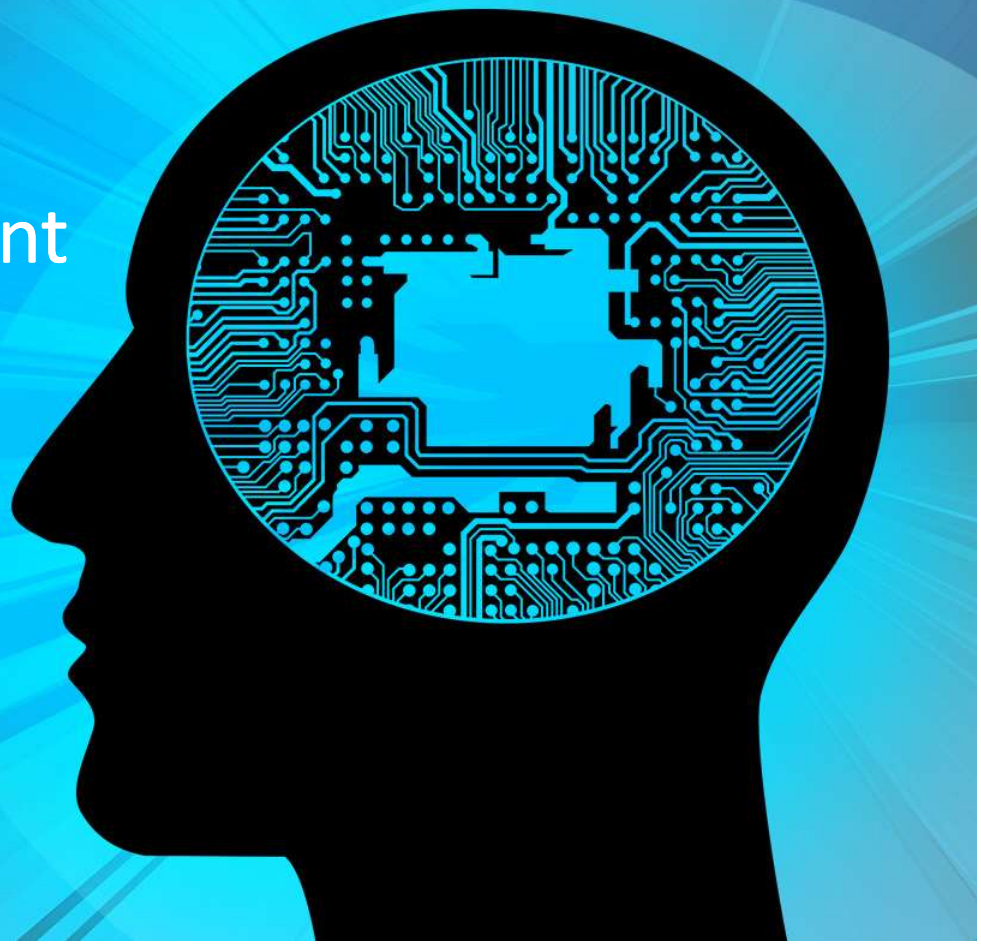
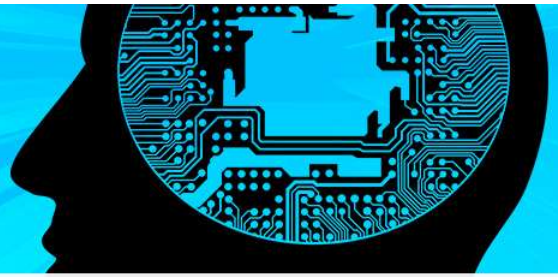
Attention based recurrent neural networks for keyword spotting

Nicolò Dal Fabbro

Davide Masiero



Index



- Introduction
- Dataset and Preprocessing
- Models
  - CNN
  - RNN
  - Attention Mechanism
- Results
- Application

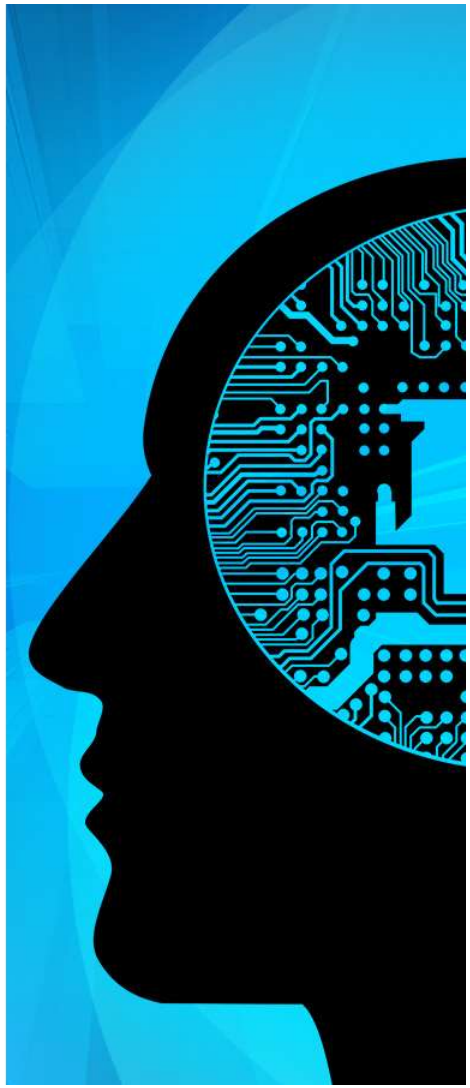
Introduction

- GOAL:

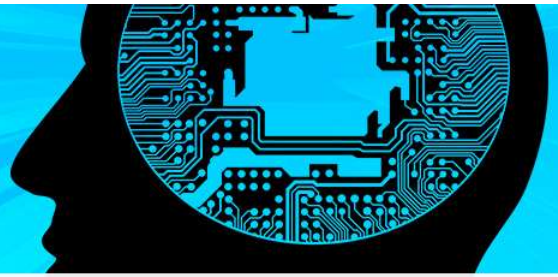
  Design a model for the Keyword Spotting task

- HOW:

  Exploring an attention-based CNN/RNN architecture

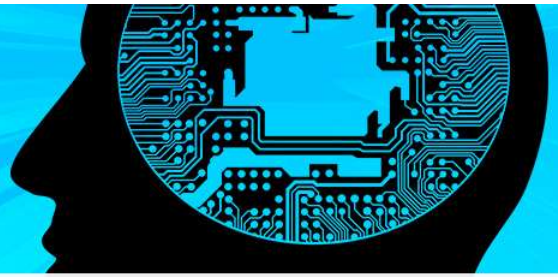


Dataset And Preprocessing



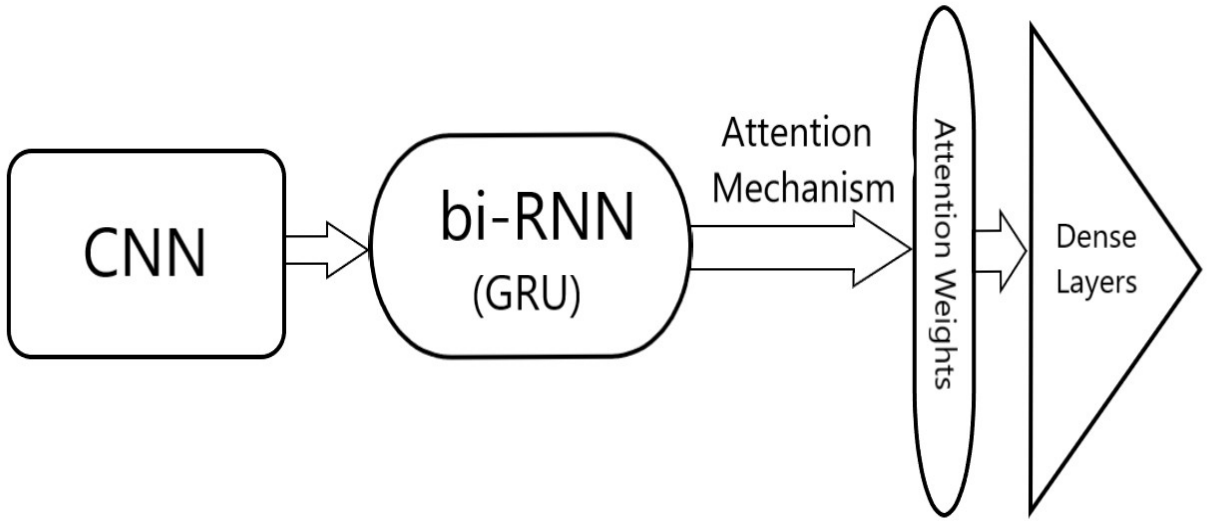
- Dataset:
  - Google Keyword Spotting Dataset V2 (http://download.tensorflow.org/data/speech_commands_v0.02.tar.gz)
  - 35 classes
  - Length of the recordings: 1sec
  - Sampling rate: 16000Hz

