Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042980

ARTICLE

Gate-Attention and Dual-End Enhancement Mechanism for Multi-Label Text Classification

Jieren Cheng[1,2], Xiaolong Chen[1,*], Wenghang Xu[3], Shuai Hua[3], Zhu Tang[1] and Victor S. Sheng[4]

[1]School of Computer Science and Technology, Hainan University, Haikou, 570228, China

[2]Hainan Blockchain Technology Engineering Research Center, Hainan University, Haikou, 570228, China

[3]School of Cyberspace Security, Hainan University, Haikou, 570228, China

[4]Department of Computer Science, Texas Tech University, Lubbock, 79409, USA

*Corresponding Author: Xiaolong Chen. Email: chen04201997@163.com

Received: 18 June 2023 Accepted: 14 September 2023 Published: 29 November 2023

ABSTRACT

In the realm of Multi-Label Text Classification (MLTC), the dual challenges of extracting rich semantic features from text and discerning inter-label relationships have spurred innovative approaches. Many studies in semantic feature extraction have turned to external knowledge to augment the model's grasp of textual content, often overlooking intrinsic textual cues such as label statistical features. In contrast, these endogenous insights naturally align with the classification task. In our paper, to complement this focus on intrinsic knowledge, we introduce a novel Gate-Attention mechanism. This mechanism adeptly integrates statistical features from the text itself into the semantic fabric, enhancing the model's capacity to understand and represent the data. Additionally, to address the intricate task of mining label correlations, we propose a Dual-end enhancement mechanism. This mechanism effectively mitigates the challenges of information loss and erroneous transmission inherent in traditional long short term memory propagation. We conducted an extensive battery of experiments on the AAPD and RCV1-2 datasets. These experiments serve the dual purpose of confirming the efficacy of both the Gate-Attention mechanism and the Dual-end enhancement mechanism. Our final model unequivocally outperforms the baseline model, attesting to its robustness. These findings emphatically underscore the imperativeness of taking into account not just external knowledge but also the inherent intricacies of textual data when crafting potent MLTC models.

KEYWORDS

Multi-label text classification; feature extraction; label distribution information; sequence generation

1 Introduction

Today, Artificial Intelligence technology is in the ascendant, and Natural Language Processing (NLP) is also growing rapidly. In the era of big data explosion, text classification, as one of the fundamental tasks in the field of NLP, has received a lot of attention based on the urgent demand of human beings for efficient text information processing techniques. Text classification [1] refers to classifying a given text according to a preset label. This text can be a sentence, a paragraph, or even a document. Text classification is also an important part of docking downstream tasks such as information retrieval [2], topic division [3], and question-answering systems [4] in the field of NLP. As

one of the complex scenarios in text classification, Multi-Label Text Classification (MLTC) needs to take into account the correlation between text feature extraction and mining labels.

In recent years, with the introduction of the Sequence Generation Model (SGM) [5], the research paradigm of Sequence-to-Sequence has been widely adopted in the field of MLTC. In this framework, the model is split into two parts: Encoder and Decoder. The Encoder module is dedicated to extracting semantic features, and the Decoder module is dedicated to mining the correlations between labels and classifying them. Currently, there is a growing body of research on semantic feature extraction to enhance the model's understanding of text by introducing exogenous knowledge. But the problem with such exogenous knowledge is that it inevitably brings noise along with new knowledge to the model. If the noise is not handled properly, it can backfire. However, this problem can be effectively solved by exploiting some inherent and intrinsic information of the text itself, such as statistical features. Compared with exogenous knowledge, this endogenous knowledge has the advantage of being naturally compatible with the corresponding classification tasks [6]. However, there are incompatibility issues between statistical features and semantic features in terms of scale and dimensionality, and not all the information in statistical features is worthy to be referenced by semantic features, so a high-quality fusion strategy is needed to combine the two features. In addition, sequence generation models are often used to mine the correlations between labels. SGM proposes that the multi-label classification problem can be transformed into a sequence generation problem, to effectively mine the correlations between labels. However, there are problems of information loss and wrong propagation [7] when decoding text feature vectors in this way, which brings certain troubles for the model to continuously generate correct labels.

To address the above issues, we propose a VFS model composed of V-Net, F-Net and S-Net, where V-Net refers to the Variational Encoding Network, F-Net refers to Feature Adaptive Fusion Network, and S-Net refers to the Sequence Enhancement Generation Model. We draw inspiration from Adaptive Gate Network (AGN) [6] and design V-Net and F-Net, which can better adapt to MLTC tasks. In V-Net, we reconstruct the original label statistical features and map them into a continuous vector space, which also address the problem of the mismatch between the original label statistical features and the semantic feature dimensions. In the F-Net module, we propose a Gate-Attention mechanism to enable statistical and semantic features to be fused across scales, and to reallocate attention weights during fusion, allowing statistical information that is not worth learning from the current semantic features to be released weights to more important statistical information. Compared to other fusion strategies, the Gate-attention mechanism enables the model to autonomously discern information from statistical features, thus reducing noise interference. In the S-Net module, we proposed a Dual-end enhancement mechanism, which introduces original hidden vectors to the input end of Long Short Term Memory (LSTM) cells for reference, and uses an attention mechanism to enhance the weight of important information on the output end, effectively alleviating the problem of information loss and error transmission during LSTM propagation. The main contributions of this paper are as follows:

- We propose a novel label distribution information extraction module, which can fully capture the mapping relationship between label and text, and thus form a unique distributed representation of text.
- We design a feature fusion strategy, which integrates the label distribution information into the original semantic feature vector of the text based on the attention mechanism.
- A large number of experiments have been carried out on two datasets, and the experimental results fully prove the effectiveness of our proposed framework.
- We propose a novel label sequence generation module, which transforms the multi-label classification problem into a label sequence generation problem and fully exploits the correlation between labels.

