# End-to-end Learning, with or without Labels

Corinne Jones, Vincent Roulet, Zaid Harchaoui
University of Washington
{cjones6, vroulet, zaid}@uw.edu

## Abstract

We develop a framework **unifying unsupervised and supervised** end-to-end learning. The objective function adjusts gracefully to the amount of supervision, reducing to a **clustering** objective when only unlabeled data is available and to a **classification** objective when only labeled data is available.

## X-Supervised Objective

Consider observations $x_i, i = 1, \dots, n$, whose labels $y_i^\star \in \{0,1\}^k$ may or may not be observed. Given a deep network $\phi$, we optimize over the unknown labels and the parameters $V$ of the network and $W, b$ of a classifier:

$$\min_{Y \in \mathcal{C}', V, W, b} \frac{1}{n} \sum_{i=1}^{n} \left\| y_i - W^T \phi(x_i; V) - b \right\|_2^2 + \alpha \sum_{j=1}^{m} \|V_j\|_F^2 + \lambda \|W\|_F^2 - \rho \sum_{i=1}^{n} \|\phi(x_i; V) - \bar{\phi}\|_2^2$$

where

$$\mathcal{C}' := \{Y \in \{0,1\}^{n \times k} : \; Y \mathbb{1}_k = 1, y_i = y_i^\star \;\text{ for } i \in \mathcal{S}, \; n_{\min} \mathbb{1}_k \leq Y^T \mathbb{1}_n \leq n_{\max} \mathbb{1}_k\}$$

and $\bar{\phi} = 1/n \sum_{i=1}^n \phi(x_i; V)$. This extends the objective of Bach and Harchaoui (2007) to the deep setting.

The additional penalty and constraints are interpreted as follows:

- $\rho \sum_{i=1}^n \|\phi(x_i; V) - \bar{\phi}\|_2^2$ : Prevents mapping all observations to the same embedding, i.e., $\phi(x_1; V) = \phi(x_2; V) = \cdots = \phi(x_n; V)$
- $n_{\min} \mathbb{1}_k \leq Y^T \mathbb{1}_n \leq n_{\max} \mathbb{1}_k$ : Prevents assigning all observations to the same cluster, i.e., $y_1 = y_2 = \cdots = y_n$

No supervision

$M =$

| | | | | |
|---|---|---|---|---|
| 1 | | | | |
| | 1 | | | |
| | | 1 | | |
| | | | 1 | |
| | | | | 1 |

$\min_{M,V}$ Trace $MA(\Phi(V))$
$s.t.$ $M \in \mathcal{C}'$

Limited supervision

$M =$

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 0 | | |
| 1 | 1 | 0 | | |
| 0 | 0 | 1 | | |
| | | | 1 | |
| | | | | 1 |

$\min_{M,V}$ Trace $MA(\Phi(V))$
$s.t.$ $M \in \mathcal{C}'$,
$M_{ij}$ fixed for $(i,j) \in \mathcal{K}$

Full supervision

$M =$

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 |

$\min_{V}$ Trace $MA(\Phi(V))$
$M$ fixed

Example equivalence matrix $M = YY^T$ and problem for varying levels of supervision. For simplicity, $\alpha = \rho = 0$.

After optimizing over $W$ and $b$ in closed form, we obtain the problem

$$\begin{aligned} \min_{M,V} \quad & \lambda \operatorname{tr}[MA(\Phi(V))] + \alpha \sum_{j=1}^{m} \|V_j\|_F^2 - \rho \sum_{i=1}^{n} \|\phi_i(V) - \bar{\phi}\|_2^2 \\ \text{s.t.} \quad & M_{ij} = m_{ij} \quad \forall\, (i,j) \in \mathcal{K} \\ & n_{\min} \mathbb{1}_n \leq M \mathbb{1}_n \leq n_{\max} \mathbb{1}_n \\ & n_{\min} \mathbb{1}_n \leq M^T \mathbb{1}_n \leq n_{\max} \mathbb{1}_n \end{aligned} \tag{1}$$

where

- $M = YY^T$
- $\phi_i(V) = \phi(x_i; V)$ and $\Phi(V) = (\phi_1(V), \dots, \phi_n(V))^T$.
- $A(\Phi) = \Pi_n \left(\Pi_n \Phi \Phi^T \Pi_n + n\lambda \, \mathrm{I}\right)^{-1} \Pi_n$ and $\Pi_n = \mathrm{I}_n - \mathbb{1}_n \mathbb{1}_n^T / n$
- $\mathcal{K}$ is the set of indices of known entries $m_{ij}$ of $M := YY^T$ .

## Optimization

We optimize over $V$ by repeating the following steps on mini-batches:

1. Obtain an approximate solution $M_t$ of the assignment problem for V fixed by matrix balancing (a generalization of Sinkhorn's algorithm).
2. Compute the gradient of the objective with respect to V by backpropagating through the computations and perform a gradient step.

