Gradient-based learning applied to document recognition

By LeCun, Yann, et al

2017010698 수학과 오서영

1. Introduction

Machine Learning with Neural Network

-> important role in pattern recognition
-> hand-designed heuristics + automatic learning

Pattern Recognition

1. First module : feature extractor

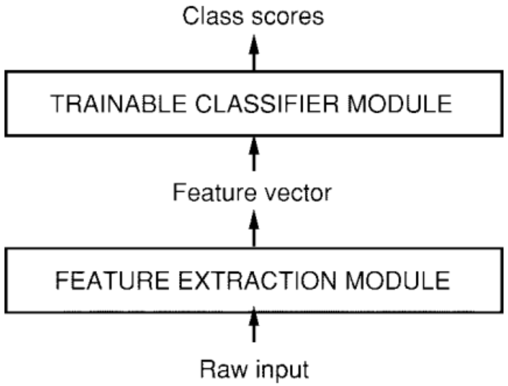
- Input patterns -> low dimensional vectors
- Distortions of the input -> do not change their nature
- Hand-craft feature extractor -> **prior knowledge**

2. Second module : Classifier

- general-purpose and trainable

-> problem

: accuracy is largely determined by the ability of the feature extractor



2. Learning from Data

Gradient-based learning

Learning machine computation

$$Y^p = F(Z^p, W)$$

Z^p p-th input pattern
W collection of adjustable parameters
Y^p class label (output)

Loss function

$$E^p = \mathcal{D}(D^p, F(W, Z^p))$$

D^p desired output

-> finding the value of **W** that minimize $E_{\text{train}}(W)$

-> W is updated $W_k = W_{k-1} - \epsilon \frac{\partial E(W)}{\partial W}.$

3. Gradient Back-Propagation

Usefulness of Gradient-based learning

was not widely realized until the following three events occurred.

1. The presence of local minima in the loss function does not seem to be a major problem in practice

2. Efficient procedure = Back-propagation
to compute the gradient in a non-linear system composed of several layers of processing

3. Back-prop applied to multi-layer neural networks can solve complicated learning tasks

4. Globally Trainable Systems

Most practical pattern recognition systems are composed of multiple modules
- Each module must be continuous and differentiable almost everywhere with respect to the internal parameters of the module

$$X_n = F_n(W_n, X_{n-1})$$

X_n input vector representing the output of the module
W_n Vector of tunable parameters
X_{n-1} Module's input vector

Partial derivatives of E^p **with respect to** W_n **and** X_{n-1}

$$\frac{\partial E^p}{\partial W_n} = \frac{\partial F}{\partial W}(W_n, X_{n-1})\frac{\partial E^p}{\partial X_n}$$
$$\frac{\partial E^p}{\partial X_{n-1}} = \frac{\partial F}{\partial X}(W_n, X_{n-1})\frac{\partial E^p}{\partial X_n}$$

5. Convolutional Networks

Traditional model of pattern recognition,
Hand-designed feature extractor gathers relevant information from input
And eliminates irrelevant information

Trainable classifier then categorized the resulting feature vector into classes

Network could be fed with almost **raw** inputs
-> fully connected feed-forward network -> **problems**

5. Convolutional Networks

1. Typical images are large, often with several hundred variables (pixels) -> **large number of parameters** increases the capacity of the system and therefore requires a larger training set

2. Handwriting ~ size, slant, position variations
for individual characters

3. Variables are spatially or temporally nearby are highly correlated
-> Deficiency of fully connected architectures is that the topology of the input is entirely ignored.

5. Convolutional Networks

CNN
- Local receptive fields, shared weights, sub-sampling

1. Local receptive fields
Neurons can extract elementary visual features such as oriented edges, end-points, corners
-> These features are combined by the subsequent layers in order to detect higher-order features

2. shared weights
Units in a layer shares the same set of weights.
The set of outputs of the units in such a plane is called a **feature map**. Units in a feature map are all constrained to perform the same operation on different parts of the image.
A complete convolutional layer is composed of several feature maps, so that multiple features can be extracted at each location

5. Convolutional Networks

3. sub-sampling
A simple way to reduce the precision with which the position of distinctive features are encoded in a feature map is to reduce the spatial resolution of the feature map.
-> thereby reducing the resolution of the feature map and reducing the sensitivity of the output to shifts and distortions