Decompose-and-Compose: A Compositional Approach to Mitigating Spurious Correlation

Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdanparast, Hamidreza Yaghoubi Araghi, Mahdieh Soleymani Baghshah

Sharif University of Technology

Spurious Correlation

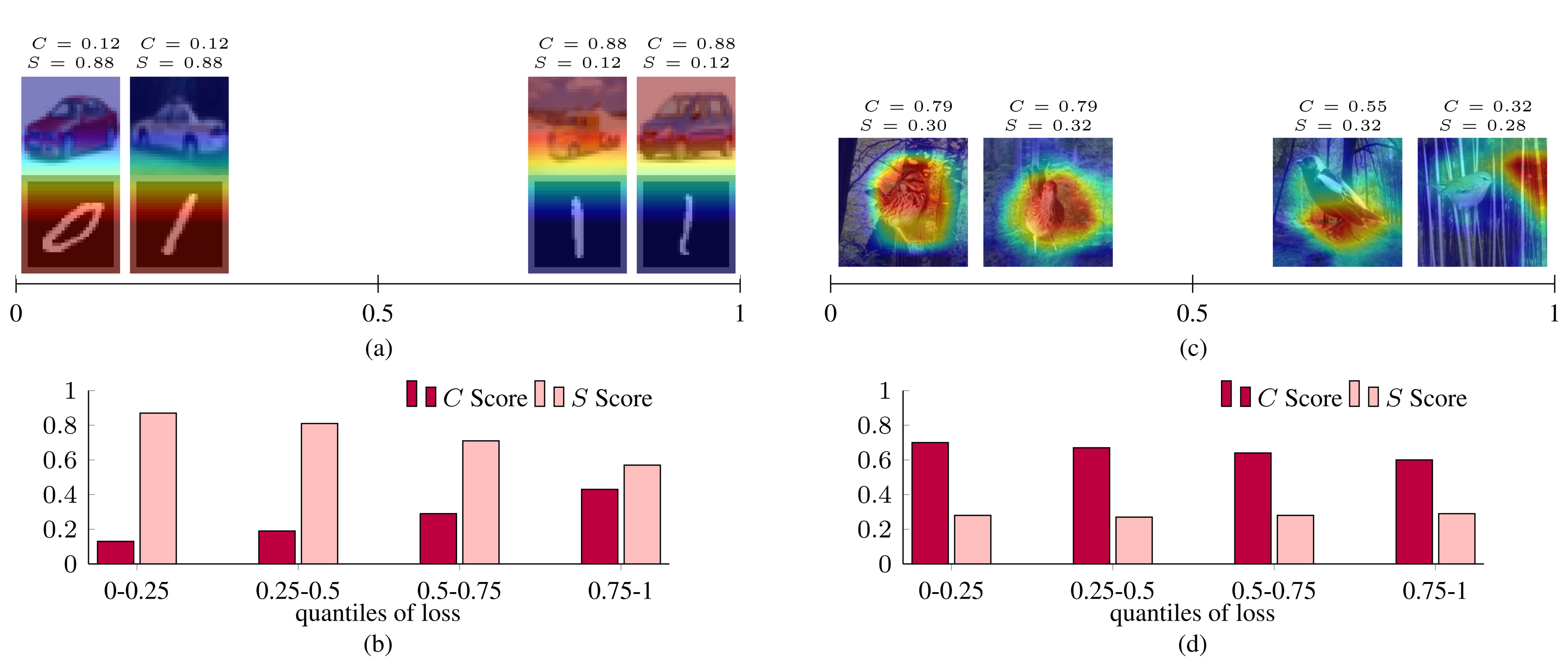
When training a deep model on a classification task, there is a chance that the model relies on spurious correlation between parts of images and the label[2].

However, the test samples may not come from the train data distribution!

When such correlation is abscent in the test data, the models accuracy drops. The Goal is to train a classifier that is robust to spurious correlation.

