Domain Generalization using Causal Matching

Divyat Mahajan, Shruti Tople, Amit Sharma

Microsoft Research

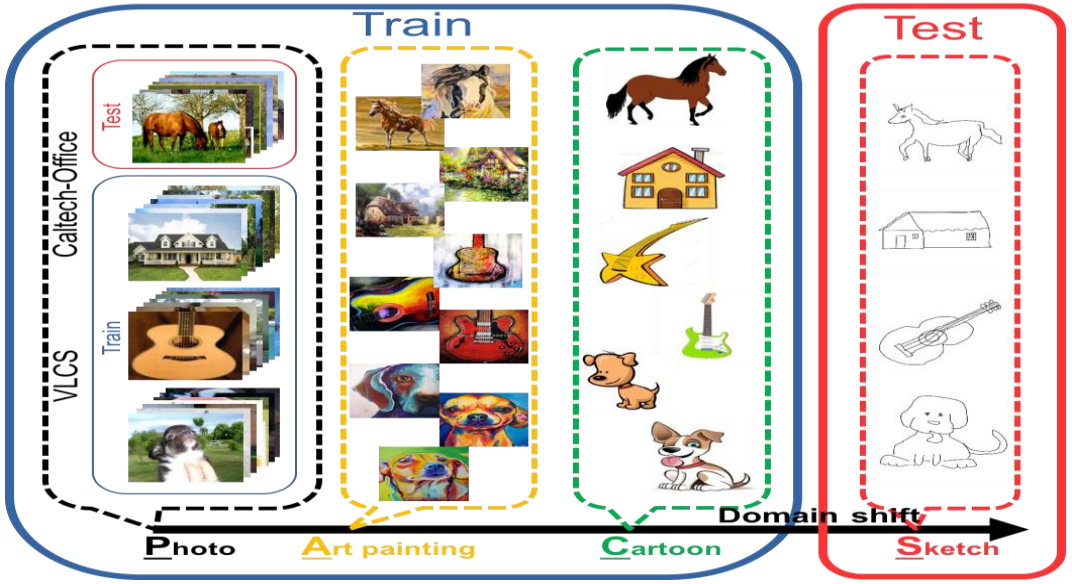
Spurious Correlations

	Common training examples		Test examples
Waterbirds	y: waterbird a: water background 	y: landbird a: land background 	y: waterbird a: land background 
CelebA	y: blond hair a: female 	y: dark hair a: male 	y: blond hair a: male 
MultiNLI	y: contradiction a: has negation (P) The economy could be still better. (H) The economy has never been better.	y: entailment a: no negation (P) Read for Slate's take on Jackson's findings. (H) Slate had an opinion on Jackson's findings.	y: entailment a: has negation (P) There was silence for a moment. (H) There was a short period of time where no one spoke.

[Link to paper]

Domain Generalization

- Domain Generalization (DG) aims to learn a single classifier that generalizes well to data from unseen domains/ distributions
- Training Data: $(d_i,\ x_i, y_i)_{i=1}^{n} \sim (D_m, X, Y)$ where $d_i \in D_m$ and $D_m \subset D$ is a set of m domains
- Covariate Shift Assumption: $p_{d_i}(y|x_s) = p_{d_j}(y|x_s)$ for any two domains d_i, d_j
- Need to capture the invariant mechanism $p(y|x_s)$ or identify the stable features x_s

