

Deep Learning Audio

Lecture 5

Pavel Severilov

Moscow Institute of Physics and Technology

2022

Outline

1. Keyword spotting (KWS)
2. KWS problems and practices

Outline

1. Keyword spotting (KWS)
2. KWS problems and practices

KWS: introduction

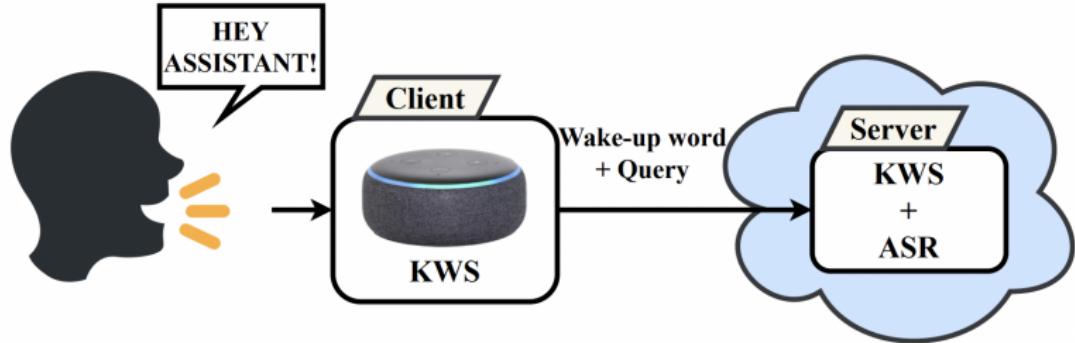
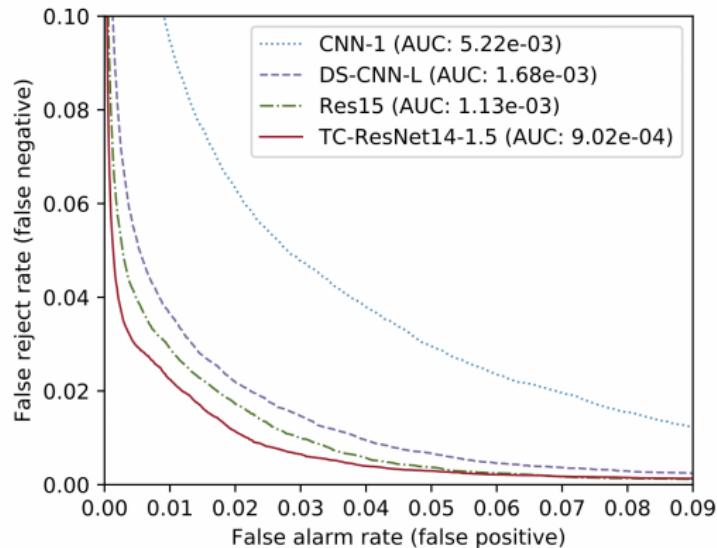


Figure: Typical voice assistant client-server framework

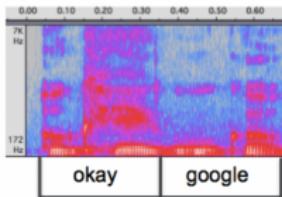
What do we need? Lightweight robust ASR model

KWS: quality

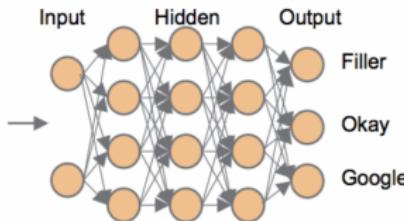


- ▶ False alarm rate (false positive rate): $\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$
- ▶ False Reject Rate (false negative rate): $\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}$
- ▶ FA/FR per hour, lower curves are better

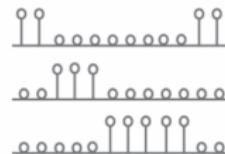
DNN based KWS



(i) Feature Extraction



(ii) Deep Neural Network



(iii) Posterior Handling

- ▶ Feature extraction: time windows w_{smooth} : 10 future frames + 30 frames in the past
- ▶ Raw posteriors from DNN are noisy, so smooth them over w_{smooth}

$$p'_{ij} = \frac{1}{j - h_{smooth} + 1} \sum_{k=h_{smooth}} p_{ij},$$

$h_{smooth} = \max\{1, j - w_{smooth} + 1\}$ – the index of the first frame within the smoothing window

Chen et al., Small-footprint keyword spotting using deep neural networks, IEEE International Conference on Acoustics, Speech and Signal Processing, 2014

