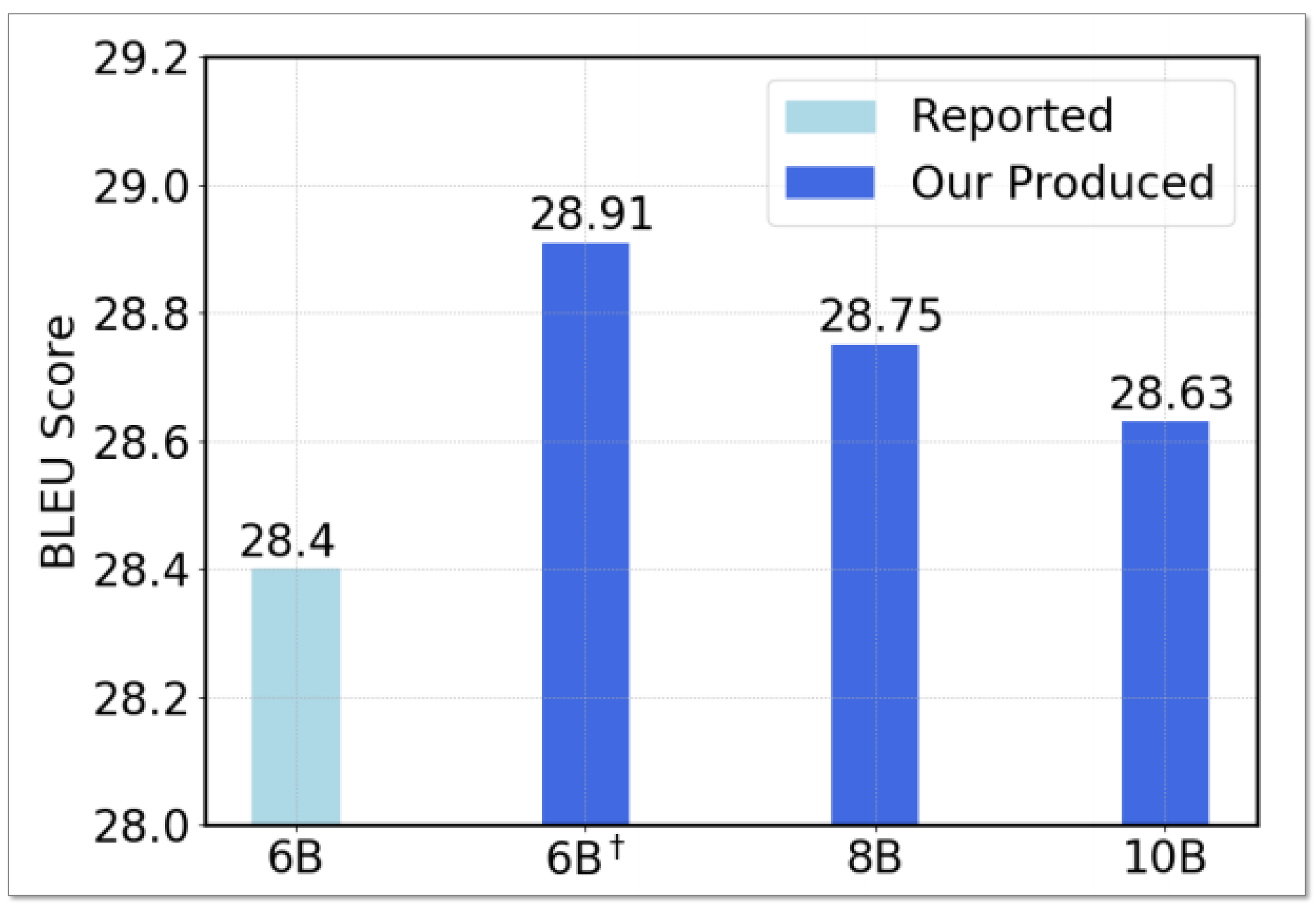
Depth Growing for Neural Machine Translation

[1]Lijun Wu, [2]Yiren Wang, [3]Yingce Xia, [3]Fei Tian, [3]Fei Gao,
[3]Tao Qin, [1]Jianhuang Lai and [3]Tie-Yan Liu
[1]Sun Yat-sen University; [2]University of Illinois at Urbana-Champaign; [3]Microsoft Research Asia

