

A Compact DNN: Approaching GoogLeNet-Level Accuracy of Classification and Domain Adaptation

Chunpeng Wu^{1*}, Wei Wen¹, Tariq Afzal², Yongmei Zhang², Yiran Chen³, and Hai (Helen) Li³

¹Electrical and Computer Engineering Department, University of Pittsburgh, Pittsburgh, PA 15260

²LG San Jose Lab, Santa Clara, CA 95054

³Electrical and Computer Engineering Department, Duke University, Durham, NC 27708

{chunpeng.wu,wei.wen}@pitt.edu, {tariq.afzal,jenny.zhang}@lge.com

{yiran.chen,hai.li}@duke.edu

Abstract

Recently, DNN model compression based on network architecture design, e.g., SqueezeNet, attracted a lot of attention. Compared to well-known models, these extremely compact networks don't show any accuracy drop on image classification. An emerging question, however, is whether these compression techniques hurt DNN's learning ability other than classifying images on a single dataset. Our preliminary experiment shows that these compression methods could degrade domain adaptation (DA) ability, though the classification performance is preserved. In this work, we propose a new compact network architecture and unsupervised DA method. The DNN is built on a new basic module Conv-M that provides more diverse feature extractors without significantly increasing parameters. The unified framework of our DA method will simultaneously learn invariance across domains, reduce divergence of feature representations and adapt label prediction. Our DNN has 4.1M parameters—only 6.7% of AlexNet or 59% of GoogLeNet. Experiments show that our DNN obtains GoogLeNet-level accuracy both on classification and DA, and our DA method slightly outperforms previous competitive ones. Put all together, our DA strategy based on our DNN achieves state-of-the-art on sixteen of total eighteen DA tasks on popular Office-31 and Office-Caltech datasets.

1. Introduction and Motivation

The success of deep neural networks (DNNs) encourages extensive applications on various types of platforms, e.g., self-driving cars and VR headsets. To overcome the hardware constraints, DNN model compression techniques, from learning based [1, 2, 3] to network architecture de-

sign [4, 5, 6], recently attracted a lot of attention. Interestingly, most of these extremely compact DNN models do not show accuracy drop on image classification. A critical question emerges, however, other than classifying images on a single dataset, whether the compression methods hurt DNN's learning ability.

In this work, we attempt to bridge the gap between compressed DNN architecture and its domain adaptation (DA) ability. The DA ability is to evaluate whether a machine learning model can capture the *covariate shift* [7] between source and target domains, and adapt itself to remove the divergence. A model with outstanding semi-supervised or unsupervised DA ability can greatly reduce the requirement of manually labeled examples for real-world applications.

We observe DA accuracy degradation from model compression methods based on architecture design, e.g., a DNN with GoogLeNet-level [9] classification accuracy only obtains AlexNet-level [8] DA accuracy. Table 1 shows our experimental results. SqueezeNet [4] and FaConvNet [5] are used to compare with AlexNet as they are respectively the smallest DNN model achieving AlexNet-level and GoogLeNet-level accuracy on image classification, to our best knowledge. The popular dataset ImageNet'12 [10] is adopted as image classification benchmark. Three standard DA tasks on Office-31 [11] dataset are adopted, and the unsupervised DA method used for all DNNs in Table 1 is GRL [12]. The DNNs are pre-trained on ImageNet'12, and then fine-tuned for all DA tasks. There is a big DA accuracy difference between AlexNet and SqueezeNet though the two networks have almost the same classification accuracy. FaConvNet, which outperforms AlexNet by 12.9% on classification, also slightly lags behind AlexNet on DA.

Intuitively, increasing parameters will lead to better accuracy. Our following experiment shows that the DA accuracy of SqueezeNet and FaConvNet can be improved, but can not reach the same level as their classification by solely

*Part of this work was done while C. Wu was an intern at LG San Jose Lab.

Table 1: Image classification and unsupervised DA accuracy of DNN models on Office-31 dataset.

	#Parameter	Classification	Task1	Task2	Task3
			AMAZON WEBCAM	DSLR WEBCAM	WEBCAM DSLR
AlexNet [8]	61 M	57.2	73.0	96.4	99.2
FaConvNet [5]	2.8 M	70.1	71.8	94.3	98.1
SqueezeNet [4]	1.2 M	57.5	64.4	92.8	96.4
Rev-FaConvNet	4.8 M	70.3	74.1	96.5	99.2
Rev-SqueezeNet	2.2 M	57.9	66.9	93.9	98.8

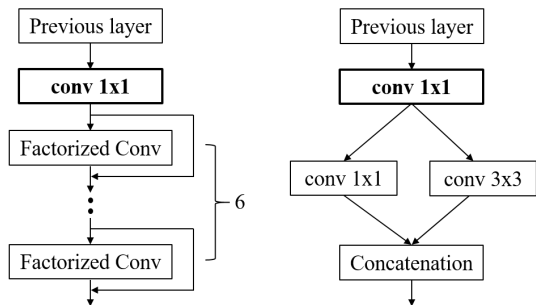


Figure 1: Basic modules adopted in FaConvNet [5] (left) and SqueezeNet [4] (right). Both modules use the “bottleneck” layer as shown in bold.

boosting parameter numbers. Specifically, without changing the structure of the two models, we increase the parameters of FaConvNet and SqueezeNet. The basic modules respectively adopted in FaConvNet and SqueezeNet are first compared, as shown in Figure 1. The shared feature of these two modules is the “bottleneck” layer $conv\ 1 \times 1$ as denoted in bold. We hence gradually increase parameters of all “bottleneck” layers in FaConvNet and SqueezeNet until no DA accuracy benefit could be obtained. The parameters in other layers (e.g., the first convolutional layer in FaConvNet and SqueezeNet) are then increased until no accuracy gain. The final DA accuracy of the adapted models *Rev-FaConvNet* and *Rev-SqueezeNet* are respectively shown in Table 1. Our expectation is that *Rev-FaConvNet*’s accuracy can be much higher than AlexNet. *Rev-FaConvNet*, however, only slightly outperforms AlexNet, with almost 70% more parameters.

