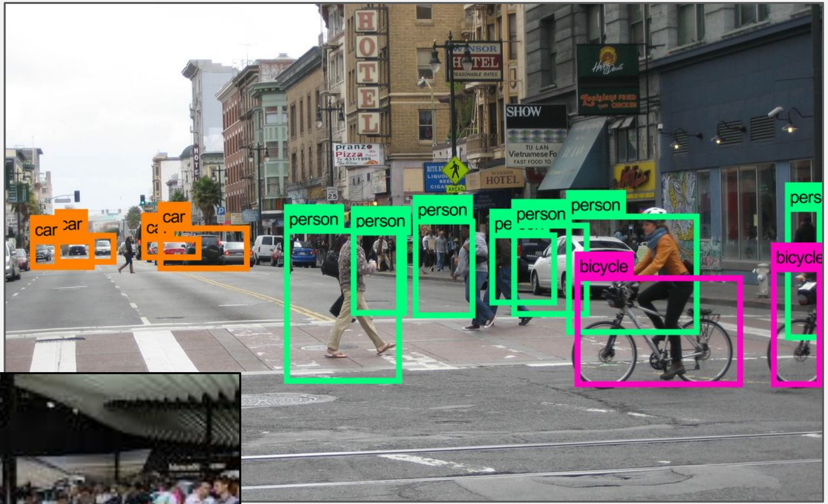
MIND THE GAP

Targeting the Distribution Gap using Augmentation

John Boudreaux, Sarah Danzi, Jennifer Mahle

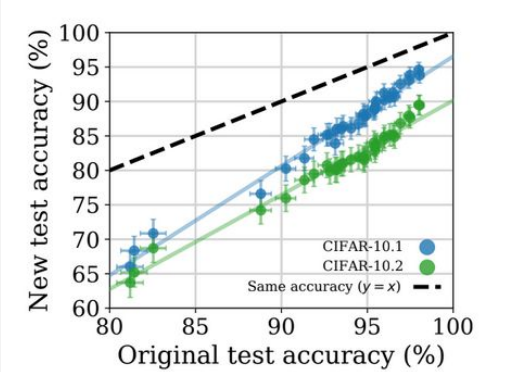
August 4, 2020

Introduction



Distribution Gap

- A distribution gap occurs when there is a systemic difference between the data distributions between datasets, causing gaps in model accuracy
- Previous research in this area shows that a even datasets pulled from the same underlying source, following the same methodology can yield a significant drop in accuracy driven by a distribution gap
- Accuracy loss could impact model performance in real life settings, like autonomous vehicles

