Deep Streaming Label Learning

Zhen Wang [1] **Liu Liu** [1] **Dacheng Tao** [1]

Abstract

In multi-label learning, each instance can be associated with multiple and non-exclusive labels. Previous studies assume that all the labels in the learning process are fixed and static; however, they ignore the fact that the labels will emerge continuously in changing environments. In order to fill in these research gaps, we propose a novel deep neural network (DNN)-based framework, Deep Streaming Label Learning (DSLL), to classify instances with newly emerged labels effectively. DSLL can explore and incorporate the knowledge from past labels and historical models to understand and develop emerging new labels. DSLL consists of three components: 1) a streaming label mapping to extract deep relationships between new labels and past labels with a novel label correlation-aware loss; 2) a streaming feature distillation propagating feature-level knowledge from the historical model to a new model; 3) a *senior student* network to model new labels with the help of knowledge learned from the past. Theoretically, we prove that DSLL admits tight generalization error bounds for new labels in the DNN framework. Experimentally, extensive empirical results show that the proposed method performs significantly better than the existing state-of-the-art multi-label learning methods to handle the continually emerging new labels.

1. Introduction

In traditional supervised learning, a single instance only possesses a single label; whereas, in many real-world tasks, one instance can be naturally associated with multiple and non-exclusive labels. For example, a document may belong to various topics (Zhang & Zhou, 2006; Nam et al., 2014); an image can contain several individuals' faces (Xiao et al., 2010); a gene may be related to multiple functions (Barutcuoglu et al., 2006; Cesa-Bianchi et al., 2012). Multi-label learning has been proposed and studied to handle such kind of tasks, with the goal of predicting a label vector $\boldsymbol{y} \in \{0,1\}^m$ for a given instance $\boldsymbol{x} \in \mathbb{R}^d$, where m is the number of labels, and d is the dimensionality of the input instance.

A straightforward solution to multi-label learning is binary relevance (BR) (Tsoumakas et al., 2010) and one-vs-all (Menon et al., 2019), which treat each label as an independent binary classification problem. However, in addition to a lack of correlations between label information, such approaches suffer from high computational and storage costs for large datasets, which would limit the prediction performance. As a result, there are two prevalent techniques to deal with multi-label learning: (1) exploring and exploiting the correlations across labels (Zhang & Zhou, 2006; Read et al., 2011; Nam et al., 2019), and (2) reducing the label space by embedding original high-dimensional label vectors into the low-dimensional representation (Tsoumakas et al., 2011; Bhatia et al., 2015; Liu & Shen, 2019). In particular, deep neural networks (DNNs) (LeCun et al., 2015; Schmidhuber, 2015) offer a good way to learn the label correlation and embedding simultaneously, significantly outperforming classical methods (Nam et al., 2014; Yeh et al., 2017; Chen et al., 2018).

Overall, existing multi-label learning studies focus on a fixed set of labels. That is, they assume that all the labels in the learning process are given at once. However, this assumption is frequently violated in real-world changing environments. In practice, with more in-depth data exploration and understanding, the number of labels gradually increases. For example, when a photograph is posted on Facebook or Twitter, the photo will be labeled continuously and differently by users who are browsing, and the classification system needs to be updated accurately according to the new labels. In the multi-label learning setting, independently learning only for new labels ignores the correlated information from past labels; integrating past labels with emerging new labels to retrain a new multi-label model would be prohibitively computationally expensive. Hence, streaming

[1]UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia. Correspondence to: Zhen Wang <zwan4121@uni.sydney.edu.au>, Liu Liu <liu.liu1@sydney.edu.au>, Dacheng Tao <dacheng.tao@sydney.edu.au>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

label learning (SLL) (You et al., 2016) has been proposed to handle this problem, which assumes that the new label vectors are a linear combination of past label vectors and that the new classifier can inherit the linear combination to obtain the correlations across labels without retraining the historical classifier. Nevertheless, the linear representation between new labels and past labels limits SLL performance, and the training knowledge from the historical classifiers is neglected.

In this paper, we propose a novel DNN-based framework, Deep Streaming Label Learning (DSLL), to effectively model the emerging new labels. DSLL has the ability to explore and utilize deep correlations between new labels and past labels without computationally intensive retraining. Furthermore, inspired by knowledge distillation (Hinton et al., 2014; Romero et al., 2015), a technique that distills knowledge or feature representations from a large *teacher* model to a smaller *student* model, DSLL develops the approach to leverage the knowledge from previous learning to model new labels more accurately. DSLL has three components: (1) a streaming label mapping to exploit deep relationships between new labels and past labels with a novel label correlation-aware loss; (2) a streaming feature distillation framework to propagate the feature-level knowledge from a past-label classifier (teacher) to a new-label classifier (student); (3) a *senior student* network to integrate the relationships and knowledge learned from the past to model the newly arrived labels.

In order to verify the effectiveness and performance of DSLL, both theoretical and experimental analyses are conducted. From the theoretical perspective, we present proofs that the generalizability of modeling emerging new labels can be significantly improved with the knowledge extracted from past labels. In contrast to linear frameworks (Yu et al., 2014), proofs for a nonlinear model can be challenging, especially in DNNs. We analyze the excess risk bounds for the proposed learning model by leveraging the latest deep learning theory. From the experimental perspective, we perform a comprehensive empirical evaluation of our model on a variety of benchmark datasets, using prevalent metrics and comparisons with well-established or state-of-the-art multi-label learning algorithms. The results show that our approach considerably outperforms other methods and demonstrates the significance of utilizing knowledge learned from past labels and their corresponding classifiers in modeling new labels.

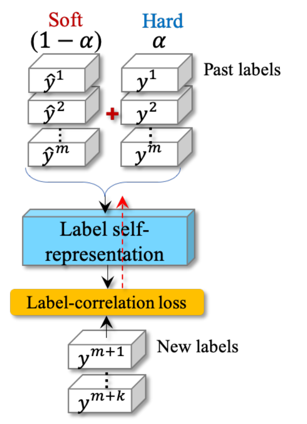
The rest of the paper is organized as follows. Section 2 summarizes the previous related work. Section 3 formulates the problem and describes the proposed model in detail. Section 4 presents the theoretical analysis. Section 5 presents the experimental results validating the model and providing several insights. We conclude in Section 6. All detailed proofs are provided in the Appendix to the Supplementary Materials. The anonymized code is available here[1].

