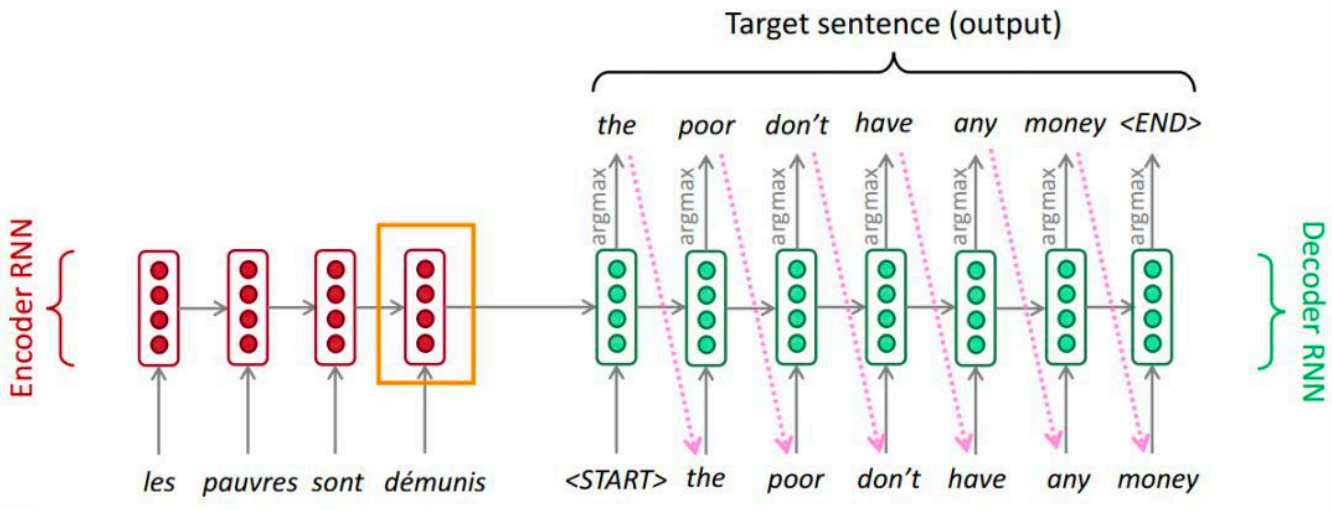
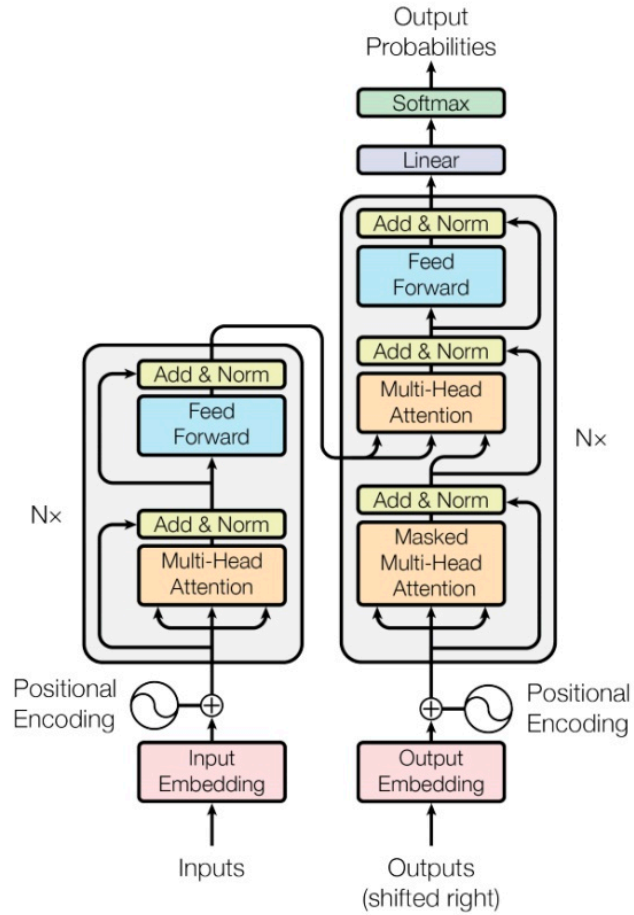
Improving Neural Machine Translation with Soft Template Prediction

Jian Yang[1] [*], **Shuming Ma**[2], **Dongdong Zhang**[2], **Zhoujun Li**[1] [†], **Ming Zhou**[2]
[1]State Key Lab of Software Development Environment, Beihang University
[2]Microsoft Research Asia
{jiaya, lizj}@buaa.edu.cn; {shumma, dozhang, mingzhou}@microsoft.com;


Target sentence (output)
the
poor
don't
have
any
money
<END>
argmax
Encoder RNN
Decoder RNN
les
pauvres
sont
démunis
<START>
the
poor
don't
have
any
money


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Template能够指导生成，如果能够利用target端的模版信息？

