progress

2020029152 김나현

Small-Footprint Keyword Spotting on Raw Audio Data with Sinc-Convolutions

Keyword Spotting (KWS)

"Hey Siri!"

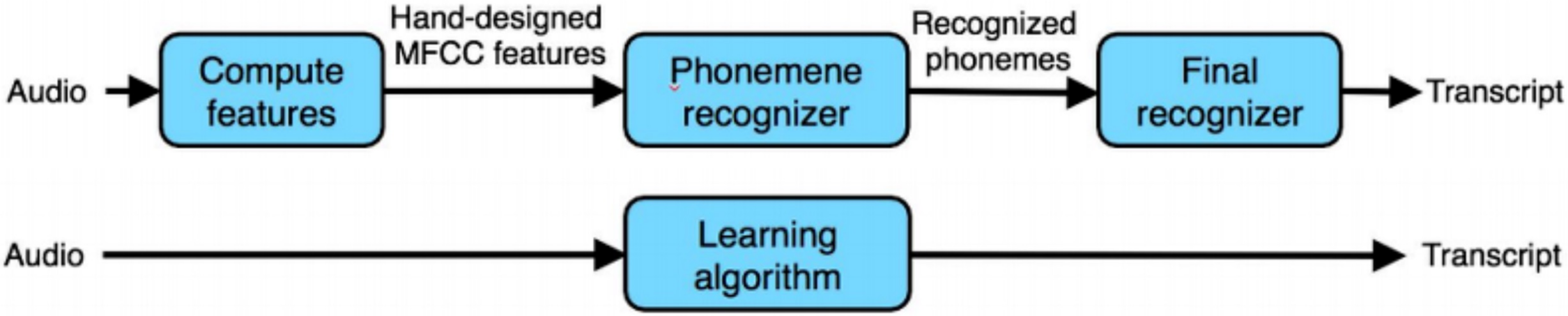
wake words

Trigger

ASR

Conventional hybrid approaches to KWS

1. divide audio signal in time frames to extract features (MFCC)
2. A neural net then estimates phoneme or state posteriors of the keyword Hidden Markov Model in order to calculate the keyword probability
3. The wake-word is then recognized when the keyword probability reaches a predefined threshold



Previous architectures

extract acoustic features
apply a neural network to
classify keyword probabilities

end-to-end architectures

extracts spectral features
using parametrized
Sinc-convolutions

Power-consuming audio preprocessing and data transfer steps are eliminate

to reduce power and memory consumption
without reducing classification accuracy.

contributions

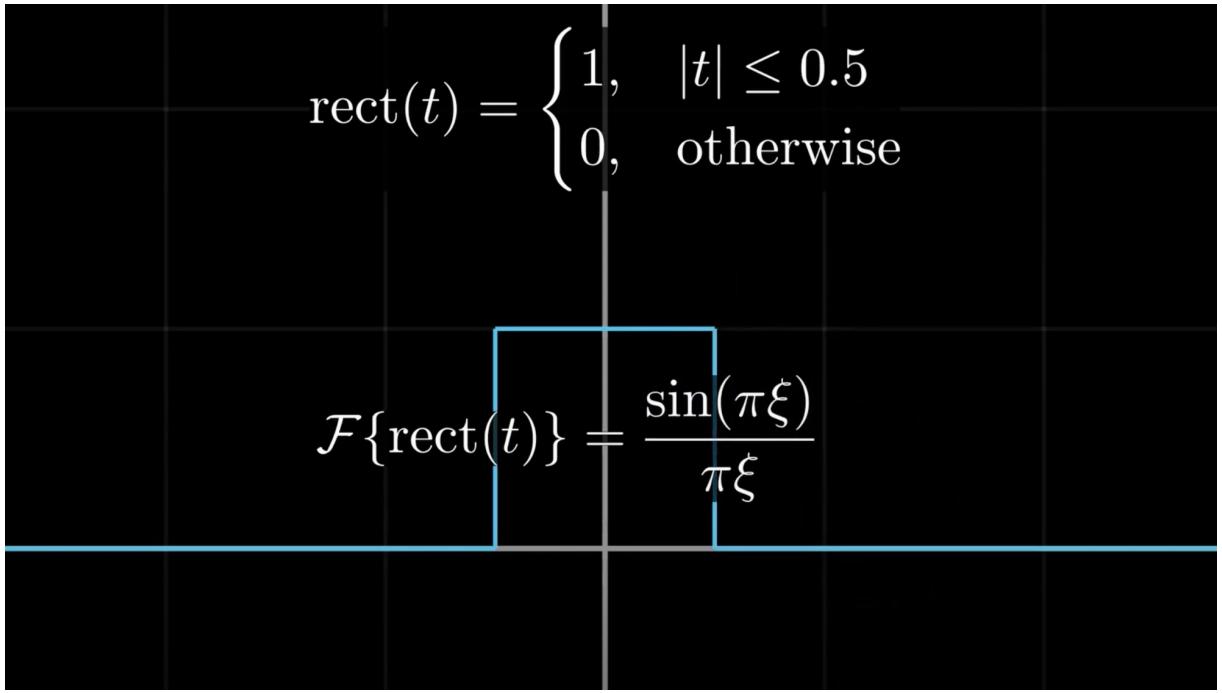
1. propose a neural network architecture tuned towards energy efficiency in microcontrollers

