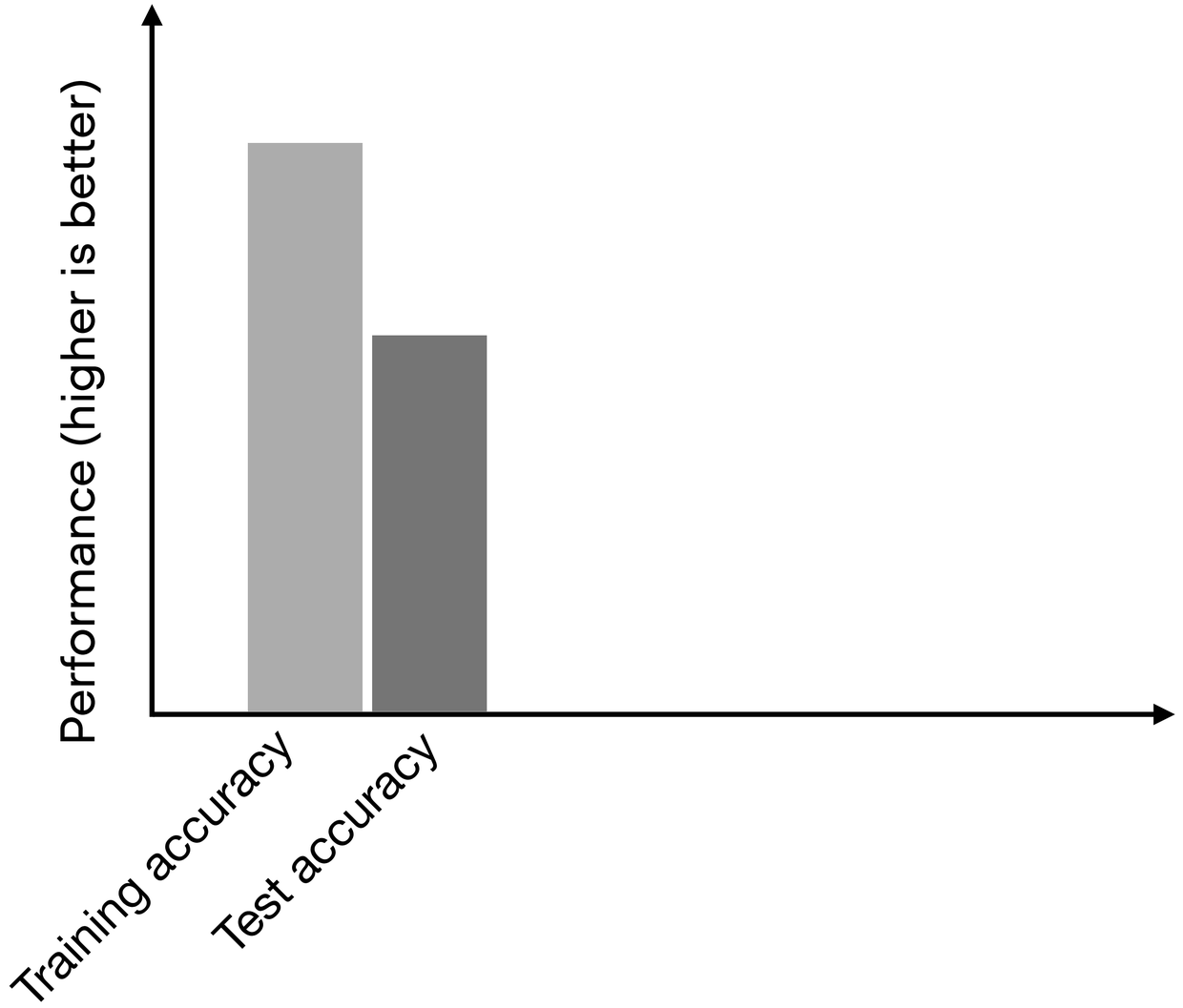
7.5

