HTCInfoMax: A Global Model for Hierarchical Text Classification via Information Maximization

Zhongfen Deng[1], **Hao Peng**[2, 3], **Dongxiao He**[4], **Jianxin Li**[2], **Philip S. Yu**[1]
[1]Department of Computer Science, University of Illinois at Chicago, Chicago, USA
[2]BDBC, Beihang University, Beijing, China
[3]School of Cyber Science and Technology, Beihang University, Beijing, China
[4]School of Computer Science and Technology, Tianjin University, Tianjin, China
{zdeng21,psyu}@uic.edu, {penghao,lijx}@act.buaa.edu.cn
hedongxiao@tju.edu.cn

Abstract

The current state-of-the-art model HiAGM for hierarchical text classification has two limitations. First, it correlates each text sample with all labels in the dataset which contains irrelevant information. Second, it does not consider any statistical constraint on the label representations learned by the structure encoder, while constraints for representation learning are proved to be helpful in previous work. In this paper, we propose HTCInfoMax to address these issues by introducing information maximization which includes two modules: text-label mutual information maximization and label prior matching. The first module can model the interaction between each text sample and its ground truth labels explicitly which filters out irrelevant information. The second one encourages the structure encoder to learn better representations with desired characteristics for all labels which can better handle label imbalance in hierarchical text classification. Experimental results on two benchmark datasets demonstrate the effectiveness of the proposed HTCInfoMax.

1 Introduction

Hierarchical text classification (HTC) is a particular subtask of multi-label text classification (Li et al., 2020). Many datasets have been proposed to study HTC for decades, such as RCV1 (Lewis et al., 2004) and NYTimes (Sandhaus, 2008), which categorize a news into several categories/labels. And all the labels in each dataset are usually organized as a tree or a directed acyclic graph. Thus, there is a label taxonomic hierarchy existing in each dataset. The goal of HTC is to predict multiple labels in a given label hierarchy for a given text.

There are two groups of existing methods for HTC: local approaches and global approaches. Local approaches usually build a classifier for each label/node (Banerjee et al., 2019), or for each parent node, or for each level of the label hierarchy(Wehrmann et al., 2018; Huang et al., 2019; Chang et al., 2020). Global approaches just build one classifier to simultaneously predict multiple labels of a given text. The earlier global approaches ignore the hierarchical structure of labels and assume there is no dependency among labels which leads to flat models such as (Johnson and Zhang, 2015). Later on, more and more works try to make use of the label taxonomic hierarchy to improve the performance by employing different strategies such as recursively regularized Graph-CNN (Peng et al., 2018), reinforcement learning (Mao et al., 2019), attentional capsule network (Peng et al., 2019), meta-learning (Wu et al., 2019) and structure encoder (Zhou et al., 2020). Many attention-based models are also proposed to learn more refined text features for text classification tasks such as (You et al., 2019; Deng et al., 2020). Among these methods, HiAGM proposed by Zhou et al. (2020) is the state-of-the-art model for HTC which designs a structure encoder that integrates the label prior hierarchy knowledge to learn label representations, and then proposes a model HiAGM with two variants (one is HiAGM-LA, the other is HiAGM-TP) based on the structure encoder to capture the interactions between text features and label representations. However, there are some limitations of HiAGM. Firstly, it utilizes the same label hierarchy information for every text sample which cannot distinguish the relevant and irrelevant labels to a specific text sample. Although HiAGM-LA can implicitly relate each text to its corresponding labels by soft attention weights, there are still irrelevant and noisy information. Secondly, for HiAGM-LA, there is no statistical constraint on the label embeddings generated by the structure encoder, while statistical constrains for representation learning are proved to be helpful by Hjelm et al. (2019).

To address the two limitations of HiAGM-LA, we propose HTCInfoMax which introduces information maximization consisting of two new modules which are text-label mutual information maximization and label prior matching on top of HiAGM-LA. Specifically, the first new module makes a connection between each text sample and its corresponding labels explicitly by maximizing the mutual information between them, and thus can filter out irrelevant label information for a specific text sample. The label prior matching module can impose some constraints on the learned representation of each label to force the structure encoder to learn better representations with desirable properties for all labels and thus also improve the quality of representations for low-frequency labels, which helps handle label imbalance issue better.

