Highway Transformer: Self-Gating Enhanced Self-Attentive Networks

Yekun Chai[1], Shuo Jin[2]

[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Pittsburgh

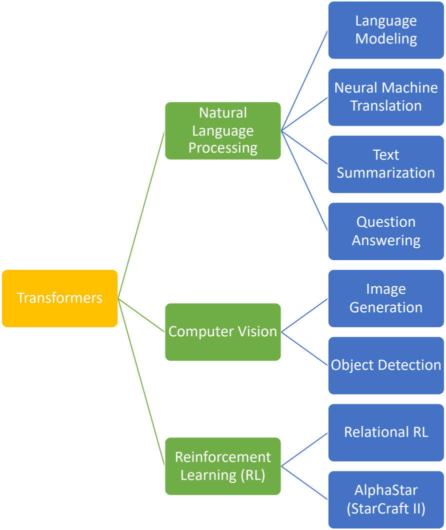
chaiyekun@gmail.com

June 14, 2020

Overview

1. Introduction
2. Methodology
3. Experiments
4. Conclusions

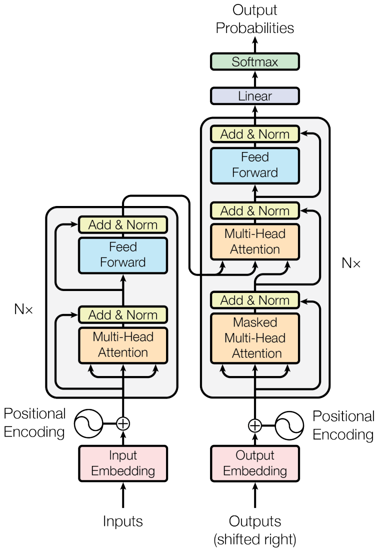
Introduction



Transformers [1]

1. Extensive applications.
2. Salient achievements.
3. Parallel training (precludes the sequence-aligned recurrence as in LSTMs).

Transformer Architecture [1]



Sublayers

1. Multi-head dot product attention.
2. Position-wise feed-forward layer.

Location-Unaware

- Absolute sinusoidal Positional Encoding.

Questions & Motivation

Q1: Are vanilla Transformers sufficient for seq-to-seq learning?

1. Previous works [2, 3, 4] leverage gating mechanisms (GLU) and Convolutional Neural Networks (CNN) to learn sequences.
2. CNNs are adept in learning *local-region features* whereas Transformers are good at modeling *global dependencies*.

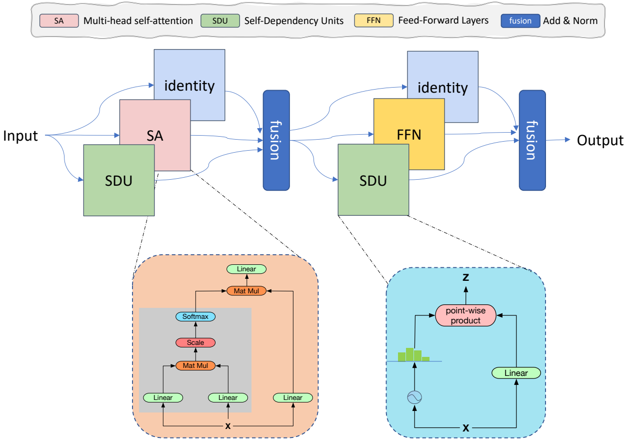
Q2: Do we need identical Transformer stacks in different depth?

- Previous work [5] claimed that self-attention models tend to capture local features in the bottom layers.

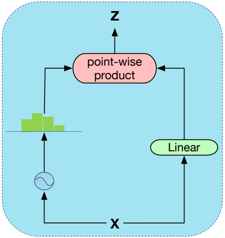
Highway Transformer Architecture

Three streams:

- Self-dependency (SDU);
- Inter-dependency (SAN / FFN);
- Identity (residual connection).



Self-Dependency Units (SDU)



$$T(\mathbf{X}) = \mathbf{\Psi}(\mathbf{X}\mathbf{W_1} + \mathbf{b_1}) \quad (1)$$

$$\mathrm{SDU}(\mathbf{X}) = T(\mathbf{X}) \odot (\mathbf{X}\mathbf{W_2} + \mathbf{b_2}) \quad (2)$$

where $T(\mathbf{X})$ indicates the *transform* gate, $\mathbf{\Psi}$ is the gate function to confine the linear projection into a fixed range, which takes the sigmoidal-curve functions such as σ and tanh.

- tanh is treated as an update gate to restrict the importance range into [-1,1].
- σ can be regarded as the input gate to modulate how much information to retain at the feature-wise level.

