Semantic-Unit-Based Dilated Convolution for Multi-Label Text Classification

Junyang Lin[1,2], Qi Su[1], Pengcheng Yang[2], Shuming Ma[2], Xu Sun[2]
School of Foreign Languages, Peking University[1]
MOE Key Lab of Computational Linguistics, School of EECS, Peking University[2]
{linjunyang, sukia, yang_pc, shumingma, xusun}@pku.edu.cn

Abstract

We propose a novel model for multi-label text classification, which is based on sequence-to-sequence learning. The model generates higher-level semantic unit representations with multi-level dilated convolution as well as a corresponding hybrid attention mechanism that extracts both the information at the word-level and the level of the semantic unit. Our designed dilated convolution effectively reduces dimension and supports an exponential expansion of receptive fields without loss of local information, and the attention-over-attention mechanism is able to capture more summary relevant information from the source context. Results of our experiments show that the proposed model has significant advantages over the baseline models on the dataset RCV1-V2 and Ren-CECps, and our analysis demonstrates that our model is competitive to the deterministic hierarchical models and it is more robust to classifying low-frequency labels[1].

1 Introduction

Multi-label text classification refers to assigning multiple labels for a given text, which can be applied to a number of important real-world applications. One typical example is that news on the website often requires labels with the purpose of the improved quality of search and recommendation so that the users can find the preferred information with high efficiency with less disturbance of the irrelevant information. As a significant task of natural language processing, a number of methods have been applied and have gradually achieved satisfactory performance. For instance, a series of methods based on machine learning have been extensively utilized in the industries, such as Binary Relevance (Boutell et al., 2004). BR treats

the task as multiple single-label classifications and can achieve satisfactory performance. With the development of Deep Learning, neural methods are applied to this task and achieved improvements (Zhang and Zhou, 2006; Nam et al., 2013; Benites and Sapozhnikova, 2015).

However, these methods cannot model the internal correlations among labels. To capture such correlations, the following work, including ML-DT (Clare and King, 2001), Rank-SVM (Elisseeff and Weston, 2002), LP (Tsoumakas and Katakis, 2006), ML-KNN (Zhang and Zhou, 2007), CC (Read et al., 2011), attempt to capture the relationship, which though demonstrated improvements yet simply captured low-order correlations. A milestone in this field is the application of sequence-to-sequence learning to multi-label text classification (Nam et al., 2017). Sequence-to-sequence learning is about the transformation from one type of sequence to another type of sequence, whose most common architecture is the attention-based sequence-to-sequence (Seq2Seq) model. The attention-based Seq2Seq (Sutskever et al., 2014) model is initially designed for neural machine translation (NMT) (Bahdanau et al., 2014; Luong et al., 2015). Seq2Seq is able to encode a given source text and decode the representation for a new sequence to approximate the target text, and with the attention mechanism, the decoder is competent in extracting vital source-side information to improve the quality of decoding. Multi-label text classification can be regarded as the prediction of the target label sequence given a source text, which can be modeled by the Seq2Seq. Moreover, it is able to model the high-order correlations among the source text as well as those among the label sequence with deep recurrent neural networks (RNN).

Nevertheless, we study the attention-based Seq2Seq model for multi-label text classification

---

[1]The code is available at https://github.com/lancopku/SU4MLC

(Nam et al., 2017) and find that the attention mechanism does not play a significant role in this task as it does in other NLP tasks, such as NMT and abstractive summarization. In Section 3, we demonstrate the results of our ablation study, which show that the attention mechanism cannot improve the performance of the Seq2Seq model. We hypothesize that compared with neural machine translation, the requirements for neural multi-label text classification are different. The conventional attention mechanism extracts the word-level information from the source context, which makes little contribution to a classification task. For text classification, human does not assign texts labels simply based on the word-level information but usually based on their understanding of the salient meanings in the source text.

For example, regarding the text "The young boys are playing basketball with great excitement and apparently they enjoy the fun of competition", it can be found that there are two salient ideas, which are "game of the young" and "happiness of basketball game", which we call "semantic units" of the text. The semantic units, instead of word-level information, mainly determine that the text can be classified into the target categories "youth" and "sports".

Semantic units construct the semantic meaning of the whole text. To assign proper labels for text, the model should capture the core semantic units of the source text, the higher-level information compared with word-level information, and then assign the text labels based on its understanding of the semantic units. However, it is difficult to extract information from semantic units as the conventional attention mechanism focuses on extracting word-level information, which contains redundancy and irrelevant details.

