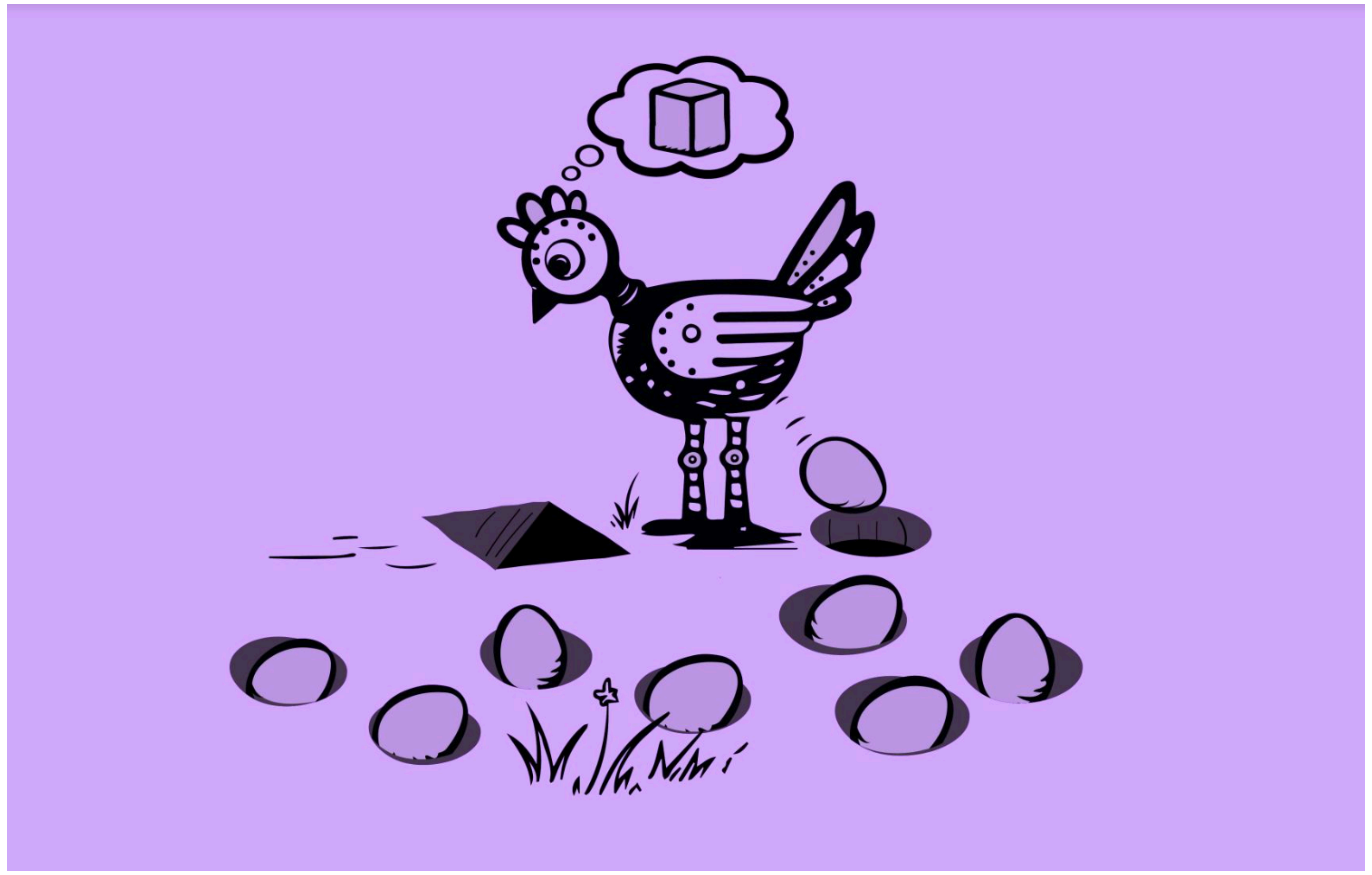
In Search of Lost Domain Generalization

Jingwen Liu



Paper by Ishaan Gulrajani and David Lopez-Paz

Feb 13

What is domain generalization?

