Entropy Minimization & DomainAdversarial Training guided by LabelDistribution Similarity for DomainAdaptation

Fangzheng Xu
China University of Mining and Technology
Yu Bao (baoyu@cumt.edu.cn)
China University of Mining and Technology
Bingye Li
China University of Mining and Technology
Zhining Hou
China University of Mining and Technology
Lekang Wang
China University of Mining and Technology

Research Article

Keywords: domain adaptation, entropy minimization, domain adversarial training

Posted Date: December 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2373132/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on May 22nd, 2023.
See the published version at https://doi.org/10.1007/s00530-023-01106-w.

Entropy Minimization & Domain Adversarial Training guided by Label Distribution Similarity for Domain Adaptation

Xu Fangzheng [a], Bao Yu [b,*] Li Bingye [a] Hou Zhining [a] and Wang Lekang [a]

[a] *Computer Department, China University of Mining and Technology, Xuzhou, Jiangsu, China*
E-mail: ts21170027a31@cumt.edu.cn

[b] *Department first, University or Company name, Jiangsu, China*
E-mails: baoyu@cumt.edu.cn, 864231907@qq.com

Abstract. In domain adaptation, entropy minimization are widely used. However, entropy minimization will bring negative transfer when the pseudo-labels are inconsistent with the real labels. We hope to increase pseudo-label accuracy to counter negative transfer in entropy minimization. To this end, we introduce domain adversarial training into entropy minimization. Furthermore, we consider the misalignment caused by domain adversarial training under severe label shift. Therefore, we propose method called entropy minimization and domain adversarial training guided by label distribution similarity. Through domain adversarial training which focus more on class-aligned divergence, our method improves pseudo-label accuracy and reduce negative transfer in entropy minimization. Extensive experiments demonstrate the effectiveness and robustness of our proposed method.

Keywords: domain adaptation, entropy minimization, domain adversarial training

1. Introduction

In the past few years, deep learning has achieved success in many fields[1–5]. However, due to the reliance on a large amount of labeled data, which is expensive, deep learning is impractical to deploy in most real-world scenarios [6]. Even if a large amount of task labeled data is available, there will still exist difference between the original training data and the data generated in the actual scene [7] as application scenarios change over time. Generally speaking, a small domain shift can also lead to a drastic reduction of model performance on test data [8]. Therefore, how to obtain a model that performs well on the target data without using labeled target data label is a very important problem.

To address this problem, the concept of Unsupervised Domain Adaptation (UDA) [9] has been proposed. UDA aims to utilize labeled source and unlabeled target data which are similar but different to learn a model with low target risk [10].

*Corresponding author. E-mail: baoyu@cumt.edu.cn.

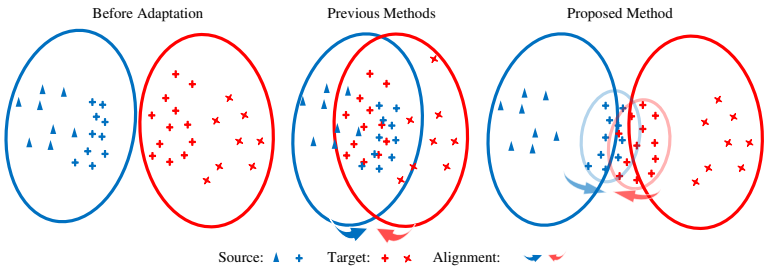

Fig. 1. Comparison between previous domain adversarial training methods and our method. Before Adaptation: there are both domain alignment divergence and class-misaligned divergence. Previous Methods: despite narrowing the domain gap, data from different classes are misaligned. Our Method: by giving higher weight to classes with similar label distribution, our method can reduce class-aligned divergence without category dislocation.

Recently, lots of works utilize entropy minimization in domain adaptation (DA) [10–14], which can increase model adaptability. Minimizing the entropy loss on the target domain can maximize the confidence of the model's predictions on target data, which has been shown to mitigate the harmful effects of domain shift [10]. However, entropy-minimization-only (EMO) do not consider target data distribution and, when pseudo-labels (labels predicted by model) are inconsistent with the true labels, it will increase the confidence in the prediction of the wrong class, which will bring negative transfer [13].

Domain Adversarial Training (DAT) is widely used in UDA [15–18]. Inspired by Generative Adversarial Networks (GANs) [19], DAT performs a minmax game between a domain discriminator and a classifier. During the training process, domain discriminator learns to minimize domain discrimination error, while feature extractor learns to maximize domain discrimination error. Through the game between the feature extractor and the domain discriminator, feature extractor learns to extract domain-invariant features, which can align the feature distribution between source and target data and improve the accuracy predicted by model for target data.

In this work, we first review domain adaptation method of EMO [10]. Secondly, the gradient update process adapting to the target data by entropy minimization is analyzed to understand the nature that entropy minimization can increase the model adaptability on target data. The results of the analysis demonstrates that model will be more confident in the prediction of target data with the update of the entropy loss gradient.However, it will lead to negative transfer when the model's predictions labels are inconsistent with the ground-truth labels, which makes its success highly dependent on initialization of model [13]. Meanwhile, in the case of severe label shift, entropy minimization leads to trivial solutions [20]. Therefore, we propose entropy minimization and DAT guided by label distribution similarity model, which aims to pay more attention to aligning class-aligned divergence between source and target data and improve model initialization performance on target data through DAT. Extensive experiments verify the effectiveness of our proposed method. In overall, our contributions are as follows:

We propose to introduce DAT into EMO to improve pseudo-labels accuracy and reduce negative transfer caused by inaccurate pseudo-labels.