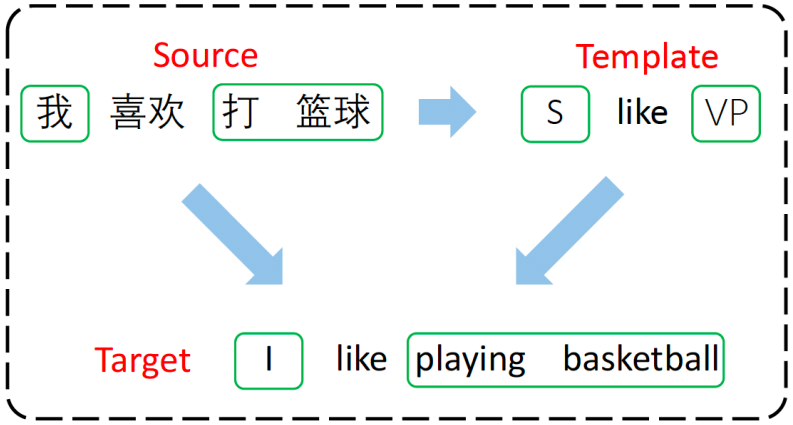


Figure 1: An example of template guided translation results. S denotes subject and VP denotes verb phrase.

Source	另一方面 , 如果 我们 反应 过度 , 将 会 被 他们 欺骗 .
Reference	on the other hand , if we overreact , we will be deceived by their trick .
Template	on the other hand , if NP VP , we will VP .
Ours	on the other hand , if we react too much , we will be hit by them .

Table 7: A Chinese-English translation example of our proposed method. VP and NP represent non-terminal nodes in the constituency-based parse tree.

$$P(Y|X) = P_{\theta_{X \to T}}(T|X) P_{\theta_{(X,T) \to Y}}(Y|X,T)$$

分成两个阶段进行训练：

1. source text => target template
2. source text + target template => target text

第一阶段：X => T

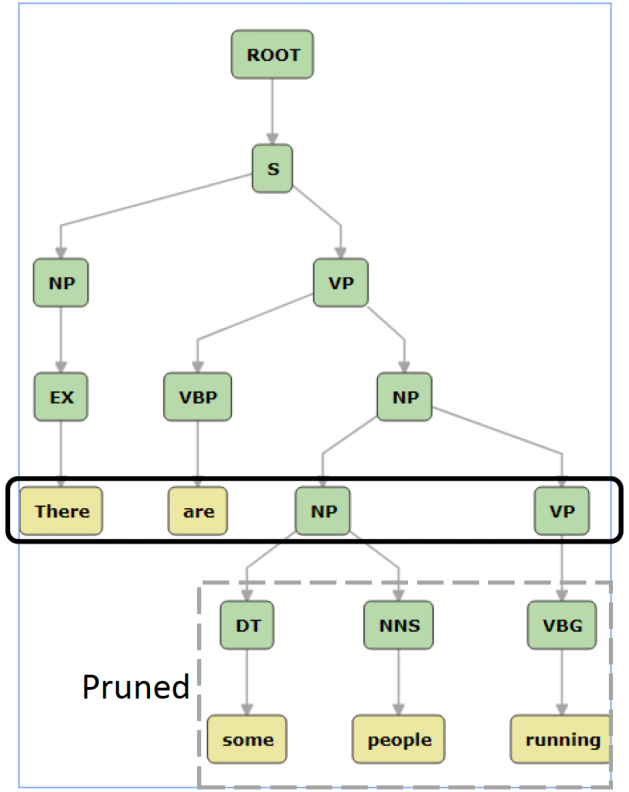
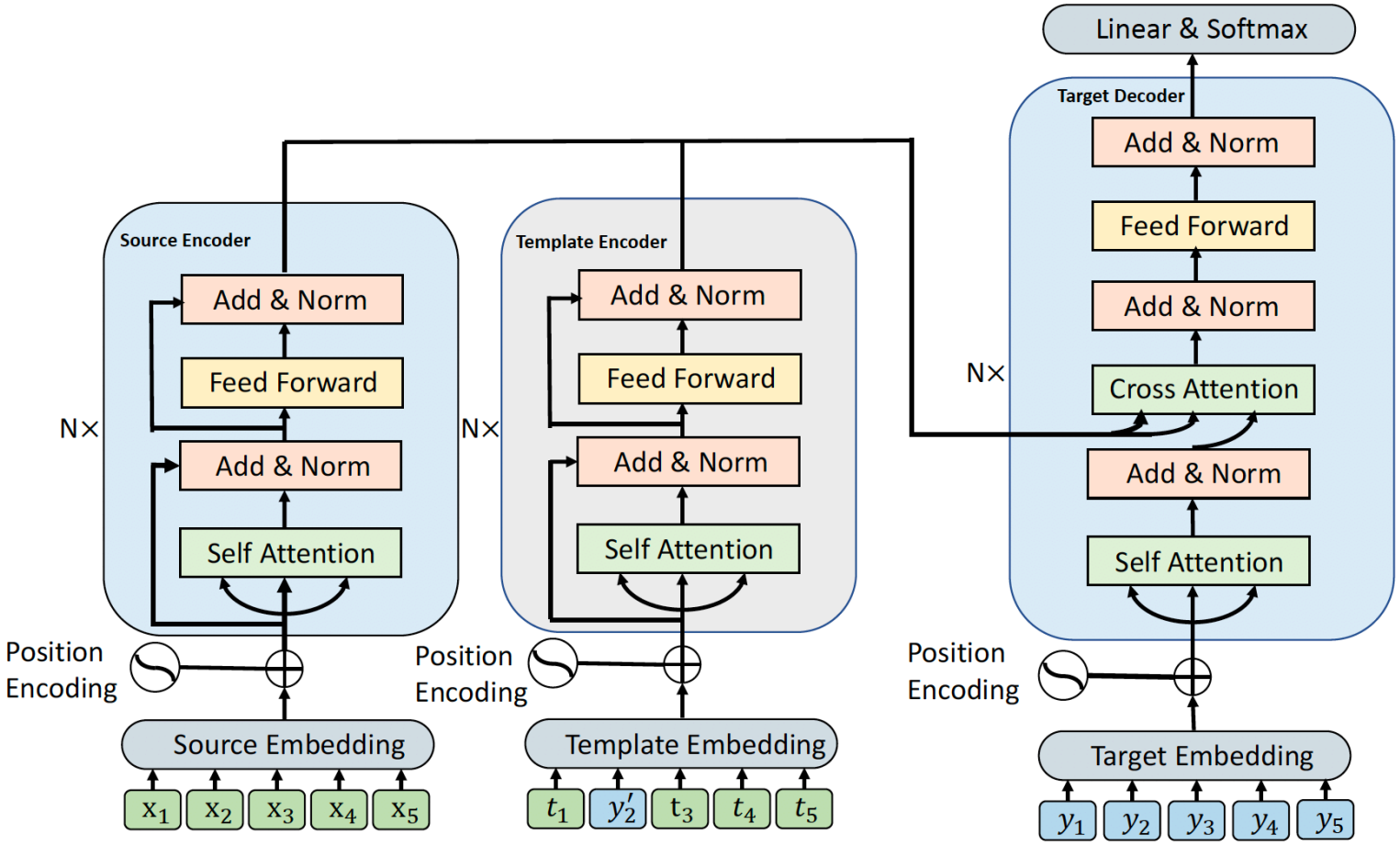


