The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Structure-Aware Feature Fusion for Unsupervised Domain Adaptation

Qingchao Chen,* **Yang Liu***
Department of Engineering Science, University of Oxford, UK
qingchao.chen@eng.ox.ac.uk, yangl@robots.ox.ac.uk

Abstract

Unsupervised domain Adaptation (UDA) aims to learn and transfer generalized features from a labelled source domain to a target domain without any annotations. Existing methods only aligning high-level representation but without exploiting the complex multi-class structure and local spatial structure. This is problematic as 1) the model is prone to negative transfer when the features from different classes are misaligned; 2) missing the local spatial structure poses a major obstacle in performing the fine-grained feature alignment. In this paper, we integrate the valuable information conveyed in classifier prediction and local feature maps into global feature representation and then perform a single mini-max game to make it domain invariant. In this way, the domain-invariant feature not only describes the holistic representation of the original image but also preserves mode-structure and fine-grained spatial structural information. The feature integration is achieved by estimating and maximizing the mutual information (MI) among the global feature, local feature and classifier prediction simultaneously. As the MI is hard to measure directly in high-dimension spaces, we adopt a new objective function that implicitly maximizes the MI via an effective sampling strategy and a discriminator design. Our STructure-Aware Feature Fusion (STAFF) network achieves the state-of-the-art performances in various UDA datasets.

Introduction

The success of deep neural network relies on a massive amount of labeled training data. However the learned representation is very sensitive to the input perturbations and dataset biases, that is, deep networks may easily fail to generalize to new dataset or environment. In practice, manual labeling of such sufficient training data for every new dataset is often prohibitive or impossible to collect. Unsupervised domain adaptation (UDA) aims to solve this problem by transferring a deep network from a source domain where sufficient labeled training data is available to a target domain where only unlabeled data is available.

The main technical difficulty of UDA is how to address the domain shift and formally reduce the distribution discrepancy across different domains. Although various distribution divergence measurements have been investigated to estimate and reduce domain discrepancy, these methods only focus on aligning high-level representations, e.g. the fully connected (FC) layer features, but without exploiting the complex multi-mode structure and local geometric spatial structures.

We *argue* that to solve UDA problem, it is far from sufficient to match only the global feature distribution, because of the following reasons: 1) The data distribution usually embody complex multi-mode structures, reflecting either the class boundaries in supervised learning or the cluster boundaries in unsupervised learning. Only matching marginal distribution without exploiting the multi-mode structure may be prone to negative transfer, especially when the corresponding mode of the distributions across domains are falsely aligned. There is no guarantee that samples from different domains with the same class label will be mapped nearby in the feature space. As a result, the discriminative structure between classes could be mixed up thus leading to a poor performance for the target domain. 2) Only matching global feature ignores the local geometric spatial structures. However, the domain discrepancy may appear at the start from the early convolutional layers, which makes any adjustment purely at the tail of the network less effective. In addition, the lack of local features for different regions, pose a major obstacle in performing a fine-grained feature alignment.

An intuitive solution for the two aforementioned issues is that performing feature alignments via multiple adversarial training at the different level of representation, including local-, global-representation and classifier predictions (mode informative). However, this is unrealistic and suffer from unstable numerical optimization due to the easy conflict gradients from a set of minimax problems coupled together, not to say the inefficient design and heavy memory consumption of multiple discriminator networks. In this paper, rather than performing multiple minimax problems simultaneously, we address the challenges from another perspective by formalizing a STructure-Aware Feature Fusion (STAFF) network. More specifically, we first **integrate** the informative content of local feature (fine-grained structural information) and classifier prediction (conveys mode structure) into the global feature and then perform a single minimax optimization to

*Equal contribution

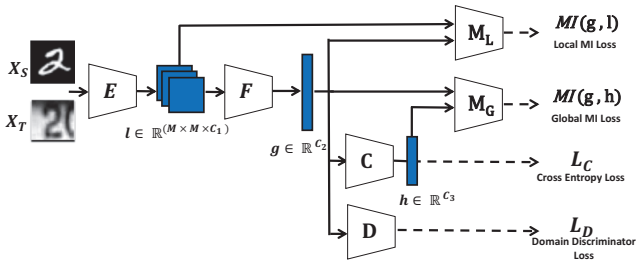

Figure 1: Overall network architecture of our proposed STructure-Aware Feature Fusion (STAFF) Network. First, in the source domain, the feature encoder E, Feature Transformer F and the content classifier C are trained to extract discriminative features from images X_s labeled by Y_s by minimizing the cross entropy loss $\mathcal{L}_C$. Second, we integrate the classifier prediction (mode structure) and local feature maps (spatial structure) into the global feature via using global and local MI discriminators M_G and M_L to maximize the mutual information MI among them. Finally, to learn domain-invariant fused features, the domain classifier D, the encoder E and the feature transformer F play an adversarial game, where D tries to discriminate whether the features are from target or source domain, while the encoder E and F tries to confuse D. The learned domain invariant feature not only describes the holistic representation of the original image but also preserves fine-grained spatial structural and discriminative mode structure.

make the global feature domain invariant. To the best of our knowledge, no previous work is able to integrate both the local feature and classifier prediction in the single global feature efficiently for adversarial learning. In addition, no proper loss function has been designed to integrate such relationships in the optimization of an adaptation network.

