Defeating the Fakes

Vocal Fake Detection using Discrete Fourier Transforms and Spectrograms in Neural Networks

Tina Tami
Lars Tijsmans

Supervisor
Zeno Geradts

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Introduction

Tech

Deepfake queen to deliver Channel 4 Christmas message [1]

23 December 2020

[1] https://www.bc.com/news/technology-55424730

Introduction

Tech

Deepfake queen to deliver Channel 4 Christmas message [1]

23 December 2020

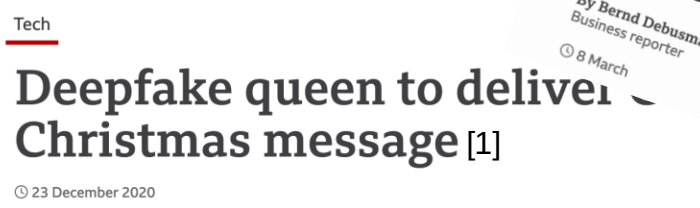
Navalny

Europese politici benaderd door deepfake van Russische oppositie [2]

22 april 2021 19:53
Aangepast: 22 april 2021 20:00

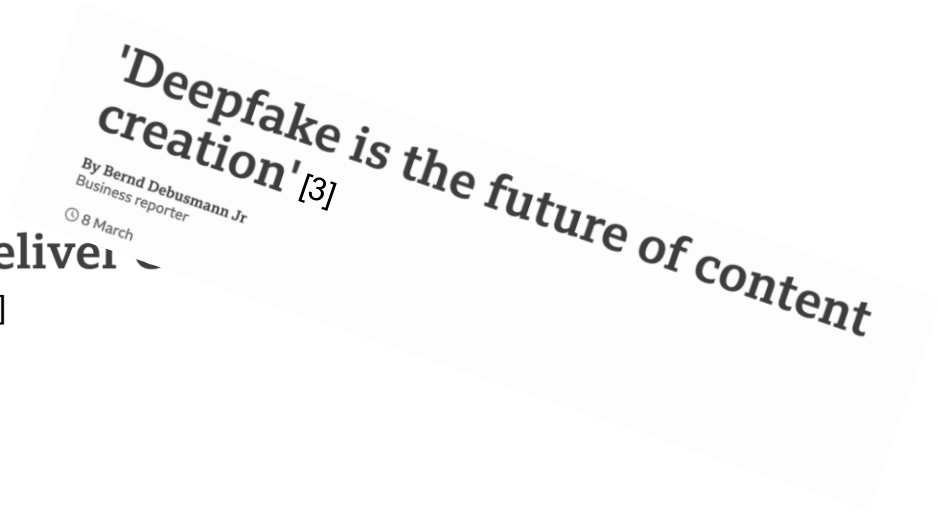
[1] https://www.bc.com/news/technology-55424730
[2] https://www.rtlnieuws.nl/tech/artikel/5226954/deepfake-rusland-navalny-europese-politici

Introduction


Tech

Deepfake queen to deliver Christmas message [1]

23 December 2020


'Deepfake is the future of content creation' [3]

By Bernd Debusmann Jr
Business reporter

8 March

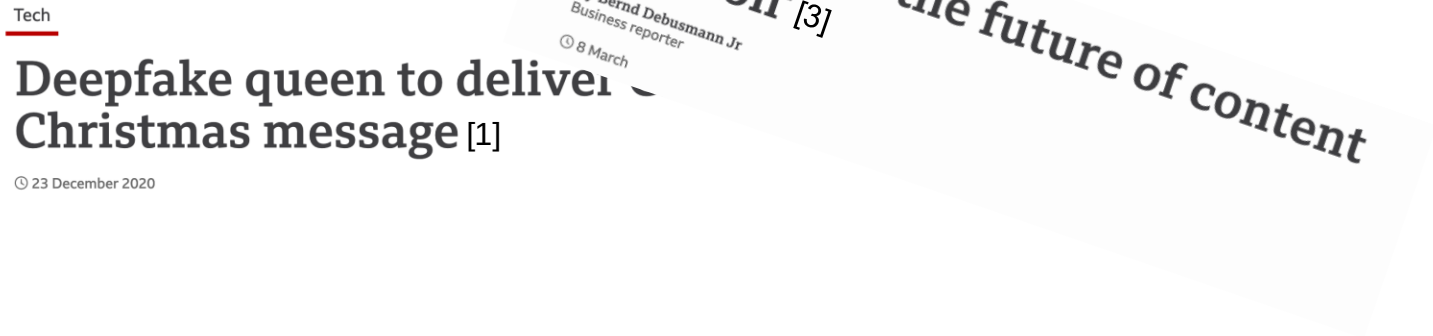

Navalny

Europese politici benaderd door deepfake van Russische oppositie [2]

22 april 2021 19:53
Aangepast: 22 april 2021 20:00

[1] https://www.bc.com/news/technology-55424730
[2] https://www.rtlnieuws.nl/tech/artikel/5226954/deepfake-rusland-navalny-europese-politici
[3] https://www.bbc.com/news/business-56278411

Introduction


Tech

Deepfake queen to deliver Christmas message [1]

