Multi-scale Octave Convolutions for Robust Speech Recognition

Joanna Rownicka, Peter Bell, Steve Renals

The Centre for Speech Technology Research,
The University of Edinburgh, UK

ICASSP 2020

Summary of the paper

- We propose a **multi-scale octave convolution** layer to learn *robust* speech representations *efficiently*.

Summary of the paper

- We propose a **multi-scale octave convolution** layer to learn *robust* speech representations *efficiently*.
- We build on OctConv proposed by Chen et al. (Facebook AI) at ICCV 2019 for Computer Vision (CV).

Summary of the paper

- We propose a **multi-scale octave convolution** layer to learn *robust* speech representations *efficiently*.
- We build on OctConv proposed by Chen et al. (Facebook AI) at ICCV 2019 for Computer Vision (CV).
  - Reduce the **spatial redundancy** of the feature maps by decomposing the output of a convolutional layer into feature maps at two different spatial resolutions, one octave apart.

Summary of the paper

- We propose a **multi-scale octave convolution** layer to learn *robust* speech representations *efficiently*.
- We build on OctConv proposed by Chen et al. (Facebook AI) at ICCV 2019 for Computer Vision (CV).
  - Reduce the **spatial redundancy** of the feature maps by decomposing the output of a convolutional layer into feature maps at two different spatial resolutions, one octave apart.
  - Improves efficiency AND accuracy (for CV).

Summary of the paper

- We propose a **multi-scale octave convolution** layer to learn *robust* speech representations *efficiently*.
- We build on OctConv proposed by Chen et al. (Facebook AI) at ICCV 2019 for Computer Vision (CV).
  - Reduce the **spatial redundancy** of the feature maps by decomposing the output of a convolutional layer into feature maps at two different spatial resolutions, one octave apart.
  - Improves efficiency AND accuracy (for CV).
- Our work: Extend the octave convolution concept to *multiple resolution groups and multiple octaves for speech recognition*.

Motivation

- low resolution processing path increases the size of the **receptive field** in the original input space

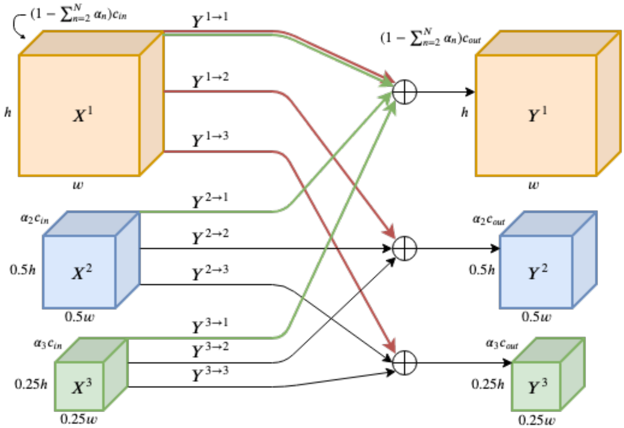
Motivation

- low resolution processing path increases the size of the **receptive field** in the original input space
- spatial average pooling in a low resolution group can be interpreted as a **low-pass filter** providing smoothed speech representations – potentially useful for noisy speech

Motivation

- low resolution processing path increases the size of the **receptive field** in the original input space
- spatial average pooling in a low resolution group can be interpreted as a **low-pass filter** providing smoothed speech representations – potentially useful for noisy speech
- enables to model the **information changing at different rates** (e.g. the characteristics of the speaker or background noise and the information necessary for phonetic discrimination)

MultiOctConv



Example of a MultiOctConv layer with 3 resolution groups.

Implementation

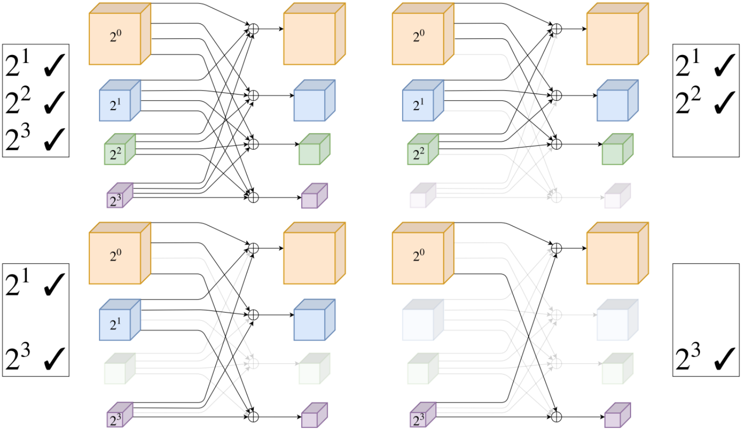
- upsampling → bilinear interpolation
- downsampling → 2D average pooling

