基于联合注意力和共享语义空间的多标签文本分类

孙坤[1]　秦博文[1]　桑基韬[1]　于剑[1]

[1]（北京交通大学计算机与信息技术学院 北京 100044）
（sunkun@bjtu.edu.cn）

Multi-label Text Classification based on Joint Attention and Shared Semantic Space

Sun Kun[1] , Qin Bowen[1] , Sang Jitao[1],and Yu Jian[1]

[1]（*School of Computer and Information Technology*, *Beijing Jiaotong University*, Beijing 1）

Abstract In the multi-label text classification task, each given document corresponds to a set of related labels. At present, it mainly faces the following three problems: (1) the joint modeling of label-text and label-label relationships is inadequate; (2) the semantic mining of the label itself is insufficient; (3) the utilization of the internal structure information of the label is ignored. To solve the above problems, this paper proposes a multi-label text classification method based on joint attention and shared semantic space. The proposed fused multi-head attention mechanism synchronously models the relationship between labels and relationship between labels and documents simultaneously, that avoids error transmission and use the interaction information between them. The proposed decouple shared semantic space embedding method improves the method of using labels semantic information, and uses the encoder of shared parameters to extract the semantic representation of labels and documents, reducing its deviation in the phase of modeling correlation. The proposed hierarchical hinting method based on prior knowledge relies on the prior knowledge in the pre-trained model to exploit the labels hierarchy information. Experimental results show that the proposed method is superior to the existing state-of-the-art multi-label text classification methods in public datasets.

Key words　multi-label text classification; attention mechanism; label representation; pre-trained model; semantic embedding

摘要　在多标签文本分类任务中，每个给定的文档都对应一组相关标签．目前主要面临以下三方面问题：（1）对标签-文本和标签-标签关系的联合建模不充分；（2）对标签本身语义的挖掘不足；（3）忽略了对标签内部结构信息的利用．对于以上问题，本文提出了一种基于联合注意力和共享语义空间的多标签文本分类方法．提出了融合多头注意力机制，该方法旨在同步地对标签与文档的关系和标签之间的关系进行建模，利用两者交互信息的同时避免误差传递．提出了解耦的共享语义空间嵌入方法，改进了利用标签语义信息的方法，使用共享参数的编码器提取标签和文档的语义表示，减少其在建模相关性阶段的偏差．提出了一种基于先验知识的层次提示方法，利用预训练模型中的先验知识增强标签层次结构信息．实验结果表明，该方法在公开数据集上优于目前最先进的多标签文本分类模型．

关键词　多标签文本分类；注意力机制；标签表示；预训练模型；语义嵌入

中图法分类号　TP391

多标签文本分类是一种重要的自然语言处理任务，在知识抽取[1]、问答[2]、情感分析[3]等领域有着广泛应用。多标签文本分类与常规文本分类的区别在于多标签文本分类允许将一篇文档同时分类为多个类别。由于文档是一个复杂的语义集合，不同的标签关注文档的不同部分。

在多标签文本分类任务上，目前的工作主要解决以下三个方面的问题：（1）探究如何从文档中充分捕捉语义信息以得到信息更丰富的文档表示，即文档表示的挖掘，这是多标签文本分类的基本问题。（2）探

收稿日期：2022-01-22
通信作者：桑基韬（jtsang@bjtu.edu.cn）

索如何获取特定标签的文档表示，即标签-文档关系的挖掘。文档本身是一个复杂的语义集合，导致文档的不同部分对于不同类别的判别的贡献存在差异。（3）探索如何利用标签之间的相关性，即标签-标签的关系的挖掘。标签与标签之间具有相关性，例如大部分多标签文本分类任务的标签之间有层次结构。近年来大部分相关工作[4,18-22]在解决第一个方面问题的基础上，主要关注对后两者之一的探索，少量工作[5-8]同时对两方面进行了探索。但这些模型在不同阶段建模两个相关性，造成了误差传递且没有利用到两者的交互信息。

