Dual Perspective of Label-Specific Feature Learning for Multi-Label Classification

Jun-Yi Hang [1 2] **Min-Ling Zhang** [1 2]

Abstract

Label-specific features serve as an effective strategy to facilitate multi-label classification, which account for the distinct discriminative properties of each class label via tailoring its own features. Existing approaches implement this strategy in a quite straightforward way, i.e. finding the most pertinent and discriminative features for each class label and directly inducing classifiers on constructed label-specific features. In this paper, we propose a dual perspective for label-specific feature learning, where label-specific discriminative properties are considered by identifying each label's own non-informative features and making the discrimination process immutable to variations of these features. To instantiate it, we present a perturbation-based approach DELA to provide classifiers with label-specific immutability on simultaneously identified non-informative features, which is optimized towards a probabilistically-relaxed expected risk minimization problem. Comprehensive experiments on 10 benchmark data sets show that our approach outperforms the state-of-the-art counterparts.

1. Introduction

Multi-label classification allows to learn from instances associated with multiple labels simultaneously (Zhang & Zhou, 2014; Liu et al., 2021). Nowadays, researches on multi-label classification have been greatly driven by real-world applications, where multi-semantic objects widely exist, such as image annotation (You et al., 2020), text categorization (Xun et al., 2020), and bioinformatics analysis (Chen et al., 2017), etc.

The most straightforward strategy for dealing with multi-label data is to exploit the identical representation of an instance in inducing classification models. However, this strategy might be suboptimal as it fails to consider that each class label may possess its own discriminative properties. With the ability to model distinct characteristics of each class label, label-specific features have become a promising strategy to facilitate the discrimination of each class label by tailoring its own features (Zhang & Wu, 2015; Huang et al., 2016b; Zhang et al., 2018; Jia et al., 2020; Yu & Zhang, 2021).

As a seminal work, LIFT (Zhang & Wu, 2015) firstly performs clustering analysis on positive and negative instances of each class label, and then heuristically constructs label-specific features via prototype-based feature transformation. While LLSF (Huang et al., 2015) employs feature selection to obtain the most pertinent feature subset for each class label under a lasso-based framework. Recent works (Hang & Zhang, 2021; Hang et al., 2022) resort to the powerful representation learning capability of deep neural networks to learn label-specific features in an end-to-end manner. It is worth noting that existing approaches focus on finding the most pertinent and discriminative features for each class label and directly inducing classifiers on constructed label-specific features.

In this paper, we attack the problem of label-specific feature learning from a dual perspective. Instead of finding the most pertinent and discriminative features for each class label as existing approaches do, we attempt to identify each label's own non-informative features and endow classifiers with immutability on these identified features. For example, to discriminate plane and non-plane images, existing approaches induce classifier on the most pertinent features, e.g. *shape*-based features. Instead, we aim to make the discrimination process immutable to variations of non-informative features, e.g. *color*-based features. We hypothesize that if non-informative features specific to each class label could be identified and their influence on the discrimination process could be eliminated, a more effective approach to learn from multi-label data could be achieved.

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China [2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China. Correspondence to: Min-Ling Zhang <zhangml@seu.edu.cn>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

As a first attempt towards this dual strategy, a novel approach named DELA, i.e. **D**ual p**E**rspective of **L**abel-specific fe**A**ture learning for multi-label classification, is presented. Under a stochastic feature perturbation framework, DELA simultaneously identifies non-informative features and induces classifiers which are immutable to these identified features. Specifically, by selectively injecting random noise to label-specific non-informative features and inducing classifiers on these perturbed features, DELA succeeds to remove classifiers' dependence on identified non-informative features. With the basic assumption that non-informative features should have no negative influence on the discrimination process when perturbed by noise, identification of non-informative features for each class label is casted as an expected risk minimization problem, which is further relaxed probabilistically to support end-to-end gradient-based optimization. We further justify DELA from the perspective of information theory and demonstrate that DELA actually optimizes an upper bound of the information bottleneck (Tishby et al., 1999). Comprehensive experiments on 10 benchmark data sets show that DELA performs better than well-established multi-label classification algorithms.

The rest of this paper is organized as follows. Section 2 briefly reviews related works. Section 3 presents details of the proposed DELA approach. Section 4 reports experimental results over a wide range of multi-label data sets. Section 5 concludes this paper.

2. Related Works

Multi-Label Classification. In the last decade, numerous approaches have been proposed to deal with multi-label classification problem (Zhang & Zhou, 2014; Liu et al., 2021). As a feasible strategy to facilitate the learning process, modelling label correlations is one of the primary focuses in recent studies. Generally speaking, these approaches can be roughly grouped into three categories, which differ in the order of label correlations considered, namely first-order approaches (Boutell et al., 2004; Zhang & Zhou, 2007), second-order approaches (Zhu et al., 2018; Sun & Zhang, 2021) and high-order approaches (Wehrmann et al., 2018; Xu & Guo, 2021). Recent works resort to deep models, such as recurrent neural networks (Wang et al., 2016; Yazici et al., 2020) and graph neural networks (Chen et al., 2019; 2022), to jointly consider the label correlation exploitation and classification model induction. Some embedding approaches (Yeh et al., 2017; Bai et al., 2020; Dahiya et al., 2021) implicitly employ label correlations via embedding and aligning features and labels in a deep latent space.

Complementary to label correlation exploitation, label-specific features have been proven to be another effective strategy to improve multi-label classification, which tackle the problem via manipulating the input space instead of the label space. Existing approaches construct the label-specific features mainly in two manners, i.e. prototype-based label-specific feature transformation and label-specific feature selection.

For the prototype-based label-specific feature transformation approaches, label-specific features are generated by treating the prototypes of each class label as the transformation bases. Under a three-stage framework, LIFT (Zhang & Wu, 2015) constructs label-specific features via querying the distances between the original instance and the cluster centers for each class label. Follow-up works enhance the three-stage framework by customized strategies, such as stabilizing the clustering process with clustering ensemble (Zhan & Zhang, 2017; Zhang & Li, 2021) or spectral clustering (Zhang et al., 2015), augmenting metric-based label-specific features with local neighbor information (Weng et al., 2018) or global topological information (Guo et al., 2019), unifying the independent three-stage framework into an end-to-end counterpart (Hang et al., 2022).

Alternatively, label-specific features can also be constructed by retaining a feature subset as the most pertinent features for each class label. LLSF (Huang et al., 2015; 2016b) presents a lasso-based framework for label-specific feature selection, where the selection process is regularized with pairwise label correlations. Subsequent studies extend this framework via imposing non-sparse constraints (Weng et al., 2020), incorporating discriminant-related regularization (Huang et al., 2018), or performing selection in a projected feature space (Yu & Zhang, 2021). Recently, CLIF (Hang & Zhang, 2021) further advances the idea to the deep learning scenario with an attractive collaborative learning strategy.

In this paper, we make a first attempt to consider label-specific discriminative properties via endowing classifiers with immutability on non-informative features, which is an unexplored direction for label-specific feature learning.

Nuisance Factor Removal. A similar concept, i.e. removal of nuisance factors, has a long history in computer vision. Early attempts include designing scale-invariant (Lowe, 1999) or rotation-invariant features (Greenspan et al., 1994), while recent approaches resort to techniques, such as data augmentation (Devries & Taylor, 2017; Cubuk et al., 2019) and representation disentangling (Tran et al., 2017; Moyer et al., 2018; Lee et al., 2021), for removal of specified factors. DELA shares the idea and generalizes it by learning to identify non-information features instead of specifying them beforehand. In domain generalization (Ahuja et al., 2020; 2021), domain-specific features are regarded as nuisance factors, which should be removed to learn domain-invariant ones for achieving good performance on unseen domains. DELA also shares similar idea but attempts to encourage the emergence of label(domain)-specific properties

rather than suppressing them.