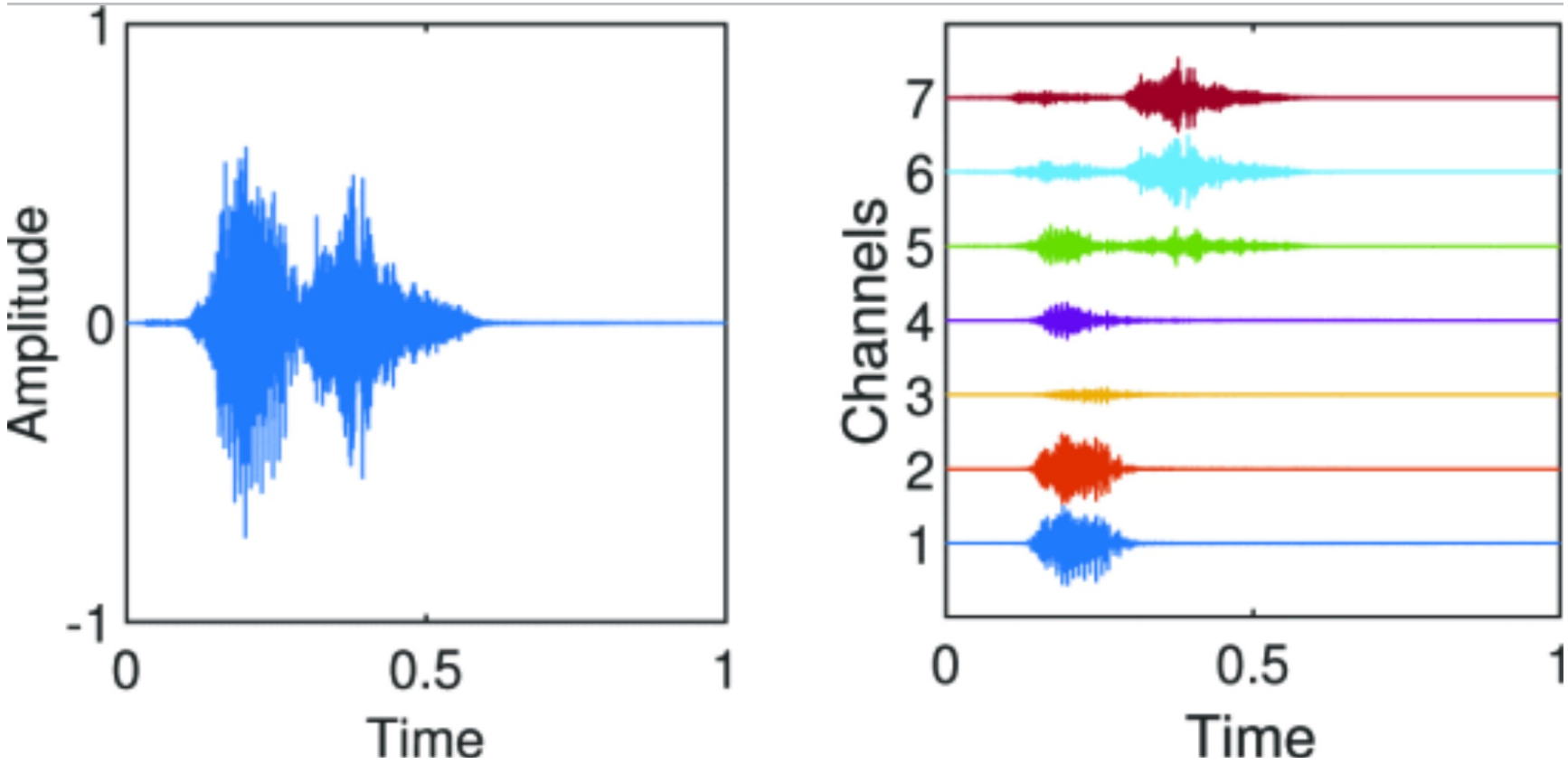
2 . Classify on raw audio employing Sinc-Convs
while reducing the number of parameters using (G)DS-Convs.