2. Related Work

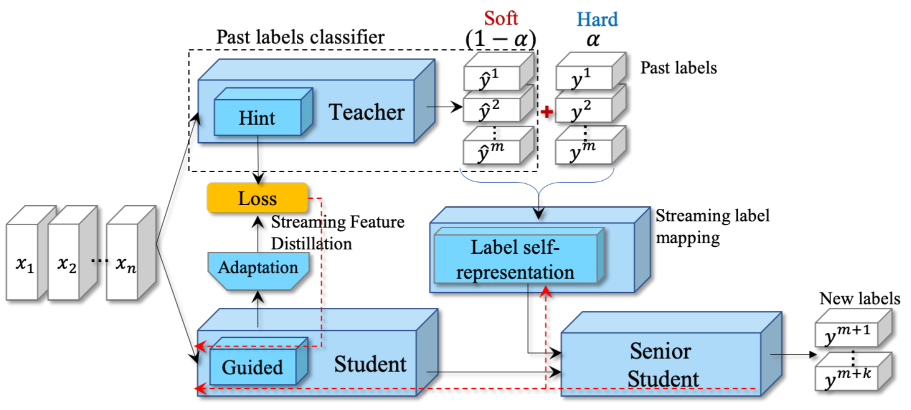
Multi-label Learning. Obvious baselines for multi-label learning (MLL) are provided by the one-vs-all technique (Menon et al., 2019) and binary relevance (Tsoumakas et al., 2010), which seek to learn an independent classifier for each label. As expected, the huge training and prediction costs for a large number of labels and the lack of ability to discover the correlations between labels are limitations. Label embedding (LE) is a popular strategy for MLL that projects the label vectors onto a lower subspace with the latent embedding of corresponding information and utilizes a decompression mapping to lift the embedded label vectors back to the original label space (Tsoumakas et al., 2011; Prabhu & Varma, 2014). SLEEC (Bhatia et al., 2015) is the state-of-the-art LE method that captures the embedding of labels by preserving the local distances between a few nearest label neighbors, and SML (Liu & Shen, 2019) further improves the convergence rate of SLEEC. Deep learning is another popular end-to-end learning approach. BP-MLL (Zhang & Zhou, 2006) introduces the neural network architecture to learn relevant features automatically. Since then, other DNN techniques have been applied to multi-label learning (Nam et al., 2014). C2AE (Yeh et al., 2017) is a state-of-the-art DNN-based approach that integrates canonical correlation analysis (Andrew et al., 2013) and autoencoder to utilize features and label embedding jointly. In summary, all these MLL studies explicitly and implicitly focus on a fixed set of labels, which is frequently violated in the changing environments. Although traditional multi-label algorithms can be adapted to handle emerging new labels, they can suffer from the lack of correlations between past labels and new labels for only learning new labels or prohibitive computational costs caused by retraining the model with whole labels.

Streaming Label Learning. To model newly arrived labels expediently and effectively, streaming label learning (SLL) (You et al., 2016) has been proposed, which aims to employ the knowledge from past labels. SLL assumes that: (1) a label is represented as a linear combination of other labels, and (2) the relationship (linear combination) between labels can be inherited by classifiers of different labels. Constrained by these hypotheses, SLL trains a linear classifier for new labels with the relationship between past labels and new labels. However, the linear representation across labels limits the SLL performance, and the training knowledge from classifiers of past labels is neglected. Note that while Zhu et al. (2018) proposed a method, in which instances with emerging new labels are regarded as outliers,

[1]https://github.com/DSLLcode/DSLL



(a) Streaming Label Mapping



(b) Deep Stream Label Learning

Figure 1. The overall DSLL architecture. The proposed framework has three components: 1) streaming label mapping, designed to capture the relationship between new labels and past labels with the proposed label correlation-aware loss, where *hard* is the ground truth of past labels and *soft* is the prediction provided by the past-labels classifier (teacher); 2) streaming feature distillation, which transforms the knowledge from the teacher to the new-label classifier (student), employing the outputs of the intermediate layer of the teacher as *hint* to guide those of the student; 3) the senior student network, which leverages the relationship and knowledge to finally model new labels. The red dotted lines denote the backpropagation path during learning.

this is different to the SLL problem.

Knowledge Distillation. Knowledge distillation (KD) aims to transfer the knowledge from an existing powerful classifier (called a *teacher*) to a lightweight classifier (called a *student*). For instance, Ba & Caruana (2014) trained a shallow student network to mimic the *logits* provided by a deep teacher network. Hinton et al. (2014) proposed a more general framework by training a small student to predict softened labels of the output of a large teacher network. FitNet (Romero et al., 2015) employed the intermediate layer representation of the teacher as a *hint* to guide student training. Due to its effectiveness, KD has been being used in many fields such as object detection (Chen et al., 2017), model compression (Heo et al., 2019), image classification (Liu et al., 2018), machine translation (Zhou et al., 2020), and meta-learning (Bai et al., 2020). In this work, we propose *streaming* feature distillation, where the past-label classifier acted as a teacher helps guide a student network to model new labels. It not only extracts knowledge from the intermediate layer of the teacher, but also takes advantage of the softened output provided by the well-trained teacher to construct the relationship between new and past labels.

3. Deep Streaming Label Learning

Conventional multi-label learning encounters the following dilemmas in the face of newly-arrived labels: (1) integrating past labels and new labels to retrain a new multi-label model requires massive computational resources; (2) independently learning for the new labels would neglect the knowledge acquired from past labels and historical models. This section will detail the proposed Deep Streaming Label Learning to handle the continuously emerged labels, which can solve these challenging problems.

3.1. Problem Formulation

First, we state the MLL with emerging new labels problem. Assume an initial training data set is denoted by $\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, where $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^{d\times 1}$ is a real vector representing an input feature (instance) and $\boldsymbol{y}_i \in \mathcal{Y} \subseteq \{0,1\}^{m\times 1}$ is the corresponding output label vector ($i \in [n]$, defined as $i \in \{1, ..., n\}$). Moreover, $y_i^j = 1$ if the j-th label is assigned to the instance $\boldsymbol{x}_i$ and $y_i^j = 0$ otherwise. The input matrix is $\mathbf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ and the initial output matrix is $\mathbf{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_n] \in \{0,1\}^{m\times n}$. By observing $\mathcal{D}$, multi-label learning aims to derive a proper learning model $\mathbf{M}_m$ that generates the prediction on a label vector for a test instance. Then new labels data arrive in a streaming fashion. For simplicity, we denote $\boldsymbol{y}_i^{new} = [y_i^{m+1}, y_i^{m+2}, ..., y_i^{m+k}]$ as the new k-labels vector for instance $\boldsymbol{x}_i$, where $k \geq 1$. The MLL problem in the changing environment is to derive a learning model on the continually emerging new labels.

