前沿技术讲习班
Advanced Technology Tutorial

Text Generation: From the Perspective of Interactive Inference

张家俊

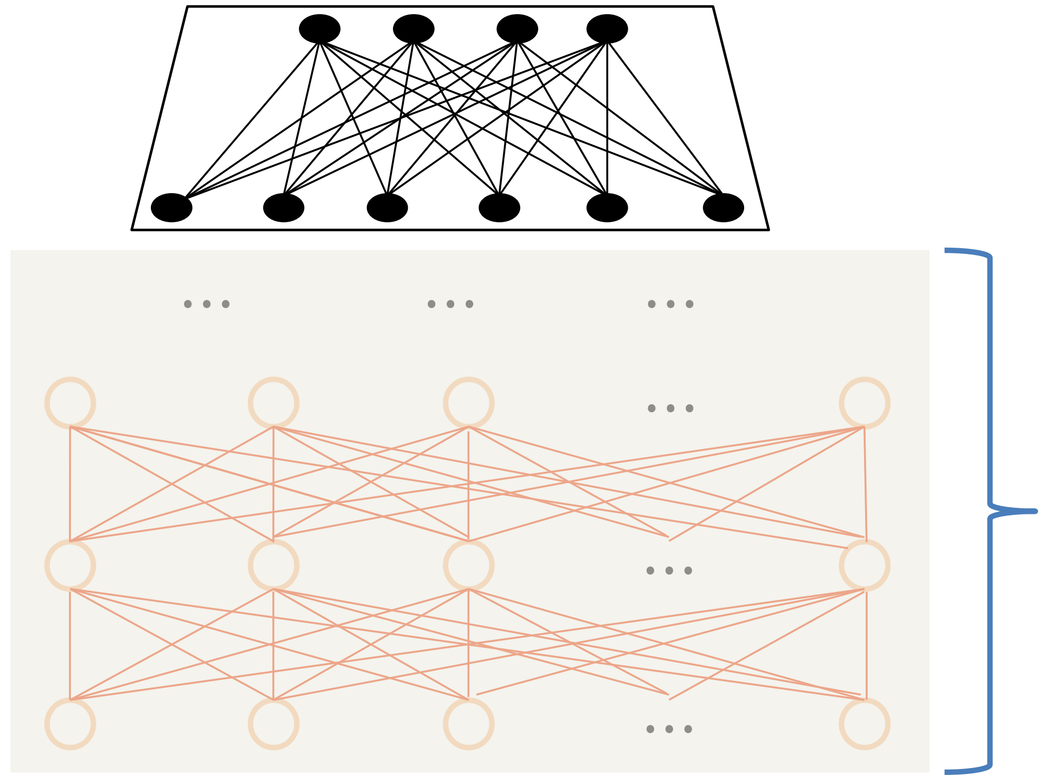
模式识别国家重点实验室
中国科学院自动化研究所
jjzhang@nlpr.ia.ac.cn
www.nlpr.ia.ac.cn/cip/jjzhang.htm

Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Two Tasks
- Summary and Future Challenges

BERT: Bidirectional Understanding

Linear Classification



Representation Learning

12 Layers

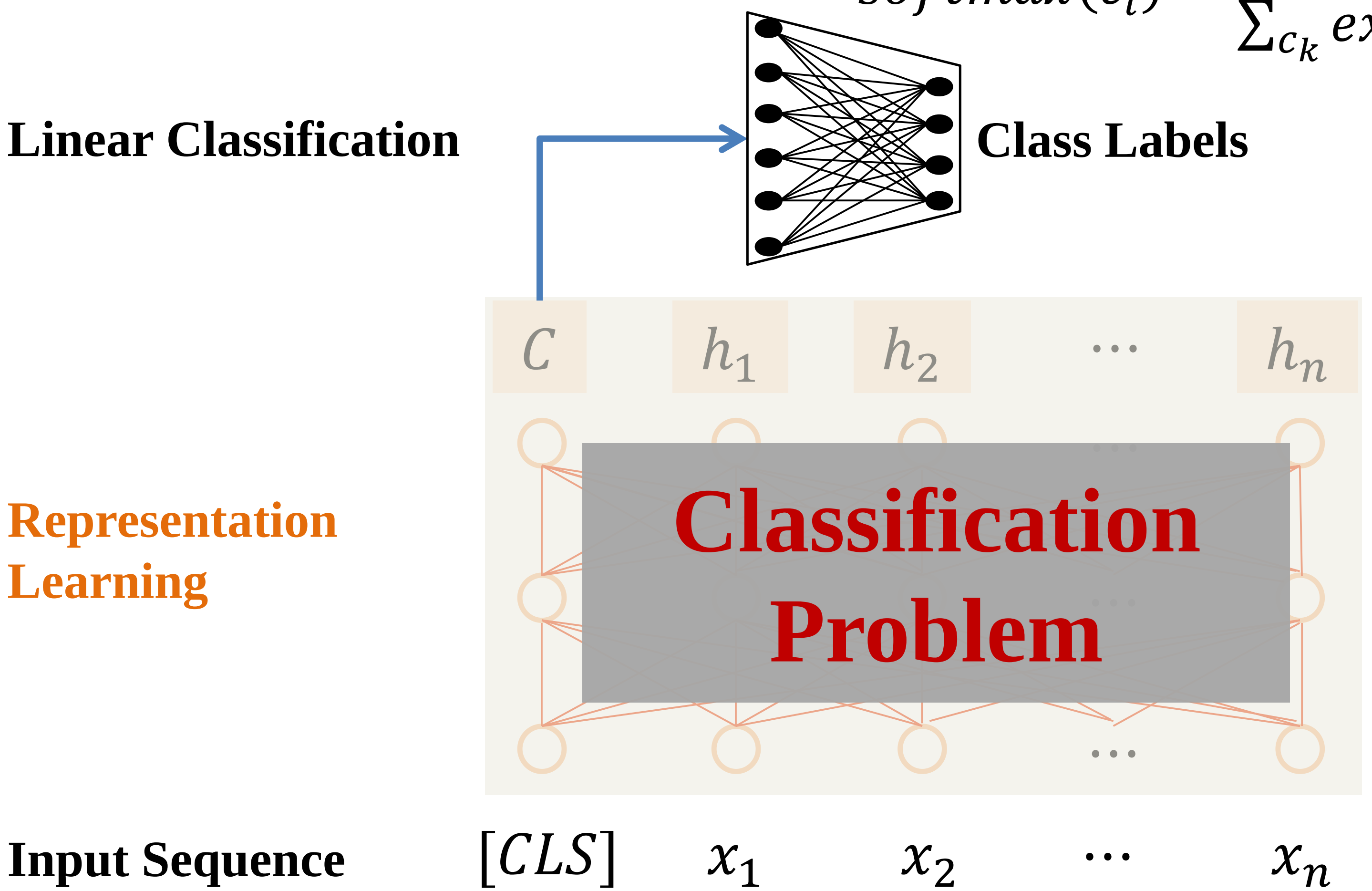
110M Params

Input Sequence x_1 x_2 x_3 $\cdots$ x_n

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *In NAACL-HLT* 2019 (**Best Paper**).

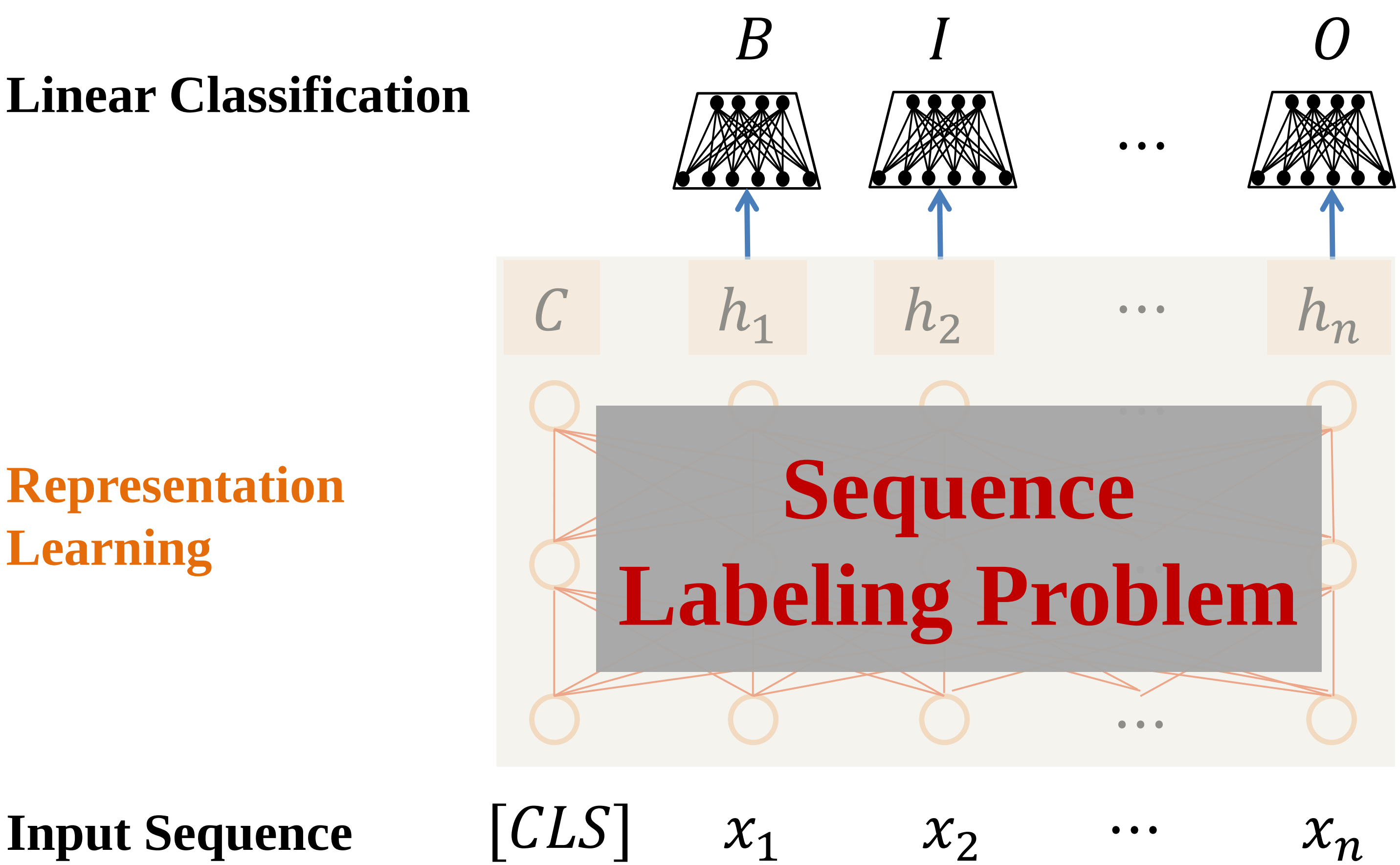
BERT for Classification

$$softmax(c_i) = \frac{exp(c_i)}{\sum_{c_k} exp(c_k)}$$



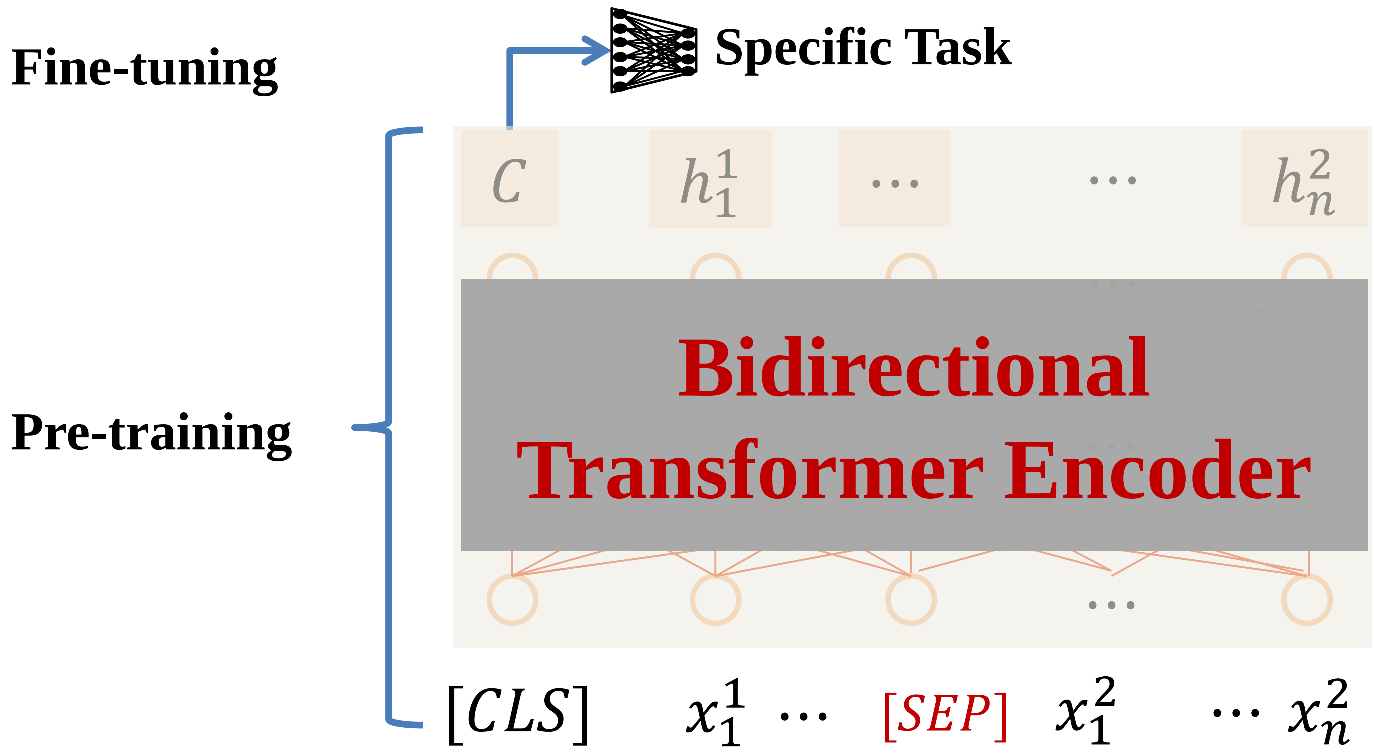
BERT for Sequence Labeling

$$softmax(c_i) = \frac{exp(c_i)}{\sum_{c_k} exp(c_k)}$$



Reasons behind BERT Success

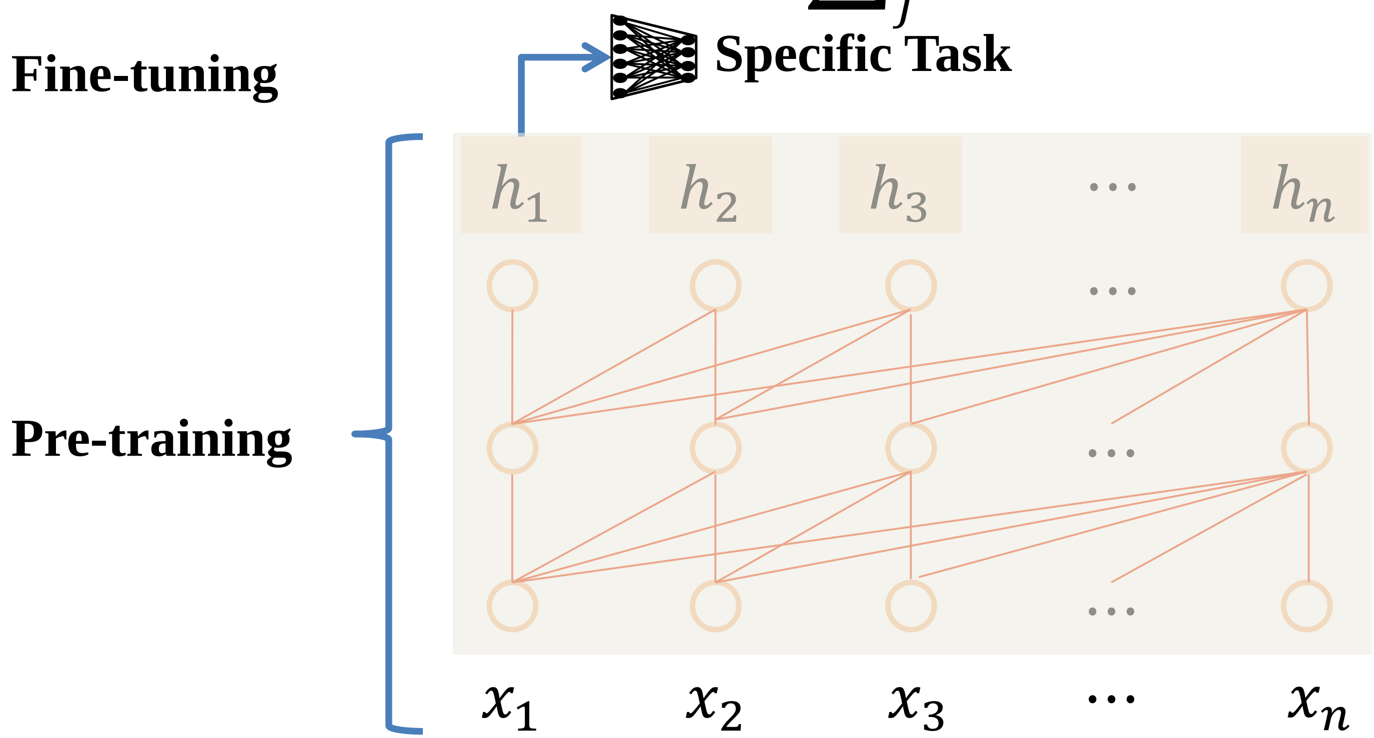
- **Corpus**: 2.5B-word Wiki and 800M-word Books
- **Architecture**: Pre-training and **Fine-tuning** Same Model
- **Model**: **Deep Bidirectional Transformer Encoder**
- **Optimization**: **Masked LM** and **Next Sentence Prediction**



BERT vs. GPT (Generative Pre-trained Transformer)

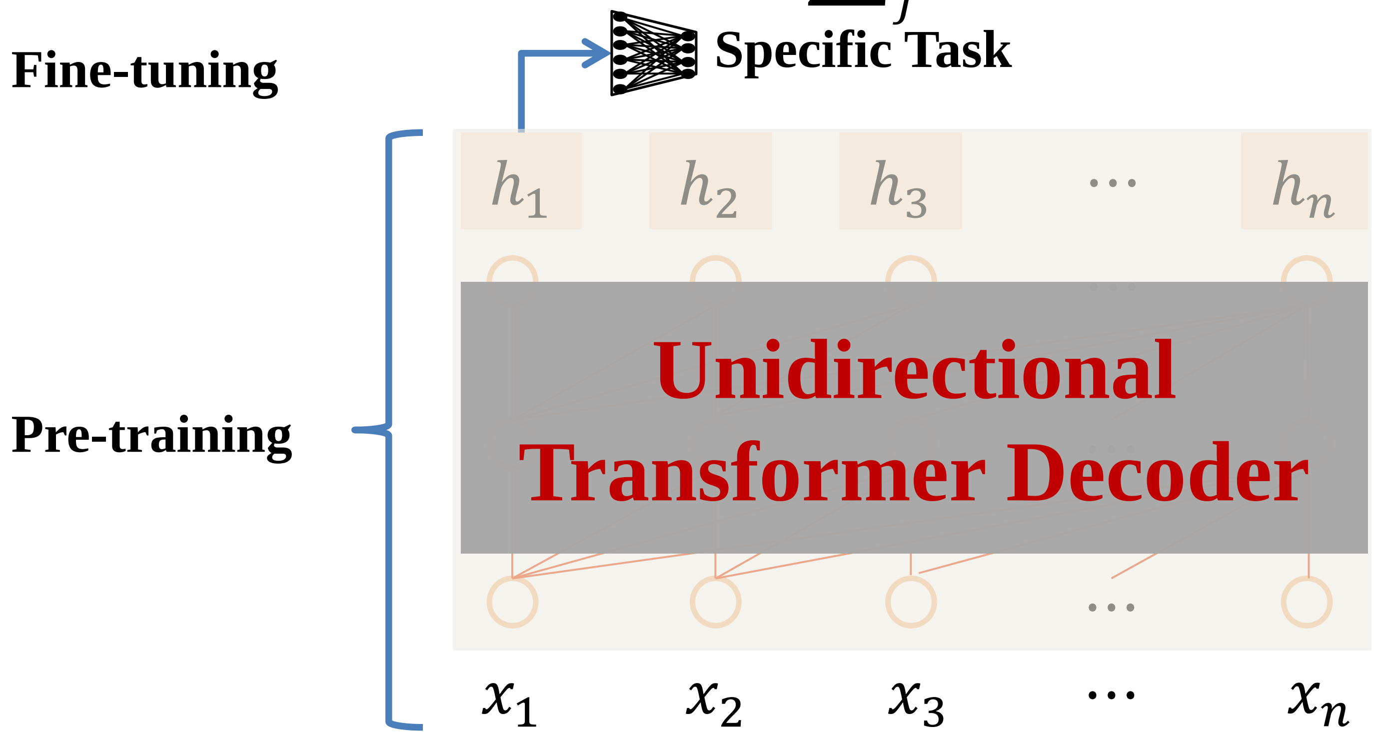
- **Architecture**: Pre-training and **Fine-tuning** Same Model
- **Model**: **Deep Unidirectional Transformer Decoder**
- **Optimization**: **Traditional Language Model**
- **Corpus**: 800M-word Books

$$\sum_j \{\log p(x_j | x_{<j}, \vec{\theta}_{Transformer})\}$$



BERT vs. GPT
(Generative Pre-trained Transformer)

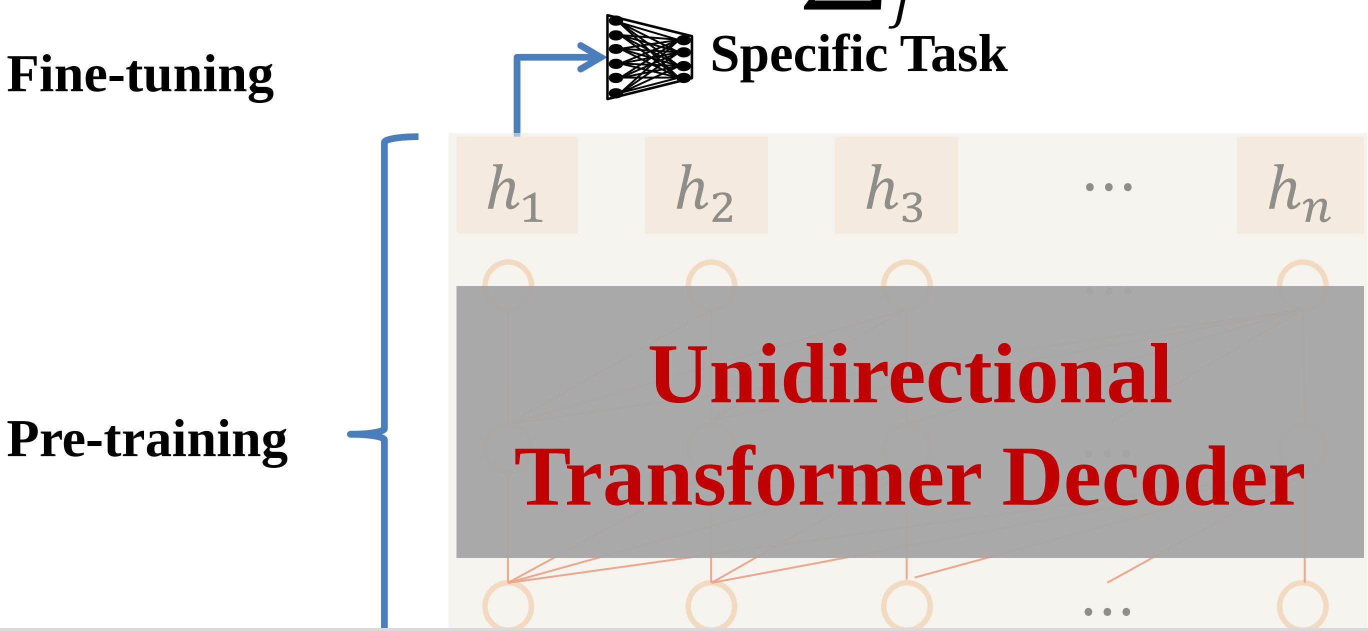
- **Architecture**: Pre-training and **Fine-tuning** Same Model
- **Model**: **Deep Unidirectional Transformer Decoder**
- **Optimization**: **Traditional Language Model**
- **Corpus**: 800M-word Books $\sum_j \{\log p(x_j | x_{<j}, \vec{\theta}_{Transformer})\}$



BERT vs. GPT (Generative Pre-trained Transformer)

- **Architecture**: Pre-training and **Fine-tuning** Same Model
- **Model**: **Deep Unidirectional Transformer Decoder**
- **Optimization**: **Traditional Language Model**
- **Corpus**: 800M-word Books

$$\sum_j \{\log p(x_j | x_{<j}, \vec{\theta}_{Transformer})\}$$



Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by Generative Pre-Training. Technical report, OpenAI.

BERT Ablation Study

	Dev Set				
Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Left-to-Right LM

Fine-tuning with BiLSTM

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

The more Layers The Better

BERT Ablation Study

Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
	Dev Set				
BERT$_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Left-to-Right LM

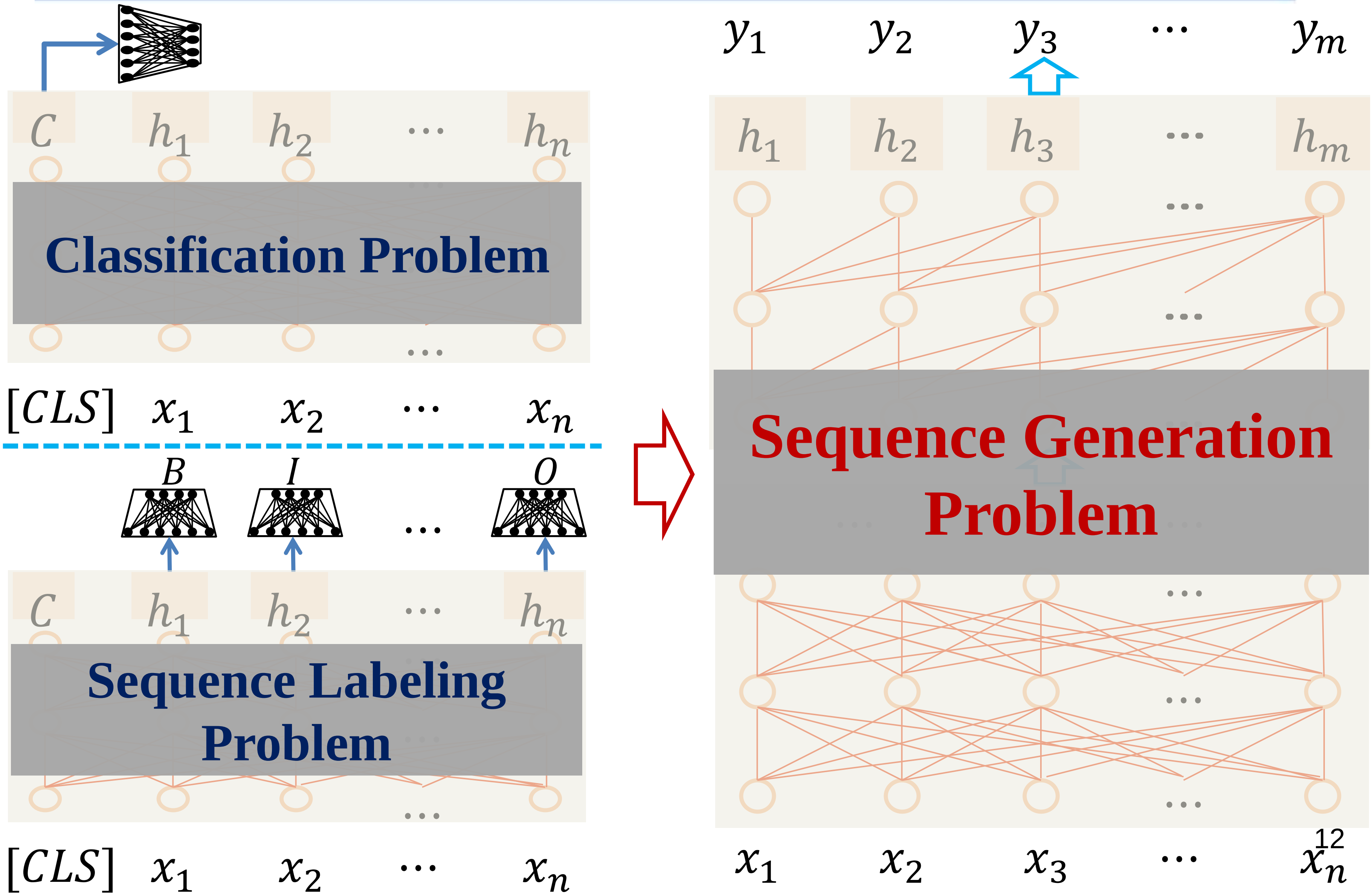
Fine-tuning with BiLSTM

Bidirectional Encoder is the Key!

