Machine Learning from Explanations

NUS Computing
National University of Singapore

Data Privacy and Trustworthy Machine Learning Research Lab

Jiashu Tao, Reza Shokri

ICML International Conference On Machine Learning

Actionable Interpretability

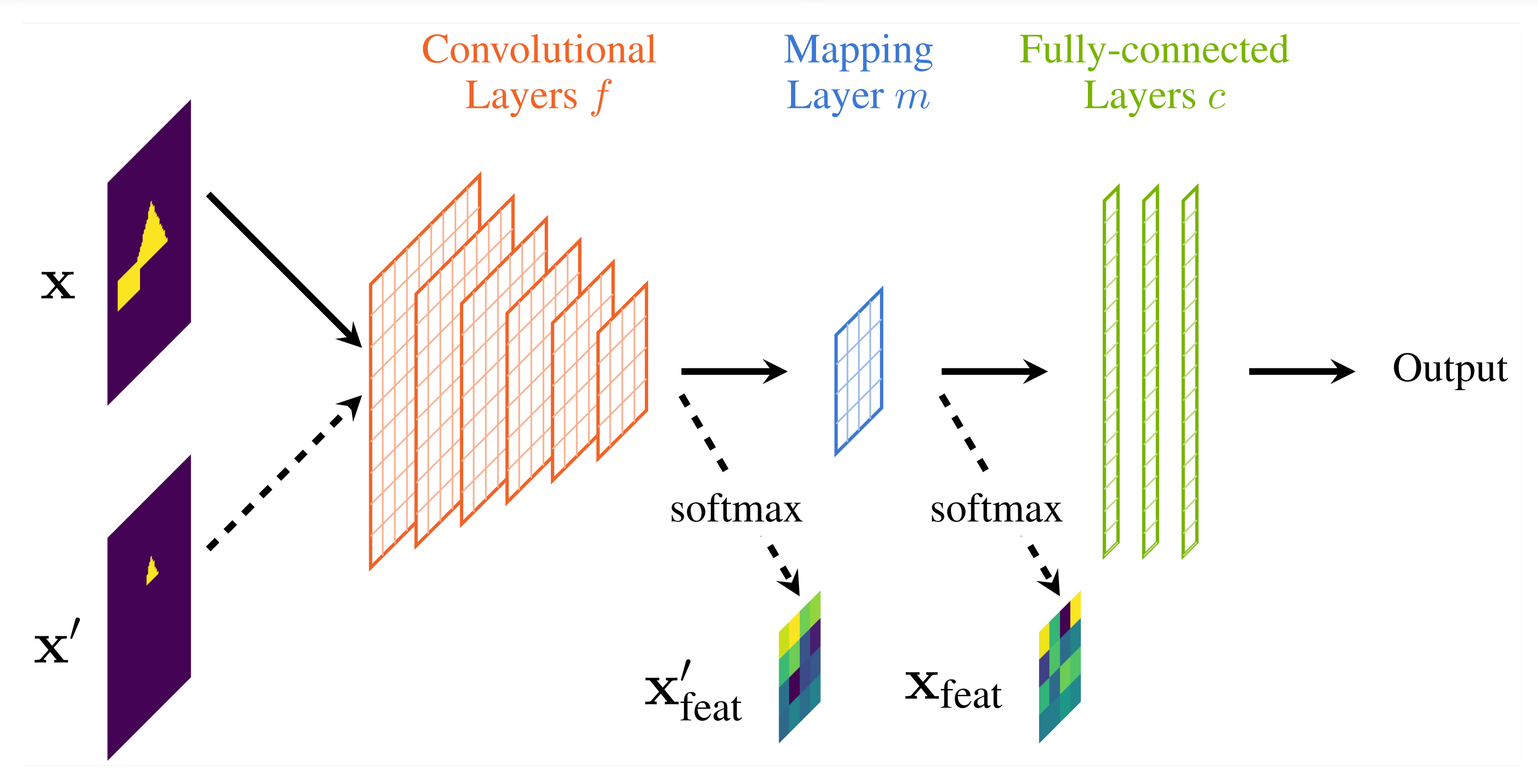
Motivation

- Training on dataset that are imbalanced or not sufficiently large tend to lead to **unstable** and **overfitted** models that rely on **spurious correlations**.
- Standard training methods rely on output label agreement, ignoring **why** models makes decisions, leading to untrustworthy models.

