LA-HCN: Label-based Attention for Hierarchical Multi-label Text Classification Neural Network

Xinyi Zhang  
Nanyang Technological University  
xinyi001@e.ntu.edu.sg

Jiahao Xu  
Nanyang Technological University  
jiahao004@e.ntu.edu.sg

Soh, Charlie Zhan Yi  
Nanyang Technological University  
charliesoh@ntu.edu.sg

Lihui Chen  
Nanyang Technological University  
elhchen@ntu.edu.sg

Abstract

Hierarchical multi-label text classification (HMTC) problems become popular recently because of its practicality. Most existing algorithms for HMTC focus on the design of classifiers, and are largely referred to as local, global, or a combination of local/global approaches. However, a few studies have started exploring hierarchical feature extraction based on the label hierarchy associating with text in HMTC. In this paper, a Neural network-based method called LA-HCN is proposed where a novel Label-based Attention module is designed to hierarchically extract important information from the text based on different labels. Besides, local and global document embeddings are separately generated to support the respective local and global classifications. In our experiments, LA-HCN achieves the top performance on the four public HMTC datasets when compared with other neural network-based state-of-the-art algorithms. The comparison between LA-HCN with its variants also demonstrates the effectiveness of the proposed label-based attention module as well as the use of the combination of local and global classifications. By visualizing the learned attention(words), we find LA-HCN is able to extract meaningful but different information from text based on different labels which is helpful for human understanding and explanation of classification results.

1 Introduction

In recent years, there has been a growing interest in hierarchical multi-label classification (HMC) which can be applied in a wide range of applications such as IPC (International Patent Classification) (Gomez and Moens, 2014), product annotation (Aly et al., 2019) and advertising recommendation (Agrawal et al., 2013). Different from normal flat classification problem where each input sample is only associated with a single label from a set of disjoint labels, the labels in HMC problem are organized in the form of a tree or a Directed Acyclic Graph (DAG) (Silla and Freitas, 2011) and each input sample is usually associated with multiple labels which is a more challenging problem.

The most straightforward approach is to convert HMC to a flat multi-label classification problem by simply ignoring the relevance between labels (Li et al., 2018; Hu et al., 2018; Aly et al., 2019). The disadvantages of flat approach is that the hierarchical information of the datasets is dropped. Another common type, local approach (Koller and Sahami, 1997), is designed to do multi-label classification at different particular regions of the label hierarchy (LCN, LCPN, LCL) and then generate the classification results based on these local predictions. Hierarchical information can be better explored in local approaches while misclassification will be easily propagated to the next level in these approaches (Punera and Ghosh, 2008). Global approaches are proposed to learn a single global model for all labels to reduce the model size and consider the entire label hierarchy at once (Kiritchenko et al., 2005, 2006). A typical category of global classifiers are developed based on flat classifiers and some modifications are made to integrate the hierarchical information of labels (Wang et al., 2009; Vens et al., 2008) into the model. Recently, more algorithms which combine the approaches mentioned above are proposed (Wehrmann et al., 2018; Mao et al., 2019).

However, all algorithms introduced above only focus on the design of hierarchical classifier while
ignoring the hierarchical feature extraction which is also important in HMC. In this work, a HMTC model with a label-based attention module is proposed for text classification. Different from Huang et al. (2019); Rojas et al. (2020) where hierarchical feature extraction is realized by applying general attention over the whole text, LA-HCN is designed to extract key information based on different labels at different hierarchical levels. Comparing with normal attention, label-based attention is more helpful for human understanding on the classification results which makes the model more explainable and interpretable. Specifically, a component-word relevance module is firstly applied at the current level to extract common important information, then each component is associated with a label at this level so that we can get the label-based word attention which will be used to generate hierarchical document embeddings for both local and global classifications. During the procedure of hierarchical feature extraction, each level will refer to features generated from its previous level (except the first level) so that the hierarchical information is fully utilized.

**Contribution** Main contributions of this work:

- Proposed a novel HMTC model which is capable of separately extracting features at different levels. During the procedure of feature extraction at the current level, the information extracted from the previous level is also applied so that the information is well inherited. Besides, both local and global classifiers are applied to reduce the effect of error-propagation between levels.

- Proposed a novel module to learn label-based word attention at each level so that the important information of each document can be captured based on individual labels which is more helpful for human understanding.