Feature Perturbation by Noise Injection. Injecting noise to perturb features has been widely applied in machine learning community. Dropout (Srivastava et al., 2014) and its extensions (Blum et al., 2015; Huang et al., 2016a; Achille & Soatto, 2018) perturb featuers by randomly dropping out neurons or layers during training to encourage redundancy in learned representation, which deviates from our goal to remove redundancy. Adversarial attack (Goodfellow et al., 2015; Madry et al., 2018; Duan et al., 2021) aims to find the most vulnerable directions to perturb the instance so that the training loss is maximized, while we attempt to identify and perturb non-informative features for expected risk minimization. Besides, feature perturbation can also be exploited to perform post-hoc explanation of prediction (Ribeiro et al., 2016; Fong & Vedaldi, 2017), while our approach perturbs features during the learning process for generalization.

3. The DELA Approach

3.1. Preliminaries

Let $\mathcal{X} = \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{l_1, l_2, \ldots, l_t\}$ denote the label space with t class labels. A multi-label example is denoted as $(\mathbf{x}, Y)$, where $\mathbf{x} \in \mathcal{X}$ is its feature vector and $Y \subseteq \mathcal{Y}$ is its set of relevant labels. Here, a t-dimensional indicator vector $\mathbf{y} = [y_1, y_2, \ldots, y_t] \in \{0,1\}^t$ is utilized to denote Y, where $y_k = 1$ indicates $l_k \in Y$ and $y_k = 0$ otherwise. Formally, multi-label classification aims to derive a multi-label prediction function $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from a multi-label data set $\mathcal{D} = \{(\mathbf{x}_i, Y_i) | i \in [m]\}$, where $[m]$ denotes the set $\{1, 2, \ldots, m\}$. Given an unseen instance $\mathbf{u} \in \mathcal{X}$, its associated label set is predicted as $h(\mathbf{u}) \subseteq \mathcal{Y}$.

3.2. Overview

DELA firstly computes a d_z-dimensional representation $\mathbf{z} \in \mathbb{R}^{d_z}$ through an embedding function $e_{\boldsymbol{\phi}} : \mathbb{R}^d \to \mathbb{R}^{d_z}$ parametrized by $\boldsymbol{\phi}$, which is shared among all the class labels. Then, a selective feature perturber injects additive random noise into representation $\mathbf{z}$ for perturbing non-informative features specific to each class label. Finally, classification is performed on the noise perturbed representations.

Learning proceeds by simultaneously identifying non-informative features and making the discrimination process immutable to identified non-informative features, with the expected risk minimizing problem as following

$$\min_{\boldsymbol{\phi}, \boldsymbol{\Pi}, \boldsymbol{\Theta}} \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[\sum_{k=1}^{t} \mathcal{L}(f_k(g_k(e_{\boldsymbol{\phi}}(\mathbf{x}); \boldsymbol{\pi}_k); \boldsymbol{\theta}_k), y_k)], \quad (1)$$

where $\boldsymbol{\Pi} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_t\}, \boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t\}$ are the sets to parametrize the selective feature perturber $g_k : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ and the classifier $f_k : \mathbb{R}^{d_z} \to \mathbb{R}$ for each class label respectively. We will describe the ingredients of DELA in detail in the next subsection.

3.3. Selective Feature Perturber

The goal of the selective feature perturber is to identify non-informative features in representation $\mathbf{z}$ and perturb them via injecting random noise in a label-wise manner, so that immutability of classification on the non-informative features can be gradually enhanced.

To instantiate the selective feature perturber, we formalize the perturbation process as

$$g_k(\mathbf{z}; \boldsymbol{\pi}_k) = \mathbf{z} + \mathbf{i}_{S_k} \odot \boldsymbol{\epsilon}, \quad with\ \boldsymbol{\epsilon} \sim p_{\boldsymbol{\vartheta}}(\boldsymbol{\epsilon}), \quad (2)$$

where $S_k \subseteq [d_z]$ denotes a subset of identified non-informative features for label l_k which is determined by parameter $\boldsymbol{\pi}_k$ and $\mathbf{i}_{S_k} \in \{0,1\}^{d_z}$ is the indicator vector of set S_k. $\boldsymbol{\epsilon}$ is a random noise variable shared among all the class labels, which is treated as an instance-dependent Gaussian one, i.e. $p_{\boldsymbol{\vartheta}}(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \sigma^2_{\boldsymbol{\vartheta}}(\mathbf{x})\mathbf{I})$[1]. With the Hadamard product $\odot$, additive random noise is selectively injected into the identified non-informative features.

Substituting Eq. (2) into Eq. (1), the expected risk minimizing problem becomes

$$\min_{\boldsymbol{\phi}, S, \boldsymbol{\vartheta}, \boldsymbol{\Theta}} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p_{\boldsymbol{\vartheta}}(\boldsymbol{\epsilon})}[\sum_{k=1}^{t} \mathcal{L}(f_k(e_{\boldsymbol{\phi}}(\mathbf{x}) + \mathbf{i}_{S_k} \odot \boldsymbol{\epsilon}; \boldsymbol{\theta}_k), y_k)], \quad (3)$$

where $S = \{S_1, \ldots, S_t\}$. The above problem is hard to solve, since the optimization over the discrete subsets of non-informative features $\{S_1, \ldots, S_t\}$ is intractable, whose choices grow exponentially in d_z. Furthermore, constraint on the level of noise is necessary to prevent collapse, i.e. insufficient perturbation, so as to endow classifiers with immutability on non-informative features. We will describe our considerations towards these two problems in detail.

3.3.1. DIFFERENTIABLE SUBSET SELECTION

To make Eq. (3) tractable, we introduce Bernoulli gates to substitute the indicator vector $\mathbf{i}_{S_k} \in \{0,1\}^{d_z}$. These Bernoulli gates can be represented by a random vector $\mathbf{b}_k \in \{0,1\}^{d_z}$, whose entries are independent and satisfy $P[b_{ki} = 1] = p_{ki}$, $i \in [d_z]$. Then, the expected risk minimizing problem can be rewritten as

$$\min_{\boldsymbol{\phi}, P, \boldsymbol{\vartheta}, \boldsymbol{\Theta}} \mathbb{E}_{p(\mathbf{x},\mathbf{y})p_{\boldsymbol{\vartheta}}(\boldsymbol{\epsilon})}[\sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{b}_k)}[\mathcal{L}(f_k(e_{\boldsymbol{\phi}}(\mathbf{x}) + \mathbf{b}_k \odot \boldsymbol{\epsilon}; \boldsymbol{\theta}_k), y_k)]]. \quad (4)$$

[1] $\boldsymbol{\vartheta}$ parametrizes the standard deviation function, which is shared among all the class labels. We can now denote $\boldsymbol{\pi}_k = [S_k, \boldsymbol{\vartheta}]$.

By introducing Bernoulli gates, the original intractable subset selection problem is converted to an optimization problem in terms of Bernoulli distribution parameters $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_t\}$. Nonetheless, the discrete property of the sampling from Bernoulli distribution prevents gradients from flowing through the discrete random nodes $\mathbf{b}_k$, thus making the problem unable to be optimized end-to-end via gradient descent.

A feasible way to circumvent this is to exploit *Gumbel-Softmax trick* (Jang et al., 2017; Maddison et al., 2017) to smooth the sampling process, where a Bernoulli random variable $b \sim Bern(p)$ is relaxed by its continuous alternative, i.e. a binary Concrete variable $c \sim BinConcrete(p, \tau)$, which can be reparameterized as

$$c = \frac{1}{1 + \exp[-(\log \alpha + l)/\tau]}, \tag{5}$$

where $\alpha = \frac{p}{1-p}$, l is a sampling from Logistic distribution, and $\tau > 0$ is a temperature parameter. In the limit $\tau \to 0$, a binary Concrete variable smoothly converges to its Bernoulli counterpart.

