ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

M.Sc. in Data Science

Deep Learning

Project - Domain Adaptation

Phevos A. Margonis - f3352317

Instructor: Themos Stafylakis

June 20, 2024

Abstract

This paper investigates the effectiveness of Domain-Adversarial Neural Networks (DANN) for unsupervised domain adaptation, comparing it against baseline models. DANN utilizes adversarial training to align feature distributions between source and target domains, enhancing model generalization. Results indicate DANN's superior performance, though the training duration significantly impacts outcomes, with extended training risking overfitting. Attempts to replicate the CORAL method revealed mixed results, underscoring the complexities of domain adaptation.

Contents

1 Introduction **2**

2 Methodology **2**

3 Results **2**

4 Discussion **3**

5 Conclusions **4**

6 Bibliography **4**

1 Introduction

In the era of data-driven decision making, machine learning models have become pivotal in extracting valuable insights from vast amounts of data. Traditionally, these models assume that both training and testing data are drawn from the same distribution. This assumption, however, often fails in real-world scenarios where the distribution of data can vary due to different acquisition conditions, temporal changes, or domain-specific peculiarities. Such discrepancies can significantly degrade the performance of conventional machine learning models when applied to new, unseen domains.

Domain adaptation, a sub-field of transfer learning [1], addresses these challenges by adapting a model trained on a source domain to perform well on a different, but related, target domain without requiring extensive labeled data in the target domain. Among the various strategies for domain adaptation, Domain-Adversarial Neural Networks (DANN) have shown promise. These networks employ adversarial training to minimize the feature distribution differences between source and target domains, thereby enhancing the model's generalization capabilities across domains.

This paper presents a study of DANN for unsupervised domain adaptation, benchmarked against two baseline models to highlight its efficacy. By integrating adversarial learning into the training process, this approach aims to not only align the domain distributions but also preserve the task-specific features crucial for the learning task at hand. Through experiments and evaluations, this research underscores the potential of adversarial methods in overcoming domain discrepancies, thereby paving the way for robust, domain-invariant machine learning models.

The subsequent sections will detail the methodologies employed, the experimental setup, and the comparative analysis of the proposed DANN model against traditional and contemporary domain adaptation techniques.

2 Methodology

We now detail the proposed model for the domain adaptation. The architecture of our DANN model is based on the principle of integrating domain adversarial training into a standard neural network framework. For the first experiment, involving adaptation from the Street View House Numbers (SVHN) dataset to the Modified National Institute of Standards and Technology (MNIST) dataset, we employ a modified LeNet architecture. This architecture is augmented with dropout and batch normalization layers to stabilize training and improve model generalization, as proposed by Srivastava et al. [4]. The LeNet serves as the feature extractor, while a domain classifier, attached via a gradient reversal layer, trains to differentiate between the source and target domains. This gradient reversal layer acts to confuse the domain classifier by multiplying the gradient by a negative scalar during backpropagation, effectively encouraging the feature extractor to generate domain-invariant features.

The training process for the DANN model follows the methodology outlined by Ganin et al. [2], where both the label predictor and domain classifier are trained simultaneously in an adversarial manner. For the SVHN to MNIST adaptation, the Adam optimizer is utilized due to its efficiency in handling sparse gradients and adaptive learning rate capabilities. In contrast, the second experiment with the Office-31 [3] dataset adaptation from Amazon to Webcam uses a pretrained ResNet-50, modified to suit domain adaptation tasks. This larger model necessitated the use of the SGD optimizer with momentum to effectively handle the complexity and size of the network, especially given the higher-dimensional feature space of the Office31 dataset compared to MNIST.

Baseline comparisons are crucial for evaluating the performance enhancement provided by the DANN model. In both experiments, two baselines are established. The first baseline, *No Adaptations (NA)*, involves training the LeNet model solely on the source dataset and testing it on the target dataset without any domain adaptation techniques. This baseline helps to illustrate the natural degradation in performance due to domain shift. The second baseline trains and tests the LeNet model on the target dataset, serving as a control to show the potential performance in the absence of domain shift. These baselines are selected to highlight the effectiveness of the DANN approach by providing contrasts both with and without the influence of domain adaptation techniques.