DAT guided by label distribution similarity is been proposed, which focus more on class-aligned divergence by reweighting the domain discrimination loss.

A specific domain adaptation model called entropy minimization and DAT guided by label distribution similarity model is been proposed, which reduces negative transfer in EMO by aligning the source and target data feature distributions.

Extensive experiments verify that our proposed model is effective and robust.

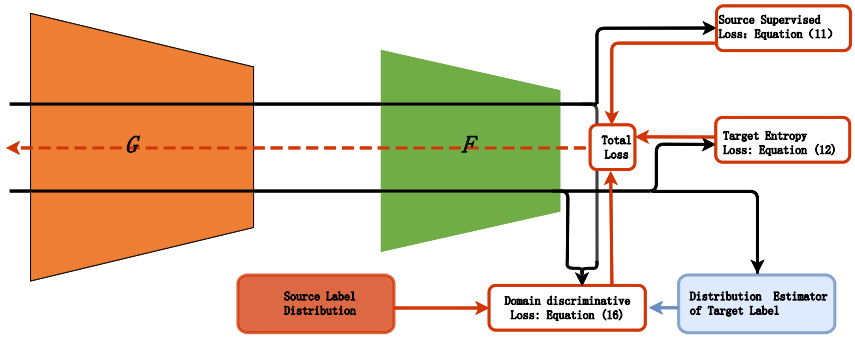

Fig. 2. Illustration of our Entropy minimization and DAT guided by Label Distribution Similarity framework, which consists of a feature extractor G and an integrated classification-discriminator F. Initially, the label distribution of the target data is assigned equally to the label distribution of the source data, and is gradually updated during the subsequent training process, which is learned from [18].The overall objective are shown in Equation.17 and Equation.18.

2. Related Words

Entropy minimization. Entropy minimization were originally proposed in [11] for semi-supervised learning, which learn decision rules from labeled and unlabeled data, and incorporate unlabeled data into standard supervised learning. Due to the effectiveness in domain adaptation, a large body of work, recently, in UDA, adopts entropy minimization. [21] passed the target classifier through low-density target regions by performing entropy minimization on unlabeled target data. [15] evaluates the transferability of target data with information entropy and reweights each training examples of the conditional domain discriminator. Since it is believed that direct entropy minimization ignores the structured output space that is beneficial for adaptation to semantic segmentation tasks, [12] indirectly minimizes entropy through adversarial training. In addition, [22] measures the distance of an instance away from the decision boundary by computing prediction entropy.

DAT. One of the most prominent works in UDA is DAT [23]. Inspired by GAN [19], DAT-based method trains feature extractors and domain discriminators in an adversarial manner. During the training process, feature extractor gradually learns to extract domain-invariant features while aligning feature distributions between source and target domains. [16] is the first to apply the idea of adversarial to the field of UDA and proposed Domain Adversarial Neural Network (DANN). Since it is believed that when data distributions embody complex multimodal structures, adversarial methods cannot align distributions if discriminator cannot perceive the cross-covariance dependency between the features and classes, [15], inspired by cycled-GAN(CGAN), proposed Conditional Domain Adversarial Networks (CDAN) which combining feature and category information for domain adaptation. To a step further, [17] combine classifier and discriminator by adding a domain discriminant neural node to the last fully connected layer of the classifier.

3. Method

Preliminary

Given labeled dataset $\mathcal{D}_s$ sampled from the distribution $p_s(x)$ and unlabeled dataset D_t sampled from the distribution $p_t(x)$, where $p_s(x) \neq p_t(x)$, The goal of UDA is to learn feature extractor $G(\cdot)$ and predictor $C(\cdot)$ with low target expectation risk $\mathcal{L}_t = \mathbb{E}_{(x_t,y_t)\in p_t}[\mathcal{L}_{cls}(C(G(x_t)), y_t)]$.

Since target data is unlabeled, true target risk is unknowable in UDA. Domain adaptation theory [24, 25] recommends to indirectly limit the upper bound of target risk

$$\mathcal{L}_t \leqslant \mathcal{L}_s + d_{\mathcal{H}}(p_s(x), p_t(x)) + \xi, \tag{1}$$

where ξ is the smaller term and $\mathcal{H}$ represents a hypothesis space, by minimizing source domain supervision risk $\mathcal{L}_s = E_{(s,y_s)\in p_s}[\mathcal{L}_{cls}(C(G(x_s), y_s)]$ and domain divergence $d_{\mathcal{H}}(D_s, D_t)$. Various proposed algorithms based on this theory all focus on the pursuit of simultaneously minimizing source domain supervision risk and domain divergence between source and target data, which mainly differ in the choice of $d_{\mathcal{H}}(D_s, D_t)$.

For source supervision risk, by training on labeled source data, model learns to minimize the source domain empirical supervision risk

$$\hat{\mathcal{L}}_s = \frac{1}{\mid D_s \mid} \sum_{(x_s,y_s)\in D_s} \mathcal{L}_{cls}(C(G(x_s)), y_s), \tag{2}$$

where $\mathcal{L}_{cls}$ is a commonly used supervision loss function such as cross-entropy, and $\mid \cdot \mid$ is number of samples in the dataset.