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

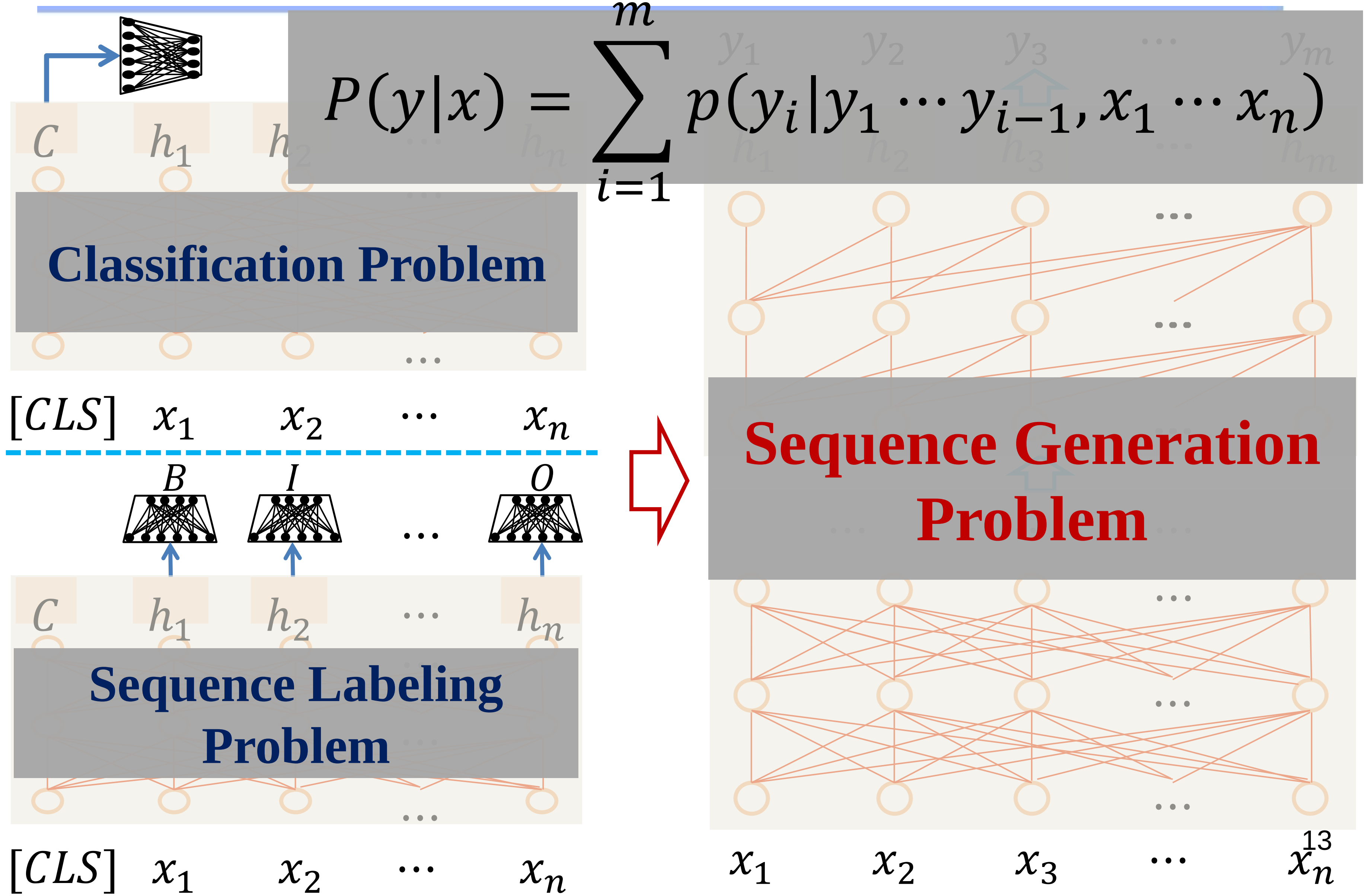
The more Layers The Better

From Understanding to Generation



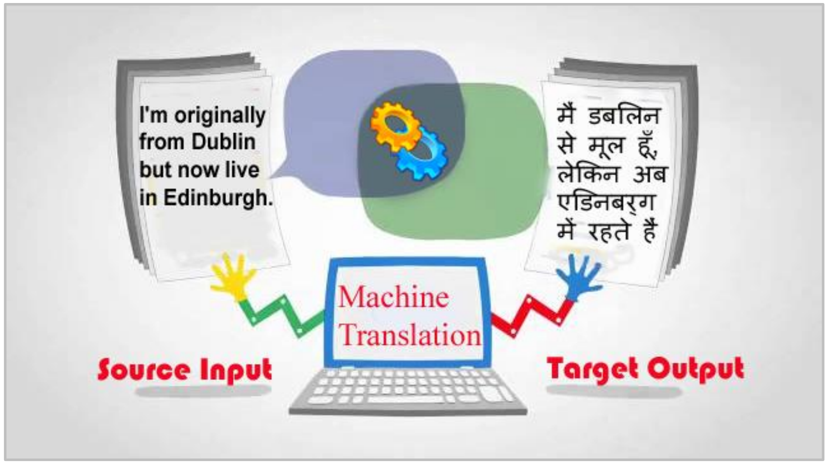
From Understanding to Generation

$$P(y|x) = \sum_{i=1}^{m} p(y_i | y_1 \cdots y_{i-1}, x_1 \cdots x_n)$$

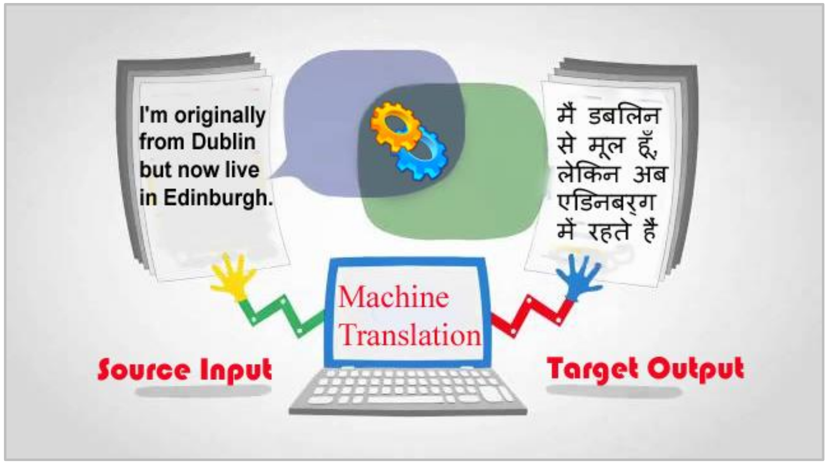


Text Generation

Text Generation

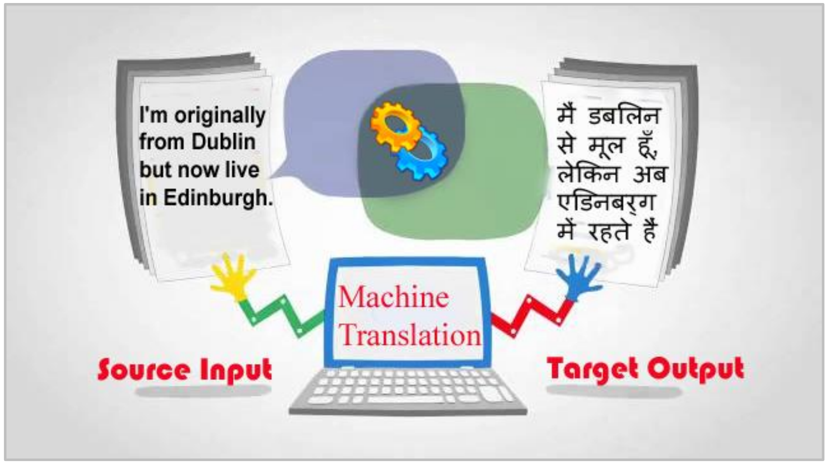


Text Generation



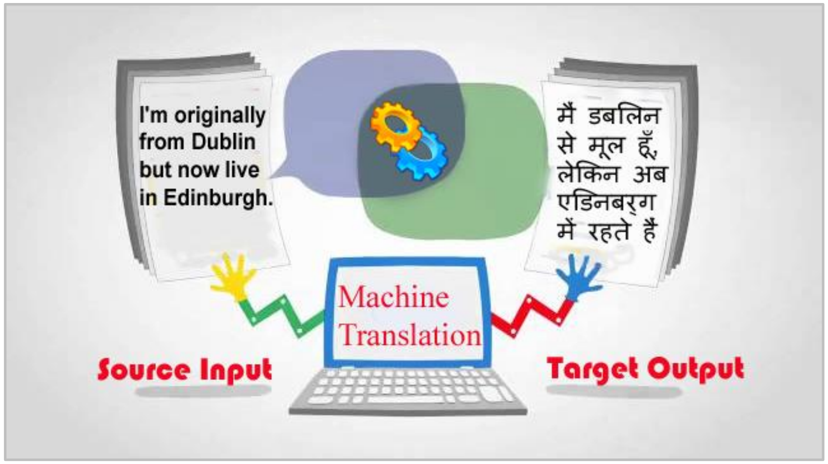
机器翻译

Text Generation



机器翻译

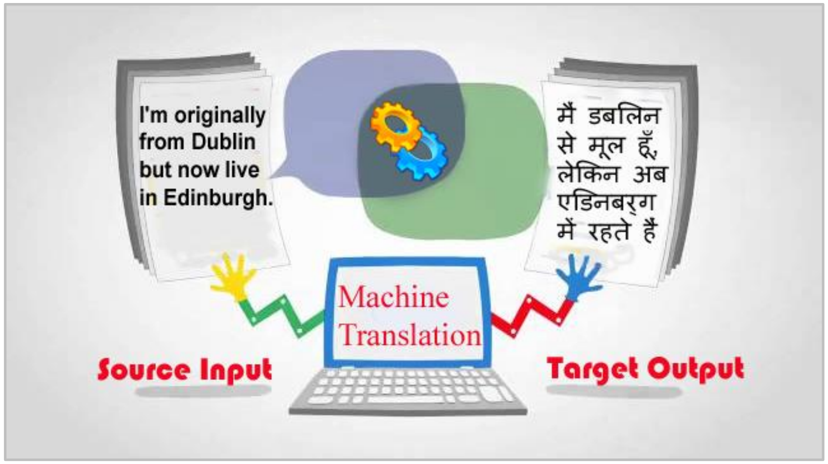
Text Generation



机器翻译

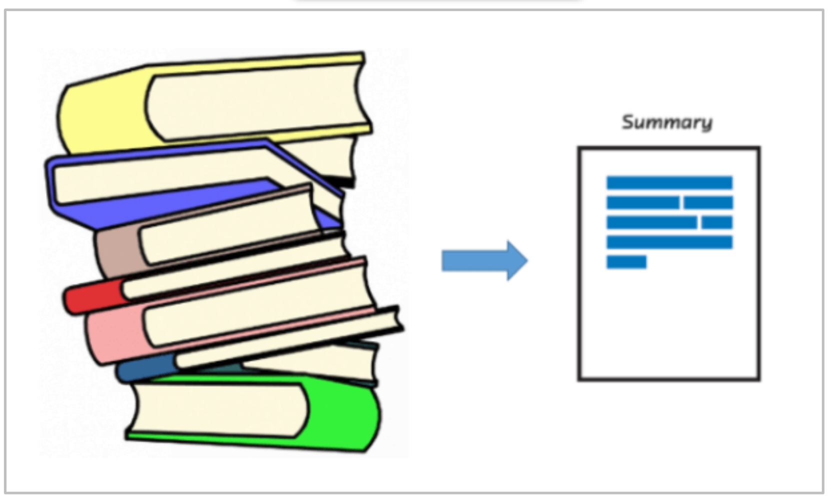
人机对话

Text Generation

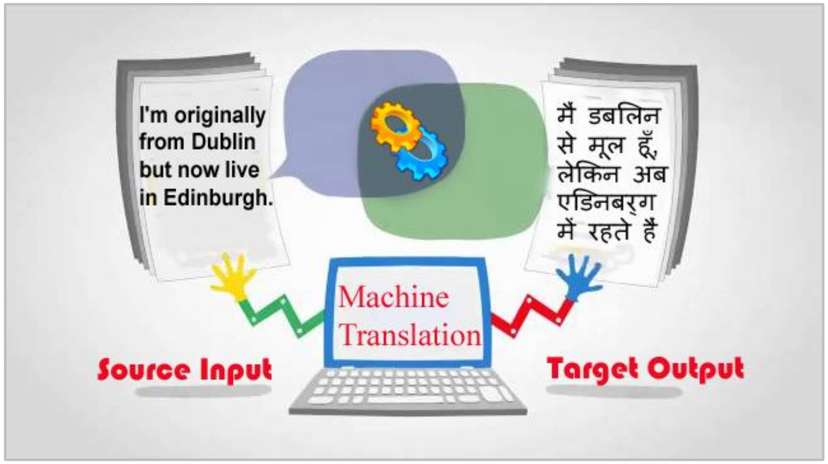


机器翻译

人机对话

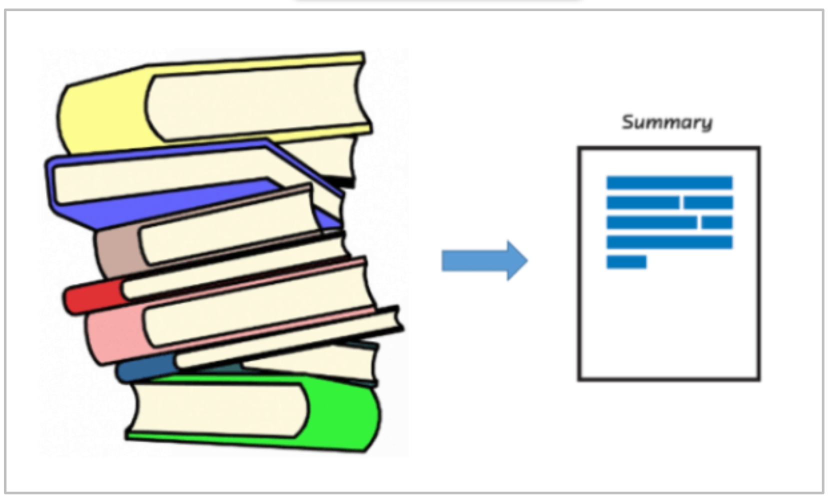


Text Generation



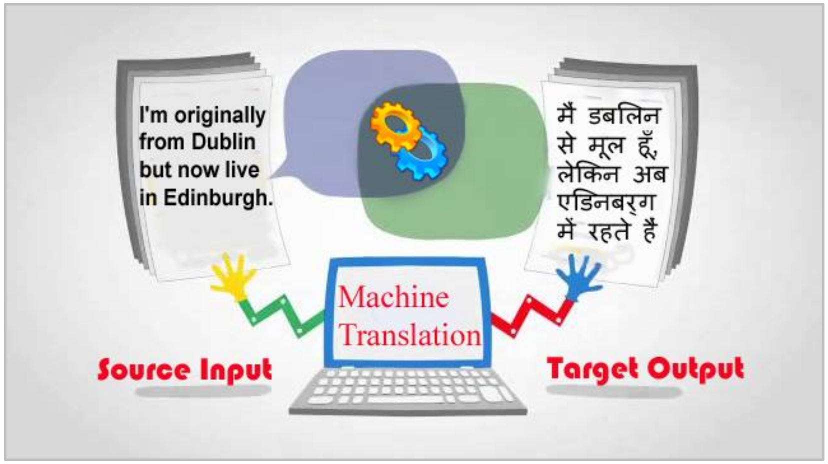
机器翻译

人机对话



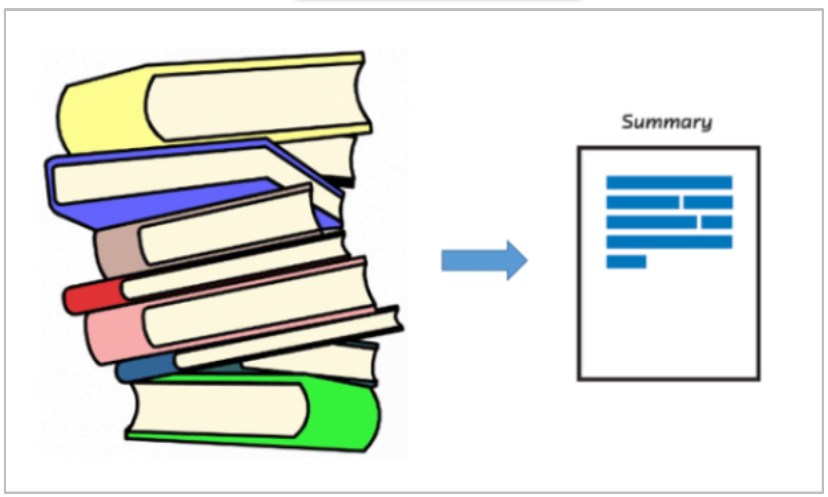
自动摘要

Text Generation

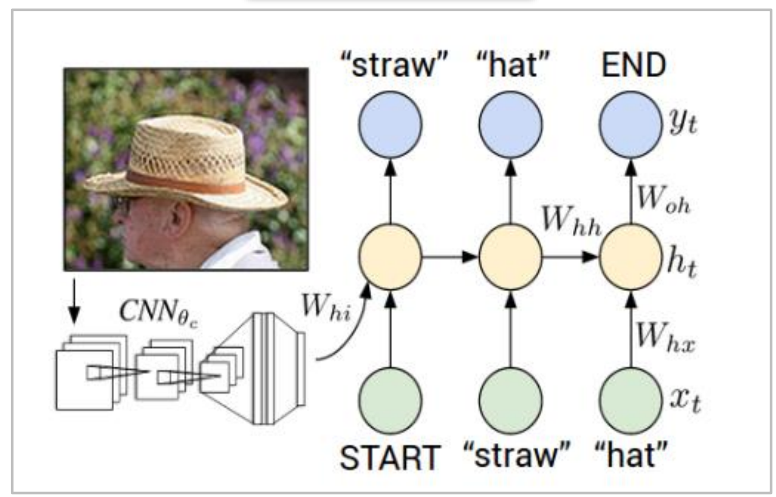


机器翻译

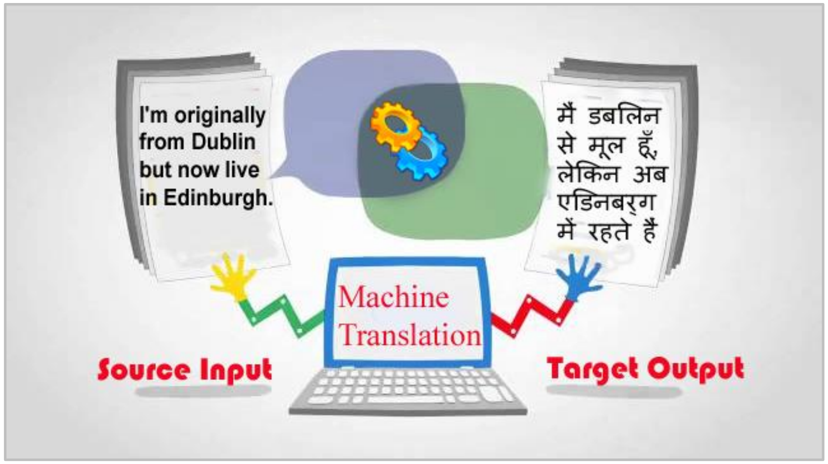
人机对话



自动摘要

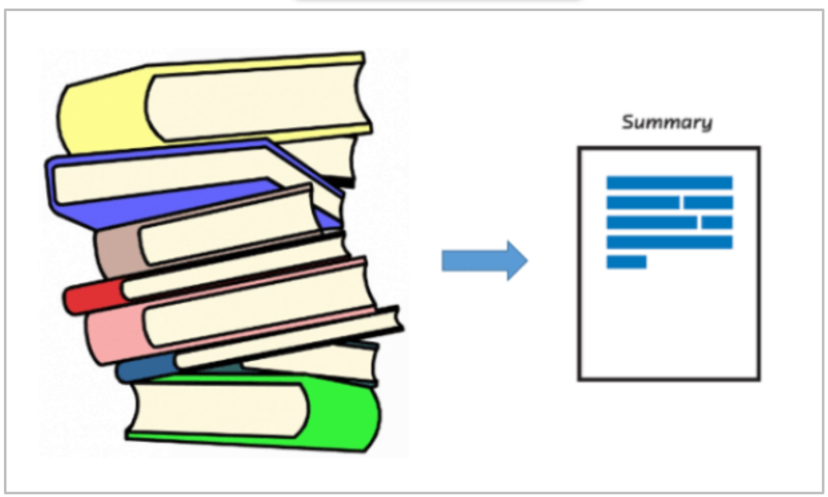


Text Generation

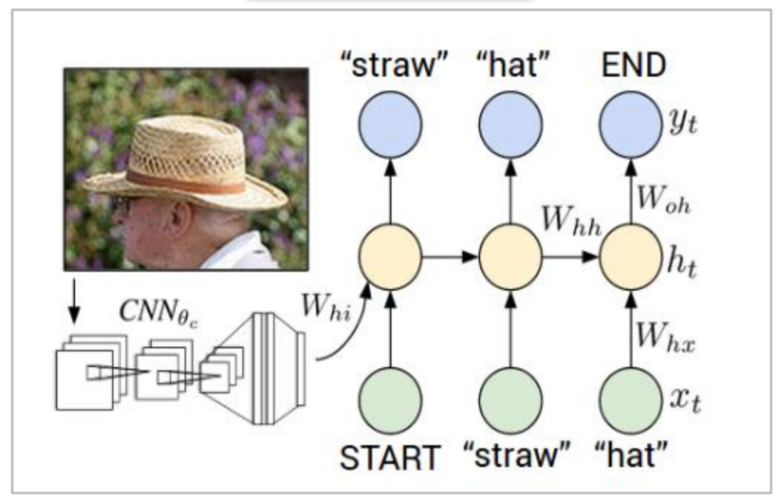


机器翻译

人机对话

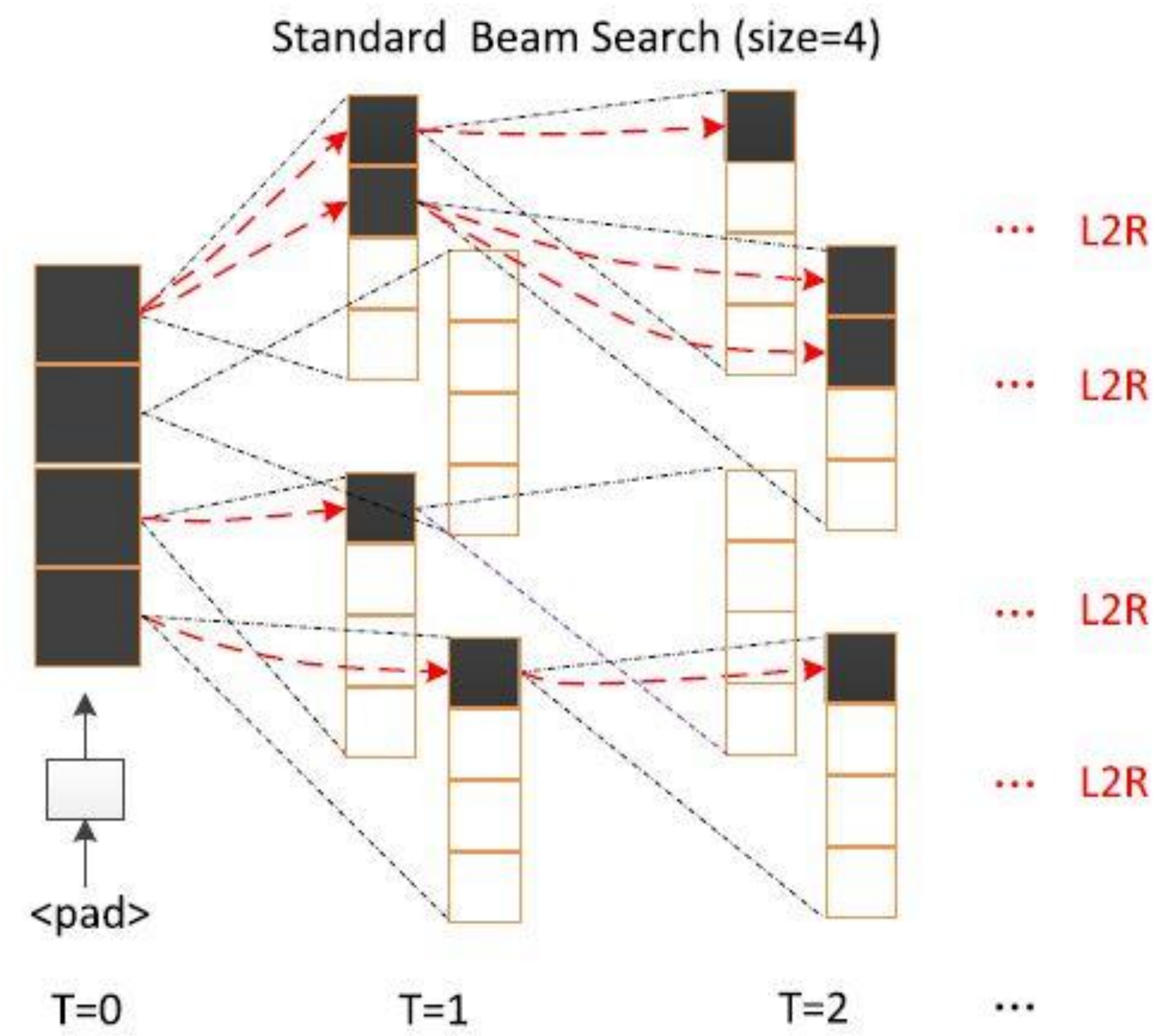


自动摘要

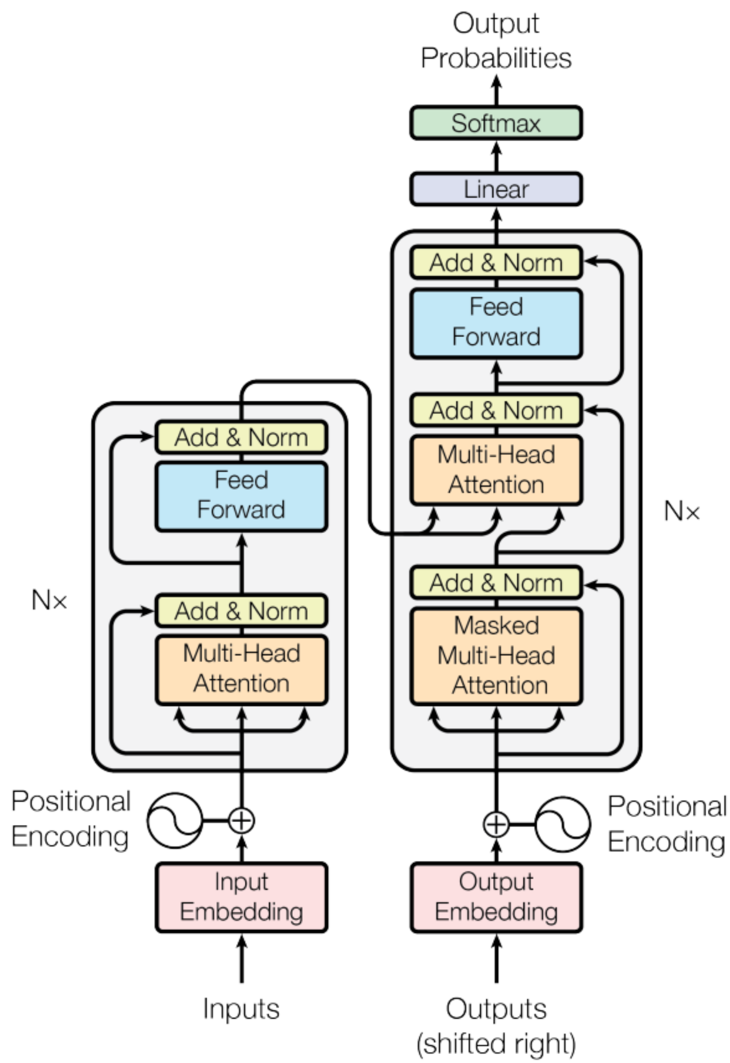


标题生成

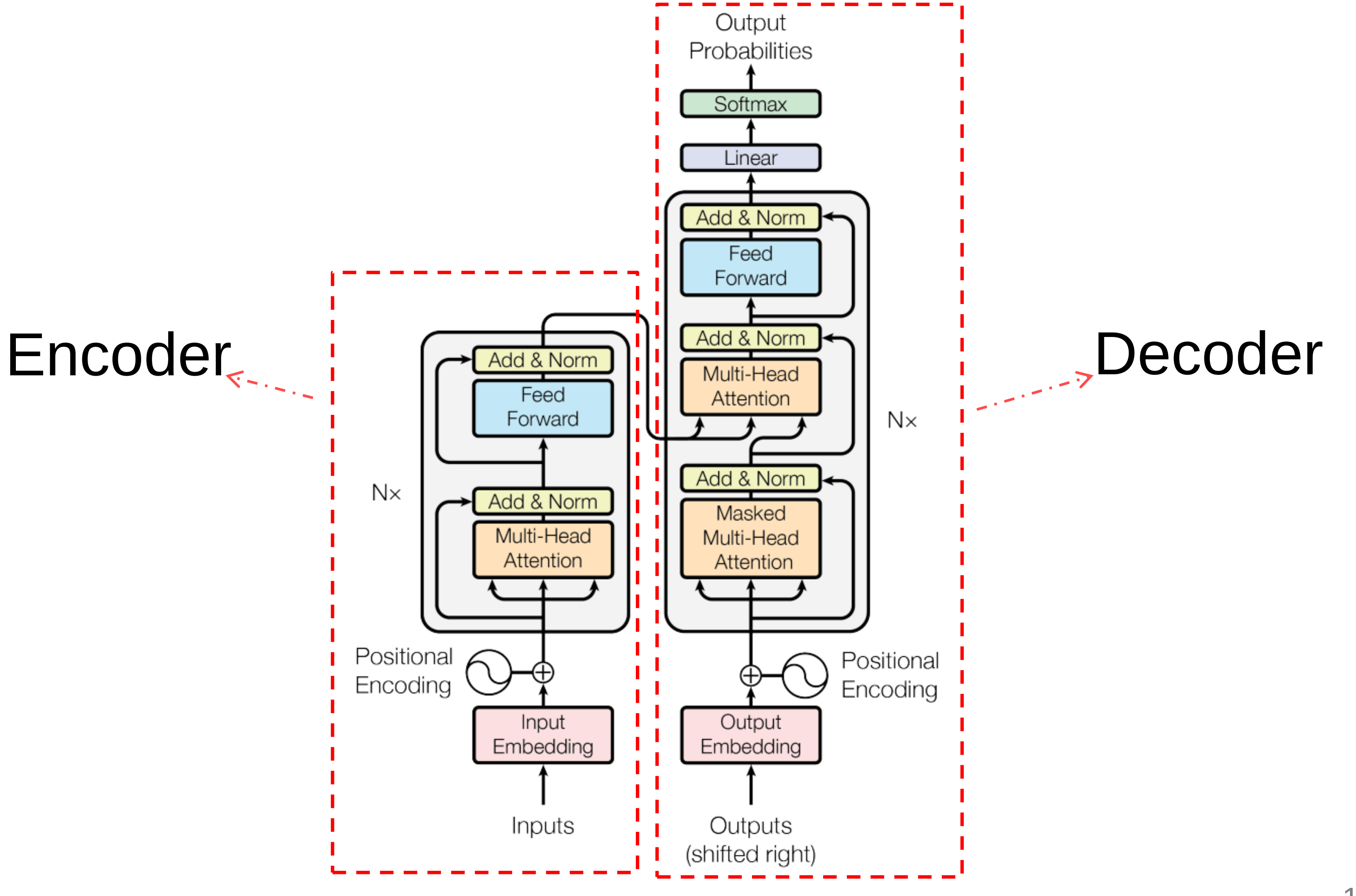
Beam Search for Unidirectional Inference



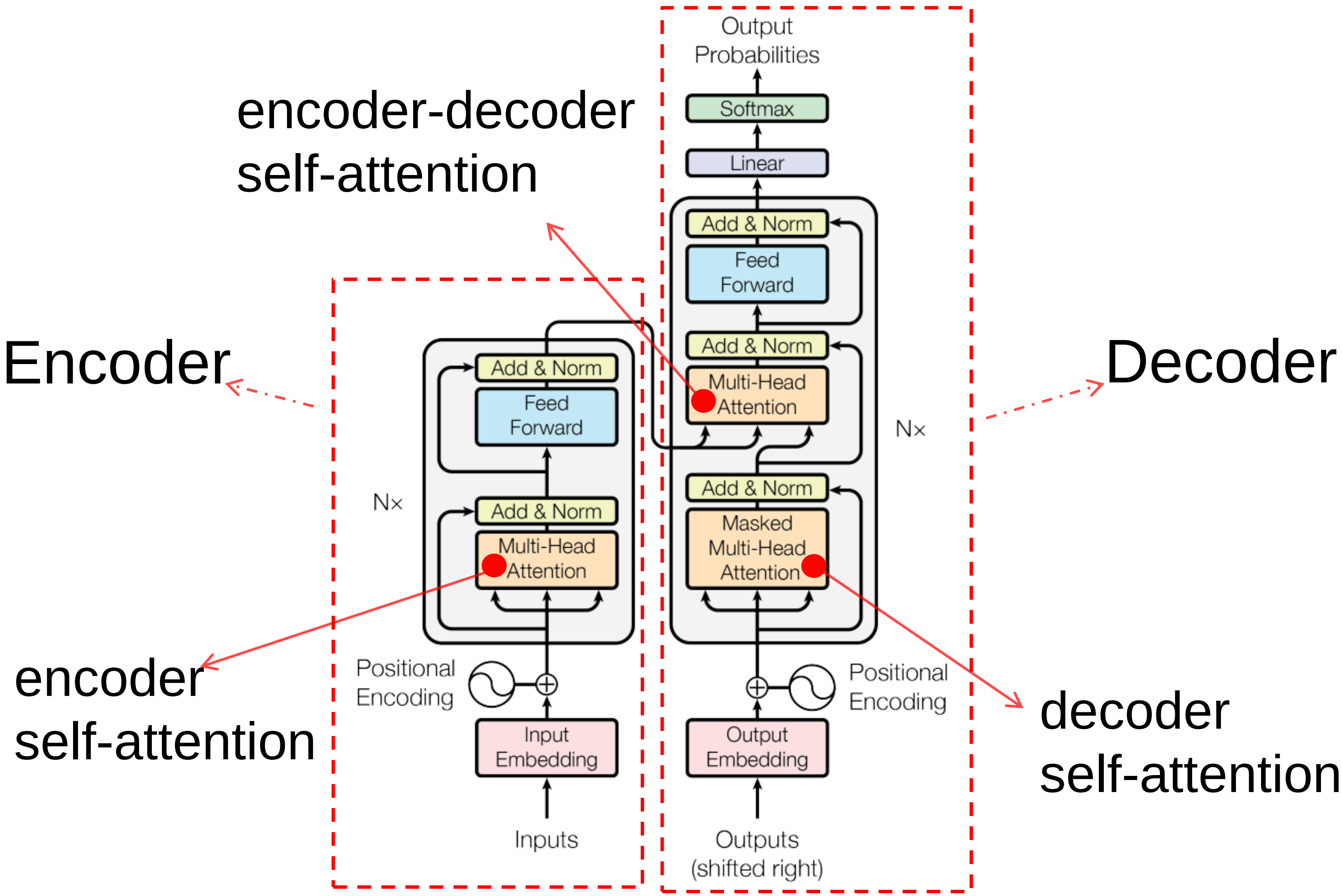
Transformer: Best Unidirectional Text Generation Framework



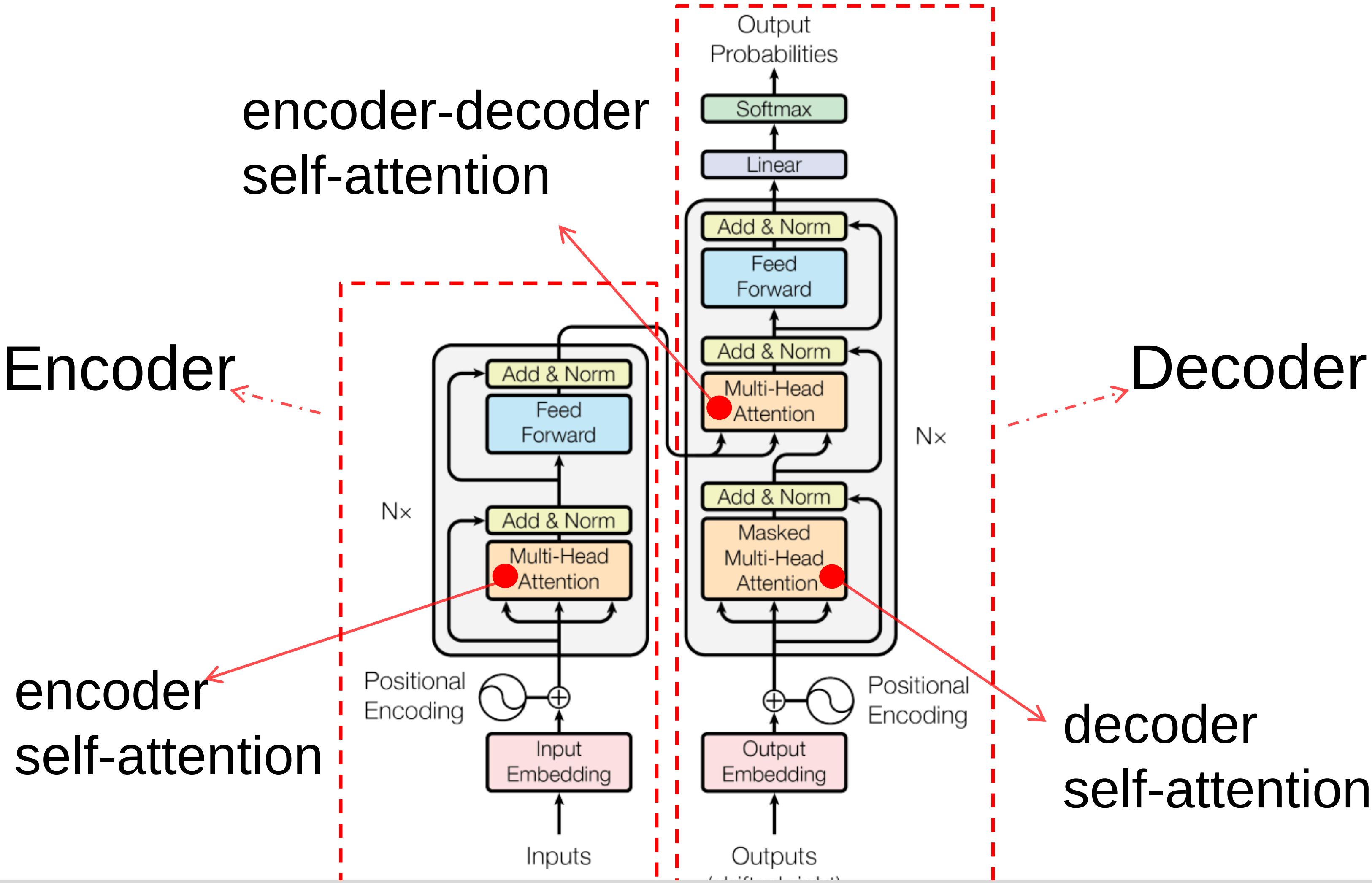
Transformer: Best Unidirectional Text Generation Framework



Transformer: Best Unidirectional Text Generation Framework



Transformer: Best Unidirectional Text Generation Framework

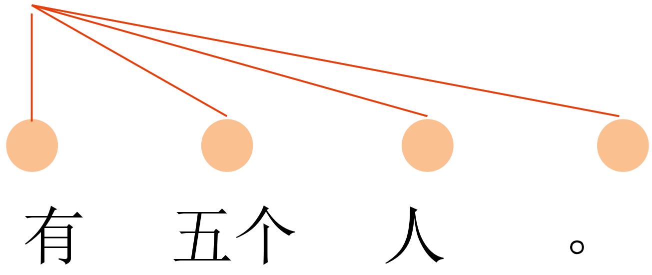


Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NIPS-2017.

