Exploiting Label Dependencies for Multi-Label Document Classification Using Transformers

Haytame Fallah, Emmanuel Bruno, Patrice Bellot, Elisabeth Murisasco

Exploiting Label Dependencies for Multi-Label Document Classification Using Transformers

Haytame Fallah
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Marseille, France
Hyperbios
Aix-en-Provence, France
haytame.fallah@lis-lab.fr

Emmanuel Bruno
Université de Toulon, Aix Marseille Univ, CNRS, LIS
Toulon, France
emmanuel.bruno@univ-tln.fr

Patrice Bellot
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Marseille, France
patrice.bellot@univ-amu.fr

Elisabeth Murisasco
Université de Toulon, Aix Marseille Univ, CNRS, LIS
Toulon, France
elisabeth.murisasco@univ-tln.fr

ABSTRACT

We introduce in this paper a new approach to improve deep learning-based architectures for multi-label document classification. Dependencies between labels are an essential factor in the multi-label context. Our proposed strategy takes advantage of the knowledge extracted from label co-occurrences. The proposed method consists in adding a regularization term to the loss function used for training the model, in a way that incorporates the label similarities given by the label co-occurrences to encourage the model to jointly predict labels that are likely to co-occur, and and not consider labels that are rarely present with each other. This allows the neural model to better capture label dependencies. Our approach was evaluated on three datasets: the standard AAPD dataset, a corpus of scientific abstracts and Reuters-21578, a collection of news articles, and a newly proposed multi-label dataset called arXiv-ACM. Our method demonstrates improved performance, setting a new state-of-the-art on all three datasets.

CCS CONCEPTS

• **Applied computing** → **Document metadata**; **Digital libraries and archives**; • **Information systems** → **Digital libraries and archives**; **Content analysis and feature selection**; *Document collection models*; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Multi-label Classification, Document Classification, BERT, Transformers, Label Dependencies

DocEng '23, August 22–25, 2023, Limerick, Ireland

ACM ISBN 979-8-4007-0027-9/23/08…$15.00
https://doi.org/10.1145/3573128.3609356

ACM Reference Format:
Haytame Fallah, Emmanuel Bruno, Patrice Bellot, and Elisabeth Murisasco. 2023. Exploiting Label Dependencies for Multi-Label Document Classification Using Transformers. In *ACM Symposium on Document Engineering 2023 (DocEng '23), August 22–25, 2023, Limerick, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3573128.3609356

1 INTRODUCTION

Multi-label classification can be considered as a generalization of the traditional binary or multi-class classification. In multi-label text classification, the goal is to associate one or more labels to the input text. It is an important task that has applications in many tasks such as research article classification and metadata generation from documents [Mustafa et al. 2021; Sajid et al. 2011] that can be used for optimizing search engine indexing.

Several methods have been proposed to tackle multi-label classification and can be split into two families: problem transformation and problem adaptation methods. By transforming the problem, transformation methods are not true multi-label approaches and thus fail to consider the label correlations that are very important for extracting all the relevant labels for a given document. In contrast, problem adaption methods try to adapt the classification algorithms to better suit the multi-label problem and are more efficient than their counterparts since they do not require multiple models to be trained or an increase of the dataset's size.

We propose in this paper a problem adaptation approach that takes advantage of label co-occurrences, in a simple yet effective way, for multi-label document classification. We mainly focus on the feed-forward neural network (FFNN), with L layers, that is usually added on top of the transformer model when fine-tuning it for a specific task. We propose a regularization term that is added to the loss function, based on the predicted labels and the label similarity matrix given by the label pairwise probability of co-occurrences using the cosine similarity measure.