3.2. Overall Structure

We propose DSLL, designed to accommodate emerging new labels. Our overall learning framework is illustrated in Figure 1. DSLL can overall be regarded as a new label classifier consisting of three components. First, we construct streaming label mapping $\mathbf{S}_m$ to capture and maintain the relationships from past labels to new labels, followed by the proposed label correlation loss function to preserve cross-label dependency (Section 3.3). As shown in Figure 1(a), $\mathbf{S}_m$ projects past label vectors $\boldsymbol{y}$ to new label vectors $\boldsymbol{y}^{new}$. In the testing process, we can only use the prediction value $\hat{\boldsymbol{y}}$ provided by the past-label classifier as the input of $\mathbf{S}_m$, since label vectors are unknown. Therefore, in the training process, we jointly employ the true past label vectors $\boldsymbol{y}$ (hard) and the prediction $\hat{\boldsymbol{y}}$ (soft) as the input of the mapping $\mathbf{S}_m$ to improve its accuracy and robustness. The shallow label mapping network has much fewer parameters than the large classifier, which can extract label correlation without computationally intensive retraining. Second, we present the streaming feature distillation, a hint-based learning approach (Romero et al., 2015), where past-label classifier serves as a teacher network to guide the student network. Streaming feature distillation encourages the feature representation of the student to mimic that of the teacher network at the intermediate layer, which boosts the final performance of the new-label classifier (Section 3.4). Besides, streaming feature distillation network, as the initialization of the new-label classifier, facilitates the convergence of DSLL. Finally, we learn integrally the label relationships obtained by $\mathbf{S}_m$ and the knowledge acquired from the teacher by training a senior student network with the cross-entropy loss function to model the newly-arrived labels (Section 3.5).

3.3. Streaming Label Mapping with Label-correlation Loss

Exploring and exploiting the relationship between labels has been demonstrated to be advantageous and crucial for boosting MLL performance (Zhang & Zhou, 2014; Yeh et al., 2017; Menon et al., 2019). We propose the streaming label mapping $\mathbf{S}_m : \mathbb{R}^{m\times 1} \rightarrow \{0,1\}^k$ to captures deep relationships between new labels and past labels effectively and inexpensively. Taking advantage of the relationships, DSLL can model new labels more accurately and avoids the huge computational cost associated with retraining all the labels to obtain label correlations. In the training process of $\mathbf{S}_m$, as shown in Figure 1(a), we use a combination of the *soft* label vector $\hat{\boldsymbol{y}}^j$ provided by the past-label classifier and the *hard* label vector $\boldsymbol{y}^j$ denoting ground truth of the past label as the input $\tilde{\boldsymbol{y}}^j$:

$$\tilde{\boldsymbol{y}}^j = \alpha \boldsymbol{y}^j + (1-\alpha)\hat{\boldsymbol{y}}^j,$$

where $j \in [m]$ and $\alpha \in [0,1]$ balance the importance of soft predictions and the true hard labels. Note that $[\tilde{\boldsymbol{y}}_1, ..., \tilde{\boldsymbol{y}}_i] = [\tilde{\boldsymbol{y}}^1, .., \tilde{\boldsymbol{y}}^j]^T$, $i \in [n]$. Distinct from KD, which places the soft labels as the target to learn (Hinton et al., 2014; Heo et al., 2019), here we employ both soft labels and hard labels as the input, aiming to make the streaming label mapping more robust in learning the relationship between labels. This is because, on the one hand, the hard past labels keep the unbroken latent relationship with newly-arrived labels, and on the other hand, in the testing process, the hard labels are unknown and $\mathbf{S}_m$ only relies on the soft labels provided by the past-label classifier. Therefore, joint training of hard labels and soft labels can improve the accuracy and robustness of the streaming label mapping and boosts DSLL performance (See Section 5.2). Specifically, $\mathbf{S}_m$ consists of a fully connected network with ReLU activation function in the hidden layers and sigmoid function in the output layer.

Inspired by Zhang & Zhou (2006), we present a novel label correlation-aware loss at the sigmoid output of the streaming label mapping, which is defined as follows:

$$\mathcal{L}_{\mathbf{S}}(\tilde{\boldsymbol{y}},\boldsymbol{y}^{new})= \sum_{i=1}^{n} \frac{1}{|Y_i^+||Y_i^-|} \sum_{(p,q)\in Y_i^+\times Y_i^-} \|(\mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^p - \mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^q) - b\|_2^2 \quad (1)$$

where Y_i^+ (or Y_i^-) denotes the index set of new labels associated (or non-associated) with $\boldsymbol{x}_i$. Formally, $Y_i^+ = \{j \mid y_i^j = 1\}$ and $Y_i^- = \{j \mid y_i^j = 0\}$, $j \in \{m+1, ..., m+k\}$; $b = \max \mathcal{Y} - \min \mathcal{Y}$ is a constant[2]. Given the past label vector $\tilde{\boldsymbol{y}}_i$ as input, $\mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^p - \mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^q$ returns the discrepancy between the output of the mapping $\mathbf{S}_m$ on one label associated with $\boldsymbol{x}_i$ ($p \in Y_i^+$) and one label not associated with it ($q \in Y_i^-$). Note that $(\mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^p) \in [-1,1]$ due to the sigmoid function after the output layer. Therefore, minimizing the label correlation-aware loss is equal to maximizing $\mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^p - \mathbf{S}_m(\tilde{\boldsymbol{y}}_i)^q$, which implicitly learns the paired correlation between different labels and enforces the preservation of cross-label dependency.

3.4. Streaming Feature Distillation

Hinton et al. (2014) proposed a knowledge distillation (KD) method using only the final output of the teacher network. Hint learning (Romero et al., 2015) demonstrates that using the intermediate representation of the teacher as a *hint* to guide that of the student can improve the final performance of the student. We propose the streaming feature distillation, a hint based learning, which transforms knowledge from the past-label classifier (as a teacher) to the new-label classifier (as a student) to boost the performance of the new-label classifier. As shown on the left side of Figure 1(b), the hint layer is represented as the output of a teacher's intermediate layer guiding a hidden layer of the student, the guided layer, to mimic it. Since the selected hint layer may have more outputs (neurons) than the guided layer, we add an adaptation matrix parameterized by $\mathbf{W}_{Adapt}$ after the guided layer to match the output size of the hint layer. Then, we train the student parameters up to the guided layer by minimizing the following loss function:

$$\mathcal{L}_H(\mathbf{R}_{Guided}, \mathbf{W}_{Adapt}) = \|\mathbf{R}_{Hint} - \mathbf{W}_{Adapt}\mathbf{R}_{Guided}\|_2^2,$$

[2]for current settings, $b = 1$

where $\mathbf{R}_{Hint}$ and $\mathbf{R}_{Guided}$ represent the output of the hint layer in the teacher and the output of the guided layer in the student, respectively. Finally, the student parameters up to the guided layer would be saved as the pre-trained parameters and kept fine-tuned with a slow learning rate in later training. Streaming feature distillation outperforms training from scratch because the pre-trained model already extracts a great deal of feature-level knowledge.

