Building Compact CNN-DBLSTM Based Character Models for HWR and OCR by Teacher-Student learning

Haisong Ding*, *Kai Chen*, Wenping Hu, Meng Cai, Qiang Huo

Microsoft Research Asia

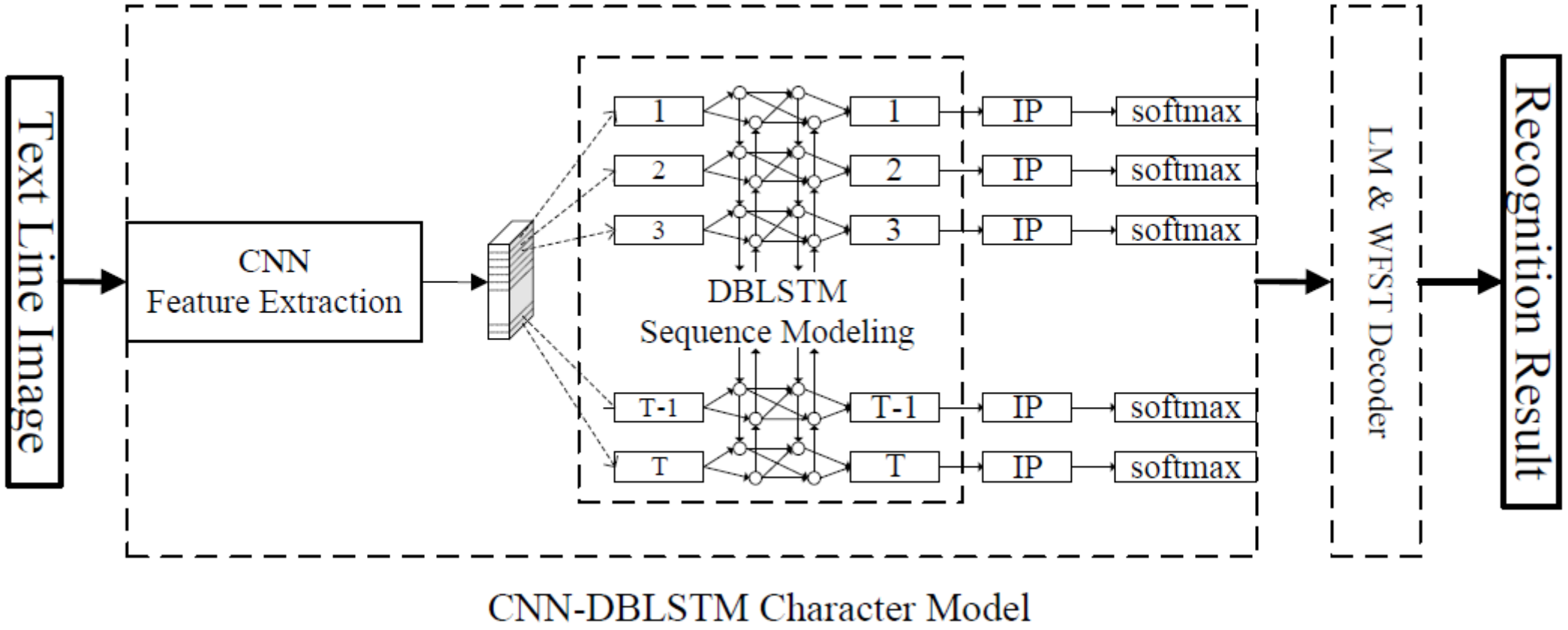
*University of Science and Technology of China

ICFHR-2018, August 2018, Niagara Falls, USA

Outline

- System Overview
- CNN Compression Method Review
- Teacher-Student Learning
- Future Work