23 December 2020

'Deepfake is the future of content creation' [3]

By Bernd Debusmann Jr
Business reporter

8 March

Navalny

Europese politici benaderd door deepfa van Russische oppositie [2]

22 april 2021 19:53
Aangepast: 22 april 2021 20:00

Scarily authentic new deep fake of Tom Cruise attracts millions of views [4]

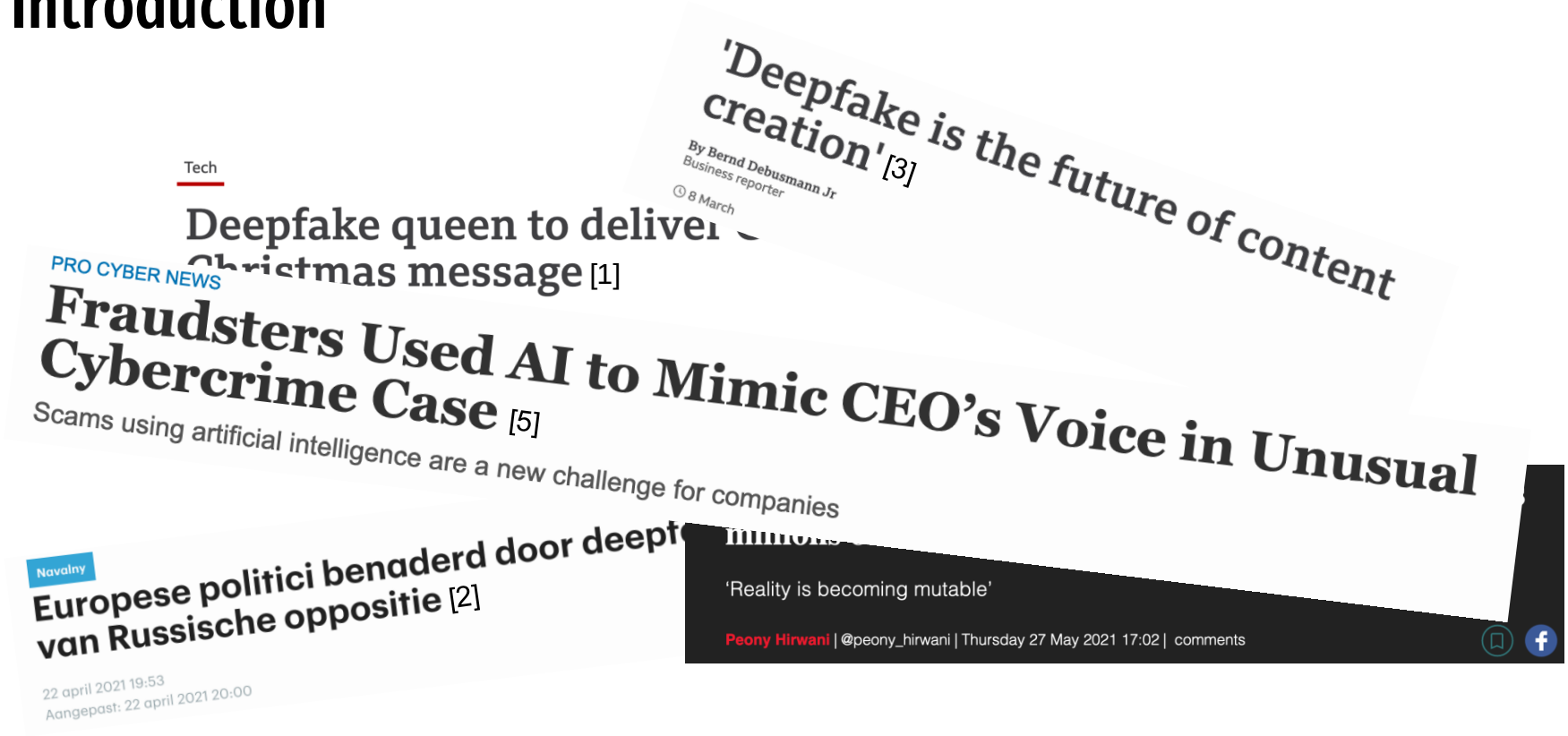
'Reality is becoming mutable'

Peony Hirwani | @peony_hirwani | Thursday 27 May 2021 17:02 | comments

[1] https://www.bc.com/news/technology-55424730
[2] https://www.rtlnieuws.nl/tech/artikel/5226954/deepfake-rusland-navalny-europese-politici

[3] https://www.bbc.com/news/business-56278411
[4] https://www.independent.co.uk/celebrity-news/tom-cruise-deep-fake-tik-tok-b1853256.html

Introduction



[1] https://www.bc.com/news/technology-55424730
[2] https://www.rtlnieuws.nl/tech/artikel/5226954/deepfake-rusland-navalny-europese-politici

[3] https://www.bbc.com/news/business-56278411
[4] https://www.independent.co.uk/celebrity-news/tom-cruise-deepfake-tik-tok-b1853256.html

[5] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Research questions

Can Fourier transformations be used to distinguish vocal fakes from real voices in a classification model?

