HOW ARTIFICIAL IS YOUR INTELLIGENCE?

Unpacking the Black Box

Nigel Cannings, CTO

Intelligent Voice®

@intelligentvox
www.intelligentvoice.com

FOR $100!

Can you “see” what this means?

What’s the problem?

The Tank Example:

In the 1980s, the Pentagon wanted to harness computer technology to make their tanks harder to attack.

Each Tank was fitted with a camera connected to a computer with the intention of scanning surrounding environments for possible threats.

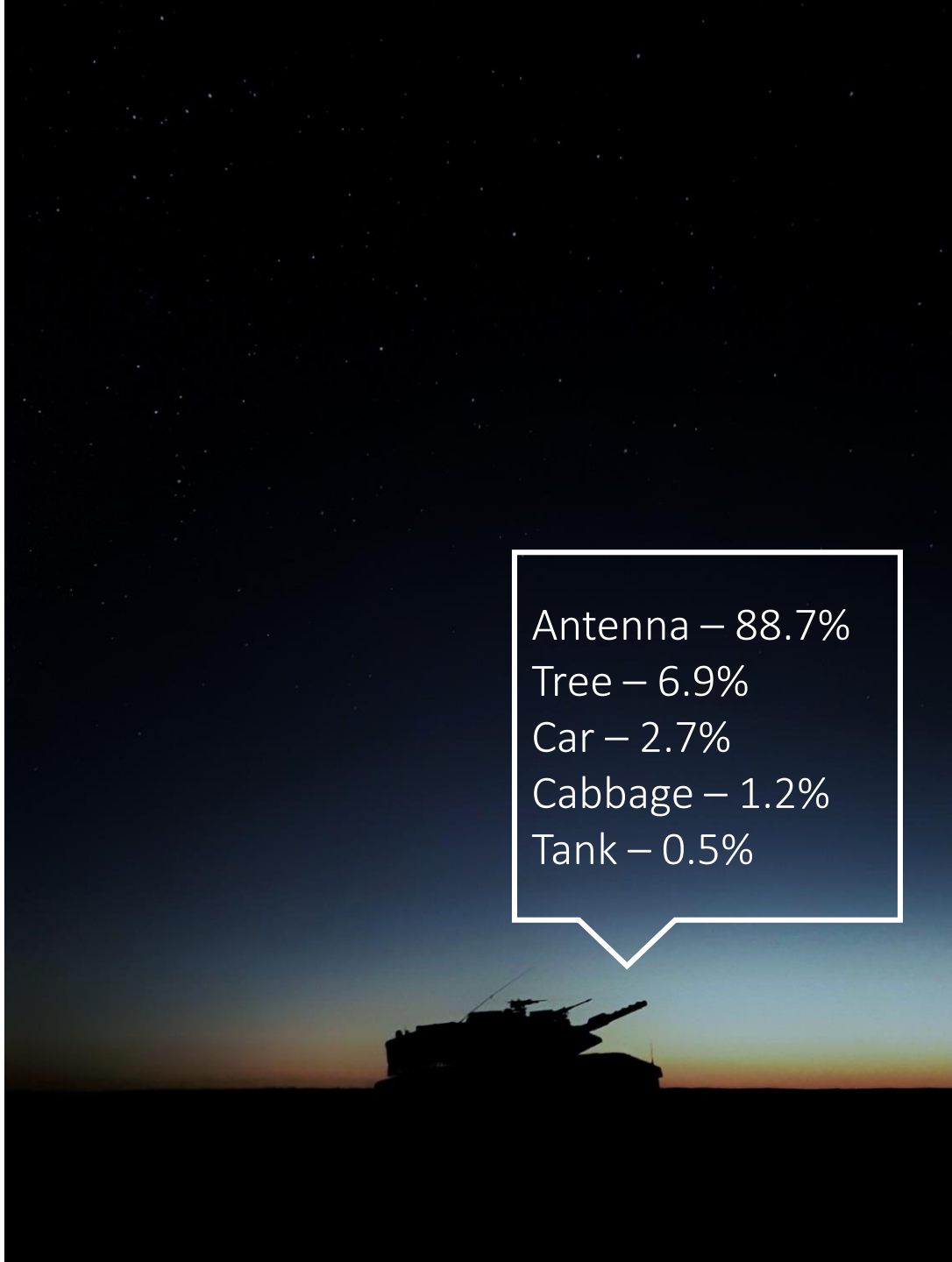
To Interpret the images they had to employ a neural network. They took 200 photos, 100 with tanks “hiding” and 100 without tanks. Half of which were used to train the network, the other half to test it.

The pentagon commissioned a further set of photos for independently testing. The results returned were random, causing some question into how the network had trained itself?

The answer was that in the original set of 200 photos, the “hiding” tank images were taken on a cloudy day whereas the images with no tanks were taken on a sunny day.

The military was now the proud owner of a multi-million dollar mainframe computer that could tell you if it was sunny or not.

Source - https://neil.fraser.name/writing/tank/



Life and Death

Image classification of potential military targets e.g. drones, satellites

The rise of CNNs as a medical diagnostic tool

Navigation and control in self-driving cars

Intelligent Voice®

Legislation

Understanding decision making of AI components is critical

These decisions can lead to loss of life, money etc

Understanding Decisions allows for Improving AI algorithms

The GDPR provides the following rights for individuals:

- The right to be **informed**.
- The right of **access**.
- The right to rectification.
- The right to erasure.
- The right to restrict processing.
- The right to **data portability**.
- The right to object.
- Rights in **relation** to automated **decision** making and **profiling**.

Intelligent Voice®

Taking Inspiration from CNNs

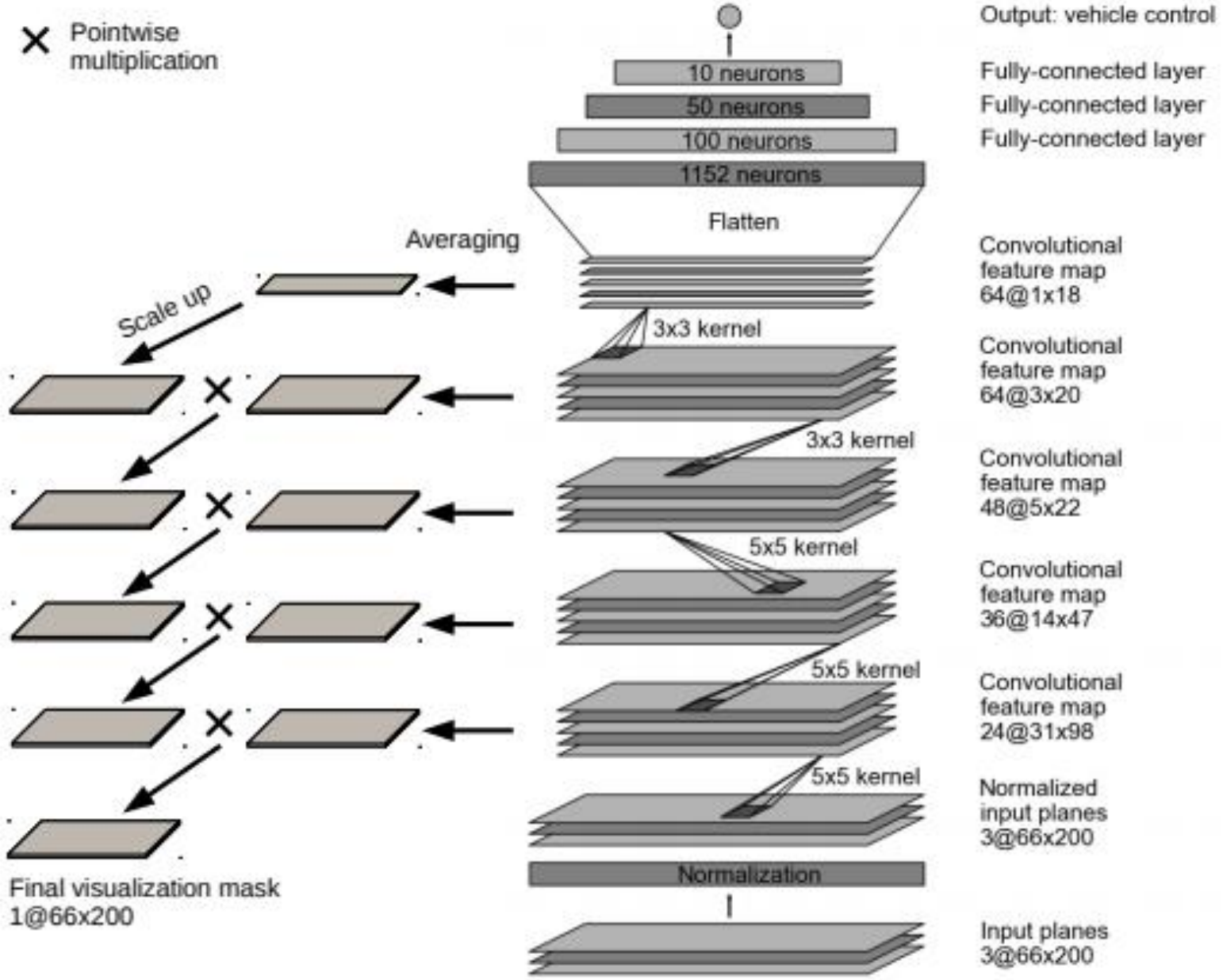


Figure 2: Block diagram of the visualization method that identifies the salient objects.

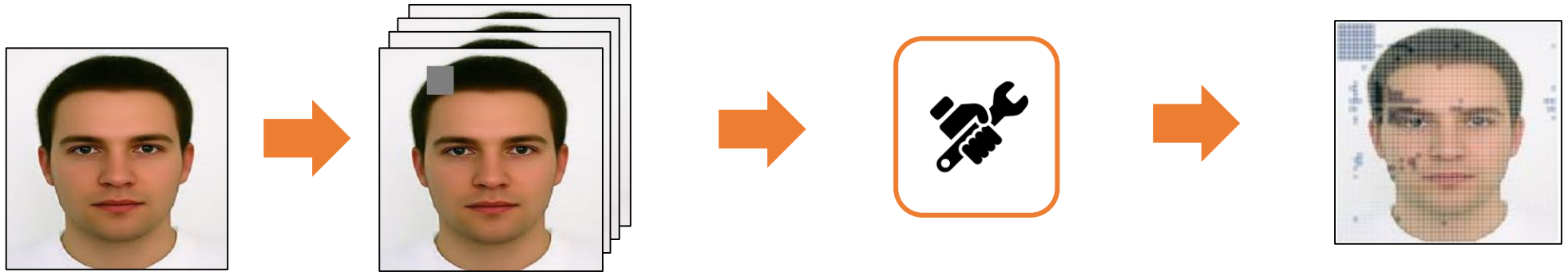
Bojarski et al., 'Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car,' arXiv:1704.07911v1, 2017.

Taking Inspiration from CNNs

Bojarski et al., ‘Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car,’ arXiv:1704.07911v1, 2017.