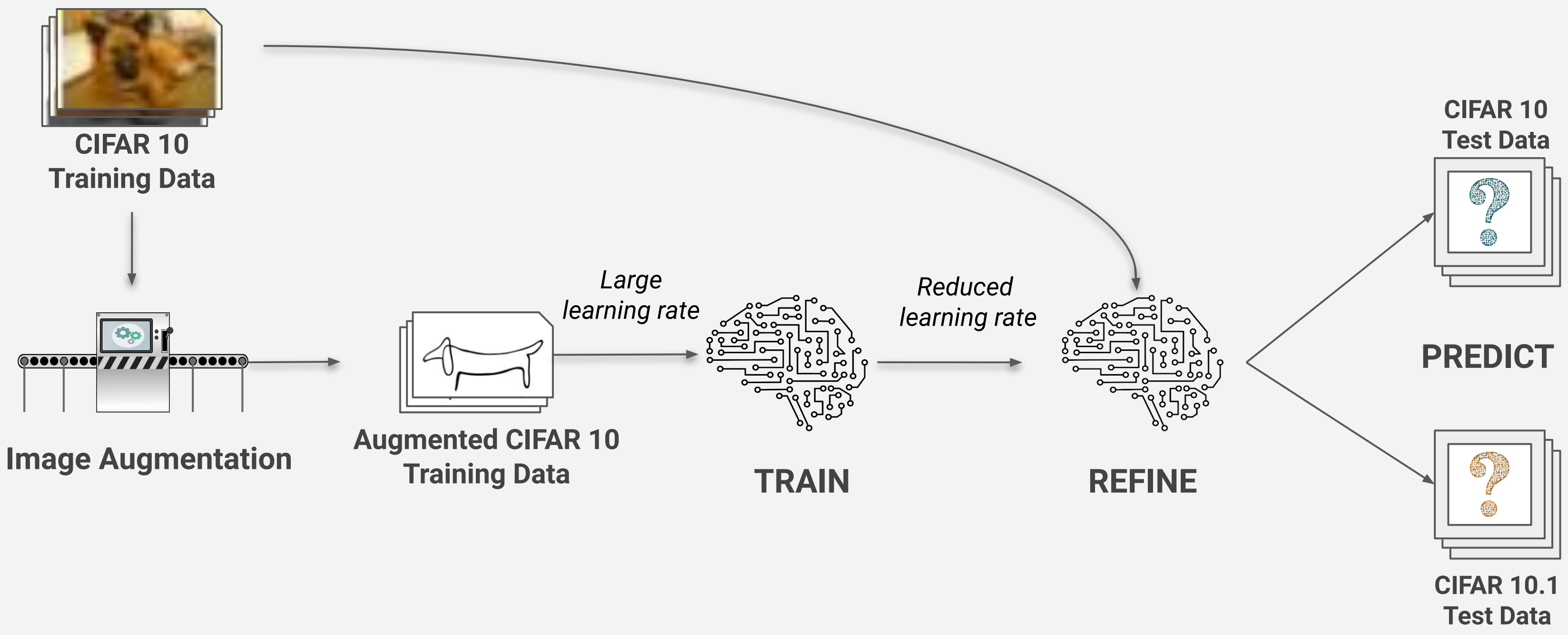


?