$$Y_1 = f(X^1; W^{1\to1}) + \texttt{upsample}(f(X^2; W^{2\to1}), 2) + \texttt{upsample}(f(X^3; W^{3\to1}), 4)$$

$$Y_2 = f(X^2; W^{2\to2}) + \texttt{upsample}(f(X^3; W^{3\to2}), 2) + f(\texttt{downsample}(X^1, 2); W^{1\to2})$$

$$Y_3 = f(X^3; W^{3\to3}) + f(\texttt{downsample}(X^1, 4); W^{1\to3}) + f(\texttt{downsample}(X^2, 2); W^{2\to3})$$

MultiOctConv versions



Results: Aurora-4

Model	OctConv	2^1	2^2	2^3	A	B	C	D	Avg.
CNN	-	-	-	-	2.19	4.68	4.22	14.53	8.69
OctCNN	L2-L15	✓	-	-	2.02	4.65	4.35	14.16	8.52
OctCNN †	L2-L15	✓	-	-	2.22	4.82	4.22	13.72	8.41
MultiOctCNN	L2-L15	✓	✓	-	**1.98**	4.51	4.11	14.00	8.37
MultiOctCNN	L2-L15	✓	-	✓	2.02	4.59	**3.92**	13.82	**8.31**
MultiOctCNN	L2-L15	✓	✓	✓	2.30	4.88	4.18	14.06	8.58
MultiOctCNN	L2-L15	-	-	✓	2.02	**4.50**	4.17	13.87	8.32
MultiOctCNN †	L2-L15	-	-	✓	2.32	4.73	4.24	**13.57**	**8.31**

† models with batch normalization after ReLU

A – clean, B – w/ noise, C – mismatched mic., D – mismatched mic. w/ noise

Unpublished results: Aurora-4

- $\alpha_n \in [0, 1]$ is a fraction of channels allocated to each group
- Previously, $\alpha_n^{(i)} =$ *const.* for $1 \leq i \leq L$
- Now, $\alpha_n^{(i)} \neq$ *const.*
- Fraction for the low resolution group **changes gradually** across the layers

$\alpha_{low}^{(2-3)}$	$\rightarrow \alpha_{low}^{(4-6)}$	$\rightarrow \alpha_{low}^{(7-9)}$	$\rightarrow \alpha_{low}^{(10-12)}$	$\rightarrow \alpha_{low}^{(13-15)}$	WER
0.125	$\rightarrow$ 0.125	$\rightarrow$ 0.125	$\rightarrow$ 0.125	$\rightarrow$ 0.125	8.31
0.9	$\rightarrow$ 0.7	$\rightarrow$ 0.5	$\rightarrow$ 0.3	$\rightarrow$ 0.1	9.67
0.7	$\rightarrow$ 0.55	$\rightarrow$ 0.4	$\rightarrow$ 0.25	$\rightarrow$ 0.1	8.76
0.5	$\rightarrow$ 0.4	$\rightarrow$ 0.3	$\rightarrow$ 0.2	$\rightarrow$ 0.1	**8.23**

Results: AMI MDM

Model	OctConv	2^1	2^2	2^3	IHM dev	IHM eval	SDM dev	SDM eval	MDM dev	MDM eval
CNN	-	-	-	-	33.4	38.3	49.1	54.0	43.9	48.0
OctCNN	L2-L15	✓	-	-	33.0	37.7	48.9	54.0	43.7	47.7
OctCNN	L1-L15	✓	-	-	**32.5**	**37.4**	**48.2**	**53.3**	**42.9**	**47.2**
MultiOctCNN	L1-L15	✓	✓	-	32.8	38.1	48.9	53.9	43.7	47.9
MultiOctCNN	L1-L15	✓	✓	✓	33.7	38.7	49.5	54.6	44.1	48.4
MultiOctCNN ‡	L1-L15	✓	✓	✓	33.2	38.3	49.3	54.5	44.0	48.5
MultiOctCNN	L1-L15	-	-	✓	32.9	38.1	49.1	54.3	43.8	48.0

‡ model without the inter-frequency exchange paths

IHM – Individual Headset Mic.
SDM – Single Distant Mic.
MDM – Multiple Distant Mic.