The objective of this work is to develop a compact DNN architecture which can achieve the same level accuracy on classification and DA. Our solution offers four important features. First, our DNN has 4.1M parameters, which is only 6.7% of AlexNet or 59% of GoogLeNet. The compactness of our network can be attributed to the use of a new module *Conv-M* which is a parameter-saving module, while extract more details based on multi-scale convolution and deconvolution, inspired by GoogLeNet’s *Inception*. Sec-

ond, our DA method consists of three components: Learning invariance across domains, reducing discrepancy of feature representations, and predicting labels. Third, experiments show that our DNN obtains GoogLeNet-level accuracy both on classification and DA. The DA accuracy gap between GoogLeNet and other compact DNNs (FaConvNet and Rev-FaConvNet) is much larger. Fourth, the unified framework of our DA method slightly outperforms previous competitive methods, and our DA method based on our DNN network achieves state-of-the-art on sixteen of total eighteen DA tasks on the popular Office-31 and Office-Caltech [13] datasets.

2. Related Work

DNN model compression with little accuracy drop on image classification traditionally are learning based. Liu *et al.* [1] zero out more than 90% of AlexNet’s parameters using a sparse decomposition, while Wen *et al.* [3] regularize a DNN model with structured sparsity based on group Lasso. Han *et al.* [2] prune the small-weight connections and retrain the DNN with the remaining connections. More recent research began to shrink a model directly based on network architecture design. SqueezeNet [4] is built on the *fire* module which feeds “squeeze” layer (1×1 convolution) into “expand” layer (a combination of 1×1 and 3×3 convolution). The basic structure of FaConvNet [5] is *Convolutional Layer as Stacked Single Basis Layer*. A popular design methodology of compact architectures extensively uses small convolutional kernels (1×1 and 3×3), especially the linear projection as the $conv\ 1 \times 1$ layer shown in bold in Figure 1. Based on the preliminary experimental result in Table 1, we argue that it is necessary to redesign the basic module of these extremely shrunk DNNs, e.g., FaConvNet and SqueezeNet, by introducing more diverse operations of feature extraction, in order to achieve high accuracy on both classification and DA. The challenge lies in that more complex feature extraction methods, e.g., multi-scale convolution, often result in the steep increase of parameters, as the basic module will be used repeatedly. The shortcut connection used in ResNet [14], for instance, can be under-

stood as a parameter-saving solution of multi-scale feature integration. We will adopt methods other than this bypass structure.

Unsupervised DA. Following the early attempt of re-weighting samples from source domain [15], Shekhar *et al.* [16] learn dictionary based representations by minimizing the divergence between the source and target domains. The subspace based methods, on the other hand, evaluate the distance between domains in a low-dimensional manifold [13] or in terms of Frobenius norm [17]. DNN based methods have been proposed recently. Glorot *et al.* [18] and Chopra *et al.* [19] learn cross-domain features using auto-encoders, followed by the label prediction. A more popular strategy is to combine feature adaptation with label prediction as a unified framework. DDC [20] introduces adaptation layers and domain confusion metric into a CNN architecture, while GRL [12] combines classifiers of label and domain using a gradient reverse layer. DAN [21] and RTN [22] focus on effectively measuring feature representations in kernel spaces. TRANSDUCTION [23] jointly optimizes the target label and domain transformation parameters. Our DA method adopts a unified framework, which can simultaneously learn invariance across domains, reduce divergence of feature representations and adapt label prediction.

DNN based image segmentation. The DNNs of segmentation and classification mainly differ in the use of up-sampling layers to recover resolution. Various up-scaling methods have been proposed and adopted, such as straightforward bicubic interpolation [24], learning based deconvolution [25], and unpooling [26, 27]. We improve the deconvolution [25] to remove artifacts that will be described in Section 3.1, and use it as a type of shape feature extractor in the basic module of our DNN. With the consideration of training convergence speed, the unpooling with fewer parameters is a better choice, compared to deconvolution, especially for small-scale and medium-scale problems. So we adopt unpooling for sample reconstruction in our DA method. In addition, different strategies have been presented to train segmentation networks. SegNet-Basic [27] is directly trained as a whole. Long *et al.* [28], on the other hand, adapt a popular classification network into a fully convolutional network (FCN), and fine-tune it for segmentation tasks. Yu *et al.* [29] show that accuracy can be further improved by plugging their context module into existing segmentation model. Our decoder design for sample reconstruction is inspired by FCN, while our structure is simpler than the multi-stream structure in FCN.

