Deep Learning for Audio

Lecture 4

Pavel Severilov

AI Masters

2024

Outline

1. Keyword spotting (KWS)

2. KWS problems and practices

Outline

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

KWS: introduction

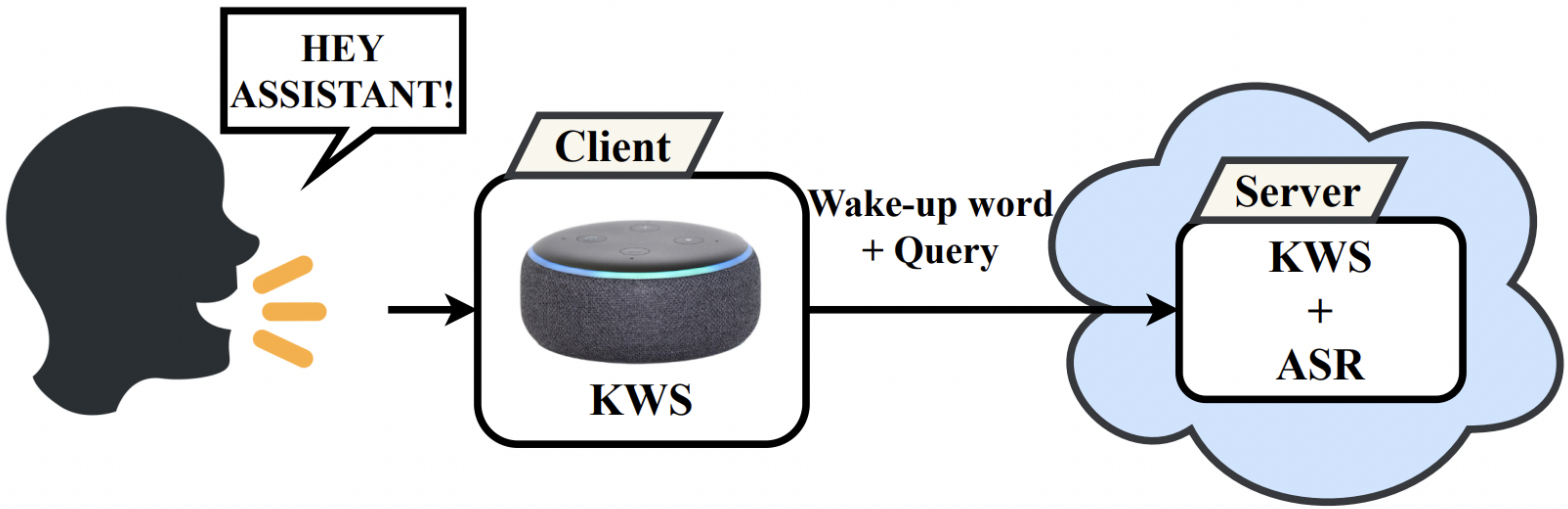
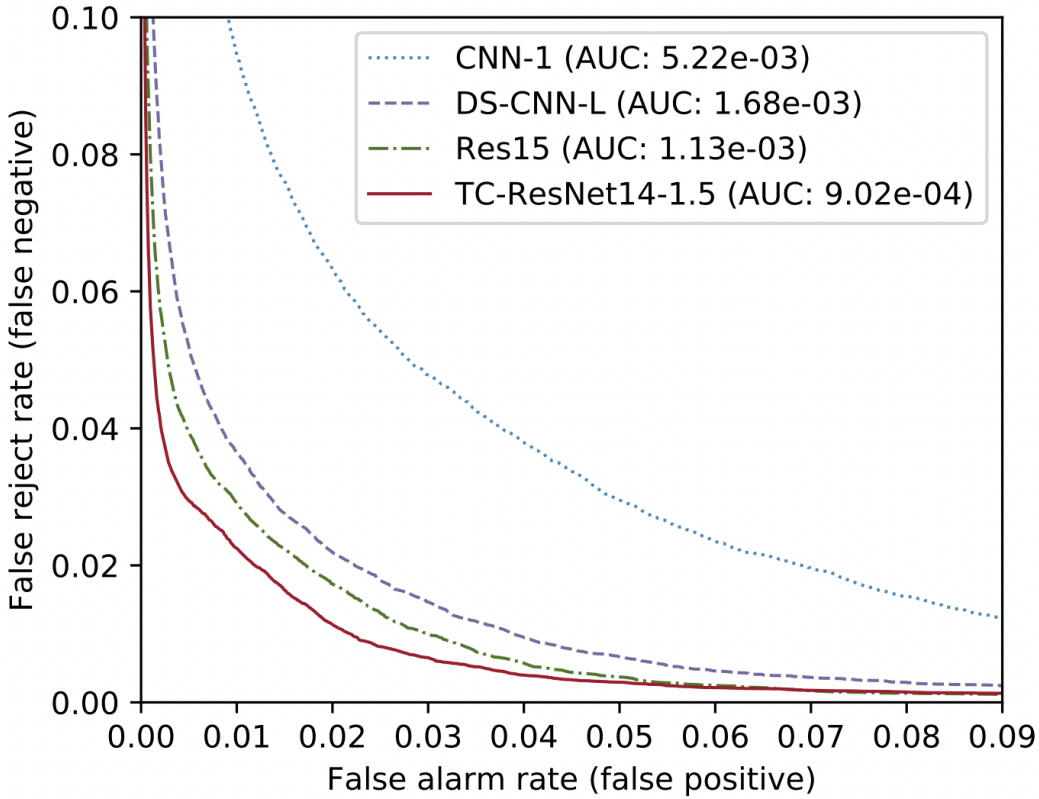


