Saliency-based Analysis of Shortcut Learning in CNNs

Rohma Anwar rohmaanwar49@gmail.com · ID 1000084894 **Abid Ali** mabid0445@gmail.com · ID 1000083978
Advanced Deep Learning Module · University of Catania (UNICT)

1. Introduction

Deep convolutional neural networks often reach high average accuracy by exploiting *spurious correlations* — superficial cues that co-occur with the label in the training data but are not causally related to it. This behaviour, known as *shortcut learning*, is dangerous because it is invisible to the standard test-set accuracy yet causes large failures whenever the correlation between the cue and the label is broken, i.e. under distribution shift. The central challenge is that a single aggregate accuracy number neither reveals whether a model has taken a shortcut nor identifies which cue it relies on.

We study this problem on *Waterbirds*, a benchmark in which the background (land or water) is deliberately correlated with the bird type (landbird or waterbird) during training but decorrelated at test time. We fine-tune an ImageNet-pretrained ResNet-18 to classify landbird vs. waterbird and then diagnose its behaviour with three complementary tools. First, we measure *per-subgroup* and *worst-group* accuracy, which expose the gap that the aggregate number hides. Second, we compute *Grad-CAM* saliency maps together with a lightweight *foreground/background attention-bias score* that quantifies how much of the model's attention falls off the bird. Third, and most importantly, we run a battery of *inference-time interventions* that edit the image (blurring, masking, or shuffling the background, and masking the foreground) and measure the causal effect on the prediction.

The key finding is unambiguous: the model reaches 83.9% overall accuracy but only **59.5%** on its worst subgroup (a waterbird on land); its saliency consistently leaks onto the background; and, decisively, masking the foreground collapses accuracy to **53.4%** while masking the background *raises* accuracy to **86.0%**. Together these show the network has partially internalised the shortcut "background → bird type." The main contribution of this work is not a new architecture but a simple, reproducible diagnostic recipe — subgroup metrics, a saliency-based bias score, and counterfactual interventions — that turns an opaque accuracy number into direct evidence of shortcut reliance.

2. Model Description

Architecture. The classifier is a *ResNet-18* (He et al., 2016). The input is a 224×224×3 image. A stem applies a 7×7 convolution (stride 2, 64 filters) followed by batch normalisation, a ReLU non-linearity, and 3×3 max-pooling (stride 2). Four residual stages follow, each containing two *basic blocks*; a basic block is two 3×3 convolutions with batch normalisation and ReLU, plus an identity (or 1×1 projection) skip connection that adds the block input to its output. The stages use 64, 128, 256 and 512 channels and halve the spatial resolution at each transition. A global average-pooling layer produces a 512-dimensional feature vector, which the original 1000-way ImageNet head replaces with a new fully-connected layer $W \in \mathbb{R}^{512\times 2}$ mapping to the two classes (landbird, waterbird). The network has approximately 11.2M parameters.

Activations and regularisation. All hidden non-linearities are ReLU. Batch normalisation is used throughout (so no dropout is required); note that it behaves differently in training, where it uses batch statistics, and in evaluation, where it uses accumulated running statistics. Regularisation comes from *L2 weight decay* (10^{-4}), *data augmentation* (random horizontal flips at training time only), and *transfer learning* from ImageNet, which provides a strong feature prior on a relatively small dataset. The whole network is fine-tuned end-to-end; the backbone is not frozen.

Learning objective. The model outputs raw logits $z = f(x)$; probabilities are obtained with the softmax and the network is trained with the *cross-entropy* loss:

$$p_{i,c} = e^{z_{i,c}} / \sum_{c'} e^{z_{i,c'}}, \qquad \mathcal{L}(\theta) = -(1/N) \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log p_{i,c}.$$

Because cross-entropy applies the log-softmax internally, the network itself outputs *logits* (softmax is applied only at evaluation time to report probabilities). Optimisation uses *Adam* (Kingma & Ba, 2015) with a constant learning rate; the checkpoint is selected on validation *worst-group* accuracy rather than average accuracy.