3. Proposed Method

Motivated by the observation described in Section 1, we propose a compact DNN architecture with a new basic module Conv-M. Our DA method gradually tunes the feature

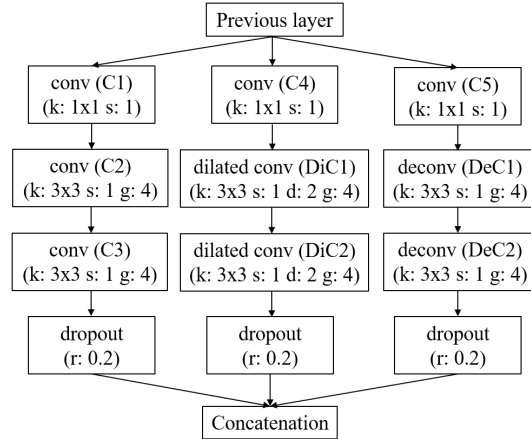


Figure 2: Module Conv-M used in our DNN. The output of *deconv* is cropped to its input size. The ReLU is adopted for all types of convolution, which is not shown in the figure for simplicity.

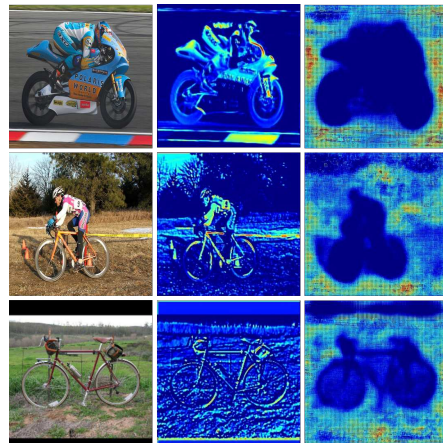


Figure 3: Visualization of activations in the same Conv-M module in our network: Convolution (middle) and deconvolution (right).

adaptation and label prediction.

3.1. DNN Architecture with Conv-M

Figure 2 shows a Conv-M module used in our DNN. According to the preliminary experiment and our analysis in Section 1, the design idea is to capture more diverse details at different levels, while using fewer parameters. To achieve this goal, the dilated convolution [29] for multi-resolution and deconvolution [25] are introduced. The dilated convolution can extract features with a larger receptive field without increasing the kernel size, *e.g.*, extracting features from a 5×5 window with a 3×3 kernel. The deconvolution is to reconstruct shapes of the input, providing distinct features from regular convolution. In addition, to decrease redundant parameters, we implement the separable convolution

Table 2: Our DNN architecture (Basic parameter settings of the module Conv-M are shown in Figure 2).

Layer	Type/Module	Output size	Filter size/Stride (If not Conv-M)	#Feature maps (Conv-M)										#Parameters
				C1	C2	C3	C4	DiC1	DiC2	C5	DeC1	DeC2		
1	input	224×224×3												
2	convolution	224×224×64	7×7/1 (x64)											9,408
3	max-pooling	112×112×64	3×3/2											
4	Conv-M	112×112×160		64	64	64	64	64	64	32	32	32		51,712
5	max-pooling	56×56×160	3×3/2											
6	Conv-M	56×56×320		128	128	128	128	128	128	64	64	64		217,088
7	Conv-M	56×56×320		128	128	128	128	128	128	64	64	64		268,288
8	max-pooling	28×28×320	3×3/2											
9	Conv-M	28×28×576		144	256	256	144	256	256	64	64	64		591,872
10	Conv-M	28×28×576		144	256	256	144	256	256	64	64	64		681,984
11	max-pooling	14×14×576	3×3/2											
12	Conv-M	14×14×688		160	256	280	160	256	280	64	128	128		783,360
13	Conv-M	14×14×688		160	256	280	160	256	280	64	128	128		826,368
14	avg-pooling	1×1×688	14×14/1											
15	linear	1×1×1000	1×1/1 (x1000)											688,000
														4.1 M

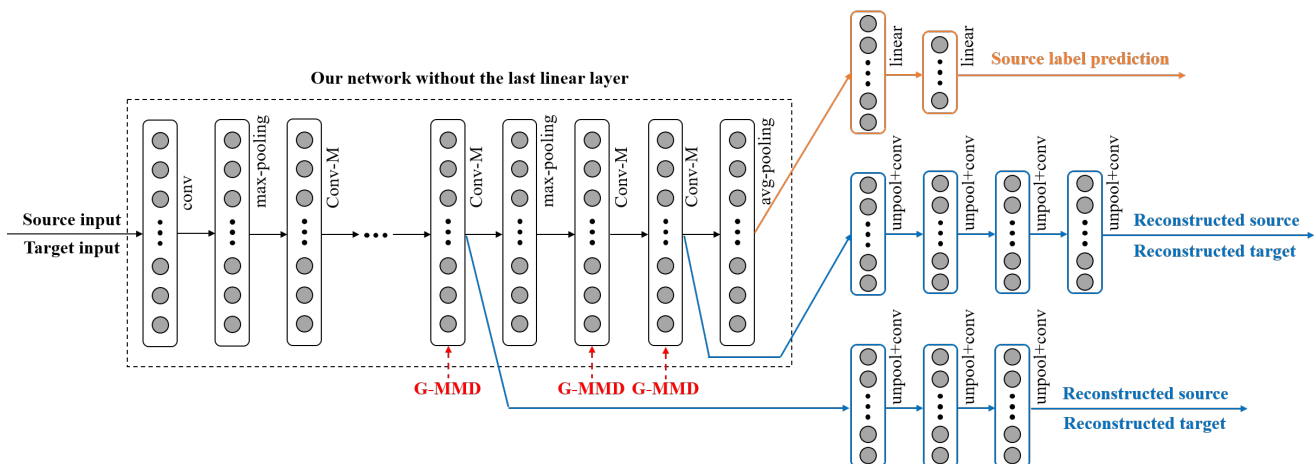


Figure 4: The unified framework of our DA method. The DNN simultaneously adapts feature representations (red and blue) and source label prediction (orange). The sampling ratio of target domain will be gradually increased during training.

inspired by separable wavelet filters [30] for all types of convolution, including deconvolution, in Conv-M.