2 Related Work

2.1 Feature Extraction

Extracting and fusing features from multiple views can help models understand the text from multiple perspectives and at a deeper level, which is also a mainstream idea in current feature extraction research. Currently, most scholars rely on information other than the input text to assist model understanding of semantics, such as Chinese Pinyin, Chinese radicals, and English parts of speech. In Chinese, Liu et al. [8] used the pinyin of Chinese characters to assist the model in understanding Chinese, while Tao et al. [9] used the association of Chinese characters to obtain information that can assist the model in understanding the text. Liu et al. [10] also fused the three characteristics of Chinese characters: font shape, font sound, and font meaning. Hong et al. [11] even calculated the similarity between characters by using strokes and sounds based on characters. In English, Li et al. [6] designed a statistical information vocabulary based on the part of speech of English words, and used it to complete deep level feature extraction of text. In addition to these, Chen et al. [12] introduced conceptual information and entity links from the knowledge base into the model pipeline through an attention mechanism. Li et al. [13] combined domain knowledge and dimension dictionaries to generate word-level sentiment feature vectors. Zhang et al. [14] improved fine-grained financial sentiment analysis tasks by combining statistical distribution methods with semantic features. Li et al. [15] improved emotion-relevant classification tasks by combining fine-grained emotion concepts and distribution learning. Li et al. [16] enabled the extraction of global semantics at both token-level and document-level by redesigning the self-attention mechanism and recurrent structure. Li et al. [17] addressed the challenge of potential inter-class confusion and noise caused by using coarse-grained emotion distribution by generating fine-grained emotion distributions and utilizing them as model constraints. However, these efforts rarely focus on the necessity and compatibility of adding information, so it is impossible possible to avoid bringing noise while bringing new knowledge to the model.

2.2 Multi-Label Text Classification

There are two types of solutions for mining the association between labels. One is based on problem transformation, which mainly transforms the data of the problem and ultimately makes it applicable to existing algorithms designed for single label classification. For example, the Binary Relevance (BR) algorithm was proposed by Boutell et al. [18], but due to not mining the correlation between labels, the classification efficiency is low. Thereafter, Read et al. [19] proposed a Classifier Chain (CC) to address this drawback. This model links all the classifiers that come before it in a chain, allowing a single trainer to train on the input space and classifiers in the chain. The Label Powerset (LP) algorithm proposed by Tsoumakas et al. [20] converts all different subsets of category labels into different categories for training. The other category is based on applicable algorithms. This category of algorithms is mainly an improvement over existing algorithms designed for single label classification, making them applicable to MLTC. Chen et al. [21] proposed a model that extracts text feature vectors from text using Convolutional Neural Network (CNN), and then sends these vectors to a Recurrent Neural Network (RNN) to output labels, named CNN-RNN. Yang et al. [5] proposed the SGM model by introducing the attention mechanism into the Sequence-to-Sequence (Seq2Seq) model and applying it to MLTC. Later, Yang et al. [22] made improvements to SGM by adding a Set Decoder module to reduce the impact of incorrect labels. Chen et al. [23] designed a MLTC model with Latent Word-Wise Label Information (MLC-LWL) to eliminate the effects of predefined label order and exposure bias in the Sequence-to-Set (Seq2Set). In terms of classification performance, models such as Seq2Seq are more advantageous.

3 Model

In this section, we will introduce the implementation details of the VFS model in detail. The overall framework is shown in Fig. 1.

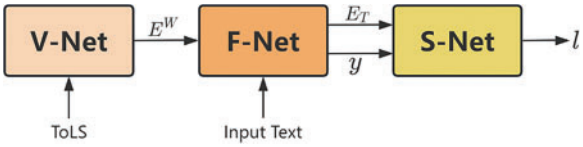

Figure 1: The overall framework of the proposed VFS

3.1 Problem Definition

MLTC refers to finding a matching subset of a text in a label set. Mathematically, give a set of text samples $T = \{t_1, t_2, \ldots, t_m\}$, and a set of labels $L = \{l_1, l_2, \ldots, l_n\}$, the goal is to learn a mapping function $\mathrm{f}: \mathrm{T} \rightarrow 2^{\mathrm{L}}$, where 2^L represents the power set of L, which contains all possible label combinations. For each text sample $\mathrm{f}: \mathrm{T} \rightarrow 2^{\mathrm{L}}$, the function f predicts a set of labels $\mathrm{f}: \mathrm{T} \rightarrow 2^{\mathrm{L}}$, which may contain zero or more labels.

3.2 V-Net: Variational Encoding Network

Due to the discrete nature of the initial label statistical features in the vector space, it is difficult to represent the statistical features in depth, and their dimensions do not match the semantic features. Therefore, we designed V-Net to reconstruct the original label statistical features to obtain a statistical feature that matches the semantic feature dimension and has deep level information. The frame diagram is shown in Fig. 2.

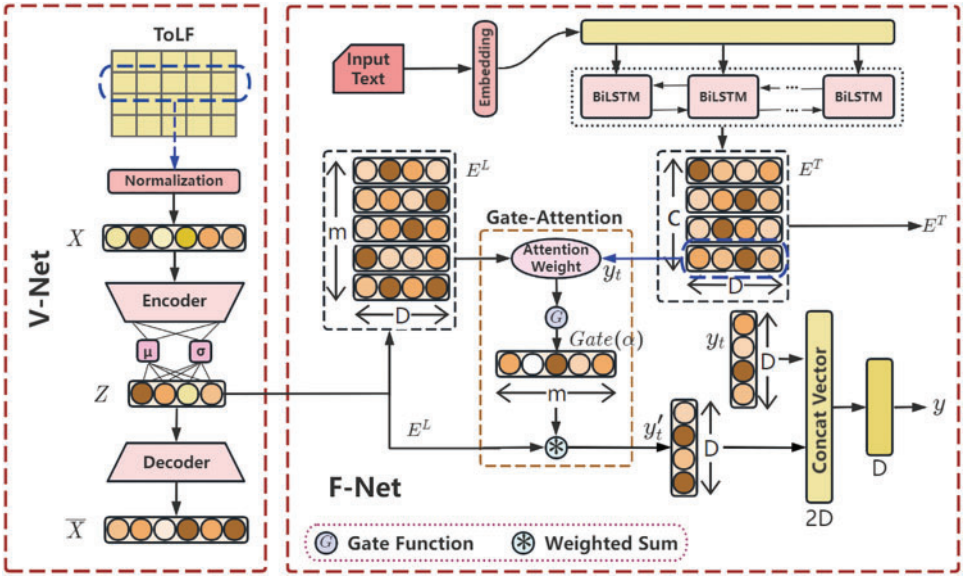

Figure 2: V-Net and F-Net frame diagram

The contribution of different words in a text to the semantics of the text varies, and the contribution of a word in different texts to the semantics of the text may also differ. Some words in the