3. achieve the competitive accuracy of 96.4%
on Google's Speech Commands test set with only 62k parameters.

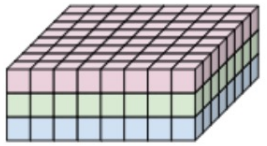
Feature Extraction using SincConvs

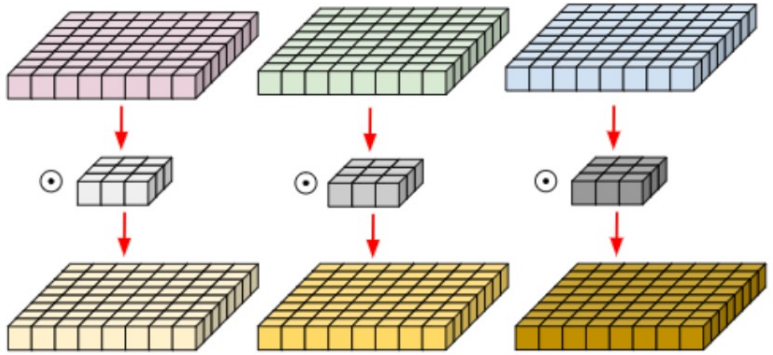
restricting the filters of the first convolutional layer of a CNN to only learn parametrized sinc functions

