Domain Generalization using Causal Matching

Divyat Mahajan 1 Shrutti Tople 1 Amit Sharma 1
1Microsoft Research, India 2Microsoft Research, UK

Causal View of Domain Generalization

- **Aim:** Learn a single classifier (f) with training data (X, Y) sampled from m domains that generalizes well to data from unseen domains/distributions
- **Assumption:** There exist stable (causal) features (X) whose relationship with outcome Y, P(Y|X), is invariant across domains
- **Notation:**
  - Representation network: Φ : X → C
  - Classification network: h : C → Y
  - Ideal solutions h∗, Φ∗ = arg min,Ω,φ E[d,x,y]|(y, h(φ(x)))] s.t. xφ = Φ∗(x) and f∗ = h∗(x)
- **Contributions:**
  - Identify conditions for failure of class-conditional invariance objective [1, 2]
  - Propose object-invariant condition for domain generalization, along with a novel approach to satisfy it in practical scenarios

Why Class-Conditional Domain Invariance Fails?

- **Domain as intervention:** For each observed x1, there are a set of counterfactual inputs x̄ where d ≠ d̄, but both have similar causal features xφ
- **Object-Invariant Condition:** X1 ∥ ∥ D/X1
  - Empirical: E[(x1),j|x1]d(∥d∥, Φ(φ(x1))) = 0, i.e. if d ≠ d̄, Ω = 0 otherwise

Perfect Match: Proposed approach for known true objects

- **Goal:** Learn a match function such that f(x, x′) = 1 when Dist(x, x′) is small
- **Assumption:** Same-class inputs are closer in true causal representation than different-class inputs
- **Simple Baseline:** Use contrastive loss to learn a representation under which same-class inputs become closer than different-class inputs
- **Our approach:** Contrastive Learning with iterative updates to positive matches to help in capturing intra-class variance across domains

MatchDG: Proposed approach for unknown true objects

- **Loss:** Empirical Risk Minimization Loss + λΩ (Object-Invariant Constraint)
- **Intuition:** Match counterfactuals (same base object pairs) instead of same class pairs to account for intra-class variability

MatchDG works even under the zero training error regime

- **Zero training error does not necessarily imply similar representations for each class, resulting in ERM unable to satisfy MatchDG penalty**
- **Methods based on comparing variation in loss across domains, like IRM [3], will be affected under zero training error**

Results: OOD accuracy on DG benchmarks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Overlap (%)</th>
<th>Top 10 Overlap (%)</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rot MNIST (5)</td>
<td>ERM</td>
<td>58.9</td>
<td>60.6</td>
<td>14.0</td>
</tr>
<tr>
<td>Rot MNIST (3)</td>
<td>MatchDG</td>
<td>60.4</td>
<td>60.6</td>
<td>14.0</td>
</tr>
<tr>
<td>Fashion MNIST (5)</td>
<td>ERM</td>
<td>69.5</td>
<td>70.0</td>
<td>16.4</td>
</tr>
<tr>
<td>Fashion MNIST (3)</td>
<td>MatchDG</td>
<td>69.5</td>
<td>70.0</td>
<td>16.4</td>
</tr>
</tbody>
</table>

Figure: Out of Domain Accuracy (OOD) Results: Brackets denote number of source domains for Rotated & Fashion MNIST

- MatchDG, Perfect Match obtain SOTA accuracy and improvement over baseline is highlighted in the case of fewer source domains
- MatchDG obtains comparable performance to SOTA approaches on more realistic benchmarks like PACS (per paper for more details)

Does MatchDG learn the causal features?

- MatchDG learns the causal features?

Figure: Results for quality of match function using following metrics: Overlap of top-1 match with the true object match, Overlap of top-10 matches with the true object match, Mean rank of the true object match in the learnt representation (lower is better)

References