In order to capture semantic units in the source text, we analyze the texts and find that the semantic units are often wrapped in phrases or sentences, connecting with other units with the help of contexts. Inspired by the idea of global encoding for summarization (Lin et al., 2018), we utilize the power of convolutional neural networks (CNN) to capture local interaction among words and generate representations of information of higher levels than the word, such as phrase or sentence. Moreover, to tackle the problem of long-term dependency, we design a multi-level dilated convolution for text to capture local correlation and long-term dependency without loss of coverage as we do not apply any form of pooling or strided convolution. Based on the annotations generated by our designed module and those by the original recurrent neural networks, we implement our hybrid attention mechanism with the purpose of capturing information at different levels, and furthermore, it can extract word-level information from the source context based on its attention on the semantic units.

In brief, our contributions are illustrated below:

- We analyze that the conventional attention mechanism is not useful for multi-label text classification, and we propose a novel model with multi-level dilated convolution to capture semantic units in the source text.
- Experimental results demonstrate that our model outperforms the baseline models and achieves the state-of-the-art performance on the dataset RCV1-v2 and Ren-CECps, and our model is competitive with the hierarchical models with the deterministic setting of sentence or phrase.
- Our analysis shows that compared with the conventional Seq2Seq model, our model with effective information extracted from the source context can better predict the labels of low frequency, and it is less influenced by the prior distribution of the label sequence.

2 Attention-based Seq2Seq for Multi-label Text Classification

As illustrated below, multi-label text classification has the potential to be regarded as a task of sequence prediction, as long as there are certain correlation patterns in the label data. Owing to the correlations among labels, it is possible to improve the performance of the model in this task by assigning certain label permutations for the label sequence and maximizing subset accuracy, which means that the label permutation and the corresponding attention-based Seq2Seq method are competent in learning the label classification and the label correlations. By maximizing the subset accuracy, the model can improve the performance of classification with the assistance of the information about the label correlations. Regarding label permutation, a straightforward method is to reorder the label data in accordance with the descending order by frequency, which shows satisfactory effects (Chen et al., 2017).

Multi-label text classification can be regarded as a Seq2Seq learning task, which is formulated as below. Given a source text $x = \{x_1, ..., x_i, ..., x_n\}$ and a target label sequence $y = \{y_1, ..., y_i, ..., y_m\}$, the Seq2Seq model learns to approximate the probability $P(y|x) = \prod_{t=1}^{m} P(y_t|y_{<t}, x)$, where $P(y_t|y_{<t}, x)$ is computed by the Seq2Seq model, which is commonly based on recurrent neural network (RNN).

The encoder, which is bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1996), encodes the source text x from both directions and generates the source annotations h, where the annotations from both directions at each time step are concatenated ($h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$). To be specific, the computations of $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are illustrated below:

$$\overrightarrow{h_i} = LSTM(x_i, \overrightarrow{h_{i-1}}, C_{i-1}) \quad (1)$$

$$\overleftarrow{h_i} = LSTM(x_i, \overleftarrow{h_{i-1}}, C_{i-1}) \quad (2)$$

We implement a unidirectional LSTM decoder to generate labels sequentially. At each time step t, the decoder generates a label y_t by sampling from a distribution of the target label set P_{vocab} until sampling the token representing the end of sentence, where:

$$P_{vocab} = g(y_{t-1}, c_t, s_{t-1}) \quad (3)$$

where $g(\cdot)$ refers to non-linear functions including the LSTM decoder, the attention mechanism as well as the softmax function for prediction. The attention mechanism generates c_t as shown in the following:

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i \quad (4)$$

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{n} exp(e_{t,j})} \quad (5)$$

$$e_{t,i} = s_{t-1}^{\top} W_a h_i \quad (6)$$

3 Problem

As we analyze the effects of the attention mechanism in multi-label text classification, we find that it contributes little to the improvement of the model's performance. To verify the effects of the attention mechanism, we conduct an ablation test to compare the performance of the Seq2Seq model without the attention mechanism and the attention-based SeqSeq model on the multi-label text classification dataset RCV1-v2, which is introduced in detail in Section 5.

Models	HL(-)	P(+)	R(+)	F1(+)
w/o attention	0.0082	0.883	0.849	0.866
+attention	0.0081	0.889	0.848	0.868

Table 1: Performances of the Seq2Seq models with and without attention on the RCV1-v2 test set, where HL, P, R, and F1 refer to hamming loss, micro-precision, micro-recall and micro-F_1. The symbol "+" indicates that the higher the better, while the symbol "-" indicates that the lower the better.