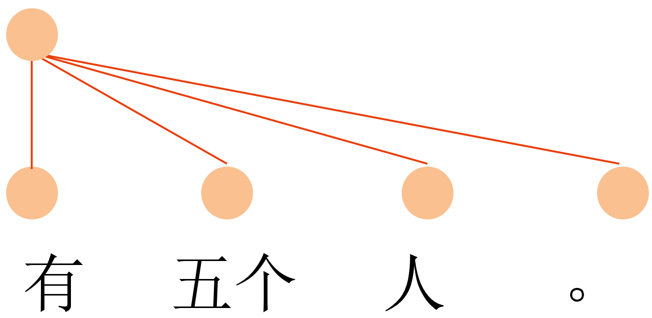
Transformer

有　五个　人　。

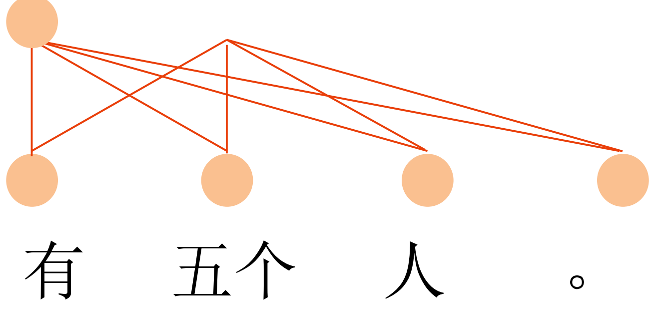
Transformer



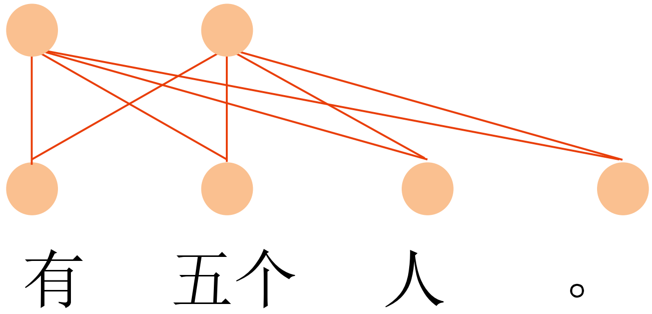
Transformer



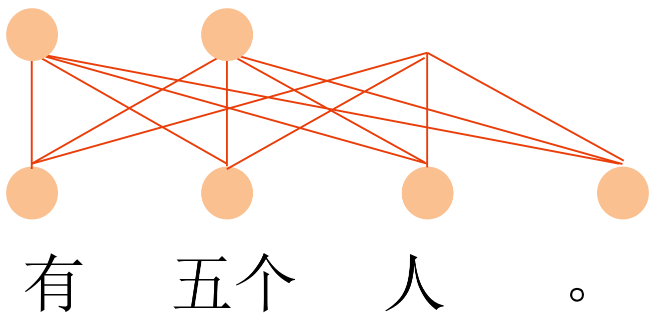
Transformer



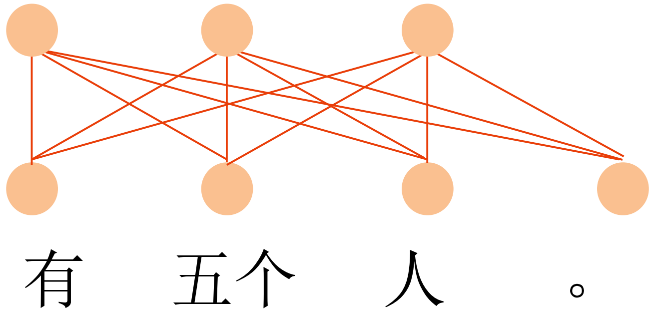
Transformer



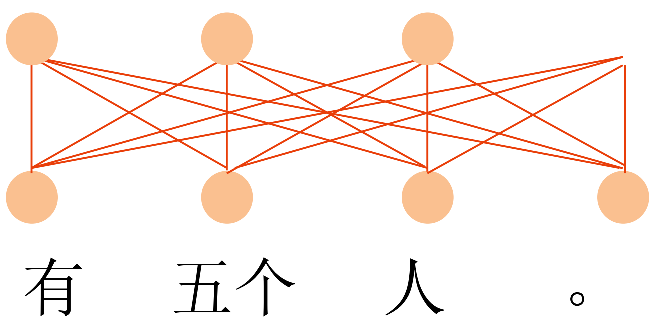
Transformer



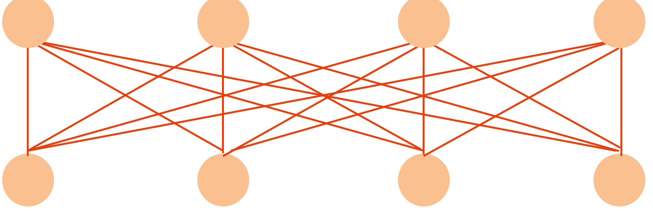
Transformer



Transformer

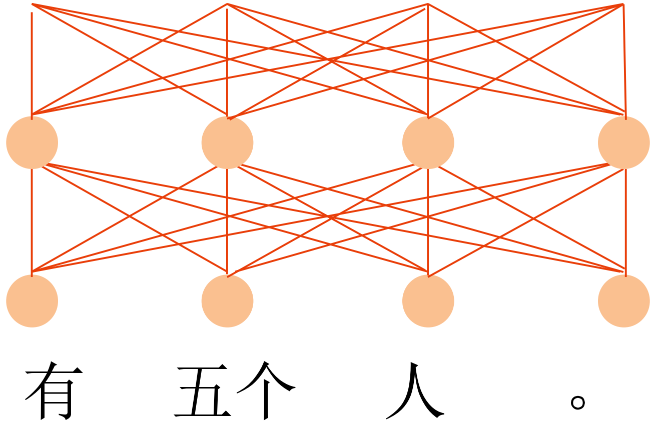


Transformer

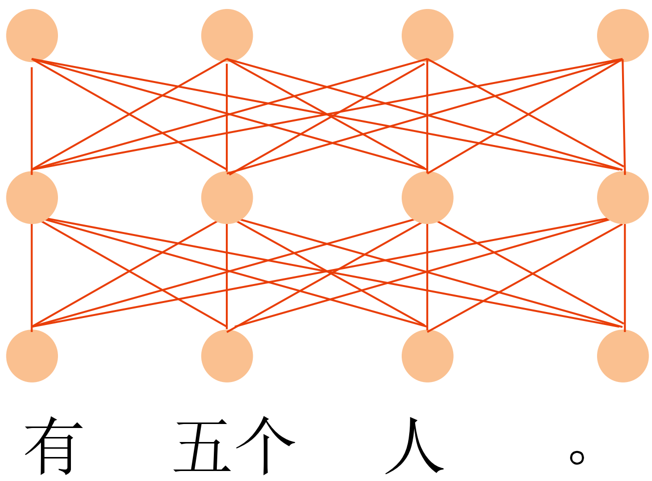


有 五个 人 。

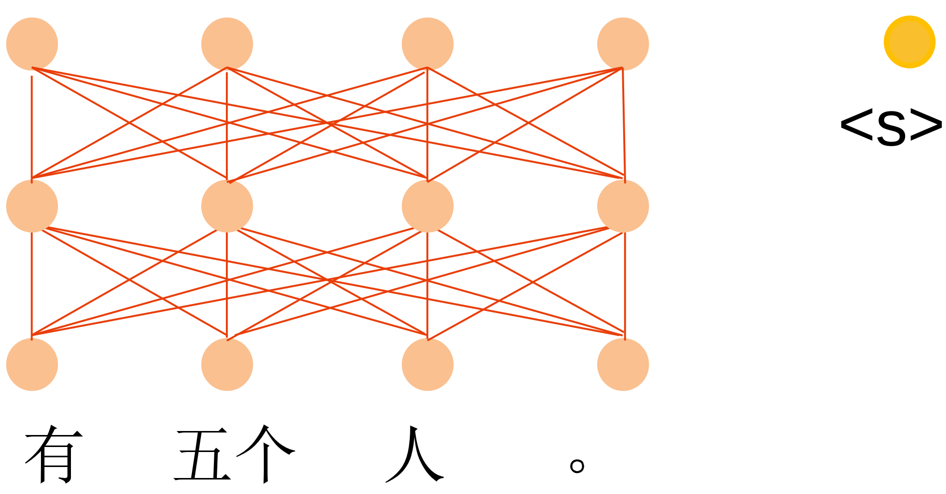
Transformer



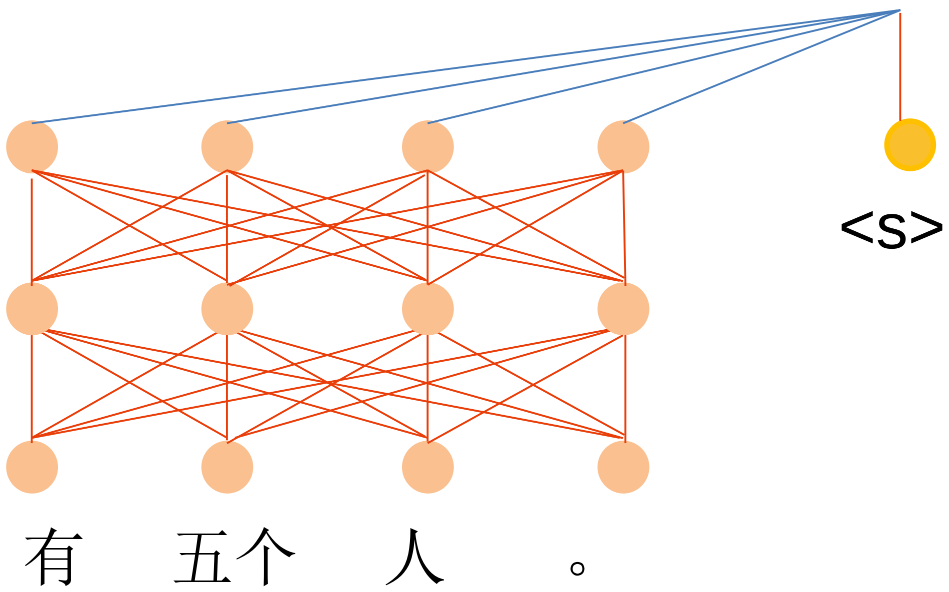
Transformer



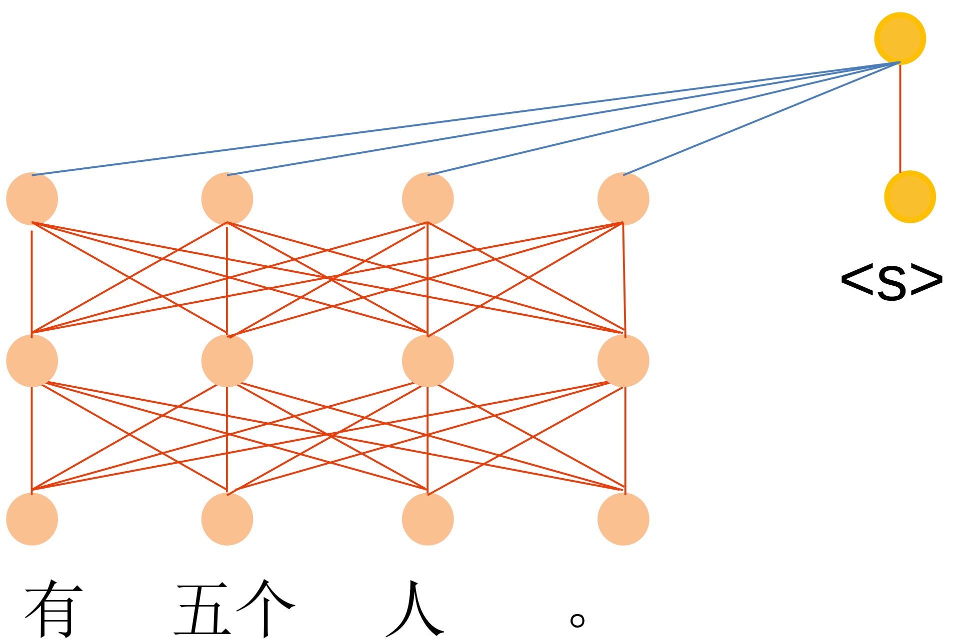
Transformer



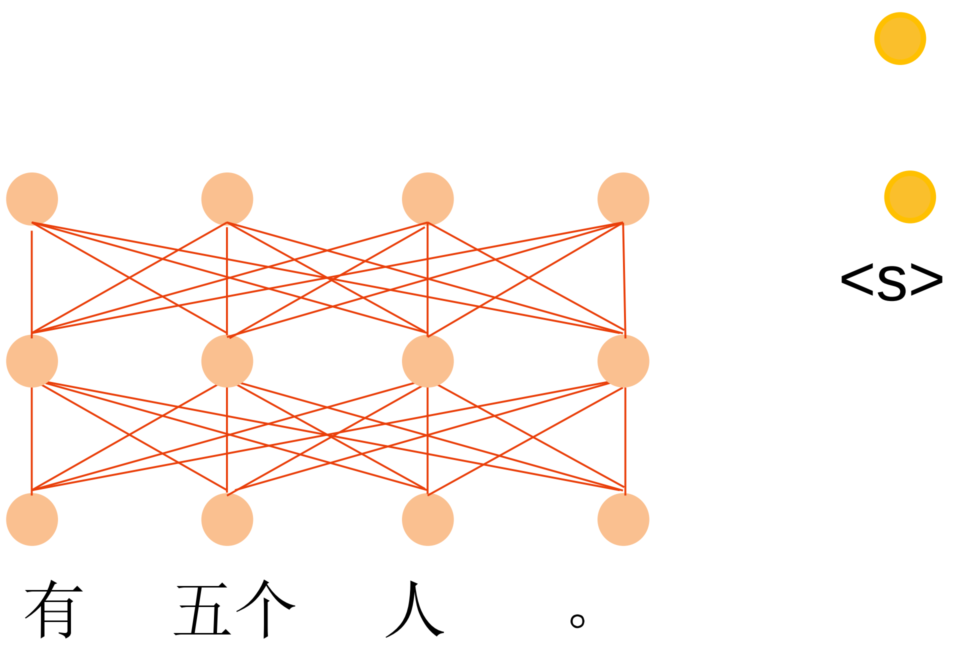
Transformer



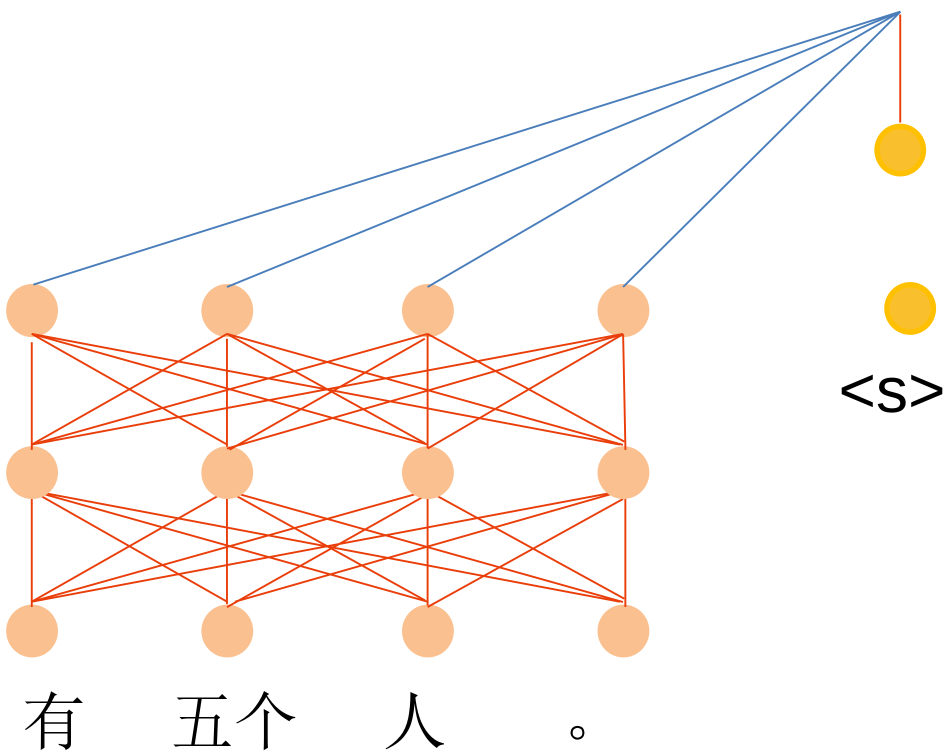
Transformer



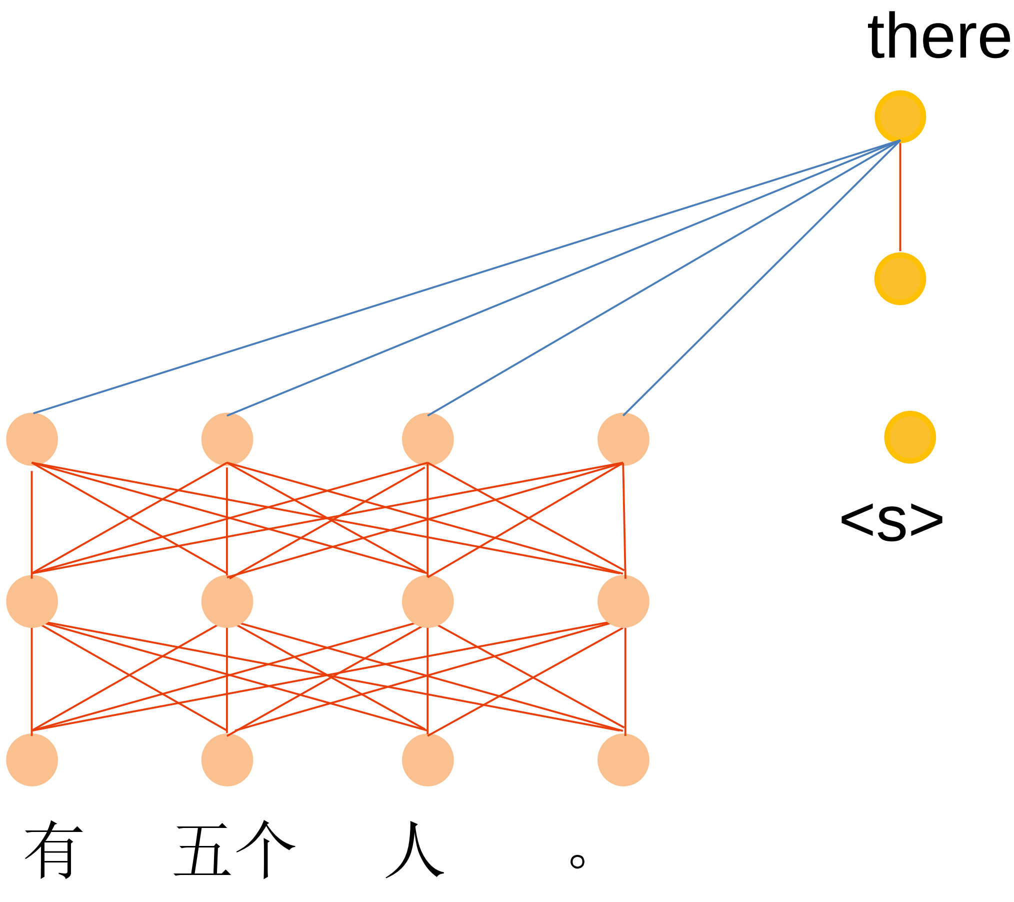
Transformer



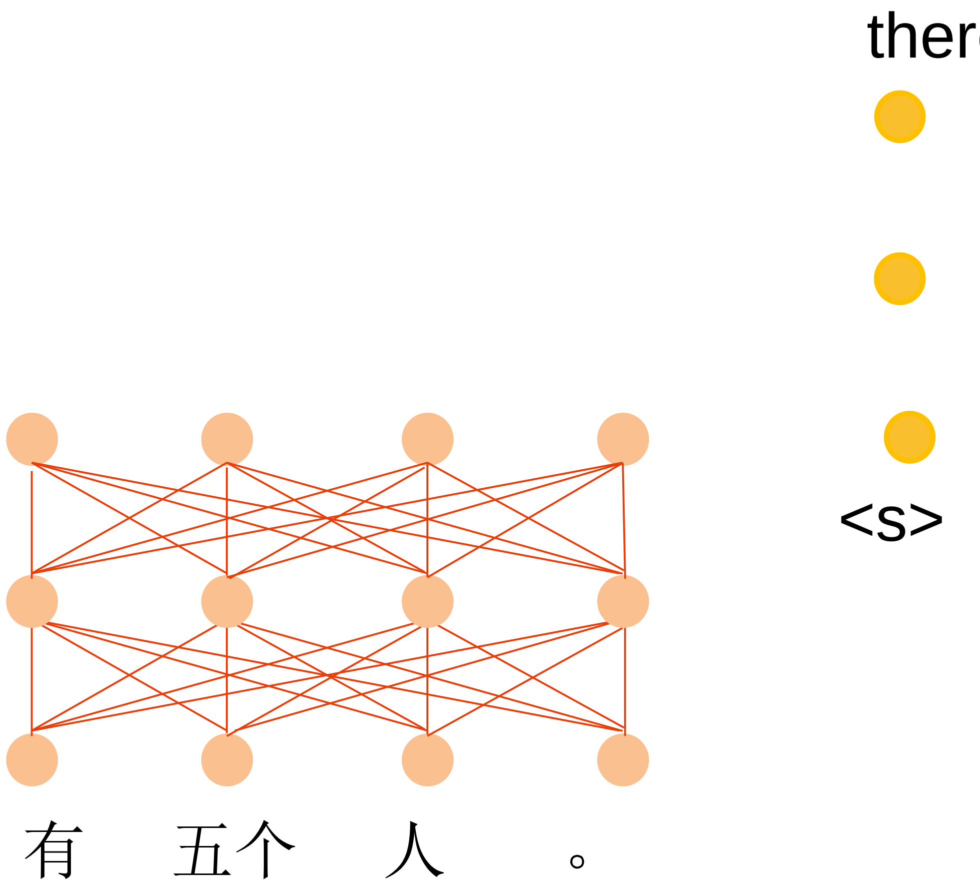
Transformer



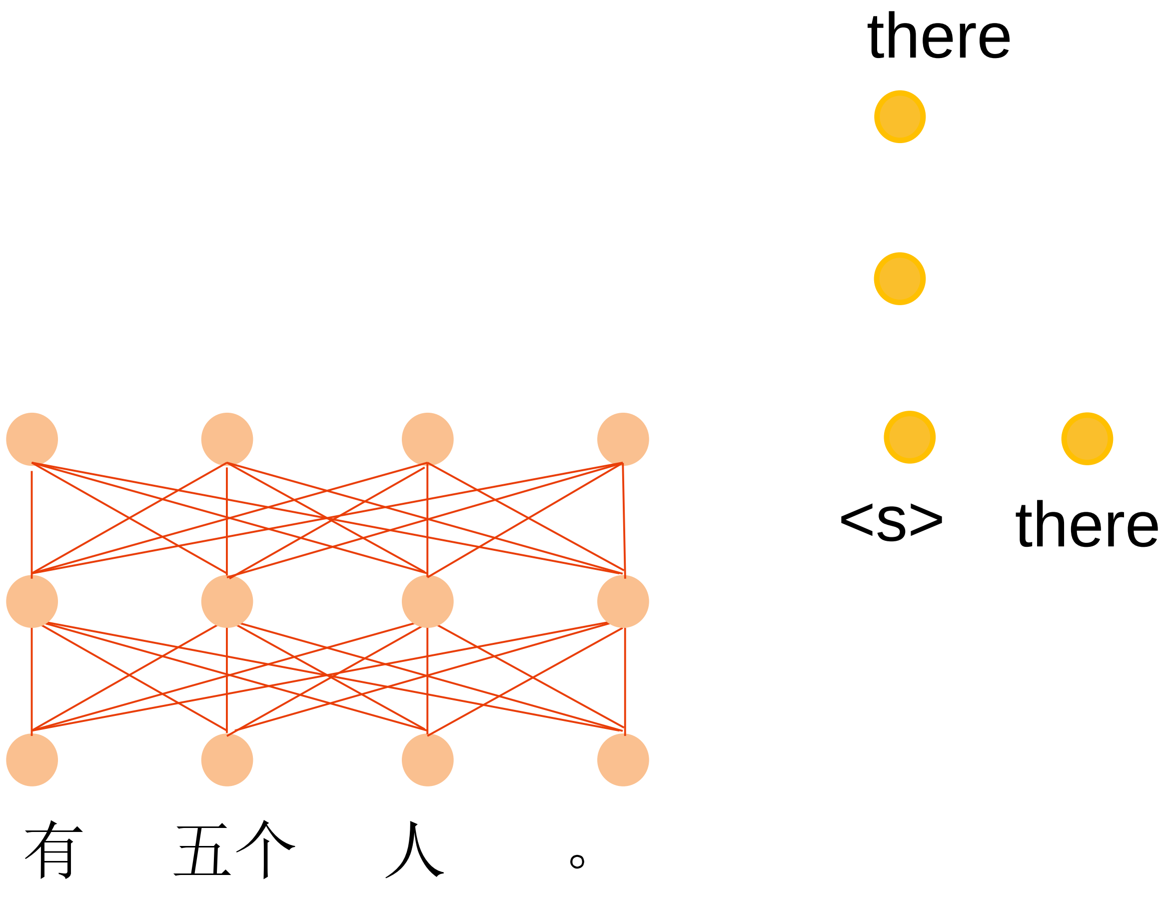
Transformer



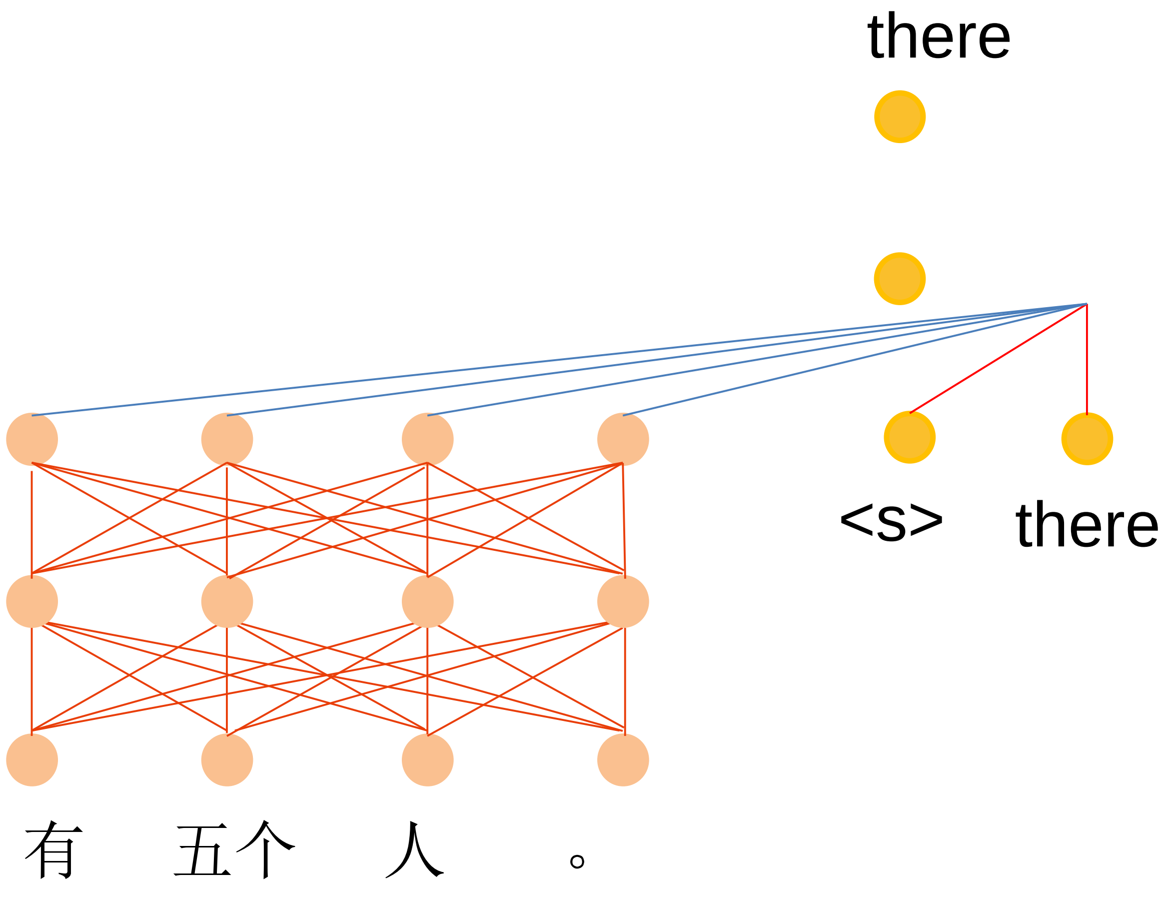
Transformer



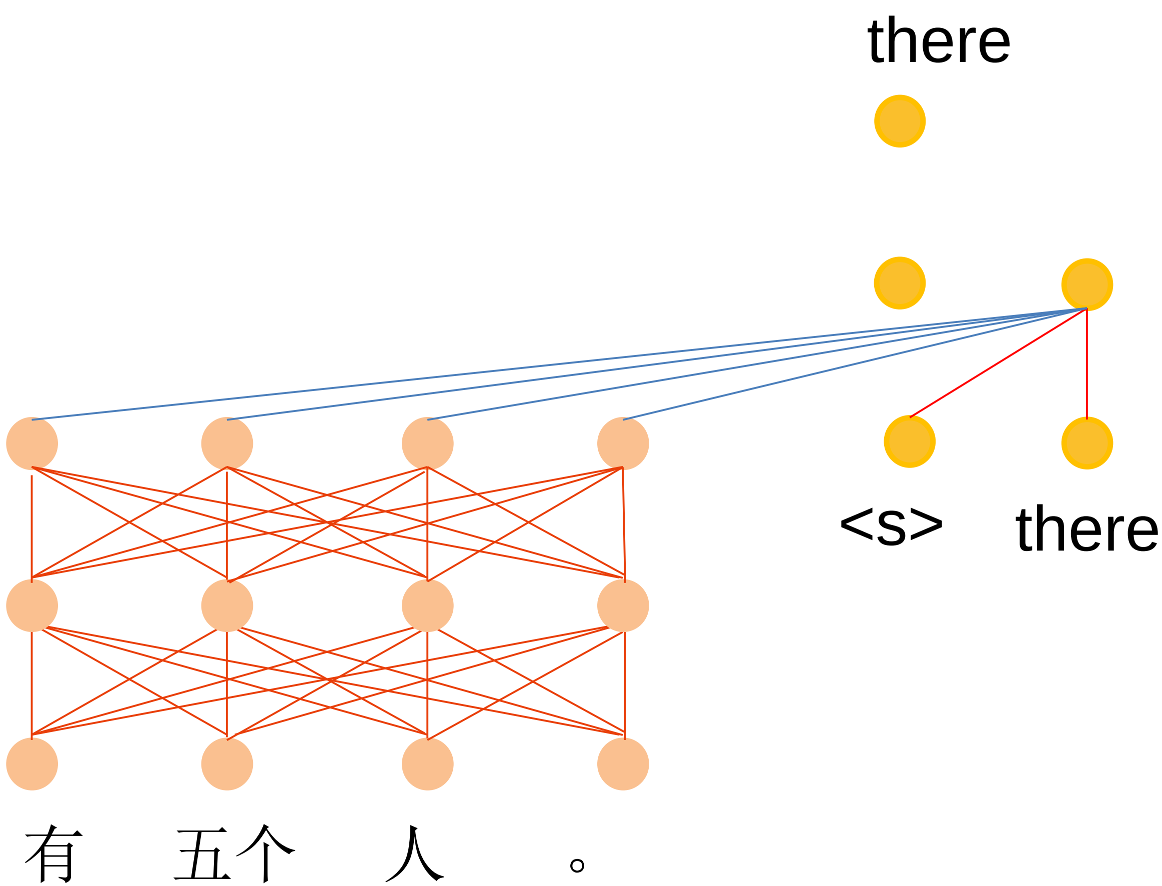
Transformer



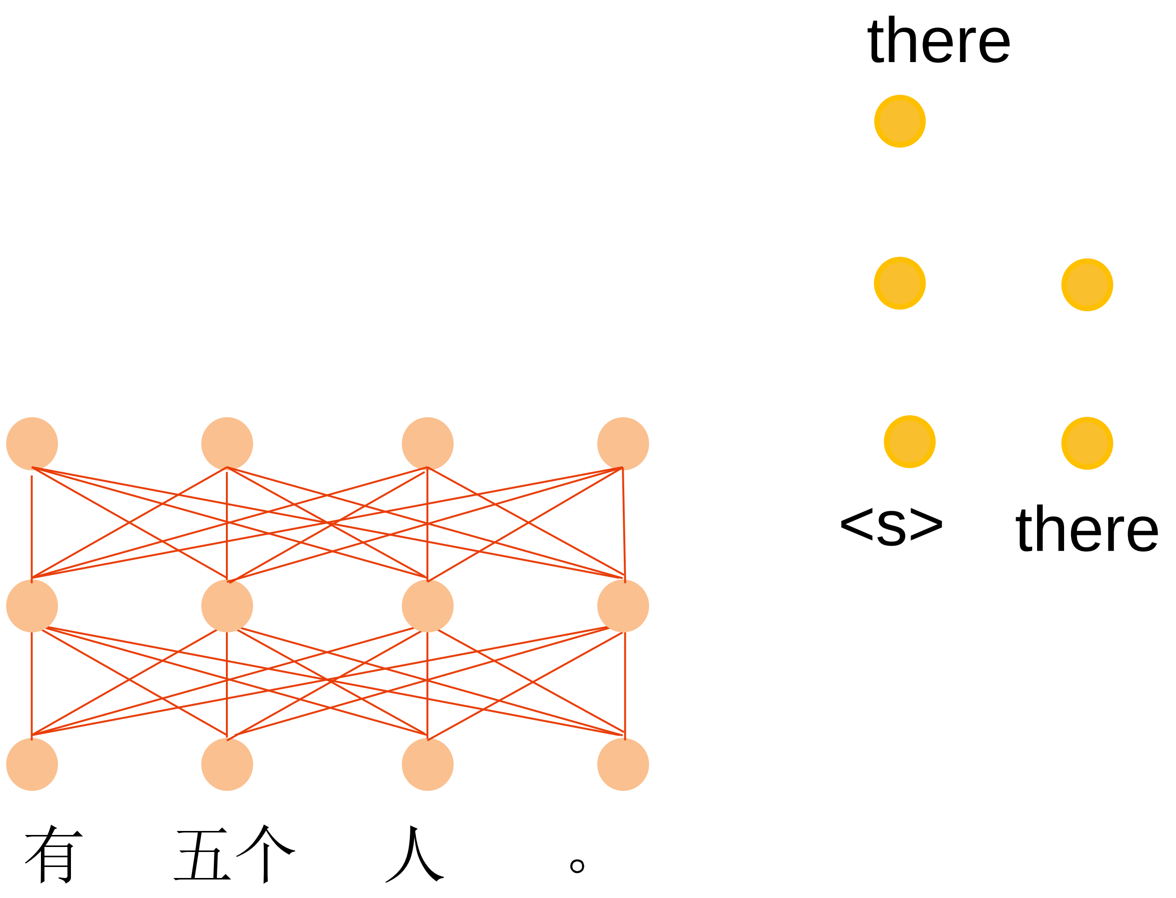
Transformer



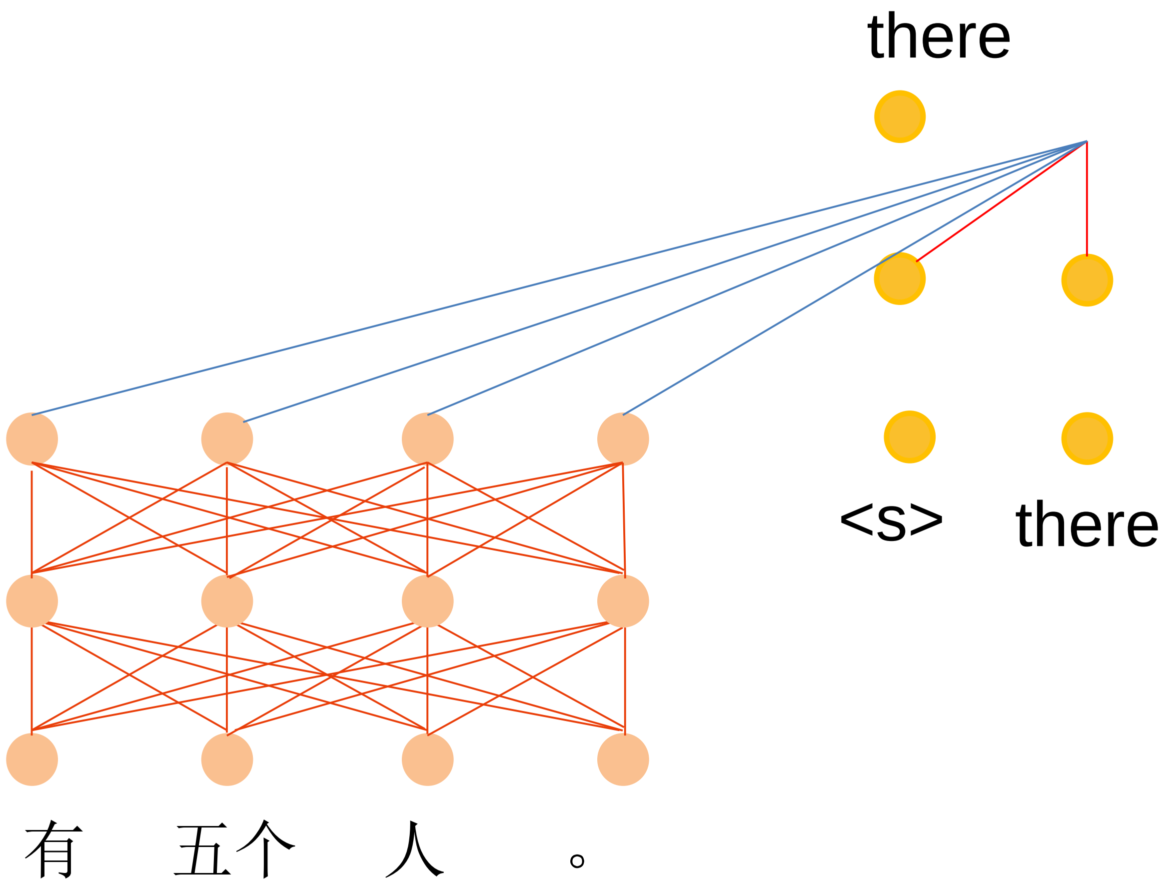
Transformer



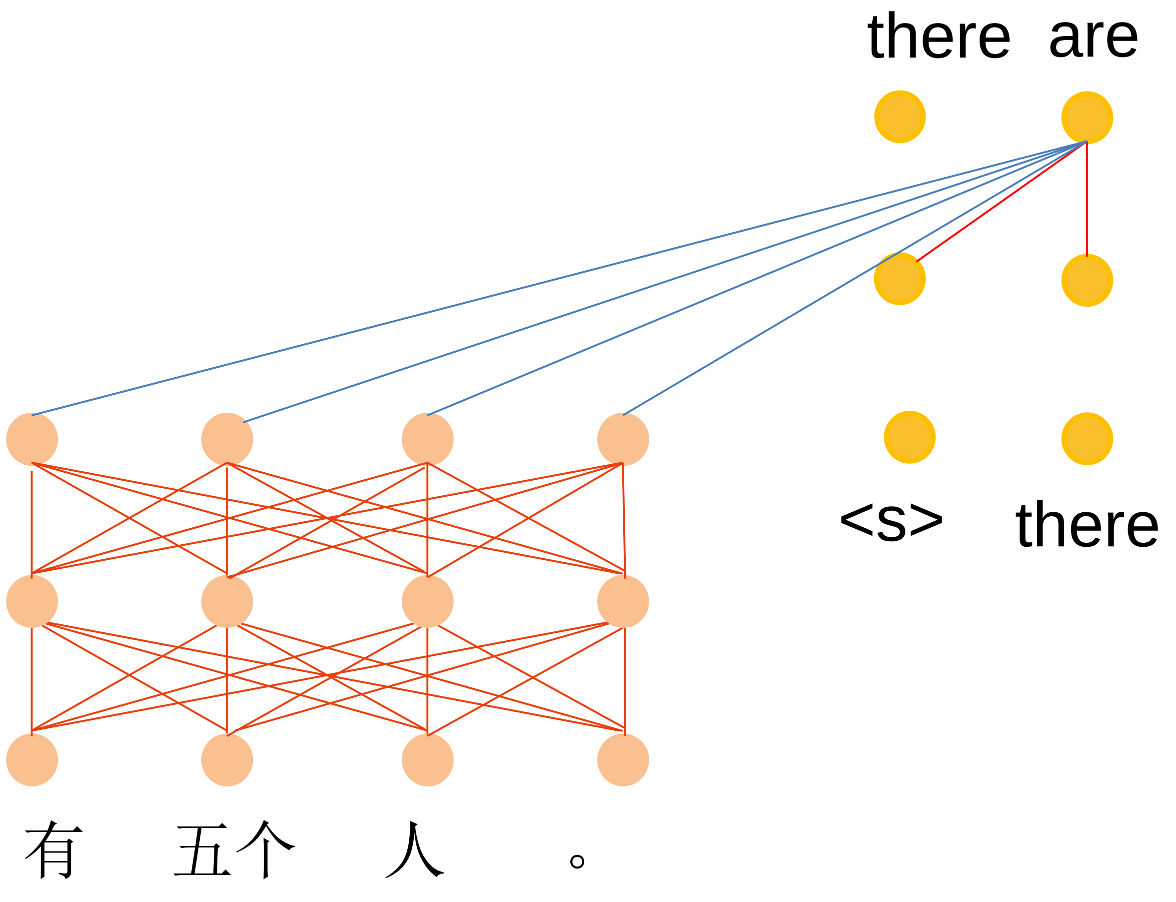
