Application Notes

Lu Liu
School of Journalism and Communication, Fudan University, Shanghai, CHINA

Yong Luo
School of Computer Science and Engineering, Nanyang Technological University, SINGAPORE

Han Hu
School of Information and Electronics, Beijing Institute of Technology, Beijing, CHINA

Yonggang Wen
School of Computer Science and Engineering, Nanyang Technological University, SINGAPORE

Dacheng Tao
School of Computer Science, The University of Sydney, Sydney, AUSTRALIA

Xin Yao
Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, CHINA

xTML: A Unified Heterogeneous Transfer Metric Learning Framework for Multimedia Applications

Abstract

Owing to the continual growth of multimodal data (or feature spaces), we have seen a rising interest in multimedia applications (e.g., object classification and searching) over these heterogeneous data. However, the accuracy of classification and searching tasks is highly dependent on the distance estimation between data samples, and simple Euclidean (EU) distance has been proven to be inadequate. Previous research has focused on learning a robust distance metric to quantify the relationships among data samples. In this context, existing distance metric learning (DML) algorithms mainly leverage on label information in the target domain for model training and may fail when the label information is scarce. As an improvement, transfer metric learning (TML) approaches are proposed to leverage information from other related domains. However, current TML algorithms assume that different domains explore the same representation; thus, they are not applicable in heterogeneous settings where the data representations of different domains vary. In this research, we propose xTML, a novel unified heterogeneous transfer metric learning framework, to improve the distance estimation of the domains of interest (i.e., the target domains in classification and searching tasks) when limited label information, complementary with extensive unlabeled data, is provisioned for model training. We further illustrate how our proposed framework can be applied to a selected list of multimedia applications, including opinion mining, deception detection and online product searching. Qualitative and quantitative comparisons between our proposed algorithm and alternative solutions demonstrate the advantages of our proposed framework in multiple performance metrics.

©ISTOCKPHOTO.COM/NATASAADZIC

Digital Object Identifier 10.1109/MCI.2020.2976187
Date of current version: 10 April 2020

Corresponding Author: Yonggang Wen (Email: ygwen@ntu.edu.sg). X. Yao is also with the School of Computer Science, University of Birmingham, UK.

1556-603X/20©2020IEEE

I. Introduction

Multimodal data plays an important role in multimedia applications, for example, sentiment (or public opinion [1]) analysis, deception detection and online product (or service) search [2], [3]. The enormous amount of data to be analyzed in these applications often come from different modalities, channels and/or feature spaces. As an example, human sentiment can be interpreted from multiple feature domains, including body actions, facial expressions, voice intonation, speech contents, etc. In another example, a tweet often consists of an image with an associated text description. Phone call scams can also be screened based on the unique voice signatures and speech contents of the intruders. A product search system usually provides an image or video of the product on its designated webpage, which may also contain some text descriptions or tags, as well as related hyperlinks. Indeed, it has been recognized that classification and search applications over these multimodal data can potentially facilitate tacklings of a range of high-impact societal problems, such as lowering crime rates and improving customer experiences.

However, existing data analysis algorithms do not perform well for these multimedia applications. This is mainly attributed to the fact that the heterogeneous data have different physical meanings and/or statistical properties [4]. Existing algorithms for handling heterogeneous data mainly fall into the following two categories:

- ❏ Multiview (or Multimodal) Learning (MVL). The goal of MVL approaches [4]–[7] is to identify an appropriate combination of multiple heterogeneous representations for prediction.
- ❏ Heterogeneous Transfer Learning (HTL). The goal of HTL [8]–[10] is to improve the learning performance for the tasks/domains of interest (i.e., target domain) by applying knowledge/skills learned from other related tasks/domains (i.e., source domains), where the data representations vary between the source and target domains.

These two alternative approaches are proposed for different scenarios. The former usually assumes that both the training and test samples have established representations in all domains and that abundant label information exists for model training. In the latter case, the label information available for training is insufficient in the target domains due to high labeling cost [11], and inference is performed in each target domain, based on a target model trained with transferred information from the source domain(s).

Our proposed xTML framework, catered for multimodal data, is particularly suitable for multimedia applications, for example, sentiment analysis, deception detection and online product search, to name a few.

In this research, we focus on the HTL algorithms, which currently face some technical challenges. A frequently utilized strategy in the HTL method is to transform the heterogeneous features into a common subspace to reduce the differences between the heterogeneous domains [10]. This is then followed by a distance metric derived from the learned transformation. This strategy is effective in some cases. However, most of these algorithms are limited in that they can deal with two domains only (i.e., one source domain and one target domain). In reality, one would expect more than two domains in many real-world applications, for example, review documents may be written in more than two languages in multilingual sentiment classification [10]. Some approaches [9], [12] have been proposed to deal with scenarios involving more than two domains. Nevertheless, these approaches are not explicitly optimized with respect to particular distance metric. Moreover, they can only learn linear transformations across candidate domains, and thus may fail when the samples lie in highly nonlinear feature spaces (e.g., visual feature spaces). A comprehensive survey of heterogeneous transfer metric learning algorithms can be found in [13], which summarizes existing efforts, provides insightful discussions, and identifies several potential future directions.

As a sequel of our survey work [13], we propose in this article a unified mathematical framework, termed as xTML, to address the aforementioned challenges for the HTL algorithms. Our proposed xTML framework outperforms comparatively existing HTL approaches from multiple aspects:

- ❏ It optimizes explicitly with respect to the distance metric; thus, the results are more appropriate for distance estimation.
- ❏ It allows knowledge transfer across an arbitrary number of heterogeneous domains; hence, the resulted model is more flexible in practice.
- ❏ It aims to learn either linear or nonlinear target metrics; therefore, the model is appropriate for many applications, including those that involve challenging visual analyses.

Specifically, we investigate different knowledge-transfer strategies and validate the designed models for a list of selected multimedia applications. Our main contribution is a unified mathematical framework that can explicitly learn either linear or nonlinear target metrics, enabling the use of heterogeneous knowledge transferred from an arbitrary number of source domains to guide multimodal classification and multimedia search applications in the target domain(s). Moreover, we substantiate the proposed unified framework with a specific mathematical formulation and alternative optimization strategies that are easy and straightforward to follow.

The remainder of this article is organized as follows. In Section II, we present the proposed xTML framework in a structured manner, with its mathematical architecture, alternative transfer strategy designs and an unified optimization approach. In Section III, we apply the proposed xTML framework to three multimedia applications, namely, public opinion mining, deception detection and online product search. Qualitative comparisons between our proposed

framework and alternative solutions, and numerical results for the online product search application are presented in Section IV. Section V summarizes this work.

II. Unified xTML Framework

In this section, we present our proposed unified xTML framework for heterogeneous transfer metric learning, in a top-down manner.

A. Framework Overview

Fig. 1 illustrates the overall architecture of our proposed xTML framework for multimodal classification or multimedia searching. The proposed architecture consists of two parts:

- ❏ Offline metric learning, which uses limited label information and large amounts of unlabeled multimodal (or multimedia) data; and
- ❏ Online classification or search, which can predict the class label or find similar items for any test data in a real-time manner, based on the learned distance metric.