57th ACL 2019

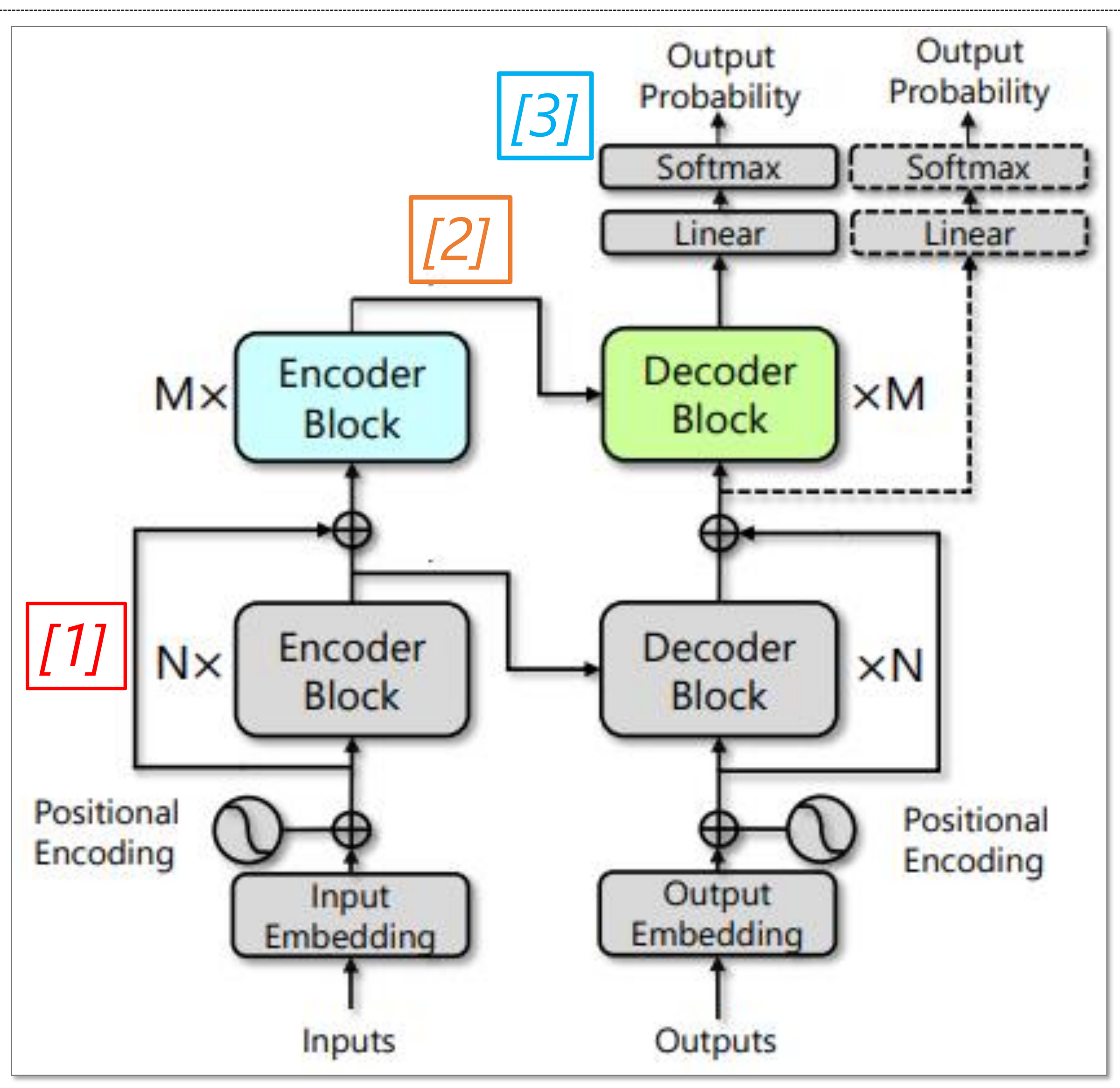
1. Motivation

- Training ***deep networks*** has been widely adopted and has ***shown effectiveness*** in image recognition, QA and text classification.
- Very deep and effective model training still ***remains challenging for NMT***.



- Instead of working on RNN/CNN structures, we propose a novel approach to construct and train ***deeper NMT models based on Transformer***.

2. Framework



3. Depth Growing

$$h_1 = \mathtt{enc}_1(x);\ h_2 = \mathtt{enc}_2(x + h_1); \quad (1)$$

$$s_{1,t} = \mathtt{dec}_1(y_{<t}, \mathtt{attn}_1(h_1)),\ \forall t \in [l_y]; \quad (2)$$

$$s_{2,t} = \mathtt{dec}_2(y_{<t} + s_{1,<t}, \mathtt{attn}_2(h_2)), \quad (3)$$

- *[1] Cross-module residual connections*
- *[2] Hierarchical encoder-decoder attention*
- *[3] Depth-shallow decoding*

4. Two-stage Training

- Stage-1: The bottom modules (enc_1 and dec_1) are trained and subsequently fixed.
- Stage-2: Only the top modules (enc_2 and dec_2) are trained and optimized.

Discussion:

- ➢ Training complexity is reduced compared with jointly training, which eases optimization difficulty.
- ➢ We only have a "single" model grown to be a well-trained deeper one, which outperforms the "ensemble" models.

