More Than Meets The Eye: Semi-supervised Learning Under Non-IID Data

Saul Calderon-Ramirez* and Luis Oala*
*Equal contribution
sacalderon@itcr.ac.cr, luis.oala@hhi.fraunhofer.de

RobustML Workshop
ICLR

Motivation and context

Mismatch between distributions P_l and P_u of **labelled data** S_l **and unlabelled data** S_u in semi-supervised deep learning (SSDL)

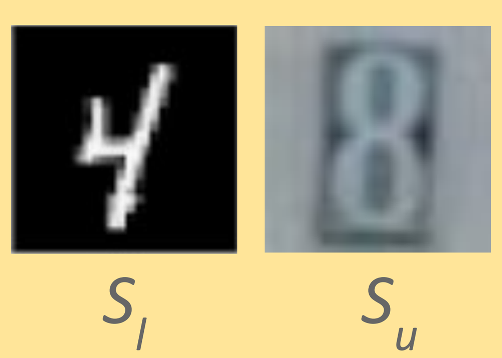
Question

Given labelled data S_l ...

...which unlabelled data S_u should we choose?

In absence of explicit data models, **semantic similarity matching** is often used:

"Numbers with numbers"



"Clothing with clothing"

...

Contributions

- Quantitative **impact estimation** of **distribution mismatch** for MixMatch SSDL
- Deep data set dissimilarity measures (**DeDiMs**): a simple and quantitative decision heuristic for S_u **selection** ***before*** **SSDL training**

Proposed method

Deep data set dissimilarity measure (DeDiM)

$$\widehat{d}_j = \sum_{r=1}^{n'} \delta_g \left(p_{r,a}, p_{r,b}\right)$$

$\widehat{d}_j$ estimated dissimilarity for the sample *j*

$\sum_{r=1}^{n'}$ sum over all *n'* dimensions in feature space

δ_g distance measure, *g=c* for cosine distance

$p_{r,a}$ approximate density functions for feature *r* of data set *a*

$p_{r,b}$ approximate density functions for feature *r* of data set *b*

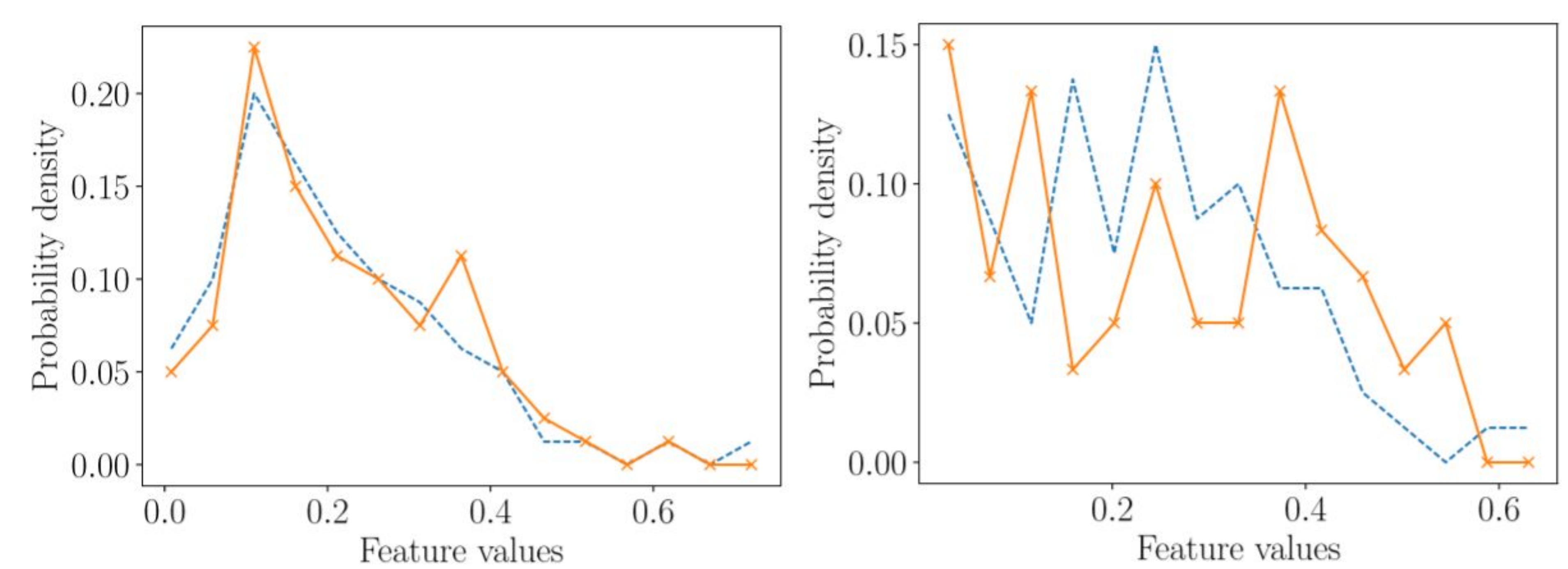


Fig 1. A specific feature density of a model trained with MNIST labelled data (orange and continuos inboth plots), and ImageNet and SVHN unlabelled data (left and right column, respectively, with the blue dashed line in both)