- How do discrete Fourier transform and spectrograms compare in accuracy?
- How do discrete Fourier transform and spectrograms compare in computational performance?
- How does the model perform on different datasets?

Related work

- 'DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices' [1]
- 'A Machine Learning Model to Detect Fake Voice' [2]
- 'A Fourier transform based audio watermarking algorithm' [3]

[1] Run Wang et al. (May 2020)
[2] Yohanna Rodriguez-Ortega, Dora Ballesteros, and Diego Renza (Oct. 2020)
[3] Kafi Redouane, Kahlessenane Fares, Euschi Salah, and Khaldi Amine (Sept. 2020)

Datasets

ASVSpoof2019

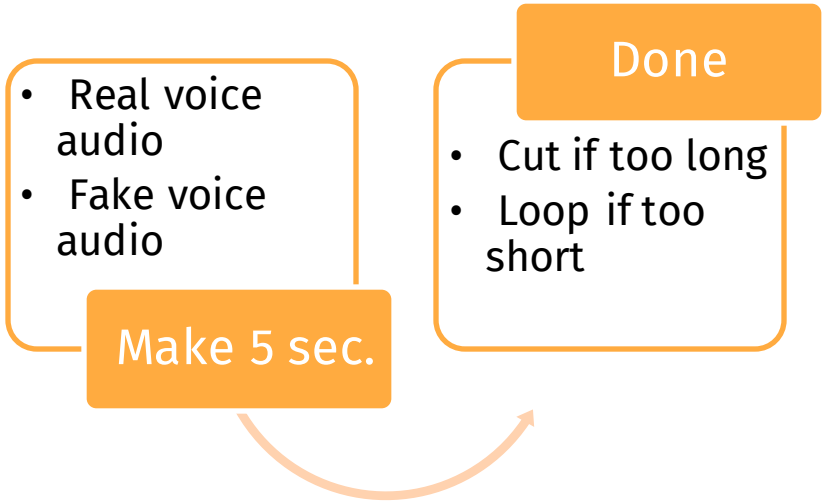
- 7.1GB
- Telephony and VoIP quality
- 20 speakers
- English
- Labeled
- Real and fake voices
- Fake voices: several generation methods

RTVCSpeech
using LibriSpeech ASR corpus

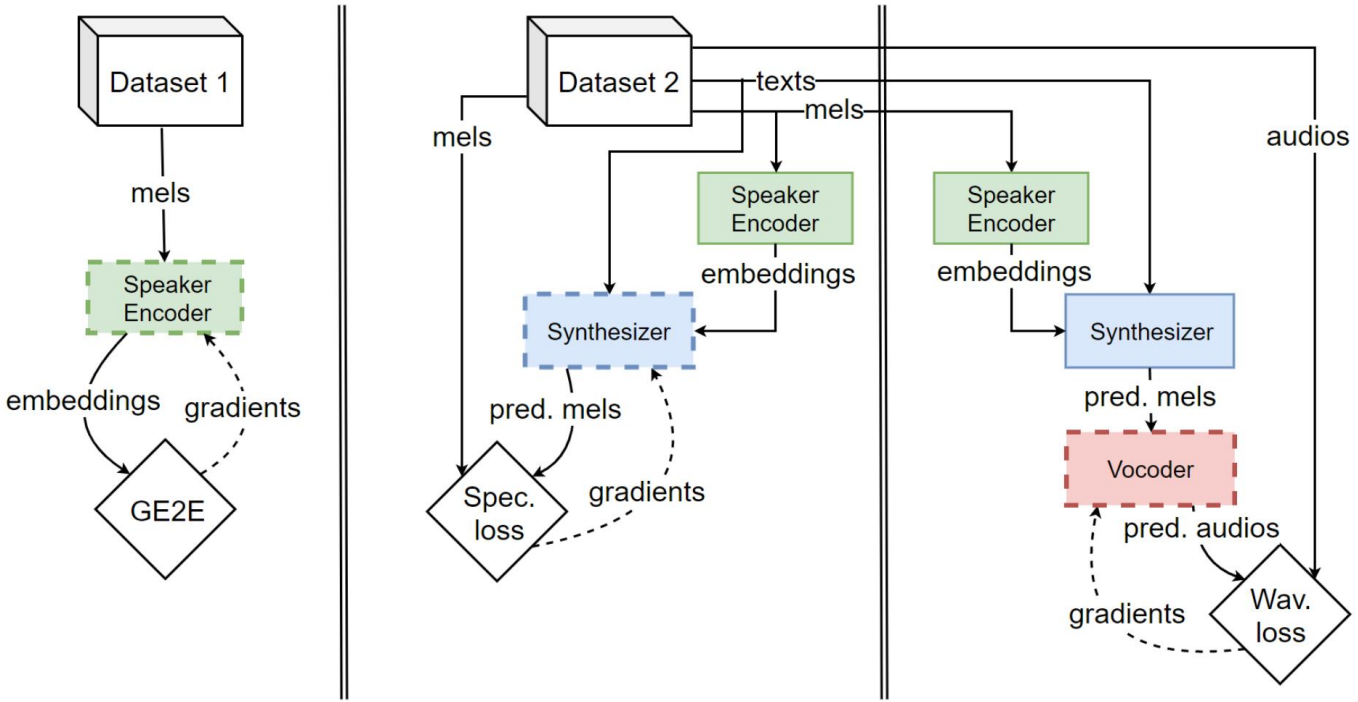
- 100 hours
- Clean speech
- 251 speakers
- English audio books
- Transcribed
- Real voices: LibriSpeech
- Fake voices: RTVC

