Soft Contextual Data Augmentation for Neural Machine Translation

[1,*]Fei Gao, [2,*]Jinhua Zhu, [3]Lijun Wu, [4]Yingce Xia, [4]Tao Qin,
[1]Xueqi Cheng, [2]Wengang Zhou and [4]Tie-Yan Liu
[1]Institute of Computing Technology, Chinese Academy of Sciences;
[2]University of Science and Technology of China;
[3]Sun Yat-sen University; [4]Microsoft Research Asia

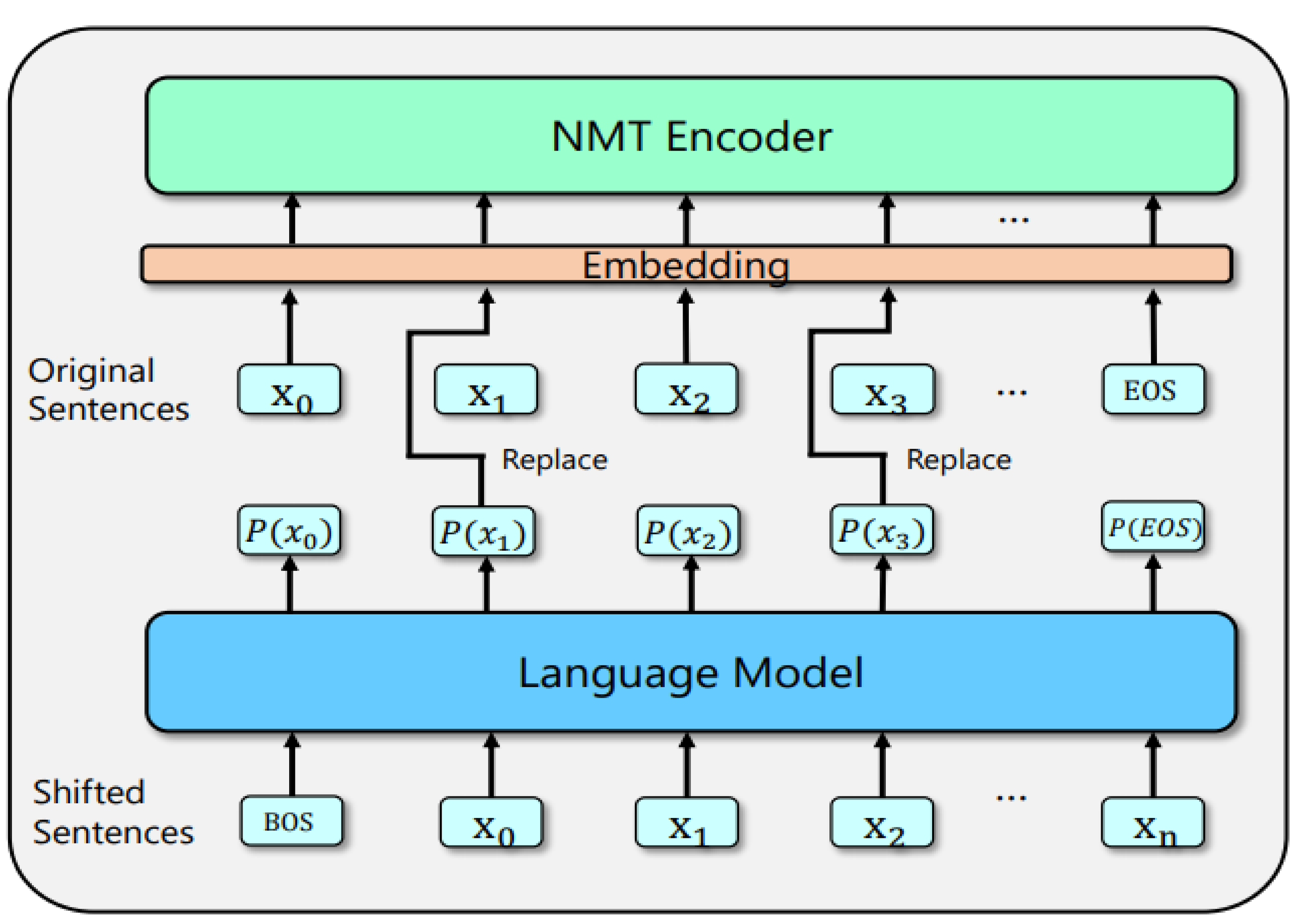
57th ACL 2019

1. Motivation

- Study of data augmentation in natural language tasks is still very limited.
- Current random transformation methods such as Swap, Dropout and Blank can result in significant changes in semantics.
- Recent contextual augmentation methods can not utilize all potential candidates.
- We propose **soft contextual data augmentation** for NMT by leveraging language models, which can not only keep semantics for source sentences, but also leverage all possible augmented data.