We relax the Bernoulli gates $\mathbf{b}_k$ to the above binary Concrete gates $\mathbf{c}_k$ so that the gradients w.r.t. the distribution parameters $\{\mathbf{p}_1, \ldots, \mathbf{p}_t\}$ are well-defined by the chain rule. In the forward pass we discretize the continuous samplings from binary Concrete gates by rounding, i.e. $\mathbf{d}_k = round(\mathbf{c}_k)$, and in the backward pass we use a straight-through gradient estimator to approximate $\nabla_{\mathbf{p}_k}\mathbf{c}_k \approx \nabla_{\mathbf{p}_k}\mathbf{d}_k$.

3.3.2. Constraint on Noise Distribution

Let $\mathbf{z}_k = e_{\boldsymbol{\phi}}(\mathbf{x}) + round(\mathbf{c}_k) \odot \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim p_{\boldsymbol{\vartheta}}(\boldsymbol{\epsilon})\, and\, \mathbf{c}_k \sim p(\mathbf{c}_k)$, we can regard the above equation as the reparameterization form of the random variable $\mathbf{z}_k$, i.e. label l_k's perturbed stochastic representation, which follows an implicit distribution $\mathbb{E}_{p(\mathbf{c}_k)}[p_{\boldsymbol{\phi},\boldsymbol{\vartheta}}(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)]$.

From a probabilistic perspective, we propose to constrain the noise distribution $p_{\boldsymbol{\vartheta}}(\boldsymbol{\epsilon})$ by penalizing the expected discrepancy between $p_{\boldsymbol{\phi},\boldsymbol{\vartheta}}(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)$ and an instance-agnostic prior distribution $q(\mathbf{z}_k)$, which can be formalized as

$$\mathbb{E}_{p(\mathbf{c}_k)}[KL(p_{\boldsymbol{\phi},\boldsymbol{\vartheta}}(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)||q(\mathbf{z}_k))], \tag{6}$$

where $KL(\cdot||\cdot)$ denotes the KL-divergence. The conditional distribution $p_{\boldsymbol{\phi},\boldsymbol{\vartheta}}(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)$ describes the extent to which the original representation is perturbed by noise conditioning on currently identified non-informative features, and the prior distribution $q(\mathbf{z}_k)$ reflects the target level of noise which is sufficient to remove classifiers' dependence on non-informative features. We set $q(\mathbf{z}_k)$ as a standard Gaussian in this paper.

Substituting the reparameterization form of the perturbed stochastic representation and the constraint on noise distribution for each class label, the overall objective is defined as

$$\min_{\boldsymbol{\phi},P,\boldsymbol{\vartheta},\boldsymbol{\Theta}} \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{c}_k)} \Big[\mathbb{E}_{p(\mathbf{z}_k|\mathbf{x},\mathbf{c}_k)}[\mathcal{L}(f_k(\mathbf{z}_k; \boldsymbol{\theta}_k), y_k)] + \beta \cdot KL(p(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)||q(\mathbf{z}_k)) \Big], \tag{7}$$

where β is a trade-off parameter and $p_{\boldsymbol{\phi},\boldsymbol{\vartheta}}(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)$ is abbreviated to $p(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)$ for clarity.

3.4. Information Theory Explanation

We further provide an information theoretic insight of DELA and demonstrate that DELA actually optimizes towards an upper bound of the information bottleneck when the risk function $\mathcal{L}(\cdot, \cdot)$ is instantiated by cross entropy loss.

The information bottleneck defines a optimal information transportation process $\mathbf{x} \to \mathbf{h} \to y$ by

$$\min -I(\mathbf{h}; y) + \beta \cdot I(\mathbf{h}; \mathbf{x}), \tag{8}$$

where $\mathbf{h}$ is the internal representation of a model (e.g. neural networks) to predict the target variable y based on the input variable $\mathbf{x}$, and $I(\cdot, \cdot)$ denotes the mutual information operator. The goal of the information bottleneck is to learn an optimal representation $\mathbf{h}$ which is maximally expressive about the target variable y and maximally compressive about the input variable $\mathbf{x}$. In other words, any information irrelevant to target prediction will be dropped during the information transportation process $\mathbf{x} \to \mathbf{h}$. In DELA, we perform label-specific feature learning by making the discrimination process immutable to non-informative features with explicit noise injection during the above information transportation process, which shares motivation with the information bottleneck.

To show the connection theoretically, we firstly derive an upper bound for the information bottleneck.

Theorem 3.1. *For any random variant $\mathbf{c} \sim p(\mathbf{c})$, the information bottleneck can be upper bounded as follows*

$$\begin{aligned} & -I(\mathbf{h}; y) + \beta \cdot I(\mathbf{h}; \mathbf{x}) \\ \leq\ & \mathbb{E}_{p(\mathbf{x},y)} \mathbb{E}_{p(\mathbf{c})} \Big[\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{c})}[-\log q(y|\mathbf{h})] \\ & + \beta \cdot KL(p(\mathbf{h}|\mathbf{x}, \mathbf{c})||q(\mathbf{h})) \Big]. \end{aligned} \tag{9}$$

Proof of Theorem 3.1 can be found in the appendix A. Extending it into multi-label classification scenario, the upper bound of label-wise information bottlenecks can be formal-

ized as

$$\begin{aligned}
&\sum_{k=1}^{t} -I(\mathbf{z}_k; y_k) + \beta \cdot I(\mathbf{z}_k; \mathbf{x}) \\
&\leq \sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{x}, y_k)} \mathbb{E}_{p(\mathbf{c}_k)} \Big[\mathbb{E}_{p(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)} [-\log q(y_k|\mathbf{z}_k)] \\
&\qquad + \beta \cdot KL(p(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k) || q(\mathbf{z}_k)) \Big] \\
&= \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \sum_{k=1}^{t} \mathbb{E}_{p(\mathbf{c}_k)} \Big[\mathbb{E}_{p(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)} [-\log q(y_k|\mathbf{z}_k)] \\
&\qquad + \beta \cdot KL(p(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k) || q(\mathbf{z}_k)) \Big],
\end{aligned} \tag{10}$$

where the last equation corresponds to the overall objective of DELA when we instantiate $\mathcal{L}(\cdot, \cdot)$ in Eq. (7) by cross entropy loss[2], which provides a theoretical justification for DELA.

4. Experiments

4.1. Experimental Setup

4.1.1. DATA SETS

For comprehensive performance evaluation, ten benchmark multi-label data sets with diversified multi-label properties are employed in this paper. Table 1 summarizes detailed properties of each data set $\mathcal{D}$, including the number of examples ($|\mathcal{D}|$), number of features ($dim(\mathcal{D})$), number of class labels ($L(\mathcal{D})$), feature type ($F(\mathcal{D})$), label cardinality ($LCard(\mathcal{D})$, i.e. average number of labels per instance). Following (Zhang & Wu, 2015), we perform dimensionality reduction for rcv-s1 and tmc2007 by retaining the top 2% features with highest document frequency. For iaprtc12, espgame and mirflickr, the local descriptor *DenseSift* is used.

4.1.2. EVALUATION METRICS

Six widely-used evaluation metrics for multi-label classification are employed to evaluate the performance of each approach, including *Average precision*, *Macro-averaging AUC*, *Hamming loss*, *One-error*, *Coverage* and *Ranking loss*. Detailed definitions on these metrics can be found in (Zhang & Zhou, 2014).

