Jasper: An End-to-End Convolutional Neural Acoustic Model

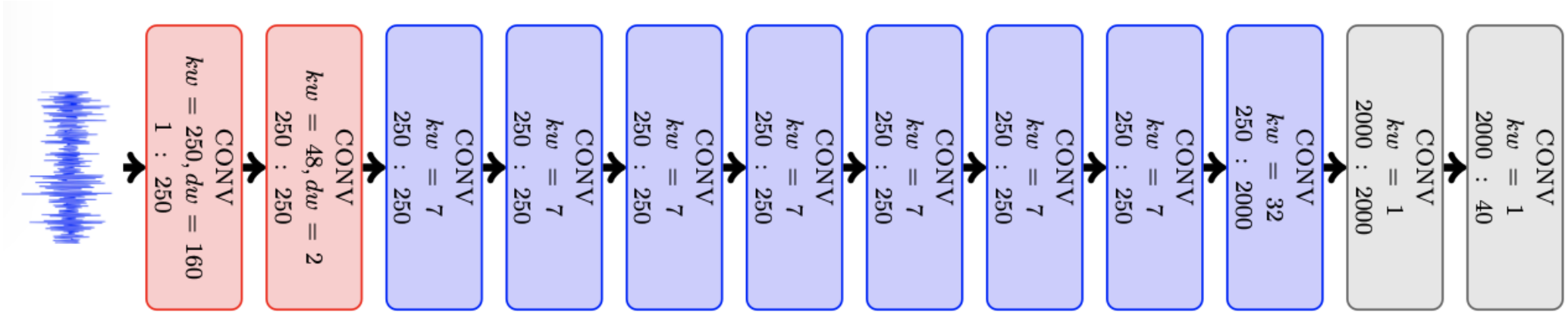
Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen Nguyen, Ravi Teja Gadde

NVIDIA, Santa Clara, USA

New York University, New York, USA

End-to-End ASR with only 1D-CNN

Collobert, R., Puhrsch, C., & Synnaeve, G. (2016). Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*.