2. Framework

- We show the architecture of our soft contextual data augmentation approach in encoder side for source sentences. The decoder side for target sentences is similar.



3. Soft word

- Soft version of a word, w is a distribution over the vocabulary of $|V|$ words:

$$-P(w) = \left(p_1(w), p_2(w), \dots, p_{|V|}(w)\right)$$

- The embedding of the soft word w is:

$$-e_w = P(w)E = \sum_{j=0}^{|V|} p_j(w)E_j$$

- We leverage a pre-trained language model to compute $P(w)$:

$$-p_j(x_t) = LM(w_j | x_{<t})$$

4. Two-stage Training

- Stage-1: First use the same training corpus of the NMT model to pretrain language models.
- Stage-2: Then randomly choose words in the training data with probability γ and replace it by its soft version to train one NMT model.

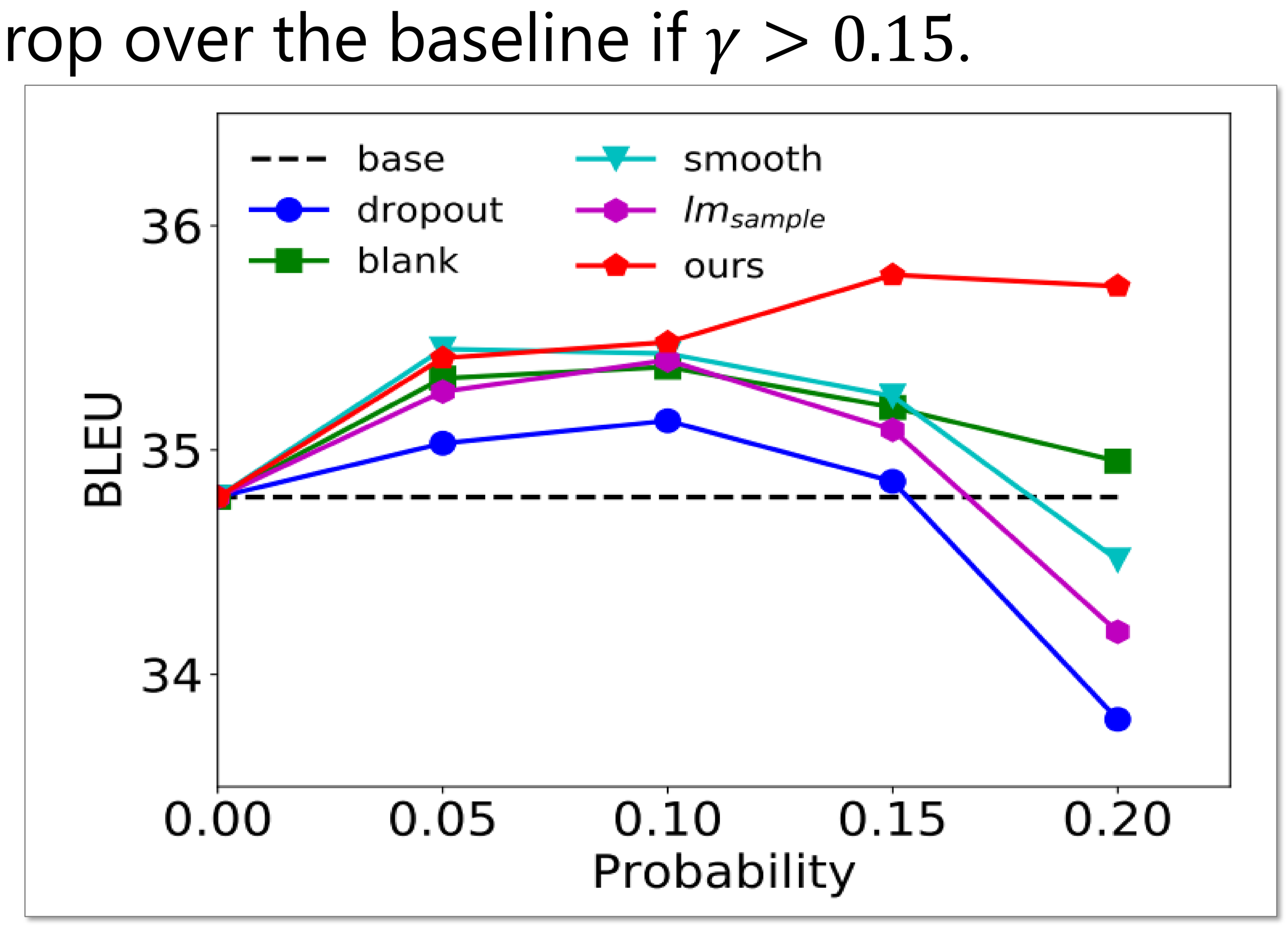
5. Experiments

- **Overall Results**
