MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

Speech Technology 2022

Lecture #6
Keyword Spotting. Part 2

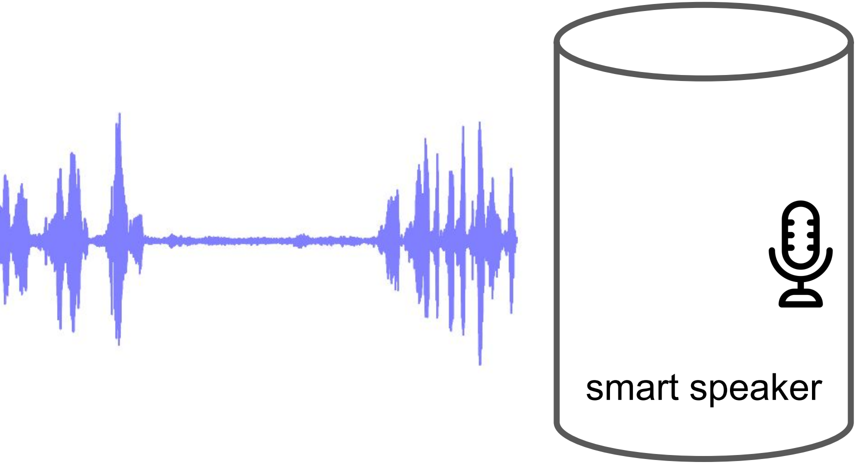
@georgygospodinov

Plan

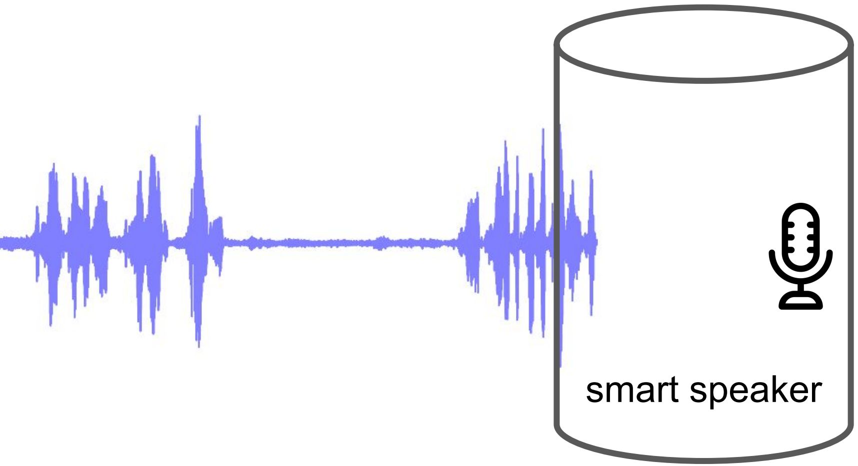
- Keyword Spotting
  - Transfer Learning
  - Multitask Learning
  - Model Compression
  - Cascade Systems
  - Streaming KWS