To adapt model to target data, a large number of methods utilize source data and unlabeled target data to minimize domain divergence. Recently entropy minimization has achieved good results. By minimizing the empirical Shannon entropy predicted by the model for the target data

$$\hat{\mathcal{L}}_{em} = \frac{1}{\mid D_t \mid} \sum_{x_t\in D_t} \mathcal{H}(softmax(C(G(x_t)))), \tag{3}$$

where $\mathcal{H}(\cdot) = -\langle x, log(s)\rangle$ and $softmax(\cdot)$ is a function that turns the output vector of the model into a probability prediction vector, entropy minimization maximize the model's prediction confidence on target data. Thus, standard EMO methods seek to solve the following problems:

$$min[\hat{\mathcal{L}}_s + \lambda \cdot \hat{\mathcal{L}_{em}}], \tag{4}$$

where λ is the trade-off between source supervision risk and entropy loss of target data.

In the following section, we firstly analyze the gradient of model predictions while adapting to the target data by EMO, and find that as the gradient is updated, the model predictions tend to approach the pseudo-labels. Therefore, when the pseudo-labels and the real labels are inconsistent, entropy minimization will bring about negative transfer. Secondly, we propose an ideal optimization direction which counter negative transfer in EMO by increasing the accuracy of pseudo-labels. Then, we implement the proposed optimization direction by introducing DAT. Finally, we further optimize the proposed method and propose entropy minimization and DAT guided by label distribution similarity.

3.1. Motivation

Theory 3.1 *Predictions of target data tend to pseudo-labels as the gradient is updated when optimizing by entropy minimization.*

Proof: Assuming $(x_t, y_t) \in D_t(x, y)$, the output of the model to the instance x_t is $a = [a_1, a_2, \dots a_K]^T$, where K is the number of categories. The predicted probability distribution the for is $\hat{y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_K]^T$, where $\hat{y}_i = \frac{e^{a_i}}{\sum_{k=1}^{K} e^{a_k}}$. The information entropy is:

$$\mathcal{H}(\hat{y}) = -\sum_{k=1}^{K} \hat{y}_k \cdot log\hat{y}_k. \tag{5}$$

Therefore, we obtain the gradient of $\mathcal{H}(\hat{y})$ with respect to the model output a_i:

$$\nabla_{a_i}\mathcal{H}(\hat{y}) = -\sum_{k=1}^{K}(1 + log(\hat{y}_k)) \cdot \frac{d_{\hat{y}_k}}{d_{a_i}} \tag{6}$$

$$= -(1 + log\hat{y}_i)(1 - \hat{y}_i) \cdot \hat{y}_i - \sum_{k \neq i}(1 + log\hat{y}_k)(-\hat{y}_k \cdot \hat{y}_i) \tag{7}$$

$$= -\hat{y}_i(log\hat{y}_i + \mathcal{H}(\hat{y})). \tag{8}$$

Then with backpropagation, the model output components are updated:

$$a_i += \eta \cdot \hat{y}_i(log\hat{y}_i + \mathcal{H}(\hat{y})), \tag{9}$$

where η is positively related to the learning rate. Due to $\eta, \hat{y}_i > 0$, we have a_i increases when $log\hat{y}_i + \mathcal{H}(\hat{y}) > 0$, i.e. $\hat{y}_i > e^{-\mathcal{H}(\hat{y})}$ and a_i decreases when $\hat{y}_i < e^{-\mathcal{H}(\hat{y})}$. At the same time, since the function $f(x) = x \cdot (logx + \mathcal{H}(\hat{y}))$ is a monotonically increasing function in the interval $x > e^{-\mathcal{H}(\hat{y})}$, the larger the probability value of the prediction term is, the greater addition is. Taken together, we obtain that predictions of target data tend to pseudo-labels as the gradient is updated when optimizing by entropy minimization.

According to Theory 3.1, we know that when the pseudo-labels are inconsistent with the true labels, entropy minimization will only lead to reinforcing such errors and bring out negative transfer. Therefore, how to increase accuracy of pseudo-labels is the key to reducing the negative transfer optimized by entropy minimization. So, we naturally thought, can further regularization be applied to promote the accuracy of pseudo-labels? Meanwhile, entropy minimization only focuses on confidence of model prediction on target data, without considering distribution alignment between source and target data which is important issue in UDA.

3.2. Introducing Domain Adversarial Training into Entropy Minimization

Based on the above analysis, we propose to increase the prediction accuracy of pseudo-labels and align feature distribution between source and target data through DAT. DAT is a cross-domain distribution alignment method that is widely used in UDA [16]. By introducing the domain discriminator $D(\cdot)$, DAT aligns the source and target data marginal distributions during the minmax-game between the domain discriminator $D(\cdot)$ and the feature extractor $C(\cdot)$. Empirical goals of DAT are as follows:

$$\max_{F} \min_{G} \mathcal{L}_{adv} = \frac{1}{\mid D_s \mid + \mid D_t \mid} \sum_{x \in D_s \cup D_t} \mathcal{L}_{bce}(D(G(x), y_d)), \tag{10}$$

where $\mathcal{L}_{bce}$ is a commonly used binary cross-entropy loss function, and y_d is the domain label.

3.3. Further Optimization of DAT

3.3.1. An integrated classification-discriminator

Although this simple DAT by introducing a domain discriminator can greatly reduce the domain divergence, separated adversarial design of the domain discriminator and classifier has shortcomings. First, feature distributions can only be aligned to a certain extent, because model capacity of feature extractor is large enough to compensate for feature distributions that are less aligned. More importantly, since domain discriminator cannot perceive class boundaries, it cannot fine-grained aligned distributions, which may lead to mode collapse in DAT[17, 26].