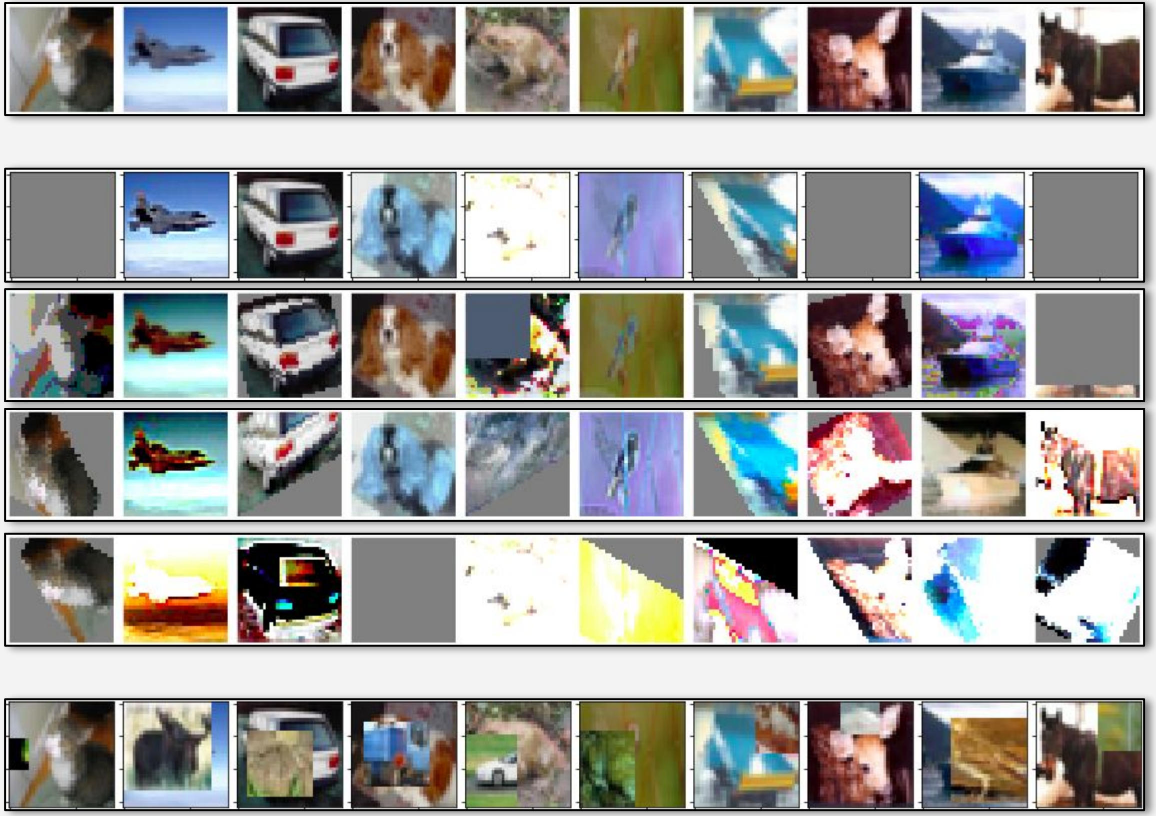
Data, Models, Augmentation

- To analyze the distribution gap, we used CIFAR-10 and CIFAR10.1 data, both of which have 10 classifications of images pulled from the Tiny Images data source
- We selected two different data augmentation methodologies to test in reducing the distribution gap by increasing a model's ability to generalize:
    - **CutMix**: A patch of one image overwrites the corresponding pixels in another and labels are weighted accordingly, increasing regularization
    - **RandAugment**: Randomly selects a configurable number of transformations to apply
- We used four previously published models to assess the distribution gap

