型

序列生成模型Seq2Seq(Sequence to Sequence)是一种Encoder-Decoder结构,最早应用于机器翻译任务中,并在当时取得了巨大成功。其主要思想是通过神经网络将原始输入的可变长序列映射到另一可变长度的序列中。其中,使用的神经网络通常是RNN,常用的有LSTM神经网络和门控循环单元网络(Gated Recurrent Unit, GRU)。Seq2Seq模型结构如图1所示,主要包括三部分:

- 1) 编码器(Encoder):读取原始语言序列,将其编码成为一个固定长度的具有原始语言信息的向量。
- 2) 中间状态变量:对所有输入内容的集合。
- 3) 解码器(Decoder):根据中间状态变量,得到解空间的



概率分布,最终生成输出可变量序列。



图1 序列生成模型结构

Fig. 1 Architecture of sequence generation model

受序列生成模型启发,有学者创新性地多标签分类问题看作标签序列生成问题。在Encoder层使用双向长短期记忆(Bi-directional Long Short-Term Memory, Bi-LSTM)神经网络+Attention结构捕获语义信息,在Decoder层的每一时刻都进行一次标签序列生成,预测的标签集合由各个时刻生成的标签组成。

序列生成模型利用LSTM顺序地生成标签以捕获标签之间的相关性,但也是由于其顺序结构,上一时刻的输出对下一时刻的标签生成具有重要影响,如果上一时刻包含错误信息,那么下一时刻的标签输出大概率也是错误的。为了尽可能将上一时刻的正确信息传导下去,本文受胶囊网络<sup>[21-22]</sup>启发,使用动态路由聚合解码器结构中的信息,以提升文本语义信息传递的聚合效果,从而更好地降低错误信息的叠加。

2.2 基于动态路由的解码器结构及动态路由算法优化

为解决传统卷积神经网络无法捕捉图像特征位置相对关系的缺点,胶囊网络<sup>[21]</sup>应运而生。在文本处理中,胶囊网络中的动态路由过程能捕捉部分-部分、部分-整体的位置信息<sup>[22]</sup>,也能更好地聚合文本语义信息<sup>[23]</sup>。本文将动态路由过程应用于序列生成模型的解码器结构中,具体结构如图2所示。

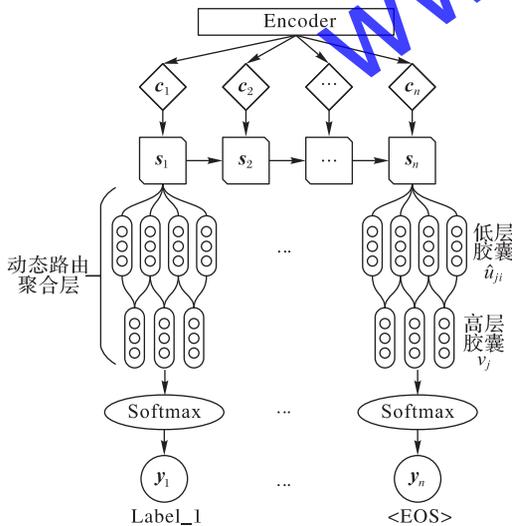


图2 基于动态路由的解码器结构

Fig. 2 Decoder based on dynamic routing

图2中省略了Encoder层和中间语义变量  $c_i \in \{c_1, c_2, \dots, c_n\}$  的详细内容,在计算得到Decoder层隐含变量  $s_i \in \{s_1, s_2, \dots, s_n\}$  后,将其输入动态路由聚合层(具体的路由优化算法如算法1所示),最后输出到Softmax层进行分类并得到解空间的标签概率分布,每一时刻的输出为解空间中概

率最大的标签。其中,标签的预测以<EOS>标志为结束。此外,动态路由聚合层的参数是全局共享的,这样能减弱累积误差产生的影响。

同时,本文探索了两种策略以优化动态路由的聚合效果。首先,为解决动态路由过程中的类别分布稀疏问题,本文使用sparsemax代替动态路由中的softmax<sup>[24]</sup>,如式(1):

$$\text{sparsemax}(z) := \underset{p \in \Delta^{K-1}}{\text{argmin}} \|p - z\|^2 \quad (1)$$