3.5. Learning from the Past

The *senior student* network is designed to integrally digest the knowledge learned from past labels and the historical classifier. As shown in Figure 1(b), the senior student receives the outputs of both the streaming label mapping and the student network as inputs and generates the prediction for new labels. In particular, we add one hidden layer after the streaming label mapping as a nonlinear transformation $\mathbf{S}_t$, which can transform the outputs of the streaming label mapping to a higher space. Overall, training the senior student can be considered as the process from instance $\boldsymbol{x}_i$ and the output of streaming label mapping to the corresponding new label vector $\boldsymbol{y}_i^{new}$. In the learning stage, the parameters of the senior student, the transformation $\mathbf{S}_t$ and the student (included the pre-trained guided layer) would be trained simultaneously by minimizing the cross-entropy loss as follows:

$$\mathcal{L}_{CE} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=m+1}^{m+k} -y_i^j \log(\hat{y}_i^j) - (1-y_i^j)\log(1-\hat{y}_i^j),$$

where $\hat{y}_i^j$ is the prediction of jth label corresponding to the instance $\boldsymbol{x}_i$ generated by the senior student, and y_i^j indicates the ground truth.

The red dotted line from the senior student in Figure 1(b) denotes the backpropagation path. The streaming label mapping is kept frozen to preserve the relationships between new labels and past labels. The student parameters up to the guided layer could be trained at a slow learning rate to fine-tune. Interestingly, we find that having a transformation $\mathbf{S}_t$ is practical for achieving effective label mapping, especially in the case of dimension raising (see Section 5.2).

4. Theoretical Analysis

In this section, we theoretically analyze the excess risk bounds for our learning model and provide a generalization error bound for the new-label classifier DSLL. Our analysis demonstrates that the generalizability for modeling the emerging new labels can be improved with the knowledge extracted from past labels. In contrast to linear frameworks (Yu et al., 2014), it is a bigger challenge to provide proofs for nonlinear models, especially DNNs.

For notational simplicity, we denote $\boldsymbol{y}_i^{new} = [y_i^{m+1}, y_i^{m+2}, ..., y_i^{m+k}] \in \{0,1\}^k$ as the new k-labels vector for instance $\boldsymbol{x}_i$, and $\boldsymbol{x}_i^* = [x_i^1, x_i^2, ..., x_i^d, y_i^1, y_i^2, ..., y_i^m]$ as the i-th tuple input vector combined from $\boldsymbol{x}_i$ and the past label vector $\boldsymbol{y}_i$. We use $\phi(x)=\max\{0,x\}$ to represent the ReLU function, and extend it to vectors $v \in \mathbb{R}^d$ by letting $\phi(\boldsymbol{v}) = (\phi(v_1), ..., \phi(v_d))$. We use $\mathbb{1}_{event}$ to denote the indicator function for $event$. We define the following network as the basic composition unit of the DSLL model,

$$\begin{aligned} g_{i,0} &= \mathbf{A}\boldsymbol{x}_i^* & h_{i,0} &= \phi(\mathbf{A}\boldsymbol{x}_i^*) & &\text{for } i \in [n] \\ g_{i,l} &= \mathbf{W}_l h_i^{l-1} & h_i^l &= \phi(\mathbf{W}_l h_i^{l-1}) & &\text{for } i \in [n], l \in [L] \\ \hat{\boldsymbol{y}}_i &= \mathbf{B}h_i^L & & & &\text{for } i \in [n], \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{u_1 \times (d+m)}$ is the weight matrix for the input layer, $\mathbf{W}_l \in \mathbb{R}^{u_{l+1} \times u_l}$ is the weight matrix for the l-th hidden layer, $\mathbf{B} \in \mathbb{R}^{k \times u_L}$ is the weight matrix for the output layer, and u_l is the number of neurons in the hidden layer l. For notational convenience, we use h_i^{-1} to denote input vector, $\mathbf{W}_0$ to denote the corresponding weight matrix for the input layer, and u_0 to denote the input data dimension.

Recently, there has been some work on deep learning theory (Zou et al., 2019; Cao & Gu, 2020; Du et al., 2019). Allen-Zhu et al. (2019) present a more simple and intuitive analysis. To simplify the proof as (Allen-Zhu et al., 2019), we define the following *diagonal sign matrix* to linearize the DNN corresponding to each instance $\boldsymbol{x}_i$.

Definition 4.1 (diagonal sign matrix). *For each $i \in [n]$ and $l \in \{0, 1, ..., L\}$, we denote by $\mathbf{D}_i^l$ the diagonal sign matrix where $(\mathbf{D}_i^l)_k^k = \mathbb{1}_{(\mathbf{W}_l h_i^{l-1})_k \geq 0}$ for each $k \in [u_l]$.*

As a result, we have $h_i^l = \mathbf{D}_i^l \mathbf{W}_l h_i^{l-1} = \mathbf{D}_i^l g_i^l$ and $(\mathbf{D}_i^l)_k^k = \mathbb{1}_{(g_i^l)_k \geq 0}$. DSLL integrates three effective networks: Streaming Student $\mathbf{W}_\mathcal{S}$, Streaming Label Mapping $\mathbf{W}_\mathcal{Y}$, and Senior Student $\mathbf{W}_\Delta$. These three networks constitute the DSLL model $\mathbf{W}$ structure, defined as

$$\mathbf{W} = \mathbf{W}_\Delta \left[\mathbf{W}_\mathcal{S}^T, \mathbf{W}_\mathcal{Y}^T\right]^T,$$

which uses $\boldsymbol{x}^*$, the tuples of data $\boldsymbol{x}$ and past labels $\boldsymbol{y}$, as the input to predict the new labels $\boldsymbol{y}^{new}$. According to Definition 4.1, each input tuple $\boldsymbol{x}_i^*$ corresponds to a network parameter in which the prediction is obtained by

$$\hat{\boldsymbol{y}_i}^{new} = f_i(\boldsymbol{x}_i^*, \mathbf{W}) = \mathbf{B}\mathbf{D}_i^L \mathbf{W}_L, ..., D_i^1 \mathbf{W}_1 \mathbf{D}_i^0 \mathbf{A}\boldsymbol{x}_i^*, \quad i \in [n].$$

DSLL learns a classifier $\hat{\mathbf{W}}$ by minimizing the *empirical risk*,

$$\hat{\mathcal{L}}(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=m+1}^{m+k} \ell(y_i^j, f_i^j(\boldsymbol{x}_i^*, \mathbf{W})),$$

where $\ell(\cdot)$ denotes the loss function, and we can obtain $\hat{\mathbf{W}} = \text{argmin}_{\mathbf{W} \in \mathcal{W}} \hat{\mathcal{L}}(\mathbf{W})$.