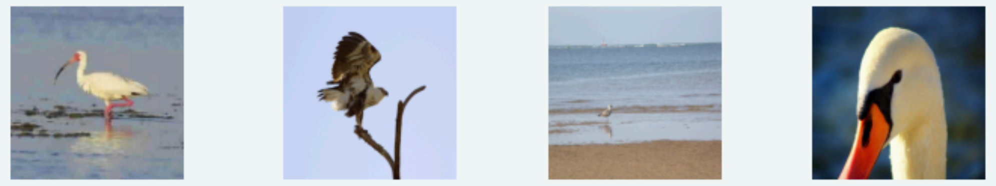
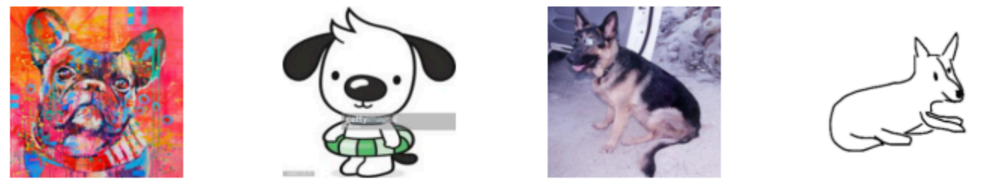
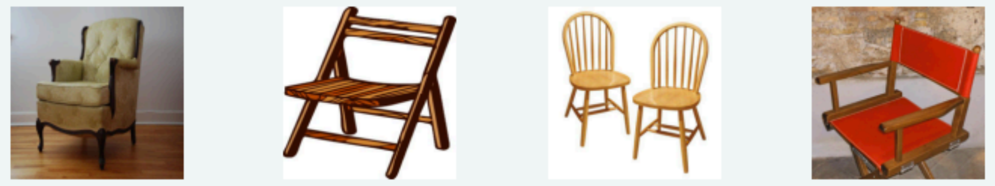
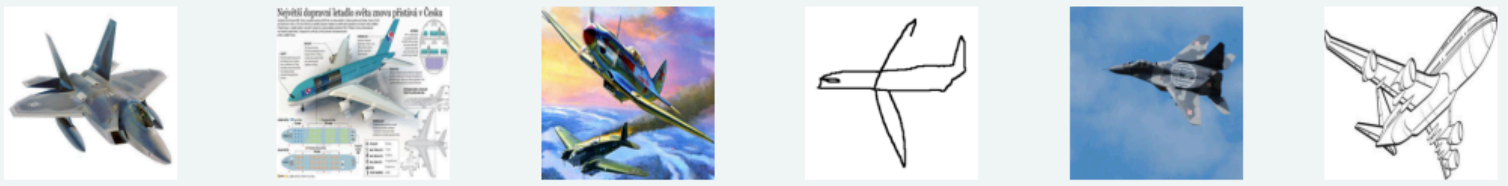
Classical Supervised Learning

- Dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ iid from $P(X, Y)$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,\infty)$
- Goal: find a predictor $f : \mathcal{X} \to \mathcal{Y}$ that minimizes $\mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)]$
- Approach: ERM minimize $\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$

Domain Generalization Problem

- k different domains: for each $j \in \{1, \ldots, k\}$ Dataset $D^j = \{(x_i^j, y_i^j)\}_{i=1}^{n_j}$ iid from $P(X^j, Y^j)$
- Goal: out-of-distribution generalization find a predictor f perform well at unseen test domain d_{test}
- Need to assume some invariances across train and test domains

