文章编号:1671-6833(2022)06-0030-06

基于图文注意力融合的主题标签推荐

冯皓楠, 何智勇, 马良荔

(中国人民解放军海军工程大学 电子工程学院, 湖北 武汉 430000)

摘 要: 为了解决社交媒体平台上的信息超载问题,帮助用户快速捕捉所需信息,对基于多模态内容的标签推荐问题进行研究。针对不同模态间的异质性差异,采用共注意力机制进行跨模态内容的特征建模与融合;针对多标签分类方法只能推荐出数据集标签空间中标签的不足,采用 Seq2Seq 框架生成新的标签序列,并通过一种聚合策略将分类方法的推荐结果聚合到生成的标签序列中,得到 2 种方法的统一推荐模型。在大规模数据集上的实验结果表明:多模态方法比单模态方法更具优势,所提出的统一推荐模型的 $F1$ 值比仅使用单模态的对比模型高 9.44 百分点;生成新标签序列的方法也优于传统的分类方法,所提出的标签序列生成模型的 $F1$ 值比对比模型 COA 高 3.41 百分点;所提出的统一推荐模型 UNIFIED-CO-ATT 的 $F1$ 值比 GEN-CO-ATT 模型高 1.25 百分点,其效果优于其他对比模型。所提出的模型综合了分类方法和生成方法的特点,可以使推荐的标签同时具有准确性和新颖性。

关键词: 共注意力机制; 标签分类; 标签生成; 统一模型; 多模态推荐

中图分类号:TP301.6;TP391.1 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2022.03.001

0 引言

社交媒体平台(如 Twitter)上提供了大量的文本、图片及视频数据,这些数据的爆发式增长已经远远超过了人们的接收理解能力。如何消化大量嘈杂的社交媒体数据,提取其中的重要内容,为用户推荐其所需的快速访问信息已经成为一个新的挑战。用户在社交媒体平台发布文本、图片和视频数据时,会使用一种特定形式的元数据标签(hashtag),它是一串以符号#为前缀的字符,一般可以用来描述帖子中的关键词或主题。

表 1 展示了一个用户在 Twitter 上为帖子内容配上标签的示例。通过帖子文本及其配套图片的耦合效应指示帖子的主题内容并且推荐一系列能反映帖子的主要关注点的标签是目前研究的热点。然而,前人的研究主要集中在文本特征的使用上[1],但社交媒体的语言风格本质上是非正式的、碎片化的,为了丰富语境,本文分析利用了帖子中配套的图片内容。

现有的研究主要是针对单模态的标签推荐或关于多模态标签推荐的分类算法的研究,但从实际应用的角度出发,生成数据集标签空间中不存在的标签至关重要。因此,本文进行了多模态标签序列生成模型(GEN-CO-ATT)的研究,并进一步提出了多模态标签推荐算法的分类方法和生成方法的统一模型(UNIFIED-CO-ATT)。

表 1 Twitter 数据集中的一个真实帖子示例

Table 1 A real post example from Twitter dataset

类别	内容
文本	stitches came out today ! the vet says i only have to wear my cone of shame for 7 more days!
图片	
标签	#dogs are family;#dogs of Twitter

本文旨在为新型社交平台设计一种完整而有效的标签推荐方法,采用共注意力机制对多模态内容进行建模融合,并采用 Seq2Seq 框架生成新的标签序列(GEN-CO-ATT);同时,针对分类方法和生成方法的特点,采用复制机制的扩展方法将分类模型的结果聚合到序列生成模型的输出中,并通过 2 个模块端到端的联合训练得到一个统一

收稿日期:2021-08-25;**修订日期**:2021-11-20

基金项目:"十三五"预研项目(41412010801)

通信作者:何智勇(1981—),男,山东博兴人,中国人民解放军海军工程大学讲师,博士,主要从事人工智能研究,E-mail:moonmon_pub@outlook.com。