Key Ideas

- Curate (expert) explanations on a subset of training data that explain the reasons.
- Aligning model's latent features with the given explanation masks via KL divergence.
- Alternating the optimization of the cross-entropy loss and the KL divergence in a two-stage optimization scheme to ensure both label and reasoning agreement.

Training ML Models from Explanations



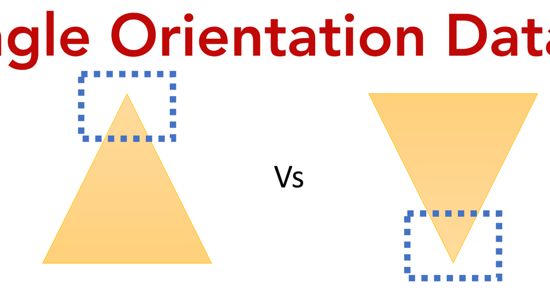
Algorithm 1 Two-stage optimization

Require: Input data $\mathbf{x}$, model $h = c \circ m \circ f$ consists of feature extractor f, mapping layer m, and fully connected layers c, target y, explanation $e(\mathbf{x})$, learning rates η_1 and η_2 for cross entropy loss and feature map loss

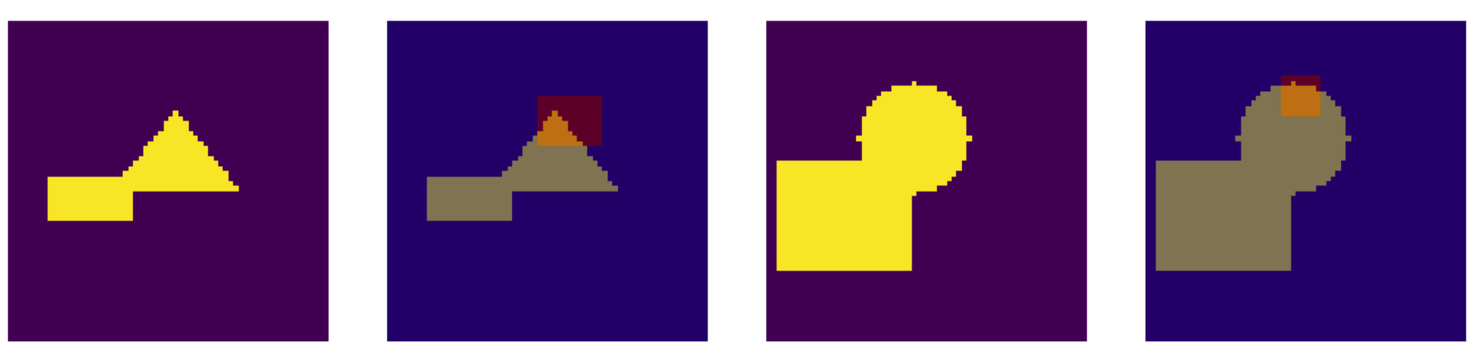
1: $\mathcal{L}_{\text{CE}} \leftarrow -y\log(h(\mathbf{x})) - (1-y)\log(1-h(\mathbf{x}))$
2: $\theta_h \leftarrow \theta_h - \eta_1 \nabla_{\theta_h} \mathcal{L}_{\text{CE}}$
3: $\mathbf{x}' \leftarrow \mathbf{x} \otimes e(\mathbf{x})$
4: $\mathbf{x}'_{\text{feat}} \leftarrow \text{softmax}(f(\mathbf{x}'))$
5: $\mathbf{x}_{\text{feat}} \leftarrow \text{softmax}(m(f(\mathbf{x})))$
6: $\mathcal{L}_{\text{feat}} \leftarrow KL(\mathbf{x}'_{\text{feat}} \,\|\, \mathbf{x}_{\text{feat}})$
7: $\theta_m \leftarrow \theta_m - \eta_2 \nabla_{\theta_m} \mathcal{L}_{\text{feat}}$

