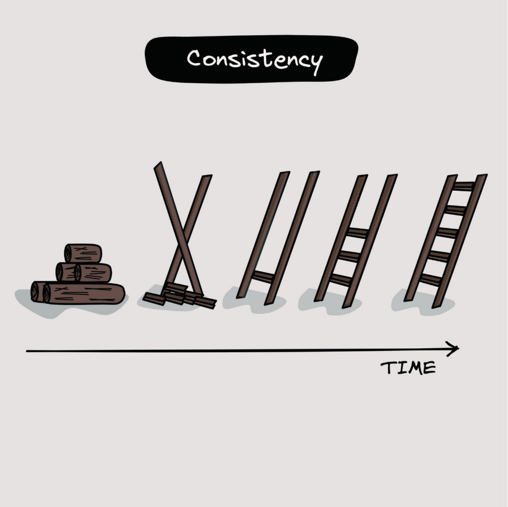
Case Study: Speech data and CNN

M.Tech. Data Science, 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

×

Filter (weights)

-1	1	-1
2	3	2
1	1	1

3x3 Filter

⇓ Output without padding

$n - f + 1$

$= 6 - 3 + 1$

$= 4$

Output Size

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

4x4

Stride = 1
(step size)

With Padding

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

4 x 4 ⟶ 6 x 6
padding

×

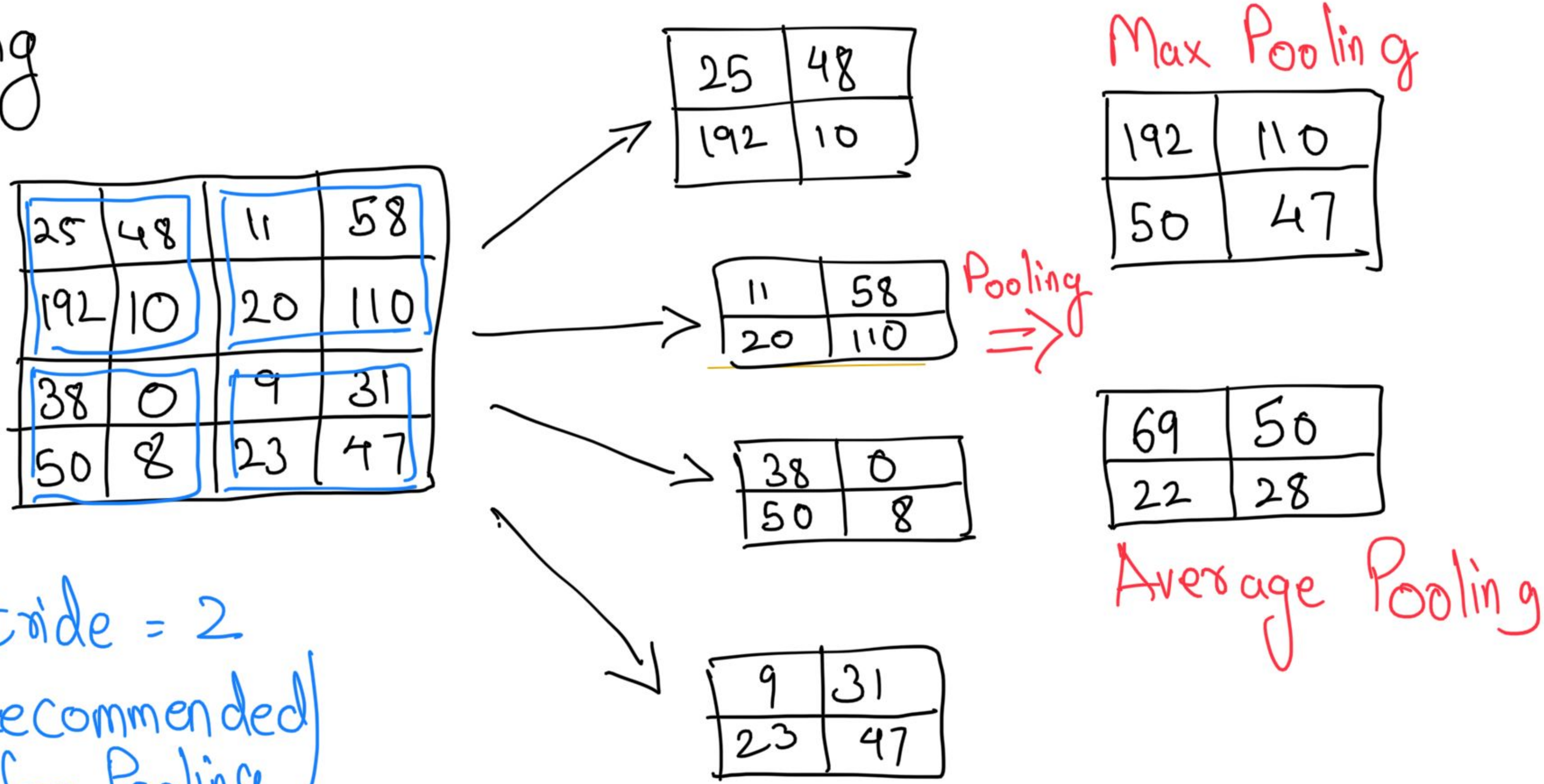
filter

1	-1
1	-1

⇓

4 x 4

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

Max Pooling

192	110
50	47

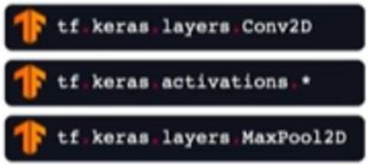
69	50
22	28

Average Pooling

Stride = 2

(Recommended for Pooling)

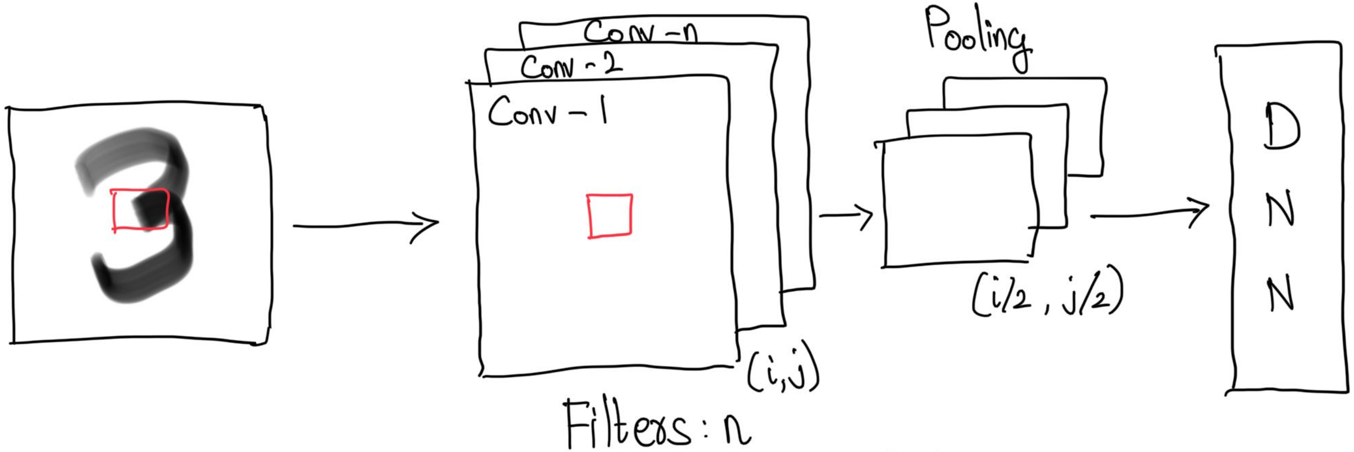
Convolution Neural Network (CNN) for Classification