Keyword Verification

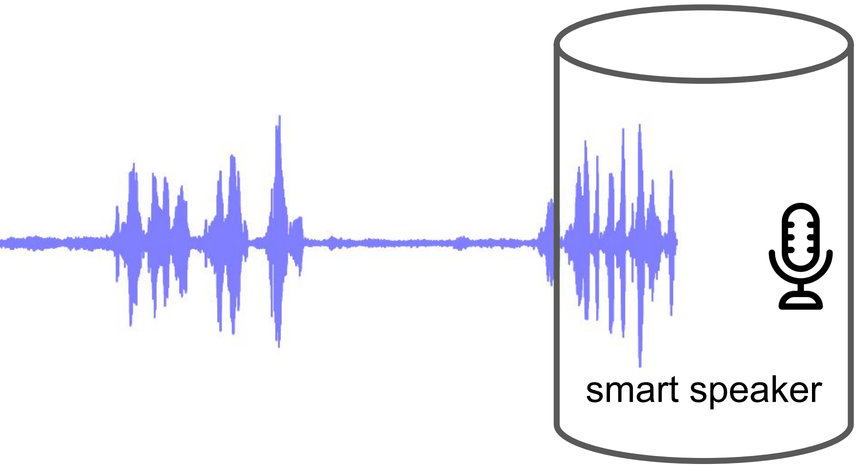
Keyword Spotting



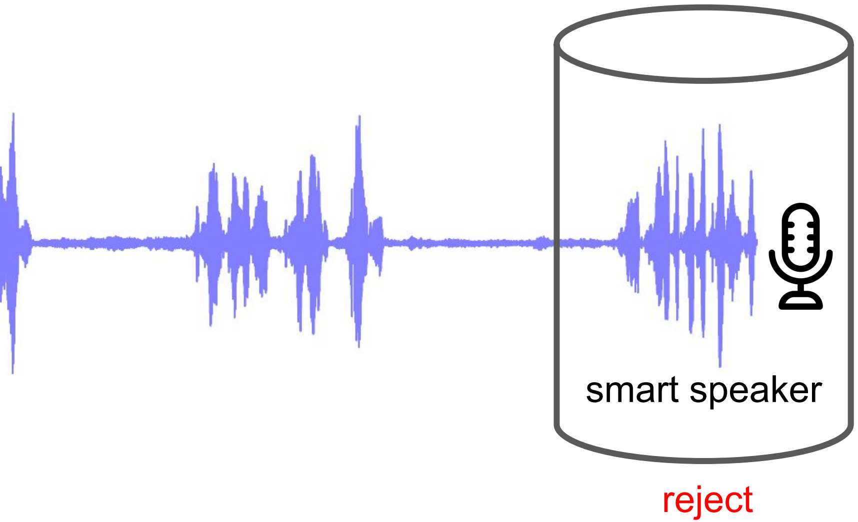
Keyword Spotting



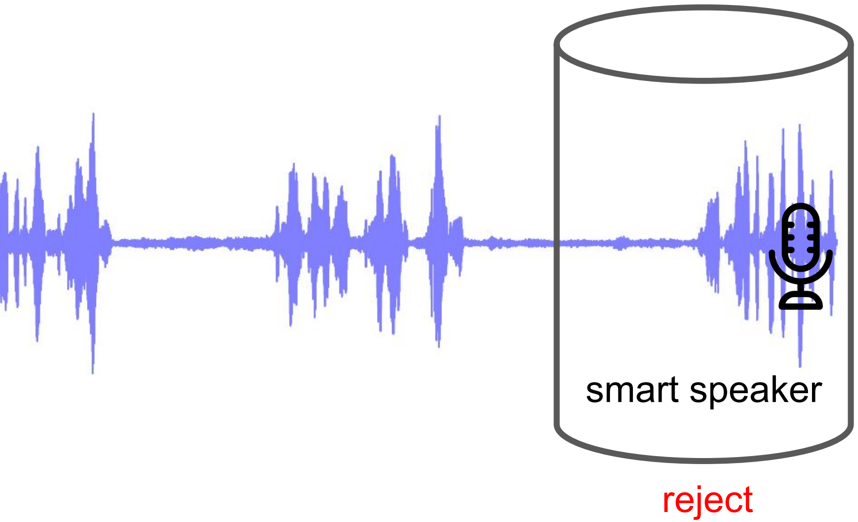
Keyword Spotting



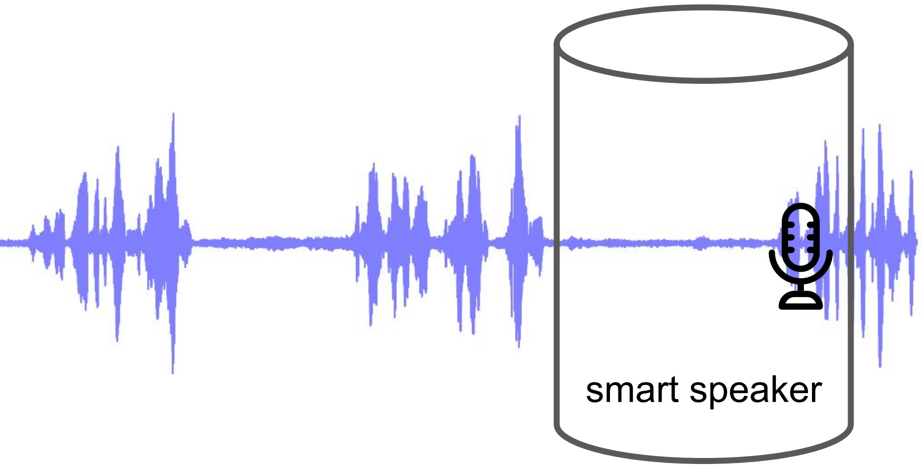
Keyword Spotting



Keyword Spotting

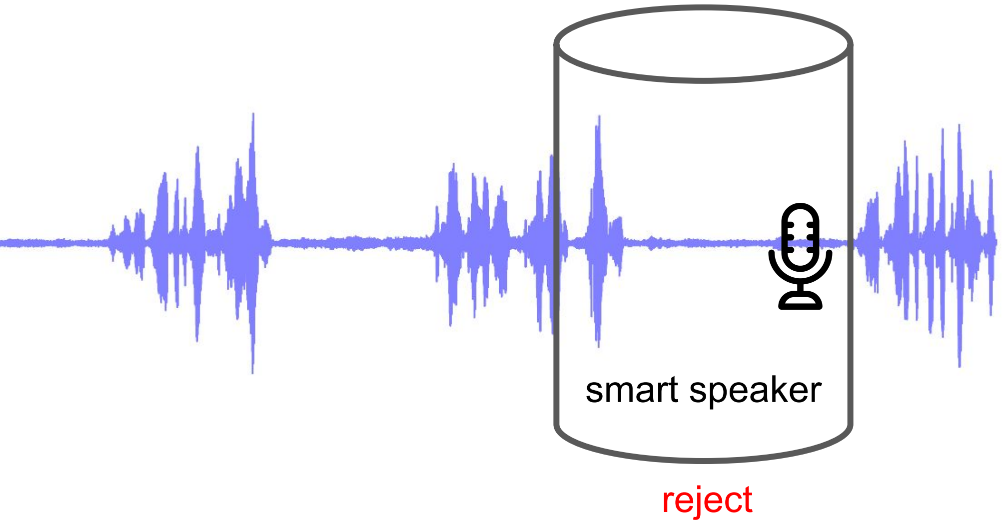


Keyword Spotting

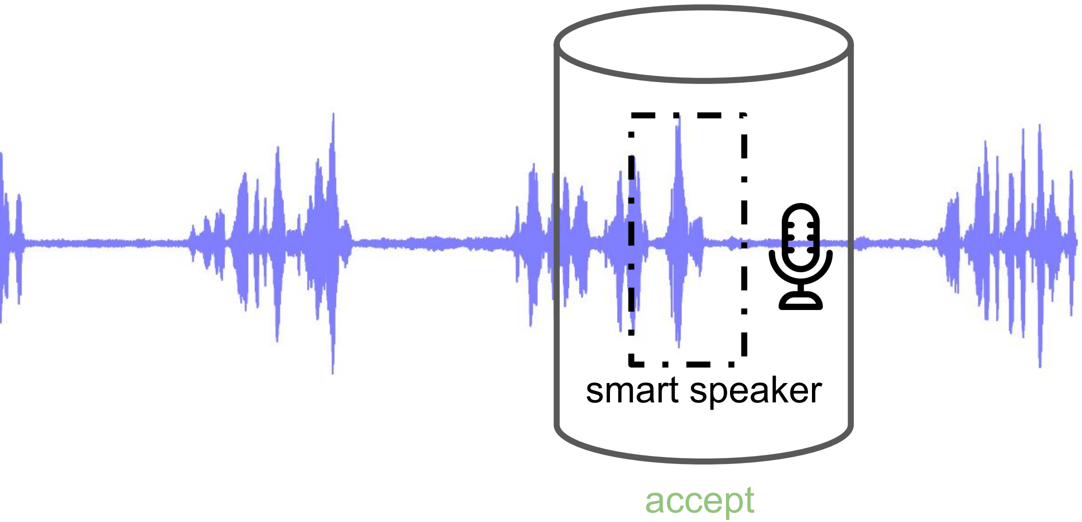


reject

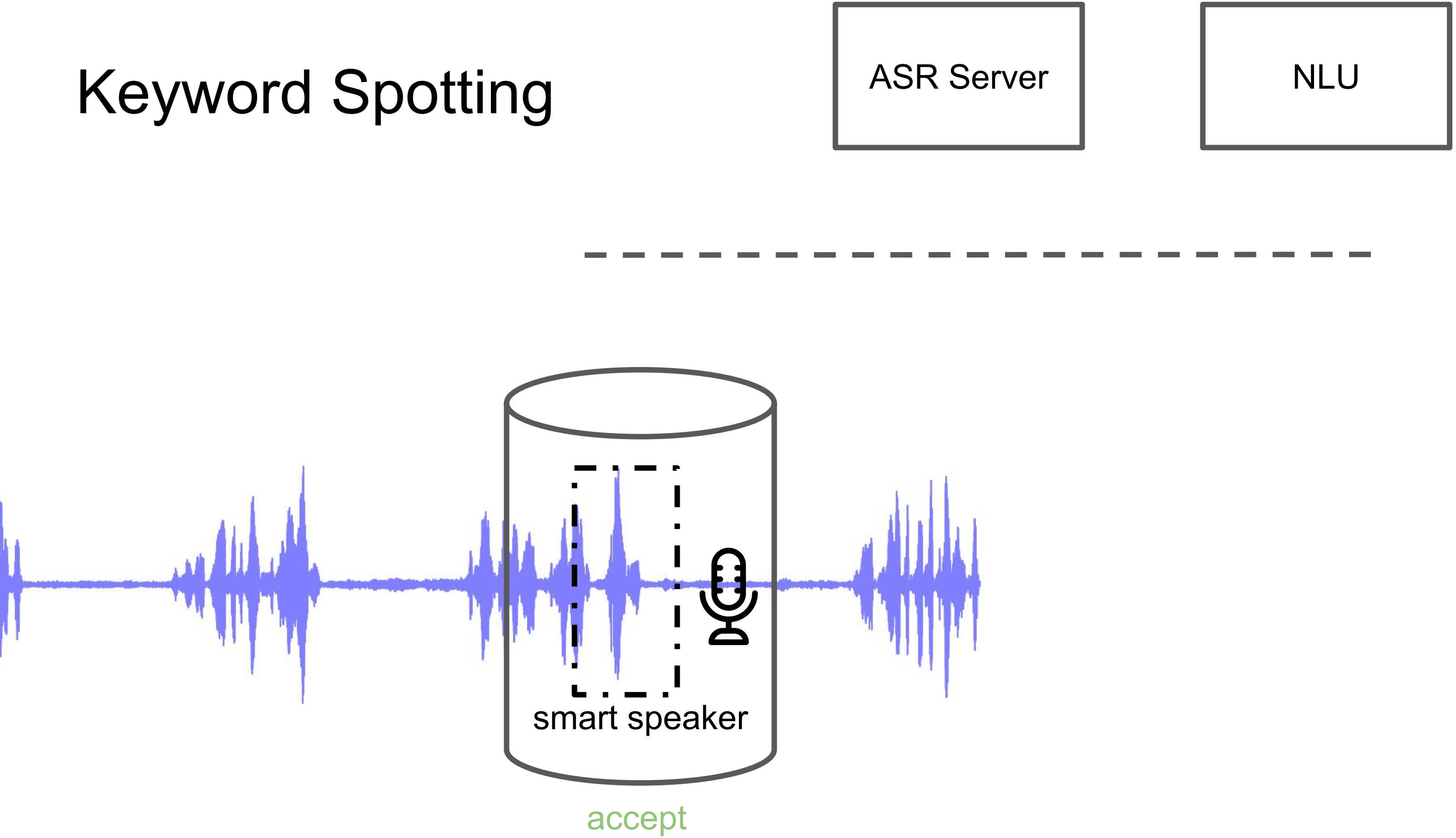
Keyword Spotting



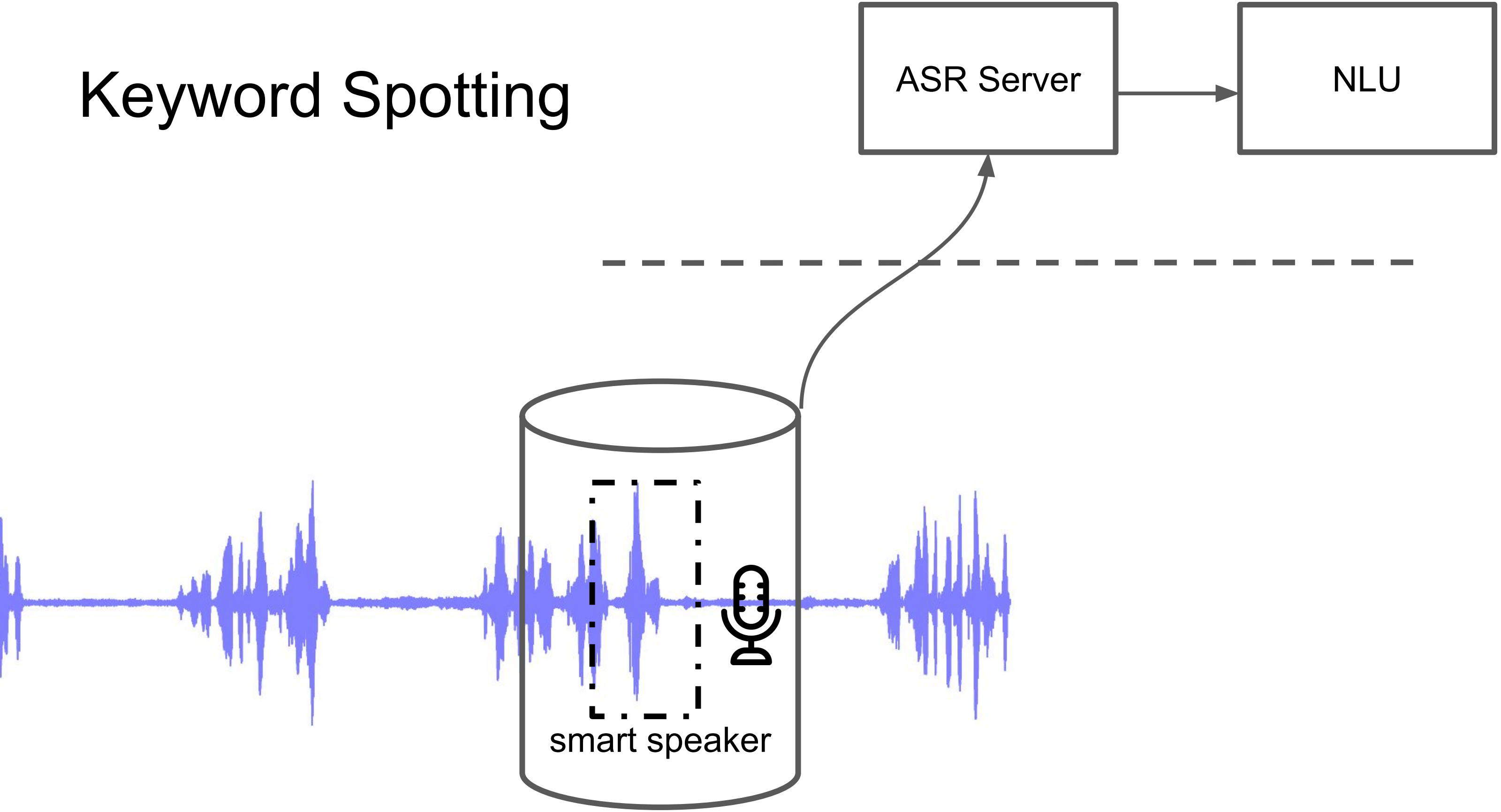
Keyword Spotting



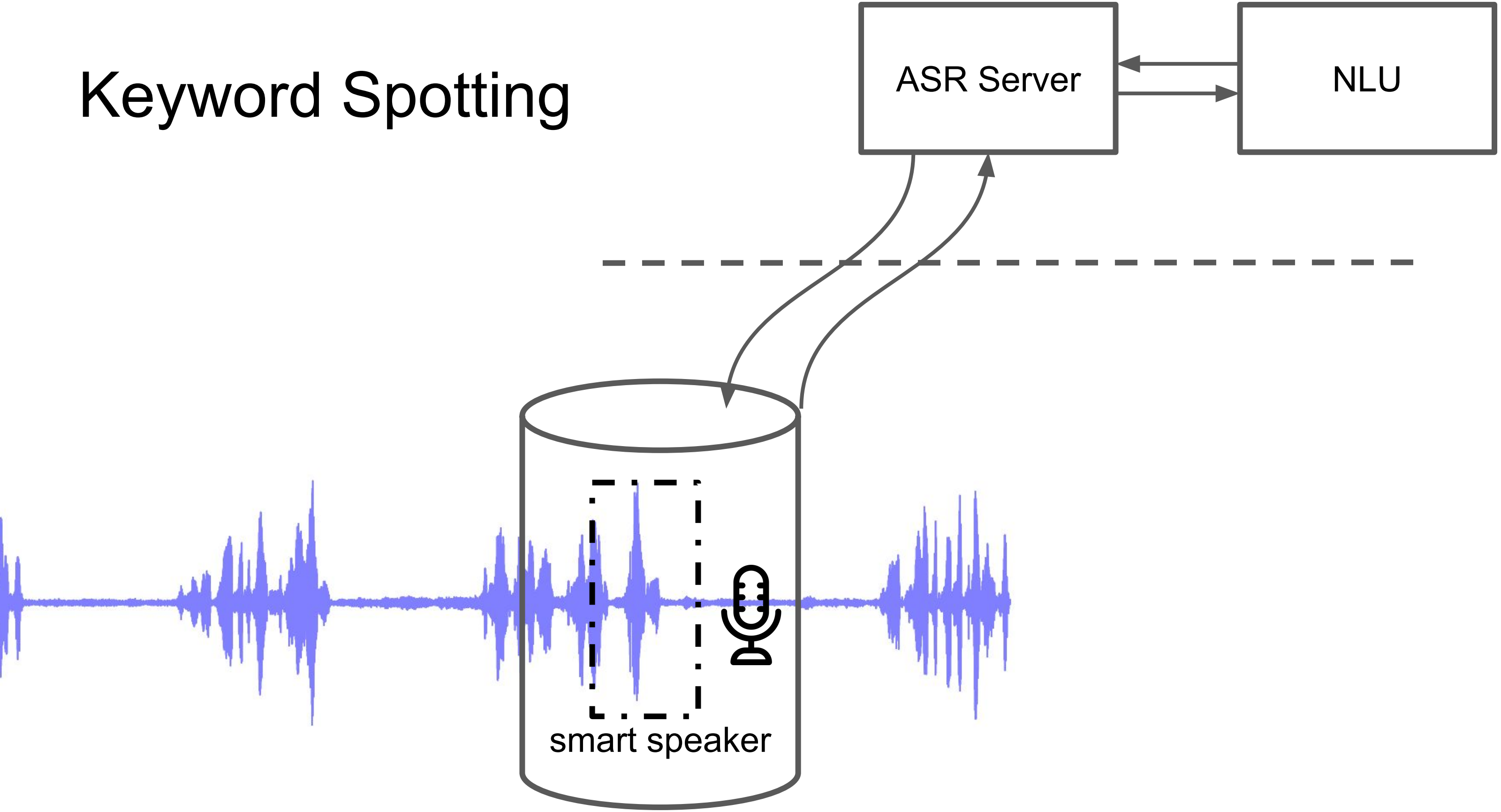