  - IWSLT14 {De, Es, He}->En (transformer_base) and WMT14 En->De (transformer_big).

	IWSLT			WMT
	De → En	Es → En	He → En	En → De
Base	34.79	41.58	33.64	28.40
+Swap	34.70	41.60	34.25	28.13
+Dropout	35.13	41.62	34.29	28.29
+Blank	35.37	42.28	34.37	28.89
+Smooth	35.45	41.69	34.61	28.97
$+LM_{sample}$	35.40	42.09	34.31	28.73
Ours	**35.78**	**42.61**	**34.91**	**29.70**

Table 1: BLEU scores on four translation tasks.

- **Analysis** (On IWSLT14 De->En)
  - Our method can observe a consistent BLEU improvement within a large probability range.
  - While other methods can easily lead to performance drop over the baseline if $\gamma > 0.15$.



Code
https://github.com/teslacool/SCA
Contact
Fei Gao: gaofei17b@ict.ac.cn
Jinhua Zhu: teslazhu@mail.ustc.edu.cn

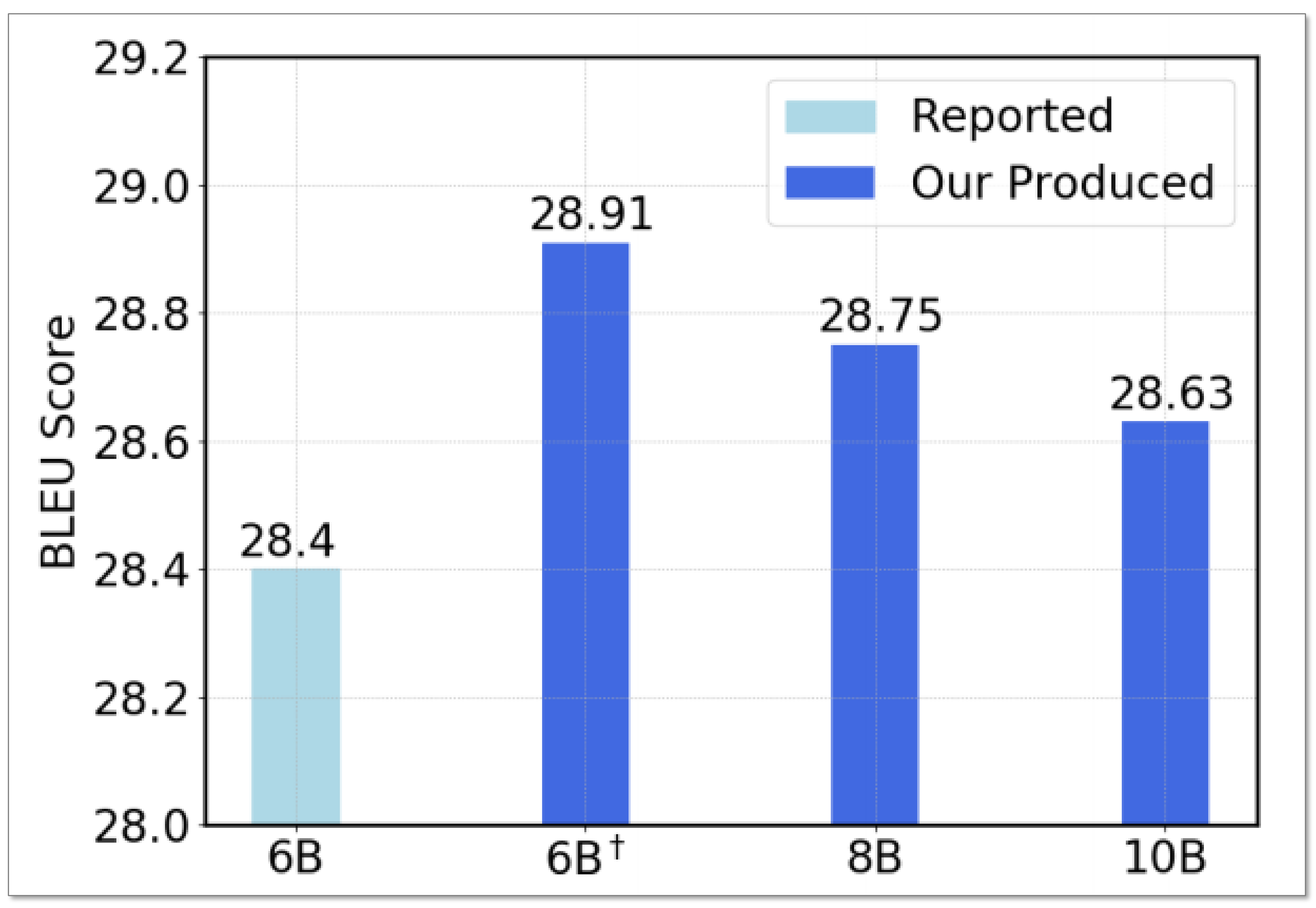
Depth Growing for Neural Machine Translation