4.1.3. IMPLEMENTATION DETAILS

We implement DELA with the same architecture of encoder and decoder as MPVAE (Bai et al., 2020). Specifically, the embedding function $e_{\boldsymbol{\phi}}$ is instantiated by a fully-connected neural network with ReLU activations, where the hidden dimensionalities are set to $[256; 512; 256]$. The standard deviation function $\sigma_{\boldsymbol{\vartheta}}$ to parametrize the noise distribution is a four-layer fully-connected neural network, which shares the first three layers with $e_{\boldsymbol{\phi}}$. Classifiers f_k, $k \in [t]$ are implemented as three-layer fully-connected neural networks, where the hidden dimensionalities are set to $[256; 512]$ and the first two layers are shared among all the class labels. To parametrize the binary Concrete gates, we employ a two-layer fully-connected neural network to produce the distribution parameters $\{\mathbf{p}_1, \ldots, \mathbf{p}_t\}$ and use $\tau = 2/3$ as suggested by (Maddison et al., 2017).

In all experiments, We consider cross entropy loss to instantiate the risk function $\mathcal{L}(\cdot, \cdot)$, as it allows to build connection between DELA and the information bottleneck. To compute the overall objective in Eq. (7), we conduct Monte Carlo sampling to estimate the expectations in terms of $p(\mathbf{c}_k), p(\mathbf{z}_k|\mathbf{x}, \mathbf{c}_k)$ with sampling number $L = 1$ and analytically calculate the KL-divergence term between two Gaussian distributions. For network optimization, Adam with a batch size of 128, weight decay of 10^{-4}, momentums of 0.999 and 0.9 is employed.

Table 1. Characteristics of the experimental data sets.

Dataset	$\|\mathcal{D}\|$	$dim(\mathcal{D})$	$L(\mathcal{D})$	$F(\mathcal{D})$	$LCard(\mathcal{D})$	Domain
corel5k	5000	499	374	Nominal	3.522	Images[1]
rcv1-s1	6000	944	101	Numeric	2.880	Text[1]
Corel16k-s1	13766	500	153	Nominal	2.859	Images[1]
delicious	16105	500	983	Nominal	19.020	Text[1]
iaprtc12	19627	1000	291	Numeric	5.719	Images[2]
espgame	20770	1000	268	Numeric	4.686	Images[2]
mirflickr	25000	1000	38	Numeric	4.716	Images[2]
tmc2007	28596	981	22	Nominal	2.158	Text[1]
mediamill	43907	120	101	Numeric	4.376	Video[1]
bookmarks	87856	2150	208	Nominal	2.028	Text[1]

[1] http://mulan.sourceforge.net/datasets.html
[2] http://lear.inrialpes.fr/people/guillaumin/data.php

4.2. Comparative Studies

DELA[3] is compared against six well-established multi-label classification approaches with parameter configurations suggested in respective literatures:

- LIFT (Zhang & Wu, 2015): A prototype-based label-specific feature transformation approach under independent three-stage framework. [$r = 0.1$]
- LLSF (Huang et al., 2016b): LLSF performs label-specific feature selection in a lasso-based framework with feature-sharing between closely-related labels. [grid search for $\alpha, \beta \in \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ and $\gamma = 0.01$]
- C2AE (Yeh et al., 2017): A deep label embedding approach, which jointly embeds features and labels via integrating deep canonical correlation analysis and

[2] Bernoulli distribution $q(y_k|\mathbf{z}_k)$ is parametrized by prediction $f_k(\mathbf{z}_k; \boldsymbol{\theta}_k)$.

[3] Code package is publicly available at: http://palm.seu.edu.cn/zhangml/files/DELA.rar.

Table 2. Predictive performance of each comparing approach (mean±std. deviation) in terms of *Average precision*, *Macro-averaging AUC* and *Hamming loss*. ↑ (↓) indicates the larger (smaller) the value, the better the performance. Best results are highlighted in **boldface**

Data Sets	*Average precision* ↑						
	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA	DELA
corel5k	0.2911±0.0137	0.2996±0.0133	0.2915±0.0100	0.3285±0.0130	0.3147±0.0106	0.3240±0.0127	**0.3382±0.0103**
rcv1-s1	0.5921±0.0145	0.6129±0.0116	0.6147±0.0100	0.6332±0.0136	0.6246±0.0094	0.6340±0.0151	**0.6391±0.0153**
Corel16k-s1	0.3168±0.0059	0.3428±0.0053	0.3297±0.0042	0.3646±0.0063	0.3516±0.0070	0.3654±0.0087	**0.3675±0.0062**
delicious	0.3833±0.0061	0.3587±0.0079	0.3648±0.0075	0.4042±0.0065	0.3832±0.0069	0.4046±0.0072	**0.4082±0.0053**
iaprtc12	0.3853±0.0058	0.3597±0.0041	0.3900±0.0061	0.4326±0.0047	0.4366±0.0049	0.4357±0.0068	**0.4490±0.0050**
espgame	0.2808±0.0072	0.2770±0.0046	0.2741±0.0072	0.3074±0.0059	0.3095±0.0072	0.3138±0.0045	**0.3162±0.0059**
mirflickr	0.6516±0.0041	0.6477±0.0039	0.6627±0.0064	0.6849±0.0053	0.6857±0.0025	0.6900±0.0040	**0.6960±0.0043**
tmc2007	0.8207±0.0048	0.8130±0.0050	0.7977±0.0048	0.8297±0.0032	0.8189±0.0024	0.8286±0.0057	**0.8363±0.0037**
mediamill	0.7417±0.0058	0.7275±0.0051	0.7266±0.0051	0.7669±0.0062	0.7650±0.0061	0.7833±0.0055	**0.7883±0.0049**
bookmarks	0.5119±0.0044	0.4920±0.0035	0.4707±0.0046	0.5104±0.0050	0.4928±0.0036	0.5022±0.0036	**0.5191±0.0036**

Data Sets	*Macro-averaging AUC* ↑						
	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA	DELA
corel5k	0.7119±0.0112	0.6568±0.0141	0.7021±0.0122	0.7522±0.0150	0.7306±0.0154	0.7509±0.0143	**0.7546±0.0185**
rcv1-s1	0.9241±0.0103	0.9062±0.0100	0.9131±0.0084	0.9368±0.0078	0.9320±0.0044	0.9362±0.0067	**0.9374±0.0079**
Corel16k-s1	0.6937±0.0097	0.6614±0.0075	0.7212±0.0131	0.7867±0.0130	0.7657±0.0105	**0.7885±0.0120**	0.7872±0.0098
delicious	0.7919±0.0044	0.7509±0.0047	0.7830±0.0052	0.8272±0.0039	0.8107±0.0043	0.8255±0.0024	**0.8305±0.0030**
iaprtc12	0.8216±0.0034	0.8114±0.0041	0.8290±0.0039	0.8735±0.0025	0.8788±0.0027	0.8746±0.0027	**0.8810±0.0033**
espgame	0.7554±0.0059	0.7515±0.0057	0.7360±0.0046	0.7959±0.0055	0.7914±0.0064	0.7997±0.0065	**0.8015±0.0047**
mirflickr	0.8091±0.0077	0.8196±0.0042	0.8213±0.0046	0.8461±0.0040	0.8436±0.0045	0.8488±0.0043	**0.8538±0.0045**
tmc2007	0.9229±0.0035	0.9225±0.0040	0.8993±0.0052	0.9307±0.0038	0.9274±0.0048	0.9299±0.0046	**0.9356±0.0037**
mediamill	0.8302±0.0080	0.7874±0.0110	0.8172±0.0069	0.8627±0.0083	0.8703±0.0086	0.8723±0.0089	**0.8836±0.0067**
bookmarks	0.8984±0.0030	0.8857±0.0037	0.8403±0.0040	0.9106±0.0015	0.9024±0.0028	0.9086±0.0017	**0.9117±0.0022**

