Spoken Keyword Spotting

Vineeth S

July 2020

Project Overview

- Spoken Keyword Spotting is the task of identifying predefined words (called as keywords) from speech.

Project Overview

- Spoken Keyword Spotting is the task of identifying predefined words (called as keywords) from speech.
- For the of this project, we will be using the Google Speech Commands Dataset[5].

Project Overview

- Spoken Keyword Spotting is the task of identifying predefined words (called as keywords) from speech.
- For the of this project, we will be using the Google Speech Commands Dataset[5].
- Speech Commands dataset has 65,000 one-second long utterances of 30 short words recorded by people with different demographics.

Project Overview

- Spoken Keyword Spotting is the task of identifying predefined words (called as keywords) from speech.
- For the of this project, we will be using the Google Speech Commands Dataset[5].
- Speech Commands dataset has 65,000 one-second long utterances of 30 short words recorded by people with different demographics.
- For this project we will be using a single word as a keyword. The dataset also contains two names — Marvin and Sheila — out of which we use **Marvin** as our hotword (keyword).

KWS System Pipeline

- The input to the system is via Tensorflow Dataset Object which provides efficient handling of large sized data, generating features in an ad-hoc fashion.

KWS System Pipeline

- The input to the system is via Tensorflow Dataset Object which provides efficient handling of large sized data, generating features in an ad-hoc fashion.
- The input feature for the system is the log Mel Filterbank energies of the speech signal calculated with a window of length $25ms$ and stepsize $10ms$.

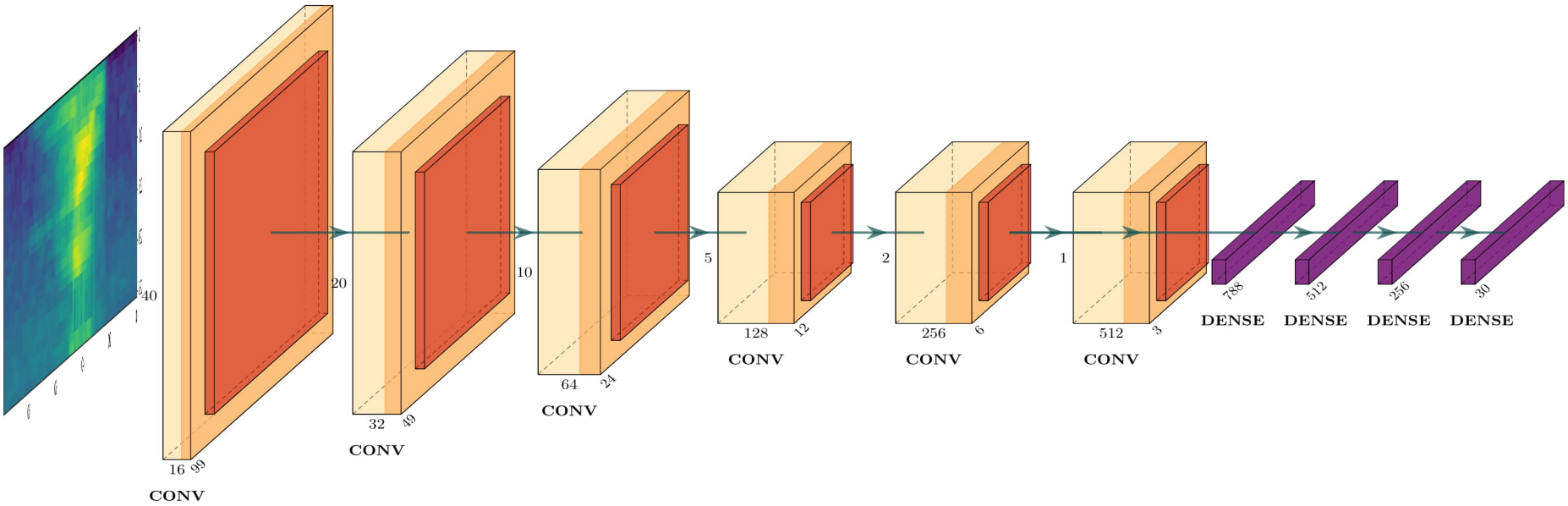
KWS System Pipeline

- The input to the system is via Tensorflow Dataset Object which provides efficient handling of large sized data, generating features in an ad-hoc fashion.
- The input feature for the system is the log Mel Filterbank energies of the speech signal calculated with a window of length $25ms$ and stepsize $10ms$.
- The keyword detection system comprises of two models — a feature (embedding) extractor and a SVM classifier.