So what is a good integration? Successful integration, in this case, should be able to distill common part while ignoring rest, that is the local-, global-representation and classifier predictions should be predictive of the others with low uncertainty. To achieve this goal, our STructure-Aware Feature Fusion (STAFF) network tries to regularize the global representation so that the mutual information (MI) between its class predictions and its local feature maps can be maximized simultaneously. Afterward, only performing one adversarial training on this single global representations. In this way, the learned domain invariant global feature not only describes the holistic representation of the original image but also preserves fine-grained structural information and keep mode informative. As the mutual information between two variable is known that hard to measure directly in high-dimension spaces, we adopt a new objective function that implicitly maximizes the MI via an encoder-discriminator architecture and an effective sampling strategy. The overall framework of STAFF is shown in Figure.1.

Our main contributions are summarized in the following:

1. We are the first to integrate the valuable information conveys in classifier prediction and local feature maps into the global feature representation and then perform a single adversarial game. The learned domain invariant feature (global feature) not only describes the holistic representation of the original image but also preserves fine-grained structural information and mode informative.

2. A successful global feature integration is achieved by maximizing the mutual information between its class predictions and its local feature maps simultaneously. We adopt a new objective function, mutual information discriminator and sampling strategy to estimate and maximize the MI.

3. Our approach achieves the state-of-the-art performance on the domain adaptation benchmarks, including handwritten digit dataset, Office-31 dataset and Office-Home dataset.

Related Work

Distribution Matching Methods in UDA

Deep features have been proved transferable, disentangled and invariant of underlying different data variations (Long et al. 2018). However, cross-domain discrepancy of representations still exist and current deep adaptation networks adopt variants of MMD (Long et al. 2015; Venkateswara et al. 2017), the adversarial training strategy (Ganin et al. 2016; Bousmalis et al. 2016; Liu, Breuel, and Kautz 2017; Chen et al. 2018) or transportation plan modelling (Courty et al. 2017; Chen et al. 2018; Damodaran et al. 2018) to measure and reduce domain discrepancies.

A number of works were proposed to translate image styles between domains directly, namely the pixel-level adaptation (Bousmalis et al. 2016; Liu, Breuel, and Kautz 2017). Most recent GAN based methods successfully explored to transfer style from source to target domain and back again (Liu, Breuel, and Kautz 2017; Russo et al. 2017), however these pixel-level image translation methods require more investigations when adapting domains with large discrepancy, for example the office-home dataset (Venkateswara et al. 2017).

Multi-level Domain Alignment

It is challenging to reduce the multi-level feature discrepancies in an effective yet efficient manner. Previous adaptation networks have investigated either to utilize multiple networks and loss functions or integrate multi-level feature into a unified one and constrain the feature learning using a single loss function.

JAN (Long et al. 2016) may be the first to consider matching the joint distribution of global feature and the label predictions using tensor product, however, JAN does not integrate and adapt the local feature discrepancies. Most recently, CDAN (Long et al. 2018) explored to integrate label prediction and the global features using random projection matrix. This is a very efficient procedure however it still does not consider integrating the local feature structure. In addition, no explicit loss function has been investigated to preserve the information contents in the feature integration operation.

Another line of research adopt multi-level domain discriminator networks (Zhang et al. 2018) or multiple loss functions (Damodaran et al. 2018). Zhang et al. (Zhang et al. 2018) first perform alignments on both local convolutional and the

global feature using multiple domain discriminators but without aligning the class predictions.

Damodaran et al. (Damodaran et al. 2018) proposed multiple OT losses to match joint global feature and label distributions, however, they failed to consider the local structure information in the domain distribution alignment. To the best of our knowledge, STAFF may be the first to integrate the global feature, local structure and label prediction in one representation using MI loss functions for adversarial training and the domain discrepancy alignment.

Mutual Information Estimation

Estimating MI on the continuous and high-dimensional feature space is extremely difficult. However, it is also very useful for unsupervised representation learning by maximizing the MI between input and output of the model, as being widely used for independent component analysis(Hyvarinen 1999). Most recently, the pioneer work Mutual Information Neural Estimation (Belghazi et al. 2018) explored to robustly estimate the MI using neural networks and Deep Info Max (DIM) (Hjelm et al. 2018) applied this method to learn a meaningful representation for unsupervised classification. To the best of our knowledge, we may be the first to explore the MI maximization for UDA problem, especially for multi-level feature integration and discrepancy alignment.