Example Datasets

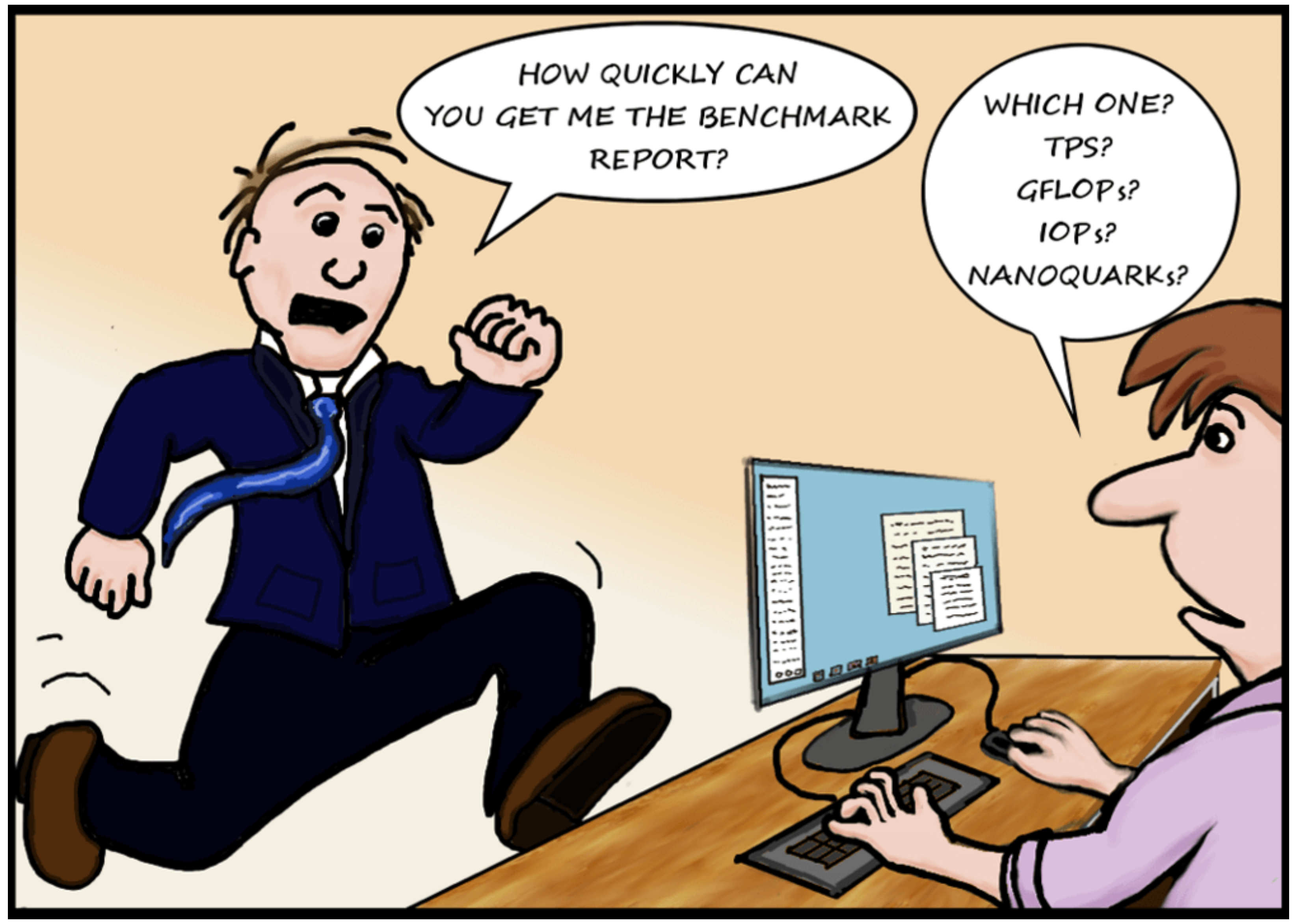
Dataset	Domains
Colored MNIST	+90%, +80%, -90%  *(degree of correlation between color and label)*
Rotated MNIST	0°, 15°, 30°, 45°, 60°, 75° 
VLCS	Caltech101, LabelMe, SUN09, VOC2007 
PACS	Art, Cartoon, Photo, Sketch 
Office-Home	Art, Clipart, Product, Photo 
Terra Incognita	L100, L38, L43, L46  *(camera trap location)*
DomainNet	Clipart, Infographic, Painting, QuickDraw, Photo, Sketch 

Lots of Algorithms, but …

- Empirical Risk Minimization (ERM)
- Group Distributionally Robust Optimization (DRO)
- DANN
- Invariant Risk Minimization (IRM)
- …

🤩🤩🤩

- All evaluated under different datasets and model selection methods
- Need a standardized and rigorous benchmark to make fair comparisons



What could go wrong?

Model Selection

- Need to choose hyperparameters
- Choose between different architecture variants
- But no validation data $\approx$ test data
- What's the correct way of doing model selection?

Training-domain validation set

- For each $j \in \{1,\ldots,k\}$, split the data set $D^j = \{(x_i^j, y_i^j)\}_{i=1}^{n_j}$ into training and validation subsets
- Combine the validation subsets of each domain
    - create an overall validation set
- Choose the model that does the best on this overall validation set
- Assumes training sample and test sample following similar distributions

