End-to-End Framework for Key word Spotting

Students:

Ivancich Stefano

Masiero Luca

Supervisors:

Rossi Michele

Meneghello Francesca

DEPARTMENT OF INFORMATION ENGINEERING

UNIVERSITY OF PADOVA

22 September 2020

Outline

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

- The Problem
- The Solution
- What we tried
- Architecture 1: 1D CNN on RAW data
- Architecture 2: DSConv
- Architecture 3: Ensemble
- Performances comparison against other papers
- Conclusions

The Problem

Key word spotting:

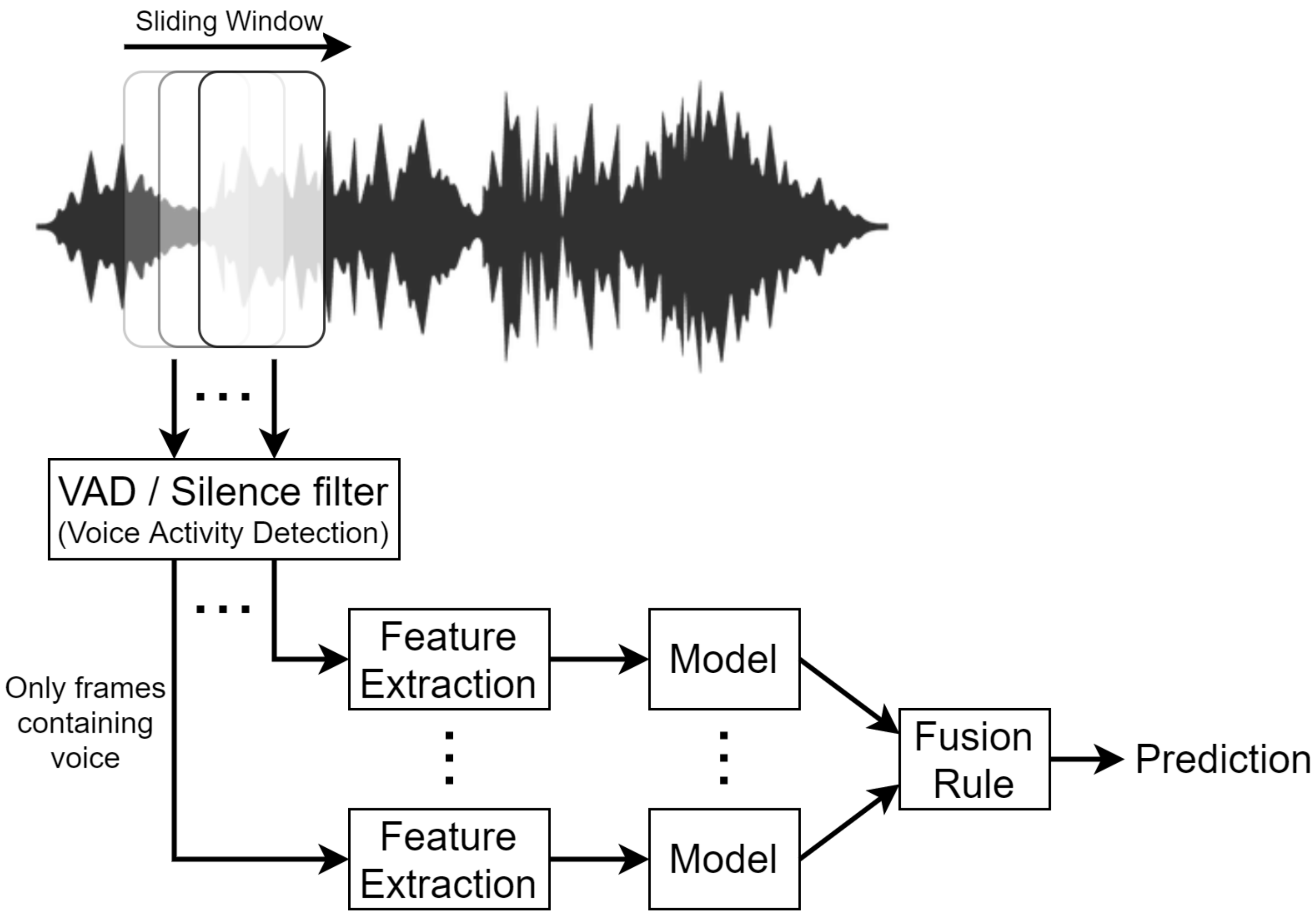
- Detect a relatively small set of predefined keywords (10 or 21) in a stream of user utterances.
- **Application**: Mobile phone, smart home device, consumer and robotics.
- **Constraints**: small footprint and fast (Real Time).