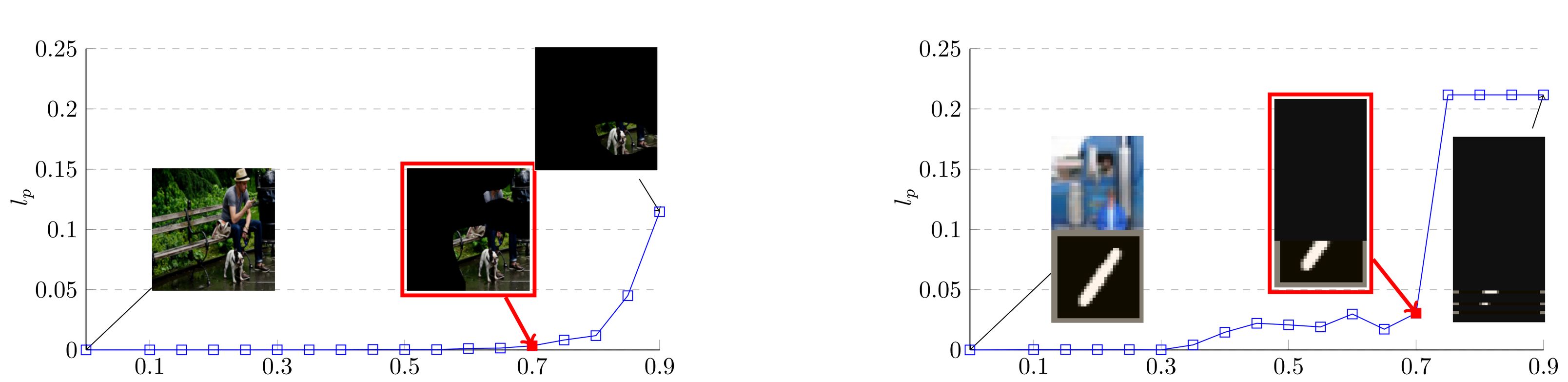
The Focal Regions of Models

Based on the easiness of inferring the label from the spurious parts, the models may attend more either to the spurious parts or the parts that are the real causes of the label.



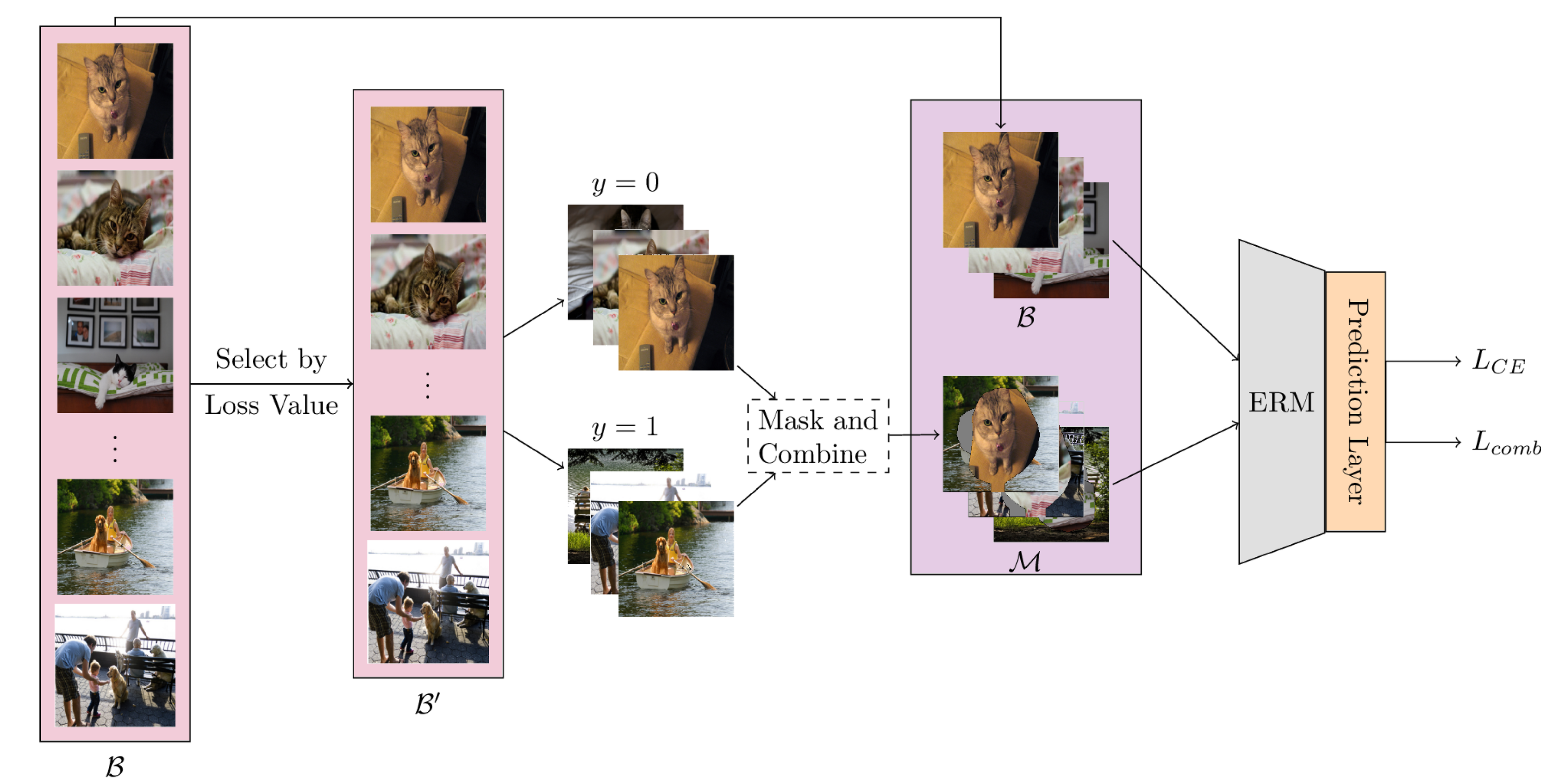
The amount of a model's attention on the C (core) and S (spurious) parts varies significantly among two different datasets exhibitting spurious correlation. This is more evident in samples on which the model has a low loss.

