BE WARE OF WHAT YOU HEAR

Deepfake Audio Detection

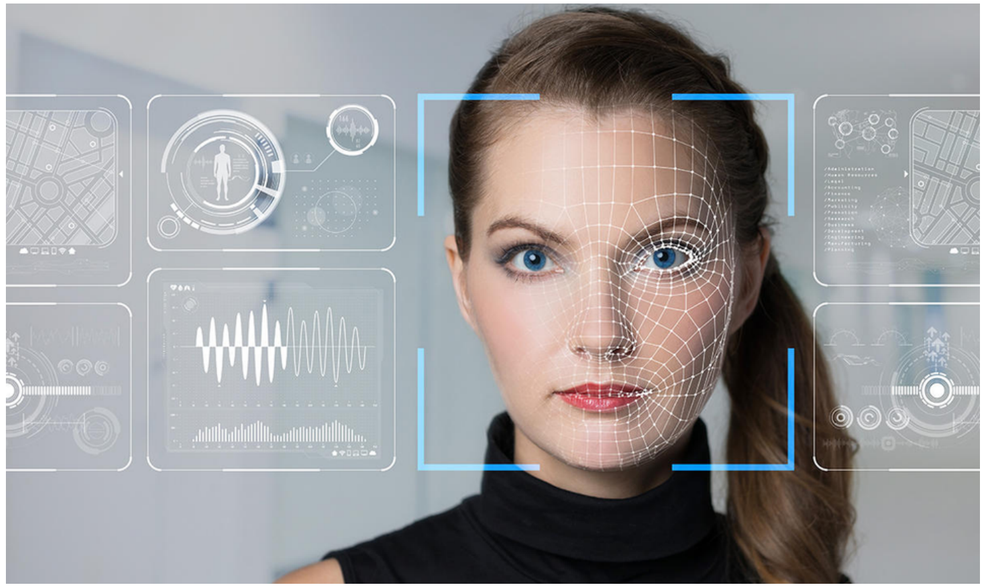
the voice may not belong to a human

mentor: Brandon Wu, Jerry Liao
members: Ching-yuan Pai, Yipin Peng, Andy Chen, Evian Chen

Introduction

What is Deepfake?

Deepfake is a technique for human image synthesis based on artificial intelligence.



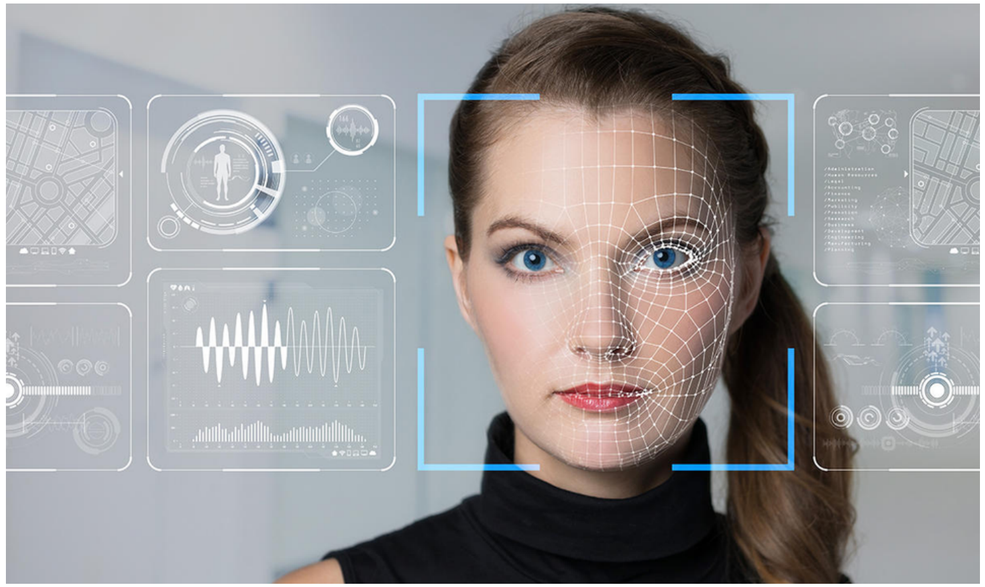
Generation system

- ITTS (text-to-speech synthesis) systems
- IVC (voice conversion) system
- ITTS-VC systems
- limplemented with various waveform generation methods including classical vocoding, GriffinLim, generative adversarial networks, neural waveform models, waveform concatenation, waveform filtering, spectral filtering, and their combination.

Introduction

What is Deepfake?

Deepfake is a technique for human image synthesis based on artificial intelligence.



Generation system

- ITTS (text-to-speech synthesis) systems
- IVC (voice conversion) system
- ITTS-VC systems
- limplemented with various waveform generation methods including classical vocoding, GriffinLim, generative adversarial networks, neural waveform models, waveform concatenation, waveform filtering, spectral filtering, and their combination.

Examples

TTS (text-to-speech synthesis) systems

[Thies et al. 2019]

Examples

TTS (text-to-speech synthesis) systems and neural network-based system

Reference

REAL

FAKE

Examples

TTS (text-to-speech synthesis) systems and neural network-based system

1. There were many editions of these works still being used in the nineteenth century.

2. Take a look at these pages for crooked creek drive.

REAL

FAKE

Examples

TTS (text-to-speech synthesis) systems and neural network-based system

1. There were many editions of these works still being used in the nineteenth century.

2. Take a look at these pages for crooked creek drive.

REAL

FAKE

Pros and Cons

of AI generated audio

Pros

1. Make the voice (siri, Miss Google) more realistic and nature
2. Assist in editing and repairing videos


THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Podcasts | Latest Headlines | More

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & Work Style Sports

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

Cons

1. Election Manipulation
2. Celebrity Pornography
3. Scams and Hoaxes
4. Financial Fraud
5. Social Engineering
6. Automated Disinformation Attacks
7. Identity Theft

Pros and Cons

of AI generated audio

Pros

1. Make the voice (siri, Miss Google) more realistic and nature
2. Assist in editing and repairing videos


THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Podcasts | Latest Headlines | More

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & Work Style Sports

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

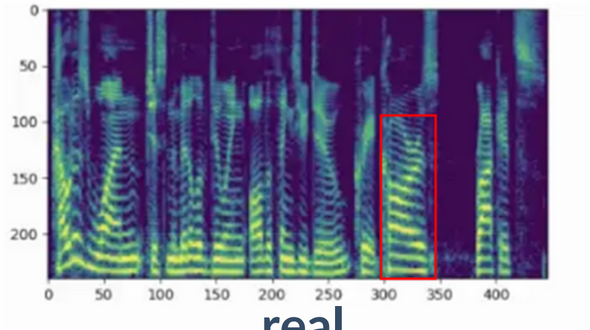
Cons

1. Election Manipulation
2. Celebrity Pornography
3. Scams and Hoaxes
4. Financial Fraud
5. Social Engineering
6. Automated Disinformation Attacks
7. Identity Theft

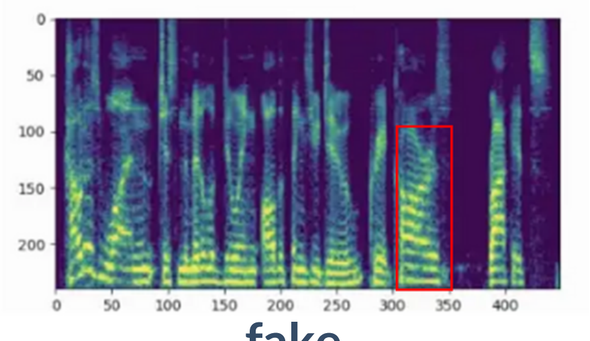
How to prevent?

Objectives

- Using Convolutional Neural Network (CNN) model for audio deepfake detection.
- Comparing the effects from training spectrograms and waveforms.
- Building a simple UI for users.



real



fake

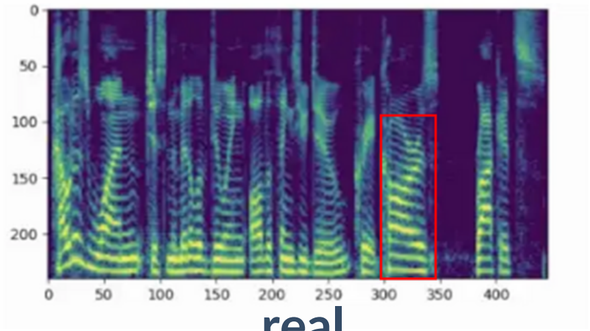
Flow chart

- Data preprocessing
  - Separate dataset into train data and validation data
  - Transform data type: .flac→.wav → .png (waveform, sectrogram)
  - labeling
- Train
- Validate (Acc: 90 up)
- Save model

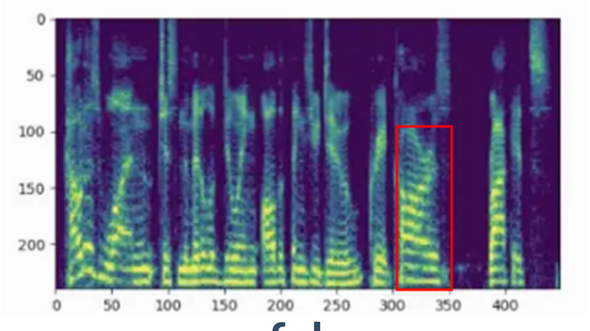
How to prevent?

Objectives

- Using Convolutional Neural Network (CNN) model for audio deepfake detection.
- Comparing the effects from training spectrograms and waveforms.
- Building a simple UI for users.



real



fake

Flow chart

- Data preprocessing
  - Separate dataset into train data and validation data
  - Transform data type: .flac→.wav → .png (waveform, sectrogram)
  - labeling
- Train
- Validate (Acc: 90 up)
- Save model and import in website

Data Introduction

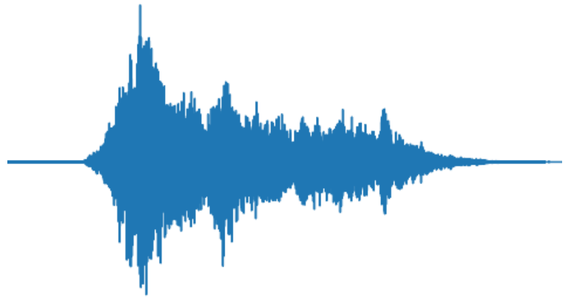
different data types

Flac format

Flac, free lossless audio codec, is an audio coding format for lossless compression of digital audio. Flac is royalty-free and is considered the preferred format for downloading and storing hi-res albums.

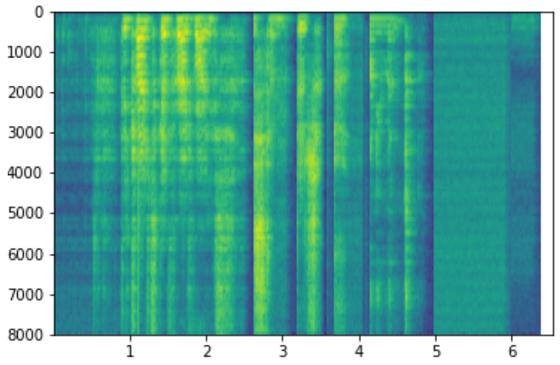
Waveform

waveform is a graphical representation of the shape and form of a signal moving. For sound, the term describes a depiction of the pattern of sound pressure variation (or amplitude) in the time domain.