Metrics:

- Accuracy
- # of parameters
- Prediction speed (milliseconds)

The Solution

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



What we tried

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

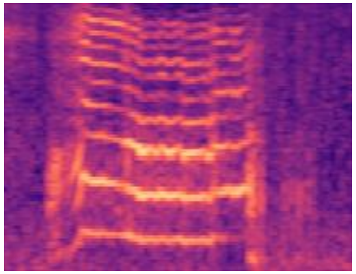
Datasets: (Google Speech Dataset V2)

- 10 commands (yes, no, up, down, left, right, on, off, stop, go)
- 21 commands (…, zero, one, two, three, four, five, six, seven, eight, nine, **unknown**)

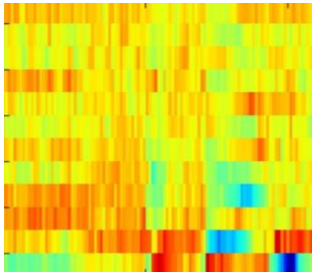
Data Preprocessing techniques:

- No preprocessing (Raw Waveform)
- 80 Mel spectrogram
- 40 MFCC
- 40 MFCC + 40delta + 40deltadelta (=120)

80 Mels



40 MFCC

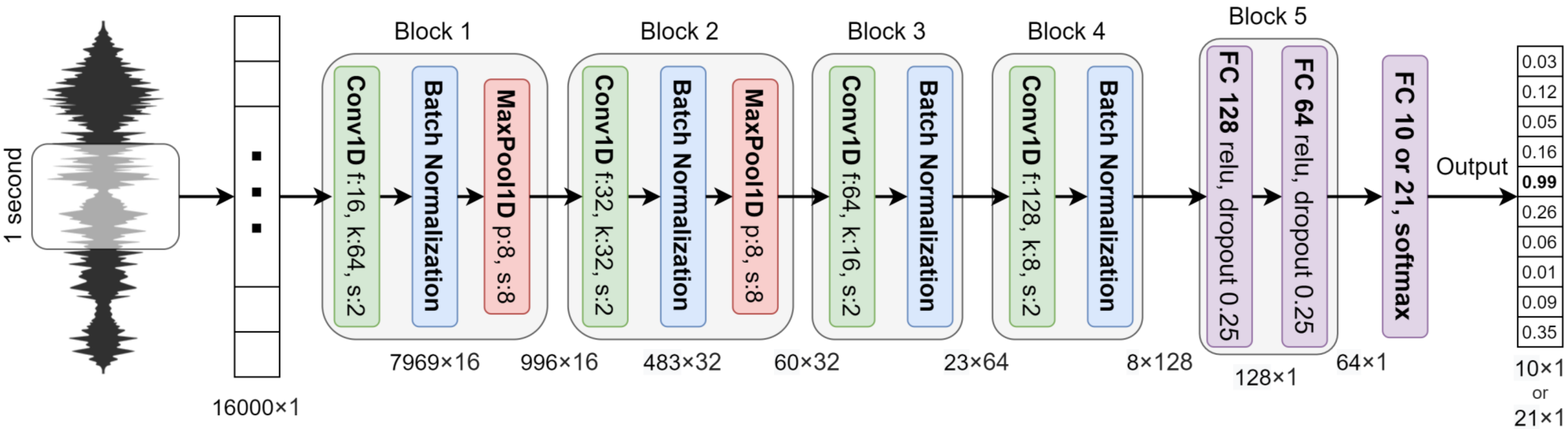


Learning Architectures:

- 1D CNN on RAW data
- DSConv (Small - Medium - Large)
- Ensemble (Small - Medium - Large)