Adaptive Masking



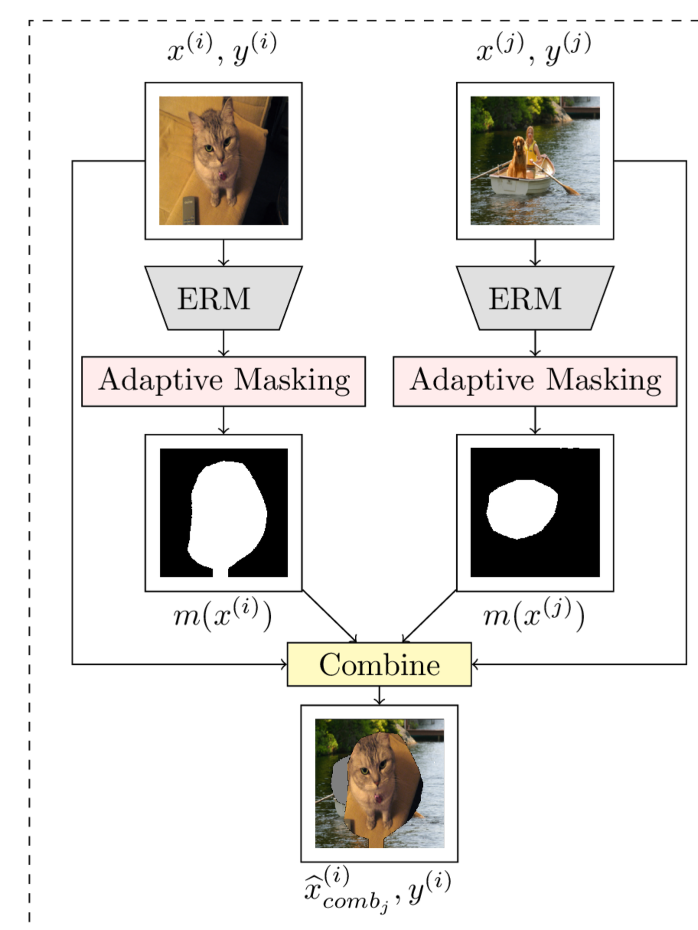
When the most predictive pixels are masked, the loss increases rapidly. Therefore, the optimal masking portion is just before this point!

Decompose and Compose



DaC

- Choose one of these two assumptions: The mask, or its inverse capture the core parts more.
  - The correct choice is determined by the worst validation group accuracy.
- Select low-loss images.
- Mask and Combine: Merge the core part of one image with the spurious part of another.
- Retrain Prediction Layer with the augmented data.



Mask and Combine

Justification

- By combining low-loss (majority) samples, we create samples representing minority groups.
- The augmented data is more group-balanced.
- retraining the last layer on the balanced augmented data makes the model more robust to spurious correlation.