CNN based KWS

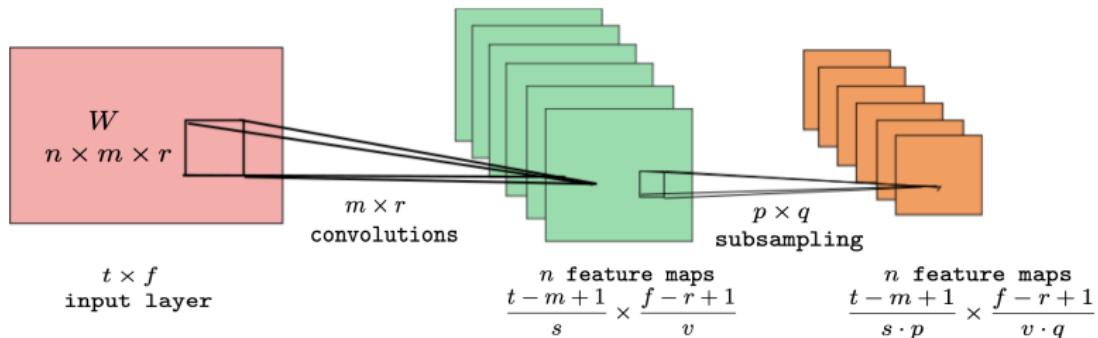
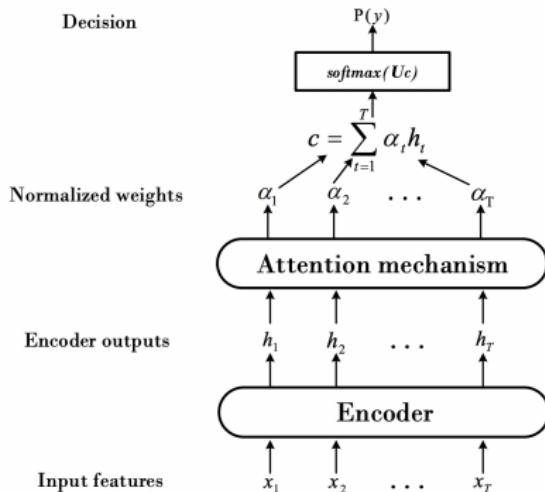


Figure: CNN architecture

- ▶ Input signal $V \in \mathbb{R}^{t \times f}$, t and f – input feature dimension in time and frequency.
- ▶ Weight matrix $W \in \mathbb{R}^{(m \times r) \times n}$ is convolved with input V
- ▶ Filter of size $m \times r$, stride by s in time and v in frequency
- ▶ Pooling size $p \times q$

Sainath et al., Convolutional neural networks for small-footprint keyword spotting, INTERSPEECH, 2015

Attention-based KWS



► Encoder:
RNN/GRU/LSTM/CRNN

$$\mathbf{h} = \text{Encoder}(\mathbf{x})$$

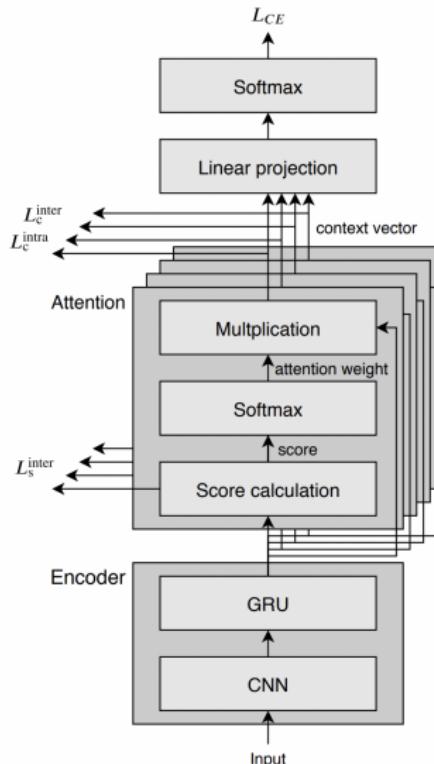
► Soft attention:

$$e_t = v^T \tanh(\mathbf{W}\mathbf{h}_t + \mathbf{b}).$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^T \exp(e_j)}.$$

$$\mathbf{c} = \sum_{t=1}^T \alpha_t \mathbf{h}_t.$$

Multihead attention based KWS



- ▶ Retrieves richer information than a single-head attention which only summarizes the whole sequence into one context vector
- ▶ The diversity is not guaranteed by its natural form → use Orthogonality regularization