Jasper structure

- 1D CNN
- Batch Normalization
- ReLU
- Dropout
- Residual Connection

Jasper structure

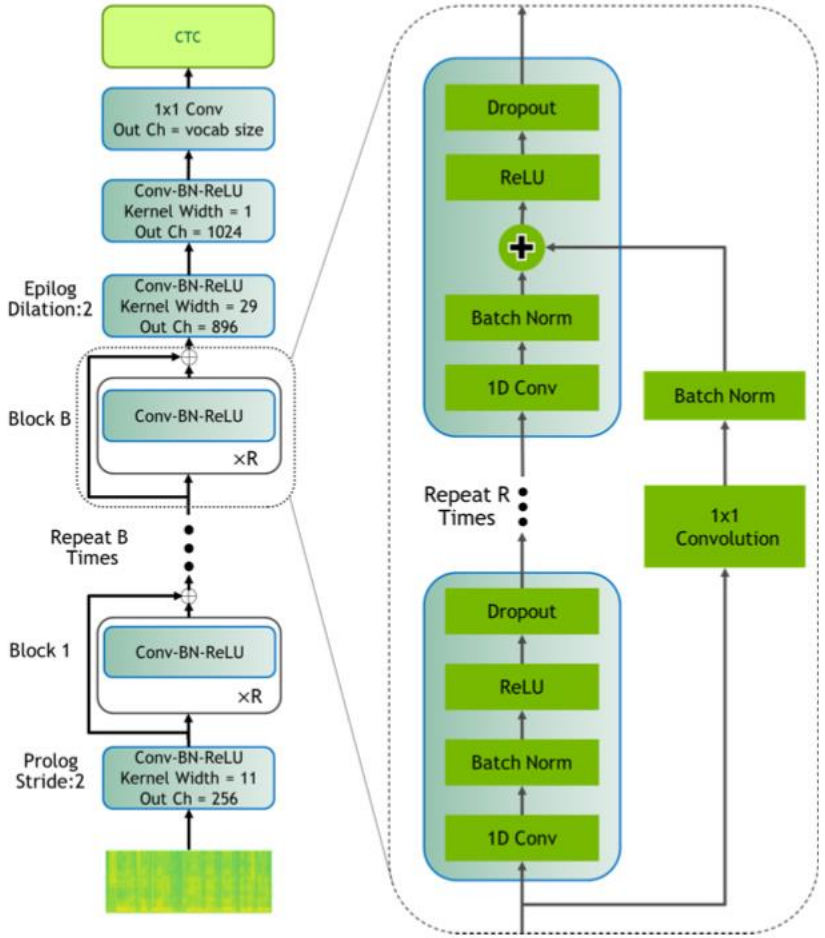


Figure 1: *Jasper BxR model: B - number of blocks, R - number of sub-blocks.*

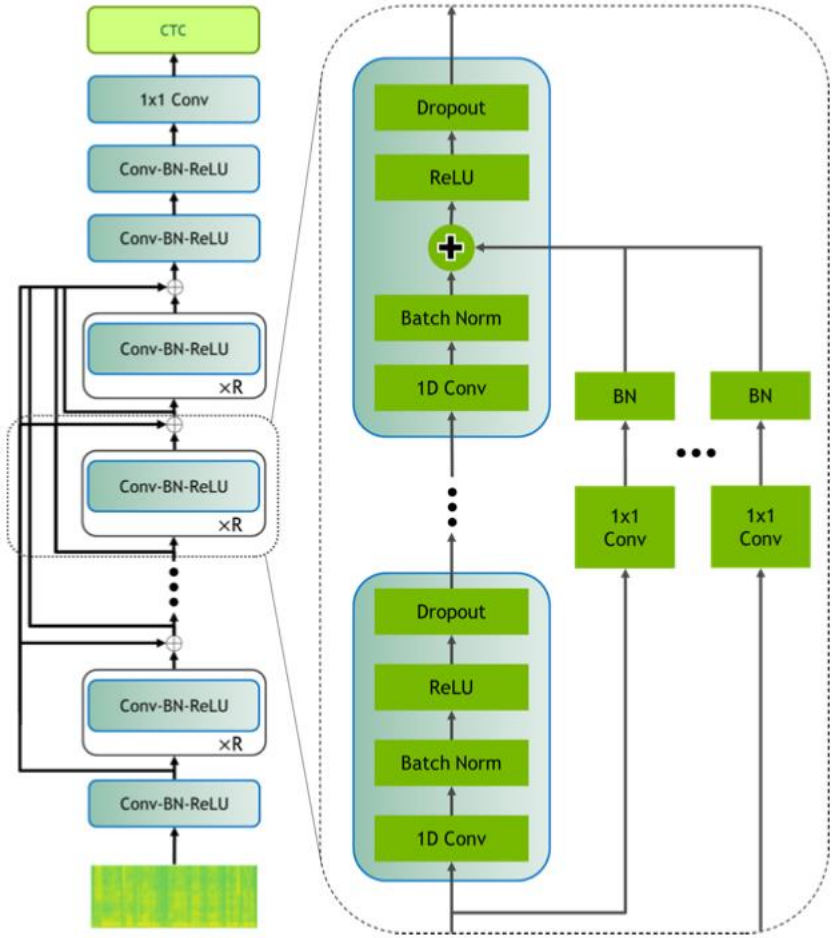
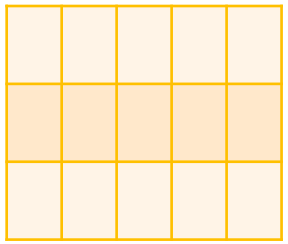
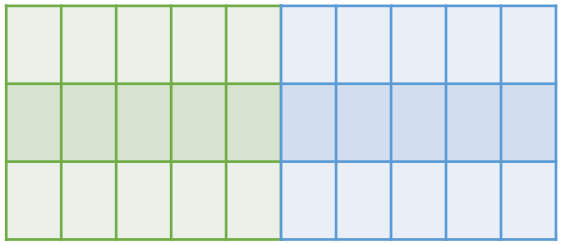


