運用響應式知識蒸餾機制增進中文多標籤文本分類效能
Enhancing Chinese Multi-Label Text Classification Performance with Response-based Knowledge Distillation

黃思齊 Szu-Chi Huang, 曹程富 Cheng-Fu Cao, 廖柏勛 Po-Hsun Liao
李龍豪 Lung-Hao Lee, 李柏磊 Po-Lei Lee, 徐國鎧 Kuo-Kai Shyu
國立中央大學 電機工程學系
Department of Electrical Engineering, National Central University
{110521103, 110521109, 110521086}@cc.ncu.edu.tw
{lhlee, pllee, kkshyu}@ee.ncu.edu.tw

摘要

資料類別不平衡存在長尾標籤問題，單獨的多標籤分類模型一次預測所有類別，針對個別標籤的最佳化十分困難，對於出現次數較少的長尾標籤效能通常不佳。本論文提出一種響應式知識蒸餾機制，將多個最佳化的二元模型作為教師網路，單一多標籤模型做為學生網路，改善多標籤模型在非平衡標籤的資料集分類效能。實驗資料來自 2,724 個中文健康照護文本，人工標記文章內容橫跨 9 個類別， 總共標籤數量是 8,731，平均每個樣本有 3.2 個標籤。實驗設定採用 5 折交互驗證，比較 TextRNN、TextCNN、HAN 和 GRU-att 模型，使用知識蒸餾機制與否的效能差異，結果顯示透過知識蒸餾機制能夠顯著提升單一多標籤分類模型的 micro-F1 約 2 至 3 %、macro-F1 約 4 至 6 %、weighted-F1 約 3 至 4 %，以及 subset accuracy 約 1 至 2 %。

Abstract

It's difficult to optimize individual label performance of multi-label text classification, especially in those imbalanced data containing long-tailed labels. Therefore, this study proposes a response-based knowledge distillation mechanism comprising a teacher model that optimizes binary classifiers of the corresponding labels and a student model that is a standalone multi-label classifier learning from distilled knowledge passed by the teacher model. A total of 2,724 Chinese healthcare texts were collected and manually annotated across nine defined labels, resulting in 8731 labels, each containing an average of 3.2 labels. We used 5-fold cross-validation to compare the performance of several multi-label models, including TextRNN, TextCNN, HAN, and GRU-att. Experimental results indicate that using the proposed knowledge distillation mechanism effectively improved the performance no matter which model was used, about 2-3% of micro-F1, 4-6% of macro-F1, 3-4% of weighted-F1 and 1-2% of subset accuracy for performance enhancement.

關鍵字：多標籤分類、長尾標籤、二元相關、知識蒸餾
Keywords: Multi-label classification, long-tailed labels, binary relevance, knowledge distillation

1 介紹

多標籤文本分類 (Multi-Label Text Classification) 廣泛用於許多應用，例如：廣告系統 (Agrawal et al., 2013)、情緒分析(Myagmar et al., 2019)、推薦系統 (Guo et al., 2016) 、幽默辨識 (Kao et al., 2021) 以及問答系統 (Chen et al., 2021)等。多標籤文本分類主要有兩種方法類型 (Tsoumakas and Katakis, 2007)，分別是 (1) 問題轉換：透過訓練多個二元分類模型來達成多重標籤分類，以及 (2) 訓練單一模型進行多重標籤分類。問題轉換類型可以透過調整類別權重(class weights) 的方式，針對單一類

別進行最佳化，提供更高的彈性 (Banerjee et al., 2019)，而當資料標籤不平衡時，訓練單一模型則難以最佳化各類別的預測效能。但訓練單一模型能夠依據標籤間的關聯性，或是潛在架構對模型進行優化，而問題轉換法則無法有效利用標籤間的相關性。

