Case Study: Speech data and CNN

M.Tech. Artificial Intelligence, Second Year, NMIMS

By,

Bilal Hungund, Data Scientist, Halliburton

Convolution & operation

0	1	0	1	0	1
0	1	0	1	0	1
0	1	0	1	0	1
0	1	0	1	0	1
0	1	0	1	0	1
0	1	0	1	0	1

6x6 image

$\times$

Filter (weights)

-1	1	-1
2	3	2
1	1	1

3x3 Filter

$\Downarrow$ Output without padding

$n - f + 1$

$= 6 - 3 + 1$

$= 4$

Output Size

4	4	4	4
4	..	..	..
:	:	:	:
:	:	:	:

4x4

Stride = 1 (Step size)

With Padding

0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1

4x4 $\xrightarrow[\text{padding}]{}$ 6x6

Padding

$\times$

filter

1	-1
1	-1

$\Downarrow$

4x4

Pooling

25	48	11	58
192	10	20	110
38	0	9	31
50	8	23	47

25	48
192	10

11	58
20	110

38	0
50	8

9	31
23	47

Pooling ⇒

Max Pooling

192	110
50	47

69	50
22	28

Average Pooling

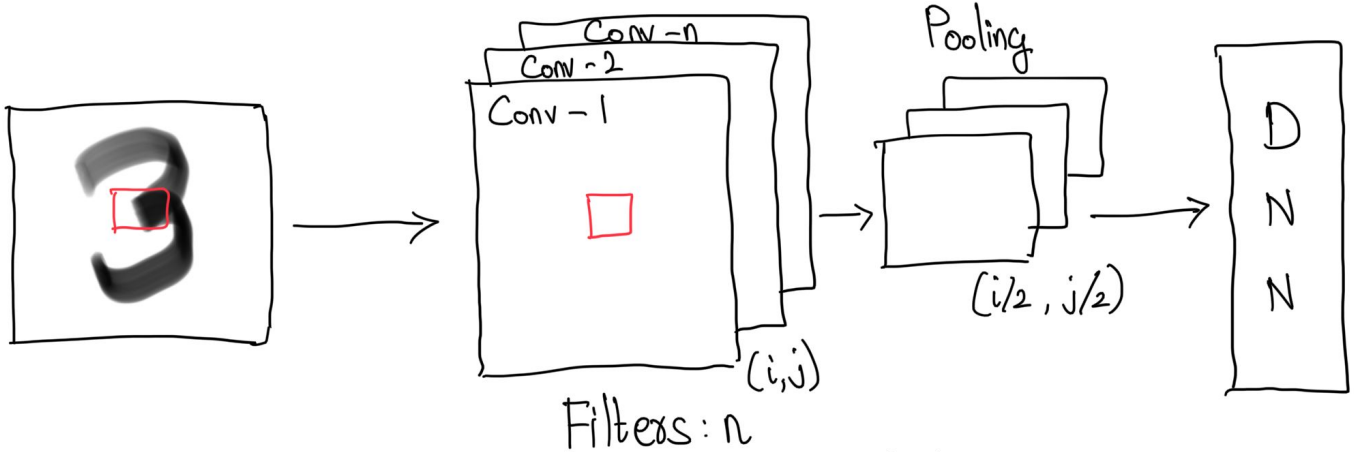
Stride = 2
(Recommended for Pooling)

Convolution Neural Network (CNN) for Classification

tf.keras.layers.Conv2D

tf.keras.activations.*

tf.keras.layers.MaxPool2D



1) Convolution: Filters to generate feature maps

2) Non-linearity: often relu

3) Backpropagation

4) Pooling: Downsampling feature maps

MFCC (Mel-frequency Ceptral Coefficients)

Mel Spectrogram

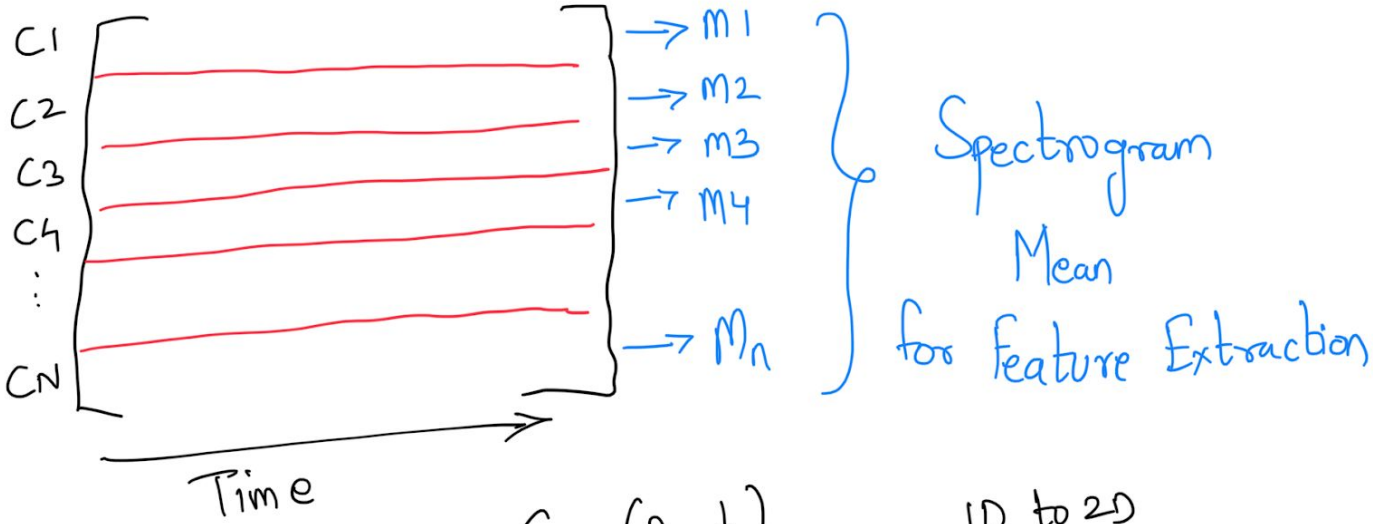
→ Spectrogram converted to Mel scale

→ Widely used in deep learning

→ Powerful tool to extract the feature from speech

→ Process includes: Fourier Transform, discrete cosine transforms and overlapping windows

→ It helps for classification problems such as genre classification, disease detection related to speech and etc.



$C = (n, t)$

$M = (n, \ ) \longrightarrow (k, \overbrace{n_1, n_2}^{\text{1D to 2D}}, \underline{\#\text{channels}})$

$k \downarrow$ #samples

#channels $\Downarrow$ CNN Model

CNN in Speech Data

→ Create features using MFCCs & Mel Spectrogram

→ Average of matrix