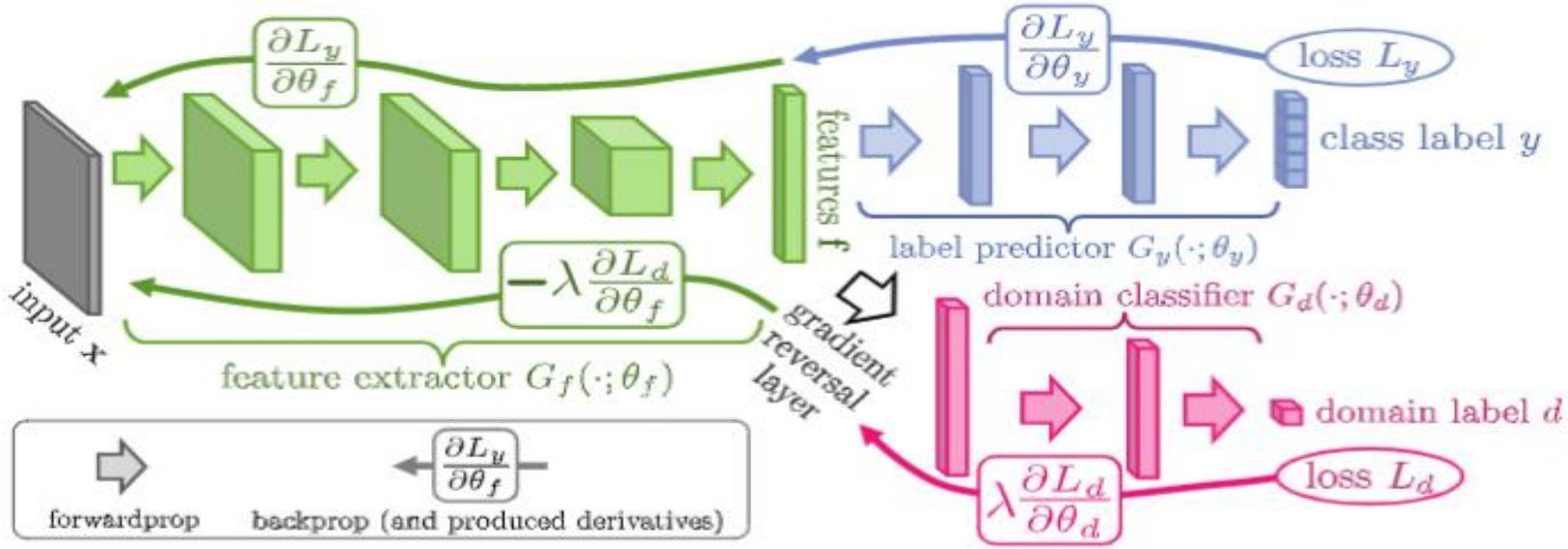


Li et al. 2017

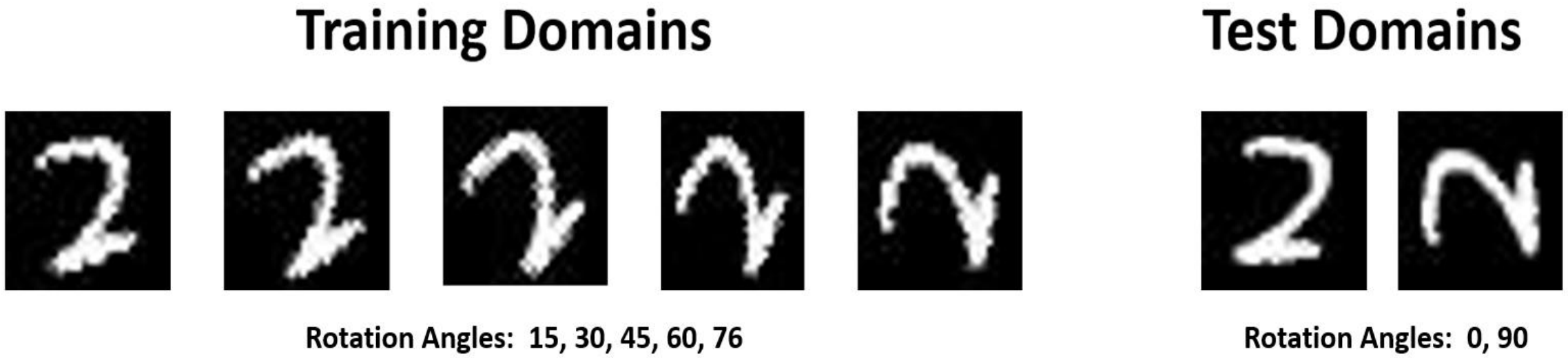
Feature Distribution Matching

- Domain Invariant Representations
  - Bad if labels and domain are correlated (Class Imbalance)
- Class Conditional Version
  - But does the distribution of invariant features need to be the same across domains?
  - Variance in the distribution due to different noise levels across domains

Domain Adversarial Training (Ganin et al.)