Data Sets	*Hamming loss* ↓						
	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA	DELA
corel5k	0.0094±0.0001	0.0094±0.0001	0.0124±0.0003	0.0094±0.0001	0.0094±0.0001	0.0094±0.0003	**0.0093±0.0002**
rcv1-s1	**0.0259±0.0009**	0.0263±0.0011	0.0408±0.0016	0.0270±0.0011	0.0267±0.0011	0.0269±0.0011	0.0266±0.0009
Corel16k-s1	0.0187±0.0002	**0.0186±0.0002**	0.0233±0.0005	0.0188±0.0003	0.0188±0.0003	0.0187±0.0002	**0.0186±0.0002**
delicious	0.0180±0.0001	0.0184±0.0002	0.0248±0.0006	**0.0177±0.0001**	0.0179±0.0001	0.0179±0.0002	0.0178±0.0001
iaprtc12	0.0189±0.0001	0.0190±0.0002	0.0420±0.0010	0.0184±0.0002	0.0181±0.0002	0.0184±0.0002	**0.0180±0.0002**
espgame	0.0179±0.0003	0.0173±0.0002	0.0572±0.0015	0.0174±0.0002	0.0176±0.0002	0.0173±0.0003	**0.0172±0.0002**
mirflickr	0.1019±0.0009	0.1005±0.0009	0.1259±0.0037	0.0969±0.0011	0.0965±0.0008	0.0959±0.0015	**0.0945±0.0012**
tmc2007	0.0603±0.0007	0.0607±0.0013	0.0632±0.0014	0.0586±0.0006	0.0587±0.0010	0.0590±0.0012	**0.0572±0.0009**
mediamill	0.0291±0.0003	0.0304±0.0002	0.0348±0.0004	0.0281±0.0004	0.0279±0.0004	0.0271±0.0005	**0.0260±0.0004**
bookmarks	0.0086±0.0001	0.0087±0.0001	0.0106±0.0001	0.0087±0.0001	**0.0085±0.0001**	0.0087±0.0001	0.0086±0.0001

autoencoder. [search for $\alpha \in \{0.1, 1, 2, 5, 10\}$]

- MPVAE (Bai et al., 2020): MPVAE employs a variational autoencoder to align features and labels in a probabilistic latent space and explicitly learns a shared covariance matrix to model the label correlations. [$\lambda_1 = \lambda_2 = 0.5$, $\lambda_3 = 10$, $\beta = 1.1$]
- CLIF (Hang & Zhang, 2021): A deep approach for label-specific feature learning, which finds the most discriminative features for each class label with the guidance of collaboratively learned label semantics. [grid search for $\lambda \in \{10^{-5}, 10^{-4}, \ldots, 1, 2, 5, 10\}$ and $d_e \in \{64, 128, 256\}$]
- PACA (Hang et al., 2022): A prototype-based deep label-specific feature transformation approach, which learns prototypes, label-specific features and classifiers in a unified probabilistic framework. [grid search for $\alpha \in \{1, 2, 5, 10, 20, 50\}$ and $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10\}$]

For our DELA approach, the trade-off parameter β is searched in $\{10^{-5}, 10^{-4}, \ldots, 10\}$. For fair comparison, all deep approaches share the same neural network structure. Grid search is conducted to find the best learning rate and learning rate decay schedule. We take out 10% examples in each data set as hold-out validation set for hyperparamter searching and perform ten-fold cross validation on the remaining 90% examples to evaluate above approaches.

Table 2 and Table 3 report detailed experimental results in terms of each evaluation metric. Table 5 further reports results of the *Wilcoxon signed-ranks test* (Wilcoxon, 1992) at 0.05 significance level to analyze whether DELA performs statistically better than other comparing algorithms. Based on these results, it is impressive to observe that:

- Across all evaluation metrics, DELA achieves the best performance in 92% cases over all the 10 data sets.
- As shown in Table 5, DELA significantly outperforms deep label embedding approaches C2AE and MPVAE in all evaluation metrics. The superior performance of DELA against C2AE and MPVAE indicates that it is a

Table 3. Predictive performance of each comparing approach (mean±std. deviation) in terms of *One-error*, *Coverage* and *Ranking loss*. ↑ (↓) indicates the larger (smaller) the value, the better the performance. Best results are highlighted in **boldface**

Data Sets	*One-error* ↓						
	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA	DELA
corel5k	0.6804±0.0251	0.6460±0.0231	0.6462±0.0145	0.6229±0.0339	0.6320±0.0246	0.6280±0.0279	**0.6178±0.0260**
rcv1-s1	0.4106±0.0194	0.4220±0.0161	0.4383±0.0210	0.4078±0.0330	0.4102±0.0192	0.4026±0.0239	**0.4006±0.0215**
Corel16k-s1	0.6764±0.0126	0.6398±0.0092	0.6442±0.0090	0.6331±0.0155	0.6420±0.0126	0.6283±0.0168	**0.6268±0.0099**
delicious	0.3339±0.0153	0.3537±0.0145	0.3374±0.0178	0.3070±0.0183	0.3194±0.0180	0.3091±0.0156	**0.3061±0.0138**
iaprtc12	0.4649±0.0182	0.4827±0.0118	0.4684±0.0145	0.4279±0.0134	0.4275±0.0097	0.4228±0.0114	**0.4104±0.0105**
espgame	0.6549±0.0211	0.6371±0.0140	0.6511±0.0122	0.6070±0.0118	0.6033±0.0173	**0.5933±0.0101**	0.5938±0.0135
mirflickr	0.3076±0.0106	0.3025±0.0089	0.2848±0.0110	0.2740±0.0105	0.2702±0.0083	0.2695±0.0113	**0.2622±0.0089**
tmc2007	0.2125±0.0076	0.2245±0.0094	0.2296±0.0082	0.2031±0.0072	0.2003±0.0036	0.2013±0.0093	**0.1945±0.0067**
mediamill	0.1757±0.0122	0.1590±0.0040	0.1643±0.0072	0.1422±0.0046	0.1421±0.0060	0.1339±0.0043	**0.1288±0.0038**
bookmarks	0.5115±0.0044	0.5319±0.0054	0.5408±0.0070	0.5165±0.0068	0.5337±0.0050	0.5242±0.0050	**0.5079±0.0042**

Data Sets	*Coverage* ↓						
	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA	DELA
corel5k	0.2906±0.0101	0.4367±0.0228	0.3228±0.0126	0.2314±0.0103	0.2403±0.0105	0.2324±0.0115	**0.2214±0.0086**
rcv1-s1	0.1231±0.0124	0.1245±0.0120	0.1040±0.0078	0.0932±0.0091	0.0992±0.0068	0.0938±0.0096	**0.0866±0.0085**
Corel16k-s1	0.3247±0.0050	0.3243±0.0071	0.3049±0.0068	0.2372±0.0055	0.2499±0.0067	0.2331±0.0072	**0.2330±0.0049**
delicious	0.4809±0.0132	0.6150±0.0093	0.5108±0.0062	0.4058±0.0063	0.4208±0.0051	0.4002±0.0043	**0.3943±0.0049**
iaprtc12	0.3080±0.0120	0.3770±0.0047	0.2940±0.0037	0.2356±0.0036	0.2223±0.0042	0.2297±0.0053	**0.2207±0.0035**
espgame	0.4026±0.0274	0.4157±0.0090	0.3821±0.0049	0.3179±0.0059	0.3203±0.0055	0.3082±0.0071	**0.3029±0.0054**
mirflickr	0.3086±0.0038	0.3205±0.0043	0.3075±0.0041	0.2741±0.0031	0.2757±0.0033	0.2732±0.0035	**0.2686±0.0046**
tmc2007	0.1193±0.0028	0.1270±0.0025	0.1511±0.0059	0.1144±0.0023	0.1161±0.0026	0.1160±0.0033	**0.1110±0.0023**
mediamill	0.1517±0.0088	0.1671±0.0031	0.1760±0.0029	0.1233±0.0033	0.1239±0.0030	0.1164±0.0024	**0.1143±0.0028**
bookmarks	0.1293±0.0060	0.1510±0.0032	0.1905±0.0040	0.1189±0.0028	0.1270±0.0019	0.1183±0.0023	**0.1105±0.0019**