The overall algorithm, called *XSDC* for "X-Supervised Discriminative Clustering", where "X" can be "un", "semi" or "-" (hence covering all cases), is as follows:

Algorithm 1: XSDC

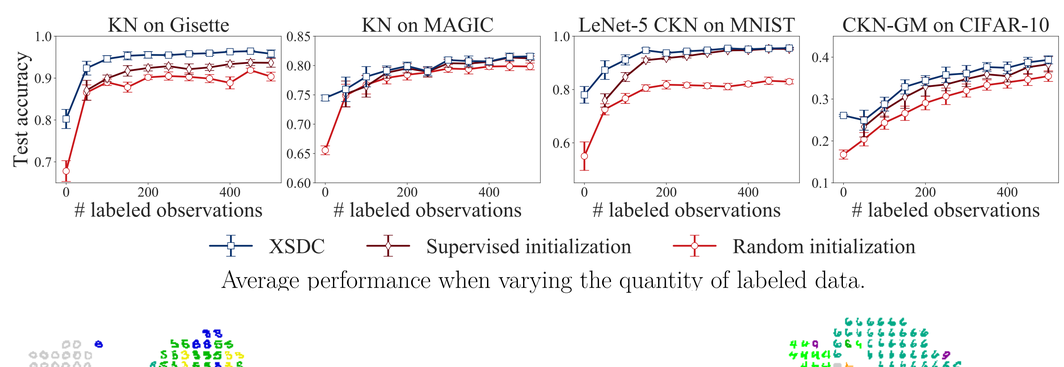
**Input:** Labeled data $X_\mathcal{S}, Y_\mathcal{S}$
  Unlabeled data $X_\mathcal{U}$
  Number of iterations $T$
**Initialize:** $V_1 \leftarrow$ Optimize (1) over $V$ using $X_\mathcal{S}, Y_\mathcal{S}$
**for** $t = 1, \dots, T$ **do**
  $X_t, Y_t \leftarrow$ Draw mini-batch of samples
  $M_t \leftarrow \text{MatrixBalancing}(A(\Phi(X_t; V_t)), Y_t Y_t^T)$
  $V_{t+1} \leftarrow \text{GradientStep}(\Phi(X_t; V_t), M_t, V_t)$
**end for**
$\hat{Y}_\mathcal{U} \leftarrow \text{NearestNeighbor}(\Phi(X; V_T), Y_\mathcal{S})$
$\hat{W}, \hat{b} \leftarrow \text{RegLeastSquares}(X; [Y_\mathcal{S}, \hat{Y}_\mathcal{U}])$
**Output:** $\hat{Y}_\mathcal{U}, V_T, \hat{W}, \hat{b}$

## Acknowledgements

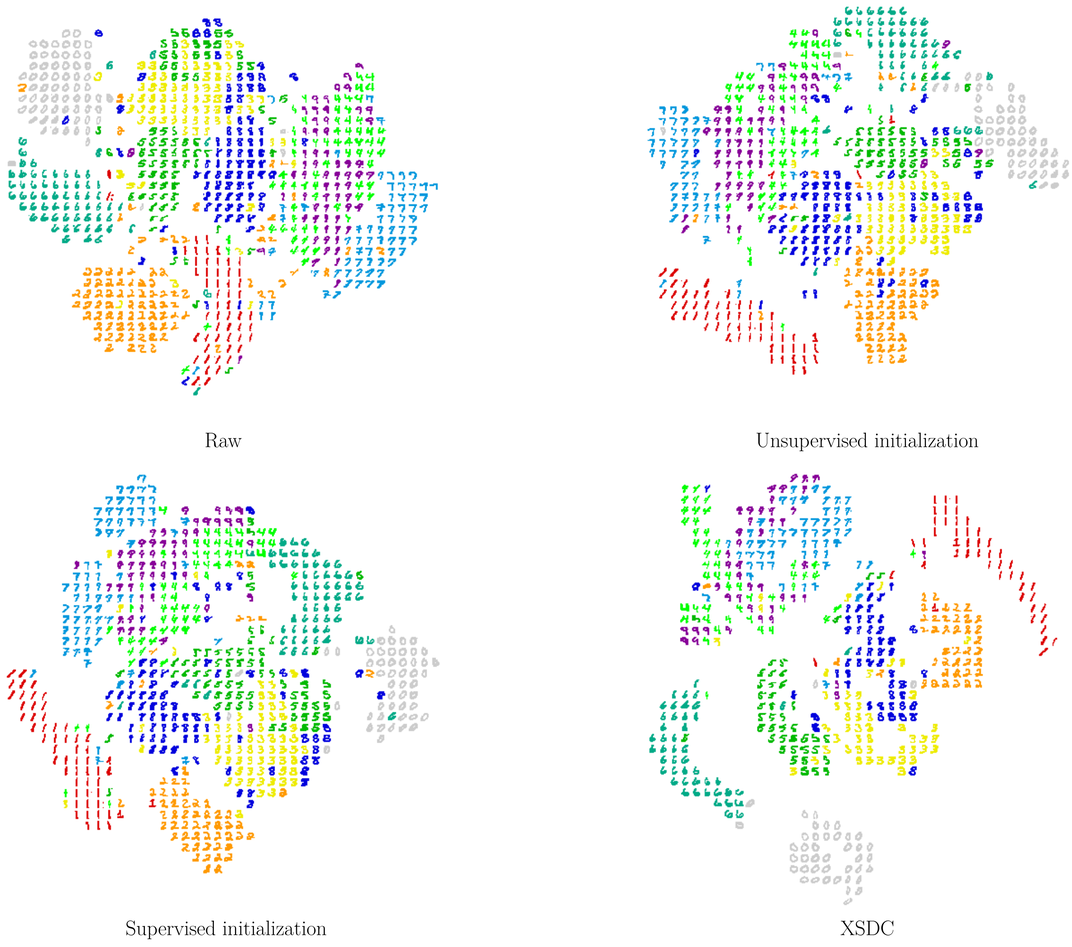
This project was supported by the National Science Foundation and the "Learning in Machines and Brains" program of the Canadian Institute for Advanced Research.