text are associated with the corresponding labels of the text, which means that when the probability of a word appearing on a label is high or low, the word can be considered to contribute significantly to the label classification of the sentence. We first define a text $T_i = \{w_1, w_2, \ldots, w_c\}$ with a length of c, which corresponds to a set $L_i = \{l_1, l_2, \ldots, l_d\}$ containing d labels. After stacking the statistics in order, we can obtain a Table of Label Frequency (ToLF) corresponding to all words. We can obtain a vector $\xi^w = [\xi_1, \xi_2, \ldots, \xi_n]$ representing a word and a vector $\zeta^l = [\zeta_1, \zeta_2, \ldots, \zeta_m]$ representing a label from ToLS, where n and m both represent dimensions.

Not all high-frequency words contribute significantly to the semantics of a text, so we will first filter these words. We believe that a word with semantic contribution should have a normal distribution over all texts, so words that do not belong to the normal distribution will be filtered out by us first and will not be used subsequently.

The vector dimensions of the original statistical features do not match the semantic features of the text, and the vectors constructed based on this positional relationship are difficult to represent fine-grained semantics. For this reason, we use an Auto-Encoder to reduce the dimension of the original distributed representation vector. However, in order to make the distribution of the feature vectors more consistent with the real scene and reduce the interference of noise, we use a Variational Auto-Encoder (VAE) [24] to achieve this process.

If the statistical vector of a label is known to be $\zeta^l = [\zeta_1, \zeta_2, \ldots, \zeta_m]$, then the statistical matrix of all labels is $\zeta^L = [\zeta^{l_1}, \zeta^{l_2}, \ldots, \zeta^{l_n}]$, where n represents the dimension of $\zeta^L \in \mathbb{R}^{m \times n}$. Unlike ordinary Auto-Encoder, VAE becomes a model that fits the probability distribution. Assuming that the intermediate vector z follows a standard multivariate Gaussian distribution, I represents the identity matrix, and the calculation process is shown in formula (1):

$$p(z) = N(0, I) \tag{1}$$

So for VAE, the encoder samples an intermediate vector z from the prior distribution $p(z)$, and then the decoder samples the $\overline{X}$ from the posterior distribution $p(X|z)$ according to the intermediate vector z. In order to facilitate the learning and training of the neural network, θ is parameterized, then the calculation process of the decoder model is shown in formula (2):

$$p_\theta(X|z) = N\left(\mu', \sigma'^2 * I\right) = N\left(X|\mu'(z;\theta), \sigma'^2(z;\theta) * I\right) \tag{2}$$

where μ represents the mean and σ represents the standard deviation. For the encoder, its task is mainly to fit a distribution $p_\theta(X)$ close to the real distribution $p(X)$. Then $p_\theta(X)$ is:

$$p_\theta(X) = \int_z p_\theta(X|z)\, p(z)\, dz \approx \frac{1}{m} \sum_{j=1}^{m} p_\theta\left(X|z_j\right) \tag{3}$$

However, if a large number of z_j are sampled from $p(z)$ to obtain $p_\theta(X)$, the requirements for the vector dimensions of X and Z are too high, which is not suitable for neural network training. Therefore, we can assume a posterior distribution $p_\theta(z|X)$, and get $p_\theta(z|X)$ according to the Bayesian formula:

$$p_\theta(z|X) = \frac{p_\theta(X|z)\, p(z)}{p_\theta(X)} = \frac{p_\theta(X|z)\, p(z)}{\int_{\hat{z}} p_\theta\left(X|\hat{z}\right) p\left(\hat{z}\right) d\hat{z}} \tag{4}$$

However, for the denominator in the above formula, it is still necessary to sample a large number of z_j from $p(z)$, so the Φ parameterized encoder is fit the $p_\Phi(z|X)$ distribution to approximate the $p_\theta(z|X)$ distribution. In addition, because $p_\theta(X|z)$ and $p(z)$ both obey the multivariate Gaussian distribution, it can be obtained that the posterior distribution $p_\theta(z|X)$ also obeys the multivariate Gaussian distribution. So:

$$p_{\Phi}(z|X) = N\left(z|\mu(X;\Phi), \sigma^2(X;\Phi) * I\right) \tag{5}$$

However, the neural network cannot backpropagate the sampling function when training the model through the loss function, so it is necessary to sample an e_i from the standard multivariate Gaussian distribution $N(0, I)$ first, and then calculate z_i:

$$z_i = \mu_i + \sigma_i \odot e_i \tag{6}$$

where $\odot$ represents the element-wise product operation.

This module is trained independently in the entire model, and only the intermediate vector Z needs to be taken out for subsequent use in this paper. The input to the VAE is ζ^t, so that the feature matrix $E^L \in \mathbb{R}^{m \times D}$ representing the label can be obtained, where D represents the reconstructed statistical feature vector dimension.

3.3 F-Net: Feature Adaptive Fusion Network

The F-Net needs to complete the extraction of text semantics and fuse it with statistical features from the V-Net. Due to the scale incompatibility between statistical and semantic features, and the presence of noise in statistical features. To this end, we designed a Gate-Attention mechanism to assign weights to statistical features and filter them. After weighted summation, we obtain a feature vector that can represent the text with high quality. Finally, we will perform vector stitching on both. The frame diagram is shown in Fig. 2.