Multihead attention based KWS: orthogonality

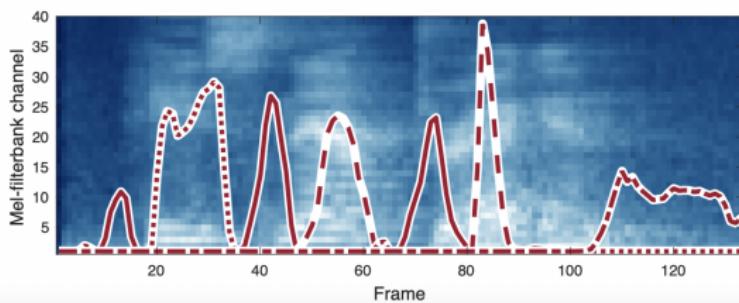
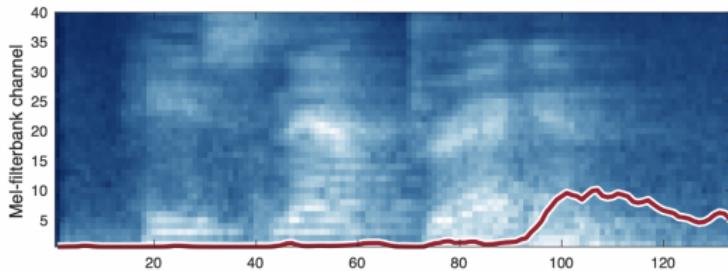


Figure: Attention weights overlaid on Mel-spectrogram with the configurations of single head attention (top), 4-head attention with regularization (bottom)

Multihead attention based KWS: orthogonality

- Context matrix and the score matrix from normalized vectors (H – number of attention heads):

$$\mathbf{C}^{(n)} = \left[\bar{\mathbf{c}}_1^{(n)}, \bar{\mathbf{c}}_2^{(n)}, \dots, \bar{\mathbf{c}}_H^{(n)} \right], \quad \mathbf{E}^{(n)} = \left[\bar{\mathbf{e}}_1^{(n)}, \bar{\mathbf{e}}_2^{(n)}, \dots, \bar{\mathbf{e}}_H^{(n)} \right]$$

- Inter-head orthogonality regularization

$$\mathcal{L}_c^{\text{inter}} = \frac{1}{N_p} \sum_{n=1}^N \frac{y^{(n)}}{H(H-1)} \| \mathbf{C}^{(n)T} \mathbf{C}^{(n)} - \mathbf{I}_H \|_F^2$$

$$\mathcal{L}_s^{\text{inter}} = \frac{1}{N_p} \sum_{n=1}^N \frac{y^{(n)}}{H(H-1)} \| \mathbf{E}^{(n)T} \mathbf{E}^{(n)} - \mathbf{I}_H \|_F^2$$

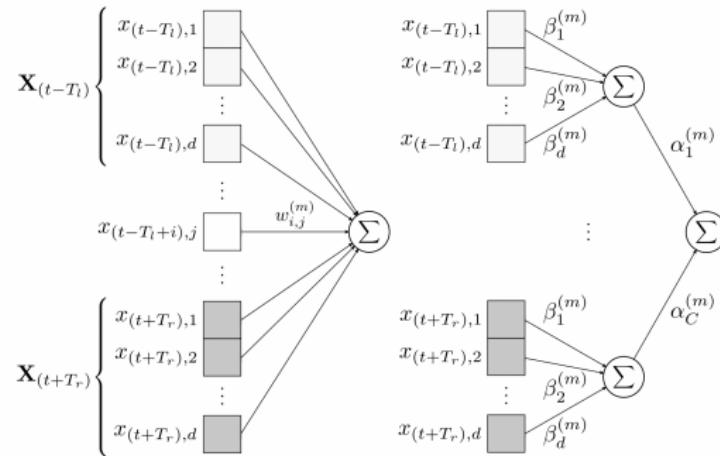
- Intra-head non-orthogonality regularization

$$\mathcal{L}_c^{\text{intra}} = \frac{1}{H} \sum_{i=1}^H \frac{1}{N_p(N_p-1)} \| \mathbf{Y} (\tilde{\mathbf{C}}_i^T \tilde{\mathbf{C}}_i - \mathbf{I}_N) \mathbf{Y} \|_F^2$$

- Cross entropy loss with the regularization terms

$$\theta^* = \arg \min \{ \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_c^{\text{inter}} - \lambda_2 \mathcal{L}_c^{\text{intra}} + \lambda_3 \mathcal{L}_s^{\text{inter}} \}$$

Single Value Decomposition Filter (SVDF)