Leave-one-domain-out cross-validation

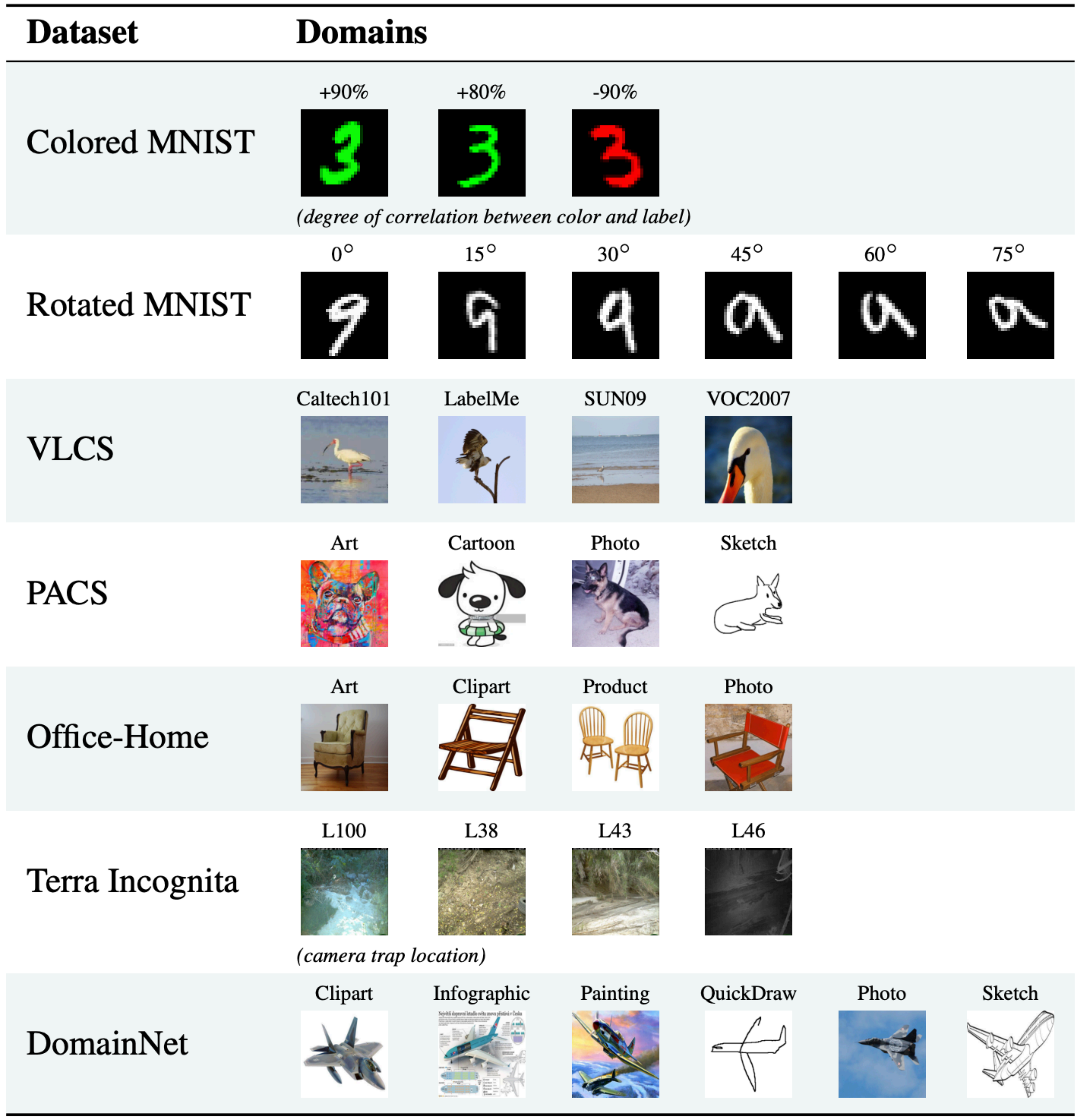
- For each hyperparameter set, train k models, each leaving one domain dataset outside of the training set
- Evaluate each model on its held-out domain and average the accuracies over k models
- Pick the hyperparamter set that has the best performance on the averaged accuracy
- Retrain the model using all k domains
- Assume training and test domain are drawn from a meta-distribution over domains

Test-domain validation set (oracle)

- Validation set $\sim$ test distribution
- Query access
- Limit the number of queries i.e. at most 20 queries in this paper

DOMAINBED

- Datasets



- Model selection criteria
- Train-domain validation set
- Leave-one-domain-out cross-validation
- Test-domain oracle validation

