INT307 Multimedia Security System

Information Retrieval with Deep Learning

Shengchen Li

Xi'an Jiaotong-Liverpool University

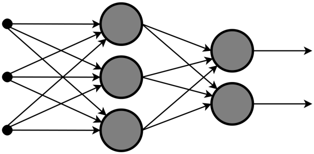
07th Oct 2022

Aims

- Master the working principle of deep learning
- Understand basic knowledge related to deep learning
- Master the framework of multimedia information retrieval via machine learning

Recall INT104

- The boundaries between classes are not necessary linear but can be approximate as a combination of linear functions.



- Could be single layer or multiple layer
- There is a threshold process after the output of each neuron, which is named as activation function

Artificial Neural Networks

- Data
- Input Layer
- Hidden Layer
- Output Layer

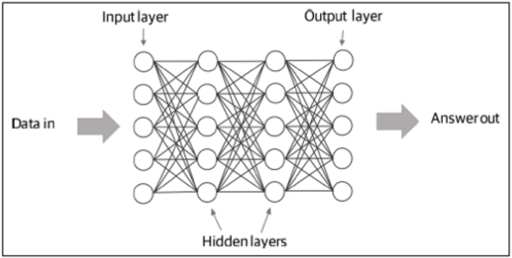


Figure 3-1. A multilayer perceptron

- Feature Extraction
- Classification

Forward Propagation

- Neurons effectively represent a mapping between feature spaces
- In neural networks, the mapping is represented as weighted sums with activation functions

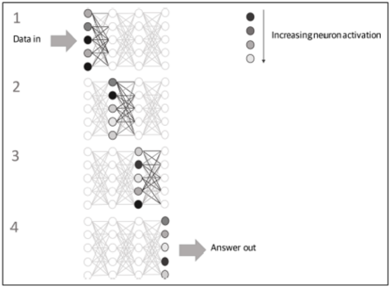


Figure 3-2. Incremental calculation steps in a multilayer perceptron

Forward Propagation

- Diagram 28×28
- 784 input neurons
- Two hidden layers with 56 neurons each
- RELU as activation functions

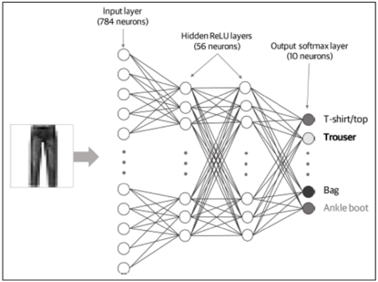
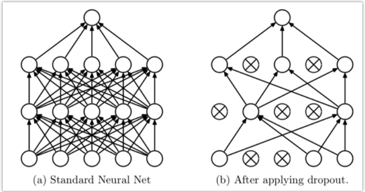


Figure 3-7. DNN architecture for image classification using Fashion-MNIST data

Common Tricks

- Dropout



- Batch Normalisation

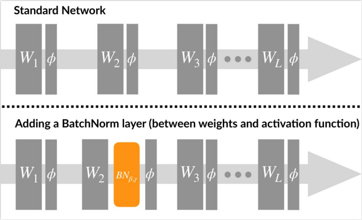


Image Processing with Deep Learning

- Scene classification
- Object detection and localisation
- Semantic segmentation
- Facial recognition

Figure 4-2. An example of image segmentation[1]

Recall

How images are presented by computer systems?