Preprocessing: RTVCSpeech

Preprocessing: ASVSpoof



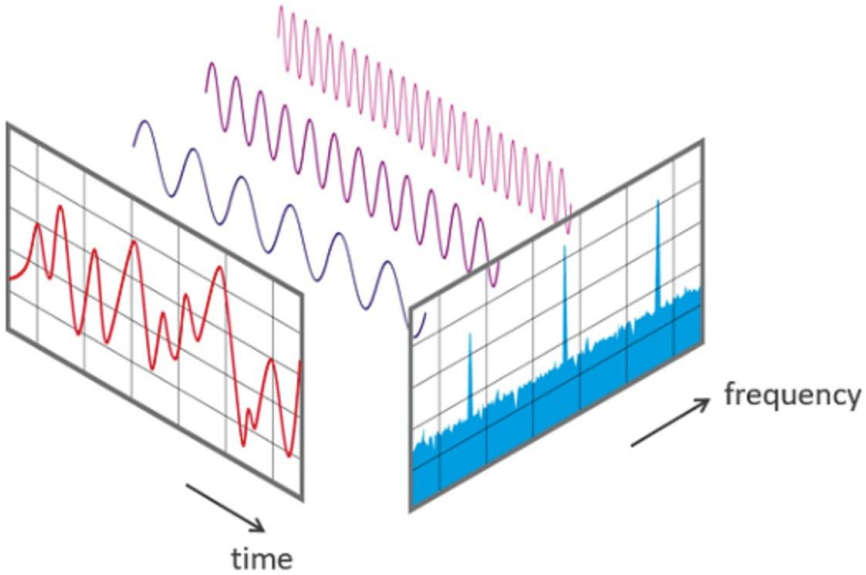
Real Time Voice Cloning (RTVC)



Discrete Fourier transform

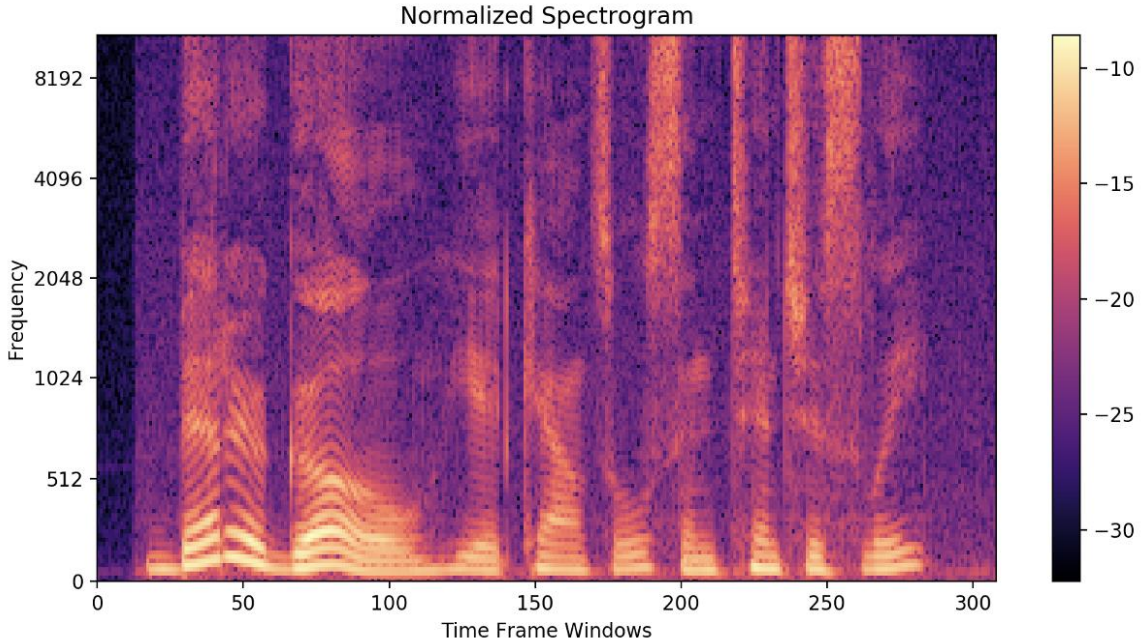
$$
\begin{aligned}
X_k &= \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N} kn} \\
&= \sum_{n=0}^{N-1} x_n \cdot \left[\cos\left(\frac{2\pi}{N} kn \right) - i \cdot \sin\left(\frac{2\pi}{N} kn \right) \right],
\end{aligned}
$$