Figure: Typical voice assistant client-server framework

What do we need? Lightweight robust ASR model

López-Espejo et al., Deep Spoken Keyword Spotting: An Overview, IEEE Access, 2021

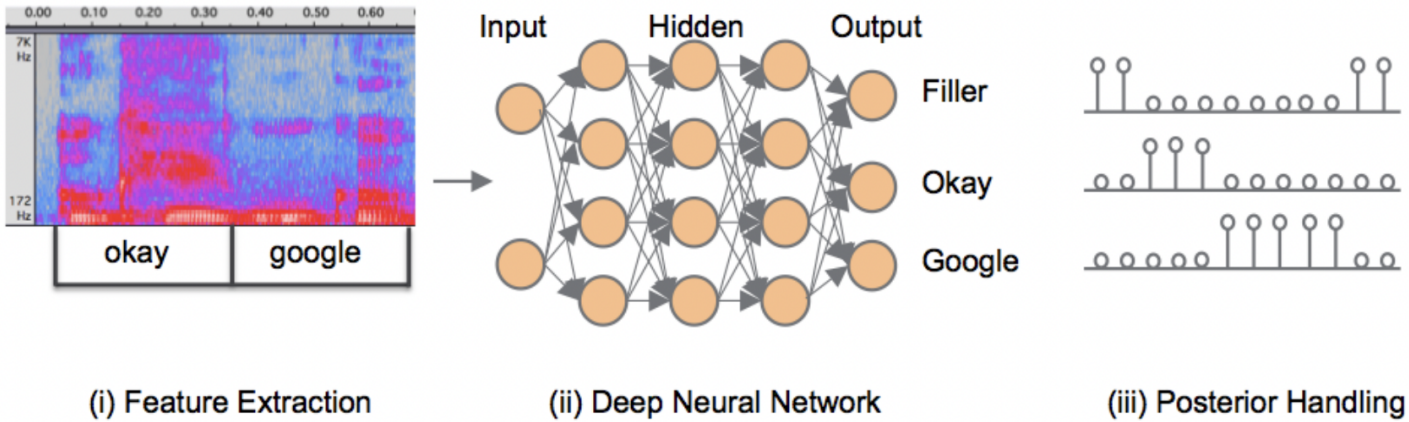
KWS: quality



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

Choi et al., Temporal Convolution for Real-time Keyword Spotting on Mobile Devices, 2019

DNN based KWS



- 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}

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

$h_{smooth} = \max\{\mathbf{1}, j - w_{smooth} + 1\}$ – 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