Experimental Approach



Training Datasets



CIFAR-10 Training Dataset

RandAugment N=1, M=20

N=2, M=5

N=2, M=20

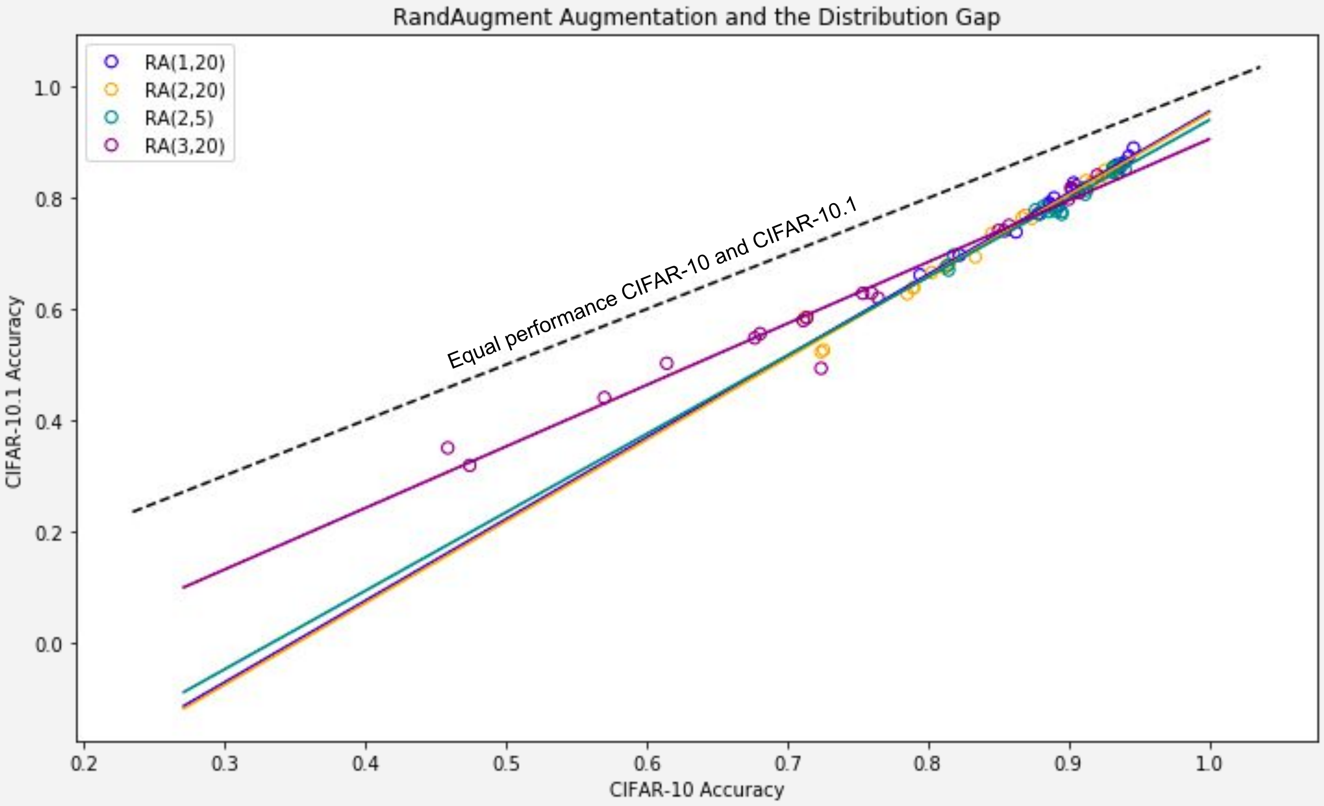
N=3, M=20

Cut Mix

Experiment Permutations

<table>
<tr><th colspan="2">TRAINING</th><th rowspan="2">REFINE</th><th rowspan="2">INFERENCE</th></tr>
<tr><th>Training Data</th><th>Validation Data</th></tr>
<tr><td rowspan="2">CIFAR 10</td><td rowspan="2">CIFAR 10</td><td rowspan="2">-</td><td>CIFAR 10</td></tr>
<tr><td>CIFAR 10.1</td></tr>
<tr><td rowspan="16">RandAugment(CIFAR 10, N=x, M=y)
$\forall\ (x,y) \in \{(1,20),(2,20),(3,20),(2,5)\}$</td><td rowspan="8">RandAugment(CIFAR 10, N=x, M=y)</td><td rowspan="4">-</td><td>CIFAR 10</td></tr>
<tr><td>CIFAR 10.1</td></tr>
<tr><td>RandAugment(CIFAR 10, N=x, M=y)</td></tr>
<tr><td>RandAugment(CIFAR 10.1, N=x, M=y)</td></tr>
<tr><td rowspan="4">CIFAR 10</td><td>CIFAR 10</td></tr>
<tr><td>CIFAR 10.1</td></tr>
<tr><td>RandAugment(CIFAR 10, N=x, M=y)</td></tr>
<tr><td>RandAugment(CIFAR 10.1, N=x, M=y)</td></tr>
<tr><td rowspan="8">CIFAR 10</td><td rowspan="4">-</td><td>CIFAR 10</td></tr>
<tr><td>CIFAR 10.1</td></tr>
<tr><td>RandAugment(CIFAR 10, N=x, M=y)</td></tr>
<tr><td>RandAugment(CIFAR 10.1, N=x, M=y)</td></tr>
<tr><td rowspan="4">CIFAR 10</td><td>CIFAR 10</td></tr>
<tr><td>CIFAR 10.1</td></tr>
<tr><td>RandAugment(CIFAR 10, N=x, M=y)</td></tr>
<tr><td>RandAugment(CIFAR 10.1, N=x, M=y)</td></tr>
<tr><td rowspan="4">CutMix(CIFAR 10, Alpha=1)</td><td rowspan="4">CIFAR 10</td><td rowspan="2">-</td><td>CIFAR 10</td></tr>
<tr><td>CIFAR 10.1</td></tr>
<tr><td rowspan="2">CIFAR 10</td><td>CIFAR 10</td></tr>
<tr><td>CIFAR 10.1</td></tr>
</table>