- We evaluate LA-HCN against both classical and state-of-the-art neural network based HMC algorithms on 4 datasets and LA-HCN achieves top 1 on all the datasets. The ablation study shows the effectiveness of different classifiers applied in LA-HCN and the learned label-based attention is able to give reasonable interpretation of the prediction results.

2 **Algorithm**

**Problem Definition** Given a set of documents \( \mathcal{X} = \{X_1, X_2, \ldots, X_K\} \) as well as their corresponding labels \( \mathcal{Y} = \{Y_1, Y_2, \ldots, Y_K\} \) where each document is represented with a sequence of words \( X_i = \{w_1, w_2, \ldots, w_N\} \) and \( Y_i = \{l_1, l_2, \ldots\} \) contains all hierarchical labels associated with the document \( X_i \). The target of this work is to learn a mapping \( \mathcal{P} : \mathcal{X} \rightarrow \mathcal{L} \) to predict the corresponding labels for each document by analyzing the document content as well as the hierarchical structure information of labels.

Here, the label hierarchy is represented as \((\mathcal{L}, \leq_h)\), where \( \mathcal{L} \) denotes the set of labels and \( \leq_h \) denotes a partial order (which is a tree in this work) representing the superclass relationship: \( \forall l_1, l_2 \in \mathcal{L} : l_1 \leq_h l_2 \) if and only if \( l_1 \) is a superclass of \( l_2 \). In the hierarchical classification problem, \( l_i \in Y_i \Rightarrow \exists l_j \leq_h l_i : l_j \in Y_i \). Besides, we also define the level of a label as the number of its direct or indirect superclasses (we set a virtual label root as the root of all label hierarchy) and the level of a label \( l_i \) is represented as \( \psi(l_i) = h \).

**Framework** The overall structure of LA-HCN is shown in Fig. 1. For each hierarchical level, a label-based attention module is applied to generate both local and global document embeddings for local and global classification respectively. The details of the 1) Label-based attention module as well as the 2) Classification module will be introduced in the following section.

2.1 **Label-based Attention Module**

This module is designed for extracting both local and global document features based on different labels at each hierarchical level. As introduced in Section 1, although learning label-based attention is more meaningful for human understanding while there are still some unsolved issues:

- Huge number of labels: due to the large number of labels in some datasets, separately building the attention module for each label is space consuming. Besides, some labels are only associated with very few input training samples. It is hard to apply a small number of samples to learn quite meaningful attention parameters for each label which will increase the difficulty of label-based attention learning.

- Label relevance: There are relevance among sibling labels and some information might be
Figure 1: The overall structure of the proposed method. Given the input document $X_i$, a text encoder is applied to get word embeddings ($I_{X_i}$) which are used at all hierarchical levels. At the level $h$, a Label-based Attention Module is applied to generate the processed label-based document embeddings ($\tilde{D}^h_{X_i}$) and the global document embedding ($v^h_{X_i}$) based on $I_{X_i}$ together with the local document embedding ($\tilde{v}^h_{X_i-1}$) from the previous level. Then $\tilde{D}^h_{X_i}$ is applied to do local classification. Global document embeddings ($v^h_{X_i}$) extracted from each level are concatenated to do global classification. Finally, the prediction results from local and global classifiers are combined to generate the final results.

Equally important for multiple labels. Totally ignoring the relevance among labels leads to inefficient learning and information loss. So we should not directly independently learn attention for individual labels.

To address the problems mentioned above, we propose the label-based attention module which uses the component-based attention as the bridge to connect label and document content. This module mainly consists of two sub-modules: 1) Component-Word Relevance generation and 2) Label-Component Association. Besides, to make use of information extracted from higher levels to lower levels, we also proposed 3) Hierarchical Information Integration sub-module.

**Component-Word Relevance.** This module is designed for learning the relevance between components and words at the level $h$ so that the important information of input document can be reflected based on each component. Here, components at the level $h$ are represented with a set of vectors: $C^h = \{v^h_{c_1}, v^h_{c_2}, \ldots, v^h_{c_m}\}$ which are trainable parameters and randomly initialized in the beginning and the relevance between a word $w_i$ and a component $c_j$ is denoted as $S^h_{w_i,c_j}$ which can be calculated through:

$$S^h_{w_i,c_j} = f_w(v^h_{w_i})^T v^h_{c_j}$$  \hspace{1cm} (1)

where $f_w(\cdot)$ is a single-layer MLP and $S^h \in \mathbb{R}^{N \times |C^h|}$ denotes the component-word relevance value matrix. $v^h_{w_i}$ represents the word embedding of word $w_i$ at the level $h$ and more details will be introduced later.