Figure 3: The constituency-based parse tree of the example sentence. Given the target sentence and definite depth of the tree, we gain the sub-tree by pruning the nodes deeper than 4 in this case. Then, the sub-tree can be converted to the soft target template "There are NP VP" from left to right.

训练完成之后，对训练数据进行beam search解码得到top-K个最好的template，构造下一阶段的训练数据。
原来：(x1, y1)
现在：(x1, T1, y1), (x1, T2, y1), … , (x1, TK, y1),
极端情况下K=1，相当于greedy decoding。

第二阶段：X + T => Y



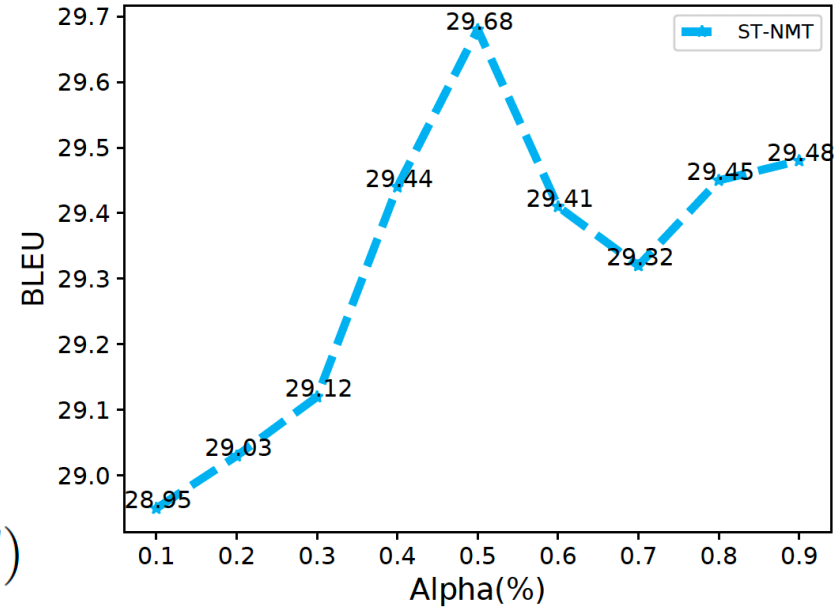
$$Z = \beta Z^{X,Y} + (1-\beta) Z^{T,Y}$$

$$\beta = \sigma(W_Y Z^{X,Y} + U_T Z^{X,T})$$

第二阶段：X + T => Y

a % $$L_{\theta_{X\to Y}}(D) = \sum_{X,Y\in D} \log P_{\theta_{X\to Y}}(Y|X)$$

(1-a) % $$L_{\theta_{(X,T)\to Y}}(D) = \sum_{X,Y\in D} \log P_{\theta_{(X,T)\to Y}}(Y|X,T)$$