Results

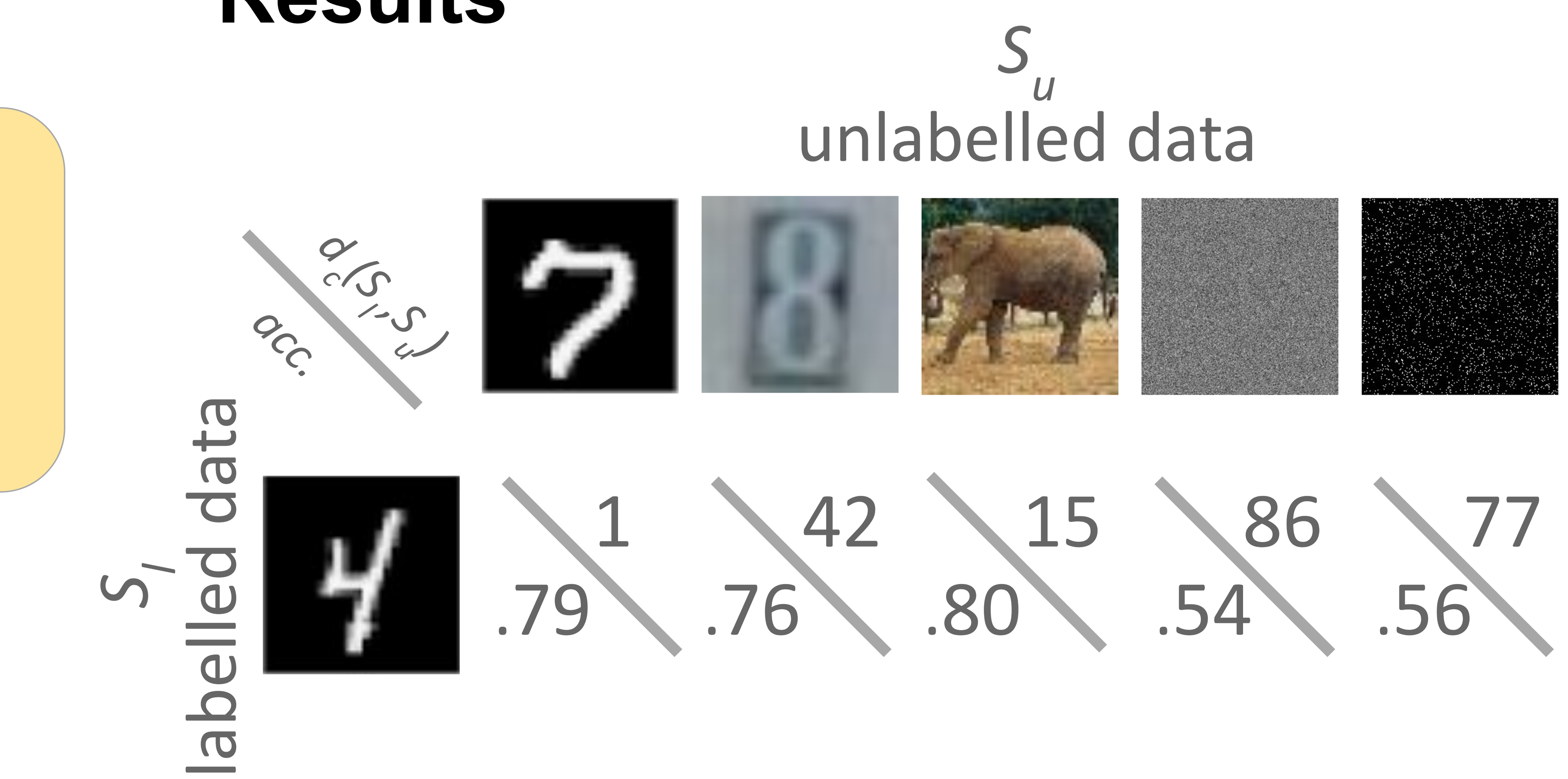


Fig 2. SSDL accuracy (acc.) and cosine deep data set dissimilarity measures between labelled and unlabelled data ($d_c(S_l,S_u)$). Full results on more data sets in the paper.

S_l	n_l	d_{ℓ_1}	d_{ℓ_2}	d_{JS}	d_C
MNIST	60	-0.876	-0.898	-0.969	-0.944
	100	-0.805	-0.83	-0.786	-0.948
	150	-0.794	-0.822	-0.81	-0.944
CIFAR-10	60	-0.823	-0.853	-0.944	-0.921
	100	-0.826	-0.878	-0.966	-0.947
	150	-0.808	-0.838	-0.952	-0.927
FashionMNIST	60	-0.2	-0.268	-0.735	-0.789
	100	-0.264	-0.326	-0.781	-0.824
	150	-0.286	-0.347	-0.785	-0.827

Tab 1. Correlation results for the dissimilarity measures between S_l and S_u with OOD contamination and SSDL accuracy

Conclusions

- **Semantic similarity matching** between labelled and unlabelled data is **not a reliable recipe** for successful SSDL
- Deep data set dissimilarity measures (**DeDiMs**) offer a **simple, quantitative and practical** decision heuristic for S_u **selection** ***before*** **SSDL training**

De Montfort University Leicester
TEC Tecnológico de Costa Rica
Fraunhofer HHI

Full paper https://arxiv.org/abs/2104.10223

Code https://github.com/luisoala/non-iid-ssdl