Our proposed xTML framework works as follows. In the offline metric learning stage, each unlabeled multimodal data point is represented in all M heterogeneous domains. The distance between any two instances $\mathbf{x}_{mi}$ and $\mathbf{x}_{mj}$ in the mth domain is given by the formula of $d_{\phi_m}(\mathbf{x}_{mi}, \mathbf{x}_{mj}) = \|\phi_m(\mathbf{x}_{mi}) - \phi_m(\mathbf{x}_{mj})\|_2^2$ for some learned function ϕ_m. We project the various representations of each unlabeled instance into a common subspace utilizing the set of mapping functions $\{\phi_1, \phi_2, \ldots, \phi_M\}$. In this way, knowledge derived from the limited label information can then be transferred across all domains, allowing them to enhance each other during the learning of the mapping functions and the distance metrics. This process leads to an optimal distance metric parameterized by the learned ϕ_m^* for the mth domain. In the online classification or search stage, the test data point is represented in the mth domain. By adopting the learned metric for existing classifications (such as k-nearest neighbor (kNN)) or retrieval (such as learning to rank [14]) models, appropriate class labels or the most closely related items can be obtained for the test instance. Notice that no target domain is specified in this model and the metrics are learned for all domains simultaneously. At the same time, this approach is also capable of supporting the case in which the representations of input test instances can be of arbitrary types. For example, when the target domain is specified, we propose to learn the target mapping by "pushing" the target subspace to be close to an integration of all the remaining subspaces. For example, we can learn the target mapping by minimizing the divergence between the induced representation of the unlabeled multimodal data point by the target mapping and a linear combination of the induced representations of the same data point by the other mappings.

Our proposed xTML framework extends the traditional distance metric learning [15], [16] in several aspects, including:

- ❏ *Heterogeneous Knowledge Transfer across Multiple Domains*. Knowledge can be transferred across an arbitrary number of heterogeneous domains, and the different domains enhance each other in metric learning. It follows that more reliable metrics can be obtained in this manner than by learning them separately, especially

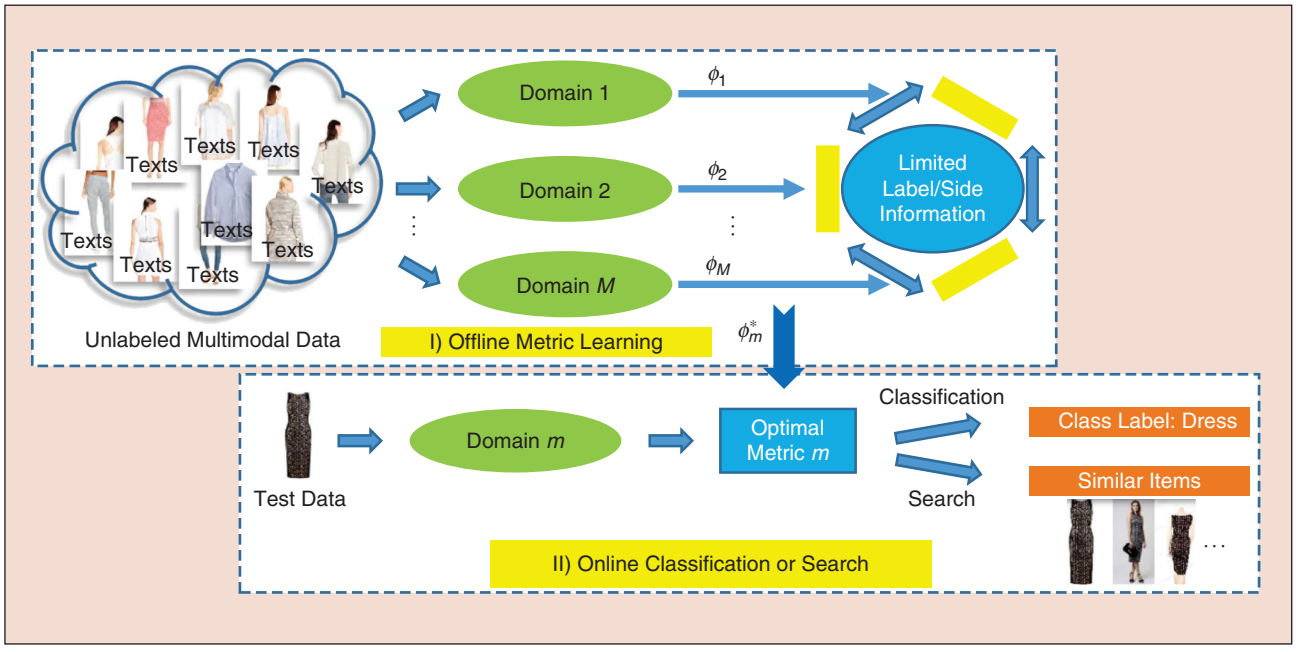

FIGURE 1 Architecture of our proposed xTML framework. The core of this framework is an offline metric learning module, in which arbitrary linear or nonlinear distance metrics are learned for an arbitrary number of heterogeneous domains simultaneously. Limited labeled data are provided in each domain, while large amounts of unlabeled multimodal data exist for establishing domain connections. Complex interactions between all the different domains can be exploited to derive an enhanced distance metric for each domain. Based on this enhanced metric, considerably better classification or search performances can be obtained in each domain.

when only limited label information is provided for the domains.

- *Mapping Function Flexibility*. The proposed framework can learn either linear or nonlinear mapping functions, in response to the varying data characteristics.

The novelty of our proposed xTML framework, compared to other existing heterogeneous transfer learning approaches, lies in the fact that arbitrary structures of the data distribution can be adapted into an arbitrary number of models, based on the mappings $\{\phi_m\}$. For example, it includes the nonlinear metric learning using the kernel trick in [20] as a special case by letting $\phi = U\psi$, where ψ is a mapping induced by a reproducing kernel based on the theory of reproducing kernel Hilbert spaces (RKHS) [21]. It also allows other nonlinear function learning techniques, such as gradient boosting regression tree (GBRT) [22] and deep learning [23], to be incorporated for nonlinear metric learning. In addition, since there are multiple domains included in the optimization, it is critical to exploit the relevance of different domains to avoid zero or negative transfer, where attention [24] and adversarial learning [25] mechanisms can be incorporated. For example, the attention scheme can be used to assign different weights to different source samples, allowing highly correlated source samples to contribute more to the target model training. Similarly, adversarial learning can be used to select source samples that are difficult to discriminate from the target samples. Such source samples are more useful during the target model training. In Fig. 2, we introduce a taxonomic approach for understanding how our proposed xTML stands out in the category of the HTL research. HTL algorithms can be characterized by the number of domains (two versus multiple) they address and the nature of their distance functions (linear versus nonlinear). Our framework exploits the nonlinear and complex interactions between an arbitrary number of domains, making it versatile for diverse multimedia applications.

