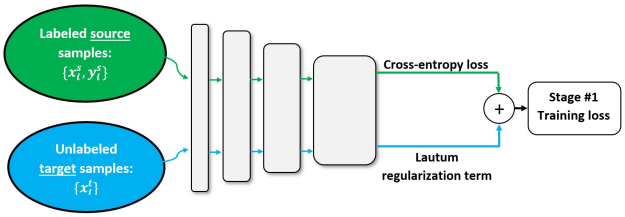
Lautum Regularization for Semi-Supervised Transfer Learning

Daniel Jakubovitz, Raja Giryes
School of Electrical Engineering
Tel Aviv University
Tel Aviv, Israel
danielshaij@mail.tau.ac.il
raja@tauex.tau.ac.il

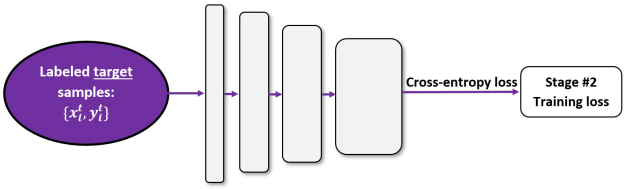
Miguel R. D. Rodrigues
Department of Electronic and Electrical Engineering
University College London
London, United Kingdom
m.rodrigues@ucl.ac.uk

Abstract

Transfer learning is a very important tool in deep learning as it allows propagating information from one "source dataset" to another "target dataset", especially in the case of a small number of training examples in the latter. Yet, discrepancies between the underlying distributions of the source and target data are commonplace and are known to have a substantial impact on algorithm performance. In this work we suggest a novel information theoretic approach for the analysis of the performance of deep neural networks in the context of transfer learning. We focus on the task of semi-supervised transfer learning, in which unlabeled samples from the target dataset are available during the network training on the source dataset. Our theory suggests that one may improve the transferability of a deep neural network by imposing a Lautum information based regularization that relates the network weights to the target data. We demonstrate the effectiveness of the proposed approach in various transfer learning experiments.



(a) Pre-transfer training stage.



(b) Post-transfer training stage.

Figure 1: Our semi-supervised transfer learning technique applying Lautum regularization. Omitting the blue part in the first training stage (top) gives standard transfer learning.

1. Introduction

Machine learning algorithms have lately come to the forefront of technological advancements, providing state-of-the-art results in a variety of fields [3]. However, alongside their incredible performance, these methods suffer from sensitivity to data discrepancies - any inherent difference between the training data and the test data may result in a substantial decrease in performance. Moreover, to obtain good performance a large amount of labeled data is necessary for their training. Such a substantial amount of labeled data is often either very expensive or simply unobtainable.

One popular approach to mitigate this issue is using "transfer learning", where training on a small labeled "target" dataset is improved by using information from another large labeled "source" dataset of a different problem. A common method for transfer learning uses the result of training on the source as initialization for training on the target, thereby improving the performance on the latter [2].

Transfer learning has been the focus of much research attention along the years. Plenty of different approaches have been proposed to encourage a more effective transfer from a source dataset to a target dataset, many of them aim at obtaining better system robustness to environment changes, so as to allow an algorithm to perform well even under some variations in the settings (e.g. changes in lighting conditions in computer vision tasks). Sometimes this is achieved at the expense of diminishing the performance on the original task or data distribution. Other works take a more targeted approach and directly try to reduce algorithms' generalization error by decreasing the difference in their performance on specific source and target datasets [6].

In addition, it is often the case that the target dataset has

a large number of samples, though only a few of those samples are labeled. In this scenario a semi-supervised learning approach could prove to be beneficial by making good use of the available unlabeled samples for training.

In this work we focus on the task of semi-supervised transfer learning. The problem we address is related to the field of domain adaptation, however we make a distinction between domain adaptation and transfer learning, where the former refers to the case of two sources of data with the same content (e.g. the MNIST $\rightarrow$ SVHN case) whereas the latter refers to the case of two sources of data which are completely different in both content and "styling". Another relevant difference is that labeled data from the target distribution is typically available in the transfer learning case, yet less so in the domain adaptation case.

Plenty of works exist in the literature on transfer learning, semi-supervised learning and using information theory for the analysis of machine learning algorithms. The closest work to ours is [1] in which an information theoretic approach is used in order to decompose the cross-entropy *train* loss of a machine learning algorithm into several separate terms. However, unlike this work we propose a different decomposition of the cross-entropy *test* loss and make the relation to semi-supervised transfer learning.