其中:  $\Delta^{K-1} = \{p \in \mathbf{R}^K \mid \mathbf{1}^T p = 1, p \geq 0\}$ ,表示  $(K-1)$  维单形,从  $\mathbf{R}^K$  到  $\Delta^{K-1}$  的映射能够更有效地将实际权重向量转化为概率分布。sparsemax将输入向量的欧氏距离投影转化为概率单形,这种投影方法使得sparsemax更适用于类别稀疏的情况。其次,为加强低层胶囊到高层胶囊的连接强度,引入高层胶囊权重系数  $w_j$  ( $j$  表示第  $j$  个高层胶囊),  $w_j$  是高层胶囊  $v_j$  的模(具体计算方法参见算法1中描述),并用于修正下一次低层胶囊对高层胶囊的连接强度,在迭代过程中提升对分类结果有重要影响的胶囊权重。

依据以上两点改进,设计动态路由优化算法如算法1所示。其中,squash表示非线性激活函数<sup>[21]</sup>。

算法1 动态路由优化算法。

输入 低层胶囊  $\hat{u}_{ji}$ , 迭代次数  $r$ , 低层胶囊所处网络层  $l$ ;  
输出 高层胶囊  $v_j$ 。

- 1) 对所有第  $l$  层胶囊  $i$  连接第  $l+1$  层胶囊  $j$  的可能性  $b_{ij} : b_{ij} \leftarrow 0$
- 2) for  $k = 1$  to  $r$  do
  - 对第  $l$  层的所有胶囊  $i$ : 设置中间变量  $c_i \leftarrow w_j \cdot \text{sparsemax}(b_i)$
  - 对第  $l+1$  层的所有胶囊  $j: s_j \leftarrow \sum_i c_{ij} \hat{u}_{ji}$
  - 对第  $l+1$  层的所有胶囊  $j: v_j \leftarrow \text{squash}(s_j), w_j = |v_j|$
  - 对第  $l$  层的所有胶囊  $i$  和对第  $l+1$  层的所有胶囊:  $b_{ij} \leftarrow b_{ij} + \hat{u}_{ji} \cdot v_j$
- 3) end
- 4) 返回  $v_j$

综上所述,本文将胶囊网络中的动态路由算法进行优化,然后将其与解码器结构融合,构建了如图2所示的基于动态路由的解码器结构。

2.3 DR-SGM 架构

利用2.2节所提出的解码器结构及优化的路由算法,设计基于动态路由的RNN序列生成模型框架(DR-SGM),如图3所示,其中D/R胶囊图标具体细节即图2所展示内容。模型主要由以下几个部分组成:

1) 输入层。对原始文本进行预处理,然后使用word2vec词嵌入技术将其转换为数字表示的词向量,模型的输入为多个词向量组合得到句子向量。

2) Encoder层。假设输入的句子中含有  $m$  个单词,向量化后该句子可表示为  $(e_1, e_2, \dots, e_p, \dots, e_m)$ , 其中  $e_i$  表示该句子中第  $i$  个词对应的词向量。Encoder层使用Bi-LSTM+Attention机制,具体计算过程见式(2)~(5):

$$\bar{h}_i = \overline{\text{LSTM}}(\bar{h}_{i-1}, e_i) \quad (2)$$



$$\bar{h}_i = \overline{\text{LSTM}}(\bar{h}_{i+1}, e_i) \quad (3)$$

$$h_i = [\bar{h}_i, \bar{h}_i] \quad (4)$$

$$\alpha_{ii} = \frac{\exp(v_a^T \tanh(W_a s_i + U_a h_i))}{\sum_{j=1}^m \exp(v_a^T \tanh(W_a s_i + U_a h_j))} \quad (5)$$

其中:  $h_i$  表示第  $i$  个单词对应 Encoder 层中的隐含状态, 它由  $i$  时刻前向 LSTM  $\bar{h}_i$  和反向 LSTM  $\bar{h}_i$  联结而成;  $\alpha_{ii}$  表示在  $t$  时刻, Attention 机制为第  $i$  个单词分配的权重;  $W_a, U_a, v_a^T$  都是权重系数。

