When Do Deep Ensembles Improve Robustness to Spurious Correlations?

CSE3000 Research Project • Controlled background-shortcut benchmark (MNIST-on-CIFAR)

Jaouad Hidayat

Delft University of Technology (EEMCS) • Supervisor: Wendelin Böhmer • Examiner: David Tax

Contact: J.Hidayat@student.tudelft.nl • J.W.Bohmer@tudelft.nl

Code / PDF

Key message: On an OOD test that **keeps backgrounds but breaks label associations**, ensembles help most when the background shortcut is **moderately strong**.

1 Why this matters

- **Shortcut learning:** models can learn an easy cue that breaks under shift.
- **Here:** during training, the **background alone predicts the digit label**.
- **Baseline:** deep ensembles (no labels for the shortcut are needed).

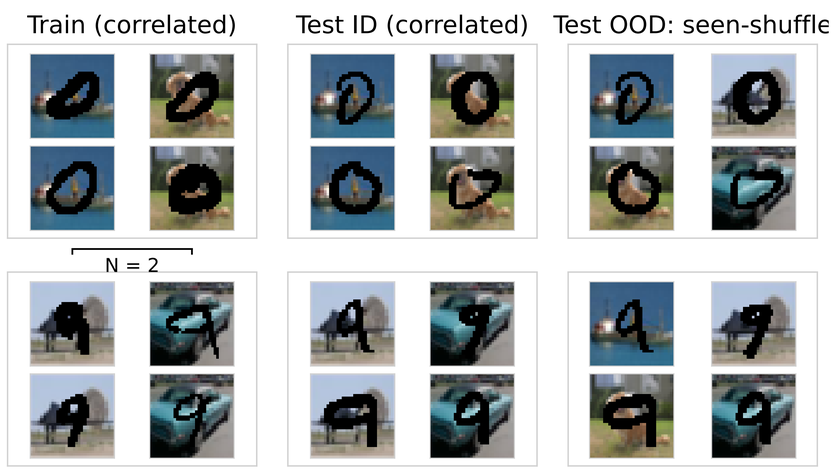
RQ: When does increasing ensemble size *M* improve OOD robustness with a **background shortcut**, and does it reduce **background-following**?

Defs: *N* = backgrounds per digit; *M* = ensemble size; ID/OOD = in/out-of-distribution.

Takeaway: We test when ensembles help under a background shortcut, and we measure shortcut following directly.

2 Benchmark and evaluation

Fig. 1: Dataset and test splits

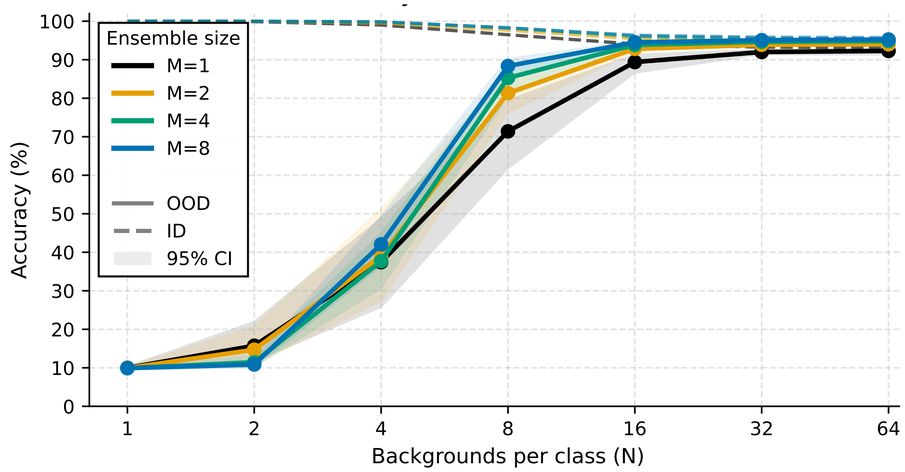


- MNIST digit placed on a CIFAR-10 background.
- Each digit class gets its own set of *N* backgrounds.
- Larger *N* makes the background cue harder to memorize.
- OOD tests:
  - **Seen-shuffle:** same backgrounds, **wrong label mapping** (tests shortcut following).
  - **Unseen-bg: new** backgrounds (shortcut removed + background shift).
- Train *M* models with different seeds; average their probabilities.