Our goal is to show that the learned $\hat{\mathbf{W}}$ is generalizable, i.e.,

$$\mathcal{L}(\hat{\mathbf{W}}) \leq \inf_{\mathbf{W} \in \mathcal{W}} \mathcal{L}(\mathbf{W}) + \epsilon,$$

where $\mathcal{L}(\mathbf{W})$ is the *population risk* of a classifier, defined

Table 1. Statistics of five real-world datasets.

Dataset	Domain	#Training	#Testing	#Feature	#Labels	#Card-Features	#Card-Labels
yeast	biology	1,500	917	103	14	103.00	4.24
MirFlickr	image	10,417	2,083	1,000	38	539.33	4.74
Delicious	web	12,920	3,185	500	983	18.17	19.03
EURlex	text	15,479	3,869	5,000	3,993	5.31	236.69
Wiki10	text	14,146	6,616	101,938	30,938	673.45	18.64

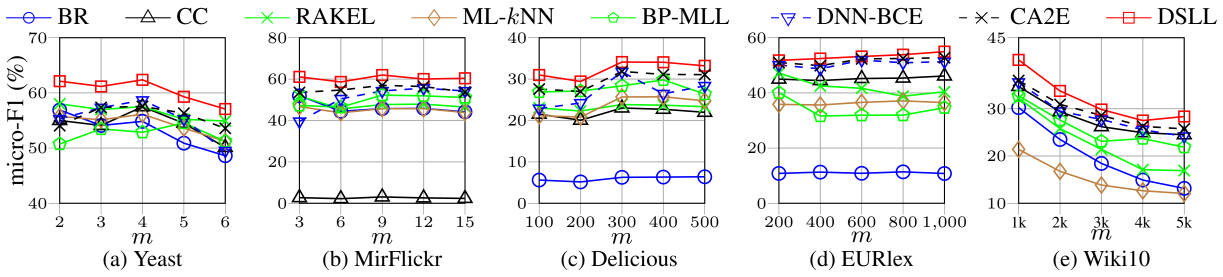


Figure 2. Performance comparison of learning new labels with different batch sizes by considering 50% of labels as past labels. m indicates the number of new labels.

as:

$$\mathcal{L}(\mathbf{W}) := \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^{new})}\left[\!\!\left[\sum_{j=m+1}^{m+k} \ell(\boldsymbol{y}^j, f^j(\boldsymbol{x}^*, \mathbf{W})) \right]\!\!\right].$$

And we give the following theorem.

Theorem 1. *Suppose we learn a new classifier* $\mathbf{W}$ *in terms of* k *new labels using formulation* $\hat{\mathbf{W}} = \operatorname{argmin}_{\mathbf{W}\in\mathcal{W}} \hat{\mathcal{L}}(\mathbf{W})$ *over a set of* n *training instances. Then, with a probability of at least* $1-\delta$*, we have*

$$\mathcal{L}(\hat{\mathbf{W}}) \leq \inf_{\mathbf{W}\in\mathcal{W}} \mathcal{L}(\mathbf{W}) + \mathcal{O}\left(\sqrt{\frac{k}{n}}\left(\gamma_x + \gamma_y\sqrt{k}\right)\right) + \mathcal{O}\left(k\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right),$$

where γ_x *and* γ_y *are real constant numbers related to the number of neurons in the output layer. We assume, without loss of generality, that* $\mathbb{E}[\![\|\boldsymbol{x}\|_2^2]\!] \leq 1$, $\mathbb{E}[\![\|\boldsymbol{y}^{new}\|^2]\!] \leq k$.

Refer to Appendix A in the Supplementary Materials for the proof. Theorem 1 demonstrates that knowledge learned from past labels helps to improve the generalizability to model new labels without compromising model accuracy.

5. Experiments

In this section, we conduct extensive experiments to demonstrate the performance of the proposed method with respect to handling new labels. All the computations are performed on a 64-Bit Linux workstation with 10-core Intel Core CPU i7-6850K 3.60GHz processor, 256 GB memory, and 4 Nvidia GTX 1080 Ti GPUs.

5.1. Experimental Setup

Datasets. We use five readily available multi-label benchmark datasets from different application domains, including three regular-scale datasets[3,4] (Yeast, MirFlickr and Delicious) and two large-scale datasets[5] (EURlex and Wiki10). The statistics of these five real-world data sets are summarized in Table 1.

Baselines. In our experiments, we compare the proposed methods with ten well-established or state-of-the-art multi-label learning algorithms:

- *Binary relevance* (BR) (Tsoumakas et al., 2010) is a set of m independent logistic regression classifiers.
- *Classifier chains* (CC) (Read et al., 2011) transforms the multi-label learning problem into a chain of binary classification problems to incorporate label dependencies into the classification process.
- RAKEL (Tsoumakas et al., 2011) considers a set of k labels in a multi-label training set as one of the classes of a new single-label classification task.
- ML-kNN (Zhang & Zhou, 2007) identifies each unseen instance's k nearest neighbors in the training set and utilizes the maximum a posteriori principle to determine the label set for the unseen instance.
- SLEEC (Bhatia et al., 2015) is a well-known embedding-based method, which learns the label embedding by preserving the pairwise distances between a few nearest label neighbors.

[3]http://mulan.sourceforge.net/
[4]https://github.com/chihkuanyeh/C2AE
[5]http://manikvarma.org/downloads/XC/XMLRepository.html

Table 2. Ranking performance of each comparison algorithm for learning new labels with different batch sizes by regarding 50% of labels as past labels. #label denotes the number of new labels. ↓(↑) means the smaller (larger) the value, the better the performance.