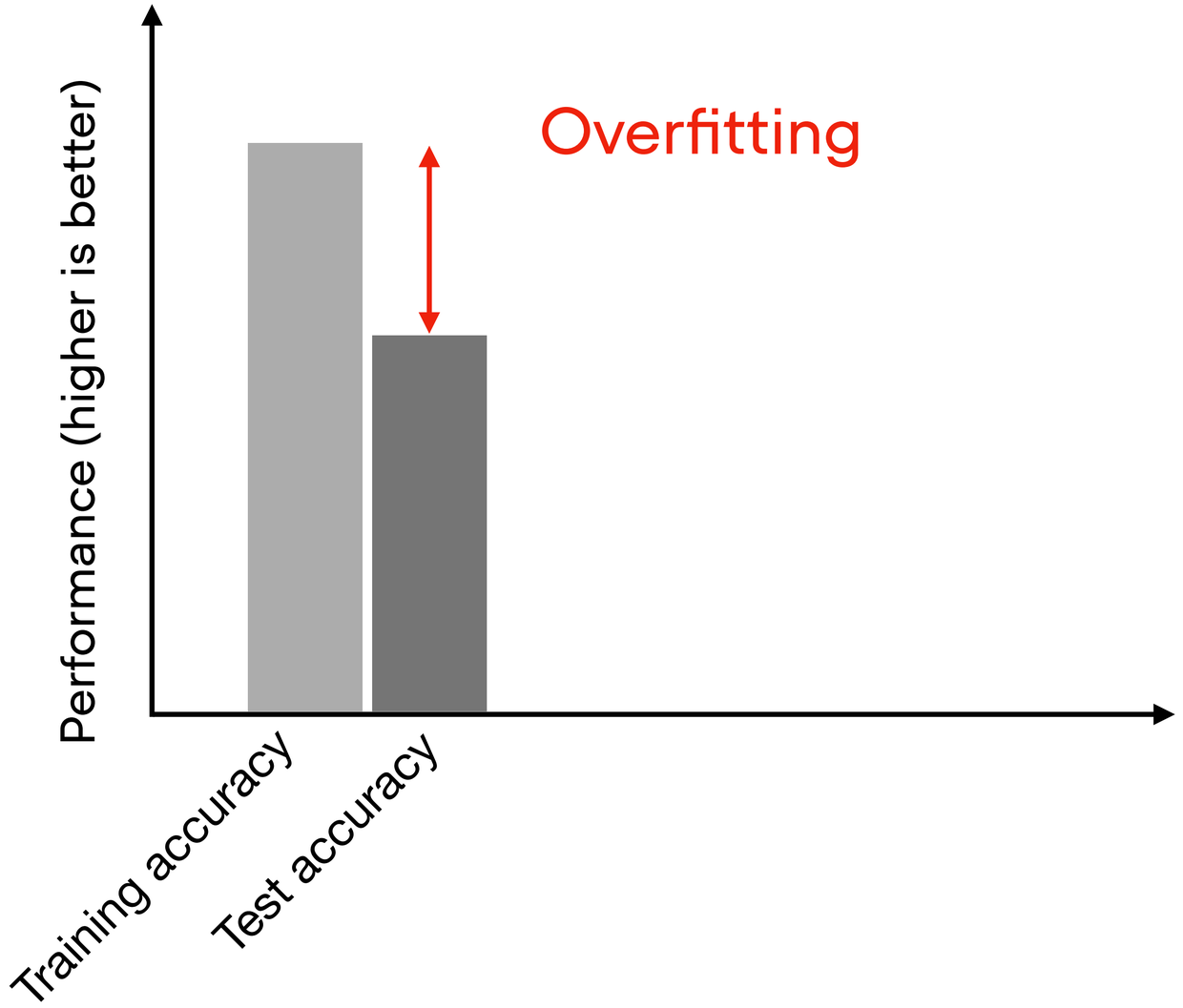
Improving Predictions with Data Augmentation