We visualize activations of convolution (middle) and deconvolution (right) in the same Conv-M module in our network in Figure 3. Appearance details are extracted by convolution, while deconvolution tends to describe the completed shapes. Therefore, the features extracted by convolution and deconvolution are *complementary* so as to benefit DA. In addition, the shapes captured by deconvolution are more *generic* for a class of object compared to the appearance details extracted by convolution, which facilitates our DA strategy to explore divergence between classes for knowledge transfer.

The detailed design of Conv-M in Figure 2 shows that

the input feature maps from the previous layer are respectively processed by regular convolution (*conv*), dilated convolution (*dilated conv*) and deconvolution (*deconv*) in three branches. Their outputs will be concatenated together. The pipelines of these three branches are: *C1-C2-C3-dropout*, *C4-DiC1-DiC2-dropout*, and *C5-DeC1-DeC2-dropout*. All of the three branches start with a 1×1 convolution as linear projection. The parameters k and s are kernel size and stride. The dilation factor d indicates that the receptive field is $(2^{d+1} - 1) \times (2^{d+1} - 1)$. The group number g for separable convolution indicates that feature maps between two adjacent layers are separated into g groups. The dropout ratio r is fixed to 0.2. The output of deconvolution is cropped to its input size. ReLU is adopted for all nine convolutions, which is not shown in Figure 2. The parameter number of

Conv-M is computed as follows. Let $N_P, N_{C1}, N_{C2}, N_{C3}, N_{C4}, N_{DiC1}, N_{DiC2}, N_{C5}, N_{DeC1}$ and N_{DeC2} denote the feature map numbers of $C1, C2, C3, C4, DiC1, DiC2, C5, DeC1$ and $DeC2$. The parameter number of the first branch in Conv-M is:

$$N_P \cdot N_{C1} + \frac{N_{C1} \cdot N_{C2} \cdot k_{C2}^2}{g_{C2}} + \frac{N_{C2} \cdot N_{C3} \cdot k_{C3}^2}{g_{C3}}. \quad (1)$$

The parameter number of the second branch is:

$$N_P \cdot N_{C4} + \frac{N_{C4} \cdot N_{DiC1} \cdot k_{DiC1}^2}{g_{DiC1}} + \frac{N_{DiC1} \cdot N_{DiC2} \cdot k_{DiC2}^2}{g_{DiC2}}. \quad (2)$$

The parameter number of the third branch is:

$$N_P \cdot N_{C5} + \frac{N_{C5} \cdot N_{DeC1} \cdot k_{DeC1}^2}{g_{DeC1}} + \frac{N_{DeC1} \cdot N_{DeC2} \cdot k_{DeC2}^2}{g_{DeC2}}. \quad (3)$$

Our DNN architecture is shown in Table 2, which generally consists of *convolution*, alternating *max-pooling* and *Conv-M*, *avg-pooling* and *linear*, as listed in the second column *Types/Module*. Note that the last linear layer is for image classification only and will be removed when conducting DA tasks. To fairly compare with other DA methods in Section 4, we include this layer into the estimation of total parameters as shown in the table. The *Output size* in the third column is multiplication of height, width and number of feature maps at each layer. Specific parameters of a non Conv-M layer are listed in the fourth column *Filter size/Stride*, while those of Conv-M are in the fifth column *#Feature maps (Conv-M)*. As the basic settings of Conv-M are represented in Figure 2, the fifth column only shows the feature map number of all nine convolutions: $C1, C2, C3, C4, DiC1, DiC2, C5, DeC1$ and $DeC2$. For each of these nine convolutions, the feature map numbers between two max-pooling layers are same, and generally increased with the model depth. The raw pixels of input images are processed by a regular convolution with a kernel size of 7×7 which is much larger than the 1×1 and 3×3 kernels used in Conv-M. Our preliminary experiment shows that for input image data, convolution with a smaller kernel (e.g., 3×3) will degrade the classification accuracy by 1.5%~2.5%. For Conv-M, on the other hand, using larger kernels (e.g., 5×5) can only improve the performance by slightly 0.3%~0.8%. The final column *#Parameters* in Table 2 lists the parameter numbers at each layer. Dominant parameter consumers are the two Conv-M modules (39%) between the fourth max-pooling and the avg-pooling. The total number of parameters of our DNN is 4.1M.

3.2. Unsupervised Domain Alignment

Our DA method simultaneously adapts feature representations and source label prediction as shown in Figure 4,

given input data sampled from both source and target domains. The sampling ratio of target domain will be gradually increased during training. Formally, three terms are minimized in the unified framework: The reconstruction error of source and target samples (blue) for invariance learning, the discrepancy of hidden representations on layers between domains (red), and the prediction error of source labels (orange). For our DNN shown in Table 2, the last linear layer with 1000 neurons will be removed in DA tasks. Extra layers, as shown in orange and blue in Figure 4, are added during domain alignment training, while only the layers related to label prediction (orange) will be kept for testing.