3 Results

The datasets used in this study are crucial for evaluating different challenges and scenarios in domain adaptation and image recognition. The SVHN dataset is composed of 600,000 RGB images, each 32×32 pixels, derived from real-world imagery of house numbers. This dataset is particularly challenging due to the variability in lighting, font, and background, making it a robust benchmark for digit classification models. The MNIST dataset, another staple in digit classification tasks, comprises 70,000 images of handwritten digits that have been size-normalized and centered, providing a more controlled environment compared to SVHN.

The Office-31 dataset is employed to assess the model's performance across more diverse domain shifts, featuring images from three distinct domains: Amazon (product images from online retailers), DSLR (high-resolution images taken with a DSLR camera), and Webcam (low-resolution images with significant noise and artifacts). This dataset includes 31 categories common in office environments, with varying numbers of images per category across the domains, offering a complex set of conditions for testing domain adaptation methods.

The results (see Table 1) underscore the utility of domain-adaptive training. In the SVHN to MNIST adaptation, the *No Adaptation* baseline achieves 70% accuracy, whereas the DANN model significantly improves performance, reaching 78% accuracy on the MNIST test set. This demonstrates the DANN's effectiveness in leveraging labeled data from both domains to generalize better on the target domain. In contrast, training solely on the target yields the highest accuracy of 99%, indicating optimal performance when the model is trained and tested within the same domain. Similarly, for the Amazon to Webcam adaptation, the *No Adaptation* strategy reaches 74% accuracy. The application of the DANN model boosts this to 80%, once again illustrating its advantage for domain adaptation, though direct training on the target remains the most effective method at 99% accuracy. This comparative analysis highlights the effectiveness of DANN in bridging the domain gap when direct training on the target is not feasible, offering a significant improvement over models that do not adapt between domains.

4 Discussion

In this section, we delve into some of the unexpected findings and challenges encountered during the implementation and testing of the Domain-Adversarial Neural Networks (DANN) model, as well as in attempts to replicate established domain adaptation techniques.

One of the more perplexing outcomes arose while experimenting with the DANN model, particularly when fine-tuning a pretrained ResNet for more than 40 epochs. Contrary to expectations, extending the training duration adversely affected the adversarial component of the DANN model. Specifically, the domain classifier began to lose its ability to discriminate between the source and target classes effectively. This anomaly resulted in a significant increase in the total loss, suggesting a breakdown in the adversarial training process (see Figure 1). A potential explanation for this phenomenon could be that the extended training allowed the feature extractor to overfit to the source domain, thereby embedding domain-specific cues into the features that the domain classifier could not negate. This issue underscores the delicate balance required in tuning the adversarial training phase, where too little training might not align the domains sufficiently, and too much might lead to destabilization of the model's ability to generalize.

Additionally, our attempt to replicate the Correlation Alignment (CORAL) (see algorithm 1) method by Sun et al. [5], yielded mixed results. Utilizing a pretrained AlexNet to extract features and applying a linear SVM for classification (referred to as the NA-fc7 baseline in the original study) replicated the findings from the paper accurately. However, efforts to replicate the *CORAL-fc7* experiment, which involved adapting the distribution of the source features to match the target by aligning their second-order statistics (see. Figure 2), did not result in the anticipated improvement. In fact, the performance was inferior to the NA baseline. This discrepancy could be attributed to several factors, including potential differences in feature normalization or SVM parameter settings not detailed in the original study, or variability in the underlying feature distribution caused by differences in the initial weights of the pretrained network.

These observations highlight the complexities and challenges inherent in domain adaptation research. They stress the importance of parameter tuning, model selection, and the need for robust methodologies that can adapt to a wide range of scenarios. Further research is required to better understand the dynamics of extended adversarial training and to refine the implementation of statistical alignment methods like CORAL, ensuring they are robust across different applications and datasets.