用两个objective的原因是：可能存在一些low-quality的template，这样可能可以降低一下它们带来的影响

实验结果

De → En	BLEU
GNMT (Wu et al., 2016)	31.44
RNMT+ (Chen et al., 2018)	34.51
ConvS2S (Gehring et al., 2017)	30.41
LightConv (Wu et al., 2019)	34.80
DynamicConv (Wu et al., 2019)	35.20
Rerank-NMT (Liu et al., 2016)	34.82
Transformer (our implementation)	34.43
ST-NMT (our proposed)	**35.24**

Table 2: BLEU-4 scores (%) on IWSLT14 De→En task. The result of our model is statistically significant compared to the other baselines ($p < 0.05$).

En → De	BLEU
GNMT (Wu et al., 2016)	24.61
ConvS2S (Gehring et al., 2017)	25.16
Transformer (Vaswani et al., 2017)	28.40
RNMT+ (Chen et al., 2018)	28.49
Rerank-NMT (Liu et al., 2016)	27.81
ABD-NMT (Liu et al., 2016)	28.22
Deliberation Network (Xia et al., 2017)	29.11
SoftPrototype (Wang et al., 2019b)	29.46
SB-NMT (Zhou et al., 2019a)	29.21
SBSG (Zhou et al., 2019b)	27.45
Insertion Transformer (Stern et al., 2019)	27.41
Transformer (our implementation)	29.25
ST-NMT (our proposed)	**29.68**

Table 3: BLEU-4 scores (%) on WMT14 En→De task. The result of our model is statistically significant compared to the other baselines ($p < 0.05$).

Template个数的影响：
增加template数量效果下降，robustness上升

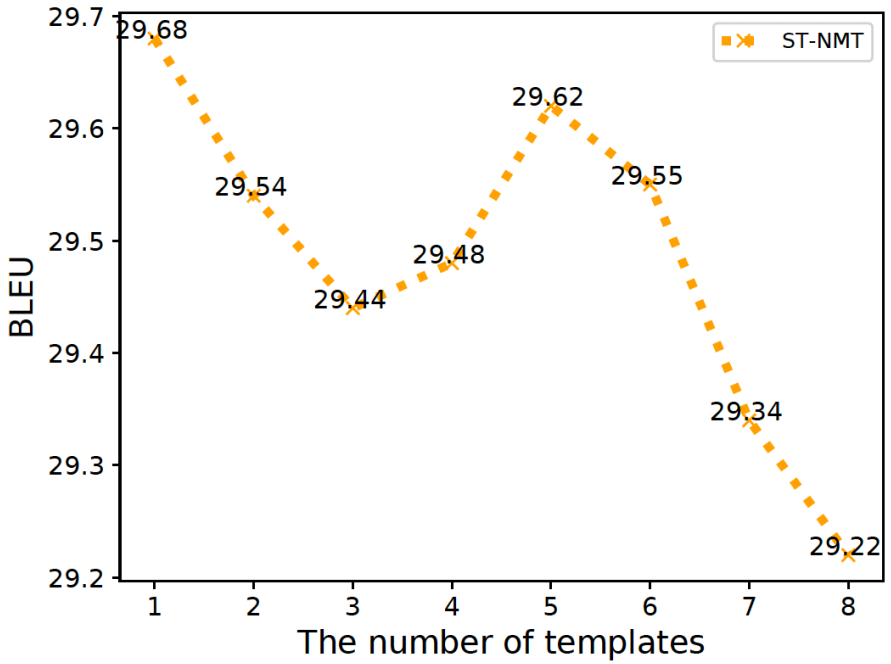


Figure 4: The effect of the multiple templates. We feed the the top-K results of the beam search as multiple templates and source sentence to generate the target translation.

树的深度如何决定？

$$d = \min(\max(L \times \lambda, \gamma_1), \gamma_2)$$

λ	MT03	MT05	MT08	MT12
0.10	45.92	45.01	36.55	35.34
0.15	**46.56**	**46.04**	**37.53**	**35.99**
0.20	46.02	45.20	37.08	35.82
0.25	46.27	44.83	36.88	35.64
0.30	46.08	45.02	36.72	35.54
0.35	46.22	44.92	36.84	35.51
0.40	46.32	45.40	36.94	35.61

Table 5: The results of the different depth on NIST2003, NIST2005, NIST2008 and NIST2012.