First of all, we extract feature vectors of the input text via a bidirectional LSTM [25]. Timing data w_t of the t-th time step will be passed into two LSTM units. Therefore, we can obtain hidden vectors from both directions of output:

$$\overrightarrow{y_t} = LSTM\left(\overrightarrow{y_{t-1}}, w_t\right) \tag{7}$$

$$\overleftarrow{y_t} = LSTM\left(\overleftarrow{y_{t+1}}, w_t\right) \tag{8}$$

Therefore, we can obtain the final hidden representation of the t-th time step by concatenating the hidden states from both directions, $y_t = [\overrightarrow{y_t}; \overleftarrow{y_t}]$ and the future matrix of the entire text, $E^T = [y_1, y_2, \ldots, y_c]$, where c denotes the length of text, $y_c \in \mathbb{R}^{1 \times D}$ denotes the last hidden state, and D denotes the dimension semantic features.

We propose a Gate-Attention mechanism that combines statistical features with semantic features. We regard y_c as *query*, E^L as *key* and *value* at the same time to implement the attention mechanism. First, We obtain the attention weight for each $e^l_\epsilon \in E^L$, where $E^L = [e^l_1, e^l_2, \ldots, e^l_m]$ and e^l_ϵ denotes the ϵ-th vector in E^L $(1 \le \epsilon \le m)$.

$$\alpha' = \left[\alpha'_1, \ldots, \alpha'_\epsilon, \ldots, \alpha'_m\right], \alpha'_\epsilon = f\left(y_c, e^l_\epsilon\right) \tag{9}$$

where $\alpha' \in \mathbb{R}^{1 \times m}$ denotes the attention weight. Besides, f denotes the distance function which is stated as an element-wise dot product operation in this paper. Then, we obtain $\alpha = [\alpha_1, \ldots, \alpha_\epsilon, \ldots, \alpha_m]$ via to normalize α' with the *softmax* function:

$$\alpha_\epsilon = \frac{\exp\left(\alpha'_\epsilon\right)}{\sum_{i=1}^{m} \exp\left(\alpha'_i\right)}, where \sum_{\epsilon=1}^{m} \alpha_\epsilon = 1 \tag{10}$$

However, in order to reduce the impact of irrelevant labels in understanding text, we have designed a gate mechanism. Under this mechanism, labels whose contribution cannot reach the threshold will be released with a weight, and this weight will be assigned to other labels.

$$Gate\,(\alpha_\epsilon) = \begin{cases} (\exp(\gamma)+1)\alpha_\epsilon, & \alpha_\epsilon \geq sigmoid(\vartheta) \\ 0, & other \end{cases} \tag{11}$$

where γ and ϑ both denote hyper-parameters. Besides, $sigmoid(\vartheta)$ denotes the threshold value at which the contribution meets the requirements and $\exp(\gamma)$ denotes the compensation of the model for satisfying the statistical future. Finally, *Gate* denotes the gate function as a filter to extract necessary information.

Afterward, we can obtain the attentive representation y_c' through attentive weighted sum as:

$$y_c' = \sum_{\epsilon=1}^{m} Gate(\alpha_\epsilon) e_\epsilon^l \tag{12}$$

where α_ϵ denotes the ϵ-th dimensional value of $\alpha \in \mathbb{R}^{1\times m}$ $(1 \leq \epsilon \leq m)$.

Thereafter, in order to systematically integrate the vectors about text representation obtained by these two methods, y_C and y_C' are concatenated.

$$Y = concat\left(y_c, y_c'\right) \tag{13}$$

where $Y \in \mathbb{R}^{1\times 2D}$ represents the direction after concatenating, and this has the advantage of retaining all information [26]. Then, the potential correlation between y_c and y_c' is learned through a fully connected layer neural network, and its dimension is reduced to D:

$$y = FC\,(Y) \tag{14}$$

3.4 S-Net: Sequence Enhancement Generation Network

After obtaining the feature vector y containing statistical information, it is necessary to parse it through LSTM and assign appropriate labels. To address the problem of error transmission and information loss during LSTM parsing, we designed a Dual-end enhancement mechanism to enhance the information at both the input and output ends of LSTM. The overall structure of the model is shown in Fig. 3.

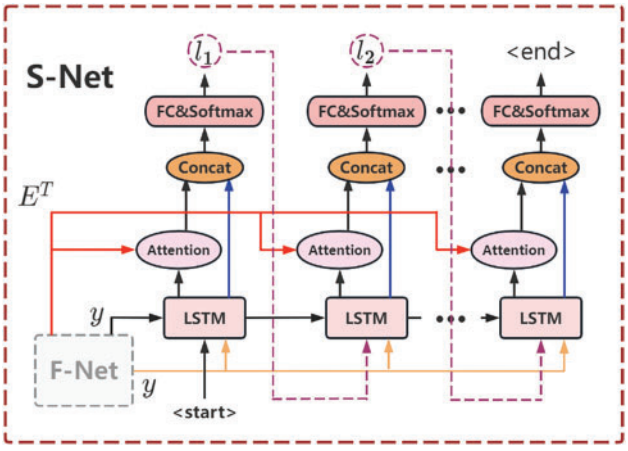

Figure 3: S-Net structure diagram

First, we will equally share the feature vector y from the F-Net with each LSTM unit, which can reduce the erroneous impact of hidden information from the previous layer.

$$h_t = LSTM\,(h_{t-1}, L_{t-1}, y) \tag{15}$$

where L_{t-1} denotes an embedded representation of the label output from the previous layer, t denotes the t-th time step.

