

Data Set Augmentation

Lecture Notes in Deep Learning: <http://www.cedar.buffalo.edu/~srihari/CSE676>

Topics in Data Augmentation

1. More data is better
2. Augmentation for classification
3. Caution in data augmentation
4. Injecting noise
5. Benchmarking using augmentation
6. Ex: Heart disease diagnosis using deep learning

More data is better

- Best way to make a ML model to generalize better is to train it on more data
- In practice amount of data is limited
- Get around the problem by creating synthesized data
- For some ML tasks it is straightforward to synthesize data








Augmentation for classification

- Data augmentation is easiest for classification
 - Classifier takes high-dimensional input \mathbf{x} and summarizes it with a single category identity y
 - Main task of classifier is to be invariant to a wide variety of transformations
- Generate new samples (\mathbf{x}, y) just by transforming inputs
- Approach not easily generalized to other problems
 - For density estimation problem
 - it is not possible generate new data without solving density estimation

Effective for Object Recognition

- 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 images a few pixels can greatly improve performance
 - Even when designed to be invariant using convolution and pooling
- Rotating and scaling are also effective

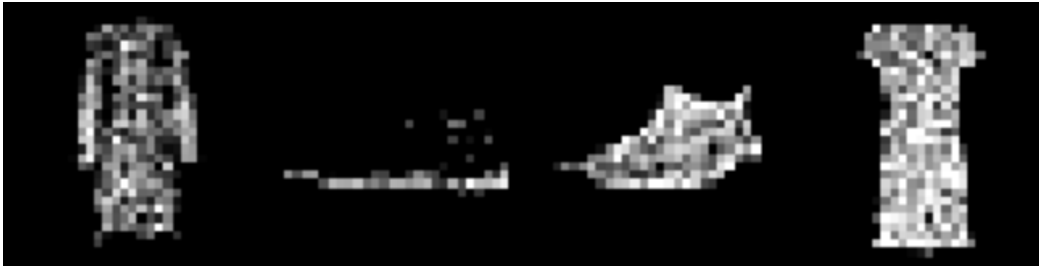
Main data augmentation methods

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

Remark: data is usually augmented on the fly during training.

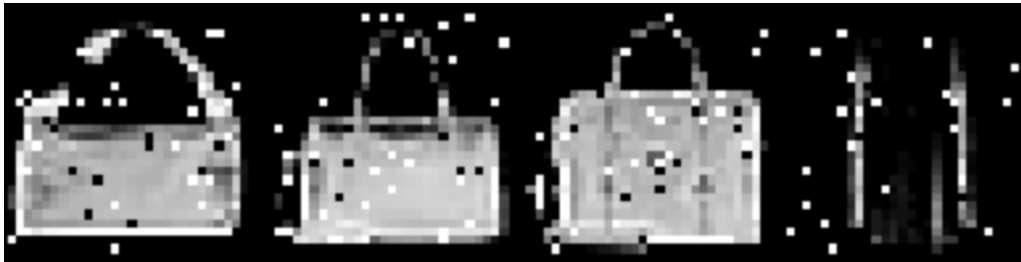
Fashion MNIST Noisy Images

Gaussian Noise



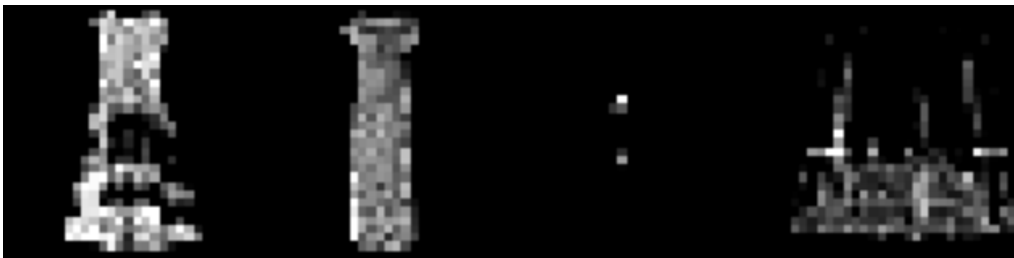
Noise on the objects only and not in the background.

Salt and Pepper Noise



Mixture of black and white noise on objects as well as background.

Speckle Noise



https://debuggercafe.com/adding-noise-to-image-data-for-deep-learning-data-augmentation/?fbclid=IwAR2KWCN5HExb8EmpW5Z6vsM_7y6n8le6-rcHvEwia6pze3DLM9hZEKU1arc

Caution in Data Augmentation

- Not apply transformation that would change the class
- OCR example: 'b' vs 'd' and '6' vs '9'
 - Horizontal flips and 180 degree rotations are not appropriate ways
- Some transformations are not easy to perform
 - Out of plane rotation cannot be implemented as a simple geometric operation on pixels

Injecting noise

- Injecting noise into the input of a neural network can be seen as data augmentation
- Neural networks are not robust to noise
- To improve robustness, train them with random noise applied to their inputs
 - Part of some unsupervised learning, such as denoising autoencoder
- Noise can also be applied to hidden units
- Dropout, a powerful regularization strategy, can be viewed as constructing new inputs by multiplying by noise

Benchmarking using augmentation

- Hand-designed data set augmentation can dramatically improve performance
- When comparing ML algorithms A and B, same data set augmentation should be used for both
 - If A performs poorly with no dataset augmentation and B performs well with synthetic transformations of the input, reason may be the data set rather than algorithm
- Adding Gaussian noise is considered part of ML while cropping input images is not

Acknowledgemnts

1. Goodfellow, I., Bengio, Y., and Courville, A., Deep Learning, MIT Press 2016
2. Yee, C-H., “Heart Disease Diagnosis with Deep Learning: State-of-the-art results with 60x fewer parameters”
<https://blog.insightdatascience.com/heart-disease-diagnosis-with-deep-learning-c2d92c27e730>