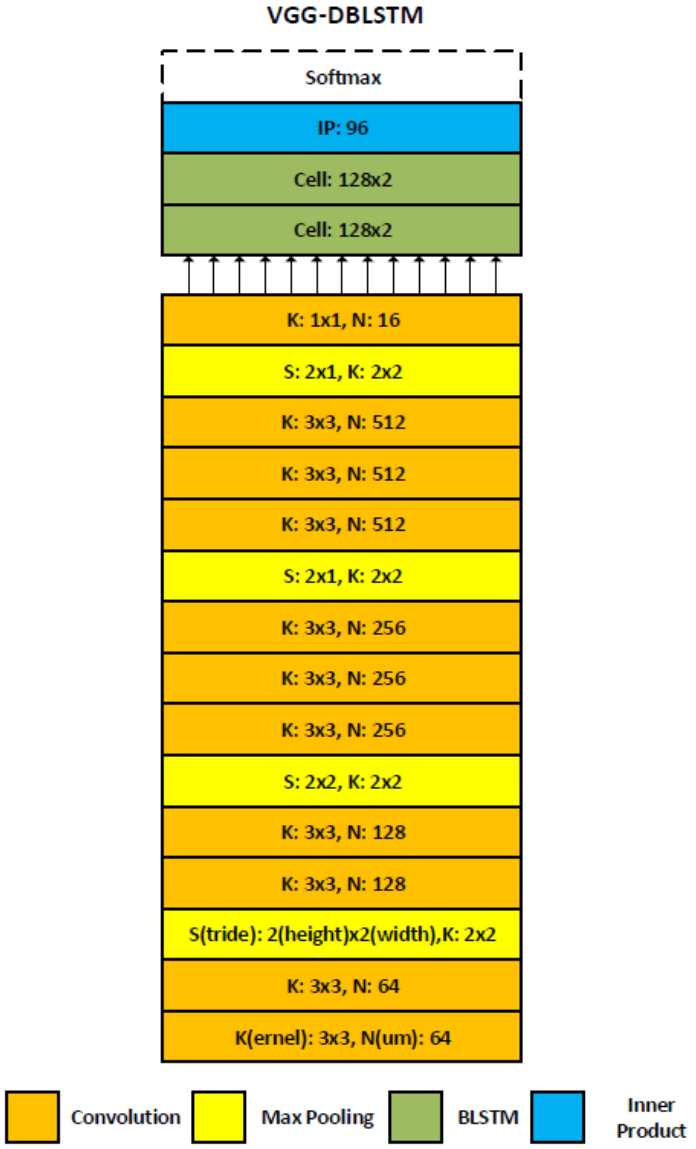
System Overview



CNN-DBLSTM Character Model

*IP: Inner-Product Layer

System Overview



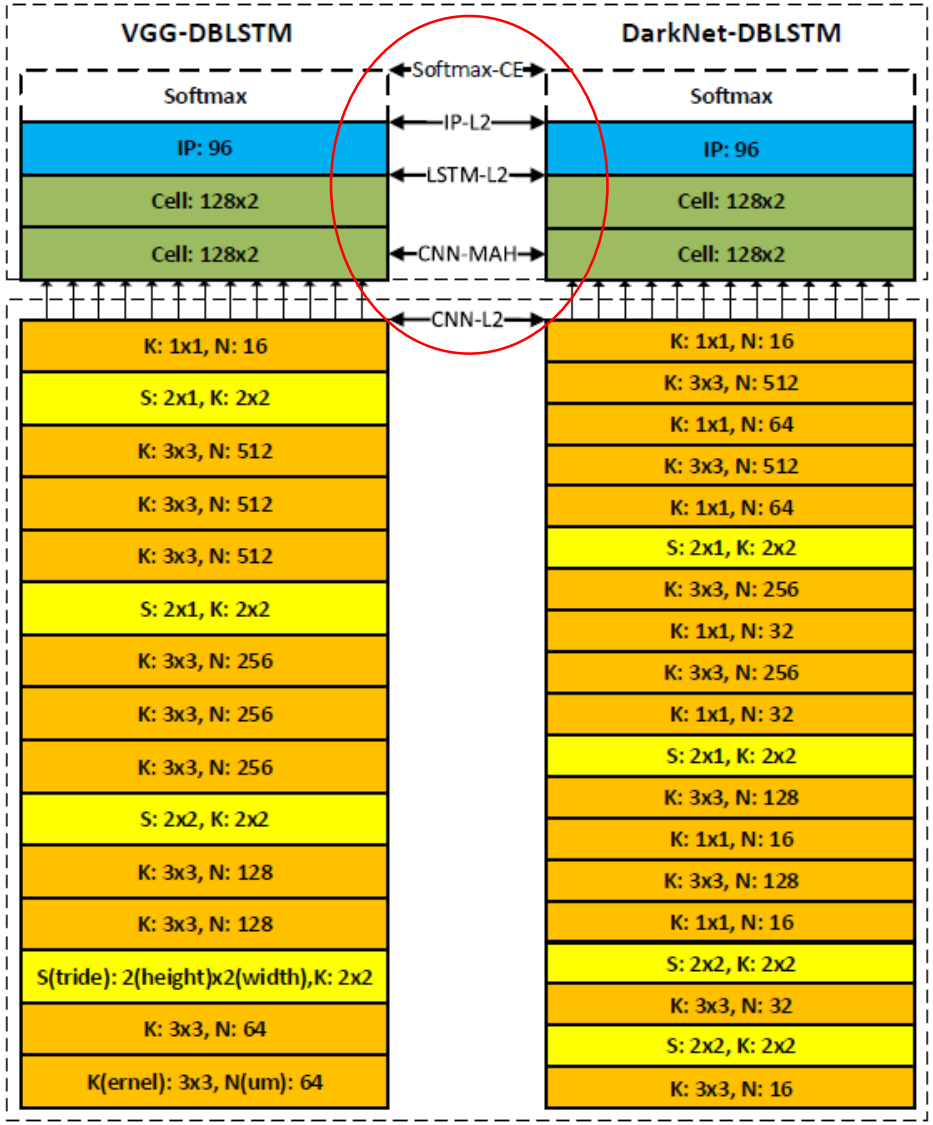
Model		Latency		Model Param.	
		(ms/line)	%	#	%
CNN	Conv3x3	197.43	97.68	7.64M	92.54
	Conv1x1	0.089	0.044	8.0e-3M	0.097
	ReLU	1.25	0.62	\	\
	MaxPooling	0.51	0.25	\	\
DBLSTM		2.84	1.41	0.62M	7.48
Total		**202.12**	100	8.26M	100

