Adversarial Discriminative Domain Adaptation (workshop extended abstract)

Eric Tzeng
University of California, Berkeley
etzeng@eecs.berkeley.edu

Judy Hoffman
Stanford University
jhoffman@cs.stanford.edu

Kate Saenko
Boston University
saenko@bu.edu

Trevor Darrell
University of California, Berkeley
trevor@eecs.berkeley.edu

1 Introduction

Many recent domain adaptation methods attempt to learn transformations that map both domains into a common feature space. This is generally achieved by optimizing the representation to minimize some measure of domain shift such as maximum mean discrepancy (Tzeng et al., 2014; Long & Wang, 2015) or correlation distances (Sun et al., 2016). Adversarial adaptation methods, which are related to generative adversarial learning (Goodfellow et al., 2014), have become an increasingly popular incarnation of this type of approach which seeks to minimize an approximate domain discrepancy distance through an adversarial objective with respect to a domain discriminator.

In this work, we propose a unified view of recent adversarial domain adaptation methods, allowing us to effectively examine the different factors of variation between the approaches and clearly view the similarities they each share. By comparing their properties such as weight-sharing, base models, and adversarial losses, we are able to facilitate understanding of the effect these choices have on the resulting adaptation method. We use this insight to propose a simple yet novel and powerful unsupervised adversarial adaptation method, Adversarial Discriminative Domain Adaptation (ADDA), and show state-of-the-art visual adaptation results on the standard *Office* adaptation dataset. Additional discussion and results are available in the long form of this report (Tzeng et al., 2017).

2 Adversarial domain adaptation: a unified view

Existing adversarial adaptation methods all share a basic core idea: representations effective for adaptation are learned via the inclusion of an adversarial loss. However, these methods vary considerably in the particular details of their instantiations. Some methods seek to minimize the domain distance in a latent discriminatively learned recognition space (Tzeng et al., 2015; Ganin et al., 2016) while others look to minimize a domain distance in pixel space (Liu & Tuzel, 2016). Other approaches include optimizing both spaces simultaneously (Donahue et al., 2016). Some argue that a single shared representation should be learned (Ganin et al., 2016), while others claim that only part of the representation should be shared to enable the most effective adaptation (Liu & Tuzel, 2016; Yoo et al., 2016). Then there are more subtle differences in terms of the adversarial learning objective, using either a minimax loss (Goodfellow et al., 2014; Ganin et al., 2016), an inverted label minimax loss (Goodfellow et al., 2014), or a combination confusion loss (Tzeng et al., 2015). Each subsequent algorithm presents a new setting across these factors, but offers limited motivation or connection to prior work. This inherently limits our ability to understand the crucial components of each algorithm and, more importantly, determine how to combine them. We unify the existing methods and highlight their variations in Figure 1.

All methods learn mappings from source and target inputs to a common high-level feature space. The mappings are instantiated by a base network that can be discriminative, or can include both a discriminative and a generative component. Although there are many similarities to generative adversarial networks, we use the more general term “mapping,” since many adversarial adaptation methods do not rely on an actual image generator. The output of this mapping is fed into both a classifier (trained on source) and an adversarial loss that encourages a common feature space. The adversarial loss is minimized when a discriminator network cannot distinguish the domain label of

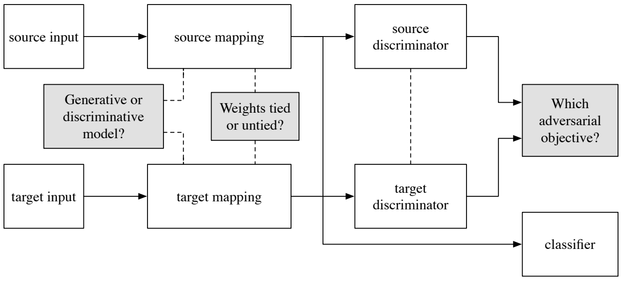

Figure 1: Adversarial adaptation methods can be viewed as instantiations of the same framework with different choices regarding their properties.