Deconvolution by Occlusion

Iterate over regions of the image, set a patch of the image to greyscale, and look at the probability of the class:



Take an image

Occlude successive parts of the image with a greyscale square centred on every pixel

Get the classification accuracy for each pixel location

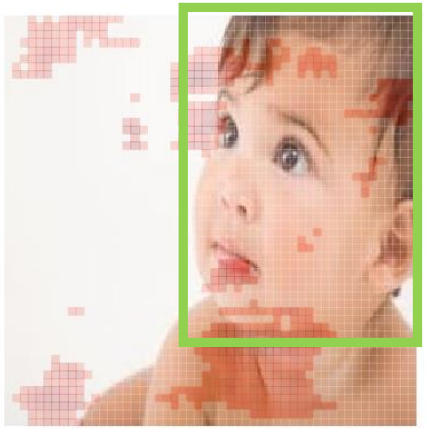
Threshold the results and overlay on original image.

Zeiler & Fergus, 'Visualizing and Understanding Convolutional Networks,' arXiv:1311.2901v3, 2013.

Age Recognition
Feeding back deconvolution results

- Misclassification
- Diagnose the problem using Deconvolution
- Crop the image
- Correct Classification

0 – 2	14.70%
4 – 6	84.48%
8 – 13	0.75%
15 – 20	0.02%
25 – 32	10.94%
38 – 43	1.49%
48 – 53	0.30%
60 -	0.11%



Classes (Age Range):	0 - 2	4 - 6	8 - 13	15 - 20	25 - 32	38 – 43	48 – 53	60 –
	72.91%	12.46%	1.53%	0.26%	10.94%	1.49%	0.30%	0.11%

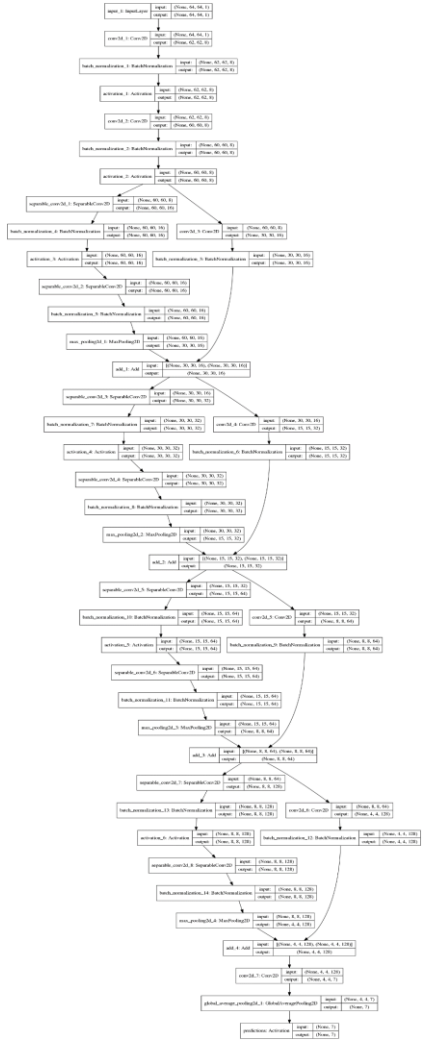
Levi, Gil, and Tal Hassner. "Age and gender classification using convolutional neural networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34-42. 2015.

Facial Emotion Recognition

- Facial emotion recognition architecture
- Performs segmentation of images to extract faces from scenes
- Classifies each face into 7 emotion classes:

  Angry, Disgust, Fear, Happy, Sad, Surprise, Bored

- We downloaded a trained model (Arriaga et al., 2017) and investigated 2 deconvolution approaches to understanding the CNN classifications:
  - GradCAM (Selavaraju, 2016) – Guided Backpropagation of activation maps
  - Deconvolution by occlusion (Zeiler, 2013).

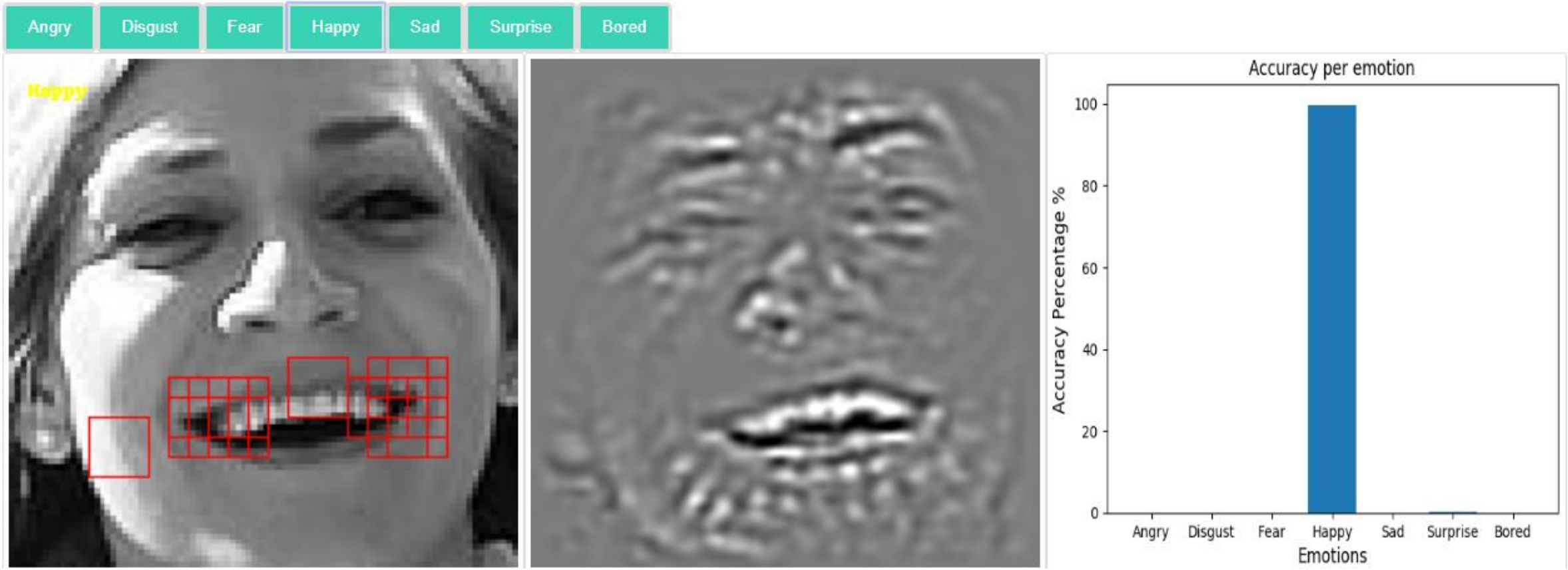


Selvaraju, R. R., Das, A., Vendantam, R., Cogswell, M., Parikh, D., Batra, D., 'Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization, arXiv:1610.02391v1, 2016.

Zeiler & Fergus, 'Visualizing and Understanding Convolutional Networks,' arXiv:1311.2901v3, 2013.

Arriaga, O., Plöger, P.G., Valdenegro, M. Real-time Convolutional Neural Networks for Emotion and Gender Classification, arXiv:1710.07557v1, 2017.

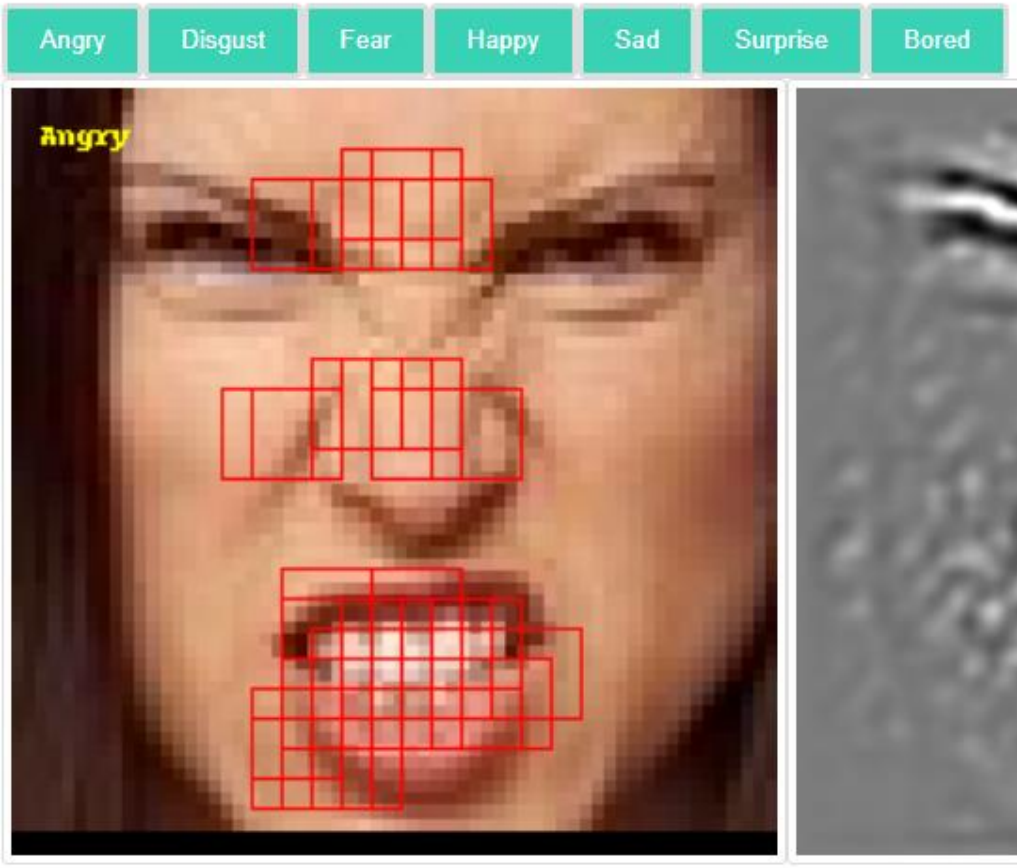
Facial Emotions
Comparing Activation and Occlusion



