



ROHINI

COLLEGE OF ENGINEERING & TECHNOLOGY

Approved by AICTE and Affiliated to Anna University (An ISO Certified Institution) | Accredited with A+ Grade by NAAC
Recognized under Section 2(f) of University Grants Commission, UGC ACT 1956
(AUTONOMOUS)

DATASET AUGMENTATION









The best way to make a machine learning model generalize better is to train it on more data. Of course, in practice, the amount of data we have is limited. One way to get around this problem is to create new data and add it to the training set.

Data augmentation is easiest for classification, Classifier takes high-dimensional input x and summarizes it with a single category identity y . Main task of classifier is to be invariant to a wide variety of transformations. We can generate new samples (x,y) just by transforming inputs.

This Approach not easily generalized to other problems, Example density estimation problem. It is not possible generate new data without solving density estimation.

Data set augmentation very effective for the classification problem of object recognition. Images are high-dimensional and include a variety of variations, may easily simulated. translating the training images a few pixels in each direction can greatly improve performance. Many other operations such as rotating the image or scaling the image have also proven quite effective.

One must be careful not to apply transformations that would change the correct class. For example, optical character recognition tasks require recognizing the difference between 'b' and 'd' and the difference between '6' and '9', so horizontal flips and 180° rotations are not appropriate ways of augmenting datasets for these tasks.

Original	Flip	Rotation	Random crop
			
<ul style="list-style-type: none"> • Image without any modification 	<ul style="list-style-type: none"> • Flipped with respect to an axis for which the meaning of the image is preserved 	<ul style="list-style-type: none"> • Rotation with a slight angle • Simulates incorrect horizon calibration 	<ul style="list-style-type: none"> • Random focus on one part of the image • Several random crops can be done in a row
Color shift	Noise addition	Information loss	Contrast change
			
<ul style="list-style-type: none"> • Nuances of RGB is slightly changed • Captures noise that can occur with light exposure 	<ul style="list-style-type: none"> • Addition of noise • More tolerance to quality variation of inputs 	<ul style="list-style-type: none"> • Parts of image ignored • Mimics potential loss of parts of image 	<ul style="list-style-type: none"> • Luminosity changes • Controls difference in exposition due to time of day

Injecting noise into the input of a neural network can be seen as data augmentation. For some regression tasks it is still possible to solve even if small random noise is added to the input. Neural networks are not robust to noise. To improve robustness, train them with random noise applied to their inputs. Input noise injection is part of some unsupervised learning algorithms such as the denoising autoencoder. Noise can also be applied to hidden units, which can be seen as doing dataset augmentation at multiple levels of abstraction.