1D CNN on RAW data

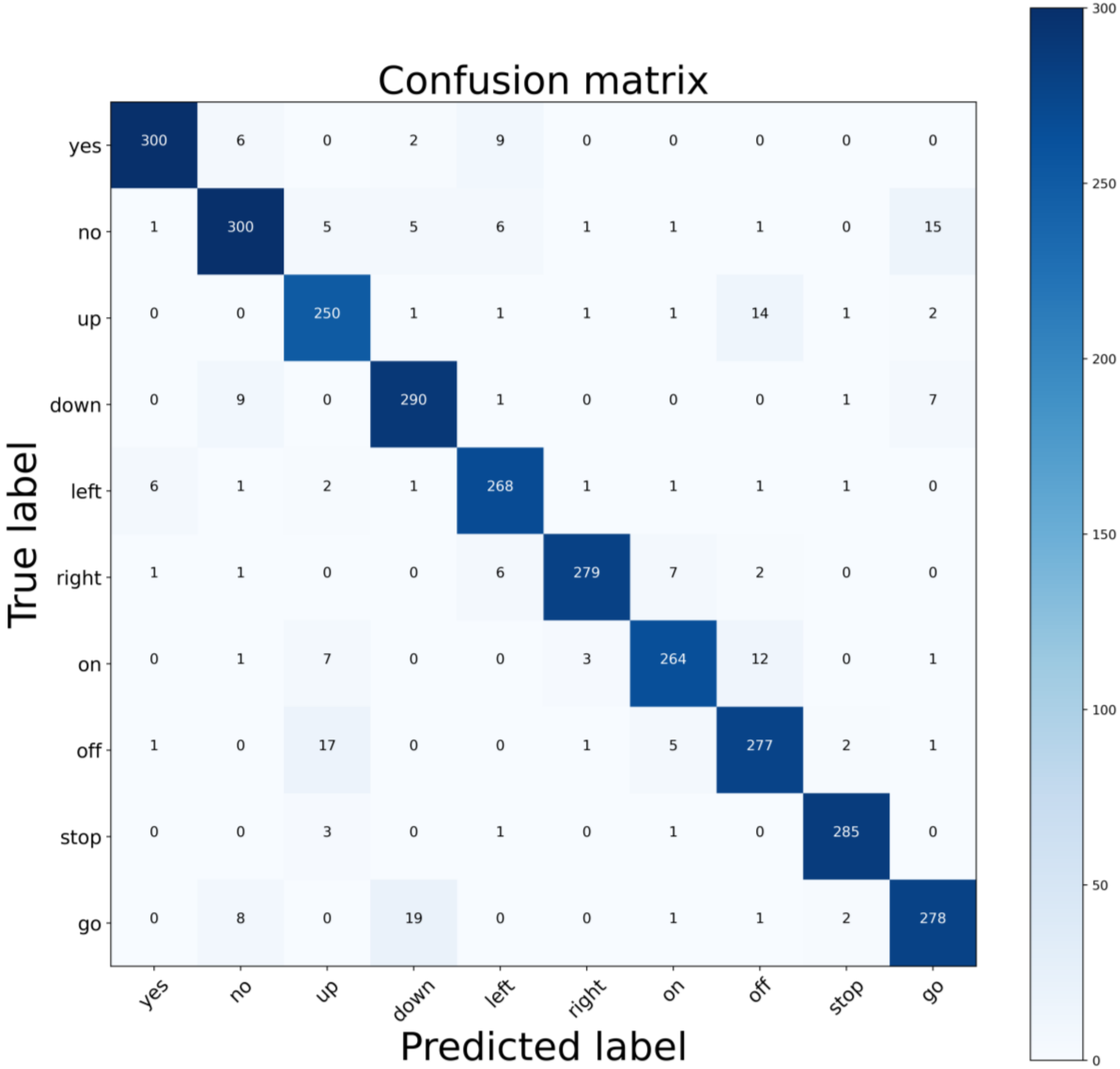
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



	10 commands (30k - 3k - 3k)	21 commands (84k - 9k - 11k)
Accuracy	93.0	89.1
# parameters	257,018	257,733
Speed (ms)	28.71	28.25