Figure 2: *Jasper Dense Residual*

Jasper Dense Residual



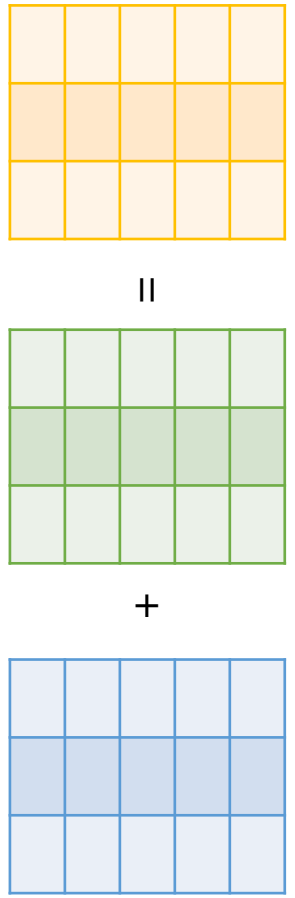
nn.Linear



current following

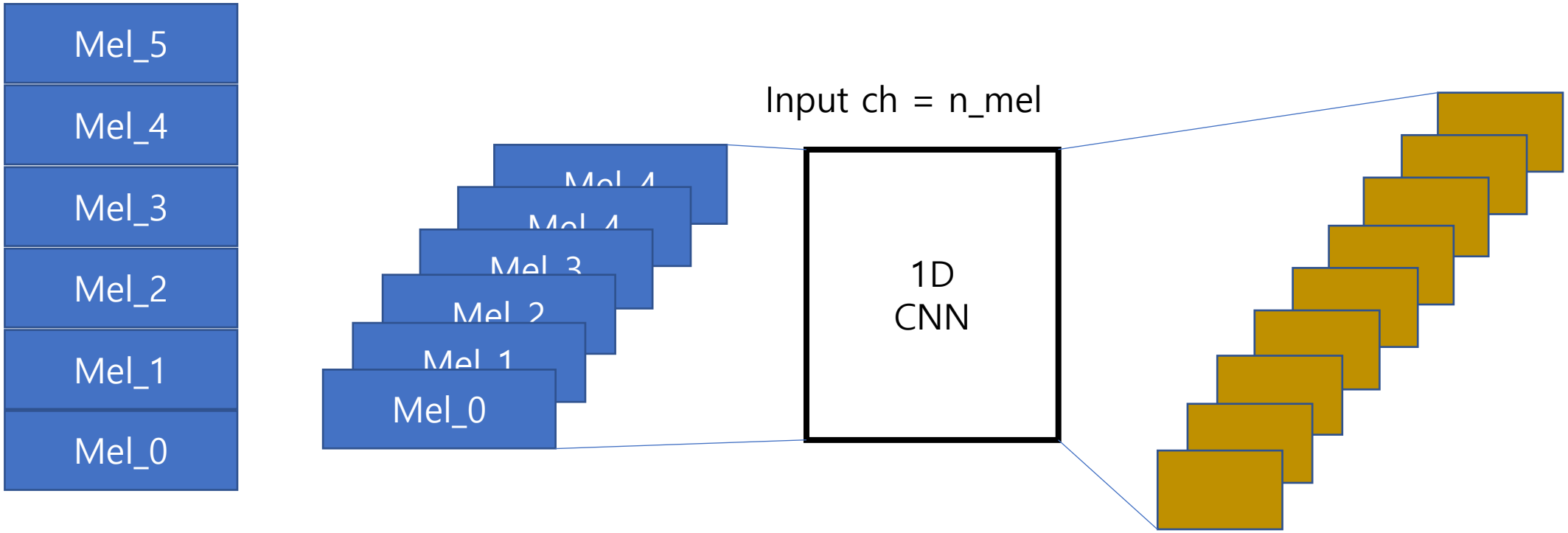
DenseNet, DenseRNet

nn.Linear



Jasper Dence Residual

End-to-End ASR with only 1D-CNN



Jasper blocks

Table 1: *Jasper 10x5: 10 blocks, each consisting of 5 1D-convolutional sub-blocks, plus 4 additional blocks.*

# Blocks	Block	Kernel	# Output Channels	Dropout	# Sub Blocks
1	Conv1	11 *stride=2*	256	0.2	1
2	B1	11	256	0.2	5
2	B2	13	384	0.2	5
2	B3	17	512	0.2	5
2	B4	21	640	0.3	5
2	B5	25	768	0.3	5
1	Conv2	29 *dilation=2*	896	0.4	1
1	Conv3	1	1024	0.4	1
1	Conv4	1	# graphemes	0	1