Occlusion

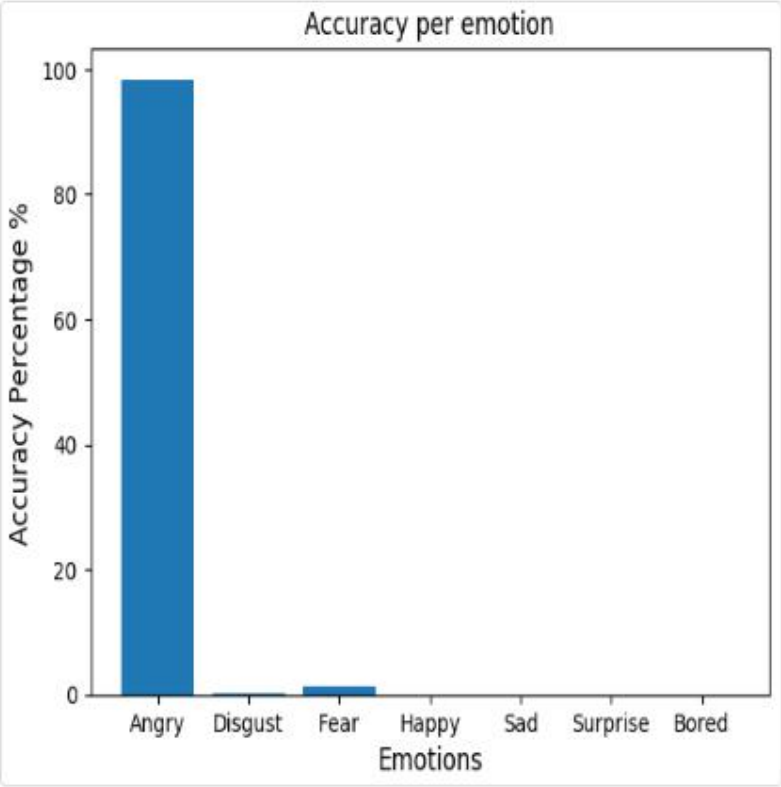
GradCAM

Facial Emotions
Comparing Activation and Occlusion

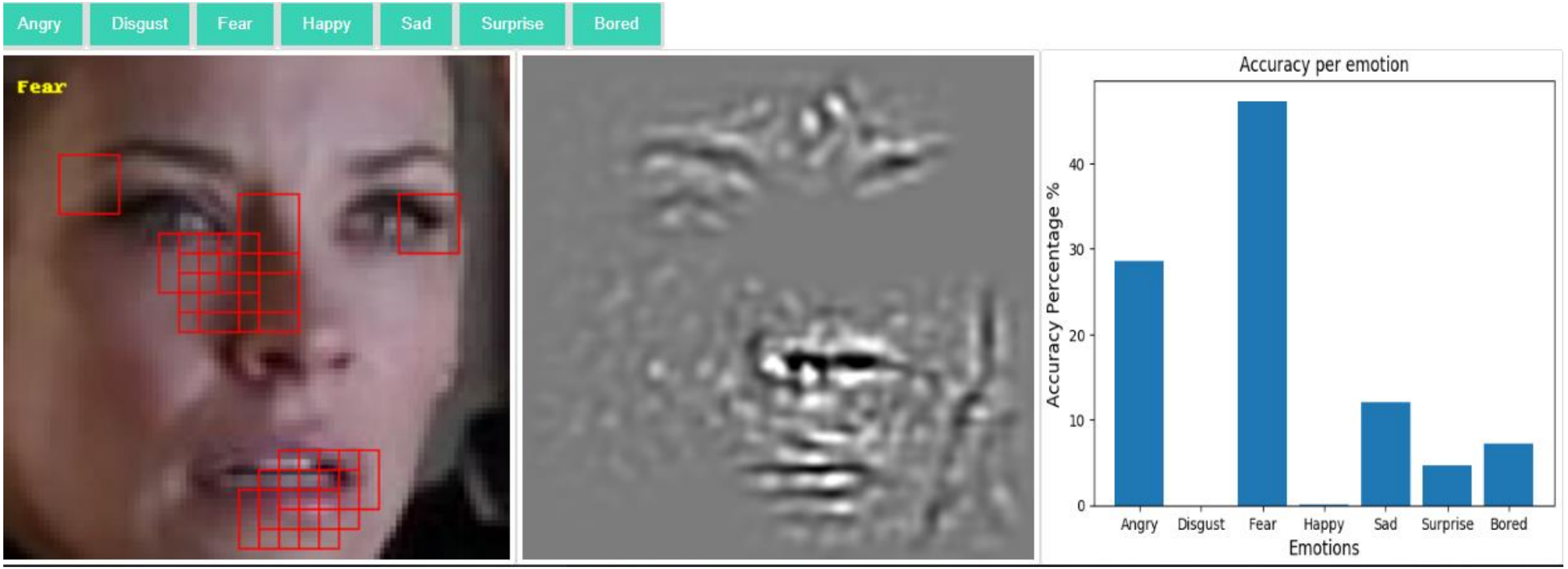


Occlusion

GradCAM



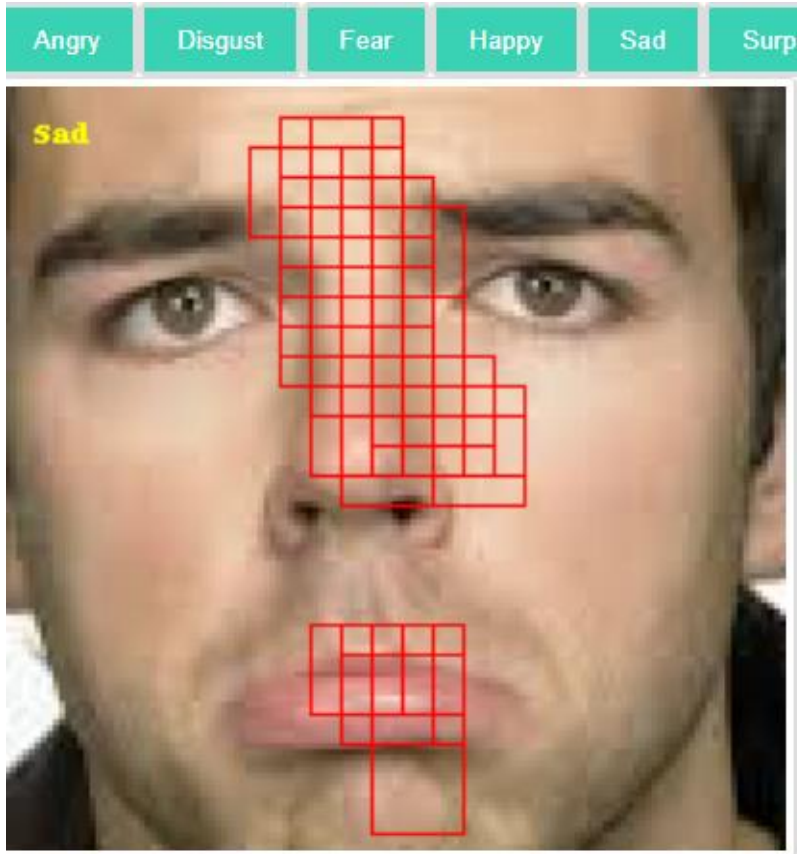
Facial Emotions
Comparing Activation and Occlusion