KWS System Pipeline

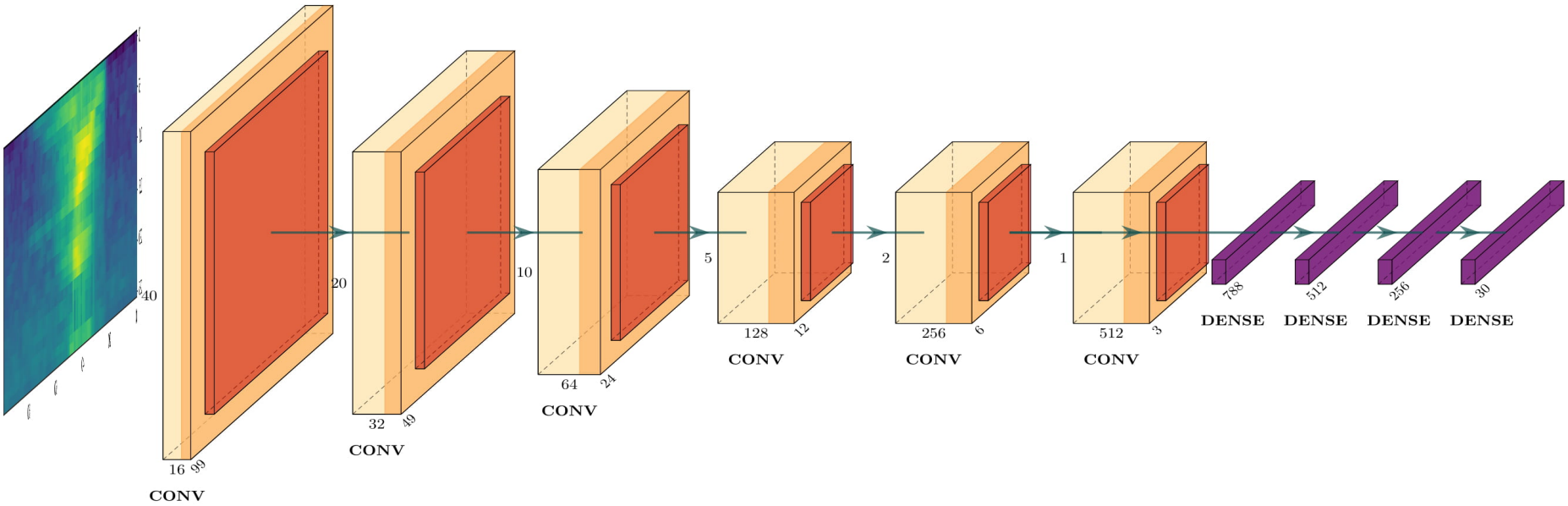
- The input to the system is via Tensorflow Dataset Object which provides efficient handling of large sized data, generating features in an ad-hoc fashion.
- The input feature for the system is the log Mel Filterbank energies of the speech signal calculated with a window of length $25ms$ and stepsize $10ms$.
- The keyword detection system comprises of two models — a feature (embedding) extractor and a SVM classifier.
- The output of the system would be a binary value predicting whether the given speech sequence is a keyword.