To this end, we adopt an adversarial design similar to [18, 27], which does not add an additional domain discrimination network, but by adding 1 neural node for domain discrimination in the last fully connected layer of the classifier, which constitutes an integrated classification-discriminator $F(\cdot)$ to predict category and domain information at the same time. $F(\cdot)$ outputs a K+1-dimensional vector, where the first K dimensions are used to predict class information, and the last dimension is activated by the $Sigmod-function$ to predict domain information. So, our optimization function is rewritten as follows:

$$\mathcal{L}_s = \frac{1}{\mid D_s \mid} \sum_{(x_s, y_s) \in D_s} \mathcal{L}_{cls}(F_1^K(D(x_s)), y_s), \tag{11}$$

$$\mathcal{L}_{em} = \frac{1}{\mid D_t \mid} \sum_{x_t \in D_t} \mathcal{H}(softmax(F_1^K(G(x_t)))), \tag{12}$$

$$\max_{G} \min_{F} \mathcal{L}_{adv} = \frac{1}{\mid D_s \mid + \mid D_t \mid} \sum_{x \in D_s \cup D_t} \mathcal{L}_{bce}(sigmod(F_{K+1}(G(x))), y_d), \tag{13}$$

where F_i^j represents a slice of the model $F(\cdot)$ output vector $[i, j)$ dimension.

3.3.2. DAT guided by label distribution similarity

DAT can efficiently align marginal distributions. However, previous DAT may introduce a new concept shift when existing the cross-domain label shift, which introduces negative transfer [27]. [28] decompose the empirical domain divergence measure into class-aligned and class-misaligned divergence, i.e.:

$$d_{\mathcal{H}\triangle\mathcal{H}}(p_s(x), p_t(x)) = 2 \sup_{h,h'\in\mathcal{H}} \mid \xi^{\cap}(h,h') + \xi^{-}(h,h') \mid, \tag{14}$$

where $\xi^{\cap}(h,h')$ is class-aligned divergence and $\xi^{-}(h,h')$ is class-misaligned divergence. Aligning marginal distributions with label shift will result in negative transfer due to forced alignment of class-misaligned divergence which is detrimental to domain adaptation[28].

To this end, we propose DAT guided by label distribution similarity. Comparison between our proposed and other DAT-based method is shown in Fig.1. Before adaptation, there are both class-aligned divergence and class-misaligned divergence between domains, which are caused by label distribution intersection and difference, respectively. Despite narrowing inter-domain divergence, previous methods may confuse data from different classes together due to the forced alignment of class-misaligned divergence. Our proposed method pays more attention to class-aligned divergence by reweighting the discrimination loss, which can significantly reduce class misalignment during DAT without causing alignment of class-misaligned divergence.

To elaborate our method, we first introduce the inter-domain label distribution similarity or class alignment metric vector $\vec{s} = \{s^1, s^2, ..., s^K\}^T$, where

$$\vec{s^i} = \frac{\min(p_s(y=i), \hat{p}_t(y=i))}{\max((p_s(y=i), \hat{p}_t(y=i)) + 1e-6} \tag{15}$$

stands for the i-th class label distribution similarity or class alignment metric. Since the estimation of the empirical vector of the initial target label distribution is inaccurate, we set $\hat{p}_t(y) = p_s(y)$ initially and adopted the same update strategy as [18] where $\hat{p}_t(y) = \frac{1}{1+\alpha}\{\hat{p}_t(y) + \alpha p_s(y)\}$.

Further, in order to force DAT to focus on the alignment of class alignment divergence, we preferentially give larger weights to category data which label distributions is close by reweighting the discrimination loss of each training data through

$$\mathcal{L}_{adv} = \frac{1}{\mid D_s \mid + \mid D_t \mid} \sum_{(x_s,y_s)\in D_s} \omega \cdot \mathcal{L}_d(D(G(x)), y_d). \tag{16}$$

where $\omega = 1 + e^{\vec{s}\cdot\vec{y}}$ and $\vec{y}$ is the one-hot encoded or predicted probability vector of the sample.

3.4. Approach Overall

In overall, proposed framework consists of a feature extractor $G(\cdot)$ and an integrated classification-discriminator $F(\cdot)$, which is illustrated in Fig.2. First, through supervised training on source data, model learns the ability to classify task data. Second, by reweighting domain discrimination loss with source data label distribution and estimated target data label distribution, $G(\cdot)$ aligns class-alignment divergence and learns to extract domain-invariant features. Finally, model fits the target data by minimizing the

Table 1
Details on JNU Bearing dataset

Transfer Tasks	Source Domain	Target Domain	Health Conditions
A->B	600rmp	800rpm	NC, IF
A->C	600rmp	1000rpm	OF, RF
B->C	800rpm	1000rpm	

Shannon entropy predicted by the model for the target data. By combining Eq.11, Eq.12 and Eq.16, our optimization objectives are as follows:

$$\min_{F}\{\mathcal{L}_s + \lambda_1\mathcal{L}_{em} + \lambda_2\mathcal{L}_{adv}\}, \lambda_1, \lambda_2 > 0, \tag{17}$$

$$\max_{G}\{-\mathcal{L}_s - \lambda_1\mathcal{L}_{em} + \lambda_2\mathcal{L}_{adv}\}. \tag{18}$$

4. Experiments

We evaluate our proposed method on three commonly used fault diagnosis datasets. We report the test accuracy results of our method, which are compared with state-of-the-art methods: Conditional Domain Adversarial Network (CDAN) [15], Domain Adversarial Neural Network (DANN) [16], Adversarial Discriminative Domain Adaptation (ADDA) [29], Deep Adaptation Network (DAN) [30] and Unsupervised Adversarial Domain Adaptive Networks based on Minimum Domain Spacing (MDS_ADAN) [31]. The code used in the experiment is submited at https://github.com/fonderxu/Transfer-Learning.git and implemented by PyTorch.

