mathematics

MDPI

Article

A Novel Ensemble Strategy Based on Determinantal Point Processes for Transfer Learning

Ying Lv [1], Bofeng Zhang [2,3,*], Xiaodong Yue [1] and Zhikang Xu [1]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[2] School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China
[3] School of Computer Science and Technology, Kashi University, Kashi 844000, China
* Correspondence: bfzhang@shu.edu.cn or bfzhang@sspu.edu.cn

Abstract: Transfer learning (TL) hopes to train a model for target domain tasks by using knowledge from different but related source domains. Most TL methods focus more on improving the predictive performance of the single model across domains. Since domain differences cannot be avoided, the knowledge from the source domain to obtain the target domain is limited. Therefore, the transfer model has to predict out-of-distribution (OOD) data in the target domain. However, the prediction of the single model is unstable when dealing with the OOD data, which can easily cause negative transfer. To solve this problem, we propose a parallel ensemble strategy based on Determinantal Point Processes (DPP) for transfer learning. In this strategy, we first proposed an improved DPP sampling to generate training subsets with higher transferability and diversity. Second, we use the subsets to train the base models. Finally, the base models are fused using the adaptability of subsets. To validate the effectiveness of the ensemble strategy, we couple the ensemble strategy into traditional TL models and deep TL models and evaluate the transfer performance models on text and image data sets. The experiment results show that our proposed ensemble strategy can significantly improve the performance of the transfer model.

Keywords: transfer learning; ensemble strategy; determinantal point processes; domain adaptation

MSC: 68T01

Citation: Lv, Y.; Zhang, B.; Yue, X.; Xu, Z. A Novel Ensemble Strategy Based on Determinantal Point Processes for Transfer Learning. *Mathematics* **2022**, *10*, 4409. https://doi.org/10.3390/math10234409

Academic Editor: Liangxiao Jiang

Received: 17 October 2022
Accepted: 21 November 2022
Published: 23 November 2022

1. Introduction

Although traditional machine learning has a performance advantage in applications with rich annotation data and has been successfully applied in many applications [1,2], its performance is limited in applications with little annotation data. In addition, traditional machine learning is limited by the assumption that the training and test data obey independent identical distributions. This assumption does not hold in many real-world applications [3,4], such as medical image analysis, autonomous driving, etc.

A better solution to the above problem is transfer learning (TL), which aims to borrow knowledge from different but related source domains to train a transfer model for the target task [3,5]. It tries to achieve models with the ability to learn by analogy, e.g., the skill of learning to ride a bicycle can be used to learn to ride a motorcycle, and the ability to learn to play the piano can be used to learn to play the guitar. Existing TL methods can be divided into four types, namely instance-based [6,7], feature-based [8,9], model-based [10,11], and deep learning-based [12,13].

Most TL approaches focus on reducing inter-domain differences to extract more useful shared knowledge for target domain task and thus increase the transferability of the model. However, they ignore the limitations of using single transfer model. Because the differences between the source and target domains cannot be completely avoided, the shared knowledge extracted from the source domain is limited. The single transfer model trained based on the knowledge is often unstable when dealing with out-of-distribution

(OOD) data. This may lead to negative transfer. In addition, ensemble learning has been successfully applied to traditional machine learning models to improve the robustness of the models, but has received less attention in transfer learning.

For transfer learning tasks, in addition to considering the diversity of instances, the transferability of the instance is more important due to the distribution difference between source and target domains. The existing submodular function methods [14,15] do not guarantee that the selected subsets are suitable for the target domain, which can lead to negative transfer of the trained model.

To solve these problems, we propose a novel parallel ensemble strategy based on improved Determinantal Point Processes (DPP). The DPP provides a class of precise probabilistic models for sample selection problems [16,17]. In the strategy, we first measure the transferability of instances in source domain based on evidence theory and use the transferability to rewrite the correlation matrix of DPP. It makes the training subset obtained by DPP sampling highly adaptable with respect to the target domain. Second, the base transfer models are trained based on the subsets. Finally, we calculate weights using the adaptability of subsets to ensemble the base models. In addition, the proposed ensemble strategy is independent of the transfer algorithm, and it can be used as a general ensemble strategy. The contributions are summarized as follows.

- Designing an ensemble strategy with Determinantal Point Processes, which enhance transfer performance and stability of models in cross domain.
- Extending the DPP with transferability and diversity to make it suitable for transfer learning.
- The ensemble strategy can be seen as a generic technique, which can be applied to different transfer algorithms.

The paper is organized as follows. Section 2 introduces related work. Section 3 introduces the ensemble strategy based on DPP. Section 4 presents the experimental results to validate that the proposed ensemble strategy is effective to improve the performances of multiple kinds of transfer learning methods. The conclusion is given in Section 5.

2. Related Work

According to a literature survey [2,3,18], most previous transfer learning (TL) methods can be organized into instance-based methods, feature-based methods, classifier-based methods, and deep learning-based methods.

In instance-based methods, most methods aim to estimate the instance weight by feature distribution matching across different domains. Jiang and Zhai [19] proposed an intuitive instance weighted method, which calculates the distribution difference between source and target instances by four parameters. Dai et al. [6] proposed a TrAdaBoost to tune instance weights based on a Boosting algorithm. In [20,21], the authors utilize the kernel mean matching (KMM) to calculate the weight for reducing the difference between source domain and target domain. Long et al. [22] proposed the Transfer Joint Matching (TJM) method by minimizing the maximum mean discrepancy (MMD). Yan et al. [23] proposed a weighted maximum mean discrepancy (WMMD) for transfer learning.

In feature-based methods, a feature transformation strategy is often adopted in transfer learning. It transforms each original feature into a new feature representation for transfer learning. The objective is to learn a new feature representation with some distribution matching metrics between source and target domains. Pan and Yang [18] firstly introduced MMD (maximum mean discrepancy) to design a transfer method called the transfer component analysis (TCA). Further improving the MMD, Long et al. [24] designed joint distribution adaptation (JDA), which measures the difference of joint distribution between domains. Gong et al. [25] proposed a geodesic flow kernel (GFK) based on manifold learning. Sun et al. [26] proposed a correlation alignment (CORAL) for transfer learning.

The classifier-based methods focus on classifier adaptation. Yang et al. [27] modified the support vector machine so that it can be adapted to the transfer learning task. Duan et al. [28] proposed a multiple kernel learning (MKL) for transfer learning. Based on

MKL, Duan et al. proposed the DTSVM [29] and DTMKL [10]. Long et al. and Cao et al. designed ARTL [30] and DMM [31] by manifold regularization in model training.

In deep learning-based method, the fine-tuning pre-trained strategy has become a common strategy in transfer learning [12,32]. It utilizes the well-trained models (CNN, transform model and Bert model) on a large dataset (e.g., ImageNet) as the base and uses target domain to fine-tune weights. In addition, there are also some deep migration models designed based on adding constraints on domain differences in the loss function, such as coupled approximation neural network [33], joint adaptation network [34], manifold embedded distribution alignment [35], normalized squares maximization [36], and multi-representation adaptation network [37].

3. Ensemble Strategy Based on DPP Sampling

In this section, we first improve the DPP sampling and apply sampling with DPP to generate subsets from source domain. Second, we train the base model for the target domain based on the selecting of the subsets. Finally, we combine the models using weight, in which the weight of the base model is computed by the adaptability of subsets. For clarity, the notations are summarized in Table 1.