Speech Commands Classifier



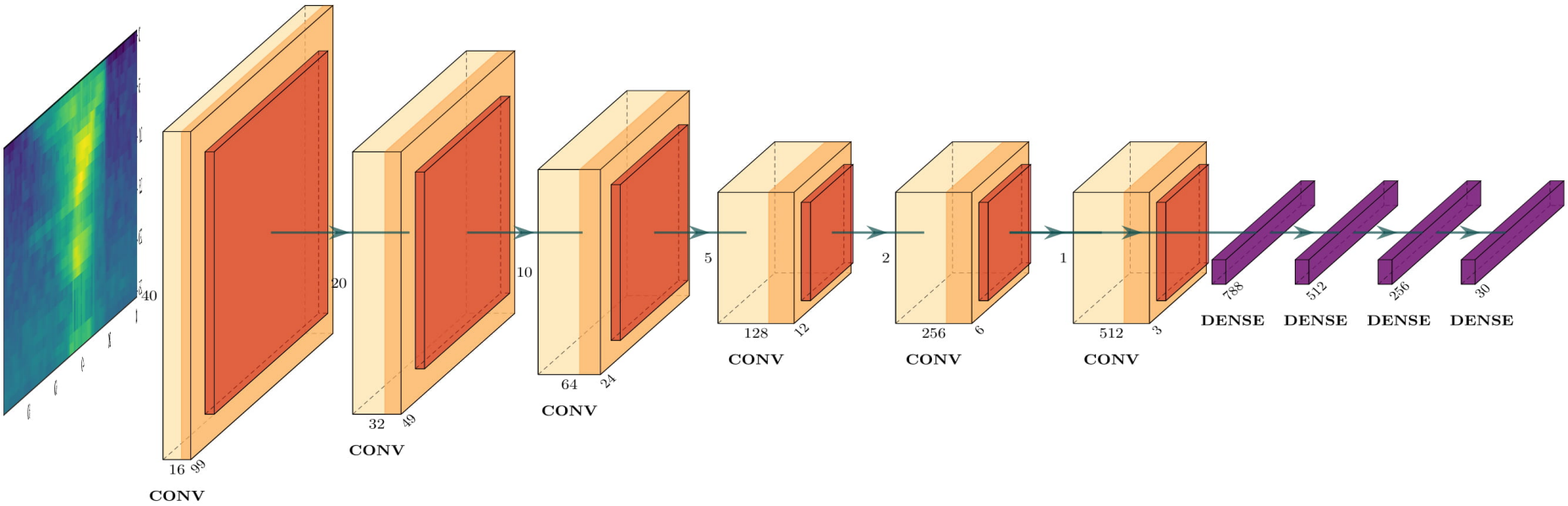
- The above is a word classifier for Speech Commands dataset.

Speech Commands Classifier



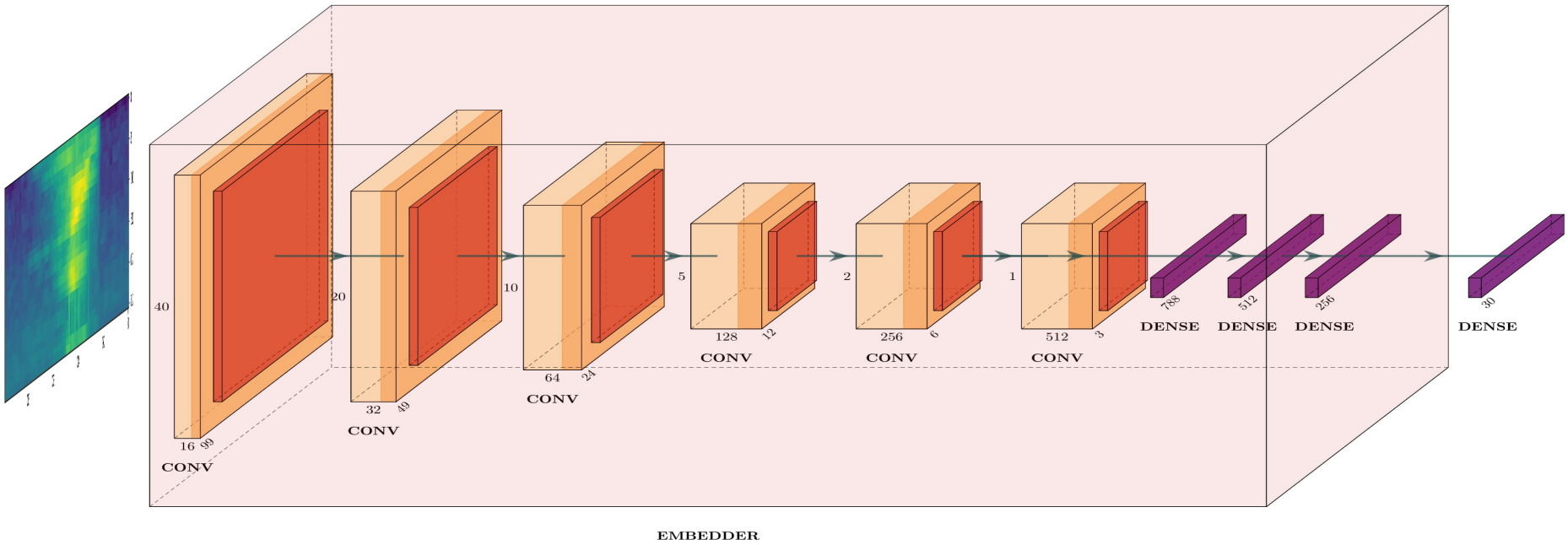
- The above is a word classifier for Speech Commands dataset.
- This model achieves an training accuracy of 96.59% and validation accuracy of 95.56%.