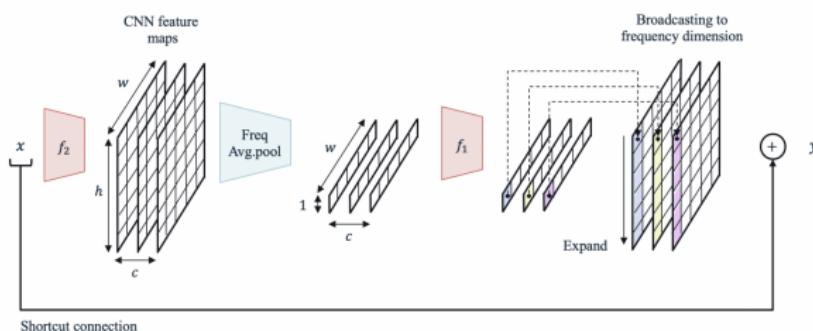
Before: $a_t^{(m)} = f \left(\sum_{i=0}^{C-1} \sum_{j=1}^d w_{i,j}^{(m)} x_{(t-T_l+i),j} \right)$

After:

$$a_t^{(m)} \approx f \left(\sum_{i=0}^{C-1} \alpha_i^{(m)} \sum_{j=1}^d \beta_j^{(m)} x_{(t-T_l+i),j} \right), \quad w_{i,j}^{(m)} \approx \alpha_i^{(m)} \beta_j^{(m)}$$

Broadcasted Residual Learning

Broadcasted Residual Learning



Normal block

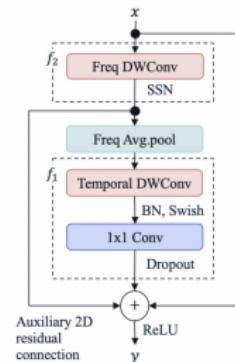


Figure: Left, Broadcasted Residual Learning $y = x + BC(f_1((f_2(x))))$, number of channels c .

Right, BCResBlock $y = x + f_2(x) + BC(f_1(\text{avgpool}(f_2(x))))$.

Convolutions in audio

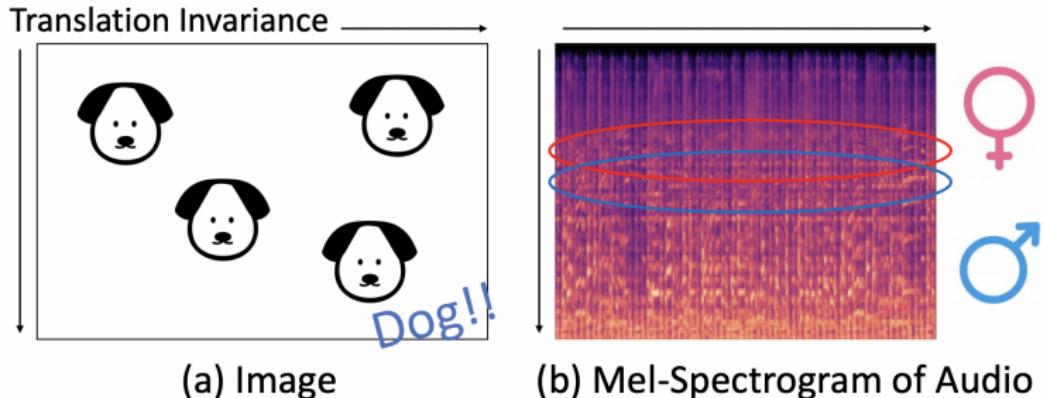


Figure: 2D convolution on image and audio input: unlike image processing, the feature in different audio frequency bands has different information.

Subspectral Normalization (SSN)

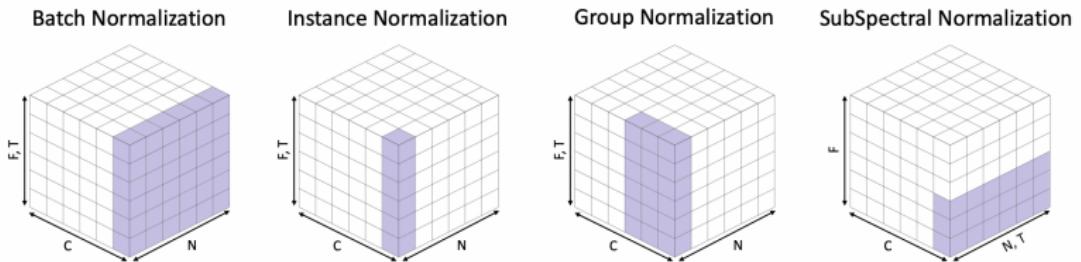


Figure: Normalization methods on Frequency-Time audio input, N – batch axis, C – channels, F – frequency, T – time axis. SSN splits the frequency dimension into multiple sub-bands and normalizes each group

