

Rewiring the Transformer with Depth-Wise LSTMs

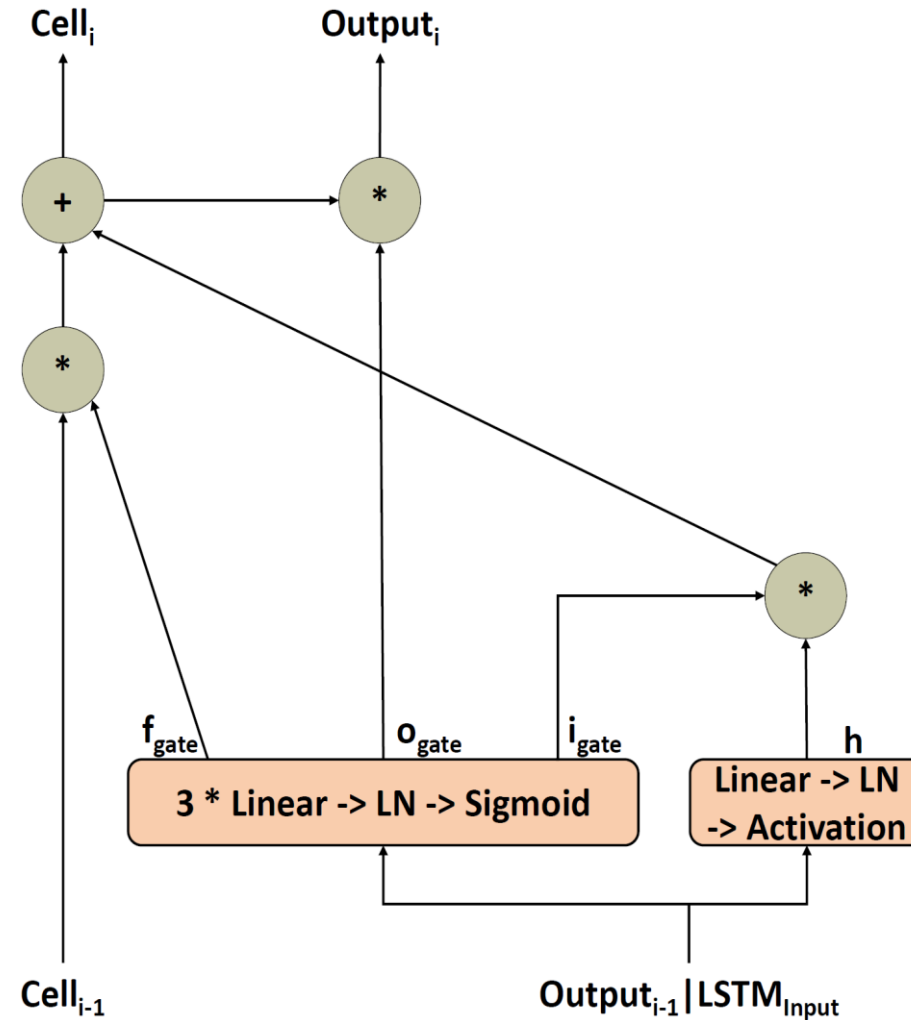
Hongfei Xu^{1,2}, Yang Song¹, Qihui Liu³, Josef van Genabith^{*2}, Deyi Xiong⁴