Invariance learning. The error minimization of reconstructing input source and target samples is to force the DNN to learn more cross-domain features. The asymmetrical encoder-decoder architecture is adopted for sample reconstruction, as shown in Figure 4. The encoder is our pre-trained DNN without the avg-pooling and last linear layers, while the decoder (blue) with fewer layers (compared to the encoder) consists of alternating un-pooling and regular convolution. The un-pooling in the decoder is to up-sample input feature maps using indexes obtained from the corresponding max-pooling layer in the encoder. The encoder is responsible for feature extraction, while the decoder is for restoring resolution. Our preliminary experiment shows that the asymmetrical structure only slightly decreases the final accuracy (averagely 0.4%) but significantly accelerates the training speed, compared to symmetrical design. In addition, two decoders on different scales are introduced.

Representation discrepancy reduction. Instead of using parametric criteria such as *Kullback-Leibler divergence* to further reduce the cross-domain divergence, we adopt a non-parametric method to estimate the feature distribution distance between domains. Specifically, we minimize the maximum mean discrepancies (MMD) by Gretton *et al.* [31]. The MMD is defined as:

$$L_M = \left\| \frac{1}{N_s} \sum_1^{N_s} \psi(x_s) - \frac{1}{N_t} \sum_1^{N_t} \psi(x_t) \right\|_{\mathcal{H}}^2, \quad (4)$$

where x_s and x_t are respectively input source and target, and N_s and N_t denote corresponding sample numbers. The function $\psi(\cdot)$ is a non-linear feature mapping. \mathcal{H} is a universal reproducing kernel Hilbert space. The MMD criteria is denoted as G-MMD in our method, as we adopt the Gaussian kernel. As shown in Figure 4, the G-MMD loss (red) is added to the last three Conv-M layers in our DNN.

Source label prediction. As shown in Figure 4, we add two linear layers (orange), and the neuron numbers of the second one is specified for the dataset. No significant accuracy benefit is observed by adding linear layers more than two in our preliminary experiment.

Table 3: The comparison of our network and popular DNNs on ImageNet’12 classification accuracy and parameter numbers.

Method	#Parameters	Top-1	Top-5
AlexNet [8]	61 M	57.2	80.3
GoogLeNet [9]	7 M	68.7	88.9
VGG16 [32]	134 M	71.9	90.6
Our network	4.1 M	68.9	89.0

4. Experiments

Our DNN is trained on the benchmark dataset ImageNet’12 [10] and compared with well-known models on total parameter numbers and classification accuracy. Following the standard pipeline, we then fine-tune our trained model for unsupervised DA tasks on two popular datasets according to our DA method. The DA accuracy is compared with competitive methods.

4.1. ImageNet Classification

We train our DNN on ImageNet’12 dataset, and set the parameters of our training solver according to the *quick_solver.prototxt* in Caffe [33]. The batch size is 64. Table 3 compares the classification accuracy (*Top-1*, *Top-5*) and parameter numbers (*#Parameters*) of our DNN and AlexNet [8], GoogLeNet [9], and VGG16 [32]. For AlexNet and GoogLeNet, we directly use the trained models provided by Caffe. The VGG16’s result is obtained from the original paper [32]. Our DNN achieves GoogLeNet-level accuracy, while the total parameter numbers (4.1M) is only 59% of GoogLeNet.

4.2. Unsupervised DA

Office-31. This standard benchmark consists of 4,652 images of 31 categories collected from three distinct domains [11]: *AMAZON* (A), *WEBCAM* (W) and *DSLR* (D). The samples of these three domains are respectively downloaded from *amazon.com*, taken by web camera and taken by digital SLR camera in an office environment with different photographic settings. All six DA tasks between the three domains will be adopted for completeness: $A \rightarrow W$, $D \rightarrow W$, $W \rightarrow D$, $W \rightarrow A$, $A \rightarrow D$ and $D \rightarrow A$.

Office-Caltech. It is a popular dataset [13] composed of 10 overlapping categories from the Office-31 and Caltech-256 (C) [36] datasets. All twelve DA tasks are used: $A \rightarrow W$, $D \rightarrow W$, $W \rightarrow D$, $A \rightarrow D$, $D \rightarrow A$, $W \rightarrow A$, $A \rightarrow C$, $W \rightarrow C$, $D \rightarrow C$, $C \rightarrow A$, $C \rightarrow W$ and $C \rightarrow D$. The Office-31 dataset is more challenging as it has more categories of images, while Office-Caltech provides more DA tasks to observe the dataset bias [37].

Methods. We compare our method with the nine pre-

vious competitive DA methods: TCA [35], GFK [34], SA [17], DLID [19], DDC [20], DAN [21], GRL [12], TRANSDUCTION [23] and RTN [22]. TCA and GFK are conventional methods, while the others are DNN based.

Networks. Five DNNs are used in our experiments: AlexNet (61M), Rev-FaConvNet (4.8M), our DNN (4.1M), GoogLeNet (7M) and FaConvNet (2.8M). DA methods DAN, GRL, TRANSDUCTION and RTN originally use pre-trained AlexNet, according to their papers. Rev-FaConvNet achieves much better DA accuracy compared to SqueezeNet, Rev-SqueezeNet and FaConvNet as shown in our preliminary experiments in Table 1. FaConvNet, Rev-FaConvNet and our DNN all reach GoogLeNet-level classification accuracy. In this work, we use GoogLeNet and FaConvNet as baselines for comparison.