Data Sets	*Ranking loss* ↓						
	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA	DELA
corel5k	0.1219±0.0039	0.1907±0.0107	0.1576±0.0075	0.1038±0.0052	0.1082±0.0048	0.1063±0.0065	**0.0963±0.0038**
rcv1-s1	0.0490±0.0056	0.0497±0.0046	0.0428±0.0032	0.0390±0.0041	0.0426±0.0025	0.0392±0.0039	**0.0344±0.0035**
Corel16k-s1	0.1636±0.0017	0.1611±0.0042	0.1638±0.0052	0.1239±0.0038	0.1303±0.0043	0.1219±0.0037	**0.1202±0.0028**
delicious	0.0985±0.0020	0.1449±0.0046	0.1197±0.0021	0.0884±0.0019	0.0933±0.0017	0.0880±0.0014	**0.0856±0.0016**
iaprtc12	0.1023±0.0033	0.1241±0.0019	0.1045±0.0023	0.0794±0.0021	0.0772±0.0019	0.0784±0.0021	**0.0738±0.0014**
espgame	0.1568±0.0087	0.1668±0.0047	0.1634±0.0027	0.1314±0.0043	0.1368±0.0038	0.1281±0.0045	**0.1246±0.0036**
mirflickr	0.1125±0.0020	0.1196±0.0028	0.1120±0.0038	0.0939±0.0015	0.0986±0.0015	0.0958±0.0022	**0.0937±0.0019**
tmc2007	0.0449±0.0019	0.0489±0.0017	0.0629±0.0023	0.0416±0.0013	0.0432±0.0014	0.0428±0.0022	**0.0395±0.0015**
mediamill	0.0412±0.0017	0.0496±0.0011	0.0537±0.0014	0.0342±0.0012	0.0343±0.0011	0.0320±0.0010	**0.0309±0.0009**
bookmarks	0.0813±0.0037	0.0947±0.0025	0.1271±0.0028	0.0767±0.0020	0.0838±0.0013	0.0765±0.0017	**0.0700±0.0014**

promising direction to facilitate multi-label classification with the strategy of label-specific features.

- Meantime, DELA achieves much better performance against other approaches based on label-specific features. Specifically, DELA is statistically superior to deep approach CLIF in terms of all evaluation metrics, and achieves statistically superior or at least comparable performance against PACA. These consistently better results demonstrate the effectiveness of our dual perspective for label-specific feature learning.

4.3. Further Analyses

4.3.1. ABLATION STUDIES

In ablation studies, we employ ten-fold cross validation on all the 10 data sets to validate the superiority of DELA against its variants. Table 4 summarizes the p-value statistics of the Wilcoxon signed-ranks test at 0.05 significance level and Table 6 shows the detailed experimental results in terms of Average precision.

Table 4. Summary of the Wilcoxon signed-ranks test for DELA against its variants at 0.05 significance level. p-values are shown in the brackets.

DELA against	DELA-sn	DELA-nn	DELA-fs
Average precision	**win** [0.0020]	**win** [0.0020]	**win** [0.0039]
Macro-averaging AUC	**tie** [0.1055]	**win** [0.0098]	**win** [0.0371]
Hamming loss	**win** [0.0078]	**win** [0.0469]	**tie** [0.0996]
One error	**win** [0.0332]	**win** [0.0020]	**win** [0.0098]
Coverage	**win** [0.0098]	**win** [0.0020]	**win** [0.0020]
Ranking loss	**win** [0.0039]	**win** [0.0020]	**win** [0.0020]

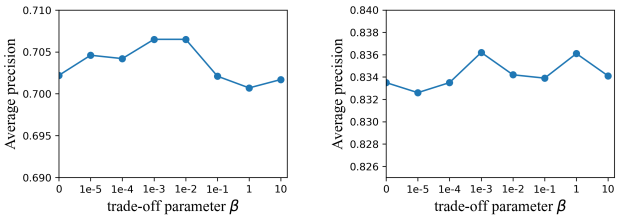
Consideration on label-specific discriminative properties. DELA accounts for each label's own discriminative properties via perturbing label-specific non-informative features in the shared representation $\mathbf{z}$ and inducing classifiers on these perturbed representations. To validate the effectiveness of the above consideration, we implement two variants named DELA-sn and DELA-nn. DELA-sn removes the identification process of label-specific non-informative features and merely perturbs $\mathbf{z}$ with a noise $\boldsymbol{\epsilon}$ shared among all class labels, while DELA-nn further removes the noise and directly induces classifiers on the shared representation $\mathbf{z}$. Results

Table 5. Summary of the Wilcoxon signed-ranks test for DELA against other comparing approaches at 0.05 significance level. p-values are shown in the brackets.

DELA against	LIFT	LLSF	C2AE	MPVAE	CLIF	PACA
Average precision	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]
Macro-averaging AUC	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**tie** [0.0059]
Hamming loss	**win** [0.0352]	**win** [0.0313]	**win** [0.0020]	**win** [0.0078]	**win** [0.0117]	**win** [0.0020]
One-error	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0039]
Coverage	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]
Ranking loss	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]	**win** [0.0020]

Table 6. Predictive performance of DELA and its variants (mean±std. deviation) in terms of *Average precision*.

Data Sets	*Average precision* ↑			
	DELA	DELA-sn	DELA-nn	DELA-fs
corel5k	**0.3382±0.0103**	0.3338±0.0117	0.3321±0.0135	0.3313±0.0124
rcv1-s1	**0.6391±0.0153**	0.6365±0.0149	0.6114±0.0132	0.6355±0.0122
Corel16k-s1	**0.3675±0.0062**	0.3627±0.0048	0.3610±0.0048	0.3635±0.0055
delicious	**0.4082±0.0053**	0.4078±0.0057	0.3872±0.0066	0.3897±0.0065
iaprtc12	**0.4490±0.0050**	0.4401±0.0046	0.4255±0.0057	0.4257±0.0062
espgame	**0.3162±0.0059**	0.3128±0.0049	0.3087±0.0059	0.3099±0.0056
mirflickr	**0.6960±0.0043**	0.6951±0.0069	0.6900±0.0057	0.6922±0.0061
tmc2007	**0.8363±0.0037**	0.8306±0.0050	0.8302±0.0049	0.8304±0.0045
mediamill	0.7883±0.0049	0.7815±0.0055	0.7849±0.0040	**0.7908±0.0071**
bookmarks	**0.5191±0.0036**	0.5104±0.0038	0.5067±0.0033	0.5123±0.0039



(a) mirflickr **(b) tmc2007**

Figure 1. Validation performance of DELA with varying trade-off parameter β in terms of *Average precision*.

in Table 4 show the consideration on label-specific discriminative properties is statistically effective.

