Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Apprentissage de dépendances entre labels pour la classification multi-labels à l'aide de transformeurs

Haytame Fallah[1, 3, *] Elisabeth Murisasco[2, *]
Emmanuel Bruno[2, *] Patrice Bellot[1, *]
(1) Aix-Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
(2) Université de Toulon, Aix-Marseille Université, CNRS, LIS, Toulon, France
(3) Hyperbios, Aix-en-Provence, France
(*) prénom.nom@lis-lab.fr

RÉSUMÉ

Dans cet article, nous proposons des approches pour améliorer les architectures basées sur des transformeurs pour la classification de documents multi-labels. Les dépendances entre les labels sont cruciales dans ce contexte. Notre méthode, appelée DepReg, ajoute un terme de régularisation à la fonction de perte pour encourager le modèle à prédire des labels susceptibles de coexister. Nous introduisons également un nouveau jeu de données nommé "arXiv-ACM", composé de résumés scientifiques de la bibliothèque numérique arXiv, étiquetés avec les mots-clés ACM correspondants.

ABSTRACT

Exploiting Label Dependencies for Multi-Label Document Classification Using Transformers.

In this paper, we propose approaches to improve transformer-based architectures for multi-label document classification. Dependencies between labels are crucial in this context. Our method, called DepReg, adds a regularization term to the loss function to encourage the model to predict labels that may coexist. We also introduce a new dataset named "arXiv-ACM", consisting of scientific abstracts from the arXiv digital library, labeled with the corresponding ACM keywords.

MOTS-CLÉS : Classification multi-labels, Transformeurs, Dépendances entre labels.

KEYWORDS: Multi-label Classification, Transformers, Label Dependencies.

1 Introduction

La classification multi-labels peut être considérée comme une généralisation de la classification binaire ou multi-classes traditionnelle. Le but est d'associer un ou plusieurs labels au texte d'entrée. C'est une tâche importante pour de nombreuses applications telles que la classification d'articles de recherche (Mustafa *et al.*, 2021; Sajid *et al.*, 2011) ou la réponse aux questions (Sahu *et al.*, 2019; Wu *et al.*, 2019).

Plusieurs méthodes ont été proposées pour résoudre la classification multi-labels. Elles peuvent être divisées en trois familles : transformation de problème (Luaces *et al.*, 2012; Tsoumakas *et al.*, 2010), adaptation de problème et méthodes d'ensemble (Saini & Ghosh, 2017; Tsoumakas & Vlahavas, 2007). Dans la configuration multi-labels, il est nécessaire de trouver des caractéristiques discriminantes pour identifier chacun d'entre eux dans le texte donné, mais dans de nombreuses applications, des

dépendances peuvent exister entre les labels cibles.

Nous proposons dans cet article une approche d'adaptation de problème qui consiste dans l'exploitation de la corrélation des labels pour la classification de documents multi-labels qui tire parti de leurs cooccurrences de manière simple mais efficace. Pour cette approche, nous proposons un terme de régularisation qui est ajouté à la fonction de perte de la tâche de classification, basé sur les labels prédits et la matrice de similarité de labels. Une matrice calculée en utilisant la similarité cosinus entre les coocurrences de chaque label par rapport aux autres.

Nous proposons un nouveau jeu de données multi-labels arXiv-ACM, qui diffère de ceux présents dans la littérature de part sa grande cardinalité ainsi qu'une meilleure distribution des échantillons par nombre de labels. Il est construit à partir de résumés d'articles scientifiques de la bibliothèque numérique arXiv et appariés avec les mots-clés ACM de niveau 2 fournis par les auteurs.

2 Travaux connexes

Diverses méthodes ont été proposées pour modéliser la dépendance entre les labels en utilisant des structures hiérarchiques (Alaydie *et al.*, 2012; Yang *et al.*, 2016), des graphes et réseaux conditionnels (Guo & Gu, 2011; Zhang & Zhang, 2010), des cooccurrences de labels, ou encore une combinaison de ces approches (Wu *et al.*, 2018). Toutefois, ces méthodes tendent à privilégier les relations verticales entre les labels plutôt que les dépendances latérales.