B. Fundamental Theory

The key decision in our proposed xTML framework is how to conduct effective knowledge transfer across heterogeneous domains. Specifically, due to the differences in various instance spaces, the main challenges involve obtaining a common representation of heterogeneous domains, and demonstrating how that common representation preserves the correlation between different domains. These pressing challenges can be addressed under a generalized notion of Mahalanobis distance [26]. In distance metric learning (DML), it is common to learn the Mahalanobis distance, denoted as follows:

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T A(\mathbf{x}_i - \mathbf{x}_j), \quad (1)$$

where A is a metric parameter that can be factorized as $A = UU^T$ due to its positive semidefinite property. By applying this factorization, we can obtain $d_A(\mathbf{x}_i, \mathbf{x}_j) = \|U\mathbf{x}_i - U\mathbf{x}_j\|_2^2$. DML can be conducted in the feature space determined by a mapping ψ, i.e., $d_A(\mathbf{x}_i, \mathbf{x}_j) =$

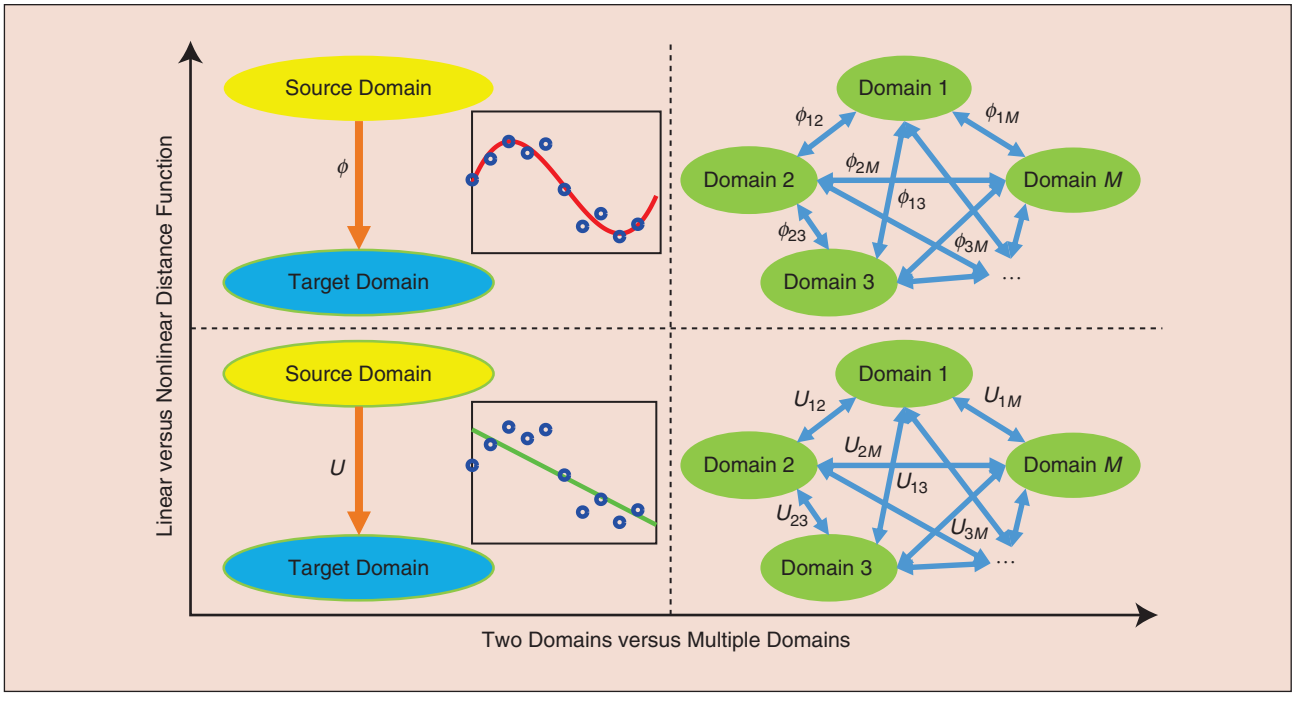

FIGURE 2 Illustration of the novelty of the proposed xTML framework. The existing HTL approaches consider only linear interactions between two domains [17], linear interactions between multiple domains [12], or nonlinear interaction between two domains [18], [19]. In comparison, our proposed xTML framework aims to exploit nonlinear and complex interactions between an arbitrary number of domains. In the two-domain case, the source domain has considerably more labeled data than that of the target domain; thus, the source domain is utilized to help the target metric learning. In the multiple-domain case, the labeled data are limited in all domains; hence, the different domains are reinforced by helping each other during metric learning.

$\|U\psi(\mathbf{x}_i) - U\psi(\mathbf{x}_j)\|_2^2$, enabling us to better exploit the structure of the data distribution. It follows that the distance can be further denoted as

$$d_\phi(\mathbf{x}_i, \mathbf{x}_j) = \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|_2^2, \qquad (2)$$

where $\phi(\cdot) = U\psi(\cdot)$ is an integrated mapping function that can be either linear or nonlinear. Leveraging this general form of distance measurement, we propose three alternative strategies to enable knowledge transfer, as presented in next subsection.

C. Computation Models for Knowledge Transfer

Using the generalized notion of distance measurement, we introduce three alternative computation paradigms (cf. Fig. 3) in our proposed xTML framework, to enable knowledge transfer across domains, as follows:

- *Representation-based Model*: All the M original representations $\{\mathbf{x}_{mn}^U\}_{n=1}^{N^U}$, $m = 1, 2, \ldots, M$ of the unlabeled instances are projected into a common subspace as $\{\mathbf{z}_{mn}^U\}_{n=1}^{N^U}$, $m = 1, 2, \ldots, M$. The transformed representations should be close to each other because they belong to the same instance. By maximizing the covariance (or equivalently minimizing the divergence) of the transformed representations, we can find a subspace for knowledge transfer.
- *Distance-based Model*: This model does not explicitly find a subspace but aims to minimize the distances between the mapped instances of different domains, i.e., $d_{mij}^U = \|\phi_m(\mathbf{x}_{mi}^U) - \phi_m(\mathbf{x}_{mj}^U)\|_2^2$, $m = 1, 2, \ldots, M$, so that they are close to each other. In this way, all the different domains help one another to narrow the hypothesis space of the distance metrics parameterized by $\{\phi_m\}$.
- *Kernel-based Model*: This model does not explicitly find a common subspace, but minimizes the divergence of the kernels $k_{mij}^U = (\phi_m(\mathbf{x}_{mi}^U))^T \phi_m(\mathbf{x}_{mj}^U)$, $m = 1, 2, \ldots, M$, induced by the mappings of different domains. This minimization also serves as a penalty for the metric hypothesis space of each domain.

In these paradigms, we treat all domains in an equal manner and conduct knowledge transfer across them. If the target domain is specified, the target mapping can be learned by minimizing the divergence between the target representation (resp. distance or kernel) and a chosen/learned integration of all the remaining representations (resp. distances or kernels).

In this article, we next develop a mathematical framework to quantify the effectiveness of these alternative computational models. A quantitative measurement of these computing paradigms in the xTML framework provides a solid foundation for the unified mathematical framework.