4.1. Dataset

CWRU Bearing Dataset [32]. The first dataset comes from Case Western Reserve University Bearing Data Center, and is one of the most used bearing datasets for machine fault diagnosis. The experiment uses the driving end data at the sampling frequency of 12kHz, which includes four health conditions: Normal Condition (NC), Inner Fault (IF), Outer Fault (OF) and Rolling Fault (RF).The damage diameters are 0.007 inches, 0.014 inches, 0.021 inches, respectively, which means it is a 10-classification task. Four domains are formed by applying four kinds of loads to the motor: 0 hp, 1 hp, 2 hp, and 3 hp.The design of transfer task on four domains is the same as [31].

Planetary Gearbox Dataset. The second dataset comes from QPZZ-II Rotating Machinery Vibration Analysis and Fault Diagnosis Test Platform System. The sampling frequency is 2.56HZ. The vibration sensor used has a total of 9 channels. The experiment selects the data of one channel for detection. There are 4 kinds of health conditions: Normal Condition (NC), Gear Pitting (GP), Mixed Fault Of Gear Pitting And Pinion Wear (GP+GW), and Pinion Wear (GW). Under the condition of 880rpm rotating speed, four kinds of loads are applied to the experimental platform to form the data domain: 0A, 0.05A, 0.1A and 0.2A. Similarly, we adopt the same transfer task design as [31].

JNU Bearing Dataset. The third set of data comes from The Centrifugal Fan System for Rolling Bearing Fault Diagnosis Testbed of JiangNan University. The health status is the same as that of CWRU bearing dataset. It is worth noting that the number of normal state data in the dataset is more than the

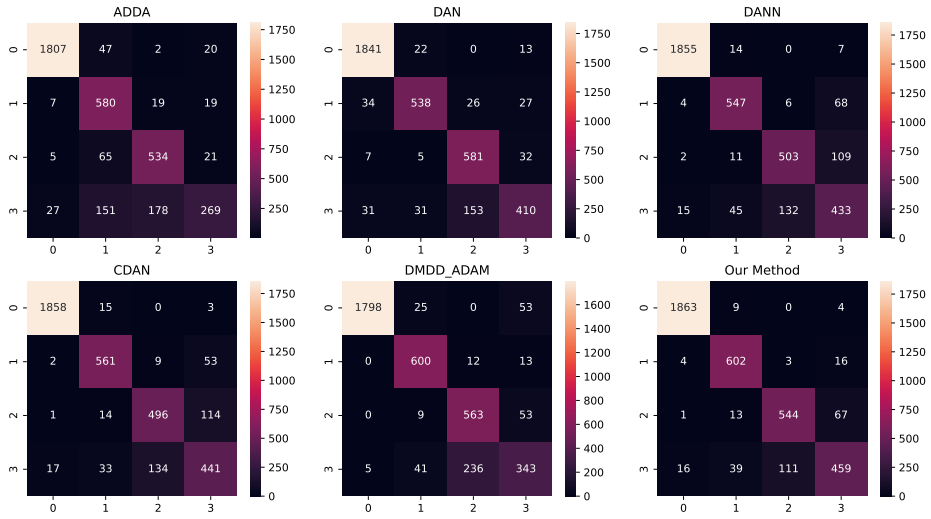

Fig. 3. The heat map of the confusion matrix on transfer task A->C of JNU Dataset. The vertical and the horizontal axis shows the true label and predicted label respectively.

sum of the other three state data, which means that it is an unevenly distributed dataset. We obtained data domains for 3 different motor speeds: 600rpm, 800rpm and 1000rpm.The details of the transfer tasks constituted by this dataset are shown in Tab.1.

4.2. Implementation Details

For a fair comparison, all methods employ the same feature extraction network, and we follow the standard evaluation protocol for UDA [33] that all labeled source and unlabeled target data are used for training. The batch size of the dataloader is set to 220, 205, and 220 for CWRU, Planetary Gearbox, and JNU, respectively. The total number of epochs on all datasets is set to 20, and the iterations of each epoch are 100 except that it is set to 200 on the Jiangnan University bearing dataset. The optimizer follows the standard back-propagation process, using Adam as the optimizer, and its initial learning rate lr_{init} and weight decay $lr\_{weight-decay}$ are both set to: 1e-4. The learning rate adjustment formula is: $lr\_{new} = \frac{lr\_{init}}{(1+lr\_{gamma} \cdot float(x))^{-lr\_{decay}}}$, where lr_{gamma} and lr_{decay} are: 1e-3 and 0.75 represent initial learning, x is the number of iterations. We adopt a progressive domain adaptation strategy to reduce the negative transfer initially optimized by the minimization entropy, by reweighting entropy loss $\lambda_1 = 0.1 \cdot (\frac{2}{1+e^{-10 \cdot \frac{epoch\_current}{epoch\_current}}} - 1)$, where $epoch\_{current}$ and $epoch\_{total}$ represent the current epoch in training and the total training epoch, respectively and set $\lambda_2 = 1$ throughout.

