Towards Discriminability and Diversity: Batch Nuclear-norm Maximization under Label Insufficient Situations

Shuhao Cui[1,2] Shuhui Wang[1*] Junbao Zhuo[1,2] Liang Li[1] Qingming Huang[1,2,3] Qi Tian[4]

[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Peng Cheng Laboratory, Shenzhen, China [4]Noah's Ark Lab, Huawei Technologies
{cuishuhao18s, wangshuhui, liang.li}@ict.ac.cn, junbao.zhuo@vipl.ict.ac.cn,
qmhuang@ucas.ac.cn, tian.qi1@huawei.com

Abstract

The learning of the deep networks largely relies on the data with human-annotated labels. In some label insufficient situations, the performance degrades on the decision boundary with high data density. A common solution is to directly minimize the Shannon Entropy, but the side effect caused by entropy minimization, i.e., reduction of the prediction diversity, is mostly ignored. To address this issue, we reinvestigate the structure of classification output matrix of a randomly selected data batch. We find by theoretical analysis that the prediction discriminability and diversity could be separately measured by the Frobenius-norm and rank of the batch output matrix. Besides, the nuclear-norm is an upperbound of the Frobenius-norm, and a convex approximation of the matrix rank. Accordingly, to improve both discriminability and diversity, we propose Batch Nuclear-norm Maximization (BNM) on the output matrix. BNM could boost the learning under typical label insufficient learning scenarios, such as semi-supervised learning, domain adaptation and open domain recognition. On these tasks, extensive experimental results show that BNM outperforms competitors and works well with existing well-known methods. The code is available at https://github.com/cuishuhao/BNM.

1. Introduction

Deep neural networks have achieved large success in most computer vision applications. Despite the success already achieved, deep models in visual learning tasks largely rely on vast amounts of labeled data, where the labeling process is both time-consuming and expensive. Without sufficient amount of labeled training data, as a common consequence, spurious predictions will be made even with a subtle departure from the training samples. Actually, in most applications, there exists large discrepancy between training data and real-world testing data. The discrepancy could lead to annoying ambiguous predictions, especially under label insufficient situations. In this paper, we focus on enhancing the model learning capability by reducing the ambiguous predictions.

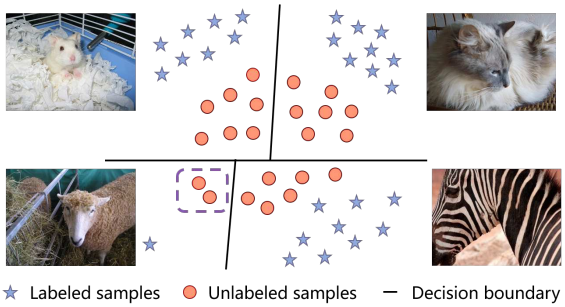


Figure 1. Illustration of problem. Beyond the discriminability, we also focus on the minority category with few samples bounded in the dotted frame at the bottom left part. The influence of minority category tends to be reduced in direct entropy minimization, resulting in the degradation of the category prediction diversity.

When deep models are applied to unlabeled examples, the prediction discriminability is always low, due to the large data density near the decision boundary. To reduce the large data density, most methods resort to classical Shannon Entropy theory [41]. In semi-supervised learning, the prediction entropy for unlabeled examples is directly minimized by [13], and further combined with Virtual Adversarial Training [30] for better results. In domain adaptation, the entropy minimization is also utilized in [27, 47] to strengthen the discriminability on the unlabeled domain. Meanwhile, from another aspect, to encourage the prediction diversity, methods in [42, 55] utilize a balance con-

*Corresponding author.

straint to equilibrate the distribution between different categories. Methods in [58, 57] take the model predictions as pseudo-labels of the statistical distribution, to further enhance the robustness of prediction on categories with less data instances.

However, entropy-based methods suffer from the side effect of the entropy minimization, *i.e.*, reduction of prediction diversity. Entropy minimization pushes the examples to nearby examples far from the decision boundary. Since there are more examples in majority categories, examples are prone to be pushed into majority categories, including the examples actually belonging to minority categories. As shown in Figure 1, the points in the dotted frame will be misclassified, if all the samples are classified into the other categories, resulting in the reduced prediction diversity. Towards higher prediction diversity, the balance constraint requires prior knowledge on minority categories, but it is difficult to obtain. Some methods rely on the prior knowledge estimated by pseudo-labels, to increase the prediction probability on minority categories. Nevertheless, it appears that the guidance provided by prior knowledge is usually less useful and straightforward towards more accurate prediction.

In this paper, we reinvestigate the above issues, and try to enforce the prediction discriminability and diversity of the unlabeled data. We start by looking at the structure of classification output matrix of a randomly selected data batch. We find by theoretical analysis that the discriminability and diversity could be measured by the Frobenius-norm and rank of the batch output matrix, respectively. The nuclear-norm of a matrix is bounded by the Frobenius-norm of the matrix. Maximizing nuclear-norm ensures large Frobenius-norm of the batch matrix, which leads to increased discriminability. The nuclear-norm of batch matrix is also a convex approximation of the matrix rank, which refers to the prediction diversity. Accordingly, we present Batch Nuclear-norm Maximization (BNM), an approach to maximize the nuclear-norm of the batch output matrix. Thus BNM could simultaneously enhance the discriminability and diversity of the prediction results.

