Label-Specific Dual Graph Neural Network for Multi-Label Text Classification

Qianwen Ma[1,2], **Chunyuan Yuan**[1,2], **Wei Zhou**[1*] **and Songlin Hu**[1,2]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{maqianwen,yuanchunyuan,zhouwei,husonglin}@iie.ac.cn

Abstract

Multi-label text classification is one of the fundamental tasks in natural language processing. Previous studies have difficulties to distinguish similar labels well because they learn the same document representations for different labels, that is they do not explicitly extract label-specific semantic components from documents. Moreover, they do not fully explore the high-order interactions among these semantic components, which is very helpful to predict tail labels. In this paper, we propose a novel label-specific dual graph neural network (LDGN), which incorporates category information to learn label-specific components from documents, and employs dual Graph Convolution Network (GCN) to model complete and adaptive interactions among these components based on the statistical label co-occurrence and dynamic reconstruction graph in a joint way. Experimental results on three benchmark datasets demonstrate that LDGN significantly outperforms the state-of-the-art models, and also achieves better performance with respect to tail labels.

1 Introduction

Automatically labeling multiple labels of documents is a fundamental and practical task in natural language processing. Recently, with the growth of data scale, multi-label text classification(MLTC) has attracted more attention, since it is often applied to many fields such as sentiment analysis (Liu and Chen, 2015; Li et al., 2016), emotion recognition (Wang et al., 2016; Jabreel and Moreno, 2019), web page tagging (Jain et al., 2016) and so on. However, the number of labels and documents and the complex relations of labels render it an unsolved and challenging task.

Existing studies for multi-label text classification mainly focus on learning enhanced document representation (Liu et al., 2017) and modeling label dependency (Zhang et al., 2018; Yang et al., 2018; Tsai and Lee, 2019) to improve classification performance. Although they have explored the informative words in text content, or considered the label structure and label semantics to capture label correlations, these models cannot distinguish similar labels well (e.g., the categories *Prices* vs *Consumer Prices* in Reuters News).

The main reason is that most of them neglect the semantic connections between labels and input documents and they learn the same document representations for different labels, which cannot issue the label similarity problem. More specifically, they do not explicitly consider the corresponding semantic parts of each label in the document.

Recently, some studies (You et al., 2019; Xiao et al., 2019; Du et al., 2019) have used attention mechanism to explore the above semantic connections, and learn a label-specific document representation for classification. These methods have obtained promising results in MLTC, which shows the importance of exploring semantic connections. However, they did not further study the interactions between label-specific semantic components which can be guided by label correlations, and thus these models cannot work well on predicting tail labels which is also a challenging issue in MLTC. To handle these issues, a common way to explore the semantic interactions between label-specific parts in document is to utilize the statistical correlations between categories to build a label co-occurrence graph for guiding interactions.

Nevertheless, statistical correlations have three drawbacks. First, the co-occurrence patterns between label pairs obtained from training data are incomplete and noisy. Specifically, the label co-occurrences that appear in the test set but do not appear in the training set may be ignored, while

*Corresponding Author

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3855–3864
August 1–6, 2021.

some rare label co-occurrences in the statistical correlations may be noise. Second, the label co-occurrence graph is built in global, which may be biased for rare label correlations. And thus they are not flexible to every sample document. Third, statistical label correlations may form a long-tail distribution, i.e., some categories are very common while most categories have few of documents. This phenomenon may lead to models failing to predict low-frequency labels. Thus, our goal is to find a way to explore the complete and adaptive interactions among label-specific semantic components more accurately.

In this paper, we investigate: (1) how to explicitly extract the semantic components related to the corresponding labels from each document; and (2) how to accurately capture the more complete and more adaptive interactions between label-specific semantic components according to label dependencies. To solve the first challenge, we exploit the attention mechanism to extract the label-specific semantic components from the text content, which can alleviate the label similar problem. To capture the more accurate high-order interactions between these semantic components, we first employ one Graph Convolution Network (GCN) to learn component representations using the statistical label co-occurrence to guide the information propagation among nodes (components) in GCN. Then, we use the component representations to reconstruct the adjacency graph dynamically and re-learn the component representations with another GCN, and thus we can capture the latent interactions between these semantic components. Finally, we exploit final component representations to predict labels. We evaluate our model on three real-world datasets, and the results show that the proposed model LDGN outperforms all the comparison methods. Further studies demonstrate our ability to effectively alleviate the tail labels problem, and accurately capture the meaningful interactions between label-specific semantic components.