*elapsed time is evaluated on 1 Core i7-6700 CPU core

Ways to Compress CNN

- Pruning
- Quantization
- Teacher-Student Learning
- Tensor Decomposition

Teacher-Student Learning



Model construction pipeline:

- Train a VGG-DBLSTM with CTC criterion from scratch as teacher model
- Distill a DarkNet-DBLSTM using teacher-student learning with specified criterion:

Criterion	Distillation Position	Metric
Softmax-CE	Outputs of Softmax layer	cross entropy
IP-L2	Outputs of IP layer	L2 distance
LSTM-L2	Outputs of last LSTM layer	
CNN-MAH	Feedforward inputs of 1st LSTM layer	
CNN-L2	Outputs of last conv layer	

During distillation, keep LSTM and IP layers fixed.

- Fine-tune DarkNet-DBLSTM with CTC criterion to get final model.

Loss Functions (1/2)

$$\mathcal{L}^{(\text{Softmax-CE})} = -\sum_{n=1}^{N}\sum_{i=1}^{V} P_{n,i}^{(T)} \log P_{n,i}^{(S)}$$

$$\mathcal{L}^{(\text{IP-L2})} = \sum_{i=1}^{N} \|\mathbf{a}_n^{(T)} - \mathbf{a}_n^{(S)}\|_2^2$$

$$\mathcal{L}^{(\text{LSTM-L2})} = \sum_{n=1}^{N} \|\mathbf{l}_n^{(T)} - \mathbf{l}_n^{(S)}\|_2^2$$

$\{\mathbf{a}_1^{(*)}, \mathbf{a}_2^{(*)}, \cdots, \mathbf{a}_N^{(*)}\}$: output sequence of IP layer

$\{\mathbf{l}_1^{(*)}, \mathbf{l}_2^{(*)}, \cdots, \mathbf{l}_N^{(*)}\}$: output sequence of last BLSTM

$* = \mathbf{T}\text{eacher}, \mathbf{S}\text{tudent}$

$$P_n^{(T)} = \frac{\exp\left(\mathbf{a}_n^{(T)}/\tau\right)}{\sum_{i=1}^{V}\exp\left(a_{n,i}^{(T)}/\tau\right)}$$

$$P_n^{(S)} = \frac{\exp\left(\mathbf{a}_n^{(S)}/\tau\right)}{\sum_{i=1}^{V}\exp\left(a_{n,i}^{(S)}/\tau\right)}$$

τ : temperature

Loss Functions (2/2)