Discrete Fourier transform



https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft

Spectrograms



https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520

Experiments

1. RTVCSpeech as training and test set

2. RTVCSpeech as training set, ASVSpoof as test set

3. ASVSpoof as training and test set

4. ASVSpoof as training set, RTVCSpeech as test set

RTVCSpeech	**Real**	**Fake**
Training set	2116	2116
Test set	528	528

ASVSpoof	**Real**	**Fake**
Training set	2064	2116
Test set	516	528

Experiments

GPU

Convolutional Neural Network

Nvidia RTX 2080

CPU

Fully Connected Neural Network

Intel Xeon 2.10GHz

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Models

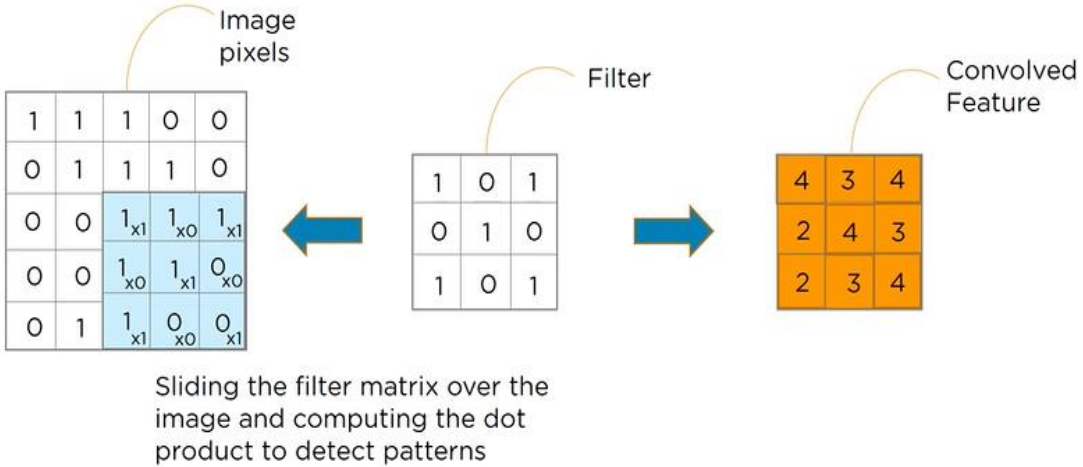
2D Convolutional Neural Network

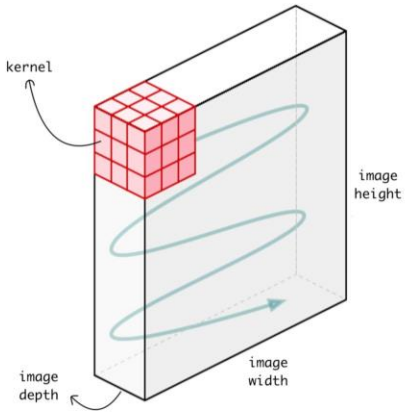
- Image classification
- Network employs a mathematical operation called convolution

Fully Connected Neural Network

- All the neurons in a layer are connected to those in the next layer
- Learns features from all the combinations of the features of the previous layer

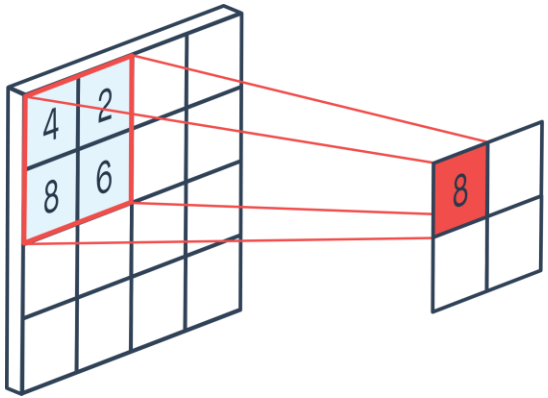
Models – 2D CNN: Convolution layer

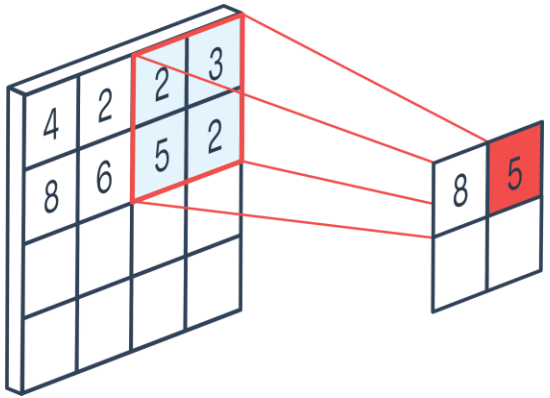




https://www.slideshare.net/Simplilearn/convolutional-neural-network-tutorial-cnn-how-cnn-works-deep-learning-tutorial-simplilearn/Simplilearn/convolutional-neural-network-tutorial-cnn-how-cnn-works-deep-learning-tutorial-simplilearn