模版是不是真的发挥了作用？

$$ratio = \frac{\sum_{w \in T} \min\left(Count_y(w), Count_t(w)\right)}{\sum_{w \in T} Count_t(w)}$$

λ	MT03	MT05	MT08	MT12
0.15	79.4	81.6	78.6	77.6

Table 6: The ratio(%) of overlapping words between the predicted soft target template and the translation on NIST2003, NIST2005, NIST2008 and NIST2012.

Hard-Coded Gaussian Attention for Neural Machine Translation

Weiqiu You*, **Simeng Sun***, **Mohit Iyyer**
College of Information and Computer Sciences
University of Massachusetts Amherst
{wyou,simengsun,miyyer}@cs.umass.edu

Intuition

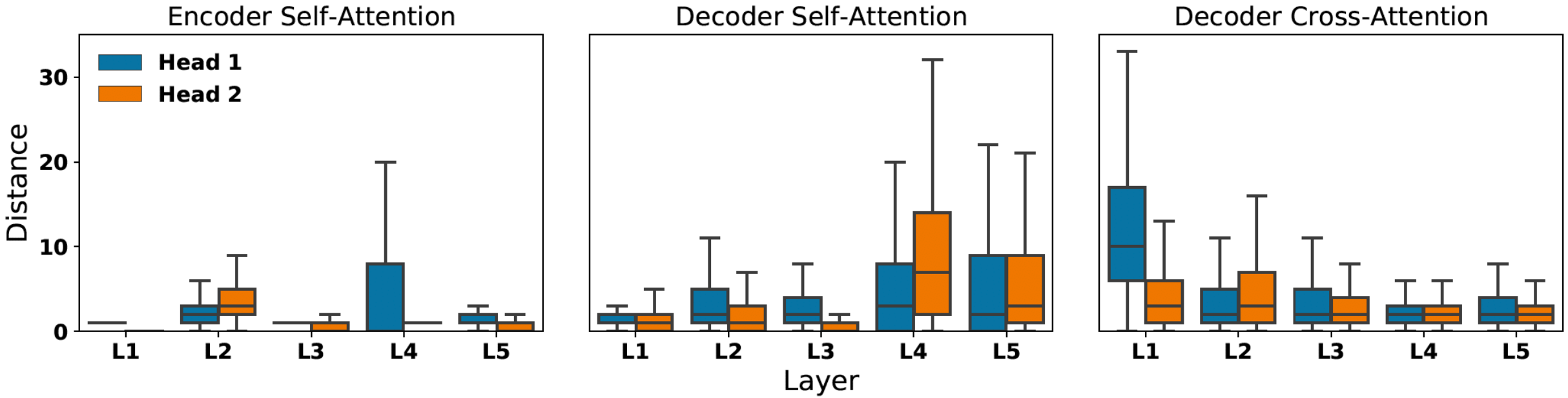


Figure 2: Most learned attention heads for a Transformer trained on IWSLT16 En-De focus on a local window around the query position. The x-axis plots each head of each layer, while the y-axis refers to the distance between the query position and the argmax of the attention head distribution (averaged across the entire dataset).

Self-attention => hard-coded gaussian

$$\mathrm{Attn}(i, \mathbf{V}) = \mathcal{N}(f(i), \sigma^2)\mathbf{V}.$$

其他paper：

1. Synthesizer
2. Lite TransFormer
3. Dynamic Convolution
4. …

Transformer:

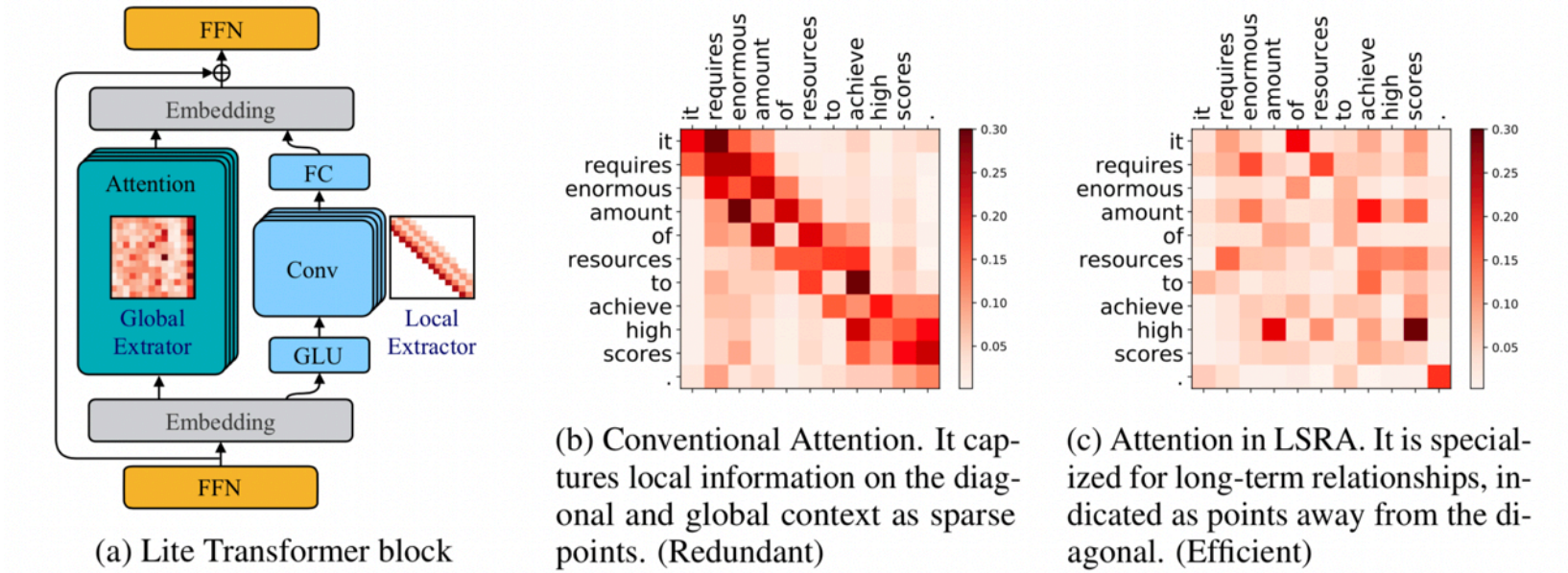
1. self-attention 冗余
2. self-attention 往往更关注local