Occlusion

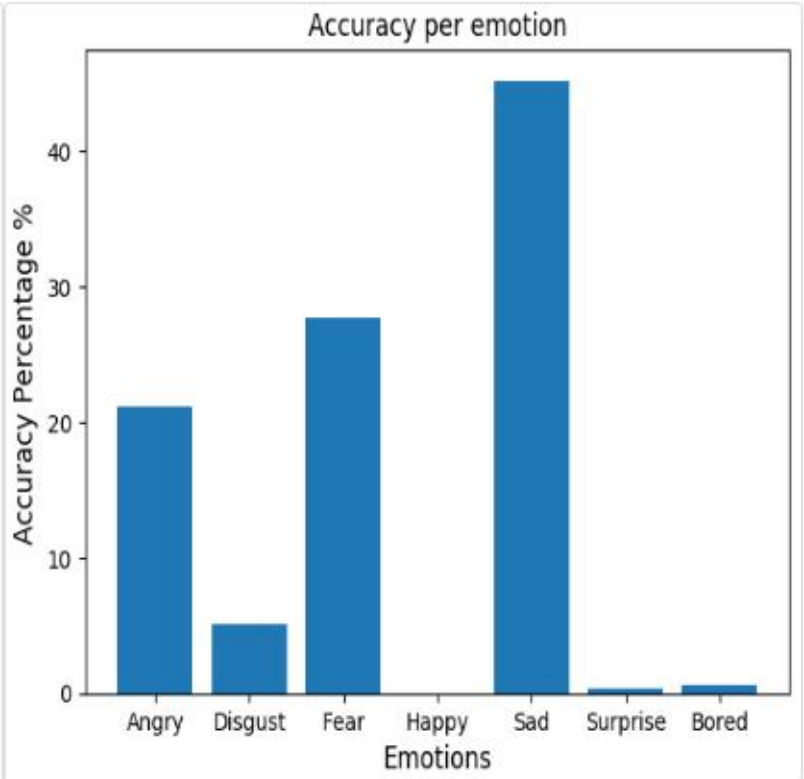
GradCAM

Facial Emotions
Comparing Activation and Occlusion

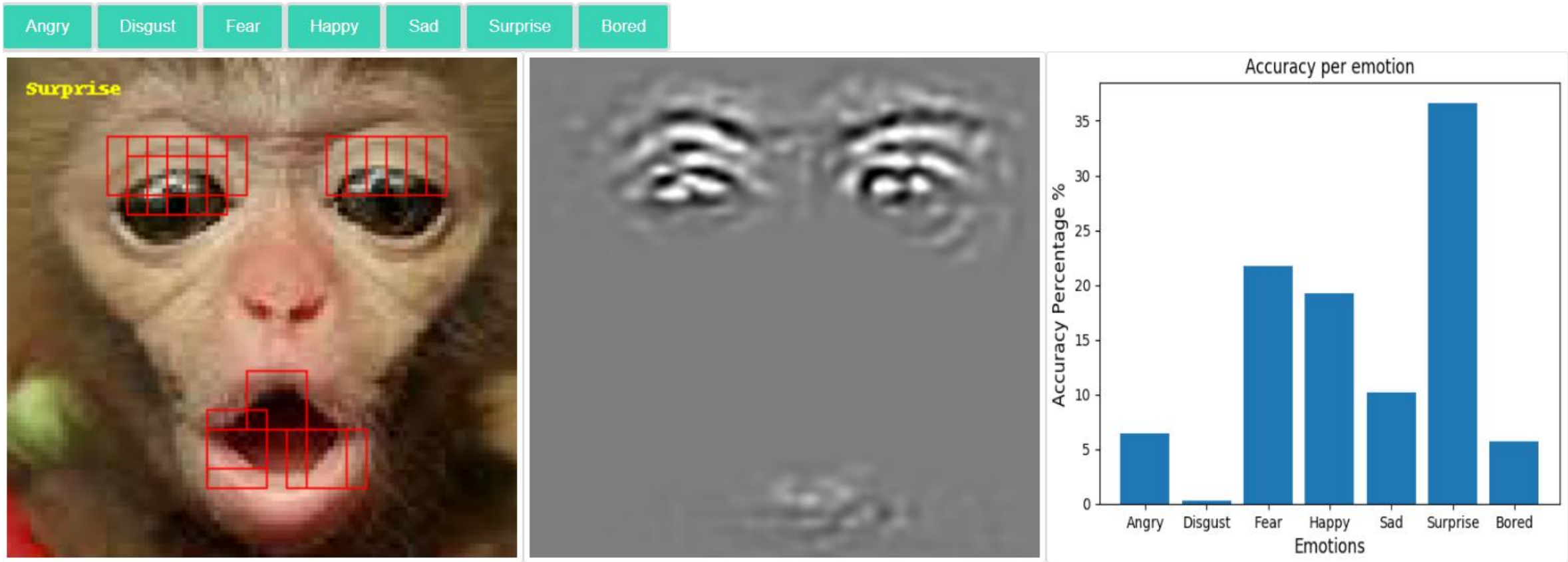


Occlusion

GradCAM



Facial Emotions
Comparing Activation and Occlusion



Occlusion

GradCAM

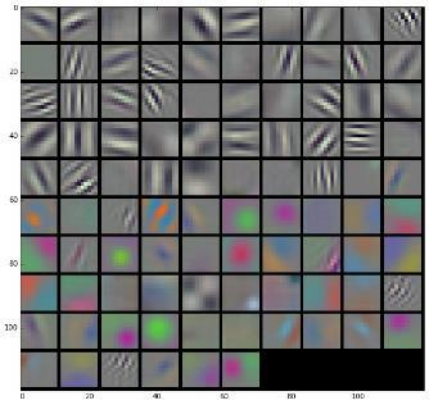
Live Demo

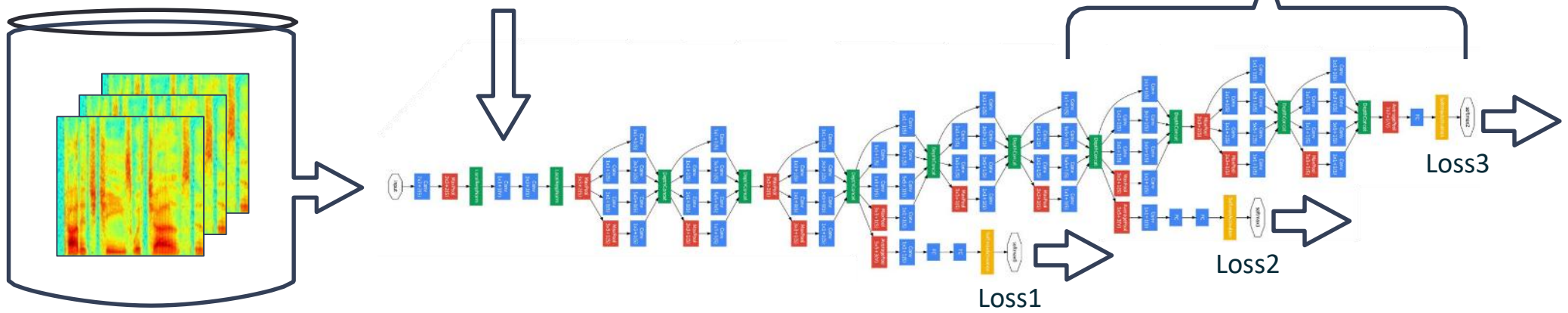
GoogLeNet Processing

Apply convolutions to extract primitives such as edges, formant ridges etc

GoogLeNet:
Szegedy et al., 'Going deeper with convolutions,' arXiv:1409.4842v1, 2014.

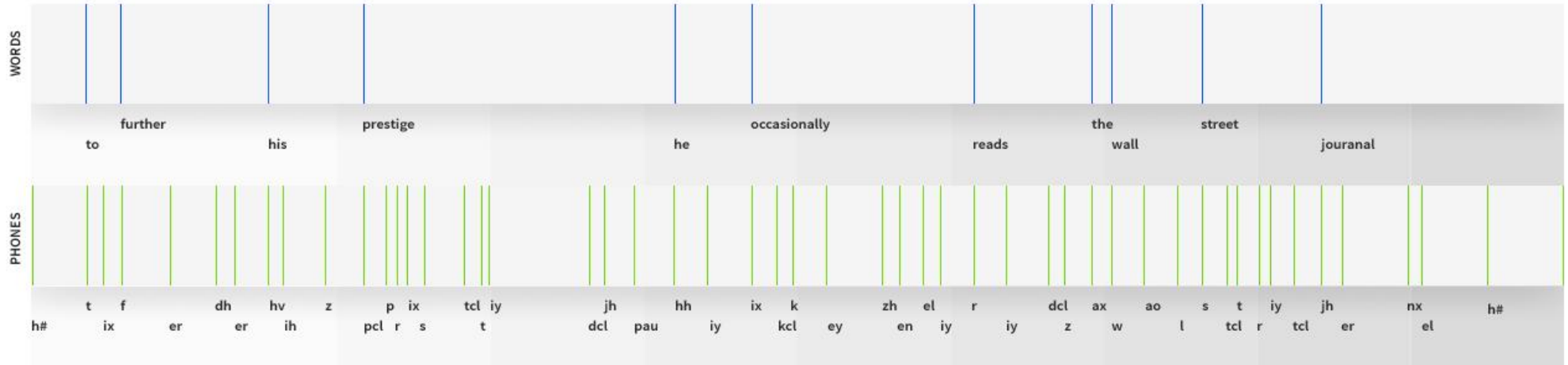
Database:
1,417,588 spectrograms for training
222,789 spectrograms for validation
294,101 spectrograms for testing





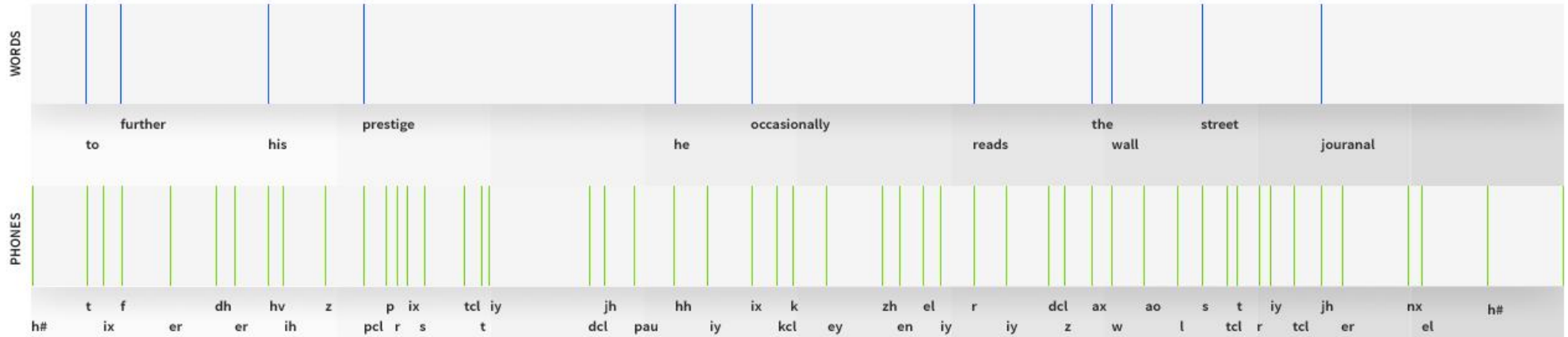
Glackin, Cornelius, Gerard Chollet, Nazim Dugan, Nigel Cannings, Julie Wall, Shahzaib Tahir, Indranil Ghosh Ray, and Muttukrishnan Rajarajan. "Privacy preserving encrypted phonetic search of speech data." In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pp. 6414-6418. IEEE, 2017.

Pick an utterance from the TIMIT test set

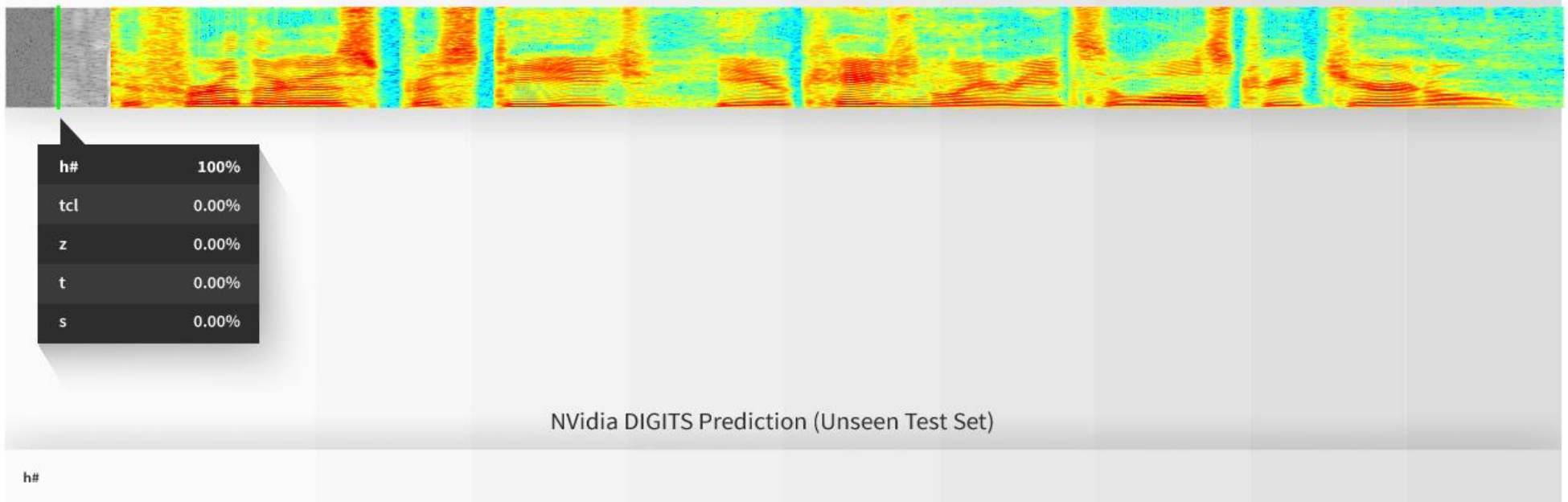


Intelligent Voice®

Pick an utterance from the TIMIT test set

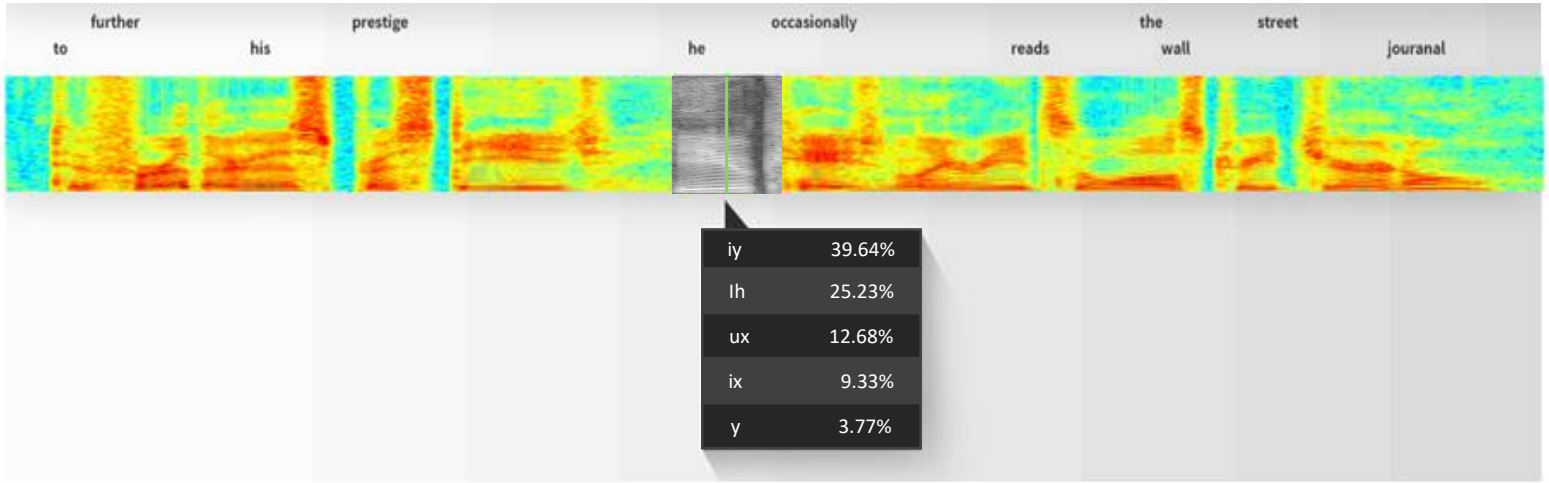


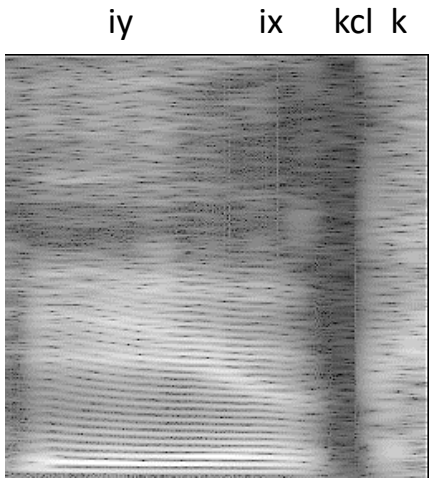
Taking a spectogram of the test utterance, we can then slide a window across the spectogram,
and test our trained NVidia DIGITS deep convolutional neural network