Second, we also enhanced the output of each LSTM unit. We use the Attention mechanism to refer different labels to different important words. This model will be used to obtain the future matrix E^T from F-Net as *query* and *value*, hidden state h_t of improved LSTM unit as the *key*. Therefore, we can obtain the attention weight representation β_t:

$$\beta_t = Softmax\left(\left(E^T\right)^T h_t\right) \tag{16}$$

where $E^T \in \mathbb{R}^{c\times D}$ needs to be transposed first. Afterward, we can obtain the attentive representation h'_t through attentive weighted sum as:

$$h'_t = \sum_{i=1}^{c} \beta_{t(i)} y_i \tag{17}$$

We will concatenate the h'_t calculated by the Attention mechanism and the hidden state h_t of the LSTM output:

$$H_t = concat\left(h'_t, h_t\right) \tag{18}$$

Compared to h_t, H_t increases the reference to important words in the understanding of labels, which can reduce the impact of insufficient information transmission at the upper level. After that, H_t is passed into the fully connected neural network to further learn the deep connection between h'_t and h_t, and the corresponding label is output through the *softmax* function.

4 Structure

4.1 Dataset Description

This experiment uses two publicly available English datasets, AAPD and RCV1-2, to train and test the model. Each dataset will be divided into three parts: training set, verification set, and test set. The AAPD dataset is a collection of 55840 abstracts and corresponding subject categories collected and collated by Li et al. [6] on the internet, with a total of 54 labels, which can predict the corresponding subject of academic papers based on a given summary. The RCV1-2 dataset is from a Reuters news column, compiled and collected by Lewis et al. [27]. With a total of 804414 news stories, each news story is assigned multiple themes, with a total of 103 themes. The details of the two datasets are given in Table 1. Including N_{train} training set is the total number of samples, N_{test} is testing samples, total L is the total number of labels, $\overline{L}$ is label number, average every sample have $\hat{L}$ is average each label has a label number, W_{train} is the average number of words, each training set sample W_{test} is test sample average word count.

Table 1: Details of the datasets

Dataset	N_{train}	N_{test}	L	$\overline{L}$	$\hat{L}$	W_{train}	W_{test}
AAPD	54840	1000	54	2.41	2444.04	163.42	171.65
RCV1-2	781265	23149	103	3.18	729.67	259.47	269.23

To test the effect of the model on texts with different numbers of labels, the label distributions of AAPD and RCV1-2 were also calculated, and the results were shown in Fig. 4.

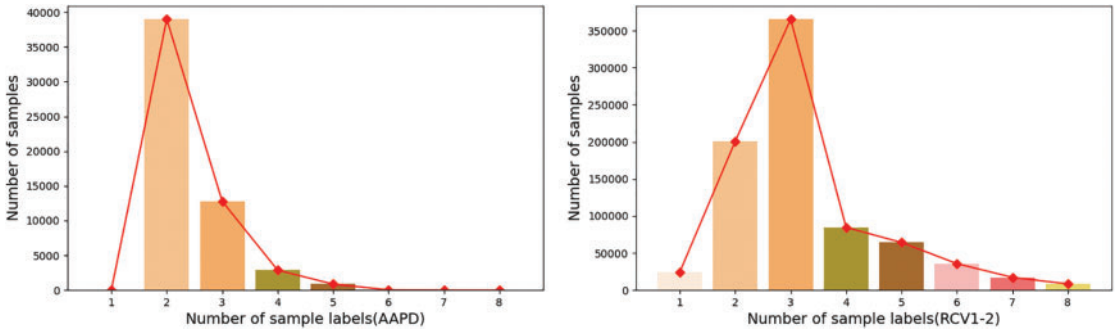

Figure 4: Dataset label distribution

4.2 Experimental Details

We set the sample length of the training set to 500, fill in <pad> if this is not enough, and cut the rest The AAPD vocabulary is 30,000 in length and the RCV1-2 vocabulary is 50,000 in length. The word embedding dimension D is set to 256, the length of the V-Net intermediate vector is set to 256, the length of Bi-LSTM for the F-Net is set to 500, and the length of LSTM for the S-Net is set to 10. To prevent overfitting, the dropout mechanism is used with the drop rate of 0.5. Adam optimizer was used, and the learning rate was 0.001. Finally, the V-Net is trained separately and the results are screened for subsequent use.

4.3 Comparison Methods

We compare our proposed method with the following baselines:

- **BR [13]:** This method converts multi-label classification into multiple binary classification tasks and trains the binary classifier for each label.
- **CC [14]:** This method converts multi-label classification into a chain binary problem.
- **LP [15]:** Treats each label combination as a new class and transforms the MLTC problem into a multi-class classification.
- **CNN-RNN [16]:** The model uses CNN to capture local features of text, RNN to capture global features, and finally fuses into a feature vector that contains both types of information.
- **SGM [5]:** The method is a sequentially generated model that uses the LSTM-based Seq2Seq model with an attention mechanism, while the decoding phase uses global embedding to obtain inter-label dependencies.
- **SGM with Global Embedding (SGM-GE) [5]:** Employs the same sequence-to-sequence model as SGM with a novel decoder structure to tackle the MLTC problem.
- **Seq2Set [17]:** Improvements have been made to SGM, including a Set Decoder module to reduce the impact of mislabeling.
- **Multi-Label Reasoner (ML-Reasoner) [28]:** This model designs a multi label classification algorithm based on reasoning, reducing the dependence of the model on label order.
- **Seq2Seq Model with a Different Label Semantic Attention Mechanism (S2S-LSAM) [29]:** This model generates fusion information containing label and text information through the interaction between label semantics and text features in the label semantic attention mechanism.
- **Spotted Hyena Optimizer with Long Short Term Memory (SHO-LSTM) [30]:** The Spotted Hyena Optimizer algorithm is used to optimize the LSTM network.
- **MLC-LWL [18]:** This model uses the topic model of labels to construct effective word-by-word label information and combines the label information carried by words with the label context information through a gated network.

- **Label-Embedding Bi-Directional Attentive (LBA) [31]:** The paper proposes a Label-Embedding Bi-Directional Attentive model by fully leveraging fine-grained token-level text representations and label embeddings.
- **Counter Factual Text Classifier (CFTC) [32]:** The paper achieves causality-based predictions by eliminating correlation bias in MLTC tasks, significantly improving the model's performance, and effectively eliminating correlation bias in the datasets.