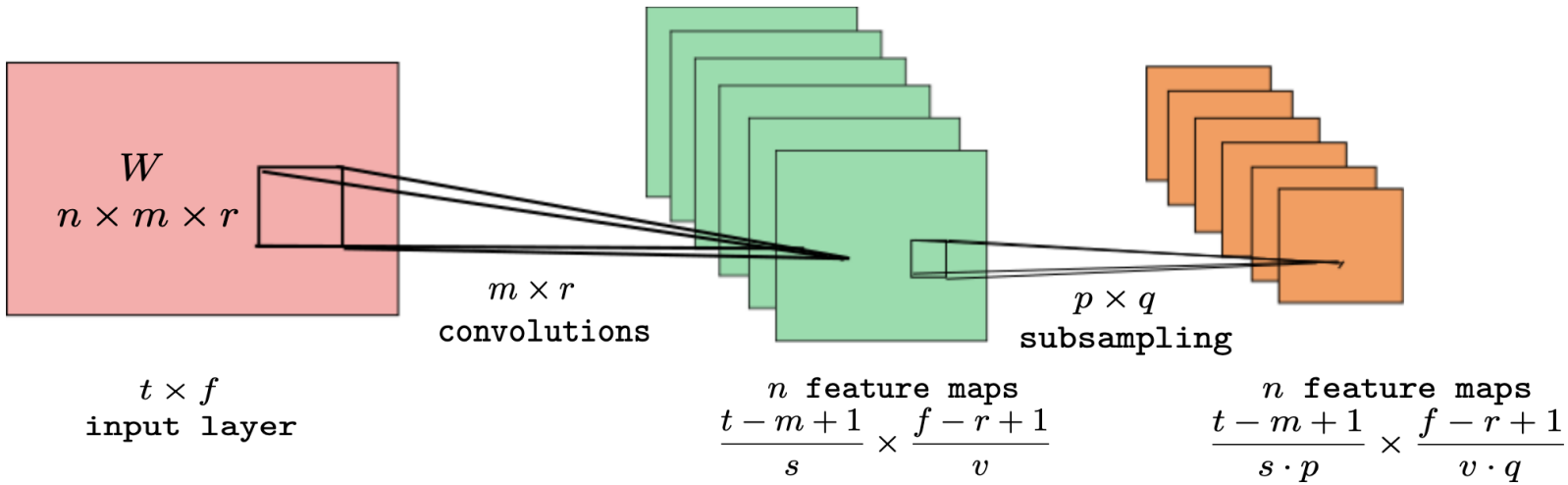
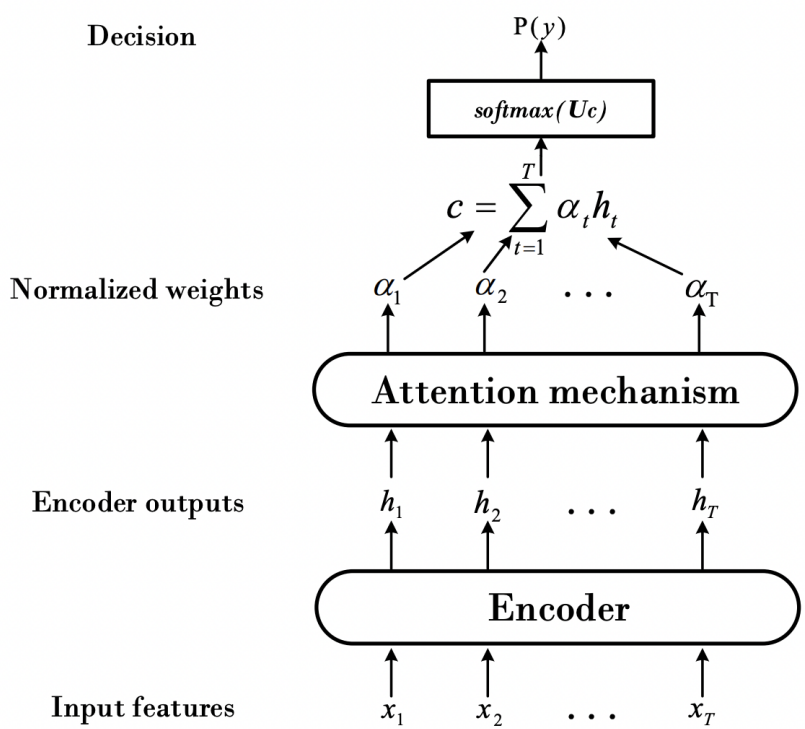