For pre-processing

For post-processing

Activation, Norm

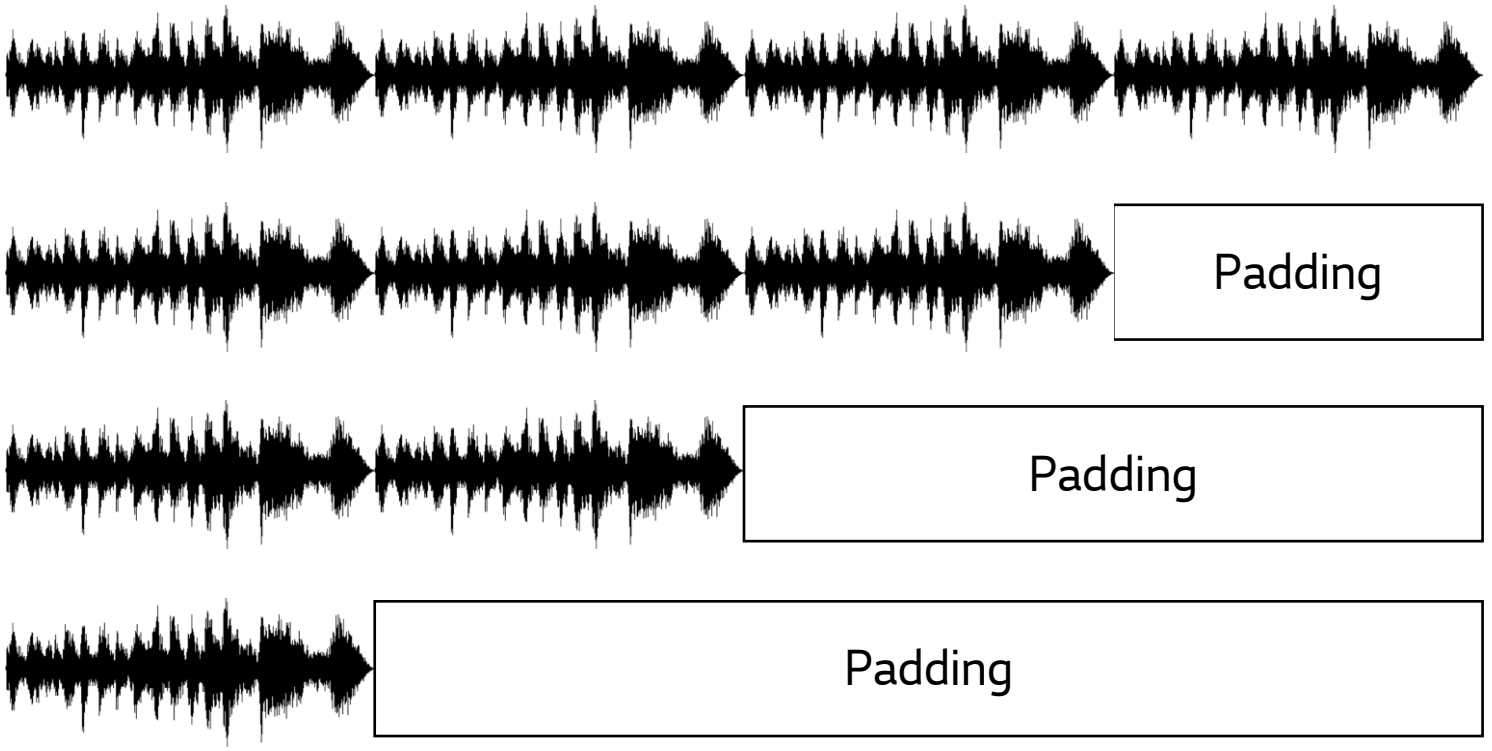
Activation

- ReLU
- cReLU
- lRelU

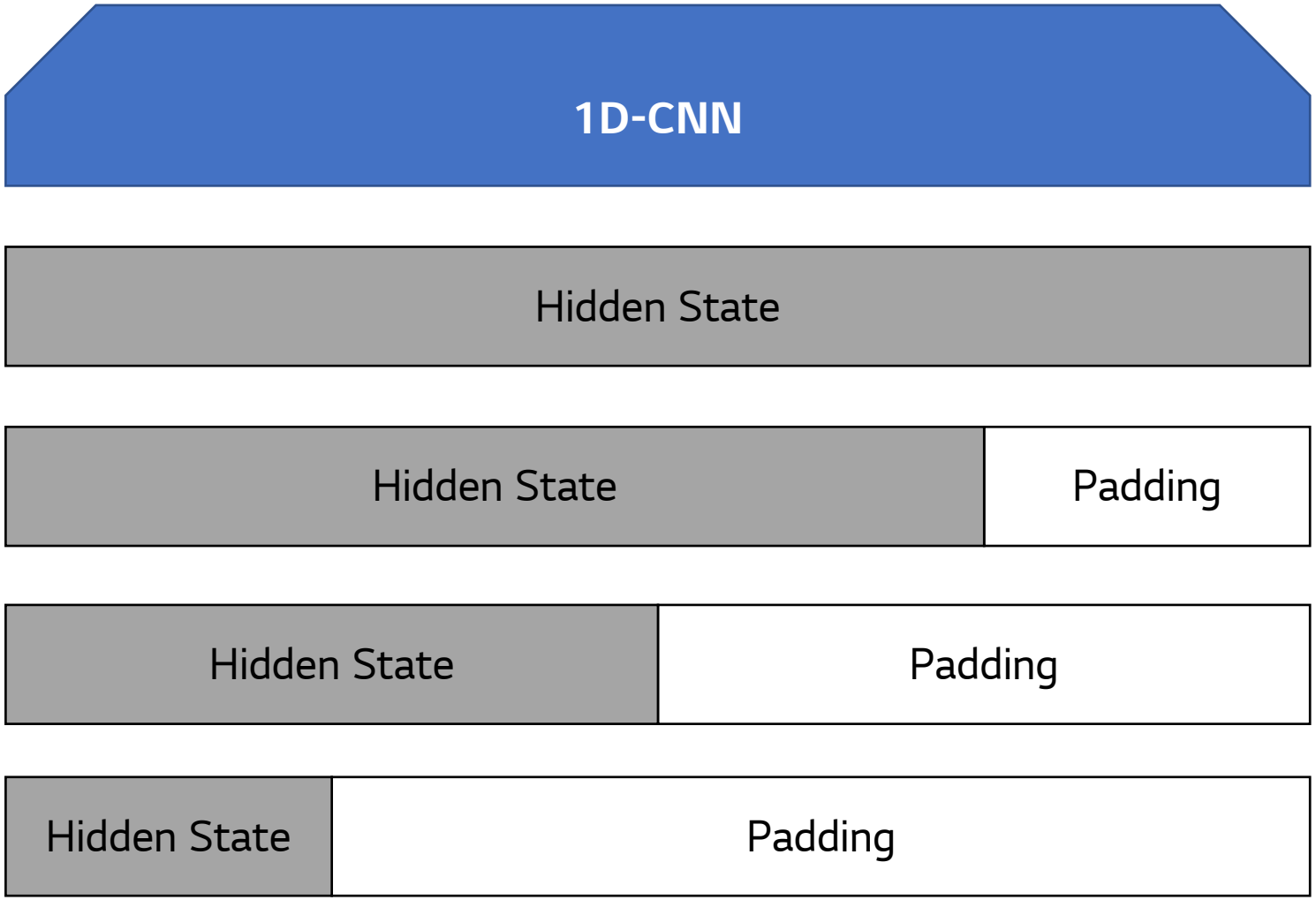
Normalization

- Weight Normalization
- Layer Normalization
- Batch Normalization

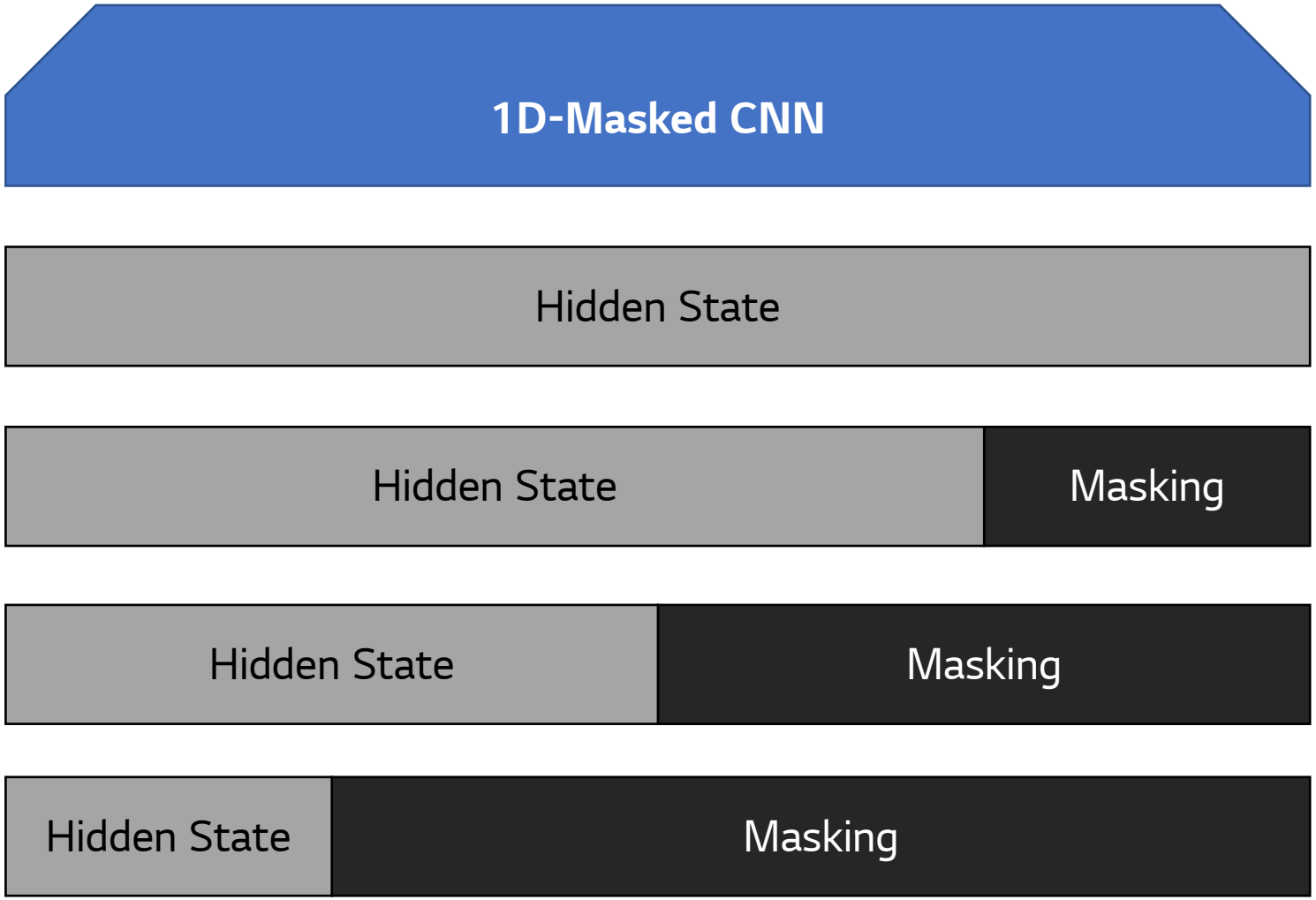
Dataloader