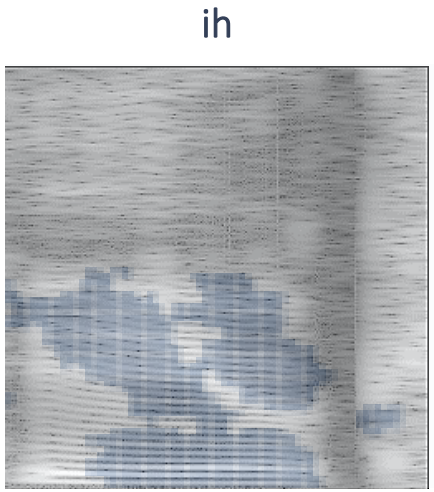


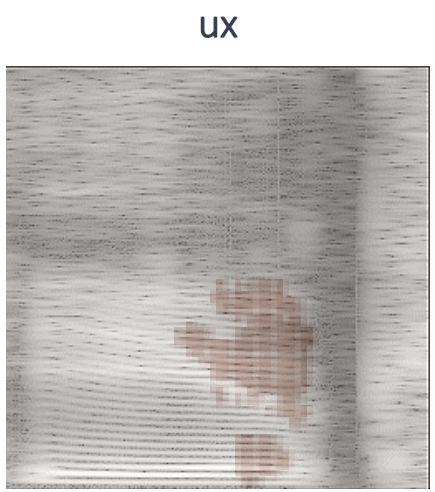
Intelligent Voice®

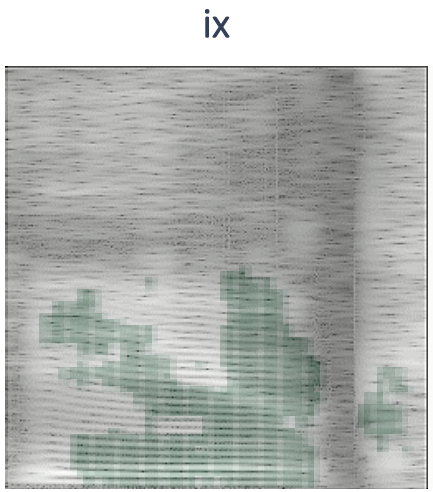
What Does the CNN See?

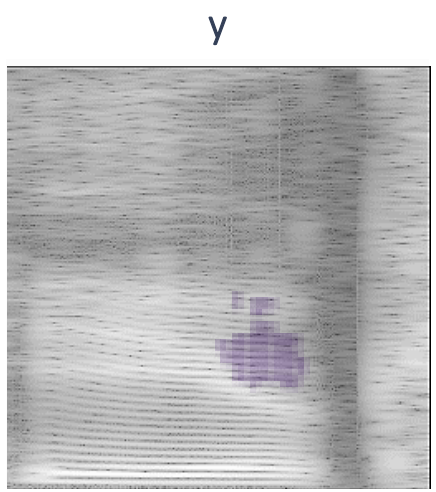






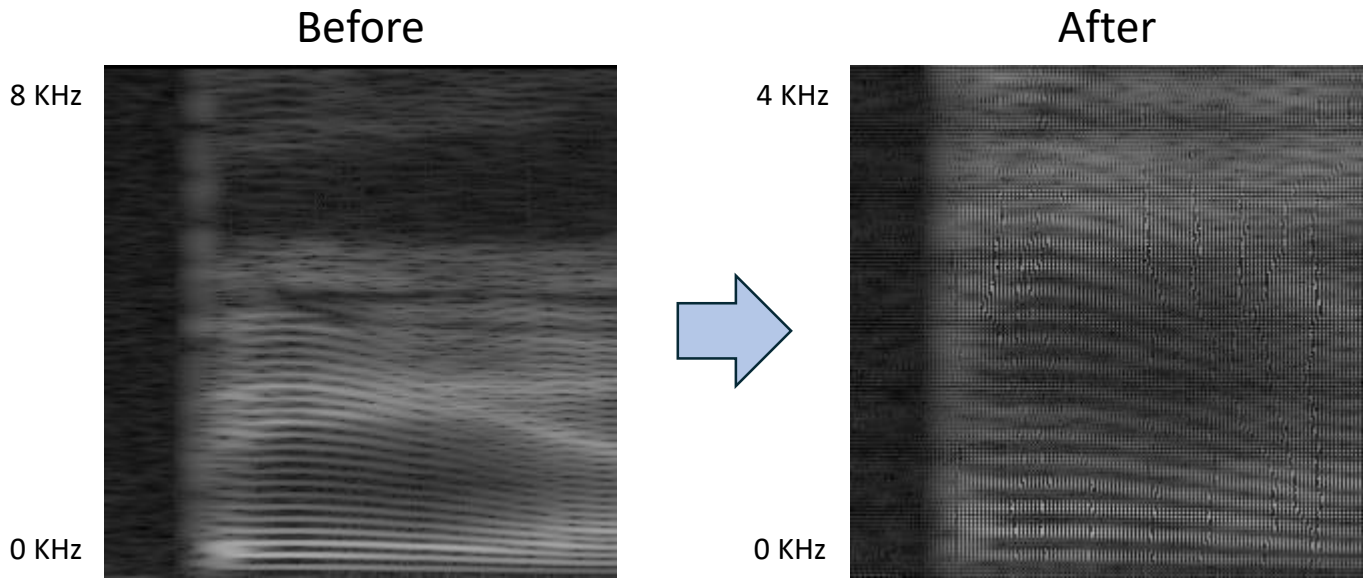






Making Use of Deconvolution Insight

- Deconvolution shows that the CNN’s automated feature extraction focuses on the first 4KHz
- Fricative sounds, like "s", can contain higher frequencies but they can reliably be identified at lower frequency range
- By concentrating on 0-4KHz range with the same resolution of spectrogram image we can improve classification accuracy by a couple of points.



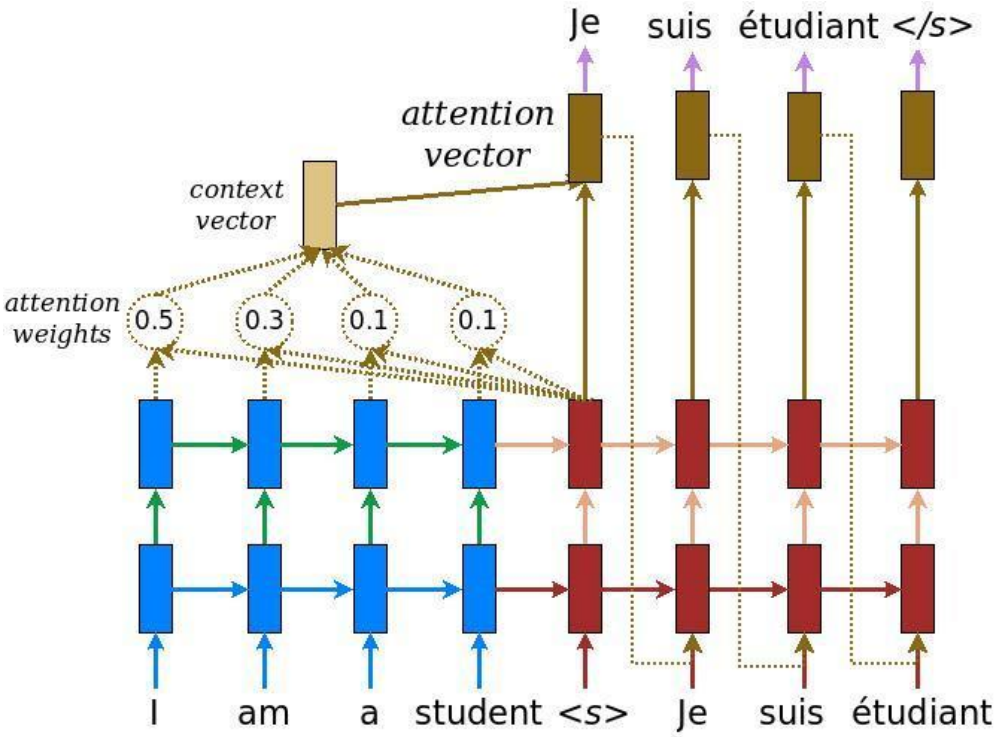
RNN Explainability

Before the attention mechanism RNN sequence to sequence models had to compress the input of the encoder into a fixed length vector

Without attention a sentence of hundreds of words the compression led to information loss resulting in inadequate translation.

Attention mechanism extends memory of the RNN seq2seq model by inserting a context vector between the encoder and decoder.

The context vector takes all cells' outputs as input to compute the probability distribution of source language words for each single word decoder wants to generate.



To build context vector, loop over all encoders' states to compare target and source states to generate scores for each state in encoders. Then use softmax to normalize all scores, which generates the probability distribution conditioned on target states.

Finally, weights are introduced to make context vector easy to train to train.