知識蒸餾 (knowledge distillation) 是一種將巨型的教師網路 (teacher net) 學習到的「知識」，轉移到較精簡的學生網路 (student net) 的深度學習技術 (Hinton et al., 2015)。我們假設最佳化多個二元分類器網路，在各類別上擁有比單一多標籤模型更佳的效能，將最佳化的多個二元分類器模型作為教師網路，單一多標籤模型作為學生網路進行訓練，目標是讓單一多標籤模型可以透過教師網路，在非平衡標籤資料上學習知識，用以改善單一多標籤模型在非平衡標籤上效能低落的缺點。

本篇論文提出利用知識蒸餾機制增進多標籤文本分類效能的方法，在中文健康照護文本上，驗證數個不同深度學習模型包含 TextRNN (Liu et al., 2016)、TextCNN (Liu et al., 2017)、HAN (Yang et al., 2016)、GRU-Att (Banerjee et al., 2019) 的效能差異，實驗結果顯示使用知識蒸餾機制與原先的單一多標籤分類模型相比，提升 2~3%的 micro-F1、4~6%的 macro-F1、3~4%的 weighted-F1 以及 1~2%的 subset accuracy。

2 相關研究

2.1 知識蒸餾

大規模的機器學習通常使用非常複雜的模型訓練，大量地計算導致模型參數量過大，同時也耗費計算資源。因此，Hinton (2015) 提出一種稱為「知識蒸餾」的模型壓縮方式，將繁瑣大型模型訓練完成後的知識，以「蒸餾」(distillation) 的方式，轉移到更適合使用的小型模型。知識蒸餾的種類可分為以下三種類型 (Gou et al., 2021)：

(1) Response-based

蒸餾的知識主要來源是教師模型的最終輸出層，通過使用損失函數 (loss function) 取得學生和教師模型的之間的差異，並在訓練過程中將損失最小化，最終讓學生模型能夠做出跟教師模型一樣的預測。

(2) Feature-based

主要是從教師模型神經網路的中間層中取得特徵，通過最小化教師和學生模型的特徵損失函數，訓練學生模型學習與教師模型相同的特徵。

(3) Relation-based

與 Feature-based 相似同樣都是從教師模型中取得特徵，差別在於先將特徵建立為圖形、矩陣、或是概率分佈圖，再與學生模型比較關係的差異，藉此訓練學生模型。

2.2 多標籤文本分類

多標籤文本分類與多元文本分類 (multi-class text classification) 都是具有兩個以上類別的分類任務。不同之處在於，多元分類每個標籤都是相互排斥的，分類標的樣本是從多個類別選擇一個。多標籤分類則可以為每個樣本分配多個不同標籤，而這些標籤並不相互排斥，這種分類方式也比較接近人類的思考，可以從多個角度來描述一件事物。

TextRNN 模型 (Liu et al., 2016) 採用長短期記憶神經網路 (Long Short-Term Memory, LSTM) 進行訓練。模型包含三個控制閘：輸入、輸出、遺忘，分別對應寫入、讀取以及遺忘的功能。透過這三個控制閘，模型能夠額外考慮前文的關係，使得當前的輸出不只受上一層輸入的影響，也受到同一層前一個輸出（即前文）的影響。

HAN 模型 (Yang et al., 2016) 使用 GRU 神經網路進行訓練。訓練時將 GRU 神經網路當作編碼器(word encoder)對每句話的詞語先進行注意力機制 (attention)，再疊上一層同樣的結構，用 GRU 對所有句子進行注意力機制，最後輸出分類結果。

TextCNN 模型 (Liu et al., 2017) 使用靜態（有微調）和非靜態（無微調）通道表示句子，並經過多個過濾器和特徵映射進行捲積，方便捕獲豐富的語義，同時採用動態最大池化 (dynamic max pooling)，生成多個特徵後，均分特徵的資訊，最後使用 dropout 和 Sigmoid 輸出分類。

GRU-Attention 模型 (Banerjee et al., 2019) 的結構便是在經過 GloVe (Pennington et al, 2014) 的預訓練後，將單詞序列輸入 GRU 之後進行注意力機制，最後線性結合最大池化(max-