3) 中间语义层。每一时刻的中间语义向量  $c_i$  由 Encoder 层中隐含状态  $h_i$  计算得到, 其计算公式如式(6):

$$c_i = \sum_{j=1}^m \alpha_{ij} h_j \quad (6)$$

4) Decoder 层。  $t$  时刻 Decoder 层的隐含向量  $s_i$  首先由中间语义向量  $c_i$  计算得到, 公式如(7)。

$$s_i = \text{LSTM}(s_{i-1}, [g(y_{i-1}); c_{i-1}]) \quad (7)$$

其中:  $g(y_{i-1})$  代表概率分布  $y_{i-1}$  中最高概率标签的全局嵌入<sup>[5]</sup>。

然后将隐含向量  $s_i$  输入动态路由聚合层, 即图 3 中的 D/R 胶囊图标。

$$dr_i = DR(s_i) \quad (8)$$

其中,  $DR$  代表动态路由过程, 具体内容见 2.2 节。

5) 输出层。输出层在每一个时刻都会输出一个标签概率分布  $y_i$ , 每次取最高概率标签作为当前时刻的“标签序列生成”,  $y_i$  的计算公式如下:

$$o_i = W_o f(W_d dr_i + V_d c_i) \quad (9)$$

$$y_i = \text{Softmax}(o_i + I_i) \quad (10)$$

其中:  $W_o, W_d$  和  $V_d$  是权重系数;  $I_i$  是为了保证不预测重复标签的掩码向量;  $f$  是非线性激活函数<sup>[5]</sup>。

由图 3 可知, DR-SGM 模型在 SGM+GE (SGM+Global Embedding) 模型<sup>[5]</sup>的基础上进行了改进, 首先, 使用 sparsemax 和迭代权重  $w$  优化动态路由策略; 然后, 添加动态路由层, 使用优化的动态路由算法改进解码器结构, 以强化语义聚合效果, 捕获文本关系, 削弱因顺序结构造成的累积误差; 最后, 在以上工作的基础上, 构建基于动态路由的序列生成模型。

算法 2 基于 DR-SGM 的多标签文本分类算法。

输入 多标签文本数据集  $(x^{(n)}, y^{(n)}) (n = 1, 2, \dots, N)$ , 训练轮数  $r$ ;

输出 预测标签集合  $\hat{y}$ 。

- 1) 对原始文本进行预处理: 去停用词、分词; 对标签进行数字化处理
- 2) 使用预训练的 300 维 word2vec 词向量, 将输入句子转换为词向量矩阵输入 DR-SGM 模型中
- 3) for  $i = 1$  to  $r$  do:
- 4) 通过 Encoder 层计算得到中间语义向量  $c_i$
- 5) 在 Decoder 层对  $c_i$  解码得到 Decoder 隐含层隐含向量  $s_i$ , 将  $s_i$  胶囊化为矢量向量后输入动态路由聚合层, 使用动态路由优化算法计算, 最后经过 softmax 得到输出序列
- 6) 获取输出序列中概率最大的未生成的 Label 标签, 将其加

- 7) end
- 8) 更新模型各层参数, 其中动态路由由聚合层共享全局参数, 使用 Adam 优化器优化神经网络
- 9) 输出  $\hat{y}$

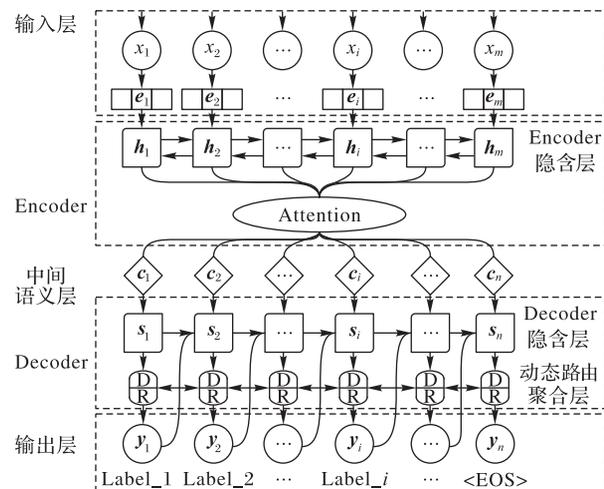


图 3 基于动态路由的 RNN 序列生成模型 (DR-SGM) 架构

Fig. 3 Architecture of RNN sequence generation model based on Dynamic Routing (DR-SGM)