4.4 Experimental Results

4.4.1 Comparative Experiments

We compared the proposed the VFS model with all baseline models on the AAPD dataset and the RCV1-2 dataset, and the results are shown in Table 2. The results show that our proposed model has achieved excellent performance, with the best performance in three indicators. On the AAPD dataset, our proposed the VFS model achieves a reduction of 5.55% hamming-loss and an improvement of 1.41% micro-F_1 score over the best model MLC-LWL in baselines. Although our model is 7.03% micro-precision score less than MLC-LWL, achieves an improvement of 2.17% over the model SHO-LSTM. We get the results of the proposed method and the baselines on the RCV1-2 test set. Similar to the experimental results on the AAPD test set, the VFS model achieves a reduction of 8.22% hamming-loss and an improvement of 0.68% micro-F_1 score over the model MLC-LWL. Based on these results, the significant advantages of our proposed model can be fully demonstrated. Where HL, P, R and F1 denote hamming-loss [33], micro-precision, micro-recall and micro-F_1 [34]. In addition, the symbol "+" denotes that the higher the value is, the better the model performs. The symbol "−" and the symbol "+" indicate opposite meanings.

Table 2: Comparison between our methods and all baselines on two datasets

Models	AAPD				RCV1-2			
	HL (−)	P (+)	R (+)	F1 (+)	HL (−)	P (+)	R (+)	F1 (+)
BR	0.0316	0.644	0.648	0.646	0.0086	0.904	0.816	0.858
CC	0.0306	0.657	0.651	0.654	0.0087	0.887	0.828	0.857
LP	0.0312	0.662	0.608	0.634	0.0087	0.896	0.824	0.858
CNN-RNN	0.0278	0.718	0.618	0.664	0.0085	0.889	0.825	0.856
SGM	0.0251	0.746	0.659	0.699	0.0081	0.887	0.850	0.869
SGM-GE	0.0245	0.748	0.675	0.710	0.0075	0.897	0.860	0.878
Seq2Set	0.0247	0.739	0.674	0.705	0.0073	0.900	0.858	0.879
ML-Reasoner	0.0238	0.761	0.684	0.720	0.0081	**0.912**	0.847	0.878
S2S-LSAM	0.0238	0.762	0.673	0.715	0.0072	0.893	**0.869**	0.881
SHO-LSTM	0.0248	0.737	0.684	0.710	0.0079	0.897	0.862	0.880
MLC-LWL	0.0234	**0.810**	0.633	0.711	0.0073	0.910	0.854	0.881
LBA	0.0228	0.774	0.669	0.718	0.0073	0.900	0.859	0.880
CFTC	0.0237	0.770	0.666	0.714	0.0074	0.893	0.861	0.880
VFS (ours)	**0.0221**	0.753	**0.691**	**0.721**	**0.0067**	0.906	**0.869**	**0.887**

4.4.2 Ablation Experiment

In addition, we used the classic model SGM in the field of MLTC as the baseline model, and compared the Encoder of the SGM model with VF and Decoder with S-Net, respectively. The results are shown in Fig. 5. From the figure, it can be seen that replacing Encoder with VF and Decoder with S-Net can improve the effect of the SGM model, and the combination of VF and S-Net has the best effect. This fully demonstrates the respective effectiveness of VF and S-Net.

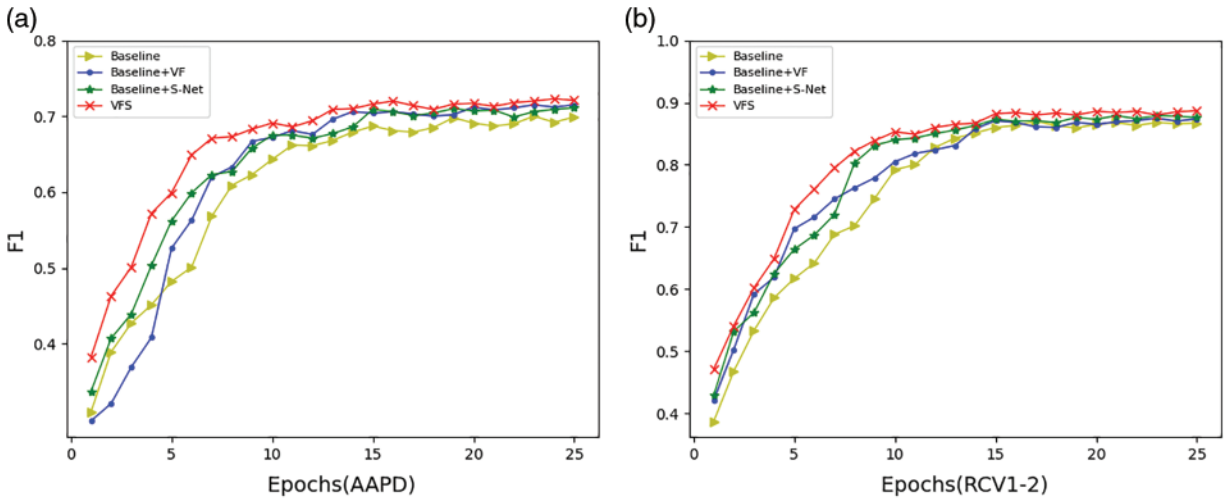

Figure 5: Comparison diagram of ablation experiment

4.4.3 Analysis of Label Length Impact

In order to explore the impact of label length on experimental results, we selected samples with label lengths of 2 to 7 from the RCV1-2 test set and tested them on models SGM and VFS, respectively. The results are shown in Fig. 6. From the figure, it can be seen that both models achieve optimal results when the label length is 3, whether it is HL or F1. Since then, as the label length increases, the model effect has become worse, indicating that the longer the label length, the greater the difficulty of classification. However, it can also be seen that the performance degradation of the VFS model is lower than that of SGM when faced with an increase in labels. This indicates that the VFS has better robustness than SGM.