Algorithm 1: Correlation Alignment (CORAL)

```
Output: Aligned source feature matrix $\mathbf{X'_s}$
Compute the covariance matrices:
    $\mathbf{C_s} = \mathrm{cov}(\mathbf{X_s})$, $\mathbf{C_t} = \mathrm{cov}(\mathbf{X_t})$;
Regularize the covariance matrices:
    $\mathbf{C'_s} = \mathbf{C_s} + \lambda\mathbf{I}$, $\mathbf{C'_t} = \mathbf{C_t} + \lambda\mathbf{I}$;
Compute the square root inverse of $\mathbf{C'_s}$:
    $\mathbf{C_s}^{-\frac{1}{2}} = \mathbf{C'_s}^{-\frac{1}{2}}$;
Compute the square root of $\mathbf{C'_t}$: $\mathbf{C_t}^{\frac{1}{2}} = \mathbf{C'_t}^{\frac{1}{2}}$;
Remove complex components if any:
    $\mathbf{C_s}^{-\frac{1}{2}} = \Re(\mathbf{C_s}^{-\frac{1}{2}})$, $\mathbf{C_t}^{\frac{1}{2}} = \Re(\mathbf{C_t}^{\frac{1}{2}})$;
Align the source features: $\mathbf{X'_s} = \mathbf{X_s}\mathbf{C_s}^{-\frac{1}{2}}\mathbf{C_t}^{\frac{1}{2}}$;
return $\mathbf{X'_s}$
```

5 Conclusions

In conclusion, our study underscores the efficacy of Domain-Adversarial Neural Networks (DANN) in addressing the challenges of unsupervised domain adaptation. Through comparative analysis with baseline models, DANN demonstrated superior performance by effectively minimizing feature distribution differences between source and target domains while preserving task-specific features. However, the delicate balance in adversarial training duration emerged as a critical factor, with extended training potentially leading to overfitting on the source domain. Additionally, attempts to replicate the CORAL method highlighted the complexities in achieving consistent improvements across different experimental setups. These findings emphasize the need for meticulous parameter tuning and robust methodologies in domain adaptation to ensure generalizability and effectiveness across diverse real-world scenarios. Further research is recommended to refine these approaches, particularly in understanding the nuances of adversarial training and statistical alignment methods.

6 Bibliography

[1] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation, 2020.

[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks, 2016.

[3] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. volume 6314, pages 213–226, 09 2010. ISBN 978-3-642-15560-4. doi: 10.1007/978-3-642-15561-1_16.

[4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014.

[5] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation, 2015.

	Accuracy	
Method	**SVHN**	**Amazon**
	MNIST	**Webcam**
No Adaptation	70%	74%
Paper Implementation	74%	73%
Our DANN	**78%**	**80%**
Train on Target	99%	99%

Table 1: Results for SVHN to MNIST and Amazon to Webcam, for the problem of domain adaptation using our implementation of the DANN algorithm