MAGNET (Pal *et al.*, 2020), un réseau de neurones en graphes qui utilise les plongements de mots de BERT, met en œuvre le mécanisme d'attention pour capturer la structure de dépendance entre les labels. LW-PT (Liu *et al.*, 2020) introduit une nouvelle tâche de pré-entraînement de classification de documents par labels et entraîne des encodeurs de documents par labels. Ces deux méthodes parviennent à avoir de bonnes performances en score F1 pour les ensembles de données AAPD et Reuters (cf. section 4) tout en utilisant des LSTMs pour l'extraction des caractéristiques textuelles.

Dans le domaine de la vision par ordinateur, et plus précisément dans la détection d'objets qui est essentiellement une tâche de classification multi-labels, (Cheng *et al.*, 2021) utilise un modèle transformeur et une attention au niveau des pixels pour capturer les dépendances spatiales dans une image, ainsi qu'un jeton spécifique à l'objet qui est ajouté à une étape ultérieure dans le modèle. Ce jeton est utilisé pour prédire le nombre de labels pour une instance donnée et peut aider le modèle à établir une correspondance plus robuste entre les caractéristiques d'entrée et les labels à prédire.

Une approche alternative consiste dans l'utilisatiion des fonctions de perte d'équilibrage pour améliorer la performance des labels sous-représentés (Huang *et al.*, 2021). Cependant, cela peut réduire l'information apportée par les cooccurrences de labels qui peuvent corriger les biais des modèles et améliorer la performance de la classification pour les labels moins fréquentes.

Dans Liu *et al.* (2022) (CNLE), les labels sont encodés sous forme de plongements de mots et alimentés avec la séquence de texte dans un réseau de co-attention (Seo *et al.*, 2016). Les représentations de texte en fonction des labels et les labels en fonction du texte sont ensuite utilisées pour la classification multi-labels en la considérant comme un problème de génération de séquences. Le mécanisme d'attention tient compte de la relation entre les labels, mais son efficacité est limitée dans les cas où les labels ne sont pas des mots complets mais plutôt des abréviations ou des codes, comme "cs.it" dans le jeu de données AAPD. Dans de tels cas, les labels sont initialisées avec une distribution

de probabilité plutôt que des plongements pré-entraînés, ce qui ne fournit pas une représentation optimale pour les labels. Dans la Section 4, nous montrons que notre approche ne dépend pas de la nature des labels et contribue à une amélioration des performances pour tous les ensembles de données étudiés.

3 Régularisation par Dépendance (DepReg)

Nous présentons ici notre méthode de Régularisation par Dépendance (DepReg), qui ajoute un terme de régularisation à la fonction de coût du modèle transformeur pour incorporer les informations de cooccurrence des labels.

Nous construisons une matrice de similarité S en utilisant la matrice de cooccurrence C.

$$S = CS_\theta(C, C) \tag{1}$$

Le terme de régularisation L_{DepReg} est calculé comme le produit scalaire entre le vecteur de dissimilarité D_{sim} et le vecteur de prédiction transposé $\hat{Y}^T$.

$$D_{sim} = 1 - CS_\theta(S, \hat{Y}) = 1 - \frac{S \cdot \hat{Y}}{|S| \cdot |\hat{Y}|} \tag{2}$$

$$L_{DepReg} = D_{sim} \cdot \hat{Y}^T \tag{3}$$

Le terme de régularisation est ajouté à la fonction de coût principale lors de l'apprentissage pour encourager le modèle à faire des prédictions cohérentes avec les dépendances des labels présentes dans le jeu de données :

$$L_{total} = L_{BCE} + \lambda_{reg}.L_{DepReg} \tag{4}$$

où L_{BCE} est la perte d'entropie croisée binaire (Binary Cross Entropy) et λ_{reg} est un hyperparamètre contrôlant le poids du terme de régularisation dans la perte totale.