1D CNN on RAW data

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

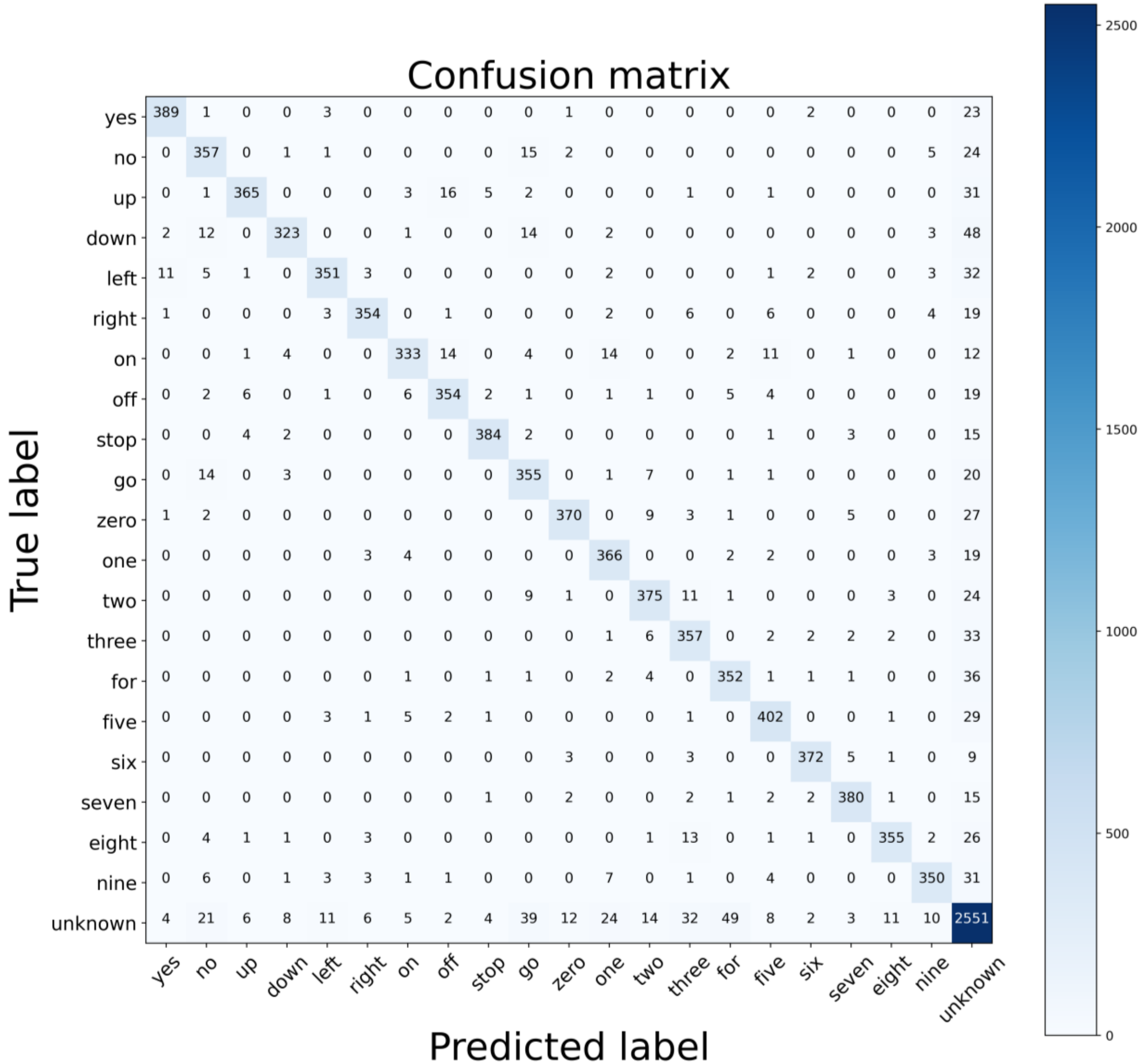


Top Mistakes:

- 31 Up-Off
- 26 Down-Go
- 23 Go-No

1D CNN on RAW data

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



Top Mistakes:

- 85 Four-Unknown
- 65 Three-Unknown

Actually it confuses

- Four - For
- Three - Tree

Separable CONV

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Separable convolution: uses less parameters, less memory and less computations than regular convolutional layers, and perform better.

There are two types of Separable convolution: Spatial and depthwise

Spatial separable convolution: divides a kernel into two, smaller kernels.

