Cross-View Representation Learning for Multi-View Logo Classification with Information Bottleneck

Jing Wang
School of Information Science and Engineering
Shandong Normal University
Shandong, China
jingwang1551@163.com

Yuanjie Zheng*
School of Information Science and Engineering
Shandong Normal University
Shandong, China
zhengyuanjie@gmail.com

Jingqi Song
School of Information Science and Engineering
Shandong Normal University
Shandong, China
songjingqi163@163.com

Sujuan Hou*
School of Information Science and Engineering
Shandong Normal University
Shandong, China
sujuanhou@sdnu.edu.cn

ABSTRACT

Multi-view logo classification is a challenging task due to the cross-view misalignment of logo image varies under different viewpoints, large intra-classes and small inter-classes variation of logo appearance. Cross-view data can represent objects from different views and thus provide complementary information for data analysis. However, most existing multi-view algorithms usually maximize the correlation between different views for consistency. Those methods ignore the interaction among different views and may cause semantic bias during the process of common feature learning. In this paper, we investigate the information bottleneck (IB) to the multi-view learning for extracting the different view common features of one category, named Dual-View Information Bottleneck representation (Dual-view IB). To the best of our knowledge, this is the first cross-view learning method for logo classification. Specifically, we maximize the mutual information between the representations of the two views to achieve the preservation of key features in the classification task, while eliminating the redundant information that is not shared between the two views. In addition, due to the unbalance of samples and limited computing resources, we further introduce a novel Pair Batch Data Augmentation (PB) algorithm for Dual-view IB model, which applies augmentations from a learned policy based on replicates instances of two samples within the same batch. Comprehensive experiments on three existing benchmark datasets, which demonstrate the effectiveness of the proposed method that outperforms the methods in the state of the art. The proposed method is expected to further the development of cross-view representation learning.

*Corresponding Author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475704

CCS CONCEPTS

• **Computing methodologies** → **Image representations**; **Object recognition**.

KEYWORDS

multi-view logo classification; information bottleneck; representation learning; data augmentation

ACM Reference Format:
Jing Wang, Yuanjie Zheng, Jingqi Song, and Sujuan Hou. 2021. Cross-View Representation Learning for Multi-View Logo Classification with Information Bottleneck. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475704

1 INTRODUCTION

With the rapid development of multimedia information, the amount of images data on the Internet continues to grow. Each view of images data may have distinct properties and potentially be noisy. Logo classification has attracted increasing attention in the multimedia for its various real-world applications, such as brand visibility analysis [33, 49], copyright infringement detection, video advertising research [14, 15, 36], and automated computation of logo-related statistics on social media [19, 34]. In reality, the data of multimedia logo images has two characteristics in logo classification: (1) Large number of categories, unbalanced samples, and few samples in some categories. (2) Large intra-class and small inter-classes variation of logo appearance. There are many very similar categories in the same root directory. The examples of similar samples between classes show in Fig. 1. Specifically, 'Molson Canadian' and 'Molson Golden' are two similar logos belonging to one company. 'Caloi' and 'Kuota' are bicycles of different brands with very similar appearance, while the car logos 'Bentley' and 'Morgan' have similar backgrounds and similar graphics. Therefore, to achieve high-performance logo classification, the classification method requires robust and discriminative feature representation from similar categories.

According to Bengio *et al.* [7, 8], the success of a classification algorithm largely depends on representation learning because different representations can hide the different explanatory factors

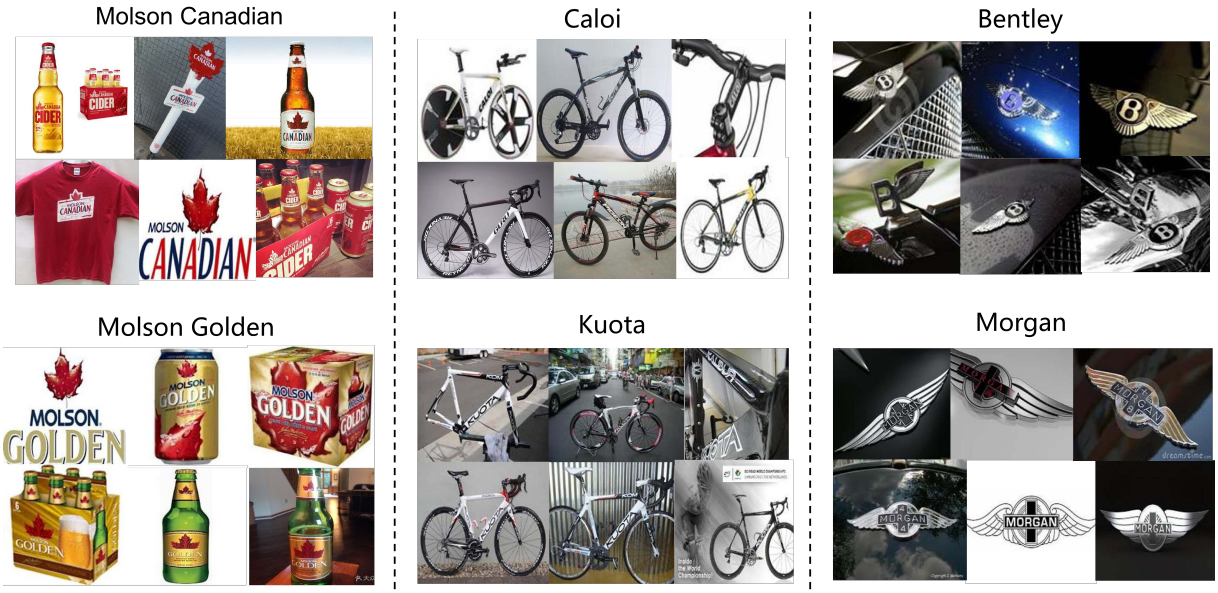

Figure 1: Examples of similar samples between classes.

behind the data. Logo classification requires robust feature representation from similar categories. Thus, it is important to introduce multi-view learning to multimedia classification. Recently, multi-view learning [50] has attracted widespread attention from deep learning researchers because it can consider more complete information than single-view. However, they often fail to take advantage of the high-order information among different views. Additionally, multi-view correlations among multiple views often vary drastically, which makes multi-view representation challenging. Considering that the single-view information bottleneck (IB) information-theoretic principle has shown good performance in robust representation and image classification, we rely on the IB to save the same class-relevant information. The key of IB is to compress and retain effective latent features for each category, using category discriminative features to distinguish from other categories. Since for real-world data, the relation between multiple views is likely to be nonlinear and complex.