The motivation behind this method lies in the idea that when restrictions are imposed on each neuron that corresponds to a label, the label dependencies will be learned by the model throughout the transformer layers, rather than added as additional knowledge as a post-processing step. By using external information on label

co-occurrences/similarities, we encourage the model to predict labels that are likely to co-occur or to not consider labels that are not present with each other, leading to improved performance. This approach can be a considered high-order method since the prediction of a label is influenced by every other label. We evaluate our approach using a BERT-based architecture [Devlin et al. 2019] on the Reuters dataset [Lewis et al. 2004], a collection of news articles as well as scientific articles from the AAPD dataset [Yang et al. 2018]. We show that it leads to a gain in classification prediction on all these datasets.

Among the papers treating the multi-label problem. AAPD [Yang et al. 2018], Reuters [Lewis et al. 2004] and PubMed [Tsatsaronis et al. 2015] are the most used datasets. But the average number of labels per instance is low. This is particularly relevant in the context of digital libraries where research papers and scientific publications often contain multiple subjects, with approximately half of them having more than one topic [Mustafa et al. 2021]. To address this, we propose a new high cardinality multi-label dataset, arXiv-ACM, built from scientific article abstracts from the arXiv digital library and paired with the level-2 ACM keywords provided by the authors.

2 RELATED WORK

In this paper, we aim to adapt deep learning-based approaches which can contribute to a significant increase in performance. The use of a single model without the need for prior data transformation is an efficient method to try to address the multi-label problem. This eliminates the need for separate models for each label or to change the dataset to a substantial size, reducing the complexity and computational cost, while still delivering improved performance. Furthermore, explicitly capturing the dependency between labels has been shown to improve the multi-label classification performance [Zhang and Zhou 2007].

Conditional graphs and networks [Guo and Gu 2011; Zhang and Zhang 2010] are a good way to model label dependencies. In these methods, semantic hierarchical dependencies and label co-occurrences, or a mix of both [Wu et al. 2018], are used to construct dependency networks that are then combined with the main machine learning model through various methods.

In Liu et al. [2022], labels are encoded as embeddings and fed with the text sequence into a co-attention network [Seo et al. 2016]. The attention mechanism accounts for the relationship between labels, but its effectiveness is limited in cases where the labels are abbreviations or codes. In Section 5 we show that our approach is not dependent on the nature of the labels and contributes to an improvement in performance across all the studied datasets.

3 DEPENDENCY EXPLOITATION USING TRANSFORMERS

Co-occurrences of labels provide valuable information as they can capture the correlations between labels. We present a new method for incorporating label co-occurrence where we add a regularization term to the loss function. We use co-occurrence information that we extract from the training dataset by building a co-occurrence probability matrix $C^{n\times n}$, with n being the number of target labels.

Dependency Regularization Term (DepReg): To incorporate label dependencies in the learning process, we add a regularization term to the loss function. This is done in order to encourage predictions of labels that frequently co-occur and conversely, discourage the model from making predictions of labels that do not. With that motivation, we compute an n-dimensional dissimilarity vector D_{sim} between the label co-occurrence matrix C and the activations of the output layer of the classification FFNN $A^{[L]}$, and this by using the cosine similarity CS_θ as shown in equation 1. The dissimilarity vector represents an estimate of the extent to which the prediction aligns with the distribution of label co-occurrences. Each element of D_{sim} serves as an indicator of whether a label should be present/absent alongside other labels based on their co-occurrences.

$$D_{sim} = 1 - CS_\theta(C, A^{[L]}) = 1 - \frac{C \cdot A^{[L]}}{\|C\| \cdot \|A^{[L]}\|} \tag{1}$$

The regularization term L_{DepReg} is then computed as the dot product between the dissimilarity vector D_{sim} and the transposed activation vector ${A^{[L]}}^T$. This encourages the model to decrease the activations for labels that have a high dissimilarity score, while not imposing severe penalties when the dissimilarity score is low and the label's activation is high. By incorporating this regularization term, we obtain a comprehensive measure of how effectively the prediction follows the co-occurrence distribution of the labels.

$$L_{DepReg} = D_{sim} \cdot {A^{[L]}}^T \tag{2}$$

This dependency term is then added to the main loss function during training, the Binary Cross Entropy denoted as L_{BCE}, encouraging the model to make predictions that are consistent with the label dependencies of the dataset :

$$L_{total} = L_{BCE} + \lambda_{reg}.L_{DepReg} \tag{3}$$

The Dependency Regularization Term (DepReg) helps the model to avoid making predictions that go against the co-occurrence information while still allowing it to make predictions based on its learned patterns in the training data. λ_{reg} is a hyperparameter that controls the weight of the regularization term in the total loss.