Model

The overall framework is illustrated in Figure 1, where arrows indicate the forward propagation direction. STructure-Aware Feature Fusion (STAFF) network is composed of the following components, including the encoder E, feature transformer F, the global and local Mutual Information Estimators M_G and M_L, the content classifier C and the domain classifier D.

Assuming the source image X_S and the discrete content label Y_S are drawn from a source domain distribution $P_S(X, Y)$, as well as target images X_T drawn from target domain distribution $P_T(X)$ without label observations. Since direct supervised learning on the target images is not possible, UDA instead learns a content classifier C driven by source labels only and then adapts the model to the target domain.

Specifically, the source image is first mapped by the encoder to the latent local representation, i.e., a set of feature maps $E(X_S) \in \mathbb{R}^{M\times M\times C_1}$. Then the feature transformer first performs a global pool over the spatial regions and then transform the feature to its latent global feature representation $F(E(X_S)) \in \mathbb{R}^{C_2}$. Afterwards, the content classifier works cooperatively with the Encoder E and Feature transformer F to minimize the content classification loss for source images $\mathcal{L}_C$, which is a conventional cross-entropy loss between ground truth Y_S and prediction $C(F(E(X_S)))$:

$$\min_{E,F,C} \mathcal{L}_C. \tag{1}$$

The general recipe to solve the UDA problem is to regularize the learning of encoder and feature transformer, so as to match the marginal distribution between $P(X_S)$ and $P(X_T)$. Most of the existing UDA approach makes the *hypothesis* that: once the marginal distribution is matched, the source content classifier can be applied to the target features for label prediction. Under this hypothesis, we can formulate the following adversarial training objective to minimize the feature discrepancy:

$$\max_{E,F} \min_{D} \mathcal{L}_D. \tag{2}$$

As can be seen, the encoder E, feature transformer F and domain classifier D play an adversarial game on the domain classification loss L_D, where E and F tries to minimize the cross-domain divergence so that D fails to correctly classify which domain the sample comes from no matter how hard D tries. Ideally, at the end of the competition, D can perform no better than a random guess, which means the learned global feature representation is domain invariant. To simplify the notations, we denote $l \in \mathbb{R}^{(M\times M\times C_1)}$, $g \in \mathbb{R}^{C_2}$ and $h \in \mathbb{R}^{C_3}$ to represent the local convolution feature map $E(X)$, global feature $F(E(X))$ and the classifier prediction $C(F(E(X)))$ respectively.

However, as shown in the introduction section, the *hypothesis* aforementioned is problematic due to two reasons and the goal of this paper is to integrate multi-level features in the global feature and align it only using single adversarial training. More specifically, we made novel designs in global and local MI discriminators M_G and M_L to achieve this goal. Mathematical formally, we learn the parameters of E , F and M_G to maximize the mutual information between global feature g and inductive classifier prediction h to make global g is mode-aware as :

$$\max_{E,F,M_G} \mathcal{MI}(g, h). \tag{3}$$

Meanwhile we want to maximize the mutual information between global feature g and local feature l to make global g preserves the useful local structure information as:

$$\max_{E,F,M_L} \mathcal{MI}(g, l). \tag{4}$$

More details about the fundamentals of MI estimation, training strategy of the MI discriminator M_G and maximizing the MI between global feature and inductive classifier prediction will be discussed in the next section. The training strategy of local MI discriminators M_L for maximizing MI between global and local representation is discussed afterwards. Finally, the overall optimization objective functions are summarized.

Maximize MI between global representation and classifier prediction

MI is a well-known unsupervised learning loss function, with the aim of maintaining the information contents between variable X and Y. As shown in Eq.(5), MI measures the Kullback-Leibler (KL) divergence between the joint distribution $P(X, Y)$ and the product of their marginal distributions $P(X)P(Y)$.

$$\mathcal{I}(X, Y) = \mathbb{KL}(P(X, Y)||P(X)P(Y)) \tag{5}$$

The MI is small when the two variables X and Y are statistically independent, while is large when two variables preserve the same information content. Although the MI between

two random variables is hard to measure directly in high-dimension spaces, some recent studies (Belghazi et al. 2018; Hjelm et al. 2018) proved that an implicit estimation of MI can be achieved with an encoder-discriminator architecture.

We attempt to use the network design as shown in Figure. 2 to maximize the MI between the global feature g and its associative classifier prediction h. More specifically, this relies on a sampling strategy that draws positive and negative samples from the joint distribution $P(g, h)$ and from the marginal product $P(g)P(h)$ respectively. In our case, the positive samples (g_1, h_1) are the features of the same input, while the negative samples (g_1, h_2) are obtained from different inputs. That is, a set of n positive and negative pairs can form a mini-batch $X = \{X_P, X_N\}$. Given g_1, cooperatively trained with F and C, the global MI discriminator M_G aims to distinguish whether the other input (h_1 or h_2) are from the same input image or not.