Speech Commands Classifier



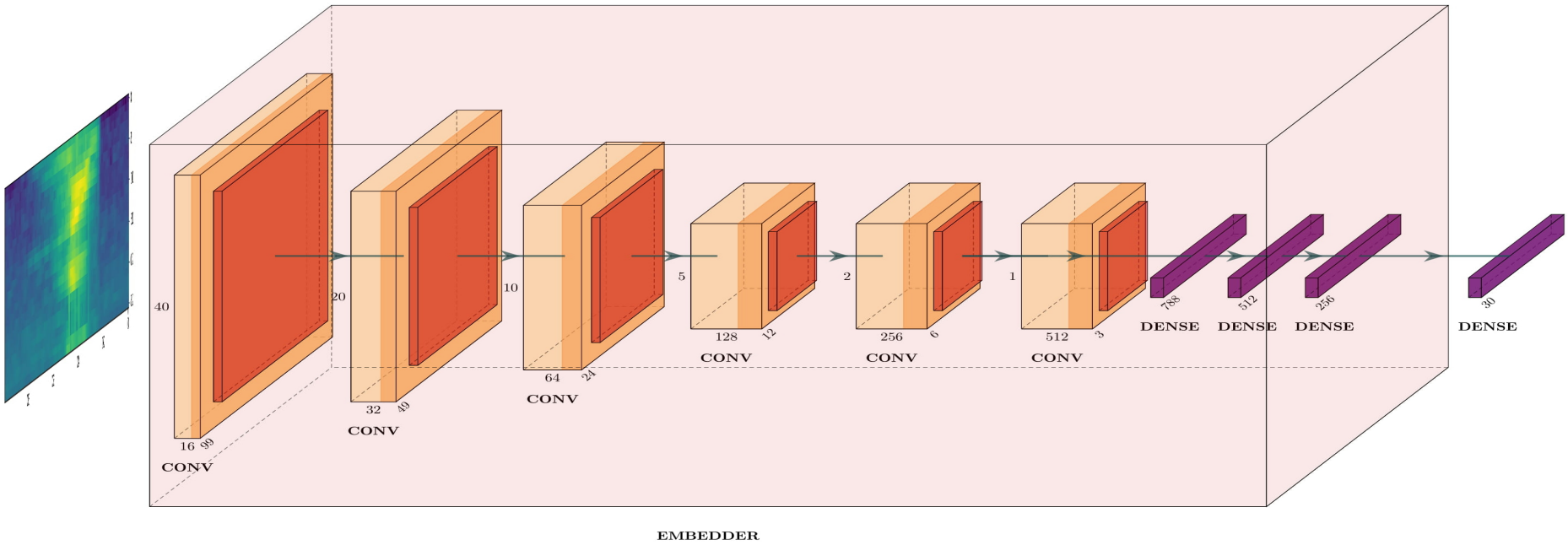
- The above is a word classifier for Speech Commands dataset.
- This model achieves an training accuracy of 96.59% and validation accuracy of 95.56%.
- We are having a deep CNN architecture with $\sim$ 1000k parameters and sample processing time of $\sim$ 1.75ms.