¹Zhengzhou University, ²DFKI and Saarland University,

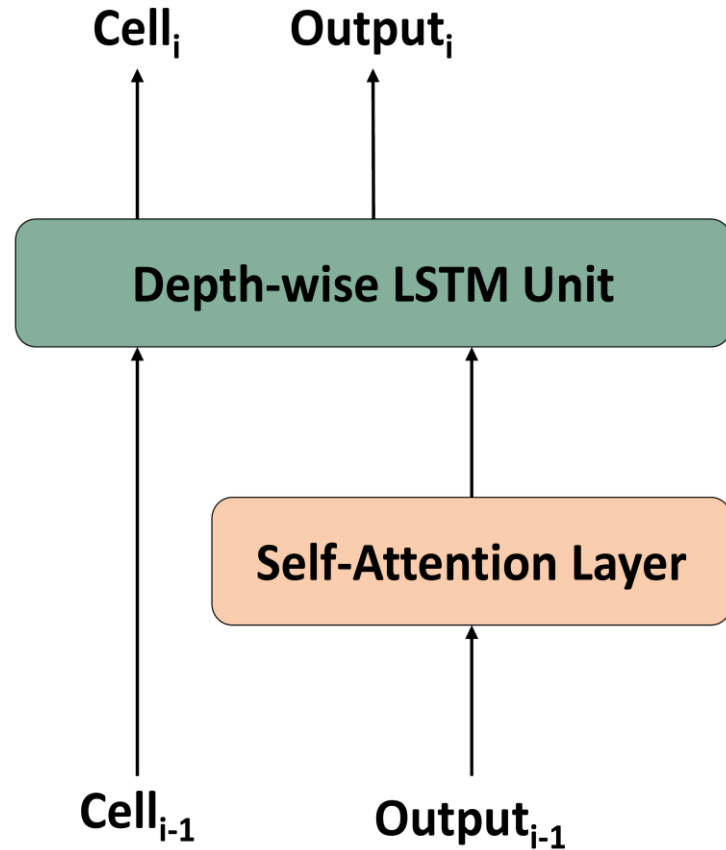
³China Mobile Online Services, ⁴Tianjin University

LREC-COLING 2024

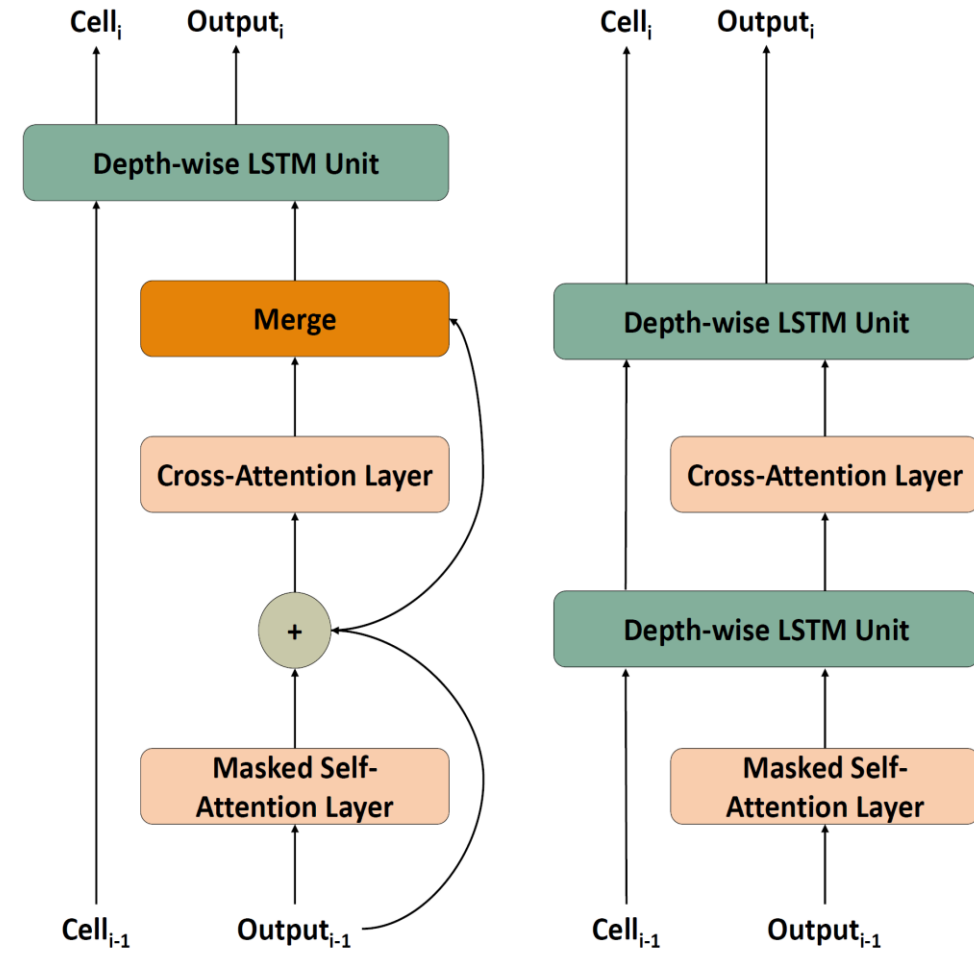
Depth-wise LSTM



Encoder/Decoder layers with depth-wise LSTM



Encoder layer



(a)

(b)

Decoder layer

Results – WMT 14 Base/Big Models

Models	En-De	En-Fr
Transformer Base	27.55	39.54
with depth-wise LSTM	28.53[†]	40.10[†]
Transformer Big	28.83	41.92
with depth-wise LSTM	29.58[†]	43.11[†]

Table 1: Results on WMT 14 En-De and En-Fr. [†] indicates $p < 0.01$ in the significance test.