We apply BNM to semi-supervised learning, domain adaptation and unsupervised open domain recognition to validate its effectiveness. Experiments show that our method outperforms other well-established methods on four datasets. Among these tasks, we achieve state-of-the-art results on unsupervised open domain recognition. For further validation, we observe that BNM leads to better average prediction diversity on the batch samples in the experiments. Our contribution is summarized as follows:

- We theoretically prove that the discriminability and diversity of the prediction output can be measured by Frobenius-norm and rank of the batch output matrix.
- We propose Batch Nuclear-norm Maximization, which is a new learning paradigm that achieves better discriminability and diversity under label insufficent learning situations.
- We achieve promising performance on semi-supervised learning, domain adaptation and open domain recognition.

2. Related Work

In this paper, we analyze the label insufficient situations on three typical tasks, *i.e.*, semi-supervised learning [20, 31, 21, 48], domain adaptation [14, 25] and unsupervised open domain recognition. Semi-supervised learning [44, 30] leverages the unlabeled examples to improve the robustness of model. Domain adaptation [18, 39, 26, 56] reduces the domain discrepancy between labeled and unlabeled examples. Unsupervised Open Domain Recognition [55] considers a more realistic circumstance than domain adaptation, where some categories in unlabeled domain are unseen in labeled domain.

Among the tasks, they all face the problem of the rich data distribution near the decision boundary. To reduce the ambiguous predictions, most methods resort to the Shannon Entropy [41] to model uncertainty. In semi-supervised learning, reducing the entropy of the classification responses for unlabeled domain is adopted in [13]. Meanwhile, the entropy minimization is further applied to Virtual Adversarial Training in [30] and implicitly modeled by pseudo-labels [2]. In domain adaptation, entropy minimization is utilized in [27, 47] to obtain reliable prediction on unlabeled examples. The entropy minimization is further modified into maximum squares loss in [6] to lower the influence of easy-to-transfer samples.

To maintain prediction diversity on the minority categories, a direct thought is resorting to imbalanced learning [16]. Existing imbalanced learning methods such as [42, 55] enforce the ratio of predictions on minority categories to be appropriately higher. However, they demand prior knowledge on the category distribution. Without the prior knowledge, the predictions are taken as pseudo-labels in [57, 58] to approximate the category distribution, while our method is performed in a data driven manner that is free from any form of prior knowledge. From another aspect of increasing diversity, Determinantal Point Processes (DPPs) [23] act probabilistically to capture the balance between quality and diversity within sets, but suffer from the large computation time.

In this paper, we analyze the problem from the perspective of matrix analysis [4, 3], which has already been widely applied to many computer vision tasks. As a popular perspective, matrix completion is based on the assumption that the noisy data brings extra components to the matrix. To

reduce the influence of the extra components, minimizing nuclear-norm of the matrix has been applied to image denoising [15], image restoration [9] and many other tasks. In comparison to the above methods, we aim to explore the extra information in the matrix, thus the nuclear-norm of the matrix is maximized towards more prediction diversity. Recently, BSP [7] penalizes the largest singular values of the batch feature matrix to boost the feature discriminability for domain adaptation, while we analyze the batch classification response matrix to increase both the prediction discriminability and diversity.

3. Method

3.1. Measuring Discriminability with F-norm

In the training process of a deep neural network, we start by looking at the prediction outputs on a data batch with B randomly selected unlabeled samples. Denote the number of categories as C, and we represent the batch prediction output matrix as $\mathrm{A} \in \mathbb{R}^{B \times C}$, which satisfies:

$$\begin{aligned} \sum\nolimits_{j=1}^{C} \mathrm{A}_{i,j} = 1 \quad & \forall i \in 1...B \\ \mathrm{A}_{i,j} \geq 0 \quad & \forall i \in 1...B, j \in 1...C, \end{aligned} \tag{1}$$

where deep methods could achieve well-performed response matrix A by training with sufficient number of labeled samples. However, in label insufficient situations, the discrepancy between labeled and unlabeled data might result in the high-density regions of the marginal data distribution near the task-specific decision boundary. Since the ambiguous samples are easy to be misclassified, we focus on optimizing prediction results of unlabeled samples by increasing the discriminability.

Actually, higher discriminability means less uncertainty in the prediction. To measure the uncertainty, most methods resort to Shannon Entropy [41], which is always denoted as entropy for simplicity. The entropy could be calculated as follows:

$$H(\mathrm{A}) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} \mathrm{A}_{i,j} log(\mathrm{A}_{i,j}). \tag{2}$$

Same as [13, 30, 27, 47], we could directly minimize $H(\mathrm{A})$, towards lessened uncertainty and more discriminability. When $H(\mathrm{A})$ reaches the minimum, only one entry is 1 and other $C-1$ entries are 0 in each row of A_i, *i.e.*, $\mathrm{A}_{i,j} \in \{0,1\} \quad \forall i \in 1...B, j \in 1...C$. The minimum exactly satisfies the highest prediction discriminability of A, where each prediction A_i is fully determined.