E.g. divide a 3x3 kernel into a 3x1 and 1x3 kernel. So, instead of doing one convolution with 9 multiplications, we do two convolutions with 3 multiplications each (6 in total) to achieve the same effect.

Problem: not all kernels can be "separated" (mathematically) into two.

Depthwise Separable convolution: uses kernels that cannot be "factored" into two smaller kernels. It deals not just with the spatial dimensions, but also with the depth dimension (number of channels).

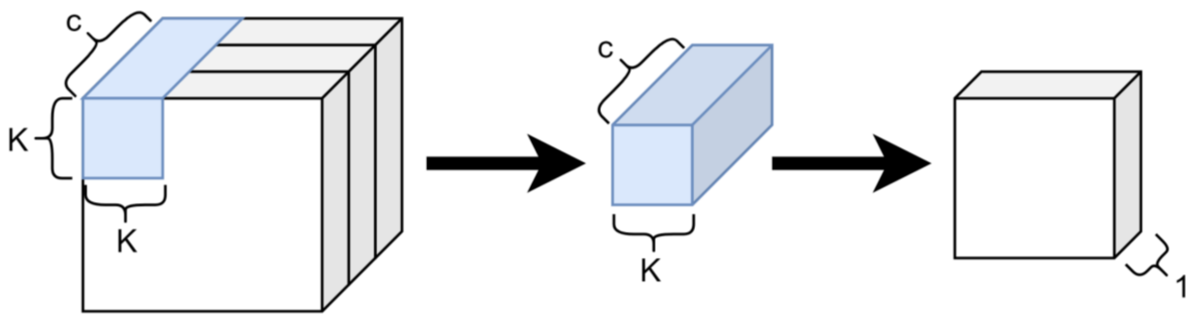
Splits a kernel into 2 separate kernels that do two convolutions:

- the depthwise convolution
- the pointwise (1x1) convolution.