Results – Ablation

Approaches	BLEU	Para.(M)	Speed
Transformer	27.55	62.37	750.58
Depth-wise RNN	23.24	68.67	737.60
Depth-wise LSTM	28.53	70.25	674.96

Table 2: Ablation study of depth-wise approaches on WMT 14 En-De.

Merging	BLEU	Para.(M)	Speed
Concat	28.26	78.90	649.27
Add	28.53	70.25	674.96
2 Depth-wise LSTMs	28.81	100.18	581.13

Table 4: Results of merging operations for decoder layer on WMT 14 En-De.

LSTM FFN	Hidden size	BLEU	Para.(M)	Speed
1-layer (Eq. 5)	512	27.84	45.05	742.19
2-layer (Eq. 6)	2048	28.53	70.25	674.96
	1586	28.20	62.37	683.67

Table 3: Ablation study of LSTM hidden computation on WMT 14 En-De.

Sharing	BLEU	Para.(M)
All	26.94	44.00
Gate	28.53	70.25
None	28.25	87.59

Table 5: Results of sharing LSTM parameters on WMT 14 En-De.

Results – Deep models

Models	Layers		En-De	Cs-En	Para.(M)	Speed
	Encoder	Decoder				
Transformer Base						
TA (Bapna et al., 2018)*	16		28.39	29.36	93.87	711.78
DLCL (Wang et al., 2019)	30	6	29.3		137.97	577.30
ODE (Li et al., 2022a)	24		30.29		119.17	565.86
Layer Aggregation (Dou et al., 2018)		6	28.63	None	111.10	667.57
EM Routing (Dou et al., 2019)		6	28.81		144.80	561.28
SDU (Chai et al., 2020)*		6	28.22		78.13	664.20
Luna (Ma et al., 2021)		6	27.8		77.60	None
DSI (Zhang et al., 2019)		20	28.67		149.54	298.50
LCPI (Xu et al., 2020a)		24	29.20	29.88	194.66	229.90
Transformer Big						
Layer Aggregation (Dou et al., 2018)		6	29.21		356.38	264.55
EM Routing (Dou et al., 2019)		6	28.97	None	490.38	221.70
MC (Wei et al., 2020)		18	30.56		798.23	70.37
ODE (Li et al., 2022a)	12	6	30.77		288.46	315.91
		3	26.36		40.33	1209.62
		6	27.55		62.37	750.58
Transformer Base		12	28.12		106.47	429.00
		18	28.60		150.57	299.81
		24	29.02		194.66	229.90
		3	27.38		46.63	1121.16
		6	28.53		70.25	674.96
Transformer Base with depth-wise LSTM		12	29.26		122.23	379.83
		18	29.41	30.27	172.63	277.21
		24	29.18	30.02	223.02	202.40
Transformer Big with depth-wise LSTM		12	30.69	30.57	452.04	181.58
+ experiment settings of Li et al. (2022a)	12	6	31.12	31.25	338.75	316.15
+ 1-layer LSTM FFN (Eq. 5)			30.83	30.96	288.41	363.60

Table 6: Results of Deep Transformers. “*” indicates reproduction of the approach.

Results – MNMT (OPUS-100)

Models	Direction	BLEU ₉₄	WR	BLEU ₄
Transformer	En→xx	18.75	-	14.73
	xx→En	27.02		22.50
Transformer + LALN + LALT (Zhang et al., 2020)	En→xx	20.81	-	17.45
	xx→En	27.22		23.30
Depth-wise LSTM	En→xx	23.38	98.94	20.47
	xx→En	28.41	79.79	26.68

Table 7: Results of multilingual NMT.

Take-away

- selectively aggregating different layer representations of the Transformer may improve the performance, and propose to use depth-wise LSTMs to connect stacked (sub-) layers of Transformers.
- show how Transformer layer normalization and feed-forward sub-layers can be absorbed by depth-wise LSTMs, while connecting pure Transformer attention layers by depth-wise LSTMs.
- experiments on MT prove the effectiveness.

Contact: hfxunlp@foxmail.com