Figure: CNN architecture

- 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 $p \times q$

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

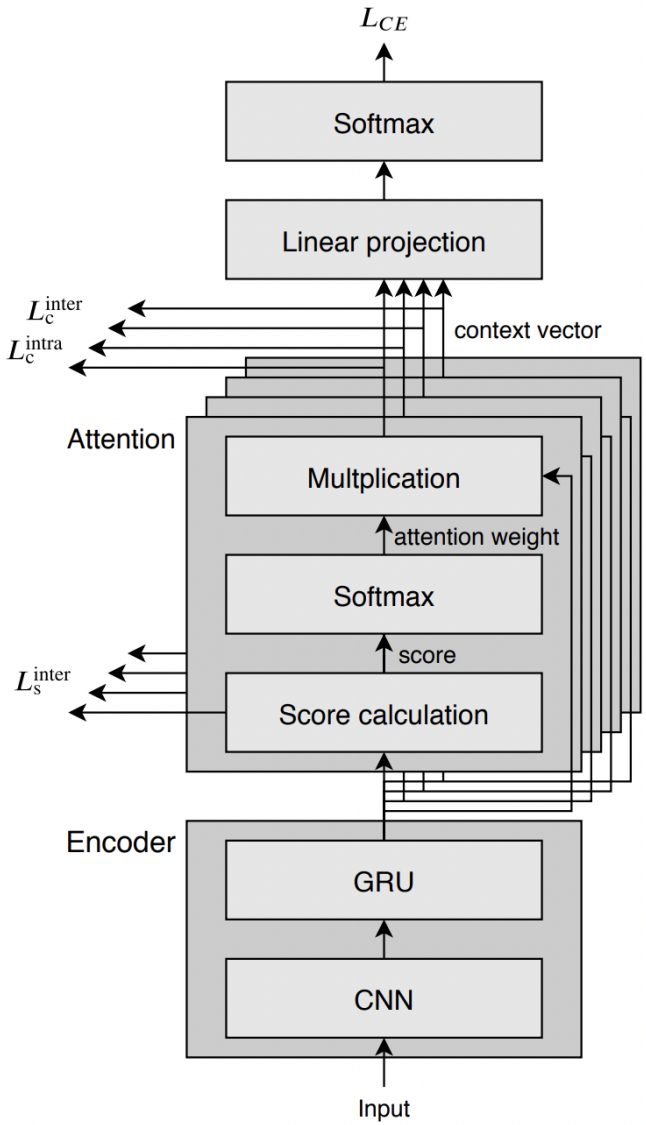
Attention-based KWS



- Encoder: RNN/GRU/LSTM/CRNN
  $\mathbf{h} = \mathit{Encoder}(\mathbf{x})$
- Soft attention:
  $e_t = v^T \tanh(\mathbf{W}\mathbf{h_t} + \mathbf{b})$.
  $\alpha_t = \frac{exp(e_t)}{\sum_{j=1}^{T} exp(e_j)}$.
  $\mathbf{c} = \sum_{\mathbf{t=1}}^{\mathbf{T}} \alpha_\mathbf{t} \mathbf{h_t}$.

Shan et al., Attention-based End-to-End Models for Small-Footprint Keyword Spotting, 2018

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 $\rightarrow$ use Orthogonality regularization

Lee et al., Orthogonality Constrained Multi-Head Attention for Keyword Spotting, IEEE Automatic Speech Recognition and Understanding Workshop, 2019

Multihead attention based KWS: orthogonality

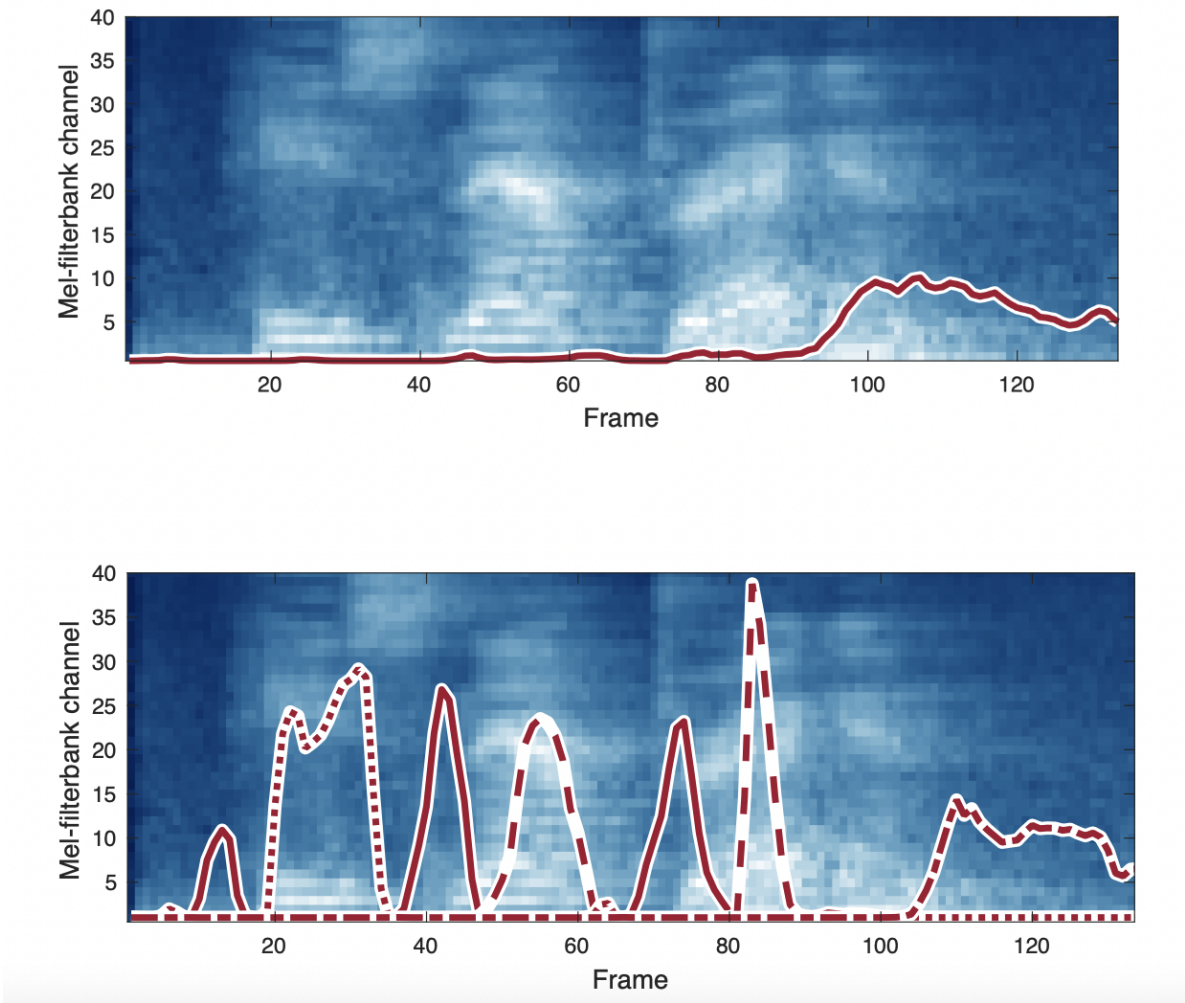


Figure: Attention weights overlaid on Mel-spectrogram with the configurations of single head attention (top), 4-head attention with regularization (bottom)