Filter and Convolution

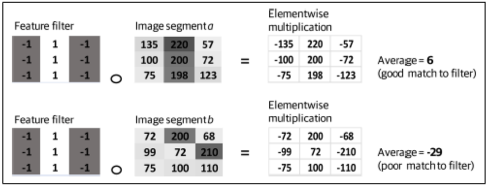


Figure 4-4. Application of a simple 3 x 3 filter to two different image segments

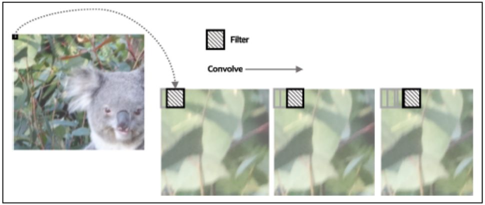


Figure 4-3. A convolutional filter is applied iteratively across an image

Convolutional Layers

- Kernel
  - Size
  - Padding
  - Hop
- Feature Maps

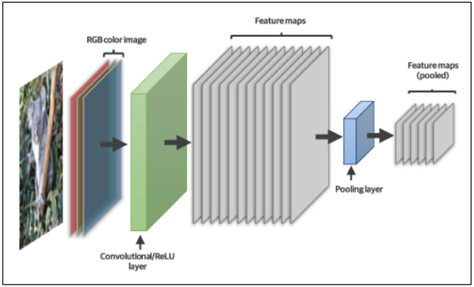


Figure 4-5. Typical layering pattern in a CNN

Question

1. Why convolution?
2. Why multiple feature maps?

Convolutional Neural Network

- Convolutional Layers
- Pooling Layers
- Fully Connected Layers
- Classifier

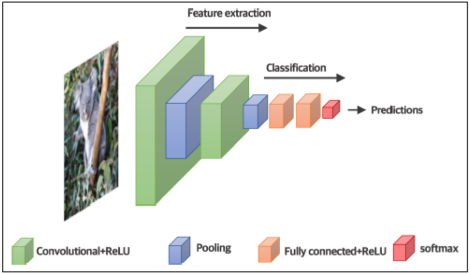


Figure 4-6. An example CNN image classification architecture

Common Image Processing Networks

- VGG
  - VGG-16
  - VGG-19
- Inception
  - 22 Layers
  - With many variations

Recall: Audio Representation

1. Waveform
2. Sampling
3. Quantisation
4. Linear Transform*

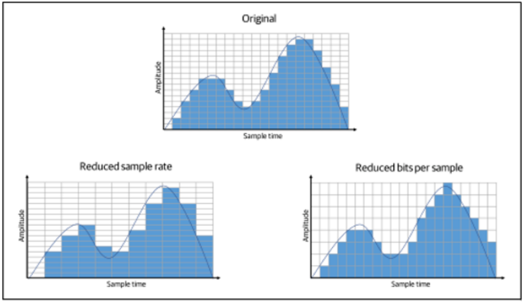


Figure 4-8. The effect of reduced sample rate and bits per sample in digital audio

Time – Frequency Transform

- Usually waveform is transformed to time-frequency domain before being processed
- Commonly used time-frequency transforms are:
  - DFT (FFT)
  - DCT
  - Mel-Spectrogram (MFCC)
  - Wavelet Transform

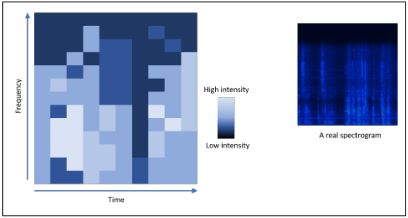


Figure 4-9. A spectrogram depicts changing intensities at different frequencies over time

Recurrent Neural Network

- Recurrent Neural Network is commonly used to process sequential media
- RNN features time slice analysis
- Commonly used time-frequency transforms are:
  - LSTM (Long Short Time Memory)
  - GRU (Gated Recurrent Unit)