此外，多标签文本分类任务的标签有两个重要的特点：（1）标签本身也是文本，具有丰富的语义信息。（2）标签具有丰富的结构信息，如共现关系和层级关系。以 AAPD 数据集[4]为例，“Logic in Computer Science”与“Programming Languages”常常同时出现，这两个标签都是“Computer Science”的子标签。有些工作开始关注到标签本身的语义信息[6-10]，但提取标签语义表示时与文档的语义表示不处于同一语义空间，或提取时两者过于耦合互相干扰，导致后续对两者关系建模的时候存在偏差。最后，标签本身具有层次结构，该信息对标签之间相关性的建模有很大帮助，但目前的工作没有注意到这个问题。

针对上述问题，在标签-文档关系和标签-标签关系建模方面，本文提出了一种融合多头注意力机制，以注意力的形式同时对两种关系进行建模，避免了误差传递并使它们的信息可以同步交互。对于标签语义的表示提取中存在的问题，本文提出了一种解耦的共享语义空间嵌入方法，通过使用共享参数的预训练语言模型作为编码器，使标签的语义表示和文档的语义表示处于同一语义空间且两者互不干扰，充分利用模型学习的语义信息。对于标签层次结构信息的利用，本文提出了基于先验知识的层次提示方法，在标签文本预处理阶段使用可以提示层次关系的标记对描述标签上下位的词语进行分割，使模型关注到层次信息。

本文的主要贡献如下：

1) 提出了融合多头注意力机制同时建模标签-文档关系和标签-标签关系，鼓励模型学习到两者的交互信息并避免误差传递；

2) 提出了一种解耦的共享语义空间嵌入方法提取标签和文档的语义表示，使其在同一语义空间并避免了在编码阶段互相干扰；

3) 提出了基于先验知识的层次提示方法来帮助模型利用标签的层次结构建模标签的相关性；

4) 在 AAPD[4]和 RCV1-V2[11]数据集上进行实验，本文模型在两个数据集上均超过目前最先进模型，验证了本文模型的有效性。

1 相关工作

我们分别从上文提到的三个方面（文档本身的挖掘、标签-文档关系的挖掘和标签-标签关系的挖掘）对相关工作进行整理。

针对第一个方面的研究，在 2014 年 CNN 被用来提取文本表示[10]。CNN 在捕获文本关键信息和局部模式上表现很好，但它忽略了了上下文信息，尤其是长距离的依赖关系。Liu 等人[13]引入 RNN 来获取上下文信息。然而，RNN 提取语义特征时存在偏好，文本后面的词比前面的词的影响更大。故 Chen 等人[14]提出结合 RNN 和 CNN 进行文本表示，使用双向 RNN 捕获上下文信息，CNN 捕获局部特征。为了得到每个单词对文本表示的贡献，Ashish 等人[15]提出利用注意力对文本进行编码。随着 transformer[15]和 BERT[16]的发展，Sun 等人[17]探究了如何更好得将 BERT 应用在文本分类任务中。上述工作都只关注了如何提取文本表示，并没有考虑标签信息与该任务的联系。

为了捕获标签与文档之间的联系，You 等人[18]提出了一种基于树标签的模型 AttentionXML，使用自注意力机制捕获和每个标签最相关的部分。标签是由几个词构成的自然语言文本，具有语义信息。Wang 等人[19]和 Pappas 等人[20]关注到了这一问题，使用词嵌入表示对标签进行编码使标签具有语义信息，然后拼接文档表示和标签语义表示作为特征进行分类。这类方法并没有考虑到文档与标签之间的语义相关性，Xiao 等人[21]使用注意力机制计算标签与文档的语义相关性，获取标签特定的文档表示。Zhang 等人[6]使用 transformer 提取标签和文档的全局语义表示。

上述工作都没有考虑到标签之间的相关性对多标签文本分类的影响，例如有利于低频类别的学习。Kurata 等人[22]注意到这个问题，提出利用标签共现信息初始化模型权重，从而考虑了标签相关性。Yang 等人[4]将多标签文本分类建模成标签序列生成问题，并提出了一种解码器捕获标签相关性。用序列生成建模标签相关性存在曝光偏差的问题，且没有考虑到标签本身的语义特征。Zhang 等人[5]提出了两个标签共现预测任务辅助学习标签相关性。Guo 等人[7]利用标签的词共现信息构造了标签词的异构网络，使用图嵌入的方法获取标签表示。Ma 等人[8]提出一种标签特定的对偶图神经网络解决相似标签难以区分的问题。