The function of M_G contains two operations: 1) to project the classifier prediction $h \in \mathbb{R}^{C_3}$ to a vector $\hat{h} \in \mathbb{R}^{C_2}$ using a linear transformation $W_h \in \mathbb{R}^{C_2 \times C_3}$; 2) measure the similarity between g and $\hat{h}$ (with the same dimension to g) via a dot product. Mathematical formally, the function M_G can be represented as

$$M_G(x, y) = g^T W_h h \tag{6}$$

Various objective functions can be used to maximize $MI(g, h)$. The simplest formulation as did in (Brakel and Bengio 2017; Hjelm et al. 2018), adopting the standard binary cross-entropy (BCE) loss as shown in (7) where the output of M_G is activated by a sigmoid function.

$$\mathbb{E}_{X_P}[log\sigma(M_G(g_1, h_1))] + \mathbb{E}_{X_N}[log(1 - \sigma(M_G(g_1, h_2)))] \tag{7}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$. Rather than optimizing exact KL divergence as defined by MI, the BCE estimate a Jenson-Shannon (JS) divergence instead. JS is more stable since it is always defined, bounded by [0,1], symmetric and more smooth.

As an alternative, the work in (Oord, Li, and Vinyals 2018) suggests that minimizing the Noise Contrasting Estimation (NCE) Loss as shown in (8) is in fact maximizing a lower bound of MI. Note that in this scenario, n samples within one mini-batch contains 1 positive pair X_P and $(n-1)$ negative pairs X_N. The work in (Oord, Li, and Vinyals 2018) has shown that the lower bound becomes tighter as n becomes larger. This loss can be regarded as the categorical cross-entropy of classifying the positive sample correctly, with $\frac{e^{(M_g(g_1,h_1))}}{\sum_{h_2 \in X} e^{(M_g(g_1,h_2))}}$ being the prediction of the model.

$$-\mathbb{E}_X[log\frac{e^{(M_G(g_1,h_1))}}{\sum_{h_2 \in X} e^{(M_G(g_1,h_2))}}] \tag{8}$$

The third alternative is to directly optimize the MI with the Mutual Information Neural Estimation (MINE) (Belghazi et al. 2018) with the objective shown in (9):

$$\mathbb{E}_{X_P}[M_G(g, h)] + \mathbb{E}_{X_N}[e^{M_G(g,\hat{h})}] \tag{9}$$

MINE explicitly computer the MI of continuous variables by exploiting a lower bound based on the *Donsker-Varadhan* representation of the KL divergence.

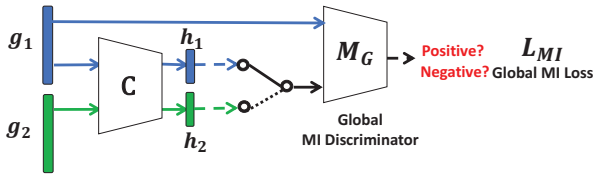

Figure 2: Network design of maximizing the mutual information between global feature g and its associative classifier prediction h.

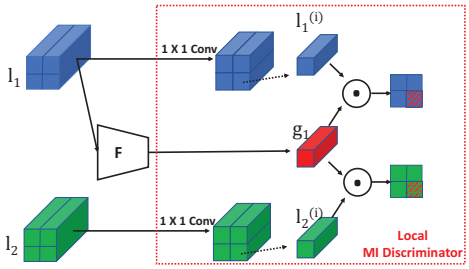

Figure 3: Network design of maximizing the mutual information between global features g and local features l. The $l_1 \in \mathbb{R}^{(M \times M \times C_1)}$ and $l_2 \in \mathbb{R}^{(M \times M \times C_1)}$ are the local feature representation extracted from E for two different images, which contains some spatial structural information. The global feature $g_1 \in \mathbb{R}^{C_2}$ is encoded by the feature transformer F for input l_1. (l_1, g_1) are drawn from the joint distribution (positive) while (l_1, g_1) are drawn from the marginals (negative). The function of the local MI discriminator M_L is shown in the red box. Both l_1 and l_2 first map to a space with the C_2 channels by 1×1 convolution. We then take the dot product between the feature at each location of the feature map with the global feature representation.

All the aforementioned objectives are based on the different approximation of KL divergence between the joint and product of marginal distributions as the definition of MI. This paper is the first to introduce MI estimation into the UDA problem, we will compare these objective functions for MI optimization in our proposed network later in Section Analysis.

Maximize MI between Global representation and Local representation

In previous section, we have discussed that at least three alternative objective functions can be utilized to implicitly maximize the MI between the global representation and classifier prediction. In this section, we present how to maximize the MI between global and local representations.