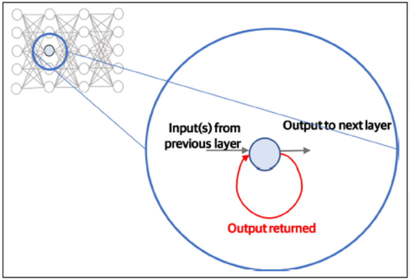


Figure 4-10. Basic concept underpinning RNN architectures

Typical Processing Chain for Audio

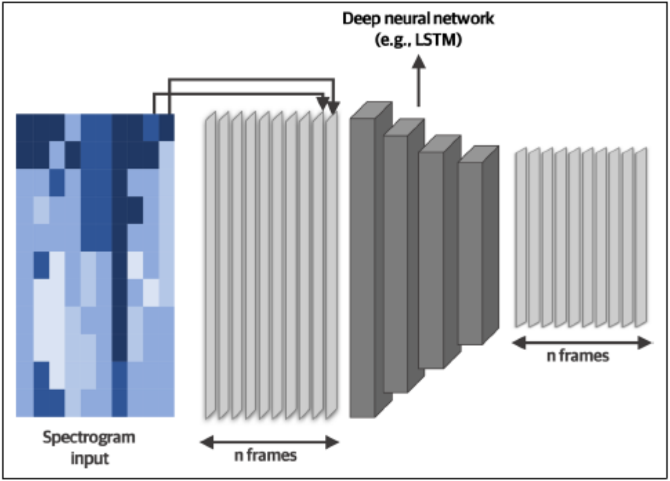
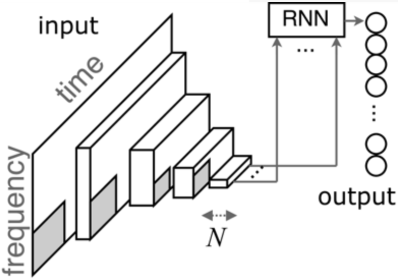


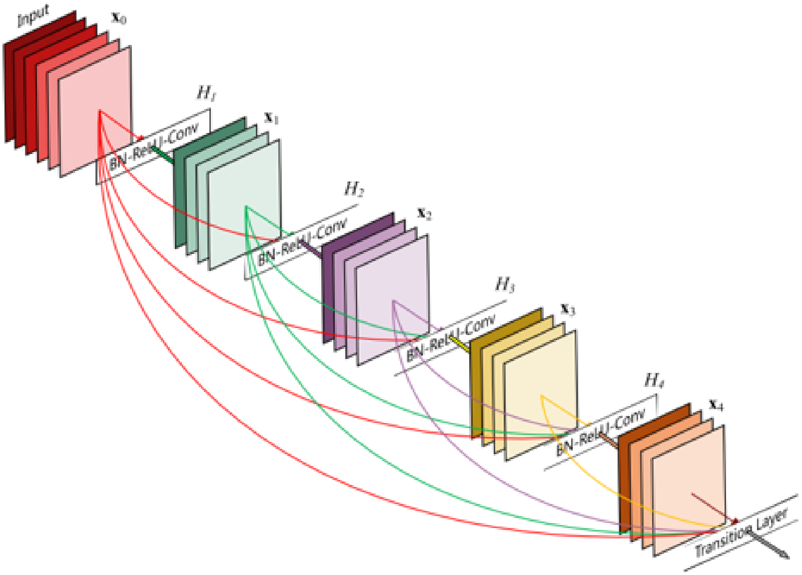
Figure 4-11. Typical processing chain for audio, with spectrogram preprocessing

Complex Networks

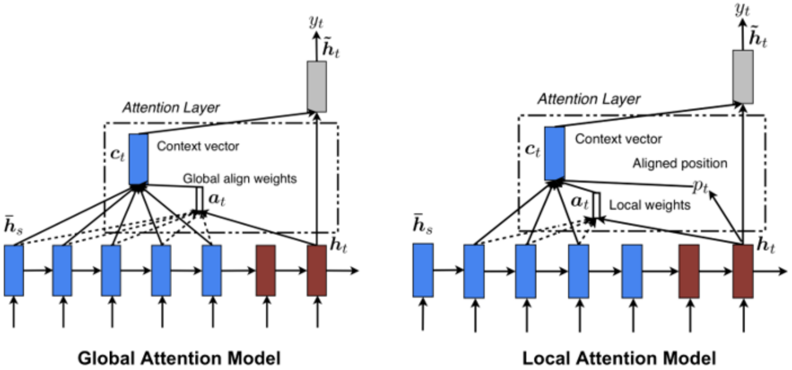
- A deep learning neural network can combine multiple types of structures
  - CNN = CNN + DNN
  - CRNN = CNN + RNN + DNN
- Discussion: Why CRNN can be considered as a way to analyse audio in multi-scale?