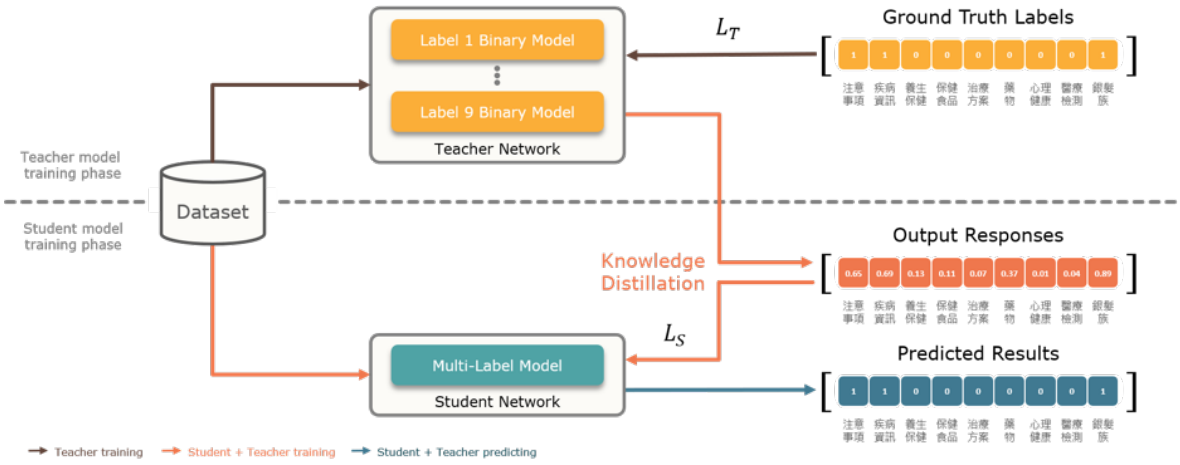


圖 1、基於知識蒸餾的多標籤分類模型訓練流程

pool) 和平均池化(mean-pool)的向量，並且輸出結果。

資料數量和分類標籤種類的增加，導致大部分資料通常集中在少數幾個標籤，多數的標籤只有極少數的資料，出現長尾 (long-tail) 效應。因此，我們提出一個響應式知識蒸餾的訓練方法，希望能改善長尾標籤的分類效能。

3 基於知識蒸餾的模型

我們提出一個基於知識蒸餾(Knowledge Distillation, KD) 的多標籤文本分類方法，目標是將二元分類器的優點應用在多標籤分類器上，用以改進單個多標籤分類器的效能。模型訓練流程如圖 1 所示，教師模型由多個二元分類模型組成，假設資料集有 n 個類別，就會有 n 個二元分類模型，學生模型則是一個多標籤分類模型。教師與學生所採用的模型架構皆相同，最大的差異在於模型最後一層的輸出數量。在訓練時，我們先訓練教師模型中的 N 個二元分類模型，針對相對應類別在驗證資料集的表現進行最佳化，以確保教師模型的效能，可以優於單個多標籤分類器的表現。然後，擁有最佳表現的教師模型會產生響應 (responses)，讓學生模型來學習。在學生模型訓練階段，學生預測目標是教師響應的機率值分佈，而不是目標標籤 (ground truth labels)，藉此，學生就能從老師的響應中得到損失值(loss)進行訓練。

我們在訓練教師模型時，透過類別權重 (class weights) 優化訓練時權重參數的更新。類別權重是一種常被用來處理資料不平衡問題的方法，對訓練集裡的每個類別加上一個權重。理論上，如果該類別的樣本數越多，那麼它被加上的權重就要越低；反之，樣本數越少的權重就給得越高。如此一來就會改變原本的損失函數，使得較多樣本數類別的損失函數變小，較少樣本數類別的損失函數變大。加上這個限制會迫使模型在訓練期間進行權重更新時，提升少樣本數類別(長尾標籤)的準確度，而樣本數多的類別則沒有明顯差異。