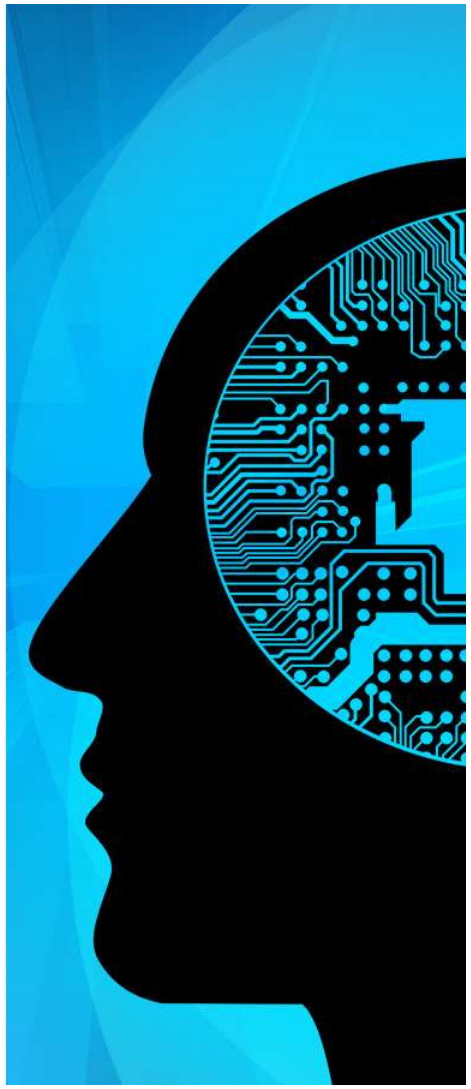
Dataset And Preprocessing



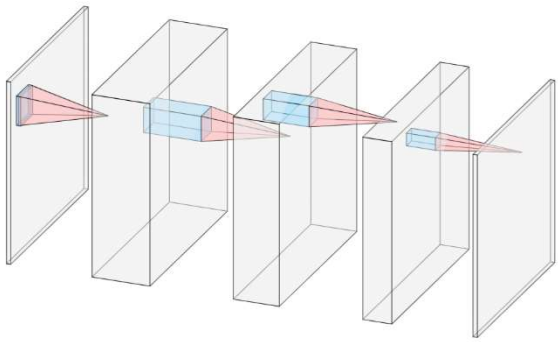
- Preprocessing and feature vectors:
  - Zero padding
  - 80 band mel scale
  - 1024 Discrete Fourier Transform points
  - Window length: 0.02s
  - Hop size: 0.01s