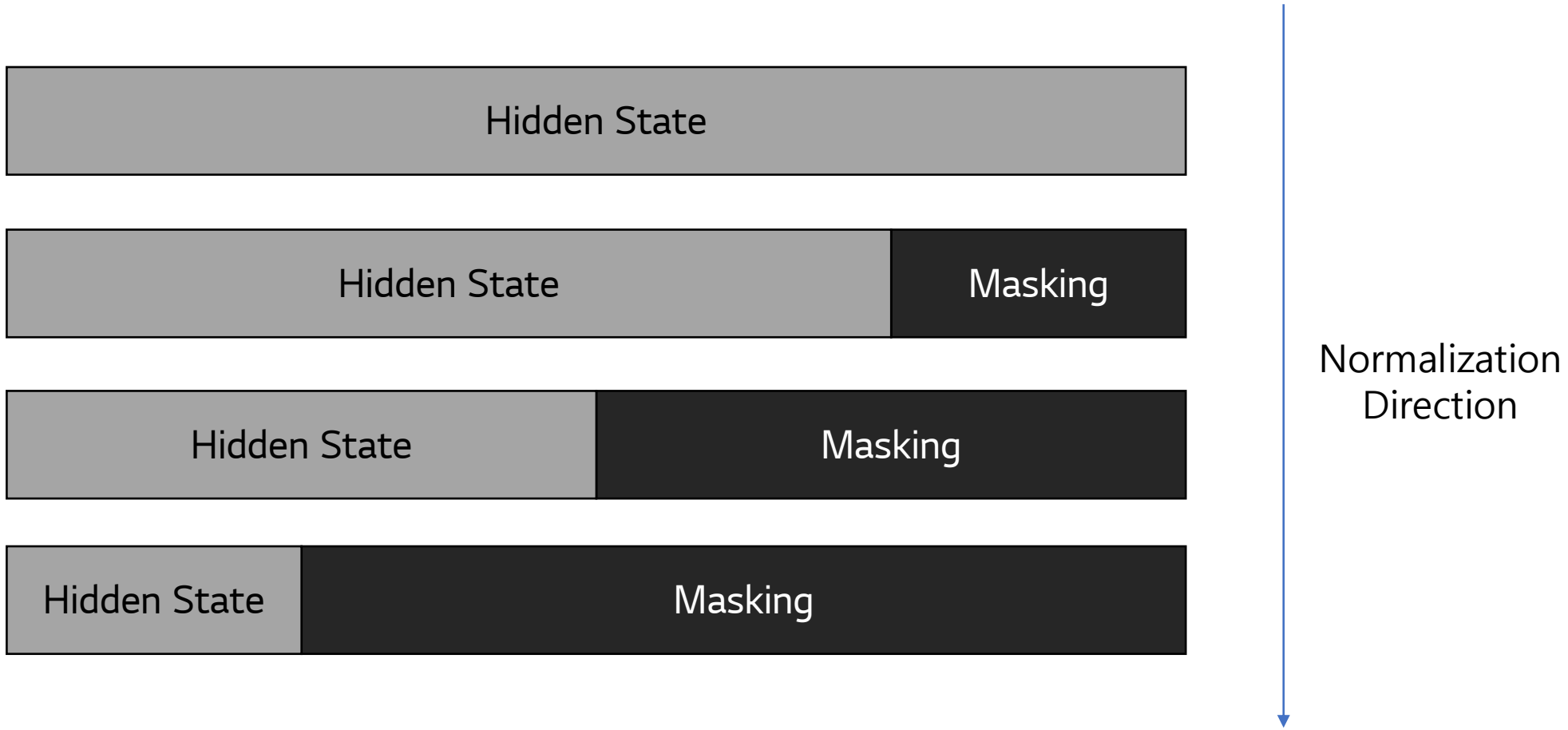
Batch padding problem



Masked CNN


1D-Masked CNN
Hidden State
Hidden State
Masking
Hidden State
Masking
Hidden State
Masking

Masked normalization



Masked conv + Masked BN

Model	Masking	Dev	
		Clean	Other
Jasper DR 10x4	None	5.88	17.62
Jasper DR 10x4	BN Mask	5.92	17.63
Jasper DR 10x4	Conv Mask	**5.66**	**16.77**