In summary, our main contributions are: 1) We propose a novel global model HTCInfoMax for HTC by introducing information maximization which includes two modules: text-label mutual information maximization and label prior matching. 2) To our best knowledge, this is the first work to utilize text-label mutual information maximization for HTC which enables each text to capture its corresponding labels' information in an effective way. 3) Also, to our best knowledge, this is the first work to introduce label prior matching for HTC which encourages the structure encoder to learn desired label representations for all labels which can better handle inherent label imbalance issue in HTC. 4) Experimental results demonstrate the effectiveness of our proposed model for HTC. 5) We release our code to enable replication, available at `https://github.com/RingBDStack/HTCInfoMax`.

2 Methodology

2.1 Our approach

The overall architecture of our model is shown in Figure 1. The major part of HTCInfoMax is the "Information Maximization" part shown in the dashed box which has two new modules: text-label mutual information maximization and label prior matching, which will be introduced in the following sections. We keep the remaining part such as text encoder, structure encoder and the predictor be the same as in HiAGM-LA (Zhou et al., 2020).

2.1.1 Text-label mutual information estimation and maximization

Good text representation is critical for predicting its corresponding labels, thus fusing label information into text feature can help improve the prediction performance. The HiAGM-LA utilizes multi-label attention to bridge the text feature of each sample with all labels' information implicitly, which can somehow help each text obtain some label information. However, irrelevant label information is also injected into the text feature by using soft attention weights. Therefore, we design a text-label mutual information maximization module to help remove irrelevant label information for each text as well as help each text capture its corresponding labels' information. In this way, the learned representation for each text incorporates useful label information which is helpful for predicting its labels.

To implement the text-label mutual information maximization, we first select the ground truth labels for each text sample in the training process, and then apply a discriminator to estimate the mutual information between text and its labels, which is also known as negative sampling estimation. Let $\mathbb{P}_T$ and $\mathbb{P}_Y$ denote the distribution of text feature outputted by the text encoder and the distribution of label representation produced by the structure encoder respectively. And the joint distribution of text and label is denoted as $\mathbb{P}_{TY} = \mathbb{P}_{Y|T}\mathbb{P}_T$. Then the positive samples are the pairs of text $\mathbf{t}$ and its corresponding labels $\mathbf{y}$ which is denoted as $(\mathbf{t}, \mathbf{y})$, in other words, these positive samples are drawn from the joint distribution of text and label. For the negative samples, we pair $\mathbf{y}$ with another text sample $\mathbf{t}'$ in the same batch which is denoted as $(\mathbf{t}', \mathbf{y})$, the negative samples can be deemed as drawn from the product of marginal distribution of text $\mathbb{P}_T$ and label $\mathbb{P}_Y$. Both positive and negative samples are fed to the discriminator D_{MI} to do classification and to estimate the mutual information $I(T;Y)$ between text and label shown in Eq. (1). $D_{MI}(\mathbf{t}, \mathbf{y})$ and $D_{MI}(\mathbf{t}', \mathbf{y})$ represents the probability score assigned to the positive and negative sample by the discriminator respectively. The goal of the text-label mutual information maximization module is to maximize $I(T;Y)$, thus the loss from this module is shown in Eq. (2).

$$\begin{aligned} I(T;Y) =& \mathbb{E}_{(\mathbf{t},\mathbf{y})\sim\mathbb{P}_{TY}}[\log D_{MI}(\mathbf{t},\mathbf{y})] + \\ & \mathbb{E}_{(\mathbf{t}',\mathbf{y})\sim\mathbb{P}_T\mathbb{P}_Y}[\log(1 - D_{MI}(\mathbf{t}',\mathbf{y}))], \end{aligned} \tag{1}$$

$$L_{MI} = -I(T;Y). \tag{2}$$

This module is inspired by Deep InfoMax (DIM) (Hjelm et al., 2019) which utilizes local and global mutual information maximization to help the encoder learn high-level representation for an image. The structure of the discriminator D_{MI} in this module can be found in the Appendix A.1.