Other functions could improve the prediction discriminability, by push A to the same minimum with $H(\mathrm{A})$. We choose to calculate Frobenius-norm (F-norm) $\|\mathrm{A}\|_F$, as follows:

$$\|\mathrm{A}\|_F = \sqrt{\sum_{i=1}^{B}\sum_{j=1}^{C} |\mathrm{A}_{i,j}|^2}. \tag{3}$$

We prove that $H(\mathrm{A})$ and $\|\mathrm{A}\|_F$ have strict opposite monotonicity and the minimum of $H(\mathrm{A})$ and the maximum of $\|\mathrm{A}\|_F$ could be achieved at the same value in Supplementary. Particularly, according to inequality of arithmetic and geometric means, the upper-bound of $\|\mathrm{A}\|_F$ could be calculated as:

$$\begin{aligned} \|\mathrm{A}\|_F &\leq \sqrt{\sum_{i=1}^{B}(\sum_{j=1}^{C} \mathrm{A}_{i,j}) \cdot (\sum_{j=1}^{C} \mathrm{A}_{i,j})} \\ &= \sqrt{\sum_{i=1}^{B} 1 \cdot 1} \quad = \sqrt{B}. \end{aligned} \tag{4}$$

The upper-bound of $\|\mathrm{A}\|_F$ could be achieved in the same A with the minimum of $H(\mathrm{A})$. Thus prediction discriminability could also be enhanced by maximizing $\|\mathrm{A}\|_F$.

3.2. Measuring Diversity with Matrix Rank

It is normal in randomly selected batch of B examples that some categories dominate the samples, while other categories contain less or even no samples. In this case, a model trained with entropy minimization or F-norm maximization tends to classify samples near the decision boundary to the majority categories. The continuous convergence to the majority categories reduces the prediction diversity, which is harmful to the overall prediction accuracy. To improve the accuracy, different from [42, 55, 58, 57], we aim to maintain the prediction diversity by analyzing the batch output matrix A to model the diversity.

To model prediction diversity, we start by looking at the fixed batch of B unlabeled samples in matrix A. The number of categories in the predictions is expected to be a constant on average. If this constant becomes larger, the prediction approach could obtain more diversity. Thus the prediction diversity could be measured by the number of predicted categories in the batch output matrix A.

We further analyze the number of categories and the predicted vectors in A. Two randomly selected prediction vectors, *i.e.*, A_i and A_k, could be linearly independent when A_i and A_k belong to different categories. When A_i and A_k belong to the same category and $\|\mathrm{A}\|_F$ is near $\sqrt{B}$, the differences between A_i and A_k are tiny. Then A_i and A_k could be approximately regarded as linearly dependent. The largest number of linear independent vectors is called the matrix rank. Thus $rank(\mathrm{A})$ could be an approximation on the number of predicted categories in A, if $\|\mathrm{A}\|_F$ is near the upper-bound $\sqrt{B}$.

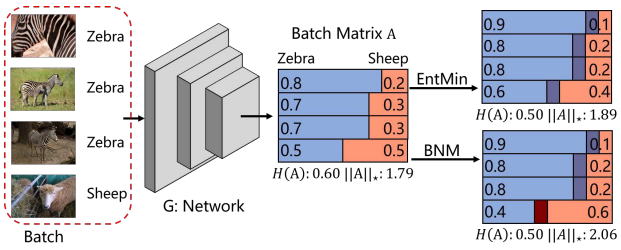


Figure 2. Illustration of comparison between the effect of BNM and Entropy Minimization (EntMin) in a toy example with two categories and batch size 4. The dark region means the increase of the variable, *i.e.*, the dark blue (red) represents the increase of blue (red) variable. $H(\mathrm{A})$ represents the entropy value and $\|\mathrm{A}\|_\star$ represent the value of nuclear-norm.

Based on the above analysis, the prediction diversity could be approximately represented by $rank(\mathrm{A})$, when $\|\mathrm{A}\|_F$ is near $\sqrt{B}$. Accordingly, we could maximize $rank(\mathrm{A})$ to maintain prediction diversity. Apparently, the maximum value of $rank(\mathrm{A})$ is $\min(B, C)$. When $B \geq C$, the maximum value is C, which firmly guarantees that the prediction diversity on this batch achieves the maximum. However, when $B < C$, the maximum value is less than C, it still enforces that the predictions on the batch samples should be as diverse as possible, though there is no guarantee that all the categories will be assigned to at least one sample. Therefore, maximization of $rank(\mathrm{A})$ could ensure the diversity in any case.

3.3. Batch Nuclear-norm Maximization

For a normal matrix, the calculation of the matrix rank is an NP-hard non-convex problem, and we could not directly restrain the rank of matrix A. Theorem in [10] shows that when $\|\mathrm{A}\|_F \leq 1$, the convex envelope of $rank(\mathrm{A})$ is the nuclear-norm $\|\mathrm{A}\|_\star$. In our situation, different from above theorem, we have $\|\mathrm{A}\|_F \leq \sqrt{B}$ as shown in Eqn. 4. Thus the convex envelope of $rank(\mathrm{A})$ becomes $\|\mathrm{A}\|_\star / \sqrt{B}$, which is also proportional to $\|\mathrm{A}\|_\star$. Meanwhile, $rank(\mathrm{A})$ could approximately represent the diversity, when $\|\mathrm{A}\|_F$ is near the upper-bound, as described in Sec. 3.2. Therefore, if $\|\mathrm{A}\|_F$ is near $\sqrt{B}$, the prediction diversity could be approximately represented by $\|\mathrm{A}\|_\star$. Also, maximizing $\|\mathrm{A}\|_\star$ could ensure higher prediction diversity.