Keyword Spotting



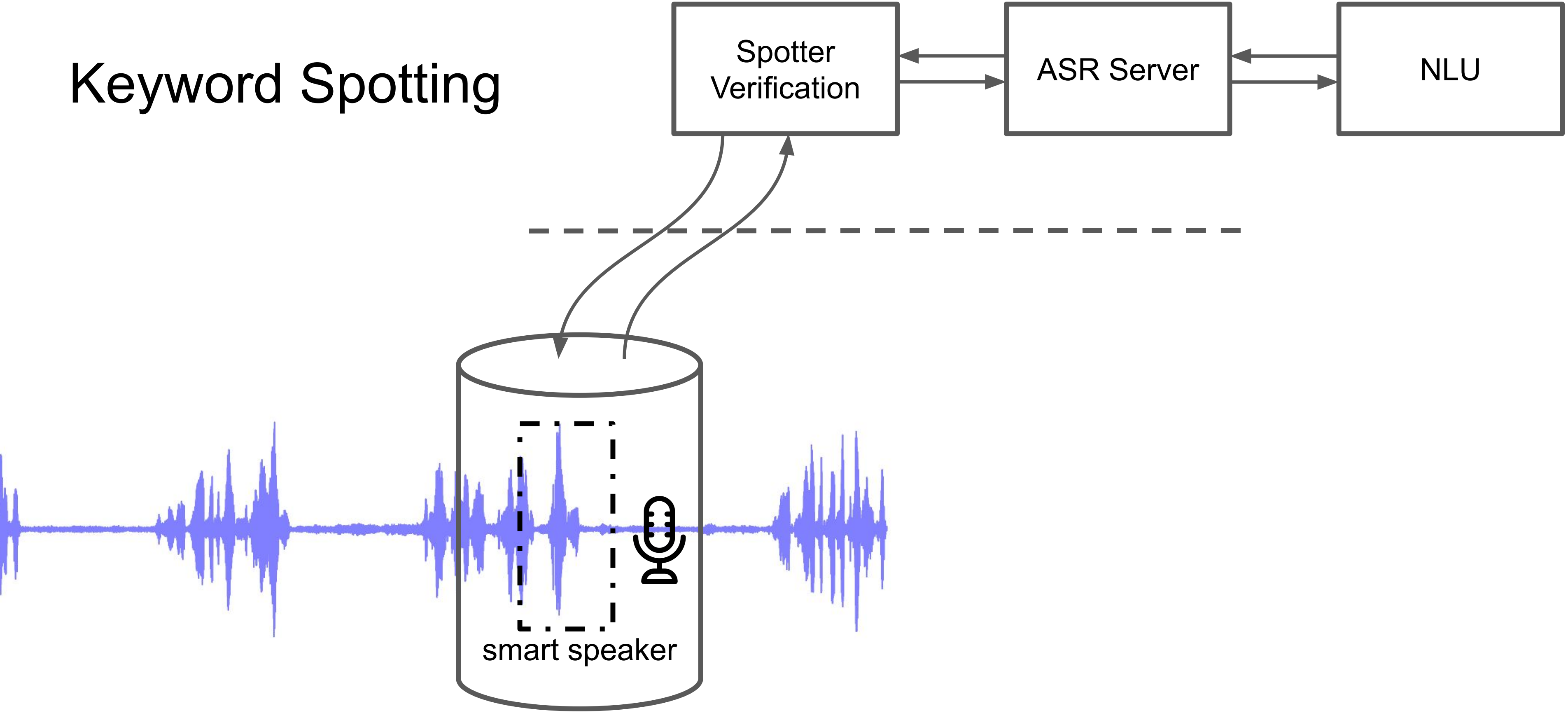
Keyword Spotting



Keyword Spotting

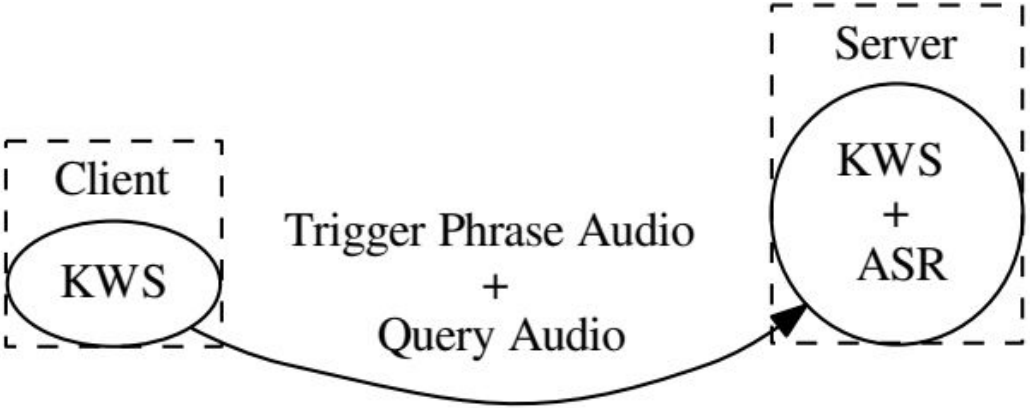


Keyword Spotting



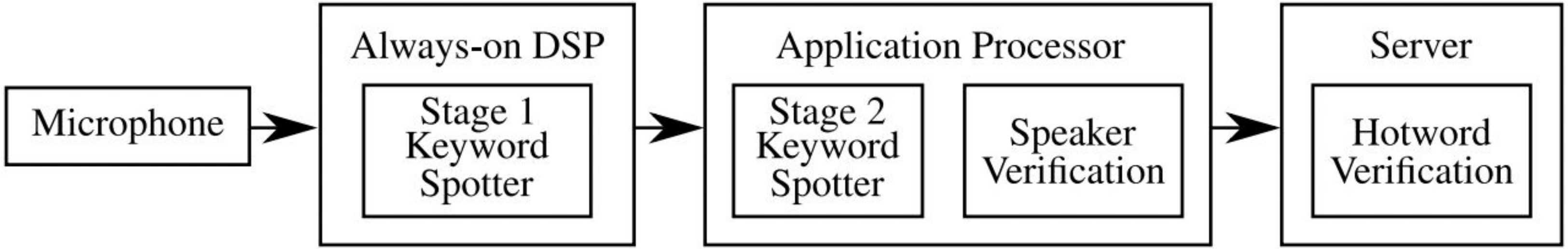
Keyword Spotting for Google Assistant Using Contextual Speech Recognition

- Google, 2017



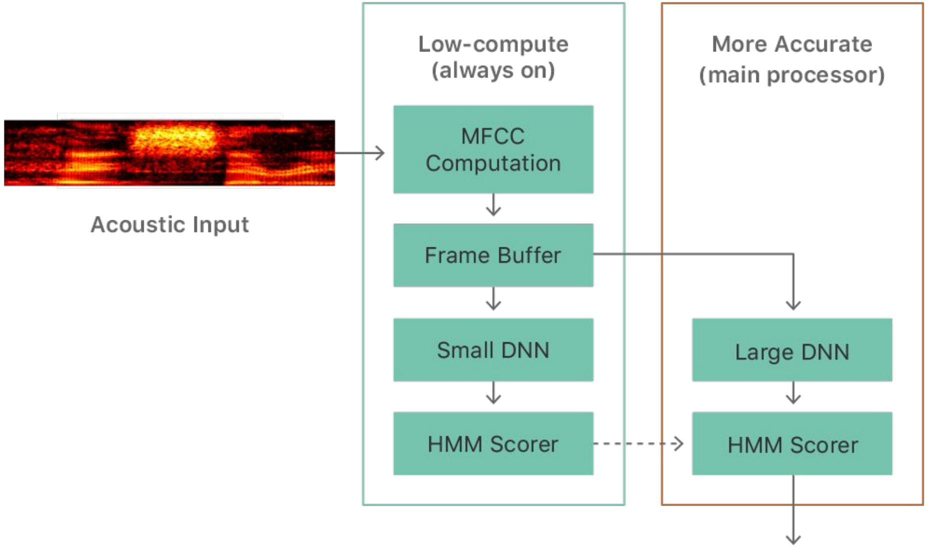
A Cascade Architecture for Keyword Spotting on Mobile Devices

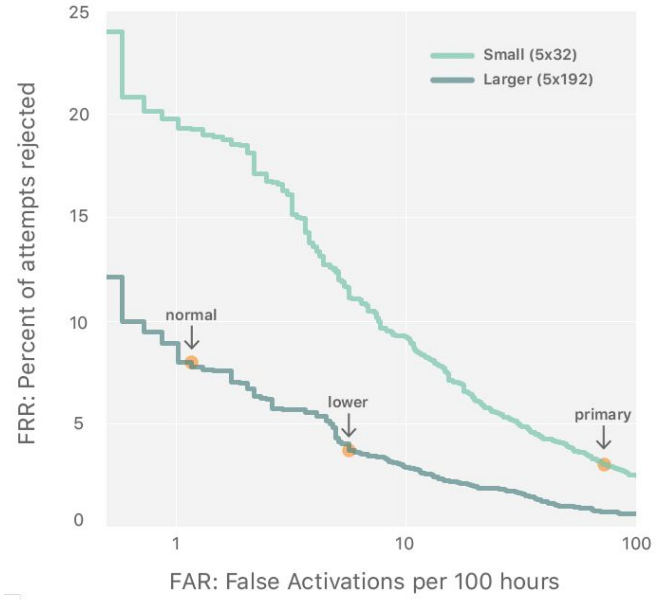
- Google, 2017



[Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant]