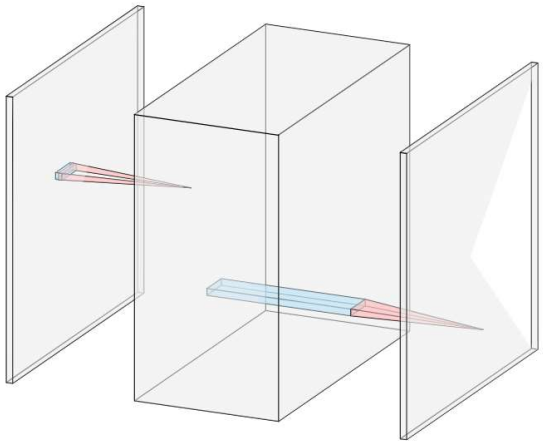
Model

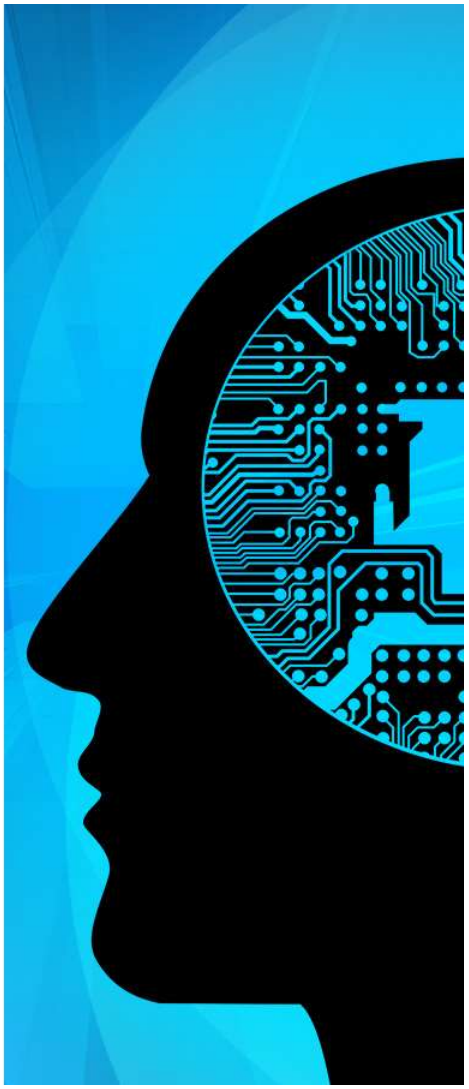

CNN
bi-RNN
(GRU)
Attention
Mechanism
Attention Weights
Dense
Layers



CNN

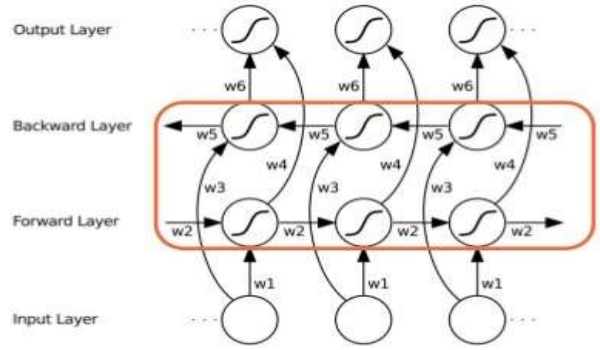


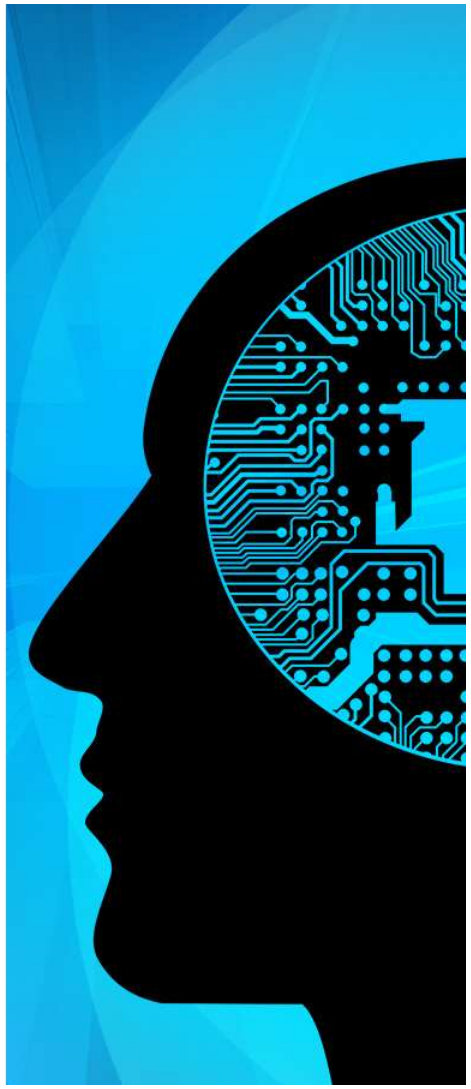




RNN

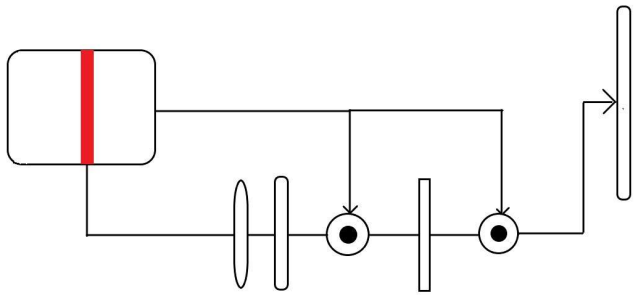
- We try different configurations of Bi-RNN:
  - GRU/LSTM
  - One or two Bidirectional RNN
  - Different number of units (32, 48, 64)
  - Passing or not the state between them