我們採用響應式(response-based) 知識蒸餾，此機制的主要想法為讓學生模型直接模仿教師模型所預測出來類別的機率分佈。我們使用 binary cross entropy 做為損失函數，因此教師模型中每個二元分類模型的損失函數 $L_T(y, z_t)$，加上類別權重 W 的損失函數變成加權平均，可以用方程式(1)表示，其中 y_i 代表第 i 個樣本的真實標籤值、z_i^t 代表教師模型第 i 個樣本的輸出值。

$$L_T(y, z_t) = -\frac{1}{N}\sum_{i=1}^{N} W \cdot y_i \cdot \log(z_{s_i}) + (1 - y_i) \cdot \log(1 - z_{s_i}) \quad (1)$$

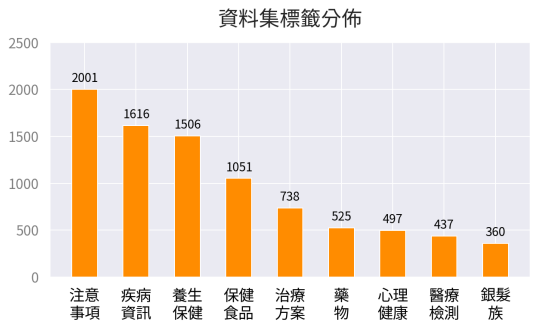


圖 2、資料集標籤分佈

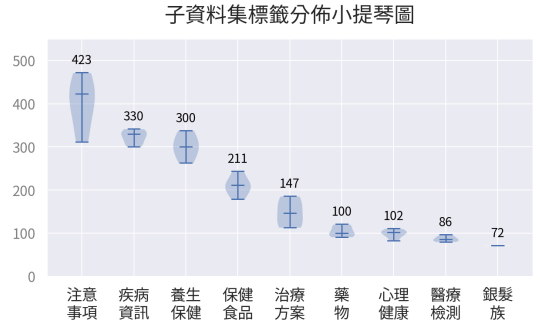


圖 3、子資料集標籤分佈

教師模型訓練完成後會對整個訓練資料集進行預測，產生各類別的機率分佈（響應），然後我們就可以用損失函數$L_S(z_t, z_s)$來訓練學生模型，其中z_s代表學生模型的輸出值。模型最後輸出的激活函數是 sigmoid，超過定義的門檻值(threshold)，即為模型的預測標籤。

$$L_S(y, z_t) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(z_{s_i}) + (1 - y_i) \cdot \log(1 - z_{s_i}) \quad (2)$$

4 實驗評估與結果

4.1 資料集

本研究實驗的資料來自網路爬取的中文健康照護網站，例如：康健雜誌、國家醫藥網路、健康醫療網等的文章內容，去除網頁標籤、圖片及詮釋資料(metadata)等雜訊，僅留下純文字內容做人工類別標記，整個資料集最終有 2,724 個文檔，橫跨 9 個類別標籤，包含：健康表現、心理健康、治療方案、醫療檢測、保健食品、注意事項、藥物和銀髮族，總共標籤數量是 8,731，平均每個樣本有 3.2 個標籤。各標籤的樣本數量如圖 2 所示，「注意事項」這個標籤出現次數最多，佔整個資料集約 73%，而出現次數最少的標籤是「銀髮族」，僅佔約 13%。前半段 4 個類別出現的平均樣本數目是 1543.5，對比後半段的 4 個類別平均出現的樣本數是 454.8，落差約 3.4 倍，雖然類別標籤僅有 9 類，文本數目也還不夠多，整個資料集的標籤分佈不明顯，但還是可以看到存在長尾效應。

我們使用五折交叉驗證 (5-fold cross-validation)，分割資料集時，從最少樣本數的標籤開始隨機平均分配樣本至各子資料集，這樣能夠保證樣本數較少的標籤能夠在交叉驗證中被均分。圖 3 為分割後的子資料集標籤分佈的小提琴圖(violin plot)，圖中標示數值為中位數，可以看到與原資料集的分佈大致相同。