D. Unified Optimization Framework

The key benefit of our proposed xTML model is its ability to leverage knowledge (such as label information) from related domains for metric learning in a target domain. In the metric learning, the label information (domain expert knowledge) is often provided in a weakly supervised way, for instance, pair- or triplet-based constraints. For example, it is common to use the must-link/cannot-link constraint in classification to indicate whether a pair of instances $(\mathbf{x}_i, \mathbf{x}_j)$ are similar or dissimilar. In a search application, the label information

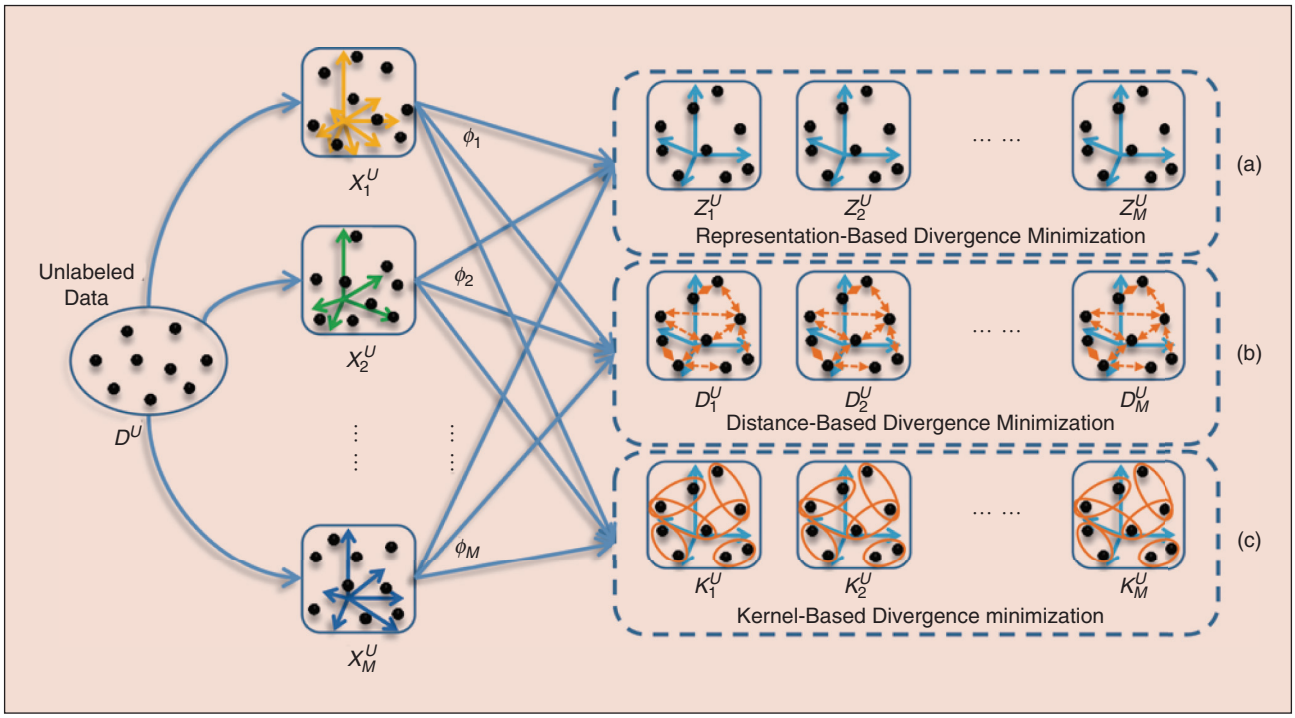

FIGURE 3 Three alternative computation paradigms: (a) representation-based model, where the knowledge transfer is conducted by maximizing the representation covariance of different domains in a common subspace; (b) distance-based model, where the knowledge transfer is conducted by minimizing the distances between corresponding sample pairs in different domains; and (c) kernel-based model, where the knowledge transfer is conducted by minimizing the divergence of the kernels between corresponding sample pairs in different domains.

is usually given by a relative constraint on a triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$, i.e., $\mathbf{x}_j$ should be more similar to $\mathbf{x}_i$ than $\mathbf{x}_k$. By introducing a loss function $g(\cdot)$, we can derive the conditional risk of a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ with the distance metric parameterized by ϕ as follows:

$$\Psi(\phi; \mathbf{x}_i, \mathbf{x}_j, y_{ij}) = g(y_{ij}[1 - d_\phi(\mathbf{x}_i, \mathbf{x}_j)]), \tag{3}$$

where y_{ij} indicates the similarity of the pair of interest. Similarly, the risk of a sample triplet is given by

$$\Psi(\phi; \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = g([d_\phi(\mathbf{x}_i, \mathbf{x}_k) - d_\phi(\mathbf{x}_i, \mathbf{x}_j)]). \tag{4}$$

Recently, angular loss [27] has been proposed to improve the robustness of metric learning algorithms, and the risk based on the angular loss is given by

$$\Psi(\phi; \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = g\left(\frac{d_\phi(\mathbf{x}_i, \mathbf{x}_j)}{2d_\phi(\mathbf{x}_k, \mathbf{x}_c)}\right), \tag{5}$$

where $\mathbf{x}_c = (\mathbf{x}_i + \mathbf{x}_j)/2$. Given M heterogeneous domains, we assume that the labeled training set $\mathcal{D}_m^L$ (of size N_m) is small in size for the mth domain but there exists a large unlabeled training set $\mathcal{D}^U$ (of size N^U), in which the data have representations in all domains. Such unlabeled data are usually easy to collect in practice [17]. It follows that the general risk minimization problem can be formulated as

$$\begin{aligned} \underset{\{\phi_m\}_{m=1}^M}{\operatorname{argmin}} \epsilon(\phi_m) = &\sum_{m=1}^{M} \Psi(\phi_m; \mathcal{D}_m^L) \\ &+ \gamma R(\phi_1, \phi_2, \ldots, \phi_M; \mathcal{D}^U), \\ \text{s.t. } &\phi_m \in \mathcal{H}(\phi_1, \phi_2, \ldots, \phi_M), \\ &m = 1, 2, \ldots, M, \end{aligned} \tag{6}$$

where $\Psi(\phi_m; \mathcal{D}_m^L)$ is the empirical risk w.r.t. ϕ_m on the training set in the mth domain, $R(\phi_1, \ldots, \phi_M; \mathcal{D}^U)$ is some regularizer to enforce information transfer across different domains, and $\mathcal{H}(\phi_1, \phi_2, \ldots, \phi_M)$ is the hypothesis space. Here, the regularizer is constructed primarily by using the unlabeled data as presented in the computation model section.

Notice that the optimization problem in (6) is a generic formulation, the following components should be specified for individual applications (scenarios):

- A *mapping function* of the specific task. This function can be either linear or nonlinear as determined by an assumed structure of the data distribution. In the linear case, the mapping function is usually given by a transformation matrix U. In the nonlinear case, we can choose $\phi(\cdot) = U\psi(\cdot)$. Alternatively, we may also directly choose ϕ to be a GBRT, a neural network, and so on.
- A chosen *computing paradigm*, for example, representation-based, distance-based or kernel-based model.
- A specific loss with respect to the distance metric associated with the application task, such as ranking based loss [28] for multimedia search.

Under this framework, our objective is to derive optimal algorithms to learn a reliable metric for the target domain that has the limited label information, and the ultimate goal is to improve the classification accuracy or search performance (e.g., mean average precision [29]) of the target application task.

The optimization problem (6) can be solved via well-established optimization algorithms. Specifically, we propose to adopt an iterative algorithm, where one parameter ϕ_m is updated in each iteration and all other $\{\phi_{m'}, m' \neq m\}$ are fixed. This iterative algorithm will be further coupled with a specific correlation maximization or divergence minimization strategy for the regularizer. We propose to adopt several alternative correlation maximization or divergence minimization strategies, for example,

1) high-order canonical correlation maximization [30],
2) Burg matrix divergence minimization [28],
3) Bregman divergence minimization [31],
4) log-determinant divergence minimization [32], or
5) Von Neumann divergence minimization [33].

The solutions to these special cases are expected to offer insights on how information is transferred across different domains.