Baseline Algorithms

- Empirical Risk Minimization (**ERM**, Vapnik [1998]) minimizes the sum of errors across domains and examples.
- Group Distributionally Robust Optimization (**DRO**, Sagawa et al. [2019]) performs ERM while increasing the importance of domains with larger errors.
- Inter-domain Mixup (**Mixup**, Xu et al. [2019], Yan et al. [2020], Wang et al. [2020]) performs ERM on linear interpolations of examples from random pairs of domains and their labels.
- Meta-Learning for Domain Generalization (**MLDG**, Li et al. [2018a]) leverages MAML [Finn et al., 2017] to meta-learn how to generalize across domains.
- Different variants of the popular algorithm of Ganin et al. [2016] to learn features $\phi(X^d)$ with distributions matching across domains:
  - Domain-Adversarial Neural Networks (**DANN**, Ganin et al. [2016]) employ an adversarial network to match feature distributions.
  - Class-conditional DANN (**C-DANN**, Li et al. [2018d]) is a variant of DANN matching the conditional distributions $P(\phi(X^d)|Y^d = y)$ across domains, for all labels y.
  - **CORAL** [Sun and Saenko, 2016] matches the mean and covariance of feature distributions.
  - **MMD** [Li et al., 2018b] matches the MMD [Gretton et al., 2012] of feature distributions.
- Invariant Risk Minimization (**IRM** [Arjovsky et al., 2019]) learns a feature representation $\phi(X^d)$ such that the optimal linear classifier on top of that representation matches across domains.

Experiment Results

Compare to the state-of-the-art for typical datasets

Dataset / algorithm	Out-of-distribution accuracy (by domain)						
Rotated MNIST	0°	15°	30°	45°	60°	75°	Average
Ilse et al. [2019]	93.5	99.3	99.1	99.2	99.3	93.0	97.2
Our ERM	95.6	99.0	98.9	99.1	99.0	96.7	**98.0**
PACS	A	C	P	S			Average
Asadi et al. [2019]	83.0	79.4	96.8	78.6			84.5
Our ERM	88.1	78.0	97.8	79.1			**85.7**
VLCS	C	L	S	V			Average
Albuquerque et al. [2019]	95.5	67.6	69.4	71.1			75.9
Our ERM	97.6	63.3	72.2	76.4			**77.4**
Office-Home	A	C	P	R			Average
Zhou et al. [2020]	59.2	52.3	74.6	76.0			65.5
Our ERM	62.7	53.4	76.5	77.3			**67.5**

Experiment Results

Model selection method: training domain validation set

Algorithm	CMNIST	RMNIST	VLCS	PACS	Office-Home	TerraInc	DomainNet	Avg
ERM	52.0 ± 0.1	98.0 ± 0.0	77.4 ± 0.3	85.7 ± 0.5	67.5 ± 0.5	47.2 ± 0.4	41.2 ± 0.2	67.0
IRM	51.8 ± 0.1	97.9 ± 0.0	78.1 ± 0.0	84.4 ± 1.1	66.6 ± 1.0	47.9 ± 0.7	35.7 ± 1.9	66.0
DRO	52.0 ± 0.1	98.1 ± 0.0	77.2 ± 0.6	84.1 ± 0.4	66.9 ± 0.3	47.0 ± 0.3	33.7 ± 0.2	65.5
Mixup	51.9 ± 0.1	98.1 ± 0.0	77.7 ± 0.4	84.3 ± 0.5	69.0 ± 0.1	48.9 ± 0.8	39.6 ± 0.1	67.1
MLDG	51.6 ± 0.1	98.0 ± 0.0	77.1 ± 0.4	84.8 ± 0.6	68.2 ± 0.1	46.1 ± 0.8	41.8 ± 0.4	66.8
CORAL	51.7 ± 0.1	98.1 ± 0.1	77.7 ± 0.5	86.0 ± 0.2	68.6 ± 0.4	46.4 ± 0.8	41.8 ± 0.2	67.2
MMD	51.8 ± 0.1	98.1 ± 0.0	76.7 ± 0.9	85.0 ± 0.2	67.7 ± 0.1	49.3 ± 1.4	39.4 ± 0.8	66.8
DANN	51.5 ± 0.3	97.9 ± 0.1	78.7 ± 0.3	84.6 ± 1.1	65.4 ± 0.6	48.4 ± 0.5	38.4 ± 0.0	66.4
C-DANN	51.9 ± 0.1	98.0 ± 0.0	78.2 ± 0.4	82.8 ± 1.5	65.6 ± 0.5	47.6 ± 0.8	38.9 ± 0.1	66.1

