Evaluating the Robustness of Deep Learning Models

Jeff Wirojwatanakul, Zian Wang, Yizheng Yu
{pwirojwatanakul,ziw080,yiy003}@ucsd.edu

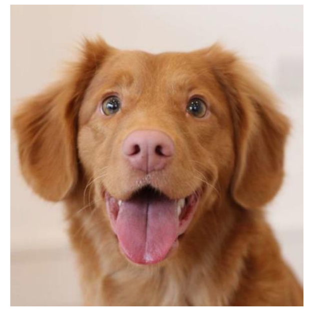
UC San Diego

Abstract

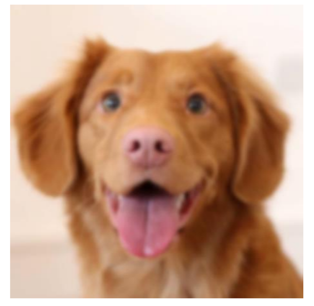
- We showed that **data augmentations** and **pretraining,** in both supervised and self-supervised ways, can increase the performance of DL models on traditional metrics as well as increase model robustness on synthetic covariate shifts.

Introduction

- The supervised machine learning setting assumes the **training** and **test data** are drawn from the **same** distribution. This often does not hold when deploying DL models. In such scenarios, the model performance may drop significantly, and is particularly worrisome for decision making systems. The main goal of this project is to investigate how DL models, **ResNet**, perform under covariate shifts, where $P_{TR}(x) \neq P_{TE}(x)$, and to find ways to improve the robustness to such shifts. To generate the covariate shift, we adjusted the brightness and performed Gaussian blur on the test images. Finally, we investigated whether the predictive probability of ResNet is representative of the likelihood to be correct.



Clean



Blurred

Darkened

Datasets

We used the pathMNIST dataset and Oxford-IIT Pets dataset. The first contains photos **of Colon Pathology** and the second contains photos of **pets**.

References

1. Hendrycks et al. Benchmarking neural network robustness to common corruptions and perturbations, 2019
2. Zhang et al. mixup: Beyond empirical risk minimization, 2017
3. Bardes et al. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2021

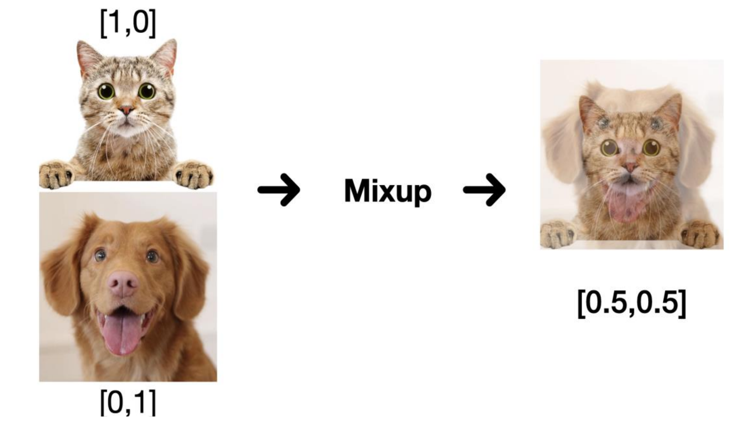
Methods

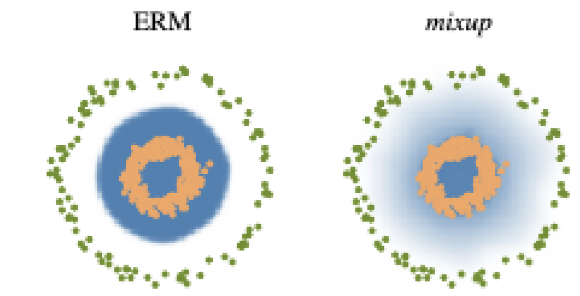
To Improve **Model Robustness**, we experimented with:

1. **Data Augmentations**: MixUp, Random Rotations, and Random Flip
2. **Pretraining on ImageNet:**
   - Supervised
   - Self Supervised: Barlow Twins

We **evaluated** our models using**: Accuracy** and **RMSE Calibration Error**

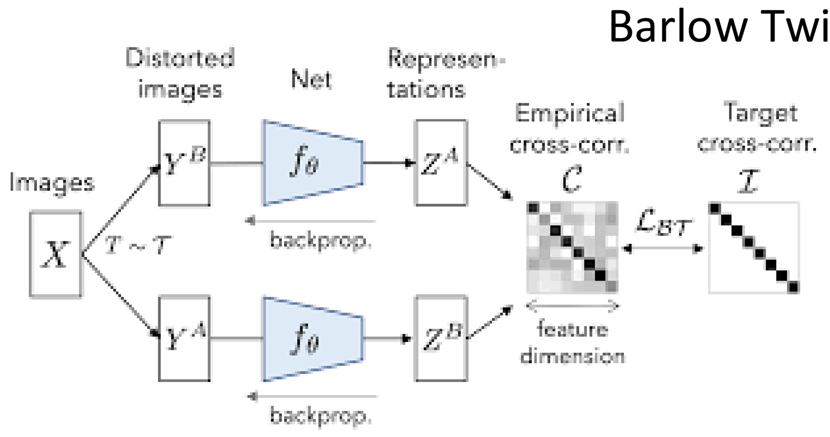
- RMSE Calibration Error: $\sqrt{\sum_{i=1}^{n} b_i\,(p_i - c_i)^2}$, where b_i is the fraction of data in bin i, p_i is the accuracy of bin i, and c_i is the confidence for bin i.