Datasets	#label	micro-AUC ↑										
		BR	CC	RAKEL	ML-kNN	SLEEC	SML	SLL	BP-MLL	DNN-BCE	C2AE	DSLL
yeast	2	0.6703	0.6750	0.6914	0.6731	0.7003	0.7114	0.7209	0.7228	0.7616	0.7624	**0.7793**
	3	0.7140	0.6972	0.7162	0.7074	0.8003	0.8287	0.7302	0.8437	0.8521	0.8590	**0.8641**
	4	0.6978	0.7096	0.7113	0.7027	0.7751	0.7955	0.7401	0.8216	0.8383	0.8384	**0.8530**
	5	0.6883	0.6979	0.7061	0.6964	0.7775	0.7889	0.7203	0.8139	0.8325	0.8318	**0.8507**
	6	0.6768	0.6743	0.7059	0.6834	0.7645	0.7751	0.7133	0.8115	0.8079	0.8199	**0.8229**
MirFlickr	3	0.6589	0.5049	0.6572	0.6296	0.5476	0.6123	0.7291	0.7448	0.7592	0.8051	**0.8122**
	6	0.7100	0.5045	0.7187	0.6624	0.6400	0.6959	0.8388	0.8476	0.8606	0.8628	**0.8713**
	9	0.7311	0.5071	0.7170	0.6754	0.7001	0.7331	0.8671	0.8692	0.8778	0.8890	**0.8972**
	12	0.7151	0.5057	0.7223	0.6692	0.6862	0.7116	0.8630	0.8600	0.8763	0.8807	**0.8886**
	15	0.7157	0.5051	0.7253	0.6611	0.6805	0.7015	0.8624	0.8584	0.8787	0.8749	**0.8873**
Delicious	100	0.7133	0.5658	0.6382	0.5707	0.7813	0.8274	0.7981	0.8776	0.8615	0.8545	**0.8820**
	200	0.7080	0.5606	0.6400	0.5692	0.7815	0.8258	0.7956	0.8626	0.8677	0.8321	**0.8888**
	300	0.7177	0.5719	0.6493	0.5925	0.7941	0.8387	0.8068	0.8792	0.8655	0.8777	**0.9037**
	400	0.7159	0.5705	0.6515	0.5916	0.7964	0.8141	0.8059	0.8736	0.8909	0.8656	**0.9048**
	500	0.7144	0.5679	0.6445	0.5868	0.7926	0.8124	0.8030	0.8762	0.8697	0.8657	**0.9011**
EURlex	200	0.6936	0.7308	0.7015	0.6255	0.8591	0.8216	0.8813	0.8130	0.8201	0.8561	**0.8884**
	400	0.6738	0.7294	0.6821	0.6228	0.8481	0.8611	0.8905	0.8257	0.8423	0.8633	**0.8928**
	600	0.6769	0.7321	0.6743	0.6256	0.8547	0.8735	0.8995	0.8218	0.8472	0.8414	**0.9001**
	800	0.6825	0.7349	0.6575	0.6283	0.8548	0.8775	0.9016	0.8289	0.8491	0.8637	**0.9034**
	1000	0.6733	0.7361	0.6708	0.6258	0.8406	0.8691	**0.9115**	0.8246	0.8431	0.8456	0.9106
Wiki10	1k	0.6293	0.6535	0.6403	0.5694	0.8092	0.8113	0.8345	0.6978	0.7830	0.8274	**0.8406**
	2k	0.5928	0.6244	0.6029	0.5510	0.7505	0.7557	0.7876	0.6545	0.6968	0.7937	**0.8013**
	3k	0.5703	0.6078	0.5815	0.5405	0.7567	0.7192	0.7417	0.6505	0.7197	0.7939	**0.8023**
	4k	0.5567	0.6008	0.5687	0.5364	0.7198	0.7105	0.7366	0.6556	0.7307	0.7944	**0.7996**
	5k	0.5502	0.5984	0.5651	0.5345	0.7087	0.7194	0.7350	0.6463	0.7228	0.7824	**0.7865**

Datasets	#label	Ranking loss ↓										
		BR	CC	RAKEL	ML-kNN	SLEEC	SML	SLL	BP-MLL	DNN-BCE	C2AE	DSLL
yeast	2	0.3064	0.2912	0.2519	0.2966	0.1527	0.1493	0.1668	0.1538	0.1494	0.1483	**0.1439**
	3	0.3059	0.3217	0.2863	0.3064	0.0927	0.0901	0.1674	0.0840	0.0818	0.0812	**0.0807**
	4	0.3409	0.3722	0.3535	0.3433	0.1137	0.1093	0.1891	0.1111	0.1077	0.1114	**0.1065**
	5	0.3694	0.4113	0.3787	0.3731	0.1245	0.1179	0.2189	0.1230	0.1163	0.1242	**0.1147**
	6	0.4080	0.4606	0.4049	0.4136	0.1558	0.1508	0.2382	0.1492	0.1536	0.1514	**0.1475**
MirFlickr	3	0.3032	0.5336	0.2842	0.3157	0.2393	0.2253	0.1179	0.1051	0.1025	0.0881	**0.0809**
	6	0.2868	0.5786	0.2735	0.3337	0.2488	0.2189	0.0586	0.0571	0.0565	0.0569	**0.0562**
	9	0.3366	0.6854	0.3194	0.3981	0.2186	0.2386	0.0672	0.0663	0.0609	0.0580	**0.0570**
	12	0.4011	0.7816	0.3784	0.4690	0.2692	0.2528	0.0877	0.0865	0.0857	0.0772	**0.0750**
	15	0.4209	0.7966	0.3873	0.4929	0.2857	0.2710	0.0899	0.0909	0.0837	0.0858	**0.0812**
Delicious	100	0.4135	0.6551	0.4550	0.6521	0.2333	0.2235	0.1438	0.0916	0.0890	0.0929	**0.0830**
	200	0.5115	0.8147	0.5603	0.7988	0.2712	0.2523	0.1771	0.1168	0.1217	0.1224	**0.1140**
	300	0.5248	0.8303	0.5710	0.7918	0.2759	0.2594	0.1758	0.1093	0.1071	0.1073	**0.1031**
	400	0.5311	0.8360	0.5711	0.7978	0.2661	0.2608	0.1778	0.1102	0.1079	**0.1050**	0.1082
	500	0.5383	0.8435	0.5841	0.8096	0.2784	0.2679	0.1822	0.1109	0.1154	0.1094	**0.1067**
EURlex	200	0.1473	0.1281	0.1403	0.1787	0.0654	0.0743	0.0198	0.0879	0.0337	0.0205	**0.0138**
	400	0.2493	0.2048	0.2386	0.2880	0.1101	0.0901	0.0299	0.1062	0.0380	0.0274	**0.0226**
	600	0.3468	0.2898	0.3494	0.4041	0.1552	0.1325	0.0404	0.1848	0.0562	0.0531	**0.0313**
	800	0.4072	0.3400	0.4384	0.4780	0.1871	0.1618	0.0483	0.1530	0.0609	0.0495	**0.0330**
	1000	0.4861	0.3960	0.4939	0.5619	0.2375	0.2085	0.0585	0.2326	0.0754	0.0721	**0.0370**
Wiki10	1k	0.4746	0.4375	0.4213	0.5520	0.0918	0.1082	0.0867	0.2259	0.1011	0.0483	**0.0421**
	2k	0.6205	0.5642	0.5705	0.6818	0.2385	0.2052	0.1259	0.3373	0.2225	0.0648	**0.0606**
	3k	0.7324	0.6577	0.6823	0.7766	0.3306	0.2573	0.1897	0.3727	0.1911	0.0715	**0.0642**
	4k	0.8167	0.7246	0.7569	0.8424	0.4037	0.3029	0.2123	0.3899	0.1831	0.0751	**0.0687**
	5k	0.8538	0.7495	0.7941	0.8698	0.4313	0.4368	0.2234	0.4525	0.1972	0.0833	**0.0816**