Pseudo-Highway Connection

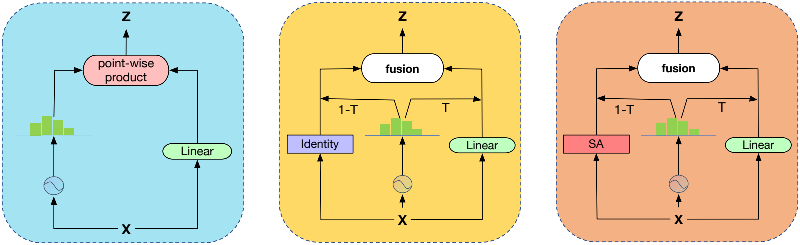
Pseudo-Highway Connection

When taking σ as the non-linearity:

$$\begin{aligned} \nabla[\mathbf{f}(\mathbf{X}) \odot \sigma(\mathbf{g}(\mathbf{X}))] = & \overbrace{\sigma(\mathbf{g}(\mathbf{X}))}^{\text{transform gate}} \odot \nabla \mathbf{f}(\mathbf{X}) \\ & + \overbrace{(1 - \sigma(\mathbf{g}(\mathbf{X})))}^{\text{carry gate}} (\sigma(\mathbf{g}(\mathbf{X})) \odot \mathbf{f}(\mathbf{X})) \end{aligned} \tag{3}$$

where the $\sigma(.)$ can be seen as the transform gate, while $(1 - \sigma(.))$ can be seen as the carry gate. This could be regarded as a form of highway networks.

Gating Variants



Self-Dependency Units (SDU) Highway Gate Gated Multi-Head Self-Attention

1. SDU $\hookrightarrow$ Self-dependency on itself by applying transform gate T.
2. Highway gate $\hookrightarrow$ Additional carry gate $(1 - T)$ on identity.
3. Gated Multi-Head Self-Attention $\hookrightarrow$ Additional carry gate $(1 - T)$ on SA.

Experiments

Transformers

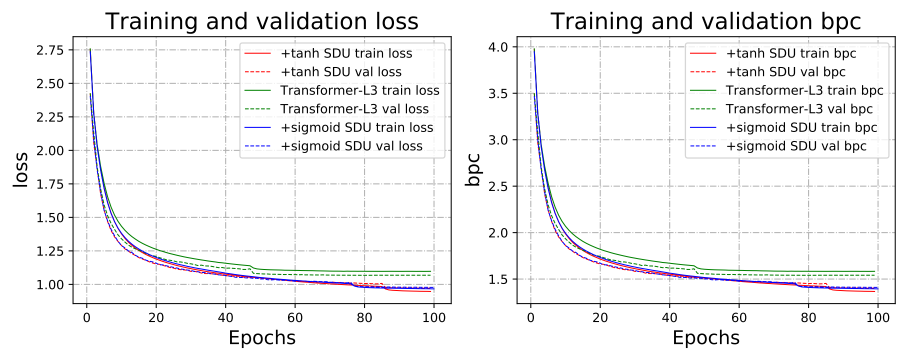
1. Vanilla Transformer
2. R-Transformer [7]
3. Transformer-XL [6]

Language Modeling Datasets

1. Char-level PTB
2. Word-level PTB
3. enwik8

Results

3-layer Transformer (T-L3) on char-level PTB

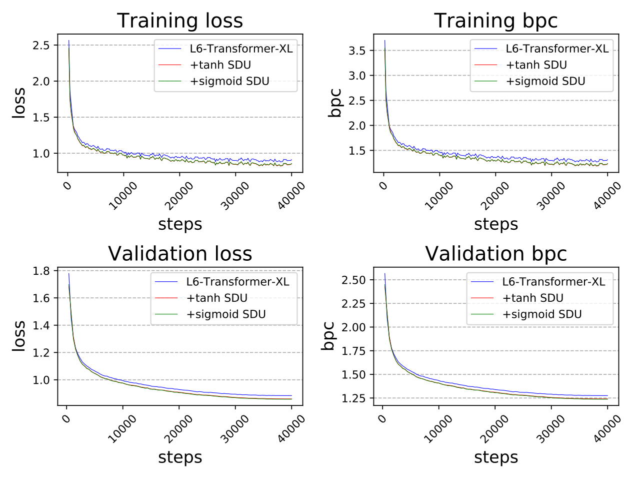


model	eval loss	eval ppl	test loss	test ppl
T-L3	1.068	1.541	1.036	1.495
$+\sigma$ SDU	0.9776	1.410$\Downarrow$	0.950	1.371$\Downarrow$
+tanh SDU	**0.9714**	**1.401**$\Downarrow$	**0.945**	**1.364**$\Downarrow$

Results

6-layer Transformer-XL (XL-L6) on *enwik8*

SDUs accelerate the convergence speed during training and evaluation process!