Table 1. Notations used in this paper.

Notation	Description
D^s, D^t	source domain, target domain
x^s, x^t	instance of source/target domain
Φ^t	evidence set
Φ^t_k	evidence subset
Ω	label spcae
$\oplus$	Dempster's combinational rule
K	kernel function
$\phi(\cdot)$	feature mapping function
$m(\cdot)$	mass function
$d(\cdot)$	distance function
L	correlation matrix

3.1. Subset Generation Based on DPP Sampling

As shown in Figure 1, the key step of the ensemble strategy is to generate the subsets with high transferability and diversity from the source domain using DPP sampling. DPP provides a class of precise probabilistic models for sample selection problems. In DPP, the probabilities of sampling are computed by correlation the matrix of items. In transfer learning, the performance of the ensemble model is determined by high transferability and diversity of subsets. To ensure that DPP sampling can increase the transferability and diversity of subsets, we reformulate the correlation matrix with the measure of samples transferability and diversity based on evidence theory.

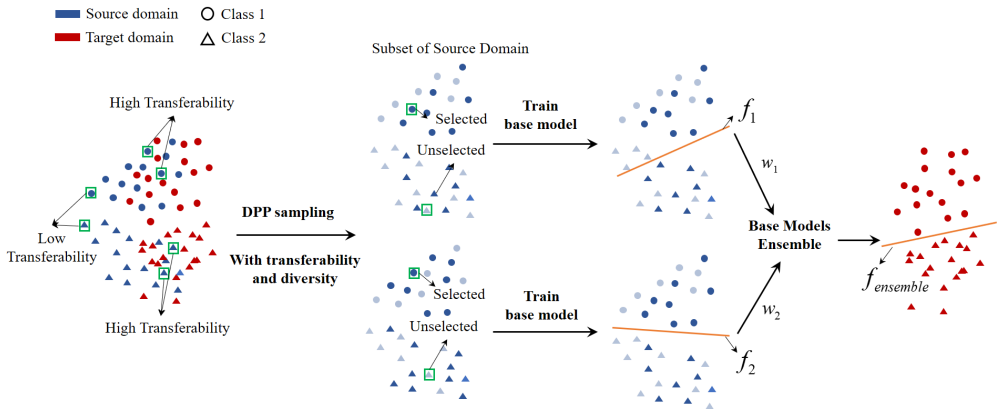

Figure 1. Ensemble strategybBased on DPP sampling for transfer learning.

To achieve this, we first revisit the selection of the subset as a stochastic sampling process according to DPP and reformulate the selection of the source domain subset with DPP sampling based on the transferability and diversity.

Give a source domain D^s, we randomly select a subset $C^s \in D^s$ from source domain by probability. The probability $P(C^s)$ of subset C^s selection is determined by transferability and diversity of source domain instance x^s in C^s.

$$P(C^S) \sim P(transferability(C^S), diversity(C^S)), \tag{1}$$

Definition 1. *DPP Sampling. Given the source domain* $D^s = \{x_1^s, x_2^s, \cdots, x_N^s\}$*, a sampling possibility of each instance in source domain* D^s *can be defined as a point process* $\mathcal{P}$*. The probability of a subset* C^S *being selected is*

$$P(C^S) = \det(K_{C^S}), \tag{2}$$

where K is referred as the marginal kernel. $K_{C^S} = \left[K_{ij}\right]_{x_i^s, x_j^s \in C^S}$ *denotes the sub-matrix indexed by the source domain instances in* C^S *and* $\det(K_{C^S})$ *is the determinant of the sub-matrix. For the empty set,* $\det(K_{C^S}) = 1$.

In fact, the marginal kernel K is difficult to construct. Referring to [38], L-ensemble can be used to construct DPP for the sampling of source domain samples.

Definition 2. *DPP Sampling with L-ensemble. A sampling possibility of each instance in source domain* D^s *can be calculated by a point process* $\mathcal{P}$ *with L-ensemble. The probability of a subset* C^S *being selected is*

$$P(C^S) = \frac{\det(L_{C^S})}{\det(L+I)}, \tag{3}$$

where L is a $N \times N$ *positive semidefinite matrix indexed by instances of* D^s*, I is the* $N \times N$ *identity matrix and* $\det(L+I) = \sum_{C' \subseteq D^s} \det(L_{C'})$ *is used to normalize the matrix determinants to probabilities.*

Definition 3. *k-DPP Sampling with L-ensemble. Suppose the sampled subset* C^S *consists of k source domain instances, the sampling probability of subset* C^S *can be written as,*

$$P^k(C^S) = \frac{\det(L_{C^S})}{\sum_{C' \subseteq D^s \wedge |C'|=k} \det(L_{C'})} \tag{4}$$

where $\left|C^S\right| = k$ *and* L_{C^S} *is the* $k \times k$ *submatrix of L indexed by* C^S.

Suppose correlation matrix $L = \sum_{i=1}^{N} \lambda_i v_i v_i^T$, λ_i refers to the eigenvalue corresponding to the eigenvector v_i, the probability of selecting a k-size subset C^S is

$$P^k(C^S) = \frac{\det(L_{C^S})}{\sum_{C' \subseteq C \wedge |C'|=k} \det(L_{C'})} = \frac{\prod_{c_i \in C^S} \lambda_i}{\sum_{C' \subseteq C \wedge |C'|=k} \{\prod_{c_j \in C'} \lambda_j\}}. \tag{5}$$

In summary, the k-DPP sampling probability is determined by the correlation matrix L. The different correlation matrices have different sampling properties. To ensure that the subset of source domain is highly transferability and diversity, we redefine the matrix L in Section 3.2.

3.2. Correlation Matrix L Construction

In this section, we define the transferability $m(\Omega|x_i^s;\Phi)$ and diversity $s(x_i^s, x_j^s)$ of source domain instances and use them to rewrite the L-correlation matrix. Decomposing the matrix L as a Gram matrix $L = B^T B$, the probability of k-DPP sampling can be calculate with the transferability $m(\Omega|x_i^s;\Phi)$ and diversity $s(x_i^s, x_j^s)$ [38]. Suppose each vector B_i in B has the form of $B_i = m(\Omega|x_i^s;\Phi) \cdot \phi_i$, in which ϕ_i is the normalized feature vector. the matrix L can be defined as

$$L = \left[L_{ij}\right]_{1 \le i,j \le N} \tag{6}$$

where $L_{ij} = m(\Omega|x_i^s;\Phi)\phi_i^T \cdot \phi_j m(\Omega|x_j^s;\Phi)$, and the inner product $\phi_i^T \cdot \phi_j \in [-1, +1]$ indicates the similarity between source domain instance x_i^s and x_j^s, we rewrite

$$L = \left\{ L_{ij} = m(\Omega|x_i^s;\Phi)s(x_i^s, x_j^s)m(\Omega|x_j^s;\Phi) | 1 \le i \le N, 1 \le j \le N \right\}. \tag{7}$$

where $s(x_i^s, x_j^s) = \phi_i^T \cdot \phi_j$.

3.2.1. Transferability Measure of Instance with Evidence Theory

The evidence theory is a generalization of Bayesian theory to subjective probabilities [39–41]. In the evidence theory, a mass function $m(\cdot)$ is constructed to assign masses to the elements of the power set of a frame of discernment Ω. For classification tasks, Ω is the class label space and the mass function is used to assign masses to the subsets of class labels. For the mass assigned on the whole label space $m(\Omega)$, it means that all the class labels have the same possibility and implies the probability of knowing nothing (ignorance) [42].

