# Data Augmentation

Limited data is a major obstacle in applying deep learning models like convolutional neural networks. Often, imbalanced classes can be an additional hindrance; while there may be sufficient data for some classes, equally important, but undersampled classes will suffer frompoor class-specific accuracy.

## What is Data Augmentation?

Definition of "data augmentation" on Wikipedia is "Techniques are used to increase the amountof data by adding slightly modified copies of already existing data or newly created synthetic data from existing data." So data augmentation involves creating new and representative data.

### Why is it important now?

Machine learning applications especially in deep learning domain continue to diversify and increase rapidly. Data augmentation techniques may be a good tool against challenges whichartificial intelligence world faces.

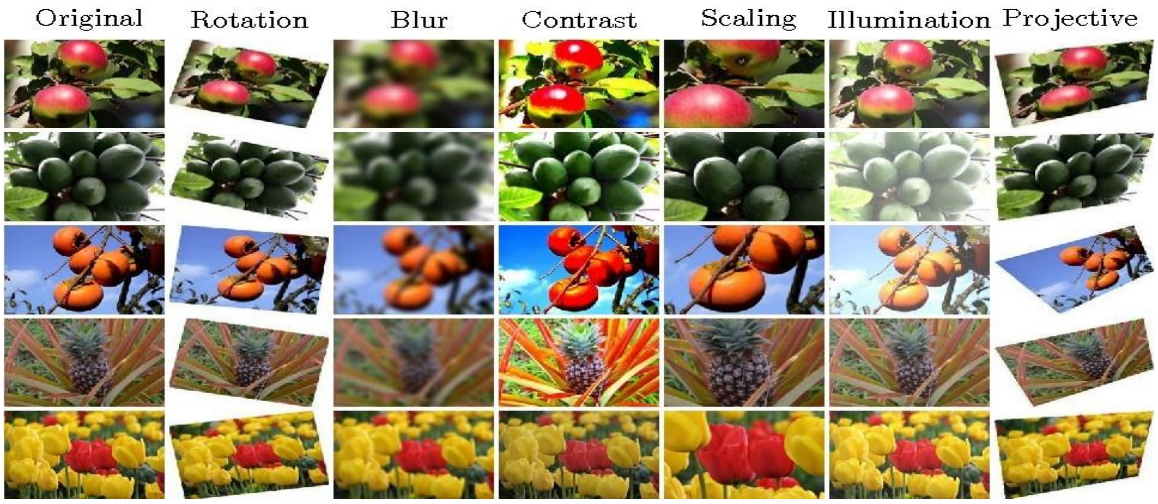
Data augmentation is useful to improve performance and outcomes of machine learning modelsby forming new and different examples to train datasets. If dataset in a machine learning model is rich and sufficient, the model performs better and more accurate.

## Used in image classification and segmentation

For data augmentation, making simple alterations on visual data is popular. In addition, generative adversarial networks (GANs) are used to create new synthetic data. Classic imageprocessing activities for data augmentation are

- padding
- random rotating
- re-scaling,
- vertical and horizontal flipping
- translation ( image is moved along X, Y direction)
- cropping
- zooming
- darkening & brightening/color modification
- grayscaling
- changing contrast
- adding noise

- random erasing




**Famous Research Papers**

**Random Erasing Data Augmentation**

Random Erasing Paper Link

**Improved Regularization of Convolutional Neural Networks withCutout**

Cutout Paper

**Data Augmentation using Random Image Cropping and Patching forDeep CNNs**

RICAP Paper

## What are the benefits of data augmentation?

Benefits of data augmentation include:

1. Improving model prediction accuracy
   - adding more training data into the models
   - preventing data scarcity for better models
   - reducing data overfitting
   - increasing generalization ability of the models
   - helping resolve class imbalance issues in classification
2. Reducing costs of collecting and labeling data

## What are the challenges of data augmentation?

1. Companies need to build evaluation systems for quality of augmented datasets. As use of data augmentation methods increases, assessment of quality of their output will be required.
2. Data augmentation domain needs to develop new research and studies to create new/synthetic data with advanced applications. For example, generation of high-resolution images by using GANs is challenging
3. If real dataset contains biases, data augmented from it will contain biases, too. So, identification of optimal data augmentation strategy is important.