Feature Extractor



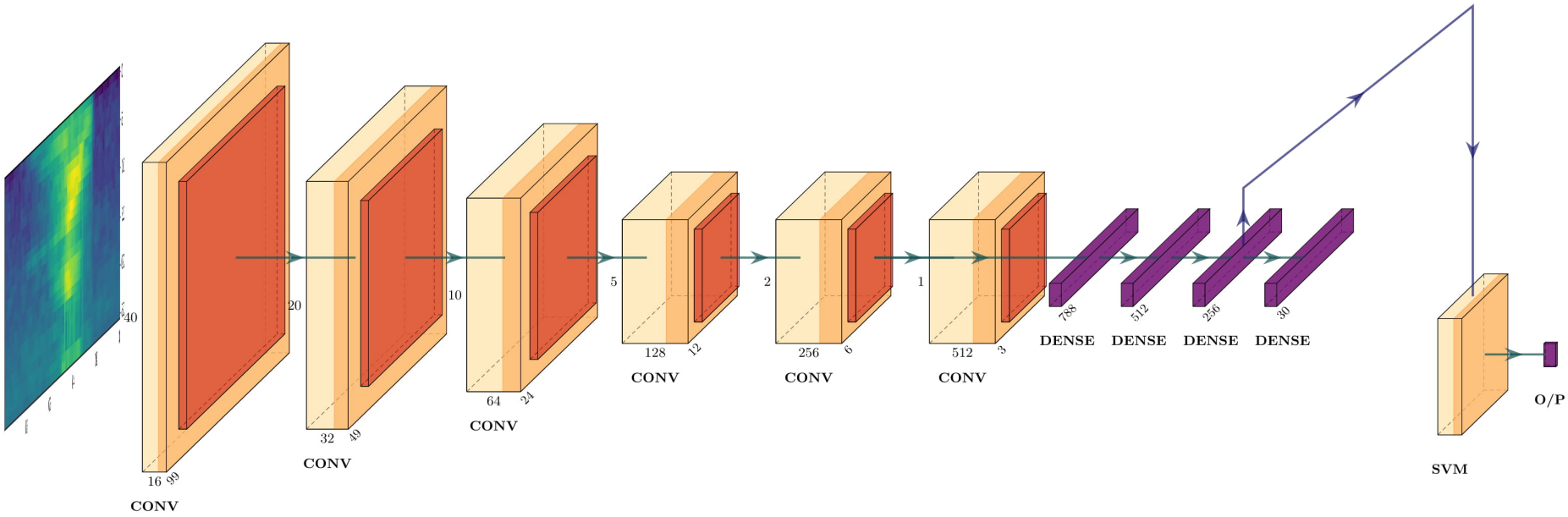
- We can interpret the deep network as a "black box" that transforms the input from one representation to another.

Feature Extractor



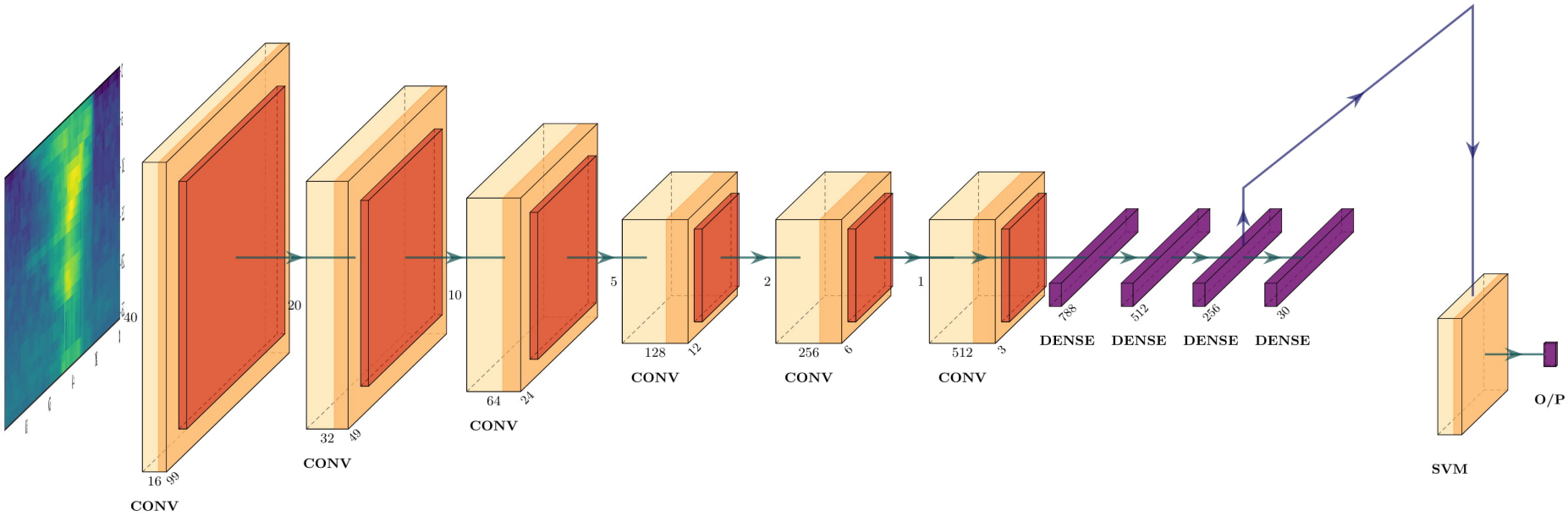
- We can interpret the deep network as a "black box" that transforms the input from one representation to another.
- Hence, the subnetwork in the box can be considered as an *embedder* that transforms the input.