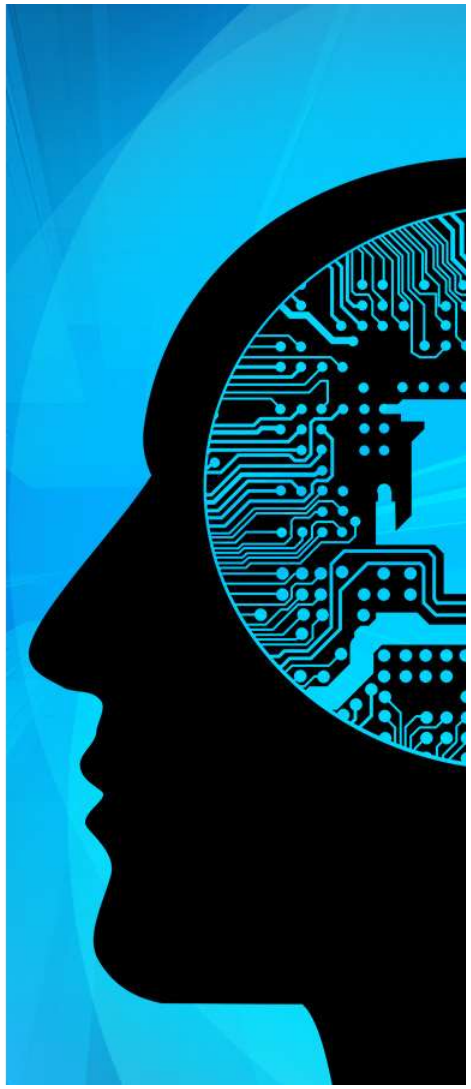




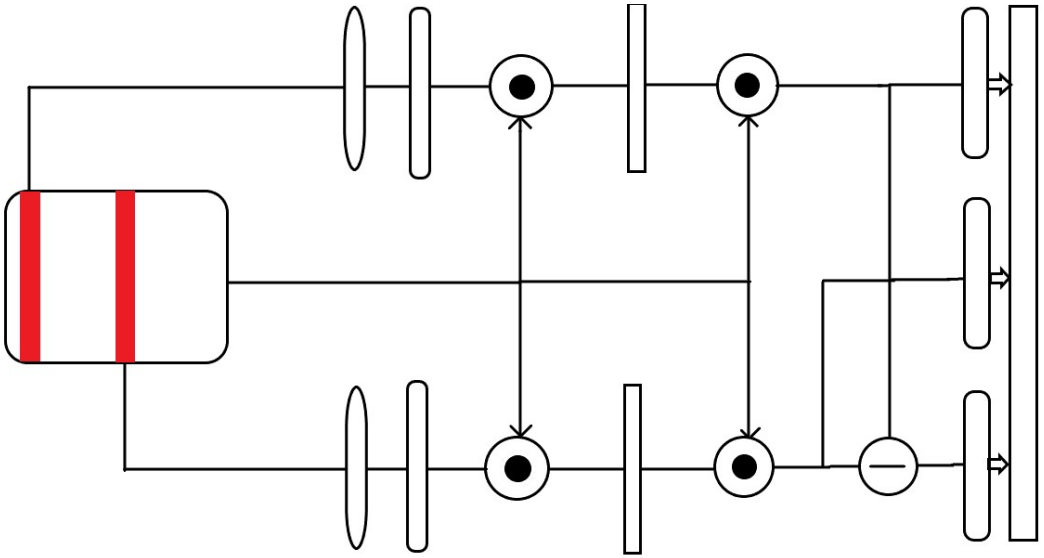
Attention Mechanism

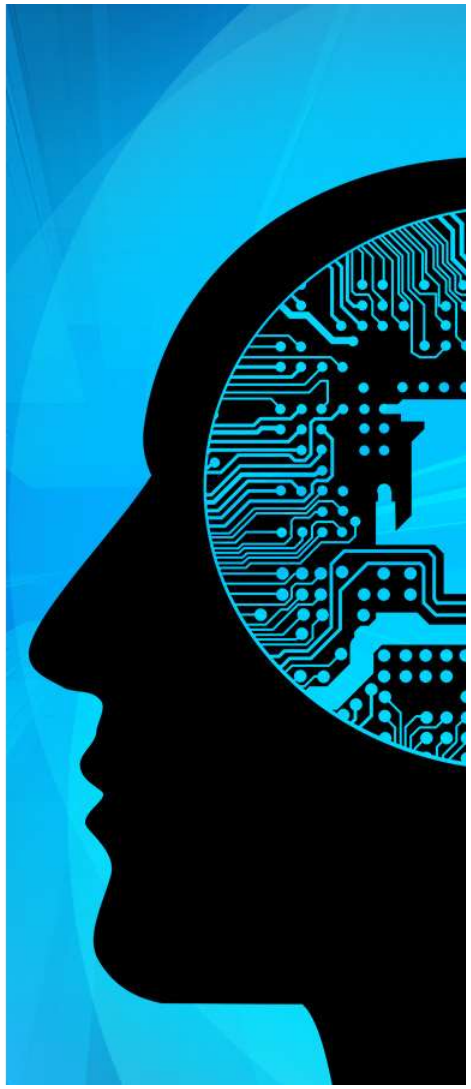
- Attention mechanism is used to train the network to understand which part of the signal is more important for the generation of the output.



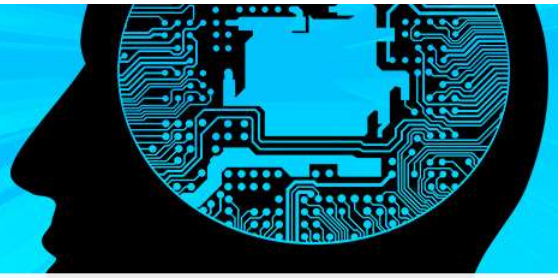


Attention Mechanism



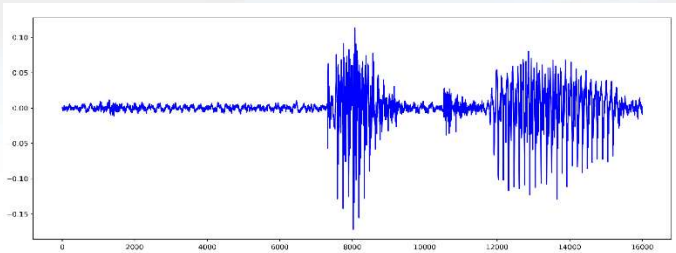


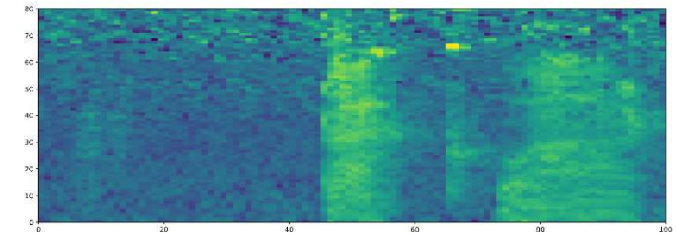
Results

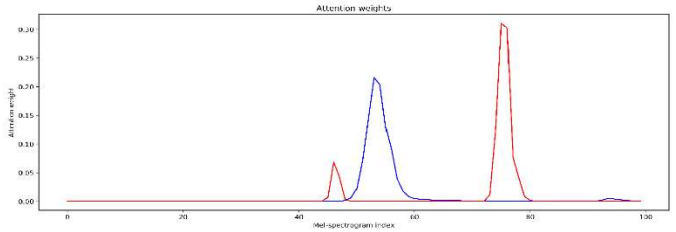


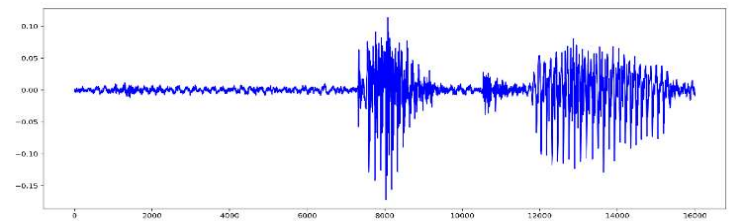
Model	Test Accuracy
CNNRNNAttState	92.55
CNNRNNAttDouglas	92.62
CNNRNNStatefulAttDouglas	92.54
PCNNRNNStatefulAttDouglas	88.26
MultiAttention	93.11
MultiAttentionDouglas	92.50
MultiAttentionDiff	**94.10**
Douglas (paper)	93.90
Douglas (by us)	92.71