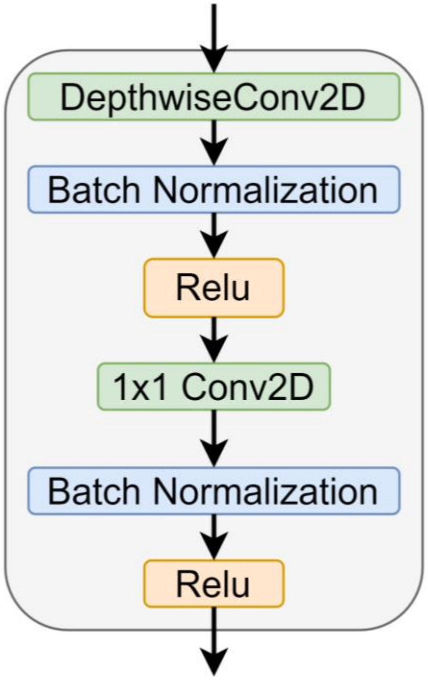
Depth-wise Sep. CONV

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

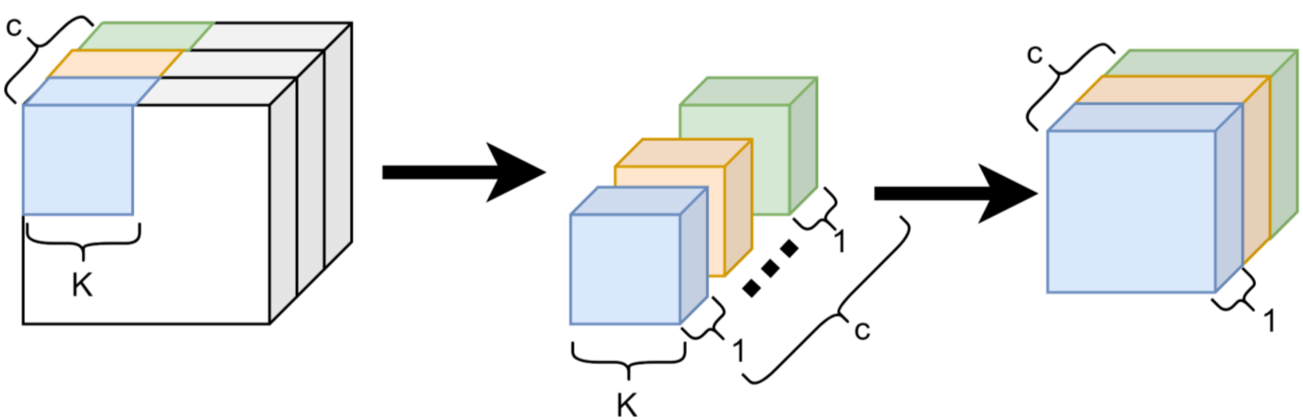
1 Filter Normal Convolution



Separaple Depthwise Convolution

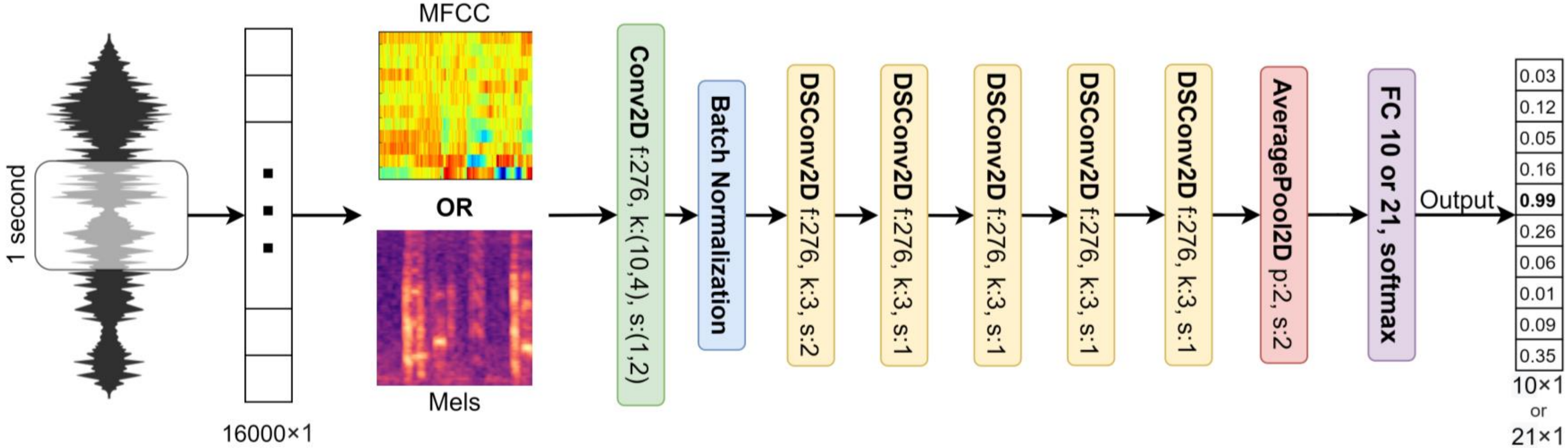


DepthWise Convolution = 1 filter KxKx1 for each channel



DSConv Model - Large

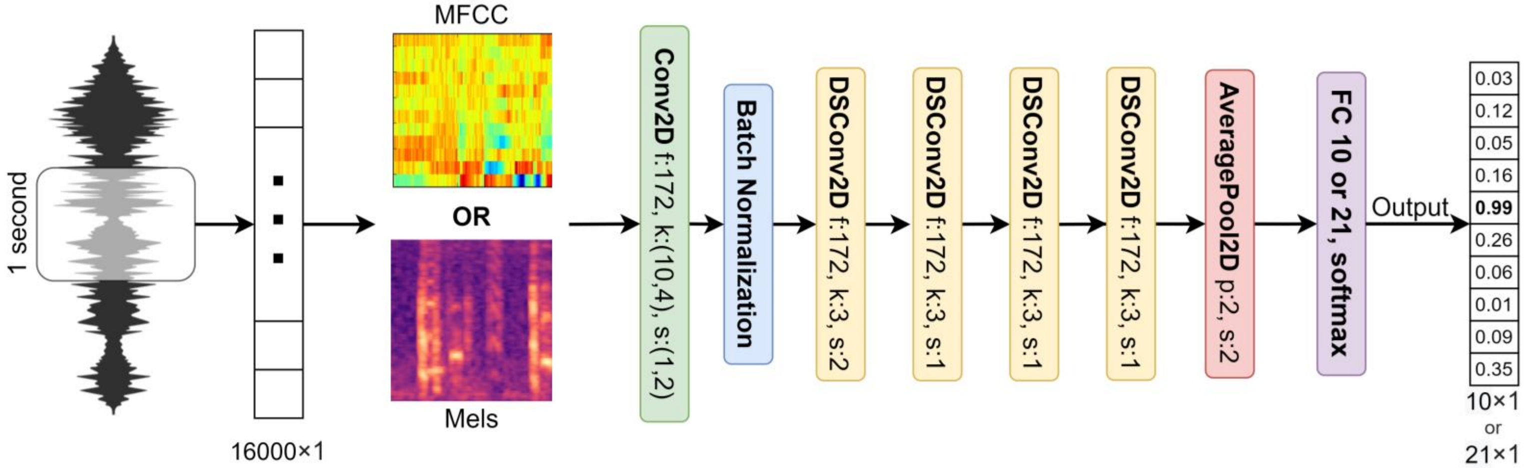
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



