Understanding the Limits of Unsupervised Domain Adaptation via Data Poisoning

Akshay Mehra[1], Bhavya Kailkhura[2], Pin-Yu Chen[3] and Jihun Hamm[1]
[1]Tulane University, [2]Lawrence Livermore National Laboratory, [3]IBM Research

Motivation

- Performance of machine learning models degrade significantly under distribution shifts i.e., when $P_{source}(x, y) \neq P_{target}(x, y)$.
- A domain invariant representation that minimizes error on the source domain may fail to reduce the error on the target domain.
- Previous works have explained this failure for scenarios such as shifts in the marginal label distributions or have empirically demonstrated these representations to increase the error of the ideal joint hypothesis.

Contributions

- To provably explain the failure of learning in the unsupervised domain adaptation (UDA) setting we propose a lower bound on the target domain error.
- Through simple examples we illustrate the success of state-of-the-art UDA methods to be dependent on the data distributions of the two domains.
- We propose mislabeled, watermarked and clean-label data poisoning attacks to gauge the robustness of UDA methods at aligning the two domains.

Necessary condition for learning in the UDA setting

Notations:

- $\mathcal{X}$ denotes the data domain and $\mathcal{D}$ denotes a distribution on this domain.
- $f: \mathcal{X} \to [0,1]$ is a deterministic labeling function, $g: \mathcal{X} \to \mathcal{Z}$ is a map from data to the representation space and $h: \mathcal{Z} \to [0,1]$ is a hypothesis in the representation space.
- $\tilde{p}(z)$ is the density function of the distribution induced by g on $\mathcal{Z}$ and $\tilde{f}(z) \coloneqq \mathbb{E}_{\mathcal{D}}[f(x)|g(x) = z]$ be the induced labeling function.
- $e(h) = \mathbb{E}_{z\sim\tilde{D}}\left[\left|\tilde{f}(z) - h(z)\right|\right]$ is the misclassification error w.r.t. the induced labeling function and $D_1(\tilde{p}, \tilde{p}') = \int_{\mathcal{Z}} |\tilde{p}(z) - \tilde{p}'(z)|\, dz$ be the total variation distance.

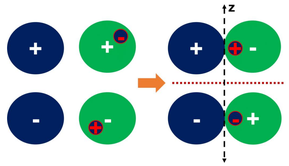
***Theorem*: (Lower bound on the target domain error in the UDA setting)**

Let $\mathcal{H}$ be the hypothesis class and $\mathcal{G}$ be the class of representation maps. Then $\forall\, h \in \mathcal{H}$ and $g \in \mathcal{G}$,

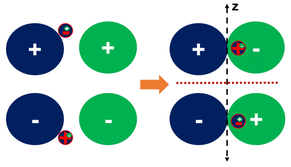
$$e_{target}(h) \geq \max\{e_{target}(\tilde{f}_{source}, \tilde{f}_{target}), e_{source}(\tilde{f}_{source}, \tilde{f}_{target})\} - (e_{source}(h) + D_1(\tilde{p}_{source}, \tilde{p}_{target})).$$

Poisoning in unsupervised domain adaptation setting

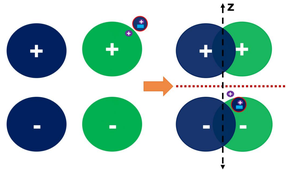
- **Poisoning using mislabeled data:**
  Poisoning with mislabeled target domain data fools UDA methods into aligning wrong classes from the two domains.



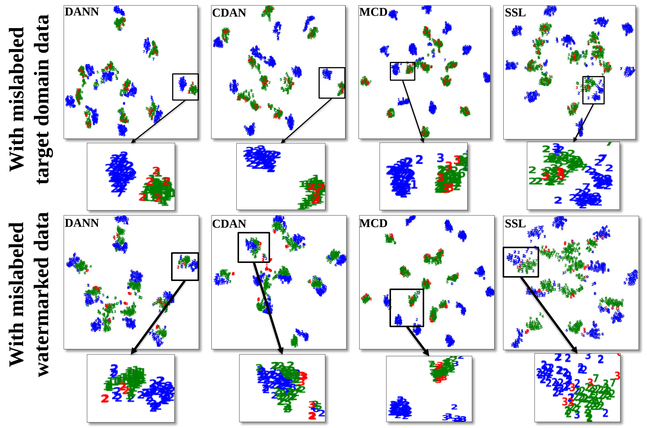
- **Poisoning using watermarked data:**
  Watermarked data looks like the data from the source domain but successfully fools UDA methods into aligning wrong classes from the two domains.



- **Poisoning using clean label data:**
  Clean label poison data are hardest to detect and can cause a target domain test point to be misclassified after domain adaptation.

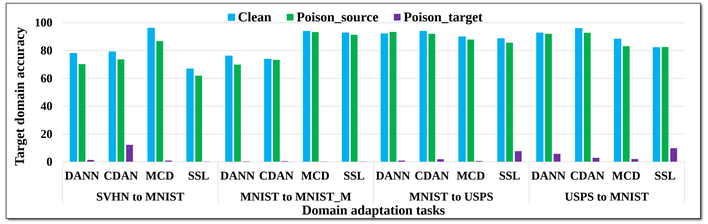


t-SNE plots of the representation learned with UDA methods

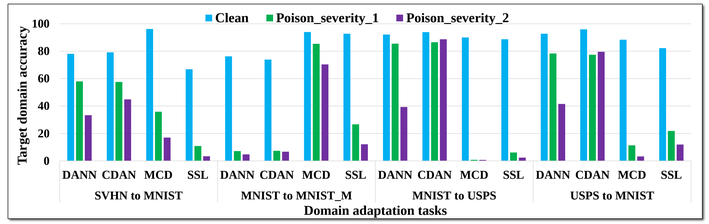


Key Results

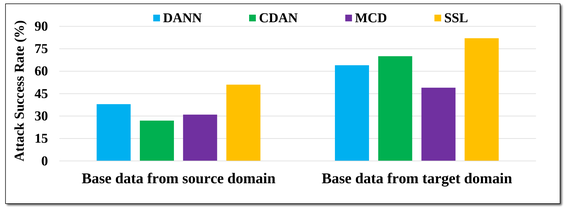
- **Poisoning using mislabeled data:**



- **Poisoning using watermarked data**



- **Poisoning using clean label data** (3 vs 8 on MNIST to MNIST_M task)



References

- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A.,Pereira, F., and Vaughan, J. W. A theory of learning from different domains.Machine learning, 79(1):151–175, 2010.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle,H., Laviolette, F., Marchand, M., and Lempitsky, V.Domain-adversarial training of neural networks.Thejournal of machine learning research, 17(1):2096–2030,2016
- Mei, Shike, and Xiaojin Zhu. "Using machine teaching to identify optimal training-set attacks on machine learners." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. No. 1. 2015.
- Zhao H., Combes R., Zhang K., and Gordon G.. On learning invariant representations for domain adaptation. In International Conference on Machine Learning, pages 7523–7532. PMLR, 2019.

This work was supported by the NSF EPSCoR-Louisiana Materials Design Alliance (LAMDA) program #OIA1946231 and by LLNL Laboratory Directed Research and Development project 20-ER-014 (LLNLCONF-824206). This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC.