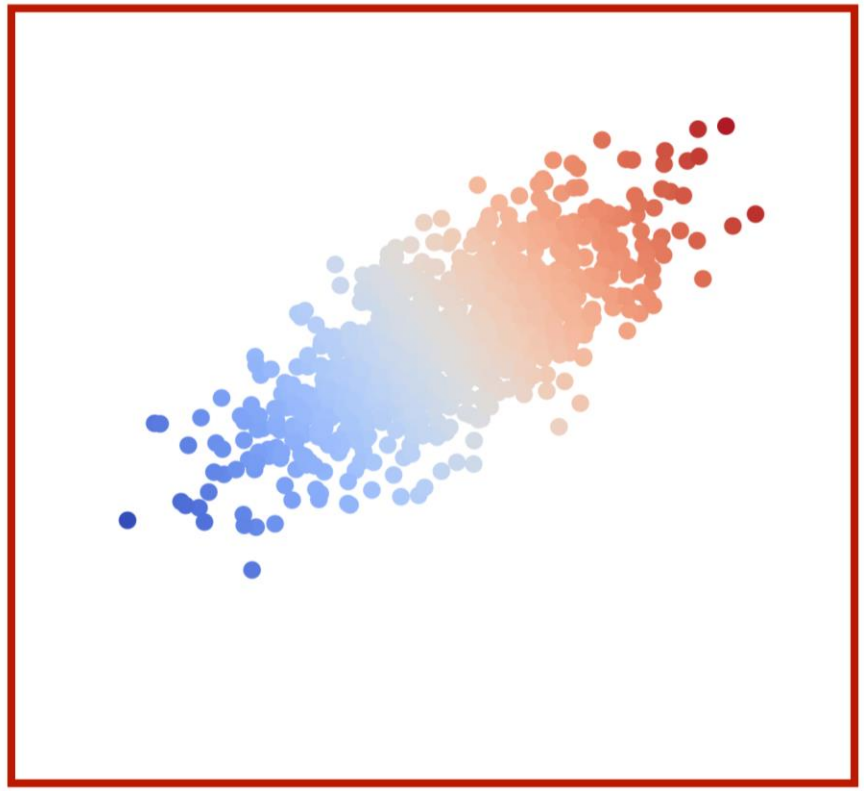
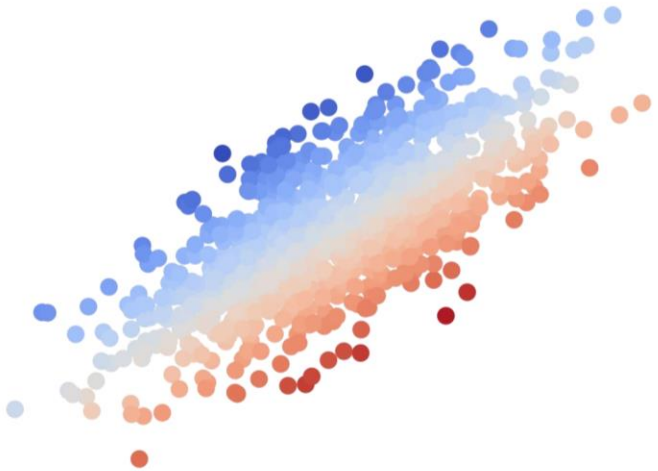
Label Alignment Regularization for Distribution Shift

Ehsan Imani, Guojun Zhang, Runjia Li, Jun Luo,
Pascal Poupart, Philip H.S. Torr, Yangchen Pan

Label Alignment Property





Problem Setup

- We have a source domain with labeled data for training a model
- We have a target domain with little to no labeled data, but we have a large amount of unlabeled data along with prior knowledge that the label alignment property holds