Models – 2D CNN: MaxPooling layer

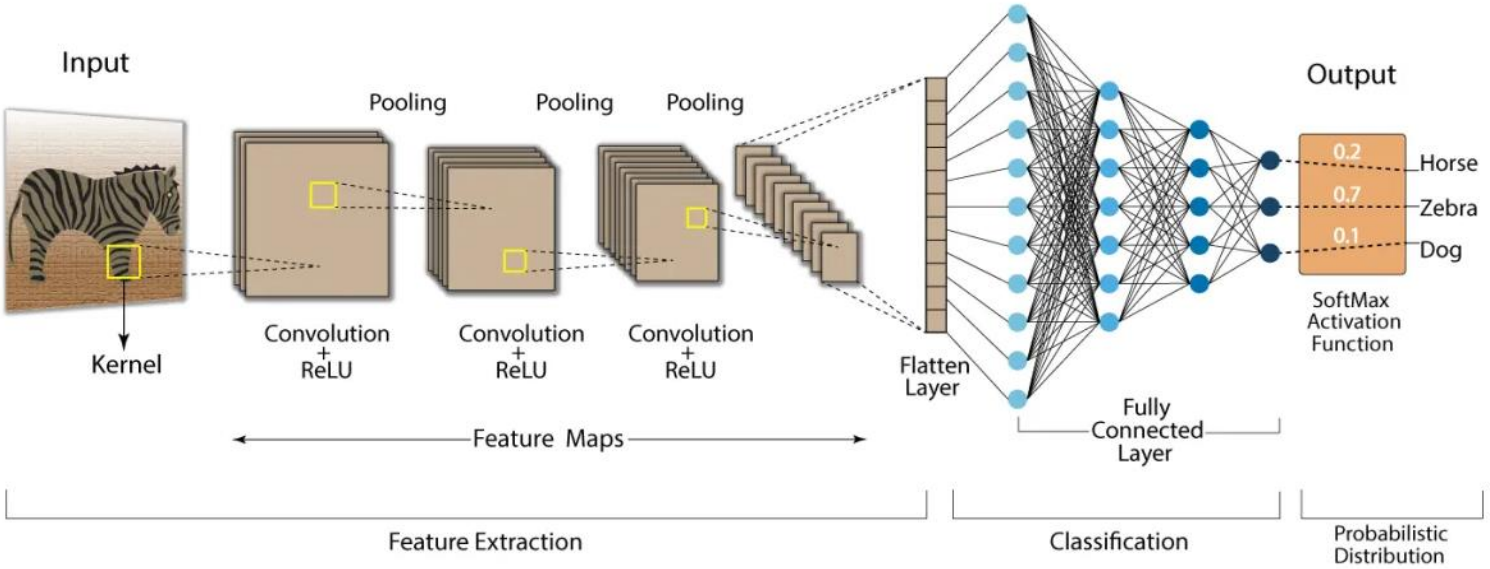




Decrease the size of the feature map by taking the largest element

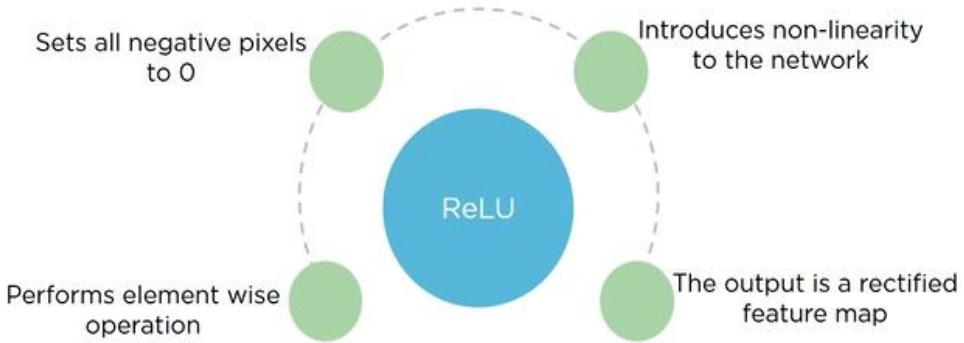
https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/2d-max-pooling-block

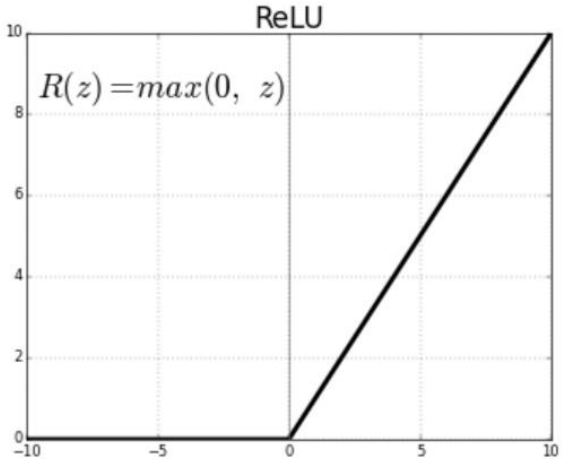
Models - 2D CNN and FCNN : Fully connected (dense) layer