In [10, 37, 43], the relationship of the range between $\|\mathrm{A}\|_\star$ and $\|\mathrm{A}\|_F$ could be expressed as follows:

$$\frac{1}{\sqrt{D}}\|\mathrm{A}\|_\star \leq \|\mathrm{A}\|_F \leq \|\mathrm{A}\|_\star \leq \sqrt{D} \cdot \|\mathrm{A}\|_F \tag{5}$$

where $D = \min(B, C)$. This shows that $\|\mathrm{A}\|_\star$ and $\|\mathrm{A}\|_F$ could bound each other. Therefore, $\|\mathrm{A}\|_F$ tends to be larger, if $\|\mathrm{A}\|_\star$ becomes larger. Since maximizing $\|\mathrm{A}\|_F$ could improve the discriminability described in Sec. 3.1, maximizing $\|\mathrm{A}\|_\star$ also contributes to the improvement on prediction discriminability.

Due to the relationship between $\|\mathrm{A}\|_\star$ and $\|\mathrm{A}\|_F$, and the fact that upper-bound of $\|\mathrm{A}\|_F$ is $\sqrt{B}$ in Eqn. 4, we could calculate the maximum of $\|\mathrm{A}\|_\star$ as follows:

$$\|\mathrm{A}\|_\star \leq \sqrt{D} \cdot \|\mathrm{A}\|_F \leq \sqrt{D \cdot B}, \tag{6}$$

where we could find that the influence factor of $\|\mathrm{A}\|_\star$ could be separated into two parts, respectively corresponding to the two inequality conditions in the equation. The first inequality corresponds to the diversity, and the second corresponds to the discriminability. When the diversity is larger, the rank of A tends to be larger and $\|\mathrm{A}\|_\star$ tends to increase. Similarly, when the discriminability becomes larger, $\|\mathrm{A}\|_F$ tends to increase and $\|\mathrm{A}\|_\star$ tends to be larger.

Based on the above findings, maximizing $\|\mathrm{A}\|_\star$ could lead to the improvement on both the prediction discriminability and diversity. Thus to improve discriminability and diversity, we propose Batch Nuclear-norm Maximization by maximizing the nuclear-norm of the batch matrix A, where A represents the classification responses on a batch of B randomly selected samples. For better comprehension of the effect of BNM, we build two toy examples, one explaining BNM in the maximum case, and another in ordinary situations.

In the first example, we assume B and C are 2. In this case, A could be expressed as:

$$\mathrm{A} = \begin{bmatrix} x & 1-x \\ y & 1-y \end{bmatrix}, \tag{7}$$

where x and y are variables. Thus the entropy, F-norm and nuclear-norm could be calculated as:

$$\begin{aligned} H(\mathrm{A}) =& -x\log(x) - (1-x)\log(1-x) - y\log(y) \\ & - (1-y)\log(1-y) \\ \|\mathrm{A}\|_F =& \sqrt{x^2 + (1-x)^2 + y^2 + (1-y)^2} \\ \|\mathrm{A}\|_\star =& \sqrt{x^2 + (1-x)^2 + y^2 + (1-y)^2 + 2|y-x|}, \end{aligned} \tag{8}$$

where the calculation of $\|\mathrm{A}\|_\star$ is described in Supplementary. For entropy and F-norm, there is no constraint limiting the relationship between x and y, thus entropy and F-norm could reach the optimal solution when:

$$\mathrm{A} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}. \tag{9}$$

But $\|\mathrm{A}\|_\star$ would reach the optimal solution when:

$$\mathrm{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{10}$$

where $\|\mathrm{A}\|_\star$ tries to maintain diversity by maximizing the prediction divergence among the data batch in this example.

The second example is shown in Figure 2, where we assume that there are only two categories, *i.e.*, zebra and sheep. In this data batch with three zebras and a goat, the zebra category is the majority category. The matrices after Entropy Minimizataion and BNM could obtain the same value on entropy as 0.50. But for the matrices, the value of nuclear-norm tends to be larger when the sheep is correctly classified. Direct entropy minimization tends to classify the batch examples into the majority category zebra. In comparison, BNM could maintain the prediction possibility of minority classes, and the image of sheep tends to be assigned with the infrequent but correct label.

Since BNM is computed by matrix operations, there remain concerns on the computational complexity. To obtain the nuclear-norm, we calculate all the singular values in the matrix A. The singular value decomposition computing on the matrix $\mathrm{A} \in \mathbb{R}^{B \times C}$ costs $O(\min(B^2C, BC^2))$ time. Since the batch size B is always small, the overall computational budget of $\|\mathrm{A}\|_\star$ is almost negligible in the training of deep networks.

3.4. Application

We apply BNM to three typical label insufficient situations, including semi-supervised learning, unsupervised domain adaptation and unsupervised open domain recognition. In the tasks, we are given labeled domain $\mathcal{D}_L$ and unlabeled domain $\mathcal{D}_U$. There are N_L labeled examples $\mathcal{D}_L = \{(x_i^L, y_i^L)_{i=1}^{N_L}\}$ in C categories and N_U unlabeled examples $\mathcal{D}_U = \{(x_i^U)_{i=1}^{N_U}\}$. In $\mathcal{D}_L$, the labels are denoted as $y_i^L = [y_{i1}^L, y_{i2}^L, ..., y_{iC}^L] \in \mathbb{R}^C$, where y_{ij}^L equals to 1 if x_i^L belongs to the jth category otherwise 0.