的标签推荐模型(UNIFIED-CO-ATT)。

1　相关工作

早期的研究工作中,通常仅将多模态内容各自建模使用,例如,Vinyals 等[2]提出先对文本和图片建模,提取高层图片特征,再将其输入 LSTM 中对图片生成字幕;何伟成[3]提出基于图卷积神经网络的个性化标签推荐算法,借助图卷积网络的表示、学习能力进行标签推荐;Yang 等[4]使用注意力机制多次查询图片,逐步推断推荐结果。但是,这些工作并没有考虑图片对文本特征提取的指导意义和二者之间的关联。

为了分析多模态内容之间的语义关联性,张素威[5]提出了一个基于异质注意力的图文融合的标签推荐模型,既强化了跨模态的共性信息,也考虑了不同模态差异信息之间的互补性。由于共注意力机制[6]可以同时考虑文本与图片对推荐结果的影响,Zhang 等[7]采用共注意力机制对文本和图片的关联建模,通过分类的方法研究了基于多模态内容的标签推荐问题。

在关键词预测方面,大部分工作是直接从源输入中提取序列[8]或从预定义的候选列表中进行分类[9],这样不会产生数据集标签空间中不存在的关键词。受到在科学文章中生成关键词方法的启发,Wang 等[10]采用 Seq2Seq 框架实现了在社交媒体平台上生成关键词;Chen 等[11]也采用了分离检索的方法来生成关键字;Wang 等[12]基于复制机制将分类方法的结果与生成方法的结果进行聚合。首先,本文应用共注意力机制对多模态内容进行建模与融合;其次,建立基于多模态内容的标签分类模型和标签序列生成模型,允许端到端的联合训练,以更好地捕捉 2 种模型的多样化结果,并通过一种聚合策略将分类方法的输出结果聚合到生成的标签序列中;最后,得到 2 种方法的统一推荐模型。

2　统一的多模态标签推荐模型

首先定义一个集合 C,输入集为 $|C|$ 个帖子的文本-图片对 $\{(x^n, I^n)\}_{n=1}^{|C|}$,为每个文本-图片对推荐一个标签集合 $y=\{y^i\}_{i=1}^{|y|}$。参考 Meng 等[13]的研究,多次复制源输入对,使每个输入对具有一个关键词。将每个输入对表示为一个三元组 (x, I, y),x、y 分别为文本内容的单词序列 $x=<x_1, x_2, \cdots, x_{l_x}>$和标签内容的单词序列 $y=<y_1, y_2, \cdots, y_{l_y}>$,其中 l_x 和 l_y 表示单词序列 x、y 的字数。

图 1 为所提出的多模态标签推荐模型的总体框架。该模型是从下往上运行的:首先,将帖子中的文本和图片编码为文本表示和图片表示,使用共注意力机制捕捉它们复杂的语义交互;其次,将学习到的多模态表示向量 $\boldsymbol{c}_{\text{fuse}}$ 用于标签的分类模型或序列生成模型,使用一种聚合策略来组合它们的输出;最后,上述整个框架可以通过多任务学习的方式联合训练为一个整体的模型。

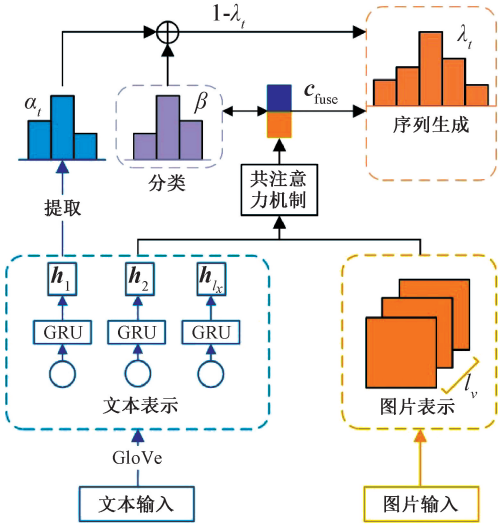


图 1　基于多模态内容的标签推荐统一模型

Figure 1　Unified model of hashtag recommendation based on multi-modal content

2.1　多模态编码

(1)学习文本表示。通过数据集预训练的查找表将文本输入序列中的每个单词 x_i 嵌入到一个高维向量中,使用双向门控循环单元(BiGRU)对嵌入后的单词 $e(x_i)$ 进行编码,表达式为