**Label-Component Association.** The target of this module is to associate each label with its related components. Labels at the level $h$ can be represented with label embeddings: $L^h = \{v^h_{l_1}, v^h_{l_2}, \ldots, v^h_{l_q}\}$ which are also trainable parameters and the relevance matrix between labels and components $\tilde{R}^h \in \mathbb{R}^{|L^h| \times |C^h|}$ is calculated as:

$$\tilde{R}^h = \text{softmax}(R^h)$$ \hspace{1cm} (2)

where $\text{softmax}(\cdot)$ is row-based softmax operation and the relevance between each pair of label $l_i$ and component $c_j$ is calculated as:

$$R^h_{c_jl_i} = f_l(v^h_{l_i})^T v^h_{c_j}$$ \hspace{1cm} (3)

where $f_l(\cdot)$ is a single-layer MLP to transform label embeddings to the same domain as component embeddings.

With the relevance matrix and association matrix defined in Eq. 1 and Eq. 2, , we can generate the label-based word attention matrix $A^h \in \mathbb{R}^{N \times L^h}$ by combining the component-word relevance matrix with the label-component association matrix through:

$$A^h = \text{softmax}(\tilde{R}^h S^h)$$ \hspace{1cm} (4)

The relevance between labels can be reflected with their shared components. Besides, the training of component embeddings is shared by all training samples which is unrelated to the number of
training samples for each label. So, the items mentioned above can be well addressed. Then, the label-based document embeddings $D_{X_i}^h \in \mathbb{R}^{L_h \times d}$ can be generated through:

$$D_{X_i}^h = \phi(A^h I_{X_i} W^h + b) \quad (5)$$

where $I_{X_i} \in \mathbb{R}^{N \times d}$ denotes the word embedding matrix of the document $X_i$ and the $i$th row of $I_{X_i}$ represents the word embedding $v_{w_i}$ of the $i$th word $w_i$. $W^h \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^{1 \times d}$ are trainable parameters and $\phi(\cdot)$ denotes an activation function (we choose to use RELU (Glorot et al., 2011) here). Generally speaking, a single-layer of MLP is applied to generate the label-based document embeddings at the level $h$. It is noted that, in this work, we have both local and global classifiers so it is required to separately generate different document embeddings for different classifiers and more details will be introduced in the Section 2.2.

Hierarchical Information Integration. From the perspective of human common sense, when there is hierarchical relationship among labels, it is more natural to carry out classification from rough to fine levels. Inspired by this idea, when we extract label-based document features at the level $h$, we will also refer to the features generated from a more rough level which is the level $h-1$. There are two ways to integrate the hierarchical information:

1) Label-based Document Embedding Masking: Given the local document embedding $\tilde{v}_{X_i}^{h-1} \in \mathbb{R}^{d \times 1}$ from the level $h-1$, the label-based document embeddings will be processed through:

$$\tilde{D}_{X_i}^h = mask(D_{X_i}^h, m_{X_i}^h) \quad (6)$$

where $mask(\cdot)$ denotes an operation that multiplying each row of $D_{X_i}^h \in \mathbb{R}^{L_h \times d}$ with the corresponding value in $m_{X_i}^h \in \mathbb{R}^{L_h \times 1}$. Here, each element of $m_{X_i}^h$ denotes the confidence score of each label at the level $h$ based on the prediction from $\tilde{v}_{X_i}^{h-1}$ and it can be calculated through:

$$m_{X_i,j}^h = sigmoid(f_m(\tilde{v}_{X_i}^{h-1}^T v_{l_j}^h)) \quad (7)$$

where $f_m(\cdot)$ represents a single-layer MLP. It is noted that $m_{X_i}^1 = 1 \in \mathbb{R}^{L_h \times 1}$, so the local document embedding from the previous level is not required for the label-based attention module at the level 1. To avoid over-filtering, LeakyRelu (He et al., 2015) is chosen as the activation function of $f_m(\cdot)$. Now, we can generate both the local $\tilde{v}_{X_i}^h$ and global $v_{X_i}^h$ document embeddings at the level $h$:

$$\tilde{v}_{X_i}^h = \text{average}(\tilde{D}_{X_i}^h)^T \quad (8)$$

$$v_{X_i}^h = \text{average}(D_{X_i}^h)^T \quad (9)$$

where $average(\cdot)$ means doing row-based average operation to matrix $\tilde{D}_{X_i}^h$ and $D_{X_i}^h$.