Experiments. Besides running previous DA methods on AlexNet, we also run the following eight experiments to quantize the contribution of our DNN and our DA method: (1) GRL (Rev-FaConvNet): Running GRL on Rev-FaConvNet; (2) GRL (Our net): Running GRL on our DNN; (3) DAN (Rev-FaConvNet): Running DAN on Rev-FaConvNet; (4) DAN (Our net): Running DAN on our DNN; (5) Our DA (Rev-FaConvNet): Running our DA method on Rev-FaConvNet; (6) Our DA (FaConvNet): Running our DA method on FaConvNet, and the result is used as a baseline; (7) Our DA (GoogLeNet): Running our DA method on GoogLeNet, and the result is used as a baseline; (8) Our DA (Our net): Running our DA method on our DNN, and this is **our final result**.

Parameter settings. We follow the specific description of all previous DA methods in their papers. The hyper-parameter of SA is selected based on cross-validation, which is consistent with other papers [12, 23]. For our DA method that is based on our pre-trained network on ImageNet’12, the *convolution* and the first three *Conv-M* shown in Table 2 are frozen, as the Office-31 and Office-Caltech datasets are rather small-scale. For all newly added layers as shown in orange and blue in Figure 4 which are trained from scratch, their learning rate is ten times higher. The learning rate policy we adopt is *poly* as described in Caffe, and the initial value is 0.0009 with the power fixed to 0.5. The batch size is 64, and the sampling ratio of target domains is uniformly increased from 30% to 70% during training. In the testing stage, the new layers for sample reconstruction are removed, as aforementioned in Section 3.2. For the remaining new layers for label prediction (orange) in Figure 4, the neuron numbers of the first linear layer is 256, while those of the second one is 31 for Office-31 dataset and 10 for Office-Caltech dataset. The G-MMD loss is added to the last three Conv-M layers of our DNN. The regularization

Table 4: Unsupervised DA accuracy of our method and previous algorithms on Office-31 dataset.

Method	#Parameters ¹	A→W	D→W	W→D	W→A	A→D	D→A
GFK [34]	-	39.8	79.1	74.6	37.1	37.9	37.9
SA [17]	-	45.0	64.8	69.9	39.3	38.8	42.0
DLID [19]	-	51.9	78.2	89.9	-	-	-
DDC [20]	-	61.8	95.0	98.5	52.2	64.4	52.1
DAN [21]	61 M	68.5	96.0	99.0	53.1	67.0	54.0
GRL [12]	61 M	73.0	96.4	99.2	53.6	72.8	54.4
TRANSDUCTION [23]	61 M	80.4	96.2	98.9	62.5	83.9	56.7
GRL (Rev-FaConvNet)	4.8 M	74.1	96.5	99.2	54.3	73.4	55.3
Our DA (Rev-FaConvNet)	4.8 M	77.0	96.5	99.2	58.4	75.9	58.1
GRL (Our net)	4.1 M	80.1	96.7	99.2	64.1	78.0	65.4
Our DA (Our net)	4.1 M	82.6	97.0	99.4	67.4	80.1	67.3
Baseline: Our DA (GoogLeNet)	7 M	83.0	96.9	99.5	67.7	80.5	67.5
Baseline: Our DA (FaConvNet)	2.8 M	73.9	96.3	99.1	54.1	73.2	55.2

¹ Most of methods will remove the last linear layer of a pre-trained network, and add extra layers for DA. According to Section 4.2, our DNN will be smaller after the change. The size of other models will also be slightly different, but the actual size is not reported in [21, 23]. We hence directly report the total parameter numbers of the pre-trained network for fair comparison.

Table 5: Unsupervised DA accuracy of our method and previous algorithms on Office-Caltech dataset.

Method	#Param. ¹	A→W	D→W	W→D	A→D	D→A	W→A	A→C	W→C	D→C	C→A	C→W	C→D
TCA [35]	-	84.4	96.9	99.4	82.8	90.4	85.6	81.2	75.5	79.6	92.1	88.1	87.9
GFK [34]	-	89.5	97.0	98.1	86.0	89.8	88.5	76.2	77.1	77.9	90.7	78.0	77.1
DDC [20]	-	86.1	98.2	100.0	89.0	89.5	84.9	85.0	78.0	81.1	91.9	85.4	88.8
DAN [21]	61 M	93.8	99.0	100.0	92.4	92.0	92.1	85.1	84.3	82.4	92.0	90.6	90.5
RTN [22]	61 M	97.0	98.8	100.0	94.6	95.5	93.1	88.5	88.4	84.3	94.4	96.6	92.9
DAN (Rev-FaConvNet)	4.8 M	94.0	99.1	100.0	92.7	92.3	92.2	85.5	84.6	82.6	92.3	90.9	90.8
Our DA (Rev-FaConvNet)	4.8 M	94.9	99.2	100.0	93.3	93.3	92.5	86.5	85.9	83.1	93.0	93.0	91.5
DAN (Our net)	4.1 M	95.0	99.2	100.0	96.0	94.8	95.2	91.6	90.4	90.7	94.4	95.0	94.3
Our DA (Our net)	4.1 M	95.6	99.7	100.0	96.8	96.0	95.6	92.5	91.6	91.4	95.3	97.2	95.3
Baseline: Our DA (GoogLeNet)	7 M	95.9	99.7	100.0	97.1	96.2	95.9	92.9	92.0	91.5	95.6	97.4	95.7
Baseline: Our DA (FaConvNet)	2.8 M	94.5	99.1	99.8	92.0	91.8	91.0	83.7	83.4	80.1	92.8	91.1	89.8

¹ Please see the footnote of Table 4 for the explanation of parameter numbers.

hyper-parameter of G-MMD loss is fixed to 0.3 across all datasets, and the bandwidth of the Gaussian kernel is the median pairwise distance [38] on training set.