The contributions of this paper are as follows:

- We propose a novel label-specific dual graph neural network (LDGN), which incorporates category information to extract label-specific components from documents, and explores the interactions among these components.
- To model the accurate and adaptive interactions, we jointly exploit global co-occurrence patterns and local dynamic relations. To make up the deficiency of co-occurrences, we employ the local reconstruction graph which is built by every document dynamically.
- We conduct a series of experiments on three public datasets, and experimental results demonstrate that our model LDGN significantly outperforms the state-of-the-art models, and also achieves better performance with respect to tail labels.

2 Model

As depicted in Figure 1, our model LDGN is composed of two major modules: 1) label-specific document representation 2) dual graph neural network for semantic interaction learning. Specifically, label-specific document representation learning describes how to extract label-specific semantic components from the mixture of label information in each document; and the dual graph neural network for semantic interaction learning illustrates how to accurately explore the complete interactions among these semantic components under the guidance of the prior knowledge of statistical label co-occurrence and the posterior information of dynamic reconstruction graph.

Problem Formulation: Let $\mathcal{D} = \{x_i, y_i\}^N$ be the set of documents, which consists of N document x_i and its corresponding label $y_i \in \{0,1\}^{|C|}$, where $|C|$ denotes the total number of labels. Each document x_i contains J words $x_i = w_{i1}, w_{i2}, \ldots, w_{iJ}$. The target of multi-label text classification is to learn the mapping from input text sequence to the most relevant labels.

2.1 Label-specific Document Representation

Given a document x with J words, we first embed each word w_j in the text into a word vector $\mathbf{e_{wj}} \in \mathcal{R}^d$, where d is the dimensionality of word embedding vector. To capture contextual information from two directions of the word sequence, we first use a bidirectional LSTM to encode word-level semantic information in document representation. And we concatenate the forward and backward hidden states to obtain the final word sequence vector $\mathbf{h} \in R^{|J| \times D}$.

After that, to explicitly extract the corresponding semantic component related to each label from documents, we use a label guided attention mechanism to learn label-specific text representation.

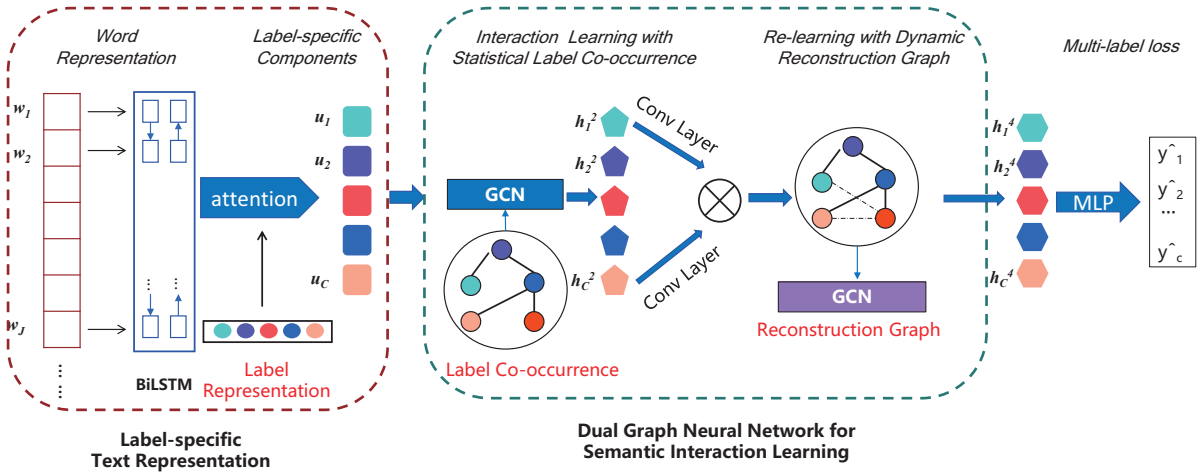


Figure 1: The architecture of the proposed network LDGN.