4 ARXIV-ACM DATASET

In this section, we present the datasets that are commonly used for multi-label text classification evaluation, as well as our newly created multi-label dataset.

The community has generated and shared numerous multi-label datasets across various domains.These datasets include[1]:

- **Reuters-21578** a collection of articles from the Reuters newswire from the year 1987. It is a dataset[2] that has been often used to evaluate models for multi-label text classification. An article can belong to one or more of the 90 domains of the dataset,
- **AAPD (arXiv Academic Paper Dataset)** a collection of the "Abstract" of scientific papers in the arXiv digital library. An article can have one or more classifications among 54 labels. We use the same training (53840), validation (1000), and test (1000) distribution as [Yang et al. 2018].

[1]All datasets, alongside the implementation code, can be downloaded on GitHub: https://anonymous.4open.science/r/DocEng-23-12/

[2]https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

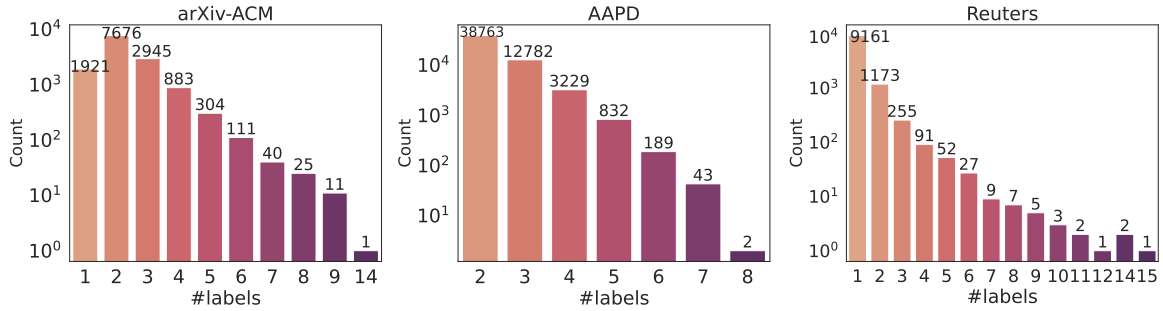

Figure 1: In Reuters-21578 and AAPD, the instances with one label are far more present than the ones with multiple labels, by building arXiv-ACM we aim to solve this problem that is present in several multi-label datasets.

Table 1: Datasets used, W is the average number of words per abstract, Card is the average number of labels per instance. Std and Med are the standard deviation and the median for the number of labels per instance.

	Labels	W	Card	Std	Med	#Train	#Valid	#Test
arXiv-ACM	64	152	2.33	1	2	9600	2157	2160
AAPD	54	163,43	2.43	0.71	2	53840	1000	1000
Reuters-21578	90	127,76	1.24	0.74	1	6770	1000	3019

These two datasets present many limitations. The most constraining one is the number of instances per label. In general, instances with one unique label are far more prevalent than multiple labels. This can introduce a classification bias, where models are more likely to learn to predict the most frequent label. This effect is amplified in unbalanced datasets, in the AAPD dataset, one pair of labels,'cs.it' and 'math.it' that always co-occur with each other, account for around 32% of the dataset's instances.

To address these limitations, we introduce in this paper a new multi-label dataset, which we call 'arXiv-ACM', with high cardinality, a reasonable size, and a better distribution of the samples per number of labels.