In the transferred classification with domain adaptation, when data comes from source domain and the label space comes from target domain, the mass assigned on the whole label space represents the transferability of source domain data with respect to the classification on target domain. Based on this, we define the transferability measure for transfer learning as follows.

Suppose that Ω is the label space of the target domain, a data instance x^s comes from source domain, Φ^t is an evidence set from the target domain for x^s, the transferability of the source domain data instance x^s about the classification on target domain is

$$transferability = m(\Omega|x^s;\Phi^t). \tag{8}$$

$m(\Omega|x^s;\Phi^t)$ denotes the ignorance probability of x^s about the classes of target domain under a given evidence set Φ^t.

Next, we introduce how to formulate $m(\Omega|x^s;\Phi^t)$. First, we construct the evidence set Φ^t according to the target domain D^t. Given a source-domain instance x^s, its evidence set Φ^t can be written as a neighborhood surrounding x^s.

$$\Phi^t = \{x_1^t, x_2^t, \cdots, x_n^t\}, \tag{9}$$

where x_i^t comes from the target domain.

To achieve this, we design the objective function of obtaining an evidence set as follows:

$$\Phi^t = \arg\min_{\Phi} \left\| \phi(x^s) - \frac{1}{|\Phi|} \sum_{x^t \in \Phi} \phi(x) \right\|_{\mathcal{H}}^2, \tag{10}$$

where ϕ is the feature mapping. The optimal evidence set Φ^t can be solved by the greedy algorithm on the labeled target domain.

Second, we further decompose Φ^t and refine the mass function to implement the uncertainty measure. Given k classes, the evidence set Φ^t can be divided into different classes,

$$\Phi^t = \{\Phi_1^t, \Phi_2^t, \ldots, \Phi_k^t\}, \tag{11}$$

where $\Phi_k^t = \{x_{k1}^t, \ldots x_{kl}^t\}$ is the evidence subset in which all the target domain instances have the class label y_k, and x_{kl}^t is the lth element in the evidence subset.

Through decomposing the evidence set, we can adopt Dempster's rule to refine the ignorance mass $m(\Omega|x;\Phi^t)$ with multilevel evidence as

$$m(\Omega|x^s;\Phi^t) = \bigoplus_{\Phi_k^t \subseteq \Phi^t} m(\Omega \mid x^s;\Phi_k^t) = \bigoplus_{\Phi_k^t \subseteq \Phi^t} \bigoplus_{x^t \subseteq \Phi_k^t} m(\Omega \mid x^s;x^t), \tag{12}$$

in which the orthogonal sum $\bigoplus$ denotes the combination operator of Dempster's rule [43]. The ignorance mass $m(\Omega|x^s;x^t)$ is calculated by

$$m(\Omega|x^s;x_i^t) = 1 - \exp(-d(x^s,x_i^t)), \tag{13}$$

where $d(\cdot)$ is defined as

$$d(x^s,x_i^t) = \mathrm{K}(x^s,x^s) - 2\mathrm{K}(x^s,x_i^t) + \mathrm{K}(x^t,x_i^t), \tag{14}$$

where $\mathrm{K}(\cdot)$ is the radial basis function kernel.

3.2.2. Diversity Measure of Instance

According to the characteristics of k-DPP sampling, if $s(x_i^s, x_j^s)$ measures the similarity of two instances, this can make it less probable that similar source domain instances are selected at the same time, thus ensuring the diversity of the selected subset. To achieve this, we adopt Normalized Mutual Information (NMI) to measure the similarity between two instances x_i^s and x_j^s.

$$s = NMI(x_i^s, x_j^s). \tag{15}$$

3.3. Model Ensemble

Algorithm 1 lists the main steps of the ensemble strategy. According to the improved DPP sampling, we can obtain m subsets $T = \{T_1, T_2, \cdots, T_m\}$ from source domain D^s. For each subset T_i, we train a base model f_i for the target domain. By repeating the process, we can get a set of base models $F = \{f_1, f_2, \cdots, f_m\}$. We combine the base model by

$$f(x) = \sum_{i=1}^{|T|} w_i f_i(x), \tag{16}$$

where w_i is the weighting and represents the transferability of the base model with respect to the target domain. To achieve this, we calculate the weight w_i using the adaptability of the subset with respect to the target domain. It can be defined as

$$w_i = \frac{\mathrm{Ent}(\mathrm{T}_i)}{\sum_{i=1}^{|T|} \mathrm{Ent}(\mathrm{T}_i)}, \tag{17}$$

where $Ent(T_i)$ is the adaptability of the subset about the target domain task. The adaptability is defined as

$$\mathrm{Ent}(T_i) = -\frac{1}{|T_i|} \sum_{x \in T_i} (m(\Omega|x)). \tag{18}$$

In addition, the proposed ensemble strategy is a general technology and the existing transfer learning (TL) methods can be embedded into it.

Algorithm 1 Ensemble strategy based on DPP sampling for transfer learning.

Input: Source domain D^s, target domain D^t;
Output: $f(x) = \sum_{i=1}^{|T|} w_i f_i(x)$;
1: **for** each x^s in D^s **do**
2: Construct an evidence set Φ for x^s with labeled data instances in target domain D^t;
3: Decompose the evidence set Φ and formulate the mass function $m(\cdot|x^s;\Phi)$ with multilevel evidences;
4: Hierarchically compute $m(\Omega|x^s;\Phi)$ to measure the transferability of x^s for transfer learning according to formula (12);
5: **end for**
6: Construct the similarity matrix of source domain instances using formula (15);
7: Construct correlation matrix L of DPP based on transferability and diversity;
8: **for** each i in $|T|$ **do**
9: Decompose L matrix into eigenvalues and eigenvectors;
10: Calculate the probability $P(T_i)$ of selecting a subset T_i;
11: Generate the subset T_i according to DPP sampling probability $P(T_i)$;
12: Calculate the transferability of subset T_i as weights w_i according to formula (17);
13: Train the transfer model f_i based on T_i;
14: **end for**
15: **return** $f(x) = \sum_{i=1}^{|T|} w_i f_i(x)$.

4. Experiments

To verify the effectiveness of the ensemble strategy, we couple the ensemble strategy into traditional TL methods and deep transfer methods, and test the transfer performance of the TL model on various kinds of data, including Amazon product reviews, Office+Caltech data sets and Office-Home data sets. The descriptions of the data sets are listed below.

4.1. Data Sets

Amazon product reviews is a cross-domain text data set for transfer learning evaluation [44]. The dataset includes four domains: books (denote B), dvds (denote D), electronics (denote E) and kitchen appliances (denote K). In each domain, there are 1000 positive reviews and 1000 negative ones. In this data set, we can construct 12 cross-domain sentiment classification tasks: $B-D$, $B-E$, $B-K$, $D-B$, $D-E$, $D-K$, $E-B$, $E-D$, $E-K$, $K-B$, $K-D$, $B-E$, where the word before '$-$' corresponds with the source domain and the word after '$-$' corresponds with the target domain.