In these tasks, the classification responses are obtained by the deep network G, *i.e.*, $\mathrm{A}_i = G(x_i)$. The classification network consists of a feature extraction network, a classifier and a softmax layer. With randomly sampled batch size B_L examples $\{X^L, Y^L\}$ on the labeled domain, the classification loss on $\mathcal{D}_L$ could be calculated as:

$$\mathcal{L}_{cls} = \frac{1}{B_L} \left\| Y^L log(G(X^L)) \right\|_1, \tag{11}$$

where the classification loss could provide initial parameters for further optimization.

For learning on the unlabeled domain $\mathcal{D}_U$, on all the three tasks, we apply the method BNM introduced in Sec. 3.3 to the classification response matrix. With randomly sampled batch size B_U examples $\{X^U\}$, the classification response matrix on D_U could be denoted as $G(X^U)$. And the loss function of BNM can be formulated as:

$$\mathcal{L}_{bnm} = -\frac{1}{B_U} \left\| G(X^U) \right\|_\star, \tag{12}$$

where the neural network G is shared between both D_L and D_U. Minimizing $\mathcal{L}_{bnm}$ could reduce the data density near the decision boundary without losing diversity, which is more effective than typical entropy minimization. Meanwhile, the gradient of nuclear-norm could be calculated according to [33], thus $\mathcal{L}_{bnm}$ could be applied to the training process of gradient-based deep networks.

To train the network, we simultaneously optimize classification loss and BNM loss, *i.e.*, $\mathcal{L}_{cls}$ and $\mathcal{L}_{bnm}$ could be simultaneously optimized and combined with the parameter λ as follows:

$$\mathcal{L}_{all} = \frac{1}{B_L} \left\| Y^L log(G(X^L)) \right\|_1 - \frac{\lambda}{B_U} \left\| G(X^U) \right\|_\star. \tag{13}$$

By enforcing diversity, the key insight of BNM may be sacrificing a certain level of the prediction hit-rate on majority categories, to enhance the prediction hit-rate on minority categories. The samples belonging to the majority classes might be misclassified as minority classes, to increase the diversity. But the classification loss on the labeled training data would penalize the wrongly encouraged diversity in a batch, since classification loss is simultaneously minimized. Asymptotically, the network tends to produce more diverse prediction given that the samples can be correctly predicted. As a consequence, BNM is particularly useful to avoid prediction degradation for learning in label insufficient situations on datasets with both balanced and imbalanced category distributions.

4. Experiments

We apply our method to semi-supervised learning, unsupervised domain adaptation, and unsupervised open domain recognition. The experiments of the three tasks are done on CIFAR-100 [22], Office-31 [38], Office-Home [46] and I2AwA [55]. The results with the notion of $method^*$ are reproduced by us in the same environment with our methods, while other results are directly reported from the original papers. We also denote the direct entropy minimization, batch Frobenius-norm maximization as EntMin, BFM in our experiments. When applied to the existing methods, we denote Batch Nuclear-norm Maximization or entropy minimization as +BNM or +EntMin.

4.1. Semi-supervised Learning

CIFAR-100 [22] is a standard benchmark dataset for semi-supervised learning. We evaluate our method of BNM on CIFAR-100 with 5000 and 10000 labeled examples respectively. We utilize the ResNet [17] model, the same backbone with [32]. The batch size is fixed to 64 in our experiments. The experiments are implemented with Tensorflow [1]. We create 4 splits for each and report the mean and variance across the accuracy on different splits.

The results are shown in Table 1. In semi-supervised learning (SSL), direct entropy minimization could improve

Table 1. Accuracy(%) on the CIFAR-100 dataset for semi-supervised learning methods.

Method	5000	10000
Temporal Ensembling [24]	-	61.35±0.51
SNTG+Π-model [28]	-	62.03±0.29
ML+CCN+VAT [50]	56.58±0.31	**64.72**±0.23
ResNet [17]	39.73±0.33	49.55±0.28
EntMin	40.92±0.18	50.36±0.20
BNM	41.59±0.27	51.07±0.24
VAT* [30]	56.63±0.18	63.62±0.18
VAT+EntMin	56.97±0.21	64.48±0.22
VAT+BNM	**57.43**±0.24	64.61±0.15

the performance, while BNM outperforms entropy minimization. The improvement of BNM applied on a simple pretrained ResNet is moderate compared to other state-of-the-art well-designed SSL methods. However, working with other SSL methods such as VAT [30], BNM demonstrates more significant improvement, which is comparable to methods with more complicated mechanism such as ML+CCN+VAT [50]. Thus BNM is more suitable for cooperation with existing SSL methods, and performs better than entropy minimization in all cases.

4.2. Domain Adaptation

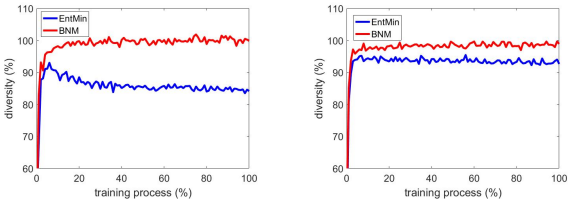
Office-31 [38] and Office-Home [46] are standard benchmarks for domain adaptation. Office-31 contains 4,652 images in 31 categories, and consists of three domains: Amazon (A), Webcam (W), and DSLR (D). We evaluate the methods across the three domains, resulting in six transfer tasks. Office-Home is a relative challenging dataset with 15,500 images in 65 categories. It has four significantly different domains: Artistic images (Ar), Clip Art (Cl), Product images (Pr), and Real-World images (Rw). There are 12 challenging transfer tasks among four domains in total.