Multihead attention based KWS: orthogonality

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

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

- Inter-head orthogonality regularization

$$\mathcal{L}_{\mathrm{c}}^{\mathrm{inter}} = \frac{1}{N_{\mathrm{P}}} \sum_{n=1}^{N} \frac{y^{(n)}}{H(H-1)} \,||\, \mathbf{C}^{(n)\mathrm{T}} \mathbf{C}^{(n)} - \mathbf{I}_H \,||_F^2$$

$$\mathcal{L}_{s}^{\mathrm{inter}} = \frac{1}{N_{\mathbf{P}}} \sum_{n=1}^{N} \frac{y^{(n)}}{H(H-1)} \,||\, \mathbf{E}^{(n)\mathrm{T}} \mathbf{E}^{(n)} - \mathbf{I}_H \,||_F^2$$

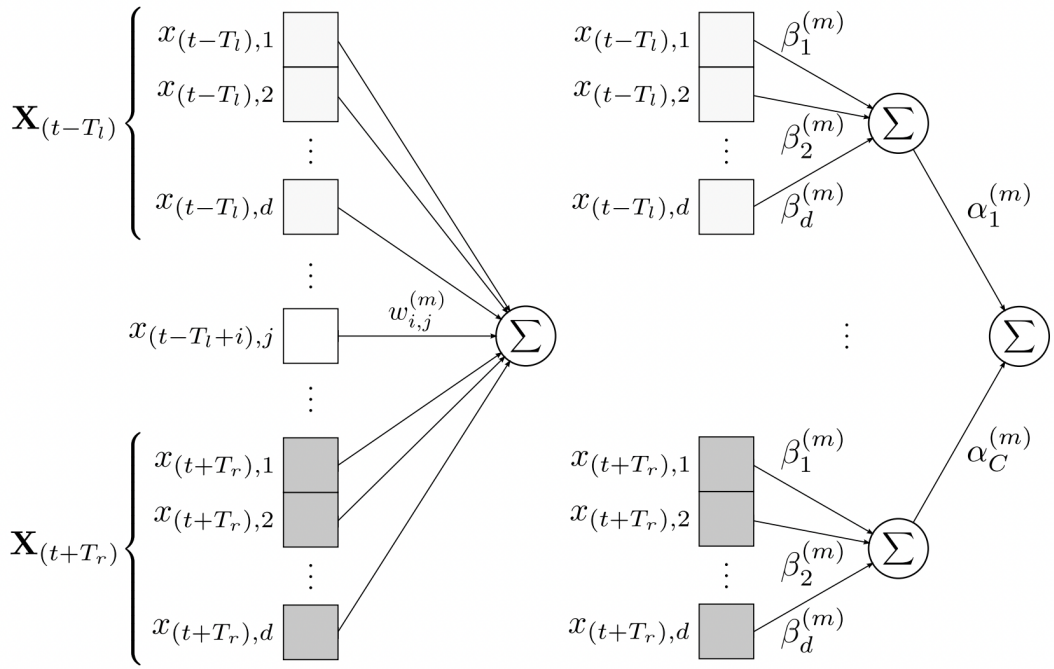
- Intra-head non-orthogonality regularization

$$\mathcal{L}_{\mathrm{c}}^{\mathrm{intra}} = \frac{1}{H} \sum_{i=1}^{H} \frac{1}{N_{\mathrm{P}}(N_{\mathrm{P}}-1)} \left\| \mathbf{Y}(\widetilde{\mathbf{C}}_i^{\mathrm{T}} \widetilde{\mathbf{C}}_i - \mathbf{I}_N)\mathbf{Y} \right\|_{\mathrm{F}}^2$$

- Cross entropy loss with the regularization terms

$$\theta^* = \arg\min\{\mathcal{L}_{\mathrm{CE}} + \lambda_1 \mathcal{L}_{\mathrm{c}}^{\mathrm{inter}} - \lambda_2 \mathcal{L}_{\mathrm{c}}^{\mathrm{intra}} + \lambda_3 \mathcal{L}_{\mathrm{s}}^{\mathrm{inter}}\}$$

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:

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

Nakkiran, Preetum et al., Compressing deep neural networks using a rank-constrained topology, INTERSPEECH 2015