The Office+Caltech data set is generated from Office and Caltech-256, which are introduced by Gong et al. [25], which are two benchmark data sets widely adopted for visual domain adaptation evaluation. It consists of 4563 images with 31 categories. Caltech-256 is a standard database for object recognition. It consists of 30,607 images with 256 categories. In experiments, we use the smaller Office+Caltech data sets. It includes four domains: Amazon (denotes A), Webcam (denotes W), DSLR (denotes D), and Caltech-256 (denotes C). The dataset includes 10 classes. There are 8 to 151 samples per category per domain, and 2533 images in total. In this dataset, we can construct 9 cross-domain classification tasks: $A-C$, $A-W$, $C-A$, $C-W$, $D-A$, $D-C$, $D-W$, $W-A$, $W-C$.

The Office-Home data set has been created to evaluate domain adaptation algorithms for object recognition using deep learning [13]. It consists of images from four different domains: Artistic images (denotes A, paintings, sketches andor artistic depictions), Clip Art (denotes C, clipart images), Product images (denotes P, images without background) and Real-World images (denotes R, regular images captured with a camera). The data set has a total of 15,500 images, and for each domain, the data set contains images of 65 object categories found typically in Office and Home settings. In this data set, we construct 12 cross-domain classification tasks: $A-C$, $A-P$, $A-R$, $C-A$, $C-P$, $C-R$, $P-A$, $P-C$, $P-R$, $R-A$, $R-C$ and $R-P$.

4.2. Experimental Study on Traditional Transfer Learning Methods

In the experiment, to verify the effectiveness of our proposed ensemble strategy, we couple it with 9 traditional Transfer Learning (TL) methods and compare the classification results with and without the ensemble strategy. In addition, we compare different ensemble strategies where the base models are trained by subsets of the source domain, which are separately generated by improved DDP sampling, information gain, and random sampling. The traditional TL methods include: transfer component analysis (TCA) [18], correlation alignment (CORAL) [26], geodesic flow kernel (GFK) [25], joint distribution adaptation (JDA) [24], kernel mean matching (KMM) [20], scatter component analysis (SCA) [45], Balanced Distribution Adaptation (BDA) [46], Manifold Embedded Distribution Alignment (MEDA) [35], and practically easy transfer learning (EasyTL) [47]. For any transfer learning (TL) method '$*$', we briefly denote the ensemble strategy based on improved DPP sampling as '$E-*$', based on information gain (IG) as "$I-*$" and based on random sampling as '$R-*$'. The abbreviations are summarized in Table 2.

Table 2. Abbreviations and descriptions used in the experiments.

Abbreviation	Description
$E-*$	The ensemble strategy with improved DPP sampling
$I-*$	The ensemble strategy with information gain
$R-*$	The ensemble strategy with random sampling

4.2.1. Experimental Setting

In each cross-domain text sentiment classification task, we apply the Bert model to extract the feature of the review texts [48]. In each cross-domain image classification task, we utilize deep convolutional activation features (DeCAF6 features) to represent all the images, in which the outputs from the 6th layer in the deep convolutional neural network are transformed to 4096 dimensional features [35]. In DPP sampling and random sampling, the number of subsets is set to 10. The size of the subset is 70% of the size of the source domain.

4.2.2. Test on Text Data

In this testing, we construct 12 cross-domain sentiment classification tasks from Amazon product reviews, i.e, $B-D$, $B-E$, $B-K$, $D-B$, $D-E$, $D-K$, $E-B$, $E-D$, $E-K$, $K-B$, $K-D$, $B-E$.

As shown in Table 3, we perform the TL methods with and without ensemble strategy to generate the sentiment classification results, respectively. It is clear to observe that TL methods with our proposed ensemble strategy achieve the best performance on all cross-domain text classification tasks. Specifically, for the TL methods, using the ensemble strategy improves the accuracies of cross sentiment classifications by 4.45%, 4.39%, 3.42%, 4.07%, 6.46%, 4.47%, 2.65%, 3.18%, respectively. In addition, we can find that the KMM method achieves the largest performance improvement 6.46% by embedding the ensemble strategy. The reason is that KMM is sensitive to domain differences and is unstable for predicting out of distribution (OOD) data. Using the ensemble strategy can enhance the performance of KMM when predicting OOD data. These results clearly demonstrate that our ensemble strategy can improve the performance of the single model on the text data set.

To further validate the effectiveness of our proposed method on text data set, we compare different ensemble strategies where the base models are trained by subsets of source domain, which are separately generated by improved DDP sampling, information gain, and random sampling. The classification accuracies are listed in Table 4. In random sampling, we randomly select 10 subsets to train the base models. The size of the subset is 70% of the size of the source domain.

Table 3. Cross-domain sentiment classification accuracies of text dataset generated by the traditional TL methods with and without ensemble strategy.

Methods	$B-D$	$B-E$	$B-K$	$D-B$	$D-E$	$D-K$	$E-B$	$E-D$	$E-K$	$K-B$	$K-D$	$K-E$	*Ave acc*
TCA	77.76	75.54	78.74	76.05	76.38	79.34	73.35	73.66	79.74	73.05	77.26	78.74	76.63
E-TCA	**81.21**	**83.31**	**84.79**	**81.95**	**82.41**	**82.13**	**76.60**	**79.89**	**82.57**	**78.10**	**79.28**	**80.80**	**81.09**
CORAL	70.76	66.21	70.00	73.05	68.70	71.96	69.90	65.71	72.35	67.45	68.61	75.68	70.03
E-CORAL	**75.79**	**73.48**	**74.77**	**76.57**	**73.51**	**74.25**	**76.90**	**71.15**	**75.36**	**73.52**	**71.35**	**76.44**	**74.42**
GFK	75.76	72.00	73.50	71.85	68.96	75.70	72.60	71.11	76.20	73.75	74.21	76.58	73.52
E-GFK	**78.00**	**77.39**	**76.42**	**79.13**	**74.98**	**79.02**	**77.23**	**73.41**	**78.05**	**75.69**	**75.08**	**78.91**	**76.94**
JDA	77.26	75.93	78.09	77.65	76.03	78.29	72.65	72.16	80.14	75.05	77.56	80.32	76.76
E-JDA	**80.09**	**82.14**	**84.23**	**80.31**	**81.71**	**81.99**	**77.38**	**77.71**	**82.88**	**80.05**	**79.81**	**81.39**	**80.81**
KMM	83.76	79.02	75.90	80.50	68.51	76.45	73.70	77.86	80.39	74.25	75.96	85.00	77.61
E-KMM	**85.98**	**81.44**	**85.47**	**85.22**	**82.79**	**85.82**	**79.00**	**84.08**	**84.52**	**82.59**	**84.16**	**87.78**	**84.07**
BDA	75.01	73.04	75.75	74.55	71.80	75.30	71.40	71.61	78.49	71.15	71.66	76.93	73.89
E-BDA	**79.22**	**77.51**	**82.41**	**80.01**	**78.06**	**78.26**	**75.70**	**74.93**	**81.76**	**77.11**	**74.88**	**80.44**	**78.36**
SCA	79.41	78.82	78.14	74.95	77.53	76.00	73.15	71.86	79.79	73.70	75.71	82.41	76.79
E-SCA	**83.52**	**79.88**	**79.80**	**82.01**	**78.30**	**78.15**	**76.45**	**74.80**	**82.22**	**74.89**	**79.00**	**84.33**	**79.45**
EasyTL	76.76	76.08	82.14	83.90	80.91	85.22	76.40	72.61	86.37	73.25	70.86	75.93	78.37
E-EasyTL	**79.44**	**83.06**	**83.21**	**85.09**	**85.66**	**85.00**	**78.13**	**77.77**	**86.93**	**77.49**	**77.81**	**79.03**	**81.55**

Table 4. Cross-domain sentiment classification accuracies of the text dataset generated by ensemble strategy with improved DPP sampling ("$E-*$"), information gain ("$I-*$") and random sampling ("$R-*$").