Our local MI maximization framework is shown in Figure.3. First we encode the input to a feature map $l \in \mathbb{R}^{(M \times M \times C_1)}$, represented as $l = \{l^{(i)}\}_{i=1}^{M^2}$ preserving the spatial structure information. After feed-forwarding l through the feature transformer F and obtaining its corresponding global features g, we can define our local MI estimator in (4) as the average MI loss between the feature $l^{(i)}$ at the spatial

location i with the global feature g:

$$\mathcal{MI}(g,l) = \frac{1}{M^2}\sum_{i=1}^{M^2}\mathcal{MI}(g,l^{(i)}). \tag{10}$$

Therefore, we can take a similar encoder-discriminator design and sampling strategy to maximize the $\mathcal{MI}(g,l)$. More specifically, we choose to sample (g_1,l_1) from the joint distribution as the positive pair, and sample (g_1,l_2) from the product of marginal distribution as the negative pair. Intuitively, cooperatively trained with E and F, the local MI discriminator M_L aims to distinguish whether the other input (l_1 or l_2) are from the same input image as g_1 or not.

The operation of the local MI discriminator M_L is slightly different from the global one M_G. The local feature map is encoded using a 1×1 convolution network and the output $\hat{l}$ has C_2 channels (i.e., same as the dimension of global feature g), with C_2 smaller than C_1. We then take the dot product between the feature at (i^{th}) location $\hat{l}^{(i)}$ with the global feature representation g for calculating the prediction of $M_L(g,l)$. To this point, all the aforementioned loss functions (7)(8)(9) can be used to implicitly maximize $\mathcal{MI}(g,l)$ by replacing the $M_G(g,h)$ by $M_L(g,l^{(i)})$. It is worth noting that because the same global representation is encouraged to have high MI with all the patches, this favors encoding the similar information shared across patches.

Optimization

This section present the complete objective of STAFF in (11). The overall loss function is a min-max problem, including the source domain classification loss L_C, domain discriminator loss L_D, global MI losses $\mathcal{MI}(g,h)$ and local MI losses $\mathcal{MI}(g,l)$. It is worth noting that $\mathcal{MI}(g,h)$ is parameterized by F,C,M_G, while $\mathcal{MI}(g,l)$ is parameterized by E,F,M_L. The hyper-parameter α,β,γ represent the weight of relevant loss functions.

$$\max_{E,F,C,M_l,M_g}\min_{D}\alpha L_D - L_C + \beta\mathcal{MI}(g,h) + \frac{\gamma}{M^2}\sum_{i=1}^{M^2}\mathcal{MI}(g,l^{(i)}) \tag{11}$$

Experiments and Results

We evaluate the proposed STAFF network with state-of-the-art deep learning based unsupervised domain adaptation methods. In this section, we first illustrate the datasets and implementation details. Then we show extensive experimental results and analysis. Our STAFF works reasonably well on all benchmarks, including Digit, Office-31 and Office-Home dataset.

Experiment Setup and Implementation Detail

Digits: We investigate three digits datasets of varying difficulties, including MNIST, USPS and the SVHN. We adopt the train-test protocol of (Russo et al. 2017) for a fair comparison with four transfer tasks: MNIST →USPS (M→U), USPS →MNIST (U→M), SVHN →MNIST, (S→M) and MNIST →SVHN (M→S). All comparison network use a variant of LeNet as the basis network, similar to the one used in (Damodaran et al. 2018). The discriminator network is composed of three FC layers with ReLU function (see details in supplementary). We fix $\alpha=1,\beta=0.01,\gamma=0.01$ for all experiments. We train our network from scratch use SGD with momentum of 0.9, learning rate of 0.002 and batch size of 128.

Table 1: Recognition rates (%) of hand-written digit dataset.

Methods	M-U	U-M	S-M	M-S
SO	90.2	61.2	59.3	26.2
MMD (Long et al. 2015)	88.5	73.5	64.8	-
DANN(Ganin et al. 2016)	95.7	90.0	70.8	-
Self-Ensemble (French et al. 2018)	88.1	92.4	93.3	42.0
GentoAdapt(Sankarayanan et al. 2018)	95.3	90.8	92.4	-
UNIT (Liu, Breuel, and Kautz 2017)	95.9	93.5	90.5	-
SBADA-GAN(Russo et al. 2017)	97.6	95.0	76.1	61.1
CDAN (Long et al. 2018)	93.9	96.9	88.5	-
Deep-JDOT(Damodaran et al. 2018)	95.7	*96.4*	*96.7*	-
TPN(Pan et al. 2019)	92.1	94.1	93.0	–
STAFF (Ours)	**98.3**	**98.1**	**97.7**	**65.8**

Office-31 and Office-Home: Office-31 is the most widely used dataset for unsupervised domain adaptation. It comprise 4110 images from 31 classes collected from three distinct domains: *Amazon* (A), *Dslr* (D), *Webcam* (W). Office-Home is a more difficult dataset than Office-31. It comprise 15,500 images from 65 classes collected from four distinct domains: *Art* (Ar), *Clip* (Cl), *Product* (Pr) and *Real-World* (Rw). We evaluate all methods on all transfer tasks for these two datasets.