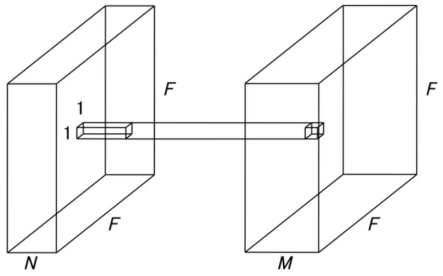



Amplitude
1
0
-1
0
0.5
1
Time
Channels
7
6
5
4
3
2
1
0
0.5
1
Time

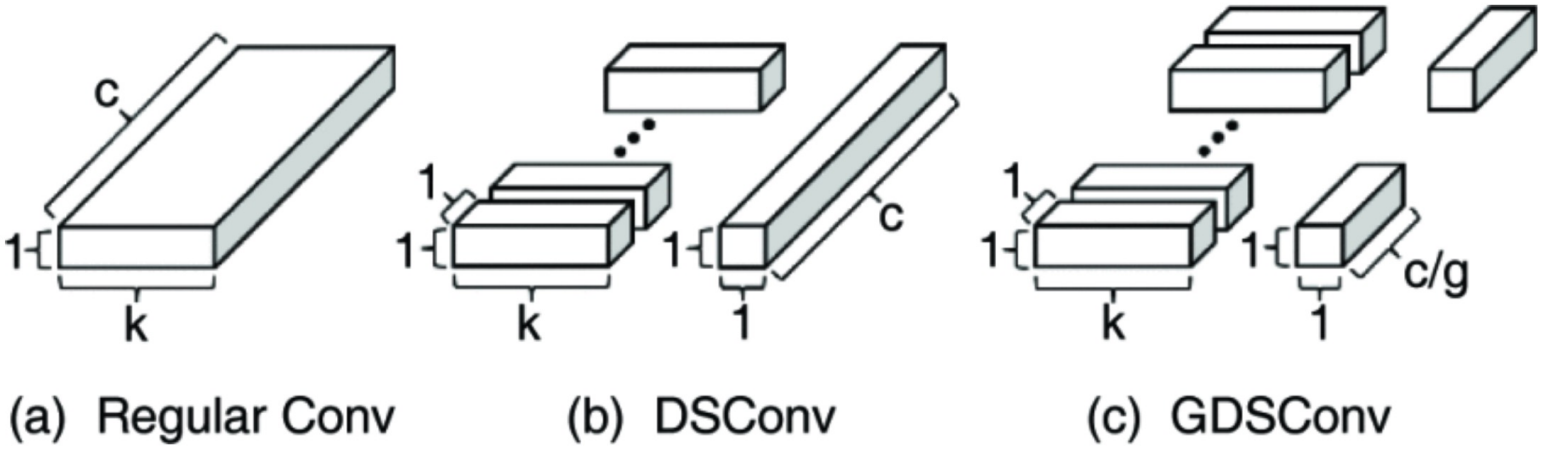
Low-Parameter GDS-Conv Layers



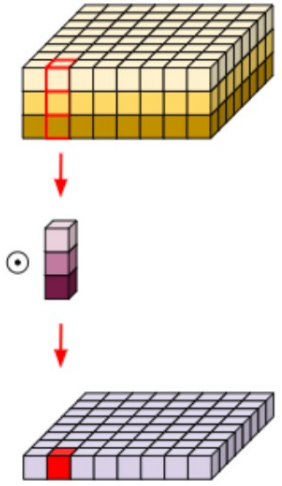




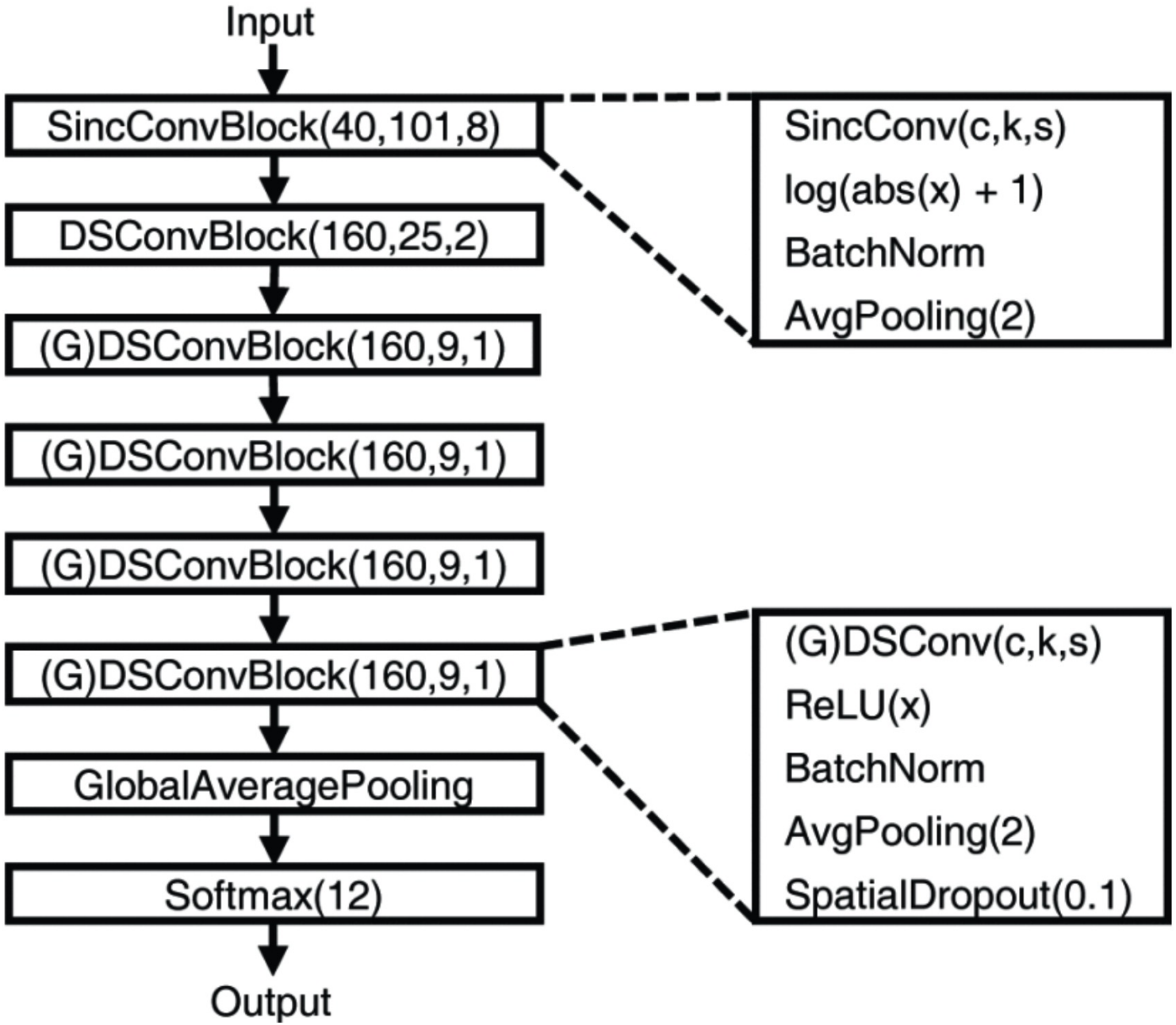
Pointwise Convolution



(a) Regular Conv (b) DSConv (c) GDSConv



Depthwise Separable Convolution



the base model has 122k parameters

After grouping,
the number of parameters
is reduced to a total of 62k

EVALUATION

- **Training on the Speech Commands Dataset**

Model	Accuracy	Parameters
DS-CNN-S [2]	94.1%	39k
DS-CNN-M [2]	94.9%	189k
DS-CNN-L [2]	95.4%	498k
ResNet15 [3]	95.8%	240k
TC-ResNet8 [4]	96.1%	66k
TC-ResNet14 [4]	96.2%	137k
TC-ResNet14-1.5 [4]	**96.6%**	305k
SincConv+DSConv	**96.6%**	122k
SincConv+GDSConv	96.4%	**62k**