Datasets with Explanations

Triangle Orientation Datasets



Fox vs Cat

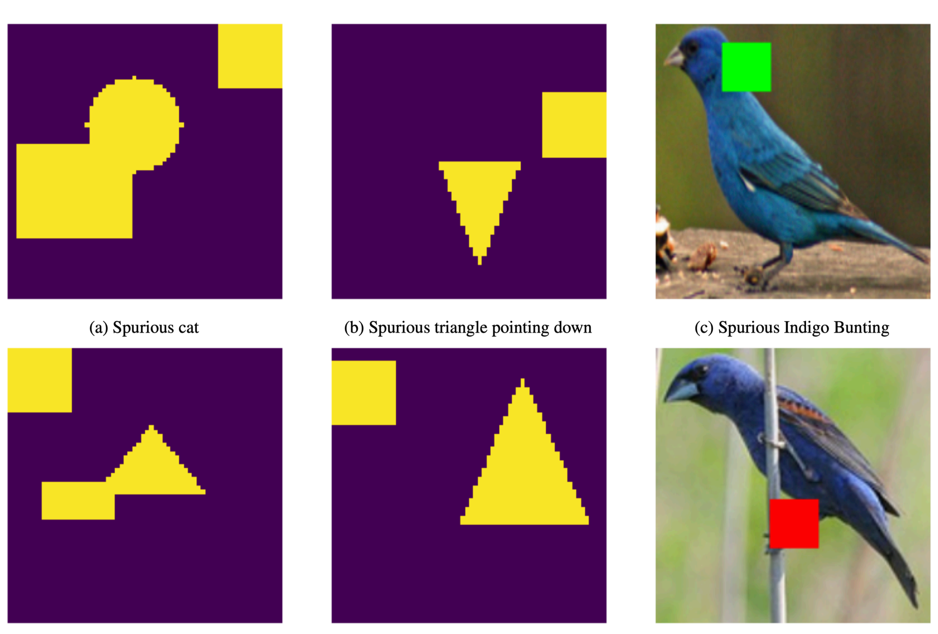


(a) Fox (b) Fox with a mask highlighting the vertex of its triangular head (c) Cat (d) Cat with a mask highlighting the arc of its round head

CUB-200 Bird

(a) An Indigo Bunting (b) An Indigo Bunting with an explanation mask on its beak (c) A Blue Grosbeak (d) A Blue Grosbeak with an explanation mask on its beak