Methods	$B-D$	$B-E$	$B-K$	$D-B$	$D-E$	$D-K$	$E-B$	$E-D$	$E-K$	$K-B$	$K-D$	$K-E$	*Ave acc*
R-TCA	75.29	79.14	78.06	75.10	74.89	77.61	71.25	74.19	77.20	74.33	77.00	79.204	76.11
I-TCA	77.12	80.45	81.11	77.09	76.54	77.00	73.62	75.04	77.68	75.76	78.4	78.93	77.40
E-TCA	**81.21**	**83.31**	**84.79**	**81.95**	**82.41**	**82.13**	**76.60**	**79.89**	**82.57**	**78.10**	**79.28**	**80.80**	**81.09**
R-CORAL	69.12	63.44	71.93	76.25	70.98	70.67	71.18	65.45	73.14	65.71	65.47	76.84	70.01
I-CORAL	69.89	64.38	72.05	77.22	71.14	72.53	71.97	68.39	74.26	66.83	67.92	76.42	71.08
E-CORAL	**75.79**	**73.48**	**74.77**	**76.57**	**73.51**	**74.25**	**76.90**	**71.15**	**75.36**	**73.52**	**71.35**	**76.44**	**74.42**
R-GFK	75.92	73.17	71.29	70.22	68.91	73.49	71.89	72.94	75.82	72.97	75.04	76.40	73.17
I-GFK	77.04	73.9	73.81	74.62	70.46	75.18	73.55	72.87	76.34	74.11	74.23	75.99	74.34
E-GFK	**78.00**	**77.39**	**76.42**	**79.13**	**74.98**	**79.02**	**77.23**	**73.41**	**78.05**	**75.69**	**75.08**	**78.91**	**76.94**
R-JDA	77.80	76.47	79.14	75.03	76.45	77.36	71.08	71.78	79.94	76.17	77.14	80.02	76.53
I-JDA	77.95	78.76	81.89	77.64	76.09	79.09	73.4	73.03	80.16	77.38	77.69	79.79	77.74
E-JDA	**80.09**	**82.14**	**84.23**	**80.31**	**81.71**	**81.99**	**77.38**	**77.71**	**82.88**	**80.05**	**79.81**	**81.39**	**80.81**
R-KMM	80.74	78.77	74.12	80.06	69.21	75.82	72.69	77.37	81.02	75.48	75.11	82.45	76.90
I-KMM	82	79.82	79.67	82.61	74.76	79.37	72.7	79.63	81.97	79.58	78.18	82.77	79.42
E-KMM	**85.98**	**81.44**	**85.47**	**85.22**	**82.79**	**85.82**	**79.00**	**84.08**	**84.52**	**82.59**	**84.16**	**87.78**	**84.07**
R-BDA	73.52	72.13	74.21	72.78	70.71	73.48	72.05	70.40	78.25	70.74	72.09	75.00	72.95
I-BDA	74.97	74.44	76.51	74.39	74.16	75.44	74.19	72.41	78.97	73.96	72.95	76.13	74.88
E-BDA	**79.22**	**77.51**	**82.41**	**80.01**	**78.06**	**78.26**	**75.70**	**74.93**	**81.76**	**77.11**	**74.88**	**80.44**	**78.36**
R-SCA	80.20	76.24	75.33	71.27	78.14	75.91	72.41	72.19	78.80	73.55	75.39	80.83	75.86
I-SCA	79.88	77.41	76.03	75.19	77.23	76.07	74.08	71.98	80.8	73.26	75.9	81.51	76.61
E-SCA	**83.52**	**79.88**	**79.80**	**82.01**	**78.30**	**78.15**	**76.45**	**74.80**	**82.22**	**74.89**	**79.00**	**84.33**	**79.45**
R-EasyTL	76.80	76.11	81.74	83.47	81.20	84.31	75.24	73.37	86.57	72.50	72.11	76.54	78.33
I-EasyTL	75.91	79.38	80.15	84.11	82.35	83.79	76.04	72.54	85.15	74.28	74.05	76.87	78.72
E-EasyTL	**79.44**	**83.06**	**83.21**	**85.09**	**85.66**	**85.00**	**78.13**	**77.77**	**86.93**	**77.49**	**77.81**	**79.03**	**81.55**

As shown in Table 4, the average classification accuracies of TL methods with the proposed ensemble strategy on 12 tasks are 81.09%, 74.42%, 76.94%, 80.81%, 84.07%, 78.36%, 79.45%, 81.55%, respectively. Comparing with the ensemble strategy based on random sampling, the improved DPP sampling improves the accuracies of different cross-domain sentiment classification tasks by 4.98%, 4.41%, 3.77%, 4.28%, 7.17%, 5.41%, 3.59%, and 3.22%. Comparing with the ensemble strategy based on information gain, the improved DPP sampling improves the accuracies of different cross-domain sentiment classification

tasks by 3.69%, 3.34%, 2.60%, 3.07%, 4.65%, 3.48%, 2.84%, and 2.83%, especially TCA, JDA, and BDA, in which the maximum mean discrepancy metric is adopted to minimize the differences between the source domain and target domain achieve greater classification improvements. Based on the reported experimental results, our proposed ensemble strategy with DPP sampling is effective and can significantly enhance the transfer performance of TL methods on text data.

4.2.3. Test on Image Data

In the test, we further validate that the proposed ensemble strategy on image data. The image cross-domain classification tasks are constructed from the Office+Caltech data set, including $A - C$, $A - W$, $C - A$, $C - W$, $D - A$, $D - C$, $D - W$, $W - A$, $W - C$.

We perform TL methods with and without the ensemble strategy to generate the image classification results, respectively. The results are shown in Table 5, in all the cross-domain image classification tasks, our proposed ensemble strategy achieves better performance than the single TL methods. The TL methods with the ensemble strategy gain a significant performance improvement of 4.96%, 2.34%, 4.79%, 5.16%, 3.52%, 2.28%, 2.06%, and 3.65% compared to the single TL methods.

Table 5. Cross-domain classification accuracies of Office+Caltech image data sets generated by the traditional TL methods with and without ensemble strategy.