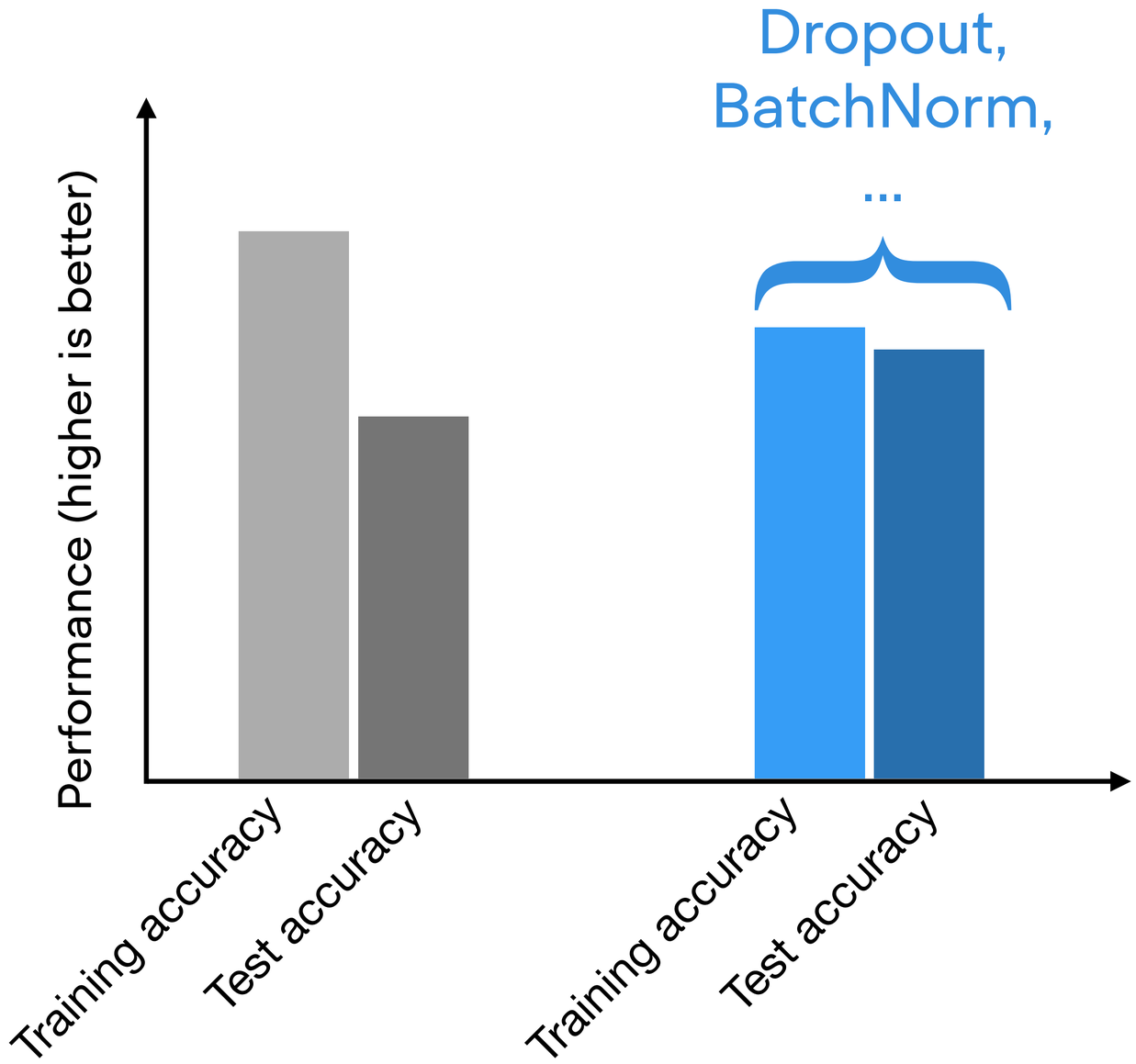
Part 1: Concepts and Examples

Sebastian Raschka and the