Impact of Dissimilarity Loss on Out of Distribution Generalization

Author
Alexandru Cristian Cazacu
a.c.cazacu@student.tudelft.nl

Supervisor & Responsible Professor
Dr. Wendelin Böhmer
j.w.bohmer@tudelft.nl

1 - Introduction

- Deep learning has made neural networks ubiquitous in all kinds of applications.
- Neural networks are black boxes, they tend to rely on **“spurious correlations” or “shortcut features”**, superficial patterns that are predictive in the training set but not causally related to the task [1].
- This leads to failing to generalize reliably on **out of distribution (OOD)** data [1][2].
- Recent research suggests that manipulation of embeddings in latent space provides promising results [3].
- The aim of this project is to define a **“dissimilarity loss”** between samples that exhibit the same shortcut features to force models to learn the causal features of the data and see if it improves OOD generalization.

2 - Research Question

Does the addition of dissimilarity loss (DL) improve OOD generalization performance?

RQ1 - How effective is adding dissimilarity loss compared to a standard baseline network trained with cross-entropy loss only?
RQ2 - What are the optimal parameters (weight, offset, etc.) for dissimilarity loss on the given dataset?
RQ3 - Can DL be applied without the labels for the spurious features?

3 - Methodology

Dataset Generation

- To encourage shortcut learning, MNIST digits are superimposed on CIFAR background images.
- Each label uses a disjunct set of background images, with set size dictated by parameter N.
- In distribution test data and out of distribution data are separated (backgrounds unseen during training).

Model

- CNN based on LeNet-5.
- Adam optimizer, cross-entropy loss, 10 epochs

4 - Proposed Dissimilarity Loss

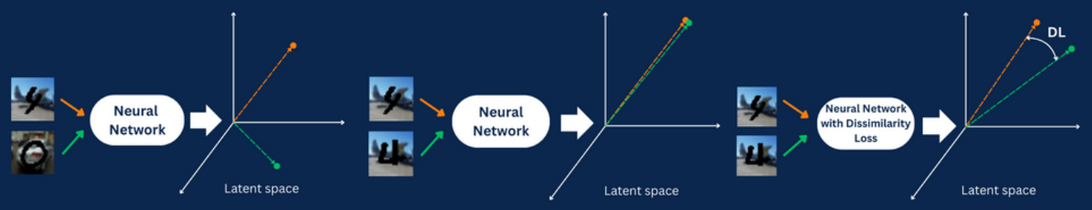
$$\mathcal{L}_{\text{dissim}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} S_{ij} \quad \text{, where}$$

$\mathbf{z}_i \in \mathbb{R}^d$ are penultimate layer embeddings

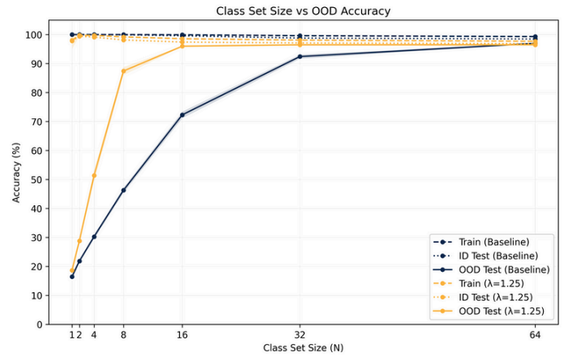
$$\hat{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} \qquad S_{ij} = \hat{\mathbf{z}}_i^\top \hat{\mathbf{z}}_j$$

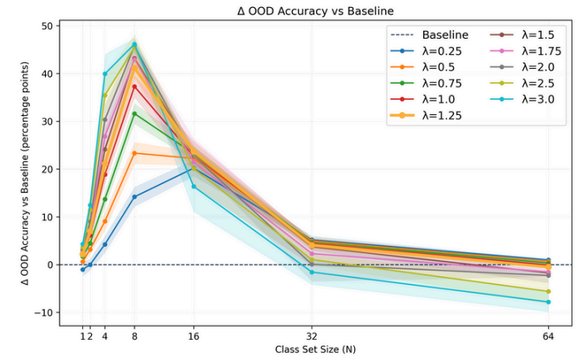
$$\mathcal{M} = \{(i,j) \mid i \neq j,\ label_i = label_j,\ bg_i = bg_j\}$$