https://developersbreach.com/convolution-neural-network-deep-learning/

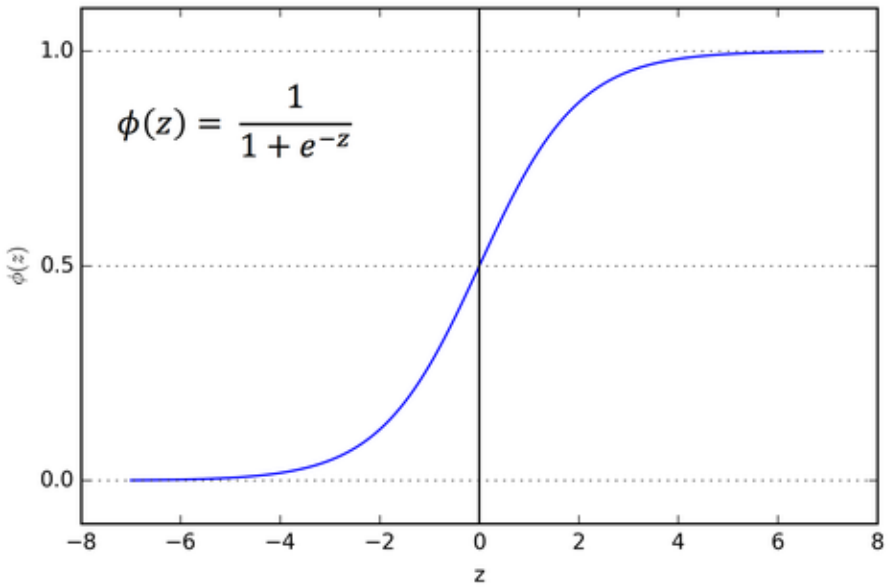
Models - 2D CNN and FCNN: ReLU (Rectified Linear Unit) activation function





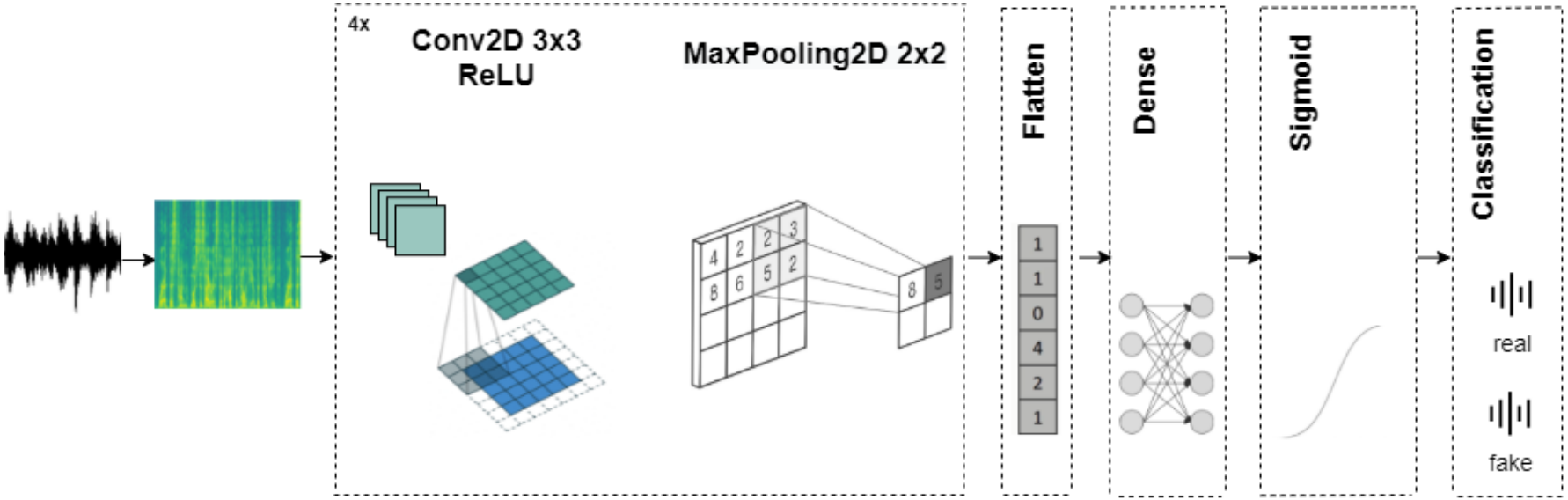
https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