Lightning AI Team

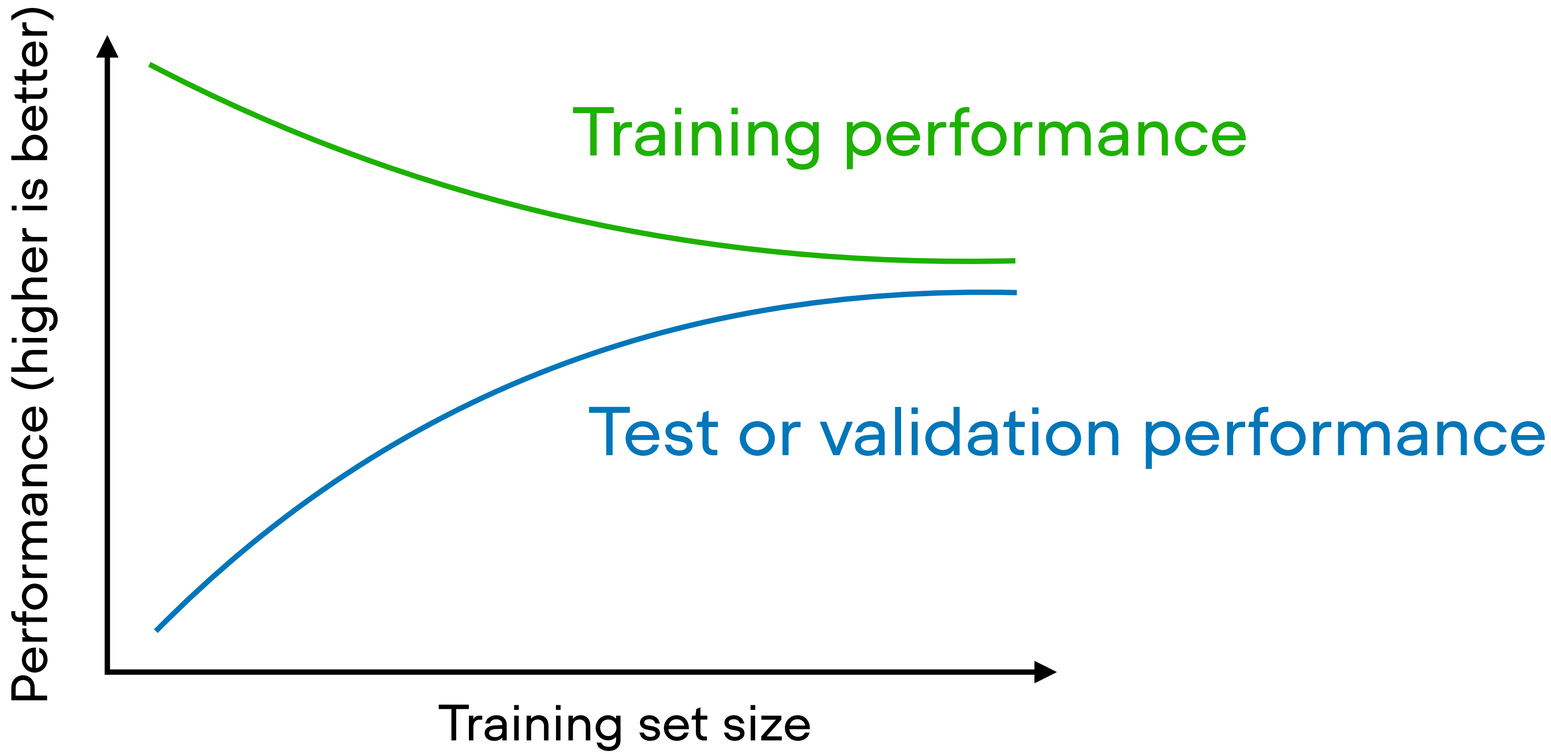
Goal:
Improving generalization performance