As is shown in Table 1, the Seq2Seq models with and without the attention mechanism demonstrate similar performances on the RCV1-v2 according to their scores of micro-F_1, a significant evaluation metric for multi-label text classification. This can be a proof that the conventional attention mechanism does not play a significant role in the improvement of the Seq2Seq model's performance. We hypothesize that the conventional attention mechanism does not meet the requirements of multi-label text classification. A common sense for such a classification task is that the classification should be based on the salient ideas of the source text. The semantic units, instead of word-level information, mainly determine that the text can be classified into the target categories "youth" and "sports". For each of a variety of texts, there are always semantic units that construct the semantic meaning of the whole text. Regarding an automatic system for multi-label text classification, the system should be able to extract the semantic units in the source text for better performance in classification. Therefore, we propose our model to tackle this problem.

4 Proposed Method

In the following, we introduce our proposed modules to improve the conventional Seq2Seq model for multi-label text classification. In general, it contains two components: multi-level dilated convolution (MDC) as well as hybrid attention mechanism.

4.1 Multi-level Dilated Convolution

On top of the representations generated by the original encoder, which is an LSTM in our model, we apply the multi-layer convolutional neural networks to generate representations of semantic units by capturing local correlations and long-term dependencies among words. To be specific, our CNN is a three-layer one-dimensional CNN. Fol-
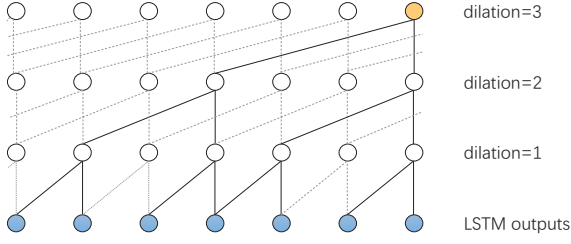
Figure 1: **Structure of Multi-level Dilated Convolution (MDC).** A example of MDC with kernel size $k = 2$ and dilation rates $[1, 2, 3]$. To avoid gridding effects, the dilation rates do not share a common factor other than 1.
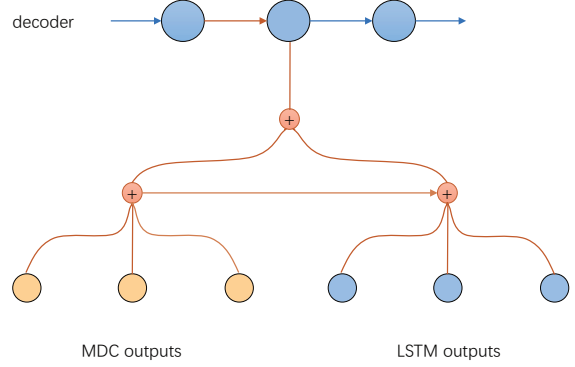

Figure 2: **Structure of Hybrid Attention.** The blue circles at the left bottom represent the source annotations generated by the LSTM encoder, the yellow circles at the right bottom represent the semantic unit representations generated by MDC, and the blue circles at the top represent the LSTM decoder outputs. At each decoding time step, the output of the LSTM attends to the semantic unit representations first, and then the new representation incorporated with high-level information attends to the source annotations.

lowing the previous work (Kalchbrenner et al., 2014) on CNN for NLP, we use one-dimensional convolution with the number of channels equal to the number of units of the hidden layer, so that the information at each dimension of a representation vector will not be disconnected as 2-dimension convolution does. Besides, as we are to capture semantic units in the source text instead of higher-level word representations, there is no need to use padding for the convolution.

A special design for the CNN is the implementation of dilated convolution. Dilation has become popular in semantic segmentation in computer vision in recent years (Yu and Koltun, 2015; Wang et al., 2018), and it has been introduced to the fields of NLP (Kalchbrenner et al., 2016) and speech processing (van den Oord et al., 2016). Dilated convolution refers to convolution inserted with "holes" so that it is able to remove the negative effects such as information loss caused by common down-sampling methods, such as max-pooling and strided convolution. Besides, it is able to expand the receptive fields at the exponential level without increasing the number of parameters. Thus, it becomes possible for dilated convolution to capture longer-term dependency. Furthermore, with the purpose of avoiding gridding effects caused by dilation (e.g., the dilated segments of the convolutional kernel can cause missing of vital local correlation and break the continuity between word representations), we implement a multi-level dilated convolution with different dilation rates at different levels, where the dilation rates are hyperparameters in our model.