2 基于共享语义空间的多标签文本分类方法

本节将详细介绍本文提出的方法。其模型如图 1 所示，文档和标签分别在共享参数的编码器得到语义表示后送入注意力模块同时学习标签-标签相关性和标签-文档相关性，根据相关性计算标签特定的文档表示，之后将其送入解码器解码，最后进行标签预测。本小节主要从以下几个方面进行介绍：文档语义表示学习、标签语义表示学习、融合多头注意力、解码器和标签预测。

2.1 基本定义

数据集$D=\{(x_i,\boldsymbol{y_i})\}_{i=1}^{N}$由 N 个文档x_i和对应的标签$\boldsymbol{y_i}\in\{0,1\}^L$组成，$L$ 为标签总数。多标签文本分类的目标就是学习一个从输入文档到最相关的数个标签的映射。

2.2 文档语义表示学习

BERT 通过掩码语言模型（Masked Language Model）和上下句预测（Next Sentence Prediction）任务在大规模无监督语料上进行预训练，以此学习到文本的语义信息和语法信息，在多个自然语言处理任务上有着出色的表现。使用 BERT 作为编码器可以获取到通过海量语料训练得到的包含丰富语义信息的表示。在本文中，我们使用共享参数的 BERT 提取文档和标签的语义表示，以使两者处于同一语义空间中。

具体来讲，对于第 i 篇文档词元化（tokenizer）后得到 j 个词元（token），即$x_i=\{w_1,w_2,\dots,w_j\}$。将词元序列输入到 BERT 中进行编码，得到每个词元的语义表示：

$$\boldsymbol{H}=BERT(w_1,w_2,\dots,w_j) \tag{1}$$

$\boldsymbol{H}\in\mathbb{R}^{j\times d_{model}}$，$d_{model}$为隐表示的维度。则$H$的每一行为每个词元的语义表示。

2.3 标签语义表示学习

对于 L 个标签，每个标签都有文本作为标签的描述，大部分工作都没有利用到这部分信息。对于标签文本集$E=(e_1,e_2,\dots,e_L)$，我们同样需要对其进行词元化。在现实场景中‘/’常常作为具有上下位关系文本的分隔符，为了使模型注意到标签中的层级信息，我们在上下级的描述文本中间加入词元‘/’。e_k处理后记为c_k。

目前最先进的利用标签描述文本的工作都将标签文本与文档同时输入到 BERT 中进行编码。这种方法有以下问题：（1）由于 BERT 是基于自注意力机制的，与文档同时输入到 BERT 时会互相干扰，难以利用预训练模型学习到的语法语义知识；（2）多标签文本的标签类别数量通常较大，而包括 BERT 在内的大部分预训练模型有输入长度限制。这会导致输入标签文本后，文档文本不能全部输入到预训练模型中，损失很多信息。部分数据集上甚至存在仅标签文本就已经超过输入长度的问题，使这种方法完全不可用。

所以我们采用与上一小节共享参数的 BERT 单独对标签文本进行编码，既可以充分利用预训练模型学习到的知识，又不受编码长度限制。这里我们利用 BERT 中的“[CLS]”符号为每个标签学习一个向量表示：

$$\boldsymbol{A_k}=BERT(c_k) \tag{2}$$

$$\boldsymbol{A}=[\boldsymbol{A_1},\boldsymbol{A_2},\dots,\boldsymbol{A_L}]^T \tag{3}$$

$\boldsymbol{A_k}\in\mathbb{R}^{d_{model}}$，$d_{model}$为隐表示的维度。则 A 中的每一行为每个标签的语义表示。