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

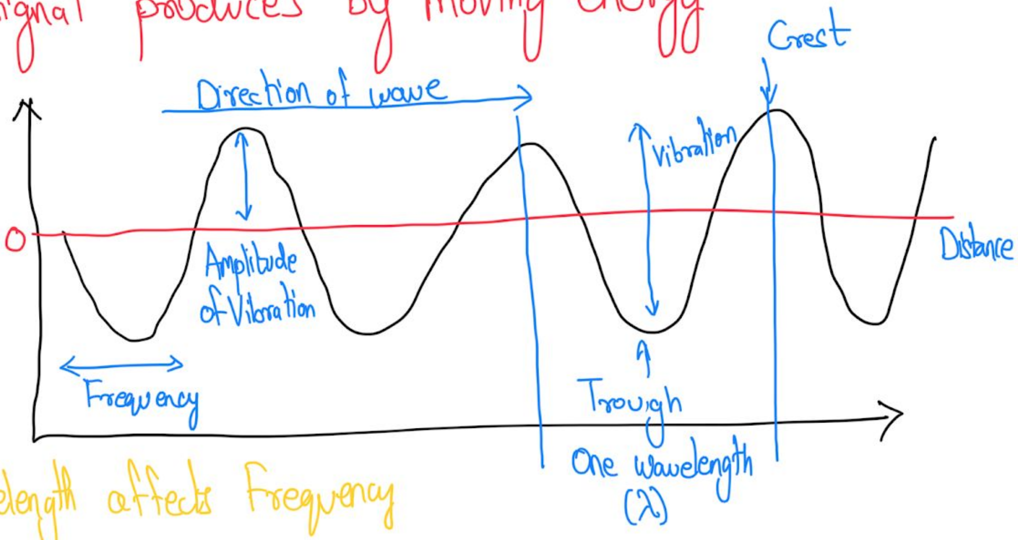
Audio Signal: (Automatic Speech Recognition)

Longitudinal vibration that produces vitality

Sound Wave:

Vibration signal produces by moving energy

Parameters
↳ Amplitude
Crest and Trough
Wavelength
Cycle
Frequency

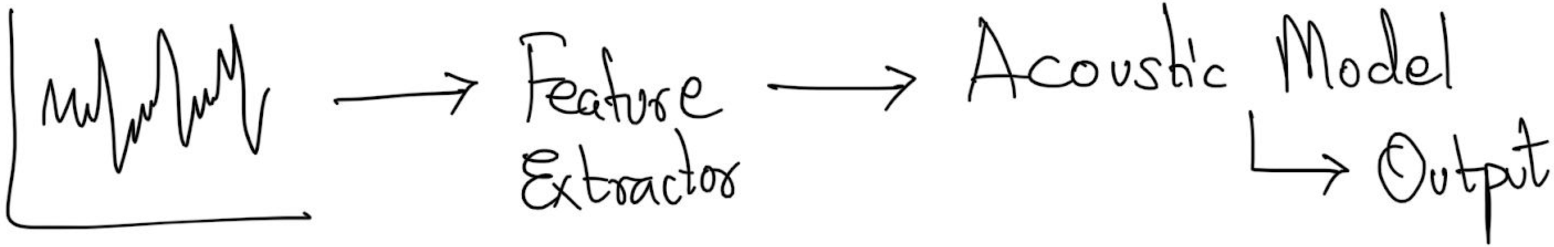


Wavelength affects Frequency

Acoustic Modelling

→ Statistical Representation of computed feature vector

HMM — Segmental Model — Super Segment Model — Neural Network



Feature Extractor → Acoustic Model ↳ Output

MFCC (Mel-frequency Ceptral Coefficients)