Effectiveness of the dual perspective. We implement a variant named DELA-fs, which performs label-specific feature selection on the shared representation $\mathbf{z}$ with the stochastic gates employed in DELA (i.e. $\mathbf{z}_k = round(\mathbf{c}_k) \odot \mathbf{z}$, $\mathbf{c}_k \sim p(\mathbf{c}_k)$ in DELA-fs). It is worth noting that DELA-fs follows the conventional perspective for label-specific feature learning, while its implementation is kept as consistent as possible with DELA, so that it provides an apple-to-apple comparison between the conventional and the dual one. As shown in Table 4 and Table 6, the superiority of our dual perspective against the conventional perspective is statistically significant.

4.3.2. PARAMETER SENSITIVITY

Figure 1 gives an illustrative example on how the performance of DELA changes when the value of the trade-off parameter β changes. Degraded performance is witnessed

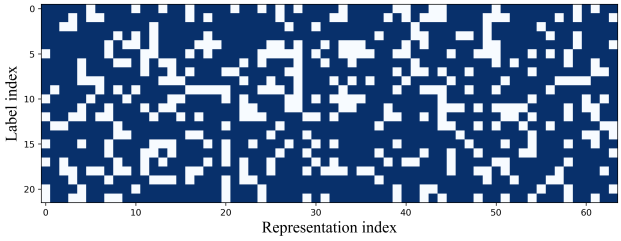


Figure 2. Visualization of identified non-informative features in DELA on tmc2007. The kth row denotes the indicator vector of the subset of non-informative features for label l_k, where the blue one denotes a non-informative feature and the white one denotes a pertinent feature.

when $\beta = 0$, which demonstrates that the proposed constraint on noise distribution does facilitate the learning process. Similar results can be observed on other data sets.

4.3.3. VISUALIZATION

Figure 2 gives an illustrative example on identified non-informative features for each class label. As can be seen, the subsets of non-informative features are quite different among class labels, which is essential to fully consider distinct discriminative properties of each class label. It is appealing to explore how to incorporate label correlations into the identification process of non-informative features, e.g. letting closely-related labels share more features (Huang et al., 2016b), which will be left for future work.

4.3.4. COMPLEXITY ANALYSES

Let b be the batch size and $\hat{d}$ denote a proxy of the hidden dimensionalities of the network, the time complexity of DELA corresponds to $\mathcal{O}(bt\hat{d}^2)$ with t class labels. Figure B.1 illustrates the empirical training and test time of each comparing approach, which shows that DELA is comparable to existing approaches in time overhead.

5. Conclusion

In this paper, we propose to tackle the problem of label-specific feature learning for multi-label classification from a novel dual perspective, where distinct discriminative prop-

erties of each class label are considered by endowing classifiers with immutability on identified label-specific non-informative features. Following this dual perspective, we present a perturbation-based approach DELA which learns to simultaneously identify non-informative features and make the discrimination process immutable to variations of these identified features via solving a probabilistically-relaxed expected risk minimization problem. Theoretical justification from an information theoretic view and comprehensive empirical studies against other well-established multi-label classification algorithms show the superiority of our approach. A nature direction for future work is to incorporate label correlations into the identification process of non-informative features and it is also interesting to explore alternative implementations towards the promising dual perspective for label-specific feature learning.

References

Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905, 2018.

Ahuja, K., Shanmugam, K., Varshney, K. R., and Dhurandhar, A. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 145–155, Virtual Event, 2020.

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems 34*, pp. 3438–3450, virtual, 2021.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

Bai, J., Kong, S., and Gomes, C. P. Disentangled variational autoencoder based multi-label classification with covariance-aware multivariate probit model. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pp. 4313–4321, Yokohama, Japan, 2020.

Blum, A., Haghtalab, N., and Procaccia, A. D. Variational dropout and the local reparameterization trick. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2575–2583, Montreal, Canada, 2015.

Boutell, M., Luo, J.-B., Shen, X.-P., and Brown, C. Learning multi-label scene classification. *Pattern Recognition*, 37 (9):1757–1771, 2004.

Chen, D., Xue, Y., Fink, D., Chen, S., and Gomes, C. P. Deep multi-species embedding. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3639–3646, Melbourne, Australia, 2017.

Chen, T., Lin, L., Chen, R., Hui, X., and Wu, H. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1371–1384, 2022.

Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y.-W. Multi-label image recognition with graph convolutional networks. In *Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5177–5186, Long Beach, CA, 2019.

Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. V. AutoAugment: Learning augmentation strategies from data. In *Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition*, pp. 113–123, Long Beach, CA, 2019.

Dahiya, K., Agarwal, A., Saini, D., K, G., Jiao, J., Singh, A., Agarwal, S., Kar, P., and Varma, M. SiameseXML: Siamese networks meet extreme classifiers with 100M labels. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2330–2340, Virtual Event, 2021.

Devries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with Cutout. *CoRR*, abs/1708.04552, 2017. URL http://arxiv.org/abs/1708.04552.

Duan, R., Chen, Y., Niu, D., Yang, Y., Qin, A. K., and He, Y. AdvDrop: Adversarial attack to DNNs by dropping information. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7506–7515, virtual, 2021.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3449–3457, Venice, Italy, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015.

Greenspan, H., Belongie, S. J., Goodman, R. M., Perona, P., Rakshit, S., and Anderson, C. H. Overcomplete steerable pyramid filters and rotation invariance. In *Proceedings of the 7th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 222–228, Seattle, WA, 1994.

Guo, Y., Chung, F., Li, G., Wang, J., and Gee, J. C. Leveraging label-specific discriminant mapping features for multi-label learning. *ACM Transactions on Knowledge Discovery from Data*, 13(2):24:1–24:23, 2019.

Hang, J.-Y. and Zhang, M.-L. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Hang, J.-Y., Zhang, M.-L., Feng, Y., and Song, X. End-to-end probabilistic label-specific feature learning for multi-label classification. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pp. 1–1, Virtual Event, 2022.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. Deep networks with stochastic depth. In *Proceedings of the 14th European Conference on Computer Vision*, pp. 646–661, Amsterdam, The Netherlands, 2016a.

Huang, J., Li, G., Huang, Q., and Wu, X. Learning label specific features for multi-label classification. In *Proceedings of the 15th IEEE International Conference on Data Mining*, pp. 181–190, Atlantic City, NJ, 2015.

Huang, J., Li, G., Huang, Q., and Wu, X. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3309–3323, 2016b.

Huang, J., Li, G., Huang, Q., and Wu, X. Joint feature selection and classification for multilabel learning. *IEEE Transactions on Cybernetics*, 48(3):876–889, 2018.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

Jia, X.-Y., Zhu, S.-S., and Li, W.-W. Joint label-specific features and correlation information for multi-label learning. *Journal of Computer Science and Technology*, 35(2): 247–258, 2020.

Lee, S., Cho, S., and Im, S. DRANet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In *Proceedings of the 34th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15252–15261, virtual, 2021.

Liu, W., Shen, X., Wang, H., and Tsang, I. W. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Lowe, D. G. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1150–1157, Kerkyra, Greece, 1999.

Maddison, C. J., Mnih, A., and Teh, Y. W. The Concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Steeg, G. V. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems 31*, pp. 9102–9111, Montreal, Canada, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, San Francisco, CA, 2016.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Sun, Y.-P. and Zhang, M.-L. Compositional metric learning for multi-label classification. *Frontiers of Computer Science*, 15(5):Article 155320, 2021.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proceedings of The 37th Annual Allerton Conference on Communications, Control and Computing*, pp. 368–377, 1999.

Tran, L., Yin, X., and Liu, X. Disentangled representation learning GAN for pose-invariant face recognition. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1283–1292, Honolulu, HI, 2017.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. CNN-RNN: A unified framework for multi-label image classification. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, Las Vegas, NV, 2016.