Methods	$A - C$	$A - W$	$C - A$	$C - W$	$D - A$	$D - C$	$D - W$	$W - A$	$W - C$	*Ave acc*
TCA	75.69	75.59	89.77	74.92	89.24	73.46	**98.30**	80.38	73.64	81.22
E-TCA	**83.41**	**81.17**	**90.14**	**89.78**	**90.28**	**79.55**	98.41	**84.51**	**78.37**	**86.18**
CORAL	83.7	74.58	89.98	78.64	85.70	79.16	**99.66**	77.14	74.98	82.62
E-CORAL	**84.75**	**81.79**	**91.25**	**81.43**	**87.88**	**81.52**	97.10	**82.33**	**76.57**	**84.96**
GFK	76.85	68.47	88.41	80.68	85.80	74.09	**98.64**	75.26	74.8	80.33
E-GFK	**81.76**	**80.25**	**90.02**	**86.66**	**88.78**	**80.04**	98.87	**82.37**	**77.35**	**85.12**
JDA	75.07	70.85	89.67	80.00	88.31	73.91	**98.31**	80.27	72.93	81.04
E-JDA	**82.72**	**81.17**	**92.45**	**87.33**	**89.91**	**78.51**	98.39	**86.85**	**78.43**	**86.20**
KMM	83.08	74.24	91.23	**80.34**	84.34	71.86	**98.98**	71.81	67.14	80.34
E-KMM	**84.81**	**78.88**	**92.35**	83.47	**85.52**	**78.16**	98.00	**80.70**	**72.77**	**83.85**
BDA	83.79	74.92	89.46	82.03	88.83	81.30	**99.31**	80.85	76.49	84.11
E-BDA	**86.51**	**76.68**	**91.94**	**86.59**	**88.78**	**83.22**	97.02	**86.70**	**80.08**	**86.39**
MEDA	87.71	85.76	91.07	84.07	92.90	**87.89**	**98.98**	93.21	86.73	89.81
E-MEDA	**87.73**	**89.63**	**92.61**	**92.71**	**93.81**	89.97	98.80	**93.67**	**87.93**	**91.87**
EasyTL	81.30	72.88	90.50	74.91	83.00	73.64	**93.22**	74.53	67.31	79.03
E-EasyTL	**83.09**	**80.54**	**90.58**	**81.25**	**85.79**	**79.11**	91.66	**79.58**	**72.54**	**82.68**

To further validate the effectiveness of our proposed ensemble strategy on the image data set, we compare different ensemble strategies where the base models are trained by subsets of source domain, which are separately generated by improved DDP sampling, information gain, and random sampling. As shown in Table 6, using the ensemble with DPP sampling, the TL methods achieve the average classification accuracies of 86.18%, 84.96%, 85.12%, 86.20%, 83.85%, 86.39%, 91.87%, and 82.68% on the cross-domain image data sets, respectively. In contrast to the ensemble strategy based on random sampling, the ensemble strategy with DPP sampling gains the significant performance improvements of 6.24%, 2.94%, 5.27%, 5.78%, 4.10%, 2.97%, 2.99%, and 4.19%. Comparing with the ensemble strategy based on information gain, the improved DPP sampling improves the accuracies of different cross-domain image classification tasks by 4.78%, 2.13%, 3.23%, 4.01%, 2.83%, 2.03%, 2.47%, and 3.53%. The experimental results reveal that our proposed ensemble strategy with DPP sampling can improve the transfer performance of TL methods on image data sets.

Table 6. Cross-domain classification accuracies of Office+Caltech image data sets generated by ensemble strategy with improved DPP sampling ("$E-*$"), information gain ("$I-*$") and random sampling ("$R-*$").

Methods	$A-C$	$A-W$	$C-A$	$C-W$	$D-A$	$D-C$	$D-W$	$W-A$	$W-C$	*Ave acc*
R-TCA	72.87	74.82	86.38	75.10	88.17	70.05	**97.28**	81.19	73.61	79.94
I-TCA	75.94	76.33	86.91	76.71	88.53	73.42	97.11	82.4	75.27	81.40
E-TCA	**83.41**	**81.17**	**90.14**	**89.78**	**90.28**	**79.55**	98.41	**84.51**	**78.37**	**86.18**
R-CORAL	83.80	74.22	87.29	78.37	84.96	80.01	**97.21**	78.00	74.31	82.02
I-CORAL	83.86	76.14	87.94	79.15	85.29	79.88	97.72	80.34	75.1	82.82
E-CORAL	**84.75**	**81.79**	**91.25**	**81.43**	**87.88**	**81.52**	97.10	**82.33**	**76.57**	**84.96**
R-GFK	77.02	70.21	86.39	80.22	84.07	74.34	**97.21**	75.52	73.69	79.85
I-GFK	78.34	75.75	87.56	82.44	86.59	76.64	97.19	77.78	74.7	81.89
E-GFK	**81.76**	**80.25**	**90.02**	**86.66**	**88.78**	**80.04**	98.87	**82.37**	**77.35**	**85.12**
R-JDA	75.43	71.43	88.53	79.29	87.19	72.46	**96.17**	80.04	73.15	80.41
I-JDA	77.95	73	89.39	81.48	87.47	74.86	97.09	82.66	75.81	82.19
E-JDA	**82.72**	**81.17**	**92.45**	**87.33**	**89.91**	**78.51**	98.39	**86.85**	**78.43**	**86.20**
R-KMM	83.22	74.63	90.17	**78.78**	83.79	71.11	**96.74**	70.04	69.26	79.75
I-KMM	83.46	75	89.61	80.35	83.91	73.71	97.56	75.52	70.06	81.02
E-KMM	**84.81**	**78.88**	**92.35**	83.47	**85.52**	**78.16**	98.00	**80.70**	**72.77**	**83.85**
R-BDA	83.33	73.81	87.35	81.97	88.91	80.05	**96.95**	81.04	77.38	83.42
I-BDA	84.08	74.87	88.23	84.66	87.52	81.45	96.07	83.52	78.81	84.36
E-BDA	**86.51**	**76.68**	**91.94**	**86.59**	**88.78**	**83.22**	97.02	**86.70**	**80.08**	**86.39**
R-MEDA	87.77	85.21	90.55	84.83	91.10	**86.64**	**95.80**	92.68	85.36	88.88
I-MEDA	87.14	84.24	91.32	87.49	91.93	87.26	96.9	92	86.37	89.41
E-MEDA	**87.73**	**89.63**	**92.61**	**92.71**	**93.81**	89.97	98.80	**93.67**	**87.93**	**91.87**
R-EasyTL	80.74	71.44	89.77	73.82	83.79	71.55	**92.56**	74.18	68.56	78.49
I-EasyTL	81.49	73.57	89.91	74.01	83.17	74.37	90.09	75.52	70.23	79.15
E-EasyTL	**83.09**	**80.54**	**90.58**	**81.25**	**85.79**	**79.11**	91.66	**79.58**	**72.54**	**82.68**

4.3. Experimental Study on Deep Transfer Model

Besides the traditional transfer learning methods, we also verify that our proposed ensemble strategy is effective to improve the deep transfer models. In the experiment, we integrate the ensemble strategy to 5 deep transfer models, including deep adaptation network (DAN) [49], deep version of manifold embedded distribution alignment (DANN) [50], joint adaptation network (JAN) [34], multi-representation adaptation network (MRAN) [37], and deep subdomain adaptation network (DSAN) [51]. We construct 12 cross-domain classification tasks from Office-Home data set.

The experiment consists of two parts: (1) Comparing the performance of the deep transfer model with and without the ensemble strategy. (2) We compared different ensemble strategies, in which the base model is trained using improved DPP sampling, information gain and random sampling to generate subsets of source domain, respectively. The results are listed in Tables 7 and 8. For any transfer learning (TL) method '$*$', we briefly denote the ensemble strategy based on improved DPP sampling as '$E-*$', based on information gain (IG) as "$I-*$" and based on random sampling as '$R-*$'. The abbreviations are summarized in Table 2.