1. Synthesizer：

F(X) 近似attention（不同时间步没有相互attend）

Random

2. Lite TransFormer：



(a) Lite Transformer block

(b) Conventional Attention. It captures local information on the diagonal and global context as sparse points. (Redundant)

(c) Attention in LSRA. It is specialized for long-term relationships, indicated as points away from the diagonal. (Efficient)

Figure 3: Lite Transformer architecture (a) and the visualization of attention weights. Conventional attention (b) puts too much emphasis on local relationship modeling (see the diagonal structure). We specialize the local feature extraction by a convolutional branch which efficiently models the locality so that the attention branch can specialize in global feature extraction (c). More visualizations are available in Figure A1.

3. Dynamic Convolution：

用卷积操作代替self-attention（捕捉local）

每个时间步生成一个新的卷积核

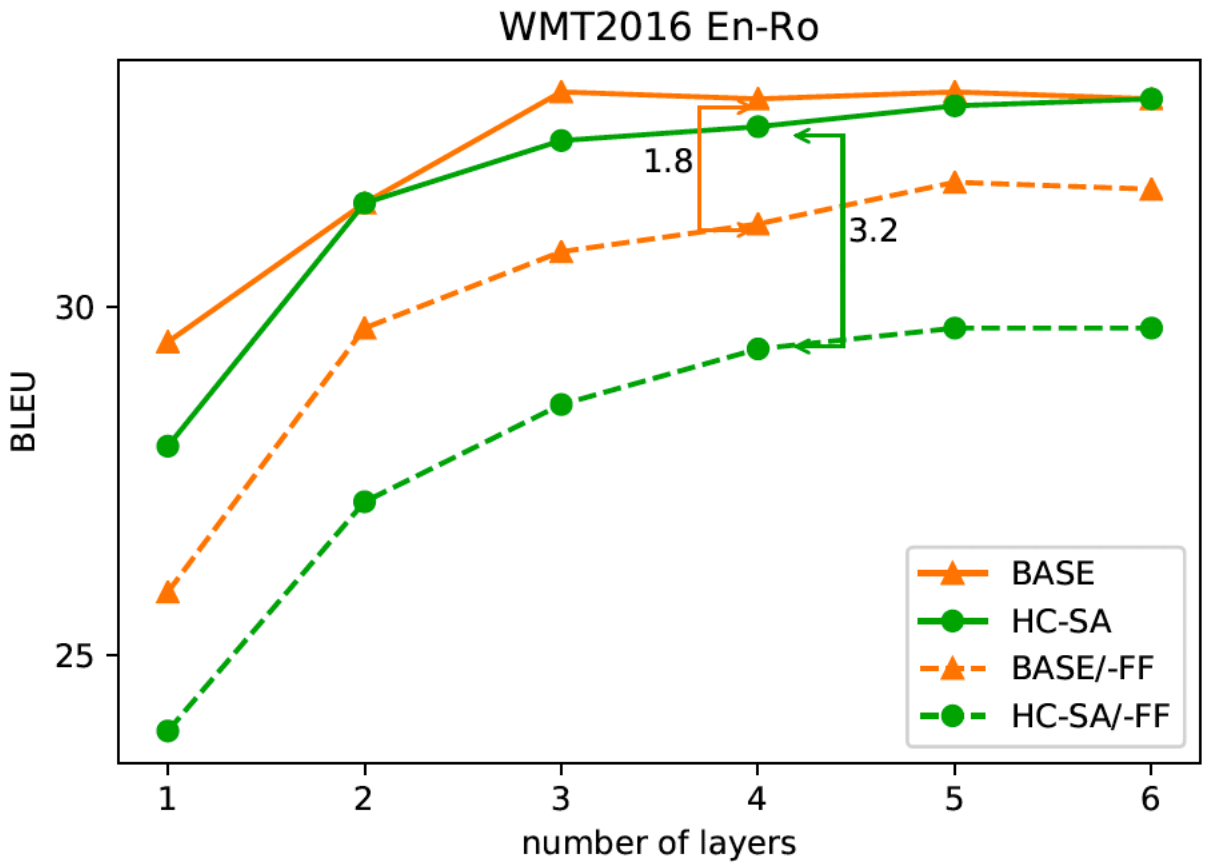
Cross attention & Experiments

HC-SA: 前面说的把encoder/decoder的self-attention替换成hard-coded gaussian
HC-ALL : 在前面基础上，把cross attention也替换掉
SH-X: 在HC-ALL基础上，保留decoder最后一层的单头attention

	BASE	HC-SA	HC-ALL	SH-X
IWSLT16 En-De	30.0	30.3	21.1	28.2
IWSLT16 De-En	34.4	34.8	25.7	33.3
IWSLT17 En-Ja	20.9	20.7	10.6	18.5
IWSLT17 Ja-En	11.6	10.9	6.1	10.1
WMT16 En-Ro	33.0	32.9	25.5	30.4
WMT16 Ro-En	33.1	32.8	26.2	31.7
WMT14 En-De	26.8	26.3	21.7	23.5
WMT14 En-Fr	40.3	39.1	35.6	37.1

为什么会去掉learned self-attention还能起作用？

猜测：可能是在hard-coded guassian的模型中，feed-forward network起到了更大的作用。



会不会对比较长的句子效果就不如learned self-attention？

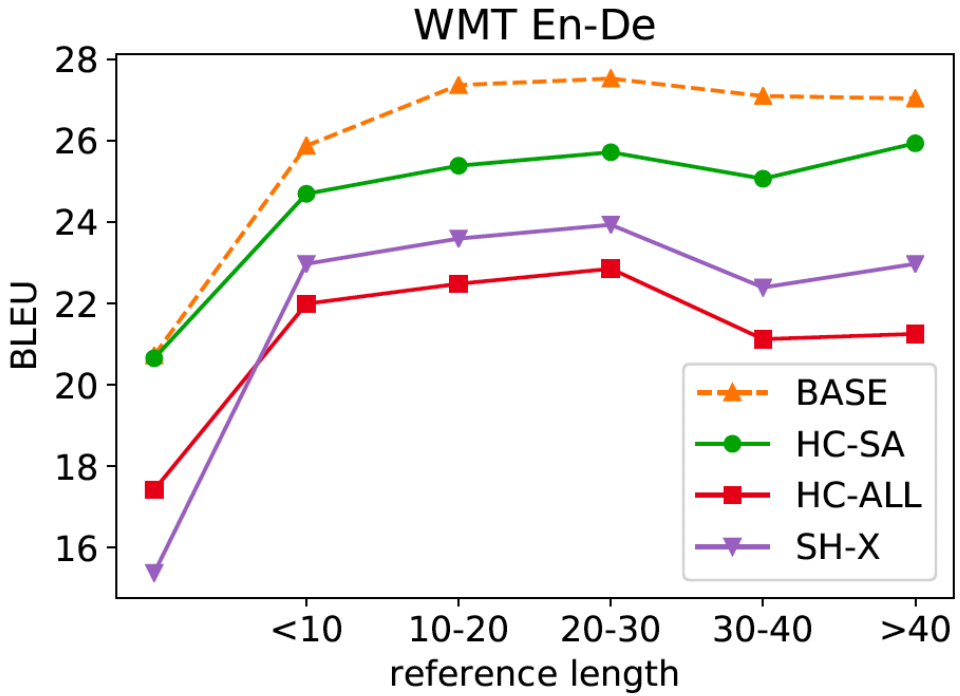


Figure 4: BLEU difference vs. BASE as a function of reference length on the WMT14 En-De test set. When cross attention is hard-coded (HC-ALL), the BLEU gap worsens as reference length increases.