In this paper, we investigate the IB method to the supervised multi-view learning for logo classification, named Dual-view information bottleneck representation (Dual-view IB). The Dual-view IB can extract the common feature of a logo category and enhance information from different views. Specifically, we maximize the mutual information between the representations of the two views to achieve the preservation of key features, while eliminating the redundant information that is not shared between the two views. In addition, due to the unbalance of samples and limited resources, we further introduce a novel Pair Batch Data Augmentation (PB) method for the Dual-view IB model, which learns a policy based on the transformers and replicates instances of two samples within the same batch. The contribution of this work is threefold,

- The proposed Dual-view IB algorithm is the first to apply information bottleneck theory and multi-view learning for logo classification, which learns quality-related features and eliminates irrelevant features to obtain more distinctive features of a logo category. It can incorporate Dual-view dynamic contributions to representation learning and derive reasonable joint latent representation consists of consistent and view-specific information.
- To solve some categories unbalanced, the Pair Batch Data Augmentation (PB) is proposed to the Dual-view IB model to further improve the robust feature representation, it can better utilize the computational resources via using larger augmented batches and transformers.
- We conduct an extensive evaluation on three datasets, including the newly proposed Logo-2K+, and the other two datasets with different scales, namely BelgaLoges and FlickrLogos-32. The experimental results verified the effectiveness of the proposed method on all these datasets.

2 RELATED WORK

Our work is closely related to three research fields, including logo classification, information bottleneck, and data augmentation.

2.1 Logo Classification

Logo classification has always been extensively studied in the field of e-commerce and multimedia [10, 13, 18, 24]. Social media websites contain large amounts of uncollected image data, from which many useful insights can be extracted. Traditional methods for logo classification rely on hand-crafted features and keypoint-based detectors [42]. Romberg *et al.* [42] presented a logo recognition technique based on feature bundling, aggregated with features into Bag of Words (BoW). Recently some works investigated the use of deep learning for logo classification, its superior performance with end-to-end pipeline automation [10, 18, 24]. Bianco *et al.* [10] used

the CNN and synthetically region proposal to generate data. In [25], the authors introduced a logo classification mechanism that combines deep representation and traditional algorithms. In addition, some works proposed logo detection methods [32, 40, 44], one standard approach is that one image is fed into one state-of-the-art deep neural detector. Deep learning can extract global features, however, it is impossible to discard irrelevant information and retain relevant information in a complex image, resulting in feature redundancy.

2.2 Information Bottleneck

With the empirical success of deep learning for image-based tasks, many researchers attempted to establish information-theoretic foundations for learning representation [4–6, 37]. Besides, people start to investigate IB theory from different views such as gaussian multivariate systems and neural network-based classification. Alemi *et al.* [2] applied a deep variational information bottleneck (VIB) algorithm to learn useful representation with a way of single-view learning. It can extract the features which are related to the target output and remove the unrelated information. In [12], the authors proposed to use IB to learn a joint latent representation. It can directly apply the IB principle by minimizing the mutual information (MI) between the data and its representation. IB in multi-view studies [17] aims to find the MI between different views. Despite the comprehensive exploration of view-specific features, the underlying shared knowledge among different views is lost.

Recently, to effectively utilize knowledge from multiple views, a variety of multi-view learning methods have been proposed. Canonical correlation analysis (CCA) was widely used multi-view learning methods [29, 30]. These methods would project the original features from the two views into a shared feature space. For image classification, information of each view has potentially noisy, multi-view learning can help improve classification performance by learning the common structure of the different views. For the IB approach [2, 12, 47], computing MI of different views requires estimation of the posterior distribution, which is computationally intractable when the model is complicated.

2.3 Data Augmentation

In the computer vision domain, image augmentation has become a common regularization technique to combat overfitting in deep CNN. Most deep learning frameworks implement based on a manually designed set of simple transformations such as mirroring and color perturbations [1, 27]. In general, designing an effective data augmentation pipeline requires domain-specific knowledge [39]. For example, Fawzi *et al.* [3] used adaptive data augmentation to choose transformations that maximize loss for the classifier. Recently, Cubuk *et al.* [16] attempted to automatically find the best augmentation functions using Reinforcement Learning. However, selecting the optimal set of transformations is often a non-trivial task. With the steady improvement of GPU performance and memory capacity, the efficiency of data augmentation operations becomes increasingly important.

In summary, our work conducts on multi-view logo classification. Different from previous work, the work mainly focuses on learning robustness of feature representation by discarding irrelevant information, obtaining inter-feature correlation via a non-linear projection of Dual-view information bottleneck. In addition, inspired by the spatial transformation [16], we further introduce Pair Batch Data Augmentation to solve the unbalanced and lacking computational resources.

In this section, we will introduce the Dual-View Information Bottleneck Representation Network (Dual-view IB) for logo classification. It aims at learning relevant representation and learning a joint representation. Specifically, our method maximizes the MI between the label and the learned joint representation, while minimizing the MI between the learned latent representation of two views and the original data representation. In addition, aiming at the Dual-view IB method, Pair Batch Data Augmentation (PB) is proposed to replicating instances of samples within the same batch with different augmentations.

Fig. 2 shows an overview of the proposed Dual-view Information Bottleneck representation Network. The pair images of one class as two input view X_1 and X_2 with the same label Y. The figure above shows the complete architecture, and the figure below shows the effect of scales of the Information Bottleneck module in the figure above. The method firstly adopted a neural network to extract features, and the feature as the input to the Information Bottleneck module, which can filter out irrelevant and noisy information from dual views. Then, using a DNN to fuse the latent representation Z_1 and Z_2 to make a prediction, it transfers knowledge from dual views and learns a joint representation Z. Finally, we use the learned discriminative representation to achieve logo classification via supervision of the category label.

3 METHOD

The parts consist of the following phases: (i). Information Bottleneck Concepts. (ii). Dual-view Information Bottleneck Representation. (iii). Pair Batch Data Augmentation Strategy.

3.1 Information Bottleneck Concepts

Each logo image can be viewed as having two different information, X and Y variables. X represents the image of class, while Y is the class label. The MI $I(X;Y)$ between the two variables is,

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= -\textstyle\sum_{x\in X}\ p(x)\ \log\ p(x) \\ &\quad + \textstyle\sum_{y\in Y}\ p(y) \sum_{x\in X}\ p(x|y)\ \log\ p(x|y) \\ &= -\textstyle\sum_{x\in X}\ p(x)\ \log\ p(x) \\ &\quad + \textstyle\sum_{y\in Y}\ \sum_{x\in X}\ p(y,x)\ \log\ p(x|y) \end{aligned} \tag{1}$$

where $H(X)$ and $H(X|Y)$ are entropy and conditional entropy respectively. $p(y,x)$ denotes joint probability distribution. Formally, the concept of single-view MI between two random variables can be formulated by,

$$I(X,Y)\ =\ \sum_{x\in X}\sum_{y\in Y} p(x,y)\ \log\ \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

where $p(x,y)$ denotes the joint distribution between two variables, and $p(x)$ and $p(y)$ denote the distribution of X and Y, respectively.