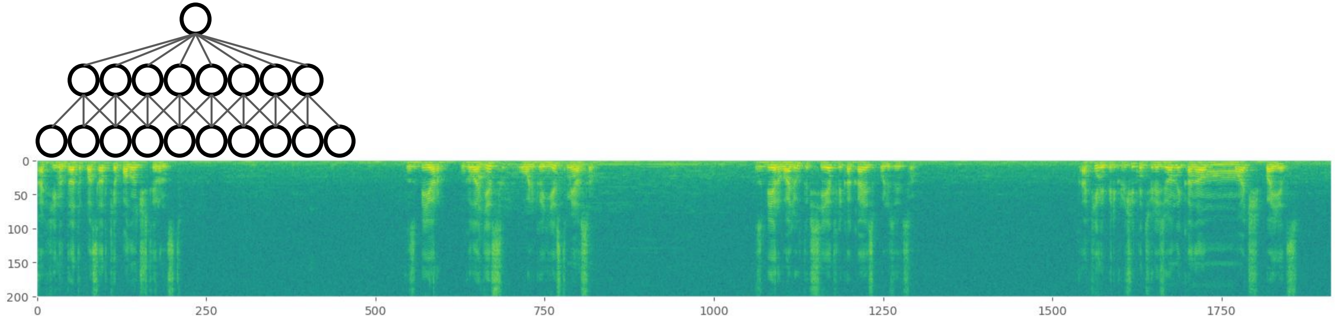
- Apple, 2017



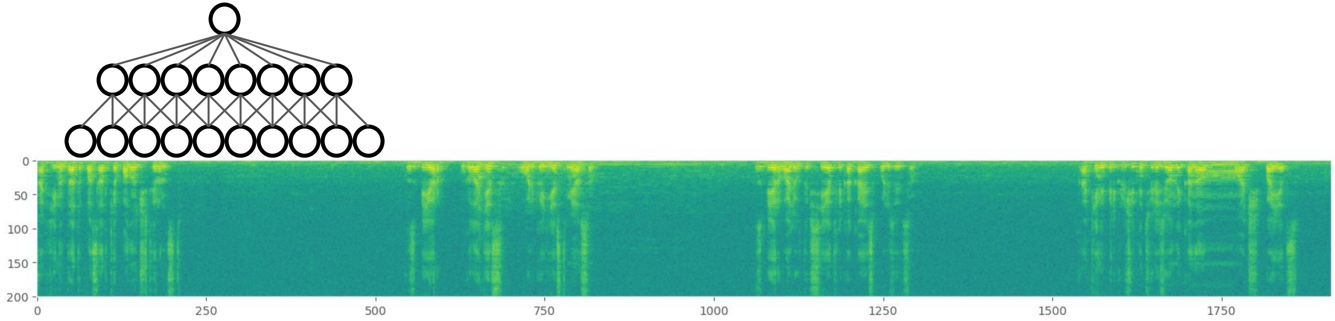


Streaming Detection

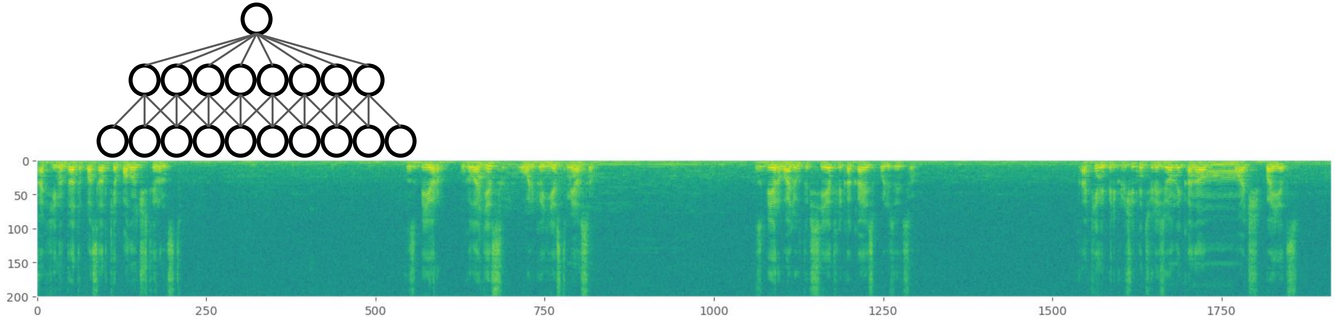
Streaming Keyword Spotting



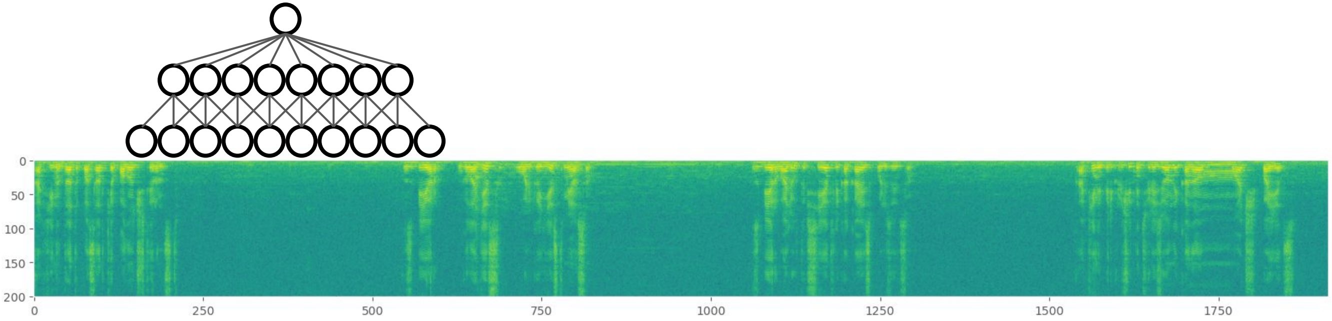
Streaming Keyword Spotting



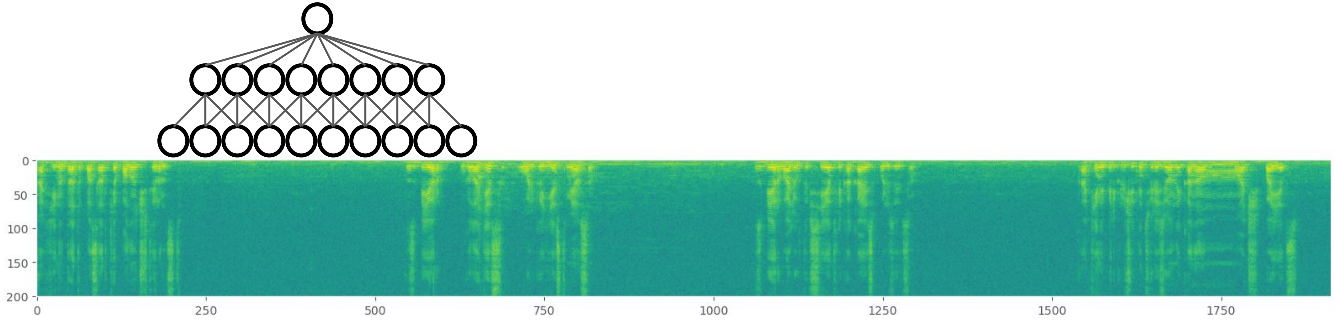
Streaming Keyword Spotting



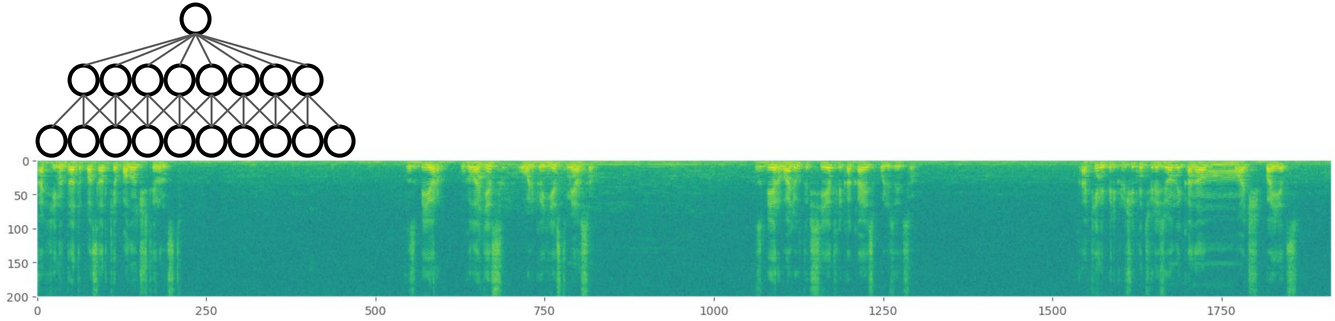
Streaming Keyword Spotting



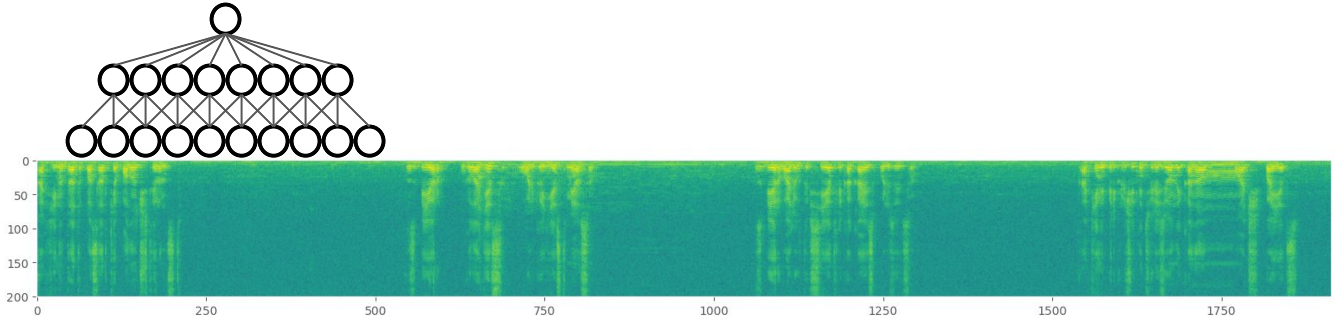
Streaming Keyword Spotting



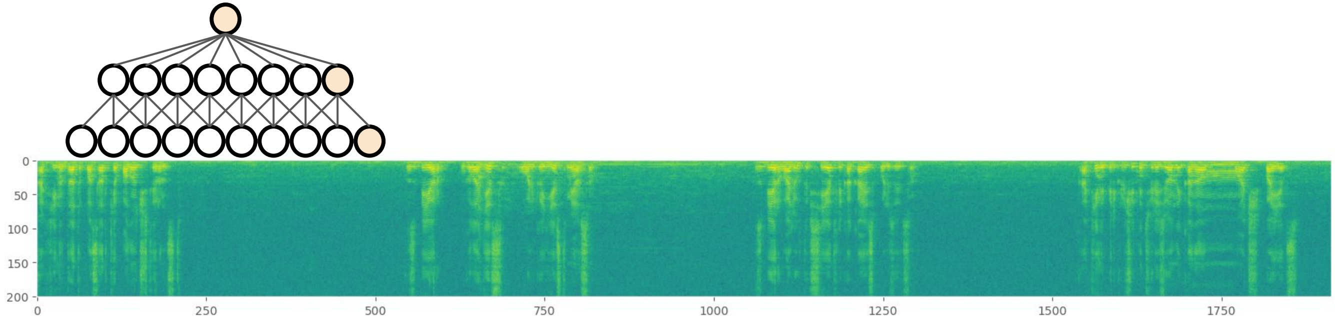
Streaming Keyword Spotting



Streaming Keyword Spotting

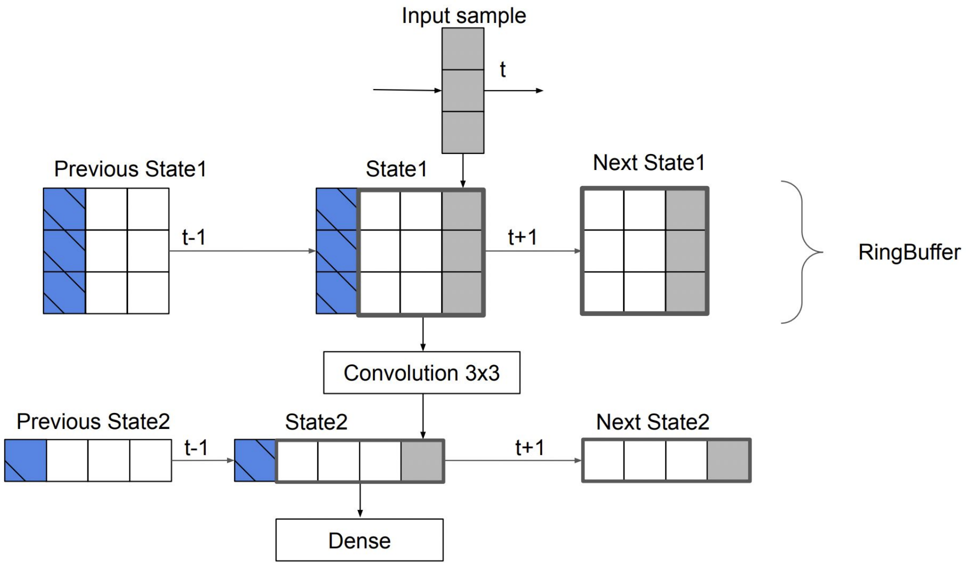


Streaming Keyword Spotting

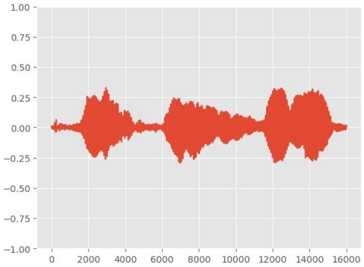


Streaming keyword spotting on mobile devices

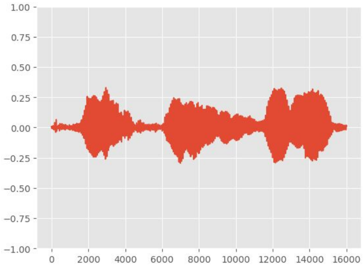
- Google, 2020
- https://cms.tinyml.org/wp-content/uploads/talks2022/Oleg-Rybakov.pdf



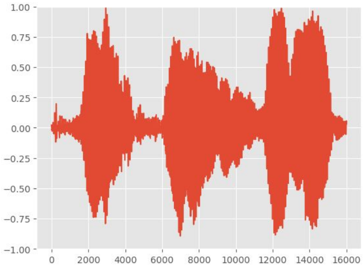
Streaming Keyword Spotting: Norm / Unnorm Gain



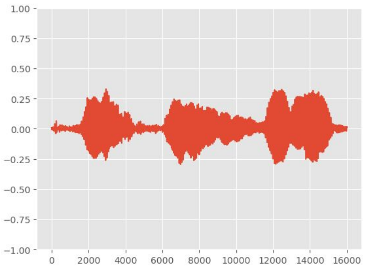
Streaming Keyword Spotting: Norm / Unnorm Gain



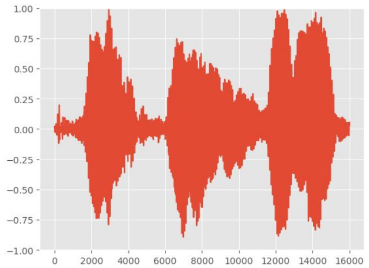
`wav / wav.abs().max()`

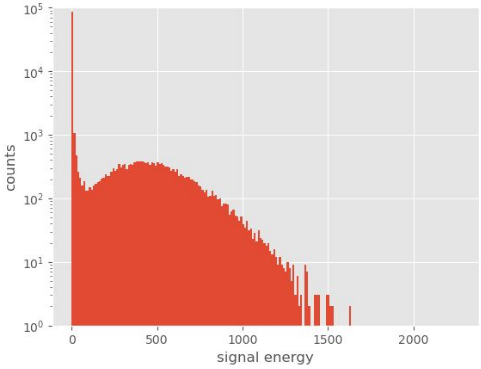


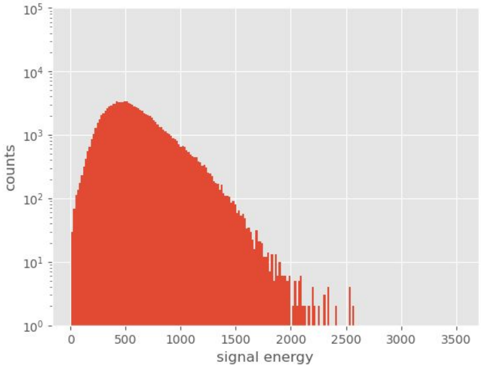
Streaming Keyword Spotting: Norm / Unnorm Gain