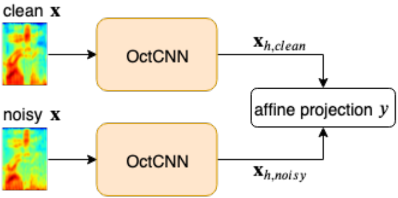
Efficiency: computational cost and memory footprint

- dependent on α, number of groups and compression rate
- with 4 groups, one octave apart (compared to a vanilla CNN)
  - **54% of computations**
  - **73% of memory**

Comparison of representations

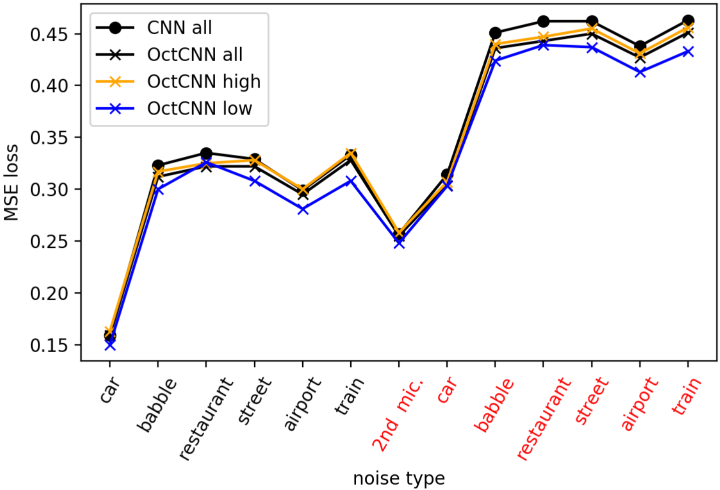
- How similar are clean and noisy hidden representations subject to an affine transformation?

$$\theta^* = \arg\min_{\theta} \frac{1}{ND} \sum_{i=1}^{N} \left\| y(\mathbf{x}_{h,clean}^{(i)}, \theta) - \mathbf{x}_{h,noisy}^{(i)} \right\|^2$$



$\mathbf{x}_{h,clean}, \mathbf{x}_{h,noisy} \in \mathbb{R}^{1 \times D}$
N – number of samples

MSE affine transformation loss



Conclusions

- Multi-scale octave CNN models for robust and efficient speech recognition

Conclusions

- Multi-scale octave CNN models for robust and efficient speech recognition
  - multiple resolution groups with a spatial reduction of more than one octave **improve the recognition**

Conclusions

- Multi-scale octave CNN models for robust and efficient speech recognition
  - multiple resolution groups with a spatial reduction of more than one octave **improve the recognition**
  - it is also more **computationally and memory efficient**

Conclusions

- Multi-scale octave CNN models for robust and efficient speech recognition
  - multiple resolution groups with a spatial reduction of more than one octave **improve the recognition**
  - it is also more **computationally and memory efficient**
  - MultiOctCNNs are the most beneficial for speech with **background noise**

Conclusions

- Multi-scale octave CNN models for robust and efficient speech recognition
  - multiple resolution groups with a spatial reduction of more than one octave **improve the recognition**
  - it is also more **computationally and memory efficient**
  - MultiOctCNNs are the most beneficial for speech with **background noise**
  - OctConv applied to the input might help with **reverberation**

Conclusions

- Multi-scale octave CNN models for robust and efficient speech recognition
  - multiple resolution groups with a spatial reduction of more than one octave **improve the recognition**
  - it is also more **computationally and memory efficient**
  - MultiOctCNNs are the most beneficial for speech with **background noise**
  - OctConv applied to the input might help with **reverberation**
- MSE affine transfromation loss as a proxy robustness measure

Conclusions

- Multi-scale octave CNN models for robust and efficient speech recognition
  - multiple resolution groups with a spatial reduction of more than one octave **improve the recognition**
  - it is also more **computationally and memory efficient**
  - MultiOctCNNs are the most beneficial for speech with **background noise**
  - OctConv applied to the input might help with **reverberation**
- MSE affine transfromation loss as a proxy robustness measure
  - OctConv design enables for **robust representation learning** especially for speech with additive noise

Thank you for your attention!

Contact: j.m.rownicka@sms.ed.ac.uk

This work was supported by a PhD studentship from the DataLab Innovation Centre, Ericsson Media Services, and Quorate Technology.