Firstly, we randomly initialize the label representation $\mathbf{C} \in R^{|C| \times d_c}$, and compute the label-aware attention values. Then, we can induce the label-specific semantic components based on the label guided attention. The formula is as follows:

$$\alpha_{ij} = \frac{\exp\left(\mathbf{h_j c_i^T}\right)}{\sum_j \exp\left(\mathbf{h_j c_i^T}\right)}, \quad (1)$$

$$\mathbf{u}_i = \sum_j \alpha_{ij} \mathbf{h_j}, \quad (2)$$

where α_{ij} indicates how informative the j-th text feature vector is for the i-th label. $\mathbf{u_i} \in R^D$ denotes the semantic component related to the label $\mathbf{c_i}$ in this document.

2.2 Dual Graph Neural Network

Interaction Learning with Statistical Label Co-occurrence To capture the mutual interactions between the label-specific semantic components, we build a label graph based on the prior knowledge of label co-occurrence, each node in which correlates to a label-specific semantic component $\mathbf{u_i}$. And then we apply a graph neural network to propagate message between nodes.

Formally, we define the label graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes refer to the categories and edges refer to the statistical co-occurrence between nodes (categories). Specifically, we compute the probability between all label pairs in the training set and get the matrix $\mathbf{A}^s \in R^{|C| \times |C|}$, where $\mathbf{A}^s_{ij}$ denotes the conditional probability of a sample belonging to category C_i when it belongs to category C_j.

Then, we utilize GCN (Kipf and Welling, 2017) to learn the deep relationships between label-specific semantic components guided by the statistical label correlations. GCNs are neural networks operating on graphs, which are capable of enhancing node representations by propagating messages between neighboring nodes.

In multi-layer GCN, each GCN layer takes the component representations from previous layer $\mathbf{H}^l$ as inputs and outputs enhanced component representations, i.e., $\mathbf{H}^{l+1}$. The layer-wise propagation rule is as follows:

$$\mathbf{H}^{l+1} = \sigma\left(\widehat{\mathbf{A}}^s \mathbf{H}^l \mathbf{W}^l\right), \quad (3)$$

where $\sigma(\cdot)$ denotes LeakyReLU (Maas et al., 2013) activation function. $\mathbf{W}^l \in R^{D \times D'}$ is a transformation matrix to be learned. $\widehat{\mathbf{A}}$ denotes the normalized adjacency matrix, and the normalization method (Kipf and Welling, 2017) is:

$$\widehat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}, \quad (4)$$

where $\mathbf{D}$ is a diagonal degree matrix with entries $D_{ij} = \Sigma_j \mathbf{A_{ij}}$

Depending on how many convolutional layers are used, GCN can aggregate information only about immediate neighbors (with one convolutional layer) or any nodes at most K-hops neighbors (if K layers are stacked). See (Kipf and Welling, 2017) for more details about GCN.

We use a two-layer GCN to learn the interactions between label-specific components. The first layer takes the initialized component representations $\mathbf{U} \in R^{|C| \times D}$ in Equation 2 as inputs $\mathbf{H^0}$; and the last layer outputs $\mathbf{H^2} \in R^{|C| \times D'}$ with D' denoting the dimensionality of final node representations.

However, the statistical label correlations obtained by training data are incomplete and noisy.

And the co-occurrence patterns between label pairs may form a long-tail distribution.

Re-learning with Dynamic Reconstruction Graph To capture the more complete and adaptive interactions between these components, we exploit the above component representations $\mathbf{H}^2$ to reconstruct the adjacency graph dynamically, which can make up the deficiency of co-occurrence matrix. And then we re-learn the interactions among the label-specific components guided by the posterior information of dynamic reconstruction graph.

Specifically, we apply two 1×1 convolution layers and dot product to get the dynamic reconstruction graph $\mathbf{A}^D$ as follows:

$$\mathbf{A}^D = f\left(\left(\mathbf{W}_a * \mathbf{H}^2\right)^T \left(\mathbf{W}_b * \mathbf{H}^2\right)\right), \quad (5)$$

where $\mathbf{W}_a \in R^{d_1\times D'}$ and $\mathbf{W}_b \in R^{d_1\times D'}$ are the weights of two convolution layers, f is the $sigmoid$ activation function. And then we normalize the reconstruction adjacency matrix as Equation 4, and obtain the normalized adjacency matrix $\widehat{\mathbf{A}}^D$ of reconstruction graph.

In a similar way as Equation 3, we apply another 2-layer GCN to learn the deep correlations between components with the dynamic reconstruction graph. The first layer of this GCN takes the component representations $\mathbf{H}^2$ as inputs, and the last layer outputs the final component representations $\mathbf{H}^4 \in R^{|C|\times D'}$.