4.3.1. Experimental Setting

We implement the experiments using PyTorch of version higher than 1.3 over a cluster of NVIDIA A100 GPUs. For the model training, we use the stochastic gradient descent (SGD) methods to optimize the network. In DPP sampling and random sampling, the number of subsets is set to 10. The size of the subset is 70% of the size of source domain.

4.3.2. Experimental Results

As shown in Table 7, using the ensemble strategy improves the classification accuracies of deep transfer model by 2.01%, 1.75%, 2.35%, 1.69%, 1.74%, respectively, compared with the single model. The performance of our proposed ensemble strategy is better than the

compared methods on most of the cross domain classification task, which indicates that our proposed ensemble strategy can be integrated into deep transfer models and further improve their performance.

Table 7. Cross-domain classification accuracies of Office-Home image data sets generated by deep transfer methods with and without ensemble strategy.

Methods	$A-C$	$A-P$	$A-R$	$C-A$	$C-P$	$C-R$	$P-A$	$P-C$	$P-R$	$R-A$	$R-C$	$R-P$	*Ave acc*
DAN	43.60	57.00	67.90	45.80	56.50	60.40	44.00	43.60	67.70	63.10	51.50	74.30	56.28
E-DAN	**45.96**	**60.05**	**69.72**	**48.51**	**58.91**	**62.04**	**46.33**	**44.86**	**68.19**	**64.51**	**53.60**	**76.77**	**58.29**
DANN	45.60	59.30	70.10	47.00	58.50	60.90	46.10	43.70	68.50	63.20	51.80	76.80	57.63
E-DANN	**45.6**	**61.13**	**72.15**	**47.9**	**60.51**	**61.84**	**48.88**	**46.09**	**70.19**	**66.65**	**52.09**	**79.41**	**59.37**
JAN	45.90	61.20	68.90	50.40	59.70	61.00	45.80	43.40	70.30	63.90	52.40	76.80	58.31
E-JAN	**46.61**	**64.08**	**70.53**	**52.94**	**62.66**	**61.99**	**48.85**	**47.39**	**73.64**	**65.55**	**54.10**	**79.56**	**60.66**
MRAN	53.80	68.60	75.00	57.30	68.50	68.30	58.50	54.60	77.50	70.40	60.00	82.20	66.23
E-MRAN	**56.66**	**70.14**	**77.63**	**59.46**	**69.78**	**70.04**	**59.14**	**55.07**	**77.93**	**73.58**	**62.21**	**83.31**	**67.91**
DSAN	54.40	70.80	75.40	60.40	67.80	68.00	62.60	55.90	78.50	73.80	60.60	83.10	67.61
E-DSAN	**56.17**	**72.68**	**75.96**	**62.26**	**69.77**	**69.89**	**64.44**	**58.83**	**79.84**	**74.09**	**62.34**	**85.92**	**69.35**

To further validate the effectiveness of our ensemble strategy on the deep transfer model, we compare different ensemble strategies where the base models are trained by subsets of source domain, which are separately generated by improved DDP sampling, information gain, and random sampling. Table 8 list the results. We can observe that our proposed ensemble strategy adopting DPP sampling achieves superior performance than the random sampling. The average accuracies of ensemble strategy with DPP sampling are higher by 2.17%, 2.58%, 2.98%, 3.22%, and 2.40% than the ensemble strategy with random sampling on each task, respectively. Comparing with ensemble strategy based on information gain, the improved DPP sampling improves the accuracies of different cross-domain image classification tasks by 1.81%, 2.19%, 2.11%, 2.12%, and 1.74%. These results further validate the effectiveness of our proposed strategy in improving the performance of deep transfer models.

Table 8. Cross-domain image classification accuracies of Office-Home image data sets generated by ensemble strategy with improved DPP sampling ("$E-*$"), information gain ("$I-*$") and random sampling ("$R-*$").

Methods	$A-C$	$A-P$	$A-R$	$C-A$	$C-P$	$C-R$	$P-A$	$P-C$	$P-R$	$R-A$	$R-C$	$R-P$	*Ave acc*
R-DAN	43.67	55.20	66.38	47.78	56.31	61.85	43.48	43.84	65.50	61.49	52.71	75.19	56.12
I-DAN	43.74	56.22	67.42	46.94	56.79	61.7	44.51	43.26	66.62	62.36	52.42	75.8	56.48
E-DAN	**45.96**	**60.05**	**69.72**	**48.51**	**58.91**	**62.04**	**46.33**	**44.86**	**68.19**	**64.51**	**53.60**	**76.77**	**58.29**
R-DANN	45.69	57.11	67.44	45.83	58.61	58.37	46.73	42.19	69.46	62.74	52.69	74.57	56.79
I-DANN	45.38	57.78	68.74	46.22	58.89	59.61	46.96	44.46	69.07	62.85	51.14	75.1	57.18
E-DANN	**45.60**	**61.13**	**72.15**	**47.90**	**60.51**	**61.84**	**48.88**	**46.09**	**70.19**	**66.65**	**52.09**	**79.41**	**59.37**
R-JAN	43.89	60.01	66.31	49.4	59.35	61.88	44.67	45.19	70.45	61.19	51.40	78.37	57.68
I-JAN	44.1	62.36	67.58	50.29	60.79	61.46	45.58	46.42	70.86	62.28	52.25	78.66	58.55
E-JAN	**46.61**	**64.08**	**70.53**	**52.94**	**62.66**	**61.99**	**48.85**	**47.39**	**73.64**	**65.55**	**54.10**	**79.56**	**60.66**
R-MRAN	53.33	68.10	73.96	55.37	66.43	67.17	56.39	52.25	75.41	68.88	58.04	81	64.69
I-MRAN	54.92	68.8	74.52	56.38	67.92	68.36	57.47	53.39	76.18	70.08	59.53	81.97	65.79
E-MRAN	**56.66**	**70.14**	**77.63**	**59.46**	**69.78**	**70.04**	**59.14**	**55.07**	**77.93**	**73.58**	**62.21**	**83.31**	**67.91**
R-DSAN	54.51	70.17	74.78	61.28	66.2	66.68	62.90	54.35	78.77	72.35	59.91	81.55	66.95
I-DSAN	55.28	70.64	74.82	61.68	68.22	67.19	63.82	55.46	79.04	73.14	60.01	82.05	67.61
E-DSAN	**56.17**	**72.68**	**75.96**	**62.26**	**69.77**	**69.89**	**64.44**	**58.83**	**79.84**	**74.09**	**62.34**	**85.92**	**69.35**

5. Conclusions

In this article, we proposed a novel ensemble strategy based on improved DPP sampling. Specifically, we first rewritten the correlation matrix of DPP with transferability and diversity.

Second, we use the improved DPP sampling to select k subsets from the source domain. Finally, we train the base models with the selected subsets and using the transferability of subsets to ensemble the base models. The proposed strategy is a general preprocessing technique. Through coupling the ensemble strategy into the transfer learning model, we can improve the robustness and generalization of the transfer model. Experiments on text and image data sets validate that our proposed ensemble strategy improves the performances of various kinds of transfer learning methods. In our work, the ensemble strategy based on improved DPP is limited by instance transferability, which can affect the performance of the ensemble strategy if there are huge differences between the source domain and target domain. Moving forward, we plan to extend the ensemble strategy to handle the transfer learning with multiple source domains and the domain adaptation on open sets.

