High-Rate Mixout: Revisiting Mixout for Robust Domain Generalization

Masih Aminbeidokhti, Heitor Rapela Medeiros
Srikanth Muralidharan, Eric Granger, Marco Pedersoli

Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
École de technologie supérieure (ETS), Montreal

DISTECH CONTROLS™

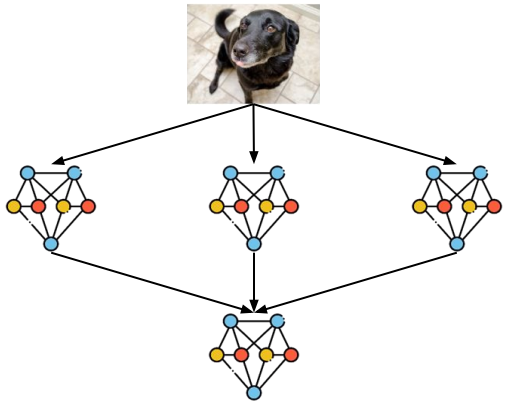
LIVIA LABORATOIRE D'IMAGERIE, DE VISION ET D'INTELLIGENCE ARTIFICIELLE

ÉTS Le génie pour l'industrie — ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Université du Québec

The Domain Generalization Bottleneck

Domain Generalization (DG) requires models to train on multiple source domains and deploy on unseen targets.

The Bottleneck: Ensembling and weight averaging boost DG, but usually require training and storing many models.



Reimagining Mixout for DG

Within each iteration **Mixout** samples a binary mask ξ ~ *Ber(p)* and swaps fine-tuned weights with pre-trained weights to preserve prior knowledge.

$$\theta^{\xi} = \theta_0 \odot (1 - \xi) + \theta \odot \xi.$$

Key Findings: DG actually prefers high swap rates.

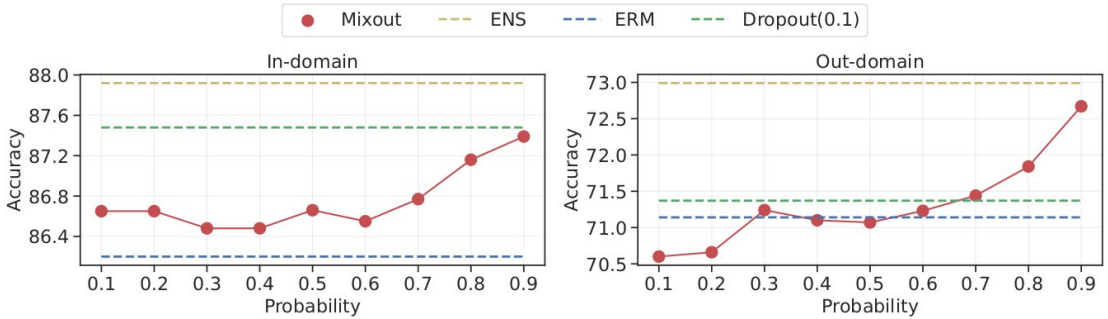
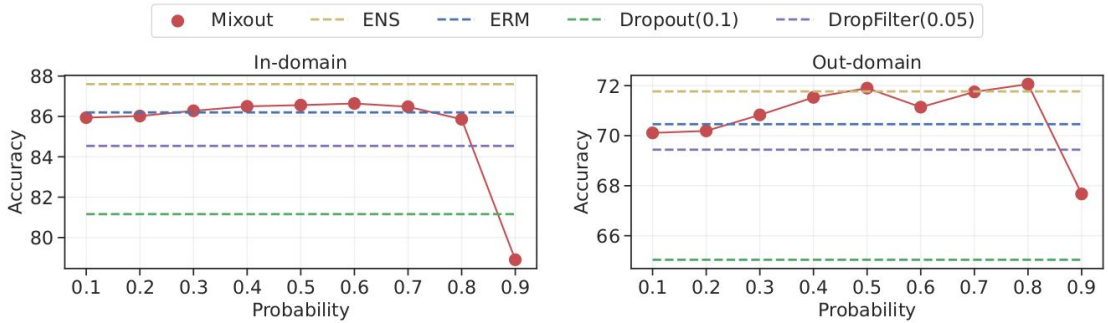


Figure: In/Out domain accuracy of Mixout and Dropout on OfficeHome dataset with ViT-S/16. Mixout sustains OOD accuracy at high rates compared to dropout..