## Results

1. XSDC improves upon using only labeled data in the cases where additional labeled data would help but is unavailable.

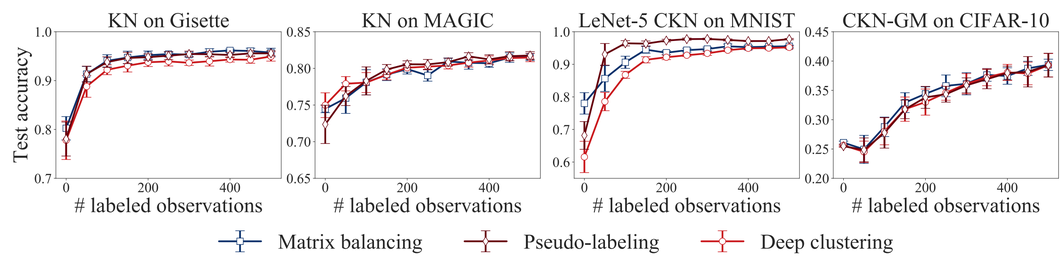



Average performance when varying the quantity of labeled data.



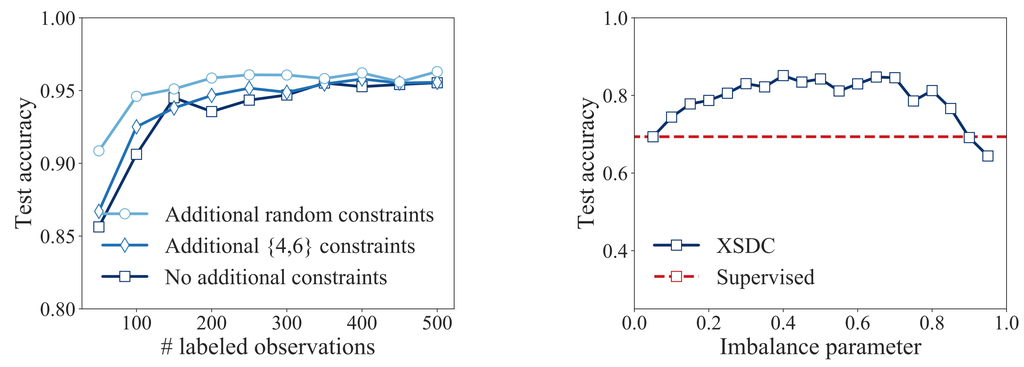

Visualizations of unlabeled MNIST features obtained when training a LeNet-5 CKN with 50 labeled points.

2. The labeling algorithm in XSDC ("matrix balancing") is typically competitive with alternative, less principled, labeling approaches.




Average performance when using matrix balancing and two alternative labeling methods, pseudo-labeling (Lee, 2013) and deep clustering (Caron et al., 2018), and varying the quantity of labeled data.

3. XSDC can use additional must-link and must-not-link constraints and can handle moderately unbalanced datasets.




Average performance when training a LeNet-5 CKN on MNIST when adding additional constraints and varying the quantity of labeled data (left) and when varying the balance of unlabeled data (right).

## References

F. R. Bach and Z. Harchaoui. DIFFRAC: a discriminative and flexible framework for clustering. In *Neural Information Processing Systems*, 2007.

M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018.

D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning Workshop on Challenges in Representation Learning*, 2013.