4.2 實驗設定

我們使用 Tensorflow 函式庫進行模型實作。所有文本使用 CKIP Transformer[1] 套件進行斷詞，模型輸入皆使用維度 300 的 Word2Vec (Mikolov et al., 2013)，並訓練在繁體中文維基百科語料庫[2]以及實驗資料集上。

我們分別比較不同深度學習模型，包含 TextRNN (Liu et al., 2016)、TextCNN (Liu et al., 2017)、HAN (Yang et al., 2016) 以及 GRU-Att (Banerjee et al., 2019)，導入知識蒸餾機制與否導致的效能差異。模型的訓練參數如下：文本長度 300 字元、Batch Size 32 (GRU-att 是 128) 、損失函數 Binary Cross Entropy、優化器 Adam、訓練迭代次數 30、早停法 patience 是 10、學習率 1e-3、輸出層激活函數 sigmoid 以及驗證集比率 15%。

每個模型比較以下三種不同訓練方法：

(1) 二元最佳化 (Binary Optimization)

[1] https://github.com/ckiplab/ckip-transformers

[2] https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2

Model	Type	Micro F1	Macro F1	Weighted F1	Subset Accuracy
TextRNN	Binary Optimization (Teacher)	72.11	62.65	71.93	13.70
	Standalone Multi-label Classifier (Student)	71.09	57.15	68.82	15.85
	Knowledge Distillation (Teacher + Student)	**74.29**	**63.76**	**73.52**	**16.24**
TextCNN	Binary Optimization (Teacher)	75.89	**68.35**	75.48	18.75
	Standalone Multi-label Classifier (Student)	74.55	63.57	72.88	18.75
	Knowledge Distillation (Teacher + Student)	**76.80**	67.68	**75.72**	**19.70**
HAN	Binary Optimization (Teacher)	75.73	69.36	75.60	17.79
	Standalone Multi-label Classifier (Student)	74.63	66.08	73.66	19.10
	Knowledge Distillation (Teacher + Student)	**77.54**	**71.00**	**77.15**	**21.22**
GRU-Att	Binary Optimization (Teacher)	73.30	64.70	73.06	14.93
	Standalone Multi-label Classifier (Student)	71.90	61.44	70.88	14.69
	Knowledge Distillation (Teacher + Student)	**75.17**	**67.63**	**75.26**	**16.52**

表 1、多標籤分類模型實驗結果

知識蒸餾機制中單獨使用教師網路 (teacher net) 的作法。採用二元相關 (binary relevance) 轉換，各類別獨立訓練分類器調整類別權重，在最佳化二元相關轉換法的模型時，我們每個類別使用不同類別權重（0.05、0.1、0.5、1、2、3 和 5）訓練出多個模型，並選擇在驗證集表現最好的模型，最佳的權重值大致與標籤數量呈反比，標籤數越多類別權重愈小，反之，標籤數越少類別權重愈大。

(2) 單獨的多標籤分類模型 (Standalone Multi-label Classifier)

可以視為知識蒸餾機制中單獨使用學生網路 (student net) 的作法。輸入為 300 維的 Word2Vec (Mikolov et al., 2013)，直接使用真實標籤值計算模型的預測誤差訓練模型。