Takeaway: **Seen-shuffle** keeps the same backgrounds but breaks label mapping, so it isolates shortcut following.

3 Main result

Fig. 2: OOD accuracy vs. background variety *N* (seen-shuffle)



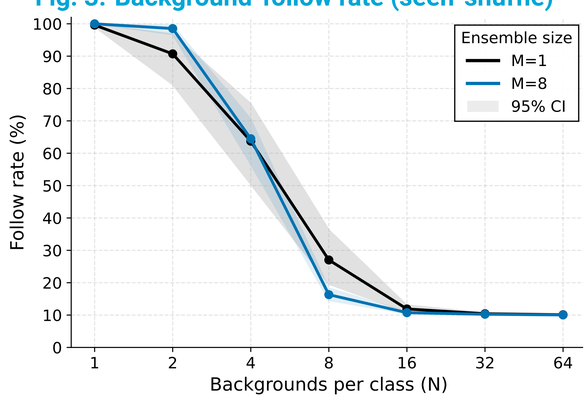
- Model: LeNet; loss: mean-squared error (MSE); $M \in \{1, 2, 4, 8\}$.
- Example ($N = 8$): 71.3% → 88.4% (**+17.1 percentage points**; 95% CI 8.4–26.8).
- Gains are smaller when shortcut learning is trivial (small *N*) or complex (large *N*).

Takeaway: Ensembling helps most when the background shortcut is not overwhelming, but not already weak.

4 Why it works: background-follow rate (BFR)

- In seen-shuffle, each image's background comes from one label's background set.
- **BFR:** fraction of predictions that match the spuriously correlated label (after averaging).
- At $N = 8$: BFR drops **26.9% → 16.3%** ($M = 1 \rightarrow 8$).
- At small *N*: BFR is near **100%** (strong background following).

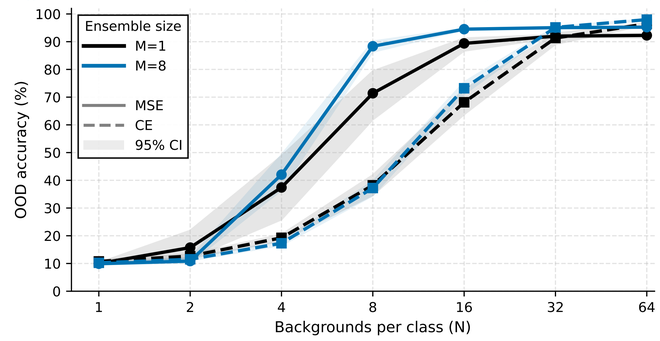
Fig. 3: Background-follow rate (seen-shuffle)



Takeaway: Where accuracy improves, the ensemble follows the background less.

5 What changes the result

Fig. 4: Loss ablation (MSE vs cross-entropy, seen-shuffle)



- The training loss changes the robustness pattern.
- At $N = 8$, MSE beats cross-entropy by 33.1 pp ($M = 1$) and 51.1 pp ($M = 8$).
- Unseen-background also changes the background distribution, so it tests more than "shortcut removal."

Takeaway: Gains are relative to the loss and type of shift.

6 Takeaways and limits

Conclusion:

- In this benchmark, the largest OOD gains occur in an **intermediate** shortcut setting (e.g., $N = 8$).
- In that setting, BFR drops (less background-following).

Limitations:

- Synthetic benchmark with a deterministic shortcut (controlled, simplified).
- Unseen-background mixes shortcut removal with a new background distribution.
- Results depend on loss, model capacity, and training choices.

Takeaway: Ensembling can help, but not in every shortcut regime.