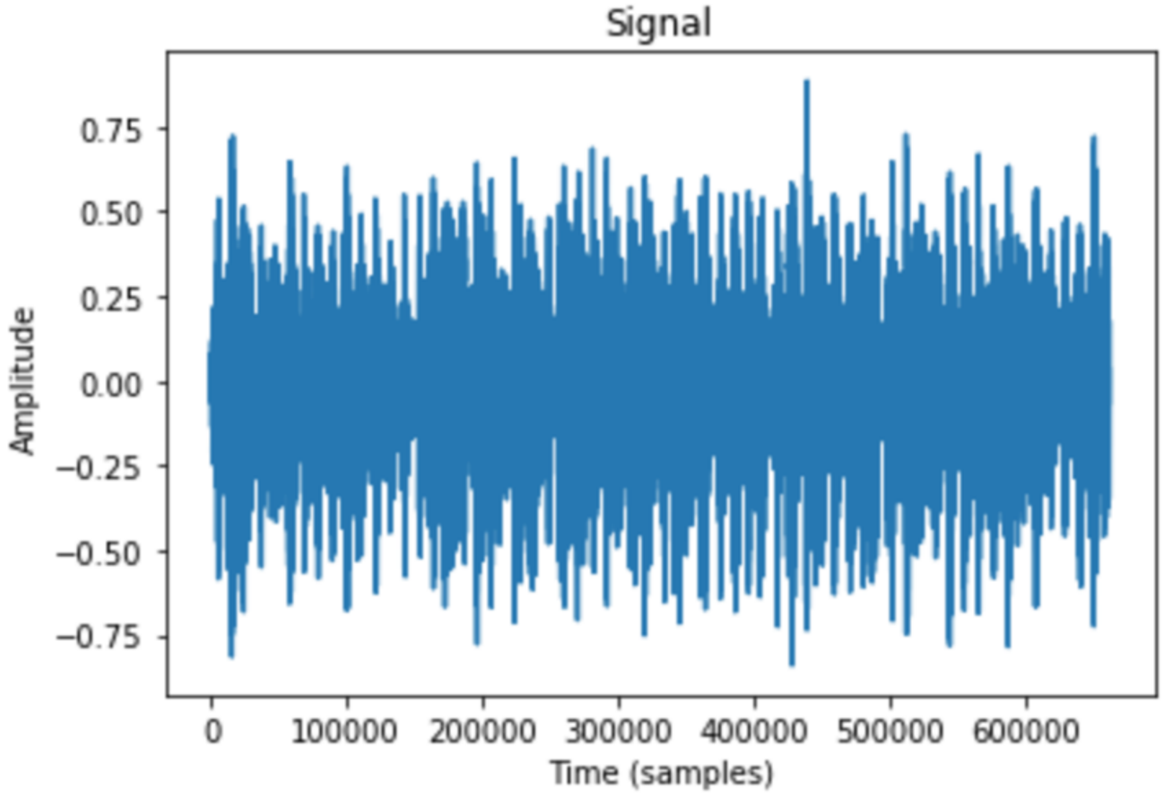


Spectrogram

Spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. It is usually depicted as a heat map, as an image with the intensity shown by varying the colour or brightness.



Feature Extraction of Audio Data



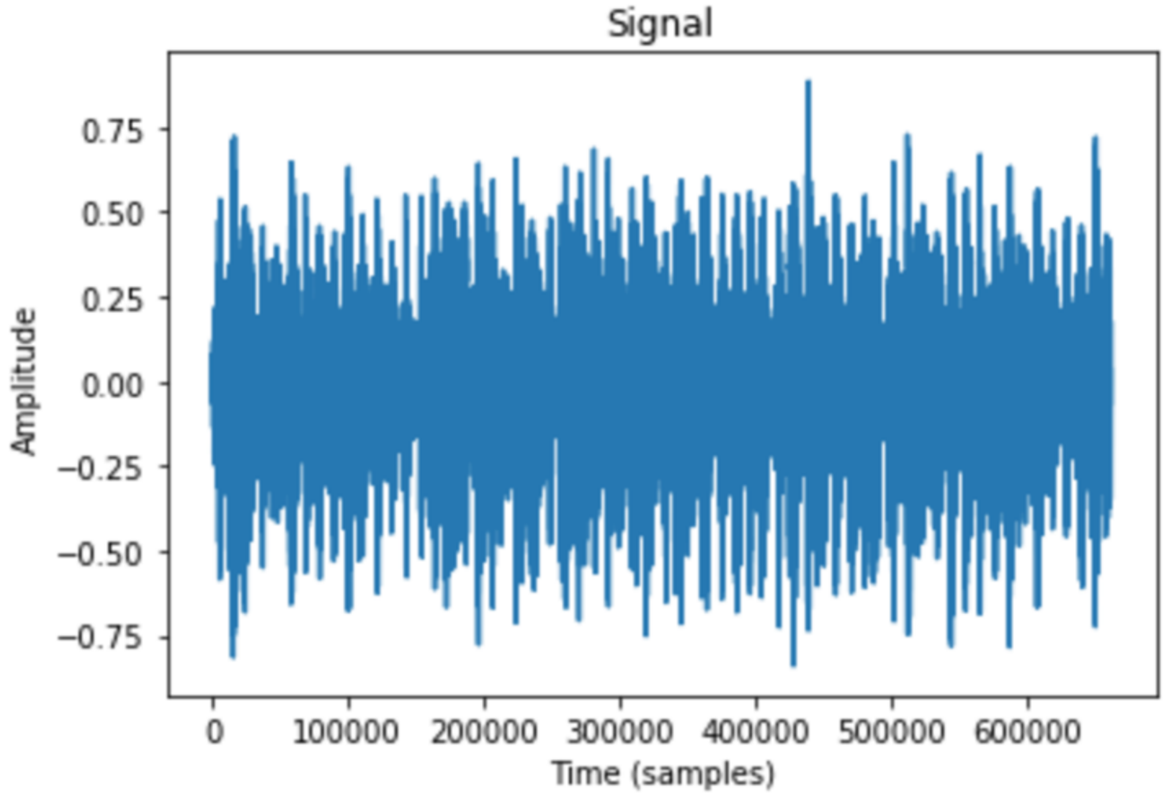
Objective

To extract useful information from raw audio data (wave form), which includes only amplitudes by time.

Transformation of Audio Data

- MFCC
- LFCC
- Spectrogram

Feature Extraction of Audio Data



Objective

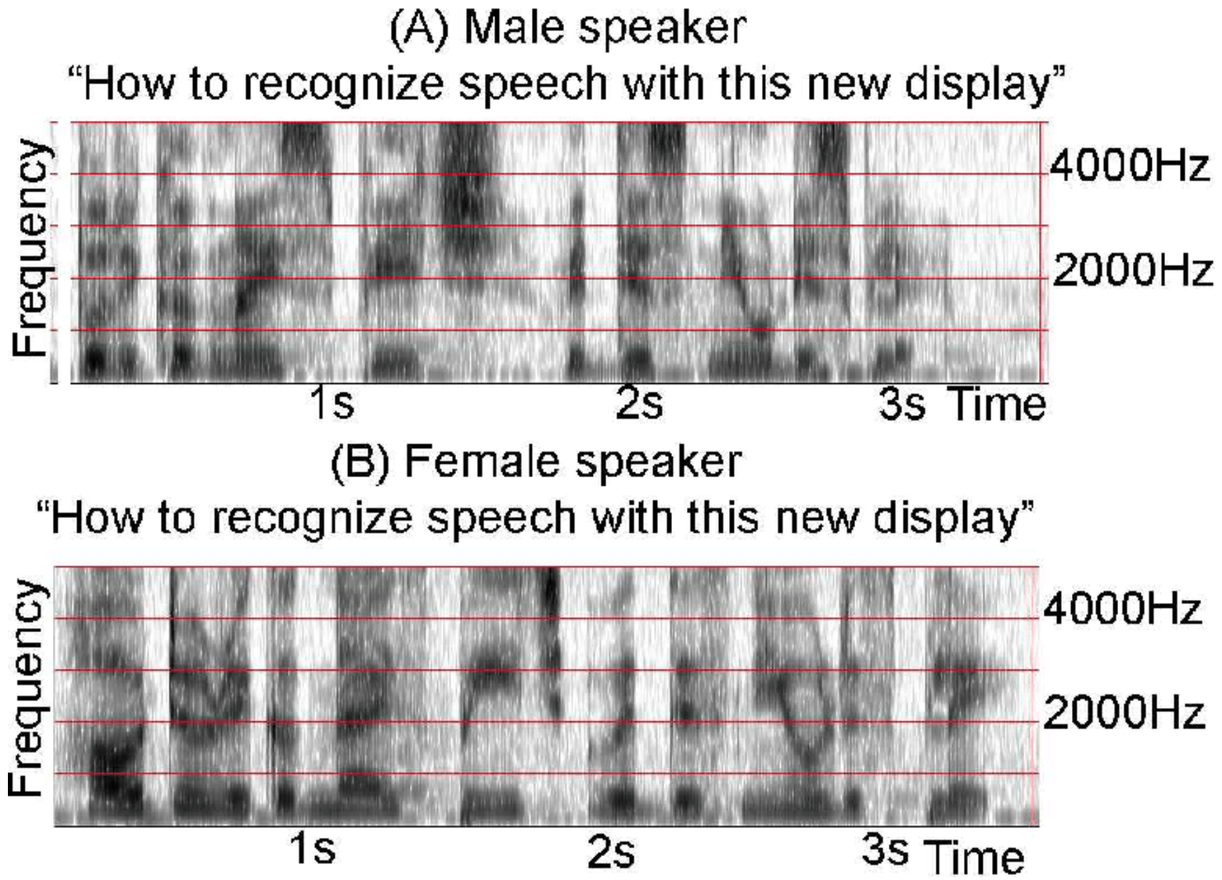
To extract useful information from raw audio data (wave form), which includes only amplitudes by time.

Transformation of Audio Data

- MFCC
- LFCC
- Spectrogram

Why Spectrogram?

To be able to identify subtle differences between audio signals.