The goal of IB is to find a latent representation of Z, the learned Z contains information about Y as much as possible, i.e. $max\ I\ (Z;Y)$ and $min\ I\ (X;Z)$. The tradeoff between the view-specific information compression and preservation is obtained by introducing a

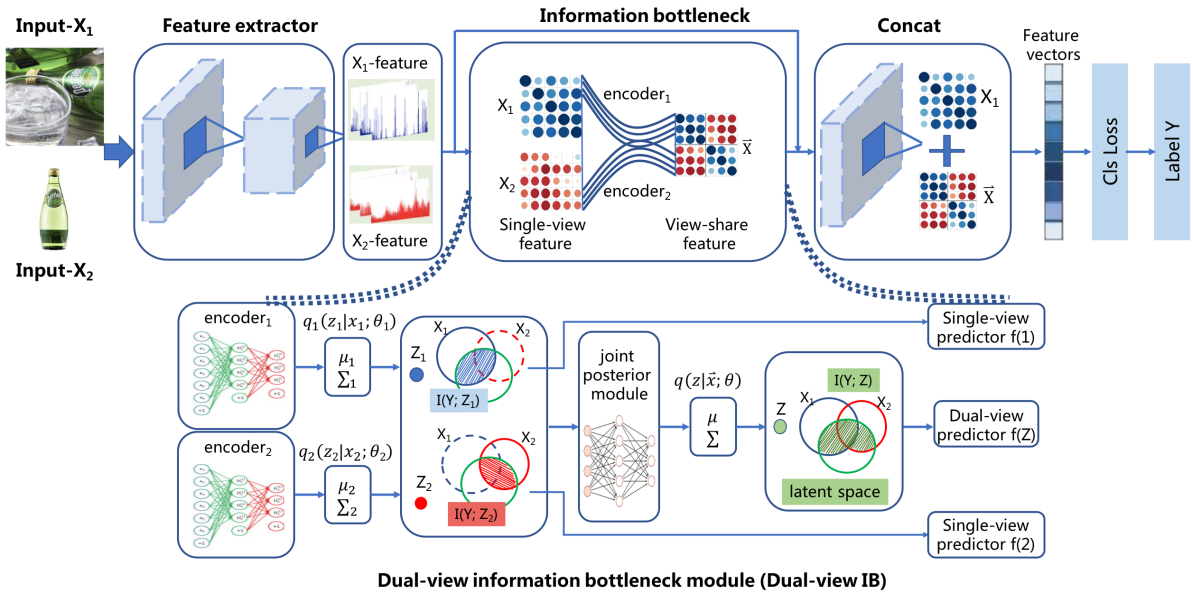

Figure 2: An overview of proposed Dual-View Information Bottleneck representations Network (Dual-view IB). The above figure shows the complete architecture of the Information Bottleneck module. And the figure below is a detailed diagram for learning joint representation.

Lagrange multiplier β, the learning parameters are defined as θ. The object function is,

$$L_{IB}(\theta) = I(Z,\ Y;\ \theta) - \beta I(Z,\ X;\ \theta) \tag{3}$$

3.2 Dual-view information bottleneck representation

In this method, we give two views X_1, X_2 and the class label Y for learning a joint representation Z, fusing the information from all views. The Dual-view IB representation is,

$$L_{DVIB}(\theta) = I(Z,\ Y;\ \theta) - \alpha I(Z_1,\ X_1;\ \theta_1) - \beta I(Z_2,\ X_2;\ \theta_2) \tag{4}$$

$$s.t.\ Z = f_\theta(Z_1,\ Z_2) \tag{5}$$

where X_1 and X_2 is two-views, namely two logo images of one class. Z_1 and Z_2 are the hidden feature representation, respectively. $f_\theta(\cdot)$ is neural network with parameters.

Notation.

Let X_1 and X_2 view ($X_{view} = X_1, X_2 = v_1, v_2$) belonging to a random variable for the input feature, and Y is a random variable for the output label. We defined a random variable for a Dual-view observation $\overrightarrow{X} = (X_1 : X_2)$ as the intermediate result of x_1 and x_2. x_1 and x_2 are realization of X_1 and X_2. As shown in Fig. 2, the Dual-view IB module contains three network parts, including dual-view encoder, integrate module, dual-view predictor. We will learn representation in a common space that leverages both marginal and joint aspects of the observed views for prediction.

The Dual-view Encoder for Representation.

We define two encoders, parameterized by $\theta = (\theta_1;\ \theta_2)$, each of which randomly maps observations from each individual view into a common latent space. In Fig. 2, we define Z as the latent space, and the marginal representation $\{Z_1, Z_2\} \in \mathbb{Z}$ be a stochastic encoding of X_1 and X_2 view-specific encoder, and joint representation $Z \in \mathbb{Z}$ be a stochastic encoding of $\overrightarrow{X}$, which is defined by the encoder $q_\theta(Z|\overrightarrow{X})$.

It combines outputs of the two view encoders. Then, the dual-view predictor estimates the label Y based on the distribution defined as $q(Y|z)$. Using these assumptions, the joint probability density function can be simplified as,

$$p(x_1,\ x_2,\ z_1,\ z_2,\ y,\ z) = p(z|\overrightarrow{x})p(z_1|x_1)p(z_2|x_2)p(x_1,\ x_2,\ y) \tag{6}$$

and the variational lower bound of the mutual information between Z and Y can be written as,

$$I(Y,\ Z) \approx \int dzdz_1dz_2 \cdot p(z|\overrightarrow{x})p(z_1|x_1)p(z_2|x_2)\ log\ q(y|z) \tag{7}$$

To learn joint aspects of the observed views for predicting the target Y, we first start with $I(Z,\ Y;\ \theta)$ in Eq.(6). We apply the IB principle on Z and $\overrightarrow{X}$ based on the following loss:

$$\begin{aligned} L_{DVIB}^{\theta,\varphi}(Y,\ \overrightarrow{X}) &= -I(Y,\ Z) - \beta I(\overrightarrow{X},\ Z) \\ &= E_{\overrightarrow{x},y\sim p(\overrightarrow{x},y)} E_{z\sim q_\theta(z|\overrightarrow{x})}\left[-\log q_\varphi(y|z)\right] \\ &+ \beta E_{\overrightarrow{x}\sim p(\overrightarrow{x})}\left[KL\left(q_\theta(Z|\overrightarrow{x})||q(Z)\right)\right] \end{aligned} \tag{8}$$

where β is a coefficient chosen to balance the two information quantities. Denoted the Kullback-Leibler divergence $KL\left(q_\theta(Z|\overrightarrow{x})||q(Z)\right)$ between the two distributions $p(Z)$ and $q(Z)$.