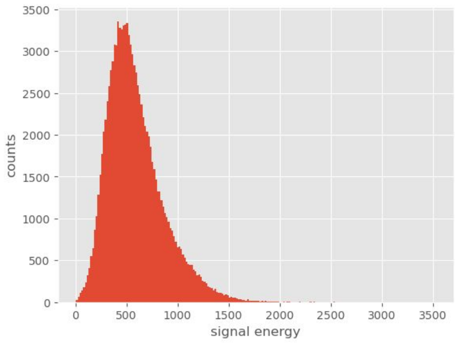


`wav / wav.abs().max()`





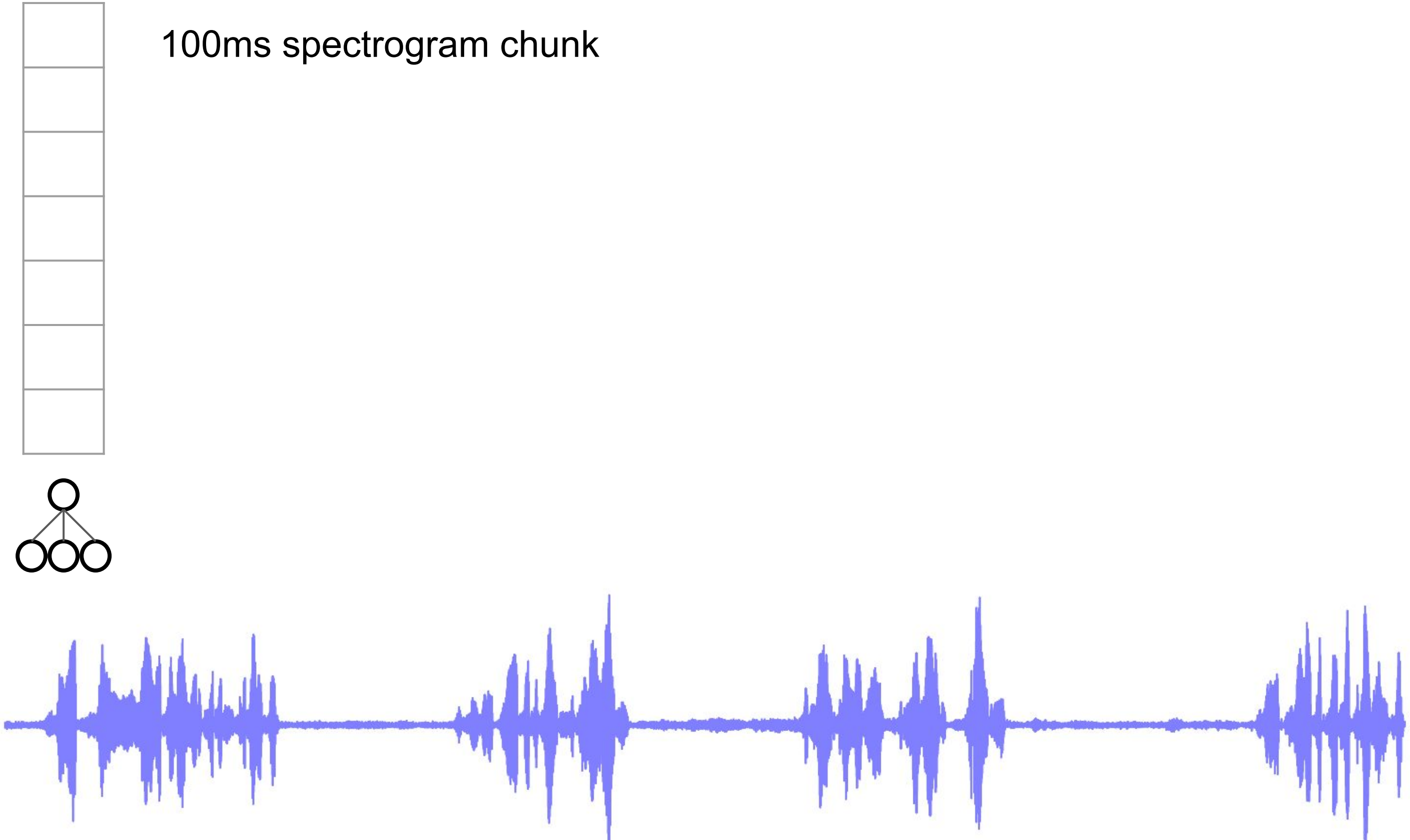




Streaming Feature Extraction

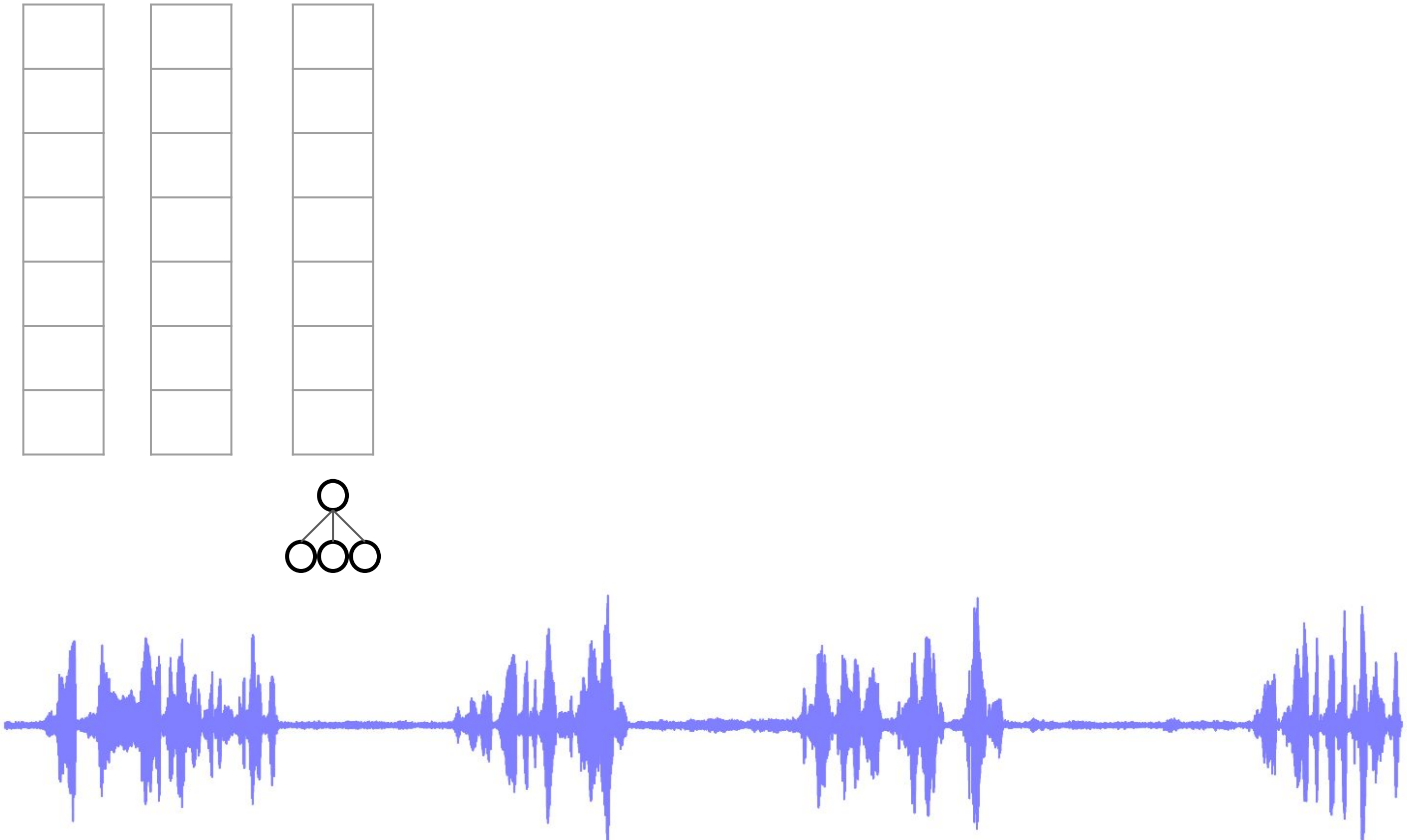
Streaming Feature Extraction

100ms spectrogram chunk

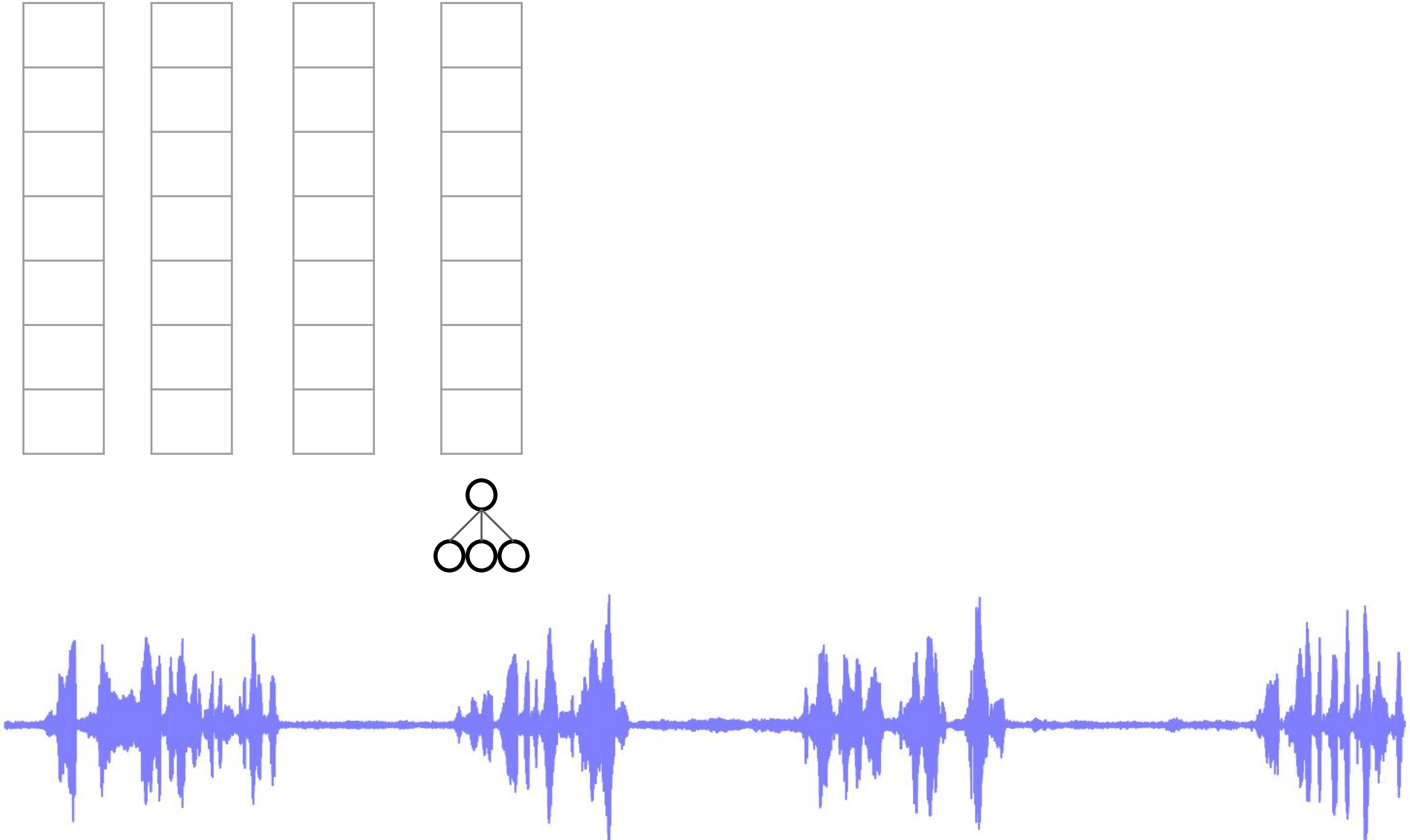


Streaming Feature Extraction

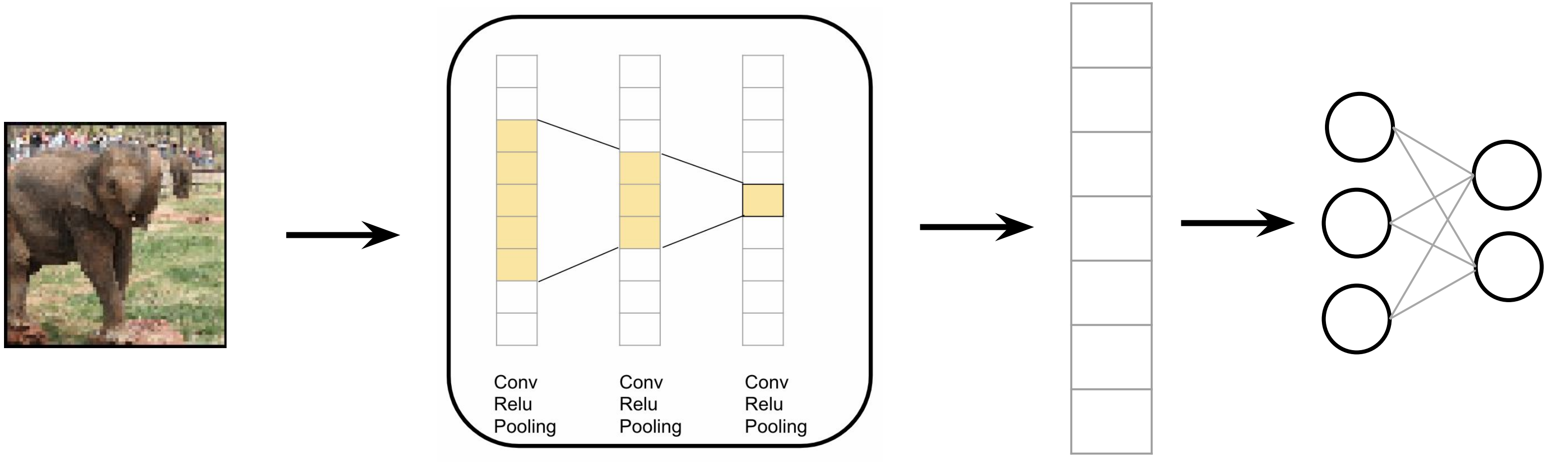
Streaming Feature Extraction



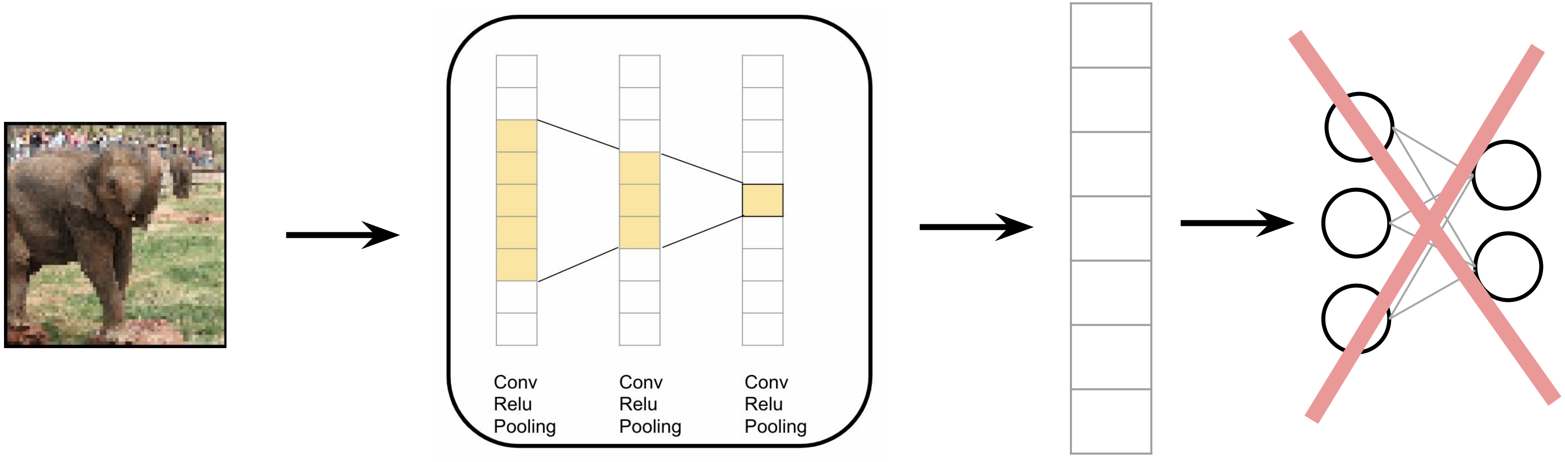
Streaming Feature Extraction



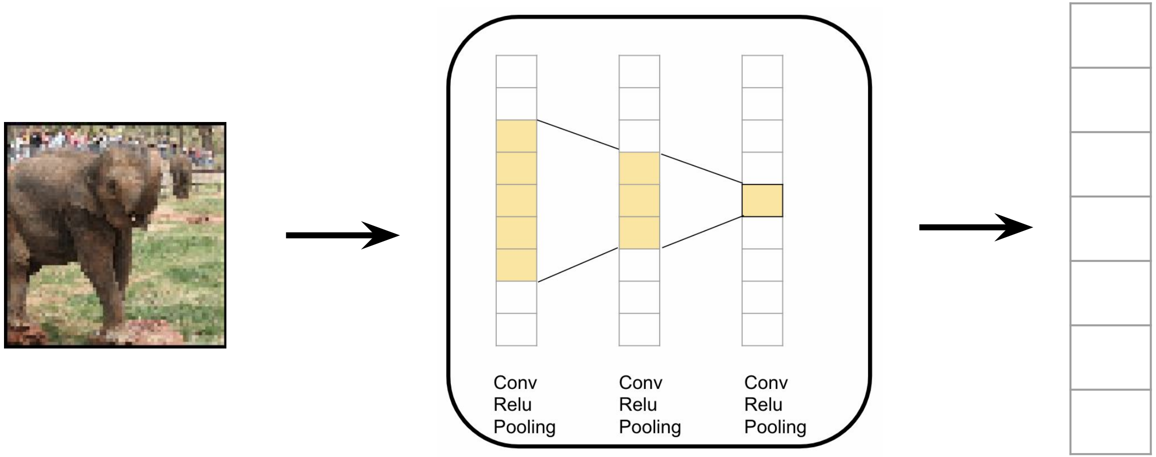