Author Contributions: Conceptualization, Y.L. and B.Z.; methodology, Y.L.; data curation, Z.X.; writing–original draft, Y.L.; writing–review–editing, B.Z. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
2. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
3. Zhang, J.; Li, W.; Ogunbona, P.; Xu, D. Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–38. [CrossRef]
4. Jiang, J.; Shu, Y.; Wang, J.; Long, M. Transferability in Deep Learning: A Survey. *arXiv* **2022**, arXiv:2201.05867.
5. Iman, M.; Rasheed, K.; Arabnia, H.R. A Review of Deep Transfer Learning and Recent Advancements. *arXiv* **2022**, arXiv:2201.09679.
6. Dai, W.; Yang, Q.; Xue, G.R.; Yu, Y. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning, Corvalis, OR, USA, 20–24 June 2007; pp. 193–200.
7. Chen, M.; Weinberger, K.Q.; Blitzer, J. Co-training for domain adaptation. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; pp. 2456–2464.
8. Courty, N.; Flamary, R.; Tuia, D.; Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1853–1865. [CrossRef]
9. Fernando, B.; Habrard, A.; Sebban, M.; Tuytelaars, T. Unsupervised visual domain adaptation using subspace alignment. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2–8 December 2013; pp. 2960–2967.
10. Duan, L.; Tsang, I.W.; Xu, D. Domain transfer multiple kernel learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 465–479. [CrossRef]
11. Karbalayghareh, A.; Qian, X.; Dougherty, E.R. Optimal Bayesian transfer learning. *IEEE Trans. Signal Process.* **2018**, *66*, 3724–3739. [CrossRef]
12. Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Bellevue, WA, USA, 2 July 2012; pp. 17–36.
13. Venkateswara, H.; Chakraborty, S.; Panchanathan, S. Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations. *IEEE Signal Process. Mag.* **2017**, *34*, 117–129. [CrossRef]
14. Wei, K.; Iyer, R.K.; Wang, S.; Bai, W.; Bilmes, J.A. Mixed robust/average submodular partitioning: Fast algorithms, guarantees, and applications. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2233–2241 .
15. Qi, J.; Tejedor, J. Robust submodular data partitioning for distributed speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2254–2258.
16. Liu, W.; Yue, X.; Zhong, C; Zhou, J. Clustering Ensemble Selection with Determinantal Point Processes. In Proceedings of International Conference on Neural Information Processing, Sydney, NSW, Australia, 12–15 December 2019; pp. 621–633.
17. Yue, X.; Xiao, X.; Chen, Y. Robust neighborhood covering reduction with determinantal point process sampling. *Knowl.-Based Systems.* **2020**, *188*, 105063. [CrossRef]
18. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2010**, *22*, 199–210. [CrossRef] [PubMed]

19. Jiang, J.; Zhai, C. Instance weighting for domain adaptation in NLP. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 264–271.
20. Huang, J.; Gretton, A.; Borgwardt, K.M.; Scholkopf, B.; Smola, A.J. Correcting Sample Selection Bias by Unlabeled Data. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 601–608.
21. Chu, W.S.; De la Torre, F.; Cohn, J.F. Selective transfer machine for personalized facial action unit detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3515–3522.
22. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer joint matching for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014, pp. 1410–1417.
23. Yan, H.; Ding, Y.; Li, P.; Wang, Q.; Xu, Y.; Zuo, W. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2272–2281.
24. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer feature learning with joint distribution adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2–8 December 2013, pp. 2200–2207.
25. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.
26. Sun, B.; Feng, J.; Saenko, K. Return of frustratingly easy domain adaptation. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
27. Yang, J.; Yan, R.; Hauptmann, A.G. Cross-domain video concept detection using adaptive svms. In Proceedings of the 15th ACM international conference on Multimedia, Augsburg, Germany, 25–29 September 2007; pp. 188–197.
28. Duan, L.; Tsang, I.W.; Xu, D.; Maybank, S.J. Domain transfer svm for video concept detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1375–1381.
29. Duan, L.; Xu, D.; Tsang, I.W.H.; Luo, J. Visual event recognition in videos by learning from web data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1667–1680. [CrossRef]
30. Long, M.; Wang, J.; Ding, G.; Pan, S.J.; Philip, S.Y. Adaptation regularization: A general framework for transfer learning. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 1076–1089. [CrossRef]
31. Cao, Y.; Long, M.; Wang, J. Unsupervised domain adaptation with distribution matching machines. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
32. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 647–655.
33. Feng, C.; Zhong, C.; Wang, J.; Sun, J.; Yokota, Y. CANN: Coupled Approximation Neural Network for Partial Domain Adaptation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, QLD, Australia, 1–5 November 2021; pp. 464–473.
34. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Deep transfer learning with joint adaptation networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2208–2217.
35. Wang, J.; Feng, W.; Chen, Y.; Yu, H.; Huang, M.; Yu, P.S. Visual domain adaptation with manifold embedded distribution alignment. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 402–410.
36. Zhang, W.; Zhang, X.; Liao, Q.; Yang, W.; Lan, L.; Luo, Z. Robust normalized squares maximization for unsupervised domain adaptation. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual, 19–23 October 2020; pp. 2317–2320.
37. Zhu, Y.; Zhuang, F.; Wang, J.; Chen, J.; Shi, Z.; Wu, W.; He, Q. Multi-representation adaptation network for cross-domain image classification. *Neural Netw.* **2019**, *119*, 214–221. [CrossRef]
38. Kulesza, A.; Taskar, B. Learning determinantal point processes. *arXiv* **2011**, arXiv:1202.3738.
39. Dempster, A.P. A generalization of Bayesian inference. *J. R. Stat. Soc. Ser. B (Methodol.)* **1968**, *30*, 205–232. [CrossRef]
40. Denoeux, T. 40 years of Dempster-Shafer theory. *Int. J. Approx. Reason.* **2016**, *79*, 1–6. [CrossRef]
41. Yager, R.R.; Liu, L. *Classic Works of the Dempster-Shafer Theory of Belief Functions*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 219.
42. Liu, W.; Yue, X.; Chen, Y. Trusted Multi-View Deep Learning with Opinion Aggregation. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, Palo Alto, CA, USA, 22 February–1 March 2022; pp. 7585–7593.
43. Shafer, G. A mathematical theory of evidence turns 40. *Int. J. Approx. Reason.* **2016**, *79*, 7–25. [CrossRef]
44. Blitzer, J.; Dredze, M.; Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 440–447.
45. Ghifary, M.; Balduzzi, D.; Kleijn, W.B.; Zhang, M. Scatter Component Analysis: A Unified Framework for Domain Adaptation and Domain Generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1414–1430. [CrossRef] [PubMed]
46. Wang, J.; Chen, Y.; Hao, S.; Feng, W.; Shen, Z. Balanced distribution adaptation for transfer learning. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 1129–1134.
47. Wang, J.; Chen, Y.; Yu, H.; Huang, M.; Yang, Q. Easy Transfer Learning By Exploiting Intra-Domain Structures. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1210–1215.

48. Bose, T.; Illina, I.; Fohr, D. Unsupervised domain adaptation in cross-corpora abusive language detection. In Proceedings of the SocialNLP 2021-The 9th International Workshop on Natural Language Processing for Social Media, Virtual, 10 June 2021.
49. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning transferable features with deep adaptation networks. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 97–105.
50. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1180–1189.
51. Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; He, Q. Deep subdomain adaptation network for image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 1713–1722. [CrossRef] [PubMed]