Instead of using the same dilation rate or dilation rates with the common factor, which can cause gridding effects, we apply multi-level dilated convolution with different dilation rates,

such as [1,2,3]. Following Wang et al. (2018), for N layers of 1-dimension convolution with kernel size K with dilation rates $[r_1, ..., r_N]$, the maximum distance between two nonzero values is $max(M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i)$ with $M_N = r_N$, and the goal is $M_2 \leq K$. In our experiments, we set the dilation rates to [1, 2, 3] and K to 3, and we have $M_2 = 2$. The implementations can avoid the gridding effects and allows the top layer to access information between longer distance without loss of coverage. Moreover, as there may be irrelevant information to the semantic units at a long distance, we carefully design the dilation rates to [1, 2, 3] based on the performance in validation, instead of others such as [2, 5, 9], so that the top layer will not process the information among overlong distance and reduce the influence of unrelated information. Therefore, our model can generate semantic unit representations from the information at phrase level with small dilation rates and those at sentence level with large dilation rates.

4.2 Hybrid Attention

As we have annotations from the RNN encoder and semantic unit representations from the MDC, we design two types of attention mechanism to evaluate the effects of information of different levels. One is the common attention mechanism, which attends to the semantic unit representations

instead of the source word annotations as the conventional does, the other is our designed hybrid attention mechanism to incorporate information of the two levels.

The idea of hybrid attention is motivated by memory networks (Sukhbaatar et al., 2015) and multi-step attention (Gehring et al., 2017). It can be regarded as the attention mechanism with multiple "hops", with the first hop attending to the higher-level semantic unit information and the second hop attending to the lower-level word unit information based on the decoding and the first attention to the semantic units. Details are presented below.

For the output of the decoder at each time step, it not only attends to the source annotations from the RNN encoder as it usually does but also attends to the semantic unit representations from the MDC. In our model, the decoder output first pays attention to the semantic unit representations from the MDC to figure out the most relevant semantic units and generates a new representation based on the attention. Next, the new representation with both the information from the decoding process as well as the attention to the semantic units attends to the source annotations from the LSTM encoder, so it can extract word-level information from the source text with the guidance of the semantic units, mitigating the problem of irrelevance and redundancy.

To be specific, for the source annotations from the LSTM encoder $h = \{h_1, ..., h_i, ..., h_n\}$ and the semantic unit representations $g = \{g_1, ..., g_i, ..., g_m\}$, the decoder output s_t first attends to the semantic unit representations g and generates a new representation s'_t. Then the new representation s'_t attends to the source annotations h and generates another representation $\tilde{s}_t$ following the identical attention mechanism as mentioned above. In the final step, the model generates o_t for the prediction of y_t, where:

$$o_t = s'_t \oplus \tilde{s}_t \tag{7}$$

For comparison, we also propose another type of attention called "additive attention", whose experimental results are in the ablation test. In this mechanism, instead of paying attention to the two types of representations step by step as mentioned above, the output of the LSTM decoder s_t attends to the semantic unit representations g and the source annotations h respectively to generate

s'_t and $\tilde{s}_t$, which are finally added element-wisely for the final output o_t.

5 Experiment Setup

In the following, we introduce the datasets and our experiment settings as well as the baseline models that we compare with.

5.1 Datasets and Preprocessing

Reuters Corpus Volume I (RCV1-v2)[2]: The dataset (Lewis et al., 2004) consists of more than 800k manually categorized newswire stories by Reuters Ltd. for research purpose, where each story is assigned with multiple topics. The total number of topics is 103. To be specific, the training set contains around 802414 samples, while the development set and test set contain 1000 samples respectively. We filter the samples whose lengths are over 500 words in the dataset, which removes about 0.5% of the samples in the training, development and test sets. The vocabulary size is set to 50k words. Numbers as well as out-of-vocabulary words are masked by special tokens "#" and "UNK". For label permutation, we apply the descending order by frequency following Kim (2014).

Ren-CECps: The dataset is a sentence corpus collected from Chinese blogs, annotated with 8 emotion tags *anger*, *anxiety*, *expect*, *hate*, *joy*, *love*, *sorrow* and *surprise* as well as 3 polarity tags *positive*, *negative* and *neutral*. The dataset contains 35096 sentences for multi-label text classification. We apply preprocessing for the data similar to that for the RCV1-v2, which are filtering samples of over 500 words, setting the vocabulary size to 50k and applying the descending order by frequency for label permutation.