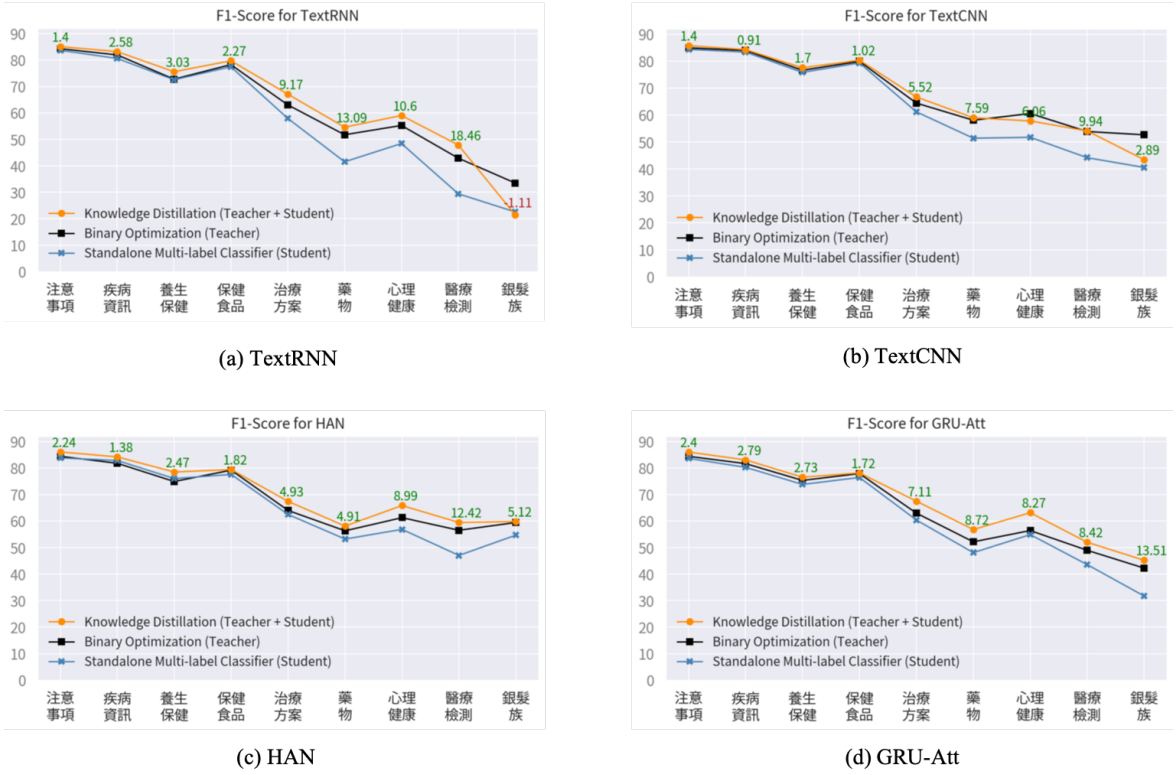
(3) 知識蒸餾機制 (Knowledge Distillation)

我們提出的基於知識蒸餾機制的多標籤分類模型訓練方式，包含教師網路和學生網路 (teacher net + student net)。先使用真實標籤值訓練出教師網絡，再以教師網路的預測值作為輸出響應，接著訓練學生網絡時，使用教師產生的響應計算預測誤差，使學生網路學習教師網路的行為。

我們使用 micro-F1、marco-F1、weighted-F1 以及 subset accuracy 評估模型的效能表現。micro-F1 不區分類別計算整體的 F1 分數。marco-F1 計算各類別 F1 分數的平均值。weighted-F1 計算各類別 F1 分數後，依據各類別的數量進行加權平均。subset accuracy 是最嚴格的評估指標，需要所有類別都是對的才判定是對，評估完全正確預測的比例。

4.3 實驗結果

表 2 為多標籤分類模型實驗結果。基於二元相關轉換訓練出來的模型，透過調整不同的類別權重進行最佳化後，在 micro-F1、macro-F1 與 weighted-F1，皆優於相同模型架構的單獨多標籤模型，表示二元最佳化的作法能夠做為教師模型，用以改進多標籤分類模型的分類效能。但二元相關轉換法訓練出的模型，



(a) TextRNN (b) TextCNN

(c) HAN (d) GRU-Att

圖 4、 模型各類別標籤效能差異

沒辦法考慮類別間的相關性，導致要準確預測樣本的所有標籤較為困難，所以 subset accuracy 比多標籤模型低。

實驗結果顯示透過我們提出的知識蒸餾機制透訓練的多標籤模型，在不同模型架構下能夠顯著提升多標籤模型的性能 (p-value <0.01)，甚至與教師模型相比，因多標籤模型能夠在訓練的過程中獲取標籤間的相關性，相較於只有 0 或 1 的標籤值，機率值分佈擁有更多資訊，使得知識蒸餾的多標籤模型具有比二元最佳化的教師模型有更好的效能。

