Hierarchical Attention Network for Multilabel Thesis Classification

Cheng-Yuan Ho
Dept. of Computer Science and Information Engineering,
Asia University,
Taichung City, Taiwan
tommyho@asia.edu.tw

Prayitno
Dept. of Computer Science and Information Engineering,
Asia University, Taichung City, Taiwan
Dept. of Electrical Engineering,
Politeknik Negeri Semarang, Indonesia
prayitno@polines.ac.id

***Abstract*—Identifying category of thesis papers is a challenging task. An automatic classification of thesis task plays an important role to reduce human workload in the process of paper thesis repository system, especially one thesis may have more than one label (multilabel). To accomplish this goal, we propose a multilabel thesis classification with hierarchical attention network. Our proposed model capture sentence context in a document with word-level representation. Experiments conducted on thesis datasets from Taiwan ministry of education artificial intelligence competition collected from arxiv repository system. The result shows that the proposed model achieved the highest F1 score compared to another machine learning model with 68%.**

***Keywords*—data mining, thesis label classification, deep learning**

I. Introduction

An automatic classification of thesis category believed as an important step to assist human in the process of paper archiving. Taiwan Ministry of Education initiative an artificial intelligence competition (AI-Cup 2019) to cultivate the development of systems for automatically evaluating a paper belong to one category or several categories [1]. For instance, in computer science thesis paper there are theoretical, engineering, empirical and others categories. Furthermore, this competition involved developing a system that given a paper thesis title and abstract, classify the categories into one or several categories. This task called a multilabel categories since one paper could be have one category such as others, or two categories such as theoretical and engineering. Fig. 1 shows an example. More information on competition task, requirements, dataset and evaluation report can be seen at the official website.

The aim of this paper is to present a description of classification system employs in Taiwan Moe AI Cup 2019 thesis classification task, summary of system results, a short information of the competition and our work plan in the future. The main contributions of this paper can be summarized as follows:

- we propose a hierarchical attention network for thesis multilabel classification. Multilabel classification task have its own challenges compare to binary or multiclass classification.
- we provide a principle way to exploit title and abstract document features in classification task. Our proposed model incorporates attention mechanism in the sentence level.
- we conduct extensive experiments on real-world arxiv dataset to demonstrate the effectiveness of proposed method for multilabel thesis classification results.

Id	D00009
Title	Robustness from structure: Inference with hierarchical spiking networks on analog neuromorphic hardware
Abstract	How spiking networks are able to perform probabilistic inference is an intriguing question, not only for understanding information processing in the brain, but also for transferring these computational principles to neuromorphic silicon circuits.$$$A number of computationally powerful spiking network models have been proposed, but most of them have only been tested, under ideal conditions, in software simulations.$$$Any implementation in an analog, physical system, be it in vivo or in silico, will generally lead to distorted dynamics due to the physical properties of the underlying substrate.$$$In this paper, we discuss several such distortive effects that are difficult or impossible to remove by classical calibration routines or parameter training.$$$We then argue that hierarchical networks of leaky integrate-and-fire neurons can offer the required robustness for physical implementation and demonstrate this with both software simulations and emulation on an accelerated analog neuromorphic device.
Authors	Petrovici/Schroeder/Breitwieser/GrÄ¼bl/Schemmel/Meier
Categories	q-bio.NC/cs.NE/stat.ML
Created	date 5/3/2018
Task 2	THEORETICAL ENGINEERING

Figure 1. An example of thesis label classification.

The remainder of the paper is organized as follows. Section 2 introduces the propose methods, data preprocessing, classification layer and evaluation metrics, and section 3 describes the results and findings. Finally, the whole paper is concluded in section 4.

II. Literature Review

In the recent years, text classification researchers find that hierarchical attention network are useful to learn a document context representation with highlighting word or sentence weight in the document [2-4].

III. Proposed Method

To develop automatic classification of thesis category, we describe the problem statement, datasets, network schema and the classifier algorithm. The system task is classifying a thesis paper category into the following categories: Theoretical paper, Engineering paper, Empirical paper, Others.