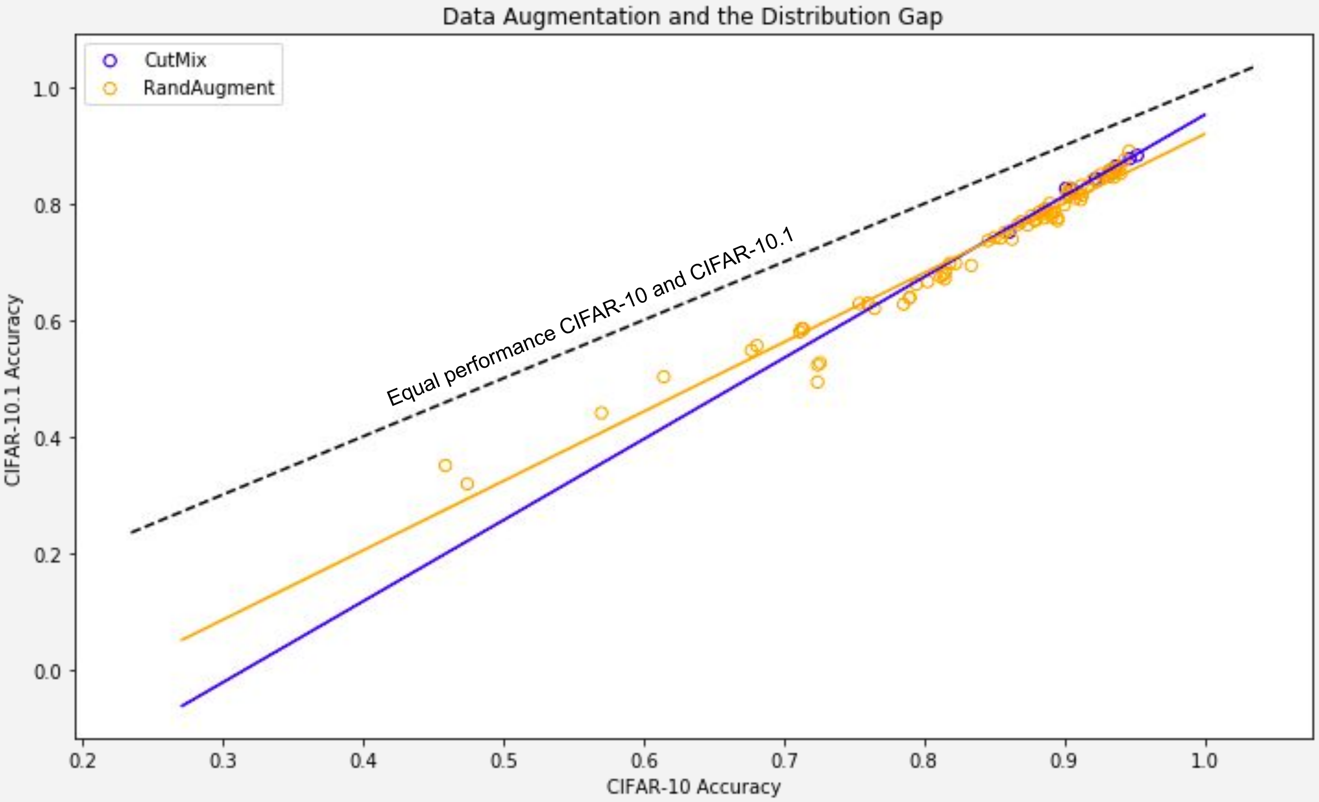
RandAugment Results



Across all different RandAugment configurations, we do not get significantly closer to equal performance on CIFAR-10 and CIFAR-10.1 test sets.

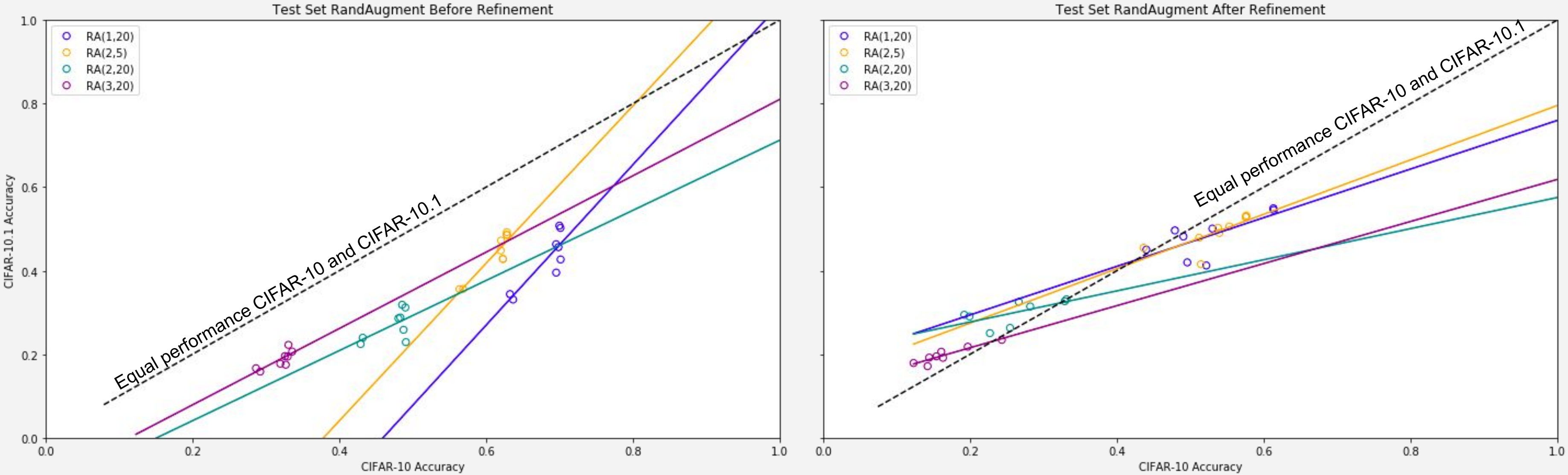
It does not appear RandAugment helps bridge the distribution gap.