$$\overrightarrow{h}_i = GRU(e(x_i), \overrightarrow{h}_{i-1}); \tag{1}$$

$$\overleftarrow{h}_i = GRU(e(x_i), \overleftarrow{h}_{i+1})。 \tag{2}$$

将前向隐藏状态 $\overrightarrow{h}_i$ 和后向隐藏状态 $\overleftarrow{h}_i$ 连接成为向量 $\boldsymbol{h}_i=[\overrightarrow{h}_i; \overleftarrow{h}_i]$,本文把它作为 x_i 的上下文感知表示,将输入序列中的所有 $\boldsymbol{h}_i$ 全部存储到一个文本向量库 $\boldsymbol{M}_{\text{text}}=\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{l_x}\} \in \mathbf{R}^{l_x \times d}$ 中,其中 d 为隐藏状态的维度。

(2)学习图片表示。采用在大规模图片库 ImageNet 上预训练后的 VGG-16 网络[14]对每个图片 I 提取 49 个卷积特征图,每个特征图通过一个线性投影层转化为一个新的向量 $\boldsymbol{v}_i$,然后存储到一个图片向量库 $\boldsymbol{M}_{\text{vis}}=\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{l_v}\} \in \mathbf{R}^{l_v \times d}$ 中,其中 l_v 为图片区域的个数。

2.2　共注意力机制

本文多模态内容的融合采用了以文本为主导

的共注意力机制[7]。这个注意力机制依次交替产生图片注意和文本注意,如图 2 所示,包括 3 个步骤:①将文本表示向量总结为单个向量;②根据文本总结向量计算图片注意力;③基于图片特征的注意力再次计算文本的注意力。具体来说,计算注意力的操作为 $\hat{x}=\text{Attention}(X,g)$,以图片(文本)特征 X 和文本/图片的注意力引导 g 作为输入,输出计算后的图片(文本)注意力向量,表达式为

$$\boldsymbol{H}=\tanh(\boldsymbol{W}_x X+\boldsymbol{W}_g g); \tag{3}$$

$$\alpha^x=\text{softmax}(\boldsymbol{\omega}_{hx}^{\mathrm{T}}\boldsymbol{H}); \tag{4}$$

$$\hat{x}=\sum \alpha_i^x x_i。 \tag{5}$$

式中:$\boldsymbol{W}_x$、$\boldsymbol{W}_g\in\mathbf{R}^{k\times d}$,$\boldsymbol{\omega}_{hx}\in\mathbf{R}^k$ 均为特征矩阵;α^x 为特征 X 的注意力权重。

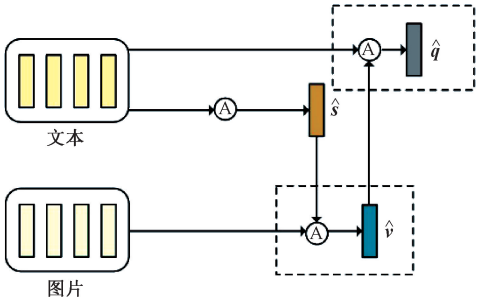


图 2 共注意力机制结构

Figure 2 Co-attention mechanism structure

考虑社交媒体数据的噪声特性,采用最大/平均池化层为每个模态获取一个整体的查询向量,将所有共注意力层的输出通过一个线性多模态融合层表示为上下文向量 $\boldsymbol{c}_{\text{fuse}}\in\mathbf{R}^d$,并输入标签分类模型和标签序列生成模型中进行标签推荐。

2.3 统一的多模态标签推荐模型

结合不同方法的特点,采用一种聚合策略将多模态标签推荐的分类方法和生成方法结合为一个统一的推荐模型。

步骤 1 标签分类。由于每个标签 y 通常只由几个单词组成,因此可以将单词视为整体标签的离散部分,并通过推荐单词来推荐标签。在分类方法中,直接将多模态上下文向量 $\boldsymbol{c}_{\text{fuse}}$ 传递到一个双层的多层感知器 *MLP* 中,然后将它映射到标签分类词汇表 V_{cls} 的分布中:

$$P_{\text{cls}}(y)=\text{softmax}(MLP_{\text{cls}}(\boldsymbol{c}_{\text{fuse}}))。 \tag{6}$$