Transformer



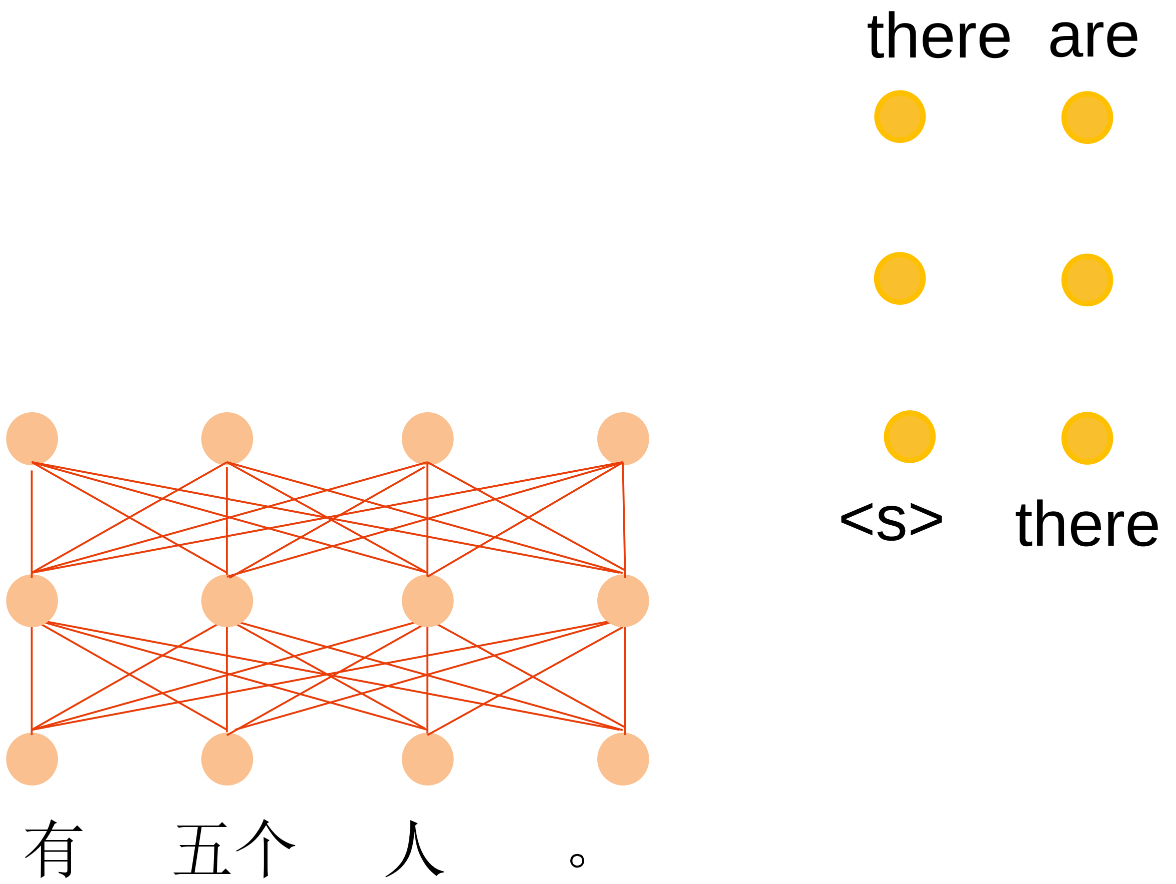
Transformer



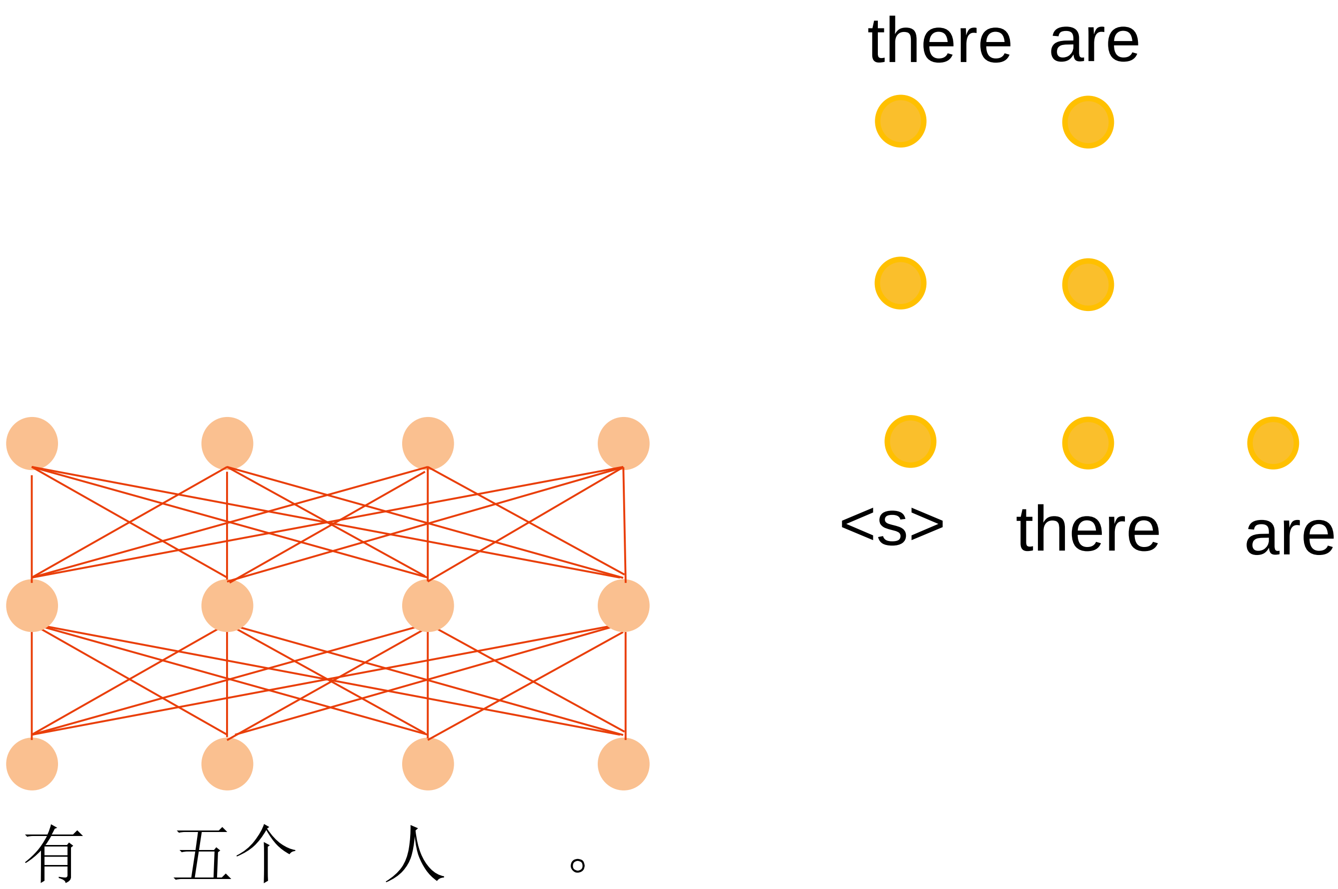
Transformer



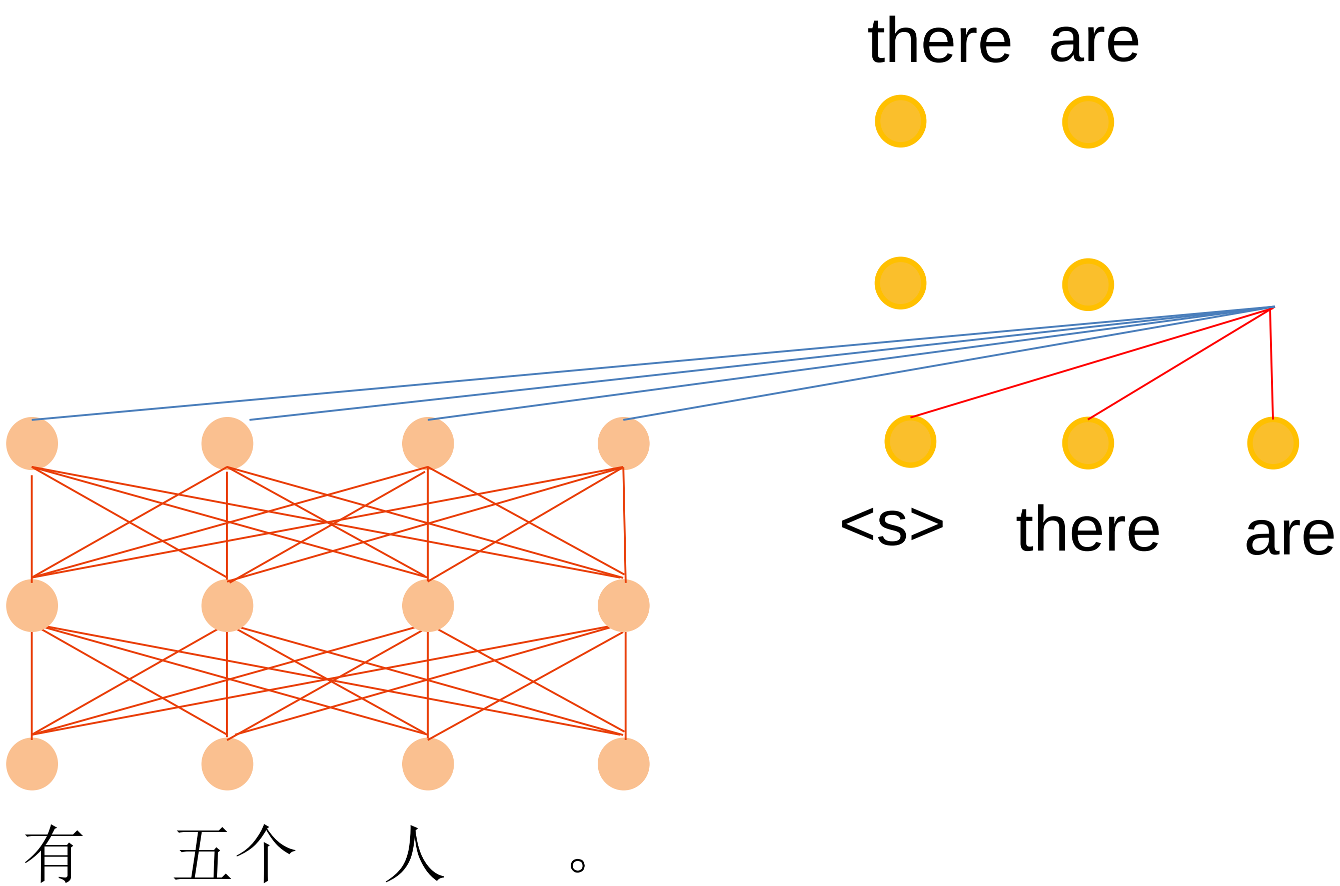
Transformer



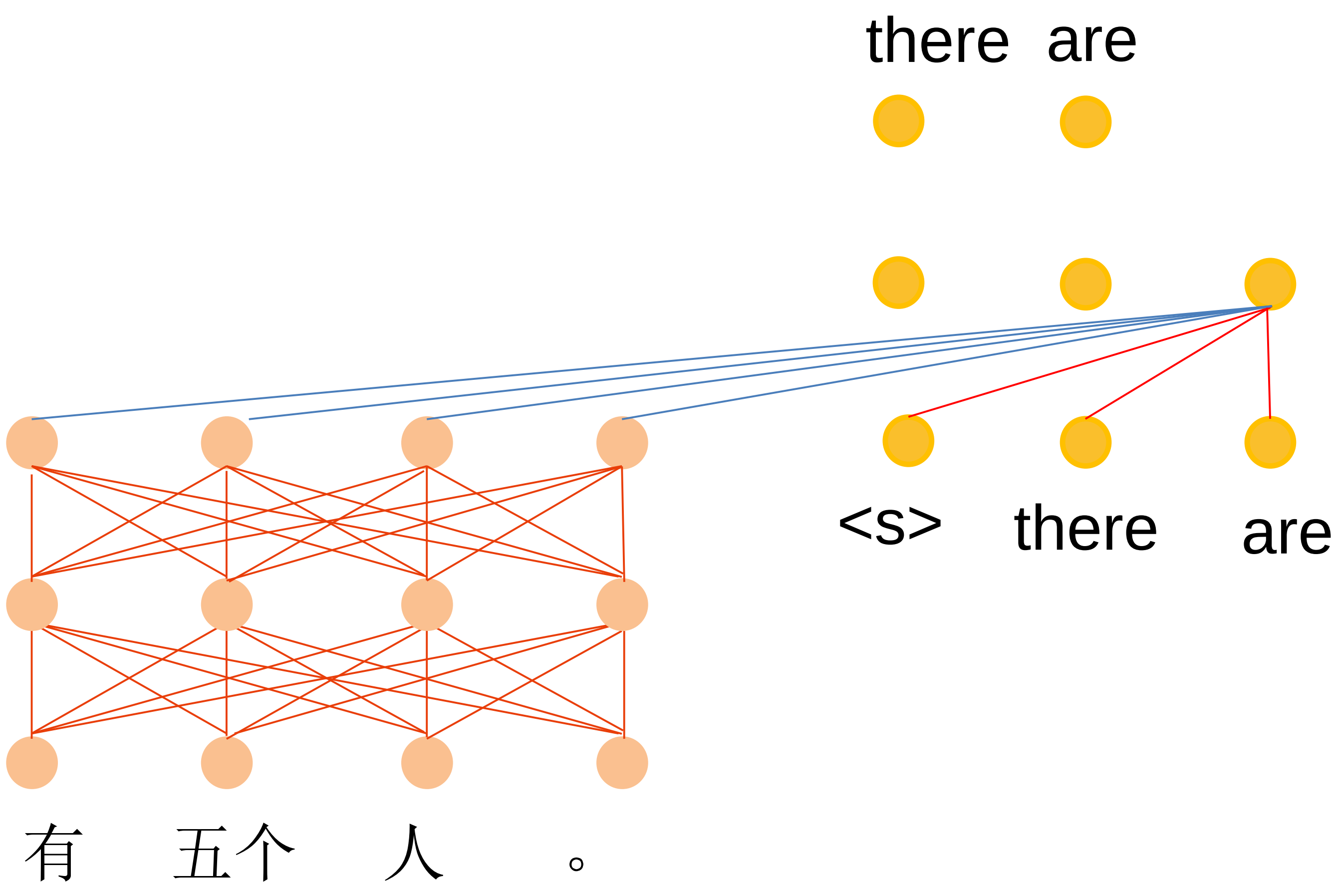
Transformer



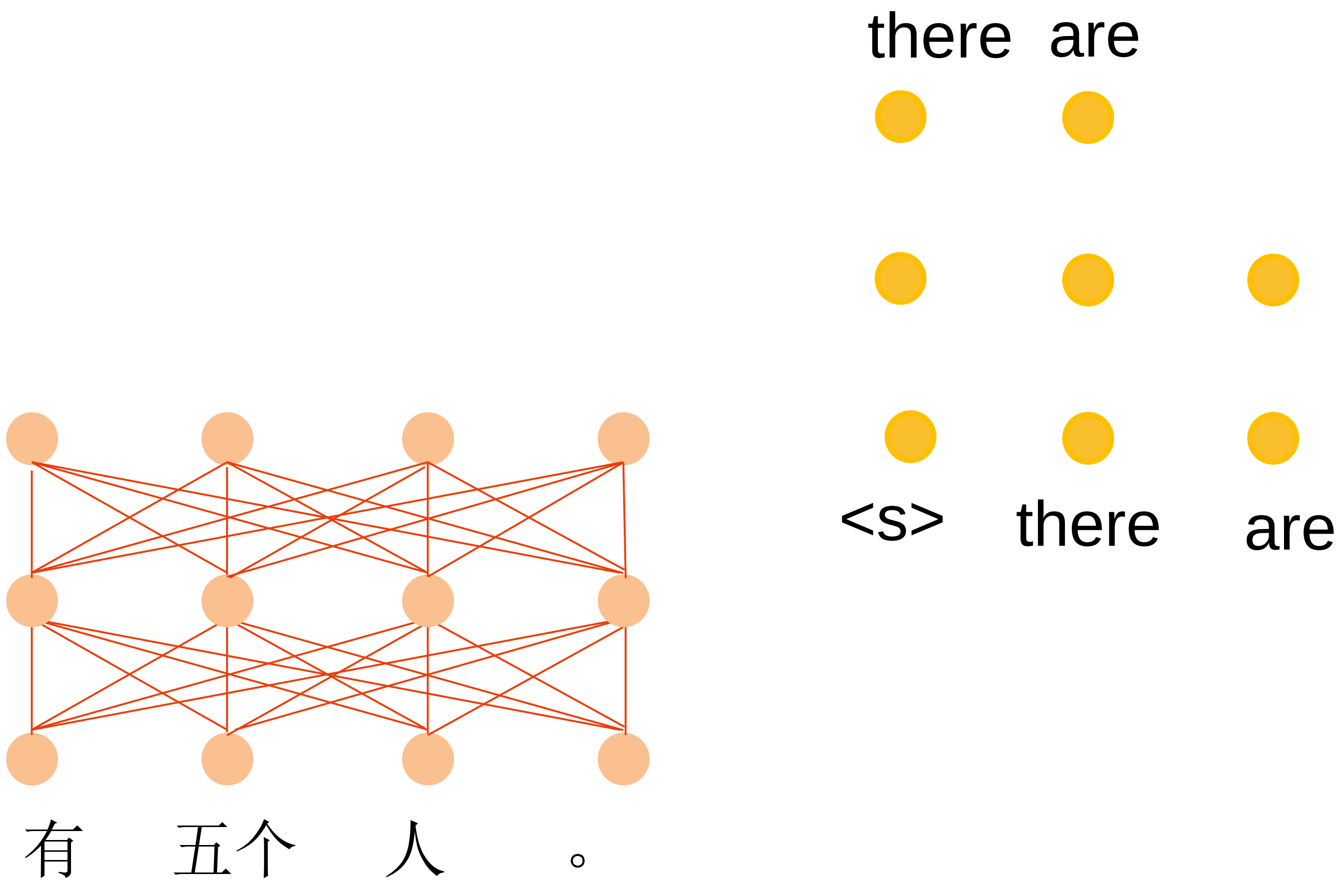
Transformer



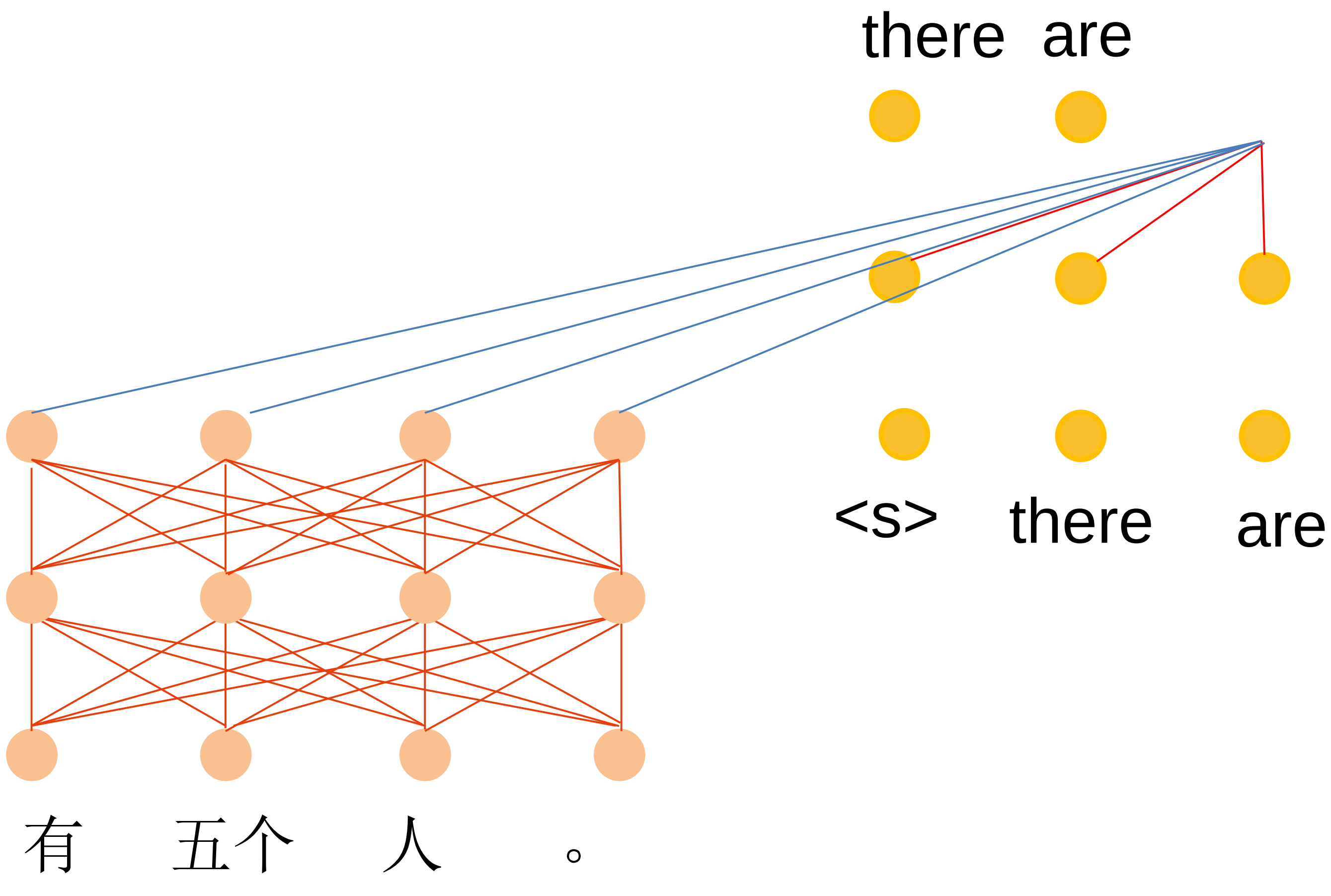
Transformer



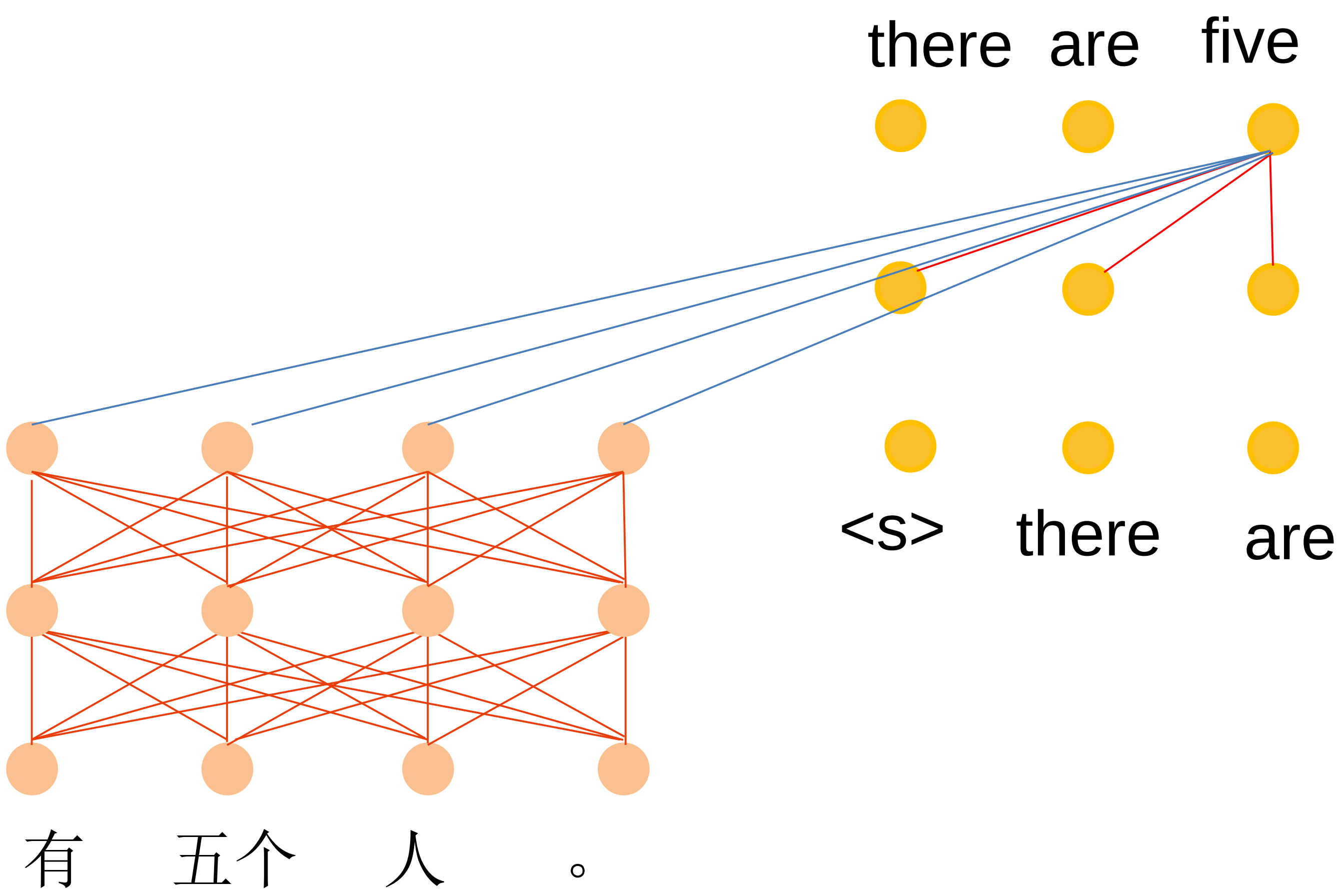
Transformer



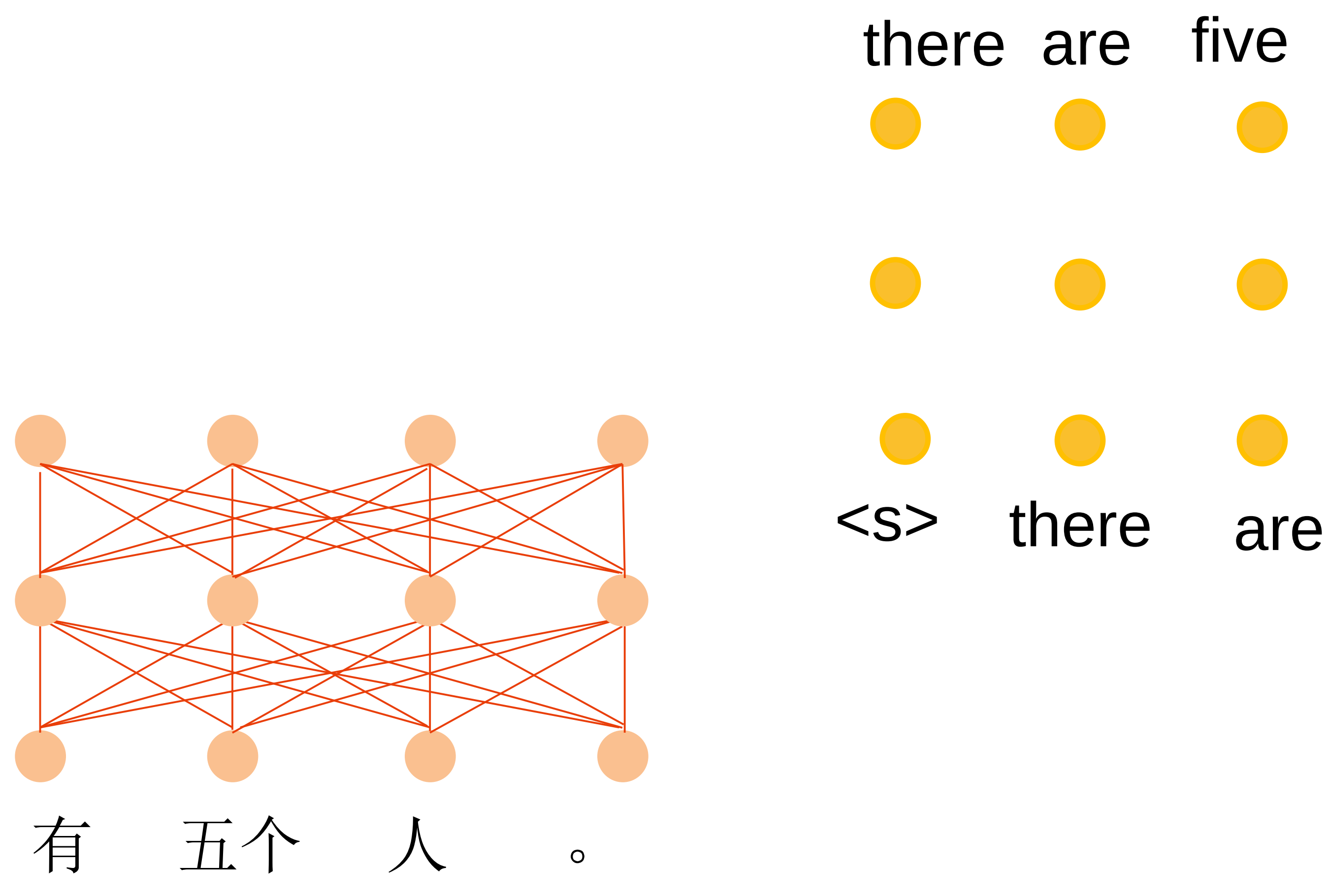
Transformer



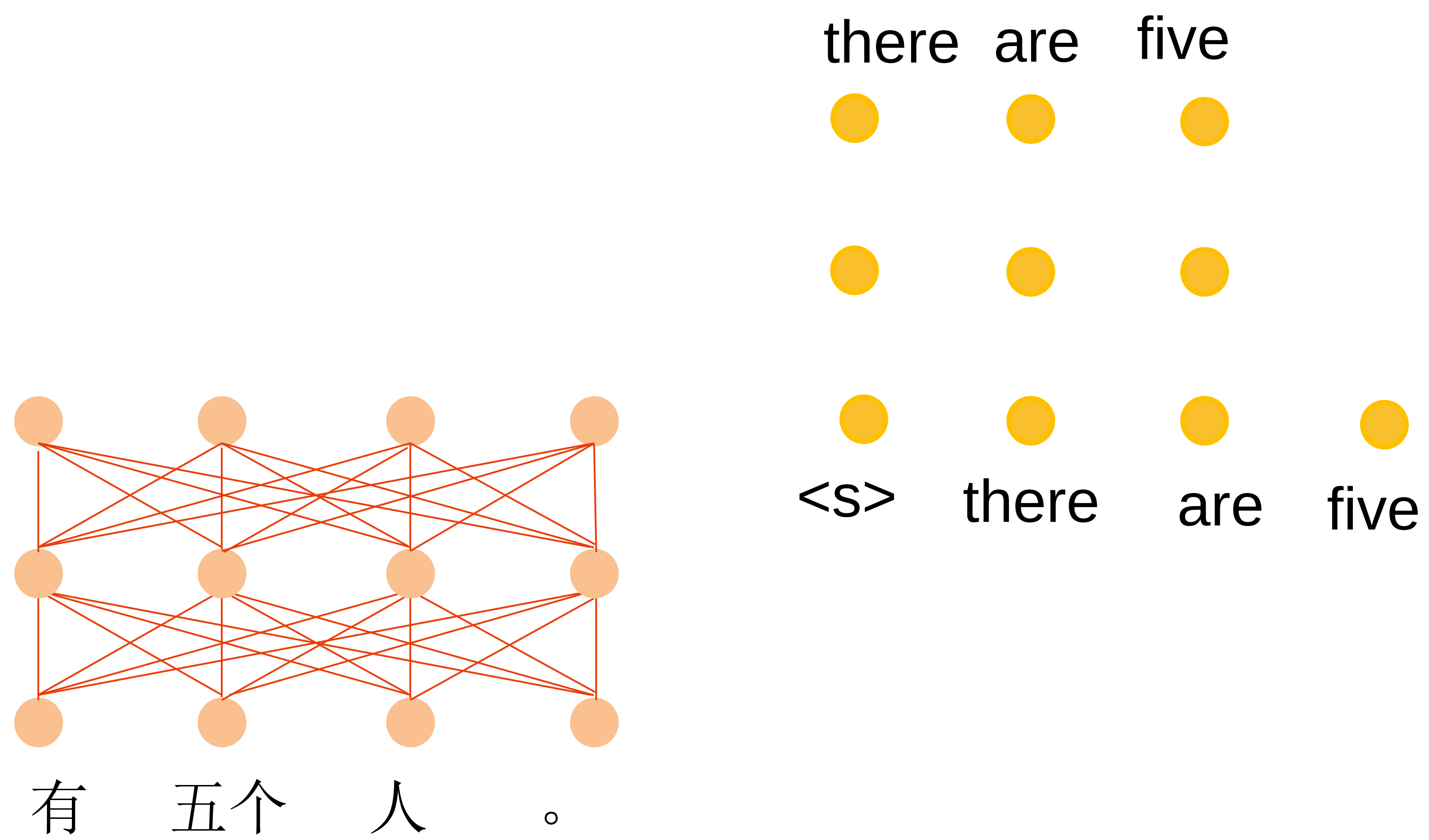
Transformer



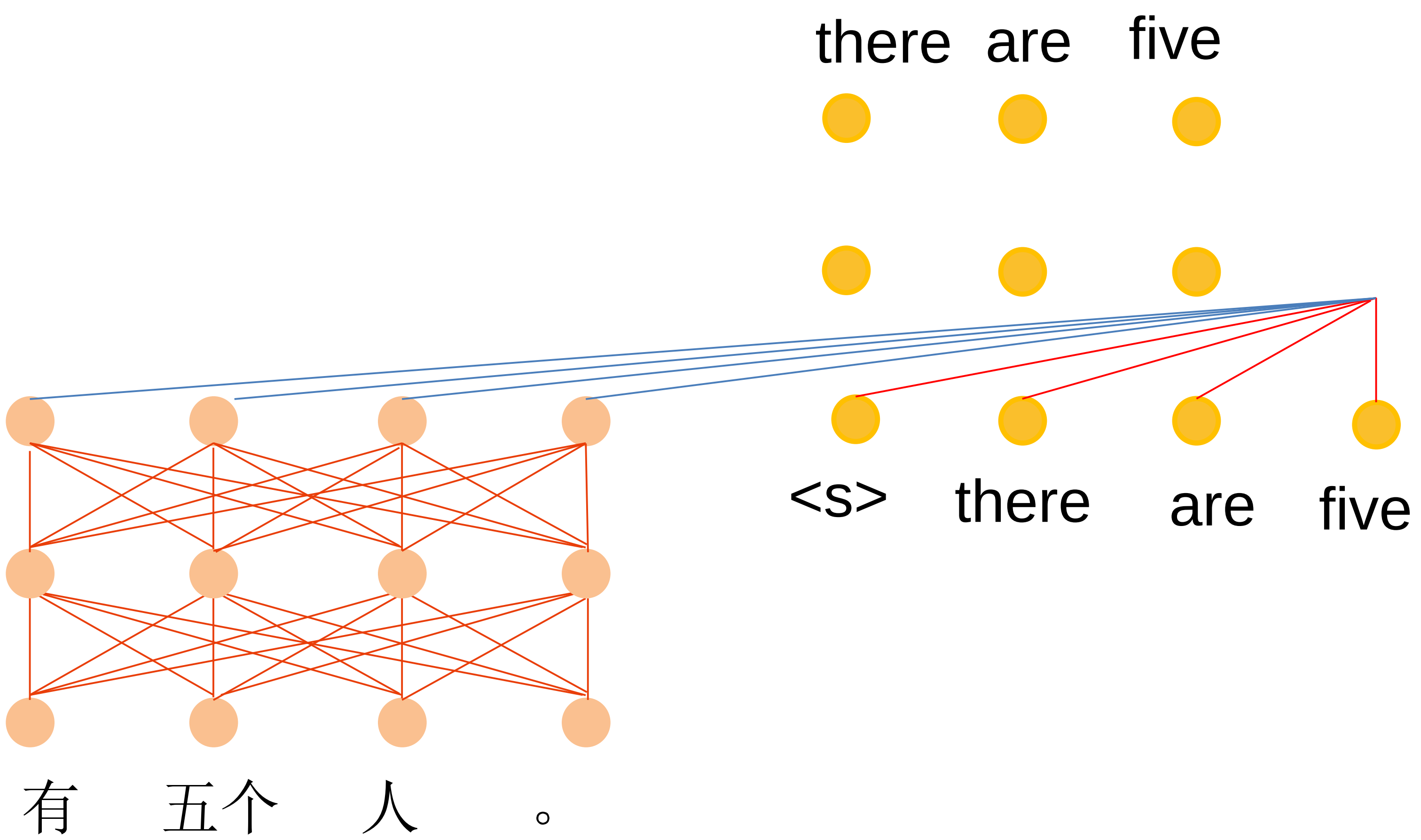
Transformer



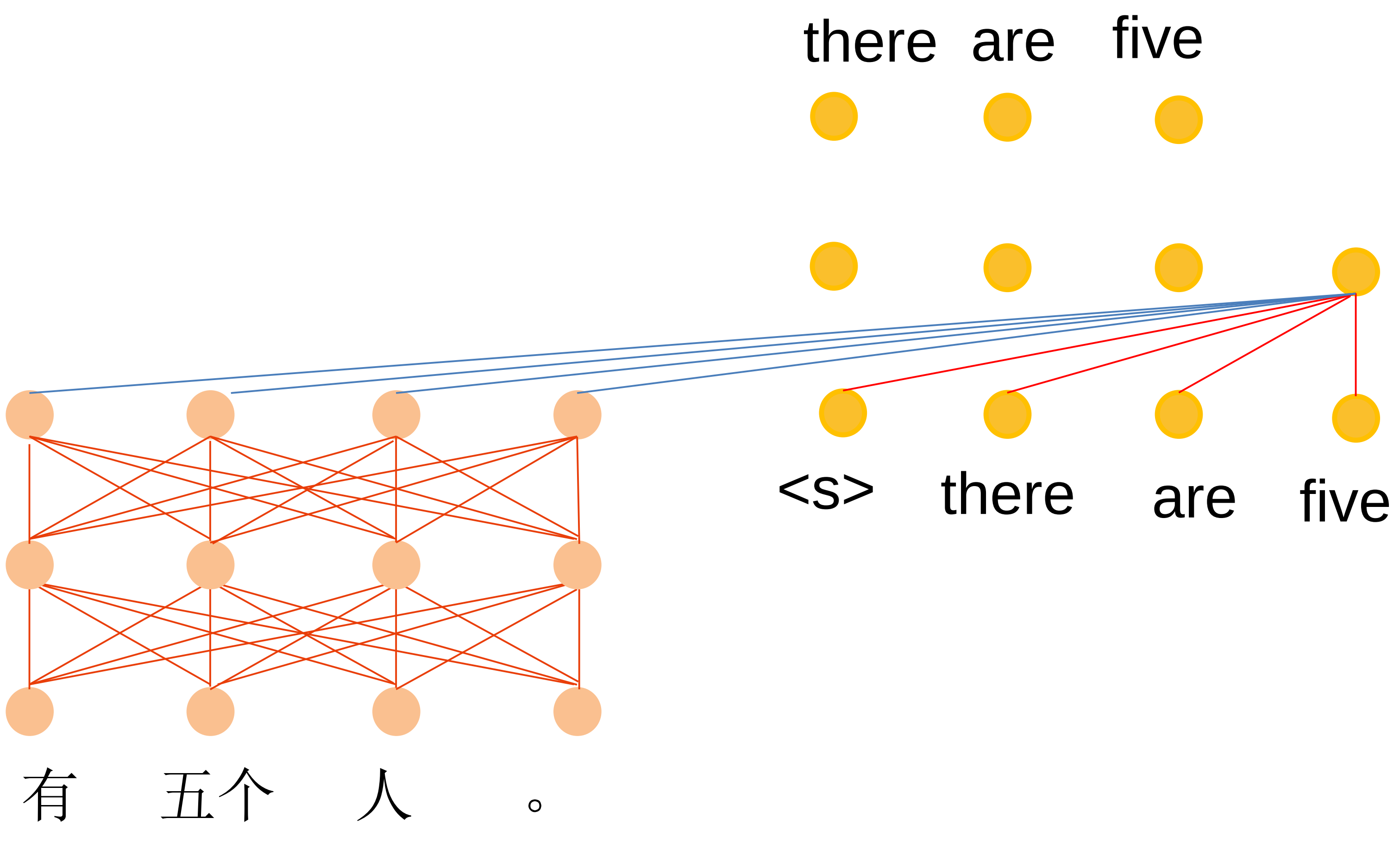
Transformer



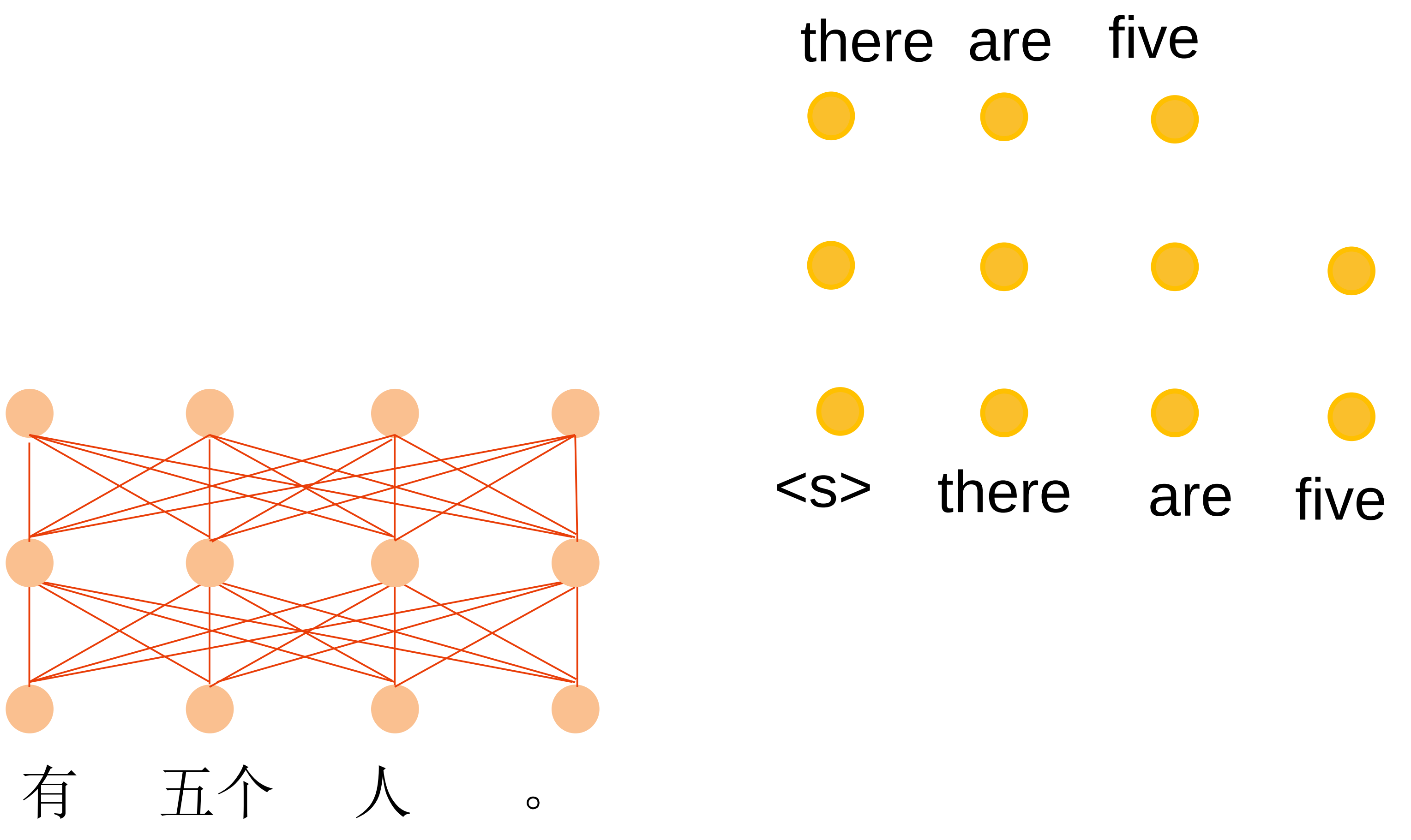
Transformer



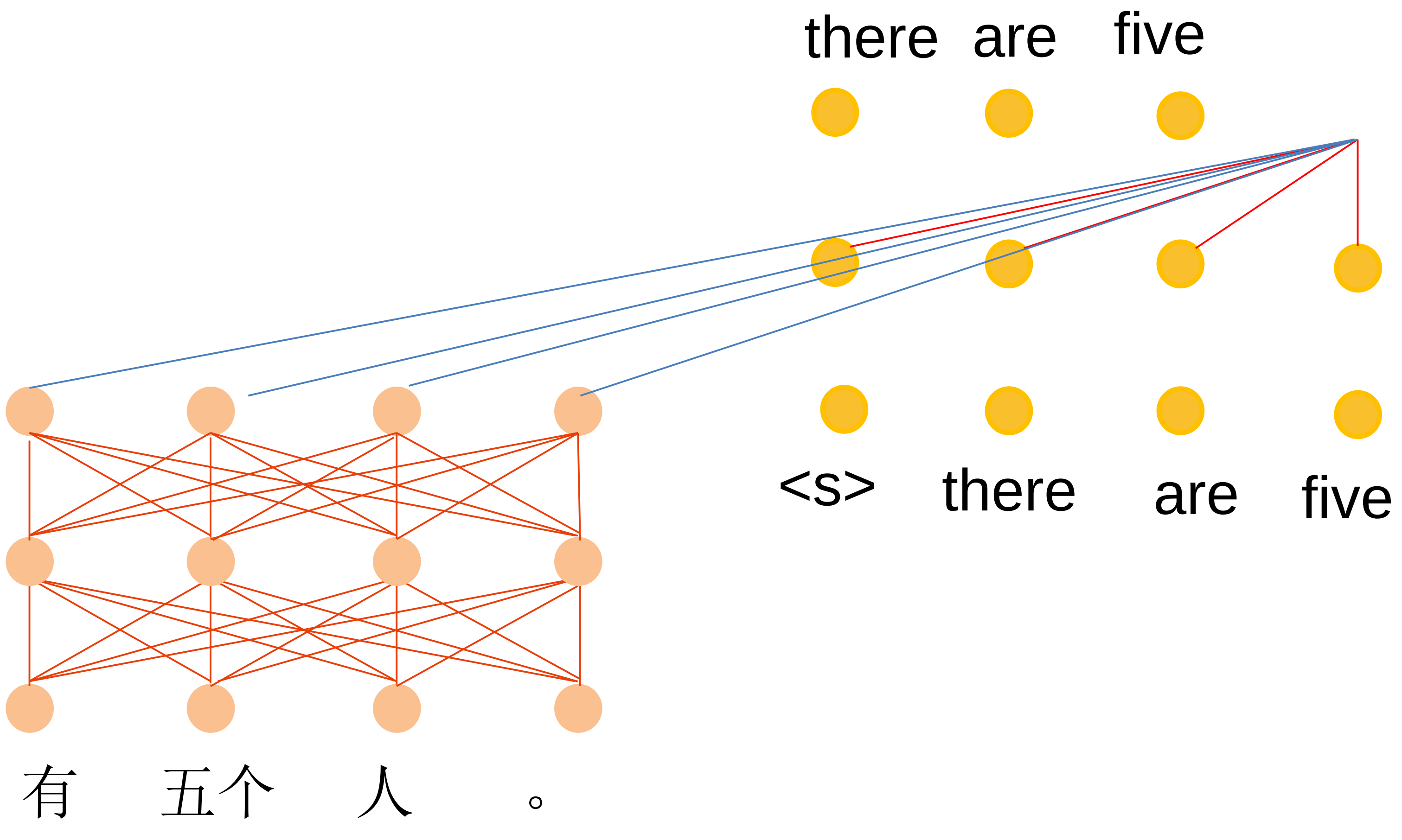
Transformer



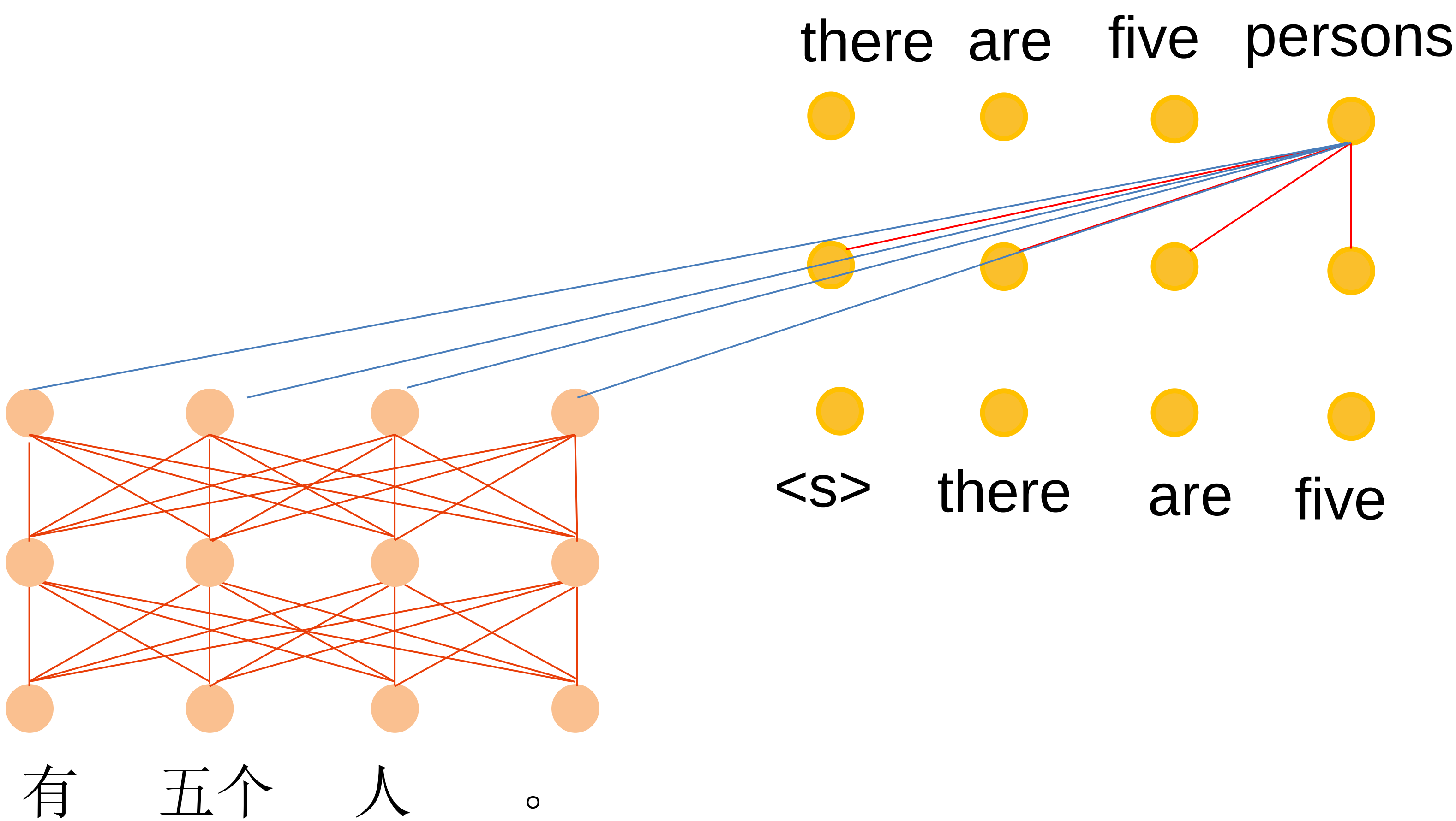
Transformer



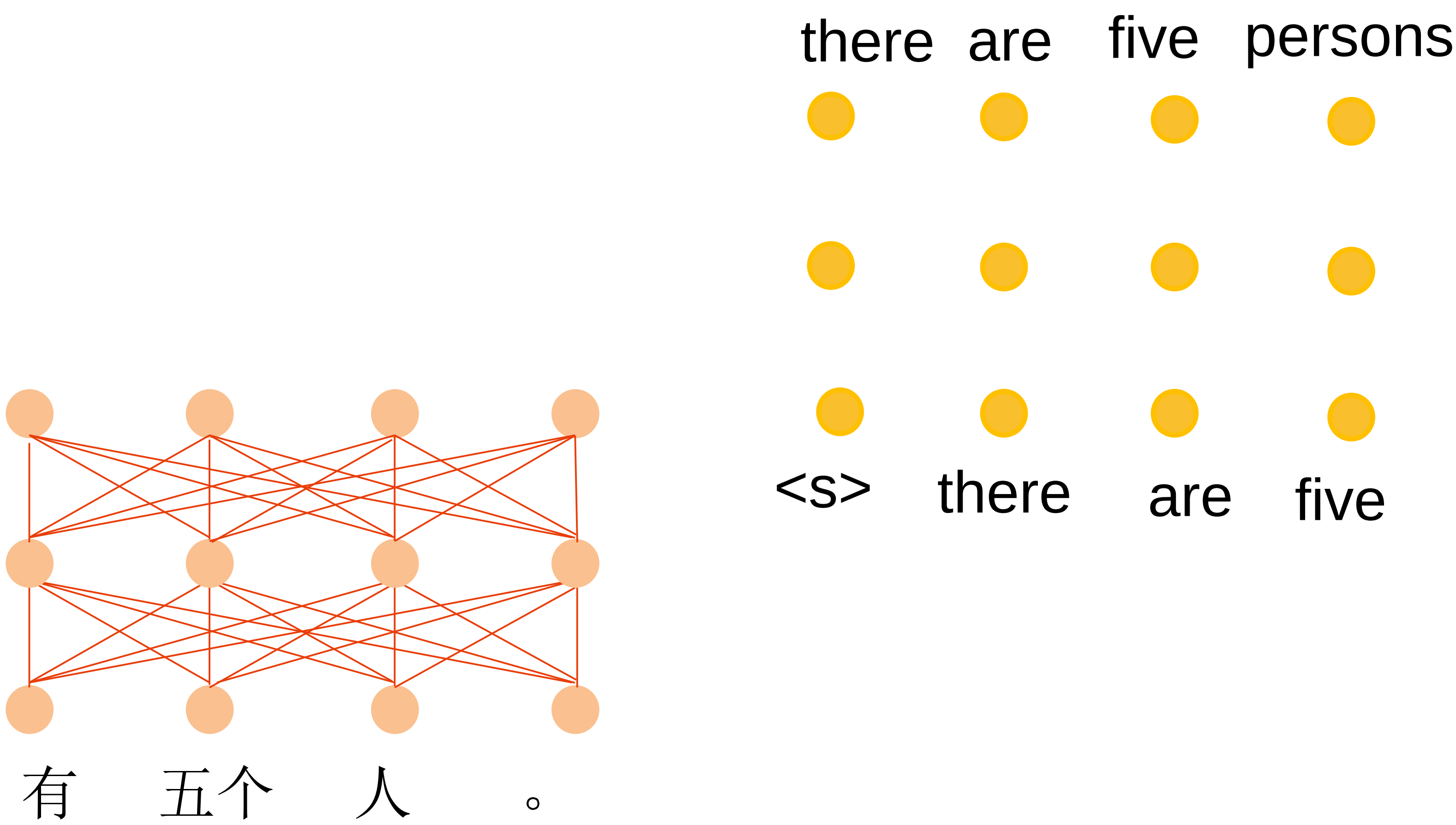
Transformer



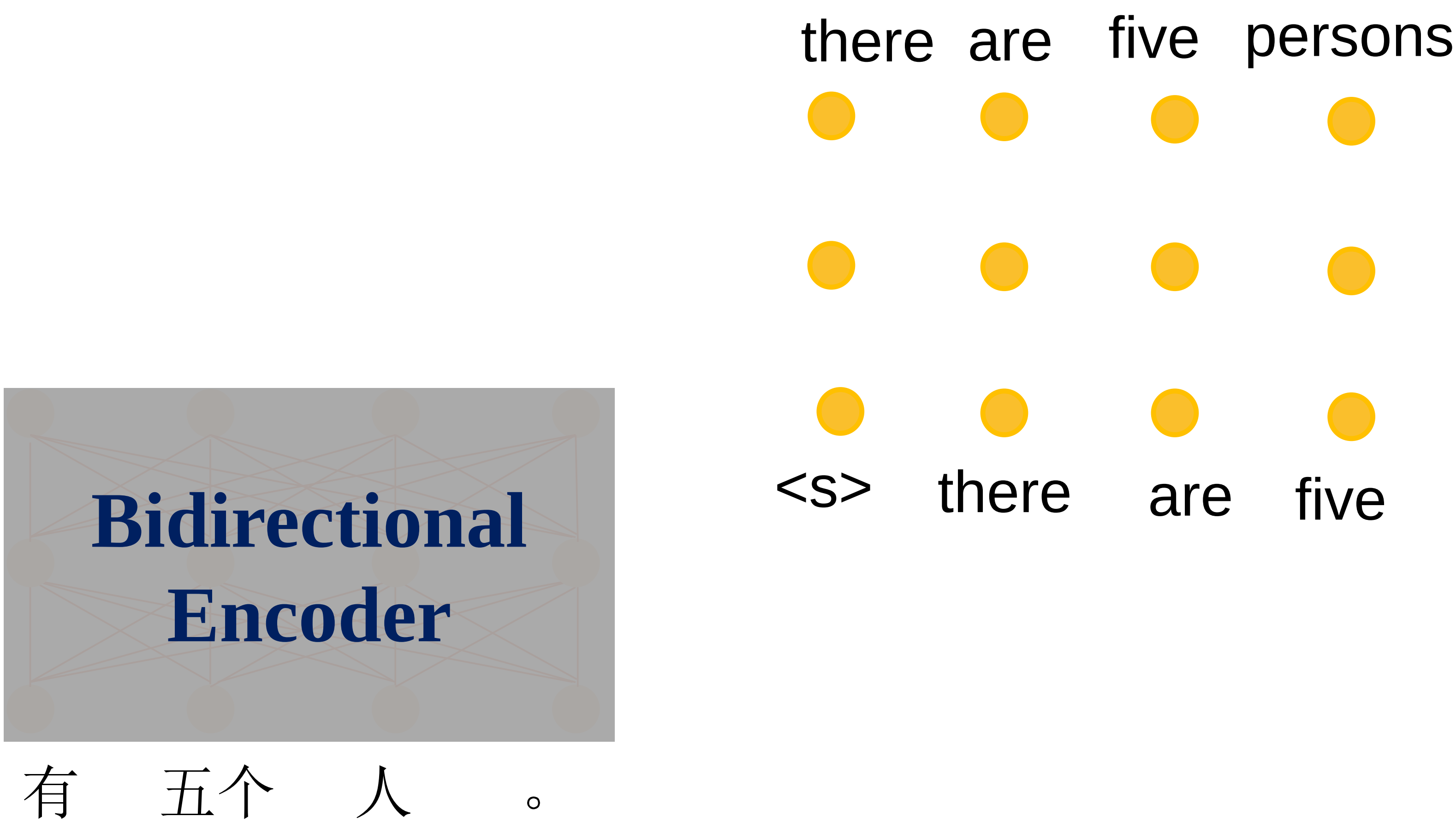
Transformer



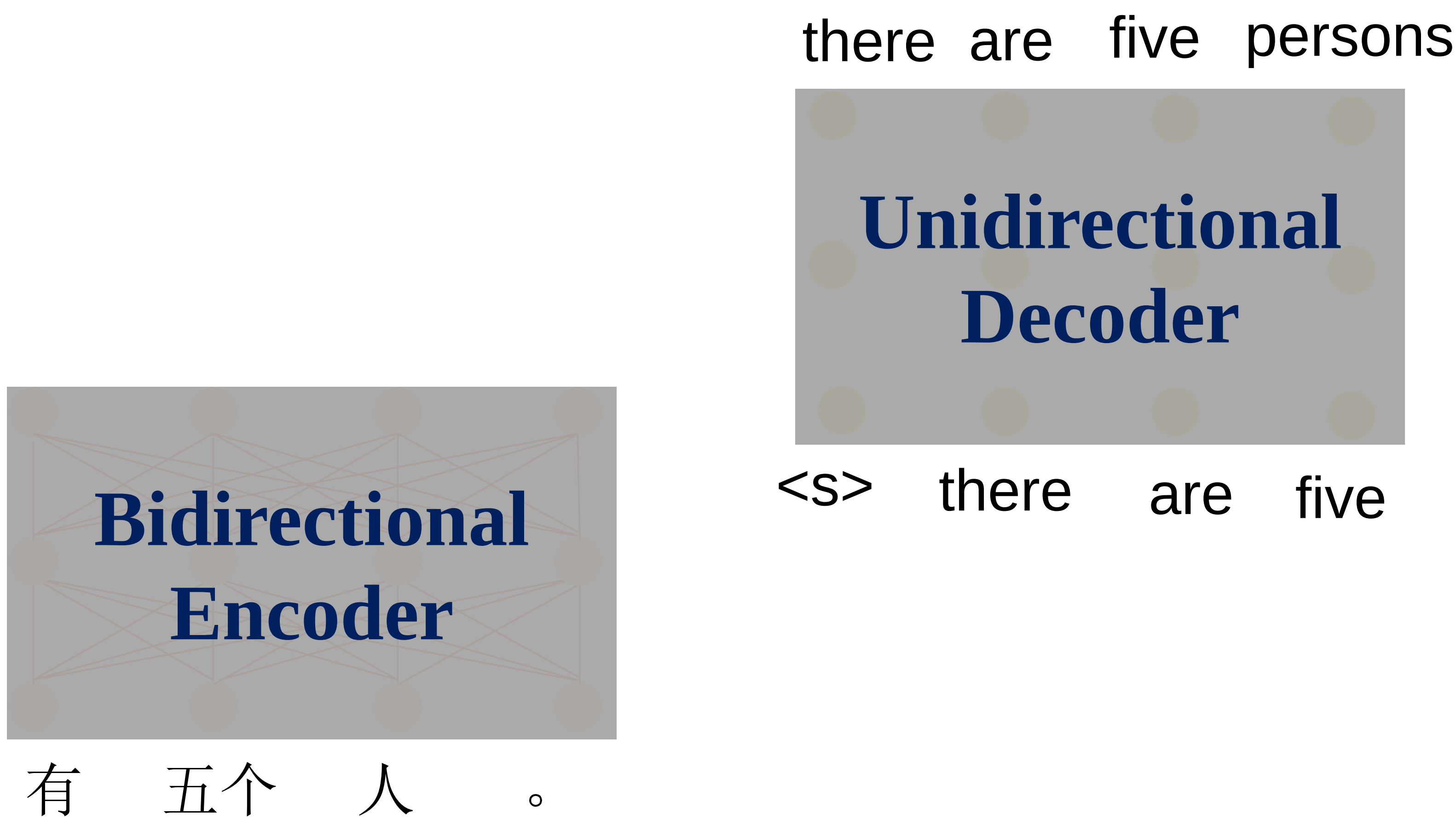
Transformer



Transformer



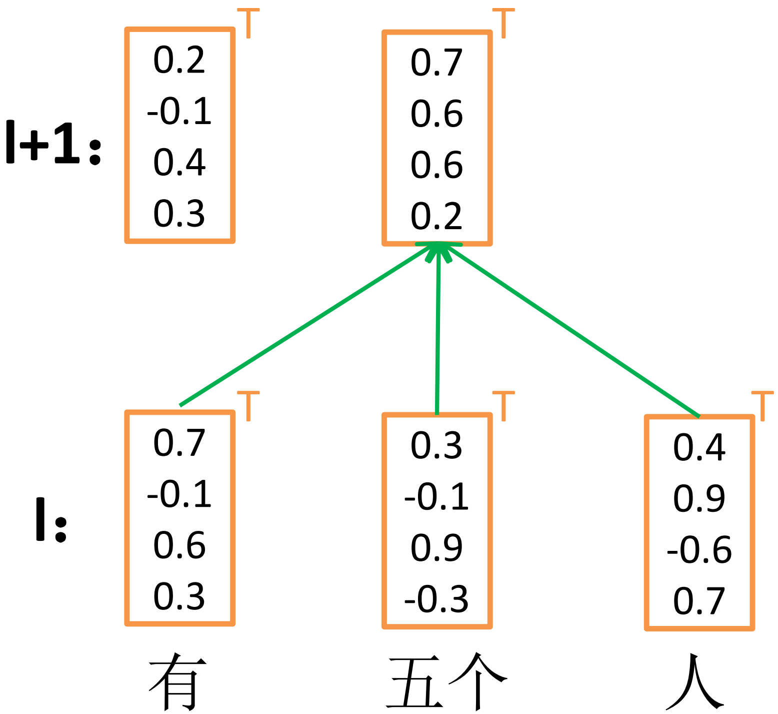
Transformer



Attention for Encoder

- Attention (Example)

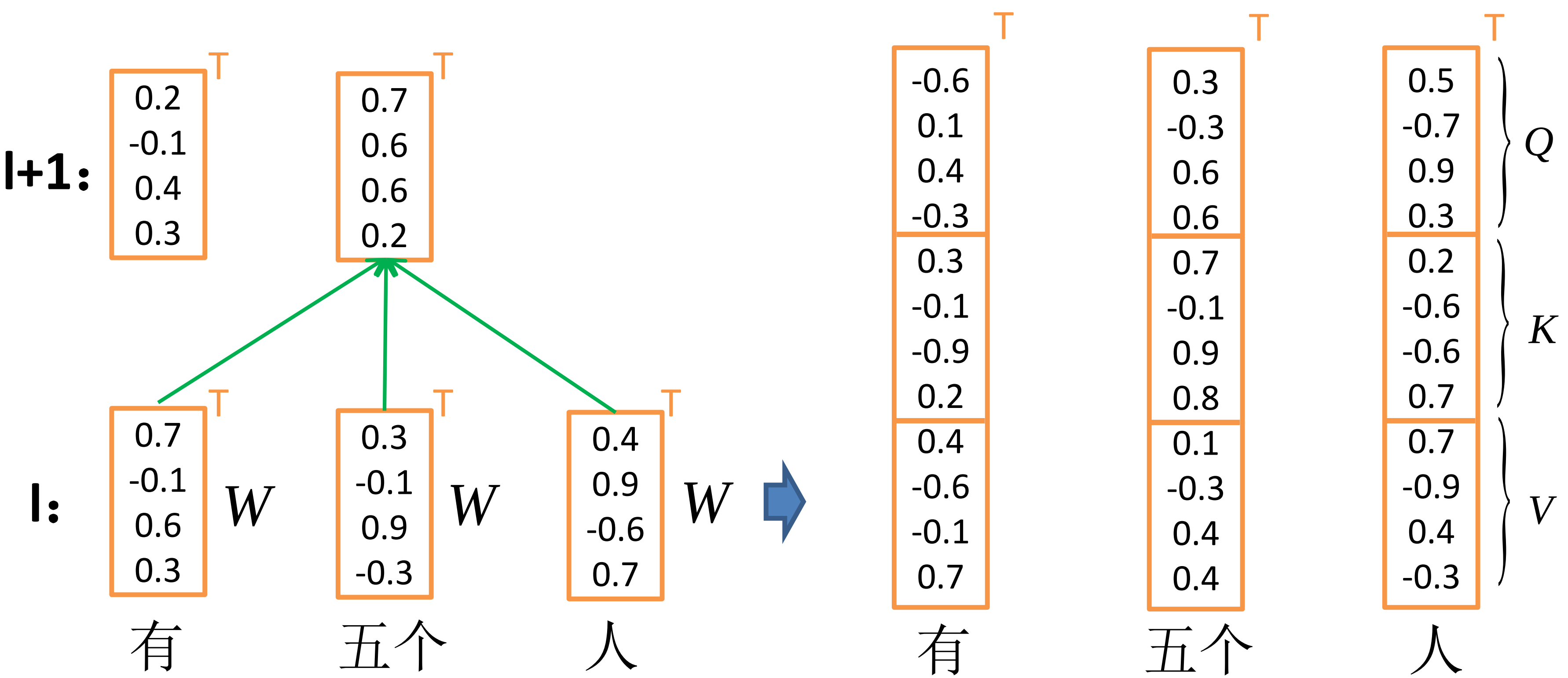
$$W \in R^{4\times 12}$$



Attention for Encoder

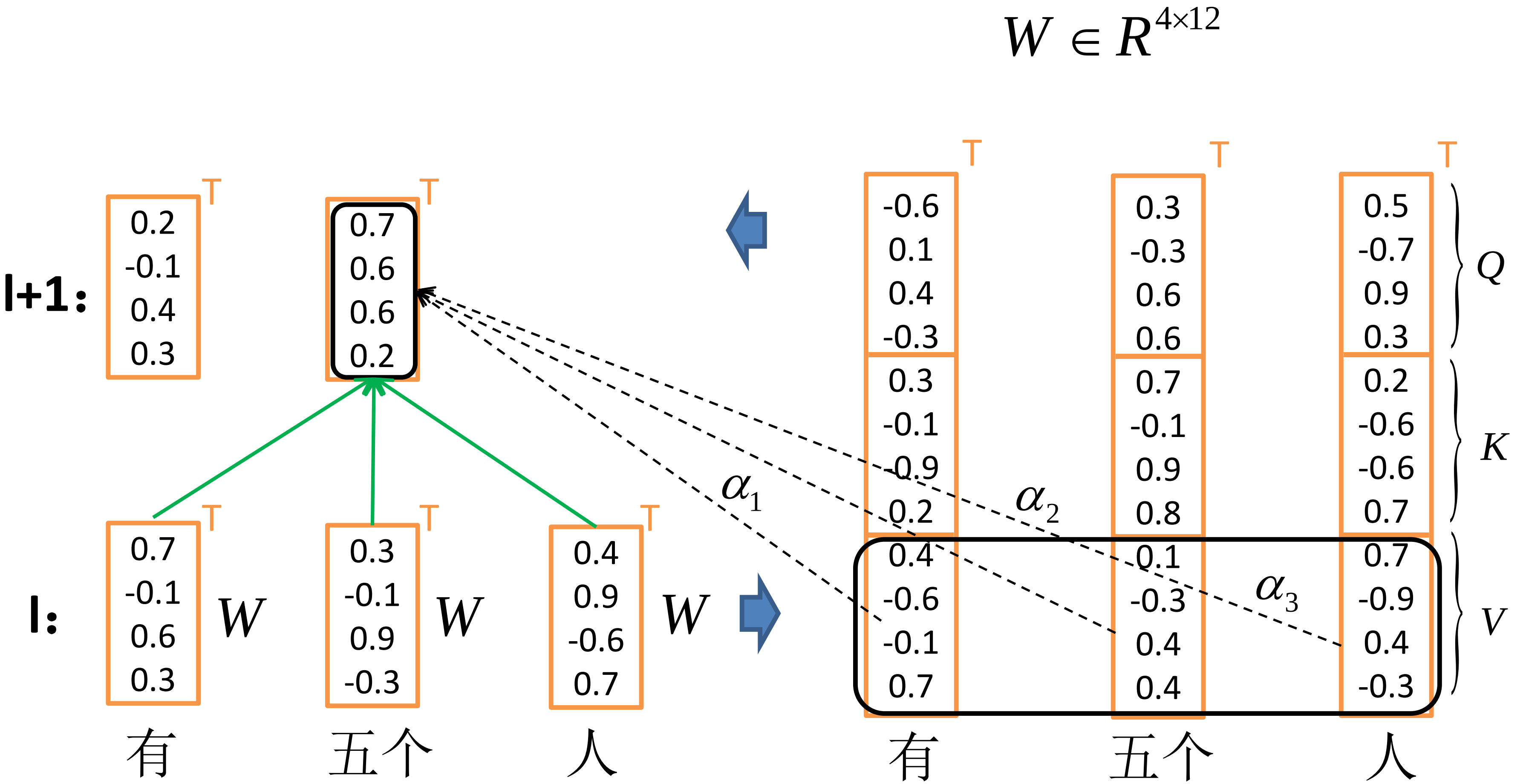