Next, we need to find the upper bound of I(Z_1, X_1). Since $p(z_1)$ is intractable, we use $r_1(z_1)$ to approximate $p(z_1)$. Similarly, the mutual information between Z_1 and X_1 is,

$$\begin{aligned} I(Z_1,\ X_1) &\approx \int dx_1 dx_2 dy dz_1 p(x_1,\ x_2,\ z_1,\ y)\ \log\ \frac{p(z_1|x_1)}{r_1(z_1)} \\ &\overset{x_1, z_1\ independent}{\approx} \int dz_1 p(z_1|x_1) \log\ \frac{p(z_1|x_1)}{r_1(z_1)} \end{aligned} \tag{9}$$

Similarly, for I(Z_2, X_2), we have

$$I(Z_2,\ X_2) \approx \int dz_2 p(z_2|x_2) \log\ \frac{p(z_2|x_2)}{r_2(z_2)} \tag{10}$$

Computing the Joint Posterior. Then, we introduce a integrate module that factorizes the joint posterior $q_\theta(Z|\overrightarrow{X})$ into a product of the marginal posteriors $q_{\theta_1}(Z|X_1)$ and $q_{\theta_2}(Z|X_2)$. C is a normalizing constant. Formally, the joint posterior can be defined as the following:

$$p(z|\overrightarrow{x}) \approx C \cdot p(z) \cdot \prod_{v_1, v_2 \in V} q_{\theta_1}(z|v_1) \overset{define}{=} q_\theta(z|\overrightarrow{x}) \tag{11}$$

The joint posterior can produce a clear distribution to be specialized in a particular aspect of the task, the $q_\theta(z|\overrightarrow{x})$ formula for calculating the joint posterior is,

$$\begin{aligned} q_\theta(z|\overrightarrow{x}) &= q_\theta(z|x_1, x_2) \\ &= \alpha_1 v_1 \cdot q_{\theta_1}(z|x_1) + \alpha_2 v_2 \cdot q_{\theta_2}(z|x_2) \end{aligned} \tag{12}$$

We assume $p(z_1|x_1)$, $p(z_2|x_2)$ and $p(z|\overrightarrow{x})$ are Gaussian. The means and variances of the Gaussian distributions are all learned from deep neural networks,

$$\begin{aligned} &q(z_1|x_1) = \mathbb{N}(\mu_1(x_1;\ \varphi_1),\ \textstyle\sum_1(x_1;\ \varphi_1)), \\ &q(z_2|x_2) = \mathbb{N}(\mu_2(x_2;\ \varphi_2),\ \textstyle\sum_2(x_2;\ \varphi_2)), \\ &q(z|z_1, z_2) = q(z|\overrightarrow{x}) = \mathbb{N}(\mu(z_1, z_2;\ \theta),\ \textstyle\sum(z_1, z_2;\ \theta)) \end{aligned} \tag{13}$$

where μ_1, μ_2, μ and $\sum_1$, $\sum_2$, $\sum$ are the networks to learn the means and variances for $p(z_1|x_1)$, $p(z_2|x_2)$ and $p(z|\overrightarrow{x})$. φ_1, φ_2, θ are network parameters for the networks to learn $q(z1|x1)$, $q(z2|x2)$ and $q(z|z1,\ z2)$, respectively. Supposed that we utilize marginal posteriors of the form, and use the prior of the form $p(z) = N(z|\mu_0,\ \Sigma_0)$,

$$q_{\theta_v}(z|x_v) = N(z|\mu_v,\ \Sigma_v) \tag{14}$$

We can derive the joint posterior as $q_\theta(z|\overrightarrow{x}) = \mathbb{N}(z|\mu, \sum)$. Hence, we can efficiently compute the joint posterior of the incomplete multi-view observations in terms of the available marginal posteriors, the formula is as follows,

$$\begin{aligned} \mu &= (\mu_0 \Sigma_0^{-1} + \textstyle\sum_{v \in V} \mu_v \Sigma_v^{-1})(\Sigma_0^{-1} + \sum_{v \in V} \Sigma_v^{-1})^{-1} \\ \textstyle\sum &= (\Sigma_0^{-1} + \textstyle\sum_{v \in V} \Sigma_v^{-1})^{-1} \end{aligned} \tag{15}$$

Building dual-specific predictors

However, training a joint posterior module is difficult, We introduce two view-specific predictors and apply the IB principle to the marginal representation. For two view, given the latent representation $z_v \in Z$ from $q_{\theta_v}(z_v|x_v)$, the predictor estimates Y, it can be solely described by the corresponding view X_1, X_2, based on the distribution defined as $q_{\Phi v}(Y_v|z_v)$. The predicting the target based on the following loss:

$$\begin{aligned} L_{DVIB}^{\theta_v, \Phi_v}(X_v,\ Y_v) &= -I(Y_v;\ Z_v) + \beta_v I(X_v;\ Z_v) \\ &\approx E_{x_v,\ y \sim p(x_v,\ y)}\ E_{z,\ q_{\theta v}(z|x_v)}[-\log q_{\Phi v}(y|z)] \\ &+ \beta_v E_{x_v \sim p(x_v)}[KL(q_{\theta v}(Z_v|x_v)\,||q(Z_v))] \end{aligned} \tag{16}$$

where β is a balancing coefficient. Minimizing encourages Z_v to become a minimal sufficient statistics of X_v for Y. It can enforce the marginal representation to learn the view aspects of the target.

Optimization and training. We trained the overall network, the encoder predictor pairs (θ; φ) and the final predictor Φ by minimizing a combination of the marginal and joint Dual-view IB losses:

$$\begin{aligned} L_{IB}^{\theta,\varphi,\Phi}(\overrightarrow{X}, Y) &= L_{DVIB}^{\theta,\varphi}(\overrightarrow{X}, Y) + \alpha L_{IB}^{\theta 1, \varphi 1}(X_1, Y) + \beta L_{IB}^{\theta 2, \varphi 2}(X_2, Y) \\ &= \textstyle\sum^N \{E_\Phi E_{\Phi 1} E_{\Phi 2} \log q(y|z)\} \\ &+ \textstyle\sum^N \{\alpha E_{\Phi 1} \log \frac{p(z_1|x_1)}{r_1(z_1)} + \beta E_{\Phi 2} \log \frac{p(z_2|x_2)}{r_2(z_2)}\} \end{aligned} \tag{17}$$

where $p(z_1|x_1)$, $p(z_2|x_2)$ are all learned from neural networks. Note that the first term is the cross-entropy between y and z. Then, we can use a deep neural network with a softmax layer as output to calculate the class probabilities and the cross-entropy loss. The total loss of logo classification is,

$$Loss_{total} = L_{IB}^{\theta,\varphi,\Phi}(\overrightarrow{X}, Y) - L_{entropy} \sum_{i=1}^{N} y_i \log(p(x_i)) \tag{18}$$