Contribution. We consider the case of semi-supervised transfer learning in which plenty of labeled examples from a source distribution are available along with just a few labeled examples from a target distribution; yet, we are provided also with a large number of unlabeled samples from the latter. This setup combines transfer learning and semi-supervised learning, where both aim at obtaining improved performance on a target dataset with a small number of labeled examples. In this work we suggest to combine both methodologies to gain the advantage of both of them. This setting represents the case where the learned information from a large labeled source dataset is used to obtain good performance when transferring to a mostly unlabeled target set, where the unlabeled examples of the target are available at the training time on the source.

To do so, we provide a theoretical derivation that leads to a novel semi-supervised technique for transfer learning. We take an information theoretic approach to examine the cross-entropy test loss of machine learning methods. We decompose the loss to several different terms that account for different aspects of its behavior. This derivation leads to a new regularization term, which we call "Lautum regularization" as it relies on the maximization of the Lautum information [7] between unlabeled data samples drawn from the target distribution and the learned model weights. Figure 1 provides a general illustration of our approach.

We corroborate the effectiveness of our approach with experiments of semi-supervised transfer learning for neural networks on image classification tasks. We examine the transfer in two cases: from the MNIST dataset to the notMNIST dataset (which consists of the letters A-J in grayscale images) and from the CIFAR-10 dataset to 10 specific classes of the CIFAR-100 dataset. We compare our results to three other methods: (1) Temporal Ensembling [5], a state-of-the-art method for semi-supervised training which we apply in a transfer learning setup; (2) the Multi-kernel Maximum Mean Discrepancy (Mk-MMD) method [4], which is popular in semi-supervised transfer learning; (3) standard transfer learning which does not use any of the unlabeled samples. The advantage of our method is demonstrated in our experimental results as it outperforms the other compared methods.

The appendices to this paper are in the supplementary material.

2. The cross-entropy loss - an information theory perspective

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a training set with N training samples that is used to train a learning algorithm with a set of weights w. We assume that given $\mathcal{D}$ (a parameter of the model), the learning algorithm selects a specific hypothesis from the hypothesis class according to the distribution $p(w_\mathcal{D})$. In the case of a neural network, selecting the hypothesis is equivalent to training the network on the data.

We denote by $w_\mathcal{D}$ the model weights which were learned using the training set $\mathcal{D}$, and by $f(y|x, w_\mathcal{D})$ the learned classification function which given the weights $w_\mathcal{D}$ and a D-dimensional input $x \in \mathbb{R}^D$ computes the probability of the K-dimensional label $y \in \mathbb{R}^K$. The learned classification function is tested on data drawn from the true underlying distribution $p(x, y)$. Ideally, the learned classification function $f(y|x, w_\mathcal{D})$ would highly resemble the ground-truth classification $p(y|x)$, and similarly $f(x, y|w_\mathcal{D})$ would highly resemble $p(x, y)$. With these notations, we turn to analyze the cross-entropy loss used predominantly in classification tasks. In our derivations we used several information theoretic measures which we present in Appendix A.

Main theoretical result. Our main theoretical result is given by the following theorem:

Theorem 1 *For a classification task with ground-truth distribution $p(y|x)$, training set $\mathcal{D}$, learned weights $w_\mathcal{D}$ and learned classification function $f(y|x, w_\mathcal{D})$, the expected cross-entropy loss of a machine learning algorithm on the test distribution is equal to*

$$\mathbb{E}_{w_\mathcal{D}} \left\{ KL(p(x, y)||f(x, y|w_\mathcal{D})) \right\} + H(y|x) - L(w_\mathcal{D}; x). \tag{1}$$

Note that KL signifies the Kullback-Leibler divergence and that we treat the training set $\mathcal{D}$ as a fixed parameter, whereas

$w_{\mathcal{D}}$ and the examined test data (x, y) are treated as random variables. We refer the reader to Appendix B for the proof of Theorem 1.

In accordance with Theorem 1, the three terms that compose the expected cross-entropy test loss represent three different aspects of the loss of a learning algorithm performing a classification task:

- **Classifier mismatch** $\mathbb{E}_{\mathbf{w}_{\mathcal{D}}}\mathbf{KL}\left(\mathbf{p}(\mathbf{x},\mathbf{y})||\mathbf{f}(\mathbf{x},\mathbf{y}|\mathbf{w}_{\mathcal{D}})\right)$: measures the deviation of the learned classification function's data distribution $f(x, y|w_{\mathcal{D}})$ from the true distribution of the data $p(x, y)$. It is measured by the KL-divergence, which is averaged over all possible instances of w parameterized by the training set $\mathcal{D}$. This term essentially measures the ability of the weights learned from $\mathcal{D}$ to represent the true distribution of the data.

- **Intrinsic Bayes error** $\mathbf{H}(\mathbf{y}|\mathbf{x})$: represents the inherent uncertainty of the labels given the data samples.