A. Problem Statement

Multilabel classification task is different from traditional supervised classification problem. Let $X=\mathscr{R}^d$ be the d-dimensional input space. Given a labeled dataset $\{(x_i,y_i)\}^m{}_{i=1}$, where $x_i \in X$ contains m instances and $y_i \in Y= \{0,1\}^q$ consists of q possible labels, the multilabel learning aims to learn a function $f:X \rightarrow Y$ that maps an input $x=X$ to output $y=Y$. In this work, the classification task problem is given a title (T) and abstract (A) of a scientific paper learn a model $f:f(T,A) \rightarrow y_i \in \{0,1\}^4$ where to classify into one, two or three labels on the following categories such as Theoretical paper, Engineering paper, Empirical paper or Others paper.

B. Data Preprocessing

Dataset. We downloaded the thesis classification datasets from Taiwan Moe AI Cup 2019 competition, which is collected from ArXiv paper repository system especially in computer science topic. The dataset provides us metadata namely id, title, abstract, author, categories, created date and

The work is supported by the Ministry of Science and Technology of Taiwan, under Grant no. MOST 108-2221-E-468-010 -, and partially supported by the Ministry of Education of Taiwan, the Artificial Intelligence Talent Cultivation Project, and by Asia University, under Grant no. 107-ASIA-UNAIR-09.

task 2. There are two datasets provided for us, first training dataset consists of 7.000 paper information and second testing dataset consists of 20.000 paper information. Table 1 shows datasets statistic provided by the competition.

TABLE I THESIS CLASSIFICATION TASK DATASETS

Datasets Statistic	
Training Set	7.000
Testing Set	20.000

To give a clear picture about the dataset, we try to explore distribution of categories in training dataset. Table 2 illustrate the paper categories distribution in training dataset. The categories distribution for each category is not in equal percentage. Category fall under 'others' become the least percentage with 4%, on the other hand, Engineering become the most percentage distribution with 48%.

TABLE II PAPER CATEGORIES STATISTIC IN TRAINING DATASET

Categories	Number of Paper	Percentage
EMPIRICAL	2.140	32%
ENGINEERING	3.391	48%
OTHERS	259	4%
THEORETICAL	3.218	46%

C. *Hierarchical Attention Network*

The hierarchical attention neural network schema for multilabel thesis classification task is shown at Figure 2. It consists several components like word encoder, word attention, sentence encoder, sentence attention and classification layer.

Inputs. Let A be text in title and abstract consisting N sentences $A=\{s_1,s_2,\ldots,s_N\}$. Each sentence $s_i=\{w^i_1,w^i_2,\ldots,w^i_{Mi}\}$ contains M_i words. Each word will be match with word vector such as GloVe for vectorization.

Pre-trained Word Vector. For developing a deep learning model we employ pre-trained word embedding such as Global Vectors for Word Representation (GloVe) developed by Stanford [5] and Fasttext sub word embedding provided by Facebook [6]. In one hand, Glove Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors). In the other hand, Fasttext was pre-trained on Wikipedia 2017, UMBC web base and statmt.org news datasets with total 16B tokens. Embedding dimension is 300, the vocabulary is 1M words.

Word Encoder. We used a recurrent neural network (RNN) based word encoder to represent the sentence in title and abstract. In theory, RNN can capture context in the long sentence dependency but in practice, the old memory will be lost as the sequence sentence becomes longer. This is caused by vanishing gradient problem. To better capture context in the long sentence Cho et.al. [7] used Gated Recurrent Units (GRU) to ensure more persistent memory in the network. We laverage GRU to encode the word sequence similar to [2]. GRU introduce two gate mechanism such as the reset gate r_t and the update gate z_t. Both reset and update gate control how information updated into the state $\tilde{h}_t$. In one hand, the update gate z_t how much past information is kept and decides how much new information will be added. On the other hand, the reset gate r_t controls how much the past state contributes to the candidate state. It will forget the previous state if r_t value is zero.