4.3. Experimental Results & Analysis

Experimental Result. Result are showed in Tab.2 (CWRU Bearing Dataset), Tab.3 (Planetary Gearbox Dataset) and Tab.4 (JNU Bearing Dataset) respectively. The average accuracy of our method outperforms all baseline methods on two of the three datasets. Specifically, our method outperforms all

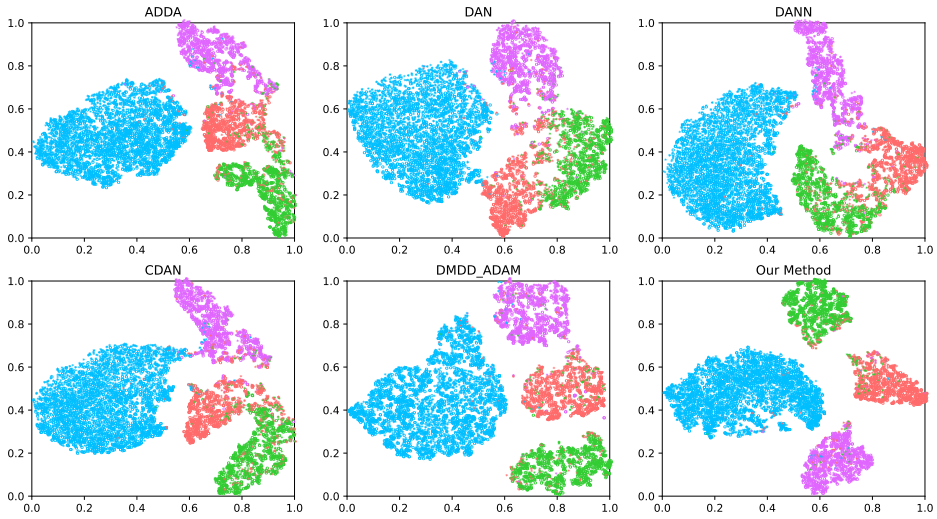

Fig. 4. Visualization of feature distribution on transfer task A->C of JNU Dataset. Different colors represent different categories of data (class 0 by DeepSkyBlue, class 1 by MediumOrchid1, class 2 by LimeGreen and class 3 by IndianRed1 respectively) and different shapes represent different domains (source data by '*' and target data by 'o', respectively).

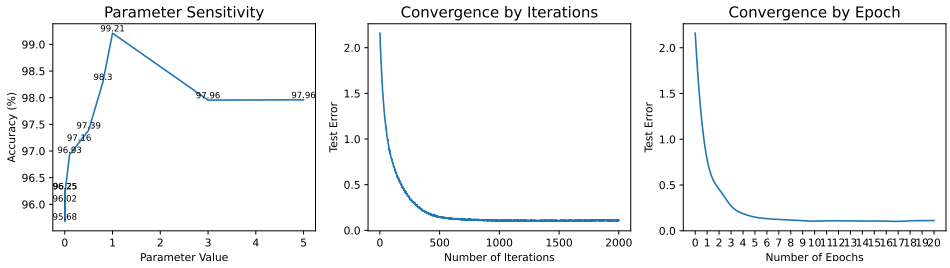

Fig. 5. Analysis on parameter sensitivity and convergence performance. Parameter sensitivity and Convergence by Iteration are analyzed on transfer task A->B of CWRU and Convergence by Epoch is explored on transfer task A->B of JNU.

comparative methods on CWRU and JNU Dataset and outperforms the third-best method CDAN by a large margin on JNU Dataset (by 0.69%). On Gearbox Dataset, the average accuracy of our method is only 0.33% lower than the best performing DMDD_ADAM, and surpasses the third best performing method on average by 1.89% mean accuracy. Furthermore, average accuracy of our method outperforms ADDA (45.82%, 31.28%, and 7.72%), DAN (1.61%, 11.44%, and 2.35%), and DANN (0.1%, 8.13%, and 1.15%) by a large margin on three datasets.

Results Analysis. In order to analyze the effectiveness and feasibility of the proposed method, we compare the confusion matrix of the classification results of the six methods. The heat map of the confusion matrix, taking the transfer task JNU[A->C] as an example, is shown in the Fig.3. Compared with other methods, our method can classify NC (category number 0), IF (category number 1) and RF (category number 3) data more accurately. At the same time, our method also shows a good classification performance on the OF (category number 3) data, which makes the average accuracy of our method

Table 2
Results On CWRU Bearing dataset

Methods	A->B	A->C	A->D	B->C	B->D	C->D	Avg
ADDA[29]	56.48	99.21	52.39	50.34	63.64	25.11	54.05
DAN[30]	96.36	98.86	97.84	99.32	97.61	99.55	98.26
DANN[16]	98.75	100.00	100.00	100.00	97.61	100.00	99.77
DMDD_ADAM[31]	95.34	99.66	99.77	100.00	100.00	99.89	99.11
CDAN[15]	98.86	100.00	100.00	100.00	100.00	100.00	99.81
Our Method	99.21	100.00	100.00	100.00	99.89	100.00	**99.85**

Table 3
Results On Planetary Gearbox dataset

Methods	A->B	A->C	A->D	B->C	B->D	C->D	Avg
ADDA[29]	86.83	61.71	87.56	73.05	48.05	52.81	68.33
DAN[30]	49.39	95.98	96.10	98.42	90.12	99.02	88.17
DANN[16]	50.00	100.00	100.00	100.00	98.90	100.00	91.48
DMDD_ADAM[31]	100.00	100.00	99.76	100.00	99.88	100.00	**99.94**
CDAN[15]	100.00	100.00	100.00	100.00	87.07	100.00	97.85
Our Method	100.00	100.00	100.00	100.00	98.42	100.00	99.74