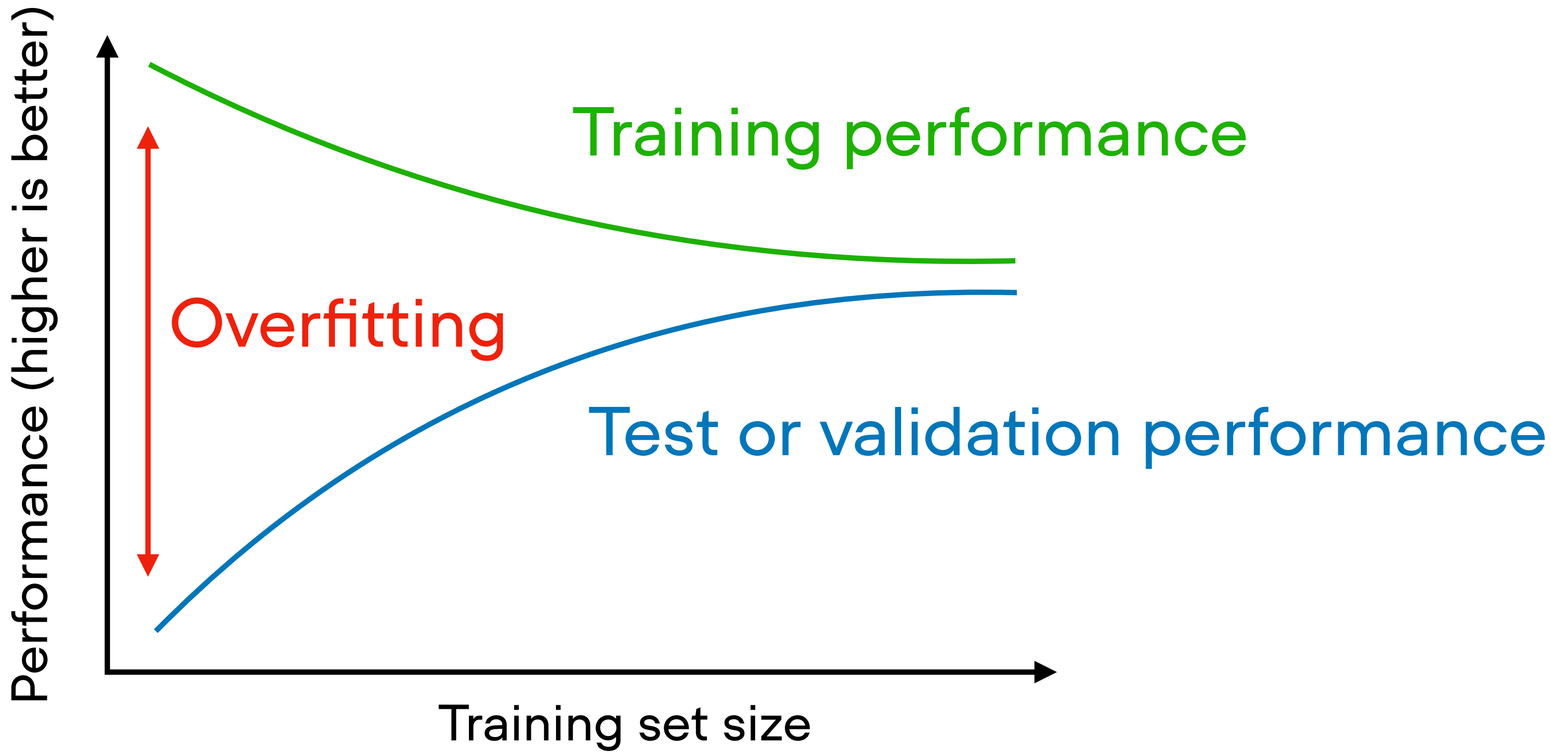

Performance (higher is better)
Training accuracy
Test accuracy


Overfitting
Performance (higher is better)
Training accuracy
Test accuracy


Dropout,
BatchNorm,
...
Performance (higher is better)
Training accuracy
Test accuracy
Training accuracy
Test accuracy

The best way to improve generalization performance is collecting additional data


Performance (higher is better)
Training performance
Test or validation performance
Training set size