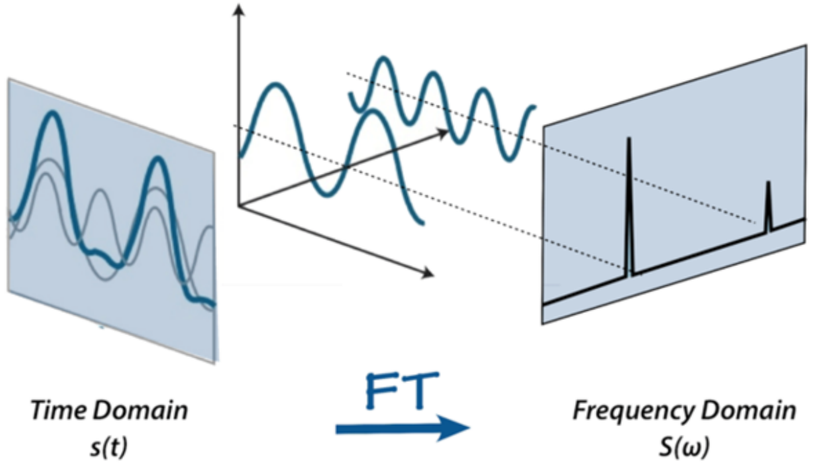
Audio Transform

fourier transform(FT) & fast fourier transform(FFT)

fourier transform

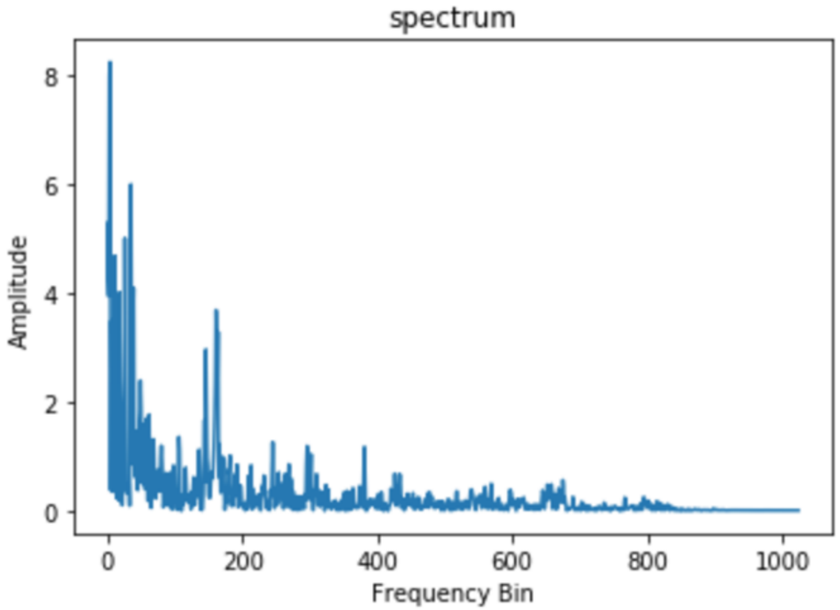
A mathematical formula that allows us to decompose an audio signal into its individual frequencies and the amplitudes of each frequency.

The result we called it spectrum



fast fourier transform

An algorithm that can efficiently compute the Fourier transform.

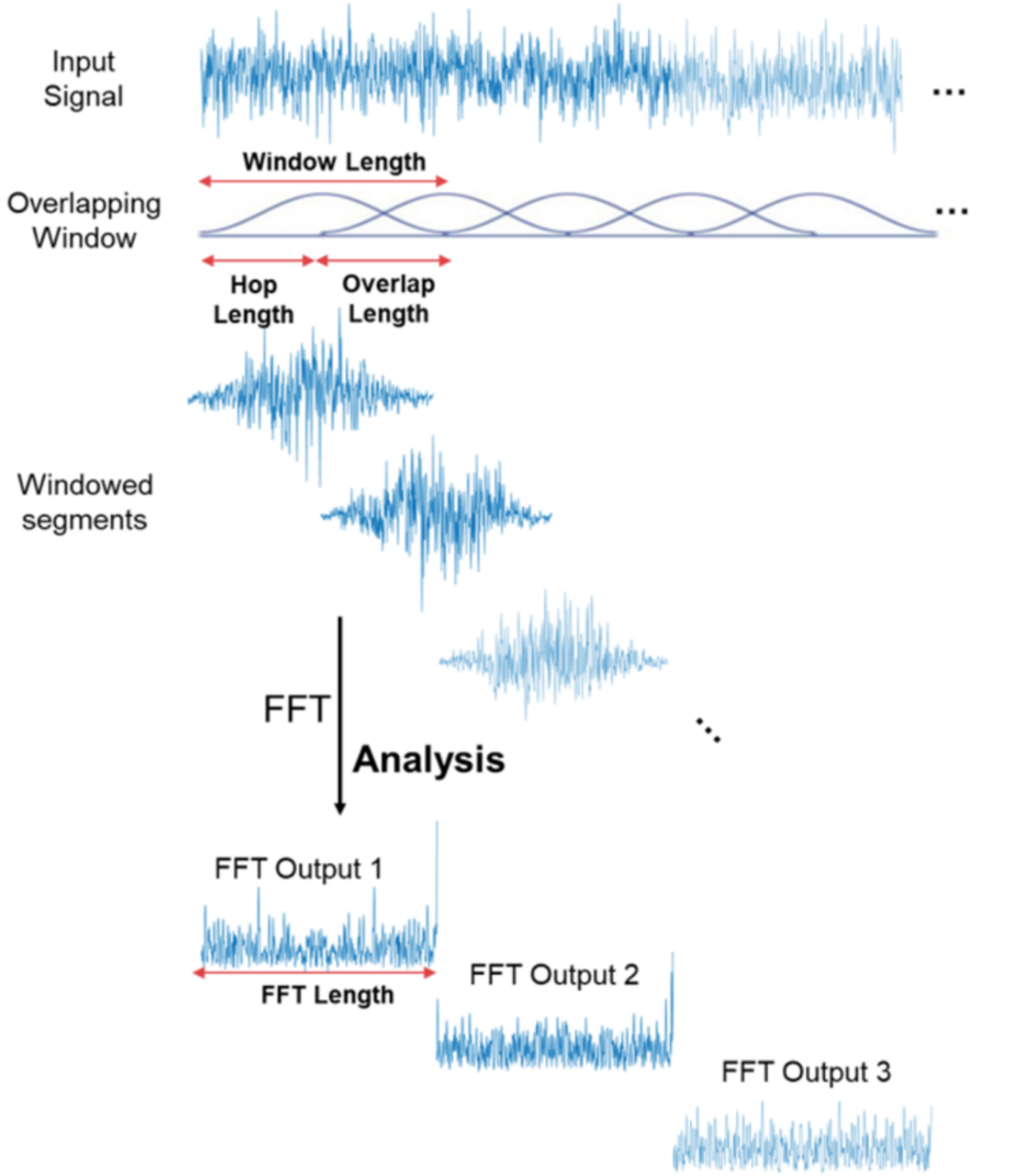


Short-Time Fourier Trasform

STFT

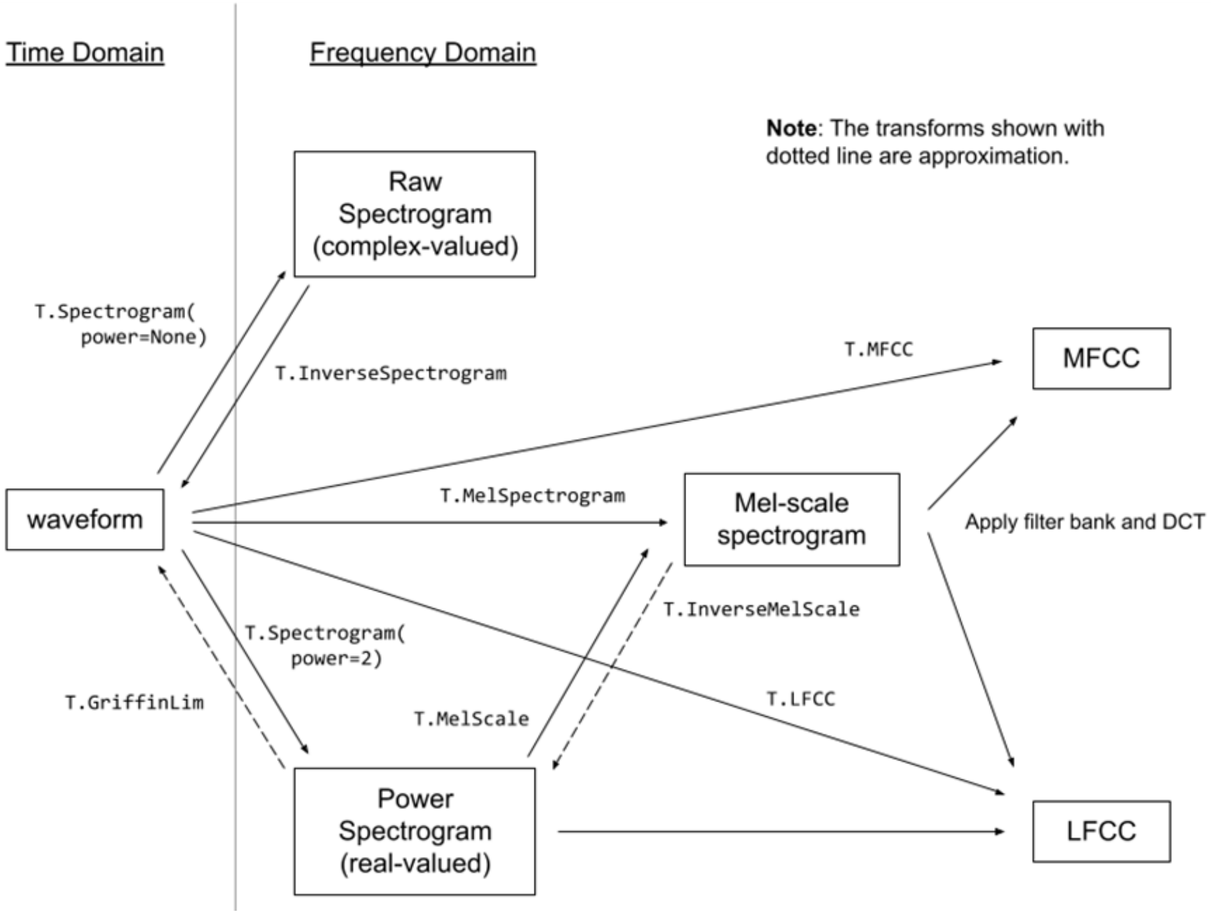
To monitor changes of frequency by time under an audio signal

- Segment audio signal by a given length of time (Window Length), and transform each chunk of signal into spectrums through algorithm FFT.
- Stack every FFT Output together to get a full picture of frequencies changing over time.



Torch Audio

PyTorch Module for Audio Data



- torchaudio.transform.spectrogram()
  →tensor
  - n_fft: size of FFT
  - win_length: window length
- torchaudio.load()
  →(signal, sample rate)
- torchaudio.transform.Resample()

Dataset Introduction

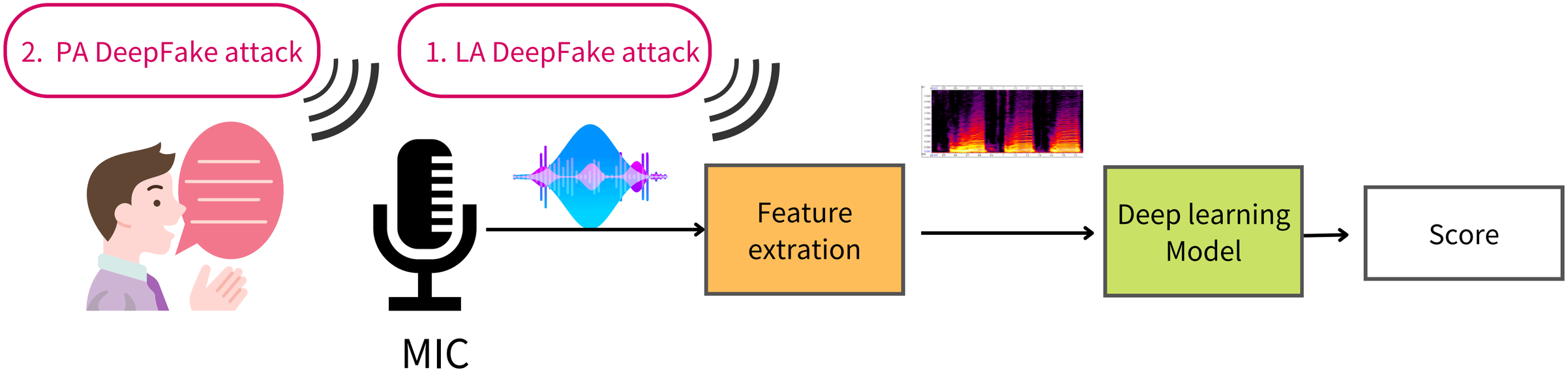
ASVspoof2019

Physical access (PA)

Bonafide utterances are made in a real, physical space in which spoofing attacks are captured and then replayed within the same physical space using replay devices of varying quality.