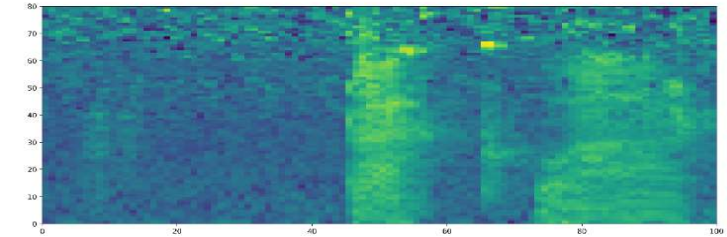
Results

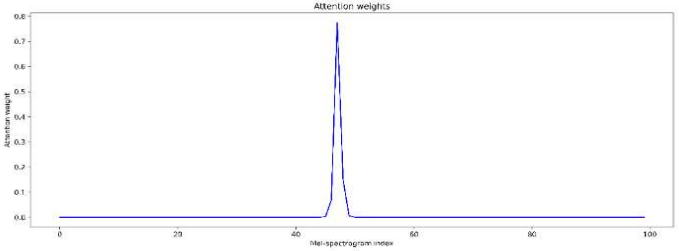




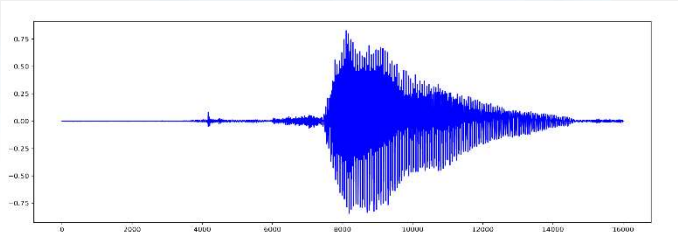

Attention weights
Attention weight
Mel-spectrogram index

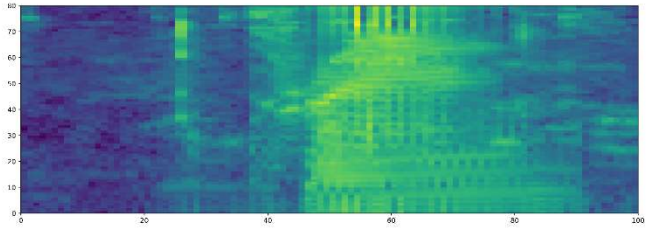


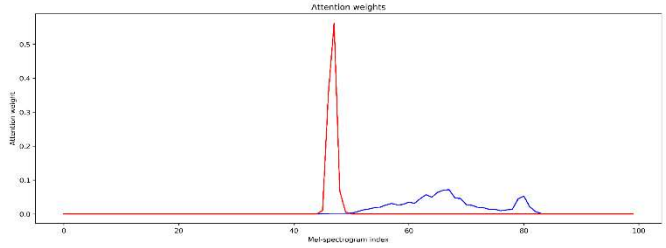


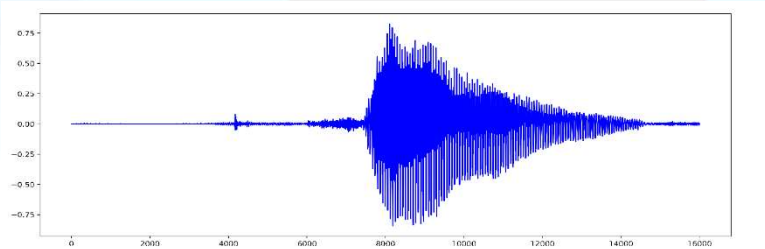

Attention weights
Attention weight
Mel-spectrogram index

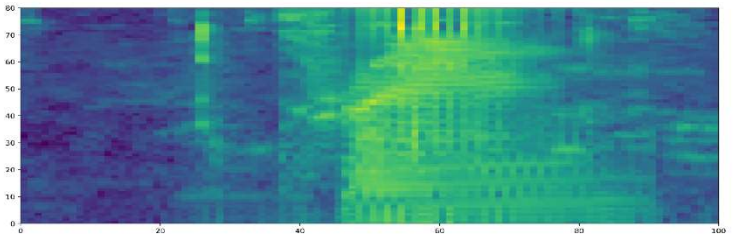
Results

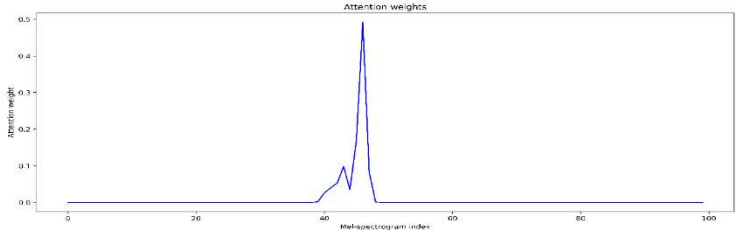




