JASPER DERIVATIVES

UCS749: SPEECH PROCESSING AND SYNTHESIS

Raghav B. Venkataramaiyer

CSED TIET Patiala India.

November 19, 2024

OUTLINE

METADATA

KEYWORDS Computer Science - Computation and Language, Computer Science - Machine Learning, Computer Science - Sound, Electrical Engineering and Systems Science - Audio and Speech Processing

INCLUDES QuartzNet, CitriNet & MatchboxNet

OUTLINE

JASPER

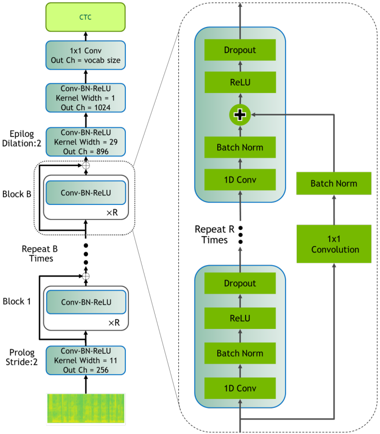


FIGURE: Jasper $B \times R$ model: B: number of blocks; R: number of sub-blocks.

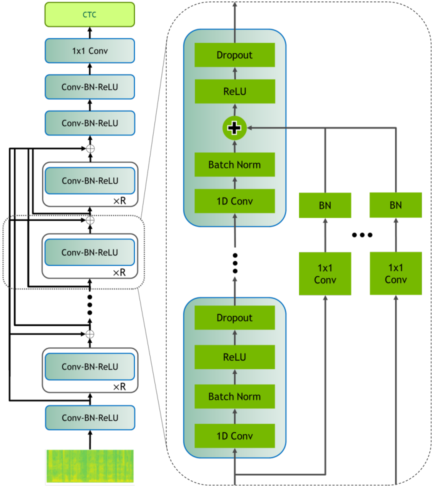


FIGURE: Jasper Dense Residual Model

OUTLINE

QUARTZNET

solves the same problem as jasper

QUARTZNET ARCHITECTURE

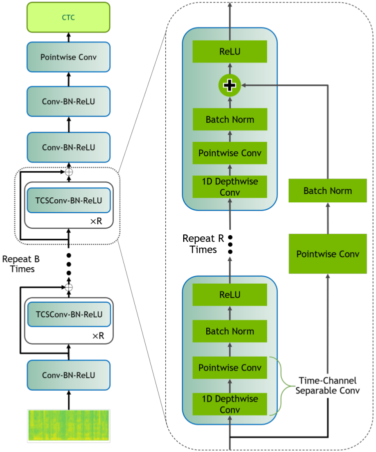
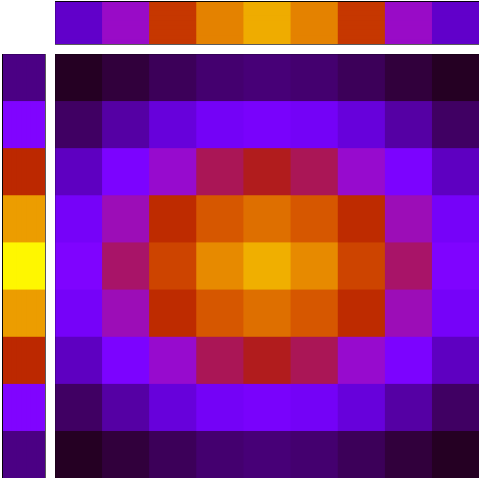


FIGURE: Quartznet Architecture

- Uses separable conv instead of conv.
- 1D Conv (along time) → 1D Conv (along frequency) → Batch Norm → ReLU instead of
- 1D Conv → Batch Norm → ReLU → Dropout.

SEPARABLE FILTERS



9×9 separable Gaussian filter with constituents as

- $\mathcal{N}(x; 0, 0.3)$ along the x-axis; and
- $\mathcal{N}(y; 0, 0.15)$ along the y-axis;

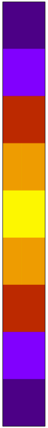
each resolved within limits $[-1, 1]$ into 9 discrete units.