3.3 Pair Batch Data Augmentation Strategy

To achieve a larger batch augmentation from two samples from dual-view without introducing additional data, the method name is Pair Batch Data Augmentation (PB), which enables to control of the gradient variance while increasing batch size. We consider a model with a loss function $L(w,\ x_1,\ x_n)$ where $\{x_1,\ ...,\ x_n\}_{n=1}^N$ is a dataset of N data sample-target pairs, where $x_n \in X$ and $T :\ X \to X$ is some data augmentation transformation applied to each example. It uses SGD with a learning rate η and batch size of B, and we suggest introducing M multiple instances of the same input sample by applying the transformation T_i, here denoted by subscript $i \in [M]$ to highlight the fact that they are different from one another. We now use the slightly modified learning rule:

$$w_{t+1} = w_t - \eta \frac{1}{MB} \sum_{i=1}^{M} \sum_{n \in (k(t))}^{M} \nabla w L(w_t,\ T_i(X_1),\ X_2) \tag{19}$$

where $k(t)$ is sampled from $[\frac{N}{B}] \triangleq \{1\ ,...,\ \frac{N}{B}\}$, $B(k)$ is the set of samples in batch k, and we assume the B divides N.

The PB method effectively using a larger MB batch at each step, which is composed of B samples augmented with M different transforms each. We note that this updated rule can be computed either by evaluating the whole MB batch. We examine the optimization dual-view samples of non-augmented datasets, using loss functions

is,

$$f(w) = \frac{1}{N}\sum_{n=1}^{N} L(w,\ x1,\ x2) \tag{20}$$

where $\{x_1,\ x_n\}_1^N$ is a dataset of N data sample target pairs and L is the loss function. We typically converge to a minimum $w*$ which is a global minimum on all data points in the training set. This means that $\forall n : \nabla w L(w*,\ x_1, x_2) = 0$. We linearize the dynamics of near $w*$ to obtain,

$$w_{t+1} = w_t - \eta\frac{1}{B}\sum_{n\in(k(t))} H_n \cdot w_t \tag{21}$$

where we assume that $w* = 0$, and denote $H_n = \nabla_w^2 L(w*,\ x_1, x_2) = 0$ as the per-sample Hessian. And the maximum over the maximal eigenvalues of $\{\langle H_k\rangle\}_{k=1}^{N/B}$

$$\langle H_k\rangle \triangleq \frac{1}{B}\sum_{n\in B(k)} H_n \tag{22}$$

$$\lambda_{max} = \max_{k\in[N/B]}\ \max_{\forall v:\|v\|=1} V\{\langle H_k\rangle\}_{k=1}^{N/B} \tag{23}$$

where the λ_{max} affects SGD, and $\lambda_{max} < \frac{2}{\eta}$ learning rate. We further research the pair images augmentation and variance reduction. Standard SGD averages the gradient over different samples, while PB additionally averages the gradient over several transformed instances $T\ (x_n)$ of the same samples.

$$\begin{aligned} &E_{n\in[N]}\left[Corr\ (\nabla_w^{(n)}),\ \nabla_w L(w,\ T(x_1),\ x_2)\right] \\ &>\ E_{n,m\in[N],\ n\neq m}\left[Corr\ (\nabla_w^{(n)}),\ \ (\nabla_w^{(n)})\right] \end{aligned} \tag{24}$$

where $\nabla_w^{(n)} \triangleq \nabla_w L(w,\ T(x_1),\ x_2)$. This implies that the λ_{max} would change less in PB than standard large-batch training, allowing the model to exhibit less of the aforementioned SGD convergence issues. Therefore, PB achieves the batch augmentation of Dual-view IB to efficiently discover more discriminative representation better for logo classification.

4 EXPERIMENTS

In this section, we demonstrate the effectiveness of our Dual-view IB model against state-of-the-art baselines, three datasets are adopted in the experiments, including Logo2K+ [46], BelgaLoges [38] and FlickrLogos-32 [41]. We present various experimental results, comparing the standard classification networks to stochastic neural networks trained by optimizing the VIB and Dual-view IB objective. Both Top-1 accuracy (Top-1 ACC.) and Top-5 accuracy (Top-5 ACC.) are adopted as evaluation metrics. Top-1 ACC. represents the probability that the test has the highest confidence and belongs to the ground-truth category, while Top-5 ACC. represents the probability that the top five categories with higher scores contain the ground-truth category. For the experiment datasets, the statistics of three datasets are shown in Table 1.

4.1 Experiments on Logo-2K+

dataset Logo-2K+, a new large-scale publicly available real-world logo dataset, is crawled from the Internet with 167,140 images and 2,341 categories. The 70%, 30% of images are randomly selected for training and testing in each logo category. The training samples are 116,998 while the test images are 50,142. Each image is represented by a vector of image features view X_1, while the other image of the same category is represented as another view X_2.

Table 1: Statistics of three dataset.

♯Datasets	♯Classes	♯Images	♯Trainval	♯Test
Logo-2K+	2,341	167,140	116,998	50,142
BelgaLoges	37	10,000	7,000	3,000
FlickrLogos-32	32	2,240	1,310	930

Table 2: Comparison of our model and baselines on Logo-2K+ (%).

Method	Top-1 Acc.	Top-5 Acc.
AlexNet	48.80	78.45
GoogLeNet	62.36	88.33
VGGNet-16	62.83	89.01
ResNet-50	66.34	91.01
ResNet-152	67.65	91.52
VGGNet-16+Efficient+LS [23]	65.45	90.12
ResNet-50+Efficient+LS [23]	66.94	91.30
ResNet-152+Efficient+LS [23]	67.99	91.68
NTS-Net(ResNet-50) [48]	69.41	91.95
DRNA-Net(ResNet-50)	71.12	92.33
DRNA-Net(ResNet-152)	72.09	93.45
VIB(ResNet-50)	75.32	94.61
Dual-view IB (ResNet-50)	**78.02**	**95.41**
Dual-view IB (ResNet-50)+PB Augment	**80.27**	**97.39**

Baselines To validate the effectiveness of the proposed classification scheme Dual-view IB, we compared it with other several baselines. The baselines use various domain-specific classification methodologies, including classic network of AlexNet [31], GoogLeNet [45], VGGNet [43] and ResNet [22]. We can regard the IB as the regularization in deep learning, which can improve the generalization ability of the model and prevent overfitting. Thus, in baselines, we also adopt some regularization methods containing Label Smoothing Regularization (LS) [23] for Logo-2K+ classification. We also compared with the single-view IB method VIB [2].