步骤 2 标签序列生成。在标签序列生成方面,使用 Seq2Seq 框架来生成新的标签序列 $y=<y_1,y_2,\cdots,y_{l_y}>$,其中生成器概率定义为

$$\prod_{t=1}^{l_y}=P(y_t\mid y<t)。 \tag{7}$$

采用一个单向的门控循环单元 GRU 解码器对生成建模过程,具体来说,解码器释放的隐藏状态 $\boldsymbol{s}_t=GRU(\boldsymbol{s}_{t-1},\boldsymbol{u}_t)\in\mathbf{R}^d$ 是基于前一个隐藏状态 $\boldsymbol{s}_{t-1}$ 和嵌入式解码器的输入 $\boldsymbol{u}_t$,$\boldsymbol{s}_t$ 由文本编码器的最后一个隐藏状态 $\boldsymbol{h}_{l_x}$ 初始化。采用共注意力机制获取文本的上下文语境向量 $\boldsymbol{c}_{\text{text}}$:

$$\boldsymbol{c}_{\text{text}}=\sum_{i=1}^{l_x}\alpha_{t,i}\boldsymbol{h}_i; \tag{8}$$

$$\alpha_{t,i}=\text{softmax}(S(\boldsymbol{s}_t,\boldsymbol{h}_i)); \tag{9}$$

$$S(\boldsymbol{s}_t,\boldsymbol{h}_i)=\boldsymbol{v}_\alpha^{\mathrm{T}}\tanh(\boldsymbol{W}_\alpha[\boldsymbol{s}_t;\boldsymbol{h}_i]+\boldsymbol{B}_\alpha)。 \tag{10}$$

式中:$S(\boldsymbol{s}_t,\boldsymbol{h}_i)$ 为得分函数,用来衡量第 t 个被解码的单词和文本编码器的第 i 个单词之间的兼容性;$\boldsymbol{W}_\alpha\in\mathbf{R}^{d\times 2d}$,$\boldsymbol{B}_\alpha$、$\boldsymbol{v}_\alpha\in\mathbf{R}^d$ 均为可训练权值。

接下来结合静态多模态向量 $\boldsymbol{c}_{\text{fuse}}$ 来构建丰富的上下文表示:

$$\boldsymbol{c}_t=[\boldsymbol{u}_t;\boldsymbol{s}_t;\boldsymbol{c}_{\text{text}}+\boldsymbol{c}_{\text{fuse}}]。 \tag{11}$$

在此基础上,采用另一个带有 softmax 函数的 *MLP* 将 $\boldsymbol{c}_t$ 映射到生成词汇表 V_{gen} 的单词分布中:

$$P_{\text{gen}}(y_t)=\text{softmax}(MLP_{\text{gen}}(\boldsymbol{c}_t))。 \tag{12}$$

为了使解码器更好地从源输入帖子中复制单词,应用复制机制[15]设置一个带有 sigmoid 激活函数的 *MLP* 软开关 $\lambda_t\in[0,1]$,它决定了模型是从词汇表 V_{gen} 中生成单词序列还是从源输入序列中提取单词,其中提取源输入序列的概率分布由文本注意力权重 $\alpha_{t,i}$ 决定。

步骤 3 聚合策略。使用复制机制的扩展方法将分类模型的输出结果聚合到标签序列生成结果中:①从分类模型中检索前 K 个预测结果,并将其转换为单词序列 $w=<w_1,w_2,\cdots,w_{l_w}>$,$l_w$ 为组合预测后的序列长度;②使用 softmax 函数将它们的分类对数归一化为一个词级分布 $\beta\in\mathbf{R}^{l_w}$,该分布表示单词从分类输出中被提取的概率。

步骤 4 统一模型的标签推荐。根据聚合后的结果得到统一的标签推荐模型。

$$P_{\text{unf}}(y_t)=\lambda_t P_{\text{gen}}(y_t)+(1-\lambda_t)\left(a\sum_{i:x_i=y_t}^{l_x}\alpha_{t,i}+b\sum_{j:w_j=y_t}^{l_w}\beta_j\right)。 \tag{13}$$