2.3 Multi-label Text Classification

After the above procedures, we concatenate the two types of component representations $\mathbf{H^O} = [\mathbf{H^2}, \mathbf{H^4}]$ and feed it into a fully connected layer for prediction: $\widehat{y} = \sigma(\mathbf{W}_1\mathbf{H}^O)$, where $\mathbf{W}_1 \in R^{2D'\times1}$ and σ is the sigmoid function.

We use $y \in \mathcal{R}^{|C|}$ to represent the ground-truth label of a document, where $y_i = 0, 1$ denotes whether label i appears in the document or not. The proposed model LDGN is trained with the multi-label cross entropy loss:

$$\mathcal{L} = \sum_{c=1}^{C} y^c \log\left(\widehat{y}^c\right) + \left(1-y^c\right)\log\left(1-\widehat{y}^c\right). \quad (6)$$

3 Experiment

3.1 Experimental Setup

Datasets We evaluate the proposed model on three benchmark multi-label text classification datasets, which are AAPD (Yang et al., 2018), EUR-Lex (Mencia and Fürnkranz, 2008) and RCV1 (Lewis et al., 2004). The statistics of these three datasets are listed in Table 1.

Dataset	N_{train}	N_{test}	L	$\overline{L}$	$\overline{W}$
RCV1	23,149	781,265	101	3.18	259.47
AAPD	54,840	1,000	54	2.41	163.42
EUR-Lex	11,585	3,865	3,954	5.32	1225.2

Table 1: Statistics of the datasets. N_{train} and N_{test} denote the number of training and testing samples respectively. L is the total number of classes, $\overline{L}$ is the average number of labels per sample and $\overline{W}$ is the average number of words per sample.

Evaluation Metric Following the settings of previous work (You et al., 2019; Xiao et al., 2019), we use precision at top K (P@k) and Normalized Discounted Cumulated Gains at top K (nDCG@k) for performance evaluation. The definition of two metrics can be referred to You et al. (2019).

Implementation Details For a fair comparison, we apply the same dataset split as previous work (Xiao et al., 2019), which is also the original split provided by dataset publisher (Yang et al., 2018; Mencia and Fürnkranz, 2008).

The word embeddings in the proposed network are initialized with the 300-dimensional word vectors, which are trained on the datasets by Skip-gram (Mikolov et al., 2013) algorithm. The hidden sizes of Bi-LSTM and GCNs are set to 300 and 512, respectively. We use the Adam optimization method (Kingma and Ba, 2014) to minimize the cross-entropy loss, the learning rate is initialized to 1e-3 and gradually decreased during the process of training. We select the best parameter configuration based on performance on the validation set and evaluate the configuration on the test set. Our code is available on GitHub[1].

3.2 Baselines

We compare the proposed model with recent deep learning based methods for MLTC, including seq2seq models, deep embedding models, and label attention based models. And it should be noted that, because of different application scenarios, we did not choose the label tree-based methods and extreme text focused methods as baseline models.

• XML-CNN (Liu et al., 2017): a CNN-based

[1]https://github.com/Makwen1995/LDGN_MLTC

Models	AAPD					EUR-Lex				
	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
XML-CNN	74.38	53.84	37.79	71.12	75.93	70.40	54.98	44.86	58.62	53.10
SGM	75.67	56.75	35.65	72.36	75.35	70.45	60.37	43.88	60.72	55.24
DXML	80.54	56.30	39.16	77.23	80.99	75.63	60.13	48.65	63.96	53.60
AttentionXML	83.02	58.72	40.56	78.01	82.31	67.34	52.52	47.72	56.21	50.78
EXAM	83.26	59.77	40.66	79.10	82.79	74.40	61.93	50.98	65.12	59.43
LSAN	85.28	61.12	41.84	80.84	84.78	79.17	64.99	53.67	68.32	62.47
LDGN	**86.24**	**61.95**	**42.29**	**83.32**	**86.85**	**81.03**	**67.79**	**56.36**	**71.81**	**66.09**

Table 2: Comparisons with state-of-the-art methods on both AAPD and EUR-Lex datasets. The experimental results of all baseline models are directly cited from paper (Xiao et al., 2019).

model which uses CNN and a dynamic pooling layer to extract high-level feature for MLTC.