[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

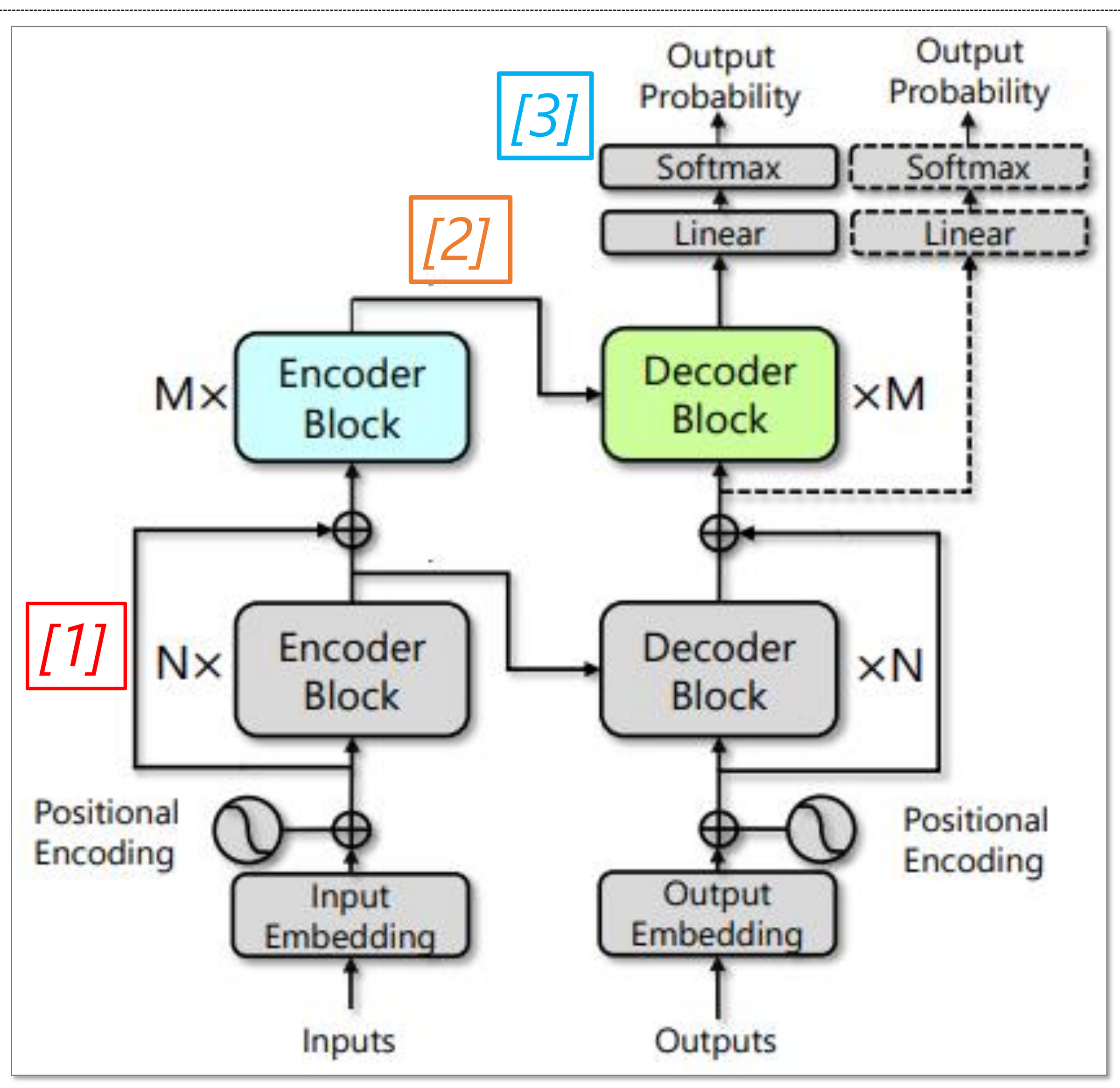
1. Motivation

- 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

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

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