Table 4
Results On JNU Bearing

Methods	A->B	A->C	B->C	Avg
ADDA[29]	81.77	85.04	91.23	86.10
DAN[30]	91.42	89.84	92.88	91.38
DANN[16]	94.16	88.99	94.59	91.59
DMDD_ADAM[31]	93.02	88.08	93.68	92.58
CDAN[15]	94.40	89.47	95.26	93.04
Our Method	94.40	92.75	94.94	**93.73**

surpass all method. To analyze effectiveness of model, visualization of feature distribution on the above tasks is been illustrated in Fig.4. The features extracted by ADDA, DAN, DANN and CDAN methods exist obvious inter-class confusion problems. This makes it difficult to be accurately classified for the data corresponding to the confused class. Meanwhile, the distance between distributions of different class features extracted by DMDD_ADAM is too close and features at the class boundary are prone to be incorrectly classified. In contrast, our method increases both the density of features of the same class and the distance between features of different classes, which improves data separability and transferability.

4.4. Experimental Analysis

1) **Parameter sensitivity**. In our work, we introduce two trade-off parameters λ_1 and λ_2 to control $\mathcal{L}_{em}$ and $\mathcal{L}_{adv}$ and set $\lambda_1 = 0.1 \cdot (\frac{2}{1+e^{-10 * \frac{epoch\_current}{epoch\_current}}} - 1)$ and $\lambda_2 = 1$ respectively. To analyze the effectiveness of adversarial components, the curve of prediction accuracy on transfer task CWRU[A->B], under different values of λ_2, is shown in Fig.5. The results show that the prediction accuracy of our method exceeds 95.68% by all values, which proves that our method is parameter robust.

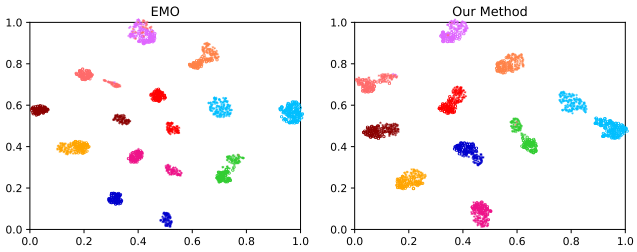

Fig. 6. Visualization of feature distribution on transfer task A->B of CWRU Dataset. Different colors represent different categories of data.

2) **Convergence**. From Fig.5, as the training progresses, the test error decreases steadily and finally converges. Note that our model carries out backpropagation under the superposition of three losses and adversarial networks are difficult to train, and the variation in test error shows that our approach is effective.

3) **Ablation experiments**. Among these 15 transfer tasks, the five most difficult ones, which come from the five tasks with the lowest average accuracy on the three high-performance model: CDAN[15], DMDD-ADAM[30] and our method, are selected to explore the effect of adversarial components in our proposed method. The results in Tab.5 show that after adding the adversarial component, the accuracy of three groups of transfer tasks has been improved and average accuracy has increased by 0.8 percentage points, which indicates the effectiveness of the adversarial component. To step further, we analyze the visualization of feature distribution, extracted by model without adding adversarial components (EMO) and our method respectively, on transfer task CWRU[A->B]. As is depicted in Fig.6, our method effectively align feature between domains without causing category dislocation which is better EMO. Based on the linear kernel Maximum Mean Discrepancy (MMD) [34] measure of feature distribution distance, the inter-domain distance are 5.7907 by EMO and 3.4025 by our method respectively, confirming our observations.

Table 5
Ablation Experiment

Value of λ_2	**CWRU[A->B]**	**Gearbox[B->D]**	**JNU[A->B]**	**JNU[A->C]**	**JNU[B->C]**	**Avg**
0	96.023	**98.902**	**94.588**	91.522	94.668	95.1406
1	**99.21**	98.4150	94.401	**92.749**	**94.935**	**95.941**

Table 6
Training cost analysis

Methods	**Total Params**	**Total Memory**	**Total MAdd**	**Total Flops**	**Total MemR+W**
ADDA[29]	3006903	0.52	25.02	12.58	12.43
DAN[30]	5366199	0.53	29.74	14.94	21.44
DANN[16]	1689014	0.5	22.4	11.27	7.39
DMDD_ADAM[31]	3006903	0.52	25.02	12.58	12.43
CDAN[15]	1697488	0.5	22.42	11.285	7.421
Our Method	1671487	0.5	22.37	11.26	7.32

4) **Training cost analysis**. The total cost analysis from parameters (Params), memory (Memory), floating-point operations (Flops), multiply-add operations (MAdd), and memory of read and write (MemR+W) is shown in Tab.6. Our method reaches the lowest value on a total of 5 indicators, which indicate we have low training cost. This is mainly due to the design of the integrated classification-discriminator, which eliminates the need to introduce an independent domain discriminant network in DAT.

5. Conclusion

Entropy minimization is powerful method in domain adaptation, but it will bring about negative transfer while pseudo-labels are inconsistent with the truth label. In this article, we suggest that the negative transfer caused by minimizing entropy can be countered by increasing the model prediction accuracy by introducing DAT into entropy minimization, while aligning distributions between source and target data. However, without guided by label distribution similarity, DAT might introduce negative transfer by the forced alignment of class-misaligned divergence. Therefore, we propose a method called entropy minimization and DAT guided by label distribution similarity. Extensive experiments demonstrate the effectiveness of our method. In the future, we will study and optimize the proposed method and verify the effectiveness of the proposed method in more fields.