- **Lautum information** between $w_{\mathcal{D}}$ and x, $\mathbf{L}(\mathbf{w}_{\mathcal{D}};\mathbf{x}) = \mathbb{E}_{\mathbf{w}_{\mathcal{D}}}\{\mathbf{KL}(\mathbf{p}(\mathbf{x})||\mathbf{p}(\mathbf{x}|\mathbf{w}_{\mathcal{D}}))\}$: represents the dependence between $w_{\mathcal{D}}$ and x. It essentially measures how much $p(x|w_{\mathcal{D}})$ deviates from $p(x)$ on average over the possible values of $w_{\mathcal{D}}$.

Our formulation suggests that a machine learning algorithm, which is trained relying on empirical risk minimization, implicitly aims at maximizing the Lautum information $L(w_{\mathcal{D}}; x)$ in order to minimize the cross-entropy loss. At the same time, the algorithm aspires to minimize the KL-divergence between the ground-truth distribution of the data and the learned classification function. The intrinsic Bayes error cannot be minimized and remains the inherent uncertainty of the task. Namely, the formulation in (1) suggests that encouraging a larger Lautum information between the data samples and the learned model weights would be beneficial for reducing the model's test error on unseen data drawn from $p(x, y)$.

3. Lautum information based semi-supervised transfer learning

We turn to show how we may apply our theory on the task of semi-supervised transfer learning. In standard transfer learning, which consists of pre-transfer and post-transfer stages, a neural network is trained on a labeled source dataset and then fine-tuned on a smaller labeled target dataset. In semi-supervised transfer learning, which we study here, we assume that an additional large set of unlabeled examples from the target distribution is available during training on the source data.

Semi-supervised transfer learning is highly beneficial in scenarios where the available target dataset is only partially annotated. Using the unlabeled part of this dataset, which is usually substantially bigger than the labeled part, has the potential of considerably improving the obtained performance. Thus, if this unlabeled part is a-priori available, then using it from the beginning of training can potentially improve the results. For using the unlabeled samples of the target dataset during the pre-transfer training on the source dataset, we leverage the formulation in (1). Considering its three terms, it is clear that by using unlabeled samples the classifier mismatch term cannot be minimized due to the lack of labels; the intrinsic Bayes error is a characteristic of the task and cannot be minimized either; yet, the Lautum information does not depend on the labels and can therefore be calculated and maximized.

When the Lautum information is calculated between the model weights and data samples drawn from the target distribution, its maximization would encourage the learned weights to better relate to these samples, and by extension to better relate to the underlying probability distribution from which they were drawn. Therefore, it is expected that an enlarged Lautum information will yield an improved performance on the target test set. Accordingly, we aim at maximizing $L(w_{\mathcal{D}}; x)$ during training. The pre-transfer maximization of the term $L(w_{\mathcal{D}}; x)$, which is computed with samples drawn from the target distribution, would make the learned weights more inclined towards good performance on the target set right from the beginning. At the same time, the cross-entropy loss at this stage is calculated using labeled samples from the source dataset. In the post-transfer stage, the cross-entropy loss is calculated using labeled samples from the target dataset, and therefore implicitly maximizes $L(w_{\mathcal{D}}; x)$ by itself. We have empirically observed that explicitly maximizing the Lautum information between the unlabeled target samples and the model weights during post-transfer training (by imposing Lautum regularization) in addition to (or instead of) during pre-transfer training does not lead to improved results.

To summarize, our semi-supervised transfer learning approach optimizes two goals at the same time: (i) minimizing the classifier mismatch $\mathbb{E}_{w_{\mathcal{D}}}\left\{KL\left(p(x, y)||f(x, y|w_{\mathcal{D}})\right)\right\}$, which is achieved using the labeled data both for the source and the target datasets during pre-transfer and post-transfer training respectively; and (ii) maximizing the Lautum information $L(w_{\mathcal{D}}; x)$, which is achieved explicitly using the unlabeled target data during pre-transfer training by imposing Lautum regularization, and in the post-transfer stage implicitly through the minimization of the cross-entropy loss which is evaluated on the labeled target data. Figure 1 summarizes our training scheme.