Table 1: Overview of adversarial domain adaption methods and their various properties. Viewing methods under a unified framework enables us to easily propose a new adaptation method, adversarial discriminative domain adaptation (ADDA).

Method	**Base model**	**Weight sharing**	**Adversarial loss**
Gradient reversal (Ganin et al., 2016)	discriminative	shared	minimax
Domain confusion (Tzeng et al., 2015)	discriminative	shared	confusion
CoGAN (Liu & Tuzel, 2016)	generative	unshared	GAN
ADDA (Ours)	discriminative	unshared	GAN

the input. The choices include whether the base mapping is generative or discriminative, whether its weights are tied or untied across domains, and which adversarial loss is used. We provide a summary of recent adversarial adaptation methods and their choices in Table 1.

Base model Because unsupervised domain adaptation generally considers discriminative tasks such as classification, previous adaptation methods have generally relied on adapting discriminative models between domains. With a discriminative base model, input images are mapped into a feature space that is useful for a discriminative task such as image classification. However, Liu and Tuzel achieve state of the art results on unsupervised MNIST-USPS using two generative adversarial networks Liu & Tuzel (2016). These generative models use random noise as input to generate samples in image space—generally, an intermediate feature of an adversarial discriminator is then used as a feature for training a task-specific classifier.

Weight sharing Previous adversarial adaptation methods learn a single, symmetric transformation by sharing weights between the source and target networks in order to map images from either domain into a common feature space. Learning a symmetric transformation reduces the number of parameters in the model. However, it may make the optimization more poorly conditioned, since the same network must handle images from two separate domains. Rozantsev et al. (2016) showed that untied but related weights can lead to effective adaptation in both supervised and unsupervised settings. As a result, some recent methods have favored untying weights (fully or partially) between the two domains, allowing models to learn parameters for each domain individually.

Adversarial loss Finally, these adaptation methods employ different adversarial loss functions for their various use cases. The gradient reversal layer of Ganin et al. (2016) optimizes the mapping to maximize the discriminator loss directly:

$$\min_{M_S,M_T}\max_{D} V(D,M_S,M_T) = \mathbb{E}_{\mathbf{x}\sim p_S(\mathbf{x})}[\log D(M_S(\mathbf{x}))] + \mathbb{E}_{\mathbf{x}\sim p_T(\mathbf{x})}[\log(1-D(M_T(\mathbf{x})))] \quad (1)$$

where $p_S(\mathbf{x})$ and $p_T(\mathbf{x})$ represent the source and target distributions, respectively, M_S and M_T represent our source and target mappings, which may or may not be identical, and D represents the discriminator. This optimization corresponds to the true minimax objective for generative adversarial networks. However, this objective can be problematic, since early on during training the discriminator converges quickly, causing the gradient to vanish.

When training GANs, rather than directly using the minimax loss, it is typical to train the generator with the standard loss function with inverted labels (Goodfellow et al., 2014). This splits the optimization into two independent objectives, one for the generator and one for the discriminator. The parallel objectives for adversarial adaptation are thus:

$$\max_{D} \mathbb{E}_{\mathbf{x}\sim p_S(\mathbf{x})}[\log D(M_S(\mathbf{x}))] + \mathbb{E}_{\mathbf{x}\sim p_T(\mathbf{x})}[\log(1 - D(M_T(\mathbf{x})))] \tag{2}$$

$$\max_{M_T} \mathbb{E}_{\mathbf{x}\sim p_T(\mathbf{x})}[\log D(M_T(\mathbf{x}))]. \tag{3}$$

This objective has the same fixed-point properties as the minimax loss but provides stronger gradients to the target mapping. We refer to this modified loss function as the "GAN loss function" for the remainder of this paper. Note that, in this setting, we use independent mappings for source and target, denoted as M_S and M_T, and learn only M_T adversarially. This mimics the GAN setting, where the real image distribution remains fixed, and the generating distribution is learned to match it.