KWS System



- We consider the output of the penultimate layer (256 dimension) as an *embedding* of the input feature.

KWS System



- We consider the output of the penultimate layer (256 dimension) as an *embedding* of the input feature.
- We then train an One Class SVM (OC-SVM), used popularly for outlier detection, with these embedding as input.

One Class SVM

- We train the OC-SVM using validation dataset of Google Speech Commands (not using the entire training dataset). We test the OC-SVM using test dataset.

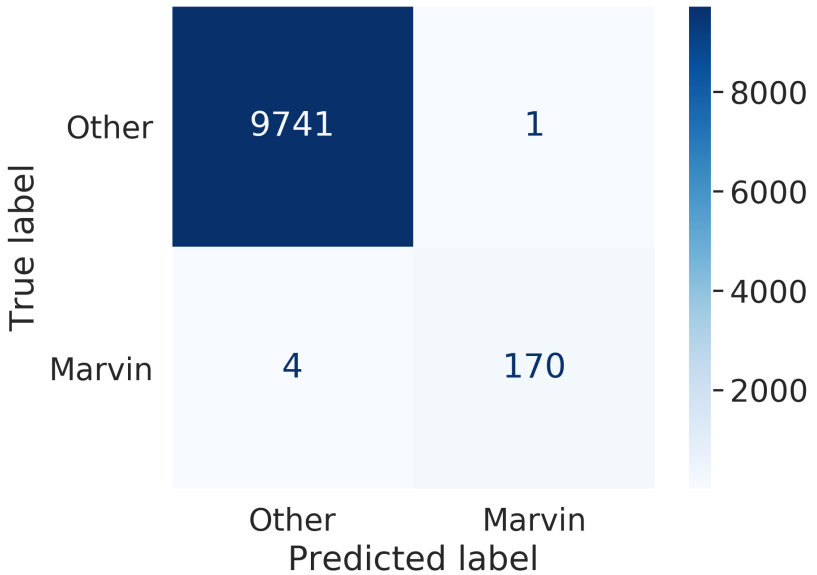
One Class SVM

- We train the OC-SVM using validation dataset of Google Speech Commands (not using the entire training dataset). We test the OC-SVM using test dataset.
- We select the hyperparameters of OC-SVM using tuning with *scikit-optimize* library.