Results

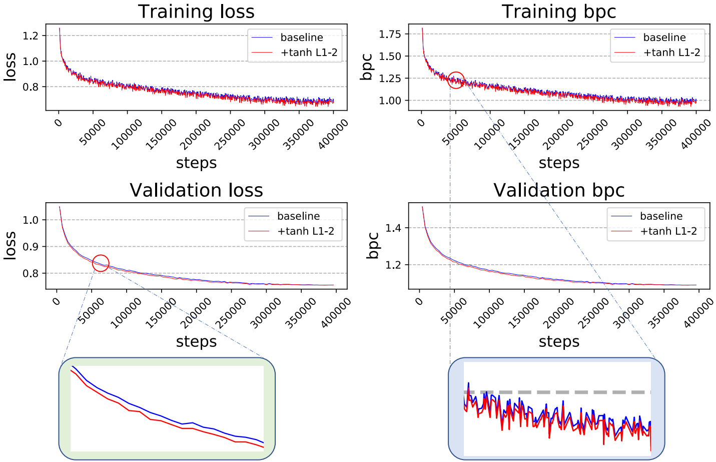
Ablation Study: XL-L6 on *enwik8*

model	**eval loss**	**eval bpc**	**test loss**	**test bpc**
L6-XL	0.8843	1.276	0.86	1.24339
+tanh SDU	0.8602	1.241⇓	**0.84**	1.21424⇓
+σ SDU	**0.8577**	**1.237**⇓	**0.84**	**1.21123**⇓
+highway gate	0.8692	1.254⇓	0.85	1.22177⇓
+gated MHDPA	0.8682	1.253⇓	0.85	1.22398⇓
		Ablation study		
+tanh L1-6\FFN	0.8720	1.258⇓	0.85	1.22866⇓
+tanh L1-3	**0.8660**	**1.249**⇓	**0.85**	**1.22039**⇓
+tanh L3-6	0.8852	1.277⇓	0.86	1.24420⇓
+σ L1-6\FFN	0.8752	1.263⇓	0.85	1.23332⇓
+σ L1-3	0.8792	1.268⇓	0.86	1.23589⇓
+σ L3-6	0.8843	1.276⇓	0.86	1.24261⇓

Results

12-layer Transformer-XL (XL-L12) on enwik8.

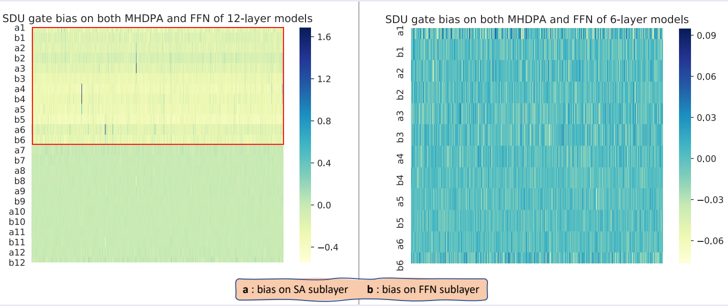
- Our experiments showed that SDU on **shallow layers** could accelerate the convergence process.



Results

Visualization of learned baises on SDUs

Shallow layers of Transformers may attend to different semantics from top layers.



Conclusions

- Self-Gating Units (SDU) allows for the pseudo-highway information flow, leading to the better convergence during training/evaluation process.
- It is compatible and scalable to common Transformer variants, including Transformer-XL and R-Transformer.
- Low layers in the Transformer stacks may pay more attention to local features [5], and the SDU components can be applied on the bottom layers for deep Transformer models.

References

Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin (2017). Attention is All you Need, *NeurIPS*

Dauphin, Yann N and Fan, Angela and Auli, Michael and Grangier, David (2017), Language modeling with gated convolutional networks, *ICML*

Gehring, Jonas and Auli, Michael and Grangier, David and Yarats, Denis and Dauphin, Yann N (2017), Convolutional sequence to sequence learning, *ICML*

Wu, Felix and Fan, Angela and Baevski, Alexei and Dauphin, Yann N and Auli, Michael (2019), Pay less attention with lightweight and dynamic convolutions, *ICLR*

Baosong Yang and Zhaopeng Tu and Derek F. Wong and Fandong Meng and Lidia S. Chao and Tong Zhang (2018), Pay less attention with lightweight and dynamic convolutions, *EMNLP*

References

Dai, Zihang and Yang, Zhilin and Yang, Yiming and Cohen, William W and Carbonell, Jaime and Le, Quoc V and Salakhutdinov, Ruslan (2019), Transformer-xl: Attentive language models beyond a fixed-length context, *ACL*

Wang, Zhiwei and Ma, Yao and Liu, Zitao and Tang, Jiliang (2019), R-transformer: Recurrent neural network enhanced transformer, *arXiv*

Thanks