2.4 融合多头注意力

每个标签关注的文档内容是不同的，可以通过文档和标签的语义相关性对其进行建模，同时标签的描述文本也蕴含了丰富的标签间关系。具有这一阶段，我们使用多头注意力机制同时建模文档与标签的相关性和标签与标签的相关性。为了在得到标签特定的文档表示的同时获取到标签间关系，我们使用标签表示作为查询向量，标签表示和文档表示拼接后作为键向量和值向量：

$$\mathbf{Q}=\mathbf{A} \tag{3}$$

$$\mathbf{K}=\mathrm{Concat}(\boldsymbol{A},\boldsymbol{H}) \tag{4}$$

$$\mathbf{V}=\mathrm{Concat}(\boldsymbol{A},\boldsymbol{H}) \tag{5}$$

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^T}{\sqrt{d_k}}\right)\cdot\mathbf{V} \tag{6}$$

$$\boldsymbol{head_i}=Attention(\boldsymbol{Q}\cdot\boldsymbol{W_i^Q},\boldsymbol{K}\cdot\boldsymbol{W_i^K},\boldsymbol{V}\cdot\boldsymbol{W_i^V}) \tag{7}$$

$$\boldsymbol{C}=Concat(\boldsymbol{head_1},..,\boldsymbol{head_h})\cdot\boldsymbol{W^o} \tag{8}$$

其中，$\mathbf{W_i^Q}\in\mathbb{R}^{d_{model}\times d_k},\mathbf{W_i^K}\in\mathbb{R}^{d_{model}\times d_k},\mathbf{W_i^V}\in\mathbb{R}^{d_{model}\times d_v},\mathbf{W_i^Q}\in\mathbb{R}^{hd_v\times d_{model}}$为可学习矩阵。h为注意力头数，$d_k=d_v=d_{model}/h$，最终得到$\mathbf{C}\in\mathbb{R}^{L\times d_{model}}$。

2.5 解码器

文本的解码器受 transformer 结构启发，由残差结构、前馈神经网络和层归一化组成，具体计算如下：

$$\mathbf{C}'=\mathrm{LN}(\mathbf{C}) \tag{9}$$

$$\mathbf{C}''=\mathrm{FNN}(\mathbf{C}') \tag{10}$$

$$\mathbf{M}=\mathrm{LN}(\mathbf{C}''+\mathbf{C}') \tag{11}$$

其中，LN 为层归一化操作，FNN 为前馈神经网络，最终得到的$\mathbf{M}\in\mathbb{R}^{L\times d_{model}}$为标签对应的文档表示，第 i 个标签对应的文档表示为$\boldsymbol{M_i}\in\mathbb{R}^{1\times d_{model}}$。

2.6 标签预测

我们使用前馈神经网络和激活函数作为分类器：

$$y_i=\sigma(\boldsymbol{M_i}\cdot\boldsymbol{W_i^y}+b_i)$$

其中，$\boldsymbol{W_i^y}\in\mathbb{R}^{d_{model}\times 1}$，$b_i\in\mathbb{R}$，$\sigma$为 sigmoid 函数，$y_i$为第 i 个标签的概率。

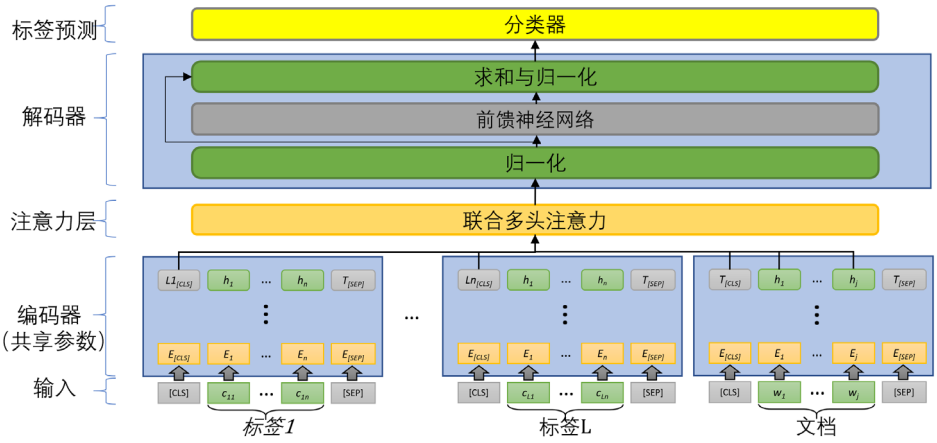

Fig. 1 Illustration of model structure

图 1 模型结构示意图

本文采用多标签交叉熵损失：

$$loss = -\sum_{i=1}^{L} y_i \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i) \quad (12)$$