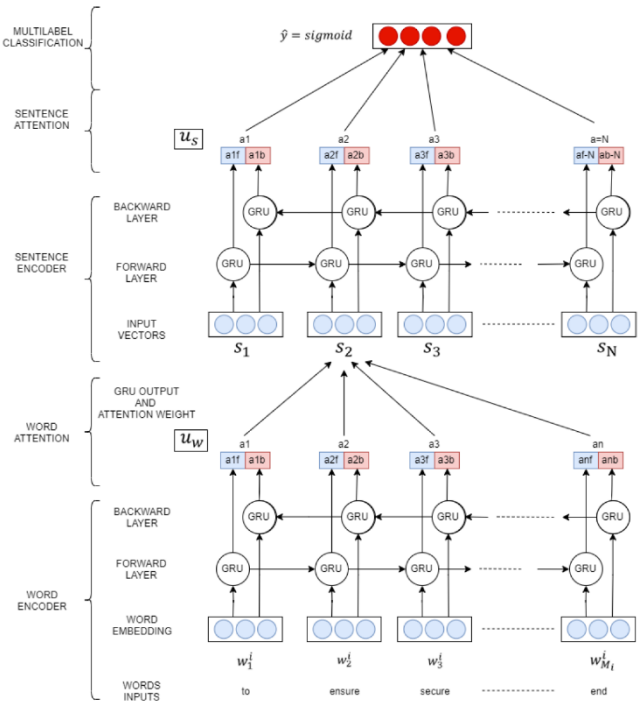

Figure 2. Hierarchical attention network for multilabel thesis classification

$$h_t=(1-z_t)\odot h_{t-1}+z_t\odot \tilde{h}_t$$
$$z_t=\sigma(W_t x_t+U_z h_{t-1}+b_z)$$
$$\tilde{h}_t=tanh(W_h x_t+r_t\odot(U_h ht-1)+b_h)$$
$$r_t=\sigma(W_r x_t+U_r h_{t-1}+b_r)$$

Furthermore, we use bidirectional GRU to capture word context from forward direction and backward direction. The forward GRU direction $\overrightarrow{GRU}$ read the sentence s_i from word w^i_1 to w^i_{Mi} and backward GRU direction $\overleftarrow{GRU}$ which reads sentence s_i from w^i_{Mi} to w^i_1:

$$\overrightarrow{h^i_t} = \overrightarrow{GRU}\left(w^i_t\right), t \in \{1,\ldots,M_i\}$$
$$\overleftarrow{h^i_t} = \overleftarrow{GRU}\left(w^i_t\right), t \in \{M_i,\ldots,1\}$$

We get the word context by concatenating the forward hidden state $\overrightarrow{h^i_t}$ and the backward hidden state $\overleftarrow{h^i_t}$ i.e.
$h^i_t = [\overrightarrow{h^i_t} \oplus \overleftarrow{h^i_t}]$

Word Attention. In a sentence, not all word gives same contribution in the sentence meaning. To capture the weight, we proposed word attention in the model. Attention model is small dense neural network with formula as follows:

$$u_{it} = tanh(W_w h_{it} + b_w)$$
$$\alpha_{it} = \frac{exp(u^T_{it}, v_w)}{\sum_t exp(u^T_{it}, v_w)}$$
$$s_i = \sum_t \alpha_{it} h_{it}$$

Sentence Encoder. We use the same model as word encoder to compute sentence encoder. We use a bidirectional GRU to encode sentences:

$$\overrightarrow{h_i} = \overrightarrow{GRU}\left(s_i\right), i \in \{1,\ldots,N\}$$
$$\overleftarrow{h_i} = \overleftarrow{GRU}(s_i), i \in \{N,\ldots,1\}$$

We get the sentence vector output $s_i \in \mathbb{R}^{2d\times 1}$ by concatenating the forward and backward hidden state, i.e. $s_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}]$, which captures the context from surrounding sentences around sentence s_i.

Sentence Attention. To differentiate sentence that contributes more meaning in document, we use attention mechanism same with word attention mechanism. This formula yields

$$u_i = tanh(W_s h_i + b_s)$$
$$\alpha_i = \frac{exp(u_i^T, v_s)}{\sum_i exp(u_i^T, v_s)}$$
$$v = \sum_i \alpha_i h_i$$

Ground Truth Label. We also convert the target from string data type into multilabel vector, we employ the MultiLabelBinarizer function from scikit-learn framework [8].