Jasper DR 10x4	Conv+BN Mask	5.80	16.97

NovoGrad

$$v_t^l = \beta_2 \cdot v_{t-1}^l + (1 - \beta_2) \cdot ||g_t^l||^2$$

$$m_t^l = \beta_1 \cdot m_{t-1}^l + \frac{g_t^l}{\sqrt{v_t^l + \epsilon}}$$

$$m_t^l = \beta_1 \cdot m_{t-1}^l + \frac{g_t^l}{\sqrt{v_t^l + \epsilon}} + d \cdot w_t$$

$$w_{t+1} = w_t - \alpha_t \cdot m_t$$

NovoGrad

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla_\theta J(\theta)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)(\nabla_\theta J(\theta))^2$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

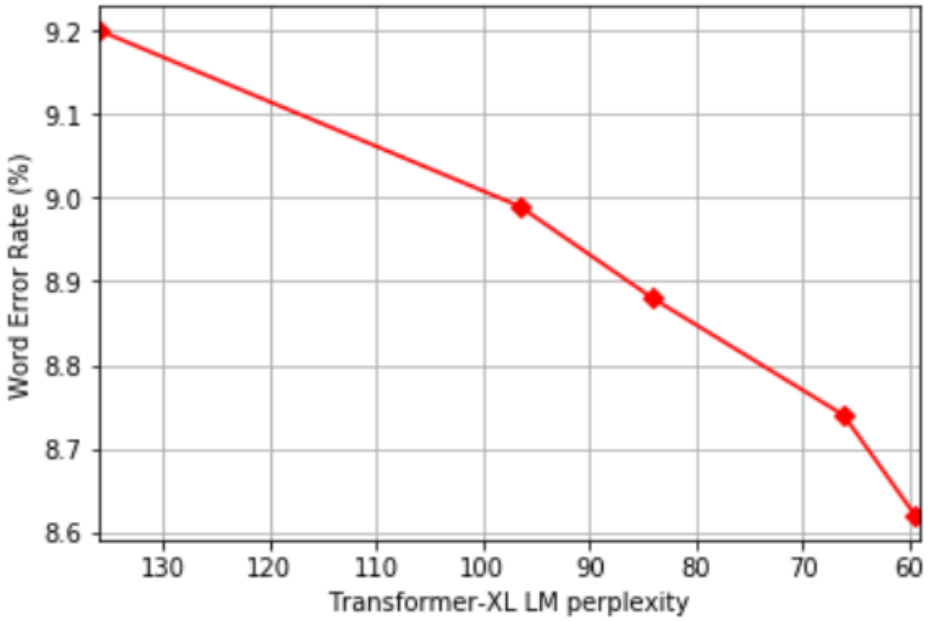
$$\theta = \theta - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t$$