- SML (Liu & Shen, 2019) is the state-of-the-art embedding-based method, which further improves the convergence rate of SLEEC.
- SLL (You et al., 2016) is the original streaming label learning method, which leverages the linear relationship between past labels and new labels.
- BP-MLL (Zhang & Zhou, 2006) introduces a shallow neural network as the multi-label model trained to minimize a ranking loss. Here, we deepen the network to boost performance.
- DNN-BCE (Nam et al., 2014) makes use of ReLU, AdaGrad, Dropout, and other deep learning techniques to train a DNN-based model.
- C2AE (Yeh et al., 2017) is the state-of-the-art deep learning method for multi-label classification, which integrates the DNN architectures of canonical correlation analysis and autoencoder to learn the deep latent space.

The codes of baseline methods are provided by the authors or scikit-multilearn (Szymański & K., 2017).

Evaluation Metrics. The predictive accuracy of multi-label

Table 3. The proposed method component comparison, including streaming feature distillation (KD, Section 3.4), cross-entropy loss (CE), our proposed label correlation-aware loss (LC) for label mapping (Section 3.3), and transformation layer $\mathbf{S}_t$ after label mapping (Section 3.5). $\mathbf{L}_{soft}^{hard}$ represents using both hard and soft label vectors in label mapping, while $\mathbf{L}_{hard}$ and $\mathbf{L}_{soft}$ respectively represent the cases where only one of the label vectors is used. ↓(↑) means the smaller (larger) the value, the better the performance.

	Baseline	KD	CE	LC	LC-$\mathbf{S}_t$	$\mathbf{L}_{hard}$	$\mathbf{L}_{soft}$	$\mathbf{L}_{soft}^{hard}$	KD+LC-$\mathbf{S}_t$+$\mathbf{L}_{soft}^{hard}$
Average precision ↑	0.3824	0.4018	0.4113	0.4182	0.4287	0.4159	0.4427	0.4592	0.4925
Coverage ↓	0.2415	0.2349	0.2245	0.2221	0.2188	0.2352	0.2289	0.2237	0.2153

learning can be evaluated in different ways. Bipartition-based metrics assess a model's ability to predict sets of labels, whereas ranking-based metrics evaluate MLL methods, which instead produce rankings of labels. Therefore, our proposed method was evaluated in terms of both bipartition-based metrics, i.e., Hamming loss, macro-F1, micros-F1, and instance-F1, and ranking-based metrics, i.e., Precision@k, coverage, ranking loss, average precision (AP), macro-AUC, and micro-AUC (Wu & Zhou, 2017; Jain et al., 2016). Note that the coverage score is 1 greater than the one given in (Tsoumakas et al., 2010) to extends it to handle the cases where an instance has zero true labels; it is then normalized by the number of labels such that all the evaluation metrics vary between [0,1]. The details of these metrics can be referred in Appendix B.1.

Settings. To handle newly arrived labels, we investigated the models with different batch sizes following previous settings (You et al., 2016). The batch size refers to the number of new labels emerging in a dynamic environment. In particular, for each dataset, we selected randomly 50% (for Delicious 483) labels as past labels and constructed a classifier for the past labels. Then, the remaining labels were treated as new labels with different batch sizes. MLL methods can be extended to handle these new labels through independent modeling. More detailed settings are provided in the Appendix B.

5.2. Results

The results will show the performance evaluation in terms of ranking and classification. Then, we shall describe the ablation studies for different components in DSLL.

Ranking performance evaluation. First, we evaluate the results obtained by DSLL and other methods for dealing with new labels using the ranking-based metrics. Limited by space, we only show the average results of two measures: ranking loss and micro-AUC (Table 2). The results evaluated by Precision@k, AP, macro-AUC, and coverage can be found in the Appendix B.2 of the Supplementary Materials. The results show that DSLL largely outperforms the other approaches with respect to handling newly arrived labels. In this setting, traditional multi-label learning methods could not utilize the knowledge from past labels and classifiers. Although SLL takes advantage of the relationship between past labels and new labels, the linear representation limits its performance. Nevertheless, DSLL not only learns the correlations across new and past labels but also distills the knowledge from the past classifier.

Classification performance evaluation. Second, we compare the classification results as evaluated by bipartition-based metrics, where the prediction value of a label belongs to $\{0,1\}$ rather than a ranking value or a probability value. Figure 2 shows the micro-F1 results of learning new labels with different batch sizes. Other results of the bipartition-based metrics (Hamming loss, macro-F1, and instance-F1) can be found in the Supplementary Materials. We can see that DSLL consistently outperforms other multi-label learning methods with respect to all measures. Several machine learning approaches, e.g., BR and ML-kNN, perform poorly on datasets with a large number of labels because the inter-relationships between labels are not effectively employed. Note that SLEEC, SML, and SLL only focus on ranking results; if adopting a naive thresholding policy (e.g., a threshold of 0.5), they underperform compared to baselines, i.e., BR (Nam et al., 2017). Hence, classification performance evaluation excludes the results of SLEEC, SML, and SLL.