Models - 2D CNN and FCNN: Sigmoid activation function

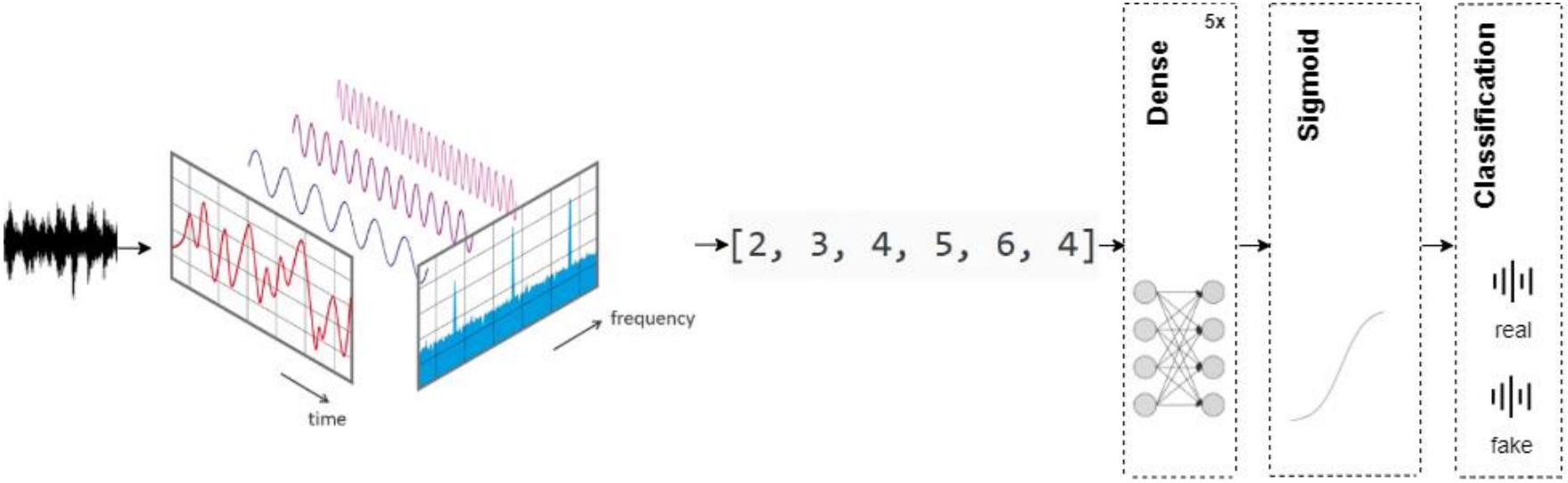


https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

Models: 2D Convolutional Neural Network



Models: Fully Connected Neural Network



Results – spectrograms: RTVCSpeech as training set

RTVCSpeech as test set

Predicted

Actual		**Real**	**Fake**
	Real	527	1
	Fake	0	528

	Real	**Fake**
Accuracy	0.999	0.999
F1-score	0.999	0.999

ASVSpoof as test set

Predicted

Actual		**Real**	**Fake**
	Real	516	0
	Fake	386	142

	Real	**Fake**
Accuracy	0.630	0.630
F1-score	0.728	0.424

Time on GPU: 20s for 1 epoch
Time on CPU: 655s for 1 epoch

Results – spectrograms: ASVSpoof as training set

ASVSpoof as test set

Predicted

Actual	**Real**	**Fake**
Real	514	2
Fake	0	528

	Real	**Fake**
Accuracy	0.998	0.998
F1-score	0.998	0.998

RTVCSpeech as test set

Predicted

Actual	**Real**	**Fake**
Real	198	330
Fake	26	502

	Real	**Fake**
Accuracy	0.663	0.663
F1-score	0.527	0.738

Time on GPU: 20s for 1 epoch
Time on CPU: 655s for 1 epoch

Results – DFT : RTVCSpeech training set

RTVCSpeech as test set

Predicted

Actual		Real	Fake
	Real	449	61
	Fake	81	467

	Real	Fake
Accuracy	0.886	0.886
F1-score	0.863	0.868

ASVSpoof as test set

Predicted

Actual		Real	Fake
	Real	175	334
	Fake	123	413

	Real	Fake
Accuracy	0.563	0.563
F1-score	0.434	0.664

Time on GPU: 1s for 1 epoch
Time on CPU: 1s for 1 epoch

Results – DFT : ASVSpoof as training set

ASVSpoof as test set

Predicted

Actual		**Real**	**Fake**
	Real	457	70
	Fake	83	435

	Real	**Fake**
Accuracy	0.854	0.854
F1-score	0.857	0.850

RTVCSpeech as test set

Predicted

Actual		**Real**	**Fake**
	Real	271	235
	Fake	152	400

	Real	**Fake**
Accuracy	0.634	0.634
F1-score	0.583	0.675

Time on GPU: 1s for 1 epoch
Time on CPU: 1s for 1 epoch

Conclusion

Spectrograms: accuracy

Training set

Test set		RTVCSpeech	ASVSpoof
	RTVCSpeech	0.999	0.663
	ASVSpoof	0.630	0.998

How do discrete Fourier transform and spectrograms compare in accuracy?

Using spectrograms results in a higher accuracy.

Discrete Fourier Transform: accuracy

Training set

Test set		RTVCSpeech	ASVSpoof
	RTVCSpeech	0.886	0.634
	ASVSpoof	0.563	0.854

Conclusion

Spectrograms: accuracy

Training set

Test set		RTVCSpeech	ASVSpoof
	RTVCSpeech	0.999	0.663
	ASVSpoof	0.630	0.998

How do discrete Fourier transform and spectrograms compare in accuracy?

Using spectrograms results in a higher accuracy.

Discrete Fourier Transform: accuracy

Training set

Test set		RTVCSpeech	ASVSpoof
	RTVCSpeech	0.886	0.634
	ASVSpoof	0.563	0.854

How does the model perform on different datasets?

ASVSpoof as training set gives better results. Training and testing on the same dataset gives better results.

Conclusion

How do discrete Fourier transform and spectrograms compare in computational performance?

	GPU	**CPU**
2D CNN	20s for 1 epoch	655s for 1 epoch
FCNN	1s for 1 epoch	1s for 1 epoch

Conclusion

Can Fourier transformations be used to distinguish vocal fakes from real voices in a classification model?

Yes, but...

Discussion and future work

- Improving the neural networks
  - Transfer learning
  - Adding explainability
  - Better vocal fakes