We adopt ResNet-50 [17] pre-trained on ImageNet [8] as our backbone. The batch size is fixed to 36 in our experiments. The experiments are implemented with PyTorch [34]. BNM loss is directly combined with classification loss with the parameter λ fixed to 1. When BNM is combined with existing methods, the parameter λ is fixed to 0.1. For each method, we run four random experiments and report the average accuracy.

The results on Office-31 and Office-Home are shown in Table 2 and 3. On both Office-31 and Office-Home, as we expected, BFM obtains similar results with EntMin, while BNM achieves substantial improvement on average over other entropy-based methods. Surprisingly, BNM obtains superior results compared with popular alignment-based comparison methods. The results show that Batch Nuclear-norm Maximization is effective for domain adaptation, especially on the difficult tasks where the baseline accuracy is relatively low. Besides, we add BNM to existing CDAN [26], denoted as CDAN+BNM. The results of CDAN+BNM outperforms CDAN and CDAN+EntMin by a large margin, which shows that BNM could cooperate well with other methods. In summary, BNM could not only be regarded as a basic simple method for domain adaptation, but also an effective module contributing to existing methods.

Table 2. Accuracies (%) on Office-31 for ResNet50-based unsupervised domain adaptation methods.

Method	A→D	A→W	D→W	W→D	D→A	W→A	Avg
ResNet-50 [17]	68.9	68.4	96.7	99.3	62.5	60.7	76.1
GFK [12]	74.5	72.8	95.0	98.2	63.4	61.0	77.5
DAN [25]	78.6	80.5	97.1	99.6	63.6	62.8	80.4
DANN [11]	79.7	82.0	96.9	99.1	68.2	67.4	82.2
ADDA [45]	77.8	86.2	96.2	98.4	69.5	68.9	82.9
MaxSquare [6]	90.0	92.4	**99.1**	**100.0**	68.1	64.2	85.6
Simnet [36]	85.3	88.6	98.2	99.7	73.4	71.8	86.2
GTA [40]	87.7	89.5	97.9	99.8	72.8	71.4	86.5
MCD [39]	92.2	88.6	98.5	**100.0**	69.5	69.7	86.5
CBST [58]	86.5	87.8	98.5	**100.0**	70.9	71.2	85.8
CRST [57]	88.7	89.4	98.9	**100.0**	70.9	72.6	86.8
EntMin	86.0	87.9	98.4	**100.0**	67.0	63.7	83.8
BFM	87.7	86.9	98.5	**100.0**	67.6	63.0	84.0
BNM	90.3	91.5	98.5	**100.0**	70.9	71.6	87.1
CDAN [26]	**92.9**	**93.1**	98.6	**100.0**	71.0	69.3	87.5
CDAN+EntMin	92.0	91.2	98.7	**100.0**	70.7	71.0	87.3
CDAN+BNM	**92.9**	92.8	98.8	**100.0**	**73.5**	**73.8**	**88.6**



(a) Ar → Cl (b) Ar → Pl

Figure 3. Diversity ratio on Office-Home for domain adaptation, calculated as the predicted diversity divided by the ground truth diversity. The predicted (ground truth) diversity is measured by the average number of predicted (ground truth) categories in randomly sampled batches.

To validate that BNM could maintain the diversity in domain adaptation compared with entropy minimization, we show the diversity ratio in Office-Home on tasks of Ar → Cl and Ar → Pr in Figure 3. The diversity is measured by the mean matrix rank, i.e., mean number of predicted categories in randomly sampled batch. Thus the diversity ratio is measured by the mean predicted category number dividing the mean ground-truth category number. As shown in Figure 3(a), the diversity ratio of BNM is larger than that of the EntMin by a large margin in Ar → Cl. This phenomenon is normal since the rich samples near the decision boundary are mainly classified into the majority categories, reducing the diversity in the batch examples. As shown in

Table 3. Accuracies (%) on Office-Home for ResNet50-based unsupervised domain adaptation methods.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 [17]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [25]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [11]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
MCD [39]	48.9	68.3	74.6	61.3	67.6	68.8	57	47.1	75.1	69.1	52.2	79.6	64.1
SAFN [52]	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
Symnets [54]	47.7	72.9	78.5	**64.2**	71.3	74.2	**64.2**	48.8	79.5	**74.5**	52.6	82.7	67.6
MDD [53]	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	**60.2**	82.3	68.1
EntMin	43.2	68.4	78.4	61.4	69.9	71.4	58.5	44.2	78.2	71.1	47.6	81.8	64.5
BFM	43.3	69.1	78.0	61.3	67.4	70.9	57.8	44.1	78.9	72.1	50.1	81.0	64.5
BNM	52.3	**73.9**	**80.0**	63.3	72.9	**74.9**	61.7	49.5	79.7	70.5	53.6	82.2	67.9
CDAN [26]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
CDAN+EntMin	54.1	72.4	78.3	61.8	71.8	73.0	62.0	52.3	79.7	72.0	57.0	83.2	68.1
CDAN+BNM	**56.2**	73.7	79.0	63.1	**73.6**	74.0	62.4	**54.8**	**80.7**	72.4	58.9	**83.5**	**69.4**

Table 4. Accuracies (%) on I2AwA for ResNet50-based unsupervised open domain recognition methods.