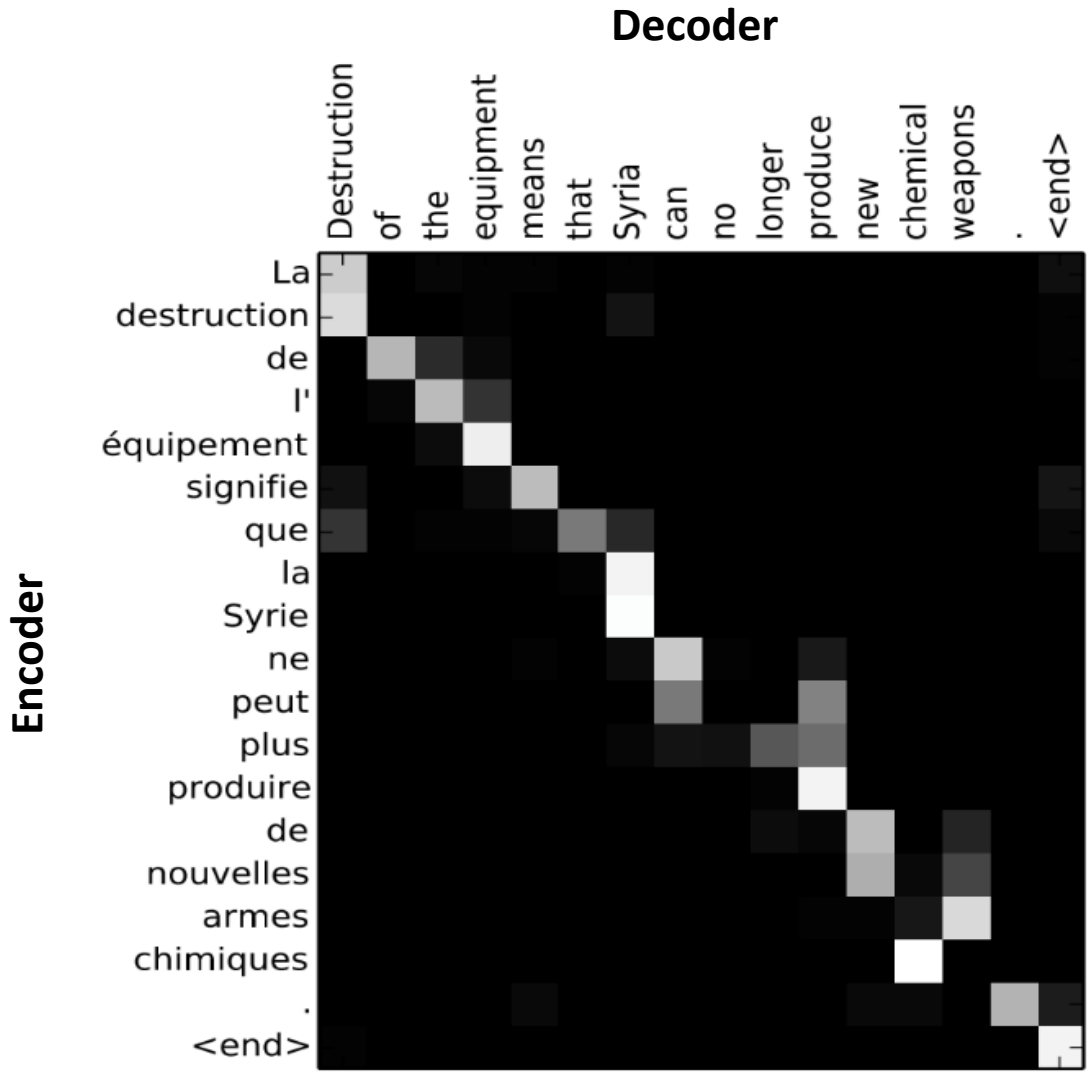
RNN Explainability

There are many variants in the attention mechanism e.g. soft, hard, additive, etc.

This development in the state-of-the-art with seq2seq RNNs also provides insight into the how these models make decisions.

The attention mechanism was developed for seq2seq models but is now also being used for providing insight into CNN-RNN models.

Matrix shows that while translating from French to English, the network attends sequentially to each input state, but sometimes it attends to two words at time while producing an output, as in translation “la Syrie” to “Syria”.



Attention with CNN/RNN Architecture

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

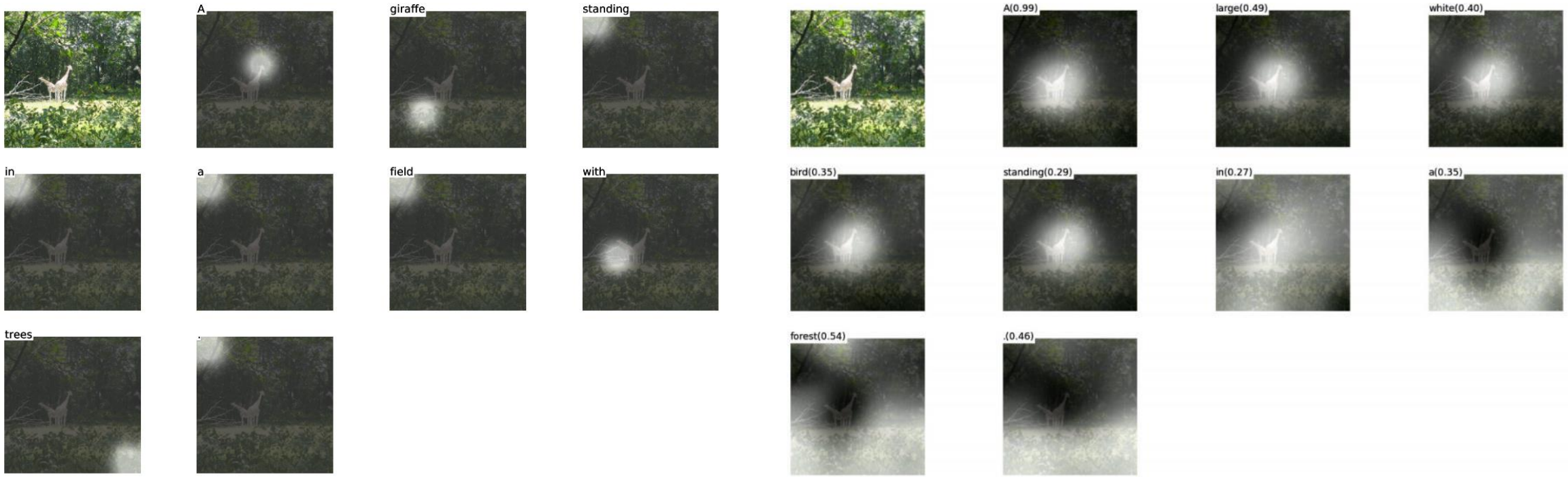
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Xu et al., 'Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,' arXiv:1502.03044v3, 2016.

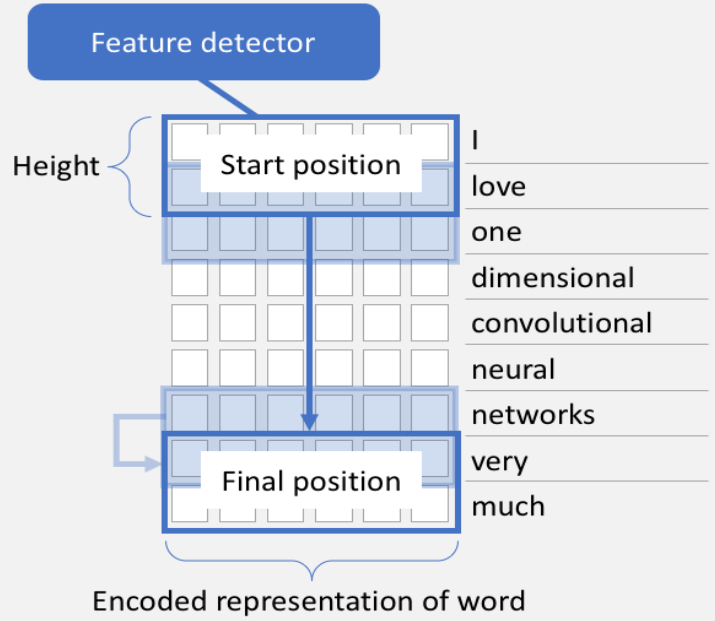
Attention with CNN/RNN Architecture



Xu et al., 'Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,' arXiv:1502.03044v3, 2016.

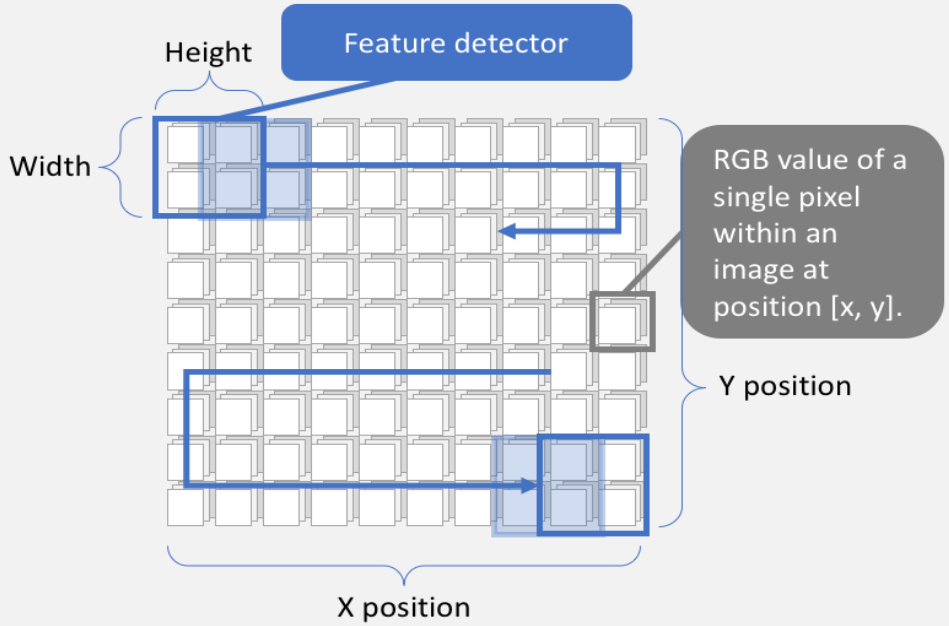
Can Replace RNN Cells With 1-D Convolutions

1D Convolutional - Example



In this example for natural language processing, a sentence is made up of 9 words. Each word is a vector that represents a word as a low dimensional representation. The feature detector will always cover the whole word. The height determines how many words are considered when training the feature detector. In our example, the height is two. In this example the feature detector will iterate through the data 8 times.

2D Convolutional - Example

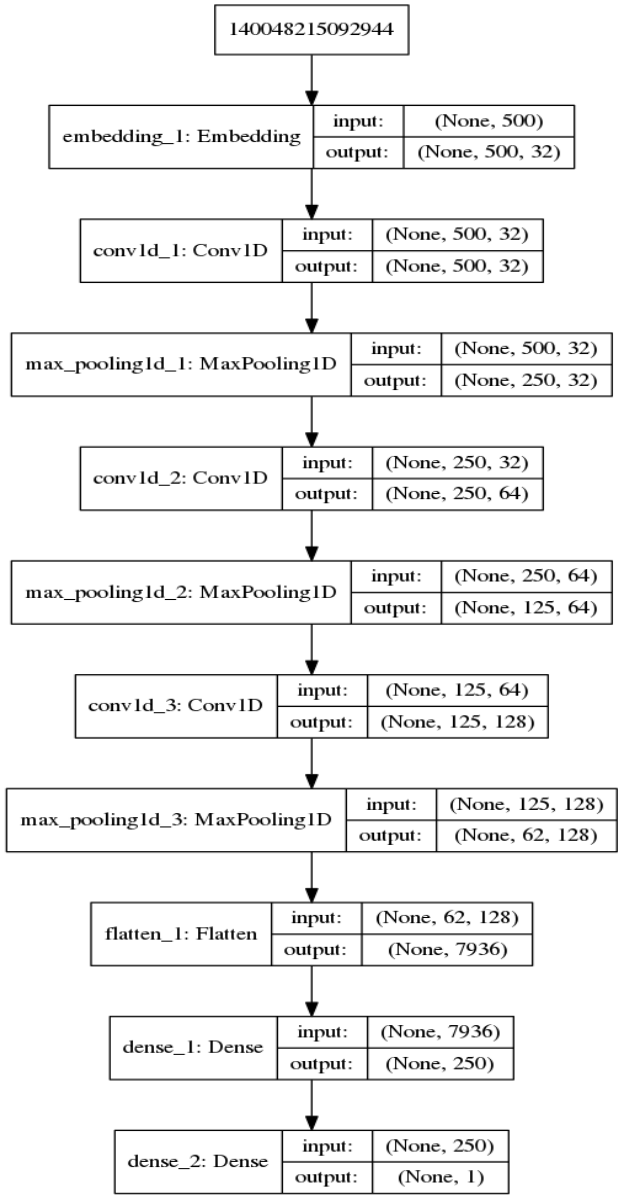


In this example for computer vision, each pixel within the image is represented by its x- and y position as well as three values (RGB). The feature detector has a dimension of 2 x 2 in our example. The feature detector will now slide both horizontally and vertically across the image.