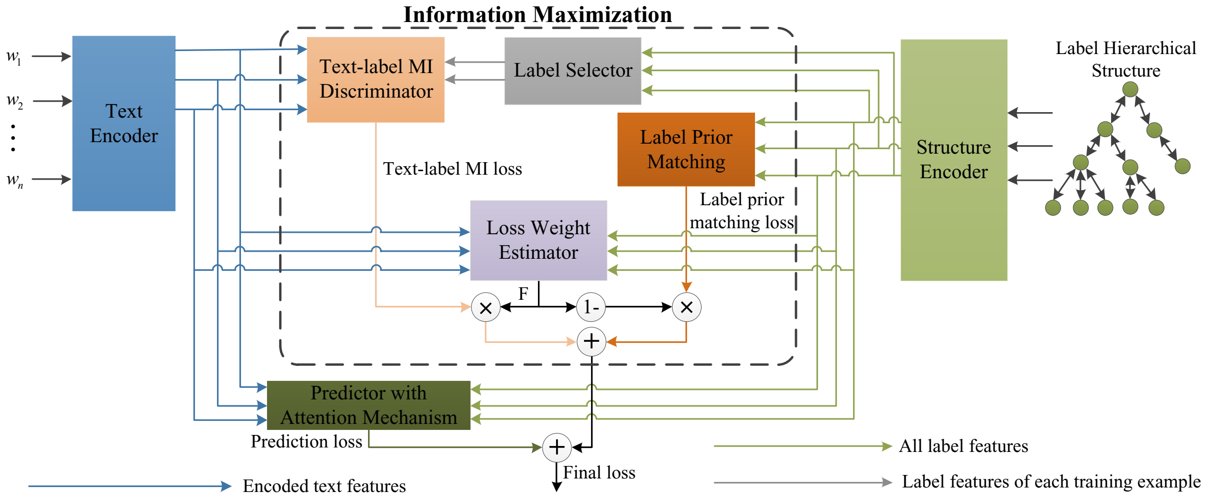

Figure 1: The architecture of our model HTCInfoMax.

2.1.2 Label prior matching

There is an inherent label imbalance issue in HTC, thus the learned label embeddings by the model for low-frequency labels are not good because of underfitting caused by less training examples. The label prior matching imposes some statistical constrains on the learned representation of each label which can help the structure encoder learn better label representations with desirable characteristics for all labels. This also improves the quality of representations for low-frequency labels, which helps handle the label imbalance situation better in terms of improvement of Macro-F1 score.

To implement the label prior matching mechanism, we use a method similar to adversarial training in adversarial autoencoders (Makhzani et al., 2015) but without a generator to force the learned label representation to match a prior distribution. We denote the prior as $\mathbb{Q}$ and the distribution of label representation learned by the structure encoder as $\mathbb{P}$. Specifically, a discriminator network D_{pr} is employed to distinguish the representation/sample drawn from the prior (i.e., real sample which is denoted as $\tilde{\mathbf{y}}$) from the label embedding produced by the structure encoder (i.e., fake sample which is denoted as $\mathbf{y}$). For each label, we utilize D_{pr} to calculate its corresponding prior matching loss l_{pr} , which is shown in Eq. (3).

$$l_{pr} = -(\mathbb{E}_{\tilde{\mathbf{y}}\sim\mathbb{Q}}[\log D_{pr}(\tilde{\mathbf{y}})] + \mathbb{E}_{\mathbf{y}\sim\mathbb{P}}[\log(1 - D_{pr}(\mathbf{y}))]), \tag{3}$$

This loss aims at pushing the distribution $\mathbb{P}$ of learned representation for a label towards its prior distribution $\mathbb{Q}$. The final label prior matching loss is the average of losses from all the labels which is shown in Eq. (4), N is the number of labels.

$$L_{pr} = \frac{1}{N}\sum_{i=1}^{N} l_{pr}^{i}. \tag{4}$$

This idea is inspired by DIM which matches the representation of an image to a prior, but different from DIM, it trains the structure encoder to learn desired representations for all labels by imposing the constraints on each label's representation.

An uniform distribution on the interval [0, 1) is adopted as the label prior distribution $\mathbb{Q}$ in the label prior matching module. The reason for choosing the uniform distribution is that it works well as a prior in DIM for generating image representations. And the improvement of Macro-F1 score in the experimental results of hierarchical text classification further verifies the suitability of using the uniform distribution as the label prior. The detailed structure of the discriminator D_{pr} can be found in the Appendix A.2.

2.1.3 Final loss of HTCInfoMax

A loss weight estimator is adopted to learn the weights for text-label mutual information loss and label prior matching loss by using learned text features $\mathbf{t}$ and all labels' representation $\mathbf{y}$, shown in Eq. (5), and both W_1 and W_2 are trainable parameters.

$$F = \text{sigmoid}(W_1\mathbf{t} + W_2\mathbf{y}), \tag{5}$$

And the loss from the predictor is the traditional binary cross-entropy loss L_c (Zhou et al., 2020). Then the final objective function of HTCInfoMax is the combination of all the three losses as follows:

$$L = L_c + F \times L_{MI} + (1 - F) \times L_{pr}. \tag{6}$$

3 Experiment

3.1 Datasets and evaluation metrics

Following HiAGM (Zhou et al., 2020), we use RCV1-V2 (Lewis et al., 2004) and Web of Science (WOS) (Kowsari et al., 2017) benchmark datasets to evaluate our model and adopt the same split of RCV1-V2 and WOS as HiAGM. The statistics of the two datasets are shown in Table 1.