2) Word Embedding Enrichment: When we calculate the component-word relevance matrix $S^h$, the local document embedding from the level $h-1$ will be concatenated with each word embedding of the input document to generate the word embeddings at the level $h$, formally:

$$v_{w_j}^h = \tilde{v}_{X_i}^{h-1} \| v_{w_j} \quad (10)$$

So far, the procedure of generating label-based hierarchical document embeddings is introduced and the details of the label-based attention module is shown in Fig. 2.

2.2 Classification

We simultaneously apply local classification and global classification in this work.

Local Classification. To preserve the document information based on labels, we choose to apply the processed label-based document embeddings $\tilde{D}_{X_i}^h$ instead of $v_{X_i}^h$ to do local classification. Formally, at the level $h$, the probability that an input document $X_i$ is associated to the label $l_j$ can be calculated as:

$$p_{X_i,j} = \text{sigmoid}(\tilde{D}_{X_i,j}^h v_{l_j}) \quad (11)$$

where $\tilde{D}_{X_i,j}^h \in \mathbb{R}^{1 \times d}$ represents the $j$th row of the label-based document embedding matrix. Then the local loss is measured as follows:

$$\mathcal{L}_{local} = - \sum_{i=1}^{K} \sum_{l=1}^{H} (z_{l_i} \cdot \log(p_{X_i,j}) + (1 - z_{l_i}) \cdot \log(1 - p_{X_i,j})) \quad (12)$$

where $z_{l_i} = 1$ iff the document $X_i$ belongs to the label $l_j$ otherwise $z_{l_i} = 0$. 

Global Classification. To reduce the effect of error-propagation introduced from local classifiers and make full use of information extracted from other levels, we simultaneously optimize both local and global loss. For global classification, the global document embeddings extracted from all levels are applied to generate the final global document embedding \( v_{X_i} \), formally:

\[
v_{X_i}^g = (v_{X_i}^1 \parallel v_{X_i}^2 \parallel \cdots v_{X_i}^H)
\]

(13)

and the global loss is measured as follows:

\[
O_{global} = - \sum_{i=1}^{K} \sum_{j=1}^{M} (z_{ij} \cdot \log(p_{X_i,j}^g) + (1 - z_{ij}) \cdot \log(1 - p_{X_i,j}^g))
\]

(14)

where \( p_{X_i,j}^g \) is the jth element of the global prediction results \( p_{X_i}^g \in \mathbb{R}^{M \times 1} \) and \( M = \sum_H |\mathcal{L}^h| \), which represents the probability that document \( X_i \) is associated to the label \( l_j \). The global prediction results are calculated through:

\[
p_{X_i}^g = f_g(v_{X_i}^g)
\]

(15)

where \( f_g(\cdot) \) is a two-layer MLP.

So, the final loss we optimize in this work is:

\[
O_{all} = O_{local} + O_{global}
\]

(16)

Then the final prediction results can be generated by combining both local and global prediction results which is:

\[
p_{X_i}^{all} = \alpha p_{X_i} + (1 - \alpha) p_{X_i}^g
\]

(17)

where

\[
p_{X_i} = (p_{X_i}^1 \parallel p_{X_i}^2 \parallel \cdots p_{X_i}^H)
\]

(18)

where \( \alpha \) is the manually selected combination parameter which we select as 0.5 here.

3 Experiments

3.1 Experiment Setup

Datasets We test our model on four public datasets with different properties, WIPO-alpha\(^1\), BlurbGenreCollection(BGC)\(^2\), Enron(Klimt and Yang, 2004) and Reuters(Lewis et al., 2004). The detailed description of the experimental datasets is shown in Table 1.

As for BGC, Enron and Reuters, the lowest labels of each input sample are not required to be at the leaf nodes which makes them more difficult to be analyzed. WIPO-alpha is the only mandatory-leaf(Bi and Kwok, 2012) dataset while its large data size also increases its difficulty of analysis from another angle.