Logical access (LA)

Bonafide and spoofed utterances generated using text-to-speech (TTS) and voice conversion (VC) algorithms are communicated across telephony and VoIP networks with various coding and transmission effects.



Model Training

by physical access senario dataset

01. flac to wave

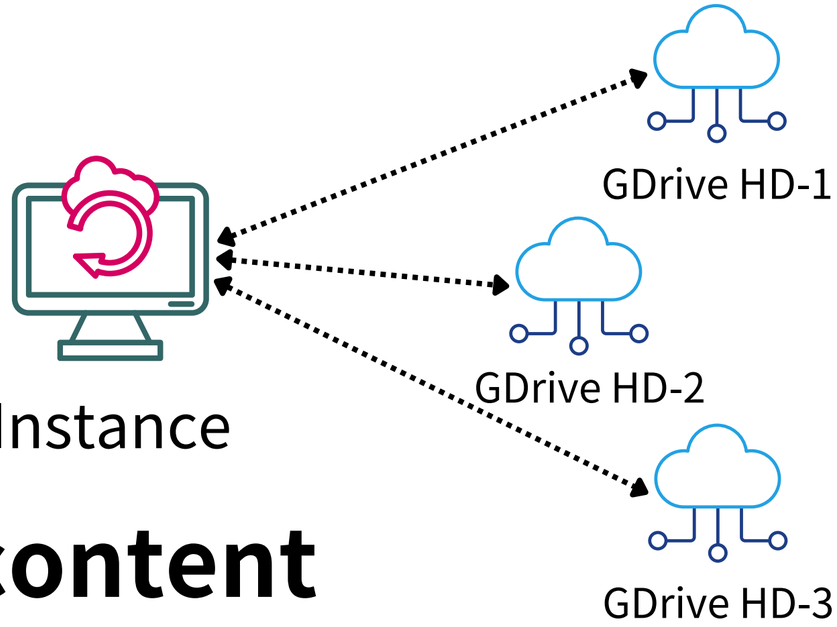
02. wave to waveform and spectrogram

03. training by VGG-like model and ResNet50

Since ResNet 50 can be pretrained, our strategy is converting all audio files into images. Additionally, spectrogram and waveform contain various sound features, which can be extracted by model, making it reach a higher accuracy.

Colab IO problem

Why file loading so long?



Solution

Using !Zip and !CP to accelator IO

!zip LA.zip

!cp LA.zip /content

!unzip LA.zip

Your data is stored in multiple GDrive HD

There is a high chance causing loading failed and run-time error

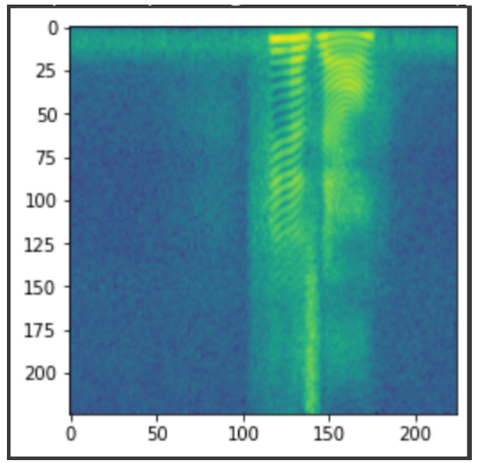
File loading time from 1HR -> 5 Min!

Data PreProccessing

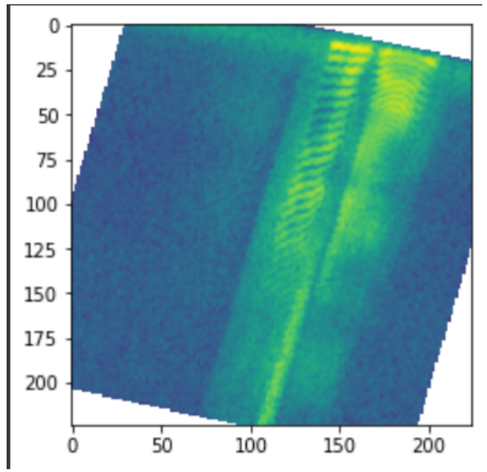
data introduction

Human: 2530， DeepFake: 22800 ACC:89%

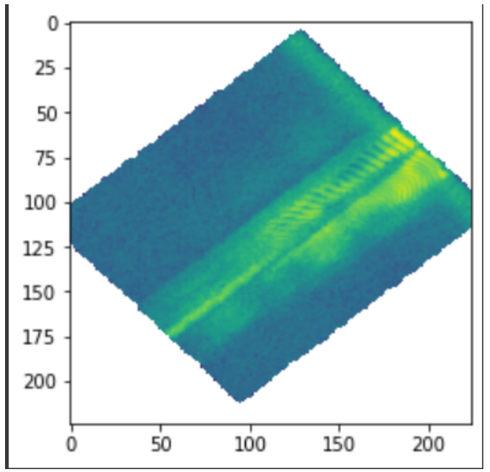
	Predict Human	Predict DeepFake
Human	0%	100%
DeepFake	0%	100%



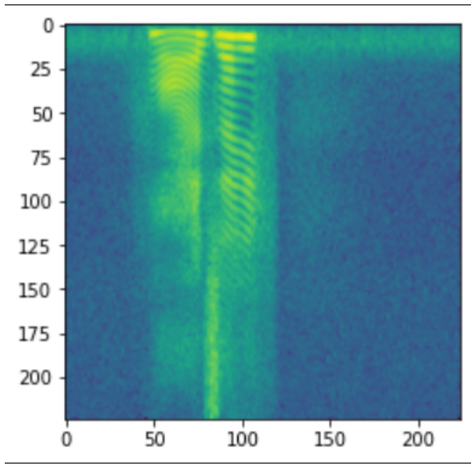
Resize: 224



RandomRotation



RandomAffine



HorizontalFlip

Human: 20640， DeepFake: 22800

```
class customDataset(Dataset):
  def __init__():
  def __getitem__(self, index):
    return self.data[index], self.label[index]
  def __len__(self):
```

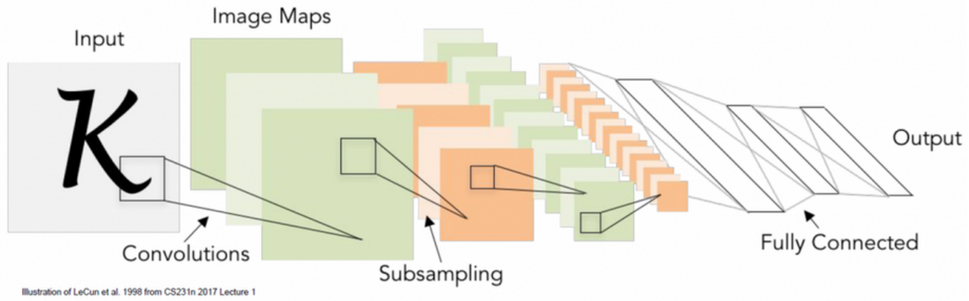
VGG-like CNN Model

data training

27-Layer CNN model training

Human: 20640， DeepFake: 22800 Total ACC: 90%

	Predict Human	Predict DeepFake
True Human	91%	9%
True DeepFake	10%	90%



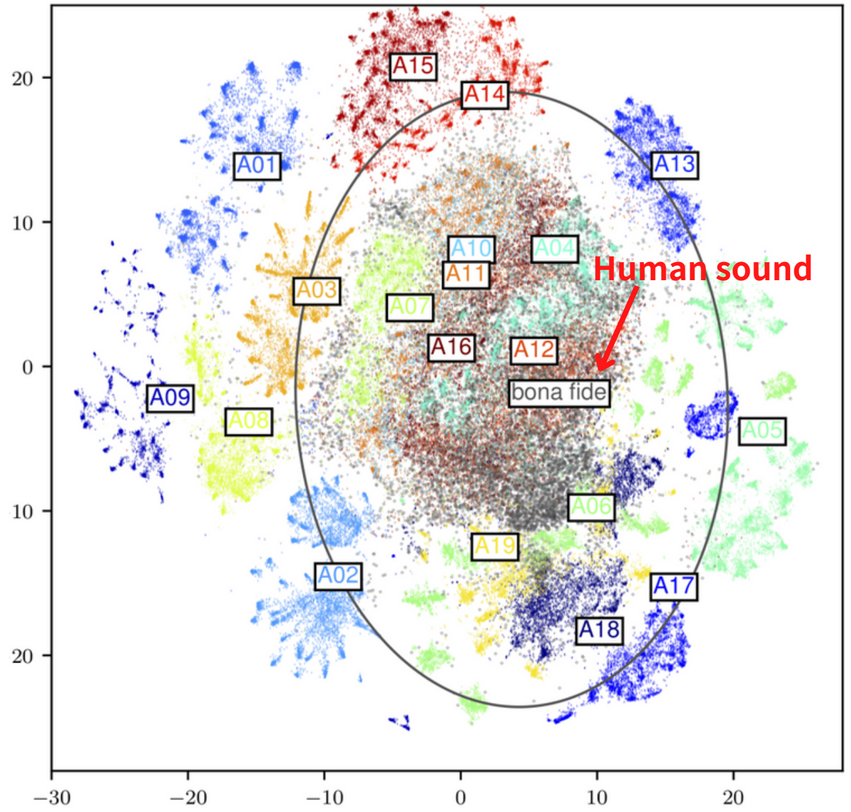
Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 16, 224, 224]	160
BatchNorm2d-2	[-1, 16, 224, 224]	32
ReLU-3	[-1, 16, 224, 224]	0
Conv2d-4	[-1, 16, 224, 224]	2,320
BatchNorm2d-5	[-1, 16, 224, 224]	32
ReLU-6	[-1, 16, 224, 224]	0
MaxPool2d-7	[-1, 16, 112, 112]	0
Conv2d-8	[-1, 32, 112, 112]	4,640
BatchNorm2d-9	[-1, 32, 112, 112]	64
ReLU-10	[-1, 32, 112, 112]	0
Conv2d-11	[-1, 32, 112, 112]	9,248
BatchNorm2d-12	[-1, 32, 112, 112]	64
ReLU-13	[-1, 32, 112, 112]	0
MaxPool2d-14	[-1, 32, 56, 56]	0
Conv2d-15	[-1, 64, 56, 56]	18,496
BatchNorm2d-16	[-1, 64, 56, 56]	128
ReLU-17	[-1, 64, 56, 56]	0
MaxPool2d-18	[-1, 64, 28, 28]	0
Conv2d-19	[-1, 64, 28, 28]	36,928
BatchNorm2d-20	[-1, 64, 28, 28]	128
ReLU-21	[-1, 64, 28, 28]	0
MaxPool2d-22	[-1, 64, 14, 14]	0
Conv2d-23	[-1, 128, 14, 14]	73,856
BatchNorm2d-24	[-1, 128, 14, 14]	256
ReLU-25	[-1, 128, 14, 14]	0
MaxPool2d-26	[-1, 128, 7, 7]	0
Conv2d-27	[-1, 128, 7, 7]	147,584
BatchNorm2d-28	[-1, 128, 7, 7]	256
ReLU-29	[-1, 128, 7, 7]	0
Flatten-30	[-1, 6272]	0
Linear-31	[-1, 2]	12,546

Does multiply work?