- Attention (Example)

$$W \in R^{4\times 12}$$



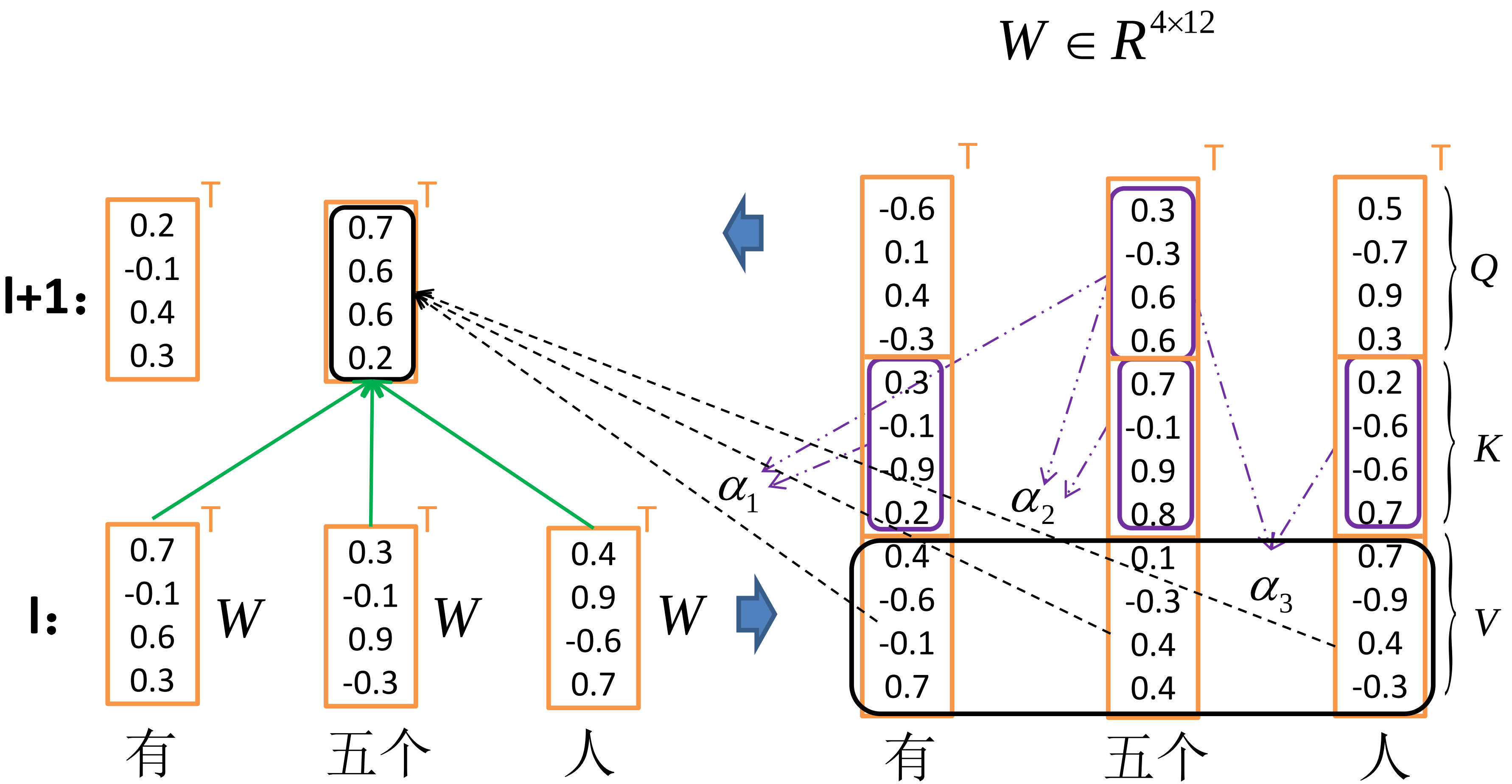
Attention for Encoder

- Attention (Example)



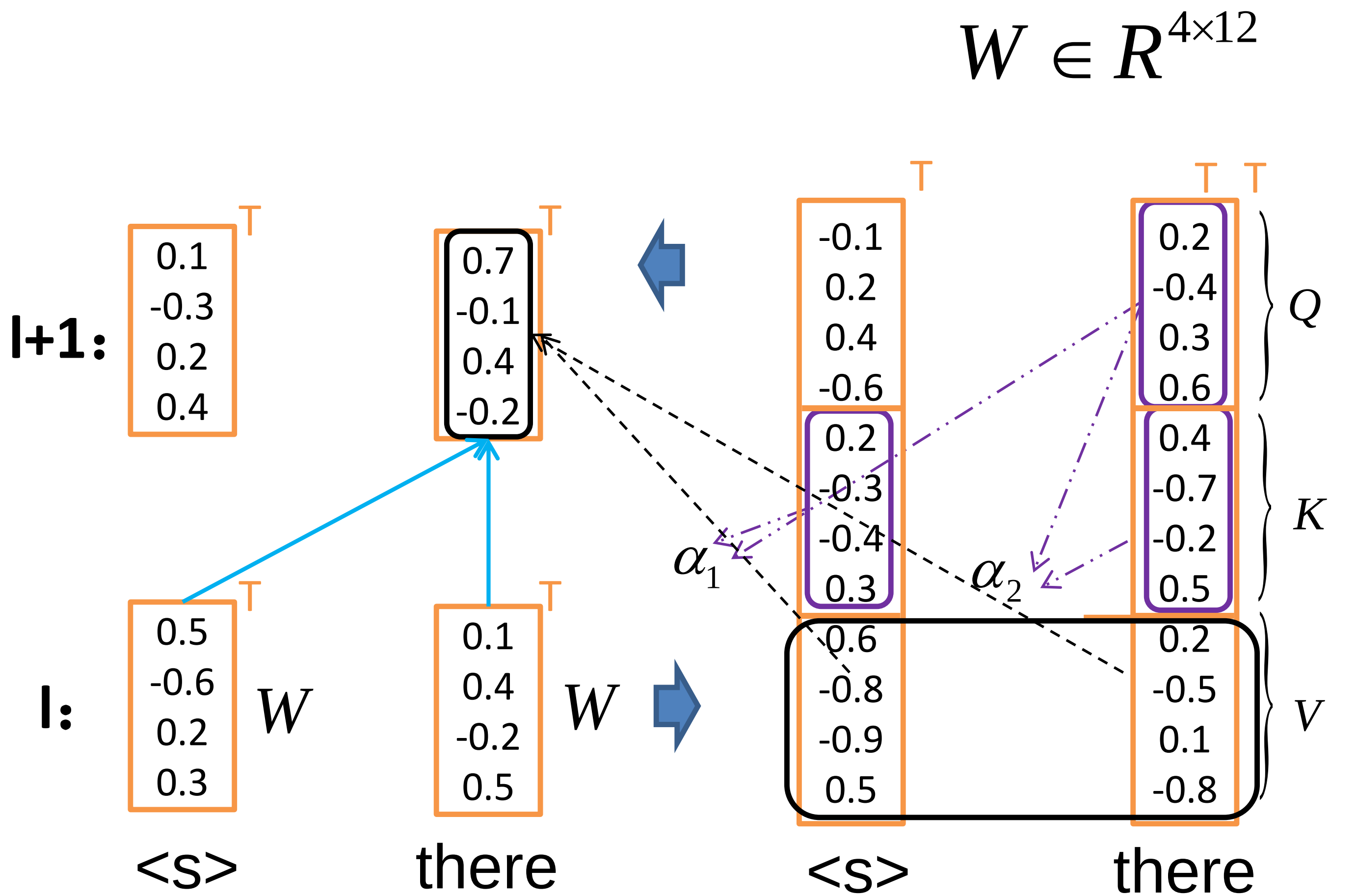
Attention for Encoder

- Attention (Example)



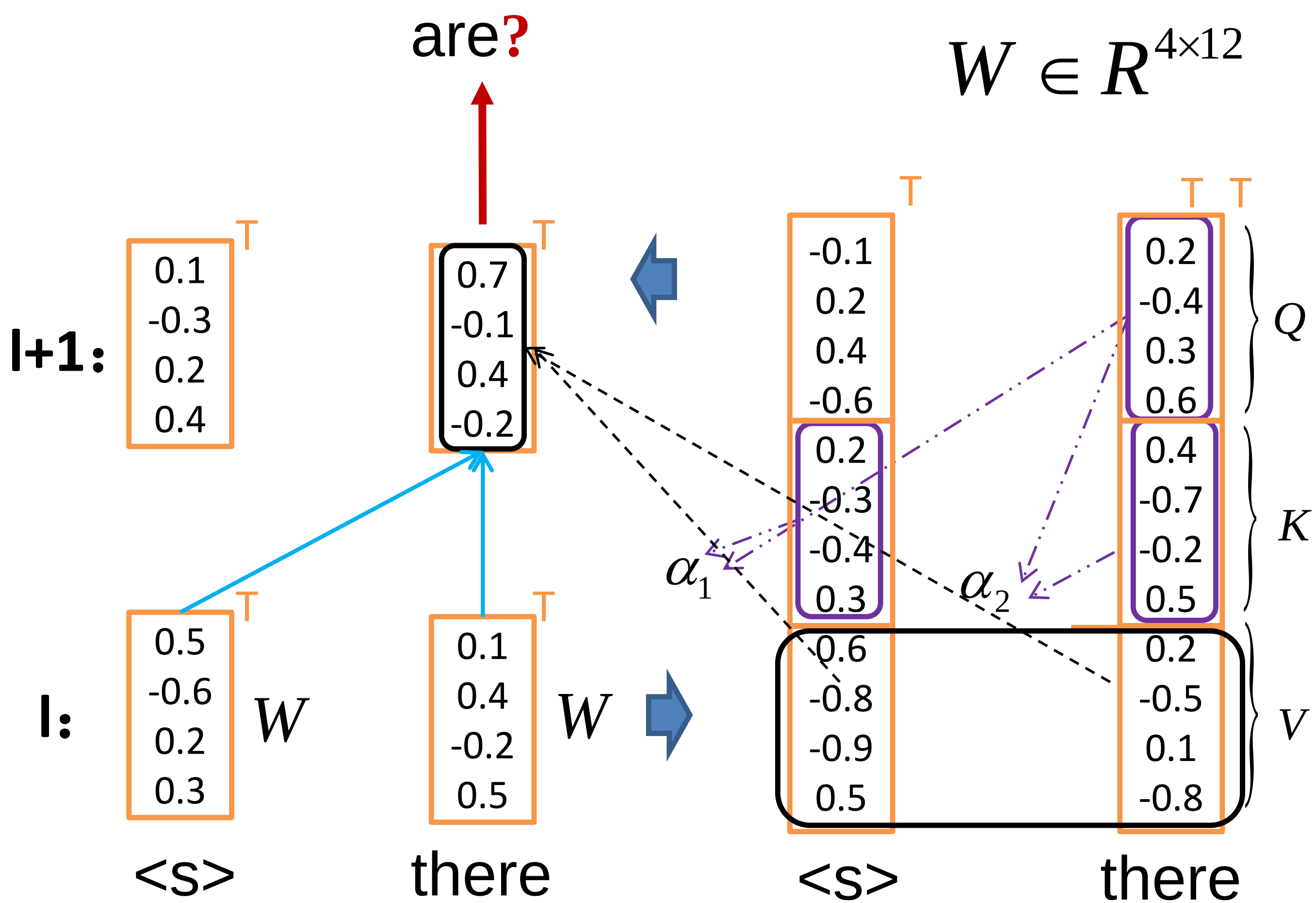
Attention for Decoder

- Attention (Example)



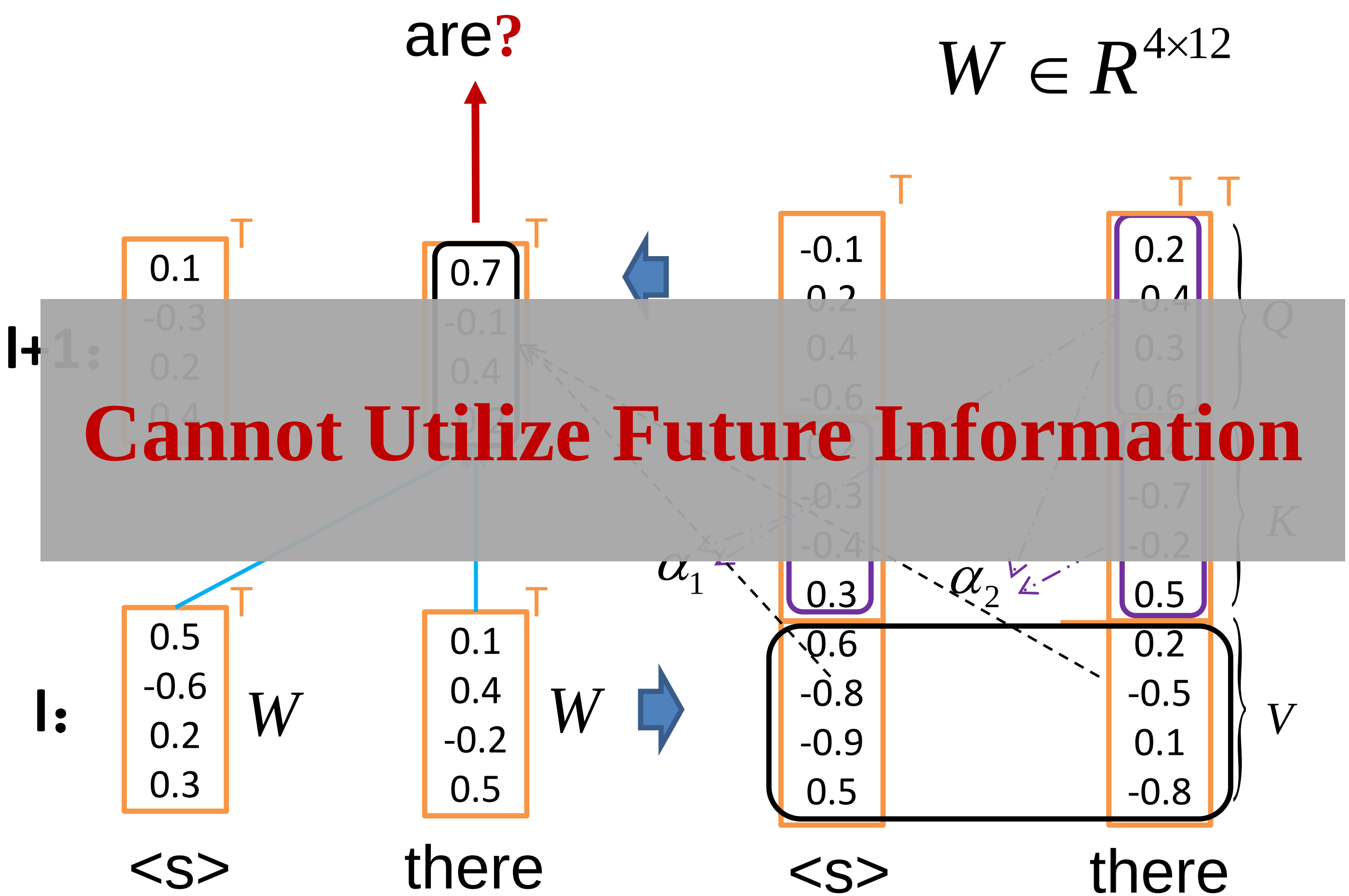
Attention for Decoder

- Attention (Example)



Attention for Decoder

- Attention (Example)



Problems for Unidirectional Inference

$x_0 \ x_1 \ \cdots \ \ x_j \ \ \cdots \ x_m$

Problems for Unidirectional Inference

Left-to-Right Decoding

$$y_0^{l2r} \longrightarrow y_1^{l2r} \longrightarrow \cdots \quad y_i^{l2r} \quad \cdots \quad \cdots$$

$$x_0 \;\; x_1 \;\; \cdots \;\; x_j \;\; \cdots \;\; x_m$$

Problems for Unidirectional Inference

Left-to-Right Decoding

$$y_0^{l2r} \longrightarrow y_1^{l2r} \longrightarrow \cdots \quad y_i^{l2r} \qquad \cdots \quad \cdots$$

$$x_0 \;\; x_1 \;\; \cdots \;\; x_j \;\; \cdots \;\; x_m$$

Problems for Unidirectional Inference

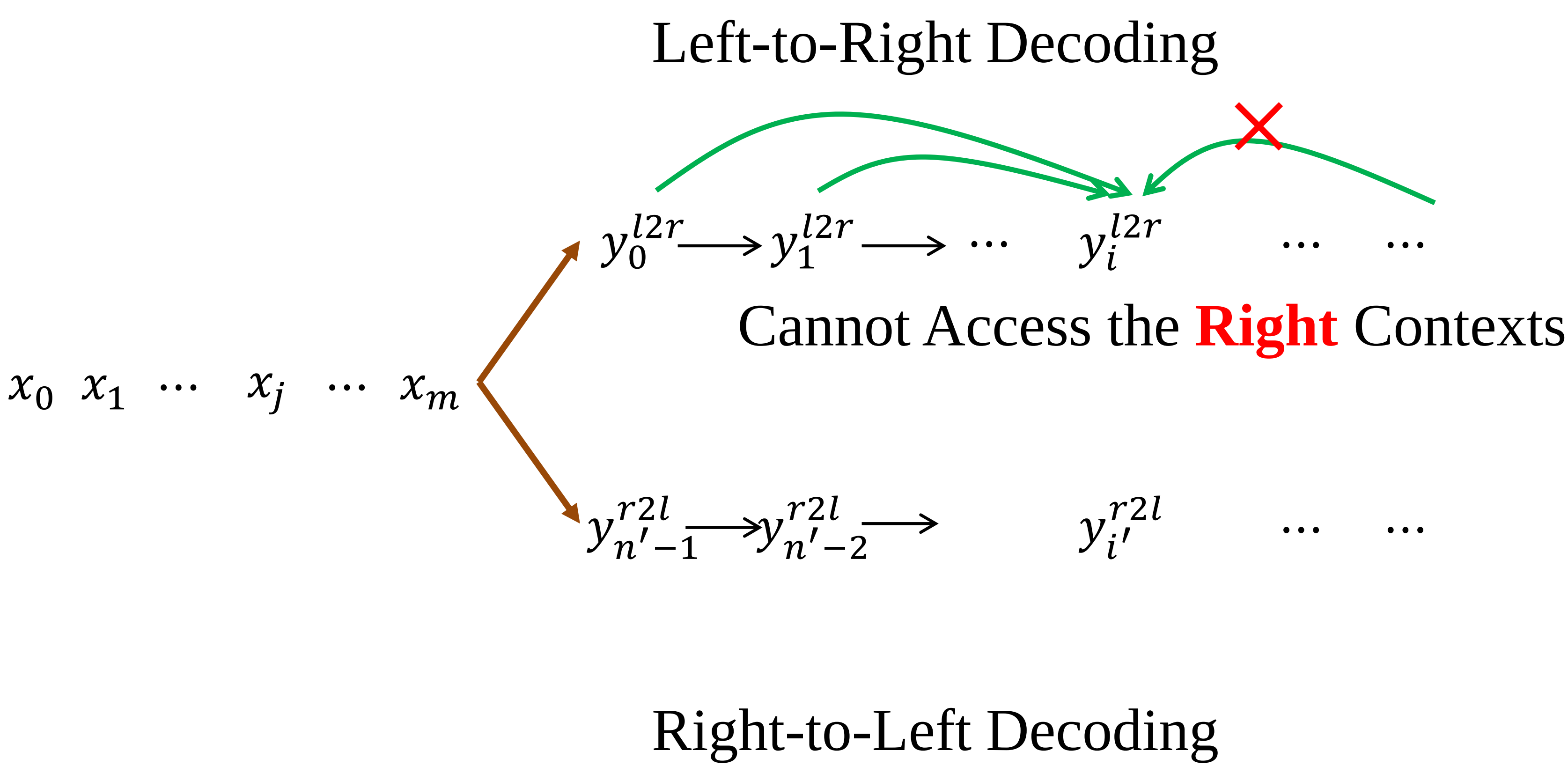
Left-to-Right Decoding

$$y_0^{l2r} \longrightarrow y_1^{l2r} \longrightarrow \cdots \quad y_i^{l2r} \quad \cdots \quad \cdots$$