\(^1\)WIPO-alpha is available at https://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/index.html

\(^2\)BGC is available at https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/blurb-genre-collection.html
Table 1: Statistics of experimental datasets. $|L|$ denotes the total number of labels in the label hierarchy. Hierarchy denotes the number of labels at each of the hierarchical level, e.g., the dataset WIPO-alpha includes 4 levels of labels, and the number of labels at each level is 8, 114, 451, and 4363, respectively.

| Dataset    | $|L|$ | Hierarchy                  | Training | Validation | Testing |
|------------|------|----------------------------|----------|------------|---------|
| WIPO-alpha | 5,229| 8,114,451,4363             | 45,105   | 15,035     | 15,035  |
| BGC        | 146  | 7,46,77,16                 | 58,715   | 14,785     | 18,394  |
| Enron      | 56   | 3,40,13                    | 692      | 296        | 660     |
| Reuters    | 101  | 4,55,42                    | 2,100    | 900        | 3,000   |

Evaluation Metrics The result of each model for each input sample is represented as a number of scores and each score indicates the probability that the input sample belongs to the corresponding category. The final prediction results can be generated by thresholding the probabilities. So, choosing a proper threshold is critical to get a good performance while this is not the focus of this work. To avoid the influence of threshold choosing, we choose to use the area under the average precision-recall curve ($AU(\overline{PRC})$) to reflect the performance of an algorithm which is also widely acceptable in HMTC domain (Cerri et al., 2016).

3.2 Compared Methods
To provide comprehensive evaluation, we compare our proposed method with a number of state-of-the-art neural network-based HMTC algorithms as well as the variants of LA-HCN.

- **HMCN-F, HMCN-R** (Wehrmann et al., 2018): The feedforward and recurrent version of HMCN which is the first neural network-based HMC method that combines both local and global information to do hierarchical classification. During the training procedure, label hierarchy information is preserved by penalizing hierarchical violations.

- **Cap.Network** (Aly et al., 2019): It is the first work that introduces capsules into HMC tasks. By associating each label in the hierarchy with a capsule and applying routing mechanism, Cap.Network indicates the effectiveness of capsules in feature identification and combination. The label hierarchy information is directly applied to do post-processing(label correction) during the prediction stage.

- **HARNN** (Huang et al., 2019): It is the most related algorithm to LA-HCN which is designed to optimize both local and global loss simultaneously. Besides, attention mechanism is applied at each hierarchical level to extract level-based document features for better local classification performance. At the prediction stage, local and global prediction scores are combined to generate the final results.

The selected baselines are all state-of-the-art neural network-based HMTC algorithms which cover both flat approach (Cap.Network) and hybrid-approaches (HARNN, HMCN). To demonstrate the effectiveness of our proposed label-based attention mechanism and the combination of local and global approaches, two variants of LA-HCN are also tested. Here, LA-HCN$_G$ and LA-HCN$_L$ represent the variants of LA-HCN which only optimize global and local loss respectively. For fair comparison, the input features of HMCN are replaced with the output of Bi-LSTM or pre-trained word embeddings which leads to slightly better performance of HMCN than what is reported in their paper.

3.3 Implementation Details
Training, testing and validation split is already provided for BGC, Enron and Reuters so we just follow the original settings. As for WIPO-alpha, we randomly select 20% of samples as testing and validation sets respectively.

As for WIPO-alpha and BGC, Bi-LSTM, a variant of LSTM (Hochreiter and Schmidhuber, 1997) is applied as the base text encoder since the raw text of these two datasets are available. However, only processed word statistic information is provided for Enron and Reuters, so we simply use pre-trained word embeddings as the input text features. More details can be find in the Appendix.

3.4 Experimental Results
Classification Results. We compare the performance of LA-HCN with other neural network-based state-of-the-art HMTC methods and the results are shown in Table 2.
**Algorithms** | WIPO-alpha | BGC | Enron | Reuters  
---|---|---|---|---
Cap.Network | OOM | 0.7613 | 0.713 | 0.6562  
HMCN-F | 0.517 | 0.805 | 0.721 | 0.675  
HMCN-R | 0.528 | 0.804 | 0.739 | 0.674  
HARNN | 0.573 | 0.822 | 0.745 | 0.667  
LA-HCN$_L$ | 0.553 | 0.815 | 0.726 | 0.645  
LA-HCN$_G$ | 0.568 | 0.825 | 0.745 | 0.747  
LA-HCN | **0.595** | **0.832** | **0.755** | **0.742**

Table 2: Performance comparison on the four datasets(\(AU(\overline{PRC})\)). OOM denotes out of memory.