5.2 Experiment Settings

We implement our experiments in PyTorch on an NVIDIA 1080Ti GPU. In the experiments, the batch size is set to 64, and the embedding size and the number of units of hidden layers are 512. We use Adam optimizer (Kingma and Ba, 2014) with the default setting $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. The learning rate is initialized to 0.0003 based on the performance on the development set, and it is halved after every epoch

---

[2] http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm

Models	HL(-)	P(+)	R(+)	F1(+)
BR	0.0086	0.904	0.816	0.858
CC	0.0087	0.887	0.828	0.857
LP	0.0087	0.896	0.824	0.858
CNN	0.0089	**0.922**	0.798	0.855
CNN-RNN	0.0085	0.889	0.825	0.856
S2S	0.0082	0.883	0.849	0.866
S2S+Attn	0.0081	0.889	0.848	0.868
Our Model	**0.0072**	0.891	**0.873**	**0.882**

Table 2: Performance on the RCV1-V2 test set. HL, P, R, and F1 denote hamming loss, micro-precision, micro-recall and micro-F_1, respectively ($p < 0.05$).

Models	HL(-)	P(+)	R(+)	F1(+)
BR	**0.1663**	**0.649**	0.472	0.546
CC	0.1828	0.572	0.551	0.561
LP	0.1902	0.556	0.517	0.536
CNN	0.1726	0.628	0.512	0.565
CNN-RNN	0.1876	0.576	0.538	0.556
S2S	0.1814	0.587	0.571	0.579
S2S+Attn	0.1793	0.589	0.573	0.581
Our Model	0.1782	0.593	**0.585**	**0.590**

Table 3: Performance of the models on the Ren-CECps test set. HL, P, R, and F1 denote hamming loss, micro-precision, micro-recall and micro-F_1, respectively ($p < 0.05$).

of training. Gradient clipping is applied with the range [-10, 10].

Following the previous studies (Zhang and Zhou, 2007; Chen et al., 2017), we choose hamming loss and micro-F_1 score to evaluate the performance of our model. Hamming loss refers to the fraction of incorrect prediction (Schapire and Singer, 1999), and micro-F_1 score refers to the weighted average F_1 score. For reference, the micro-precision as well as micro-recall scores are also reported. To be specific, the computations of Hamming Loss (HL) micro-F_1 score are illustrated below:

$$HL = \frac{1}{L} \sum \mathbb{I}(y \neq \hat{y}) \qquad (8)$$

$$microF_1 = \frac{\sum_{j=1}^{L} 2tp_j}{\sum_{j=1}^{L} 2tp_j + fp_j + fn_j} \qquad (9)$$

where tp_j, fp_j and fn_j refer to the number of true positive examples, false positive examples and false negative examples respectively.

5.3 Baseline Models

In the following, we introduce the baseline models for comparison for both datasets.

- **Binary Relevance (BR)** (Boutell et al., 2004) transforms the MLC task into multiple single-label classification problems.

- **Classifier Chains (CC)** (Read et al., 2011) transforms the MLC task into a chain of binary classification problems to model the correlations between labels.

- **Label Powerset (LP)** (Tsoumakas and Katakis, 2006) creates one binary classifier for every label combination attested in the training set.

- **CNN** (Kim, 2014) uses multiple convolution kernels to extract text feature, which is then input to the linear transformation layer followed by a sigmoid function to output the probability distribution over the label space.

- **CNN-RNN** (Chen et al., 2017) utilizes CNN and RNN to capture both global and local textual semantics and model label correlations.

- **S2S** and **S2S+Attn** (Sutskever et al., 2014; Bahdanau et al., 2014) are our implementation of the RNN-based sequence-to-sequence models without and with the attention mechanism respectively.

6 Results and Discussion

In the following sections, we report the results of our experiments on the RCV1-v2 and Ren-CECps. Moreover, we conduct an ablation test and the comparison with models with hierarchical models with the deterministic setting of sentence or phrase, to illustrate that our model with learnable semantic units possesses a clear advantage over the baseline models. Furthermore, we demonstrate that the higher-level representations are useful for the prediction of labels of low frequency in the dataset so that it can ensure that the model is not strictly learning the prior distribution of the label sequence.