SEPARABLE FILTERS

$$F_x = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}$$

$$\mu_x = 5 \qquad \sigma_x = \frac{4}{3}$$

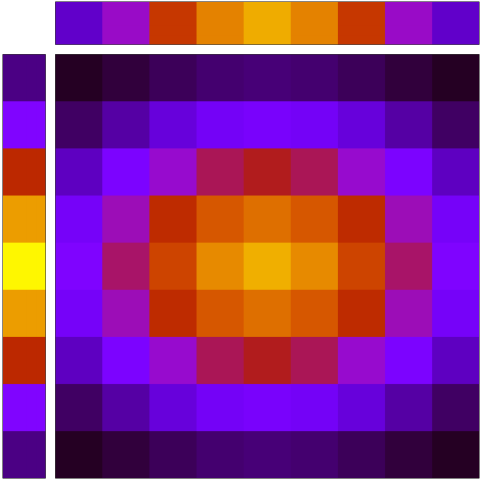
SEPARABLE FILTERS



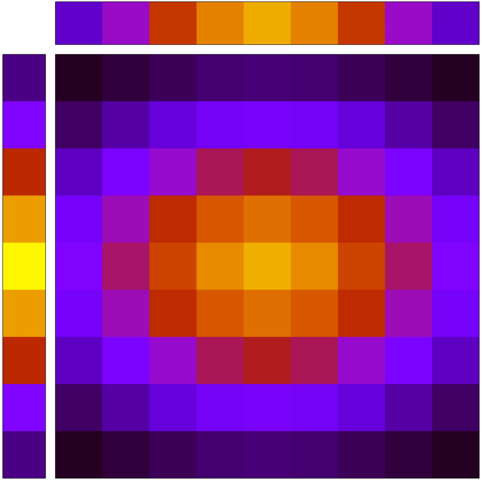
$$G_y = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}$$

$$\mu_y = 5 \qquad \sigma_y = \frac{2}{3}$$

SEPARABLE FILTERS



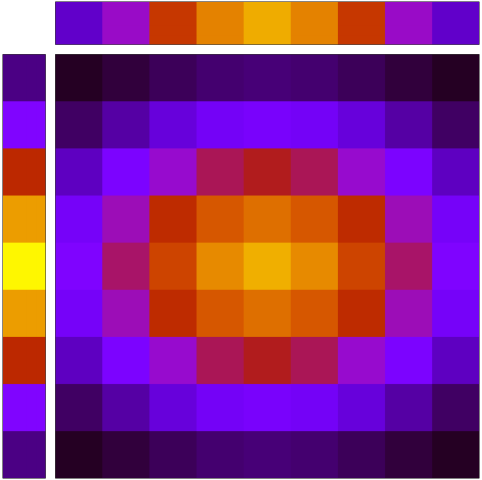
$$M_{xy} = F_x \times G_y$$



If F and G constitute a separable filter M so that $M(x, y) = F(x) \times G(y)$, the following equivalence for convolution $\otimes$ over a given signal X, holds true,

$$
\begin{aligned}
M \otimes X &\equiv F \otimes G \otimes X \\
&\equiv G \otimes F \otimes X
\end{aligned}
$$

LEARNABLE SEPARABLE CONV



If M is a learnable 2D filter with size $k \times k$; the num params is k^2

If M is also separable into F and G, then, for a given signal X,
$M \otimes X \equiv F \otimes G \otimes X$,
with F and G bearing k params each. Hence, num params becomes $2k$.

SEPARABILITY IN QUARTZNET

In Jasper, for a given convolution layer,

- Let inputs be $\mathbf{x} \in \mathbb{R}^{C_{\text{in}} \times t}$;
- Let outputs be $\mathbf{y} \in \mathbb{R}^{C_{\text{out}} \times t'}$;
- Let 1D conv filter be of size k;
- num input time steps $= t$;
- num input channels $= C_{\text{in}}$;
- num output channels $= C_{\text{out}}$;
- num params required for each channel of ouptut $= k \times C_{\text{in}}$
- num params required in total $= k \times C_{\text{in}} \times C_{\text{out}}$

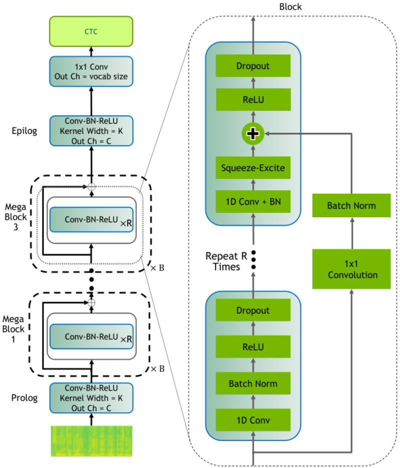
In Quartznet, the same op is implemented in 2 layers,

- $F(\mathbf{x}) \to \mathbf{y}' : \mathbf{y}' \in \mathbb{R}^{C_{\text{in}} \times t'}$
- $G(\mathbf{y}') \to \mathbf{y}$
- F convolves over the time axis, separately for each frequency (input channel); hence num params $= k \times C_{\text{in}}$
- G convolves over the frequency axis, with same kernel for each time step; hence num params $= C_{\text{in}} \times C_{\text{out}}$
- Total num params $= k \times C_{\text{in}} + C_{\text{in}} \times C_{\text{out}}$

