Lautum Regularization for Semi-Supervised Transfer Learning

Daniel Jakubovitz and Raja Giryes, School of EE, Tel Aviv University, Israel
Miguel R. D. Rodrigues, Department of EE, University College London, UK

Outline

- In Semi-Supervised Transfer Learning the task is to make good use of the available unlabeled samples, alongside any available labeled samples, for effective transfer from a labeled source dataset to a mostly unlabeled target dataset.
- This work proposes to **improve the transferability of DNNs by imposing a Lautum based regularization that relates the network weights to the target data. Our theory relies on an information theoretic decomposition of the cross-entropy test loss [1].**
- We demonstrate the effectiveness of Lautum regularization in various Semi-Supervised Transfer Learning experiments.

The Cross-Entropy Loss

- The differential entropy of a random variable X is defined by:

$$H(X) = -\int p(x)\log p(x)\,dx$$

- The Lautum information [2] between two random variables X, Y is defined by:

$$L(X;Y) = \iint p(x)p(y)\log\left\{\frac{p(x)p(y)}{p(x,y)}\right\}dx\,dy$$

- Our theory suggests that for a classification task with ground-truth distribution $p(y|x)$, training set D, learned weights w_D and learned classification function $f(y|x,w_D)$, the expected cross-entropy loss of a machine learning algorithm on the *test* distribution is equal to

$$\mathbb{E}_{\boldsymbol{w_D}}\{\boldsymbol{KL}(\boldsymbol{p}(\boldsymbol{x},\boldsymbol{y})||\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{w_D}))\} + \boldsymbol{H}(\boldsymbol{y}|\boldsymbol{x}) - \boldsymbol{L}(\boldsymbol{w_D};\boldsymbol{x})$$

Lautum Regularization

- In a Semi-Supervised Transfer Learning setting we aim at improving the post-transfer accuracy on the *target* test set by promoting a larger Lautum information between the learned weights and the unlabeled target training samples: $\boldsymbol{L}(\boldsymbol{w_D};\boldsymbol{x^t})$.
- We use a regularization term during <u>pre-transfer</u> training, which is *subtracted* from the cross-entropy loss of the labeled *source* samples:

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

- <u>Post-transfer</u> training is done using labeled *target* samples only:

$$Loss = \sum_{i=1}^{N}\sum_{k=1}^{K} -y_{ik}^{t}\log f_k\left(x_i^t|w_D\right)$$

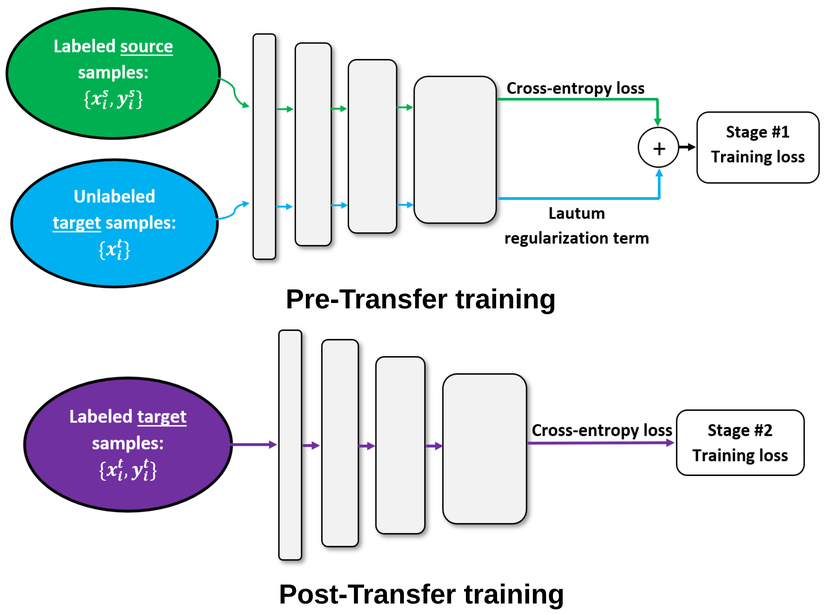
Result Interpretation

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** $\mathbb{E}_{\boldsymbol{w_D}}\{\boldsymbol{KL}(\boldsymbol{p}(\boldsymbol{x},\boldsymbol{y})||\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{w_D}))\}$: measures the deviation of the learned classification function's data distribution from the true distribution of the data.
- **Intrinsic Bayes error** $\boldsymbol{H}(\boldsymbol{y}|\boldsymbol{x})$: represents the inherent uncertainty of the labels given the data samples.
- **Lautum information between $\boldsymbol{w_D}$ and $\boldsymbol{x}$, $\boldsymbol{L}(\boldsymbol{w_D};\boldsymbol{x})$:** represents the dependence between w_D and x. It essentially measures how much $p(x|w_D)$ deviates from $p(x)$ on average over the possible values of w_D.

2. Promoting a larger Lautum information between the learned weights and the unlabeled samples is expected to reduce the cross-entropy test loss.

Training Scheme



Experimental Results

- Two Semi-Supervised Transfer Learning image classification tasks were examined: MNIST → notMNIST and CIFAR-10 → CIFAR-100 (10 classes), each for three different partitions to labeled and unlabeled *target* training sets.
- We compared the performance of Lautum regularization to standard transfer learning (not using any unlabeled samples), Temporal Ensembling [3] and Mk-MMD [4].
- Lautum regularization outperforms all the other tested methods over all partitions: it achieves a higher target test set accuracy.

Target test set accuracy comparisons

Method	Source → Target	# Labeled Target	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%**

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

References

- [1] Alessandro Achille and Stefano Soatto. Emergence of Invariance and Disentanglement in Deep Representations. Journal of Machine Learning Research, 19, 2018.
- [2] D. P. Palomar and S. Verdu. Lautum Information. IEEE Trans. Inform. Theory, 54(3):964–975, March 2008.
- [3] Samuli Laine and Timo Aila. Temporal Ensembling for Semi-Supervised Learning. In ICLR, 2017.
- [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, pages 1205–1213. Curran Associates, Inc., 2012.