Random 3000 model

Human: 2530， DeepFake: 3000　Total ACC: 91%

- DeepFake sound file was generated from 13 types of AI model Named from A07-A19



Graph from ***ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan*** shows that different DeefFake models performed.

	CNN-Multiply	CNN-3000
Human	Acc 0.9110	Acc 0.9547
A07	Acc 0.9988	Acc 0.9833
A09	Acc 0.9988	Acc 0.9996
A10	Acc 0.8115	Acc 0.9580
A11	Acc 0.9424	Acc 0.9656
A12	Acc 0.7912	Acc 0.9762
A13	Acc 0.7597	Acc 0.9909
A14	Acc 0.9278	Acc 0.9996
A15	Acc 0.9174	Acc 0.9657
A16	Acc 0.9424	Acc 0.9656
A17	Acc 0.9364	Acc 0.7322
A18	Acc 0.5345	Acc 0.2441
A19	Acc 0.9846	Acc 0.8459
Total:	90%	91%

Model Setup

ResNet50

ResNet is a powerful CNN that contains 150+ layers and overcome the problem of vanishing gradient. ResNet includes skip connection feature which enables training of mutiple deep layers.

Why ResNet?

- training a deep CNN is difficult
- more expressive, less difference
- vanishing gradient
- won the ImageNet challenge in 2015

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Model Setup

ResNet50

ResNet is a powerful CNN that contains 150+ layers and overcome the problem of vanishing gradient. ResNet includes skip connection feature which enables training of mutiple deep layers.

Why ResNet?

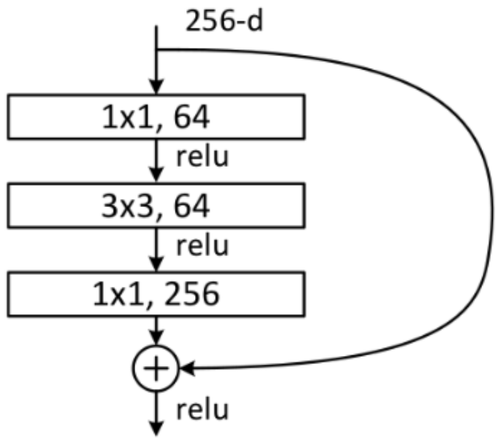
- training a deep CNN is difficult
- more expressive, less difference
- vanishing gradient
- won the ImageNet challenge in 2015

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

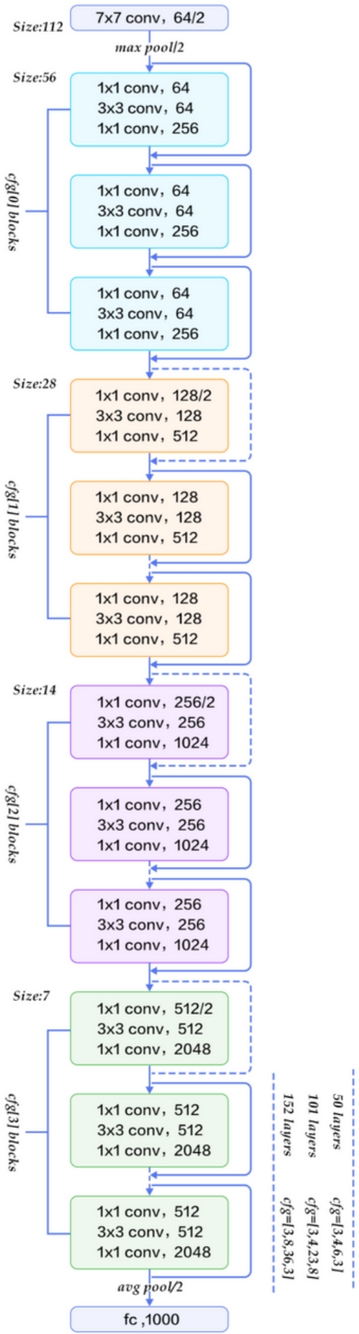
Model Setup

How ResNet50 Works?

- 7×7 kernel, 64 kernels with a 2-sized stride.
- A max pooling layer with a 2-sized stride.
- 3×3,64 kernel, 1×1,64 kernels, 1×1,256 kernels, iterated 3 times. (9)
- 1×1,128 kernels, 3×3,128 kernels,1×1,512 kernels, iterated 4 times.(12)
- 1×1,256 cores, 2 cores 3×3,256, 1×1,1024, iterated 6 times. (8)
- 1×1,512 cores, 3×3,512 cores, 1×1,2048 cores iterated 3 times. (9)
- Average pooling
- Fully connected



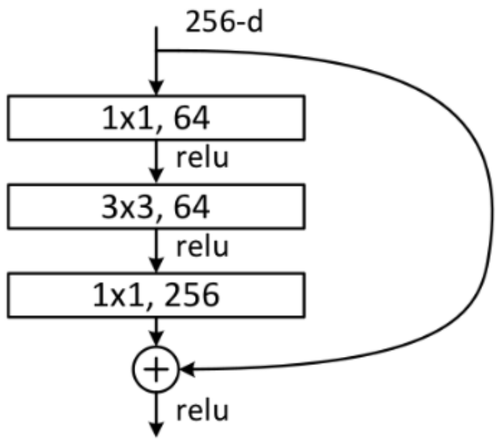
shortcut connection



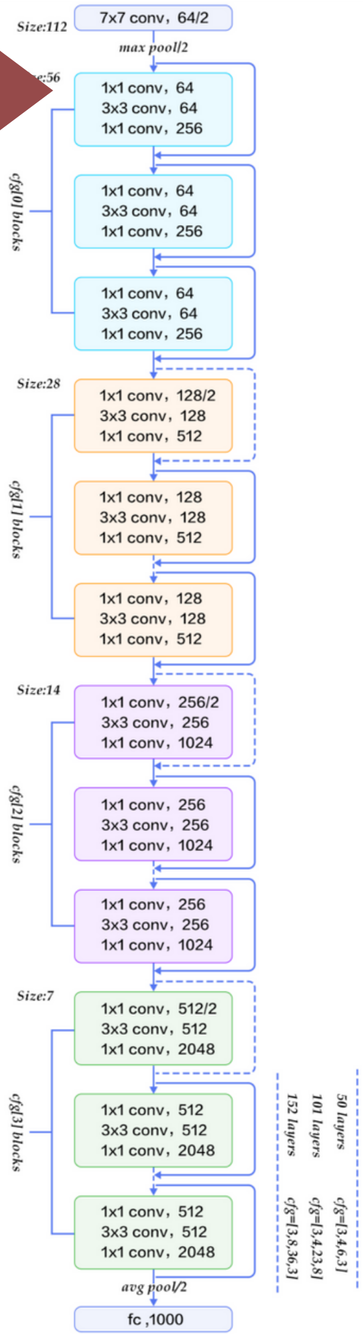
Model Setup

How ResNet50 Works?

- 7×7 kernel, 64 kernels with a 2-sized stride.
- A max pooling layer with a 2-sized stride.
- 3×3,64 kernel, 1×1,64 kernels, 1×1,256 kernels, iterated 3 times. (9)
- 1×1,128 kernels, 3×3,128 kernels,1×1,512 kernels, iterated 4 times.(12)
- 1×1,256 cores, 2 cores 3×3,256, 1×1,1024, iterated 6 times. (8)
- 1×1,512 cores, 3×3,512 cores, 1×1,2048 cores iterated 3 times. (9)
- Average pooling
- Fully connected



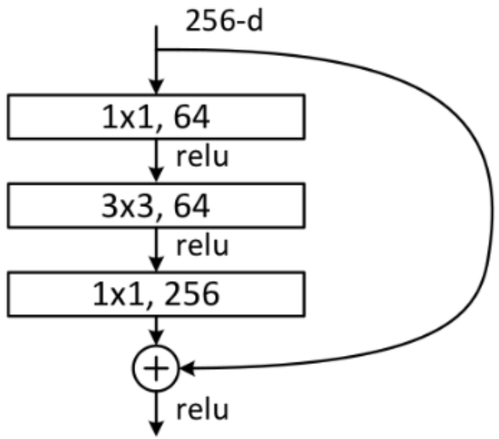
shortcut connection



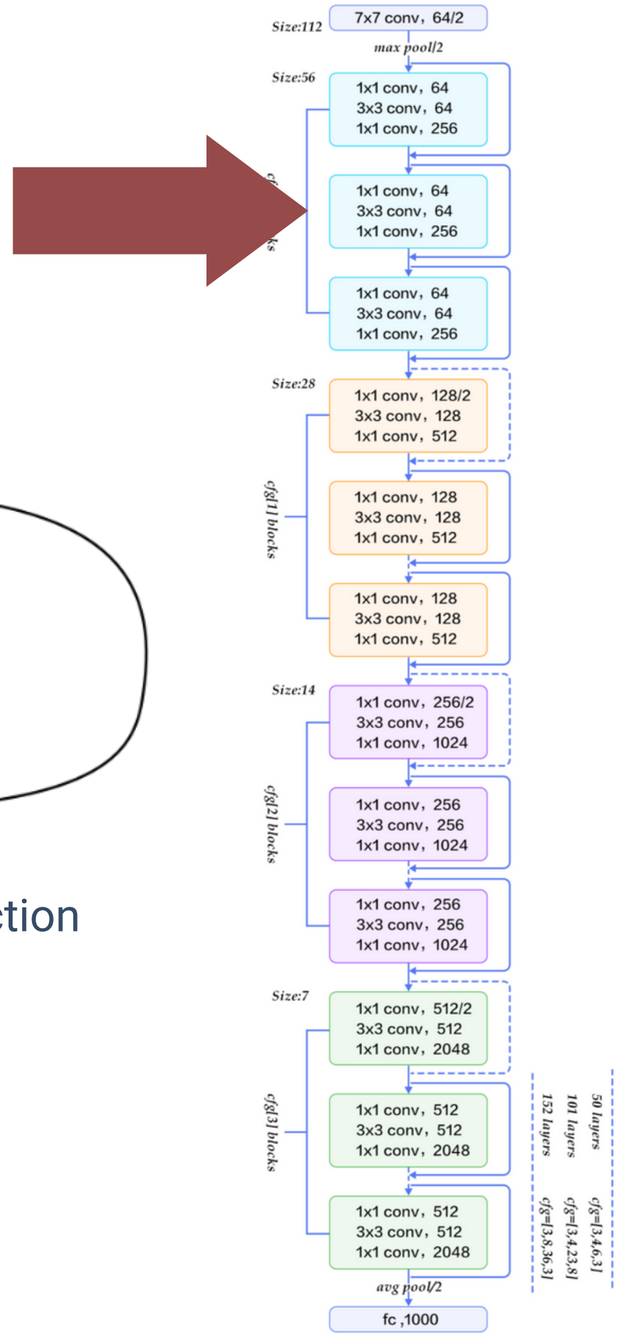
Model Setup

How ResNet50 Works?