Le terme de régularisation par dépendance (DepReg) aide le modèle à éviter de faire des prédictions qui vont à l'encontre des informations de cooccurrence tout en lui permettant de faire des prédictions basées sur les motifs appris dans les données d'entraînement.

4 Expériences et Résultats

Pour l'évaluation des méthodes proposées, nous utilisons l'implémentation de HuggingFace du modèle BERT (Devlin *et al.*, 2019) en version *uncased-base*. Nous ajoutons un réseau de neurones (FFNN) avec $L = 2$ couches et utilisons l'entropie croisée binaire (BCE) pour la version de base de BERT et pour notre approche DepReg. AdamW est l'optimiseur utilisé avec un λ_{reg} de 0.2.

Jeux de données :

— **AAPD** (arXiv Academic Paper Dataset) est une collection de "résumés" d'articles scientifiques de la bibliothèque numérique arXiv. Un article peut avoir une ou plusieurs classifications parmi 54 labels. Nous utilisons la même distribution d'entraînement (53840), de validation (1000)

et de test (1000) que (Yang *et al.*, 2018). Cet ensemble de données présente de nombreuses limites. La plus contraignante est le nombre d'instances par labels. En général, les instances avec un seul label sont beaucoup plus courantes que celles avec plusieurs.

— Pour remédier à ces limitations, nous introduisons dans cet article un nouveau jeu de données multi-labels, que nous appelons "arXiv-ACM", avec une grande cardinalité, une taille raisonnable et une meilleure distribution des échantillons selon le nombre de labels. **ArXiv-ACM** est composé de résumés d'articles en informatique extraits via l'API arXiv [1], publiés entre 1998 et 2021. Ces résumés ont ensuite été associés aux mots-clés ACM [2] fournis par les auteurs des articles lors de la soumission. Seuls les mots-clés de deuxième niveau ont été considérés, car le premier niveau est trop large et les niveaux suivants sont trop spécifiques. Nous avons ensuite filtré les labels qui ont moins de 20 instances pour obtenir 64 labels.

Baselines :

— **MAGNET** (Pal *et al.*, 2020) : classification de texte multi-labels utilisant un réseau de neurones en graphes avec mécanisme d'attention pour capturer les dépendances entre les labels,

— **CNLE** (Liu *et al.*, 2022) : un modèle de transformeur qui introduit les plongments des labels en plus de ceux du texte, liés par une co-attention pour obtenir une représentation contextualisée de la séquence d'entrée par les labels de classification,

Résultats : Comme le montre le tableau 1, l'utilisation des informations de dépendance contenues dans la matrice de cooccurrence des labels entraîne une augmentation du score micro-F1 pour les deux jeux de données avec notre approche.

SVM peut être considérée comme la meilleure approche non neuronale, mais elle est inférieure aux autres méthodes testées. La précision plus élevée a un coût de rappel plus faible, réduisant ainsi le score micro F1. La précision seule n'est pas un facteur fiable pour mesurer les performances dans les tâches de classification.

AAPD - En raison de la nature du vocabulaire utilisé dans les résumés scientifiques, ce jeu de données est complexe. Les scores de performance montrent que les modèles ont du mal à associer les sujets à leurs résumés correspondants. L'apprentissage de la dépendance obtient cette fois-ci une augmentation du rappel avec un score micro-F1 de 73,81.

arXiv-ACM - Ce jeu de données partage la nature scientifique des documents avec AAPD. Malgré une distribution plus équilibrée du nombre de labels par instance, les scores sont les plus bas pour les modèles testés. La méthode DepReg présente le gain le plus élevé en terme de précision avec un gain de 1,76 par rapport à la version base de BERT. Cette augmentation notable, associée à un gain de rappel, contribue au meilleur score micro-F1 pour cet ensemble de données (58,08), avec la plus forte augmentation par rapport à la version de base de BERT.

L'augmentation des performances obtenue par l'approche d'apprentissage de dépendance que nous proposons peut être expliquée par le fait que la prédiction d'un label est influencée par la prédiction de tous les autres labels en utilisant les co-occurrences. Dans certains cas, ces informations aident à prédire des labels qui n'auraient pas été prédites autrement (augmentation du rappel). D'autre part, les dépendances entre les labels peuvent réduire le nombre de faux positifs en réduisant le biais que le modèle peut avoir pour les labels fréquentes dans l'ensemble de données, contribuant ainsi à une amélioration de la précision.