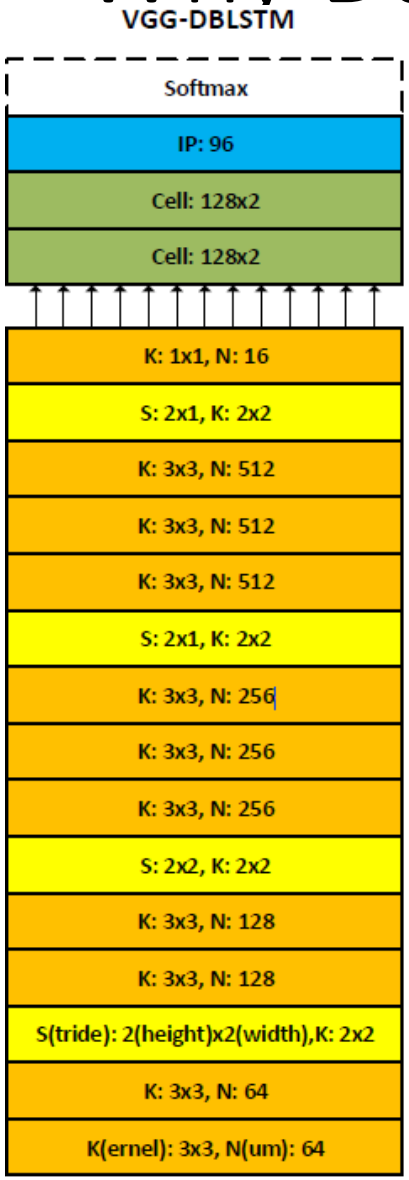
$$\mathcal{L}^{(\text{CNN-L2})} = \sum_{n=1}^{N} \|\mathbf{x}_n^{(T)} - \mathbf{x}_n^{(S)}\|_2^2$$

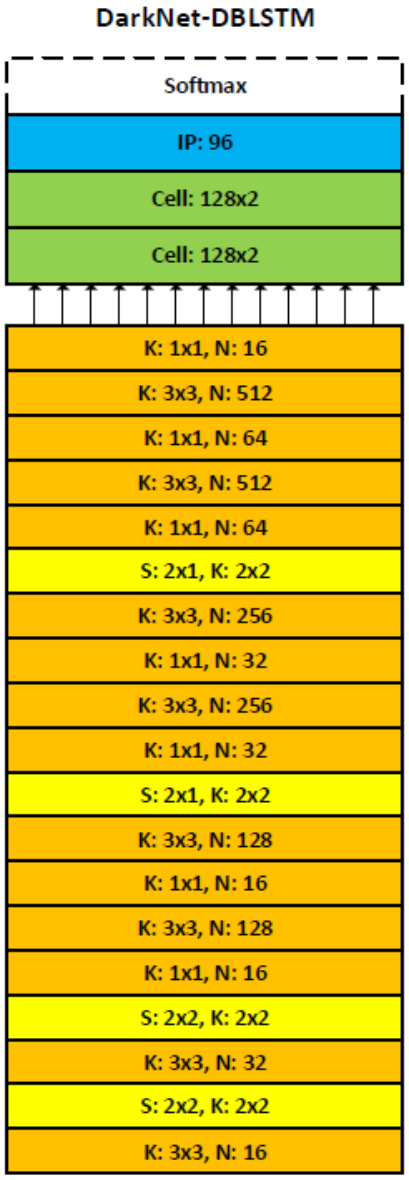
$\{\mathbf{x}_1^{(*)}, \mathbf{x}_2^{(*)}, \cdots, \mathbf{x}_N^{(*)}\}$: output sequence of CNN part

$$\begin{aligned}\mathcal{L}^{(\text{CNN-MAH})} &= \sum_{n=1}^{N} \|\mathbf{W}(x_n^{(T)} - x_n^{(S)})\|_2^2 \\ &= \sum_{n=1}^{N} (x_n^{(T)} - x_n^{(S)})^T \mathbf{W}^T \mathbf{W} (x_n^{(T)} - x_n^{(S)})\end{aligned}$$

$\mathbf{W}$: feed forward weight matrixes of first BLSTM layer

Why DarkNet?





Comparison of VGG-DBLSTM and DarkNet-DBLSTM in terms of model parameters, computation cost, and runtime latency

Model	Params		GLOPs		Runtime	
	#	Cr	#	Sr	Latency	Speedup
VGG-DBLSTM	8.26M	1.00	11.81	1.00	202.12	1.00
DarkNet-DBLSTM	1.47M	**5.62**	0.69	17.04	**14.19**	**14.24**

Cr: compression ratio
Sr: theoretical speedup ratio

Experimental Setup – HWR Task

- Training set:
  - 283k handwriting text line images extracted from whiteboard and handwritten note images
- Validation set:
  - 4k text line images
- Test set:
  - E2E: 4,028 text line images extracted from 288 whiteboard and handwritten note images
  - IAM: 1,861 text line images extracted from IAM Handwriting English Sentence dataset