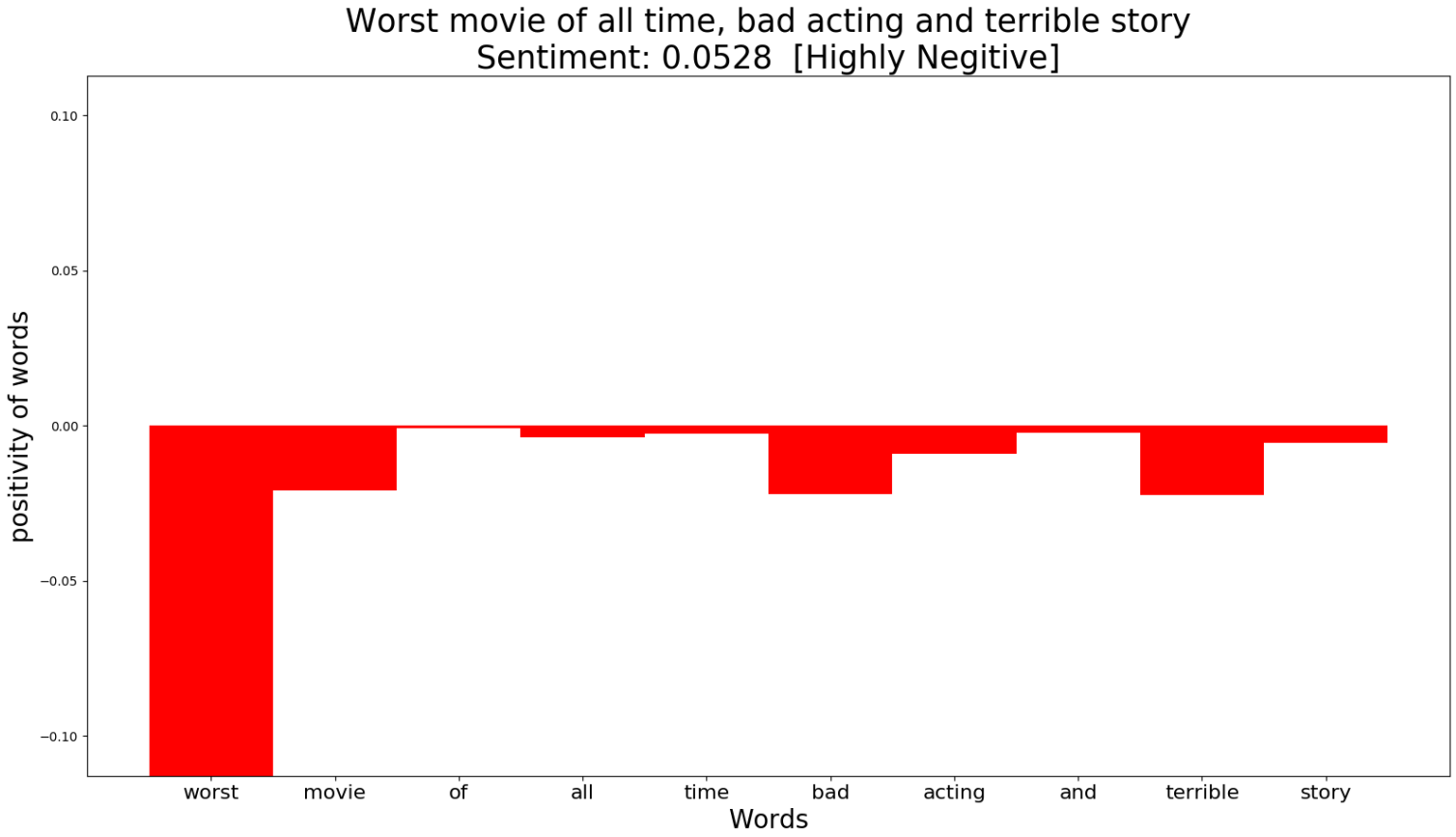
Ackerman, N. Introduction to 1D Convolutional Neural Networks in Keras for Time Sequences, Medium, 2019.

Explaining Sentiment

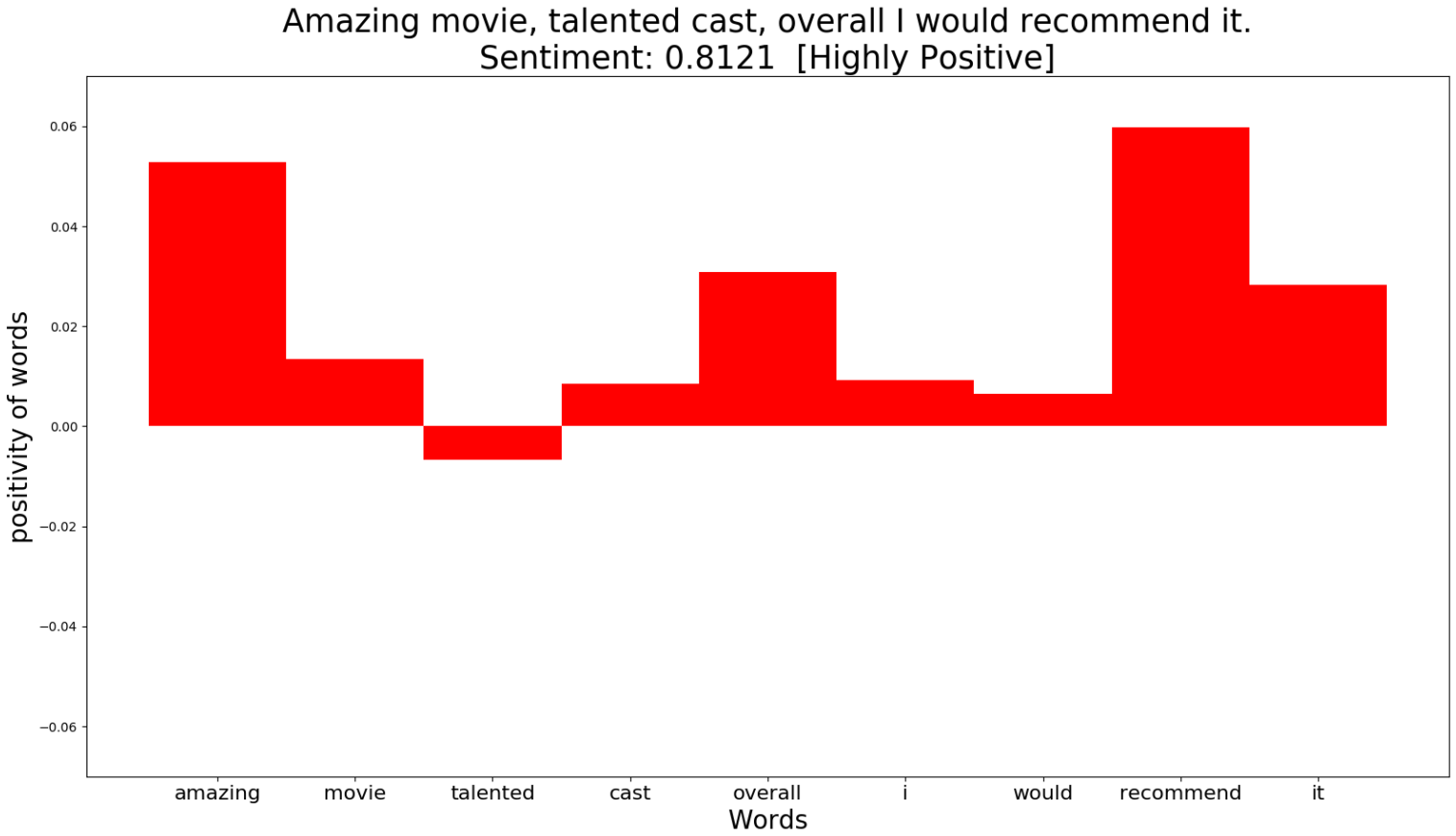
- Example 1-D convolution architecture
- Famous IMDB sentiment analysis dataset
  - Movie Reviews: 0 – 10 Score
- Typical LSTM-based approach has been improved with Conv 1-D cells
- We can the apply the occlusion principle to Conv 1-D cells
- Provides a way to explain text classification