Based on NVIDIA GTX TITAN X, the inference speed of SqueezeNet and Rev-SqueezeNet is faster than that of FaConvNet, Rev-FaConvNet and our network, though they cannot obtain GoogLeNet-level classification and DA. Specifically, Rev-SqueezeNet is 22% slower than that of SqueezeNet, and Rev-FaConvNet decreases the speed of FaConvNet by 12%. Our network consumes 11% less time compared to FaConvNet.

Table 4 and Table 5 respectively summarize the DA accuracy on Office-31 and Office-Caltech datasets. Both tables are separated into four groups by rows. The first group is the previous DA methods based on AlexNet. The second group compares previous and our DA methods on Rev-FaConvNet, while the third group compares DA methods on our DNN. The fourth group provides result of our DA

method on GoogLeNet and FaConvNet as baselines. The results in the two tables are analyzed from the following three aspects:

First, our DNN approaches GoogLeNet’s DA accuracy on the same DA method, while the gap between GoogLeNet and previous compact DNNs (FaConvNet and Rev-FaConvNet) is much larger, according to the four observations: *Our DA (Our net)*, *Our DA (GoogLeNet)*, *Our DA (FaConvNet)* and *Our DA (Rev-FaConvNet)* in Table 4 and Table 5. Though FaConvNet, Rev-FaConvNet and our DNN all obtain GoogLeNet-Level classification accuracy, only our DNN has matched accuracy on both classification and DA. Moreover, our DNN (4.1M) is smaller than Rev-FaConvNet (4.8M). Our DNN also outperforms AlexNet using the same DA method, as the comparison of *GRL* and *GRL (Our net)* in Table 4 shows.

Second, our DA method outperforms GRL and DAN, based on the same DNN, according to the four comparisons:

Table 6: Contribution of non-regular convolution in our Conv-M module on Office-31 dataset.

	#Parameter	Classification	A→W	D→W	W→D	W→A	A→D	D→A
Our DA (Our net1)	4.1 M	62.2	74.2	96.5	99.2	56.2	74.1	56.0
Our DA (Our net)	4.1 M	68.9	82.6	97.0	99.4	67.4	80.1	67.3

Table 7: DA accuracy of our method without including specified component on Office-31 dataset.

Method	A→W	D→W	W→D	W→A	A→D	D→A
No G-MMD	76.7	96.5	99.2	62.0	77.5	64.7
No recons.	79.6	95.4	99.3	64.4	77.3	62.1
All	82.6	97.0	99.4	67.4	80.1	67.3

Table 8: DA accuracy of our method without including specified component on Office-Caltech dataset.

Method	A→W	D→W	A→D	A→C	W→C	D→C
No G-MMD	91.1	99.6	93.4	90.9	87.1	87.8
No recons.	93.9	99.4	95.0	88.7	89.8	86.6
All	95.6	99.7	96.8	92.5	91.6	91.4

GRL (Rev-FaConvNet) and Our DA (Rev-FaConvNet) in Table 4, GRL (Our net) and Our DA (Our net) in Table 4, DAN (Rev-FaConvNet) and Our DA (Rev-FaConvNet) in Table 5, and DAN (Our net) and Our DA (Our net) in Table 5.

Third, put all together, our DA method based on our DNN achieves state-of-the-art on sixteen of total eighteen DA tasks on two datasets, as shown on the last row of these two tables (Our DA (Our net)). The other two is $A \rightarrow D$ in Table 4 and $A \rightarrow W$ in Table 5. We boost the accuracy of task $D \rightarrow A$ by 10.6% compared to TRANSDUCTION, as shown in Table 4. On Office-31 dataset, the accuracy gap between the tasks $D \rightarrow W$ and $W \rightarrow D$ is 2.4%, while the gap between $A \rightarrow W$ and $W \rightarrow A$ greatly increases to 15.2%, indicating larger appearance difference between domains A and W. The domain difference between A and D is also larger than that between D and W. In other words, on Office-31 dataset, transfer (in two directions) between D and W is relatively easier for our DA method, while other two are more difficult, which is consistent with the results from previous DA methods. On Office-Caltech dataset, the bilateral transfer between C and W gets the largest accuracy gap (5.6%) in our DA method, as shown in Table 5.

4.3. Sensitivity Analysis

Convolution in Conv-M. To validate the contribution of non-regular convolution (dilated convolution and improved deconvolution) in our Conv-M module, we replace all non-regular convolution with regular ones and keep the 3×3 kernel size unchanged. The first row *Our DA (Our net1)* in Table 6 shows the result, and the second row *Our DA (Our net)* is our original solution. Significant accuracy drop can

be observed on classification and almost all DA tasks. The comparison in Table 6 indicates the importance of features extracted by dilated convolution and improved deconvolution in our Conv-M.