Model selection method: Leave-one-domain-out cross-validation

Algorithm	CMNIST	RMNIST	VLCS	PACS	Office-Home	TerraInc	DomainNet	Avg
ERM	34.2 ± 1.2	98.0 ± 0.0	76.8 ± 1.0	83.3 ± 0.6	67.3 ± 0.3	46.2 ± 0.2	40.8 ± 0.2	63.8
IRM	36.3 ± 0.4	97.7 ± 0.1	77.2 ± 0.3	82.9 ± 0.6	66.7 ± 0.7	44.0 ± 0.7	35.3 ± 1.5	62.9
DRO	32.2 ± 3.7	97.9 ± 0.1	77.5 ± 0.1	83.1 ± 0.6	67.1 ± 0.3	42.5 ± 0.2	32.8 ± 0.2	61.8
Mixup	31.2 ± 2.1	98.1 ± 0.1	78.6 ± 0.2	83.7 ± 0.9	68.2 ± 0.3	46.1 ± 1.6	39.4 ± 0.3	63.6
MLDG	36.9 ± 0.2	98.0 ± 0.1	77.1 ± 0.6	82.4 ± 0.7	67.6 ± 0.3	45.8 ± 1.2	42.1 ± 0.1	64.2
CORAL	29.9 ± 2.5	98.1 ± 0.1	77.0 ± 0.5	83.6 ± 0.6	68.6 ± 0.2	48.1 ± 1.3	41.9 ± 0.2	63.9
MMD	42.6 ± 3.0	98.1 ± 0.1	76.7 ± 0.9	82.8 ± 0.3	67.1 ± 0.5	46.3 ± 0.5	39.3 ± 0.9	64.7
DANN	29.0 ± 7.7	89.1 ± 5.5	77.7 ± 0.3	84.0 ± 0.5	65.5 ± 0.1	45.7 ± 0.8	37.5 ± 0.2	61.2
C-DANN	31.1 ± 8.5	96.3 ± 1.0	74.0 ± 1.0	81.7 ± 1.4	64.7 ± 0.4	40.6 ± 1.8	38.7 ± 0.2	61.1

Model selection method: Test-domain validation set *(oracle)*

Algorithm	CMNIST	RMNIST	VLCS	PACS	Office-Home	TerraInc	DomainNet	Avg
ERM	58.5 ± 0.3	98.1 ± 0.1	77.8 ± 0.3	87.1 ± 0.3	67.1 ± 0.5	52.7 ± 0.2	41.6 ± 0.1	68.9
IRM	70.2 ± 0.2	97.9 ± 0.0	77.1 ± 0.2	84.6 ± 0.5	67.2 ± 0.8	50.9 ± 0.4	36.0 ± 1.6	69.2
DRO	61.2 ± 0.6	98.1 ± 0.0	77.4 ± 0.6	87.2 ± 0.4	67.7 ± 0.4	53.1 ± 0.5	34.0 ± 0.1	68.4
Mixup	58.4 ± 0.2	98.0 ± 0.0	78.7 ± 0.4	86.4 ± 0.2	68.5 ± 0.5	52.9 ± 0.3	40.3 ± 0.3	69.0
MLDG	58.4 ± 0.2	98.0 ± 0.1	77.8 ± 0.4	86.8 ± 0.2	67.4 ± 0.2	52.4 ± 0.3	42.5 ± 0.1	69.1
CORAL	57.6 ± 0.5	98.2 ± 0.0	77.8 ± 0.1	86.9 ± 0.2	68.6 ± 0.4	52.6 ± 0.6	42.1 ± 0.1	69.1
MMD	63.4 ± 0.7	97.9 ± 0.1	78.0 ± 0.4	87.1 ± 0.5	67.0 ± 0.2	52.7 ± 0.2	39.8 ± 0.7	69.4
DANN	58.3 ± 0.2	97.9 ± 0.0	80.1 ± 0.6	85.4 ± 0.7	65.6 ± 0.3	51.6 ± 0.6	38.3 ± 0.1	68.2
C-DANN	62.0 ± 1.1	97.8 ± 0.1	80.2 ± 0.1	85.7 ± 0.3	65.6 ± 0.3	51.0 ± 1.0	38.9 ± 0.1	68.7

- ERM is very good
- Model selection methods matter

Some more questions

- Data augmentation pipeline
- “Right” dataset?

Thanks for listening!