Blue indicates $p(y = 1|x)$, green/orange dots are data points

Barlow Twins



$$L_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_{j \neq i} {C_{ij}}^2$$

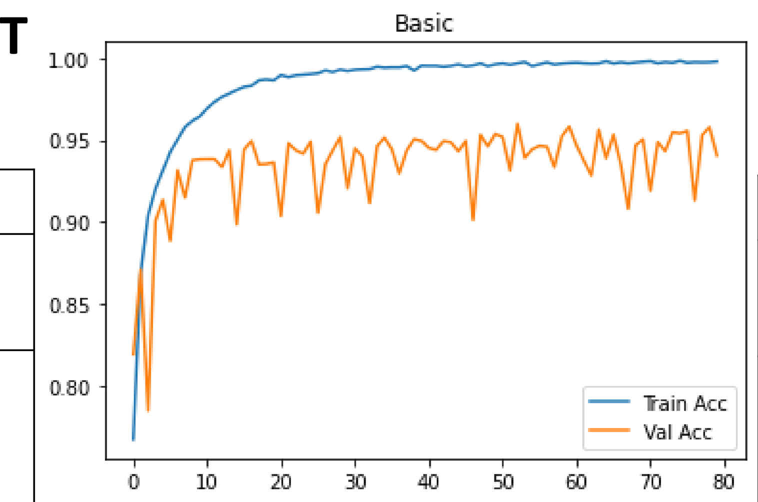
C is **the cross-correlation matrix** computed given two outputs of the neural network

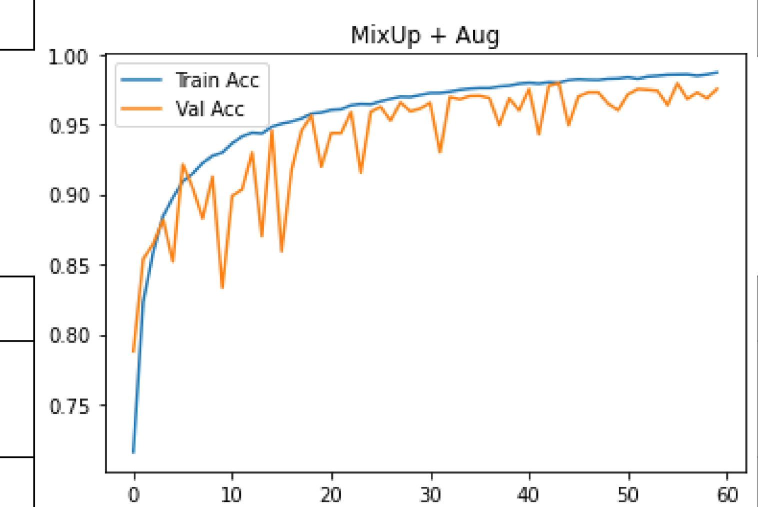
Results

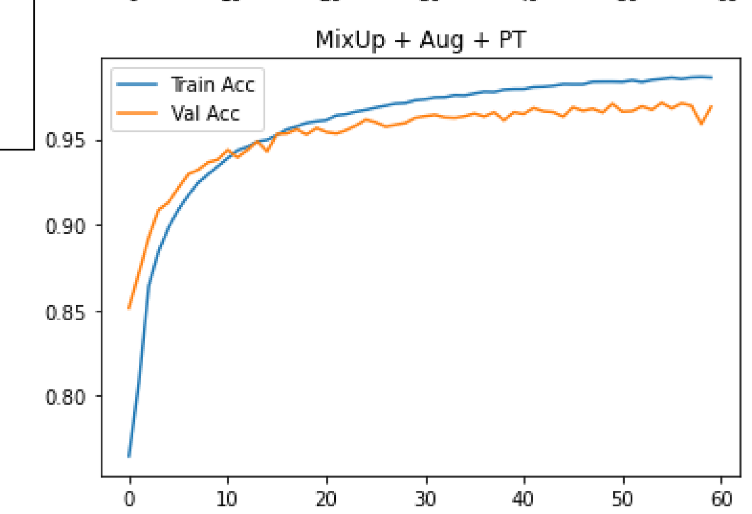
pathMNIST

Accuracy (%)			
Model	Clean	Brightness	Gaussian Blur
Basic	82.3	36.4	67.5
Mixup + Aug	85.9	48.8	68.1
Mixup + Aug + PT	**87.6**	**50.3**	**74.3**