Experimental Results – HWR Task

Model	Loss Function	IAM		E2E	
		CER	WER	CER	WER
VGG-DBLSTM	CTC	3.3	8.2	4.1	13.4
DarkNet-DBLSTM	CTC	3.8	9.0	4.6	15.1
DarkNet-DBLSTM (teacher-student learning)	CNN-L2	3.5	8.7	4.2	13.8
	CNN-MAH	**3.5**	**8.5**	**4.2**	**13.6**
	LSTM-L2	3.5	8.6	4.2	13.7
	IP-L2	3.7	8.7	4.3	13.9
	Softmax-CE (T=1)	3.6	8.8	4.4	14.2
	Softmax-CE (T=2)	3.7	9.0	4.4	14.1
	Softmax-CE (T=5)	3.7	9.0	4.5	14.4
	Softmax-CE (T=10)	3.8	9.1	4.5	14.5

*CER: Character Error Rate; WER: Word Error Rate

Analysis

Loss function values of student models trained with different teacher-student learning criteria on HWR task

Model	Loss function					
	$\mathcal{L}_{(\mathrm{Softmax-CE})}$	$\mathcal{L}_{(\mathrm{IP-L2})}$	$\mathcal{L}_{(\mathrm{LSTM-L2})}$	$\mathcal{L}_{(\mathrm{CNN-MAH})}$	$\mathcal{L}_{(\mathrm{CNN-L2})}$	$\mathcal{L}_{(\mathrm{CTC})}$
Softmax-CE	**0.166**	0.271	2.35e-3	19.583	0.101	10.686
IP-L2	0.196	0.0986	7.95e-4	0.455	4.14e-3	9.035
LSTM-L2	0.180	**0.0763**	**5.96e-4**	0.371	3.85e-3	**8.696**
CNN-MAH	0.183	0.0838	6.66e-4	**0.232**	**2.18e-3**	8.971
CNN-L2	0.201	0.0854	6.69e-4	0.260	2.26e-3	9.059

Comparison with Tucker Decomposition

- Tucker decomposition

  Decompose Conv3x3 to Conv1x1-Conv3x3-Conv1x1 to compress and accelerate CNN simultaneously

Teacher-student learning vs Tucker decomposition in terms of recognition accuracy (%), model parameters, GFLOPs and runtime latency

Model	IAM		E2E		Params		GFlops		Runtime	
	CER	WER	CER	WER	#	Cr	#	Sr	Latency	Speedup
VGG-DBLSTM	3.3	8.2	4.1	13.4	8.26M	1.00	11.81	1.00	202.12	1.00
DarkNet-DBLSTM	3.5	8.5	4.2	13.6	1.47M	5.62	**0.69**	**17.04**	**14.19**	**14.24**
VGG-TK-DBLSTM-v1	3.5	8.6	4.3	14.1	**0.99M**	**8.34**	0.74	15.92	26.96	7.50
VGG-TK-DBLSTM-v2	3.4	8.5	4.2	13.7	1.13M	7.31	1.05	11.17	32.46	6.23
VGG-TK-DBLSTM-v3	3.4	8.4	4.2	13.5	1.79M	4.61	2.35	5.03	60.37	3.35

* We have optimized runtime implementation after paper submission.

Experimental Setup – OCR Task

- Training Set
  - 1.06M printed text lines extracted from Open Image dataset and Microsoft street view images
- Validation Set
  - 131K printed text lines
- Test Sets
  - G-test: 55,258 text lines extracted from Open Image dataset
  - S-test: 44,823 text lines extracted from street view dataset
  - IC13: 1,094 text lines from ICDAR-2013 robust reading competition set
- Training configuration
  - Parallel training with Blockwise Model Update Filtering (BMUF) method on 8 GPU cards

Experimental Results – OCR Task

CER(%) and WER(%) of DarkNet-DBLSTM student model on OCR task

Model	G-test		S-test		IC13-test	
	CER	WER	CER	WER	CER	WER
VGG-DBLSTM	1.8	6.1	0.8	3.8	4.0	11.1
DarkNet-DBLSTM (from scratch)	2.2	7.1	1.1	4.7	4.4	13.2
DarkNet-DBLSTM (CNN-MAH)	1.8	6.2	0.8	3.8	4.0	11.4

Conclusion

- Teacher-Student learning unblocks the deployment of CNN-DBLSTM based character model.

- Guidance of LSTM layers helps to distill a better student model.

Future Work

- Compressing LSTM layers

- Designing more compact student models

Thanks!