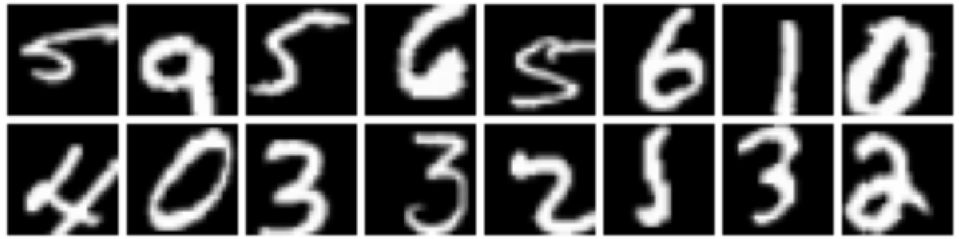

Performance (higher is better)
Training performance
Overfitting
Test or validation performance
Training set size

Random data augmentation is a way
to artificially increase the dataset size

Original training images

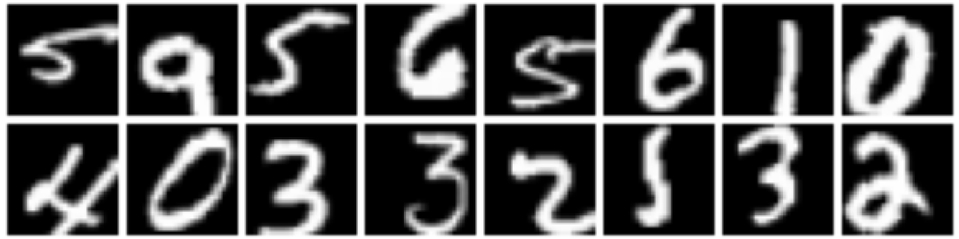
Original training images

Randomly cropped training images

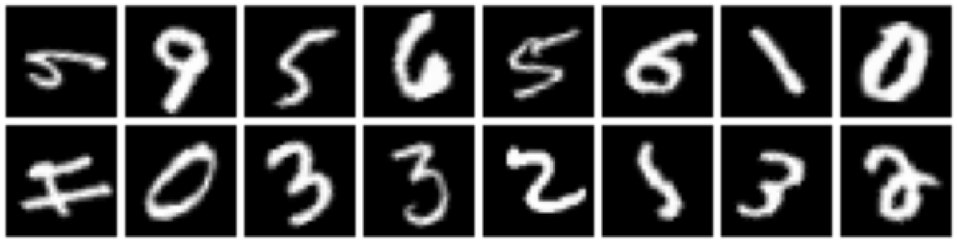


Original training images

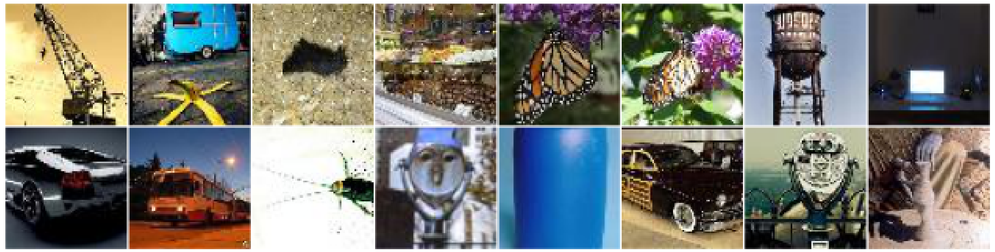
Randomly cropped training images



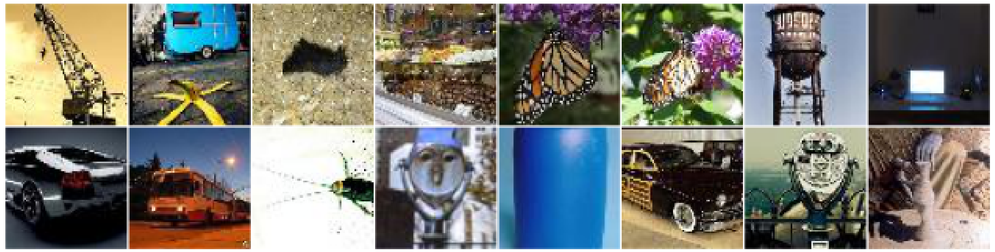