Wehrmann, J., Cerri, R., and Barros, R. C. Hierarchical multi-label classification networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5225–5234, Stockholm, Sweden, 2018.

Weng, W., Lin, Y., Wu, S., Li, Y., and Kang, Y. Multi-label learning based on label-specific features and local pairwise label correlation. *Neurocomputing*, 273:385–394, 2018.

Weng, W., Chen, Y.-N., Chen, C.-L., Wu, S., and Liu, J. Non-sparse label specific features selection for multi-label classification. *Neurocomputing*, 377:85–94, 2020.

Wilcoxon, F. *Individual Comparisons by Ranking Methods*, pp. 196–202. Springer, Berlin, Germany, 1992.

Xu, M. and Guo, L.-Z. Learning from group supervision: the impact of supervision deficiency on multi-label learning. *Science China Information Sciences*, 64(3):Article 130101, 2021.

Xun, G., Jha, K., Sun, J., and Zhang, A. Correlation networks for extreme multi-label text classification. In *Proceedings of The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1074–1082, Virtual Event, CA, 2020.

Yazici, V. O., Gonzalez-Garcia, A., Ramisa, A., Twardowski, B., and van de Weijer, J. Orderless recurrent models for multi-label classification. In *Proceedings of the 33rd IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13437–13446, Seattle, WA, 2020.

Yeh, C., Wu, W., Ko, W., and Wang, Y. Learning deep latent spaces for multi-label classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 2838–2844, San Francisco, CA, 2017.

You, R., Guo, Z., Cui, L., Long, X., Bao, Y., and Wen, S. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pp. 12709–12716, New York, NY, 2020.

Yu, Z.-B. and Zhang, M.-L. Multi-label classification with label-specific feature generation: A wrapped approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Zhan, W. and Zhang, M.-L. Multi-label learning with label-specific features via clustering ensemble. In *Proceedings of the 4th IEEE International Conference on Data Science and Advanced Analytics*, pp. 129–136, Tokyo, Japan, 2017.

Zhang, C. and Li, Z. Multi-label learning with label-specific features via weighting and label entropy guided clustering ensemble. *Neurocomputing*, 419:59–69, 2021.

Zhang, J., Fang, M., and Li, X. Multi-label learning with discriminative features for each label. *Neurocomputing*, 154:305–316, 2015.

Zhang, J., Li, C., Cao, D., Lin, Y., Su, S., Dai, L., and Li, S. Multi-label learning with label-specific features by resolving label correlations. *Knowledge-Based Systems*, 159:148–157, 2018.

Zhang, M.-L. and Wu, L. LIFT: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):107–120, 2015.

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

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

Zhu, Y., Kwok, J. T., and Zhou, Z.-H. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2018.

A. The Proof of Theorem 3.1

Theorem 3.1. *For any random variant* $\mathbf{c} \sim p(\mathbf{c})$*, the information bottleneck can be upper bounded as follows*

$$-I(\mathbf{h}; y) + \beta \cdot I(\mathbf{h}; \mathbf{x}) \leq \mathbb{E}_{p(\mathbf{x},y)}\mathbb{E}_{p(\mathbf{c})}\Big[\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{c})}[-\log q(y|\mathbf{h})] + \beta \cdot KL(p(\mathbf{h}|\mathbf{x},\mathbf{c})||q(\mathbf{h}))\Big].$$

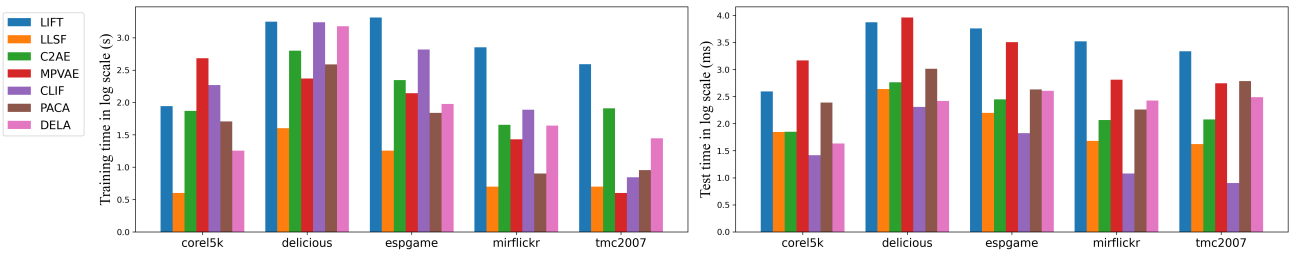
Proof.

$$\begin{aligned}
-I(\mathbf{h}; y) + \beta \cdot I(\mathbf{h}; \mathbf{x}) &\leq \mathbb{E}_{p(\mathbf{x},y)}\Big[\mathbb{E}_{p(\mathbf{h}|\mathbf{x})}[-\log q(y|\mathbf{h})] + \beta \cdot KL(p(\mathbf{h}|\mathbf{x})||q(\mathbf{h}))\Big] \\
&= \mathbb{E}_{p(\mathbf{x},y)}\Big[\mathbb{E}_{p(\mathbf{c})p(\mathbf{h}|\mathbf{x},\mathbf{c})}[-\log q(y|\mathbf{h})] + \beta \cdot \mathbb{E}_{p(\mathbf{c})p(\mathbf{h}|\mathbf{x},\mathbf{c})}[\log \frac{p(\mathbf{h}|\mathbf{x})}{q(\mathbf{h})}]\Big] \\
&= \mathbb{E}_{p(\mathbf{x},y)}\Big[\mathbb{E}_{p(\mathbf{c})p(\mathbf{h}|\mathbf{x},\mathbf{c})}[-\log q(y|\mathbf{h})] + \beta \cdot \mathbb{E}_{p(\mathbf{c})p(\mathbf{h}|\mathbf{x},\mathbf{c})}[\log \frac{p(\mathbf{h}|\mathbf{x},\mathbf{c})}{q(\mathbf{h})} - \log \frac{p(\mathbf{h}|\mathbf{x},\mathbf{c})}{p(\mathbf{h}|\mathbf{x})}]\Big] \\
&= \mathbb{E}_{p(\mathbf{x},y)}\Big[\mathbb{E}_{p(\mathbf{c})p(\mathbf{h}|\mathbf{x},\mathbf{c})}[-\log q(y|\mathbf{h})] + \beta \cdot \mathbb{E}_{p(\mathbf{c})}[KL(p(\mathbf{h}|\mathbf{x},\mathbf{c})||q(\mathbf{h})) - KL(p(\mathbf{h}|\mathbf{x},\mathbf{c})||p(\mathbf{h}|\mathbf{x}))]\Big] \\
&\leq \mathbb{E}_{p(\mathbf{x},y)}\mathbb{E}_{p(\mathbf{c})}\Big[\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{c})}[-\log q(y|\mathbf{h})] + \beta \cdot KL(p(\mathbf{h}|\mathbf{x},\mathbf{c})||q(\mathbf{h}))\Big],
\end{aligned}$$

where the first inequality is derived based on the variational approximation (Alemi et al., 2017) to the information bottleneck and the last inequality is derived based on non-negativity of the KL-divergence. □

B. Empirical Running Time Comparison

Empirical running time of each comparing approach considered in the *Comparative Studies* part of the main body is further reported here for comprehensive evaluation. Figure B.1 illustrates the empirical training and test time of each comparing approach, which shows that DELA is comparable to existing approaches in time overhead.



(a) Training time of comparing approaches

(b) Test time of comparing approaches

Figure B.1. Running time (training/test) of each comparing approach on five benchmark data sets. For histogram illustration, the y-axis corresponds to the logarithm of running time.