Transfer Learning: Computer Vision



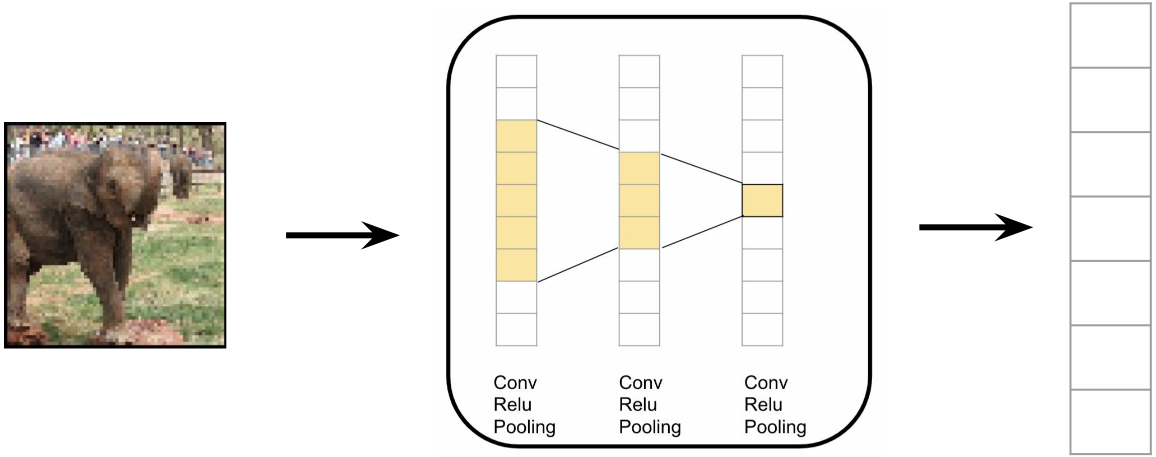
Transfer Learning: Computer Vision



Transfer Learning: Computer Vision



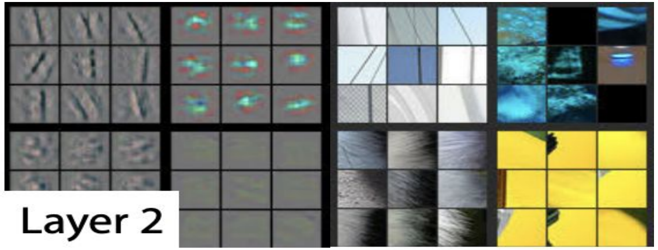
Transfer Learning: Computer Vision

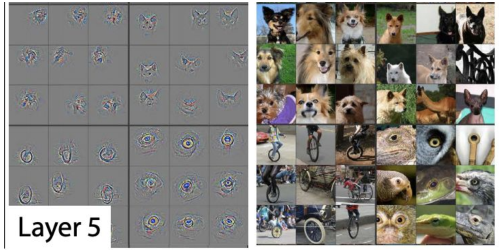


Nearest
Neighbours

Transfer Learning: Learned Features

Layer 1





Transfer Learning: Learned Features

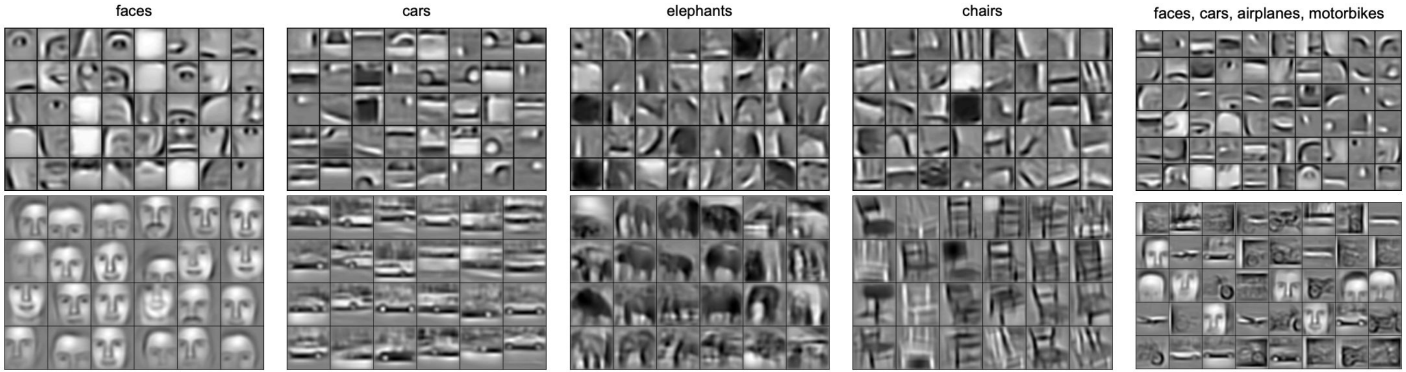
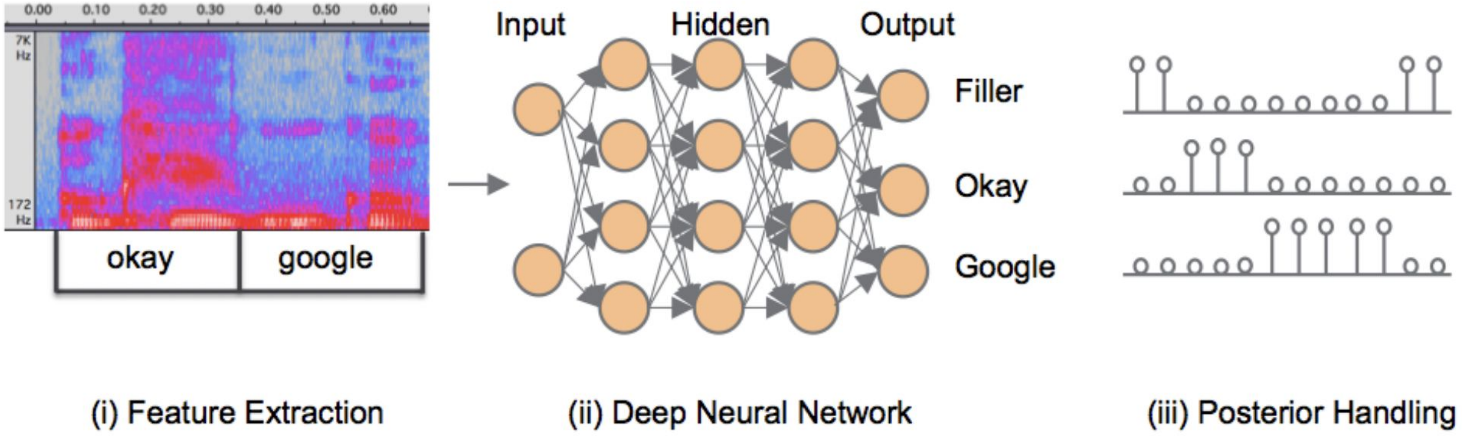


