TREATING SPURIOUS CORRELATIONS WITH ENAMOR: ENFORCING NUISANCE ATTRIBUTES TO BE MITIGATED ON THE REPRESENTATIONS

Akira Sakai * **Taro Sunagawa** * **Spandan Madan** †§ **Kanata Suzuki** * **Takashi Katoh** *

Hiromichi Kobashi * **Xavier Boix** ‡§ **Tomotake Sasaki** *§

ABSTRACT

Spurious correlations between the object category and a nuisance attribute or factor in the training distribution cause dramatic degradation of the generalization performance of Deep Neural Networks (DNNs). Such spurious correlations are induced by the data collection process which is often skewed towards objects with certain nuisance factor, such as specific viewpoints and illumination conditions. Recent works have shown that DNNs are more robust to spurious correlations when invariant representations to nuisance factors emerge in the DNN's intermediate layers. Following this observation, in this paper we propose a novel regularization technique, ENAMOR, to enforce the emergence of invariant representations with the aim to alleviate the effects of spurious correlations. We report results in four datasets with spurious correlations related to position, viewpoint and illumination conditions, namely the MNIST-Positions dataset, the iLab-Orientations dataset and two novel datasets (planed to be made available at dataset.labs.fujitsu.com) of 3D rendered cars and images of objects taken with a robotic arm in controlled illumination conditions. These datasets allow to study in detail the effects of spurious correlations in DNNs and are challenging as DNNs perform poorly. Our results demonstrate that ENAMOR allows for substantial gains of generalization performance.

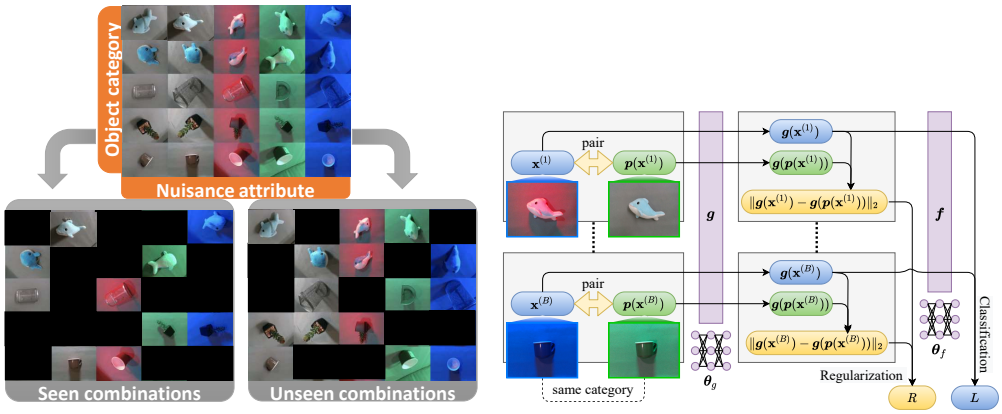

Figure 1: Problem setting (left) : We split the combinations of object categories and nuisance attributes into two partitions, namely, seen combinations (training dataset) and unseen combinations (testing dataset). Regularization term in ENAMOR (right) : Pairs of images $\mathbf{x}$ and $\boldsymbol{p}(\mathbf{x})$ that belong to the same object category are fed into the network. The Euclidean distance between the pairs of the last ReLU activity $\|\boldsymbol{g}(\mathbf{x};\boldsymbol{\theta}_g) - \boldsymbol{g}(\boldsymbol{p}(\mathbf{x});\boldsymbol{\theta}_g)\|_2$ is used as the regularization term.

*Fujitsu Laboratories Ltd., Kawasaki, Japan. {akira.sakai, sunagawa.taro, suzuki.kanata, kato.takashi_01, h_kobashi, tomotake.sasaki}@fujitsu.com (They are currently with Fujitsu Limited.)

†Harvard University, Cambridge, USA. spandan_madan@g.harvard.edu

‡Massachusetts Institute of Technology, Cambridge, USA. xboix@mit.edu

§Center for Brains, Minds and Machines, Cambridge, USA.