- 7×7 kernel, 64 kernels with a 2-sized stride.
- A max pooling layer with a 2-sized stride.
- 3×3,64 kernel, 1×1,64 kernels, 1×1,256 kernels, iterated 3 times. (9)
- 1×1,128 kernels, 3×3,128 kernels,1×1,512 kernels, iterated 4 times.(12)
- 1×1,256 cores, 2 cores 3×3,256, 1×1,1024, iterated 6 times. (8)
- 1×1,512 cores, 3×3,512 cores, 1×1,2048 cores iterated 3 times. (9)
- Average pooling
- Fully connected



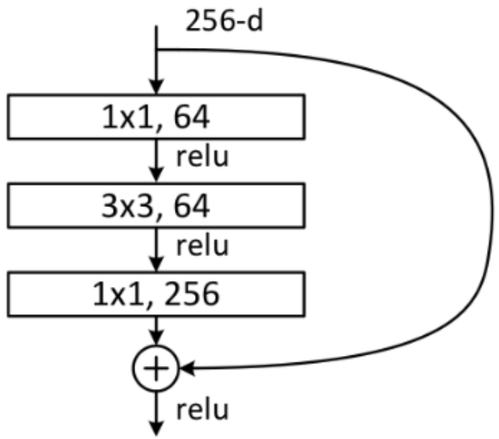
shortcut connection



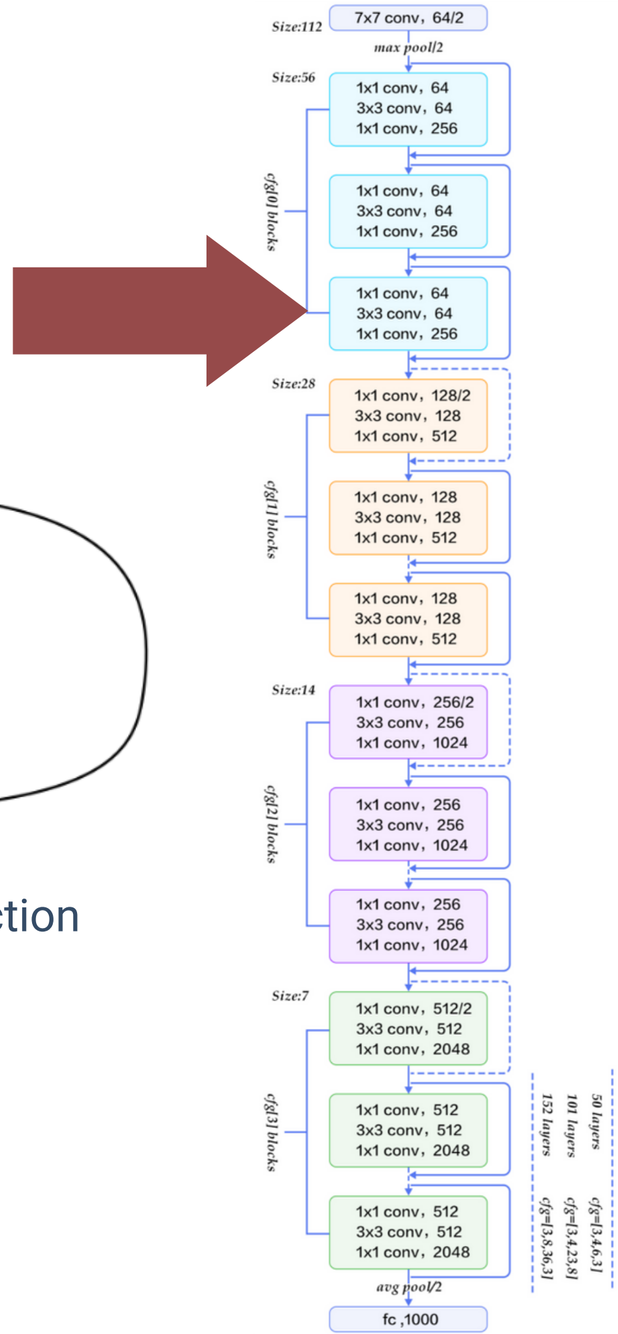
Model Setup

How ResNet50 Works?

- 7×7 kernel, 64 kernels with a 2-sized stride.
- A max pooling layer with a 2-sized stride.
- 3×3,64 kernel, 1×1,64 kernels, 1×1,256 kernels, iterated 3 times. (9)
- 1×1,128 kernels, 3×3,128 kernels,1×1,512 kernels, iterated 4 times.(12)
- 1×1,256 cores, 2 cores 3×3,256, 1×1,1024, iterated 6 times. (8)
- 1×1,512 cores, 3×3,512 cores, 1×1,2048 cores iterated 3 times. (9)
- Average pooling
- Fully connected



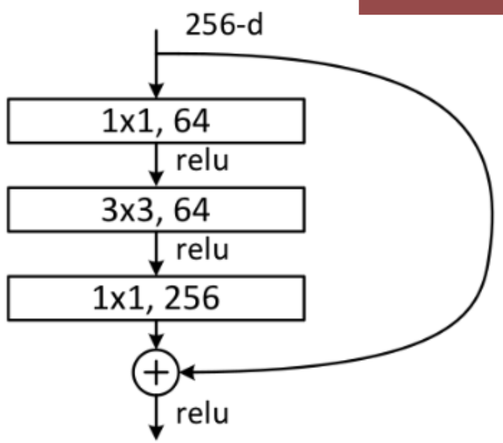
shortcut connection



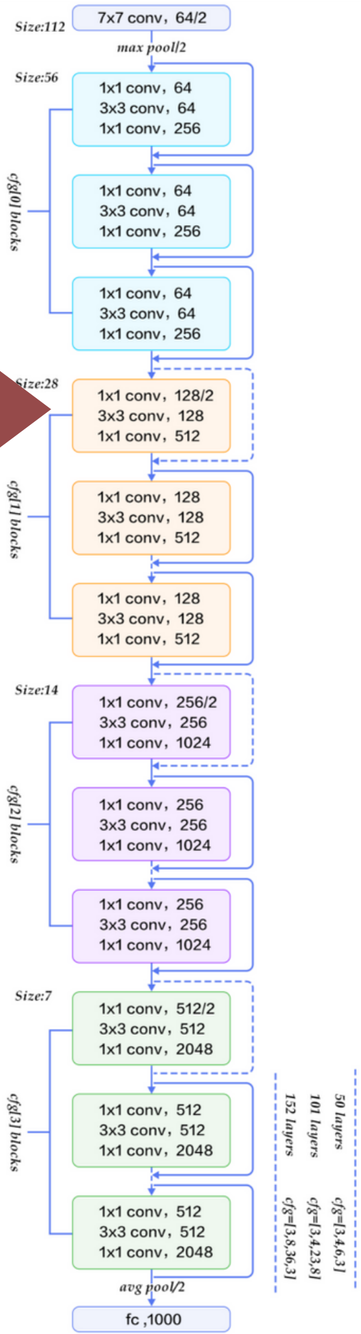
Model Setup

How ResNet50 Works?

- 7×7 kernel, 64 kernels with a 2-sized stride.
- A max pooling layer with a 2-sized stride.
- 3×3,64 kernel, 1×1,64 kernels, 1×1,256 kernels, iterated 3 times. (9)
- 1×1,128 kernels, 3×3,128 kernels,1×1,512 kernels, iterated 4 times.(12)
- 1×1,256 cores, 2 cores 3×3,256, 1×1,1024, iterated 6 times. (8)
- 1×1,512 cores, 3×3,512 cores, 1×1,2048 cores iterated 3 times. (9)
- Average pooling
- Fully connected



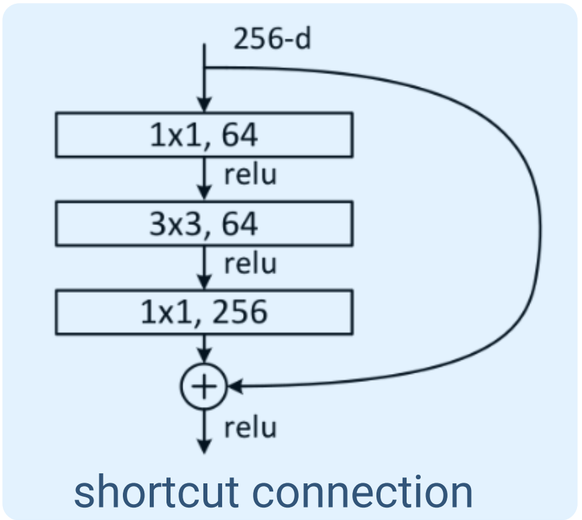
shortcut connection



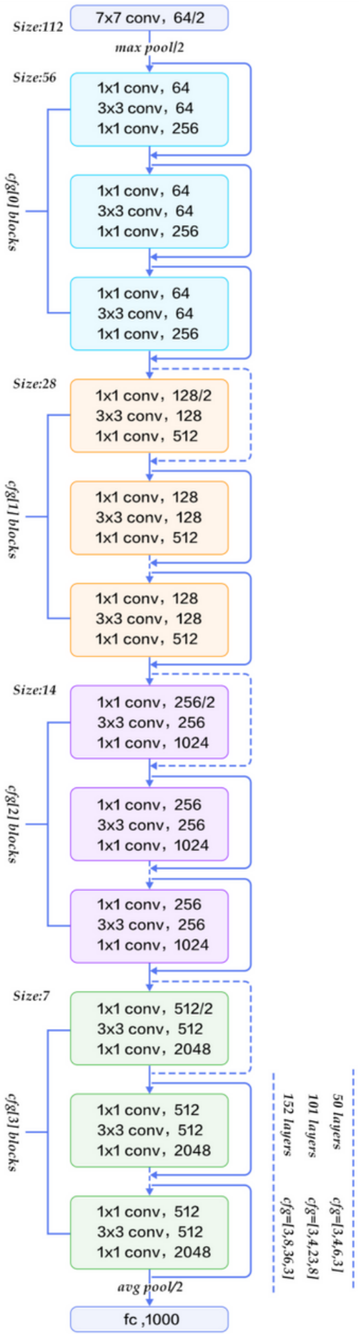
Model Setup

How ResNet50 Works?