Saliency and the bias score. To locate the model's attention we use *Grad-CAM* (Selvaraju et al., 2017) on the last convolutional block (`layer4[-1]`), targeting the *predicted* class c. Let A^k be the k-th feature map and y^c the class score. The importance weights and localisation map are

$$\alpha_k^c = (1/Z) \sum_i \sum_j \partial y^c / \partial A^k_{ij}, \quad L^c = \mathrm{ReLU}(\sum_k \alpha_k^c A^k).$$

We take the central 60% crop F as a coarse foreground proxy and define the *attention-bias score* as the fraction of the (non-negative) saliency mass that falls outside it:

$$\beta = (\sum_{u \notin F} L^c_u) / (\sum_u L^c_u).$$

A high β means the model attends to the background. **Interventions.** To move from correlation to causation we edit each test image and re-run the model under five conditions: *original*; *background blur* (Gaussian, kernel 31, outside F); *background mask* (background set to neutral grey 0.5); *background patch-shuffle* (16×16 background patches permuted); and *foreground mask* (the centre crop set to grey 0.5). For each edit we record the change in prediction and confidence relative to the original.

3. Dataset

- **Data source.** *Waterbirds* (Sagawa et al., 2020), obtained from the Hugging Face hub as `grodino/waterbirds`. It composites bird foregrounds from CUB-200-2011 (Wah et al., 2011) onto scene backgrounds from Places (Zhou et al., 2017).
- **Data size.** 4,795 training, 1,199 validation and 5,794 test images. There are two classes (landbird, waterbird) and a binary place attribute (land, water), giving four subgroups. In training the background is correlated with the label about 95% of the time (waterbirds mostly on water, landbirds mostly on land); the test split is group-balanced, with subgroup counts of 642 (waterbird-water), 642 (waterbird-land), 2,255 (landbird-land) and 2,255 (landbird-water).
- **Preprocessing.** Images are resized to 224×224 and normalised with the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225). Training applies random horizontal flipping; no augmentation is used at validation or test time.

4. Training details and procedure

4.1 Evaluation metrics

We report *overall accuracy*, *per-subgroup accuracy*, and *worst-group accuracy* (the minimum accuracy over the four subgroups), together with macro-averaged precision, recall and F1 and the confusion matrix. Macro averaging weights both classes equally regardless of their frequency. We additionally report the saliency-based bias score β and, for the intervention study, the *prediction-flip rate* (fraction of samples whose predicted class changes relative to the unedited image) and the mean *confidence drop*.

4.2 Training procedure

- **Epochs:** 15. **Batch size:** 32 (training); 64 (evaluation).
- **Validation strategy:** a fixed held-out validation split is evaluated after every epoch; both overall and worst-group accuracy are logged.
- **Stopping criteria:** no early stopping. The best checkpoint is the epoch with the highest validation worst-group accuracy (here epoch 4, 56.4%); the last epoch is also saved. Selecting on worst-group rather than average accuracy is deliberate — it is the honest measure of robustness under the train/test shift.

4.3 Training details

- **Optimizer:** Adam, learning rate 1×10^{-4}, weight decay 1×10^{-4}, default moments (β_1=0.9, β_2=0.999).
- **Hyperparameters:** constant learning rate (no schedule), batch size 32, 15 epochs; hyperparameters were fixed a priori rather than tuned, to keep the study reproducible.

- **Initialization:** the backbone is initialised from ImageNet-pretrained weights; the new two-unit classification head uses PyTorch's default (Kaiming-uniform) initialisation. The global random seed is fixed to 42.

5. Experimental Results

5.1 Training curves

Figure 1 shows the optimisation. Training loss falls close to zero and training accuracy saturates near 99%, while validation accuracy stays around 80–85%. Crucially, validation worst-group accuracy is highly unstable, swinging between 16% and 56% across epochs: the model fits the majority groups quickly but never reliably solves the conflict groups — the signature of a shortcut.