We design a regularizer to improve performance in this setup

The overall objective function

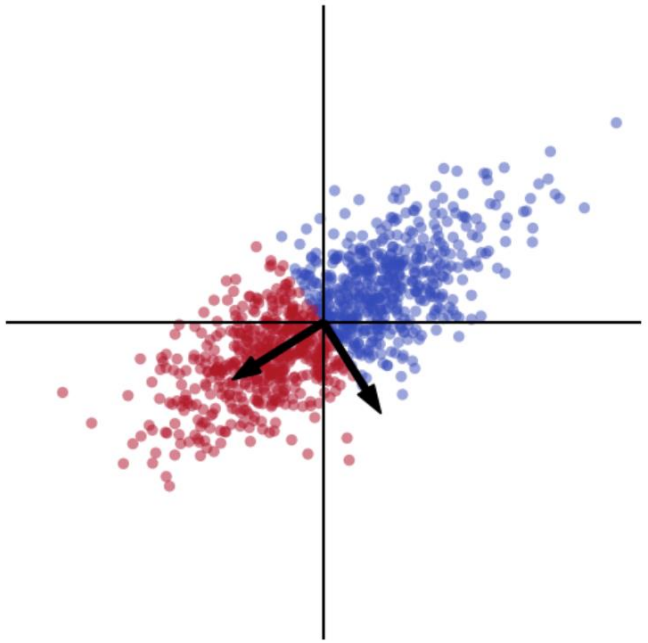
$$\min_{w} \|\Phi w - y\|^2 + \lambda \sum_{i=\tilde{k}+1}^{d} \tilde{\sigma}_i^2 (w_i^{\tilde{V}})^2 - \sum_{i=k+1}^{d} \sigma_i^2 (w_i^{V})^2$$

Training on the source domain

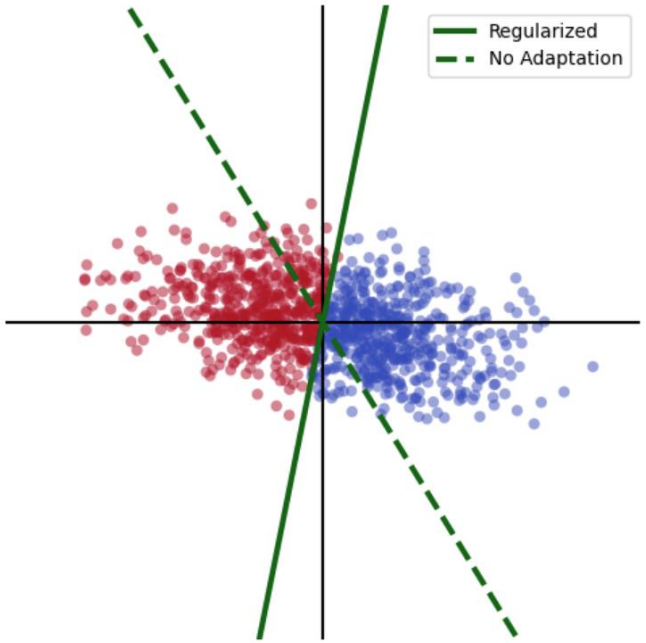
Regularizing with the target domain

Removing the regularization in the source domain

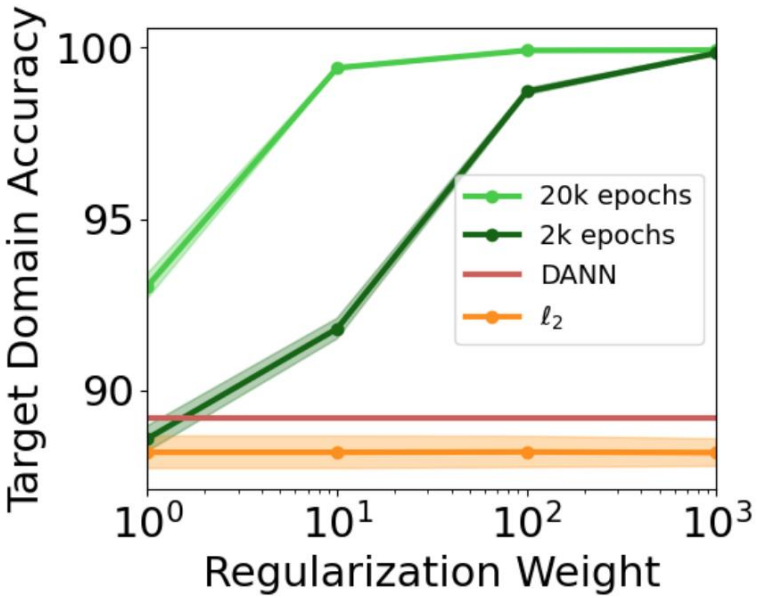
Synthetic example



(a) Source Domain



(b) Target Domain



(c) Performance

Visit our poster for

- Formal definition of Label Alignment Property (LAP)
- Evidence for LAP in pre-trained neural network representations
- Theoretical results for Label Alignment Regularizer (LAR)
- Comparison between LAR and domain-adversarial methods on cross-lingual sentiment classification