6.1 Results

We present the results of our implementations of our model as well as the baselines on the RCV1-v2 on Table 2. From the results of the conventional baselines, it can be found that the classical methods for multi-label text classification still own competitiveness compared with the

machine-learning-based and even deep-learning-based methods, instead of the Seq2Seq-based models. Regarding the Seq2Seq model, both the S2S and the S2S+Attn achieve improvements on the dataset, compared with the baselines above. However, as mentioned previously, the attention mechanism does not play a significant role in the Seq2Seq model for multi-label text classification. By contrast, our proposed mechanism, which is label-classification-oriented, can take both the information of semantic units and that of word units into consideration. Our proposed model achieves the best performance in the evaluation of Hamming loss and micro-F_1 score, which reduces 9.8% of Hamming loss and improves 1.3% of micro-F_1 score, in comparison with the S2S+Attn.

We also present the results of our experiments on Ren-CECps. Similar to the models' performance on the RCV1-v2, the conventional baselines except for Seq2Seq models achieve lower performance on the evaluation of micro-F_1 score compared with the Seq2Seq models. Moreover, the S2S and the S2S+Attn still achieve similar performance on micro-F_1 on this dataset, and our proposed model achieves the best performance with the improvement of 0.009 micro-F_1 score. An interesting finding is that the Seq2Seq models do not possess an advantage over the conventional baselines on the evaluation of Hamming Loss. We observe that there are fewer labels in the Ren-CECps than in the RCV1-v2 (11 and 103). As our label data are reordered according to the descending order of label frequency, the Seq2Seq model is inclined to learn the frequency distribution, which is similar to a long-tailed distribution. However, regarding the low-frequency labels with only a few samples, their amounts are similar, whose distributions are much more uniform than that of the whole label data. It is more difficult for the Seq2Seq model to classify them correctly while the model is approximating the long-tailed distribution compared with the conventional baselines. As Hamming loss reflects the average incorrect prediction, the errors in classifying into low-frequency labels will lead to a sharper increase in Hamming Loss, in comparison with micro-F_1 score.

6.2 Ablation Test

To evaluate the effects of our proposed modules, we present an ablation test for our model. We re-

Models	HL(-)	P(+)	R(+)	F1(+)
w/o attention	0.0086	**0.904**	0.816	0.871
attention	0.0087	0.887	0.828	0.869
MDC	0.0074	0.889	0.871	0.880
additive	0.0073	0.888	0.871	0.879
hybrid	**0.0072**	0.891	**0.873**	**0.882**

Table 4: Performance of the models with different attention mechanisms on the RCV1-V2 test set. HL, P, R, and F1 denote hamming loss, micro-precision, micro-recall and micro-F_1, respectively ($p < 0.05$).

move certain modules to control variables so that their effects can be fairly compared. To be specific, besides the evaluation of the conventional attention mechanism mentioned in Section 3, we evaluate the effects of hybrid attention and its modules. We demonstrate the performance of five models with different attention implementation for comparison, which are model without attention, one with only attention to the source annotations from LSTM, one with only attention to the semantic unit representations from the MDC, one with the attention to both the source annotations and semantic unit representations (additive) and hybrid attention, respectively. Therefore, the effects of each of our proposed modules, including MDC and hybrid attention, can be evaluated individually without the influence of the other modules.

Results in Table 4 reflect that our model still performs the best in comparison with models with the other types of attention mechanism. Except for the insignificant effect of the conventional attention mechanism mentioned above, it can be found that the high-level representations generated by the MDC contribute much to the performance of the Seq2Seq model for multi-label text classification, which improves about 0.9 micro-F_1 score. Moreover, simple additive attention mechanism, which is equivalent to the element-wise addition of the representations of MDC and those of the conventional mechanism, achieves similar performance to the single MDC, which also demonstrates that conventional attention mechanism in this task makes little contribution. As to our proposed hybrid attention, which is a relatively complex combination of the two mechanisms, can improve the performance of MDC. This shows that although conventional attention mechanism for word-level information does not influence the performance of the SeqSeq model significantly, the hybrid attention which extracts word-level infor-

Models	HL(-)	P(+)	R(+)	F1(+)
Hier-5	0.0075	0.887	0.869	0.878
Hier-10	0.0077	0.883	0.873	0.878
Hier-15	0.0076	0.879	0.879	0.879
Hier-20	0.0076	0.876	**0.881**	0.878
Our model	**0.0072**	**0.891**	0.873	**0.882**

Table 5: Performance of the hierarchical model and our model on the RCV1-V2 test set. Hier refers to hierarchical model, and the subsequent number refers to the length of sentence (word) for sentence-level representations ($p < 0.05$).

mation based on the generated high-level semantic information can provide some information about important details that are relevant to the most contributing semantic units.