4.4.4 Analysis of Attention Weight Distribution

The S-Net model can allow words that contribute more to the semantics of text to receive more attention and give them greater weight. At the same time, the weight can also reflect differences when faced with different labels. The thermal distribution table of the attention weight section is shown in Table 3. From Table 3, it can be seen that when the VFS model predicts the "cs.CV" label, the words "visual" and "movie" have gained more attention from the model, while when predicting the "cs.CL" label, the words "presence", "LSTM", and "verb" have gained more attention from the model. This shows that our proposed model can automatically assign greater weight to words that can contribute more semantic information, and there are differences in the consideration of different labels and key words in the text.

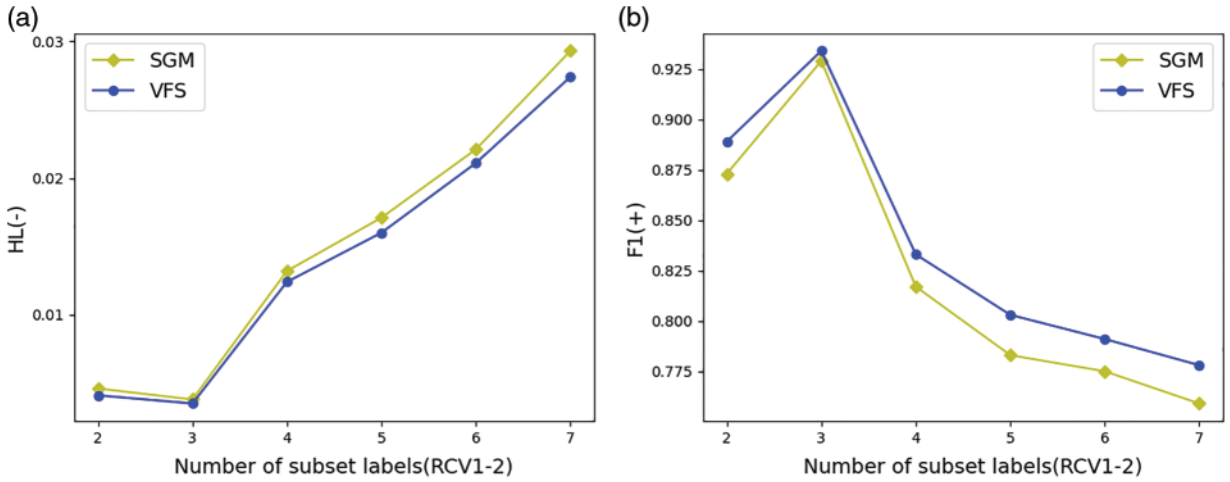

Figure 6: Comparison of effects on labels of different lengths

Table 3: Visualization of attention weight distribution

cs.CV	cs.CL
We show how to learn robust visual classifiers from the weak annotations of the sentence descriptions based on these visual classifiers.	We show how to learn robust visual classifiers from the weak annotations of the sentence descriptionss based on these visual classifiers.
We learn how to generate a description using an LSTM. We explore different design choices to build and train the LSTM.	We learn how to generate a description using an LSTM. We explore different design choices to build and train the LSTM.
We argue that it is important to distinguish verbs, objects, and places in the challenging setting of movie description.	We argue that it is important to distinguish verbs, objects, and places in the challenging setting of movie description.

5 Conclusion

In this paper, we propose a novel fusion strategy that combines statistical features with semantic features in a high-quality manner to solve the problem of mismatching between statistical and semantic features in terms of scale and dimension. Secondly, we propose an information enhancement mechanism to effectively alleviate the problems of information loss and incorrect transmission in LSTM networks. A large number of experimental results show that our proposed model is significantly superior to the baseline. Further analysis shows that our model can effectively capture the semantic contributions of important words. In future work, we plan to explore additional types of statistical features and apply them to tasks such as named entity recognition and even image classification. Although our proposed model can alleviate the impact of the increase in the number of labels to some extent, it is still difficult to cope with the prediction task of a large number of labels. Further exploration is needed in this area in the future.

Acknowledgement: None.

Funding Statement: This work was supported by National Natural Science Foundation of China (NSFC) (Grant Nos. 62162022, 62162024), the Key Research and Development Program of Hainan Province (Grant Nos. ZDYF2020040, ZDYF2021GXJS003), the Major Science and Technology Project of Hainan Province (Grant No. ZDKJ2020012), Hainan Provincial Natural Science Foundation of China (Grant Nos. 620MS021, 621QN211), Science and Technology Development Center of the Ministry of Education Industry-University-Research Innovation Fund (2021JQR017).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Xiaolong Chen, Jieren Cheng; data collection: Xiaolong Chen, Wenghang Xu, Shuai Hua; analysis and interpretation of results: Xiaolong Chen, Zhu Tang; draft manuscript preparation: Xiaolong Chen, Victor S. Sheng. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available on request from the corresponding author, Xiaolong Chen, upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Chen, J. Cheng, J. Liu, W. Xu, S. Hua *et al.,* "A survey of multi-label text classification based on deep learning," in *8th Int. Conf. on Adaptive and Intelligent Systems (ICAIS)*, Qinghai, China, pp. 443–456, 2022.

[2] Y. Chen, X. Qi and P. Zhang, "Replica wormhole and information retrieval in the SYK model coupled to Majorana chains," *Journal of High Energy Physics*, vol. 2020, no. 6, pp. 1–26, 2020.

[3] J. Chen, H. Li, L. Ma and H. Bo, "Improving emotion analysis for speech-induced EEGs through EEMD-HHT-based feature extraction and electrode selection," *International Journal of Multimedia Data Engineering and Management*, vol. 2, no. 12, pp. 1–18, 2021.

[4] J. Gomes, R. C. Mello, V. Ströel and J. F. Souza, "A hereditary attentive template-based approach for complex knowledge base question answering systems," *Expert Systems with Applications*, vol. 205, no. 6, pp. 117725, 2022.

