Extending the WILDS Benchmark for Unsupervised Adaptation

Data, code, and paper at https://wilds.stanford.edu

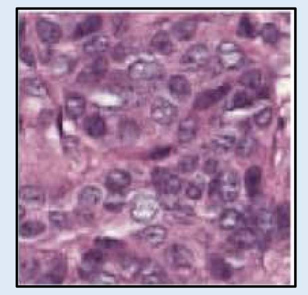
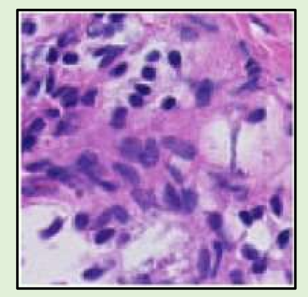
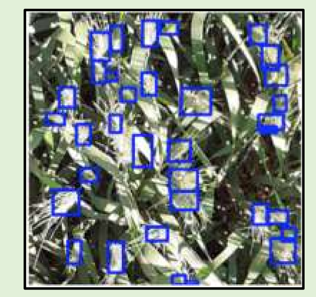
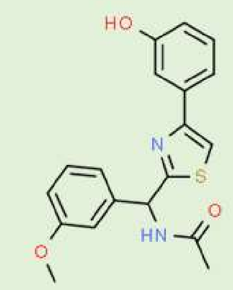
Shiori Sagawa*, Pang Wei Koh*, Tony Lee*, Irena Gao*, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, Percy Liang

STANFORD, CALTECH, INRAE, THE UNIVERSITY OF SASKATCHEWAN, THE UNIVERSITY OF TOKYO, BOSTON UNIVERSITY, AND UC BERKELEY

Overview

Unlabeled data can be a powerful source of leverage for mitigating distribution shifts. We extend 8 datasets in WILDS with unlabeled data that would be realistically obtainable in deployment.

We systematically benchmark methods that leverage unlabeled data, including domain-invariant, self-training, and self-supervised methods, and show that many methods perform similarly to or even underperform purely-supervised ERM.

	Dataset	iWildCam	Camelyon17	FMoW	PovertyMap	GlobalWheat	OGB-MolPCBA	CivilComments	Amazon
	Input (x)	camera trap photo	tissue slide	satellite image	satellite image	wheat image	molecular graph	online comment	product review
	Prediction (y)	animal species	tumor	land use	asset wealth	wheat head bbox	bioassays	toxicity	sentiment
	Domain (d)	camera	hospital	time, region		location, time	scaffold	demographic	user
	Source example							What do Black and LGBT people have to do with bicycle licensing?	Overall a solid package that has a good quality of construction for the price.
	Target example							As a Christian, I will not be patronizing any of those businesses.	I *loved* my French press, it's so perfect and came with all this fun stuff!
	Original paper	Beery et al. 2020	Bandi et al. 2018	Christie et al. 2018	Yeh et al. 2020	David et al. 2021	Hu et al. 2020	Borkan et al. 2019	Ni et al. 2019
Labeled	# domains	323	5	16 x 5	23 x 2	47	120,084	16	3,920
	# examples	203,029	455,954	141,696	19,669	6,515	437,929	448,000	539,502
Unlabeled: Source domains	# domains	-	3	11 x 5	13 x 2	18	44,930	-	-
	# examples	-	1,799,247	11,948	181,948	5,997	4,052,627	-	-
Unlabeled: Extra domains	# domains	3,215	-	-	-	53	-	1	21,694
	# examples	819,120	-	-	-	42,445	-	1,551,515	2,927,841
Unlabeled: Validation domains	# domains	-	1	3 x 5	5 x 2	11	31,361	-	1,334
	# examples	-	600,030	155,313	24,173	2,000	430,325	-	266,066
Unlabeled: Target domains	# domains	-	1	2 x 5	5 x 2	18	43,793	-	1,334
	# examples	-	600,030	173,208	55,275	8,997	517,048	-	268,761