Perfect Match

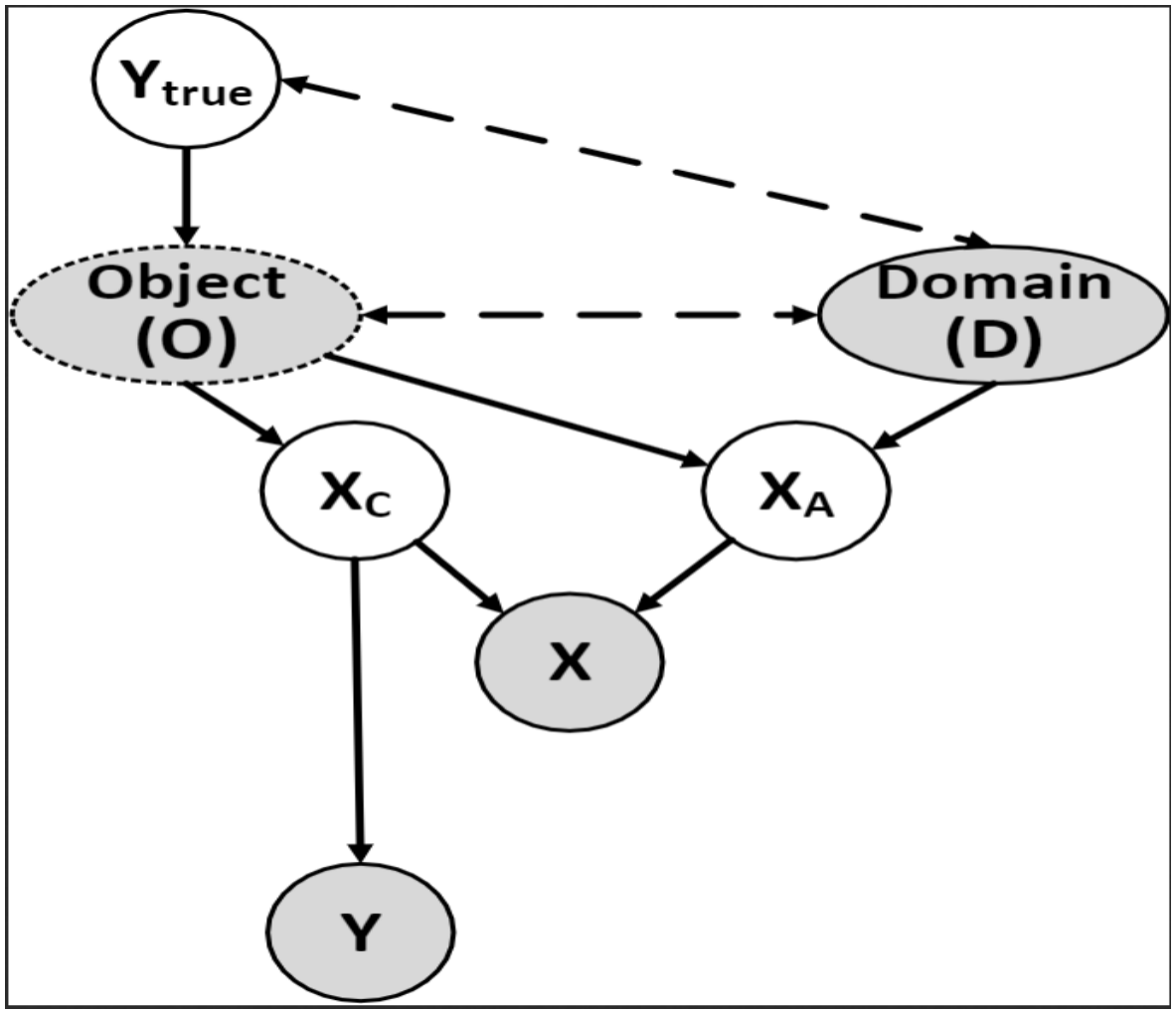


- Perfect Match:
  - Same data point rotated by different angle across domains shares the same invariant feature
  - Match feature representations for the “counterfactuals” of each data point across domains

