Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Rééquilibrer la distribution des labels tout en éliminant le temps d'attente inhérent dans l'apprentissage actif multi-label appliqué aux transformers

Maxime Arens[1, 2] Lucile Callebert[2] Jose G. Moreno[1] Mohand Boughanem[1]
(1) IRIT, Toulouse University, UMR 5505 CNRS, 31400 Toulouse, France
(2) Synapse Développement, 7 Boulevard de la Gare, 31500 Toulouse, France
maxime.arens@gmail.com

RÉSUMÉ

L'annotation des données est cruciale pour l'apprentissage automatique, notamment dans les domaines techniques, où la qualité et la quantité des données annotées affectent significativement l'efficacité des modèles entraînés. L'utilisation de personnel humain est coûteuse, surtout lors de l'annotation pour la classification multi-label, les instances pouvant être associées à plusieurs labels. L'apprentissage actif (AA) vise à réduire les coûts d'annotation en sélectionnant intelligemment des instances pour l'annotation, plutôt que de les annoter de manière aléatoire. L'attention récente portée aux transformers a mis en lumière le potentiel de l'AA dans ce contexte. Cependant, dans des environnements pratiques, la mise en œuvre de l'AA rencontre des défis pratiques. Notamment, le temps entre les cycles d'AA n'est pas mis à contribution par les annotateurs. Pour résoudre ce problème, nous examinons des méthodes alternatives de sélection d'instances, visant à maximiser l'efficacité de l'annotation en s'intégrant au processus de l'AA. Nous commençons par évaluer deux méthodes existantes, en utilisant respectivement un échantillonnage aléatoire et des informations de cycle d'AA périmées. Ensuite, nous proposons notre méthode novatrice basée sur l'annotation des instances pour rééquilibrer la distribution des labels. Notre approche atténue les biais, améliore les performances du modèle (jusqu'à une amélioration de 23% sur le score F1), réduit les disparités dépendantes de la stratégie (diminution d'environ 50% sur l'écart type) et diminue le déséquilibre des libellés (diminution de 30% sur le ratio moyen de déséquilibre). [1]

ABSTRACT

Rebalancing Label Distribution while Eliminating Inherent Waiting Time in Multi Label Active Learning applied to Transformers.

Data annotation is crucial for machine learning, notably in technical domains, where the quality and quantity of annotated data, significantly affect effectiveness of trained models. Employing humans is costly, especially when annotating for multi-label classification, as instances may bear multiple labels. Active Learning (AL) aims to alleviate annotation costs by intelligently selecting instances for annotation, rather than randomly annotating. Recent attention on transformers has spotlighted the potential of AL in this context. However, in practical settings, implementing AL faces challenges beyond theory. Notably, the gap between AL cycles presents idle time for annotators. To address this issue, we investigate alternative instance selection methods, aiming to maximize annotation efficiency by seamlessly integrating with the AL process. We begin by evaluating two existing methods in our transformer setting, employing respectively random sampling and outdated information. Following

1. Cet article a fait l'objet d'une publication en anglais à LREC-COLING 2024 (Arens *et al.*, 2024).

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 16–17.

this we propose our novel method based on annotating instances to rebalance label distribution. Our approach mitigates biases, enhances model performance (up to 23% improvement on f1score), reduces strategy-dependent disparities (decrease of nearly 50% on standard deviation) and reduces label imbalance (decrease of 30% on Mean Imbalance Ratio).

MOTS-CLÉS : apprentissage active, transformers, temps d'attente, distribution des labels.

KEYWORDS: active learning, transformers, wait time, label distribution.

Références

ARENS M., CALLEBERT L., BOUGHANEM M. & MORENO J. G. (2024). Rebalancing label distribution while eliminating inherent waiting time in multi label active learning applied to transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 13621–13632.