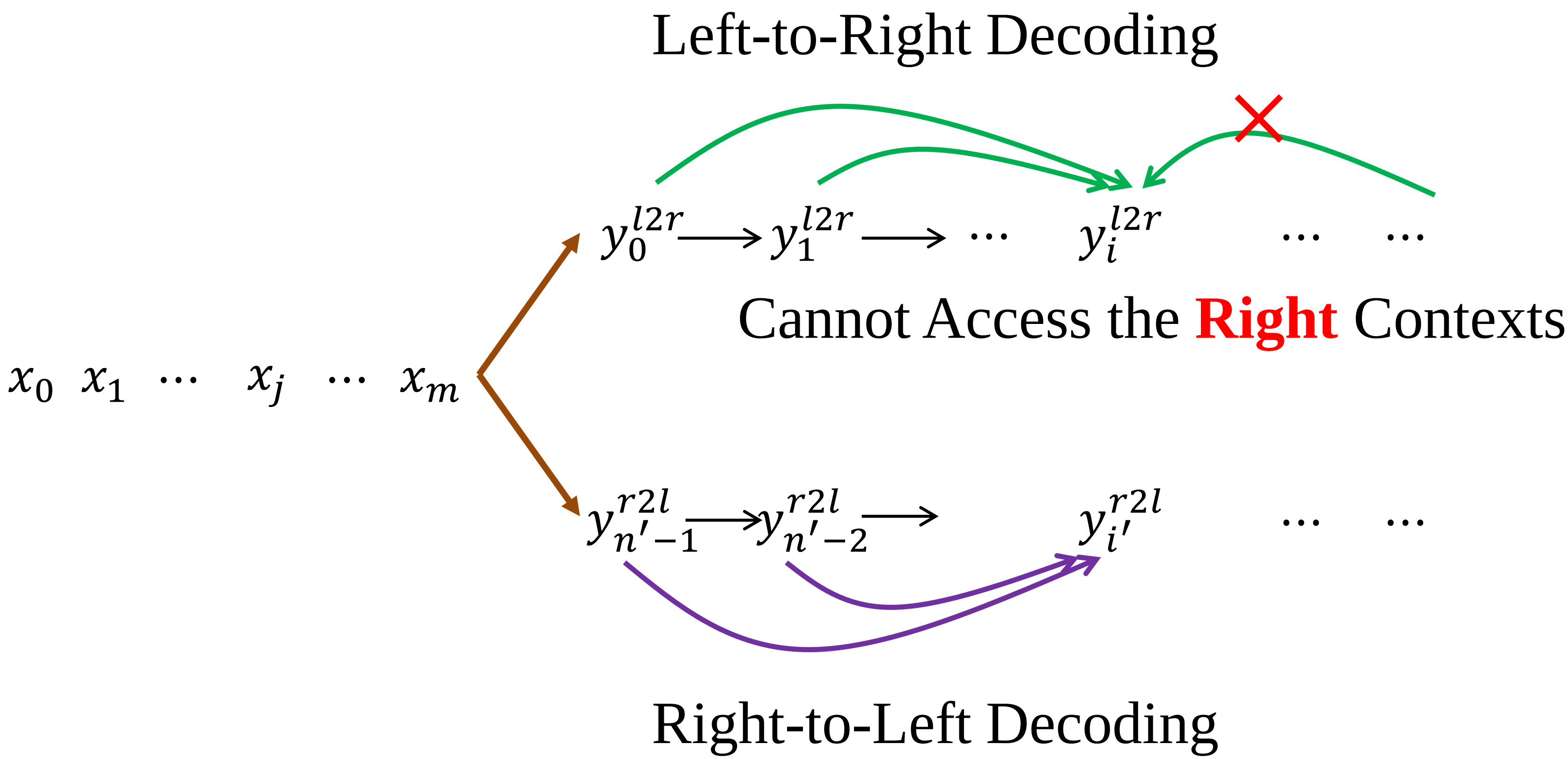
Cannot Access the **Right** Contexts

$$x_0 \quad x_1 \quad \cdots \quad x_j \quad \cdots \quad x_m$$

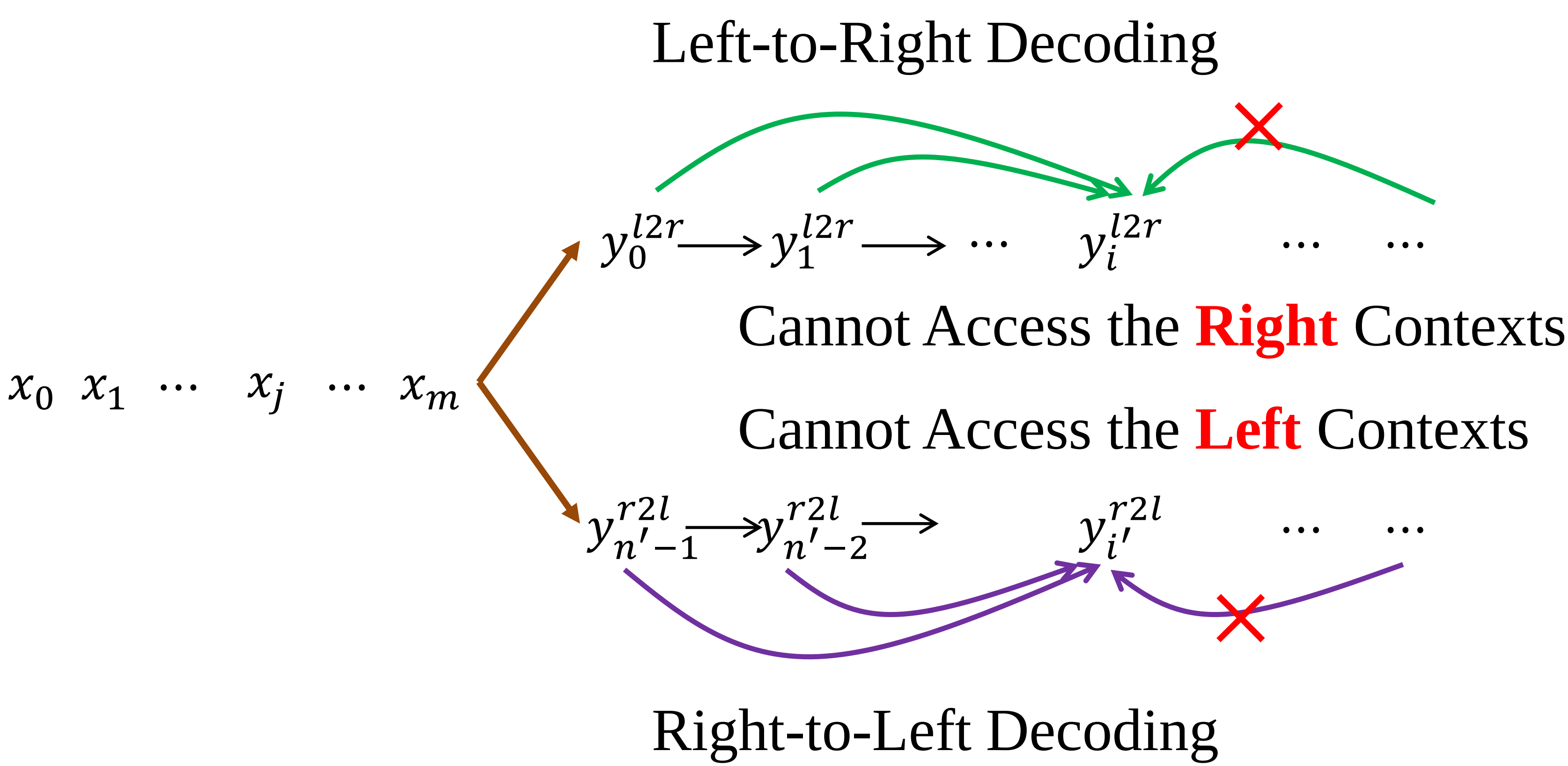
Problems for Unidirectional Inference



Problems for Unidirectional Inference



Problems for Unidirectional Inference



Problems: Unbalanced Outputs

Source	捷克 总统 哈维 卸任 新 总统 仍 未 确定
Reference	czech president havel steps down while new president still not chosen
L2R	czech president leaves office
R2L	the outgoing president of the czech republic is still uncertain

Source	他们 正在 研制 一 种 超大型 的 叫做 炸弹 之 母 。
Reference	they are developing a kind of superhuge bomb called the mother of bombs .
L2R	they are developing a super , big , mother , called the bomb .
R2L	they are working on a much larger mother called the mother of a bomb .

Problems: Unbalanced Outputs

- Statistical Analysis

Model	The first 4 tokens	The last 4 tokens
L2R	**40.21%**	35.10%
R2L	35.67%	**39.47%**

Table: Translation accuracy of the first 4 tokens and last 4 tokens in NIST Chinese-English translation tasks.

Problems: Unbalanced Outputs

- Statistical Analysis

Model	The first 4 tokens	The last 4 tokens
L2R	**40.21%**	35.10%
R2L	35.67%	**39.47%**

Table: Translation accuracy of the first 4 tokens and last 4 tokens in NIST Chinese-English translation tasks.

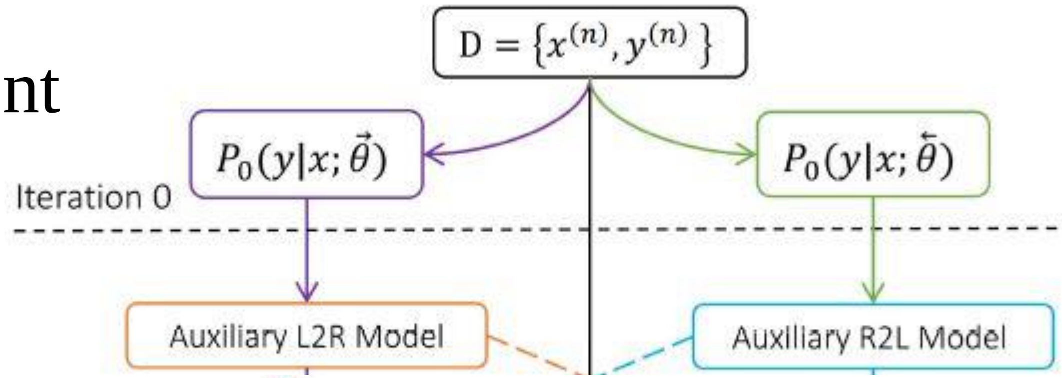
How to effectively utilize bidirectional decoding?

Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Two Tasks
- Summary and Future Challenges

Solution 1: Bidirectional Agreement from Perspective of Loss Function

- Agreement

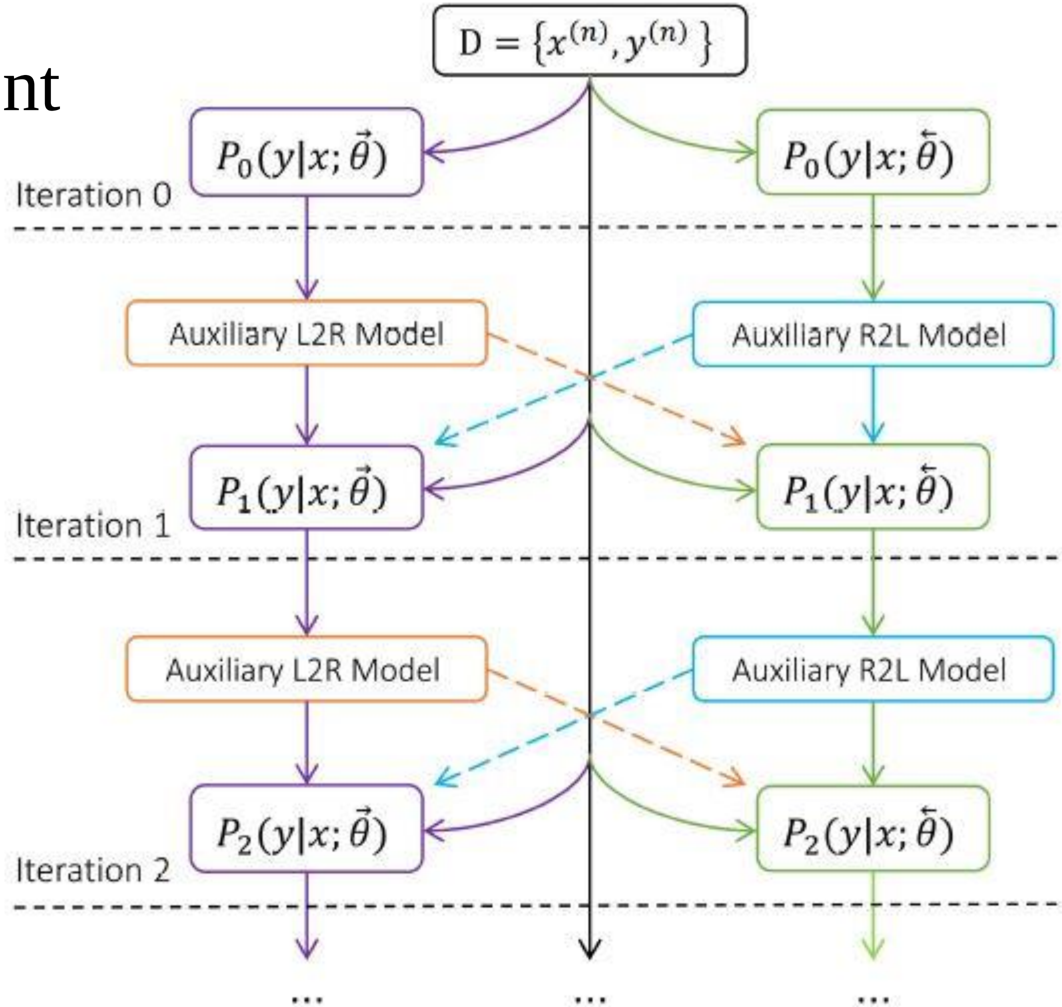


$$L(\overrightarrow{\theta}) = \sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)}; \overrightarrow{\theta})$$
$$- \lambda \sum_{n=1}^{N} \mathrm{KL}(P(y|x^{(n)}; \overleftarrow{\theta})||P(y|x^{(n)}; \overrightarrow{\theta}))$$
$$- \lambda \sum_{n=1}^{N} \mathrm{KL}(P(y|x^{(n)}; \overrightarrow{\theta})||P(y|x^{(n)}; \overleftarrow{\theta}))$$

[Liu et al., 2016] Agreement on Target-bidirectional Neural Machine Translation. NAACL.
[Zhang et al., 2019] Regularizing Neural Machine Translation by Target-Bidirectional Agreement. AAAI

Solution 1: Bidirectional Agreement from Perspective of Loss Function

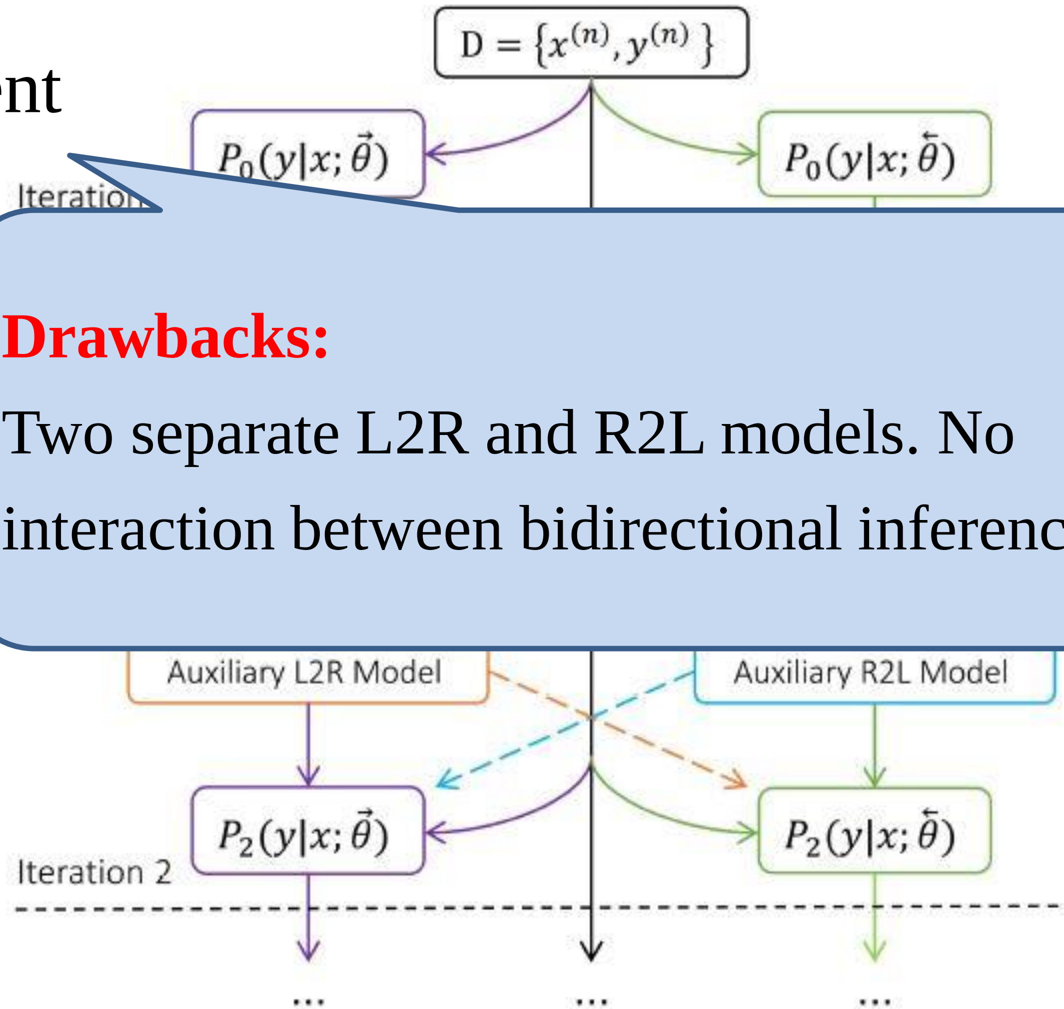
- Agreement



[Liu et al., 2016] Agreement on Target-bidirectional Neural Machine Translation. NAACL.
[Zhang et al., 2019] Regularizing Neural Machine Translation by Target-Bidirectional Agreement. AAAI

Solution 1: Bidirectional Agreement from Perspective of Loss Function

- Agreement



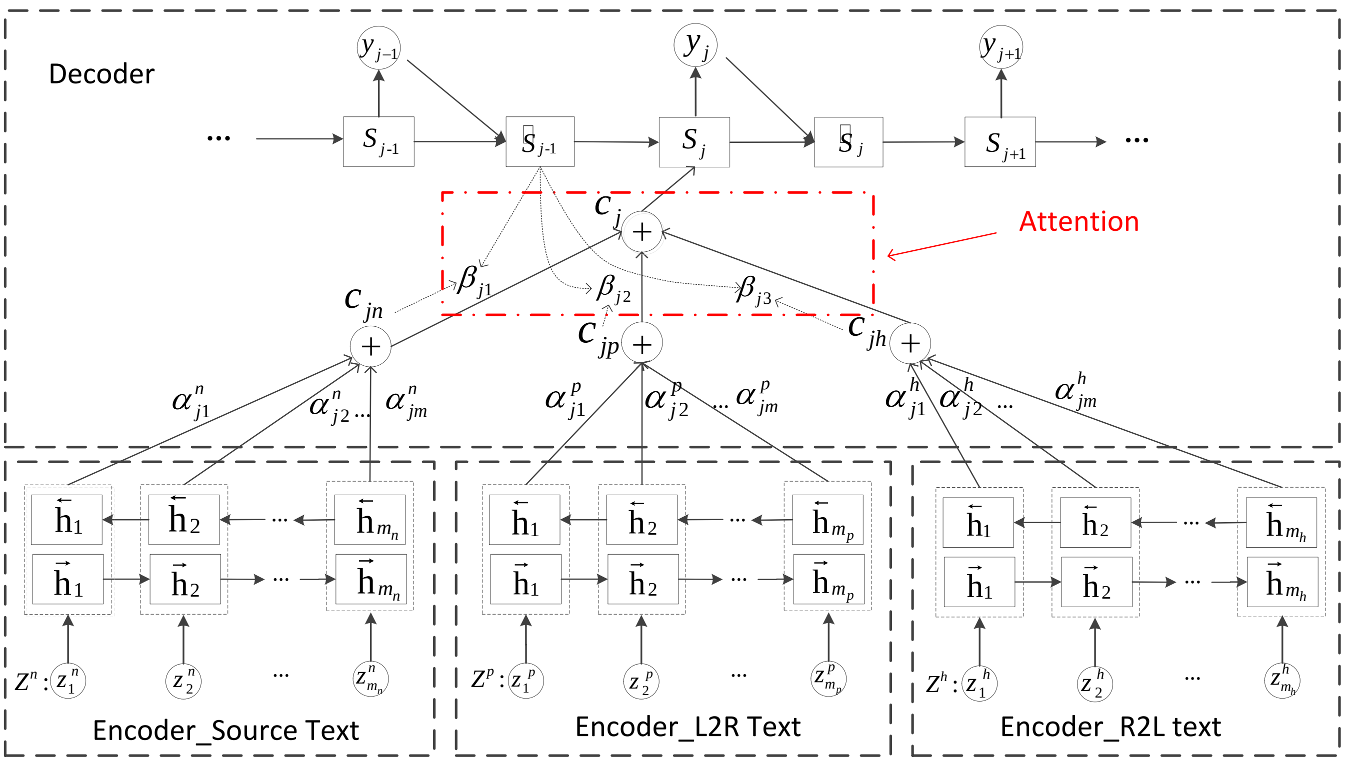
Drawbacks:

Two separate L2R and R2L models. No interaction between bidirectional inference.

[Liu et al., 2016] Agreement on Target-bidirectional Neural Machine Translation. NAACL.
[Zhang et al., 2019] Regularizing Neural Machine Translation by Target-Bidirectional Agreement. AAAI

Solution 2: Neural System Combination from the Perspective of Ensemble

- NSC-NMT



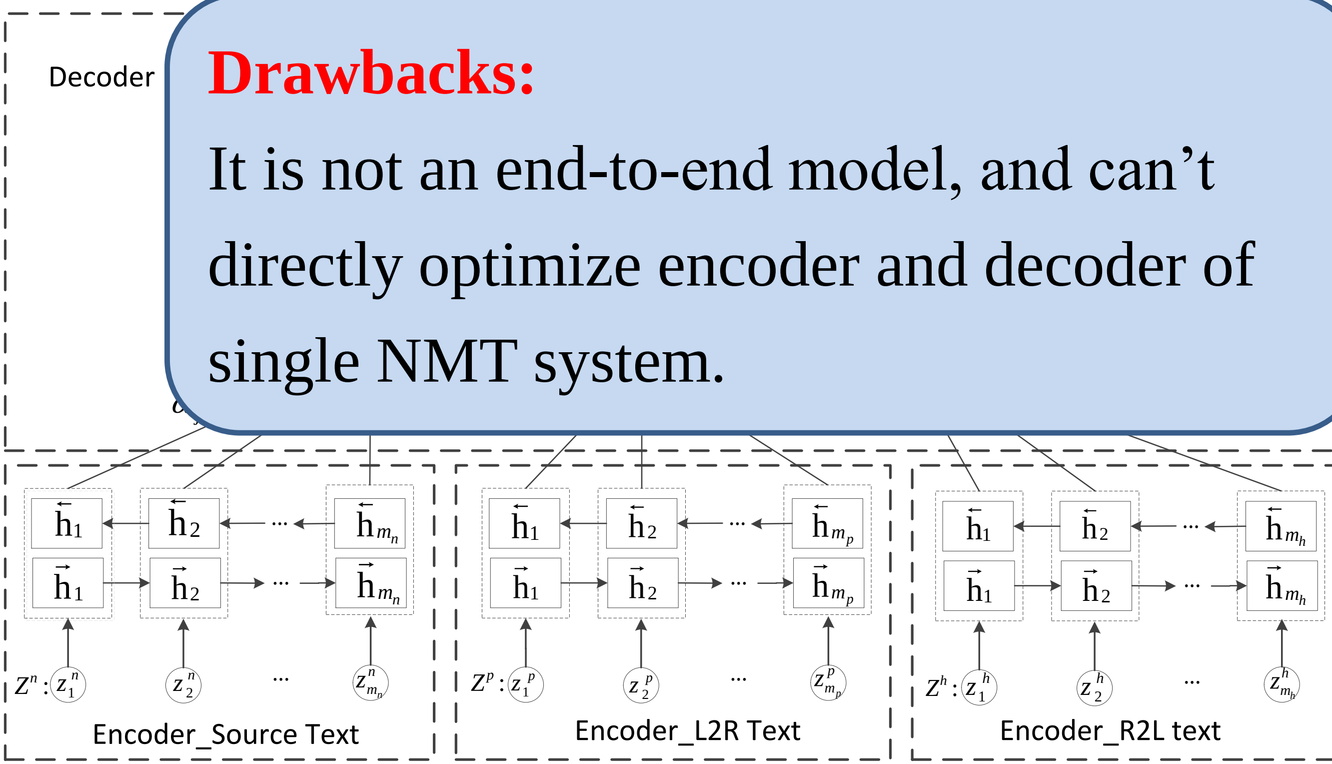
[Zhou et al., 2017] Neural System Combination for Machine Translation. ACL.

Solution 2: Neural System Combination from the Perspective of Ensemble

- NSC-NMT

Drawbacks:

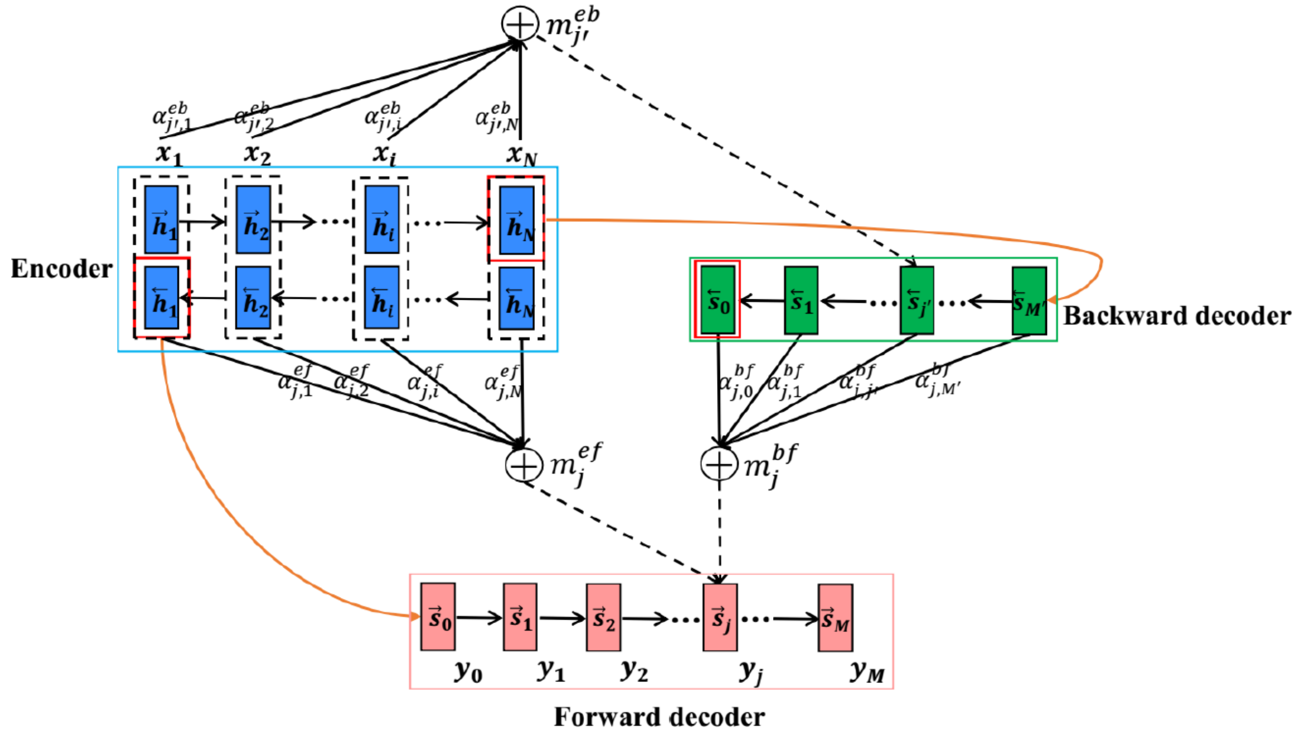
It is not an end-to-end model, and can't directly optimize encoder and decoder of single NMT system.



[Zhou et al., 2017] Neural System Combination for Machine Translation. ACL.

Solution 3: Asynchronous Bidirectional Decoding from the Perspective of Model Integration

- ABD-NMT



[Zhang et al., 2018] Asynchronous Bidirectional Decoding for Neural Machine Translation. AAAI.

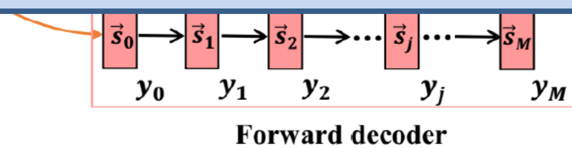
Solution 3: Asynchronous Bidirectional Decoding from the Perspective of Model Integration

- ABD-NMT

Drawbacks:

(1) This work still requires two NMT models or decoders.

(2) Only the forward decoder can utilize information of backward decoder.



[Zhang et al., 2018] Asynchronous Bidirectional Decoding for Neural Machine Translation. AAAI.

Solution 3: Asynchronous Bidirectional Decoding from the Perspective of Model Integration

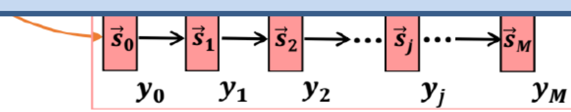
- ABD-NMT

Drawbacks:

(1) This work still requires two NMT models or decoders.

(2) Only the forward decoder can utilize information of backward decoder.

Question: How to utilize bidirectional decoding more effectively and efficiently?



Forward decoder

[Zhang et al., 2018] Asynchronous Bidirectional Decoding for Neural Machine Translation. AAAI.

Solution 4

Solution 4

Synchronous Bidirectional Neural Machine Translation

Long Zhou, Jiajun Zhang and Chengqing Zong.

Transactions on ACL 2019.

Synchronous Bidirectional Neural Machine Translation

$\langle l2r \rangle$

$x_0 \;\; x_1 \;\; \cdots \;\; x_i \;\; \cdots \;\; x_m$

$\langle r2l \rangle$

Synchronous Bidirectional Neural Machine Translation

$\langle l2r \rangle$

$x_0 \ x_1 \ \cdots \ x_i \ \cdots \ x_m$

$\langle r2l \rangle$

T_0

Synchronous Bidirectional Neural Machine Translation

$\langle l2r \rangle$ y_0^{l2r}

$x_0 \; x_1 \; \cdots \; x_i \; \cdots \; x_m$

$\langle r2l \rangle$

T_0

Synchronous Bidirectional Neural Machine Translation

$\langle l2r \rangle$

y_0^{l2r}

$x_0 \;\; x_1 \;\; \cdots \;\; x_i \;\; \cdots \;\; x_m$

$y_{n'-1}^{r2l}$

$\langle r2l \rangle$

T_0

Synchronous Bidirectional Neural Machine Translation

$\langle l2r \rangle$ y_0^{l2r}

$x_0 \; x_1 \; \cdots \; x_i \; \cdots \; x_m$

$\langle r2l \rangle$ $y_{n'-1}^{r2l}$

T_0 $\quad$ T_1

Synchronous Bidirectional Neural Machine Translation

$\langle l2r \rangle$

y_0^{l2r}

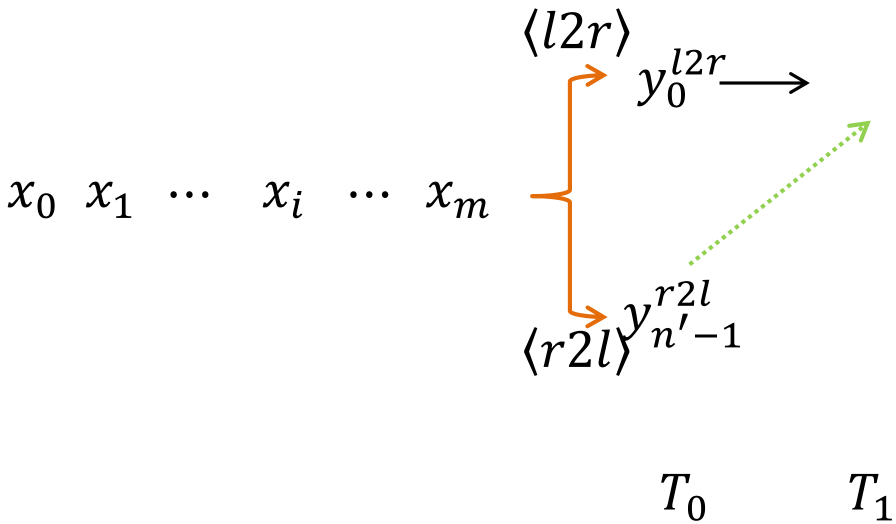
$x_0 \quad x_1 \quad \cdots \quad x_i \quad \cdots \quad x_m$

$y_{n'-1}^{r2l}$

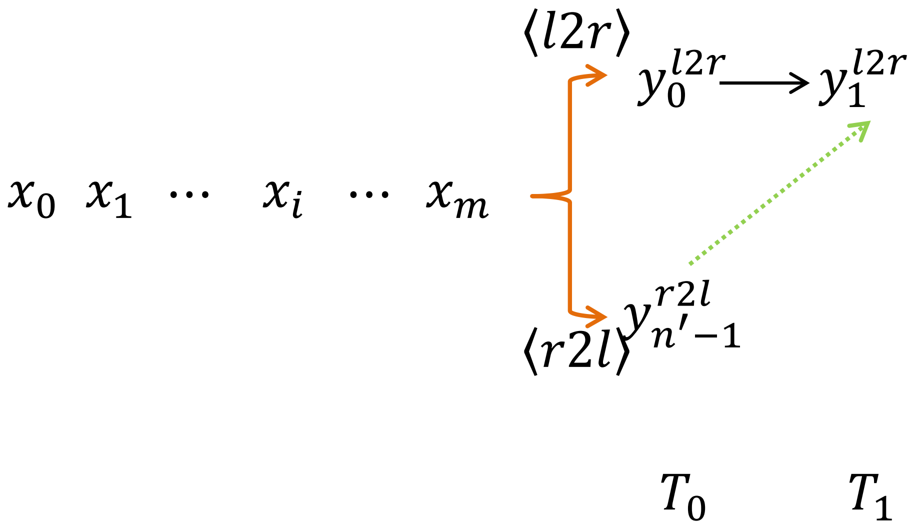
$\langle r2l \rangle$

$T_0 \qquad T_1$

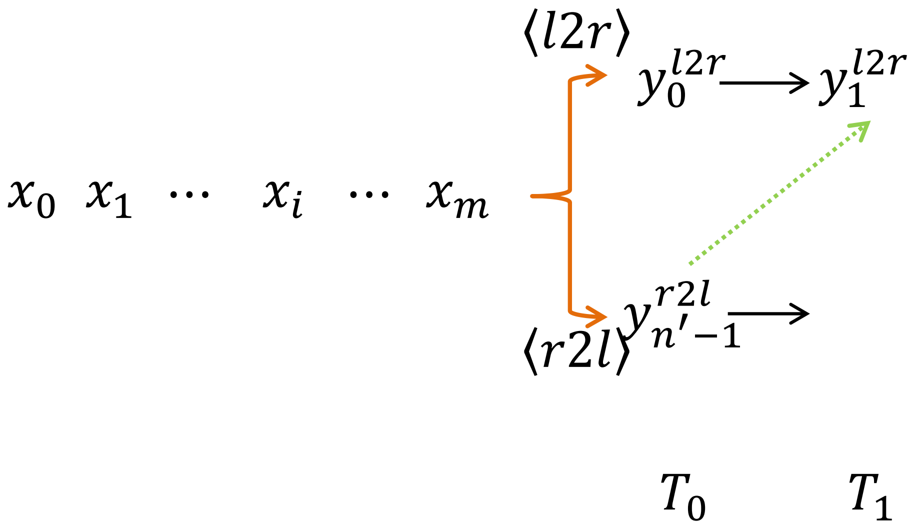
Synchronous Bidirectional Neural Machine Translation



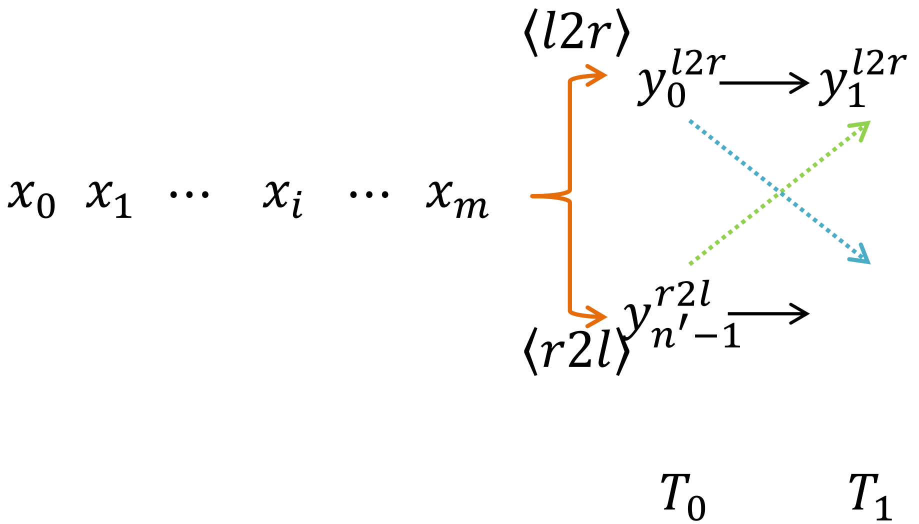
Synchronous Bidirectional Neural Machine Translation



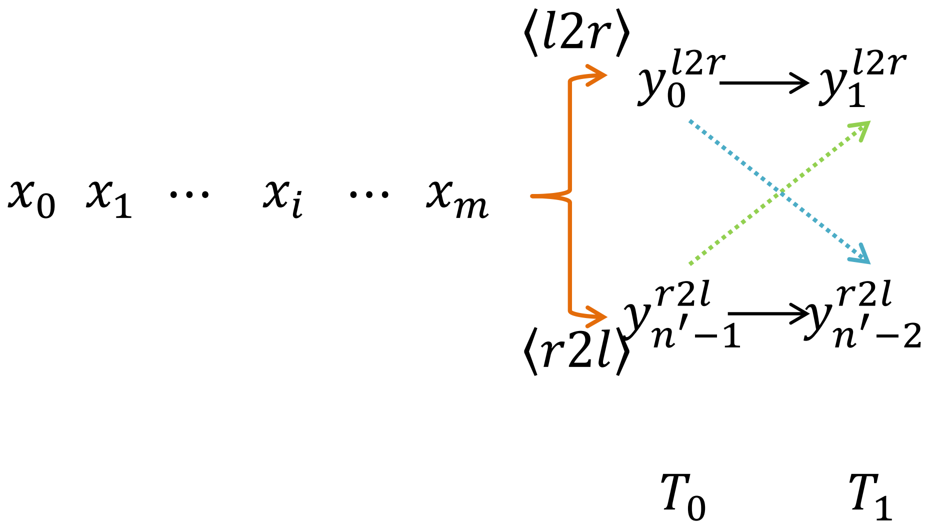
Synchronous Bidirectional Neural Machine Translation



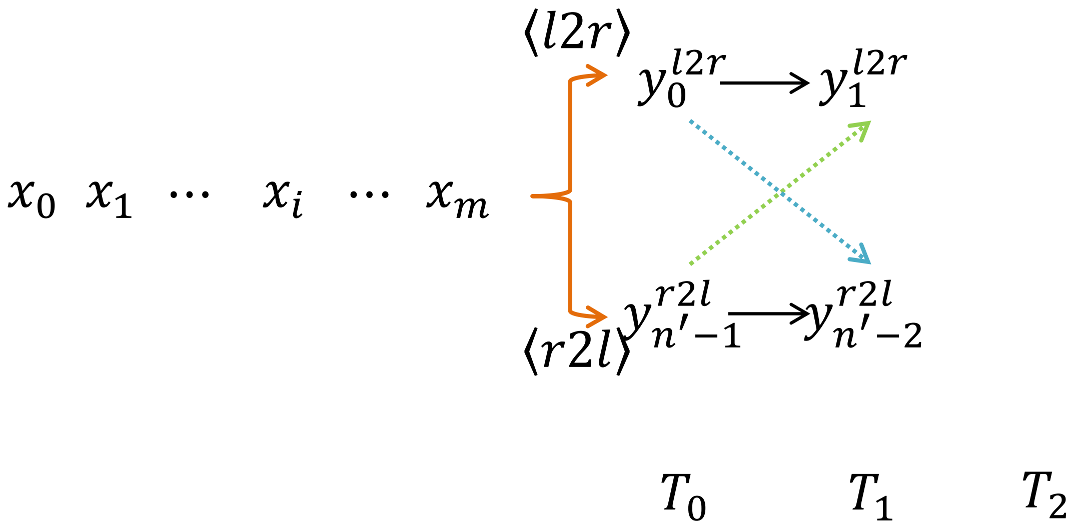
Synchronous Bidirectional Neural Machine Translation



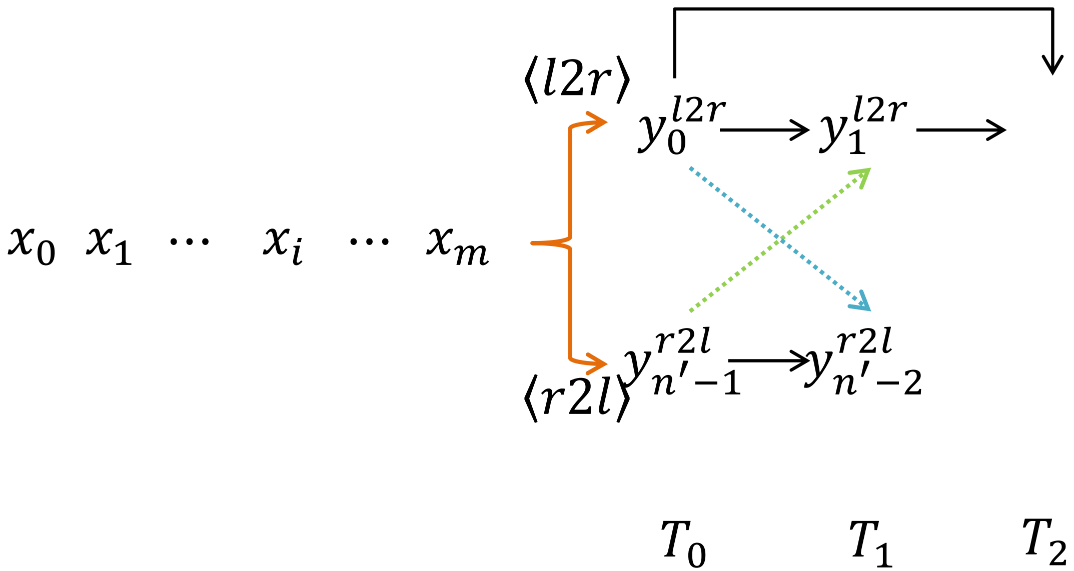
Synchronous Bidirectional Neural Machine Translation