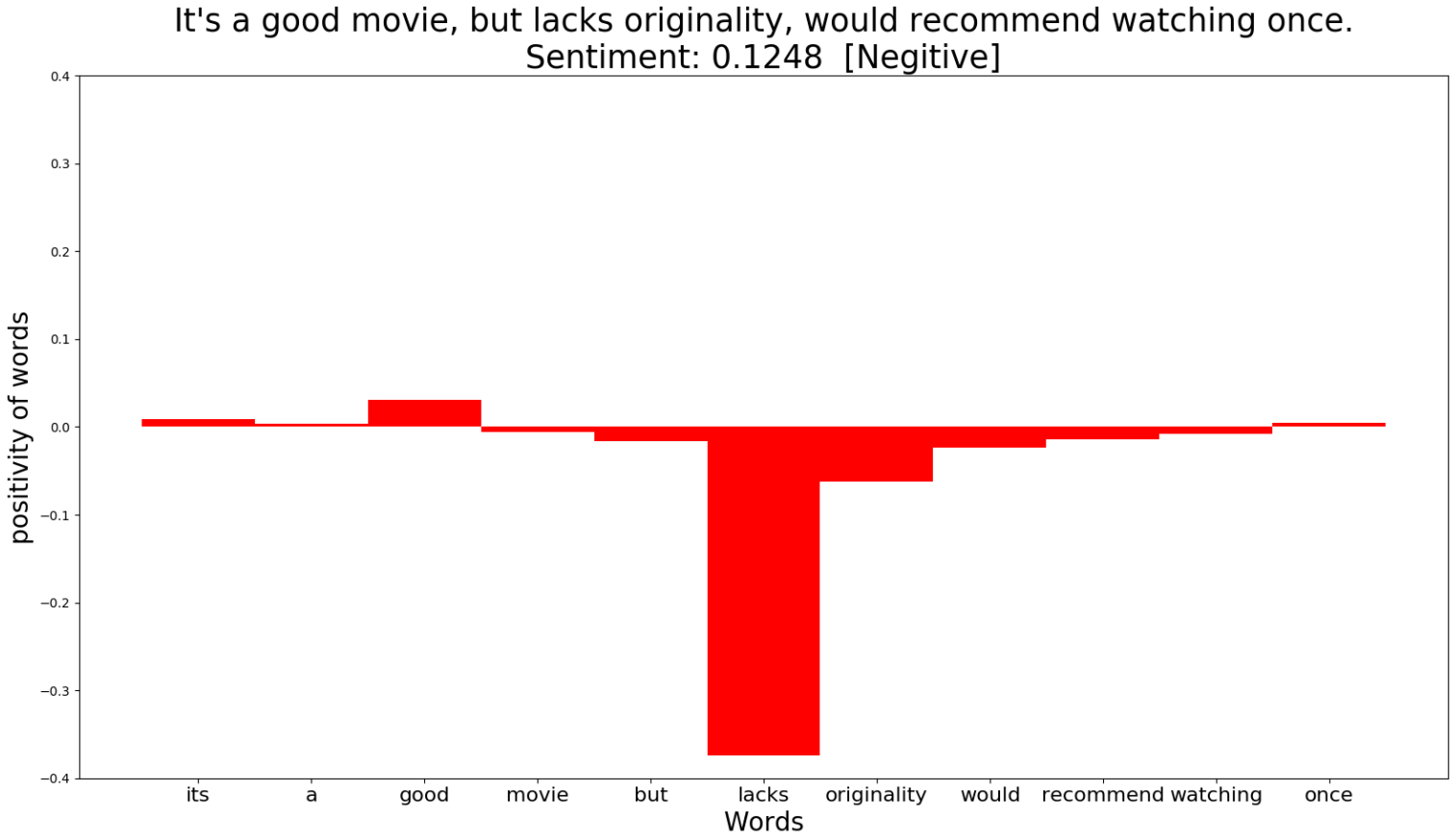
Explaining Sentiment



Explaining Sentiment

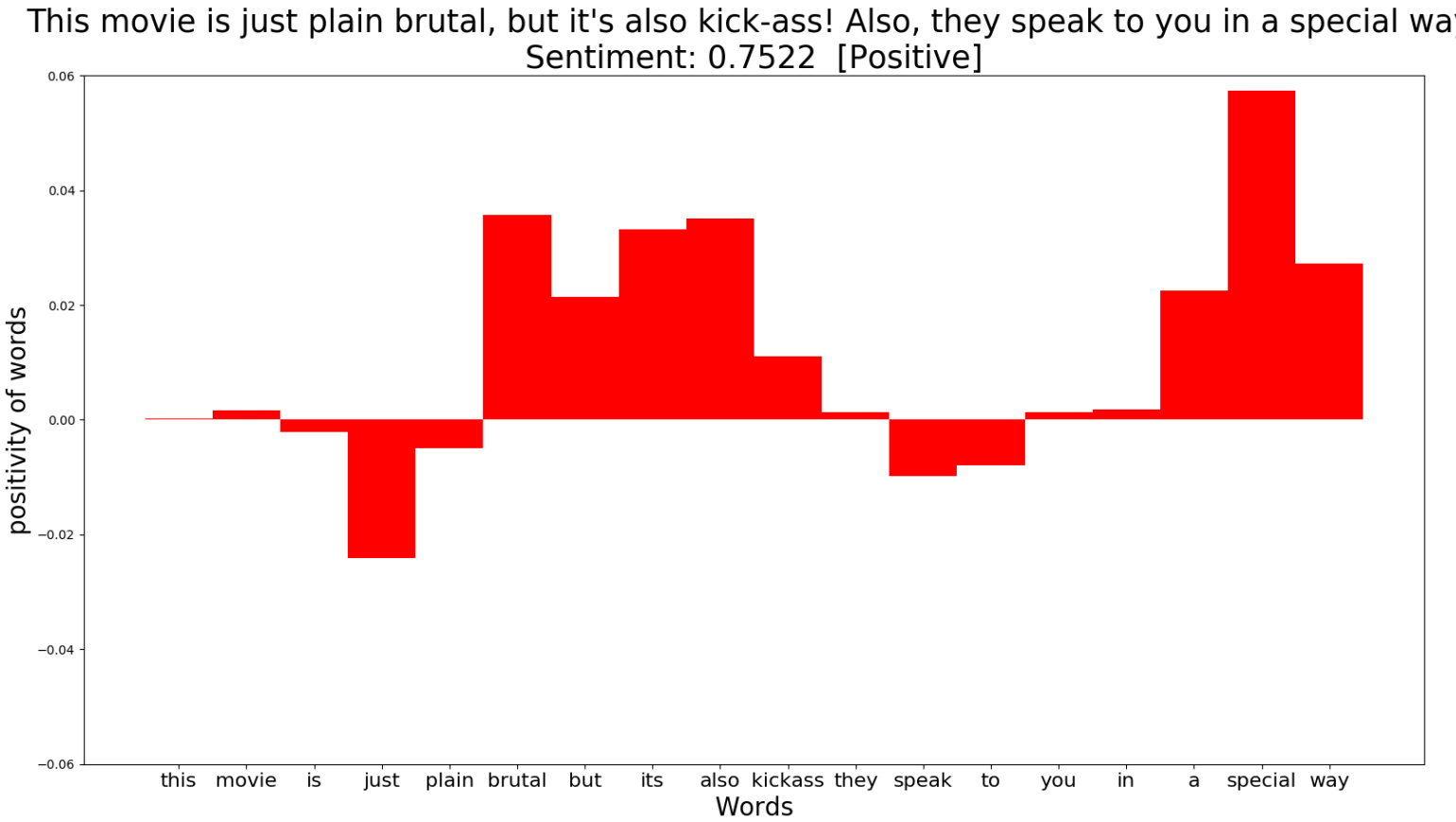


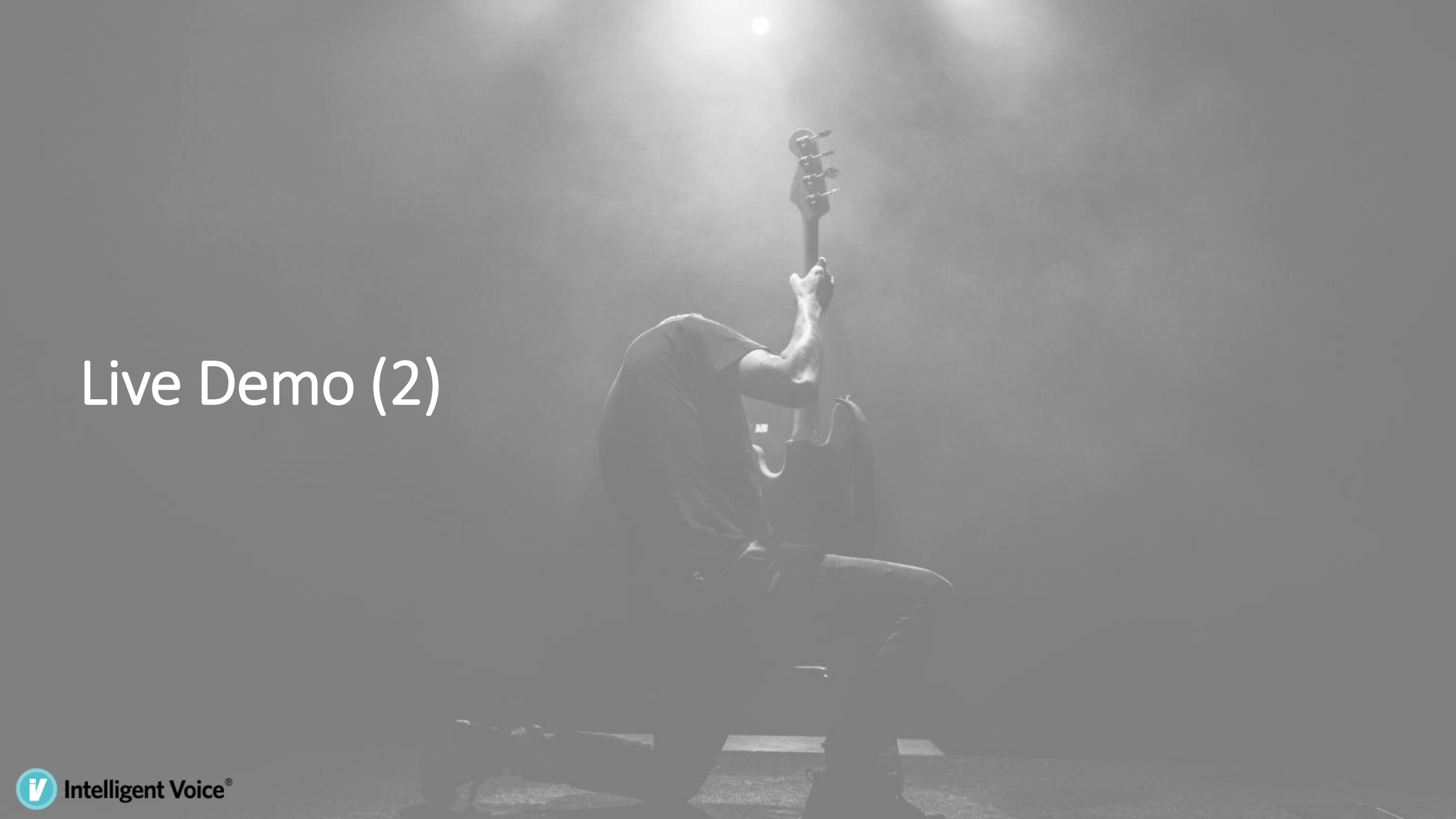
Explaining Sentiment



Intelligent Voice®

Explaining Sentiment





Live Demo (2)

Intelligent Voice®

Importance of Explainability

A pair of computer scientists at the University of California, Berkeley developed an AI-based attack that targets speech-to-text systems. With their method, no matter what an audio file sounds like, the text output will be whatever the attacker wants it to be.

They can duplicate any type of audio waveform with 99.9 percent accuracy and transcribe it as any phrase they chose at a rate of 50 characters per second with a 100 percent success rate.

Mozilla's DeepSpeech implementation was used.

Original

'without the dataset the article is useless'

Adversarial

okay google browse to evil dot com

https://nicholas.carlini.com/code/audio_adversarial_examples/

Deep Learning Roadmap for Explainability

Deep learning models with decision-making, human-facing application need to be explainable for legal reasons.

Explainability provides insight into DL models that in turn provides valuable insight into how to improve them.

We have demonstrated that CNNs can let you see what they are thinking with deconvolution.

Attention mechanisms also provides insight into RNN models and CNN-RNN models

Moving forward, explainability needs to be something that is built into deep learning architectures.

IBM Wants To Make Artificial Intelligence Fair And Transparent With AI OpenScale

Janakiram MSV Contributor ⓘ
Enterprise & Cloud
I cover Cloud Computing, Machine Learning, and Internet of Things

IBM has announced AI OpenScale, a service that aims to bring visibility and explainability of AI models for enterprises.

Explainability/transparency/interpretability are becoming really important with many companies and researchers looking into it. We need to make sure we are actually looking inside the box.

Conclusion

Nigel Cannings, CTO

Intelligent Voice®

@intelligentvox
www.intelligentvoice.com

Hewlett Packard Enterprise & Intelligent Voice

Come and check our demo at HPE Booth #1129