D. Thesis Label Classification

In the previous section we introduce word encoder, word attention, sentence encoder and sentence attention components. Furthermore, to compute the label probability in multilabel classification we employ sigmoid activation layer:

$$\hat{y}=sigmoid([\hat{t}\oplus\hat{a}]W_f+b_f)$$

$[\hat{t}\oplus\hat{a}]$denotes the concatenation of learned features for title and abstract sentences in the document. $b_f\in\mathbb{R}^{1\times2}$ is the bias weight in the network layer. In addition, for each piece of title and abstract, the classification goal is to minimize the cross-entropy loss function as follows:

$$L(\theta)=-y\log(\hat{y})-(1-y)log(1-\hat{y})$$

where θ denotes the weights hyperparameters in the network, y is the ground truth label of each class and $\hat{y}$ is the predicted class probability. The weight and bias parameters in the network are learned through adam optimizer [9], which is gradient-stochastic based optimization.

E. Evaluation Metrics

In multilabel classification measurement, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$Precision(p) = \frac{\sum_i^4 TP_i}{\sum_i^4 TP_i + FP_i}$$
$$Recall(r) = \frac{\sum_i^4 TP_i}{\sum_i^4 TP_i + FN_i}$$
$$micro\ F1 = \frac{2 \times (Precision \times Recall)}{(Precision \times Recall)}$$

IV. Results And Discussion

Submission to the competition were evaluated using F1 score metrics as shown in Fig. 3. Our submissions using machine learning and deep learning model could be shown at Tables 3 and 4, respectively.

V. Conclusion And Future Work

In this paper, we propose multilabel thesis classification using hierarchical attention network (HAN). Our best result in this thesis label classification using hierarchical attention network model with pre-trained GloVe word embedding with 300 dimensions achieved 0.68% F1 measures. This score beat the others machine learning model for text classification in this case. To improve performance, increasing a number of datasets in training phase.

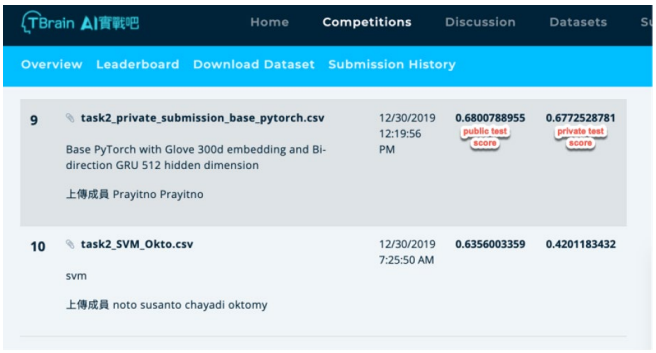

Figure 3. Submission score in competition platform.

TABLE III Machine Learning Model

Best TFIDF Text Vectorizer			Classifier Model	Inside Testing F1 Score					Outside Testing
Max_df	Max Features	Ngram		Theoretical	Engineering	Empirical	Others	Micro (Total Average)	
0.8	5.000	1	Multinomial Naive Bayer	0.68	0.67	0.02	0.00	0.57	0.6388770736
0.8	10.000	(1, 2)	Logistic Regression	0.72	0.68	0.40	0.00	0.63	0.6433260394
0.7	10.000	(1, 2)	KNN (K=4)	0.73	0.68	0.52	0.06	0.66	0.4900340301
0.8	10.000	(1, 2)	XGBoost	1.00	1.00	1.00	1.00	1.00	0.64160401

TABLE IV Deep Learning Model

Text Vectorizer	Classifier Model	Inside Testing F1 Score					Outside Testing
		Theoretical	Engineering	Empirical	Others	Micro (Total Average)	
Fasttext	BERT	0.68	0.67	0.02	0.00	0.57	0.5755225656
Glove 300 dimension	LSTM	0.65	0.60	0.05	0.00	0.51	0.52
Glove 300 dimension	Bi-directional GRU	0.86	0.73	0.71	0.17	0.69	**0.6800788955**

References

[1] Taiwan-MoE, "T-Brain AI 實戰吧 - AI CUP 2019 人工智慧論文機器閱讀競賽之論文分類." [Online]. Available: https://tbrain.trendmicro.com.tw/Competitions/Details/9.

[2] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480-1489.

[3] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19, 2019, pp. 395-405.

[4] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor Detection with Hierarchical Social Attention Network," in Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM'18, 2018, pp. 943-951.

[5] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, pp. 1532-1543.

[6] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in LREC 2018 - 11th International Conference on Language Resources and Evaluation, 2019, no. 1, pp. 52-55.

[7] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, vol. 28, no. 4, pp. 1724-1734.

[8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1-15, Dec. 2014.