Causal View of Domain Generalization

- Object (0) can be interpreted as the base person where the Domain (D) corresponds to different views that lead to creation of an image (X) for that person (O)

- Domains can be interpreted as interventions: For each observed x_i^d, there area set of counterfactual inputs $x_i^{d\prime}$ where $d \neq d$' , but both correspond to the (possibly unobserved) same object (O)



Invariance Condition from SCM

- Invariance Condition: $X_C \perp\!\!\!\perp D \mid O$
- Perfect Match:

$$f_{\texttt{perfectmatch}} = \arg\min_{h,\Phi} \sum_{d=1}^{m} L_d(h(\Phi(X)), Y) + \lambda \sum_{\Omega(j,k)=1; d \neq d'} \text{dist}(\Phi(\mathbf{x}_j^{(d)}), \Phi(\mathbf{x}_k^{(d')}))$$

- Prior work incorrectness:
  - Domain-invariant representations: $X_C \perp\!\!\!\perp D$
  - Class-conditional domain-invariant: $X_C \perp\!\!\!\perp D \mid Y_{true}$
  - Both incorrect due to backdoor path via Object O

Observational Data

- Latent base object not known generally in observational data (PACS, VLCS)
  - Perfect Match still applicable using self augmentations

- Class Conditional Approximation:
  - Data points with the same class label are likely to closer under causal features as compared to point with different class labels

- Inferring latent base objects / match function ($\Omega : X \times X \rightarrow \{0,1\}$)
  - Contrastive Loss: $Dist(\ Anchor, Positive\ Match\ ) - Dist(Anchor, Negative\ Match)$

- Iterative Contrastive Learning:
  - Initialize Ω with Random Match across domains with same class label
  - Using Ω to infer Positive Match given anchor and minimize contrastive loss
  - Update Ω based on nearest same-class pairs in the representation space

MatchDG

$$f_{\texttt{randommatch}} = \arg\min_{h,\Phi} \sum_{d=1}^{m} L_d(h(\Phi(X)), Y) + \lambda \sum_{\Omega_Y(j,k)=1; d \neq d'} \operatorname{dist}(\Phi(\mathbf{x}_j^{(d)}), \Phi(\mathbf{x}_k^{(d')})) \quad (2)$$

$$l(\mathbf{x}_j, \mathbf{x}_k) = -\log \frac{\exp\left(\operatorname{sim}(\Phi(\mathbf{x}_j), \Phi(\mathbf{x}_k))/\tau\right)}{\exp\left(\operatorname{sim}(\Phi(\mathbf{x}_j), \Phi(\mathbf{x}_k))/\tau\right) + \sum_{i=0, y_i \neq y_j}^{b} \exp\left(\operatorname{sim}(\Phi(\mathbf{x}_j), \Phi(\mathbf{x}_i))/\tau\right)} \quad (3)$$

Algorithm 1: MatchDG

Input: Dataset $(d_i, x_i, y_i)_{i=1}^n$ from m domains, τ, t

Output: Function $f : \mathcal{X} \to \mathcal{Y}$

Create random match pairs Ω_Y.

Build a $n * m$ data matrix $\mathcal{M}$.

Phase I. **while** *notconverged* **do**

- **for** $batch \sim \mathcal{M}$ **do**
  - Minimize contrastive loss (3).
- **if** *epoch % t ==0* **then**
  - Update match pairs using Φ_{epoch}.

Phase 2. Compute matching based on Φ. Minimize the loss (2) to obtain f.

MNIST Dataset

Table 1: Accuracy for Rotated MNIST & Fashion-MNIST datasets on target domains of 0° and 90°. Accuracy for CSD [23], MASF [20], IRM [6] are reproduced from their code.