### 3 基于 DR-SGM 的多标签文本分类算法

利用 DR-SGM 模型, 设计多标签文本分类算法如算法 2 所示。首先对文本进行去停用词、分词, 将其转化为词向量后进行本地结构化存储。句子转化为词向量矩阵后输入 DR-SGM 模型, 通过 Encoder 层得到各个时刻的中间语义向量  $c_i$ , 再计算出 Decoder 层隐含向量  $s_i$ , 转化为胶囊向量后作为动态路由优化算法的输入, 进行三次路由迭代。最后通过 Softmax 输出标签序列。输出层在每个时刻的输出中选择输出标签序列中最大概率的标签加入预测标签序列, 以  $\langle \text{EOS} \rangle$  为序列生成结束标志。每次训练完成后, 在测试数据集上验证模型分类效果, 对模型参数进行迭代更新, 共享动态路由层参数, 并通过 Adam 优化器优化神经网络。

## 4 实验设置

### 4.1 数据集

本文采用 RCV1-V2、AAPD 和 Slashdot 作为实验数据集: 公开数据集 RCV1-V2 是路透社公布的新闻数据集, 包含 804 414 篇新闻, 共 103 个主题; AAPD 数据集是 arxiv 网站的论文摘要数据集, 包含 55 840 个论文标题和摘要, 共 54 个主题; Slashdot 是一个社交网络数据集, 包含 24 072 个文档, 共 291 个主题。数据集的具体信息如表 1 所示。

表 1 数据集详细信息

Tab. 1 Detail of datasets

数据集	样本数量	总标签数	平均标签数	平均文本长度
RCV1-V2	804 414	103	3.24	121
AAPD	55 840	54	2.41	163
Slashdot	24 072	291	4.15	64



## 4.2 评价指标

采用  $F1$  值、汉明损失 (Hamming Loss, HL) 作为性能评价指标, 如式 (11)、(12):

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \quad (11)$$

$$HL = \frac{1}{M} \sum_{i=1}^M |Y_i \Delta Y_i^*| \quad (12)$$

其中:  $Precision$  代表准确率,  $Recall$  代表召回率,  $M$  为样本数,  $Y_i$  是真实标签集合,  $Y_i^*$  是预测标签集合,  $Y_i \Delta Y_i^*$  为样本预测标签集合和真实标签集合的对称差分。  $F1$  值越大多标签分类效果越好, 而  $HL$  是衡量样本中误分标签平均数量的指标,  $HL$  越小, 误分标签数量越少, 多标签分类模型性能越好。

## 4.3 通用设置

本文实验基于 Tensorflow 框架, 使用 Numpy、Keras 库进行开发, 编程语言是 Python 3.6。数据集被随机洗乱, 其中 90% 作为训练集, 剩余 10% 作为测试集。词向量使用预训练的 300 维 word2vec 词向量, 不在字典中的低频单词用全 0 表示。RCV1-V2 固定句子长度为 500, AAPD 数据集固定句子长度为 300, Slashdot 数据集固定句子长度为 120, 多余截去, 不足用 0 补齐。此外, dropout 设置为 0.5, 学习率设置为 0.001, 并使用 Adam 优化器和交叉熵损失函数训练数据。

## 5 实验及结果分析

### 5.1 基准模型

本文使用以下基准模型与本文构建的 DR-SGM 模型进行对比:

1) 二值分类 (Binary Relevance, BR): BR 算法将多标签分类任务分解成  $n$  个独立的二元分类问题, 每一个二元分类问题对应于标签空间中的某一特定标签。

2) 分类器链 (Classifier Chain, CC): CC 将多标签学习问题转化为一组有序的二分类问题, 其中每个二分类器的输入都要基于之前分类器的预测结果。

3) 标签幂集 (Label Powerset, LP): LP 将多标签学习问题转化为多类分类问题进行学习。它将训练数据集的标签集合每个不同的标签子集成为 labelset, 看作是单标签分类任务中多类分类问题的不同类别值, 然后利用分类器进行求解。

4) CNN-RNN<sup>[4]</sup>: 利用 CNN 捕获全局文本特征后输入 RNN 进行局部语义特征捕获, 同时考虑标签相关性。

5) 序列生成模型 (SGM)<sup>[5]</sup>: 将多标签文本分类问题转换为标签序列生成问题。

6) SGM+GE (Global Embedding, 全局嵌入)<sup>[5]</sup>: 在序列生成模型的基础上使用 Global Embedding。