Broadcasted Residual Learning

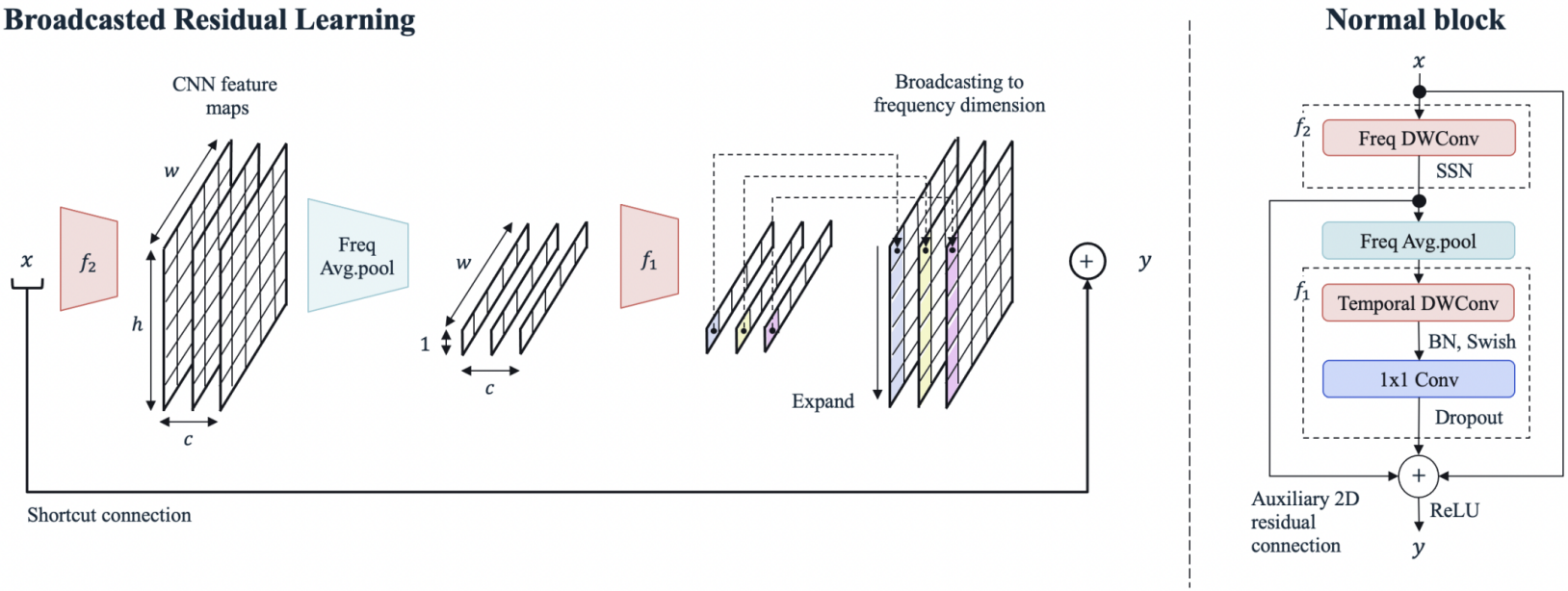


Figure: **Left, Broadcasted Residual Learning** $y = x + BC(f_1((f_2(x))))$, number of channels c.
Right, BCResBlock $y = x + f_2(x) + BC(f_1(avgpool(f_2(x))))$.

Kim, Byeonggeun et al., Broadcasted Residual Learning for Efficient Keyword Spotting., Interspeech, 2021