	10 commands (30k - 3k - 3k)			21 commands (84k - 9k - 11k)		
	80 Mels	40 MFCC	40 + Δs MFCC	80 Mels	40 MFCC	40 + Δs MFCC
Accuracy	**96.0**	95.3	95.5	93.4	**93.7**	92.7
# params	874,930	**571,330**	1,178,530	1,375,881	**738,321**	2,013,441
Speed (ms)	**33.39**	30.24	33.42	33.79	**30.87**	33.13
With FE	**41.44**	45.32	44.25	41.62	**39.66**	44.62

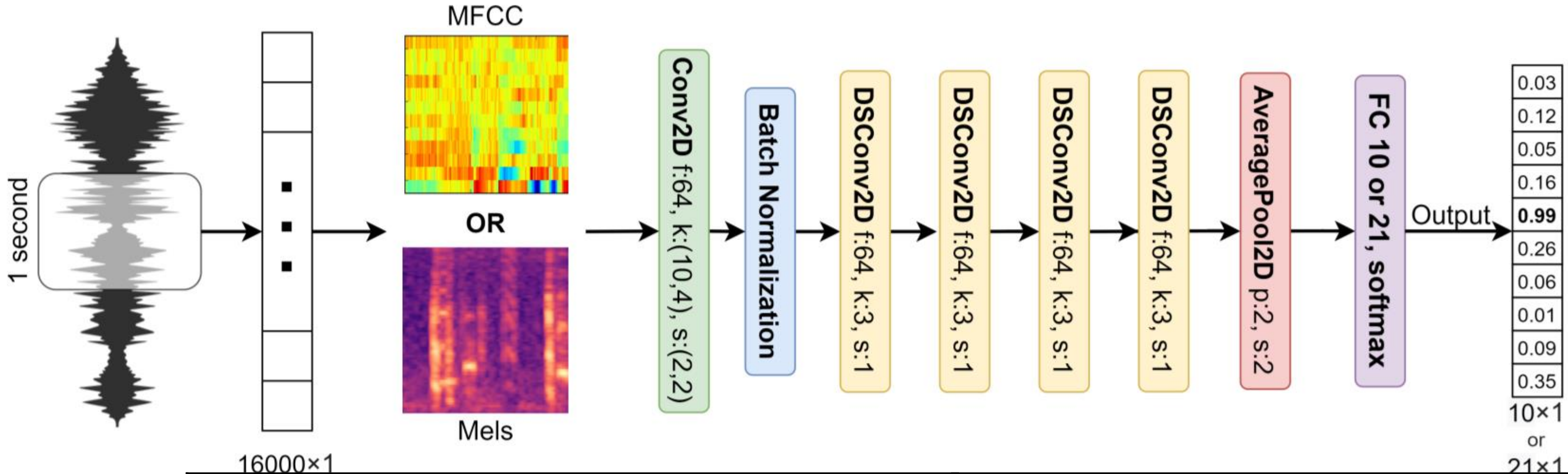
DSConv Model - Medium

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



	10 commands (30k - 3k - 3k)			**21 commands** (84k - 9k - 11k)		
	80 Mels	40 MFCC	40 + Δs MFCC	80 Mels	40 MFCC	40 + Δs MFCC
Accuracy	94.3	**95.0**	94.8	**92.7**	92.2	91.7
# params	469,398	**262,998**	675,798	832,673	**399,233**	1,266,113
Speed (ms) **With FE**	30.75 38.49	**29.55** **38.23**	30.76 41.82	**32.25** **39.72**	30.01 46.76	33.03 47.08