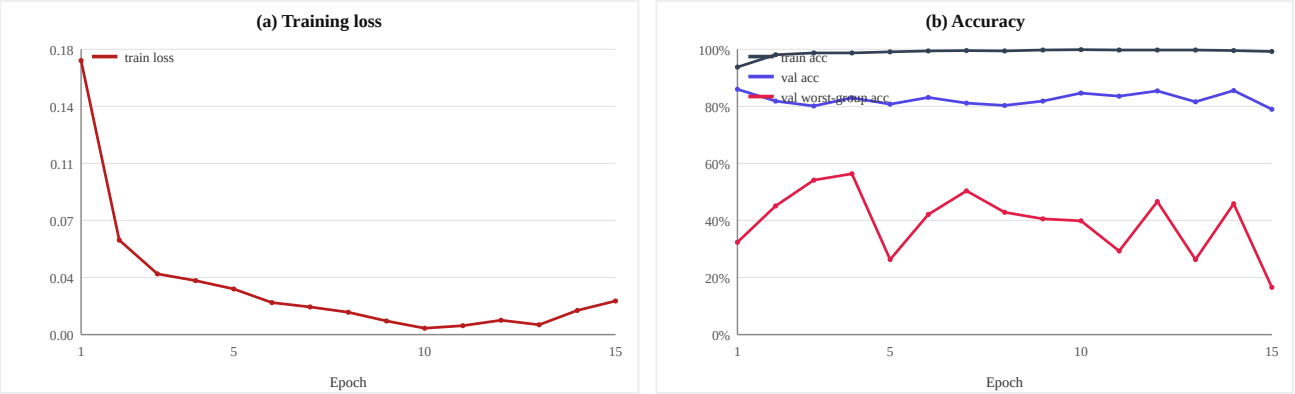

Figure 1: Training dynamics over 15 epochs. (a) Training loss. (b) Training, validation, and validation worst-group accuracy. The large, unstable gap between validation and worst-group accuracy is the shortcut signal.

5.2 Test performance

On the full 5,794-image test set the model reaches 83.9% overall accuracy (macro precision 76.9%, recall 81.2%, F1 78.6%) but only 59.5% worst-group accuracy. Table 1 shows the per-subgroup breakdown: the two *conflict* groups (a bird on the atypical background) are exactly where accuracy falls, while confidence stays high — the model is confidently wrong. Figure 2 gives the confusion matrix.

Subgroup	Count	Accuracy	Avg. confidence	Type
Landbird on land	2,255	98.6%	98.4%	majority
Waterbird on water	642	93.3%	96.9%	majority
Landbird on water	2,255	73.6%	87.6%	conflict
Waterbird on land	642	**59.5%**	90.1%	conflict (worst)

Table 1: Per-subgroup test performance. Accuracy collapses on the two conflict groups while confidence remains high.

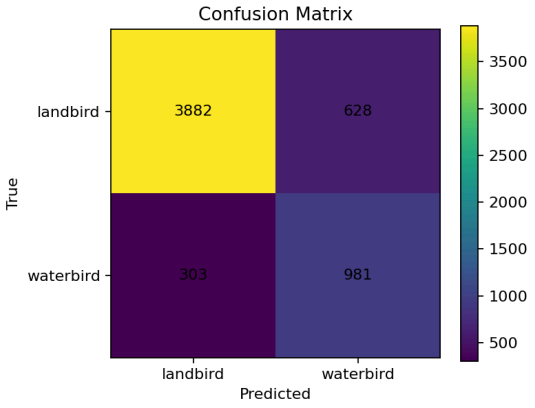

Figure 2: Test-set confusion matrix (rows: true class, columns: predicted class).

5.3 Saliency analysis

Averaged over a balanced test sample, the attention-bias score β ranges from 34.8% (waterbird-water) to 48.4% (landbird-land): a large share of the model's attention is off the bird even when it is correct. Figure 3 shows a representative failure — a waterbird photographed on land, predicted as a landbird, with the Grad-CAM heat concentrated on the land background rather than the bird.