OUTLINE

solves the same problem as jasper

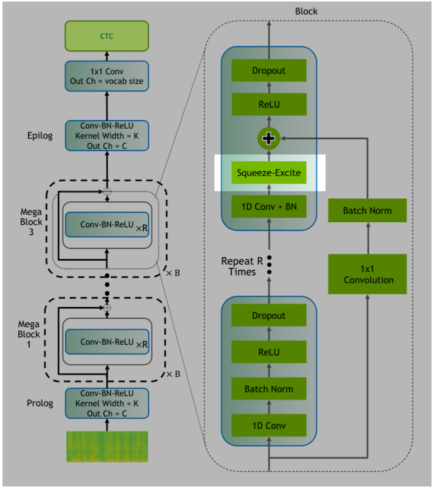
CITRINET ARCHITECTURE



Uses squeeze and excitation attention on top of *conv + bn* and before non-linear activation.

$$\mathrm{SE}\left(\mathbf{x}\right) = \mathbf{x} \otimes F_{sc} \circ F_{ex} \circ F_{sq}(\mathbf{x})$$

where, $\otimes$ is the Hadamard product (element-wise product)
Conv is Time–channel separable Conv as in Quartznet.



Uses squeeze and excitation attention on top of *conv + bn* and before non-linear activation.

$$\mathrm{SE}\,(\mathbf{x}) = \mathbf{x} \otimes F_{sc} \circ F_{ex} \circ F_{sq}(\mathbf{x})$$

where, $\otimes$ is the Hadamard product (element-wise product)
Conv is Time–channel separable Conv as in Quartznet.

$$\text{SE}\ (\mathbf{x}) = \mathbf{x} \otimes F_{sc} \circ F_{ex} \circ F_{sq}(\mathbf{x})$$

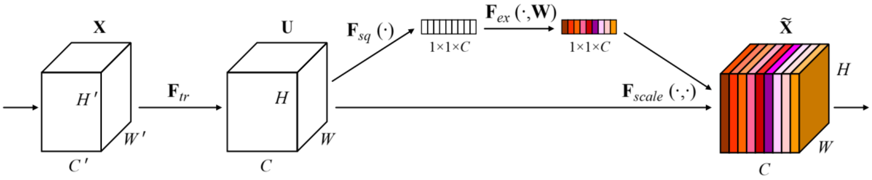


Figure 1: A Squeeze-and-Excitation block.

The image is from the original squeeze and excitation paper; and shows SE in the context of 2D Conv. But recall that in the context of speech recognition we have 1D Conv; and the same concept is extended trivially.

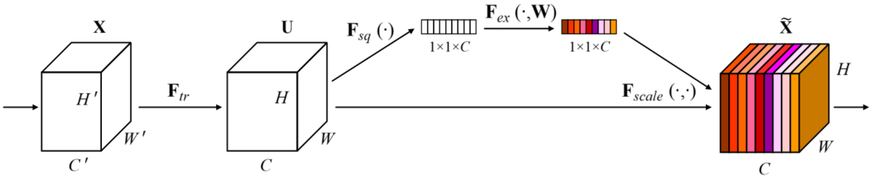


Figure 1: A Squeeze-and-Excitation block.

$F_{sq}(\mathbf{x}) = \bar{\mathbf{x}} = \frac{1}{T}\sum_t \mathbf{x}_t$ performs global average pooling; $\rightarrow 1 \times C$

$F_{ex}(\mathbf{x}) = \text{RELU}(W_1\mathbf{x} + \mathbf{b}_1)$ conv+RELU; $1 \times C \rightarrow 1 \times c$; typically $c < C$;

$F_{sc}(\mathbf{x}) = \text{sigmoid}(W_2\mathbf{x} + \mathbf{b}_2)$ conv+sigmoid; $1 \times c \rightarrow 1 \times C$;

RESULTS

Table 3: *LibriSpeech: Citrinet vs Transducers, WER(%)*

Model	LM	Test clean	Test other	Params, M
ContextNet-L [14]	-	2.10	4.60	112.7
	RNN	1.90	4.10	
Conformer-L[15]	-	2.10	4.30	118
	RNN	1.90	3.90	
Citrinet-256	-	3.78	9.6	9.8
	6-gram	3.65	8.06	
	Transf	2.75	6.87	
Citrinet-384	-	3.20	7.90	21.0
	6-gram	2.94	6.71	
	Transf	2.52	5.95	
Citrinet-512	-	3.11	7.82	36.5
	6-gram	2.40	6.08	
	Transf	2.19	5.5	
Citrinet-768	-	2.57	6.35	81
	6-gram	2.15	5.11	
	Transf	2.04	4.79	
Citrinet-1024	-	2.52	6.22	142
	6-gram	2.10	5.06	
	Transf	2.00	4.69	

Citrinet shows comparable performance with significantly lower number of params.