LA-HCN achieves top one on all the datasets against other comparison algorithms especially on Reuters and this indicates the effectiveness of LA-HCN in HMTI problem. More specifically, the hybrid-approaches(HMCN, HARNN, LA-HCN) outperform simple flat algorithms(Cap.Network) on all datasets which shows the advantages of combining local and global approaches. The performance of HARNN and LA-HCN is better than HMCN on three datasets(except Reuters) which indicates that hierarchically extracting document features based on label levels is helpful to improve the hierarchical classification performance. In addition, by comparing LA-HCN and its variants, it is easy to find that the performance of global classification is always better than that of local classification. Separately generating document embeddings for local and global classification is the reason why LA-HCN gets much better performance than other algorithms where the combined local and global classifications are also applied especially on Reuters. The use of the separately generated global document embeddings in LA-HCN can effectively reduce the error-propagation introduced by local classification. However, the performance of LA-HCN can still be improved with local label-based document embeddings which can provide more hierarchical information. Besides, usually when the label structure is more complicated, the improvement becomes more obvious (WIPO-alpha).

**Learned Label-based Attention.** One of the most important characteristic of LA-HCN is its capability of capturing important information of document based on different labels. In this experiment, we visualize the learned attention based on labels from different levels to verify its effectiveness and we choose BGC as the experimental data here. We also show the learned attention from

| Dataset | LA-HCN | HARNN | HAN | BGC | OOM | 0.832 | 0.822 | 0.749 |
|---|---|---|---|---|---|---|---|---|
| BGC | | | | | | | | |

Table 3: The classification performance of LA-HCN, HARNN and HAN on BGC(\(AU(\overline{PRC})\)). 

HARNN and HAN for comparison. 

The classification performance of the three algorithms on BGC is shown in Table 3 and the learned attention is shown in Fig. 3 and Fig. 4.

In Fig. 3, we show the attention learned from the level \(h = 3\) where the text is associated with the label *Exercise* and *Philosophy*. It is observable to find that LA-HCN pays more attention to the word "yoga" when it is classified to *Exercise* while it gives "spiritual", "buddha" and "tibetan" higher importance when it is classified to *Philosophy* which is also in line with human common sense. However, HARNN and HAN can only generate the same attention value for all labels which is difficult to understand in some case.

In Fig. 4, we show the attention learned based on the label *Nonfiction* from the level \(h = 1\). Combined with the attention learned from the level \(h = 3\) in Fig. 3, we can find that, in LA-HCN, the word "yoga" is not important at the level \(h = 1\) where it is associated with the label *Nonfiction* while it can still be captured at the level \(h = 3\) when it is associated with the label *Exercise*. However, HARNN can only select important words at the level \(h = 3\) based on the attention learned from its previous level so that the word "yoga" will be lost at the level \(h = 3\) since "yoga" is not that important for the classification at the level \(h = 1\). This also demonstrates the way that LA-HCN passing hierarchical information is more effective.

### 4 Related Work

Text classification is an important research area where HMTI also plays an important role which is widely applied in many different applications. As for the classification problem, some work focus on text encoding such as Doc2vec(Le and Mikolov, 2014), LSTM(Hochreiter and Schmidhuber, 1997) and BERT(Devlin et al., 2019) while others pay more attention to classifier design which is also the focused topic of HMTI.

As for the design of hierarchical classifier, there are two main directions: local and global approaches(Silla and Freitas, 2011). Koller and Sahami (1997) is the first type of local classifier which
Figure 3: Visualization of the learned attention(words) at the level $h = 3$ from HAN, HARNN and LA-HCN. The corresponding labels as well as the prediction probabilities are also displayed.