7) set-RNN (Adapting RNN to Multilabel Set Prediction, 自适应 RNN)<sup>[6]</sup>: 同样将多标签文本分类问题转换为标签序列生成问题, 提出新的训练和预测目标, 使 RNN 能发现最佳标签顺序。

其中: 1)~3) 是传统机器学习算法, 均使用梯度提升决策树作为基分类器; 4)~7) 是基于 RNN 的深度学习模型。

### 5.2 动态路由胶囊维数实验及分析

胶囊维数对动态路由过程有重要影响, 胶囊维数过少可能无法有效捕捉文本语义, 胶囊维数过多可能导致噪声出现。因此, 本文对动态路由的胶囊数对实验结果的影响进行了探索, 结果如表 2 所示。在 RCV1-V2 和 Slashdot 上, 低层胶囊数/高层胶囊数为 32/16 时取得较好效果。而在 AAPD 数据集上, 低层胶囊数/高层胶囊数为 16/8 时取得较好效果。就平均文本长度, RCV1-V2 和 Slashdot 数据集中文本更短小, 可能需要更多的胶囊进行语义信息捕获。此外, 并不是胶囊数越多, 性能越优, 也出现了胶囊数增多, 性能不变甚至下降的情况, 这可能是由于多余胶囊捕获了额外的无关语义信息, 从而对计算结果造成负面影响。

表 2 胶囊数对实验结果的影响

Tab. 2 Effect of the number of capsules on experimental results

低层 胶囊 数	高层 胶囊 数	RCV1-V2		AAPD		Slashdot	
		F1	HL	F1	HL	F1	HL
16	8	0.883	0.0077	<u>0.722</u>	<u>0.0242</u>	0.539	0.0670
32	16	<u>0.889</u>	<u>0.0071</u>	0.721	0.0244	<u>0.541</u>	<u>0.0667</u>
64	32	0.888	0.0071	0.719	0.0244	0.541	0.0669

### 5.3 模型评估与分析

在 RCV1-V2、AAPD 和 Slashdot 数据集上, 分别利用  $F1$  值和  $HL$  两个性能指标, 测试了上述 6 个基准模型以及本文提出 DR-SGM 模型, 实验结果如表 3 所示, “—”表示不可获取。其中  $F1$  值越大, 反映模型性能越好,  $HL$  则正好相反。

表 3 实验结果

Tab. 3 Results of experiments

模型	RCV1-V2		AAPD		Slashdot	
	F1	HL	F1	HL	F1	HL
BR	0.858	0.0086	0.646	0.0316	0.484	0.0736
CC	0.857	0.0087	0.654	0.0306	0.480	0.0728
LP	0.858	0.0087	0.634	0.0312	0.516	0.0708
CNN-RNN	0.856	0.0085	0.664	0.0278	—	—
SGM	0.869	0.0081	0.699	0.0251	0.528	0.0681
SGM+GE	0.878	0.0075	0.710	0.0245	0.532	0.0675
set-RNN	0.838	—	0.720	<u>0.0241</u>	0.538	—
DR-SGM	<u>0.889</u>	<u>0.0071</u>	<u>0.722</u>	0.0242	<u>0.541</u>	<u>0.0667</u>

从表 3 可以看出 (最佳结果在表格中用下划线标出), 在 RCV1-V2、AAPD 以及 Slashdot 数据集上, DR-SGM 模型的大部分评估标准相对于基准模型都取得了最优的效果, 只有在 AAPD 数据集上, 其  $HL$  比 set-RNN 模型略逊一筹, 低了 0.4%。然而相对于 SGM+GE 模型 (在其基础上改进), 在 RCV1-V2 数据集上, 其  $F1$  值提升了 1.25%,  $HL$  提升了 5.3%; 在 AAPD 数据集和 Slashdot 数据集上, 其  $F1$  值和  $HL$  均有一定程度提升。

从实验结果看, 深度学习方法 (包括本文提出的 DR-SGM 以及 CNN-RNN、SGM 和 set-RNN) 相较传统方法 (包括 BR、CC 和 LP 等), 无疑有着更加优秀的结果。传统方法非常依赖于特征工程, 而复杂的特征工程往往带来繁琐的工作和人工操作错误的风险。对于一些复杂的情况, 传统方法由于特征工程的局限常常无法进行处理。而深度学习方法可以自动提取特征, 完全消除了特征工程带来的负面影响。此外, 数据集