III. Multimedia Applications

Our proposed xTML framework, catered for multimodal data, is particularly suitable for multimedia applications, for example, sentiment analysis, deception detection and online product search, to name a few. In these applications, the sample representations usually stem from different sources (e.g., image, audio, text, etc.), or exist in different feature spaces (e.g., color, shape, texture, etc.). To save human labeling effort, we adopt the proposed xTML technique to learn the optimal metric for each domain (representation) by utilizing the limited label information in each domain and transferring information between different domains. The final classification (for sentiment analysis and Internet fraud detection) or search is performed based on the distances calculated using the learned metrics. Details on how to apply our proposed xTML framework are presented as follows.

A. Opinion Mining

Opinion mining [1], or basic sentiment analysis, aims to classify the polarity (positive, negative or neutral) of private states, such as opinions, evaluations or sentiments. Advanced sentiment analysis classifies the states beyond mere polarity into emotional categories, e.g., happy, angry and sad. It follows that opinion mining has a variety of applications, including marketing, customer service, public opinion analysis [34] and analysis of political debates. For example, a merchant can understand market demands of products and plan their supplies accordingly by analyzing consumers' review comments. In another example, service quality can be improved by analyzing previous service recordings. Similarly, Netizens' opinions can be gleaned from microblogs (e.g., tweets) to acquire quantitative insights for government discourse on crisis management [1].

Most of the existing works on sentiment analysis are confined to textual data. However, due to the rapid growth in Internet usage and social networks, online users (i.e., "netizens") tend to express their opinions on social media platforms in multimodal formats, including audio,

It is quite popular to shop for products that are the same or similar to those shown on TV programs. We have developed a prototype system to meet this eminent market need.

images, and videos. In addition, customer service is typically provided by voice calls. Therefore, it is crucial to develop a system that can classify sentiments from diverse modalities.

In Fig. 4, we propose to adopt appropriate computation paradigms and learn enhanced distance metrics for different modalities, by using a small number of labeled reviews in each modality in conjunction with large numbers of unlabeled multimodal reviews. The objective is to accurately predict the polarity of a person's state in terms of any modality.

B. Deception Detection

Deception detection intends to judge whether an individual is lying. The decision can be made according to either a person's physiological responses or nonphysiological traits. In the deception detection literature, the most commonly adopted approach is the polygraph test, performed by asking the subject to answer some designed questions and recording their physiological indices (such as heart rate, blood pressure, etc.) for analysis. However, polygraph results are usually not admitted as evidence in most courts because the technique has significant error rates, and it is even possible to trick the system by well-trained subjects. Some other issues with polygraph tests include that they can be performed only in the presence of the subject, and they often require the recording devices to be in direct physical contact with the subject.

Recently, an increasing number of works have focused on utilizing nonphysiological traits to detect deception. For example, unique linguistic patterns can be found in the verbal and written output of liars to discriminate them from people telling the truth. Voice stress analysis (VSA), which studies voice patterns and speech fluctuations, is also useful for deception detection and has been claimed to be better than the polygraph test. Another promising trait for noninvasive lie detection is facial microexpressions [35].

In this study, we focus on deception detection via nonphysiological features (cf. Fig. 5). Using the proposed xTML framework, we can design algorithms to learn reliable metrics for different modalities, such as surrounding texts, speech texts, vocal sounds, and images or videos, where the objective is to minimize the deception detection error while allowing the input data consisting of arbitrary types.

C. Online Product Search

It is quite popular to shop for products that are the same or similar to those shown on TV programs. We have developed a prototype system to meet this eminent market need. The goal of the system is to create an effortless TV-to-Online (T2O) experience [2], [3]. An essential module in this system is a product search function.

In the product search application, each data point may be associated with multiple modalities. For example, it is common to use an image together with surrounding texts or tags to describe a product. There may also be affiliated hyperlinks, and it is often necessary to extract various types of features to represent the image.

In this application, we adopt a ranking-based loss in the proposed xTML framework. Hence, the learned metrics are particularly suitable for product search (cf. Fig. 6). In addition, tensor-based correlation maximization [30] is introduced to explore the high-order statistics between all domains. It follows that our xTML approach encodes more correlation information in the learned metrics than does the traditional pairwise correlation maximization. As a

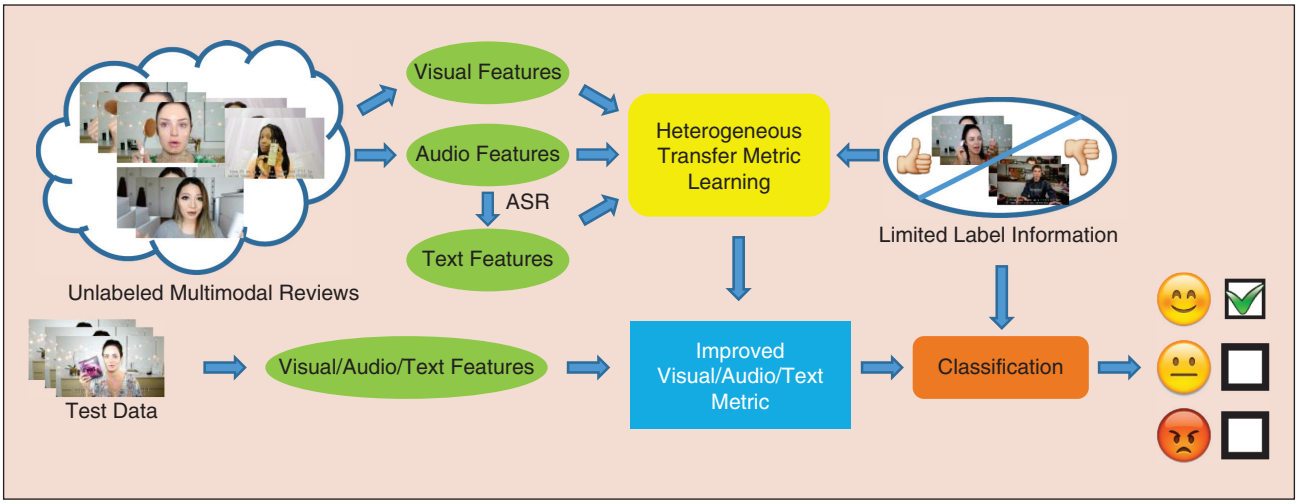

FIGURE 4 xTML application in sentiment analysis and opinion mining. Different types of features, including visual, audio and text features, are extracted from the review videos. Large amounts of unlabeled multimodal reviews together with limited labeled reviews are utilized to learn an improved metric for each feature type, allowing the polarity of a person's state to be accurately inferred.

result, the xTML approach is expected to achieve better performance in product search accuracy, as presented in the next section.

IV. Comparative Studies

In this section, we first present a qualitative analysis of our proposed framework, compared against a few existing solutions for each application. The advantages of our proposed xTML framework are further verified by a numerical analysis of the product search application.

A. Qualitative Analysis

We qualitatively compare our proposed xTML framework with four alternative solutions, list as follows:

- EU: a method to directly calculate the Euclidean distance between samples.
- RAML [28] and FRML [36]: two alternative ranking-based distance metric learning algorithms without knowledge transfer.
- MTDA [12]: a competitive heterogeneous multitask learning approach.
- RHMTML [2]: a recently proposed heterogeneous multitask metric learning approach for retrieval.