Mel Spectrogram

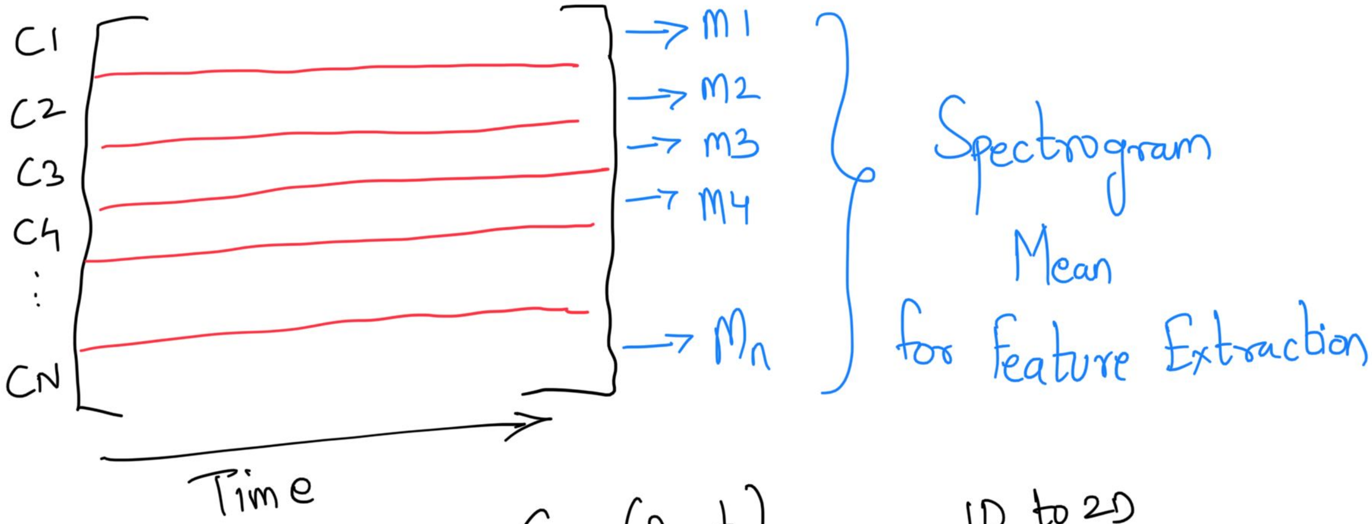
→ Spectrogram converted to Mel scale

→ 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.



$$C = (n, t)$$

$$M = (n, \ ) \longrightarrow (\underset{\downarrow \atop \#\text{samples}}{K}, \overbrace{n_1, n_2}^{\text{1D to 2D}}, \underline{\#\text{channels}})$$

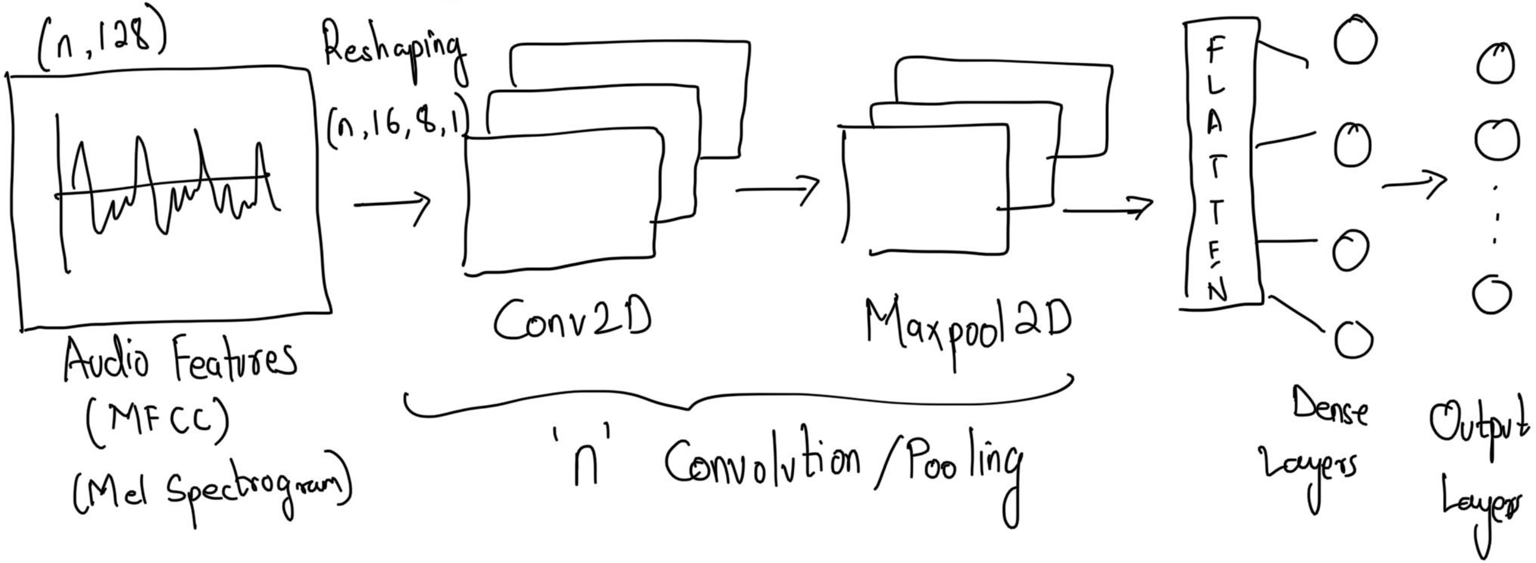
$$\Downarrow$$

CNN
Model

CNN in Speech Data

→ Create features using MFCCs & Mel Spectrogram

→ Average of matrix


(n,128)
Reshaping
(n,16,8,1)
Audio Features
(MFCC)
(Mel Spectrogram)
Conv2D
Maxpool2D
'n' Convolution / Pooling
FLATTEN
Dense Layers
Output Layers