5. Experiments

- **Overall Results**
  - WMT14 En→De and WMT14 En→Fr

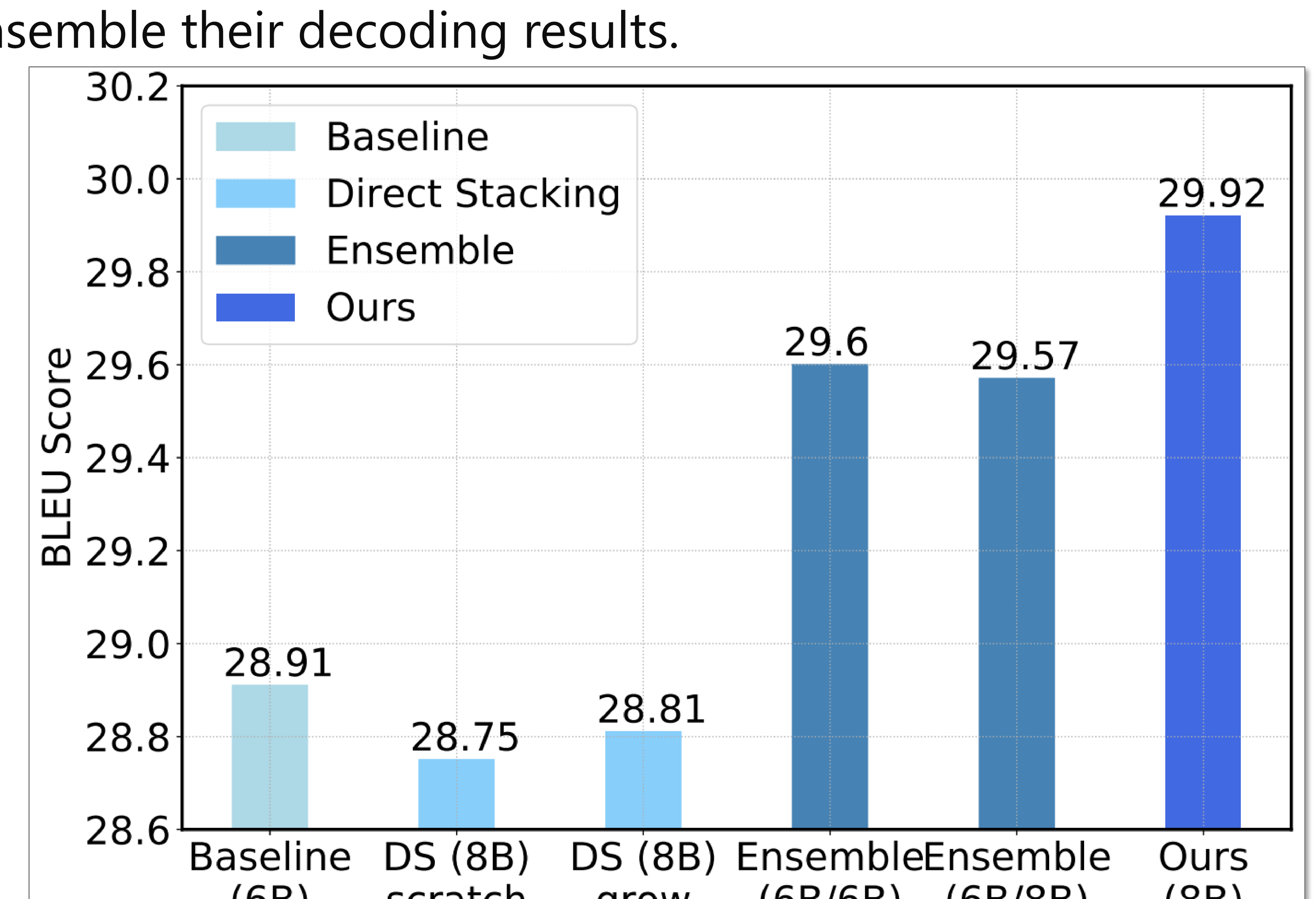
The test performances of WMT14 En→De and En→Fr.

Model	En→De	En→Fr
Transformer (6B)†	28.40	41.80
Transformer (6B)	28.91	42.69
Transformer (8B)	28.75	42.63
Transformer (10B)	28.63	42.73
Transparent Attn (16B)†	28.04	–
Ours (8B)	**29.92**	**43.27**

dagger: results reported in previous works

- *We achieve **30.07** BLEU score on En→De with 10 blocks (10B).*

- **Analysis**
  - *Directly Stacking (DS):* extend the 6-block baseline to 8-block by directly stacking 2 blocks.
  - *Ensemble Learning (Ensemble):* separately train 2 models and ensemble their decoding results.



The test performances of WMT14 En→De translation task.

Code

- https://github.com/apeterswu/Depth_Growing_NMT

Contact

- wulijun3@mail2.sysu.edu.cn (SYSU)
- yingce.xia@microsoft.com (MSRA)