- SGM (Yang et al., 2018): a sequence generation model which models label correlations as an ordered sequence.
- DXML (Zhang et al., 2018): a deep embedding method which models the feature space and label graph structure simultaneously.
- AttentionXML (You et al., 2019): a label tree-based deep learning model which uses a probabilistic label tree and multi-label attention to capture informative words in extreme-scale data.
- EXAM (Du et al., 2019): a novel framework that leverages the label information to compute the word-level interactions.
- LSAN (Xiao et al., 2019): a label-specific attention network model based on self-attention and label-attention mechanism.

The SotA model (i.e., LSAN) used BiLSTM model for text representations. For a fair comparison, we also take BiLSTM as text encoder in our model.

3.3 Experimental Results and Analysis

Table 2 and Table 3 demonstrate the performance of all the compared methods based on the three datasets. For fair comparison, the experimental results of baseline models are directly cited from previous studies (Xiao et al., 2019). We also bold the best result of each column in all tables.

From the Table 2 and Table 3, we can observe that the proposed LDGN outperforms all other baselines on three datasets. The outstanding results confirm the effectiveness of label-specific semantic interaction learning with dual graph neural network, which include global statistical patterns and local dynamic relations.

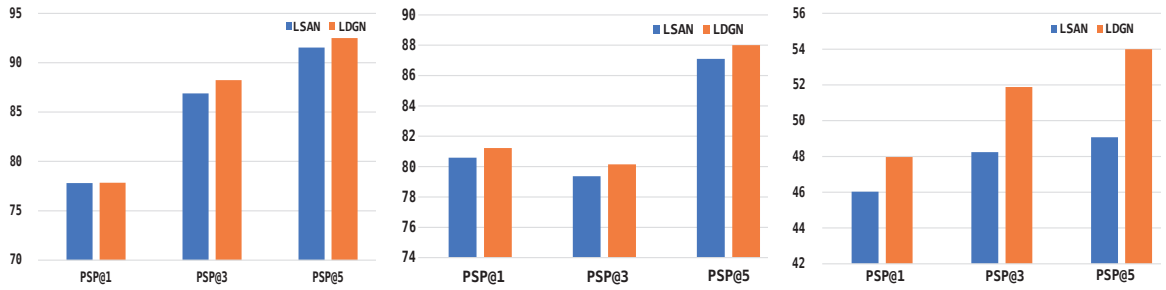
Models	RCV1				
	P@1	P@3	P@5	N@3	N@5
XML-CNN	95.75	78.63	54.94	89.89	90.77
SGM	95.37	81.36	53.06	91.76	90.69
DXML	94.04	78.65	54.38	89.83	90.21
AttentionXML	96.41	80.91	56.38	91.88	92.70
EXAM	93.67	75.80	52.73	86.85	87.71
LSAN	96.81	81.89	56.92	92.83	93.43
LDGN	**97.12**	**82.26**	**57.29**	**93.80**	**95.03**

Table 3: Comparisons with state-of-the-art methods on the RCV1 dataset. The experimental results of baselines are directly cited from (Xiao et al., 2019).

It is observed that the performance of **XML-CNN** is worse than other comparison methods. The reason is that it only exploits the text content of documents for classification but ignores the label correlations which have been proven very important for multi-label classification problem.

The label tree-based model **AttentionXML** performs better than the seq2seq method (**SGM**) and the deep embedding method (**DXML**). Although both **DXML** and **SGM** employ a label graph or an ordered sequence to model the relationship between labels, they ignore the interactions between labels and document content. And **AttentionXML** uses multi-label attention which can focus on the most relevant parts in content and extract different semantic information for each label.

Compared with other label attention based



(a) RCV1 (b) AAPD (c) EUR-Lex

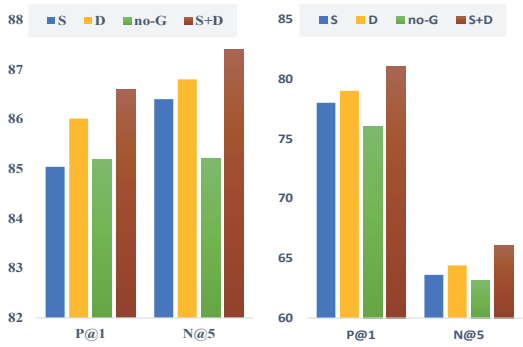
Figure 2: Performance on tail labels.

methods (**AttentionXML, EXAM**), **LSAN** performs the best because it takes the semantic correlations between document content and label text into account simultaneously, which exploits an adaptive fusion to integrate self-attention and label-attention mechanisms to learn the label-specific document representation.