Slashdot、RCV1-V2、AAPD 包含标签数分别为 291、103、54。在各种分类方法下,相对于其他数据集,拥有近三百个标签的 Slashdot 分类结果显然十分不理想,其原因可能是样本数量与标签量的不匹配。Slashdot 的标签数量是 RCV1-V2 的近 3 倍, AAPD 的 5 倍多,但是其样本量只有 RCV1-V2 的约 1/30, AAPD 的约 1/2,能够训练的样本数量不足以匹配庞大的标签数,同时文本长度短,能捕捉的语义信息少,因而造成分类评价结果差。然而数据集 RCV1-V2 的标签数是 AAPD 的约 2 倍,其分类效果却明显优于 AAPD 数据集,这可能是因为 RCV1-V2 数据集的样本数更多,大约为 AAPD 的 15 倍,因此模型能够学习到的内容更多,从而分类效果更好。由此可见,多标签文本分类方法对样本数量的依赖性很大,同时,标签数量和文本长度也是影响分类效果的重要因素。

就各种深度学习方法而言,由于在多标签文本分类任务中,标签相关性是极其重要的信息之一,捕捉标签相关性对多标签文本分类具有重大意义,而 CNN-RNN 模型中并没有考虑到标签相关性问题,但 DR-SGM 模型通过 LSTM 顺序结构处理标签序列,每一个生成的标签都充分考虑了之前标签的信息,从而取得了比它更好的效果;DR-SGM 模型是基于 SGM+GE 模型进行优化,相对于原始的 SGM 模型或 SGM+GE 模型都有一定性能上的提升,其原因可能在于,动态路由方法能够额外捕获文本中部分-部分、部分-整体的位置信息,同时因为动态路由聚合层共享了全局参数,削弱了前一时刻的文本信息对后面时刻的影响,从而降低 RNN 循环结构造成的累积误差。set-RNN 模型通过数学方法修改模型训练的方法和目标,使其能发现最佳标签顺序,误分标签数较少,Hamming Loss 指标表现更好,但是,相对于 DR-SGM,它在有效捕捉文本语义方面略微逊色,因此 F1 值结果略逊一筹。

综上所述,无论是和传统方法相比,还是和现有的深度学习方法相比,DR-SGM 都取得了有竞争力的结果。

## 6 结语

本文沿用将序列生成模型应用于多标签文本分类的思想,将多标签文本分类看作一个标签序列生成问题,不同于以往的解码器结构,本文借鉴胶囊网络思想,将动态路由应用于序列生成中的解码器结构,构建了 DR-SGM 模型。在 Encoder 层,通过使用 BiLSTM+Attention 结构最大限度捕捉语义信息;在 Decoder 层,增加了动态路由聚合层聚合文本信息,从而额外捕获了文本中部分-部分、部分-整体的位置信息,同时通过在全局范围内共享动态路由参数,在一定程度上减轻了序列生成产生累积误差的负面影响,而且,在设计的动态路由算法中,为解决路由过程中类别稀疏问题,采用 sparsemax 代替 Softmax;为加强低层胶囊到高层胶囊的连接强度,引入权重系数  $w$  在动态路由过程进行迭代加权。此外,面向多标签文本分类领域,制定了基于 DR-SGM 的多标签文本分类算法,实验结果表明,相比 7 个基准模型,本文设计的 DR-SGM 模型取得了较好的分类效果。