式中:a、b 为超参数,$a+b=1$,用于决定模型是从输入序列中提取单词还是从分类输出中提取单词。为了稳定分类输出结果的聚合,设置 a 为 1,b 为 0,输入分类器进行训练,实验完成几个批次后,将两者都设置为 0.5 以进行更进一步的训练。

2.4 联合训练目标

本文采用标准的负对数似然损失函数来定义

整个模型的训练目标。似然损失函数由多任务学习的标签分类损失和单词级序列生成损失的线性组合构成:

$$L(\theta) = -\sum_{n=1}^{N}\left[\log P_{\text{cls}}(y^n) + \gamma\sum_{t=1}^{l_n^y}\log P_{\text{unf}}(y_t^n)\right]。\quad (14)$$

式中:N 为训练文本-图片对的大小;γ 为平衡这 2 个损失的超参数,设为 1;θ 表示整个框架共享的可训练参数。从式(14)可以看出,联合训练标签分类模型有助于统一的标签推荐,不仅隐式地提供了更好的参数学习,还明确提供了更精确的输出,以供聚合策略组合到标签生成模型中。

3　实验与结果分析

本文的实验设置为 Ubuntu20.04、CPU i9-10900X、64 GB 内存、NVIDIA GeForce RTX 2090,实验环境为 python3.6、pytorch1.5。

3.1　数据收集和统计

由于缺少社交媒体平台基于多模态内容的帖子及标签的公开数据集,因此本文使用了文献[12]中公开的数据集。该数据集使用了 Twitter 高级搜索 API 查询 2019 年 1 月至 2019 年 6 月期间包含文本、图片和标签的英文帖子,并获得 53 701 条推文。本文将数据按 8:1:1随机划分为训练集、验证集和测试集。数据集的数据分割和统计信息如表 2 所示。

表 2　数据集的数据分割和统计

Table 2　Data segmentation and statistics of dataset

数据集	帖子数目	帖子的平均标签个数	标签长度
训练集	42 959	1.33	1.85
验证集	5 370	1.34	1.85
测试集	5 372	1.32	1.86

3.2　实验设置

3.2.1　评价指标

本文采用信息检索指标宏平均 $F1$ 值来评估本文模型,选取推荐概率排名前 K 的主题标签计算评价指标,例如:$F1@K$ 表示推荐概率排名前 K 的标签计算出的 $F1$ 值,其中 $K=1,3,5$。$F1@K$ 值越大表示模型性能越好。为了进一步测量标签的推荐顺序,本文对推荐概率排名前 5 的标签采用平均精度指标 MAP(mean average precision)[16]进行评价。指标得分越高表示模型性能越好。

3.2.2　参数设置

本文使用了一个有 45 000 单词的生成词汇表 V_{gen} 和 4 262 个标签的关键短语分类词汇表 V_{cls},采用 200 维的 Twitter GloVe 嵌入[17]来编码文本输入。采用两层的 BiGRU 作为编码器,一层的 GRU 作为解码器,隐藏大小设置为 300。对于图片,本文使用 VGG-16 提取 49 个特征图和 512 维的特征。在训练中,本文设置损失系数 $\gamma=1$,采用 Adam 优化器,学习率为 0.001。如果验证损失没有下降,则采用最大梯度范数为 5 的梯度裁剪方法将其衰减 0.5,通过监测验证损失的变化,采用了提前停止方法。

3.2.3　对比模型

选择 2 种对比模型 TAKG[10] 和 COA[6]。TAKG 模型是针对社交媒体平台的主题感知关键词生成模型,只使用了帖子中的文本模态信息推荐关键字;COA 模型是针对社交媒体平台的基于多模态内容的主题标签推荐模型,此模型使用共注意力机制对多模态特征建模,并使用多类分类的方法进行标签推荐。