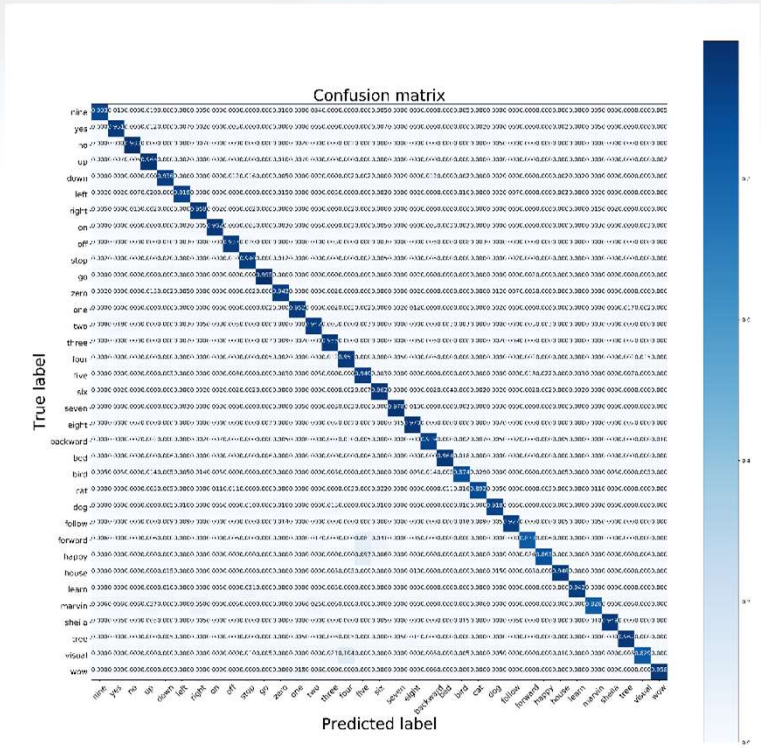

Attention weights
Attention weight
Mel-spectrogram index

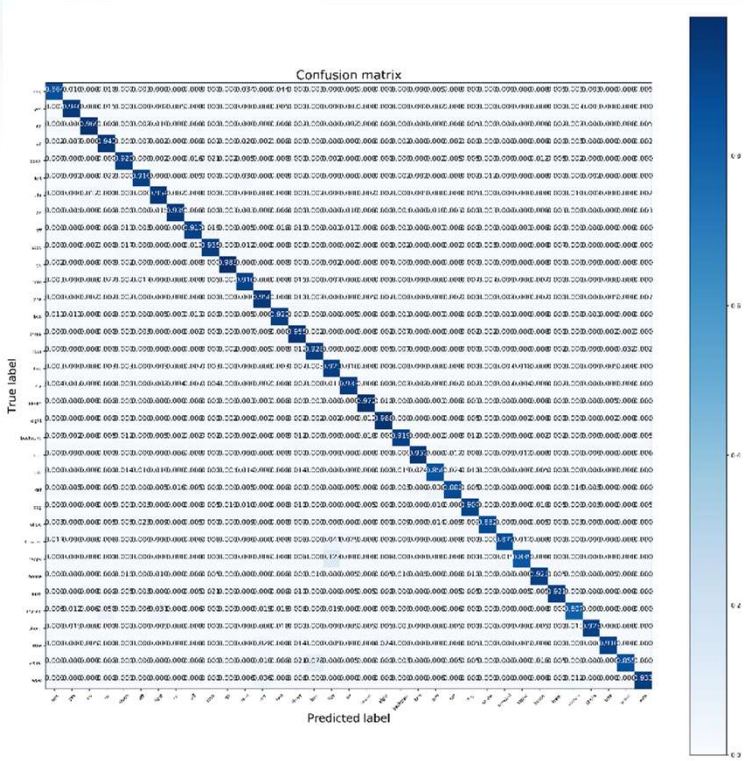





Attention weights
Attention weight
Mel-spectrogram index

Results

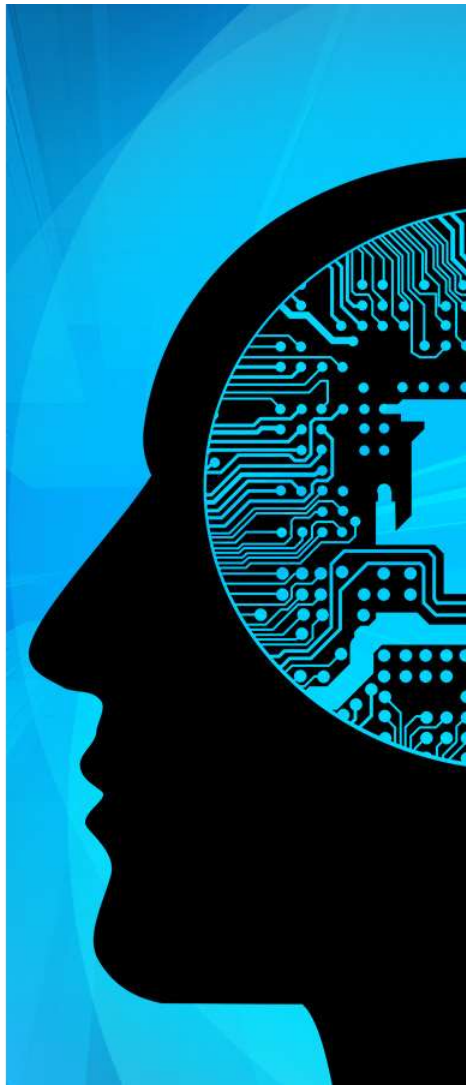

Confusion matrix
True label
Predicted label
nine
yes
no
up
down
left
right
on
off
stop
go
zero
one
two
three
four
five
six
seven
eight
backward
bed
bird
cat
dog
follow
forward
happy
house
learn
marvin
sheila
tree
visual
wow