1. https ://arxiv.org/help/api/
2. https ://www.acm.org/publications/computing-classification-system/1998/ccs98

Models	arXiv-ACM			AAPD		
	Pr.	R	F1	Pr.	R	F1
Baselines						
GradientBoost	57,99	29,87	39,43	79,73	46,8	58,98
SVM	**70,15**	39,79	50,78	**80,85**	59,98	68,86
MAGNET	57,31	53,24	55,2	72,88	66,79	69,7
CNLE	56,85	52,37	54,52	74,71	69,11	71,80
BERT	60,04	55,58	57,72	74,49	72,03	73,24
Our Dependency Learning Approaches						
BERT+ *DepReg*	61,80	**56,09**	**58,08**	75,53	**72,16**	**73,81**

TABLE 1 – Micro-précision (Pr.), micro-rappel (R) et scores micro-F1 (F1) pour les ensembles de tests arXiv-ACM et AAPD. Les meilleurs sont en bleu gras.

Models	arXiv-ACM				AAPD			
	Head	Med	Tail	3+ subset	Head	Med	Tail	3+ subset
BERT	58,95	52,11	**41,55**	54.32	73,97	69,29	**64,82**	66,47
BERT+*DepReg*	**59.80**	**56.62**	40.38	**55,39**	74,56	**69,75**	64,23	**67,18**

TABLE 2 – Scores Micro-F1 pour les labels de tête, de moyenne et de queue. Les meilleurs scores sont en bleu gras.

Head, Med, Tail - Pour étudier l'influence du déséquilibre des données, nous suggérons d'analyser les résultats d'inférence sur les labels tête (Head), moyenne (Med) et queue (Tail) selon la fréquence des labels pour arXiv-ACM (head > 600 instances, Med entre 50-600, Tail $\leq$ 50) et AAPD (head >3000 instances, Med entre 1000-3000, Tail $\leq$ 1000). Notre approche basée sur la cooccurrence de labels réalise une augmentation notable des performances, mais seulement dans les sous-ensembles tête et moyen, (cf. tableau 2).

3+ Subset - Pour une évaluation plus pratique des gains de performance en classification de documents, nous analysons également le sous-ensemble de chaque jeu de données contenant au moins 3 labels. Notre méthode de dépendance de labels obtient généralement les scores les plus élevés pour ces sous-ensembles. Ce qui montre que notre méthode capture efficacement les dépendances entre labels et conduit à une amélioration des performances.

5 Conclusion

La classification multi-label est une tâche pertinente, en particulier pour la gestion des bibliothèques numériques et l'étiquetage automatique de documents. Dans cet article, nous avons proposé une méthode simple mais efficace pour utiliser les informations de cooccurrence des labels pour permettre aux modèles à base de transformeurs d'apprendre les dépendances entre les labels. La méthode de régularisation de dépendance (DepReg) s'est avérée particulièrement efficace, améliorant les

performances du modèle BERT de base. Cependant, cette approche pénalise les labels moins fréquents en raison de leurs faibles probabilités de cooccurrence. Atténuer cet inconvénient peut conduire à un gain de performance plus notable. Enfin, nous avons introduit dans cet article un nouveau jeu de données multi-label, l'ensemble de données "arXiv-ACM", plus adapté pour tester les nouvelles approches multi-label. Notre jeu de données et tout le code de mise en œuvre seront disponibles au moment de la publication [3].

Références

ALAYDIE N., REDDY C. K. & FOTOUHI F. (2012). Exploiting Label Dependency for Hierarchical Multi-label Classification. In P.-N. TAN, S. CHAWLA, C. K. HO & J. BAILEY, Éds., *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, p. 294–305, Berlin, Heidelberg : Springer.