Reconstruction and G-MMD. Based on our DNN, Table 7 and Table 8 respectively show the contribution of two components of our DA methods (sample reconstruction and G-MMD) on Office-31 and Office-Caltech datasets. The row *No G-MMD* in two tables shows the result obtained by removing G-MMD from our DA method, while the row *No recons.* corresponds to our method without including sample reconstruction. For these two rows, lower accuracy indicates more contribution of the component. The row *All* is the regular result without removing any component, which is the same as the respective row *Our DA (Our net)* in Table 4 and Table 5. For Office-31 dataset shown in Table 7, reconstruction is more important for the transfers $D \rightarrow W$ and $D \rightarrow A$, while $A \rightarrow W$ and $W \rightarrow A$ rely more on G-MMD. Table 8 demonstrates that the contributions of reconstruction and G-MMD are almost the same.

5. Conclusion

In this paper, we present a compact DNN architecture and unsupervised DA method, based on our observation that current small DNNs (SqueezeNet and FaConvNet) have unmatched accuracy on classification and DA, e.g., a DNN with GoogLeNet-level classification accuracy only obtains AlexNet-level DA accuracy. The basic module used in our DNN, Conv-M, introduces multi-scale convolution and deconvolution without using kernels larger than 3×3 . The unified framework of our DA method learns cross-domain features by sample reconstruction and G-MMD, and simultaneously tunes label prediction. The parameter numbers of our DNN is only 59% of GoogLeNet, while experiments show that our DNN obtains GoogLeNet-level accuracy both on classification and DA. Our DA method slightly outperforms previous competitive GRL and DA. In addition, our method based on our DNN achieves state-of-the-art on sixteen of total eighteen DA tasks on the popular Office-31 and Office-Caltech datasets.

Acknowledgments. This work is in part supported by NSF CCF-1615475 and DOE SC0017030. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of grant agencies or their contractors.

References

- [1] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky. Sparse Convolutional Neural Networks. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [2] S. Han, H. Mao, and W. J. Dally. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *International Conference on Learning Representations (ICLR)*, 2016.
- [3] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning Structured Sparsity in Deep Neural Networks. *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [4] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and <0.5MB Model Size. *arXiv preprint arXiv:1602.07360*, 2016.
- [5] M. Wang, B. Liu, and H. Foroosh. Factorized Convolutional Neural Networks. *arXiv preprint arXiv:1508.04337*, 2016.
- [6] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv preprint arXiv:1606.02147*, 2016.
- [7] H. Shimodaira. Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference*, 2000.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Network. *Advances in Neural Information Processing Systems (NIPS)*, 2012.
- [9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and S. Reed. Going Deeper with Convolutions. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. F. Li. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.
- [11] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting Visual Category Models to New Domains. *European Conference on Computer Vision (ECCV)*, 2010.
- [12] Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation. *International Conference on Machine Learning (ICML)*, 2015.
- [13] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic Flow Kernel for Unsupervised Domain Adaptation. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- [15] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf. Correcting Sample Selection Bias by Unlabeled Data. *Advances in Neural Information Processing Systems (NIPS)*, 2006.
- [16] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa. Generalized Domain-Adaptive Dictionaries. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [17] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised Visual Domain Adaptation Using Subspace Alignment. *International Conference on Computer Vision (ICCV)*, 2013.
- [18] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. *International Conference on Machine Learning (ICML)*, 2011.
- [19] S. Chopra, S. Balakrishnan, and R. Gopalan. DLID: Deep Learning for Domain Adaptation by Interpolating between Domains. *International Conference on Machine Learning Workshop (ICMLW)*, 2013.
- [20] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv preprint arXiv:1412.3474*, 2014.
- [21] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning Transferable Features with Deep Adaptation Networks. *International Conference on Machine Learning (ICML)*, 2015.
- [22] M. Long, J. Wang, and M. I. Jordan. Unsupervised Domain Adaptation with Residual Transfer Networks. *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [23] O. Sener, H. O. Song, A. Saxena, and S. Savarese. Learning Transferable Representations for Unsupervised Domain Adaptation. *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [24] C. Dong, C. C. Loy, K. He, and X. Tang. Image Super-Resolution Using Deep Convolutional Networks. *arXiv preprint arXiv:1501.00092*, 2015.
- [25] H. Noh, S. Hong, and B. Han. Learning Deconvolution Network for Semantic Segmentation. *International Conference on Computer Vision (ICCV)*, 2015.
- [26] S. Hong, H. Noh, and B. Han. Decoupled Deep Network for Semi-Supervised Semantic Segmentation. *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [27] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv preprint arXiv:1511.00561*, 2015.
- [28] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

- [29] F. Yu and V. Koltun. Multi-scale Context Aggregation by Dilated Convolutions. *International Conference on Learning Representations (ICLR)*, 2016.
- [30] L. Sifre and S. Mallat. Rotation, Scaling and Deformation Invariant Scattering for Texture Discrimination. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [31] A. Gretton, K. M. Borgwardt, M. Rasch, B. Scholkopf, and A. J. Smola. A Kernel Method for the Two-Sample-Problem. *Advances in Neural Information Processing Systems (NIPS)*, 2006.
- [32] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICLR)*, 2015.
- [33] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *ACM International Conference on Multimedia*, 2014.
- [34] B. Gong, K. Grauman, and F. Sha. Connecting the DOTs with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation. *International Conference on Machine Learning (ICML)*, 2013.
- [35] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2011.
- [36] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. *Technical Report, California Institute of Technology*, 2007.
- [37] A. Torralba and A. Efros. Unbiased look at dataset bias. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [38] A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal Kernel Choice for Large-Scale Two-Sample Tests. *Advances in Neural Information Processing Systems (NIPS)*, 2012.