Residue Network



Attention


y_t
$\tilde{h}_t$
Attention Layer
c_t
Context vector
Global align weights
a_t
$\bar{h}_s$
h_t
Global Attention Model
y_t
$\tilde{h}_t$
Attention Layer
c_t
Context vector
Aligned position
p_t
a_t
Local weights
$\bar{h}_s$
h_t
Local Attention Model

Transformer

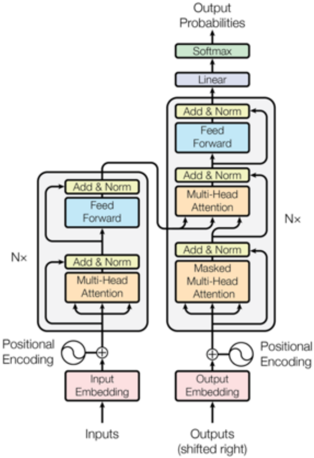
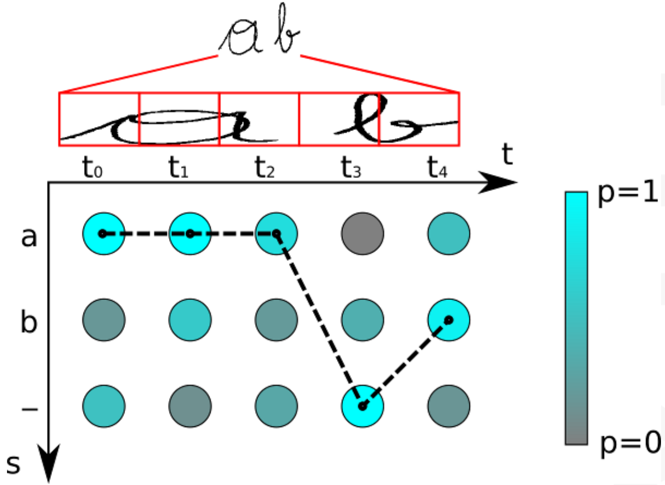
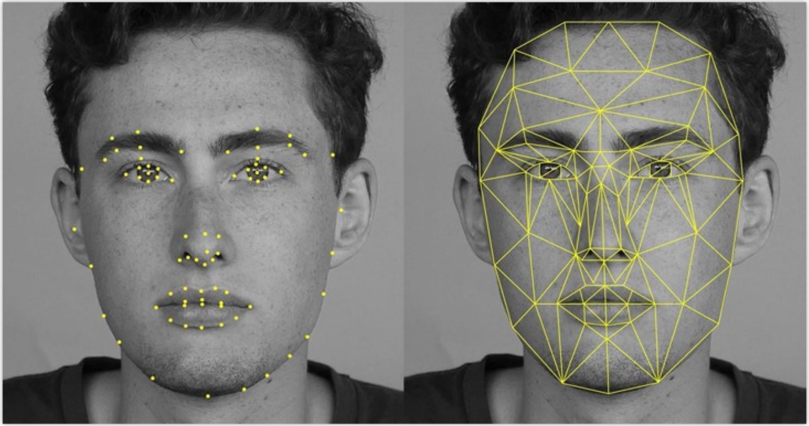


Figure 1: The Transformer - model architecture.