其中，$\hat{y}_i \in [0,1]$，$y_i \in \{0,1\}$分别表示第 i 个标签的预测概率和真实标签。

3 实验与结果分析

我们在两个公开数据集上验证了所提出算法的有效性。

3.1 实验设置

3.1.1 数据集

数据集[4]：该数据集从论文预印网站 arXiv 上收集了 55840 篇论文摘要与其对应的多个主题。主题共 54 个，每个主题都有对应学科大类。

RCV1-V2 数据集[11]：该数据集是由路透社有限公司提供的 804414 条人工分类的新闻通讯报道和所对应的多个主题组成。主题共 103 个，并具有清晰的层次结构。

Table 1 Datasets introduction

表 1 数据集简介

数据集	训练集	测试集	标签数	平均每条数据标签数
AAPD	54840	1000	54	2.41
RCV1-V2	23149	781265	103	3.18

3.1.2 评价指标

我们选择最高 k 条精度 P@k（precision at k）作为性能比较的评价指标：

$$P@k = \frac{1}{k} \sum_{l \in rank_k(\hat{y})}^{k} y^l$$

$y \in \{0,1\}^L$是文档的真实标签向量，$\hat{y} \in [0,1]^L$为文档的预测得分向量，$rank_k(\hat{y})$是预测标签分数前 k 高的索引。我们评估使用的是所有文档 P@k 值的平均数。

3.1.3 基线模型

为充分证明本文提出模型的有效性，我们选择以下几个主流模型作为基线模型：

XML-CNN[23]：使用 CNN 和动态池化层提取高层文本特征的模型。

SGM[4]：将标签相关性建模为有序序列，使用序列生成方法进行预测。

DXML[24]：同时将特征空间和标签图结构建模的深度嵌入模型。

AttentionXML[18]：使用概率标签树和多标签注意力捕获信息词的基于标签树模型。

EXAM[25]：利用标签信息捕获词级别交互的算法。

LSAN[21]：使用自注意力和标签注意力获取标签特定的文档表示的模型。

LSTR[7]：使用图嵌入表示学习标签词共现信息的模型。

LDGN[8]：使用对偶图神经网络学习类别信息的算法。

3.1.4 参数设置

本文的模型使用”bert-base-uncased”预训练模型作为编码器，隐藏层维度d_{model}为 768，注意力头d为 8，编码器外的参数全部随机初始化。本文使用 AdamW[26]进行训练，初始学习率为 0.00002，批大小为 8。为了保证公平的比较，我们采用了和之前工作相同的数据集划分[21]，并采用早停机制，如果模型效果在验证集上 5000 步没有提升，则停止训练。

Table 2 Comparisons with state-of-the-art methods on AAPD dataset

表 2 在 AAPD 数据集上与先进方法进行比较

模型	P@1(%)	P@3(%)	P@5(%)
XML-CNN	74.38	53.84	37.79
SGM	75.67	56.75	35.65
DXML	80.54	56.30	39.16
AttXML	83.02	58.72	40.56
EXAM	83.26	59.77	40.66
LSAN	85.28	61.12	41.84
LTAR	85.03	61.46	41.80
LDGN	86.24	61.95	42.29
本文模型	**87.30**	**62.13**	**42.36**

Table 3 Comparisons with state-of-the-art methods on RCV1-V2 dataset

表 3 在 RCV1-V2 数据集上与先进方法进行比较

模型	P@1(%)	P@3(%)	P@5(%)
XML-CNN	95.75	78.63	54.94
SGM	95.37	81.36	53.06
DXML	94.04	78.65	54.38
AttXML	96.41	80.91	56.38
EXAM	93.67	75.80	52.73
LSAN	96.81	81.89	56.92
LTAR	97.31	83.11	57.85
LDGN	97.12	82.26	57.29
本文模型	**97.35**	**83.21**	**58.01**