3.1. Training with Lautum regularization

We refer the reader to Appendix C for details regarding the estimation of the Lautum information. Once the Lautum

information has been estimated, our loss function for pre-transfer training is:

$$Loss = \sum_{i=1}^{N}\sum_{k=1}^{K} -y_{ik}^{s} \log f_k(x_i^s|w_{\mathcal{D}}) - \lambda L(w_{\mathcal{D}}; x^t). \quad (2)$$

Note that the the cross-entropy loss is calculated using labeled samples from the source training set (which we denote by the s superscript) whereas the Lautum regularization term is calculated using unlabeled samples from the target training set (which we denote by the t superscript). Also note that y_i represents the ground truth label of the sample x_i; $f(x_i|w_{\mathcal{D}})$ represents the network's estimated post softmax label for that sample; and $L(w_{\mathcal{D}}; x)$ is calculated as detailed in Appendix C. We emphasize that the Lautum regularization term is subtracted and not added to the cross-entropy loss since we aim at *maximizing* the Lautum information during training. Our loss function for post-transfer training consists of a standard cross-entropy loss:

$$Loss = \sum_{i=1}^{N}\sum_{k=1}^{K} -y_{ik}^{t} \log f_k(x_i^t|w_{\mathcal{D}}). \quad (3)$$

Note that at this stage the cross-entropy loss, which is calculated using labeled target samples, inherently includes the Lautum term of the target data (see Theorem 1).

4. Experiments

In order to demonstrate the advantages of semi-supervised transfer learning with Lautum regularization we perform several experiments on image classification tasks using deep neural networks (though our theoretical derivations also apply to other machine learning algorithms). We train deep neural networks and perform transfer learning from the original source dataset to the target dataset. In our experiments we use the original labeled source training set as is and split the target training set into two parts. The first part is very small and contains labeled samples, whereas the second part consists of the remainder of the target training set and contains unlabeled samples only (the labels are discarded). The performance is evaluated by the post transfer accuracy on the target test set.

We examine four different methods of transfer learning: (1) standard supervised transfer which uses the labeled samples only. (2) Temporal Ensembling semi-supervised learning as outlined in [5], applied in a transfer learning setting. Temporal Ensembling is applied in the post-transfer training stage. (3) Mk-MMD [4], which is based on 19 different Gaussian kernels with different standard deviations. Mk-MMD is applied in the pre-transfer training stage. (4) Lautum regularization - our technique as described in Section 3.

We refer the reader to Appendix D for more details about the experimental setup. Using the settings outlined in Appendix D.1 we obtained the results shown in Table 1 for the MNIST → notMNIST case, and using the settings outlined in Appendix D.2 we obtained the results shown in Table 2 for the CIFAR-10 → CIFAR-100 (10 classes) case.

The advantage of using Lautum regularization is evident from the results, as it outperforms the other compared methods in all the examined target training set splits. In general, the Temporal Ensembling method by itself does not yield very competitive results compared to standard transfer learning.

Method	Source → Target	# labeled	Accuracy
Standard	MNIST / notMNIST	50	34.02%
TE	MNIST / notMNIST	50	37.28%
Mk-MMD	MNIST / notMNIST	50	46.72%
Lautum	MNIST / notMNIST	50	**47.96%**
Standard	MNIST / notMNIST	100	57.58%
TE	MNIST / notMNIST	100	61.45%
Mk-MMD	MNIST / notMNIST	100	63.32%
Lautum	MNIST / notMNIST	100	**65.21%**
Standard	MNIST / notMNIST	200	67.78%
TE	MNIST / notMNIST	200	74.87%
Mk-MMD	MNIST / notMNIST	200	80.35%
Lautum	MNIST / notMNIST	200	**83.77%**

Table 1: target test set accuracy comparison between standard transfer learning, Temporal Ensembling (TE), Mk-MMD and Lautum regularization for different amounts of labeled training target samples, MNIST → notMNIST.

Method	Source → Target	# labeled	Accuracy
Standard	CIFAR-10 / 100	100	39.90%
TE	CIFAR-10 / 100	100	42.20%
Mk-MMD	CIFAR-10 / 100	100	45.30%
Lautum	CIFAR-10 / 100	100	**46.70%**
Standard	CIFAR-10 / 100	200	52.80%
TE	CIFAR-10 / 100	200	54.60%
Mk-MMD	CIFAR-10 / 100	200	59.30%
Lautum	CIFAR-10 / 100	200	**60.90%**
Standard	CIFAR-10 / 100	500	64.50%
TE	CIFAR-10 / 100	500	66.50%
Mk-MMD	CIFAR-10 / 100	500	68.00%
Lautum	CIFAR-10 / 100	500	**70.80%**

Table 2: target test set accuracy comparison between standard transfer learning, Temporal Ensembling (TE), Mk-MMD and Lautum regularization for different amounts of labeled training target samples, CIFAR-10 → CIFAR-100 (10 classes).

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19, 2018.

[2] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 647–655, Bejing, China, 22–24 Jun 2014. PMLR.

[3] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[4] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1205–1213. Curran Associates, Inc., 2012.

[5] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

[6] Xuhong LI, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2825–2834, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[7] D. P. Palomar and S. Verdu. Lautum information. *IEEE Trans. Inform. Theory*, 54(3):964–975, March 2008.