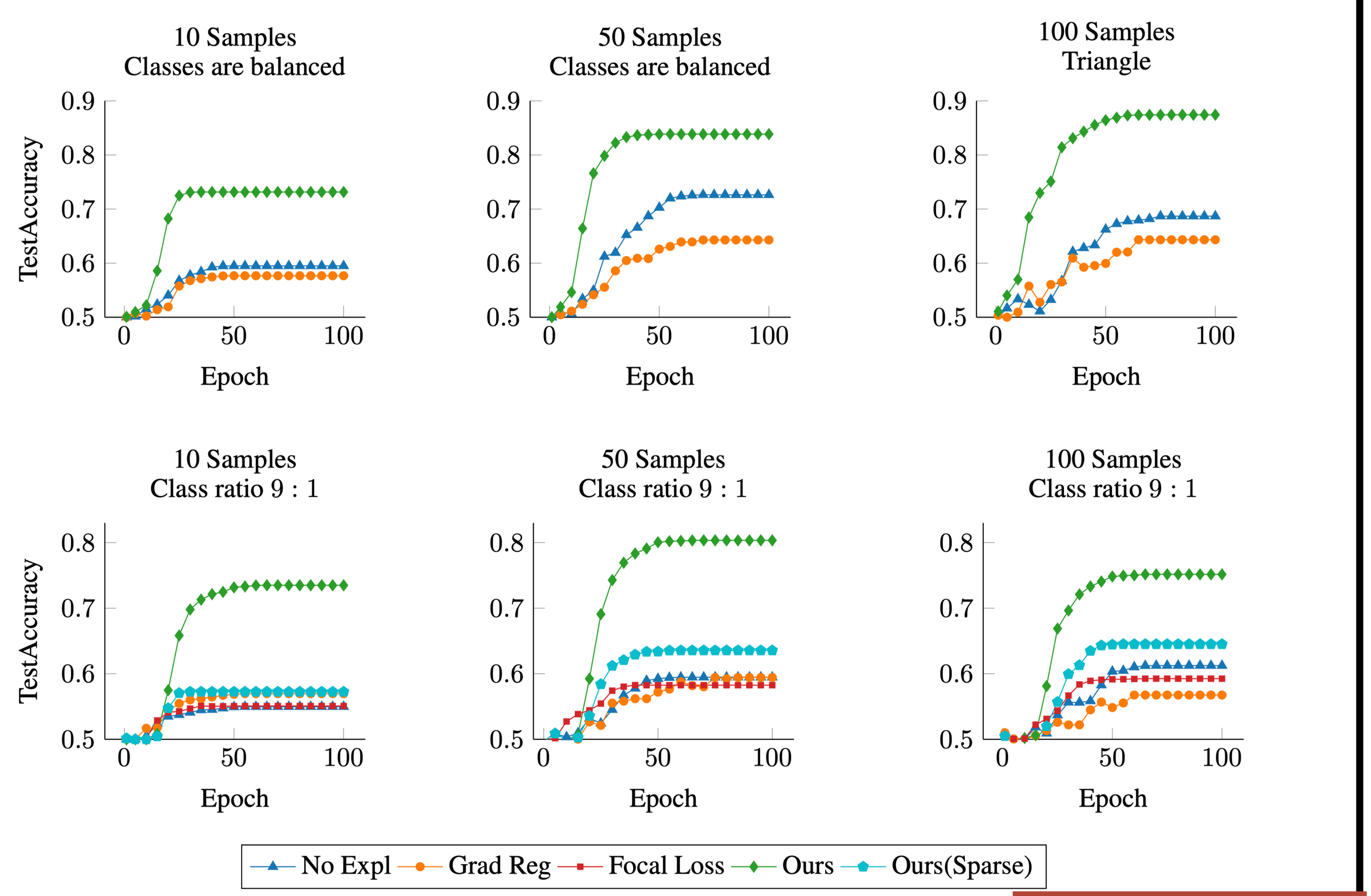
Injecting Spurious Correlations



(a) Spurious cat (b) Spurious triangle pointing down (c) Spurious Indigo Bunting (d) Spurious fox (e) Spurious triangle pointing up (f) Spurious Blue Grosbeak

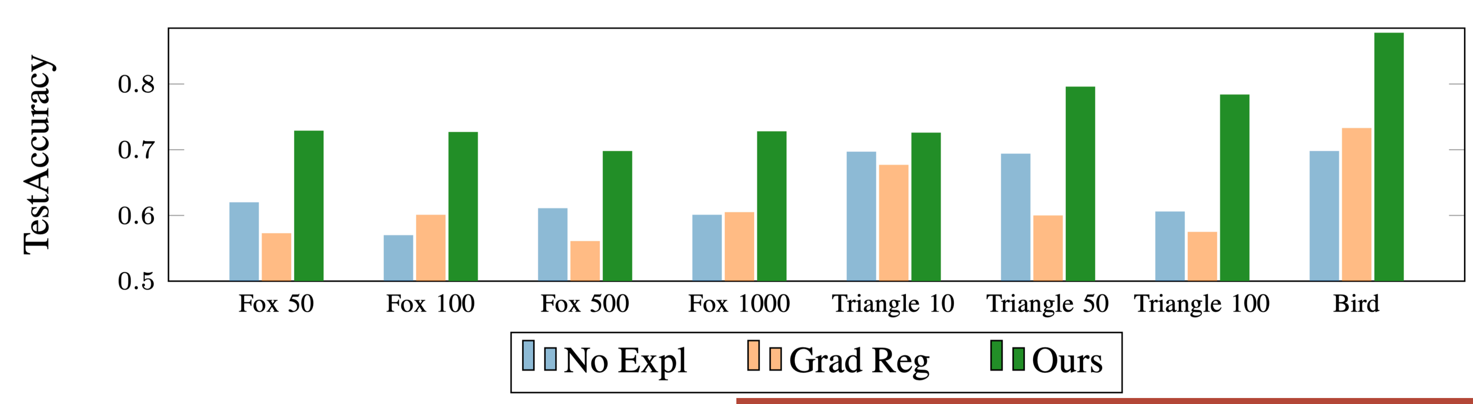
Adding spurious patches/features to training data only

Learning from Explanations Makes Models Learn Faster and Better …



Even with 10% data with explanations

… and More Robust to Spurious Correlations



This further proves the models trained in our proposed way learns the given rule from explanations