Convolutions in audio

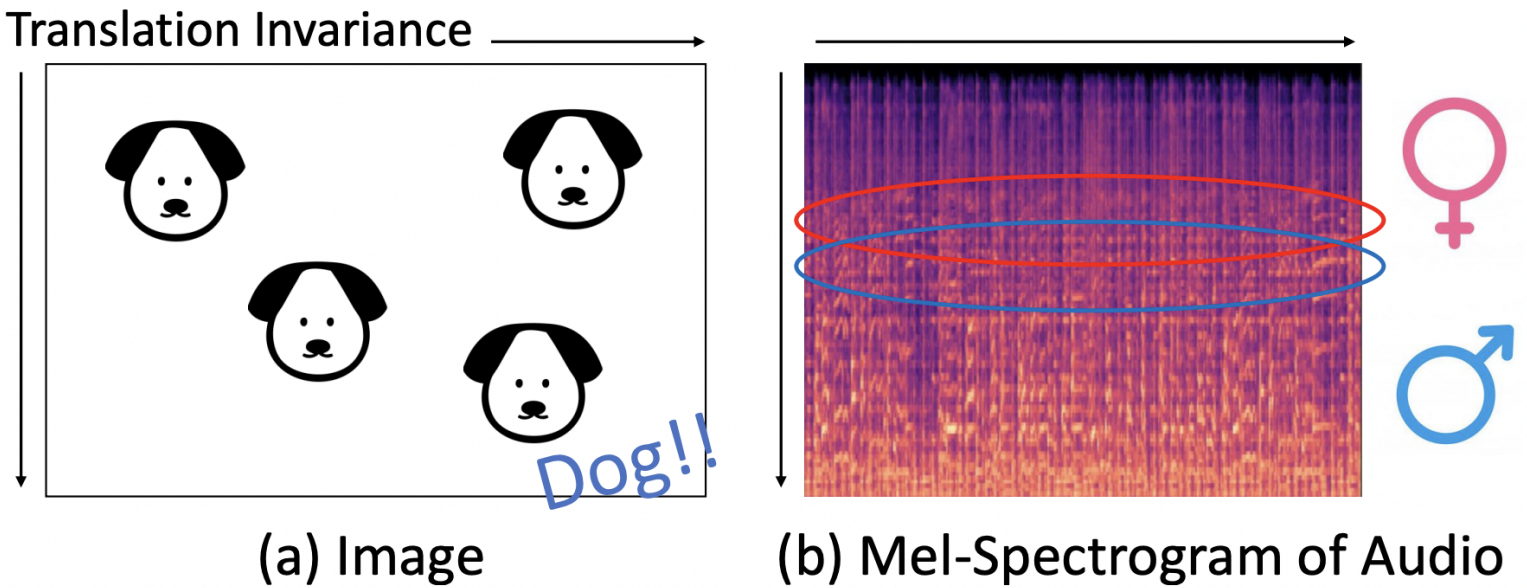


Figure: 2D convolution on image and audio input: unlike image processing, the feature in different audio frequency bands has different information.

Subspectral Normalization (SSN)

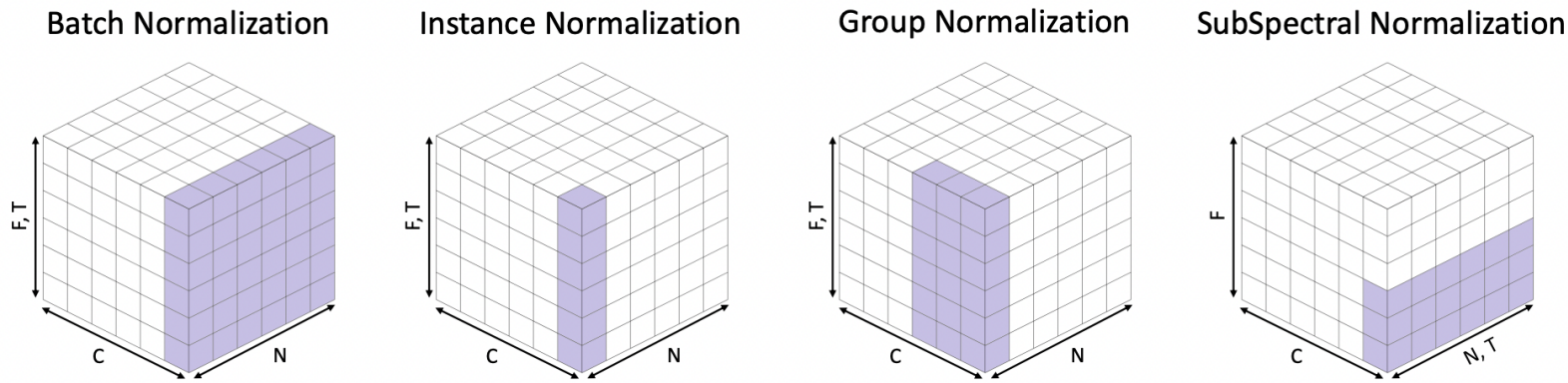


Figure: Normalization methods on Frequency-Time audio input, N – batch axis, C – channels, F – frequency, T – time axis. SSN splits the frequency dimension into multiple sub-bands and normalizes each group