Keyword Transformer based KWS

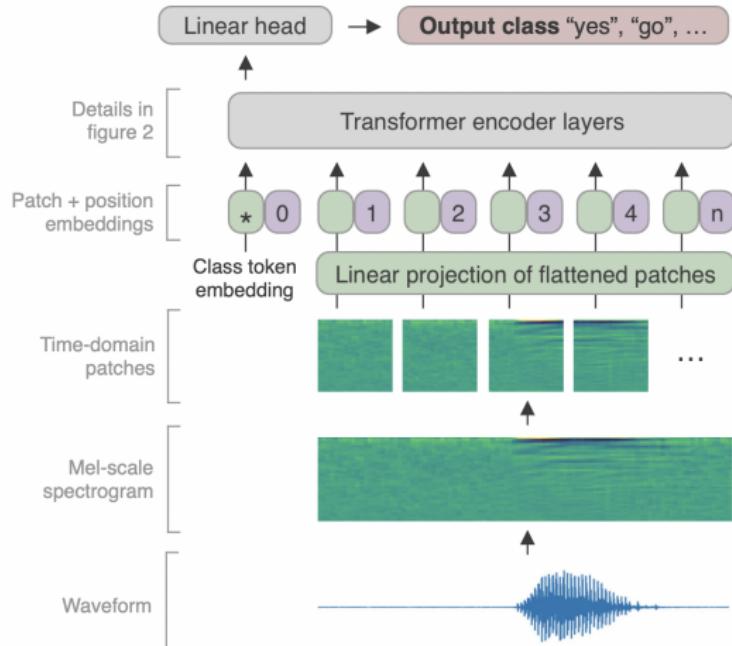


Figure: Keyword Transformer architecture

KWS: summary

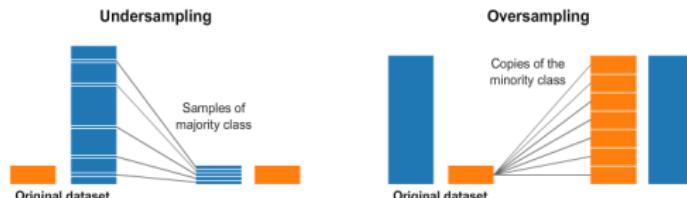
KWS Idea	Year/Company	Params	Accuracy, V1 dataset
DNN	2014, Google	~224k	91.2
CNN	2015, Google	20-70k	95.4
Attention	2018, Xiaomi	~84k	95.6
Multihead attention	2019, Qualcomm	743k	97.2
Single Value Decomposition Filter	2015/2019, Google	40-700k	96.3
Broadcasted Residual	2021, Qualcomm	9.2k	96.6
Subspectral Normalization	2021, Qualcomm	113-243k	97.5
Transformer	2021, Arm ML Research Lab	607-5361k*	97.5

Outline

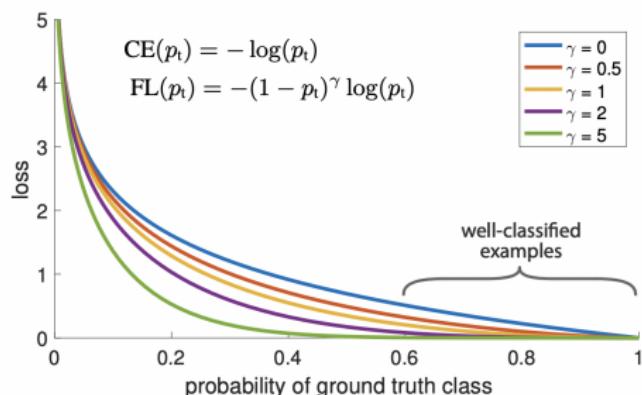
1. Keyword spotting (KWS)
2. KWS problems and practices

Class imbalance

- ▶ Oversampling



- ▶ Class weighted Cross-entropy (CE)
- ▶ Focal Loss



Robustness

- ▶ Multitask learning: Train ASR and KWS simultaneously

$$\mathcal{L} = \gamma \mathcal{L}^{(1)} + (1 - \gamma) \mathcal{L}^{(2)}, \quad 0 \leq \gamma \leq 1$$

- ▶ Augmentations
 - ▶ Pitch and Tempo
 - ▶ Time shifts
 - ▶ Resampling
 - ▶ Environment Simulation
 - ▶ Background noise
- ▶ Hard negative mining
- ▶ Federated Learning
- ▶ Clean data