OUTLINE

MATCHBOXNET

- Uses a similar architecture as jasper for keyword spotting aka speech command recognition;
- Designed for devices with low computational and memory resources;
- SoTA performance with significantly fewer params.

Also,

- Fixed length input (1 second long utterance).

MATCHBOXNET ARCHITECTURE

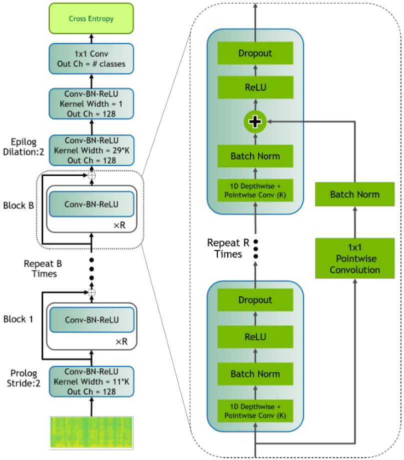


Figure 1: *MatchboxNet BxRxC model: B - number of blocks, R - number of sub-blocks, C - the number of channels.*

Table 1: *MatchboxNet-3x2x64 model has B=3 blocks, each black has R=2 time-channel separable convolutional sub-blocks with C=64 channels, plus 4 additional sub-blocks: prologue - Conv1, and epilogue - Conv2, Conv3, Conv4).*

Block	# Blocks	# Sub Blocks	# Output Channels	Kernel
Conv1	1	1	128	11
B1	1	2	64	13
B2	1	2	64	15
B3	1	2	64	17
Conv2	1	1	128	29, *dilation=2*
Conv3	1	1	128	1
Conv4	1	1	# classes	1
Soft-max				
Cross-entropy				

MATCHBOXNET ARCHITECTURE

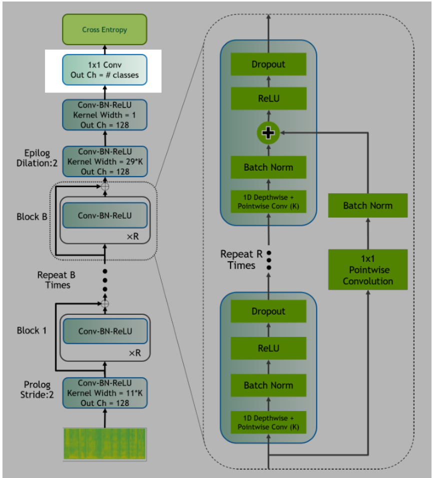


Figure 1: *MatchboxNet BxRxC model: B - number of blocks, R - number of sub-blocks, C - the number of channels.*

Table 1: *MatchboxNet-3x2x64 model has B=3 blocks, each black has R=2 time-channel separable convolutional sub-blocks with C=64 channels, plus 4 additional sub-blocks: prologue - Conv1, and epilogue - Conv2, Conv3, Conv4).*

Block	**# Blocks**	**# Sub Blocks**	**# Output Channels**	**Kernel**
Conv1	1	1	128	11
B1	1	2	64	13
B2	1	2	64	15
B3	1	2	64	17
Conv2	1	1	128	29, *dilation=2*
Conv3	1	1	128	1
Conv4	1	1	# classes	1
Soft-max				
Cross-entropy				

Table 2: *MatchboxNet on Google Speech Commands dataset v1, the accuracy is averaged over 5 trials (95% Confidence Interval).*

Model	# Parameters, K	Accuracy, %	Reference
ResNet-15	238	95.8 ± 0.351	[17]
DenseNet-BC-100	800	96.77	[32]
EdgeSpeechNet-A	107	96.80	[29]
MatchboxNet-3x1x64	77	97.21 ± 0.067	
MatchboxNet-3x2x64	93	97.48 ± 0.107	

Table 3: *MatchboxNet on Google Speech Commands dataset v2, the accuracy is averaged over 5 trials (95% Confidence Interval).*

Model	# Parameters, K	Accuracy, %	Reference
Attention RNN	202	94.30	[33]
Harmonic Tensor 2D-CNN	-	96.39	[30]
”Embedding + Head” Model	385	97.7	[31]
MatchboxNet-3x1x64	77	96.91 ± 0.101	
MatchboxNet-3x2x64	93	97.21 ± 0.072	
MatchboxNet-6x2x64	140	97.37 ± 0.110	