[5] P. Yang, X. Sun, W. Li, S. Ma, W. Wu *et al.,* "SGM: Sequence generation model for multi-label classification," in *27th Int. Conf. on Computational Linguistics (COLING'18)*, Santa Fe, NM, USA, pp. 3915–3926, 2018.

[6] X. Li, Z. Li, H. Xie and Q. Li, "Merging statistical feature via adaptive gate for improved text classification," in *35th Proc. of the AAAI Conf. on Artificial Intelligence (AAAI'21)*, Virtual Event, pp. 13288–13296, 2021.

[7] Y. Xiao, Y. Li, J. Yuan, S. Guo, Y. Xiao *et al.,* "History-based attention in Seq2Seq model for multi-label text classification," *Knowledge-Based Systems*, vol. 224, pp. 107094, 2021.

[8] S. Liu, T. Yang, T. Yue, F. Zhang and D. Wang, "PLOME: Pre-training with misspelled knowledge for Chinese spelling correction," in *59th Annual Meeting of the Association for Computational Linguistics (ACL'21)*, Bangkok, Thailand, pp. 2991–3000, 2021.

[9] H. Tao, S. Tong, K. Zhang, T. Xu, Q. Lu *et al.,* "Ideography leads us to the field of cognition: A radical-guided associative model for Chinese text classification," in *35th AAAI Conf. on Artificial Intelligence (AAAI'21)*, Virtual Event, vol. 35, pp. 13898–13906, 2021.

[10] J. Liu, J. Cheng, X. Peng, Z. Zhao, X. Tang *et al.,* "Multi-view semantic feature fusion model for Chinese named entity recognition," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 6, pp. 1833–1848, 2022.

[11] Y. Hong, X. Yu, N. He, N. Liu and J. Liu, "FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm," in *5th Workshop on Noisy User-Generated Text at EMNLP 2019(WNUT'19)*, Hong Kong, China, pp. 160–169, 2019.
[12] J. Chen, Y. Hu, J. Liu, Y. Xiao and H. Jiang, "Deep short text classification with knowledge powered attention," in *33th Proc. of the AAAI Conf. on Artificial Intelligence (AAAI'19)*, Honolulu, HI, USA, pp. 6252–6259, 2019.
[13] Z. Li, H. Xie, G. Cheng and Q. Li, "Word-level emotion distribution with two schemas for short text emotion classification," *Knowledge-Based Systems*, vol. 227, pp. 107163, 2021.
[14] H. Zhang, Z. Li, H. Xie, R. Lau, G. Cheng *et al.,* "Leveraging statistical information in fine-grained financial sentiment analysis," *World Wide Web*, vol. 25, no. 2, pp. 513–531, 2021.
[15] Z. Li, X. Li, H. Xie, F. Wang, M. Leng *et al.,* "A novel dropout mechanism with label extension schema toward text emotion classification," *Information Processing & Management*, vol. 60, no. 2, pp. 103173, 2023.
[16] X. M. Li, Z. X. Li, X. T. Luo, H. R. Xie, X. Lee *et al.,* "Recurrent attention networks for long-text modeling," in *61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, Toronto, Canada, pp. 3006–3019, 2023.
[17] Z. Li, X. Li, H. Xie, Q. Li and X. Tao, "A label extension schema for improved text emotion classification," in *20th IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT'21)*, Melbourne, Australia, pp. 32–39, 2021.
[18] M. R. Boutell, J. Luo, X. Shen and C. M. Brown, "Learning multilabel scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.
[19] J. Read, B. Pfahringer, G. Holmes and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.
[20] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.
[21] G. Chen, D. Ye, Z. Xing, J. Chen and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *2017 Int. Joint Conf. on Neural Networks (IJCNN'17)*, Anchorage, AK, USA, pp. 2377–2383, 2017.
[22] P. Yang, F. Luo, S. Ma, J. Lin, X. Sun *et al.,* "A deep reinforced sequence-to-set model for multi-label classification," in *57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, Florence, Italy, pp. 5252–5258, 2019.
[23] Z. Chen and J. Ren, "Multi-label text classification with latent word-wise label information," *Applied Intelligence*, vol. 51, no. 2, pp. 966–979, 2021.
[24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013. https://doi.org/10.48550/arXiv.1312.6114
[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[26] K. Zhang, G. Lv, L. Wang, L. Wu, E. Chen *et al.,* "DRr-Net: Dynamic re-read network for sentence semantic matching," in *33th Proc. of the AAAI Conf. on Artificial Intelligence (AAAI'19)*, Honolulu, HI, USA, pp. 7442–7449, 2019.
[27] D. Lewis, Y. Yang, T. Rose and F. Li, "RCV1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.
[28] R. Wang, R. Ridley, X. Su, W. Qu and X. Dai, "A novel reasoning mechanism for multi-label text classification," *Information Processing & Management*, vol. 58, no. 2, pp. 102441, 2021.
[29] X. Zhang, X. Tan, Z. Luo and J. Zhao, "Multi-label sequence generating model via label semantic attention mechanism," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 5, pp. 1711–1723, 2022.
[30] H. Maragheh, F. Gharehchopogh, K. Majidzadeh and A. Sangar, "A new hybrid based on long short-term memory network with spotted hyena optimization algorithm for multi-label text classification," *Mathematics*, vol. 10, no. 3, pp. 488, 2022.
[31] N. Liu, Q. Wang and J. Ren, "Label-embedding bi-directional attentive model for multi-label text classification," *Neural Process Letters*, vol. 53, no. 1, pp. 375–389, 2021.

[32] C. Fan, W. Chen, J. Tian, Y. Li, H. He *et al.,* "Accurate use of label dependency in multi-label text classification through the lens of causality," *Applied Intelligence*, vol. 50, no. 8, pp. 1–7, 2023. https://doi.org/10.1007/s10489-023-04623-3

[33] R. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[34] C. D. Manning, H. Schütze and P. Raghavan, "Support vector machines and machine learning on documents," in *Introduction to Information Retrieval*, 1st ed., Cambridge, UK: Cambridge University Press, Chapter 15, pp. 319–346, 2009.