3.2 实验结果

为方便比较，基线模型的结果直接引用前人的研究，LSTR 使用文献[7]的结果，其余基线模型采用文献[8]的实验结果。本文的模型运行 5 次，取五次结果的平均值。

表 2 和表 3 分类列出了在两个数据集上所有模型的效果，从实验结果可以看出本文提出的模型明显优于其他 8 种方法。观察到 XML-CNN 的性能比其他方法差很多，因为其只关注文本表示的提取，忽略了标签的相关性，而标签相关性已经被证明对多标签文本分类非常重要。基于标签树的方法 AttentionXML 优于 seq2seq 的方法（SGM）和深度嵌入的方法(DXML)。因为尽管 SGM 和 DXML 采用了有序序列和标签图的方法建模标签间的关系，但它们忽略了标签与文档的交互，而 AttentionXML 采用多标签注意力，可以提取到每个标签最关注的文档内容。相对于 AttentionXML 和 EXAM，LSAN 的表现，因为其同时考虑到文档本身的相关性以及文档与标签之间的相关性。但其在关注标签相关性时没有考虑到标签的语义信息，所以效果不如 LTAR 和 LDGN。LTAR 和 LDGN 都用图算法对标签间的语义关系进行建模，并同时注意到文档与标签的相关性，所以取得了之前最先进的效果。但 LTAR 和 LDGN 在不同阶段对文档-标签关系和标签-标签关系进行建模，没有利用到两者的交互信息，且标签与文档的表示提取不在同一语义空间。而本文提出的模型在注意力阶段同时对两者进行建模，且在同一语义空间提取文档和标签的表示，故我们的模型明显优于其他模型，在两个数据集上的三个指标都高于目前最先进模型。

3.3 消融实验分析

为验证本文提出的不同模块的有效性，我们做了一系列实验。（1）对于共享语义空间嵌入的编码器模块，NoShare 模型标签与文档的编码不共享参数，其他与本文模型一致；（2）对于本文提出了融合多头注意力机制，OnlyText 模型只关注文档本身的注意力；Label2Doc 模型只考虑标签对文档的语义相关性。这两个模型其余部分与本文模型一致；（3）对于本文提出的考虑标签层次结构的方法，NoHire 模型不加入标签层次感知标记，其余部分一致。表 4 列出了在 AAPD 数据集上的消融实验结果，NoShare 在三项指标上均明显低于本文模型，证明了共享语义空间编码器的有效性。OnlyText 表现最差，验证了只关注文档信息难以做好该任务。Label2Doc 的性能明显高于 OnlyText，说明了加入标签对文档关注的有效性，但仍与本文模型有较大差距，验证了融合注意力的重要性。NoHire 的性能低于本文模型，说明了基于先验知识的层次提示方法的有效性。

Table 3 Ablation results

表 3 消融实验结果

模型	P@1(%)	P@3(%)	P@5(%)
NoShare	86.20	61.75	41.78
OnlyText	85.98	61.35	41.56
Label2Doc	86.60	61.97	42.10
NoHire	87.15	62.05	42.21
本文模型	**87.30**	**62.13**	**42.36**

4 结论

本文提出了一种新的基于注意力机制和预训练

模型的多标签文本分类方法，将标签和文档之间的表示提取和相关性计算统一到一个框架中，并提出一种基于先验知识的层次提示方法利用标签层次信息，在 AAPD 和 RCV1-V2 数据集上验证了本文模型的有效性。

在未来的工作中，我们将考虑使用图神经网络等更有效的方法对标签的层次信息进行建模。此外，考虑如果将本文提出的模型应用到更复杂的任务场景中，如极端多标签文本分类任务。

作者贡献声明：作者一提出了算法思路和实验方案并完成实验和论文撰写，作者二辅助完成实验与论文撰写，作者三和作者四提出指导意见并修改论文。

参 考 文 献

[1] Gopal S, Yang Y. Multilabel classification with meta-level features[C]//Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 315-322.

[2] Kumar A, Irsoy O, Ondruska P, et al. Ask me anything: Dynamic memory networks for natural language processing[C]//International conference on machine learning. PMLR, 2016: 1378-1387.

[3] Cambria E, Olsher D, Rajagopal D. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis[C]//Twenty-eighth AAAI conference on artificial intelligence. 2014.

[4] Yang, Pengcheng, Xu Sun, et al. SGM: Sequence Generation Model for Multi-Label Classification. ArXiv:1806.04822 [Cs], June 15, 2018. http://arxiv.org/abs/1806.04822.