The GAN loss function is the standard choice in the setting where the generator is attempting to mimic another unchanging distribution. However, in the setting where both distributions are changing, this objective will lead to oscillation—when the mapping converges to its optimum, the discriminator can simply flip the sign of its prediction in response. Tzeng et al. (2015) instead proposed the domain confusion objective, under which the mapping is trained using a cross-entropy loss function against a uniform distribution, replacing Equation 3 with:

$$\max_{M_S, M_T} \sum_{d\in\{S,T\}} \mathbb{E}_{\mathbf{x}\sim p_d(\mathbf{x})}\left[\frac{1}{2}\log D(M_d(\mathbf{x})) + \frac{1}{2}\log(1 - D(M_d(\mathbf{x})))\right]. \tag{4}$$

3 Adversarial discriminative domain adaptation

The benefit of our unified view for domain adversarial methods is that it directly enables the development of novel adaptive methods. In fact, designing a new method has now been simplified to the space of making three design choices: whether to use a generative or discriminative base model, whether to tie or untie the weights, and which adversarial learning objective to use. In light of this view we can summarize our method, adversarial discriminative domain adaptation (ADDA), as well as its connection to prior work, according to our choices (see Table 1 "ADDA"). Specifically, we use a discriminative base model, unshared weights, and the standard GAN loss.

Table 2: Unsupervised adaptation performance on the Office dataset in the fully-transductive setting. ADDA achieves state-of-the-art results on all three evaluated domain shifts and demonstrates the largest improvement on the hardest shift, $A \to W$.

Method	$A \to W$	$D \to W$	$W \to D$
DDC (Tzeng et al., 2014)	0.618	0.950	0.985
DAN (Long & Wang, 2015)	0.685	0.960	0.990
DRCN (Ghifary et al., 2016)	0.687	0.964	0.990
DANN (Ganin et al., 2016)	0.730	0.964	0.992
ADDA (Ours)	**0.751**	**0.970**	**0.996**

First, we choose a discriminative base model, as we hypothesize that much of the parameters required to generate convincing in-domain samples are irrelevant for discriminative adaptation tasks. Next, we choose to allow independent source and target mappings by untying the weights. We use the pre-trained source model as an intitialization for the target representation space and fix the source model during adversarial training. In doing so, we are learn an asymmetric mapping, in which we modify the target model so as to match the source distribution. This is most similar to the original generative adversarial learning setting, where a generated space is updated until it is indistinguishable with a fixed real space. Therefore, we opt to use the inverted label GAN loss.

We note that the unified framework presented in the previous section has enabled us to compare prior domain adversarial methods and make informed decisions about the different factors of variation. Through this framework we are able to motivate a novel domain adaptation method, ADDA, and offer insight into our design decisions.

We evaluate ADDA on the standard *Office* dataset for domain adaptation, which consists of images from 3 domains: Amazon (A), DSLR (D), and Webcam (W) (Saenko et al., 2010). We use the fully transductive setting and evaluate across 3 domain shifts commonly used for evaluation. The results of this experiment are presented in Table 2. ADDA achieves state of the art on all 3 domain shifts, achieving the largest improvement on the hardest shift, $A \to W$, indicating the effectiveness of our method. For additional results and analysis of our method in other adaptation settings, we refer the reader to the long-form version of this report (Tzeng et al., 2017).

REFERENCES

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016. URL `http://arxiv.org/abs/1605.09782`.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision (ECCV)*, pp. 597–613. Springer, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*. 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *CoRR*, abs/1606.07536, 2016. URL `http://arxiv.org/abs/1606.07536`.

Mingsheng Long and Jianmin Wang. Learning transferable features with deep adaptation networks. *International Conference on Machine Learning (ICML)*, 2015.

Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *CoRR*, abs/1603.06432, 2016. URL `http://arxiv.org/abs/1603.06432`.

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

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. URL `http://arxiv.org/abs/1412.3474`.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *International Conference in Computer Vision (ICCV)*, 2015.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *CoRR*, 2017.

Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S. Paek, and In-So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision (ECCV)*, 2016.