Experimental Setup All training images are used as two views for the logo classification task, which have different image features extracted from two shared ResNet-50 networks. $Encoder_1$ and $encoder_2$ consist of a multi-layer hidden layers with ReLU activations, to learn two latent representations z_1 and z_1 for input images X_1 and X_2, respectively. The feature extractors are frozen during the training procedure for the two representations. Each training iteration used batches of size $B = 32$. The initial value of β is set to 10^{-4}. All models are trained with an initial learning rate of 0.01. α has similar settings. We estimate $I(x; z)$ and $I(y; z)$ using mutual information estimation networks trained from the final representation using of joint samples (x_1, x_2, z). The PB data augmentation is applied for Dual-view IB architecture and the model is trained for iterations, and the batch size is increased to $B = 64$. Our experiments are implemented based on PyTorch deep learning framework. All models were trained end-to-end with 2 NVIDIA GTX 1080Ti GPUs.

Result We list experimental results of baselines and our proposed method on the Logo-2K+ dataset in Table 2. In Table 2, Dual-view IB achieves the best performance: the 78.02% Top-1 accuracy and the 95.41% Top-5 accuracy with ResNet-50. The classification performance of our method far exceeds the best DRNA-Net classification network 8.18%. These experimental results show the effectiveness of IB for extracting effective features and removing redundant features. In addition, compared with the single-view VIB method, Dual-view IB improves the 2.7% Top-1 accuracy and 0.8% Top-5 accuracy, respectively. The Dual-view IB method combines the IB and multi-view learning indeed to achieve the best performance on logo classification tasks. In addition, we add the PB augmentation training strategy to the Dual-view IB model, the best performance is achieved, which is 80.27% Top-1 accuracy and 97.39% Top-5 accuracy, Comparing with the Dual-view IB model, the Top-1 accuracy of 2.25% has been improved. It illustrates the effectiveness of the PB algorithm and solves the problem of computing resources and accuracy. To increase the batch size starts by scaling nearly linearly to achieve a batch increase between 32 and 64. PB can increase the local batch size as much as possible to maximize equipment utilization.

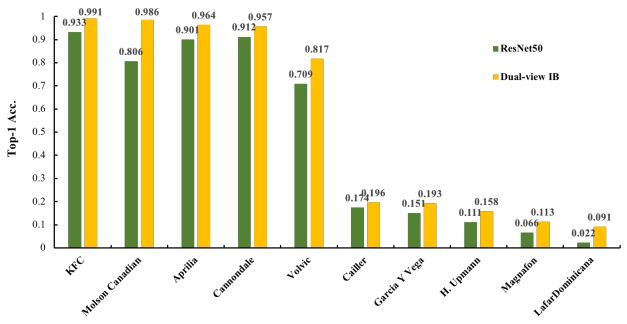

Figure 3: Selected categories from the 5 best and the 5 worst performing classes.

Qualitative analysis Taking into account the imbalance and similarity of samples from different logo categories, we will further comprehensively evaluate the performance of Dual-view IB. We select ten classes in the test phase to further evaluate our method. Particularly, we listed the Top-1 accuracy of both five best and five worst-performing classes in Fig. 3. We find that for all the classified classes, there is performance improvement for our method. We also observe that some categories can be easily classified, such as 'KFC' and 'Aprilia', and their Top-1 accuracy is above 98%. However, there are some categories, which are very hard to classify, such as 'Aprilia' and 'LafarDominicana', and their Top-1 accuracy is below 10%. In addition, for the categories with bad classification results, we observe that most of these logo categories contain fewer samples, noisy background areas, too many similar categories. For some similar categories, such as Molson Canadian, Dual-view IB can effectively extract cross-modal and cross-view image representation, which proves the effectiveness of IB and multi-view learning and achieve better classification results in large-scale categories.

Table 3: Performance comparison of methods on BelgaLoges(%).

Method	Top-1 Acc.	Top-5 Acc.
RCNN(CaffeNet) [21]	91.80	-
FRCN(VGGNet-16) [20]	87.30	-
SPPnet(ZF) [28]	87.70	-
NTS-Net(ResNet-50) [48]	93.33	96.15
DRNA-Net(VGGNet-16)	92.41	95.96
DRNA-Net(ResNet-50)	94.44	97.11
DRNA-Net(ResNet-152)	95.82	98.40
VIB (ResNet-50)	95.06	97.79
Dual-view IB (ResNet-50)	**95.76**	**97.98**
Dual-view IB (ResNet-50)+PB Augment	**96.42**	**98.81**

4.2 Experiments on BelgaLoges Baseline

Experimental Setup To evaluate the robustness and generalization ability of the Dual-view IB architecture, we explore on BelgaLoges [38]. It contains 37 logo classes and a total of 10,000 images. The two view images in all training samples are normalized before Dual-view IB training, and all models are end-to-end trained. The setting of BelgaLoges is the same as the experimental setting of Logo-2K+. To facilitate the optimization, the hyper-parameter β and α is slowly increased during training, since it starts from random initialization.

Results on BelgaLoges As shown in Table 3, We list experimental results on BelgaLoges. Our method achieves the best performance: the 96.42% Top-1 accuracy and the 98.81% Top-5 accuracy. There is about 0.6% improvement in Top-1 accuracy compared with the DRNA-Net method. These experimental results can further demonstrate the effectiveness of our proposed method. In addition, taking the dataset as an example, the effectiveness of the PB augmentation strategy will be stereotyped and analyzed. In Fig. 4, the convergence speed of the PB strategy test is significantly improved, and the final classification error has been significantly reduced, which shows that we achieve higher accuracy faster via pair batch augmentation. The PB to achieve 93% accuracy in only half iterations (28K), the larger learning rate and faster learning rate decay schedule. This result indicates not only an accuracy gain but a potential runtime improvement for given hardware.

Table 4: Comparison of our model and state-of-the-art methods on FlickrLogos-32 (%).