RMSE			
Model	Clean	Brightness	Gaussian Blur
Basic	0.145	0.603	0.278
Mixup + Aug	0.040	0.302	0.168
Mixup + Aug + PT	**0.033**	**0.299**	**0.111**



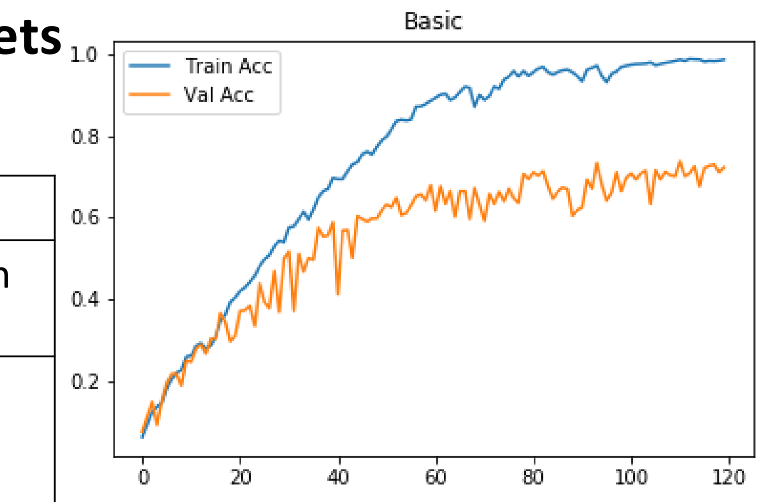


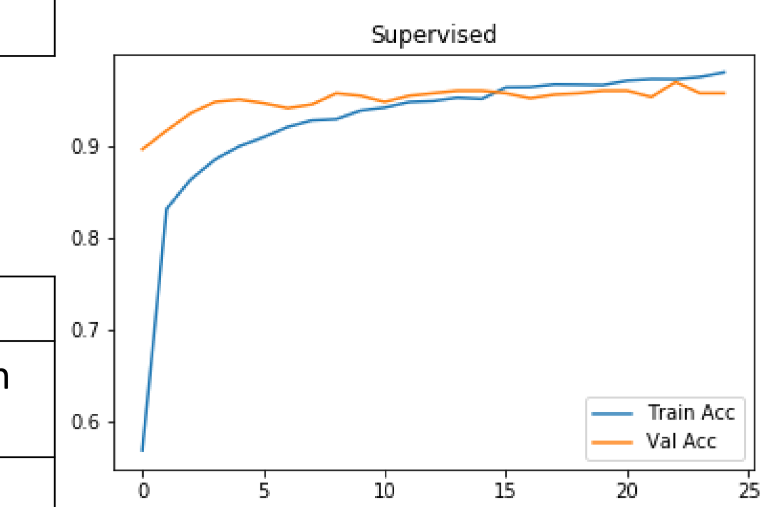


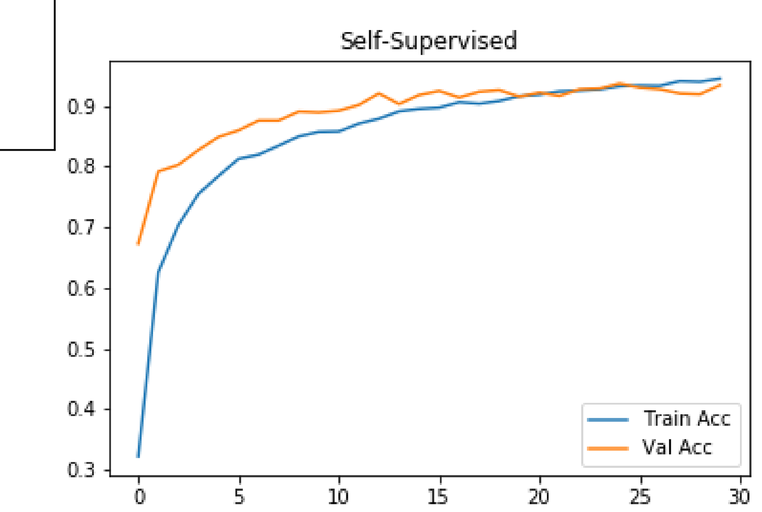
Oxford-IIT Pets

Accuracy (%)			
Model	Clean	Brightness	Gaussian Blur
No Pretraining	73.7	67.5	66.4
Supervised	**96.9**	**96.0**	**95.0**
Self-Supervised	93.6	92.4	88.4

RMSE			
Model	Clean	Brightness	Gaussian Blur
No Pretraining	**0.055**	**0.050**	**0.060**
Supervised	0.155	0.149	0.158
Self-Supervised	0.218	0.207	0.195







Conclusion

Key Takeaways:

- Pretraining and data-augmentation help increase **accuracy** on both the **clean** and **perturbed** test set.
- Both **supervised** and **self-supervised** pretraining increases the accuracy, but not model calibration.
- DL models are **more sensitive** to perturbations in the medical image domain.

Future Work: Test the model on adversarial attacks and try out newer augmentation techniques like AugMix.