All comparison networks use a ResNet-50 (pretrained from ImageNet) as base networks. We train domain discriminator D, MI estimators M_G, M_L and classifier C from scratch. Whatever module trained from scratch, its learning rate was set to be 10 times that of the fine-tuning layers. We used the following parameters $\alpha=1,\beta=0.1,\gamma=0.05$ for all experiments. The SGD with 0.9 momentum is used and the learning rate is annealed by $u_p=u_0(1+\eta p)^{-\phi}$, where p is the training progress changing from 0 to 1, and $u_0=0.01,\eta=10,\phi=0.75$ (Ganin et al. 2016). We used the conventional three-layer FC in discrminator network for both office-31 and office-home datasets.

Results and Comparisons

Digits: The results on Digit datasets of four adaptation tasks are reported in Table 1, with baseline results directly reported from the original papers if the protocol is the same. Our proposed STAFF outperforms all comparison methods on all tasks. Note that GentoAdapt, UNIT and SBADA-GAN rely on pixel-level image generation, which is specifically designed for digits and unrealistic to real-world adaptation tasks. These approaches achieve quite competitive results when the domain shift is small, while degrades a lot when the domain discrepancy is large. This may be because image-translation across domains with large discrepancy is challenging, let alone learning good domain-invariant features from these images. The approaches based on matching latent feature distribution performs fairly stable.

The four methods listed in the last four rows, all consider exploiting multi-level feature representation in reducing domain discrepancy. Deep-JDOT uses multiple loss at low and global- representation directly to minimize the domain discrepancy explicitly. CDAN integrates global feature and classifier prediction by performing either a tensor-product in Kernel space or approximately calculate cross-co-variance. Our proposed STAFF consistently outperforms them in all four adaptation tasks. We hypothesize the out-performance is because we are the first to make the domain invariant feature not only describes the holistic representation but also preserves both fine-grained local structure and mode structure simultaneously. Moreover, maximizing the mutual information among multi-level representation is an effective way to integrate features. A more detailed ablation study about the contribution of global and local MI can be found in Section Analysis.

Office-31 dataset: The results on Office-31 dataset of six transfer tasks are reported in Table 2, with results of baselines directly reported from the original papers. The proposed approach outperforms all comparison methods on all tasks. Compared with digit dataset, these tasks are more difficult as more dissimilar across domains and with much lower adaptation accuracy. It is desirable that STAFF yield larger boosts on such a difficult task, which reveals the importance of structure-aware feature fusion. Among comparison approaches, CAN(Zhang et al. 2018), JAN(Long et al. 2016), CDAN and our proposed approach all consider exploiting multi-level presentation and we all boost performance. Our STAFF achieves the best performance, which demonstrates that maximizing mutual information among multi-level representation is an effective way to integrate the information of data local spatial structure and mode structure and make them contribute to reducing the domain discrepancy. Rather than using multiple domain discriminators at a different position for distribution matching as done in CAN, our formulation is more elegant and requires only one single domain discriminator in a simple form.

Office-Home dataset: The results on Office-Home dataset of 12 transfer tasks are reported in Table 3. The proposed approach outperforms all baseline methods in 10 out of 12 transfer tasks by a large margin except the most recent SymNets (Zhang et al. 2019). The potential reason is: Office-Home dataset consists of much more class categories and we may expect more difficult multi-modal class prediction distribution to match on the target domain (the mode structure is more important). Maximizing the mutual information between the global feature and classifier prediction is an effective way to integrate feature. Compared with the SymNets results, STAFF still outperforms it and most of the sub-task results are highly competitive. The experimental results of STAFF also dedicate that combining the local and global MI losses can significantly alleviate the problem of domain discrepancies.

Table 2: Recognition rates (%) of adapting Office-31 dataset.