| Algorithm | Label | Predicted Score | Input Document |
|-----------|-------|-----------------|----------------|
| HAN       | Exercise (level: 3) | 0.503 | complete manual physical spiritual well founder om yoga center yoga body buddha mind first book give reader best intrinsically linked practice cyndi lee author bestselling series om yoga box share twenty year experience practicing shaman buddhist one country famous yoga instructor easy use guide show reader yoga level combine basic tenet buddhism meditation yoga practice book offer simple meditation program exercise sequence done anywhere addition advanced rigorous regimen written personal comfortable charismatic style cyndi lee brought class yoga body buddha mind comprehensive guide spiritual well ultimate enlightening experience |
| HARNN     | Philosophy (level: 3) | 0.753 | complete manual physical well founder om yoga center yoga body buddha mind first book give reader best intrinsically linked practice cyndi lee author bestselling series om yoga box share twenty year experience practicing shaman buddhist one country famous yoga instructor easy use guide show reader yoga level combine basic tenet buddhism meditation yoga practice book offer simple meditation program exercise sequence done anywhere addition advanced rigorous regimen written personal comfortable charismatic style cyndi lee brought class yoga body buddha mind comprehensive guide spiritual well ultimate enlightening experience |
| LA-HCN    | Exercise (level: 3) | 0.949 | complete manual physical well founder om yoga center yoga body buddha mind first book give reader best intrinsically linked practice cyndi lee author bestselling series om yoga box share twenty year experience practicing shaman buddhist one country famous yoga instructor easy use guide show reader yoga level combine basic tenet buddhism meditation yoga practice book offer simple meditation program exercise sequence done anywhere addition advanced rigorous regimen written personal comfortable charismatic style cyndi lee brought class yoga body buddha mind comprehensive guide spiritual well ultimate enlightening experience |
|           | Philosophy (level: 3) | 0.821 | complete manual physical well founder om yoga center yoga body buddha mind first book give reader best intrinsically linked practice cyndi lee author bestselling series om yoga box share twenty year experience practicing shaman buddhist one country famous yoga instructor easy use guide show reader yoga level combine basic tenet buddhism meditation yoga practice book offer simple meditation program exercise sequence done anywhere addition advanced rigorous regimen written personal comfortable charismatic style cyndi lee brought class yoga body buddha mind comprehensive guide spiritual well ultimate enlightening experience |

Figure 4: Visualization of the learned attention(words) at the level $h = 1$ from HARNN and LA-HCN. The corresponding labels as well as the prediction probabilities are also displayed.

| Algorithm | Label | Predicted Score | Input Document |
|-----------|-------|-----------------|----------------|
| HARNN     | Nonfiction (level: 1) | 1.000 | complete manual physical well founder om yoga center yoga body buddha mind first book give reader best intrinsically linked practice cyndi lee author bestselling series om yoga box share twenty year experience practicing shaman buddhist one country famous yoga instructor easy use guide show reader yoga level combine basic tenet buddhism meditation yoga practice book offer simple meditation program exercise sequence done anywhere addition advanced rigorous regimen written personal comfortable charismatic style cyndi lee brought class yoga body buddha mind comprehensive guide spiritual well ultimate enlightening experience |
| LA-HCN    | Nonfiction (level: 1) | 1.000 | complete manual physical well founder om yoga center yoga body buddha mind first book give reader best intrinsically linked practice cyndi lee author bestselling series om yoga box share twenty year experience practicing shaman buddhist one country famous yoga instructor easy use guide show reader yoga level combine basic tenet buddhism meditation yoga practice book offer simple meditation program exercise sequence done anywhere addition advanced rigorous regimen written personal comfortable charismatic style cyndi lee brought class yoga body buddha mind comprehensive guide spiritual well ultimate enlightening experience |

is proposed to explore the hierarchy by using local information and build multiple local classifiers around it, following which, a series of local approaches including LCN, LCNP as well as LCL\(^3\) based methods are proposed (Fagni and Sebastiani; Secker et al.; Costa et al., 2007). However, simply applying local approaches has the disadvantages of error-propagation. Global approaches are designed to build one classifier to explore hierarchical information from the global angle which can reduce the overall model size. Most global approaches are generally modified from the flat classification algorithms such as hAnt-Miner(Barril Otero et al., 2009), Vens et al. (2008)(decision tree based) and GMNB\(w\)U(Silla Jr. and Freitas, 2009)(Naive Bayes classifier based). Recently, more neural network based HMC algorithms which combine both local and global approaches are proposed(Mao et al., 2019; Wehrmann et al., 2018). However, these algorithms ignored the hierarchical feature extraction which is also important for HMC.

As for hierarchically extracting features, the most relevant work includes HAN(Yang et al., 2016), SMASH(Jiang et al., 2019) and HARNN(Huang et al., 2019). However, HAN proposed to extract hierarchical information based on the document structures(from words to sentences) instead of label structures. SMASH is similar to HAN except they use BERT as the base model. HARNN extracts hierarchical information at different label levels and generate multiple document embeddings for local and global classification. However, HARNN applied a general attention to extract document features which is not explicit for human understanding compared with label-based attention. Besides, the same document embedding is used in both local and global classification which reduces their performance.