Architecture-Specific Masking (CNNs vs. ViTs)

Excessive unstructured swapping can lead to information leakage and the inadvertent deactivation of significant neural network components.

Solution: Swap coherent structures (like whole kernels/filters) to respect spatial correlations.



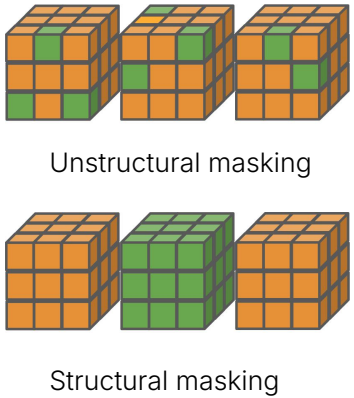


Figure: Unlike ViT-S/16, Mixout with structured masking is better for both the in and out domain performance.

The "Ensemble Effect" & Efficient Inference

Mixout Training averages over many subnetworks (the "Ensemble Effect").

Single-pass inference: We use a deterministic weight-scaling approximation instead of costly Monte-Carlo sampling.

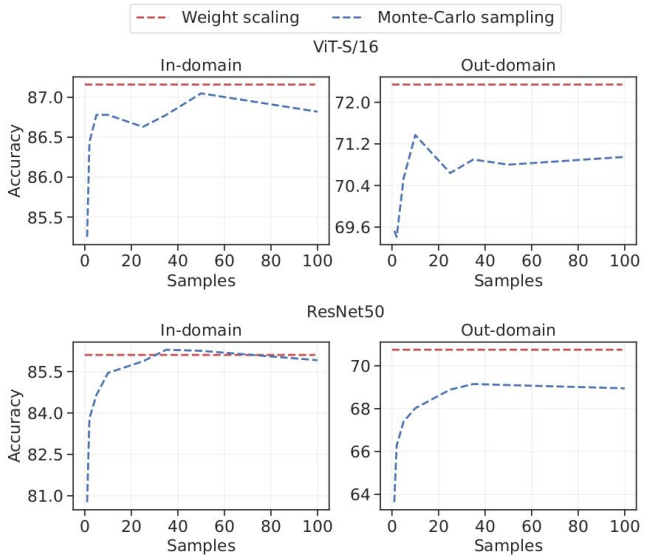


Figure: In/Out-domain accuracy of MC sampling and weight-scaling on OfficeHome.

Key Results: Maximum Accuracy, Minimum Cost

High-rate Mixout achieves ensemble-like robustness with a single training run and single-pass inference.

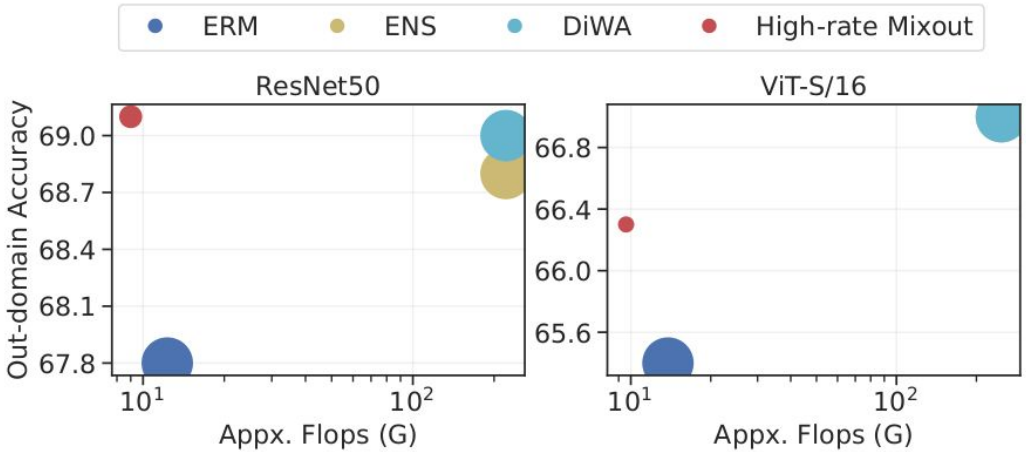


Figure: Performance versus computational (log-scale) and memory cost for the backward pass across different architectures and methods. Computation overhead relative to the ERM baseline

masih.aminbeidokhti.1@ens.etsmtl.ca

Thank You !