6.3 Comparison with the Hierarchical Models

Another method that can extract high-level representations is a heuristic method that manually annotates sentences or phrases first and applies a hierarchical model for high-level representations. To be specific, the method does not only apply an RNN encoder to the word representations but also to sentence representations. In our reimplementation, we regard the representation from the LSTM encoder at the time step of the end of each sentence as the sentence representation, and we implement another LSTM on top of the original encoder that receives sentence representations as input so that the whole encoder can be hierarchical. We implement the experiment on the dataset RCV1-v2. As there is no sentence marker in the dataset RCV1-v2, we set a sentence boundary for the source text and we apply a hierarchical model to generate sentence representations.

Compared with our proposed MDC, the hierarchical model for the high-level representations is relatively deterministic since the sentences or phrases are predefined manually. However, our proposed MDC learns the high-level representations through dilated convolution, which is not restricted by the manually-annotated boundaries. Through the evaluation, we expect to see if our model with multi-level dilated convolution as well as hybrid attention can achieve similar or better performance than the hierarchical model. Moreover, we note that the number of parameters of the hierarchical model is more than that of our model, which are 47.24M and 45.13M respec-
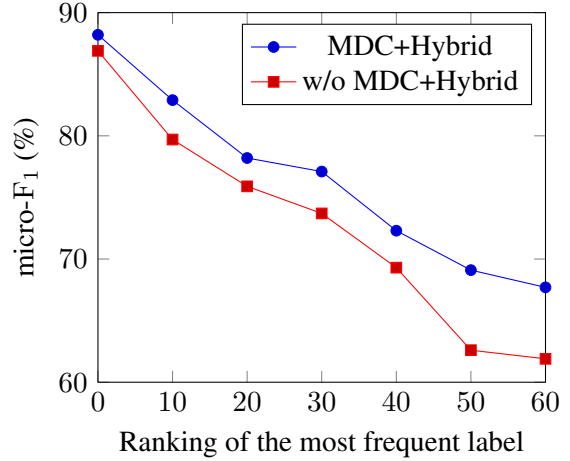

Figure 3: **Micro-F_1 scores of our model and the baseline on the evaluation of labels of different frequency.** The x-axis refers to the ranking of the most frequent label in the labels for classification, and the y-axis refers to the micro-F_1 score performance.

tively. Therefore, it is obvious that our model does not possess the advantage of parameter number in the comparison.

We present the results of the evaluation on Table 5, where it can be found that our model with fewer parameters still outperforms the hierarchical model with the deterministic setting of sentence or phrase. Moreover, in order to alleviate the influence of the deterministic sentence boundary, we compare the performance of different hierarchical models with different boundaries, which sets the boundaries at the end of every 5, 10, 15 and 20 words respectively. The results in Table 5 show that the hierarchical models achieve similar performances, which are all higher than the performances of the baselines. This shows that high-level representations can contribute to the performance of the Seq2Seq model on the multi-label text classification task. Furthermore, as these performances are no better than that of our proposed model, it can reflect that the learnable high-level representations can contribute more than deterministic sentence-level representations, as it can be more flexible to represent information of diverse levels, instead of fixed phrase or sentence level.

6.4 Error Analysis

Another finding in our experiments is that the model's performance on low-frequency label classification is lower than that on high-frequency label classification. This problem is also reflected

in our report of the experimental results on the Ren-CECps. The decrease in performance is reasonable since the model is sensitive to the amount of data, especially on small datasets such as Ren-CECps. We also hypothesize that this error comes from the essence of the Seq2Seq model. As the frequency of our label data is similar to a long-tailed distribution and the data are organized by descending order of label frequency, the Seq2Seq model is inclined to model the distribution. As the frequency distribution of the low-frequency labels is relatively uniform, it is much harder for it to model the distribution.

In contrast, as our model is capable of capturing deeper semantic information for the label classification, we believe that it is more robust to the classification of low-frequency labels with the help of the information from multiple levels. We remove the top 10, 20, 30, 40, 50 and 60 most frequent labels subsequently, and we evaluate the performance of our model and the baseline Seq2Seq model on the classification of these labels. Figure 3 shows the results of the models on label data of different frequency. It is obvious that although the performances of both models decrease with the decrease of the label frequency, our model continues to perform better than the baseline on all levels of label frequency. In addition, the gap between the performances of the two models continues to increase with the decrease of label frequency, demonstrating our model's advantage over the baseline on classifying low-frequency labels.