\(^3\)LCN, LCNP and LCL denotes local classifier per node, local classifier per parent node and local classifier per level approach respectively.
5 Conclusion

We proposed a label-based attention hierarchical multi-label text classification neural network algorithm (LA-HCN) where label-based attention are learned for the text at different hierarchical levels, after which, both local and global text embeddings are generated for local and global classification respectively. At different levels, meaningful attention can be learned based on different labels. Comprehensive experiments are conducted to show the effectiveness of LA-HCN which outperforms other neural network-based state-of-the-art HMTC algorithms on 4 datasets. Besides, the learned label-based attention is visualized which is also in line with human understanding. However, the practical meaning of learned components is not well explored so far and we will do this in our future work where the explicit label structure should also be taken into consideration.

References

Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. 2013. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In Proceedings of the 22nd International Conference on World Wide Web, WWW 13, page 1324, New York, NY, USA. Association for Computing Machinery.

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Fernando Barril Otero, Alex Freitas, and Colin Johnson. 2009. A hierarchical classification ant colony algorithm for predicting gene ontology terms. pages 68–79.

Wei Bi and James T. Kwok. 2012. Mandatory Leaf Node Prediction in Hierarchical Multilabel Classification. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 153–161. Curran Associates, Inc.

Ricardo Cerri, Rodrigo C Barros, André CPLF de Carvalho, and Yaochu Jin. 2016. Reduction strategies for hierarchical multi-label classification in protein function prediction. BMC bioinformatics, 17(1):373.

Eduardo P Costa, Ana C Lorena, André CPLF Carvalho, Alex A Freitas, and Nicholas Holden. 2007. Comparing several approaches for hierarchical classification of proteins with decision trees. In Brazilian symposium on bioinformatics, pages 126–137. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Tiziano Fagni and Fabrizio Sebastiani. On the selection of negative examples for hierarchical text categorization. Citeseer.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of Proceedings of Machine Learning Research, pages 315–323, Fort Lauderdale, FL, USA. PMLR.

Juan Carlos Gomez and Marie-Francine Moens. 2014. A survey of automated hierarchical classification of patents. In Professional search in the modern world, pages 215–249. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pages 1026–1034.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Comput., 9(8):17351780.

Jie Hu, Shaobo Li, Yong Yao, Liya Yu, Guanci Yang, and Jianjun Hu. 2018. Patent keyword extraction algorithm based on distributed representation for patent classification. Entropy, 20(2):104.

Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. Hierarchical multi-label text classification: An attention-based recurrent network approach. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 19, page 10511060, New York, NY, USA. Association for Computing Machinery.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In The World Wide Web Conference, WWW 19, page 795806, New York, NY, USA. Association for Computing Machinery.

Svetlana Kiritchenko, Stan Matwin, and A Fazel Famili. 2005. Functional annotation of genes using hierarchical text categorization. In in Proc. of the BioLINK SIG: Linking Literature, Information and Knowledge for Biology (held at ISMB-05. Citeseer.
Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In Conference of the Canadian Society for Computational Studies of Intelligence, pages 395–406. Springer.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In ECML ’04: Proceedings of the 18th European Conference on Machine Learning – LNCS 3201, pages 217–226. Springer Berlin / Heidelberg.

Daphne Koller and Mehran Sahami. 1997. Hierarchically classifying documents using very few words. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 170–178.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196.

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, 5:361–397.

Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. Deepatent: patent classification with convolutional neural networks and word embedding. Scientometrics, 117(2):721–744.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. 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 445–455.

Kunal Punera and Joydeep Ghosh. 2008. Enhanced hierarchical classification via isotonic smoothing. In Proceedings of the 17th international conference on World Wide Web, pages 151–160.

Kervy Rivas Rojas, Gina Bustamante, Marco A Sobrevilla Cabezudo, and Arturo Oncevay. 2020. Efficient strategies for hierarchical text classification: External knowledge and auxiliary tasks. arXiv preprint arXiv:2005.02473.

Andrew Secker, Matthew N Davies, Alex Alves Freitas, Edward Clark, Jonathan Timmis, and Darren R Flower. Hierarchical classification of g-protein-coupled receptors with data-driven selection of attributes and classifiers.

Carlos N Silla and Alex A Freitas. 2011. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, 22(1-2):31–72.

C. N. Silla Jr. and A. A. Freitas. 2009. A global-model naive bayes approach to the hierarchical prediction of protein functions. In 2009 Ninth IEEE International Conference on Data Mining, pages 992–997.