Method	Known	Unknown	All	Avg
zGCN [49]	77.2	21.0	65.0	49.1
dGCN [19]	78.2	11.6	64	44.9
adGCN [19]	77.3	15.0	64.1	46.2
bGCN [42]	84.6	28.0	72.6	56.3
pmd-bGCN [5]	84.7	27.1	72.5	55.9
UODTN [55]	84.7	31.7	73.5	58.2
Balance* [55]	85.9	22.3	72.4	54.1
EntMin	87.5	7.2	70.5	47.4
BFM	87.7	9.2	71.1	48.4
BNM	**88.3**	**39.7**	**78.0**	**64.0**

Table 5. Parameter Sensitivity on the I2AwA dataset for ResNet50-based unsupervised open domain recognition methods.

Method	Known	Unknown	All	AVG
BNM ($\lambda = 1$)	88.0	39.4	77.7	63.7
BNM ($\lambda = 1.5$)	88.1	**39.7**	77.9	63.9
BNM ($\lambda = 2$)	**88.3**	**39.7**	**78.0**	**64.0**
BNM ($\lambda = 3$)	87.7	39.5	77.5	63.6
BNM ($\lambda = 4$)	87.4	38.6	77.1	63.0

Figure 3(b), the diversity of BNM is still larger in Ar → Pl. But the differences between EntMin and BNM in diversity ratio are shorter than that in Ar → Cl. This results from fewer samples near the decision boundary in Ar → Pl since Ar → Pl is easier, *i.e.*, the basic accuracy of Ar → Pl is higher. Thus BNM is more effective in difficult tasks with rich data near the decision boundary.

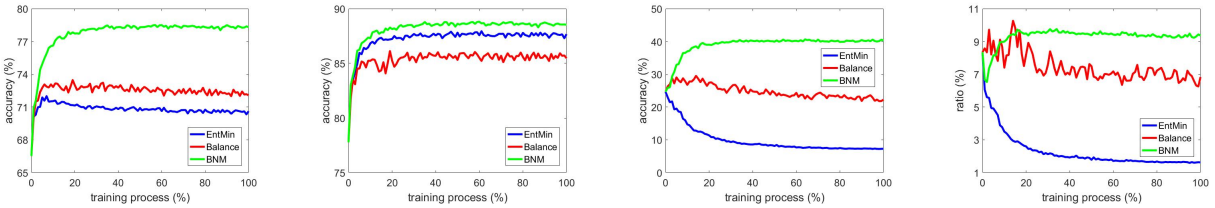
4.3. Unsupervised Open Domain Recognition

We evaluate our BNM method on I2AwA for unsupervised open domain recognition [55]. In I2AwA, the labeled domain consists of 2,970 images belonging to 40 known categories, via selecting images from ImageNet and Google image search engine. The unlabeled domain of I2AwA is AwA2 [51] which contains a total of 37,322 images. The images are totally classified into 50 categories, with the same 40 known categories as labeled domain, and the remaining 10 classes as unknown categories.

To obtain a reliable initial classification model on unknown categories, we construct the same knowledge graph for I2AwA with UODTN [55]. The graph structure is built according to the popular methods, Graph Convolutional Networks (GCN) [21, 49]. The graph nodes include all categories in the unlabeled domain and also their children and ancestors in WordNet [29]. To obtain the features of the nodes, we choose the word vectors of all categories extracted via the GloVe text model [35] trained on Wikipedia. We use ResNet-50 [17] pretrained on ImageNet [8] as our backbone, where the parameters of the last fully connected layer could be initialized by the parameters of GCN in the same categories.

For fair comparison, we perform in the same environment as UODTN [55]. The experiments are implemented with Pytorch [34]. We fix the batch size to 48 for both the labeled and unlabeled domain. We apply BNM on the classfication outputs on the total 50 categories and minimize classification loss on the known 40 categories in labeled domain and BNM loss on all the 50 categories in the unlabeled domain to train the network. We report the results of known categories, unknown categories, all categories on unlabeled domain and the average of known and unknown category accuracy. For each method, we run four random experiments and report the average classification accuracy.

The results are shown in Table 5, we achieve remarkable improvement on I2AwA. We achieve 11.4% improvement on the known categories over the baseline zGCN, and BNM surprisingly improves by 19.0% on the unknown categories over zGCN. From the overall range of the dataset, we achieve 13.3% improvement on the whole dataset and we achieve an average improvement of 15.2% improvement over zGCN. Besides, BNM outperforms the state-of-the-art



(a) All Category Accuracy (b) Known Category Accuracy (c) Unknown Category Accuracy (d) Unknown Category Ratio

Figure 4. Statistics for Entropy, Balance and Nuclear-norm in the whole training process.

UODTN [55] by 4.8%. This shows that the simple BNM is effective enough for unsupervised open domain recognition, which outperforms the combination of complex functions in UODTN [55]. We also show the parameter sensitivity experiments in Table 5. The results show BNM is relatively stable under different parameters, and we set λ fixed to 2 for further comparison.