Figure 3. Columns 1-4: the second layer bases (top) and the third layer bases (bottom) learned from specific object categories. Column 5: the second layer bases (top) and the third layer bases (bottom) learned from a mixture of four object categories (faces, cars, airplanes, motorbikes).

https://dl.acm.org/doi/10.1145/1553374.1553453

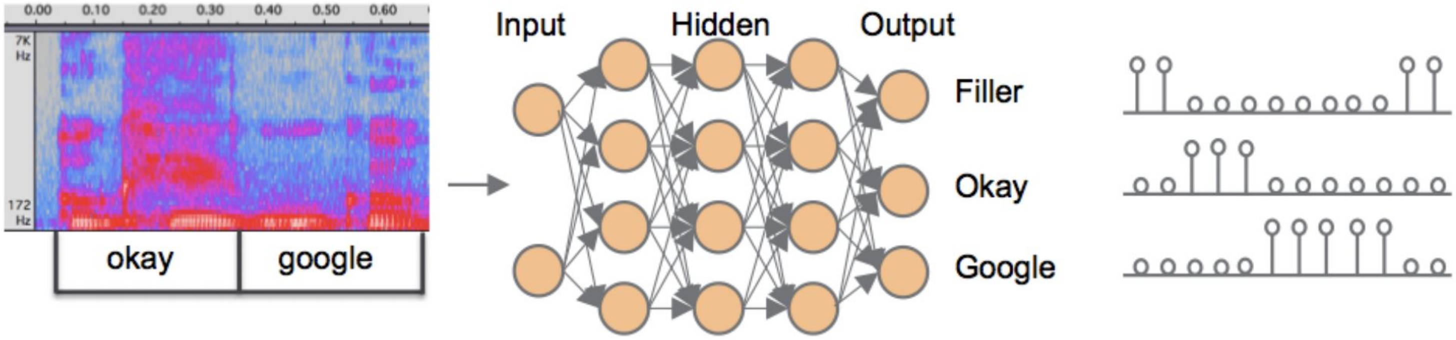
SMALL-FOOTPRINT KEYWORD SPOTTING USING DEEP NEURAL NETWORKS

- Google (2014). Cited by 532



SMALL-FOOTPRINT KEYWORD SPOTTING USING DEEP NEURAL NETWORKS

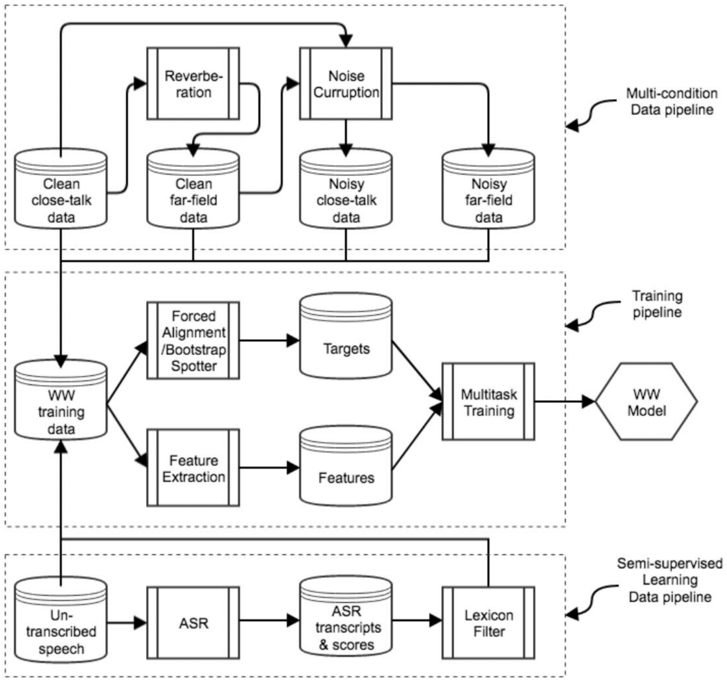
- Google (2014). Cited by 532
- **Transfer Learning**: Here, we use a deep neural network for speech recognition with suitable topology to initialize the hidden layers of the network. All layers are updated in training.



(i) Feature Extraction (ii) Deep Neural Network (iii) Posterior Handling

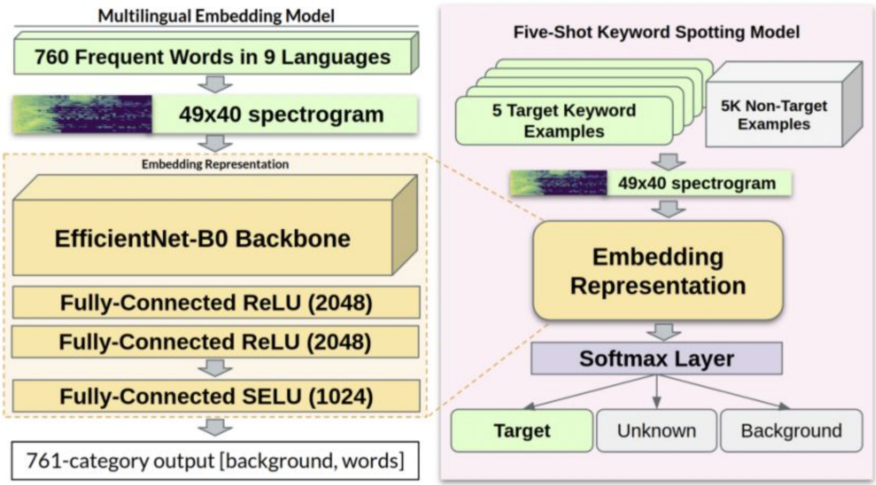
TOWARDS DATA-EFFICIENT MODELING FOR WAKE WORD SPOTTING

- Amazon Alexa, 2020
- The models are trained using transfer-learning paradigm where the weights of the DNN are initialized by an ASR acoustic model of the same architecture and size trained with ASR senone targets



Few-Shot Keyword Spotting in Any Language

- Google, 2021



Language	# words	# train	# val	val acc
English	265	518760	57640	78.95
German	152	287100	31900	79.90
French	105	205920	22880	79.16
Kinyarwanda	68	134640	14960	73.64
Catalan	80	132660	14740	87.63
Persian	35	69300	7700	85.70
Spanish	31	61380	6820	79.65
Italian	17	31680	3520	81.16
Dutch	7	13860	1540	72.60
Model	760	1455300	161700	79.81

Transfer Learning: may be useful in homework

- https://huggingface.co/SberDevices/quartznet-russian

- QUARTZNET: DEEP AUTOMATIC SPEECH RECOGNITION WITH 1D TIME-CHANNEL SEPARABLE CONVOLUTIONS (Nvidia, 2019)

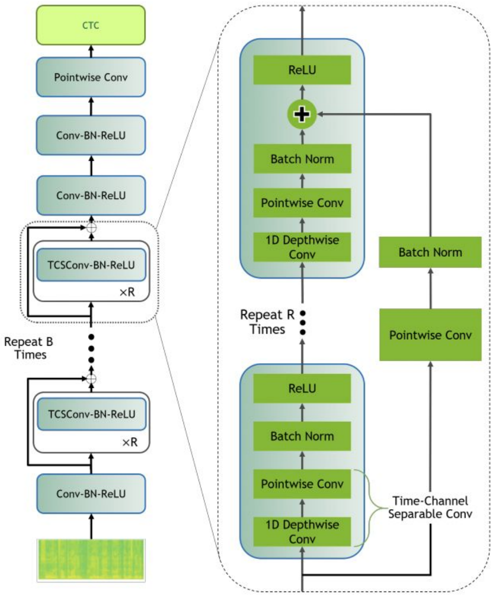
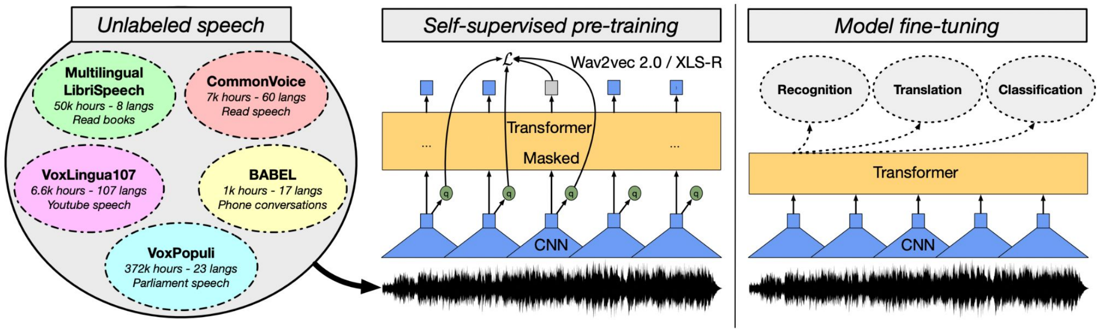


Fig. 1. QuartzNet BxR architecture

Transfer Learning: may be useful in homework



- XLS-R: SELF-SUPERVISED CROSS-LINGUAL SPEECH REPRESENTATION LEARNING AT SCALE (FAIR, 2021)

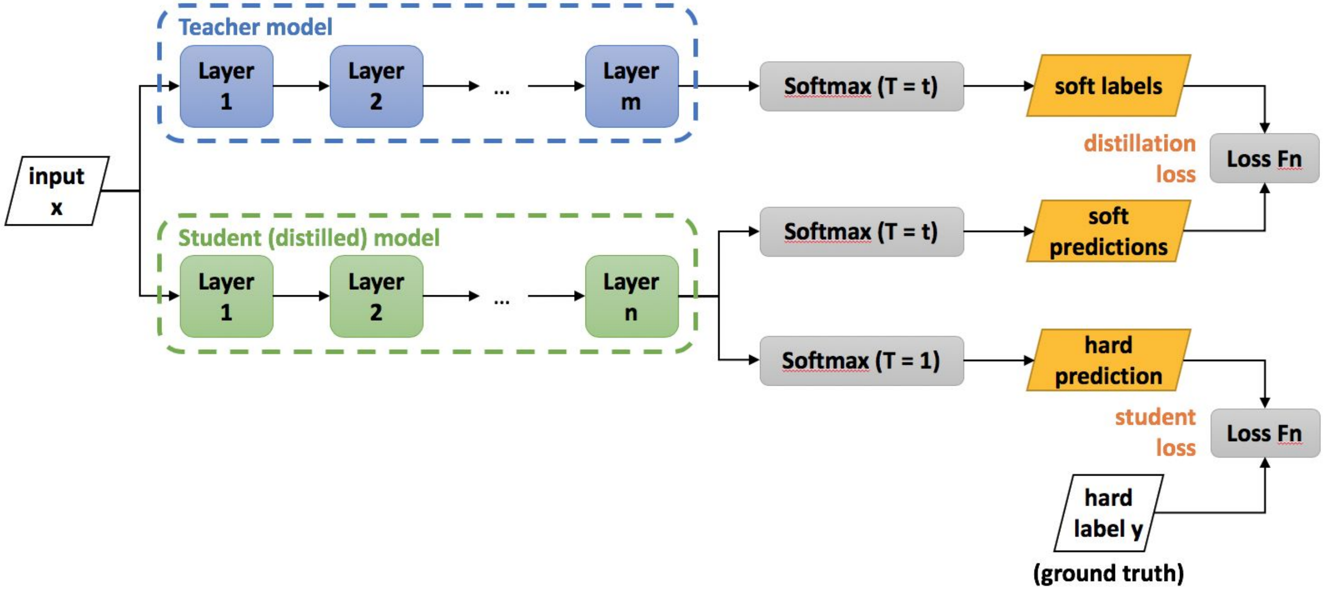
Transfer Learning: may be useful in homework

```
import transformers


w2v_backbone = transformers.Wav2Vec2Model.from_pretrained("facebook/wav2vec2-xls-r-300m")
for param in w2v_backbone.parameters():
    param.requires_grad = False

clf_head = torch.nn.Sequential(
    torch.nn.AdaptiveAvgPool1d(output_size=1),
    torch.nn.Flatten(),
    torch.nn.Linear(1024, 256),
    torch.nn.ReLU(),
    torch.nn.Linear(256, 5)
)


out = w2v_backbone(wav)['last_hidden_state']
out = out.transpose(1, 2)
logits = clf_head(out)
# tensor([[ 0.0204,  0.0275,  0.0601, -0.0196,  0.0896]], grad_fn=<AddmmBackward0>)
```