7   Related Work

The current models for the multi-label classification task can be classified into three categories: problem transformation methods, algorithm adaptation methods, and neural network models.

Problem transformation methods decompose the multi-label classification task into multiple single-label learning tasks. The BR algorithm (Boutell et al., 2004) builds a separate classifier for each label, causing the label correlations to be ignored. In order to model label correlations, Label Powerset (LP) (Tsoumakas and Katakis, 2006) creates one binary classifier for every label combination attested in the training set and Classifier Chains (CC) (Read et al., 2011) connects all classifiers in a chain through feature space.

Algorithm adaptation methods adopt specific learning algorithms to the multi-label classification task without requiring problem transformations. Clare and King (2001) constructed decision tree based on multi-label entropy to perform classification. Elisseeff and Weston (2002) adopted a Support Vector Machine (SVM) like learning system to handle multi-label problem. Zhang and Zhou (2007) utilized the k-nearest neighbor algorithm and maximum a posteriori principle to determine the label set of each sample. Fürnkranz et al. (2008) made ranking among labels by utilizing pairwise comparison. Li et al. (2015) used joint learning predictions as features.

Recent studies of multi-label text classification have turned to the application of neural networks, which have achieved great success in natural language processing. Zhang and Zhou (2006) implemented the fully-connected neural networks with pairwise ranking loss function. Nam et al. (2013) changed the ranking loss function to the cross-entropy loss to better the training. Kurata et al. (2016) proposed a novel neural network initialization method to treat some neurons as dedicated neurons to model label correlations. Chen et al. (2017) incorporated CNN and RNN so as to capture both global and local semantic information and model high-order label correlations. (Nam et al., 2017) proposed to generate labels sequentially, and Yang et al. (2018); Li et al. (2018) both adopted the Seq2Seq, one with a novel decoder and one with a soft loss function respectively.

8   Conclusion

In this study, we propose our model based on the multi-level dilated convolution and the hybrid attention mechanism, which can extract both the semantic-unit-level information and word-level information. Experimental results demonstrate that our proposed model can significantly outperform the baseline models. Moreover, the analyses reflect that our model is competitive with the deterministic hierarchical models and it is more robust to classifying the low-frequency labels than the baseline.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Fernando Benites and Elena Sapozhnikova. 2015. Haram: a hierarchical aram neural network for large-scale text classification. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pages 847–854. IEEE.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771.

Guibin Chen, Deheng Ye, Zhenchang Xing, Jieshan Chen, and Erik Cambria. 2017. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *IJCNN 2017*, pages 2377–2383.

Amanda Clare and Ross D King. 2001. Knowledge discovery in multi-label phenotype data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer.

André Elisseeff and Jason Weston. 2002. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687.

Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. 2008. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2):133–153.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML 2017*, pages 1243–1252.

Sepp Hochreiter and Jürgen Schmidhuber. 1996. LSTM can solve hard long time lag problems. In *NIPS, 1996*, pages 473–479.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aäron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *CoRR*, abs/1610.10099.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *ACL 2014*, pages 655–665.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP 2014*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In *NAACL HLT 2016*, pages 521–526.

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.

Li Li, Houfeng Wang, Xu Sun, Baobao Chang, Shi Zhao, and Lei Sha. 2015. Multi-label text categorization with joint learning predictions-as-features method. In *EMNLP 2015*, pages 835–839.

Wei Li, Xuancheng Ren, Damai Dai, Yunfang Wu, Houfeng Wang, and Xu Sun. 2018. Sememe prediction: Learning semantic knowledge from unstructured textual wiki descriptions. *CoRR*, abs/1808.05437.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *ACL 2018*, pages 163–169.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP 2015*, pages 1412–1421.

Jinseok Nam, Jungi Kim, Iryna Gurevych, and Johannes Fürnkranz. 2013. Large-scale multi-label text classification - revisiting neural networks. *CoRR*, abs/1312.5419.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *NIPS 2017*, pages 5419–5429.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *ISCA Speech Synthesis Workshop 2016*, page 125.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3):333.

Robert E Schapire and Yoram Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS, 2015*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS 2014*, pages 3104–3112.

Grigorios Tsoumakas and Ioannis Katakis. 2006. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3).

Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison W. Cottrell. 2018. Understanding convolution for semantic segmentation. In *WACV 2018*, pages 1451–1460.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *COLING 2018*, pages 3915–3926.

Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.