In conclusion, the proposed network **LDGN** outperforms sequence-to-sequence models, deep embedding models, and label attention based models, and the metrics $P@k$ and $nDCG@k$ of multi-label text classification obtain significant improvement. Specifically, on the AAPD dataset, **LDGN** increases the $P@1$ of the **LSAN** method (the best baseline) from 85.28% to 86.24%, and increases $nDCG@3$ and $nDCG@5$ from 80.84% to 83.33%, 84.78% to 86.85% , respectively. On the EUR-Lex dataset, the metric $P@1$ is boosted from 79.17% to 81.03%, and $P@5$ and $nDCG@5$ are increased from 53.67% to 56.36%, 62.47% to 66.09%, respectively. On the RCV1 dataset, the $P@k$ of our model increased by 0.3% at average, and **LDGN** achieves 1% and 1.6% absolute improvement on $nDCG@3, 5$ compared with **LSAN**. The improvements of the proposed LDGN model demonstrate that the semantic interaction learning with joint global statistical relations and local dynamic relations are generally helpful and effective, and **LDGN** can capture the deeper correlations between categories than **LSAN**.

3.4 Ablation Test

We perform a series of ablation experiments to examine the relative contributions of dual graph-based semantic interactions module. To this end, **LDGN** is compared with its three variants:(1)**S**: Graph-based semantic interactions only with statistical label co-occurrence; (2)**D**: Graph-based semantic interactions only with dynamic reconstruction graph; (3)**no-G**:Removing the dual graph



(a) AAPD (b) EUR-Lex

Figure 3: Ablation test of LDGN on two datasets.

neural network. For a fair comparison, both **S** and **D** use 4-layer GCN which is the same as **LDGN**.

As presented in Figure 3, **S** and **D** perform better than **no-G**, which demonstrates that exploring either statistical relations or dynamic relations can correctly capture the effective semantic interactions between label-specific components. **D** performs better than **S**, indicating the model with local dynamic relations is adaptive to data and has better stability and robustness, which also shows that the model with local dynamic relations can capture semantic dependencies more effectively and accurately. The performance of **S+D** (i.e., **LDGN**) combining two aspect relations obtains significant improvement, which shows dynamic relations can make up the deficiency of statistical co-occurrence and correct the bias of global correlations. Thus, it is necessary to explore their joint effects to further boost the performance.

3.5 Performance on tail labels

In order to prove the effectiveness of the proposed **LDGN** in alleviating the tail labels problem, we evaluate the performance of **LDGN** by propensity scored precision at k (PSP@k), which is calcu-

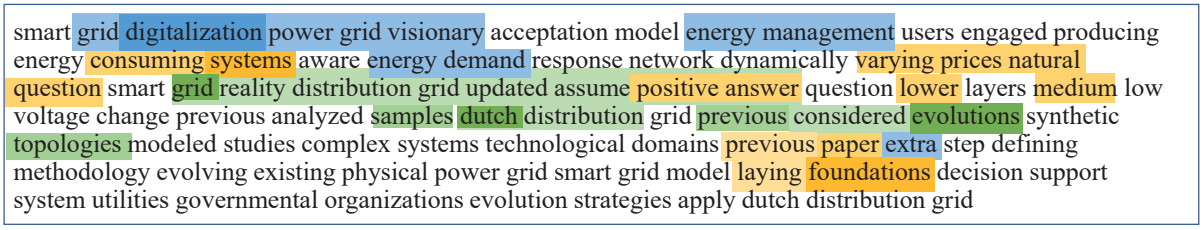


Figure 4: The Visualization of label attention weights. (The attention weights of 'physics.soc' for words are shaded in blue, and the attention scores of class *CS.CY* and *CS.CE* are shaded in green and yellow color respectively. Darker color represents higher weight score.)

lated as follow:

$$PSP@k = \frac{1}{k}\sum_{l=1}^{k}\frac{y_{rank(l)}}{P_{rank(l)}}, \qquad (7)$$

where $P_{rank(l)}$ is the propensity score (Jain et al., 2016) of label rank(l). Figure 2 shows the results of LDGN and LSAN on three datasets.