Confusion matrix
True label
Predicted label

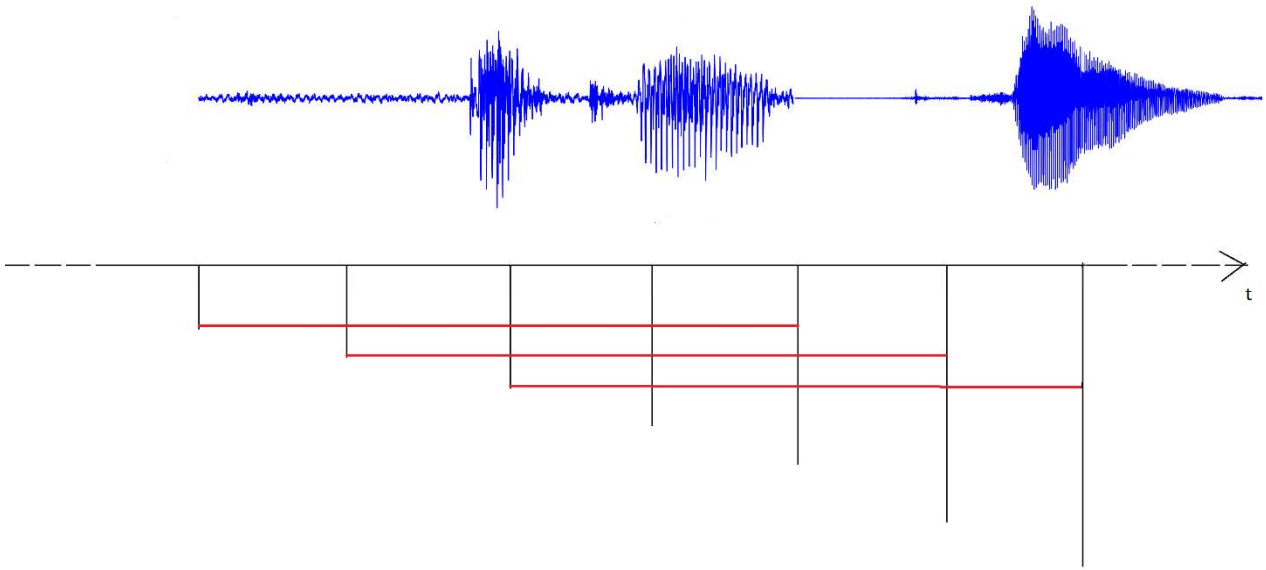
Application

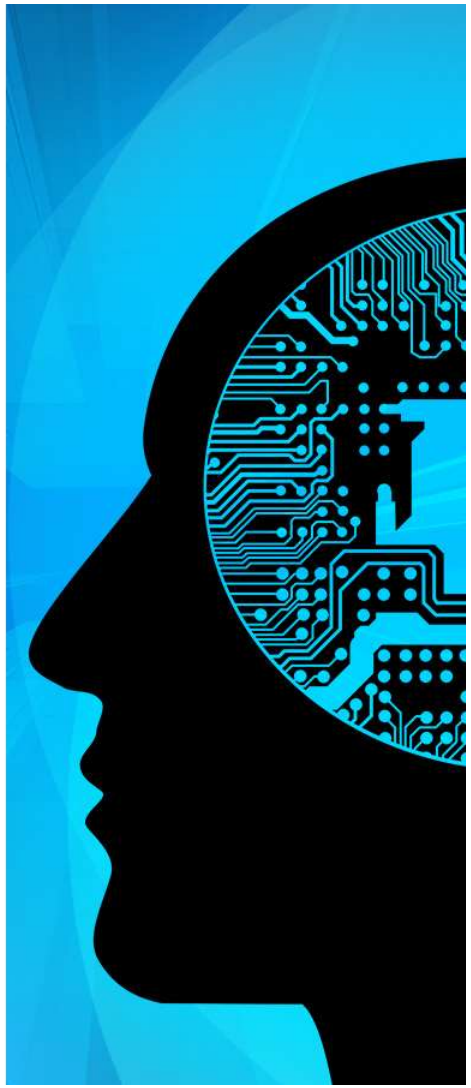
- Our application predicts words spoken by a user in a real time context.

- We re-train our best model in a dataset with noisy samples.



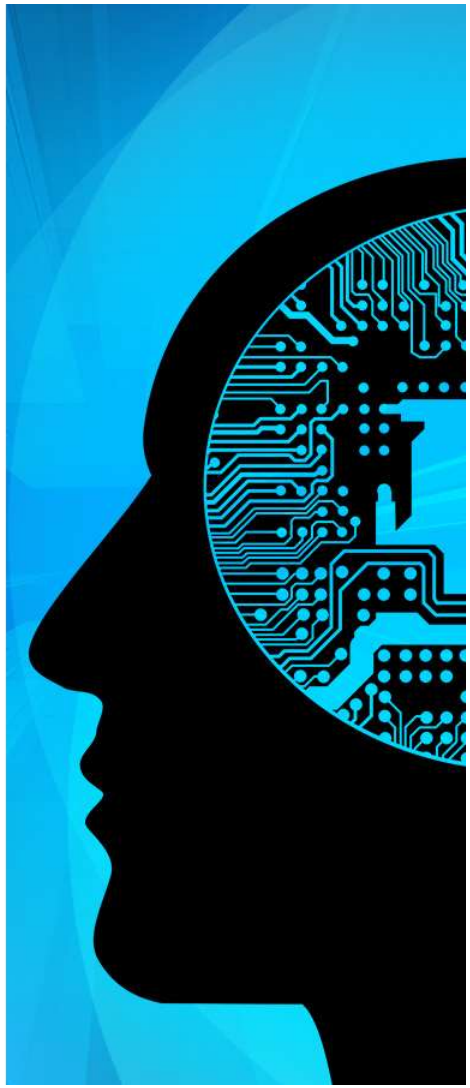
Application





Application

- To classify the words said by the user:
  - The prediction in a sequence of 3 windows must be the same
  - A threshold is applied to the score of the prediction (80%).



THE END
ENJOY THE DEMO