Figure 3: Grad-CAM for a waterbird on land. Left: input. Centre: Grad-CAM for the predicted class. Right: the 60% centre-crop foreground box. The model predicts “landbird” and its attention sits on the land background rather than the bird.

5.4 Intervention experiments

Interventions provide the causal test. They were run on a balanced 1,000-image test subset (baseline “original” accuracy 80.2%). Table 2 and Figure 4 summarise the result. Masking the *foreground* collapses accuracy to 53.4% and flips 31.8% of predictions — removing the bird removes most of the signal. In contrast, masking the *background* raises accuracy to 86.0%, *above* the 80.2% baseline, because on the conflict groups the background actively pushes the wrong class; neutralising it removes that misleading cue. Blur and patch-shuffle, which only degrade the background, cause small changes. This asymmetry is direct causal evidence that the background is driving predictions.

Condition	Accuracy	Flip rate	Avg. background saliency β
Original	80.2%	—	41.0%
Background blur	83.5%	7.7%	33.0%
Background mask	**86.0%**	10.4%	27.8%
Background patch-shuffle	83.6%	12.0%	40.6%
Foreground mask	**53.4%**	31.8%	67.4%

Table 2: Model behaviour under each intervention (1,000-image test subset). Masking the foreground hurts most; masking the background helps.

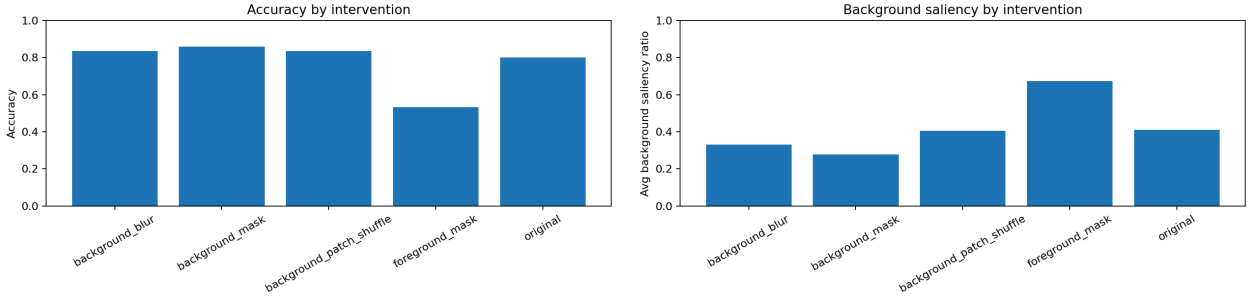

Figure 4: Left, test accuracy under each intervention; right, average background saliency under each intervention. Masking the foreground is the only edit that hurts; neutralising the background helps.

5.5 Comparison and discussion

Our model is a standard empirical-risk-minimisation (ERM) baseline; the study is diagnostic rather than a new training method. The observed pattern — high average accuracy with a large worst-group gap — matches the ERM behaviour reported by Sagawa et al. (2020), who show that group-robust training (Group DRO) raises worst-group accuracy on Waterbirds substantially (to roughly 86–90%). Relative to that ideal, our ERM model's 59.5% worst-

group accuracy quantifies the cost of the shortcut, and our interventions explain *why* it occurs. Among the intervention conditions, background masking is the strongest positive contrast (+5.8 points over the original), and foreground masking the strongest negative (−26.8 points), bracketing the model's true reliance on the two regions. The main limitation is that the 60% centre crop is a coarse foreground proxy rather than a true segmentation mask, and results come from a single ResNet-18 run; a segmentation-based foreground and multiple seeds would tighten the quantitative claims without changing the qualitative conclusion.

References

1. S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *ICLR*, 2020. arXiv:1911.08731.
2. R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
3. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *ICCV*, 2017.
4. K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CVPR*, 2016.
5. C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, Caltech, 2011.
6. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE TPAMI*, 40(6):1452–1464, 2017.
7. D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ICLR*, 2015.
8. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *CVPR*, 2009.