Method	Top-1 Acc.	Top-5 Acc.
FRCN + AlexNet [26]	75.00	-
Eggert et. al[11]	84.60	-
Bianco et. al[9]	88.40	-
Bianco et. al[10]	91.70	-
SIFT [42]	94.10	-
NTS-Net(ResNet-50) [48]	94.14	96.29
DRNA-Net(ResNet-50)	95.33	97.17
DRNA-Net(ResNet-152)	96.63	98.80
VIB (ResNet-50)	96.48	98.85
Dual-view IB (ResNet-50)+PB Augment	**97.46**	**99.06**

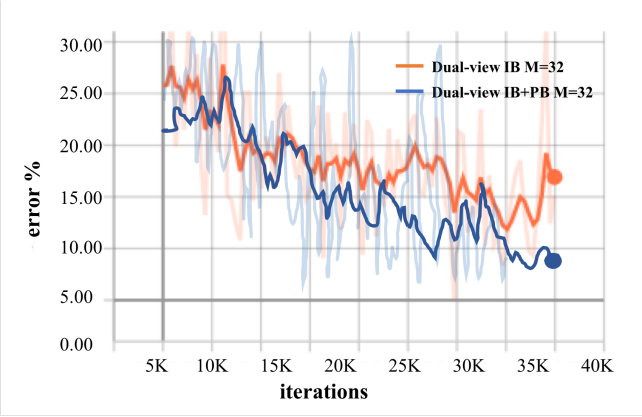

Figure 4: Impact of batch augmentation. We used the original (Orange) training regime with B = 32, and compared to PB data augmentation (Blue) with $M = 2$ creating an effective batch of $32 \cdot M$ (B = 64).

4.3 Experiments on FlickrLogos-32 Baseline

Experimental Setup Since the accurate estimation of mutual information is extremely expensive [35], we focus on relatively small and popular datasets, etc. FlickrLogos-32. It aims to uncover the difference between ResNet-50 and dual-view approaches for representation learning. FlickrLogos-32 is a publicly available and popular logo dataset with full annotations. It contains 32 different logo brands by downloading them from Flickr. The labeled set contains three different splits of train, validation, and test sets, the dataset offers 10 training data images, 30 validation images, and 30 test images.

Results on FlickrLogos-32 The classification accuracy of different methods is summarized in Table 4. We can see that our method achieves the best performance in both Top-1 and Top-5 accuracy. Compared with VIB, the proposed strategy obtains 1% improvement, which demonstrates that introducing a multi-view learning method can help improve performance. In order to illustrate the effectiveness of the joint representation extraction of the Dual-view IB method, we conduct additional visualizations for the dual-view FlickrLogos-32 dataset.

As shown in Fig. 5, we visualization of linear projection of the embedding obtained by applying the ResNet-50 and Dual-view IB encoder to the FlickrLogos-32 960 test images. The upper image shows the classification visualization effect of the baseline on the test set and the lower is our method. It can be clearly seen that Dual-view IB can learn a discriminative representation of the categories, obtaining relevant information and dropping irrelevant information, and learning nonlinear joint latent representation for logo classification.

5 CONCLUSIONS

In this paper, we have proposed the first cross-view classification method for logo images, to solve the problem of the cross-view misalignment of logo image varies under different viewpoints, large intra-class and small inter-classes variation of logo appearance. The

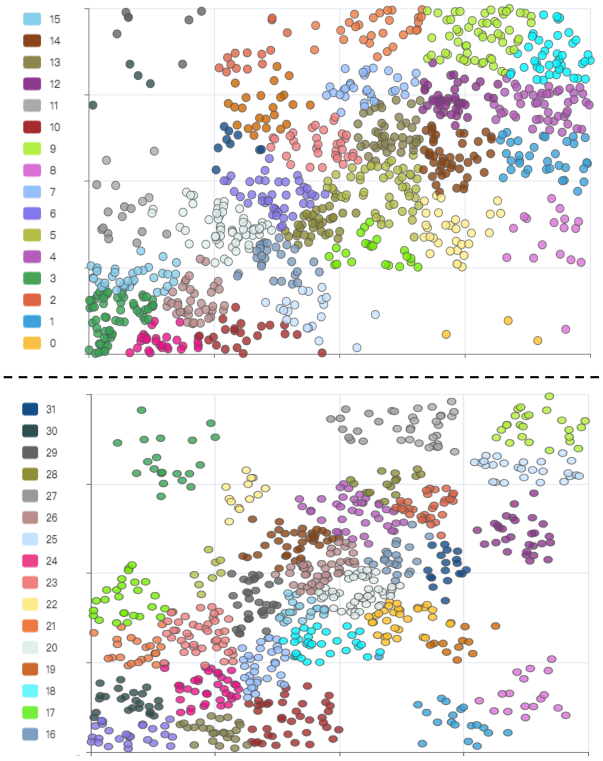

Figure 5: Linear projection of the embedding obtained by applying the ResNet-50 and Dual-view IB to the FlickrLogos-32 testset (960 images). Different colors are used to represent the 32 classes.

method learns quality-related features and eliminates irrelevant features to obtain more distinctive features of a logo category. In order to solve some categories unbalanced, the Pair Batch Data Augmentation (PB) is proposed to the Dual-view IB model to further improve the robust feature representation of data. We have shown its effectiveness on Logo-2K+ and the other two existing logo benchmarks. Future work includes the cross-modal representation learning of text and visual information in logo classification, and the multi-view IB representation learning will use for other tasks, such as person re-identification and image segmentation.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (81871508, 61773246, 62072289, 61702313, and 61972378), Major Program of Shandong Province Natural Science Foundation (ZR2019ZD04, and ZR2018ZB0419), Taishan Scholar Program of Shandong Province of China (TSHW201502038) and its 2nd round support, in part by Postdoctoral Science Foundation of China (2017M612338), in part by Shandong science and technology plan project (J17KB177).

REFERENCES

[1] Z.Hussain J.Dunnmon A.J.Ratner, H.Ehrenberg and C.Re. 2017. Learning to compose domain-specific transformations for data augmentation. *in Advances in neural information processing systems* (2017), 3236–3246.

[2] Joshua V Dillon Alexander A Alemi, Ian Fischer and Kevin Murphy. 2016. Deep variational information bottleneck. *In Proceedings of the 5th International Conference on Learning Representations*.

[3] Deepak Turaga Alhussein Fawzi, Horst Samulowitz and Pascal Frossard. 2016. Adaptive data augmentation for image classification. *In International Conference on Image Processing*, 3688–3692.

[4] Rana Ali Amjad and Bernhard Claus Geiger. 2019. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[5] Imma Boada Miquel Feixas Anton Bardera, Jaume Rigau and Mateu Sbert. 2009. Image segmentation using information bottleneck method. *IEEE Transactions on Image Processing*, 1601–1612.

[6] Brendan D. Tracey Artemy Kolchinsky and Steven Van Kuyk. 2019. Caveats for information bottleneck in deterministic scenarios. *in International Conference on Learning Representations*.

[7] Y. Bengio, S. Bengio, and J. Cloutier. 1991. Learning a synaptic learning rule. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*. 969–975.

[8] Y. Bengio, A. Courville, and P. Vincent. 2013. Representation Learning: A Review and New Perspectives. (2013), 1798–1832.

[9] Simone Bianco, Marco Buzzelli, Davide Mazzini, and Raimondo Schettini. 2015. Logo Recognition Using CNN Features. In *International Conference on Image Analysis and Processing*. 438–448.

