DROPOUT

Vincent Barra
LIMOS, UMR 6158 CNRS, Université Clermont Auvergne

Definition

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. JMLR, 15:1929?1958, 2014.

Dropout

Regularization technique.

- Remove units at random during the forward pass on each sample
- Trains an ensemble of corresponding subnetworks
- Put all neurons back during test.

LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

UNIVERSITÉ Clermont Auvergne

Definition

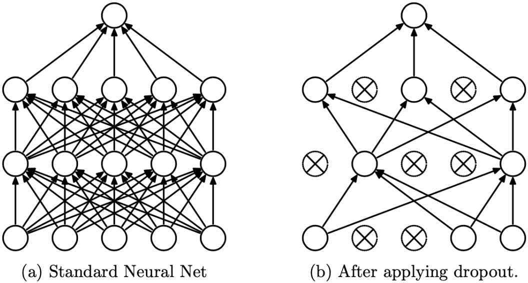


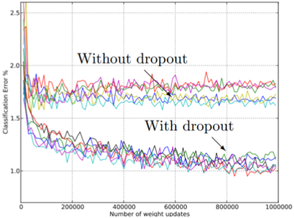
Figure 1: Dropout Neural Net Model. **Left**: A standard neural net with 2 hidden layers. **Right**: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.

Can be interpreted as set of models with heavy weight sharing.

Properties

Dropout

- Increases independence between neurons
- Distributes the representation
- Generally improves performance.



Bagging

Dropout is a bagging method

- Averaging over several models to improve generalization
- inexpensive but powerful method of regularizing a broad family of models

But: Bagging build independent models / Dropout use parameter sharing

LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

How ?

Objective: drop a neuron with probability p.
During training, for each sample, as many Bernoulli variables as neurons are sampled independently to select neurons to drop.

X: neuron activation
D independent boolean random variable of probability $1-p$.

$$\mathbb{E}(DX) = \mathbb{E}(D)\mathbb{E}(X) = (1-p)\mathbb{E}(X)$$

so either

- multiply activations by $1-p$ during test
- multiply activations by $1/(1-p)$ during train and keep the network untouched during test

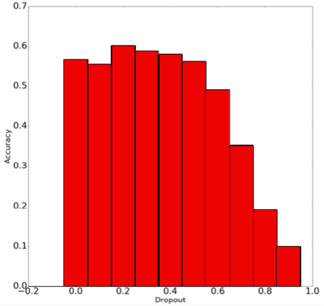
Practical implementation : drop activations at random on each sample using a `dropout` layer.

LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

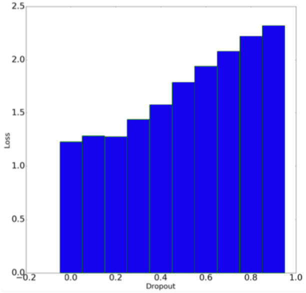
Dependance on p

Experiment:

- Conv((3×3),64) - MaxPool-Conv((3×3),128) - MaxPool - Conv((3×3),64256 - MaxPool - FC(512) - FC(512) - Dense(10).
- Trained on CIFAR10



Accuracy(p)

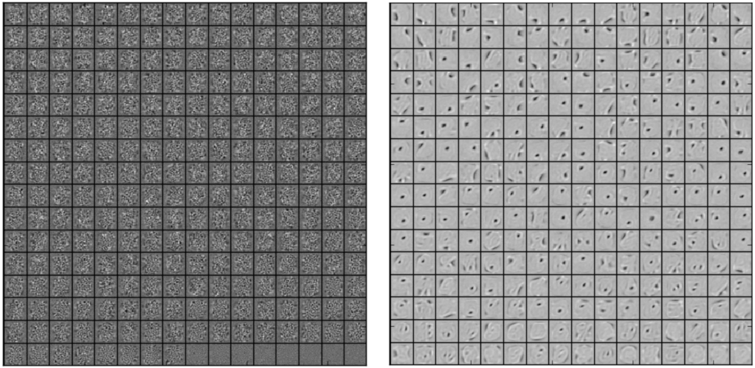


Loss(p)

LIMOS LABORATOIRE D'INFORMATIQUE, DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

UNIVERSITÉ Clermont Auvergne

Example



(a) Without dropout (b) Dropout with $p = 0.5$.

Features learned on MNIST with one hidden layer autoencoders having 256 rectified linear units (from Srivastava et al.)