3.3　实验结果

表 3 为本文模型与其他模型的实验结果对比。分析表 3 可得如下结论。

表 3　各模型的实验结果对比

Table 3　Experimental results of models　%

模型	$F1@1$	$F1@3$	$MAP@5$
TAKG	36.38	27.65	43.49
COA	41.16	31.13	47.44
GEN-CO-ATT	44.57	31.24	49.56
UNIFIED-CO-ATT	45.82	31.26	49.85

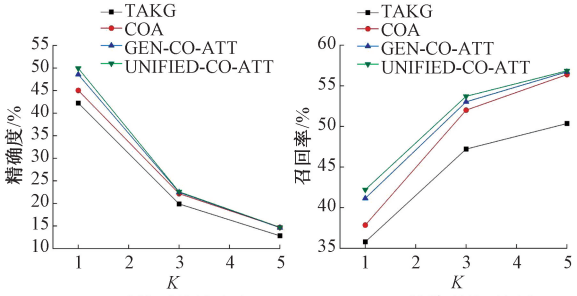
(1)多模态方法比单模态方法更具优势。所提统一推荐模型 UNIFIED-CO-ATT 的 $F1$ 值比仅使用单模态的对比模型 TAKG 高 9.44 百分点;所提标签序列生成模型 GEN-CO-ATT 相比于 TAKG 模型在 $F1@1$、$F1@3$、$MAP@5$ 上分别提升 8.19 百分点、3.59 百分点、6.07 百分点。可以看出,考虑多模态内容的模型比只考虑文本模态内容的模型有更好的表现,这说明基于 Seq2Seq 框架的标签序列生成模型能够很好地利用社交媒体平台上多模态信息的特殊性,且图片模态提供了许多文本模态中未包含的额外信息。

(2)生成新标签序列的方法也优于传统的分类方法。所提 GEN-CO-ATT 模型比基于多模态内容的多类分类方法进行主题标签推荐的模型 COA 在 $F1@1$、$F1@3$、$MAP@5$ 上分别提升 3.41 百分点、0.11 百分点、2.12 百分点。这说明基于多模态内容进行主题标签推荐的问题中,能够生

成出标签空间中不存在的主题标签是非常重要的,分类方法只能推荐出在标签空间中预定义的主题标签,有一定局限性。

(3)本文统一标签推荐模型 UNIFIED-CO-ATT 比仅使用生成方法的 GEN-CO-ATT 模型在 $F1@1$、$F1@3$、$MAP@5$ 上分别提升 1.25 百分点、0.02 百分点、0.29 百分点,即统一的标签推荐模型比仅使用分类方法的模型表现更好。这说明本文先联合训练分类模型和生成模型,再将分类结果聚合于生成方法中进行优化的聚合策略有效果。这种聚合策略使模型同时具有准确性和新颖性的特点。

图 3 为 4 种模型在 $K=1,3,5$ 时的精确度和召回率。由图 3 可以看出,模型 GEN-CO-ATT 和 UNIFIED-CO-ATT 在精确度和召回率方面也优于对比模型 TAKG 和 COA。由于测试集中每个帖子中已有的标签的平均数量为 1.32(见表 2),因此所有模型在 K 从 1 到 3 的性能比 K 从 3 到 5 的性能表现更好,同时性能也下降更快;在 $K>3$ 时,模型的性能都逐渐平稳。这可能是由于在本文使用的嘈杂的社交媒体数据集中,关键词数量大但是缺位率高的原因。



(a) 4种模型的精确度 (b) 4种模型的召回率

图 3 4 种模型在 $K=1,3,5$ 时的精确度和召回率

Figure 3 Accuracy and recall rate of 4 models with $K=1,3,5$

4 结束语

本文围绕社交媒体平台上的基于多模态内容的标签推荐问题,研究了标签序列生成模型在此问题中的性能表现,进一步提出了一个统一的标签推荐模型,将序列生成模型和分类模型的优势结合起来。此外,本文使用的先联合训练单个模型,再将分类模型结果聚合到生成模型结果中的聚合策略是有效的。在大规模数据集上的实验结果表明,本文的模型明显优于只使用文本内容生成标签的模型和仅使用分类方法推荐标签的模型。

参考文献:

[1] ZHANG Y Y, LI J, SONG Y, et al. Encoding conversation context for neural keyphrase extraction from microblog posts[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 1676-1686.