Chang et al., SubSpectral Normalization for Neural Audio Data Processing, 2021

Keyword Transformer based KWS

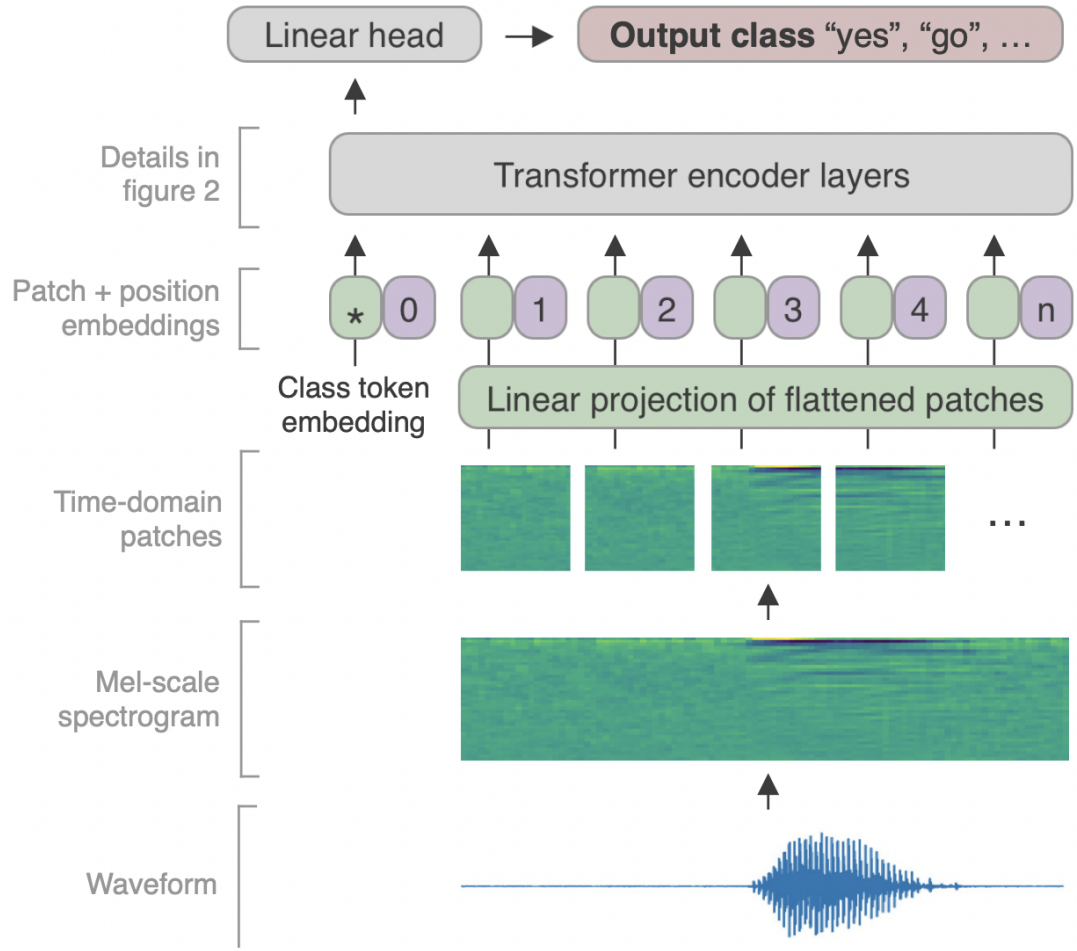


Figure: Keyword Transformer architecture

Berg et al., Keyword Transformer: A Self-Attention Model for Keyword Spotting. Interspeech, 2021

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

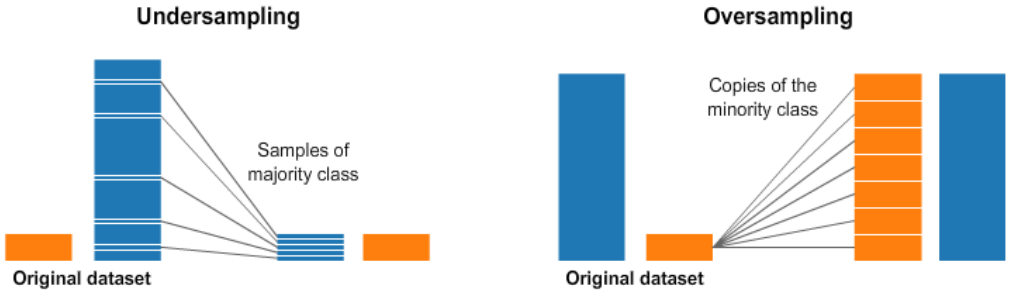
Outline

1. Keyword spotting (KWS)

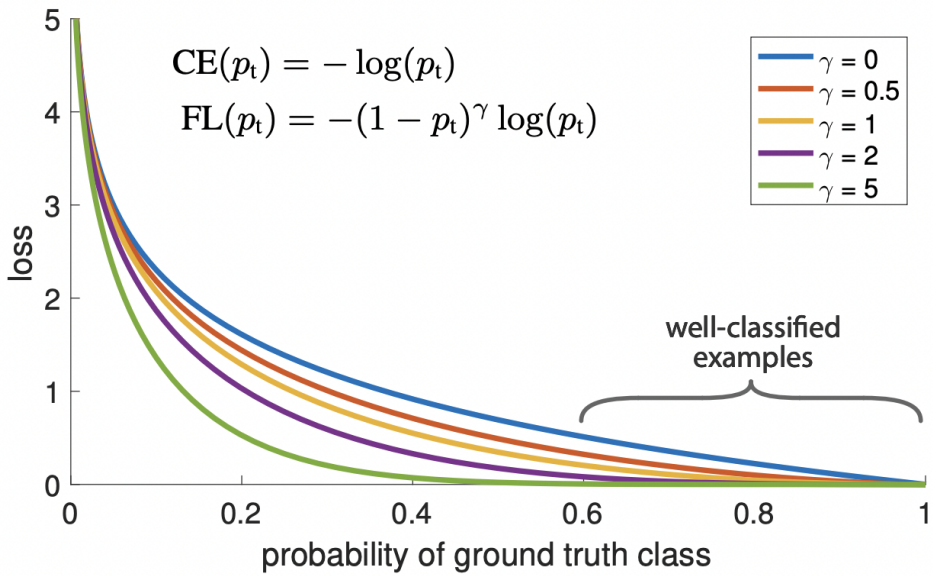
2. KWS problems and practices

Class imbalance

- Oversampling



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



Lin, Goyal et al., Focal Loss for Dense Object Detection, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, 2020

Robustness

- Multitask learning: Train ASR and KWS simultaneously

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

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