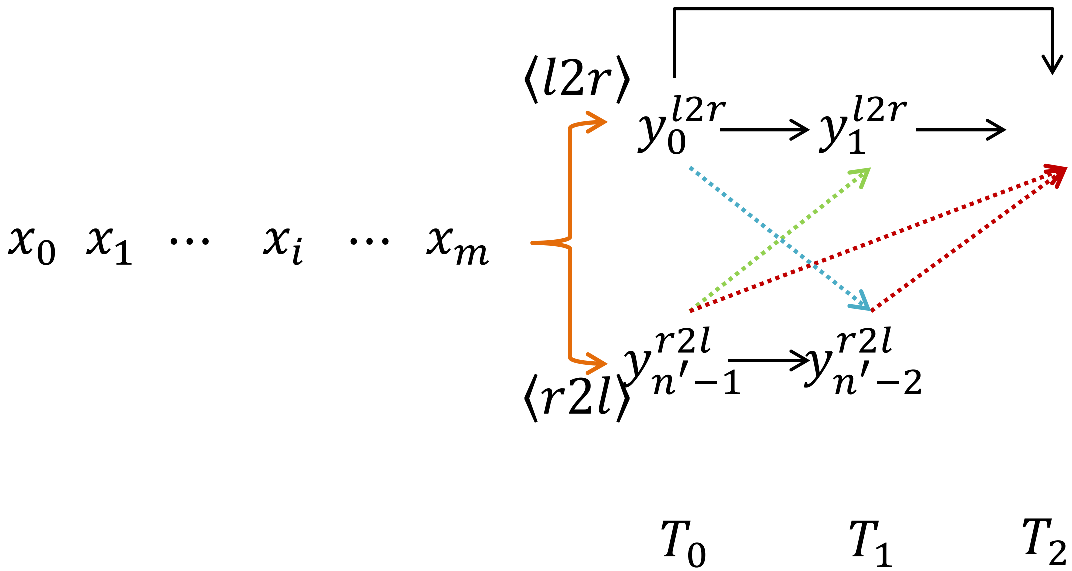
Synchronous Bidirectional Neural Machine Translation



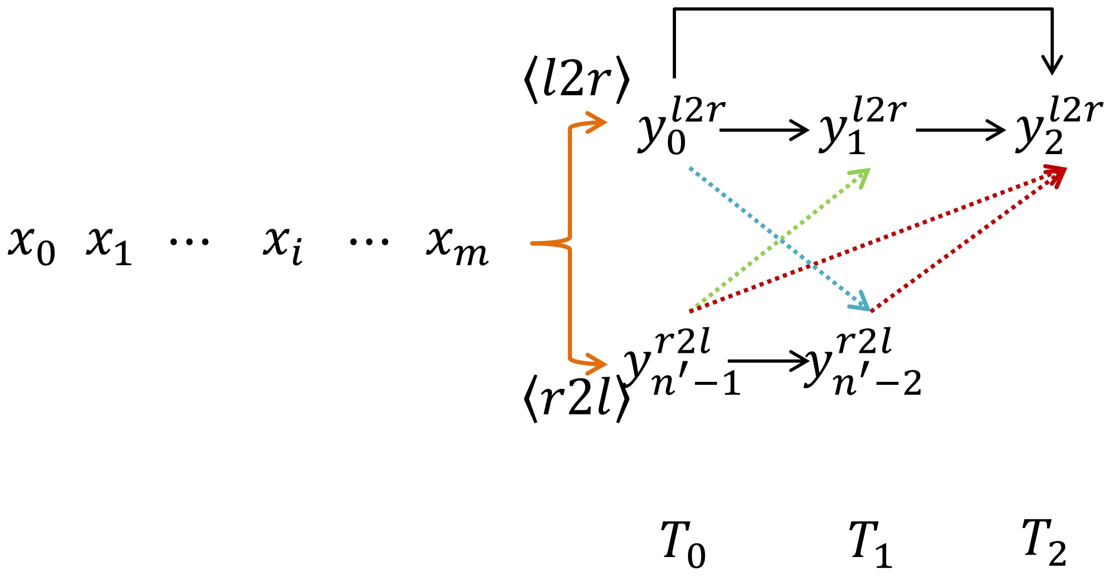
Synchronous Bidirectional Neural Machine Translation



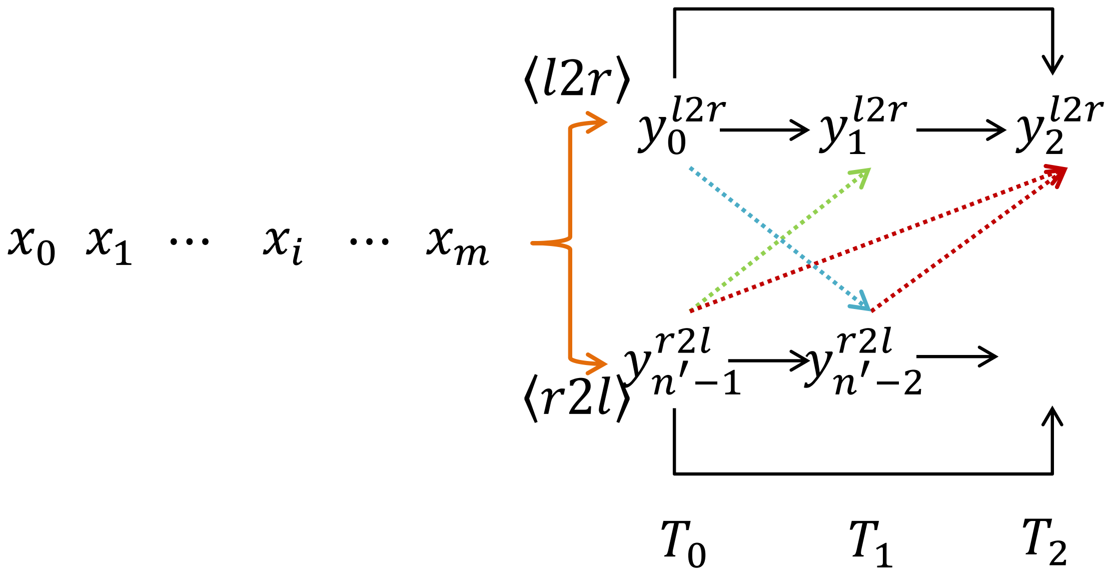
Synchronous Bidirectional Neural Machine Translation



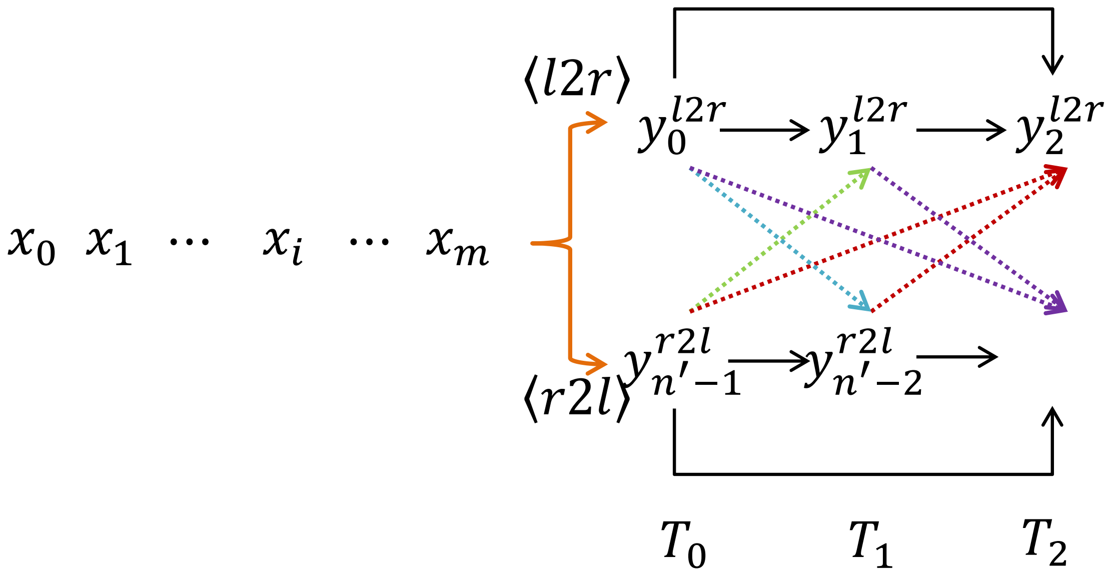
Synchronous Bidirectional Neural Machine Translation



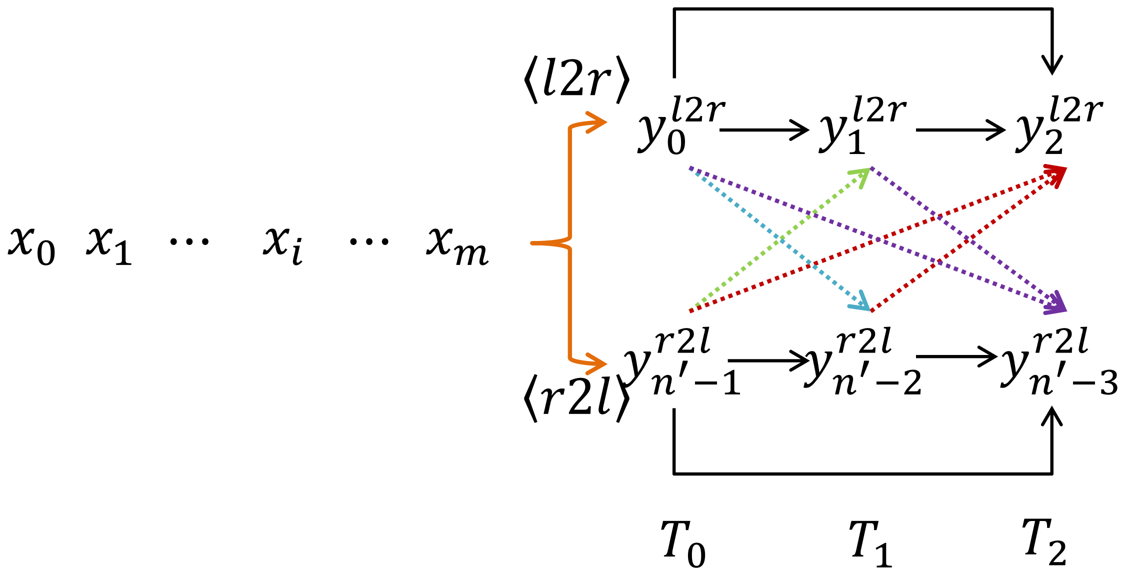
Synchronous Bidirectional Neural Machine Translation



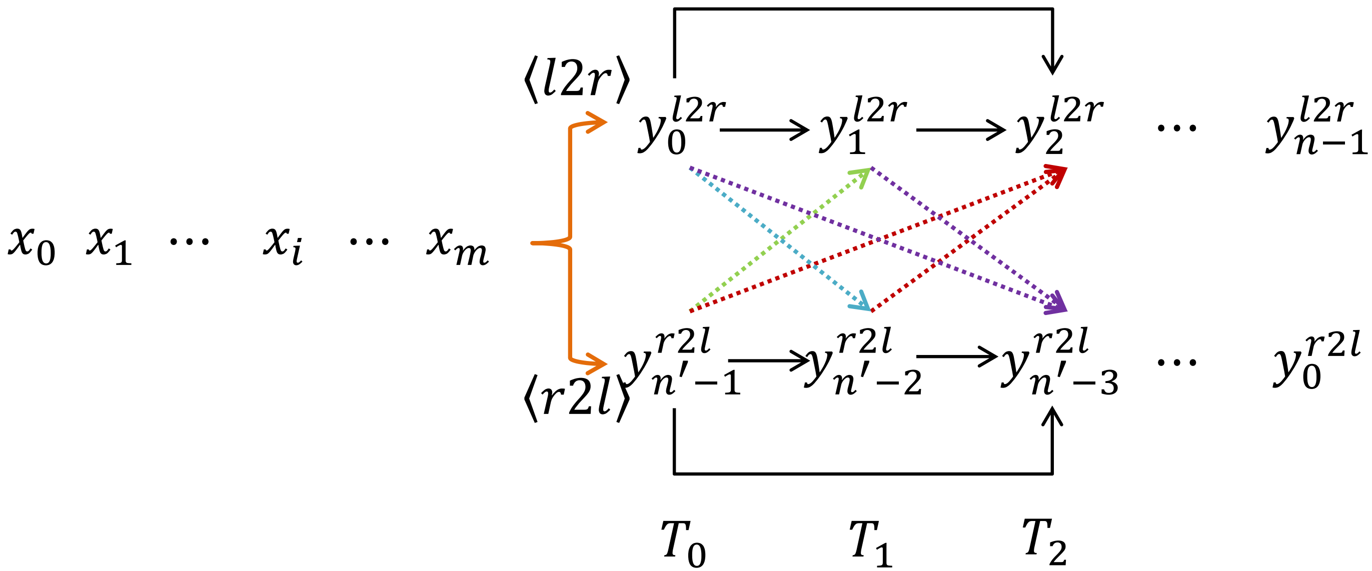
Synchronous Bidirectional Neural Machine Translation



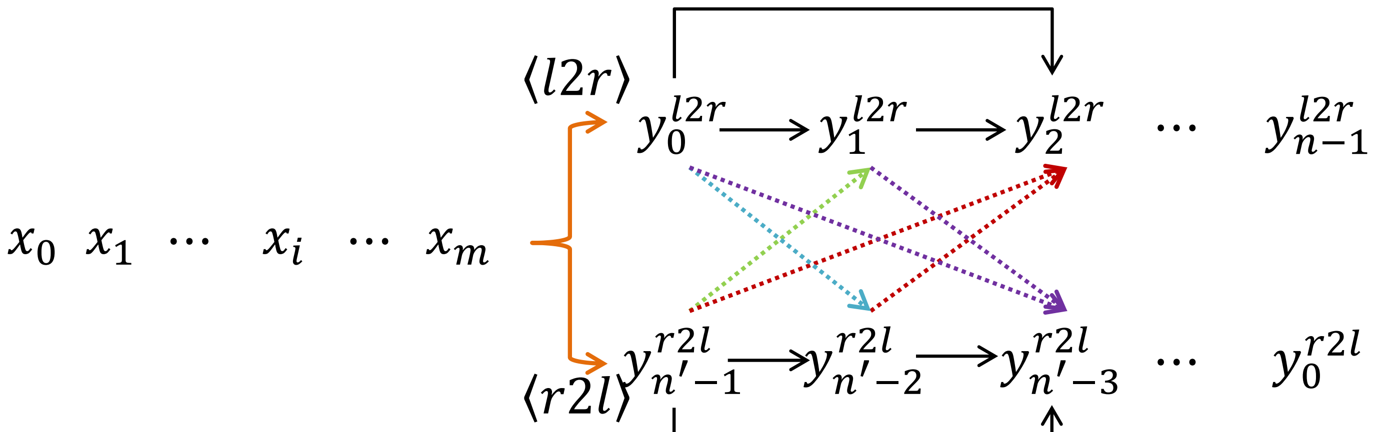
Synchronous Bidirectional Neural Machine Translation



Synchronous Bidirectional Neural Machine Translation

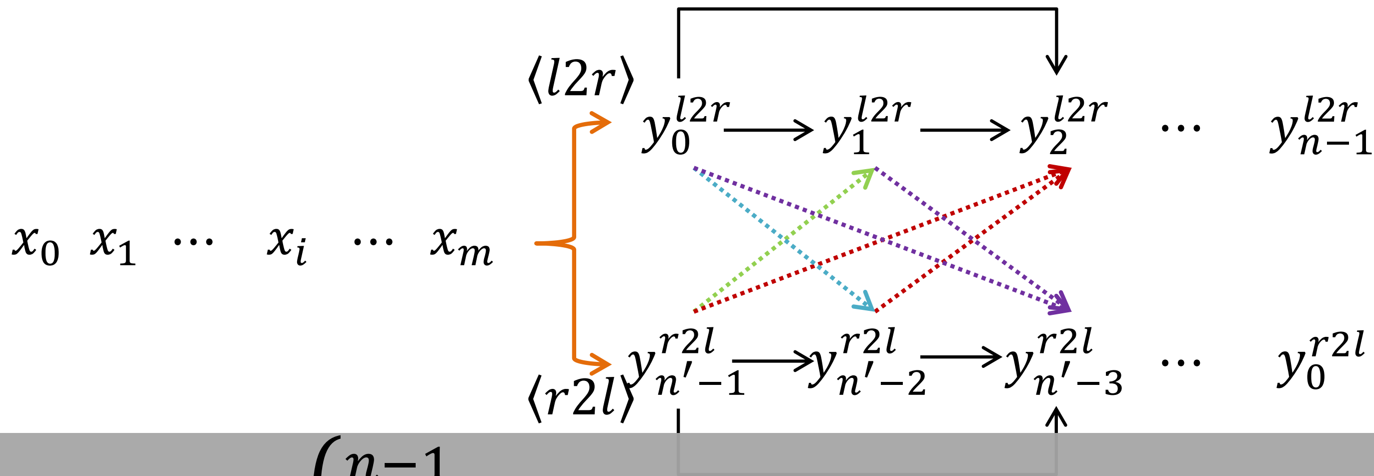


Synchronous Bidirectional Neural Machine Translation



$$P(y|x) = \begin{cases} \sum_{i=0}^{n-1} p(\vec{y}_i|\vec{y}_0 \cdots \vec{y}_{i-1}, x, \overleftarrow{\boldsymbol{y}}_{\boldsymbol{0}} \cdots \overleftarrow{\boldsymbol{y}}_{\boldsymbol{i-1}}) & if\ L2R \\ \sum_{i=0}^{n'-1} p(\overleftarrow{y}_i|\overleftarrow{y}_0 \cdots \overleftarrow{y}_{i-1}, x, \vec{\boldsymbol{y}}_{\boldsymbol{0}} \cdots \vec{\boldsymbol{y}}_{\boldsymbol{i-1}}) & if\ R2L \end{cases}$$

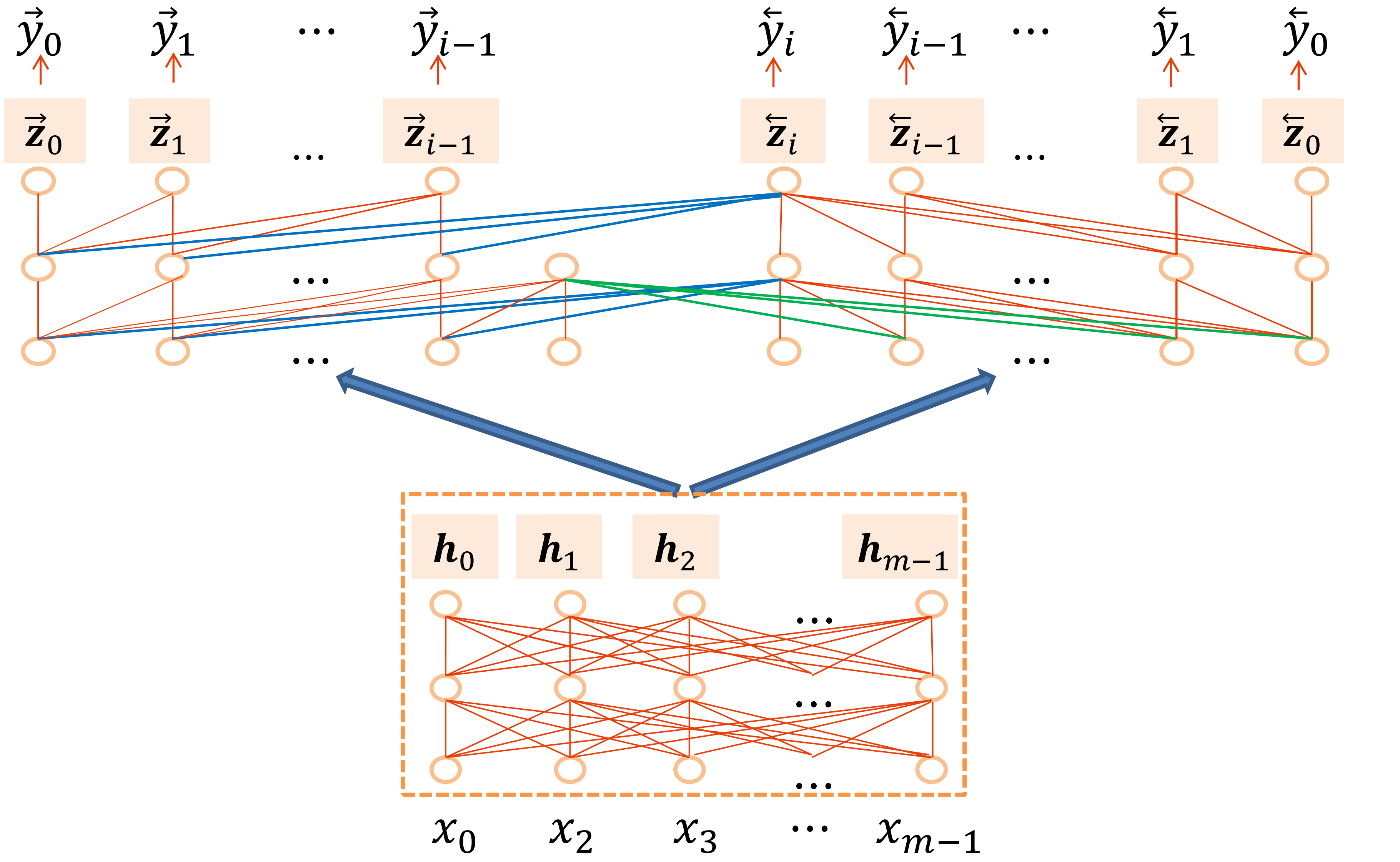
Synchronous Bidirectional Neural Machine Translation



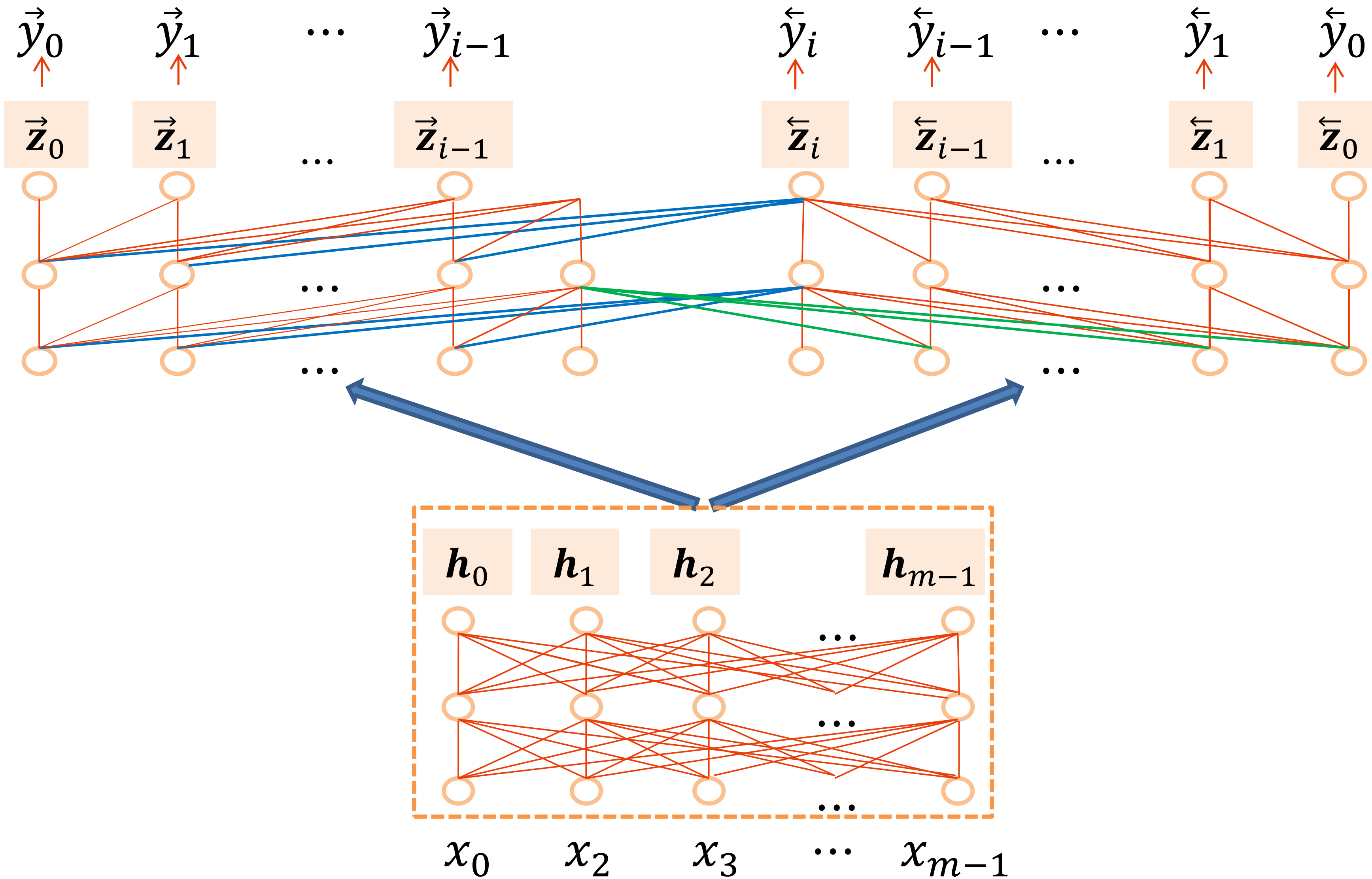
$$P(y|x) = \begin{cases} \sum_{i=0}^{n-1} p(\vec{y}_i | \vec{y}_0 \cdots \vec{y}_{i-1}, x, \overleftarrow{\mathbf{y}}_{\mathbf{0}} \cdots \overleftarrow{\mathbf{y}}_{\boldsymbol{i-1}}) \quad if\ L2R \\ \sum_{i=0}^{n'-1} p(\overleftarrow{y}_i | \overleftarrow{y}_0 \cdots \overleftarrow{y}_{i-1}, x, \overrightarrow{\mathbf{y}}_{\mathbf{0}} \cdots \overrightarrow{\mathbf{y}}_{\boldsymbol{i-1}}) \quad if\ R2L \end{cases}$$

L2R (R2L) inference not only uses its previously generated outputs, but also uses future contexts predicted by R2L (L2R) decoding.