- 7×7 kernel, 64 kernels with a 2-sized stride.
- A max pooling layer with a 2-sized stride.
- 3×3,64 kernel, 1×1,64 kernels, 1×1,256 kernels, iterated 3 times. (9)
- 1×1,128 kernels, 3×3,128 kernels,1×1,512 kernels, iterated 4 times.(12)
- 1×1,256 cores, 2 cores 3×3,256, 1×1,1024, iterated 6 times. (8)
- 1×1,512 cores, 3×3,512 cores, 1×1,2048 cores iterated 3 times. (9)
- Average pooling
- Fully connected

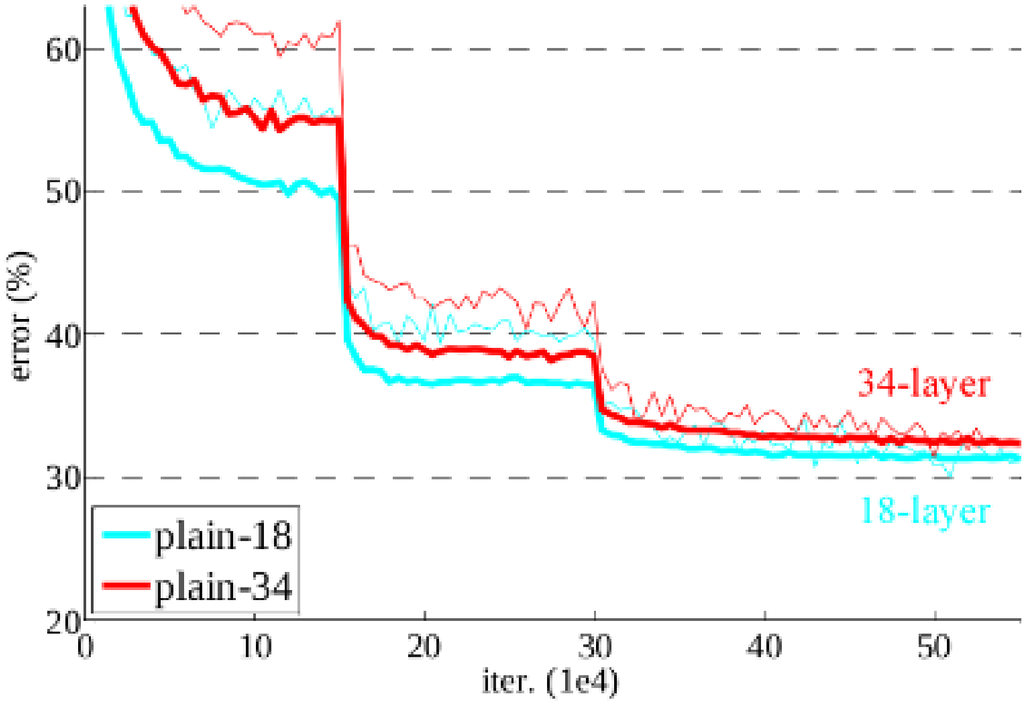


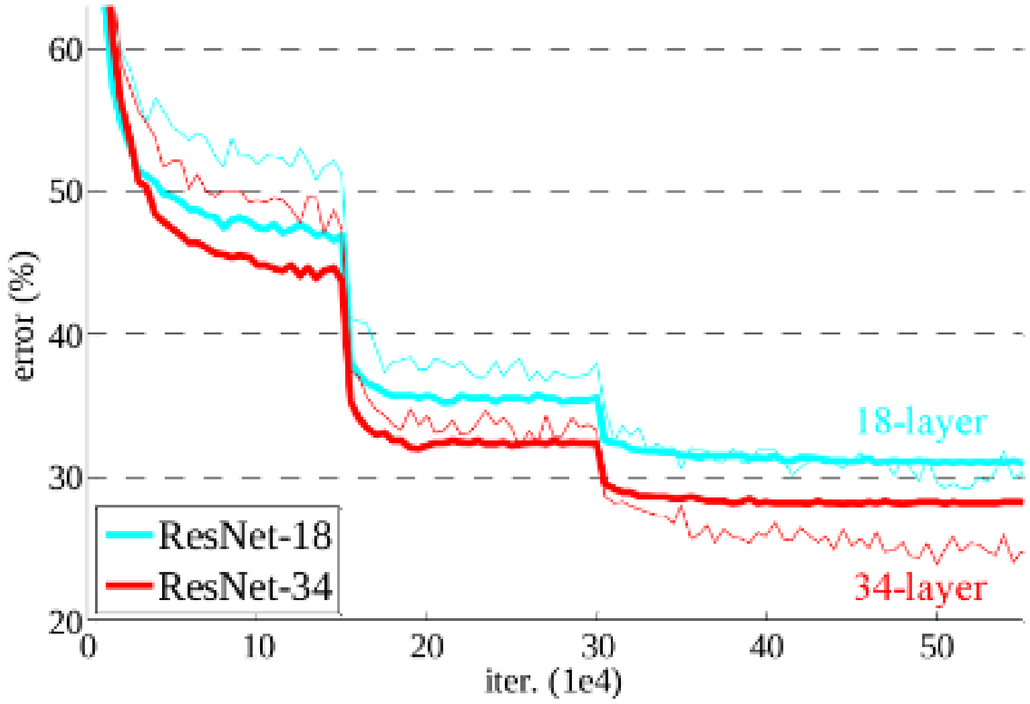
shortcut connection



Model Setup

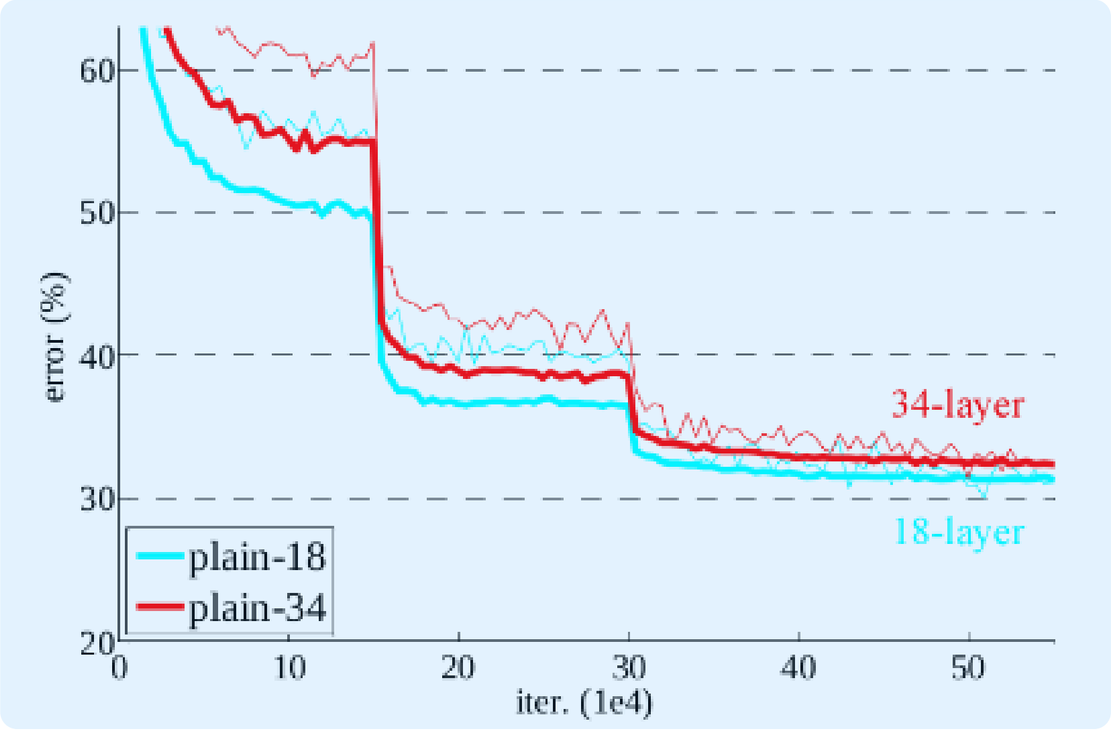
Comparison

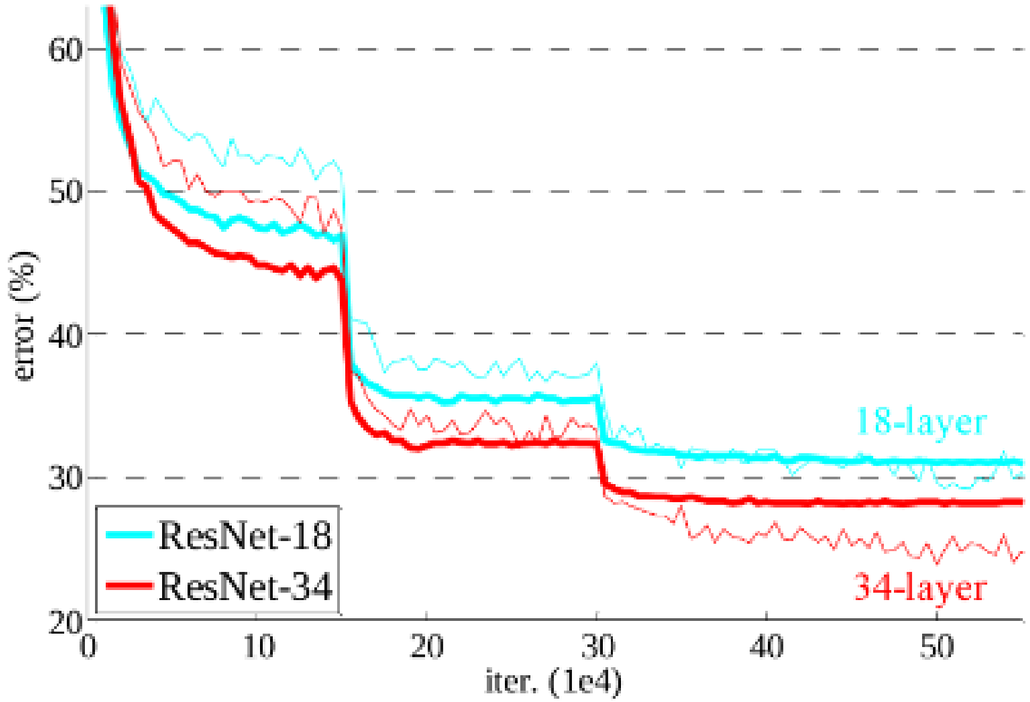




Model Setup

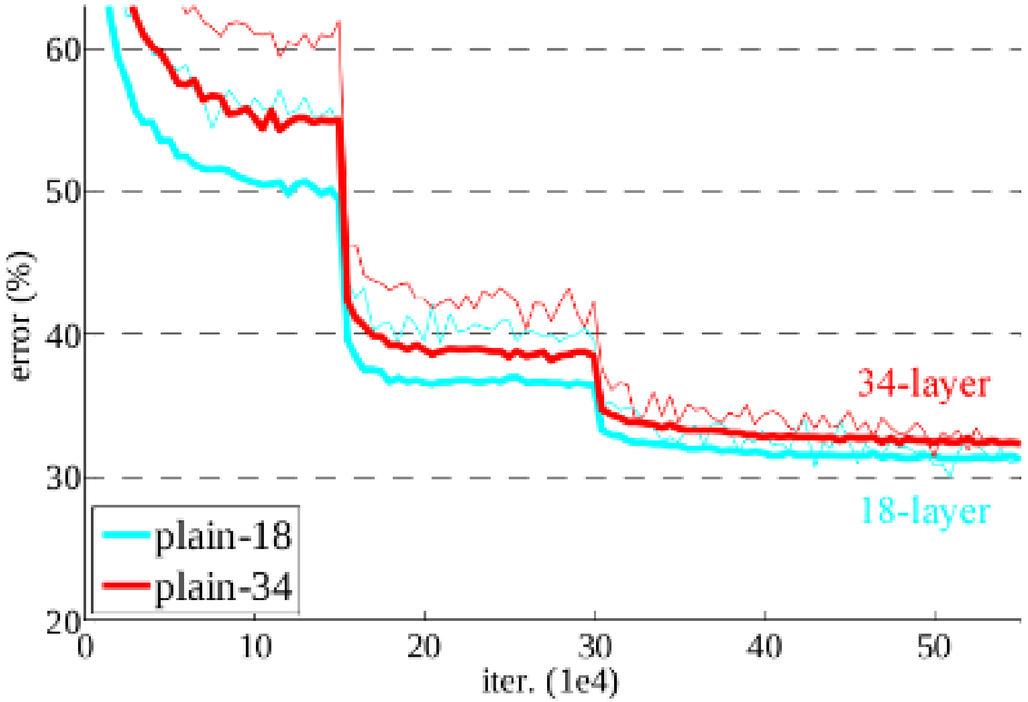
Comparison

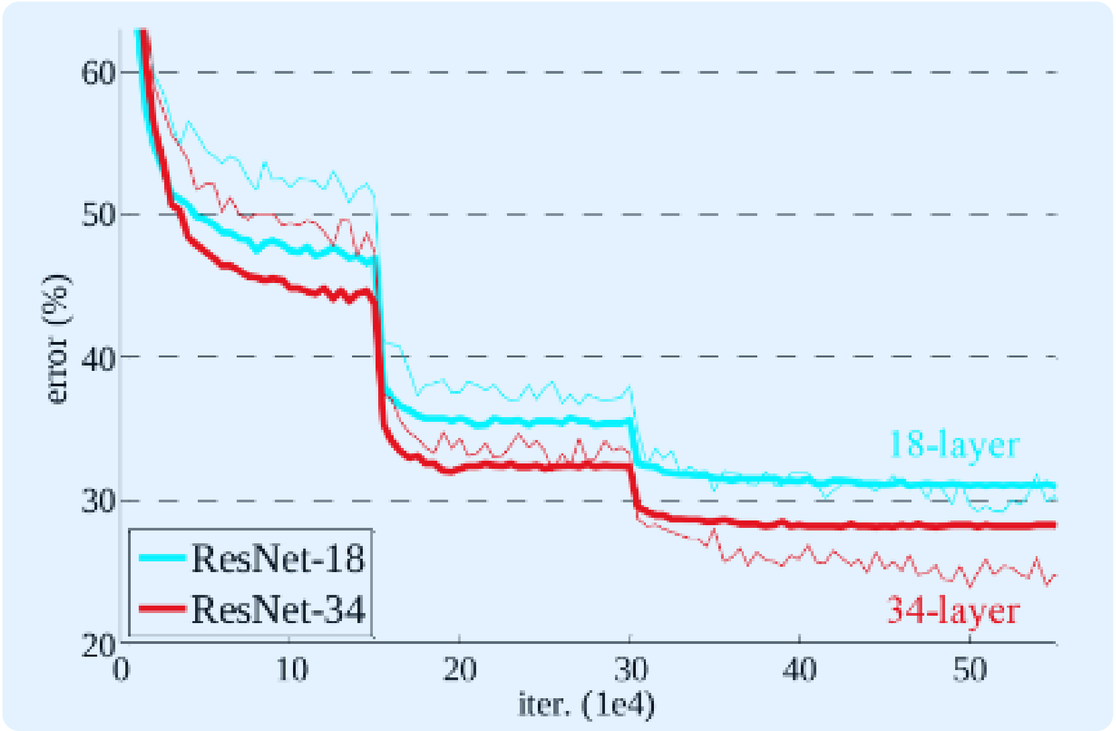




Model Setup

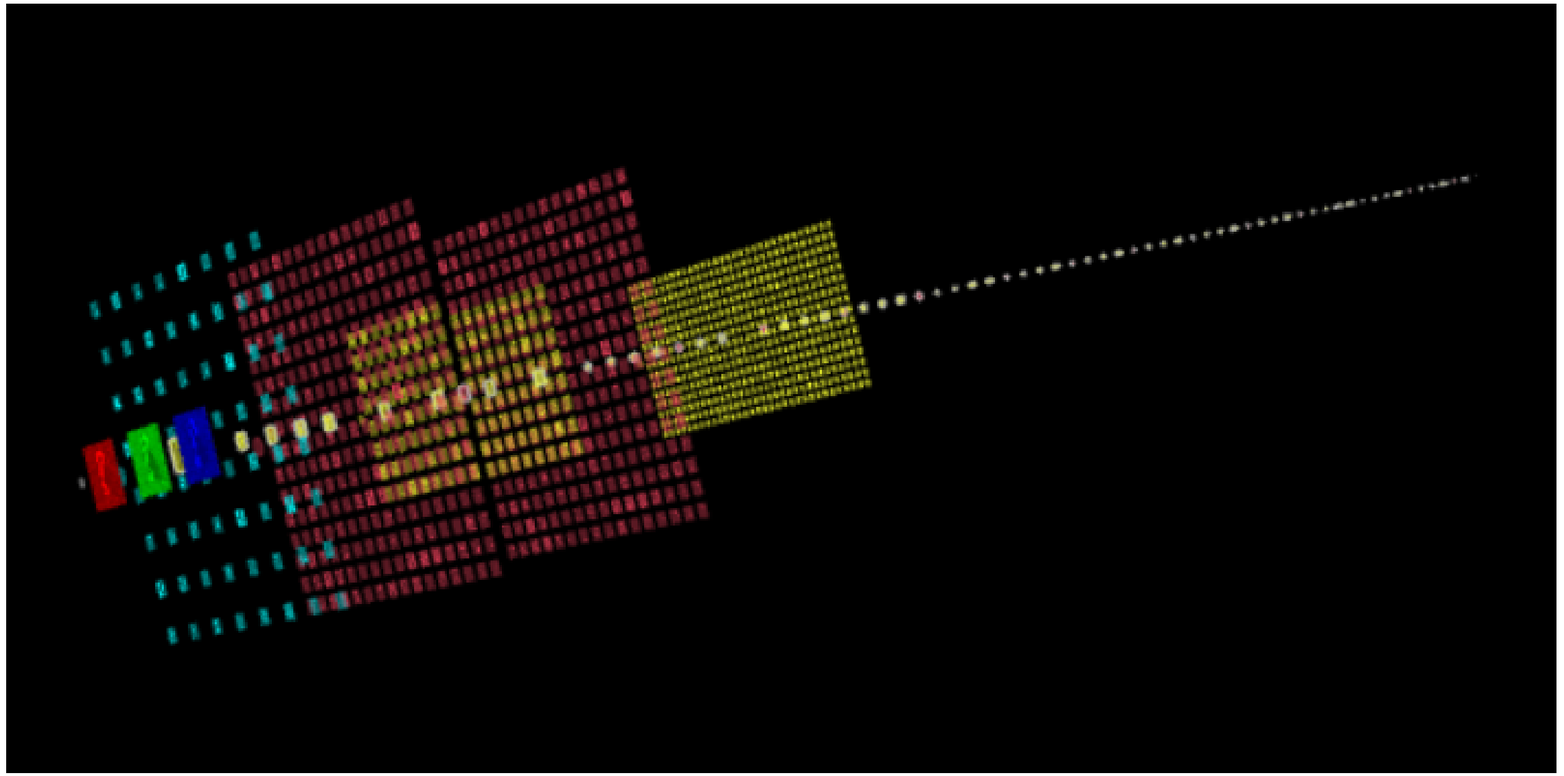
Comparison





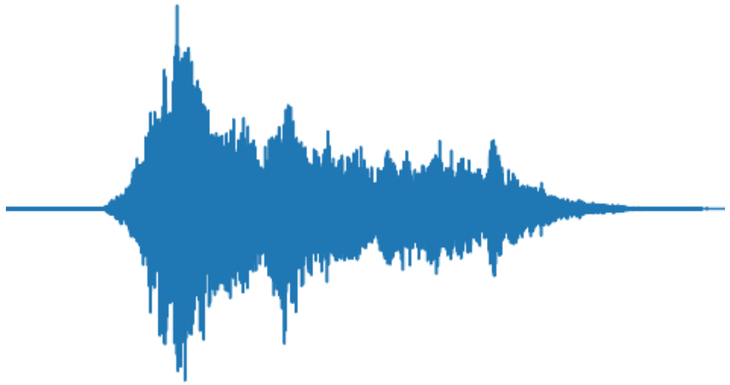
ResNet50 3D Visualization

Tensorspace.js



Performance: Waveform

47973 audio files



learning rate: 1e-4

batch size: 50

epoch: 4

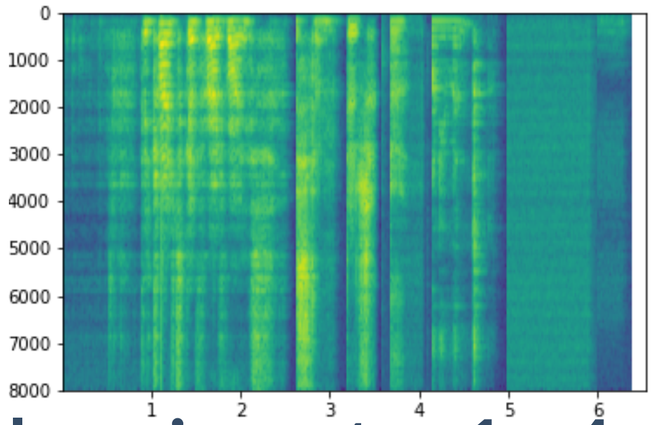
bonafide: 23492

spoof: 24481

```
epoch 1 :0.7422024011611938/acc:  0.5087443161944736
epoch 1 :0.0012585122603923082/acc:  0.8946601375772415
epoch 2 :8.205428457586095e-05/acc:  0.9466013757724145
epoch 2 :2.896679143304936e-05/acc:  0.9551708056429987
epoch 3 :6.994004070293158e-05/acc:  0.9547627375539233