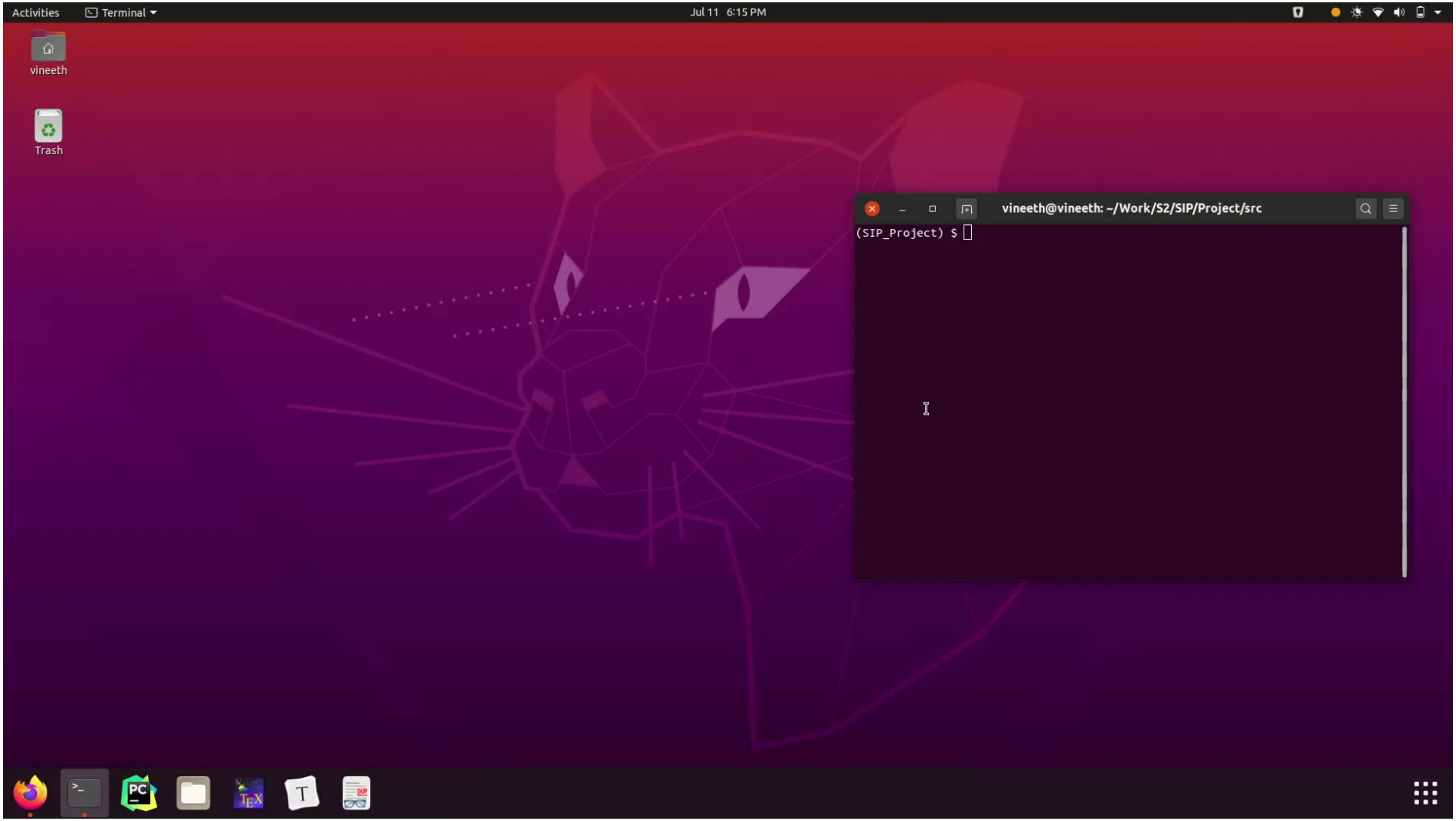
One Class SVM

- We train the OC-SVM using validation dataset of Google Speech Commands (not using the entire training dataset). We test the OC-SVM using test dataset.
- We select the hyperparameters of OC-SVM using tuning with *scikit-optimize* library.

Model size	**11.4MB**
Model size (Quantized)	978KB
Real Time Factor (RTF)	**1.7ms**
Accuracy	0.9995
Precision	0.9942
Recall (True Detection Rate)	0.9770
F1 Score	**0.9855**
Matthews Correlation Coefficient	**0.9853**
False Alarm Rate (FAR)	0.0001
False Alarm per Hour (FA/Hr)	**0.0003**
True Rejection Rates (TRR)	0.9998
False Rejection Rates (FRR)	0.0229



Demo


Activities
Terminal
Jul 11 6:15 PM
vineeth
Trash
vineeth@vineeth: ~/Work/S2/SIP/Project/src
(SIP_Project) $

Conclusions

- A small footprint reconfigurable CNN-OCSVM based KWS system is proposed in this work.

Conclusions

- A small footprint reconfigurable CNN-OCSVM based KWS system is proposed in this work.
- The raw performance numbers shows that this model outperforms many of the models in the literature.

Conclusions

- A small footprint reconfigurable CNN-OCSVM based KWS system is proposed in this work.
- The raw performance numbers shows that this model outperforms many of the models in the literature.
- However, a direct comparison is not meaningful because of the differences in the datasets and the actual keywords.

Conclusions

- A small footprint reconfigurable CNN-OCSVM based KWS system is proposed in this work.
- The raw performance numbers shows that this model outperforms many of the models in the literature.
- However, a direct comparison is not meaningful because of the differences in the datasets and the actual keywords.
- To prove this claim a lot of further work is necessary, such as performance analysis under noise and far-field conditions.

Bibliography

[1] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4087–4091.

[2] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *INTERSPEECH*, 2015.

[3] S. Tabibian, "A survey on structured discriminative spoken keyword spotting," *Artificial Intelligence Review*, 2019.

[4] J. Rownicka, P. Bell, and S. Renals, "Analyzing deep cnn-based utterance embeddings for acoustic model adaptation," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 235–241.

[5] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv e-prints*, arXiv:1804.03209, 2018.

The End

Questions? Suggestions?