Results

Method	Group Info	Waterbirds		CelebA		Metashift		Dominoes	
	train/val	Worst	Average	Worst	Average	Worst	Average	Worst	Average
DFR[3]	✗/✓✓	$92.3_{\pm 0.2}$	$93.3_{\pm 0.5}$	$88.3_{\pm 1.1}$	$91.3_{\pm 0.3}$	$72.8_{\pm 0.6}$	$77.5_{\pm 0.6}$	$\mathbf{90_{\pm 0.4}}$	$\mathbf{92.3_{\pm 0.2}}$
Group DRO[5]	✓/✓	$91.4_{\pm 1.1}$	$93.5_{\pm 0.3}$	$88.9_{\pm 2.3}$	$\mathbf{92.9_{\pm 0.2}}$	$66.0_{\pm 3.8}$	$73.6_{\pm 2.1}$	-	-
LISA[6]	✓/✓	$89.2_{\pm 0.6}$	$91.8_{\pm 0.3}$	$\mathbf{89.3_{\pm 1.1}}$	$92.4_{\pm 0.4}$	$59.8_{\pm 2.3}$	$70.0_{\pm 0.7}$	-	-
MaskTune[1]	✗/✗	$86.4_{\pm 1.9}$	$93.0_{\pm 0.7}$	79.4	89.5	$66.3_{\pm 6.3}$	$73.1_{\pm 2.2}$	$65.8_{\pm 4.7}$	$85.6_{\pm 0.7}$
CnC[7]	✗/✓	$88.5_{\pm 0.3}$	$90.9_{\pm 0.1}$	$88.8_{\pm 0.9}$	$89.9_{\pm 0.5}$	-	-	-	-
JTT[4]	✗/✓	86.7	93.3	81.1	88.0	$64.6_{\pm 2.3}$	$74.4_{\pm 0.6}$	-	-
Base (ERM)	✗/✗	$70.8_{\pm 0.5}$	$91.6_{\pm 0.1}$	41.7	96.0	$61.3_{\pm 3.4}$	$73.9_{\pm 1.5}$	$72.8_{\pm 1.6}$	$88.5_{\pm 0.3}$
DaC-C	✗/✓	$\mathbf{92.6_{\pm 0.2}}$	$94.9_{\pm 0.2}$	$76.11_{\pm 0}$	$91.35_{\pm 0.2}$	$76.0_{\pm 0.8}$	$80.0_{\pm 1.4}$	$89.0_{\pm 0.7}$	$92.2_{\pm 0.2}$
DaC	✗/✓	$92.3_{\pm 0.4}$	$\mathbf{95.3_{\pm 0.4}}$	$81.9_{\pm 0.7}$	$91.4_{\pm 1.1}$	$\mathbf{78.3_{\pm 1.6}}$	$\mathbf{79.3_{\pm 0.1}}$	$89.2_{\pm 0.1}$	$92.2_{\pm 0.3}$

Algorithm

Algorithm: Decompose-and-Compose (DaC)

Input: Model $f_\theta(.) = w \circ g_\phi(.)$; Dataset $\mathcal{D}_{tr}$; Loss function $l(.,.)$; Hyperparameters α, q, causalflag

for *epoch*$=1, 2, \dots K$ **do**
  for *batch* $\mathcal{B}$ *in* $\mathcal{D}_{tr}$ **do**
    $b \leftarrow mean(\mathcal{B})$
    $\mathcal{B}' \leftarrow q$ portion of samples in $\mathcal{B}$ with the lowest loss
    $\mathcal{M} \leftarrow \{\}$
    for *each image* $(x, y) \in \mathcal{B}'$ **do**
      Pick $(x', y') \in \mathcal{B}'$ s.t. $y \neq y'$
      $m \leftarrow \text{AdaptiveMasking}(f, x, y, l)$
      $m' \leftarrow \text{AdaptiveMasking}(f, x', y', l)$
      if *casualflag=False* **then**
        $m \leftarrow 1 - m$
        $m' \leftarrow 1 - m'$
      end
      $\hat{x}_{comb} = m \odot x$
      $+(1-m) \odot (1-m')x' + (1-m) \odot m'b$
      $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\hat{x}_{comb}, y)\}$
    end
    $L_{CE} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} l(f_\theta(x), y)$
    $L_{comb} \leftarrow \frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} l(f_\theta(x), y)$
    $L_{total} \leftarrow L_{CE} + \alpha L_{comb}$
    $w \leftarrow \text{UpdateWeights}(L_{total})$
  end
end

References

[1] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*, 2022.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. Springer-Verlag, 2018.

[3] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

[4] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 2021.

[5] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[6] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, 2022.

[7] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, 2022.

Email: fhosseini@ce.sharif.edu