We also compare the training process of EntMin, Balance and BNM loss functions in Figure 4. The prediction results on all categories, known categories and unknown categories are separately shown in Figure 4(a), 4(b) and 4(c). BNM outperforms others on All accuracy, Known category accuracy and Unknown category accuracy in the whole training process. To explore the intrinsic effect of BNM on unknown categories, we show the unknown category ratio, which is the ratio of predicting the samples in the labeled domain of I2AwA into unknown categories in Figure 4(d). Obviously, Entmin reduces the unknown category ratio by a large margin, which greatly damages the prediction diversity and accuracy on unknown categories. Though the unknown category ratio in BNM is reduced at first, it gradually raises along the training process, and it appears to be even higher than initial ratio after training. This means BNM could protect the diversity by ensuring ratio of prediction on minority categories. Though the Balance constraint could also protect ratio of prediction on minority categories, results of Balance loss seems not quite stable. Besides, the accuracy of Balance loss is much lower than BNM due to the lack of discriminability. The experimental phenomenon has steadily proved the effectiveness of BNM towards both discriminability and diversity.

4.4. Discussion

The chosen tasks are typical label insufficient situations to show the mechanism of BNM. Among the tasks and datasets, there are differences in two aspects, *i.e.*, the domain discrepancy and category balance. There exists large domain discrepancy in tasks of domain adaptation and unsupervised open domain recognition, while no domain discrepancy is assumed in semi-supervised learning. From the view of category balance, the categories are balanced in datasets of semi-supervised learning, *i.e.*, CIFAR-100. In datasets of domain adaptation, *i.e.*, Office-31 and Office-Home, the categories are imbalanced. While unsupervised open domain recognition is a learning task with extremely imbalanced category distributions, where some categories are even unseen in the labeled domain. In datasets of I2AwA, 10 categories are unknown categories, which hold a remarkable percentage of the total 50 categories.

As shown in the experiments, BNM could cooperate well with existing methods in semi-supervised learning. For domain adaptation, BNM could outperform most existing methods using losses such as adversarial loss. While in unsupervised open domain recognition, method with only the BNM loss and classification loss could even achieve state-of-the-art results. We could see the progressive progress and fitness of BNM to the tasks, from semi-supervised learning to unsupervised open domain recognition. Considering the differences between the tasks, we could obtain two conclusions on the applicability of BNM. The first is that BNM could work well in label insufficient situations. The other is that BNM outperforms entropy minimization significantly, especially when there exists rich domain discrepancy and imbalanced category distribution.

5. Conclusion

The discriminability and diversity could be separately represented by the Frobenius-norm and rank of the batch output matrix. Nuclear-norm is the upperbound of Frobenius-norm, also a convex approximation of matrix rank. Accordingly, we propose BNM method which maximizes the batch nuclear-norm to ensure higher prediction discriminability and diversity. Experiments show our method is suitable for the classification tasks under scenarios of semi-supervised learning, domain adaptation and open domain recognition. We will explore the effect of BNM on other settings and tasks in the future.

Acknowledgement. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102003, in part by National Natural Science Foundation of China: 61672497, 61620106009, 61836002, 61931008, 61771457, 61732007 and U1636214, and in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[3] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[4] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[5] Jianfei Chen, Chongxuan LI, Yizhong Ru, and Jun Zhu. Population matching discrepancy and applications in deep learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*, pages 6262–6272. Curran Associates, Inc., 2017.

[6] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2090–2099, 2019.

[7] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[9] Weisheng Dong, Guangming Shi, and Xin Li. Nonlocal image restoration with bilateral variance estimation: a low-rank approach. *IEEE transactions on image processing*, 22(2):700–711, 2012.

[10] Maryam Fazel. Matrix rank minimization with applications. 2002.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[12] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012.

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[15] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[19] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. *arXiv preprint arXiv:1805.11724*, 2018.

[20] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[21] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[23] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[26] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1647–1657, 2018.

[27] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, pages 136–144, 2016.

[28] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2018.

[29] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[30] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[31] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[32] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018.

[33] Théodore Papadopoulo and Manolis IA Lourakis. Estimating the jacobian of the singular value decomposition: Theory and applications. In *European Conference on Computer Vision*, pages 554–570. Springer, 2000.

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[35] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[36] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018.

[37] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[39] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[40] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[41] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[42] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. Transductive unbiased embedding for zero-shot learning. In *ICCV*, pages 1024–1033, 2018.

[43] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2005.

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[45] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[46] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[47] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[48] Shuhui Wang, Qingming Huang, Shuqiang Jiang, and Qi Tian. S^3mkl: Scalable semi-supervised multiple kernel learning for real-world image applications. *IEEE Trans. Multimedia*, 14(4):1259–1274, 2012.

[49] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866, 2018.

[50] Si Wu, Jichang Li, Cheng Liu, Zhiwen Yu, and Hau-San Wong. Mutual learning of complementary networks via residual correction for improving semi-supervised classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6500–6509, 2019.

[51] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. *arXiv preprint arXiv:1703.04394*, 2017.

[52] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[53] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019.

[54] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2019.

[55] Junbao Zhuo, Shuhui Wang, Shuhao Cui, and Qingming Huang. Unsupervised open domain recognition by semantic discrepancy minimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 750–759, 2019.

[56] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 261–269. ACM, 2017.

[57] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019.

[58] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.