在多标签文本分类领域,仍然有许多问题值得探索,例如序列生成模型极度依赖标签顺序,而在现实生活中标签集合

是无序的;同时多标签文本中往往存在大量样本不均衡的情况,对部分类别标签样本的偏向性会严重影响分类模型的分

## 参考文献 (References)

- [1] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 562-570.
- [2] WANG B. Disconnected recurrent neural networks for text categorization [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 2311-2320.
- [3] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2016: 1480-1489.
- [4] CHEN G, YE D, XING Z, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization [C]// Proceedings of 2017 International Joint Conference on Neural Networks. Piscataway: IEEE, 2017: 2377-2383.
- [5] YANG P, SUN X, LI W, et al. SGM: sequence generation model for multi-label classification [C]// Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 3915-3926.
- [6] QIN K, LI C, PAVLU V, et al. Adapting RNN sequence prediction model to multi-label set prediction [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019: 3181-3190.
- [7] ZHOU W, YU Y, ZHANG M. Binary linear compression for multi-label classification [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2017: 3546-3552.
- [8] 胡天磊,王皓波,尹文栋. 基于深度双向分类器链的多标签新闻分类算法[J]. 浙江大学学报(工学版), 2019, 53(11): 2110-2117. (HU T L, WANG H B, YIN W D. Multi-label news classification algorithm based on deep bi-directional classifier chains [J]. Journal of Zhejiang University (Engineering Science), 2019, 53(11): 2110-2117.)
- [9] OSOJNIK A, PANOVA P, DŽEROSKI S. Multi-label classification via multi-target regression on data streams [J]. Machine Learning, 2017, 106(6): 745-770.
- [10] 李兆玉,王纪超,雷曼,等. 基于引力模型的多标签分类算法 [J]. 计算机应用, 2018, 38(10): 2807-2811, 2821. (LI Z Y, WANG J C, LEI M, et al. Multi-label classification algorithm based on gravitational model [J]. Journal of Computer Applications, 2018, 38(10): 2807-2811, 2821.)
- [11] 刘慧婷,冷新杨,王利利,等. 联合嵌入式多标签分类算法[J]. 自动化学报, 2019, 45(10): 1969-1982. (LIU H T, LENG X Y,



- WANG L L, et al. A joint embedded multi-label classification algorithm[J]. *Acta Automatica Sinica*, 2019, 45(10):1969-1982. )
- [12] BAKER S, KORHONEN A. Initializing neural networks for hierarchical multi-label text classification[C]// *Proceedings of the 2017 Conference on Biomedical Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2017: 307-315.
- [13] KURATA G, XIANG B, ZHOU B. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence[C]// *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA: Association for Computational Linguistics, 2016: 521-526.
- [14] SHIMURA K, LI J, FUKUMOTO F. HFT-CNN: learning hierarchical category structure for multi-label short text categorization[C]// *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2018: 811-816.
- [15] YANG Y Y, LIN Y A, CHU H M, et al. Deep learning with a rethinking structure for multi-label classification[EB/OL]. [2019-03-12]. <https://arxiv.org/pdf/1802.01697.pdf>.
- [16] 宋攀, 景丽萍. 基于神经网络探究标签依赖关系的多标签分类[J]. *计算机研究与发展*, 2018, 55(8): 1751-1759. (SONG P, JING L P. Exploiting label relationships in multi-label classification with neural networks[J]. *Journal of Computer Research and Development*, 2018, 55(8): 1751-1759. )
- [17] LIU J, CHANG W C, WU Y, et al. Deep learning for extreme multi-label text classification[C]// *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 2017: 115-124.
- [18] HE Z, YANG M, GAO Y, et al. Joint multi-label classification and label correlations with missing labels and feature selection[J]. *Knowledge-Based Systems*, 2019, 163: 145-158.
- [19] BANERJEE S, AKKAYA C, PEREZ-SORROSAL F, et al. Hierarchical transfer learning for multi-label text classification[C]// *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2019: 6295-6300.
- [20] 熊涛. 基于长短时记忆网络的多标签文本分类[D]. 杭州: 浙江大学, 2017. (XIONG T. Multi-label text classification based on long short-term memory network[D]. Hangzhou: Zhejiang University, 2017. )
- [21] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[C]// *Proceedings of 31st International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2017: 3856-3866.
- [22] YANG M, ZHAO W, YE J, et al. Investigating capsule networks with dynamic routing for text classification[C]// *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2018: 3110-3119.
- [23] GONG J, QIU X, WANG S, et al. Information aggregation via dynamic routing for sequence encoding[C]// *Proceedings of the 27th International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2018: 2742-2752.
- [24] MARTINS A F T, ASTUDILLO R F. From softmax to sparsemax: a sparse model of attention and multi-label classification[C]// *Proceedings of 33rd International Conference on Machine Learning*. New York: JMLR.org, 2016: 1614-1623.
- This work is partially supported by the National Natural Science Foundation of China (51679180).
- WANG Minrui**, born in 1995, M. S. candidate. Her research interests include natural language processing.
- GAO Shu**, born in 1967, Ph. D., professor. Her research interests include intelligent computing and semantic recognition, data analysis and application.
- YUAN Ziyong**, born in 1995, M. S. candidate. His research interests include natural language processing.
- YUAN Lei**, born in 1997, M. S. candidate. Her research interests include natural language processing.