[5] Zhang X, Zhang Q W, Yan Z, et al. Enhancing Label Correlation Feedback in Multi-Label Text Classification via Multi-Task Learning[J]. arXiv preprint arXiv:2106.03103, 2021.

[6] Zhang Qian-Wen, Ximing Zhang, Zhao Yan, et al. Correlation-Guided Representation for Multi-Label Text Classification[J]. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 3363–69. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, 2021. https://doi.org/10.24963/ijcai.2021/463.

[7] Guo H, Li X, Zhang L, et al. Label-Aware Text Representation for Multi-Label Text Classification[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 7728-7732.

[8] Ma Q, Yuan C, Zhou W, et al. Label-Specific Dual Graph Neural Network for Multi-Label Text Classification[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3855-3864.

[9] Liu H, Caixia Y, and Xiaojie Wang. Label-Wise Document Pre-Training for Multi-Label Text Classification. ArXiv:2008.06695 [Cs], August 15, 2020. http://arxiv.org/abs/2008.06695.

[10] Liu H, Chen G, Li P, et al. Multi-label text classification via joint learning from label embedding and label correlation[J]. Neurocomputing, 2021, 460: 385-398.

[11] Lewis D D, Yang Y, Russell-Rose T, et al. Rcv1: A new benchmark collection for text categorization research[J]. Journal of machine learning research, 2004, 5(Apr): 361-397.

[12] Kim, Yoon. Convolutional Neural Networks for Sentence Classification. ArXiv:1408.5882 [Cs], September 2, 2014. http://arxiv.org/abs/1408.5882.

[13] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[J]. arXiv preprint arXiv:1605.05101, 2016

[14] Chen G, Ye D, Xing Z, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization[C]//2017 international joint conference on neural networks (IJCNN). IEEE, 2017: 2377-2383.

[15] Vaswani Ashish, Noam Shazeer, Niki Parmar, et al. Attention Is All You Need. In Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[16] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, et al.“BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” ArXiv:1810.04805 [Cs], May 24, 2019. http://arxiv.org/abs/1810.04805.

[17] Sun C, Qiu X, Xu Y, et al. How to Fine-Tune BERT for Text Classification?[C]// China National Conference on Chinese Computational Linguistics. Springer, Cham, 2019.

[18] You R, Zhang Z, Wang Z, et al. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification[J]. arXiv preprint arXiv:1811.01727, 2018.

[19] Wang G, Li C, Wang W, et al. Joint embedding of words and labels for text classification[J]. arXiv preprint arXiv:1805.04174, 2018.

[20] Pappas N, Henderson J. Gile: A generalized input-label embedding for text classification[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 139-155.

[21] Xiao L, Huang X, Chen B, et al. Label-specific document representation for multi-label text classification[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 466-475.

[22] 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. 2016: 521-526.

[23] 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. 2017: 115-124.

[24] Zhang W, Yan J, Wang X, et al. Deep extreme multi-label learning[C]//Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. 2018: 100-107.

[25] Du C, Chen Z, Feng F, et al. Explicit interaction model towards text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 6359-6366.

[26] Loshchilov I, Hutter F. Fixing weight decay regularization in adam[J]. 2018.

Sun Kun, born in 1997. Master candidate. His main research interests include text mining and knowledge graph.

孙坤，1997 年生，硕士研究生。主要研究方向为文本挖掘和知识图谱。

Qin Bowen, born in 1997. Master candidate. His main research interests include pre-trained language model and knowledge graph.

秦博文，1997 年生，硕士研究生。主要研究方向为预训练语言模型和知识图谱。

Sang Jitao, born in 1985. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include multimedia computing, network data mining, and trustworthy machine learning.

桑基韬，1985 年生，博士，教授，CCF 高级会员。主要研究方向为多媒体计算，网络数据挖掘和可信赖机器学习。

Yu Jian, born in 1969. PhD, professor, PhD supervisor. CCF Fellow. His main research interests include machine learning, data mining, and image segmentation.

于剑，1969 年生，博士，教授，CCF 会士。主要研究方向为机器学习，数据挖掘和图像分割。