Randomly rotated training images



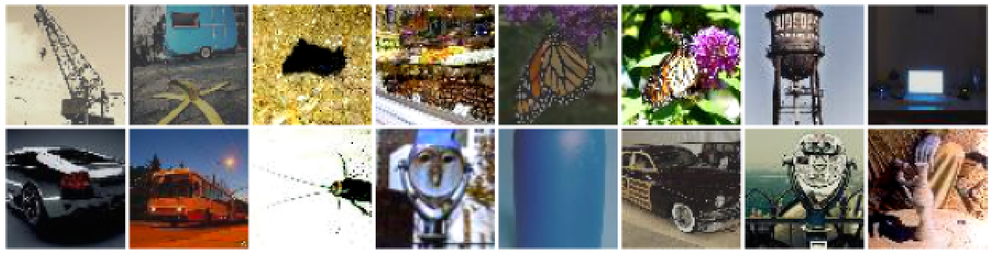
Original training images



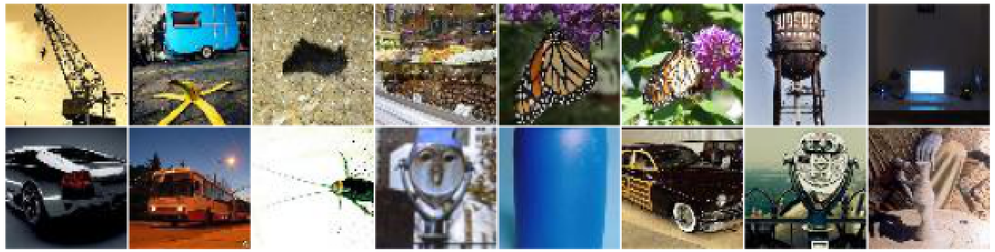
Original training images



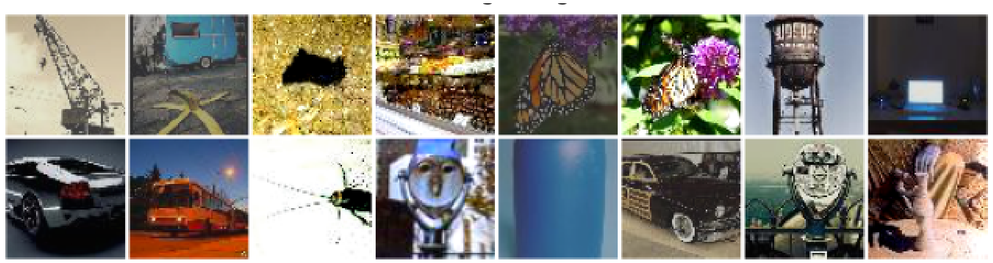
Randomly modified brightness



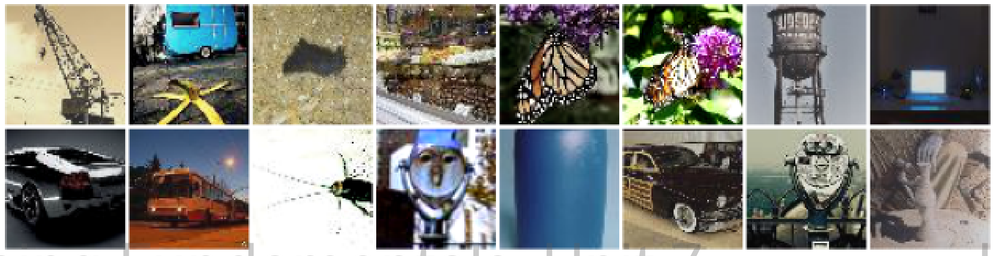
Original training images



Randomly modified brightness



Randomly modified saturation and hue



There are many, many more!

But we don't apply random transformation to test data (we want deterministic behavior)

Next: Let's try it out in practice