We further introduce a comparison framework for evaluating alternative solutions in five aspects, namely:

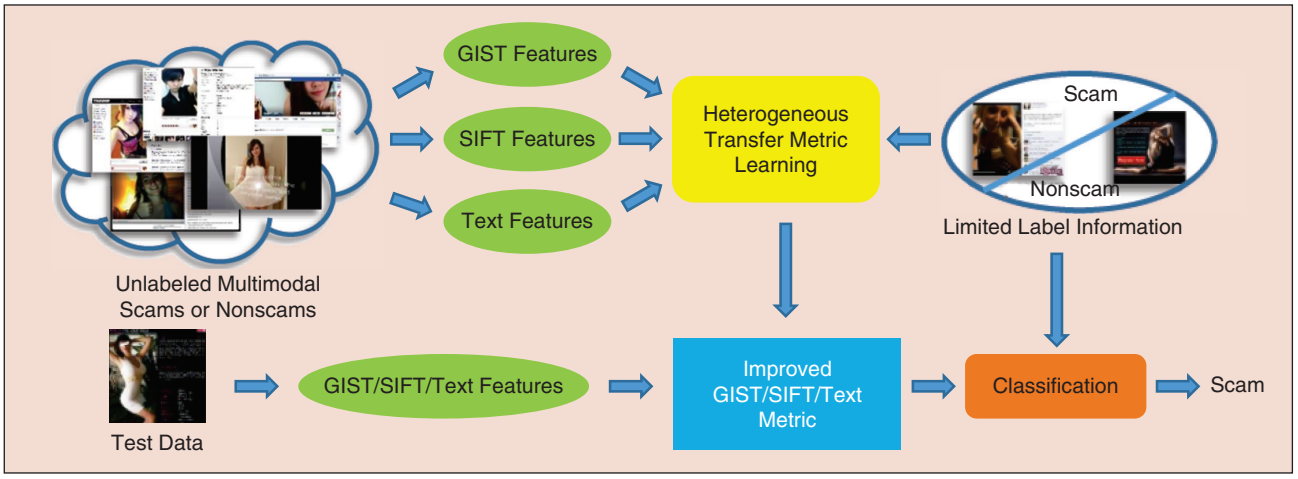

FIGURE 5 xTML application in deception detection. Different types of features, such as GIST, SIFT and text features, are extracted to represent the scam or nonscam data. By utilizing both the limited labeled data and abundant multimodal unlabeled data, improved distance metrics can be learned for all the different features simultaneously, after which satisfactory scam detection results can be obtained.

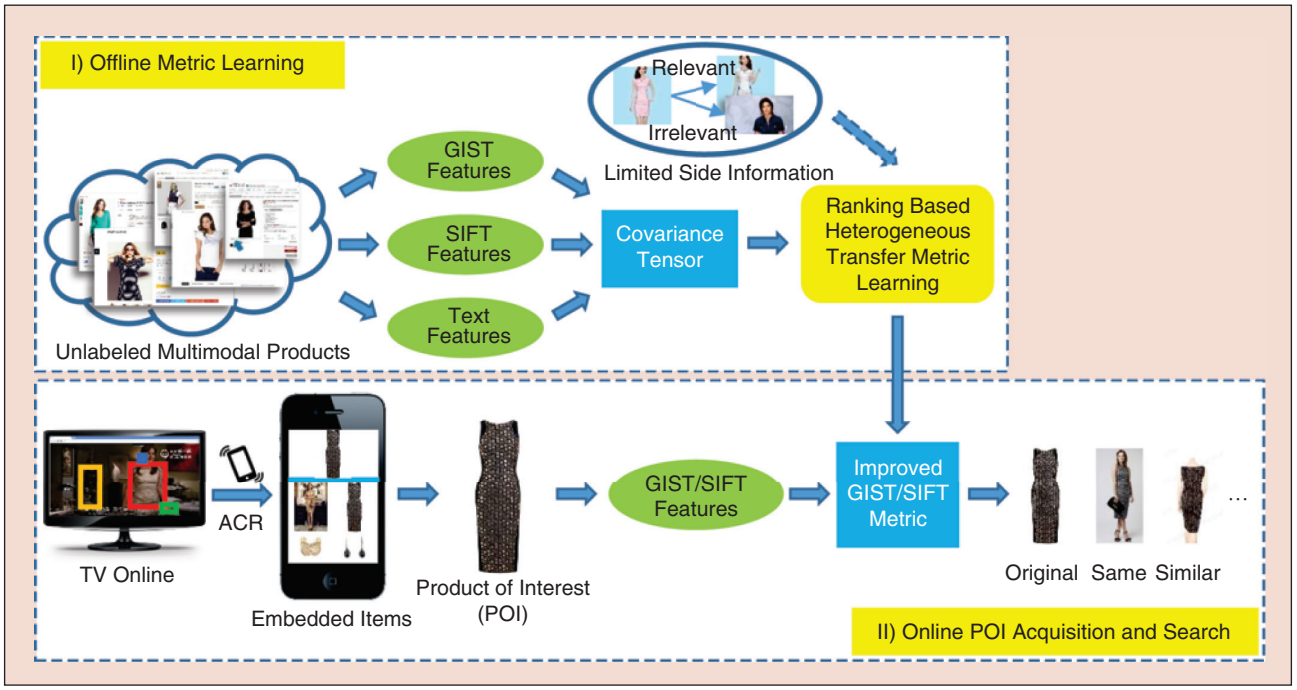

FIGURE 6 xTML application for online product search, which is a basic service in the emerging TV-to-online paradigm. Different types of features are extracted to represent the products. Large amounts of unlabeled multimodal data and limited labeled data are utilized to learn an improved distance metric for each type of feature, and ranking-based loss is adopted to make the learned metric particularly suitable for search.

Nevertheless, our method needs far less labeled samples to achieve satisfactory performance, and its performance is much better, given equal numbers of labeled samples.

1) the presence of metric learning;
2) the presence of knowledge transfer;
3) the number of labelled samples required to achieve satisfactory performance;
4) the time complexity;
5) the algorithms' performances in terms of MAP [29], AUC [37], etc.

Table 1 summarizes our comparison results. In particular, RAML and FRML learn metrics for different domains separately without knowledge transfer. MTDA aims to learn discriminative feature transformations, each of which can be used to derive a metric. Since the high-order correlation maximization paradigm is utilized, the time complexity of our xTML method is higher than those of other solutions. Nevertheless, our method needs far less labeled samples to achieve satisfactory performance, and its performance is much better, given equal numbers of labeled samples. Therefore, our proposed framework is noteworthy from several aspects, which we believe will result in better performance, as verified in the next subsection.