As shown in Figure 2(a), Figure 2(b) and Figure 2(c), the proposed **LDGN** performs better in predicting tail labels than the **LSAN** model (the best baseline) on three datasets. Specifically, on the RCV1 dataset, **LDGN** achieves 0.97% and 1.35% absolute improvement in term of $PSP@3$ and $PSP@5$ compared with **LSAN**. On the AAPD dataset, the $PSP@k$ increased by at least 0.63% up to 0.90%. And on the EUR-Lex dataset, **LDGN** achieves 1.94%, 3.64% and 4.93% absolute improvement on $PSP@1, 3, 5$ compared with **LSAN**. The reason for the improvement in the EUR-Lex dataset is more obvious is that the semantic interactions learning is more useful to capture related information in the case of a large number of labels.

The results prove that **LDGN** can effectively alleviate the problem of predicting tail labels.

3.6 Case Study

To further verify the effectiveness of our label attention module and dual graph neural network in LDGN, we present a typical case and visualize the attention weights on the document words and the similarity scores between label-specific components. We show a test sample from original AAPD dataset, and the document belongs to three categories, *'Physics and Society'* (**physics.soc**), *'Computers and Society'* (**cs.cy**) and *'Computational Engineering, Finance, and Science'* (**cs.ce**).

Visualization of Attention We can observe from the Figure 4 that different labels focus on different parts in the document text, and each label has its own concerned words. For example,

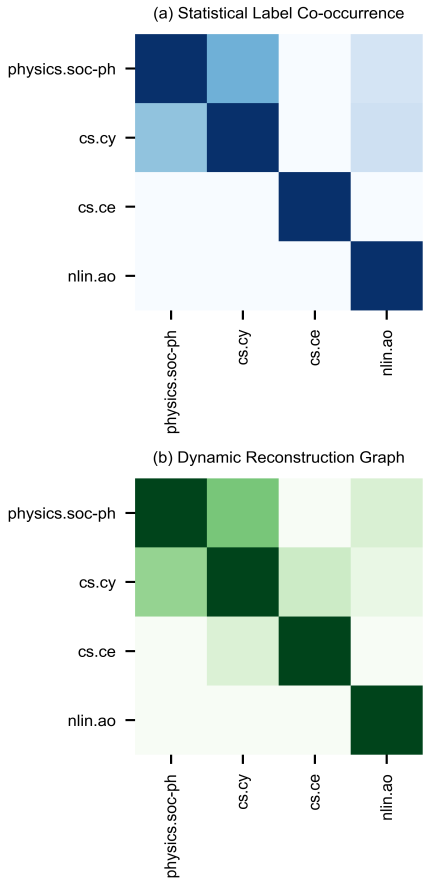


Figure 5: The Visualization of two adjacency matrices of dual GNN. Darker color represents higher weight.

the more important parts in the *'physics.soc'* category are 'digitalization power grid', 'energy management'. And the words that the *'cs.ce'* category focuses on are 'consuming systems', 'varying prices', 'laying foundations', 'lower ' and etc. For class *'cs.cy'*, the concerned words are 'samples dutch distribution', 'evolutions' and 'topologies'. The corresponding related words of the three categories can reflect the semantics of the categories.

Visualization of Interactions To gain a clearer view of the importance of our dual graph-based interactions learning module, we display two

heatmaps in Figure 5 to visualize the partial graph structure of dual GCN. The edge weights shown in the heatmaps are obtained by global label co-occurrence and local dynamic relations (i.e., computed by Equation 5), respectively.

As presented in heatmaps, different relations between categories are captured by dual GCN. In global statistical relations, *'cs.cy'* is highly linked with *'physics.soc'* and wrong label *'nlin.ao'*, while the true label *'cs.ce'* is isolated. And in local dynamic relations, *'cs.cy'* is more related to *'cs.ce'*, and the correlations between wrong label *'nlin.ao'* and true labels are reduced. This demonstrates that local dynamic relations can capture the latent relations that do not appear in global relations, and correct the bias of global correlations.

4 Related Work

Multi-label Text Classification The existing methods for MLTC mainly focus on learning enhanced document representation (Liu et al., 2017) and modeling label dependency (Nam et al., 2017; Yang et al., 2018; Tsai and Lee, 2019) to improve the classification performance.