CutMix Results



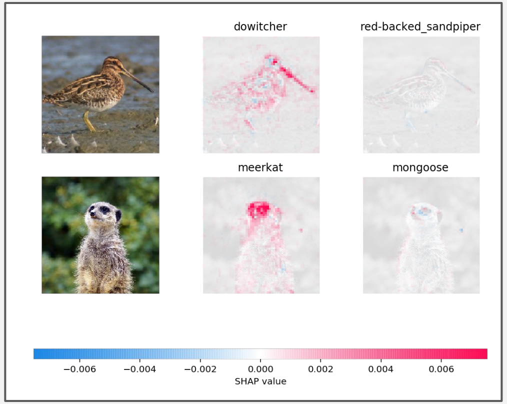
CutMix appears to give similar performance with regards to the distribution gap.

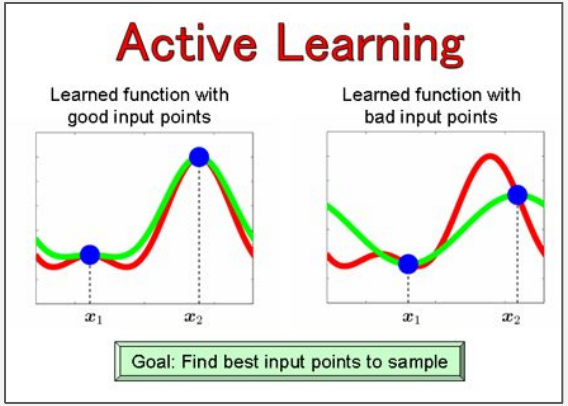
Inferencing on an Augmented Testset



Using augmentation on test set gave better results for the distribution gap, but hurt accuracy too much to be useful.

Future Work


dowitcher
red-backed_sandpiper
meerkat
mongoose
−0.006
−0.004
−0.002
0.000
0.002
0.004
0.006
SHAP value


Active Learning
Learned function with good input points
Learned function with bad input points
x_1
x_2
x_1
x_2
Goal: Find best input points to sample


SHAPENET

Conclusion

Overall, it seems that data augmentation does not effectively bridge the distribution gap seen in computer vision research without hurting overall model performance.

Backup

Citations

Cubuk, E., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. (2019). AutoAugment: Learning Augmentation Strategies From Data. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, & Quoc V. Le. (2019). RandAugment: Practical automated data augmentation with a reduced search space.

Gao Huang and Zhuang Liu and Kilian Q. Weinberger (2016). Densely Connected Convolutional NetworksCoRR, abs/1608.06993.

He, Zhuoxun & Xie, Lingxi & Chen, Xin & Zhang, Ya & Wang, Yanfeng & Tian, Qi. (2019). Data Augmentation Revisited: Rethinking the Distribution Gap between Clean and Augmented Data.

Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun (2015). Deep Residual Learning for Image RecognitionCoRR, abs/1512.03385.

Krizhevsky, A. Learning Multiple Layers of Features From Tiny Images, 2009. https://www.cs.toronto.e du/˜kriz/learning-features-2009-TR.pdf.

Lu, Shangyun, Nott, Bradley Nott, Olson, Aaron, Todeschini, Alberto, Vahabi, Hossein, Carmon, Yair, & Schmidt, Ludwig. (2020) Harder or Different? A Closer Look at Distribution Shift in Dataset Reproduction

Lundberg, Scott M and Lee, Su-In. (2017) A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30.

Recht, Benjamin, Roelofs, Rebecca, Schmidt, Ludwig, & Shankar, Vaishaal (2019). Do ImageNet Classifiers Generalize to ImageNet?CoRR, abs/1902.10811.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, & Wieland Brendel. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.

Saining Xie and Ross B. Girshick and Piotr Dollár and Zhuowen Tu and Kaiming He (2016). Aggregated Residual Transformations for Deep Neural NetworksCoRR, abs/1611.05431.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In International Conference on Computer Vision, 2019.

Sergey Zagoruyko and Nikos Komodakis (2016). Wide Residual NetworksCoRR, abs/1605.07146.