Methods	A-W	W-A	A-D	D-A	W-D	D-W	Avg
SO	73.5	59.8	76.5	56.7	99.0	93.6	76.5
DAN	80.5	62.8	78.6	63.6	99.6	97.1	80.4
RTN	84.5	64.8	77.5	66.2	99.4	96.8	81.6
DANN	82.0	67.4	79.7	68.2	99.1	96.9	82.2
JAN	86.0	70.7	85.1	69.2	99.7	96.7	84.6
CAN	81.5	63.4	85.5	65.9	99.7	98.2	82.4
CDAN	93.5	67.8	86.4	66.9	99.8	98.5	84.5
CDAN-E	94.1	69.3	92.9	71.0	**100**	98.6	87.7
SymNets	90.8	**72.5**	93.9	74.6	**100**	98.8	88.4
STAFF	**96.4**	70.2	**94.0**	**71.7**	99.8	**99.6**	**88.6**

Analysis

We first compare our method with various sub-models (selected functional modules of STAFF) to study the effectiveness of each part. We then compare our proposed Mutual information based feature integration with some other feature integration method, we not only report the recognition accuracy and $\mathcal{A}$-distance but also show T-SNE plot of their feature embedding. We also compare the performance with three alternative loss proposed in the Method Section for maximizing the mutual information, including JSD, NCE and MINE. Finally, we visualize how the local MI discriminator exploits local spatial structure information. We plot the output of local MI discriminator M_L by calculating the dot product between the feature at each location of the local feature map with the global feature representation. All ablation studies in this section are performed using the task: A$\rightarrow W$ in the office-31 dataset.

Contribution of Network Components

We conduct a detailed ablation study by examining the effectiveness of each proposed component in our network structure. As shown in Table 4, introducing different combinations of modules all boost the performance compared to two Base Adaptation models, using DANN and MMD respectively. Adding either global MI maximization loss $MI(g, h)$ or local MI maximization loss $MI(g, l)$ improves the performance by approximately over 8%, which verifies the effectiveness of leveraging the mutual information constraints to integrate multi-level features for both base divergence measurements. Also, it is observed that the global MI maximization loss contributes to the most performance gain as an individual module. It indicates that integrates valuable information from classifier prediction is very important to make the domain-invariant feature maintains discriminative capability and thus leading a better performance. By exploiting local spatial structure and mode structure simultaneously, our proposed STAFF achieves the best performance.

Comparison of Feature Integration Approach

In this section, we compare different feature integration approaches in unsupervised domain adaptation problem. We integrate both the local spatial structure and classifier prediction into the global representation in four ways. As shown in Table.5, the conventional feature concatenation (i.e., concatenation of l , g and h), concatenation after projecting l , g and h to the same dimension and our proposed MI-based integration boost the recognition performance compared to the Base adaptation model and the one using multiple discriminator networks (Multi-Adv). Our MI-based integration

Table 3: Recognition rates (%) of adapting Office-Home dataset.

Methods	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
Source Only	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	43.6	50.7	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN	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
JAN	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN+E	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
SymNet	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
STAFF (Ours)	**53.3**	71.9	**80.2**	63.1	69.8	74.1	**65.3**	**50.9**	77.8	73.1	56.6	82.4	**68.2**

Table 4: Ablation study of different network components. Global MI (GMI) and Local MI(LMI) indicate two MI losses.

Model	L_C	L_D	GMI	LMI	Acc.
Source Only	✓	×	×	×	72.3
Base Adaptation(DANN)	✓	✓	×	×	82.0
Base Adaptation(MMD)	✓	✓	×	×	80.5
Only GMI(DANN)	✓	✓	✓	×	94.2
Only GMI(MMD)	✓	✓	✓	×	88.0
Only LMI(DANN)	✓	✓	×	✓	92.2
Only LMI(MMD)	✓	✓	×	✓	88.3
STAFF(MMD)(Ours)	✓	✓	✓	✓	90.2
STAFF(DANN)(Ours)	✓	✓	✓	✓	96.4

Table 5: Comparison of feature integration approaches.

	Base Adaptation	Multi Adv	Concat	Concat +Proj	STAFF (Ours)
Acc	82.0	83.5	87.9	89.1	96.4
A-dist	1.648	1.573	1.458	1.451	1.313

Table 6: Comparison of Different Mutual Information loss.

MI loss	JSD	NCE	MINE
Accuracy	96.4	91.0	93.7

outperforms both feature concatenation methods by around 8.5%, which implies that MI-based integration is more effective to incorporate structure information.

Besides the recognition performance, we also measure the distribution discrepancy quantitatively through $\mathcal{A}$-distance (Ben-David et al. 2010). The $\mathcal{A}$-distance is calculated following: $d = 2(1 - 2\theta)$, where θ is the domain classification generalization error using the Support Vector Machine (SVM) classifier trained to discriminate the source from the target. Table.5 presents the $\mathcal{A}$-distance achieved by the base adaptation and two feature integration models. It can be observed that using our STAFF (MI-based integration) achieves the lowest $\mathcal{A}$ distance, which proves its superior performance of reducing the distribution gap more effectively. Finally, to measure the feature discriminative capability, we plotted the T-SNE (Maaten and Hinton 2008) result to visualize the 2-D embedding of the extracted features for different feature integration approaches. Figure.4(a) -(c) plot the representation of target domain images by base adaptation, feature concatenation and MI-based feature integration. Using MI-based feature integration are evidently clustered closer than other

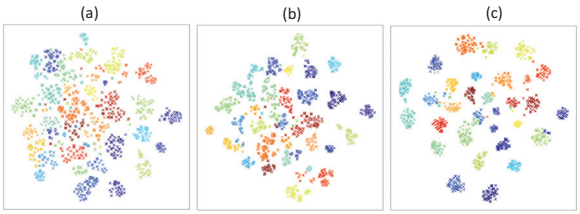

Figure 4: Visualization of TSNE plot of target image features; (a) Base Adaptation (b) Feature Concatenation (c) STAFF, MI-based Feature Integration.