B. Quantitative Analysis

In this research, we have also conducted an initial evaluation of the proposed xTML framework for the online product search application. Specifically, we choose the representation-based computing paradigm, the ranking-based loss, and the high-order canonical correlation maximization strategy. Our experiments are conducted using 10 product categories (such as book and computer) from the NUS-WIDE dataset [38]. Based on this selection, each subset consists of 23,539 images with associated tags. Three different types of features, i.e., the bag of SIFT visual words (local visual), wavelet texture (global visual) and tags (textual) are regarded as the heterogeneous domains, and we evaluate the performance using

TABLE 1 **Comparisons of different solutions (d_m and N_m are the feature dimension and the number of labeled samples of the mth domain, and r and r' are the resulting and some intermediate feature dimension, respectively. k_1, k_2 and k_3 are constants, $d' = \prod_{m=1}^{M} d_m$ and $\bar{d}_m$ is the average feature dimension of all domains).**

SOLUTIONS	METRIC LEARNING	KNOWLEDGE TRANSFER	LABEL REQUIREMENTS	COMPLEXITY	PERFORMANCE
EU	NO	NO	NONE	$O(0)$	LOW
RAML, FRML	YES	NO	LARGE	$O\left(k_1 \sum_{m=1}^{M} d_m^2\right), O\left(r^2 \sum_{m=1}^{M} d_m\right)$	MEDIUM
MTDA	DERIVE	YES	MODERATE SMALL	$O\left(\sum_{m=1}^{M} (r'(r^2 + d_m^2) + d_m^3)\right)$	MODERATE HIGH
RHMTML	YES	YES	SMALL	$O(k_2 M[rd' + k_3(r\bar{d}_m N_m^2 + r\bar{d}_m^2)] + N^U d')$	HIGH
xTML	YES	YES	SMALLEST	$O(k_2 M[rd' + k_3(r\bar{d}_m (N_m^2 + N^U) + r\bar{d}_m^2)] + N^U d')$	HIGHEST

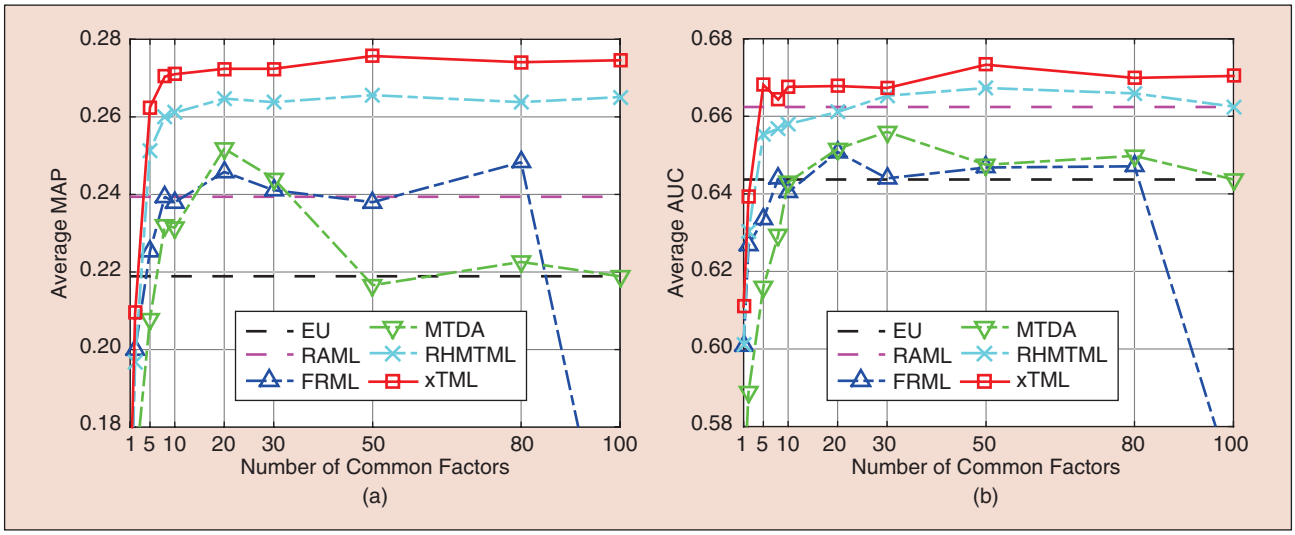

FIGURE 7 Average MAP and AUC of all domains versus the number of common factors on the NUS dataset, where the number of labeled samples for each concept is 10. Learning the metric can significantly improve the performance, and the proposed method consistently outperforms other alternative approaches.

two popular criteria: the mean average precision (MAP) [29] and the area under the ROC curve (AUC) [37].

Moreover, we have compared our proposed xTML framework to the Euclidean (EU) baseline, two popular ranking-based DML algorithms (RAML [28] and FRML [36]), and two competitive heterogeneous transfer learning approaches (MTDA [12] and RHMTML [2]). The experimental results are illustrated in Fig. 7. From the results, we make the following observations:

1) All the metric learning algorithms outperform the EU baseline significantly. This result demonstrates the effectiveness of metric learning in this application.
2) The transfer-approach MTDA is comparable to but sometimes slightly worse than are the single-domain DML algorithms without knowledge transfer. This result can be attributed to the fact that MTDA is mainly designed for classification.
3) Our proposed xTML model is superior to all other approaches in most cases and its performance is much more stable.

This numerical evaluation validates the effectiveness of our proposed computing paradigm and knowledge transfer strategy.

Finally, we compare the computational complexity of all the competing solutions in term of the training time, in Fig. 8. The EU baseline is not included because it does not require training. The results show that the time costs of the transfer approaches are higher than those of DML algorithms without transfer. Although our proposed method has the highest training time cost, some parallel computing techniques could be adopted to accelerate the training process. It should be noted that the prediction time costs of all these approaches are comparable because only the learned metric or mapping is utilized for final inference.

Our method is capable of handling an arbitrary number of heterogeneous domains and exploiting the arbitrary structures of the data distribution. In our proposed framework, large amounts of unlabeled data are utilized to bridge different domains, and we provide three alternative computing paradigms for knowledge transfer.

V. Conclusion

In this article, we proposed the xTML framework, a unified heterogeneous transfer metric learning approach, for multimodal classification and multimedia search in diverse applications, including sentiment analysis, deception detection and online product search. Compared with the existing heterogeneous transfer learning approaches, our method is capable of handling an arbitrary number of heterogeneous domains and exploiting the arbitrary structures of the data distribution. In our proposed framework, large amounts of unlabeled data are utilized to bridge different domains, and we provide three alternative computing paradigms for knowledge transfer. We also validate the effectiveness of the representation-based paradigm and correlation maximization strategy in an online product search application and report quantitative results. Moving forward, we plan to validate other paradigms and strategies and provide theoretical analyses of multiple real-world applications for multimedia practice.

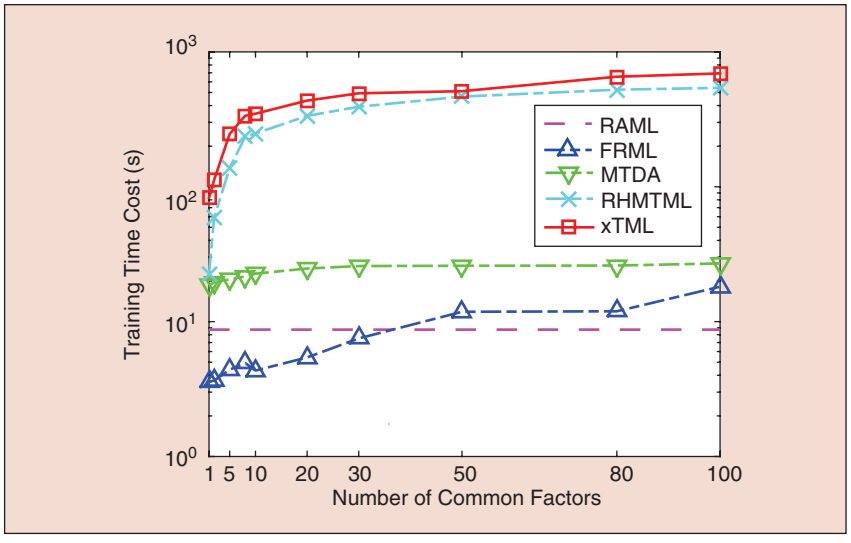

FIGURE 8 Computational time of different approaches on the NUS dataset. The time costs of the transfer approaches are higher than those of the DML algorithms without transfer, and the proposed method has the highest training cost.