Dataset	L	Depth	Avg-L	Train	Val	Test
RCV1-V2	103	4	3.24	20,834	2,315	781,265
WOS	141	2	2.0	30,070	7,518	9,397

Table 1: Statistics of datasets. L is the total number of labels in the dataset, Avg-L is the average number of labels for each sample. Depth means the maximum level of the label hierarchy.

Standard evaluation metrics including Micro-F1 (Mi-F1) and Macro-F1 (Ma-F1) score are employed to evaluate our model. In label imbalance situation, Ma-F1 can better evaluate model's performance in the perspective of not focusing on frequent labels in a certain degree.

3.2 Experimental setup

In order to make a fair comparison between our model and HiAGM, we use the same parameter settings as HiAGM and follow its implementation details which can be seen in (Zhou et al., 2020).

3.3 Experimental results

The experimental results of our model are shown in Table 2, each score is the average result of 8 runs. The results of HiAGM are referred from (Zhou et al., 2020). There are two variants of HiAGM which are HiAGM-LA and HiAGM-TP. As stated before, our model is built on top of HiAGM-LA to address its limitations. From Table 2, one can see that our model outperforms the HiAGM-LA model with either GCN or TreeLSTM as structure encoder on two datasets, which demonstrates that the introduced information maximization in our model can address the limitations of HiAGM-LA and improve the performance. This is because the label prior matching can drive the structure encoder to learn good and desired label representations that encode more useful and informative information of labels, and the text-label mutual information maximization module helps learn better representation of each text for prediction by fusing the above learned good representations of its ground truth labels while ignoring irrelevant labels' information. It is also worth nothing that the improvement of Ma-F1 on the RCV1-V2 dataset is bigger compared with that on WOS, which indicates that our model can work better on dataset with a more complicated label hierarchy as RCV1-V2 has a deeper label hierarchical structure than WOS.

Models		RCV1-V2		WOS	
		Mi-F1	Ma-F1	Mi-F1	Ma-F1
HiAGM-LA	GCN	82.21	61.65	84.61	79.37
	TreeLSTM	82.54	61.90	84.82	79.51
HiAGM-TP	GCN	***83.96***	***63.35***	***85.82***	***80.28***
	TreeLSTM	83.20	62.32	85.18	79.95
HTCInfoMax (Ours)		**83.51**	**62.71**	**85.58**	**80.05**

Table 2: Results of HTCInfoMax and HiAGM on RCV1-V2 and WOS datasets.

Although our model does not outperform all the results of HiAGM-TP, it reaches the similar performance. This indicates that information maximization is an alternative effective way to fuse the text feature and label information together to boost the performance. In addition, apart from generating text representations, our model can also generate refined label representations via information maximization which can be utilized for inference, while HiAGM-TP cannot produce such label embeddings for usage in the inference phase because it directly feeds the text feature into the structure encoder to obtain final text representation for prediction. In other words, HiAGM-TP encodes text and label information into only one feature space. However, obtaining separate text features and label features such as the ones generated by our model can help encode more semantic information of labels, which may be helpful for HTC especially when there is a large label hierarchy in the dataset.

We do not report the results of other baselines such as HFT(M) (Shimura et al., 2018), SGM (Yang et al., 2018), HiLAP-RL (Mao et al., 2019), etc. as they can be found in (Zhou et al., 2020), and our model performs better than these baselines.

3.4 Ablation study

To demonstrate the effectiveness of the two modules of information maximization, we conduct an ablation study and the results are shown in Table 3. Every score in Table 3 is the average result of 8 runs. From Table 3, one can see that HTCInfoMax outperforms the variant without text-label mutual information maximization module (i.e., HTCInfo-

Models	RCV1-V2		WOS	
	Mi-F1	Ma-F1	Mi-F1	Ma-F1
HTCInfoMax w/o MI	83.42	61.79	85.46	79.94
HTCInfoMax w/o LabelPrior	82.75	60.57	84.74	79.01
HTCInfoMax	**83.51** ↑	**62.71** ↑	**85.58** ↑	**80.05** ↑

Table 3: Ablation study results on RCV1-V2 and WOS datasets. w/o means without. Arrow ↑ indicates statistical significance ($p < 0.01$).