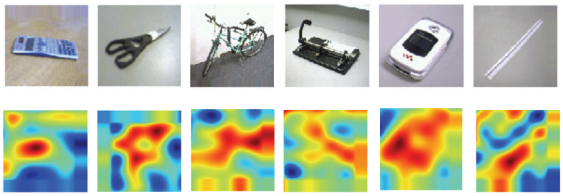

Figure 5: Visualization of MI map between local features at each spatial location and the global feature.

comparison methods. This shows the benefit of STAFF on discriminative predictions.

Comparison of MI Optimization Loss

We compare three objective functions to maximize the MI to integrate local and mode structures, including Jenson-Shannon Divergence (JSD), Noise Contrastive Estimation (NCE) and Mutual Information Neural Estimates (MINE). The performance is reported in Table.6. These numbers are all based on the same network architecture and training strategy with batch size 32. It can be seen that using JSD achieves the best performance. Another interesting observation is that with increasing batch size, the performance of the NCE loss improves a lot, which is consistent with the observation in (Oord, Li, and Vinyals 2018). It can achieve similar performance to JSD, i.e., 96.3 when the batch size is 256. For a fair comparison with others, we fixed the batch size as 32 for all comparison methods throughout the paper unless specified.

Visualize the output of local MI discriminator

We visualize the output of local MI discriminator, representing which spatial location has larger MI with the global fea-

ture. As shown in Figure 5, different regions in images have different corresponding MI value. The hotter the color, the larger the MI value. Taking the first image as an example, the calculator is highlighted with red color while the background diminishes in blue color. These results intuitively reveal that the positions with larger MI in local feature map closely link to the discriminative area (i.e., foreground object). This enables a fine-grained feature alignment, thus leading to a better performance.

Conclusion

In this paper, we proposed the Structure-Aware Feature Fusion (STAFF) module to integrate multi-level structure information into a single global feature for UDA tasks. Through maximizing the MI among multi-level features, STAFF can integrate the multi-mode structure of class predictions and the geometric structure of the local features into the global feature and then perform a single adversarial game to make it domain invariant. In this way, the learned domain-invariant feature not only describes the holistic representation of the original image but also preserves the fine-grained spatial structure and discriminative mode structure. Evaluation on extensive datasets suggests that the integrated features can characterize the multi-level domain discrepancies in a more meaningful and comprehensive manner.

Acknowledgement

We acknowledge the support of EPSRC Programme Grants Seebibyte EP/M013774/1 and CALOPUS EP/R013853/1.

References

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Hjelm, D.; and Courville, A. 2018. Mutual information neural estimation. In *International Conference on Machine Learning*, 530–539.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning* 79(1):151–175.

Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; and Krishnan, D. 2016. Unsupervised pixel-level domain adaptation with generative adversarial networks. *arXiv preprint arXiv:1612.05424*.

Brakel, P., and Bengio, Y. 2017. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*.

Chen, Q.; Liu, Y.; Wang, Z.; Wassell, I.; and Chetty, K. 2018. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7976–7985.

Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017. Joint distribution optimal transportation for domain adaptation. *arXiv preprint arXiv:1705.08848*.

Damodaran, B.; Kellenberger, B.; Flamary, R.; Tuia, D.; and Courty, N. 2018. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *The European Conference on Computer Vision (ECCV)*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17(59):1–35.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hyvarinen, A. 1999. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks* 10(3):626–634.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, 700–708.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 97–105.

Long, M.; Han, Z.; Wang, J.; and Jordan, M. 2016. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*. 1647–1657.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pan, Y.; Yao, T.; Li, Y.; Wang, Y.; Ngo, C.-W.; and Mei, T. 2019. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2239–2247.

Russo, P.; Carlucci, F. M.; Tommasi, T.; and Caputo, B. 2017. From source to target and back: symmetric bi-directional adaptive gan. *arXiv preprint arXiv:1705.08824* 3.

Sankarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8503–8512.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, W.; Ouyang, W.; Li, W.; and Xu, D. 2018. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3801–3809.

Zhang, Y.; Tang, H.; Jia, K.; and Tan, M. 2019. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5031–5040.