Ablation study. Finally, as shown in Table 3, we compare different strategies for learning knowledge from the past in DSLL to evaluate the effectiveness of various parts of our proposed framework. AP and coverage (Wu & Zhou, 2017) are used to measure the performance on the MirFlickr dataset, where 19 labels are used as past labels and 9 labels as new labels. We choose a fully-connected deep network as the baseline, and another network distills the knowledge at the intermediate layer of the past-label classifier to encourage the network to converge to a better value, denoted KD in Table 3. We compare the label correlation-aware loss in Equation (1), denoted LC, with the cross-entropy loss (CE) to achieve slightly better performance. Moreover, we find that the transformation layer $\mathbf{S}_t$ proposed in Section 3.5 is critical for effective relationship representations, which can be used with LC to obtain a 1.7% improvement. Note that a 1% improvement in multi-label classification is considered significant. In the process of streaming label mapping, we discover that the performance of using both hard and soft label vectors denoted $\mathbf{L}_{soft}^{hard}$, is better than using only one of the two, which is supported by the results shown in Table 3. This is because that the labels are unknown in the testing stage, the soft labels predicted by the past-label classifier are sent to the streaming label mapping as the input. Therefore, learning the combination of hard labels and soft labels can improve the robustness of the model in the testing. Finally, we find that our proposed overall method (KD+LC-S_t+L_{soft}^{hard}) significantly outperforms the baseline and others with part strategies.

6. Conclusion

In this paper, we propose a novel framework, Deep Streaming Label Learning (DSLL), to address the multi-label learning problem with new labels. DSLL distills the knowledge from past labels and their classifiers to effectively model new emerging labels without retraining the whole multi-label classifier, which has three components: a streaming label mapping, a streaming feature distillation, and a senior student. Our theoretical derivations prove that DSLL framework can provide a tight excess risk bound for new labels. Extensive empirical results show that DSLL significantly outperforms current state-of-the-art methods, and demonstrate the significance of leveraging knowledge learned from the past when modeling new labels.

Acknowledgements

This work was supported in part by Australian Research Council Projects FL-170100117, DP-180103424, IH-180100002, IC-190100031, and LE-200100049

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *ICML*, volume 97, pp. 242–252, 2019.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. Deep canonical correlation analysis. In *ICML*, volume 28, pp. 1247–1255, 2013.

Ba, J. and Caruana, R. Do deep nets really need to be deep? In *NeurIPS*, pp. 2654–2662. Curran Associates, Inc., 2014.

Bai, H., Wu, J., King, I., and Lyu, M. Few Shot Network Compression via Cross Distillation. In *AAAI*, 2020.

Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

Bhatia, K., Jain, H., Kar, P., Varma, M., and Jain, P. Sparse local embeddings for extreme multi-label classification. In *NeurIPS*, pp. 730–738. 2015.

Cao, Y. and Gu, Q. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *AAAI*, pp. 3349–3356, 2020.

Cesa-Bianchi, N., Re, M., and Valentini, G. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Machine Learning*, 88(1):209–241, 2012.

Chen, G., Choi, W., Yu, X., Han, T., and Chandraker, M. Learning efficient object detection models with knowledge distillation. In *NeurIPS*, pp. 742–751, 2017.

Chen, S.-F., Chen, Y.-C., Yeh, C.-K., and Wang, Y.-C. Order-free rnn with visual attention for multi-label classification, 2018.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1675–1685, 2019.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., and Choi, J. Y. A comprehensive overhaul of feature distillation. In *ICCV*, October 2019.

Hinton, G., Vinyals, O., and Dean, J. Distilling the Knowledge in a Neural Network. In *NeurIPS workshop*, 2014.

Jain, H., Prabhu, Y., and Varma, M. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *KDD*, pp. 935–944. ACM, 2016.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Regularization techniques for learning with matrices. *J. Mach. Learn. Res.*, 13:1865–1890, 2012.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Rademacher Averages*, pp. 89–121. Springer Berlin Heidelberg, 1991.

Liu, W. and Shen, X. Sparse extreme multi-label learning with oracle property. In *ICML*, volume 97, pp. 4032–4041, 2019.

Liu, Y., Sheng, L., Shao, J., Yan, J., Xiang, S., and Pan, C. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *ACM International Conference on Multimedia*, pp. 700–708, 2018.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory*, pp. 3–17, 2016.

Menon, A. K., Rawat, A. S., Reddi, S., and Kumar, S. Multi-label reductions: what is my loss optimising? In *NeurIPS*, pp. 10599–10610. 2019.

Nam, J., Kim, J., Loza Mencía, E., Gurevych, I., and Fürnkranz, J. Large-scale multi-label text classification — revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases*, pp. 437–452, 2014.

Nam, J., Loza Mencía, E., Kim, H. J., and Fürnkranz, J. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *NeurIPS*, pp. 5413–5423. 2017.

Nam, J., Kim, Y.-B., Mencia, E. L., Park, S., Sarikaya, R., and Fürnkranz, J. Learning context-dependent label permutations for multi-label classification. In *ICML*, volume 97, pp. 4733–4742, 2019.

Prabhu, Y. and Varma, M. Fastxml: a fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD*, 2014.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, Jun 2011.

Romero, A., Kahou, S. E., Montréal, P., Bengio, Y., Montréal, U. D., Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

Sammut, C. and Webb, G. I. (eds.). *Encyclopedia of Machine Learning: McDiarmid's Inequality*, pp. 651–652. 2010.

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015.

Szymański, P. and K., T. A scikit-based Python environment for performing multi-label classification. *arXiv preprint arXiv:1702.01460*, 2017.

Tsoumakas, G., Katakis, I., and Vlahavas, I. *Mining Multi-label Data*, pp. 667–685. Springer US, 2010.

Tsoumakas, G., Katakis, I., and Vlahavas, I. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 2s3(7):1079–1089, 2011.

Wu, X.-Z. and Zhou, Z.-H. A unified view of multi-label performance measures. In *ICML*, volume 70, pp. 3780–3788, 2017.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, June 2010.

Yeh, C.-K., Wu, W.-C., Ko, W.-J., and Wang, Y.-C. F. Learning deep latent spaces for multi-label classification. In *AAAI*, pp. 2838–2844, 2017.

You, S., Xu, C., Wang, Y., Xu, C., and Tao, D. Streaming Label Learning for Modeling Labels on the Fly. *arXiv preprint arXiv:1604.05449*, 2016.

Yu, H.-F., Jain, P., Kar, P., and Dhillon, I. S. Large-scale multi-label learning with missing labels. In *ICML*, pp. 593–601, 2014.

Zhang, M. and Zhou, Z. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Zhang, M.-L. and Zhou, Z.-H. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

Zhang, M.-L. and Zhou, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recogn.*, 40(7): 2038–2048, 2007.

Zhou, C., Gu, J., and Neubig, G. Understanding knowledge distillation in non-autoregressive machine translation. In *ICLR*, 2020.

Zhu, Y., Ting, K. M., and Zhou, Z. Multi-label learning with emerging new labels. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1901–1914, Oct 2018. ISSN 2326-3865. doi: 10.1109/TKDE.2018.2810872.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109:467 – 492, 2019.