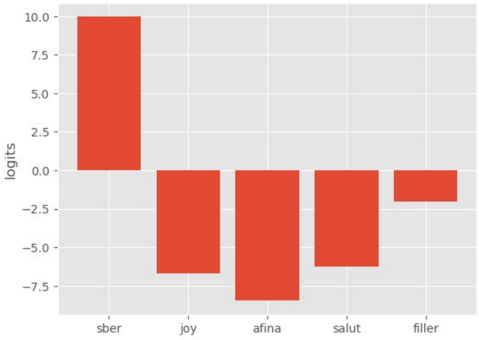
Model Compression

Distilling the Knowledge in a Neural Network

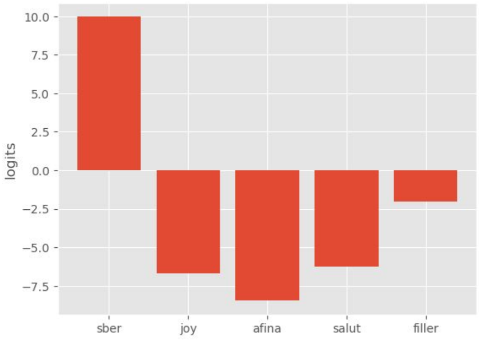
- Hinton, 2015, cited by ~ 11k
- https://intellabs.github.io/distiller/knowledge_distillation.html



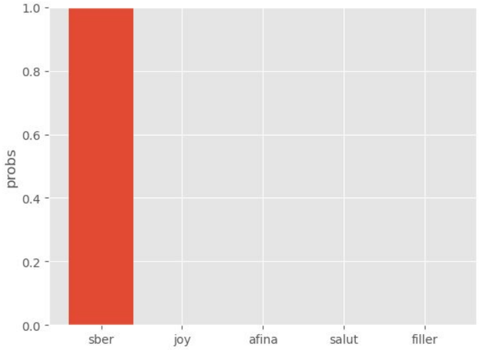
Distillation Temperature



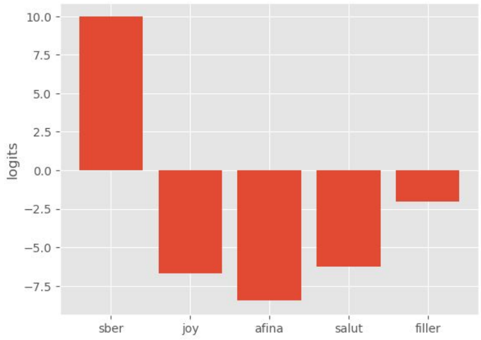
Distillation Temperature



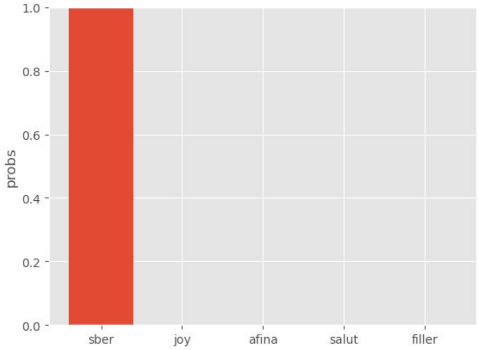
`F.softmax(logits)`



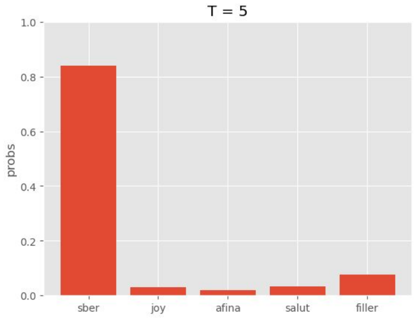
Distillation Temperature

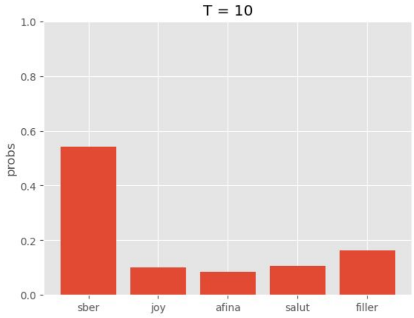


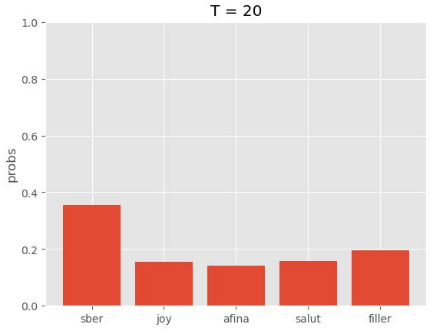
`F.softmax(logits)`



`F.softmax(logits / T)`

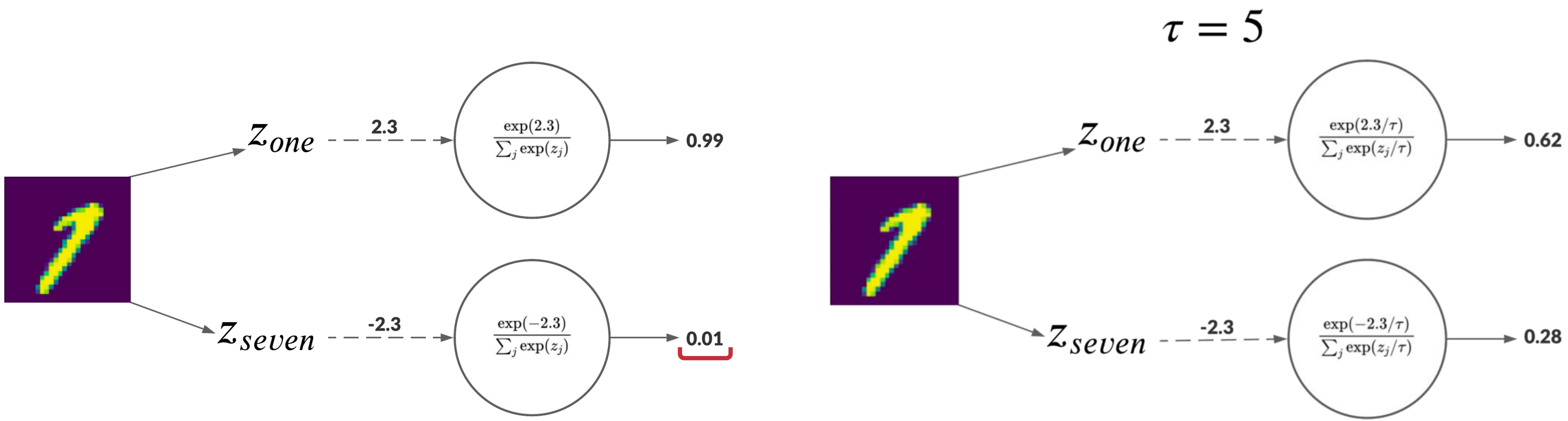






Distillation Temperature

- wandb tutorial:

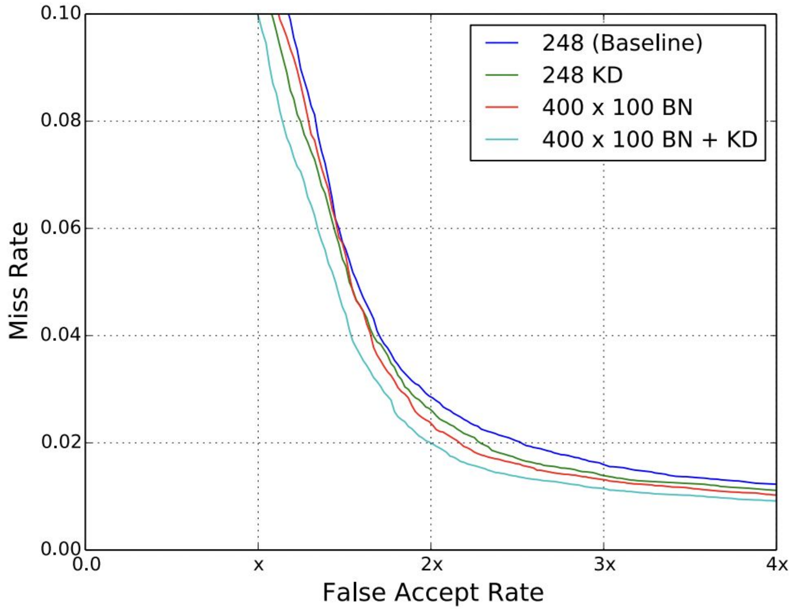


Model compression applied to small-footprint keyword spotting

- Amazon, 2016

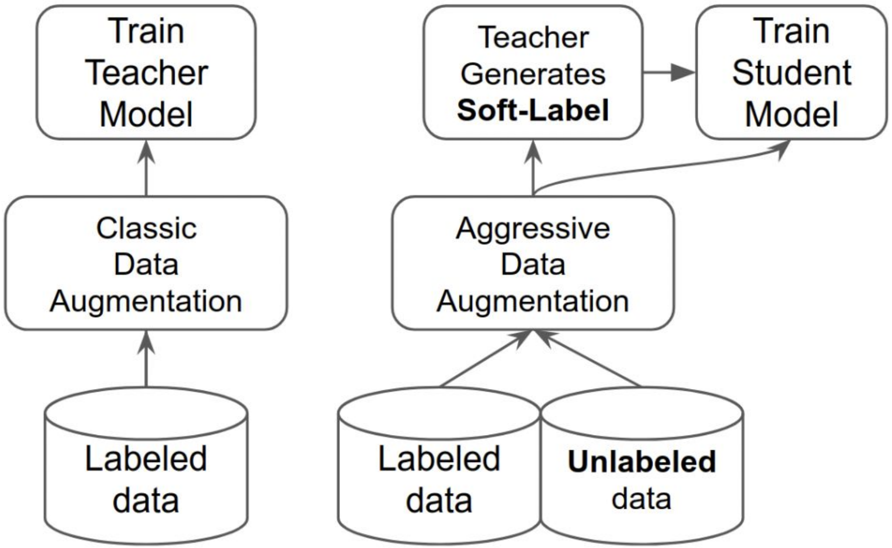
$$\lambda \sum_i \log(p_i) t_i + \frac{1-\lambda}{T^2} \sum_i \log(p_i(T)) q_i(T)$$

$$p_i(T) = \frac{p_i^{1/T}}{\sum_j p_j^{1/T}}, q_i(T) = \frac{q_i^{1/T}}{\sum_j q_j^{1/T}},$$



Noisy student-teacher training for robust keyword spotting

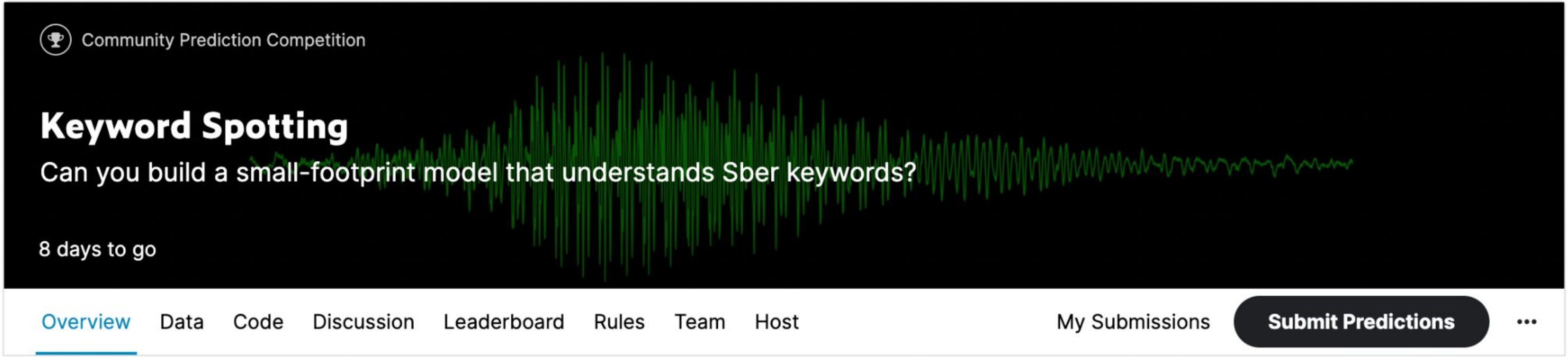
- Google, 2021



Homework

Keyword Spotting

- [Kaggle In-Class Competition]
- 100k train, 2k test
- model: <= **1e4** params, <= **1e6** MACs
- report + model-checkpoint + leaderboard submits
- deadline: 2022-10-18 17:59



Thank you for your attention!

@georgygospodinov