With the wide application of neural network methods for text representation, some innovative models have been developed for this task, which include traditional deep learning methods and Seq2Seq based methods. Liu et al. (2017) employed CNNs and dynamic pooling to learn the text representation for MLTC. However, they treated all words equally and cannot explored the informative words in documents. The Seq2Seq methods, such as MLC2Seq (Nam et al., 2017) and SGM (Yang et al., 2018), employed a RNN to encode the input text and an attention based RNN decoder to generate predicted labels sequentially. Although they used attention mechanism to capture the informative words in text content, these models cannot distinguish similar labels well. There is a big reason that most of them neglect the semantic connections between labels and document, and learn the same document representations for different labels.

Recently, some studies (You et al., 2019; Xiao et al., 2019; Du et al., 2019) have used attention mechanism to explore the interactions between words and labels, and learned a label-specific document representation for classification. These methods have obtained promising results in MLTC, which shows the importance of exploring semantic connections. However, they did not further study the interactions between label-specific semantic components which can help to predict low-frequency labels.

To handle these issues, a common way to explore the semantic interactions between label-specific parts in document, is to utilize the label graph based on statistical co-occurrences.

MLC with Label Graph In order to capture the deep correlations of labels in a graph structure, many researches in image classification apply node embedding and graph neural network models to the task of multi-label image classification. Lee et al. (2018) incorporated knowledge graphs for describing the relationships between labels, and the information propagation can model the dependencies between seen and unseen labels for multi-label zero-shot learning. Chen et al. (2019) learned label representations with prior label correlation matrix in GCN, and mapped the label representations to inter-dependent classifiers, which achieved superior performance.

However, there were few related approaches for multi-label classification of text. Zhang et al. (2018) established an explicit label co-occurrence graph to explore label embedding in low-dimension latent space.

Furthermore, the statistical label correlations obtained by training data are incomplete and noisy. And the co-occurrence patterns between label pairs may form a long-tail distribution.

Thus, our goal is to find a way to explore the complete and adaptive interactions among label-specific semantic components more accurately.

5 Conclusion

In this paper, we propose a graph-based network LDGN to capture the semantic interactions related to corresponding labels, which jointly exploits global statistical patterns and local dynamic relations to derive complete and adaptive dependencies between different label-specific semantic parts. We first exploit multi-label attention to extract the label-specific semantic components from documents. Then, we employ GCN to learn component representations using label co-occurrences to guide the information propagation among components. After that, we use the learned component representations to compute the adjacency graph dynamically and re-learn with GCN based on the reconstruction graph. Extensive experiments con-

ducted on three public datasets show that the proposed LDGN model outperforms other state-of-the-art models on multi-label text classification task and also demonstrates much higher effectiveness to alleviate the tail label problem. In the future, we will improve the proposed model in efficiency, for example we could construct a dynamic graph for a few samples rather than each sample. And besides, we will explore more information about labels for MLC classification.

Acknowledgement

We gratefully thank the anonymous reviewers for their insightful comments. This research is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDC02060400.

References

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186.

Cunxiao Du, Zhaozheng Chen, Fuli Feng, Lei Zhu, Tian Gan, and Liqiang Nie. 2019. Explicit interaction model towards text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6359–6366.

Mohammed Jabreel and Antonio Moreno. 2019. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9(6):1123.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 935–944.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1576–1585.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Xin Li, Haoran Xie, Yanghui Rao, Yanjia Chen, Xuebo Liu, Huan Huang, and Fu Lee Wang. 2016. Weighted multi-label classification model for sentiment analysis of online news. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 215–222. IEEE.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124.

Shuhua Monica Liu and Jiun-Hung Chen. 2015. A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, 42(3):1083–1093.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3.

Eneldo Loza Mencia and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *Advances in neural information processing systems*, pages 5413–5423.

Che-Ping Tsai and Hung-Yi Lee. 2019. Order-free learning alleviating exposure bias in multi-label classification. *arXiv preprint arXiv:1909.03434*.

Yaqi Wang, Shi Feng, Daling Wang, Ge Yu, and Yifei Zhang. 2016. Multi-label chinese microblog emotion classification via convolutional neural network. In *Asia-Pacific Web Conference*, pages 567–580. Springer.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 466–475.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: sequence generation model for multi-label classification. *arXiv preprint arXiv:1806.04822*.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems*, pages 5820–5830.

Wenjie Zhang, Junchi Yan, Xiangfeng Wang, and Hongyuan Zha. 2018. Deep extreme multi-label learning. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 100–107.