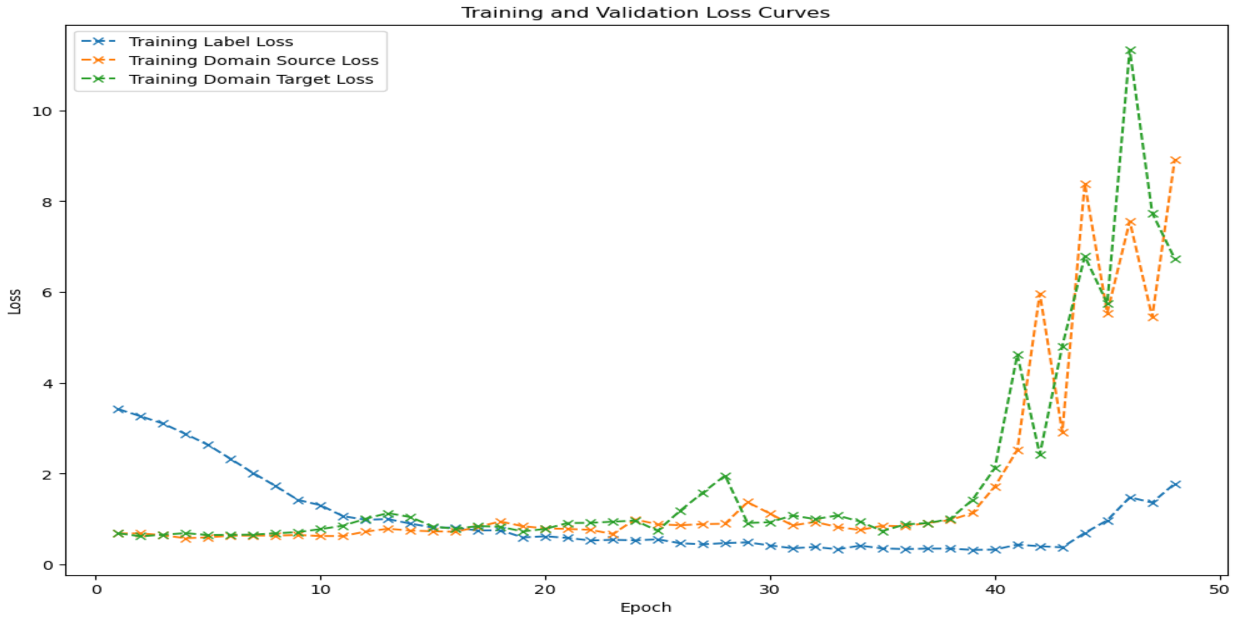

Figure 1: Classification loss for the label and the two domain classifiers as a function of training epoch, demonstrates how the adversarial component of the model is affected by the extended training duration.

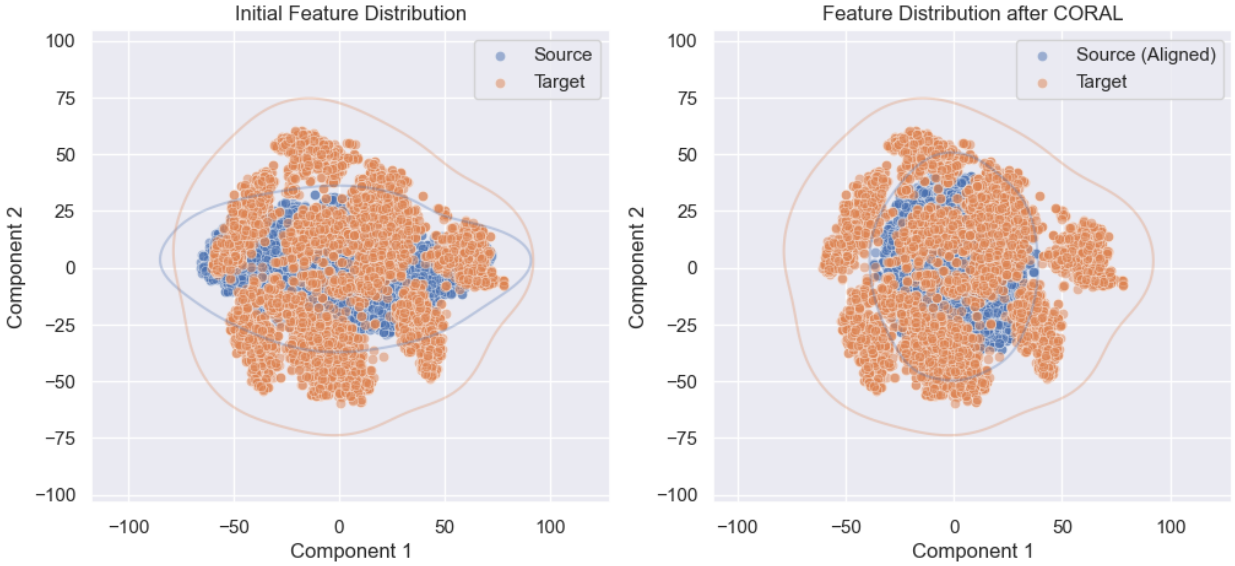

Figure 2: (**Left**) The source (SVHN) and target (MNIST) domains originally have different distribution covariances, even though the features are normalized to zero mean and unit standard deviation. This discrepancy poses a challenge for transferring classifiers trained on the source domain to the target domain. (**Right**) By first decorrelating the source domain, which involves removing its feature correlations, and then re-correlating it by adding the target domain's correlations, the source and target distributions become well aligned. However, despite these adjustments, the classifier trained on the modified source domain did not perform well in the target domain.