Experimental Results

	iWildCam2020-wilds (Unlabeled extra, macro F1) In-distribution	Out-of-distribution	FMoW-wilds (Unlabeled target, worst-region acc) In-distribution	Out-of-distribution
ERM (-data aug)	46.7 (0.6)	30.6 (1.1)	59.3 (0.7)	33.7 (1.5)
ERM	47.0 (1.4)	**32.2** (1.2)	60.6 (0.6)	34.8 (1.5)
CORAL	41.6 (2.0)	25.7 (1.2)	57.9 (3.0)	35.4 (2.7)
DANN	48.5 (2.8)	**31.9** (1.4)	57.9 (0.8)	34.6 (1.7)
Pseudo-Label	47.3 (0.4)	30.3 (0.4)	60.9 (0.5)	33.7 (0.2)
FixMatch	46.3 (0.5)	**31.0** (1.3)	58.6 (2.4)	32.1 (2.0)
Noisy Student	47.5 (0.9)	**32.1** (0.7)	61.3 (0.4)	**37.8** (0.6)
SwAV	47.3 (1.4)	29.0 (2.0)	61.8 (1.0)	36.3 (1.0)
ERM (fully-labeled)	54.6 (1.5)	44.0 (2.3)	65.4 (0.4)	58.7 (1.4)
	Camelyon17-wilds (Unlabeled target, avg acc) In-distribution	Out-of-distribution	**PovertyMap-wilds (Unlabeled target, worst U/R corr)** In-distribution	Out-of-distribution
ERM (-data aug)	85.8 (1.9)	70.8 (7.2)	0.65 (0.03)	**0.50** (0.07)
ERM	90.6 (1.2)	82.0 (7.4)	0.66 (0.04)	**0.49** (0.06)
CORAL	88.9 (1.5)	86.2 (2.9)	0.55 (0.10)	**0.46** (0.05)
DANN	86.9 (2.2)	68.4 (9.2)	0.50 (0.07)	0.33 (0.10)
Pseudo-Label	91.3 (1.3)	67.7 (8.2)	–	–
FixMatch	91.3 (1.1)	71.0 (4.9)	0.54 (0.11)	0.30 (0.11)
Noisy Student	93.2 (0.5)	86.7 (1.7)	0.61 (0.07)	0.42 (0.11)
SwAV	92.3 (0.4)	**91.4** (2.0)	0.60 (0.13)	**0.45** (0.05)
	GlobalWheat-wilds (Unlabeled target, avg domain acc) In-distribution	Out-of-distribution	**OGB-MolPCBA (Unlabeled target, avg AP)** In-distribution	Out-of-distribution
ERM	77.8 (0.2)	**51.0** (0.7)	–	**28.3** (0.1)
CORAL	–	–	–	22.0 (3.0)
DANN	–	–	–	20.4 (0.8)
Pseudo-Label	73.3 (0.9)	42.9 (2.3)	–	19.7 (0.1)
Noisy Student	78.1 (0.3)	46.8 (1.2)	–	27.5 (0.1)
	CivilComments-wilds (Unlabeled extra, worst-group acc) In-distribution	Out-of-distribution	**Amazon-wilds (Unlabeled target, 10th percentile acc)** In-distribution	Out-of-distribution
ERM	89.8 (0.8)	**66.6** (1.6)	72.0 (0.1)	**54.2** (0.8)
CORAL	–	–	71.7 (0.1)	53.3 (0.0)
DANN	–	–	71.7 (0.1)	53.3 (0.0)
Pseudo-Label	90.3 (0.5)	**66.9** (2.6)	71.6 (0.1)	52.3 (1.1)
Masked LM	89.4 (1.2)	**65.7** (2.3)	71.9 (0.4)	**53.9** (0.7)
ERM (fully-labeled)	89.9 (0.1)	69.4 (0.6)	73.6 (0.1)	56.4 (0.8)

Problem Setting

Examples in WILDS belong to different domains. The unlabeled data can be from the same domains as the labeled train split ("**source domains**"), the labeled validation split ("**validation domains**"), the labeled test split ("**target domains**"), or from disjoint domains with no labeled data ("**extra domains**").

Using WILDS

Our PyTorch-based package automates data loading and evaluation, and comes with default models and baselines. Just `pip install wilds`!

```
>>> from wilds import get_dataset
>>> from wilds.common.data_loaders import get_train_loader
>>> import torchvision.transforms as transforms
# Load the labeled data
>>> dataset = get_dataset(dataset="fmow", download=True)
>>> labeled_subset = dataset.get_subset("train", transform=transforms.ToTensor())
>>> data_loader = get_train_loader("standard", labeled_subset, batch_size=16)
# Load the unlabeled data
>>> dataset = get_dataset(dataset="fmow", unlabeled=True, download=True)
>>> unlabeled_subset = dataset.get_subset("test_unlabeled", transform=transforms.ToTensor())
>>> unlabeled_data_loader = get_train_loader("standard", unlabeled_subset, batch_size=64)
# Train loop
>>> for labeled_batch, unlabeled_batch in zip(data_loader, unlabeled_data_loader):
...     x, y, metadata = labeled_batch
...         unlabeled_x, unlabeled_metadata = unlabeled_batch
...     ...
```