CHENG X., LIN H., WU X., YANG F., SHEN D., WANG Z., SHI N. & LIU H. (2021). Mltr : Multi-label classification with transformer. *CoRR*, **abs/2106.06195**.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

GUO Y. & GU S. (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, p. 1300–1305, Barcelona, Catalonia, Spain : AAAI Press.

HUANG Y., GILEDERELI B., KÖKSAL A., ÖZGÜR A. & OZKIRIMLI E. (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*, p. 8153–8161, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.643.

LIU H., YUAN C. & WANG X. (2020). Label-Wise Document Pre-training for Multi-label Text Classification. In X. ZHU, M. ZHANG, Y. HONG & R. HE, Éds., *Natural Language Processing and Chinese Computing*, Lecture Notes in Computer Science, p. 641–653, Cham : Springer International Publishing.

LIU M., LIU L., CAO J. & DU Q. (2022). Co-attention network with label embedding for text classification. *Neurocomputing*, **471**, 61–69. DOI : https ://doi.org/10.1016/j.neucom.2021.10.099.

LUACES O., DÍEZ J., BARRANQUERO J., DEL COZ J. J. & BAHAMONDE A. (2012). Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, **1**(4), 303–313.

MUSTAFA G., USMAN M., YU L., AFZAL M., SULAIMAN M. & SHAHID A. (2021). Multi-label classification of research articles using word2vec and identification of similarity threshold. *Scientific Reports*, **11**, 21900. DOI : 10.1038/s41598-021-01460-7.

PAL A., SELVAKUMAR M. & SANKARASUBBU M. (2020). Multi-Label Text Classification using Attention-based Graph Neural Network. In *ICAART*.

SAHU T. P., THUMMALAPUDI R. S. & NAGWANI N. K. (2019). Automatic Question Tagging Using Multi-label Classification in Community Question Answering Sites. In *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, p. 63–68. DOI : 10.1109/CSCloud/EdgeCom.2019.00-17.

3. https ://github.com/hf-lis/Coria-TALN-2023

SAINI R. & GHOSH S. (2017). Ensemble classifiers in remote sensing : A review. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, p. 1148–1152. DOI : 10.1109/CCAA.2017.8229969.

SAJID N. A., ALI T., AFZAL M. T., AHMAD M. & QADIR M. A. (2011). Exploiting reference section to classify paper's topics. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, MEDES '11, p. 220–225, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2077489.2077531.

SEO M. J., KEMBHAVI A., FARHADI A. & HAJISHIRZI H. (2016). Bidirectional attention flow for machine comprehension. *CoRR*, **abs/1611.01603**.

TSOUMAKAS G., KATAKIS I. & VLAHAVAS I. (2010). Mining Multi-label Data. In O. MAIMON & L. ROKACH, Éds., *Data Mining and Knowledge Discovery Handbook*, p. 667–685. Boston, MA : Springer US.

TSOUMAKAS G. & VLAHAVAS I. (2007). Random k -Labelsets : An Ensemble Method for Multilabel Classification. volume 4701, p. 406–417.

WU B., JIA F., LIU W., GHANEM B. & LYU S. (2018). Multi-label Learning with Missing Labels Using Mixed Dependency Graphs. *International Journal of Computer Vision*, **126**(8), 875–896. DOI : 10.1007/s11263-018-1085-3.

WU H., ZHANG S., WANG J., LIU M. & LI S. (2019). Multi-label Aspect Classification on Question-Answering Text with Contextualized Attention-Based Neural Network. In M. SUN, X. HUANG, H. JI, Z. LIU & Y. LIU, Éds., *Chinese Computational Linguistics*, Lecture Notes in Computer Science, p. 479–491, Cham : Springer International Publishing.

YANG P., SUN X., LI W., MA S., WU W. & WANG H. (2018). SGM : Sequence Generation Model for Multi-label Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 3915–3926, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

YANG Z., YANG D., DYER C., HE X., SMOLA A. & HOVY E. (2016). 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*, p. 1480–1489, San Diego, California : Association for Computational Linguistics.

ZHANG M.-L. & ZHANG K. (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, p. 999–1008, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1835804.1835930.