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{dissim}}$$

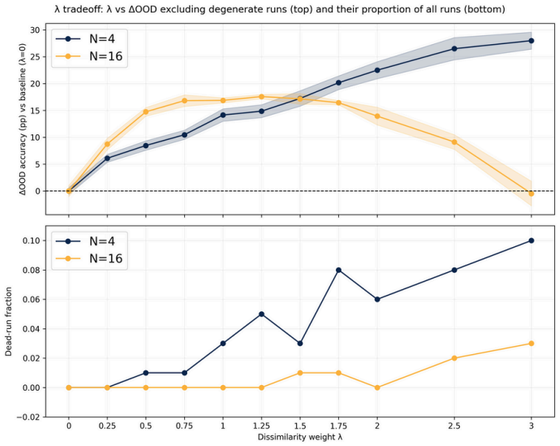


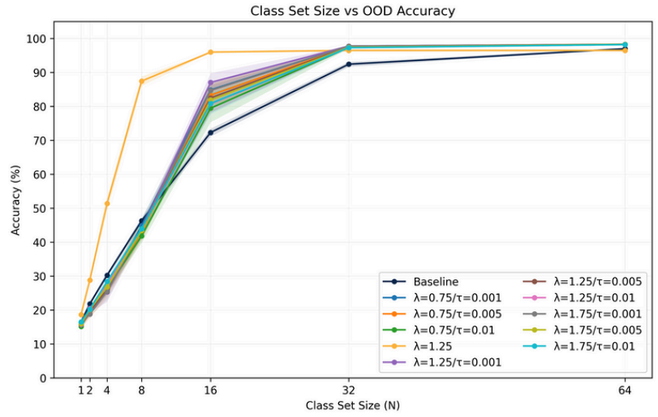
5 - Results





- Dissimilarity Loss provides a noticeable gain when N=4, N=8 and N=16.
  - At this N, spurious features are heavily exploited by the model and dissimilarity loss forces the samples’ embeddings apart encouraging the model to learn more of the causal features of the digits.
- At small N, the background is a too strong predictor of the label and dissimilarity loss provides close to no impact.
- At N >= 32, the backgrounds become varied enough such that they are no longer a strong enough indicator of the class, “shortcuts” becoming less and less reliable.
  - The additional loss introduced becomes noise and makes classification harder, not easier.
- Setting the weight of dissimilarity loss at λ = 1.25 is a good middle-ground for this task. Below 1.25, the effects become smaller, while increasing it further yields diminishing returns and decreases stability.





- We define a training run as “dead”, when it gets trapped in local minima, its ID test accuracy being < 20%.
- As λ increases, so does the proportion of the runs that gets stuck in local minima. Notice the increase in confidence intervals for OOD accuracy even with “dead” runs excluded (top plot), highlighting the decrease in stability with bigger λ.
- Lambda is a double-edged sword: small to moderate values improve robustness by discouraging shortcut learning, while large values destabilize training by overwhelming the classification signal.

- To make DL more applicable in real world scenarios we tried relaxing the same-background (spurious) constraint.
- In order to target the same set of backgrounds, we considered picking a similarity threshold (τ), such that pairs that exceed 1 - τ would get penalized, without disturbing class structure.
- This threshold is dependent on N (the strength of the shortcut), and no global, optimal option could be discerned.
- However, we still explored a few (λ, τ) pairs; (1.25, 0.001) yielded the best results, but remained inferior to the “normal” dissimilarity loss.

6 - Conclusions

- Models trained on data where shortcut features are easily available struggle to generalize on out of distribution data.
- Dissimilarity Loss counteracts spurious correlations and forces models to separate labels based on the intrinsic, causal features of the data.
- No “free lunch”:
  - In the absence of spurious features, dissimilarity loss adds noise, and makes classification harder, not easier.
  - Setting the weight too high decreases ID and OOD test accuracy.
  - DL still performs without spurious labels, but does not achieve the same results as with their inclusion.

7 - Future Work

- Further explore the hyperparameter space, other training regimes could interact with DL in non-trivial ways.
- Test dissimilarity loss on more complex datasets, such as FashionMNIST. or other types of models, such as AlexNet.
- Instead of a fixed λ, set a desired level of dissimilarity and add an optimizer which minimizes the difference between observed dissimilarity and the desired level.
- Experiment with activation functions and setting a minimum offset.
- Evaluate performance compared with other OOD generalization techniques.

References

[1] Geirhos, R. et al. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665–673.
[2] Ye, W. et al. (2025). The Clever Hans Mirage. arXiv:2402.12715.
[3] Hansen, N., Wang, X. (2021). Generalization in reinforcement learning by soft data augmentation. IEEE International Conference on Robotics and Automation (ICRA), pages 13611–13617, 2021. doi: 10.1109/

TU Delft