[10] Simone Bianco, Marco Buzzelli, Davide Mazzini, and Raimondo Schettini. 2017. Deep learning for logo recognition. *Neurocomputing*. (2017), 23–30.

[11] R. Lienhart C. Eggert, A. Winschel. 2015. On the benefit of synthetic data for company logo detection. In *ACM Conference on Multimedia Conference*. 1283–1286.

[12] D. Tao C. Xu and C. Xu. 2014. Large-margin multi-viewinformation bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1559–1572.

[13] Jingying Chen, Maylor K. Leung, and Yongsheng Gao. 2003. Noisy logo recognition using line segment Hausdorff distance. *Pattern Recognition* (2003), 943–955.

[14] ZhiQi Cheng, Yang Liu, Xiao Wu, and Xian Sheng Hua. 2016. Video ECommerce: Towards Online Video Advertising. In *ACM International Conference on Multimedia*. 1365–1374.

[15] Z. Cheng, X. Wu, Y. Liu, and X. Hua. 2017. Video eCommerce++: Toward Large Scale Online Video Advertising. *IEEE Transactions on Multimedia* (2017), 1170–1183.

[16] Dandelion Mane Vijay Vasudevan Ekin D. Cubuk, Barret Zoph and Quoc V. Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

[17] Marco Federici, Anjan Dutta, Patrick Forre, Nate Kushman, and Zeynep Akata. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *International Conference on Learning Representations*.

[18] István Fehérvári and Srikar Appalaraju. 2019. Scalable Logo Recognition Using Proxies. In *IEEE Winter Conference on Applications of Computer Vision*. 715–725.

[19] Y. Gao, F. Wang, H. Luan, and T.-S. Chua. 2014. Brand data gathering from live social media streams. In *International Conference on Multimedia Retrieval*. 169–176.

[20] Ross Girshick. 2015. Fast R-CNN. In *IEEE International Conference on Computer Vision*. 1440–1448.

[21] R Girshick, J Donahue, T Darrell, and J Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 580–587.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[23] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2018. Bag of Tricks for Image Classification with Convolutional Neural Networks. *CoRR* abs/1812.01187 (2018).

[24] Steven C.H. Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. 2015. LOGO-Net: Large-scale Deep Logo Detection and Brand Recognition with Deep Region-based Convolutional Networks. *arXiv preprint arXiv:1511.02462* (2015).

[25] Sujuan Hou, Jianwei Lin, Shangbo Zhou, Maoling Qin, Weikuan Jia, and Yuanjie Zheng. 2017. Deep Hierarchical Representation from Classifying Logo-405. *Complexity* (2017), 1–12.

[26] Forrest N. Iandola, Anting Shen, Peter Gao, and Kurt Keutzer. 2015. DeepLogo: Hitting Logo Recognition with the Deep Neural Network Hammer. *arXiv preprint arXiv:1510.02131* (2015).

[27] S.Bazrafkan J.Lemley and P.Corcoran. 2017. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access* (2017), 5858–5869.

[28] S. Ren K. He, X. Zhang and J. Sun. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*. 346–361.

[29] Sham M Kakade and Dean P Foster. 2007. Multi-view regression via canonical correlation analysis. *In International Conference on Computational Learning Theory*, 82–96.

[30] Karen Livescu Kamalika Chaudhuri, Sham M Kakade and Karthik Sridharan. 2009. Multi-view clustering via canonical correlation analysis. *ACM ICML*, 129–136.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems* (2012), 1097–1105.

[32] Yuanyuan Li, Qiuyue Shi, Jiangfan Deng, and Su Fei. 2018. Graphic logo detection with deep region-based convolutional networks. In *Visual Communications and Image Processing*. 10–13.

[33] Liu Liu, Daria Dzyabura, and Natalie Mizik. 2018. Visual Listening In: Extracting Brand Image Portrayed on Social Media. In *AAAI Conference on Artificial Intelligence*. 71–77.

[34] Xu Lu, Lei Zhu, Zhiyong Cheng, Jingjing Li, Xiushan Nie, and Huaxzhang Zhang. 2019. Flexible Online Multi-modal Hashing for Large-scale Multimedia Retrieval. 1129–1137.

[35] David McAllester and Karl Stratos. 2018. Formal Limitations on the Measurement of Mutual Information. *arXiv* (2018), 7–12.

[36] Tao Mei, Xian-Sheng Hua, Linjun Yang, and Shipeng Li. 2007. VideoSense-Towards Effective Online Video Advertising. In *ACM International Conference on Multimedia*. 1075–1084.

[37] S. Motiian and G. Doretto. 2016. Information bottleneck domain adaptation with privileged information for visual recognition. *Proc. Eur. Conf. Comput. Vis*, 630–647.

[38] Jan Neumann, Hanan Samet, and Aya Soffer. 2002. Integration of local and global shape analysis for logo classification. *Pattern Recognition Letters*. (2002), 1449–1457.

[39] Julien Mairal Nikita Dvornik and Cordelia Schmid. 2018. Modeling Visual Context is Key to Augmenting Object Detection Datasets. *ECCV*, 2.

[40] G. Oliveira, X. Frazao, A. Pimentel, and B. Ribeiro. 2016. Automatic graphic logo detection via Fast Region-based Convolutional Networks. In *International Joint Conference on Neural Networks*. 985–991.

[41] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. 2011. Scalable logo recognition in real-world images. In *ACM International Conference on Multimedia Retrieval*. 18–20.

[42] R. Lienhart S. Romberg. 2013. Bundle min-hashing for logo recognition. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. 113–120.

[43] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[44] Hang Su, Shaogang Gong, and Xiatian Zhu. 2017. WebLogo-2M: Scalable Logo Detection by Deep Learning from the Web. In *IEEE International Conference on Computer Vision Workshop*. 270–279.

[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. *CoRR, abs/1409.4842* (2015), 7–12.

[46] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang, and Shuqiang Jiang. 2020. Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification. In *AAAI Conference on Artificial Intelligence*. 6194–6201.

[47] J. Li Y. Gao, S. Gu and Z. Liao. 2007. The multi-view information bottleneck clustering. *in Advancs in Databases: Concepts, Systems and Applications*, 912–917.

[48] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. 2018. Learning to Navigate for Fine-grained Classification. In *European Conference on Computer Vision*. 438–454.

[49] Haojie Li Tat-Seng Chua Yue Gao, Yi Zhen. 2016. Filtering of Brand-Related Microblogs Using Social-Smooth Multiview Embedding. *IEEE Transactions on Multimedia* (2016), 2115–2126.

[50] Xiao Zhang, Fuzhen Zhuang, Wenzhong Li, Haochao Ying, Hui Xiong, and Sanglu Lu. 2019. Inferring Mood Instability via Smartphone Sensing: A Multi-View Learning Approach. 1401–1409.