Adam

Language model

Decoding : Beam search(2048) with N-gram

Rescoring : Transformer-XL



Experiments

Dataset : 2 domain

Reading :

- Librispeech (1000hr)
  model : Jasper DR 10x5
  opt : NovoGrad

- WSJ (80hr)
  model : Jasper 10x3
  opt : momentum

Conversation :

train – Fisher+Switchboard(2000hr)

eval
- Switchboard
- Call Home(CHM)

model : Japer DR 10x5
opt : SGD

Result

Table 5: *LibriSpeech, WER (%)*

Model	E2E	LM	dev-clean	dev-other	test-clean	test-other
CAPIO (single) [23]	N	RNN	3.02	8.28	3.56	8.58
pFSMN-Chain [25]	N	RNN	2.56	7.47	2.97	**7.5**
DeepSpeech2 [26]	Y	5-gram	-	-	5.33	13.25
Deep bLSTM w/ attention [21]	Y	LSTM	3.54	11.52	3.82	12.76
wav2letter++ [27]	Y	ConvLM	3.16	10.05	3.44	11.24
LAS + SpecAugment [4] [28]	Y	RNN	-	-	*2.5*	*5.8*
Jasper DR 10x5	Y	-	3.64	11.89	3.86	11.95
Jasper DR 10x5	Y	6-gram	2.89	9.53	3.34	9.62
Jasper DR 10x5	Y	Transformer-XL	2.68	8.62	**2.95**	8.79
Jasper DR 10x5 + Time/Freq Masks [4]	Y	Transformer-XL	2.62	7.61	2.84	7.84

Result

[4]We include the latest SOTA which was achieved by Park et al. [28] after our initial submission. We add results for Jasper with time and frequency masks similar to SpecAugment. We use 1 continuous time mask of size $T \sim U(0, 99)$ time steps, and 1 continuous frequency mask of size $F \sim U(0, 26)$ frequency bands.

weight decay as regularization. At training time, we use 3-fold speed perturbation with fixed +/-10% [32] for Lib-

Result

Table 6: *WSJ End-to-End Models, WER (%)*

Model	LM	nov93	nov92
seq2seq + deep conv [35]	-	-	10.5
wav2letter++ [27]	4-gram	9.5	5.6
wav2letter++ [27]	ConvLM	7.5	4.1
E2E LF-MMI [14]	3-gram	-	4.1
Jasper 10x3	-	16.1	13.3
Jasper 10x3	4-gram	9.9	7.1
Jasper 10x3	Transformer-XL	9.3	6.9

Table 7: *Hub5'00, WER (%)*

Model	E2E	LM	SWB	CHM
LF-MMI [14]	N	RNN	7.3	14.2
Attention Seq2Seq [36]	Y	-	8.3	15.5
RNN-T [37]	Y	4-gram	8.1	17.5
Char E2E LF-MMI [14]	Y	RNN	8.0	17.6
Phone E2E LF-MMI [14]	Y	RNN	7.5	14.6
CTC + Gram-CTC	Y	N-gram	7.3	14.7
Jasper DR 10x5	Y	4-gram	8.3	19.3
Jasper DR 10x5	Y	Transformer-XL	7.8	16.2

Contribution

- RNN 없이 CNN 만으로 병렬에 효율적인 E2E model
- ReLU + BN 조합이 가장 좋고 층이 깊어지면 Residual 필요하다
- 메모리를 덜 사용하는 Optimizer인 NovoGrad를 적용했다
- Libri test-clean 한정 SOTA 달성했다(첫 submit 당시)

Why accepted...?