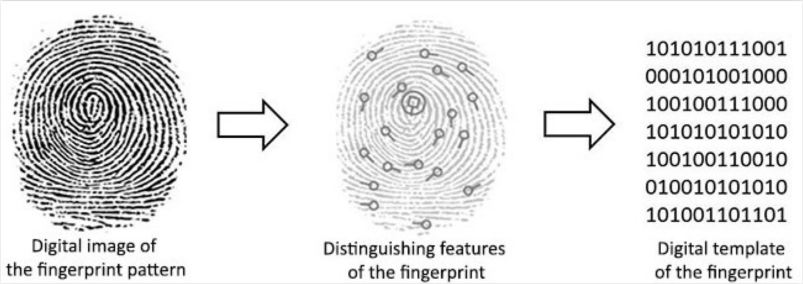
Connectionist Temporal Classification



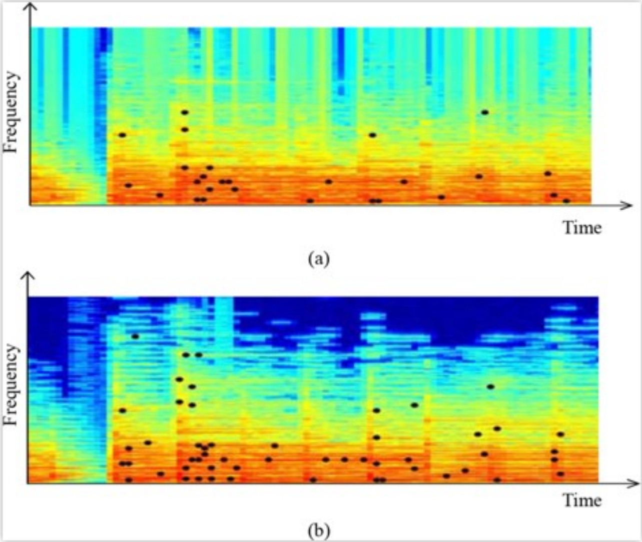
Face Recognition



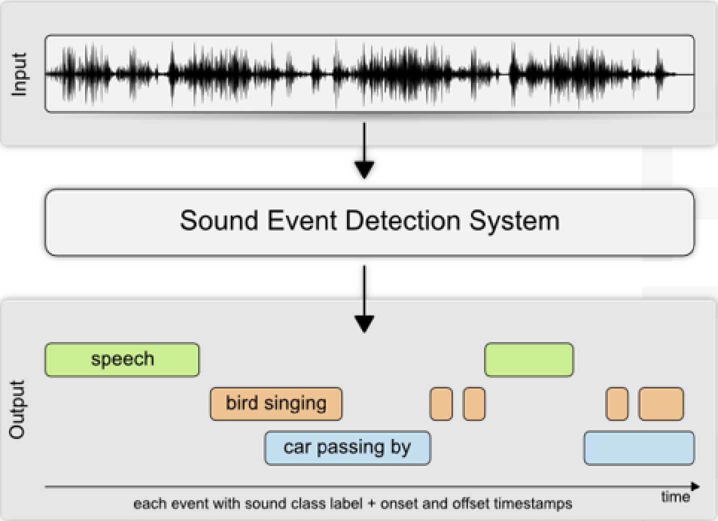
Fingerprint Recognition



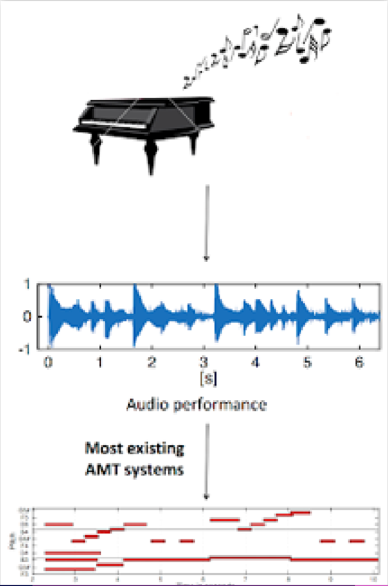
Audio Fingerprint



Audio Event Detection



Audio Event Detection



Translation