ArXiv-ACM is comprised of computer science article abstracts extracted via arXiv API[3], published between 1998 and 2021. These abstracts were then paired with the ACM keywords[4] filled by the authors of the articles during submission. Only the second-level keywords were considered, as the first level is too broad and the subsequent levels are too specific. We then filtered out the labels that have less than 20 instances for a final number of labels of 64. Table 1 and Figure 1 present this dataset alongside AAPD and Reuters, the other datasets that are covered in this paper.

5 EXPERIMENTS AND RESULTS

In this section, we outline the experimental setup and the comparison baselines used for evaluating our approach.

For the evaluation of the proposed methods, we use HugginFace's [Wolf et al. 2020] implementation of the *cased-base* version of **BERT**. We add an FFNN with $L = 2$ layers. We find that λ_{reg} of 0.2 gives the optimal results[5]. **Baselines**: We compare our approach to other neural approaches:

- **MAGNET** [Pal et al. 2020]: **M**ulti-label Text classification using **A**ttention based **G**raph Neural **NET**work, a graph network implementing the attention mechanism to capture dependencies between labels,
- **CNLE** [Liu et al. 2022]: a transformer model that introduces label embeddings alongside the text embedding, linked by a co-attention to get a contextualized representation of the input sequence by the classification labels,
- **CB-NTR** [Huang et al. 2021] that uses **C**lass **B**alanced focal loss with **N**egative **T**olerant **R**egularization as a loss function. This method is applied on the BERT transformer (base cased version) and is considered the state-of-the-art for multi-label classification on certain datasets.

As we can see in Table 2, using dependency information contained in the co-occurrence matrix of labels yields an increase in the micro-F1 score across all the datasets compared to the base version of BERT.

Reuters: The base version of BERT already manages to get good performance scores for Reuters. The class balancing focal loss achieves a small gain in both precision and recall (0.15 and 0.01 respectively), contributing to a gain in micro-F1 with 90.85 vs the base method that uses the binary cross entropy loss (90.77). DepReg outperforms the other methods with a micro-F1 score of 91.39, a maximum increase of 0.62 over the base version of BERT.

AAPD: Due to the nature of the vocabulary used in scientific abstracts, **AAPD** is a more complex dataset than Reuters. In scientific articles, various domains and disciplines might be involved.

[3]https://arxiv.org/help/api/
[4]https://www.acm.org/publications/computing-classification-system/1998/ccs98
[5]All datasets, alongside the implementation code, can be downloaded on GitHub: https://github.com/hf-lis/DocEng-23

Table 2: Micro-precision (Pr.), micro-recall (R), and micro-F1 scores (F1) for the test sets of arXiv-ACM, Reuters and AAPD, with std. values for micro-F1. Scores were averaged over 10 runs. Best scores are in bold blue.

	arXiv-ACM			Reuters-21578			AAPD		
Models	Pr.	R	F1	Pr.	R	F1	Pr.	R	F1
MAGNET	57.31	53.24	55.2 ±0.18	91.2	88.6	89.9 ±0.15	72.88	66.79	69.7 ±0.18
CNLE	56.85	52.37	54.52 ±0.15	90.9	88.7	89.8 ±0.12	74.71	69.11	71.80 ±0.22
CB-NTR	60.11	55.74	57.84 ±0.11	91.37	90.34	90.85 ±0.1	74.27	72.68	73.45 ±0.16
BERT	60.04	55.58	57.72 ±0.11	91.22	90.33	90.77 ±0.08	74.49	72.03	73.24 ±0.14
BERT+ *DepReg*	61.80	56.09	**58.08** ±0.13	92.22	90.57	**91.39** ±0.12	75.53	72.16	**73.81** ±0.17

With a more specific vocabulary and more precise words, the language modeling and classification tasks can be more difficult for the AAPD dataset. The performance scores show that the models have difficulty in learning to associate subjects with their corresponding abstracts, with the base version of BERT achieving a micro-F1 score of 73.24. Our DepReg method achieves this time an increase in micro-F1 of 0,57 compared to the Vanilla BERT.