Acknowledgment

This research is supported in part by Singapore NRF2015ENC-GDCR01001-003, administrated via IMDA, NRF2015ENC-GBICRD001-012, administrated via BCA, Youth Program of the National Social Science Fund of China under No.16CXW008, and National Natural Science Foundation of China (NSFC) under No. 61971457.

References

[1] L. Liu, *Government Discourse in Risk Society: Problems and Countermeasures*. China Radio International Publishing House, 2017.

[2] Y. Luo, Y. Wen, D. Tao, and Q. Fu, "Toward effortless TV-to-Online (T2O) experience: A novel metric learning approach," in *Proc. IEEE Global Communications Conf.*, Washington, D.C., Dec. 4–8, 2016. doi: 10.1109/GLOCOM.2016.7842335.

[3] Q. Fu, Y. Luo, Y. Wen, D. Tao, Y. Li, and L.-Y. Duan, "Toward intelligent product retrieval for TV-to-Online (T2O) application: A transfer metric learning approach," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2114–2125, Aug. 2018. doi: 10.1109/TMM.2018.2791803.

[4] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013. doi: 10.1109/TNNLS.2013.2238682.

[5] P. Xie and E. Xing, "Multi-modal distance metric learning," in *Proc. Int. Joint Conf. Artificial Intelligence*, Beijing, Aug. 3–9, 2013, pp. 1806–1812.

[6] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2355–2368, Aug. 2015. doi: 10.1109/TIP.2015.2421309.
[7] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 414–427, Jan. 2016. doi: 10.1109/TIP.2015.2495116.
[8] X. Shi, Q. Liu, W. Fan, S. Y. Philip, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *Proc. IEEE Int. Conf. Data Mining*, Sydney, Australia, Dec. 14–17, 2010, pp. 1049–1054. doi: 10.1109/ICDM.2010.65.
[9] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *Proc. Int. Joint Conf. Artificial Intelligence*, Barcelona, Spain, July 16–22, 2011, pp. 1541–1546. doi: 10.5591/978-1-57735-516-8/IJCAI11-259.
[10] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, "Heterogeneous domain adaptation for multiple classes," in *Proc. Int. Conf. Artificial Intelligence and Statistics*, Reykjavik, Iceland, Apr. 22–25, 2014, pp. 1095–1103.
[11] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sept. 2014. doi: 10.1109/TIP.2014.2332398.
[12] Y. Zhang and D.-Y. Yeung, "Multi-task learning in heterogeneous feature spaces," in *Proc. AAAI Conf. Artificial Intelligence*, San Francisco, Aug. 7–11, 2011, pp. 574–579. doi: 10.5555/2900423.2900515.
[13] Y. Luo, Y. Wen, L. Duan, and D. Tao, Transfer metric learning: Algorithms, applications and outlooks. 2018. [Online]. Available: arXiv:1810.03944
[14] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inform. Retrieval*, vol. 3, no. 3, pp. 225–331, Mar. 2009. doi: 10.1561/1500000016.
[15] E. P. Xing, M. I. Jordan, S. Russell, and A. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Advances in Neural Information Processing Systems*, 2002, pp. 505–512.
[16] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Advances in Neural Information Processing Systems*, 2005, pp. 1473–1480.
[17] G.-J. Qi, C. C. Aggarwal, and T. S. Huang, "Transfer learning of distance metrics by cross-domain metric sampling across heterogeneous spaces," in *Proc. SIAM Int. Conf. Data Mining*, Anaheim, CA, Apr. 26–28, 2012, pp. 528–539. doi: 10.1137/1.9781611972825.46.
[18] Y. Luo, Y. Wen, T. Liu, and D. Tao, "General heterogeneous transfer distance metric learning via knowledge fragments transfer," in *Proc. Int. Joint Conf. Artificial Intelligence*, Melbourne, Australia, Aug. 19–25, 2017, pp. 2450–2456. doi: 10.24963/ijcai.2017/341.
[19] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 1013–1026, Apr. 2019. doi: 10.1109/TPAMI.2018.2824309.
[20] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, "A new kernelization framework for mahalanobis distance learning algorithms," *Neurocomputing*, vol. 73, no. 10–12, pp. 1570–1579, June 2010. doi: 10.1016/j.neucom.2009.11.037.
[21] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2001.
[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001. doi: 10.1214/aos/1013203451.
[23] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, June 27–30, 2016, pp. 4004–4012. doi: 10.1109/CVPR.2016.434.
[24] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Machine Learning*, Lille, France, July 6–11, 2015, pp. 2048–2057.
[25] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Dec. 8–13, 2014, pp. 2672–2680.
[26] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, July 2013. doi: 10.1561/2200000019.
[27] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proc. IEEE Int. Conf. Computer Vision*, Venice, Italy, Oct. 22–29, 2017, pp. 2593–2601. doi: 10.1109/ICCV.2017.283.
[28] J.-E. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Anchorage, AK, June 24–26, 2008, pp. 1–8. doi: 10.1109/CVPR.2008.4587389.
[29] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *Proc. ACM SIGIR Conf. Research and Development Information Retrieval*, Amsterdam, The Netherlands, July 23–27, 2007, pp. 271–278. doi: 10.1145/1277741.1277790.
[30] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3111–3124, Nov. 2015. doi: 10.1109/TKDE.2015.2445757.
[31] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, July 2010. doi: 10.1109/TKDE.2009.126.
[32] J. Mei, M. Liu, H. R. Karimi, and H. Gao, "LogDet divergence based metric learning using triplet labels," in *Proc. ICML Workshop on Divergences and Divergence Learning*, Atlanta, GA, June 16–21, 2013.
[33] P. Yang, K. Huang, and C.-L. Liu, "Geometry preserving multi-task metric learning," *Mach. Learn.*, vol. 92, no. 1, pp. 133–175, July 2013. doi: 10.1007/s10994-013-5379-y.
[34] L. Liu, "Preliminary exploration of mechanism of action of microblog in public opinion on emergencies," *J. Bimon.*, vol. 117, no. 2, pp. 55–59, Mar. 2013.
[35] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Computer Vision*, Barcelona, Spain, Nov. 6–13, 2011, pp. 1449–1456. doi: 10.1109/ICCV.2011.6126401.
[36] D. Lim and G. Lanckriet, "Efficient learning of Mahalanobis metrics for ranking," in *Proc. Int. Conf. Machine Learning*, Beijing, June 21–26, 2014, pp. 1980–1988.
[37] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. Int. Conf. Machine Learning*, Bonn, Germany, Aug. 7–11, 2005, pp. 377–384. doi: 10.1145/1102351.1102399.
[38] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of Singapore," in *Proc. ACM Int. Conf. on Image and Video Retrieval*, Santorini Island, Greece, July 8–10, 2009. doi: 10.1145/1646396.1646452.

Digital Object Identifier 10.1109/MCI.2020.2976211