Max w/o MI) by 0.09, 0.92 points on RCV1-V2 and 0.12, 0.11 points on WOS in terms of Mi-F1 and Ma-F1 respectively, which indicates that the text-label mutual information maximization module can make each text capture its corresponding labels' information and thus improves the Mi-F1 and Ma-F1 score at the same time. When compared with the other variant (i.e., HTCInfoMax w/o LabelPrior), the improvements of the two metrics can also be observed but Ma-F1 has larger improvements by 2.14 and 1.04 points on RCV1-V2 and WOS respectively compared with Mi-F1. This demonstrates that label prior matching helps regularize the label feature space and forces the structure encoder to learn better representations with desired properties for all labels. Thus the representations of imbalanced labels are also well learned, which helps mitigate the issue of underfitting of low-frequency labels, and thus improves the Ma-F1 score more and better to handle the label imbalance issue.

4 Conclusion

We propose HTCInfoMax to address the limitations of HiAGM by introducing information maximization which includes two modules: text-label mutual information maximization and label prior matching. The label prior matching can drive the model to learn better representations for all labels, while the other module further fuses such learned label representations into text to learn better text representations containing effective label information for prediction. The experimental results demonstrate the effectiveness of HTCInfoMax.

Acknowledgment

The corresponding author is Hao Peng. The authors of this paper were supported by NSFC through grants U20B2053, 62002007, 62073012 and 61876128, S&T Program of Hebei through grant 20310101D, and in part by NSF under grants III-1763325, III-1909323, and SaTC-1930941. We thank the reviewers for their constructive feedback.

References

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, Florence, Italy. Association for Computational Linguistics.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3163–3171.

Zhongfen Deng, Hao Peng, Congying Xia, Jianxin Li, Lifang He, and Philip Yu. 2020. Hierarchical bi-directional self-attention networks for paper review rating recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6302–6314, Barcelona, Spain (Online). International Committee on Computational Linguistics.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

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*, pages 1051–1060.

Rie Johnson and Tong Zhang. 2015. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112, Denver, Colorado. Association for Computational Linguistics.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE.

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.

Qian Li, Hao Peng, Jianxin Li, Congyin Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

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, Hong Kong, China. Association for Computational Linguistics.

Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. 2018. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the 2018 World Wide Web Conference*, pages 1063–1072.

Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, Philip Yu, and Lifang He. 2019. Hierarchical taxonomy-aware and attentional graph capsule rcnns for large-scale multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Brussels, Belgium. Association for Computational Linguistics.

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. 2018. Hierarchical multi-label classification networks. In *International Conference on Machine Learning*, pages 5075–5084.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label 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 4354–4364, Hong Kong, China. Association for Computational Linguistics.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.

A Architecture Details of Information Maximization

A.1 The structure of discriminator in text-label mutual information maximization module

The discriminator D_{MI} consists of two 1D-convolutional layers with kernels of size 3 and three linear layers. The architecture of D_{MI} is shown in Figure 2 and the details of all the layers are shown in Table 4 ("-" indicates that there is no activation for the corresponding layer). As shown in Figure 2, the discriminator D_{MI} takes pairs of text representation and label representation as input. The text representations are fed to the convolutional layers first, then the label representations are concatenated with the output from the convolutional layers and fed to the following linear layers. The final linear layer produces a score for each pair of text sample and corresponding labels.

Layers	Size (Input)	Size (Output)	Activation
1D-conv layer	300	300	ReLU
1D-conv layer	300	512	-
Linear layer	812	512	ReLU
Linear layer	512	512	ReLU
Linear layer	512	1	-

Table 4: Layer details of the discriminator D_{MI}.

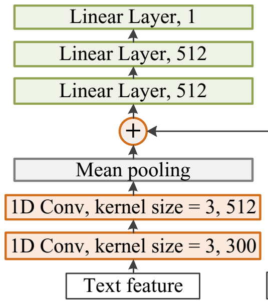

Figure 2: The structure of discriminator D_{MI}.

A.2 The structure of discriminator in label prior matching

The discriminator D_{pr} in the label prior matching module is composed of three linear layers. The details of these layers are shown in Table 5. This discriminator takes label representation as input and is applied for each label to compute its prior matching loss as stated in Section 2.1.2.

Layers	Size (Input)	Size (Output)	Activation
Linear layer	300	1000	ReLU
Linear layer	1000	200	ReLU
Linear layer	200	1	Sigmoid

Table 5: Layer details of the discriminator D_{pr}.