圖 4 為模型在不同訓練方法下，各類別標籤的 F1 (separated-F1)。我們可以看出二元相關轉換法訓練出的模型，因為每個分類器只需考慮單一類別，並且已透過不同類別權重進行最佳化，當訓練資料不平衡時，對效能的影響相對較小。

而多標籤分類模型在藥物、心理健康、醫療檢測、銀髮族等樣本數較少標籤，因為訓練資料不平衡造成分類效能明顯下降，但透過知識蒸餾訓練的多標籤模型，在這些長尾標籤相對於多樣本標籤能夠有更明顯的效能提升。

5 結論

我們提出基於知識蒸餾的多標籤分類模型訓練方法，將最佳化的二元相關轉換法作為教師網路，對多標籤分類模型進行響應式知識蒸餾，用以改善多標籤分類模型。實驗資料來自人工標記的 2,724 個多標籤中文健康照護文本，橫跨 9 個類別標籤，標籤數量是 8,731，平均每個樣本有 3.2 個標籤。實驗結果顯示使用知識蒸餾的訓練方法，在不同的深度學習模型，無論 micro-F1、macro-F1、weighted-F1 以及 subset accuracy 皆能有顯著的效能提升。

Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan, under the grant MOST 111-2628-E-008-005-MY3.

References

Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. 2013. Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd International Conference on World Wide Web*, Association for Computing Machinery, pages 13–24. https://doi.org/10.1145/2488388.2488391

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 6295–6300. http://doi.org/10.18653/v1/P19-1633

Po-Han Chen, Yu-Xiang Zeng, and Lung-Hao Lee. 2021. Incorporating domain knowledge into language transformers for multi-Label classification of Chinese medical questions. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing*. Association for Computational Linguistics, pages 265–270.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. 2021. Knowledge distillation: a survey. *International Journal of Computer Vision*, 129:1789-1819. https://doi.org/10.1007/s11263-021-01453-z

Li Guo, Bo Jin, Ruiyun Yu, Cuili Yao, Chonglin Sun, and Degen Huang. 2016. Multi-label classification methods for green computing and application for mobile medical recommendations. *IEEE Access*, 4:3201-3209. https://doi.org/10.1109/ACCESS.2016.2578638

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. arXiv preprint, arXiv:1503.02531 https://doi.org/10.48550/arXiv.1503.02531

Hao-Chuan Kao, Man-Chen Hung, Lung-Hao Lee, Yuen-Hsien Tseng. 2021. Multi-label classification of Chinese humor texts using hypergraph attention networks. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing*. Association for Computational Linguistics, pages 257–264.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, pages 115-124 https://doi.org/10.1145/3077136.3080834

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. http://dx.doi.org/10.3115/v1/D14-1162

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. pages 2873-2879. https://arxiv.org/abs/1605.05101

Batsergelen Myagmar, Jie Li, and Shigetomo Kimura. 2019. Cross-domain sentiment classification with bidirectional contextualized transformer language models. *IEEE Access*, 163219-163230. https://doi.org/10.1109/ACCESS.2019.2952360

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013, Efficient estimation of word representations in Vector Space. *arXiv Preprint* arXiv:1301.3781 https://doi.org/10.48550/arXiv.1301.3781

Grigorios Tsoumakas, and Ioannis Katakis. 2007. Multi-label classification: an overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13. https://doi.org/10.4018/jdwm.2007070101

Ran Wang, Xi'ao Su, Siyu Long, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Meta-LMTC: meta-learning for large-scale multi-label text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 8633–8646. http://doi.org/10.18653/v1/2021.emnlp-main.679

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1480–1489. http://doi.org/10.18653/v1/N16-1174