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

- *[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

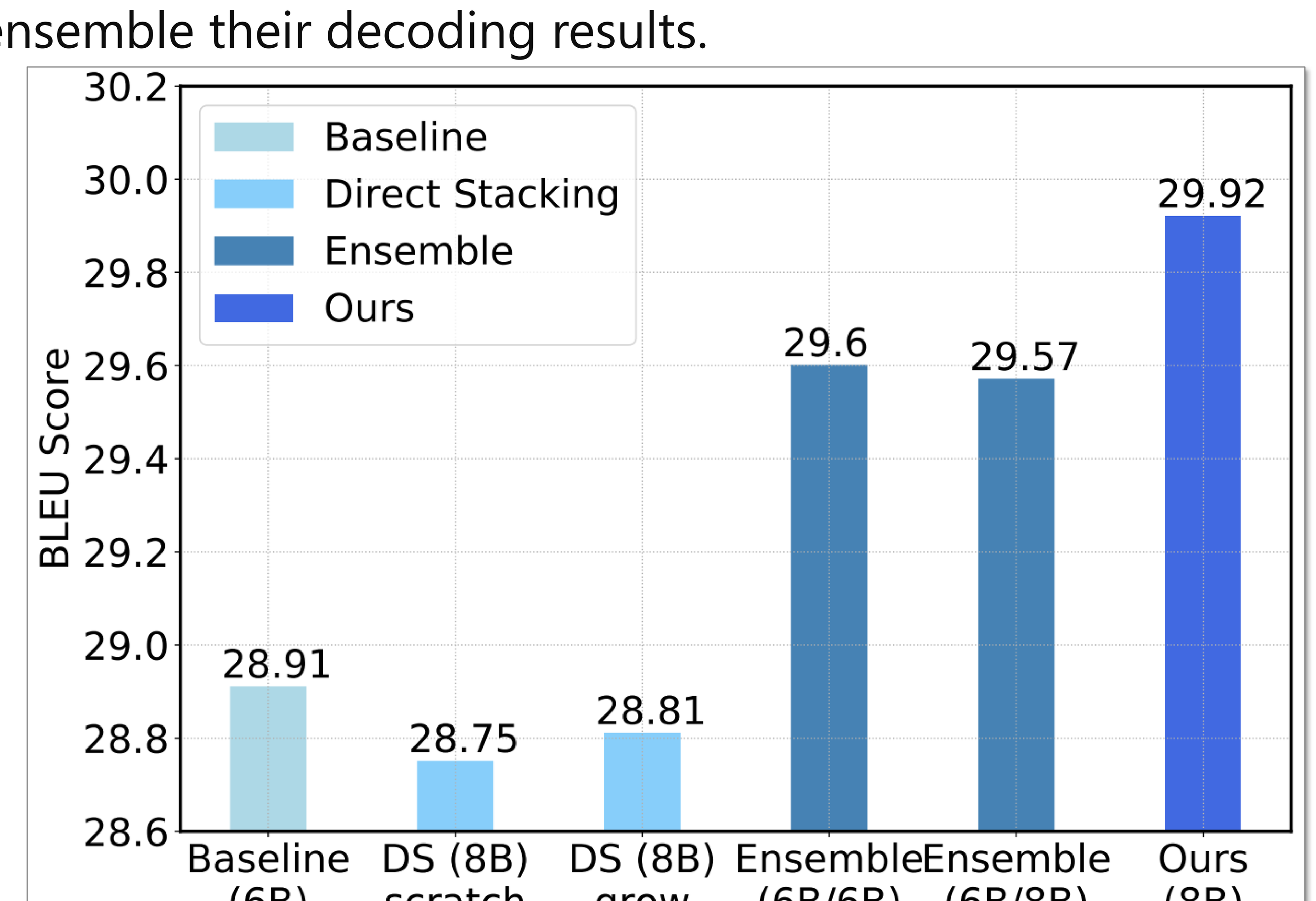
The test performances of WMT14 En→De and En→Fr.

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)