arXiv-ACM Despite a more balanced distribution of the number of labels per instance, which reduces the model's bias to predict the most frequent label, it results in the lowest scores among the models for this task. The base version of BERT manages to obtain a micro-F1 score of 57.72. The balancing loss function contributes to a small gain in both precision and recall, achieving an increase of 0.12 in the micro-F1 score. The DepReg method has the highest gain in precision with a gain of 1.76. This notable increase alongside a gain in recall contributes to the best micro-F1 score for this dataset (58.08) with the highest increase over the base version of BERT.

This gain in performance achieved by the dependency learning approach we propose can be explained by the fact that the prediction of a label is influenced by the prediction of all other labels using co-occurrences. In some instances, this information helps predict labels that would not have been predicted otherwise (increase in recall). On the other hand, label dependencies can lower the number of false positives by reducing the bias that the model can have for frequent labels in the dataset, contributing to a gain in precision.

6 CONCLUSION

Multi-label classification is a relevant task, especially for managing digital libraries and automatic labeling of documents. Unfortunately, it is not included in the most important benchmarks such as GLUE. We proposed in this paper an effective way of using the pairwise label co-occurrence information to allow transformers to learn dependencies between labels. This high-order label dependency method is model agnostic and is not limited to text classification but can be used for any other multi-label task. We have tested and shown that co-occurrences and label dependencies can be used to achieve a tangible gain in performance for multi-label text classification. Our dependency regularization approach manages to capture the dependencies between labels contributing to more accurate multi-label document classification. Finally, we have introduced in this paper a new multi-label dataset that is more suitable for testing new multi-label approaches, aiming to address the limitations of commonly used datasets in terms of the number of labels distribution and class balance.

REFERENCES

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL* (2019).

Yuhong Guo and Suicheng Gu. 2011. Multi-label classification using conditional dependency networks. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two (IJCAI'11)*. AAAI Press, Barcelona, Catalonia, Spain, 1300–1305.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 8153–8161.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research* 5, Apr (2004), 361–397.

Minqian Liu, Lizhao Liu, Junyi Cao, and Qing Du. 2022. Co-attention network with label embedding for text classification. *Neurocomputing* 471 (2022).

Ghulam Mustafa, Muhammad Usman, Lisu Yu, Muhammad Afzal, Muhammad Sulaiman, and Abdul Shahid. 2021. Multi-label classification of research articles using Word2Vec and identification of similarity threshold. *Scientific Reports* 11 (11 2021), 21900. https://doi.org/10.1038/s41598-021-01460-7

Ankit Pal, M. Selvakumar, and Malaikannan Sankarasubbu. 2020. Multi-Label Text Classification using Attention-based Graph Neural Network.

Naseer Ahmed Sajid, Tariq Ali, Muhammad Tanvir Afzal, Munir Ahmad, and Muhammad Abdul Qadir. 2011. Exploiting Reference Section to Classify Paper's Topics. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems* (San Francisco, California) *(MEDES '11)*. Association for Computing Machinery, New York, NY, USA, 220–225. https://doi.org/10.1145/2077489.2077531

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension. *CoRR* (2016).

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* 16, 1 (2015).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. Technical Report arXiv:1910.03771.

Baoyuan Wu, Fan Jia, Wei Liu, Bernard Ghanem, and Siwei Lyu. 2018. Multi-label Learning with Missing Labels Using Mixed Dependency Graphs. *International Journal of Computer Vision* 126, 8 (Aug. 2018), 875–896. https://doi.org/10.1007/s11263-018-1085-3

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence Generation Model for Multi-label Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Min-Ling Zhang and Kun Zhang. 2010. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*. Association for Computing Machinery, New York, NY, USA, 999–1008.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* (2007).