References

[1] F. Zhao, C. Zhang, N. Dong, Z. You and Z. Wu, A Uniform Framework for Anomaly Detection in Deep Neural Networks, *Neural Processing Letters* (2022), 1–22.

[2] Y.H. Bhosale and K.S. Patnaik, Application of Deep Learning Techniques in Diagnosis of Covid-19 (Coronavirus): A Systematic Review, *Neural Processing Letters* (2022), 1–53.

[3] Y. Mao, S. Wang, D. Yu and J. Zhao, Automatic image detection of multi-type surface defects on wind turbine blades based on cascade deep learning network, *Intelligent Data Analysis* **25**(2) (2021), 463–482.

[4] S. Xiao, Y. Li, Y. Ye, L. Chen, S. Pu, Z. Zhao, J. Shao and J. Xiao, Hierarchical Temporal Fusion of Multi-grained Attention Features for Video Question Answering, *Neural Processing Letters* **52**(2) (2020), 993–1003.

[5] O. Habimana, Y. Li, R. Li, X. Gu and G. Yu, Sentiment analysis using deep learning approaches: an overview, *Science China Information Sciences* **63**(1) (2020), 1–36.

[6] Y. Madadi, V. Seydi, K. Nasrollahi, R. Hosseini and T.B. Moeslund, Deep visual unsupervised domain adaptation for classification tasks: a survey, *IET Image Processing* **14**(14) (2020), 3283–3299.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: *International conference on machine learning*, PMLR, 2014, pp. 647–655.

[8] A. Ma, J. Li, K. Lu, L. Zhu and H.T. Shen, Adversarial entropy optimization for unsupervised domain adaptation, *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[9] M. Long, H. Zhu, J. Wang and M.I. Jordan, Unsupervised domain adaptation with residual transfer networks, *Advances in neural information processing systems* **29** (2016).

[10] X. Wu, S. Zhang, Q. Zhou, Z. Yang, C. Zhao and L.J. Latecki, Entropy minimization versus diversity maximization for domain adaptation, *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[11] Y. Grandvalet and Y. Bengio, Semi-supervised learning by entropy minimization, *Advances in neural information processing systems* **17** (2004).

[12] T.-H. Vu, H. Jain, M. Bucher, M. Cord and P. Pérez, Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[13] V. Prabhu, S. Khare, D. Kartik and J. Hoffman, Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8558–8567.

[14] P. Morerio, J. Cavazza and V. Murino, Minimal-entropy correlation alignment for unsupervised deep domain adaptation, *arXiv preprint arXiv:1711.10288* (2017).

[15] M. Long, Z. Cao, J. Wang and M.I. Jordan, Conditional adversarial domain adaptation, *Advances in neural information processing systems* **31** (2018).
[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, Domain-adversarial training of neural networks, *The journal of machine learning research* **17**(1) (2016), 2096–2030.
[17] L. Tran, K. Sohn, X. Yu, X. Liu and M. Chandraker, Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2672–2681.
[18] X. Liu, Z. Guo, S. Li, F. Xing, J. You, C.-C.J. Kuo, G. El Fakhri and J. Woo, Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10367–10376.
[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial networks, *Communications of the ACM* **63**(11) (2020), 139–144.
[20] B. Li, Y. Wang, T. Che, S. Zhang, S. Zhao, P. Xu, W. Zhou, Y. Bengio and K. Keutzer, Rethinking Distributional Matching Based Domain Adaptation, 2020.
[21] M. Long, H. Zhu, J. Wang and M.I. Jordan, Unsupervised domain adaptation with residual transfer networks, *Advances in neural information processing systems* **29** (2016).
[22] K. Saito, D. Kim, S. Sclaroff, T. Darrell and K. Saenko, Semi-supervised domain adaptation via minimax entropy, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8050–8058.
[23] Y. Ganin and V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.
[24] S. Ben-David, J. Blitzer, K. Crammer and F. Pereira, Analysis of representations for domain adaptation, *Advances in neural information processing systems* **19** (2006).
[25] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira and J. Wortman, Learning bounds for domain adaptation, *Advances in neural information processing systems* **20** (2007).
[26] V.K. Kurmi and V.P. Namboodiri, Looking back at labels: A class based domain adaptation technique, in: *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019, pp. 1–8.
[27] H. Tang and K. Jia, Discriminative Adversarial Domain Adaptation, *national conference on artificial intelligence* (2019).
[28] X. Jiang, Q. Lao, S. Matwin and M. Havaei, Implicit class-conditioned domain alignment for unsupervised domain adaptation, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 4816–4827.
[29] E. Tzeng, J. Hoffman, K. Saenko and T. Darrell, Adversarial discriminative domain adaptation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.
[30] M. Long, Y. Cao, J. Wang and M. Jordan, Learning transferable features with deep adaptation networks, in: *International conference on machine learning*, PMLR, 2015, pp. 97–105.
[31] Z. Ruicong, B. Yu, L. Zhongtian, W. Qinle and L. Yonggang, Unsupervised adversarial domain adaptive for fault detection based on minimum domain spacing, *Advances in Mechanical Engineering* **14**(3) (2022), 16878132221088647.
[32] K. Loparo, Case western reserve university bearing data centre website, 2012.
[33] X. Wang, L. Li, W. Ye, M. Long and J. Wang, Transferable attention for domain adaptation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 5345–5352.
[34] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.-P. Kriegel, B. Schölkopf and A.J. Smola, Integrating structured biological data by kernel maximum mean discrepancy, *Bioinformatics* **22**(14) (2006), e49–e57.