epoch 3 :1.2926899216836318e-05/acc:  0.9485834207764953
epoch 4 :8.728451575734653e-06/acc:  0.9488166025416812
epoch 4 :1.0316204679838847e-05/acc:  0.9547627375539233
```

Performance: Spectrogram

16013 audio files



learning rate: 1e-4

batch size: 40

epoch: 3

bonafide: 5400

spoof: 10613

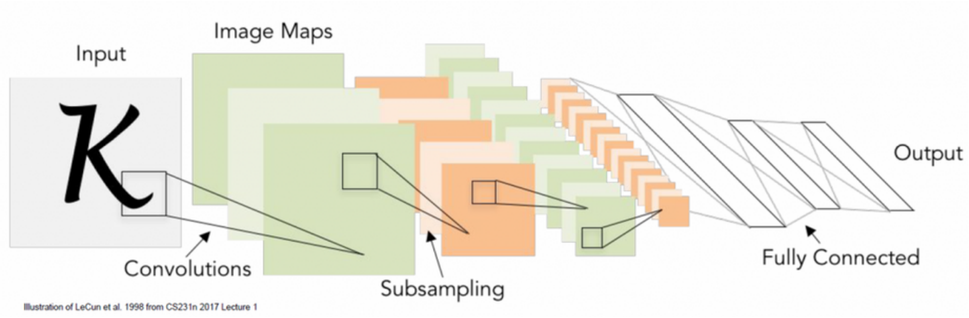
```
Epoch: 1: 0.8728336095809937/Acc:  0.3440025352559024
Epoch: 1: 0.6031295657157898/Acc:  0.6926002218348914
Epoch: 1: 0.21666458249092102/Acc:  0.8830613215021391
Epoch: 1: 0.10520239174365997/Acc:  0.9000158453493899
Epoch: 1: 0.001599683426320552 8/Acc:  0.9256853113611155
Epoch: 2: 0.006586838513612747/Acc:  0.8089050863571542
Epoch: 2: 0.0003010924556292593 5/Acc:  0.8985897639042941
Epoch: 2: 0.0001111829333240166 3/Acc:  0.8776739027095547
Epoch: 2: 0.138189867138862 6/Acc:  0.919980985580732
Epoch: 2: 0.02670418843626976/Acc:  0.8819521470448424
Epoch: 3: 0.003784365952014923/Acc:  0.9191887181112344
Epoch: 3: 0.003071725135669112/Acc:  0.8892410077642212
Epoch: 3: 0.000667512358631938 7/Acc:  0.9096815084772619
Epoch: 3: 3.151399869238958e-05/Acc:  0.91380129931865
Epoch: 3: 1.136945229518460 1e-05/Acc:  0.9207732530502297
```

Results

VGG-like model

all spetrogram

91%



ResNet50

all waveform

95%

all spectrogram

91%