Hard-coded Gaussian 会不会带来隐藏的问题？

Error type	BASE	HC-SA	HC-ALL
np-agreement	**54.2**	53.5	53.5
subj-verb-agreement	**87.5**	85.8	82.5
subj-adequacy	**87.3**	85.0	80.3
polarity-particle-nicht-del	**94.0**	91.4	83.2
polarity-particle-kein-del	**91.4**	88.3	79.9
polarity-affix-del	**91.6**	90.8	83.1
polarity-particle-nicht-ins	**92.6**	92.5	89.8
polarity-particle-kein-ins	94.8	96.7	**98.7**
polarity-affix-ins	**91.9**	90.6	84.3
auxiliary	**89.1**	87.5	85.6
verb-particle	**74.7**	72.7	70.2
compound	88.1	**89.5**	80.5
transliteration	97.6	**97.9**	93.4

Table 4: Accuracy for each error type in the LingEval97 contrastive set. Hard-coding self-attention results in slightly lower accuracy for most error types, while more significant degradation is observed when hard-coding self and cross attention. We refer readers to Sennrich (2017) for descriptions of each error type.

an example is shown here:


```
{
  "source": "Prague Stock Market falls to minus by the end of the trading day",
  "reference": "Die Prager Börse stürzt gegen Geschäftsschluss ins Minus.",
  "origin": "newstest2009.1",
  "errors": [
    {
      "type": "np_agreement",
      "contrastive": "Der Prager Börse stürzt gegen Geschäftsschluss ins Minus.",
      "distance": 2,
      "frequency": 2020
```

尝试更极端的情况

	Original	Conv (window=3)	Indexing
En-De	30.3	30.1	29.8
En-Ro	32.4	32.3	31.4

Table 5: Comparison of three implementations of HC-SA. Truncating the distribution to a three token span has little impact, while removing the weights altogether slightly lowers BLEU.

Conv： [0.242, 0.399, 0.242]
Indexing: [1, 0, 0]

如果允许保留更多的heads

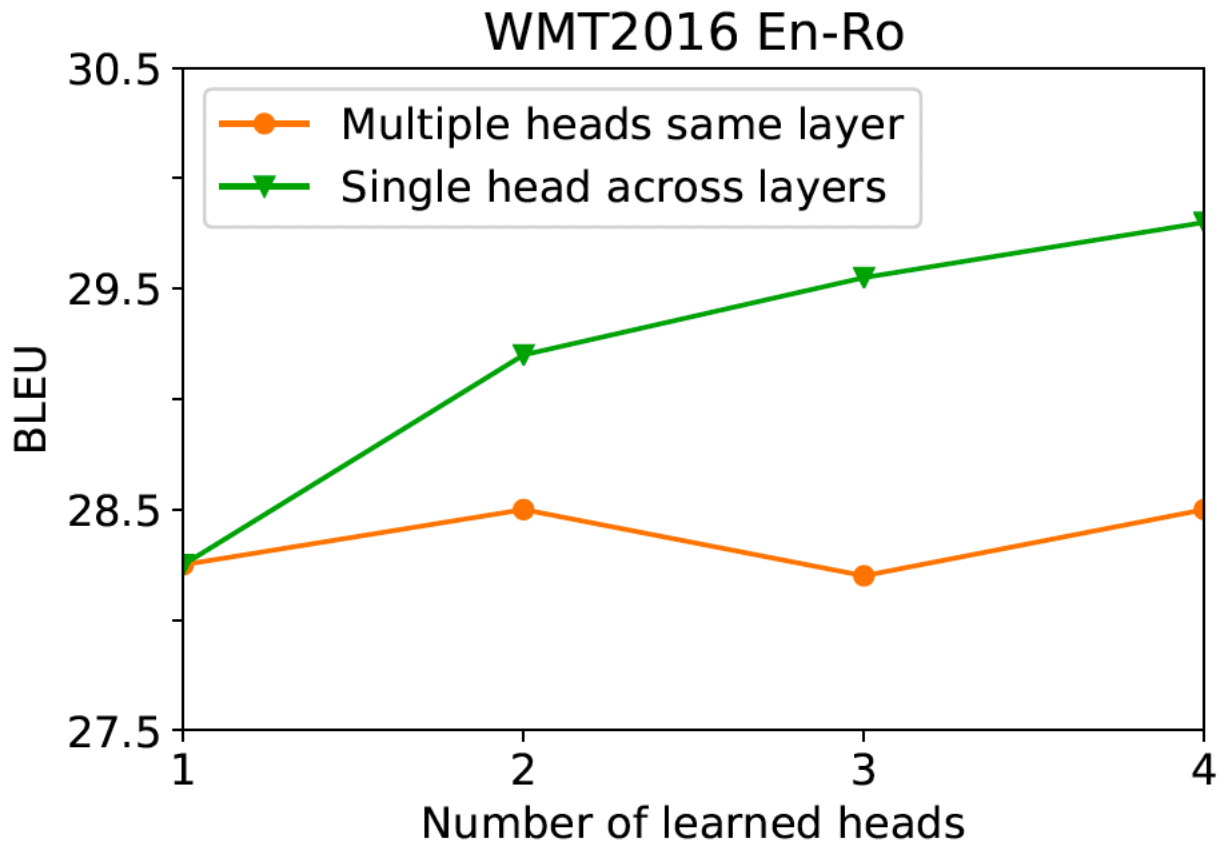


Figure 7: Adding more cross attention heads in the same layer helps less than adding individual heads across different layers.