DSConv Model - Small



	10 commands (30k - 3k - 3k)			21 commands (84k - 9k - 11k)		
	80 Mels	40 MFCC	40 + Δs MFCC	80 Mels	40 MFCC	40 + Δs MFCC
Accuracy	92.5	**92.9**	92.5	**90.0**	89.2	86.5
# params	300,618	**127,818**	473,418	604,757	**241,877**	967,637
Speed (ms) **With FE**	31.00 38.31	**29.27** **38.23**	30.54 40.23	**32.86** **37.97**	29.48 48.15	30.31 40.99

Ensemble 10-cmd

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Ensemble between the best models.

	10 commands (30k - 3k - 3k)		
	Large	**Medium**	**Small**
	DSConv L 80 Mels + DSConv L 40 MFCC + DSConv L 40 MFCC Δ	DSConv L 40 MFCC + DSConv M 80 Mels + DSConv M 40 MFCC Δ	DSConv M 40 MFCC + DSConv S 80 Mels + DSConv S 40 MFCC
Accuracy	96.8	96.4	95.6
# params	2,624,790	1,303,726	691,463
Speed (ms)	131.01	122.04	106.09

Ensemble 21-cmd

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Ensemble between the best models.

	21 commands (84k - 9k - 11k)		
	Large	**Medium**	**Small**
	DSConv L 80 Mels + DSConv L 40 MFCC + DSConv L 40 MFCC Δ	DSConv L 40 MFCC + DSConv M 80 Mels + DSConv M 40 MFCC	DSConv L 40 MFCC + DSConv M 40 MFCC + DSConv S 40 MFCC
Accuracy	95.2	95.0	94.2
# params	2,498,019	1,970,227	1,379,431
Speed (ms)	125.9	114.21	109.03

Performances 10-cmd

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

	Accuracy %	# Parameters	Speed (ms)
SincConv [Mittermaier et al.]	97.4	162K	40.35
Our Ensemble Large	**96.8**	2,624,790	131.01
Our Ensemble Medium	96.4	1,303,726	122.04
Our DSConvLarge + 80Mels	96.0	874,930	41,44
Our Ensemble Small	95.6	691,463	106.09
Our DSConvMedium + 40MFCC	95.0	262,998	38,23
Our 1D CNN on Raw data	93,0	257,018	**28,71**
Our DSConvSmall + 40MFCC	92,9	**127,818**	38,23

- **Best**: Ensemble Large
- **Smaller**: DSConvSmall + 40MFCC
- **Fastest**: 1D CNN on raw data

Performances 21-cmd

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

	Accuracy %	# Parameters	Speed (ms)
SincConv [Mittermaier et al.]	97.4	162K	40.35
Our Ensemble Large	**95.2**	2,498,019	125.9
Our Ensemble Medium	95.0	1,970,227	114.21
Our Ensemble Small	94.2	1,379,431	109.03
Our DSConvLarge + 40MFCC	93.7	738,321	39.66
Our DSConvMedium + 80Mels	92,7	832,673	39,72
Our DSConvSmall + 80Mels	90,0	604,757	37,97
Our 1D CNN on Raw data	89,1	**257,733**	**28,25**

- **Best**: Ensemble Large
- **Smaller**: 1D CNN on raw data
- **Fastest**: 1D CNN on raw data

Conclusions

Conclusions:

- The tests showed that our model was **very good at classify keywords**.
- Unfortunately we didn't beat the state-of-the-art models.
- We found that the number of convolutional layers played a key role in detecting high-level concepts
- There is not so much difference between using 80 Mels or 40 MFCCs.
- We presented different model sizes in order to fit different devices.

Future Work:

- Try different hyper-parameters during training
- Change the **structure** of the **network** using:
    - SincConv
    - GDSConv
- Build a Feature extractor with convolutional autoencoder

Thank You

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Thanks for your attention!

Any questions?