Dataset	Source	ERM	MASF	CSD	IRM	RandMatch	MatchDG	PerfMatch (Oracle)
Rotated MNIST	15, 30, 45, 60, 75	93.9 (0.67)	93.2 (0.2)	94.7 (0.2)	94.2 (0.32)	95.9 (0.18)	**96.1** (0.34)	97.5 (0.17)
	30, 45, 60	77.9 (2.44)	69.4 (1.32)	79.2 (2.47)	77.6 (1.68)	81.4 (0.77)	**86.3** (1.14)	92.0 (0.83)
	30, 45	64.6 (3.23)	60.8 (1.53)	68.7 (1.01)	63.1 (3.14)	68.4 (1.78)	**74.3** (2.47)	81.7 (2.79)
Rotated Fashion MNIST	15, 30, 45, 60, 75	78.6 (1.17)	72.4 (2.9)	78.0 (1.5)	79.6 (1.82)	79.4 (0.81)	**82.8** (0.27)	86.2 (0.69)
	30, 45, 60	33.7 (2.24)	25.7 (1.73)	37.2 (1.15)	35.5 (1.51)	38.8 (2.28)	**45.6** (1.74)	55.3 (1.54)
	30, 45	22.1 (2.36)	20.8 (1.26)	24.9 (1.78)	24.4 (1.01)	25.1 (1.89)	**34.9** (1.56)	41.4 (1.58)

Table 2: Overlap with perfect matches. top-10 overlap and the mean rank for perfect matches for MatchDG and ERM over all training domains. Lower is better for mean rank.

Dataset	Method	Overlap (%)	Top 10 Overlap (%)	Mean Rank
MNIST	ERM	18.9 (1.01)	52.4 (1.91)	25.1 (1.43)
	MatchDG (Default)	**35.1** (5.23)	**69.6** (5.97)	**14.3** (4.16)
	MatchDG (PerfMatch)	47.6 (5.61)	81.5 (4.70)	8.2 (3.17)
Fashion MNIST	ERM	3.1 (0.20)	14.4 (0.68)	190.6 (7.92)
	MatchDG (Default)	**23.9** (2.61)	**50.1** (3.29)	**79.7** (9.91)
	MatchDG (PerfMatch)	54.7 (4.38)	82.5 (3.07)	15.5 (3.54)

Chest X Ray Dataset

Details: Link

- Source Domains (NIH, ChexPert)
  - Images with class label 0 are translated vertically downwards
- Target Domains (Kaggle)
  - No spurious correlation

	NIH (Source)	Chex (Source)	RSNA (Target)
ERM	78.9 (0.34)	**84.3** (3.52)	55.2 (2.27)
IRM	**79.1** (1.01)	83.4 (2.42)	56.6 (2.04)
CSD	73.2 (3.35)	83.3 (2.03)	60.5 (0.82)
RandMatch	75.3 (1.87)	83.6 (1.84)	57.4 (1.76)
MatchDG	74.7 (0.66)	82.2 (0.68)	58.4 (0.62)
MDGHybrid	74.3 (0.91)	82.4 (1.03)	**62.6** (0.72)

Evaluation Issues with DG

- OOD accuracy evaluated on few test domains (PACS, VLCS)
  - No guarantees regarding performance on a large set of unseen domains
  - Evaluation metrics to capture the extent to which DG algorithm learnt stable features

- Membership Inference (MI) Attacks
  - Utilize overfitting of ML models to predict train vs test dataset samples
  - Stable features → Better Generalization → Good Defense against MI attacks

- Connections between DG and Privacy Attacks
  - Theoretical: Causal models (stable feature learning) leads to better defense on MI attacks
  - Empirical: Use MI attacks to evaluate DG algorithms
  - Software: Toolkit to support DG algorithms and evaluate them on various privacy attacks

DG and Trustworthy Explanations

- Counterfactual Explanations (CF) for models under distribution shifts
  - CF generated might not generalize across distributions (Rawal et al.)

- DG models could lead to more stable explanations under distribution shifts ?
  - Distributionally robust ML models might lead to CF that generalize well across distributions
  - Regularizers inspired from DG literature could help in generating robust CF

- Evaluating DG models using counterfactual explanations?
  - Metrics based on stability of counterfactual explanations across distributions to evaluate DG models