[2] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3156-3164.

[3] 何伟成. 基于图卷积神经网络的个性化标签推荐系统[D]. 广州:华南理工大学, 2020.
HE W C. Personalized tag recommender system based on graphconvolutional neural network[D]. Guangzhou: South China University of Technology, 2020.

[4] YANG Z C, HE X D, GAO J F, et al. Stacked attention networks for image question answering[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 21-29.

[5] 张素威. 社交网络多模态内容标签推荐技术研究[D]. 南京:南京大学, 2020.
ZHANG S W. Research on hashtag recommendation for multimodal contents in social networks[D]. Nanjing: Nanjing University, 2020.

[6] LU J S, YANG J W, BATRA D, et al. Hierarchical question-Image co-attention for visual question answering[EB/OL]. (2017-01-19)[2021-01-06]. https://arxiv.org/pdf/1606.00061.pdf%20.

[7] ZHANG Q, WANG J W, HUANG H R, et al. Hashtag recommendation for multimodal microblog using co-attention network[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne: International Joint Conferences on Artificial Intelligence Organization, 2017: 3420-3426.

[8] ZHANG Q, WANG Y, GONG Y Y, et al. Keyphrase extraction using deep recurrent neural networks on twitter[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 836-845.

[9] CHAN H P, CHEN W, WANG L, et al. Neural keyphrase generation via reinforcement learning with adaptive rewards[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 2163-2174.

[10] WANG Y,LI J,CHAN H P,et al. Topic-aware neural keyphrase generation for social media language[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:Association for Computational Linguistics, 2019: 2516-2526.

[11] CHEN W,CHAN H P,LI P J,et al. An integrated approach for keyphrase generation via exploring the power of retrieval and extraction[C]// Proceedings of the 2019 Conference of the North. Stroudsburg: Association for Computational Linguistics,2019:2846-2856.

[12] WANG Y,LI J,LYU M,et al. Cross-media keyphrase prediction: a unified framework with multi-modality multi-head attention and image wordings[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 3122-3132.

[13] MENG R,ZHAO S Q,HAN S G,et al. Deep keyphrase generation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017: 582-592.

[14] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2021-01-06]. https://www.researchgate.net/publication/265385906_Very_Deep_Convolutional_Networks_for_Large-Scale_Image_Recognition.

[15] BAHDANAU D, CHO K H, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. Statistics,2014,3:1-15.

[16] 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. Stroudsburg: Association for Computational Linguistics,2017:1073-1083.

[17] 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: Association for Computational Linguistics,2014:1532-1543.

Multimodal Hashtag Recommendation Based on Image and Text Attention Fusion

FENG Haonan, HE Zhiyong, MA Liangli

(School of Electronic Engineering, Naval University of Engineering, Wuhan 430000, China)

Abstract: In order to solve the information overload problem on social media platforms and help users quickly capture the required information, in this study the problem of hashtag recommendation based on multimodal content was investigated. To address the heterogeneous differences between different modalities, a co-attention mechanism was used to model and fuse features of cross-modal content, and use Seq2Seq framework was used to generate new hashtag sequences to address the deficiency that multi-label classification methods could only recommend hashtags in the hashtag space of the dataset. An aggregation strategy was used to aggregate the recommendation results of classification methods into the generated hashtag sequences to obtain a unified recommendation model for both methods. The experimental results on a large-scale dataset showed that, firstly, the multimodal approach was more advantageous than the unimodal approach, and the unified recommendation model proposed in this paper had 9.44 percentage points improvement in $F1$ value over the comparison model using unimodal approach, and 3.41 percentage points improvement over the comparison model using the classification method. Finally, the unified recommendation model UNIFIED-CO-ATT is 1.25 percentage points higher than GEN-CO-ATT in $F1$ values. The model proposed in this study could combine the advantages of classification and generation methods and could make the recommended hashtags have the advantages of accuracy and novelty at the same time.

Keywords: co-attention mechanism; hashtag classification; hashtag generation; unified model; multimodal recommendation