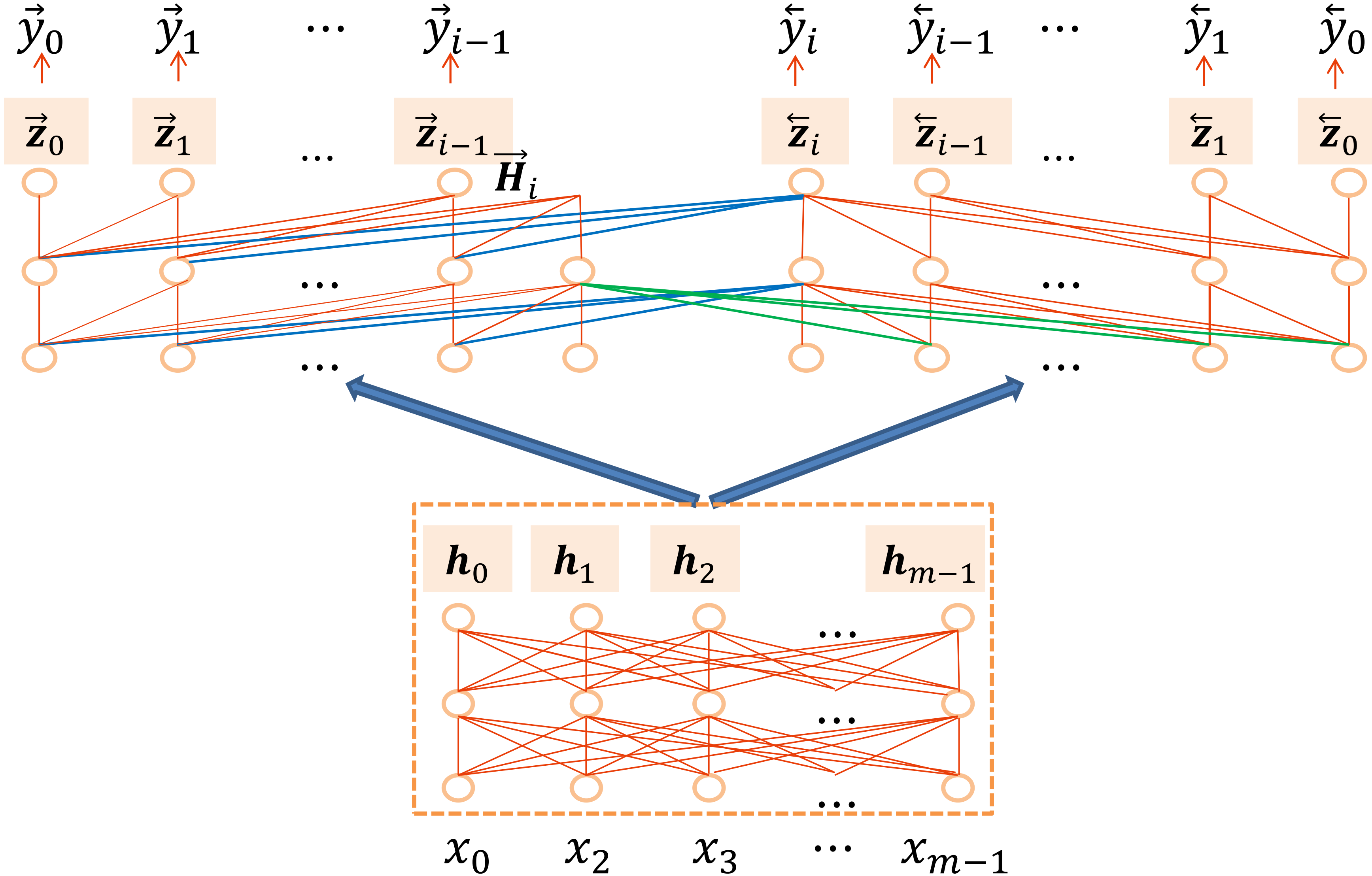
Synchronous Bidirectional Attention



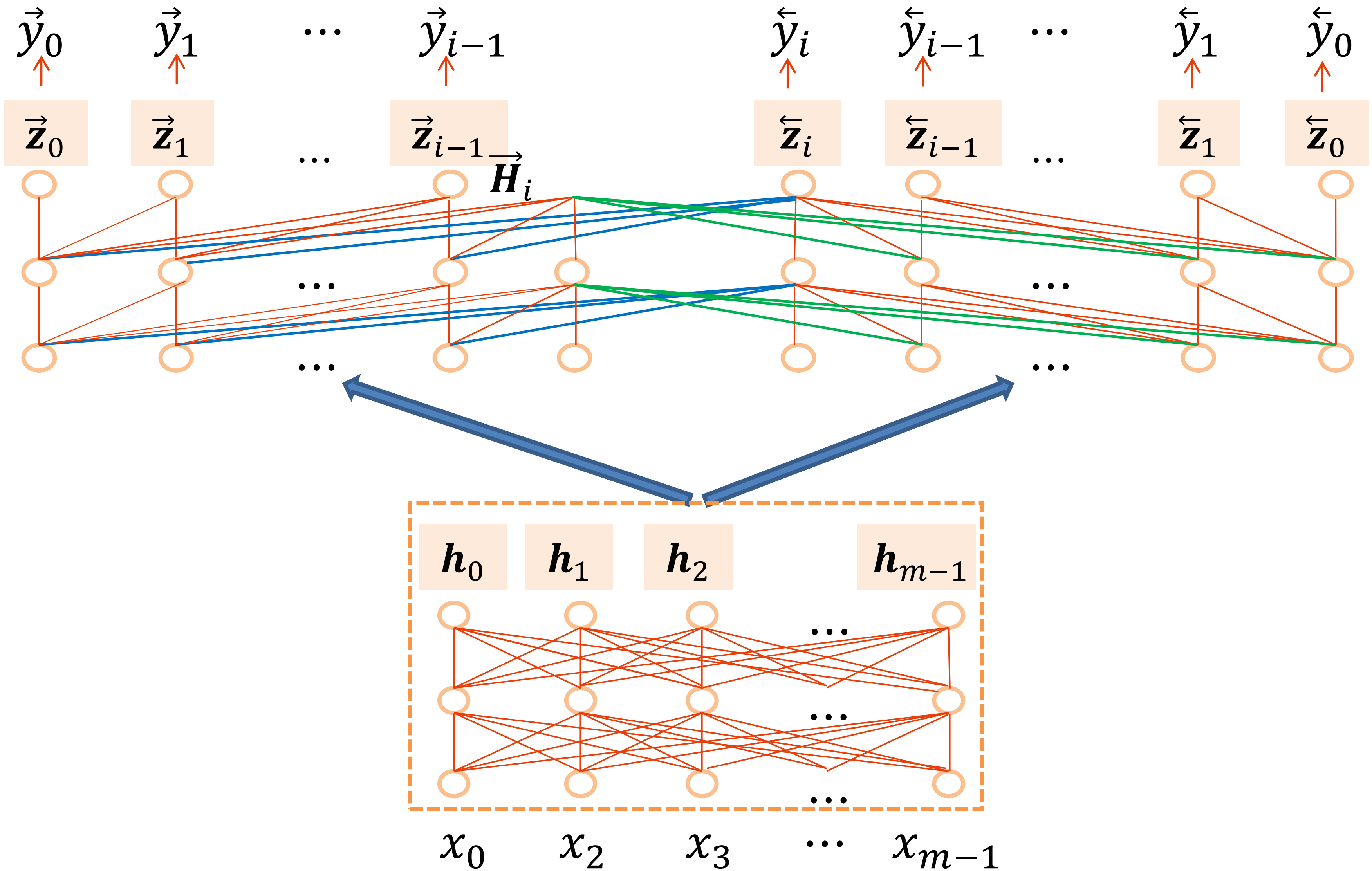
Synchronous Bidirectional Attention



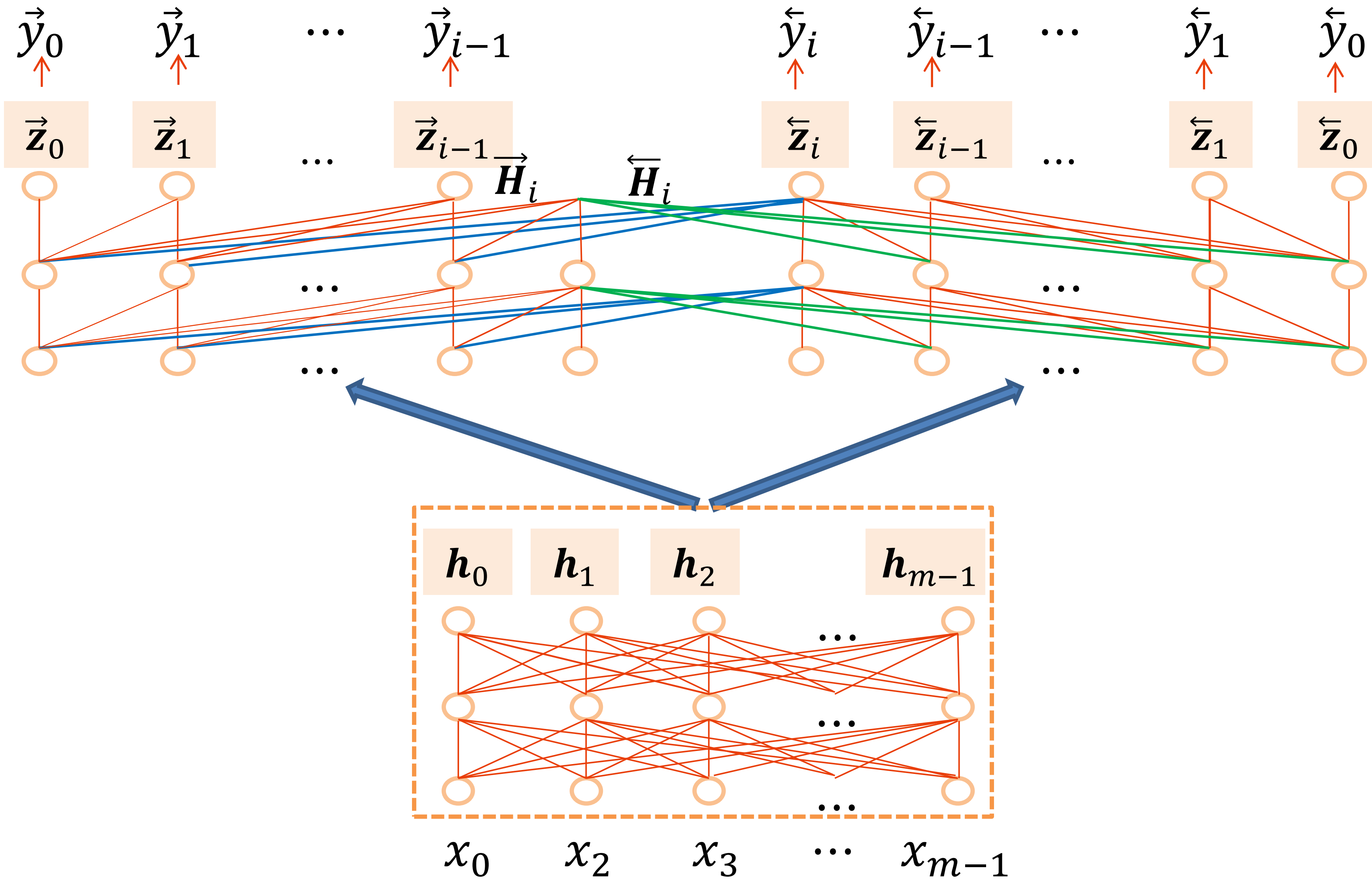
Synchronous Bidirectional Attention



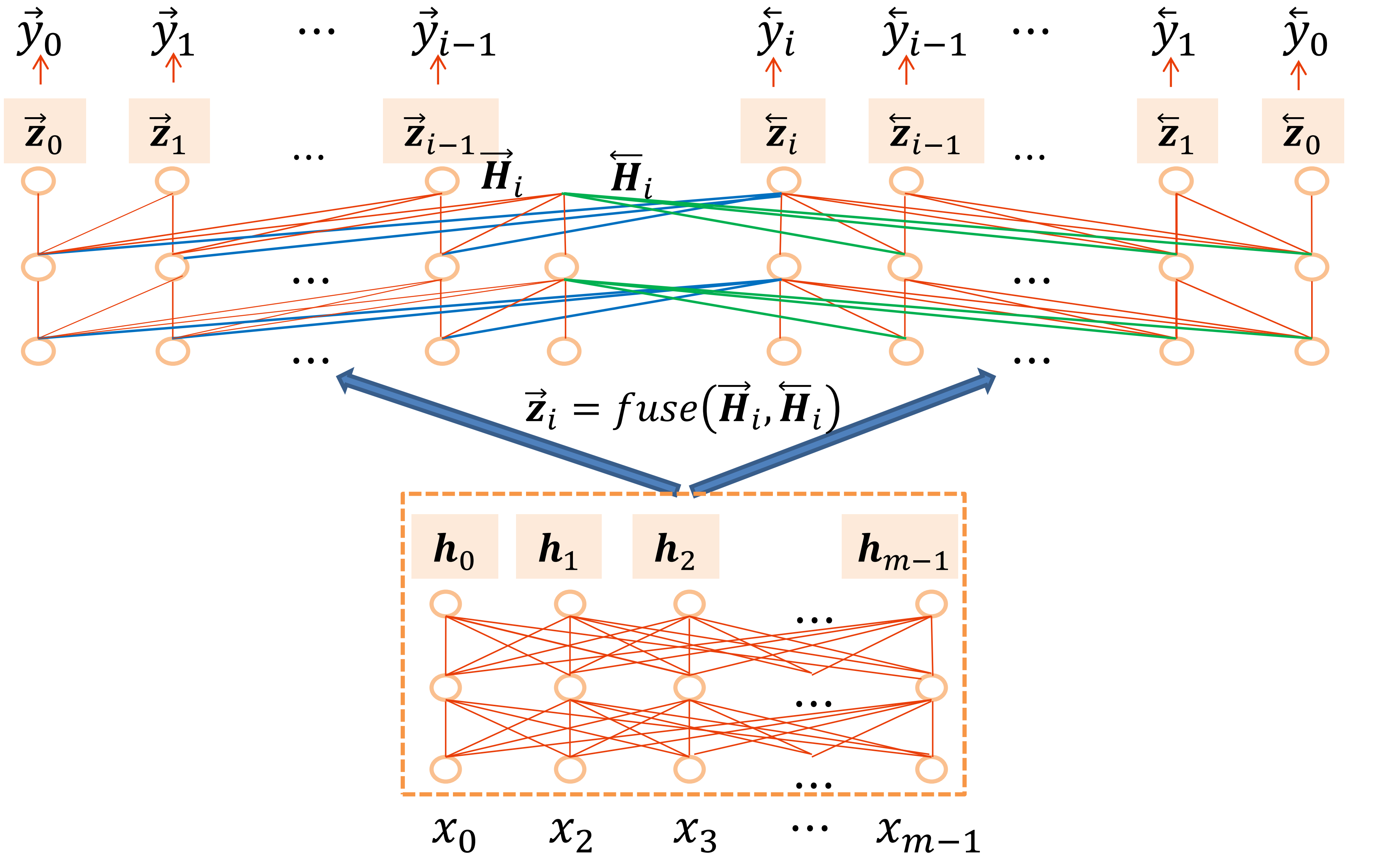
Synchronous Bidirectional Attention



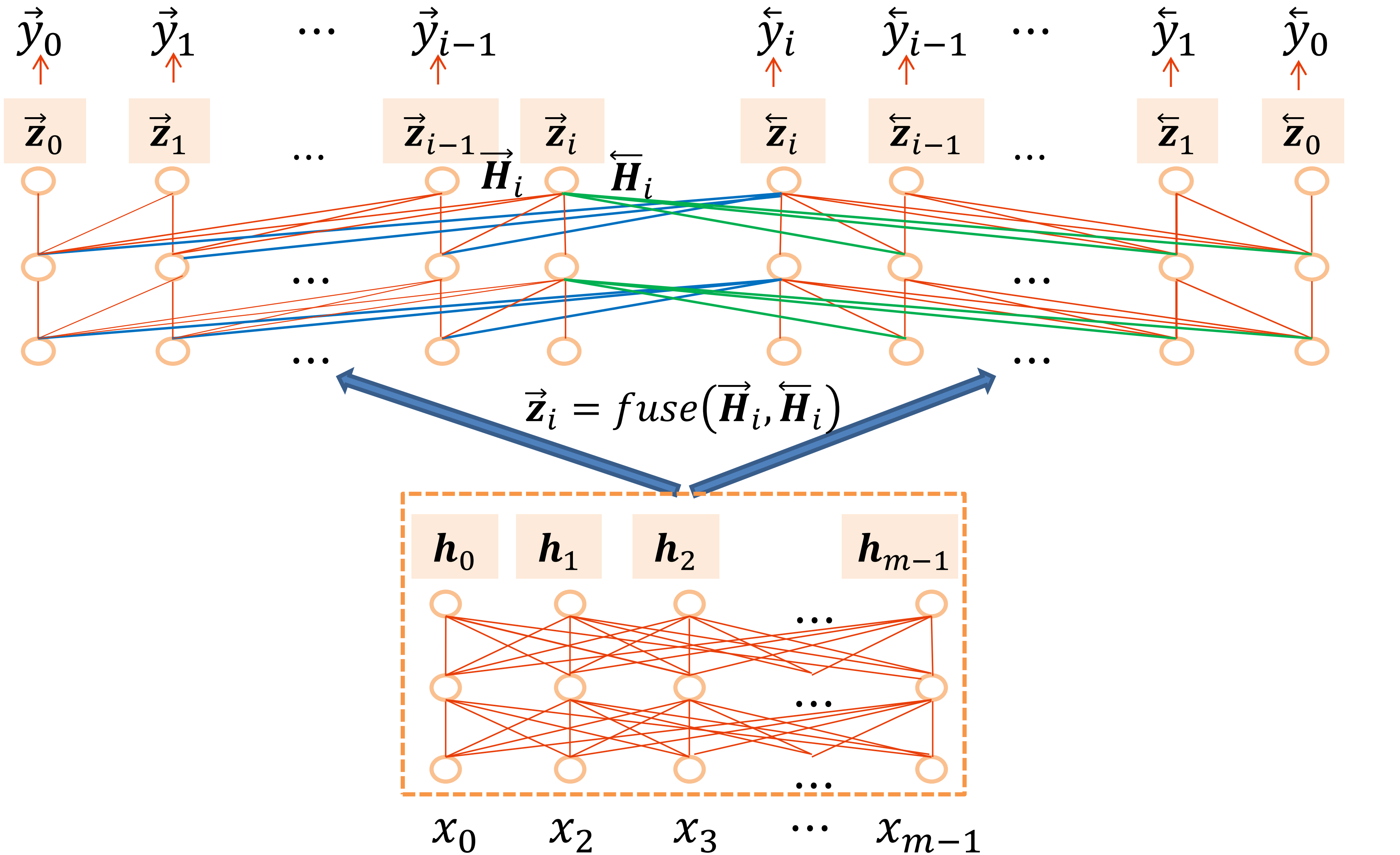
Synchronous Bidirectional Attention



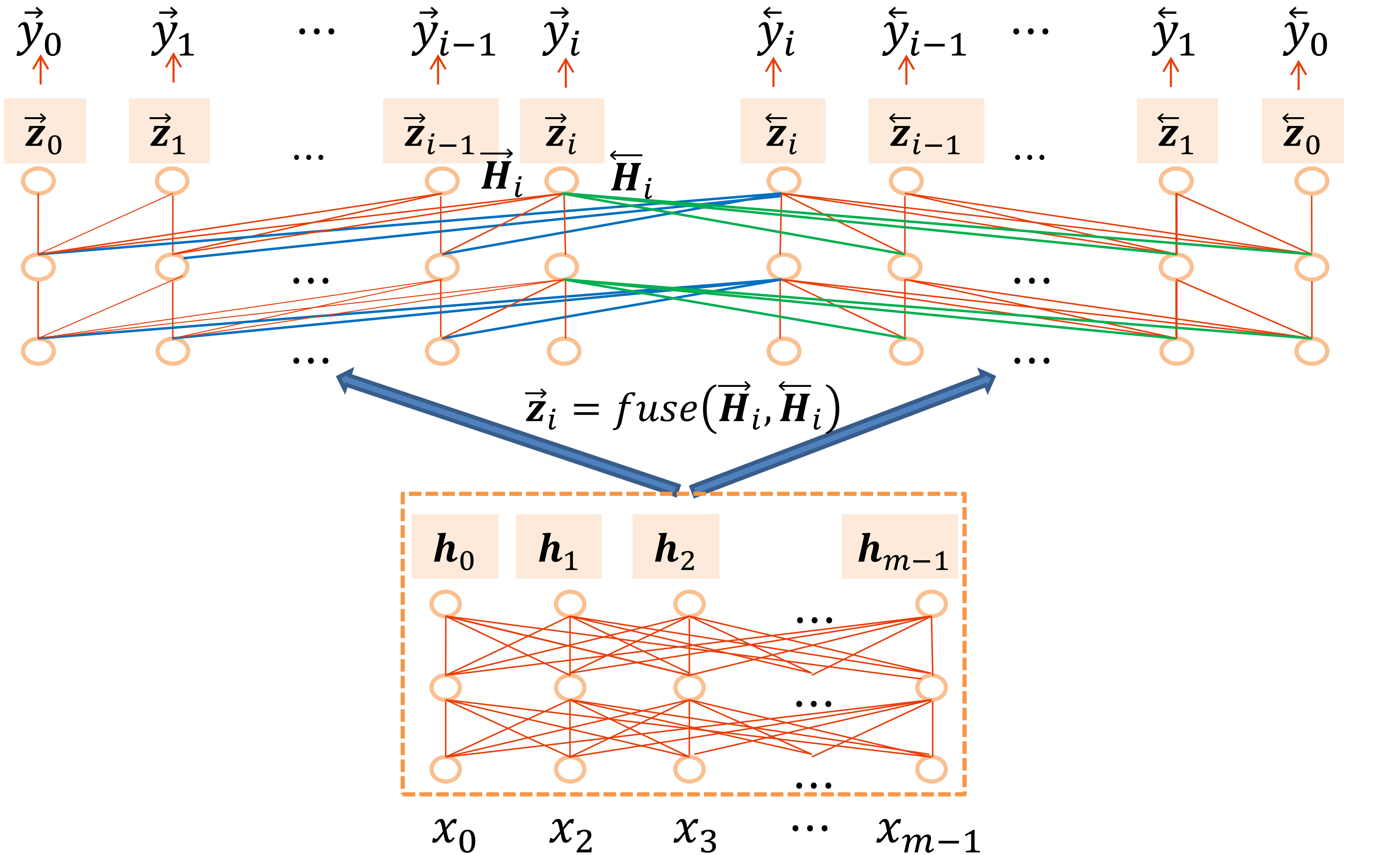
Synchronous Bidirectional Attention



Synchronous Bidirectional Attention

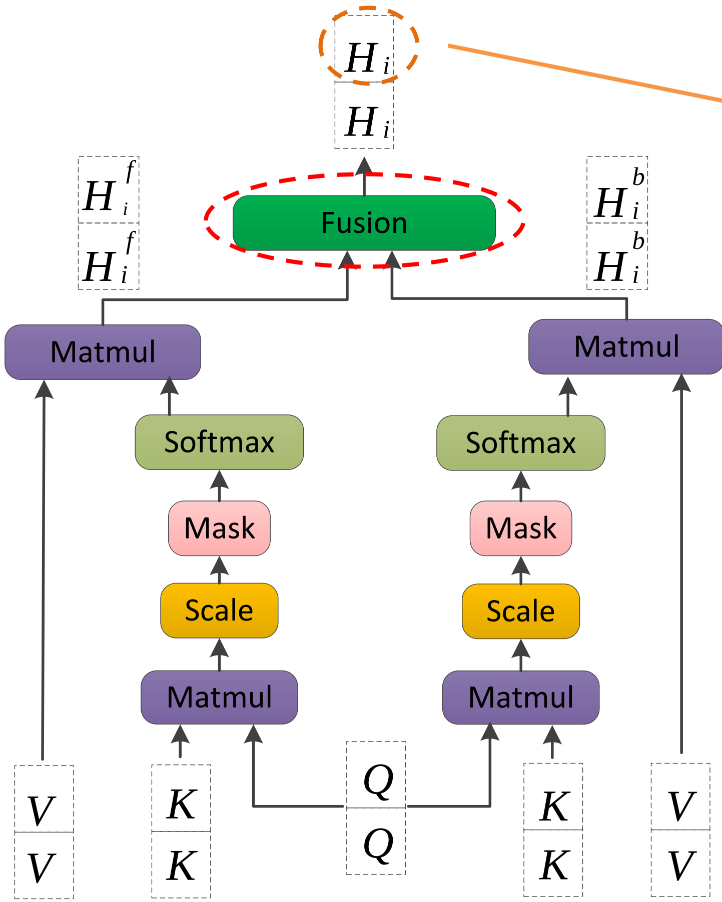


Synchronous Bidirectional Attention



Synchronous Bidirectional Attention

- Synchronous Bidirectional Dot-Product Attention

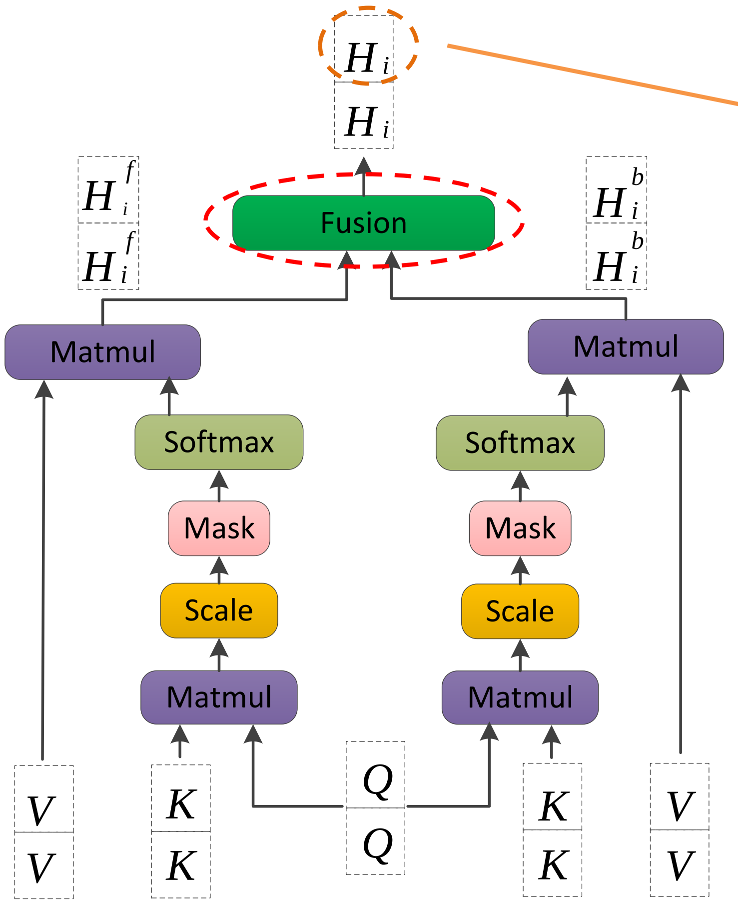


$$\overrightarrow{H}_i = Fusion(\overrightarrow{H}_i^f, \overrightarrow{H}_i^b)$$

- Linear Interpolation
- Nonlinear Interpolation
- Gate Mechanism

Synchronous Bidirectional Attention

- Synchronous Bidirectional Dot-Product Attention



$$\overrightarrow{H}_i = Fusion(\overrightarrow{H}_i^f, \overrightarrow{H}_i^b)$$

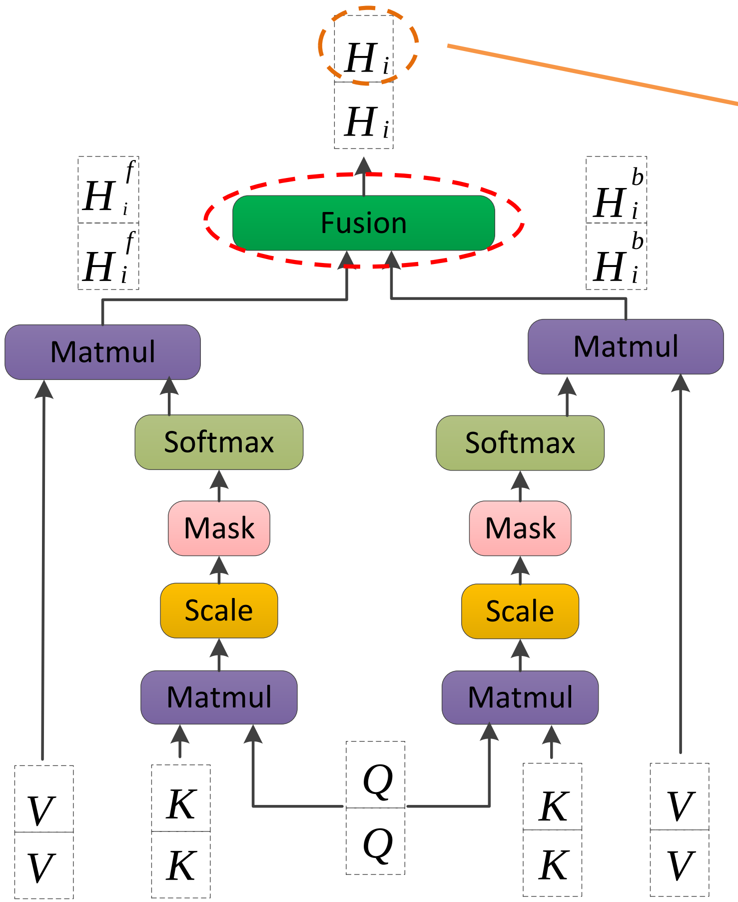
- Linear Interpolation

$$\overrightarrow{H}_i = \overrightarrow{H}_i^f + \lambda * \overrightarrow{H}_i^b$$

- Nonlinear Interpolation
- Gate Mechanism

Synchronous Bidirectional Attention

- Synchronous Bidirectional Dot-Product Attention



$$\overrightarrow{H}_i = Fusion(\overrightarrow{H}_i^f, \overrightarrow{H}_i^b)$$

- Linear Interpolation

$$\overrightarrow{H}_i = \overrightarrow{H}_i^f + \lambda * \overrightarrow{H}_i^b$$

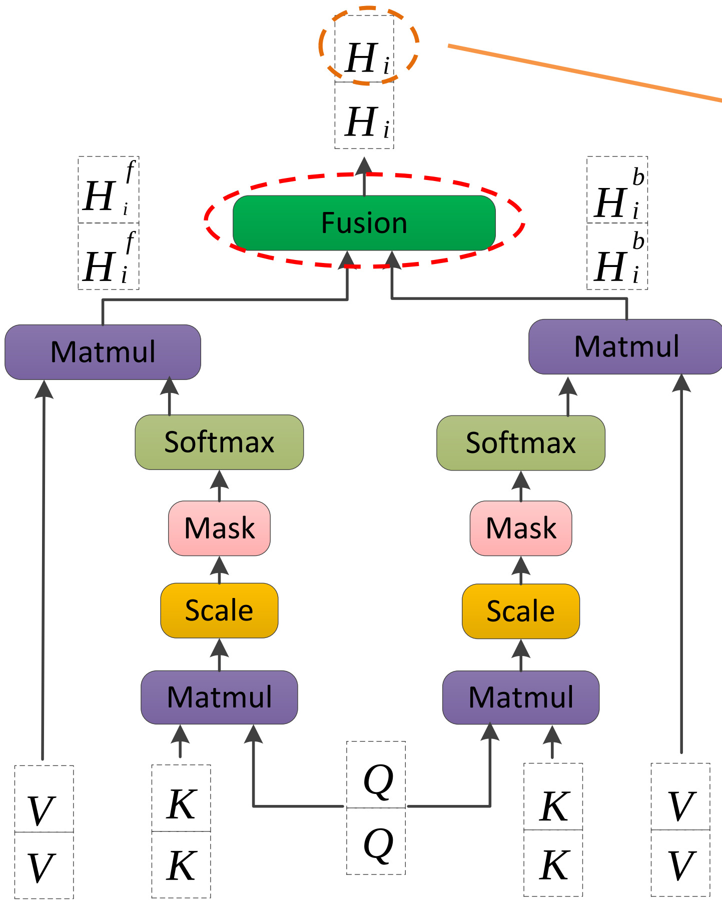
- Nonlinear Interpolation

$$\overrightarrow{H}_i = \overrightarrow{H}_i^f + \lambda * AF(\overrightarrow{H}_i^b) \quad \begin{cases} tanh \\ Relu \end{cases}$$

- Gate Mechanism

Synchronous Bidirectional Attention

- Synchronous Bidirectional Dot-Product Attention



$$\overrightarrow{H}_i = Fusion(\overrightarrow{H}_i^f, \overrightarrow{H}_i^b)$$

- ◆ Linear Interpolation

$$\overrightarrow{H}_i = \overrightarrow{H}_i^f + \lambda * \overrightarrow{H}_i^b$$

- ◆ Nonlinear Interpolation

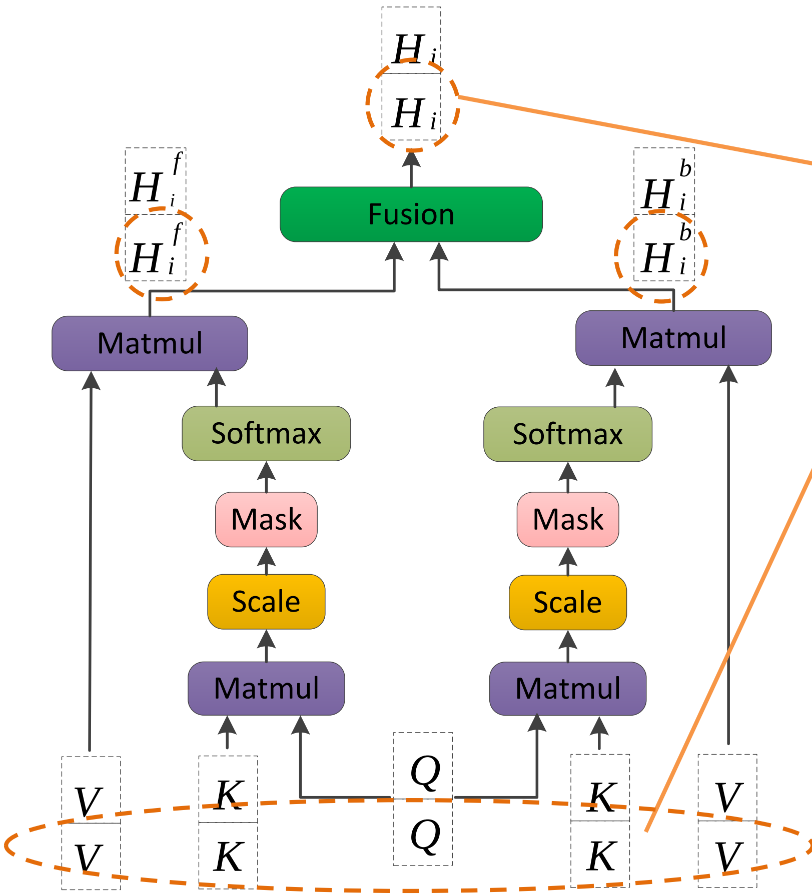
$$\overrightarrow{H}_i = \overrightarrow{H}_i^f + \lambda * AF(\overrightarrow{H}_i^b) \quad \begin{cases} tanh \\ Relu \end{cases}$$

- ◆ Gate Mechanism

$$r_t, z_t = \sigma(W^g[\overrightarrow{H}_i^f; \overrightarrow{H}_i^b])$$
$$\overrightarrow{H}_i = r_t \odot \overrightarrow{H}_i^f + z_t \odot \overrightarrow{H}_i^b$$

SBDPA

- Synchronous Bidirectional Dot-Product Attention



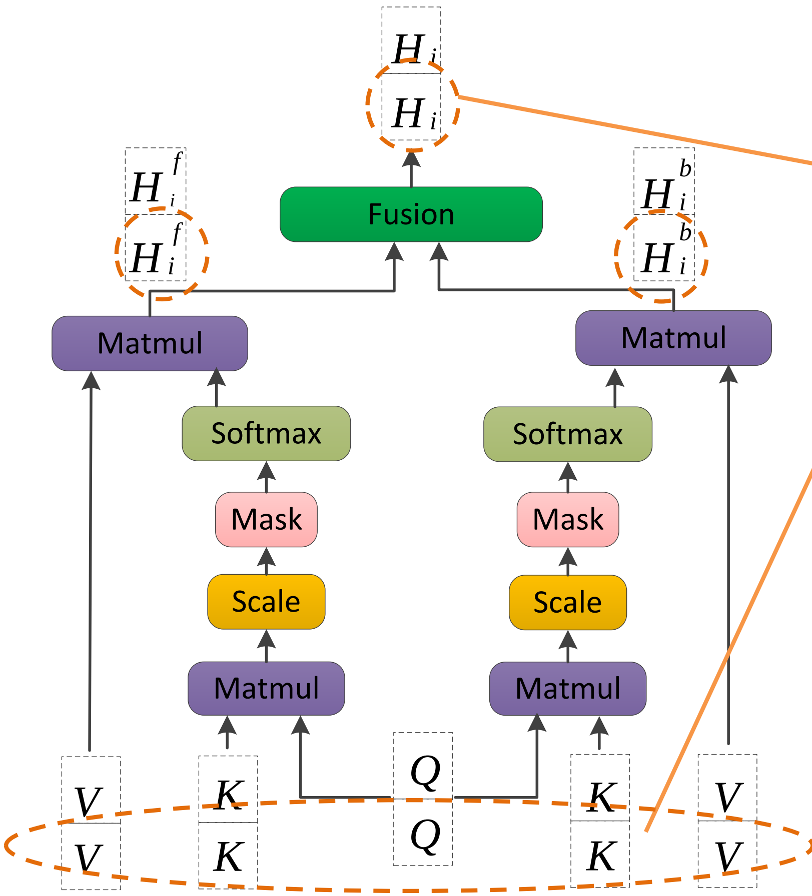
$$\overleftarrow{H}_i^f = Attention(\overleftarrow{Q}, \overleftarrow{K}, \overleftarrow{V})$$

$$\overleftarrow{H}_i^b = Attention(\overleftarrow{Q}, \overrightarrow{K}, \overrightarrow{V})$$

$$\overleftarrow{H}_i = Fusion(\overleftarrow{H}_i^f, \overleftarrow{H}_i^b)$$

SBDPA

- Synchronous Bidirectional Dot-Product Attention



$$\overleftarrow{H}_i^f = Attention(\overleftarrow{Q}, \overleftarrow{K}, \overleftarrow{V})$$

$$\overleftarrow{H}_i^b = Attention(\overleftarrow{Q}, \overrightarrow{K}, \overrightarrow{V})$$

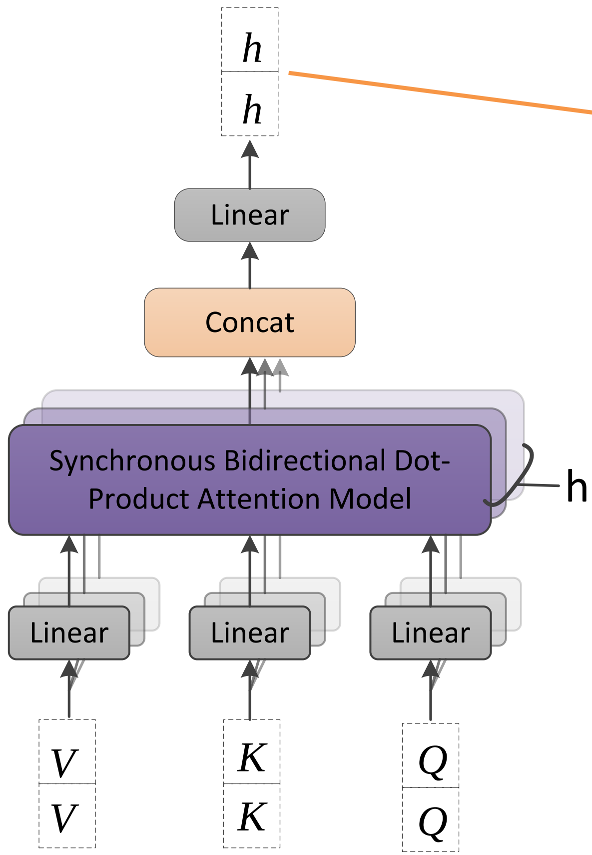
$$\overleftarrow{H}_i = Fusion(\overleftarrow{H}_i^f, \overleftarrow{H}_i^b)$$

We refer to the whole procedure as **SBDPT()**:

$$[\overrightarrow{H}_i, \overleftarrow{H}_i] = \mathrm{SBDPA}([\overleftarrow{Q}; \overrightarrow{Q}], [\overleftarrow{K}; \overrightarrow{K}], [\overleftarrow{V}; \overrightarrow{V}])$$

Synchronous Multi-Head Attention

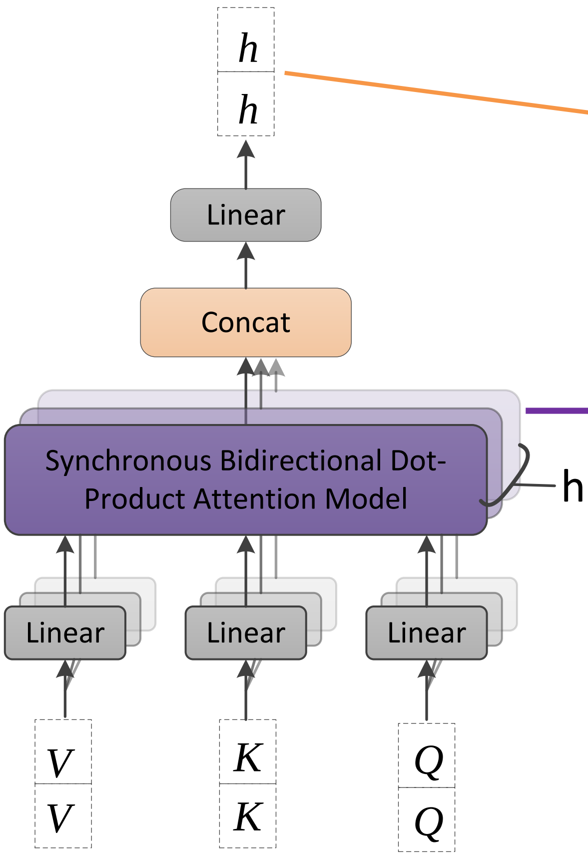
- Synchronous Bidirectional Multi-Head Attention



$$MultiHead([\overleftarrow{Q};\overrightarrow{Q}],[\overleftarrow{K};\overrightarrow{K}],[\overleftarrow{V};\overrightarrow{V}])$$
$$= Concat([\overrightarrow{H}_1;\overleftarrow{H}_1],...,[\overrightarrow{H}_h;\overleftarrow{H}_h])W^O$$

Synchronous Multi-Head Attention

- Synchronous Bidirectional Multi-Head Attention

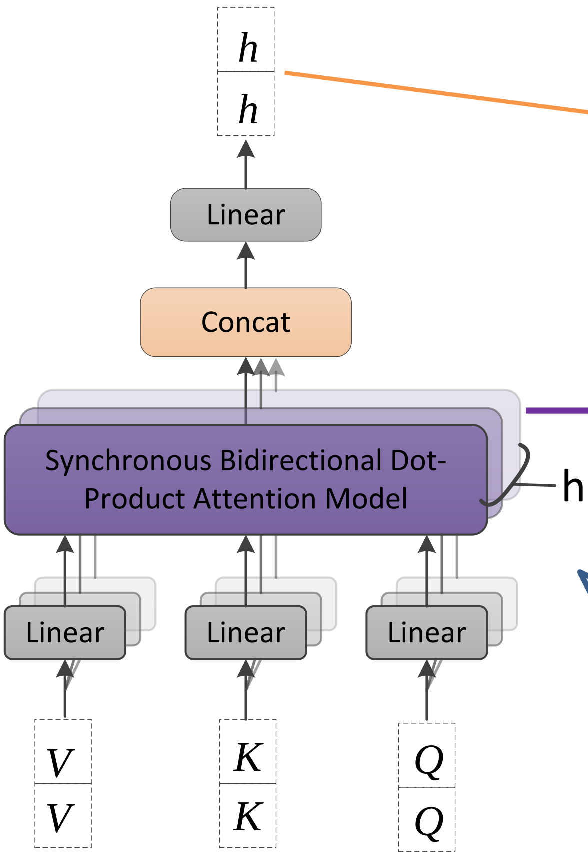


$$MultiHead([\overleftarrow{Q};\overrightarrow{Q}],[\overleftarrow{K};\overrightarrow{K}],[\overleftarrow{V};\overrightarrow{V}])$$
$$= Concat([\overrightarrow{H}_1;\overleftarrow{H}_1],\ldots,[\overrightarrow{H}_h;\overleftarrow{H}_h])W^O$$

$$[\overrightarrow{H}_i;\overleftarrow{H}_i] = \mathrm{SBDPA}([\overleftarrow{Q};\overrightarrow{Q}]\mathrm{W}_\mathrm{i}^\mathrm{Q},$$
$$[\overleftarrow{K};\overrightarrow{K}]W_i^K,[\overleftarrow{V};\overrightarrow{V}]W_i^V)$$

Synchronous Multi-Head Attention

- Synchronous Bidirectional Multi-Head Attention



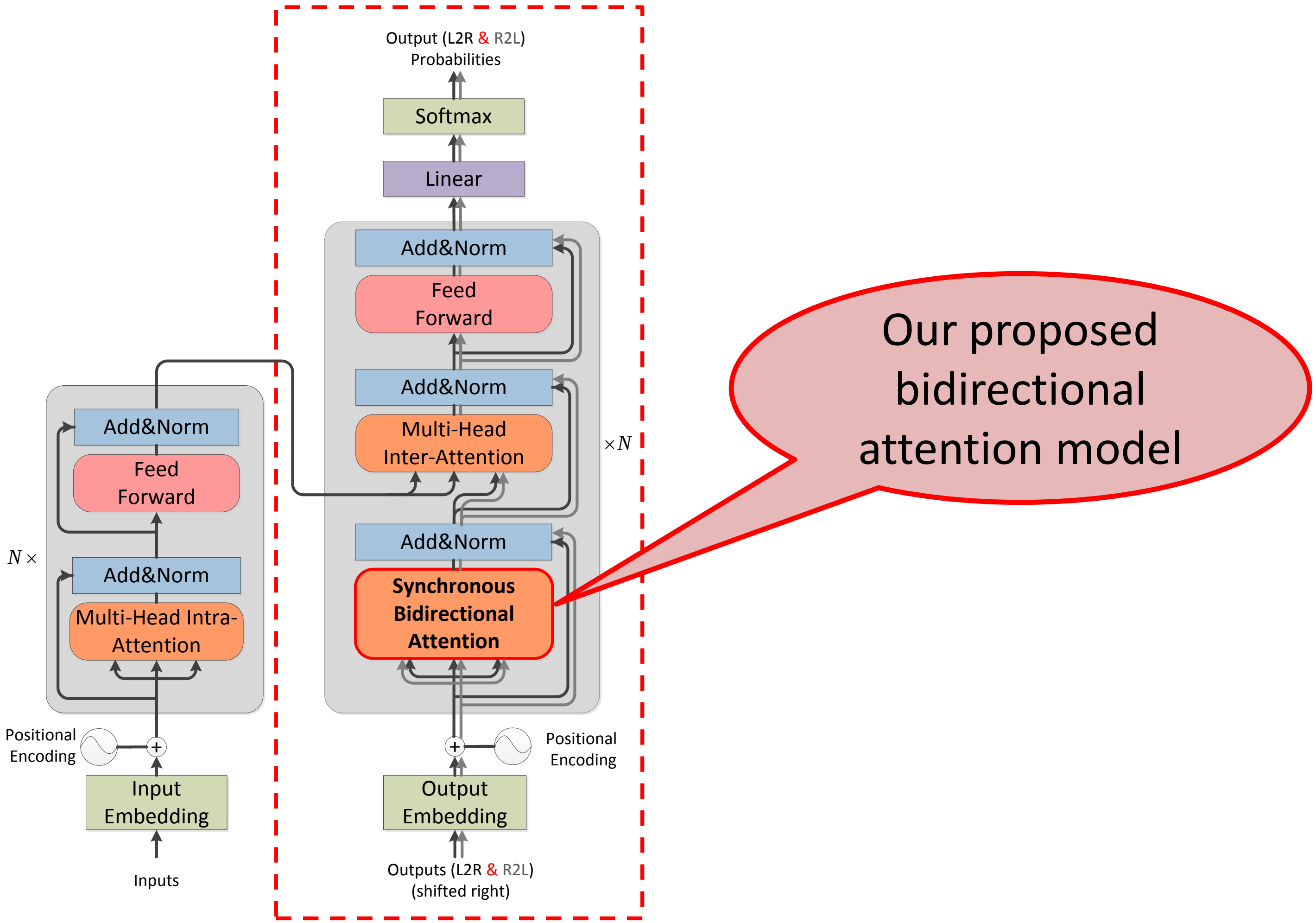
$$
\begin{aligned}
&MultiHead([\overleftarrow{Q};\overrightarrow{Q}],[\overleftarrow{K};\overrightarrow{K}],[\overleftarrow{V};\overrightarrow{V}])\\
&= Concat([\overrightarrow{H}_1;\overleftarrow{H}_1],\ldots,[\overrightarrow{H}_h;\overleftarrow{H}_h])W^O
\end{aligned}
$$

$$
\begin{aligned}
[\overrightarrow{H}_i;\overleftarrow{H}_i] = \mathrm{SBDPA}(&[\overleftarrow{Q};\overrightarrow{Q}]\mathrm{W}_i^Q,\\
&[\overleftarrow{K};\overrightarrow{K}]W_i^K,[\overleftarrow{V};\overrightarrow{V}]W_i^V)
\end{aligned}
$$

Note that all parameters are the same as standard multi-head attention model.

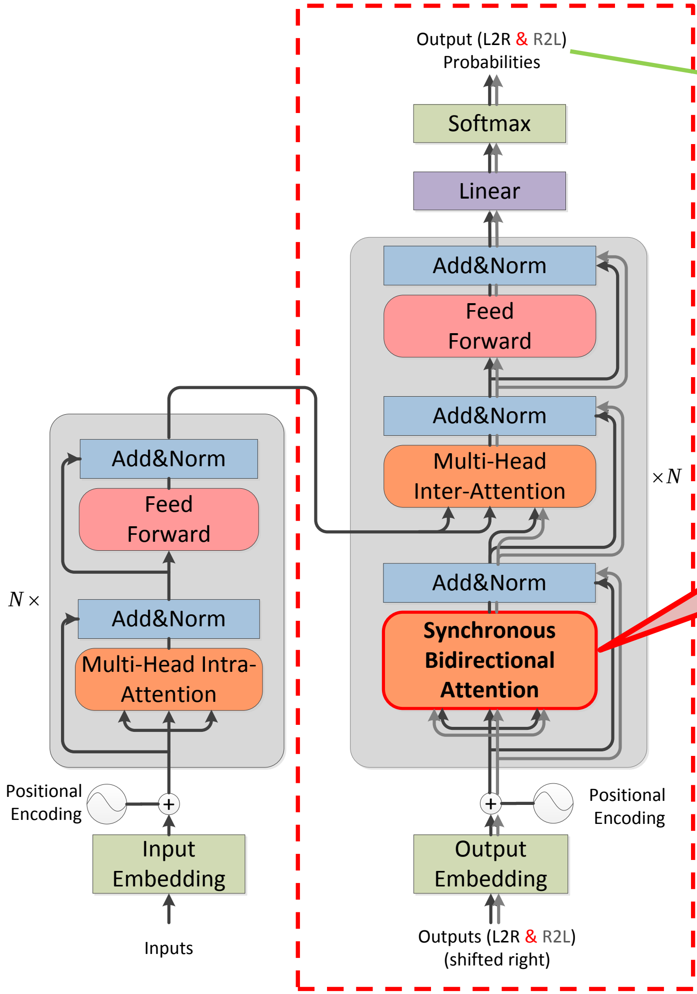
Synchronous Bidirectional Attention

- Integrating Bidirectional Attention into NMT



Synchronous Bidirectional Attention

- Integrating Bidirectional Attention into NMT

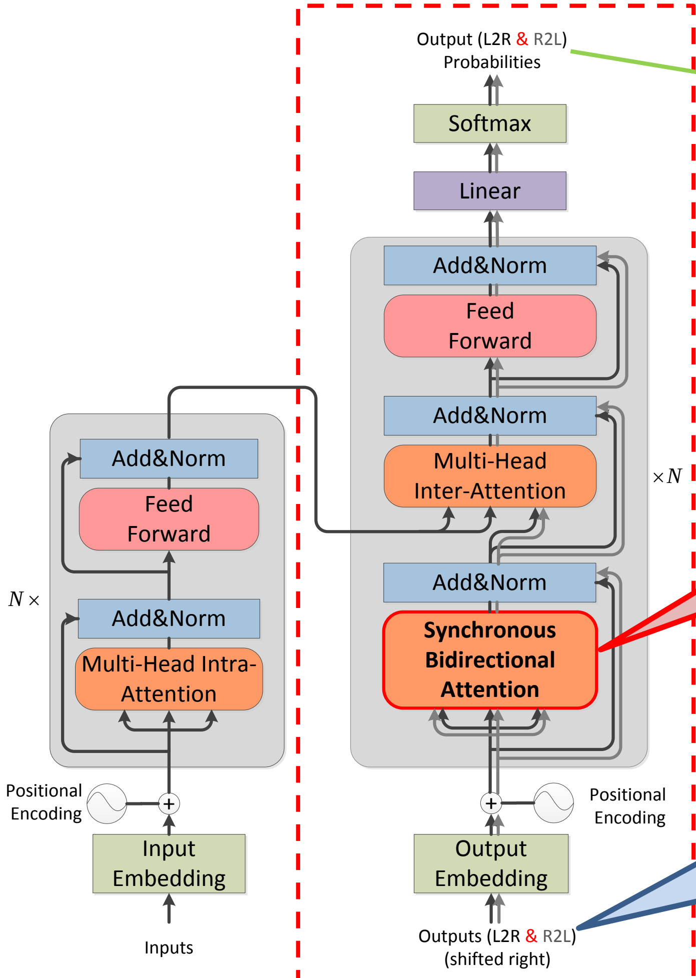


$$p(\overrightarrow{y}_j|\overrightarrow{y}_{<j},\overleftarrow{y}_{<j},x,\theta)=softmax(\overrightarrow{s}^N W)$$
$$p(\overleftarrow{y}_j|\overleftarrow{y}_{<j},\overrightarrow{y}_{<j},x,\theta)=softmax(\overleftarrow{s}^N W)$$

Our proposed
bidirectional
attention model

Synchronous Bidirectional Attention

- Integrating Bidirectional Attention into NMT

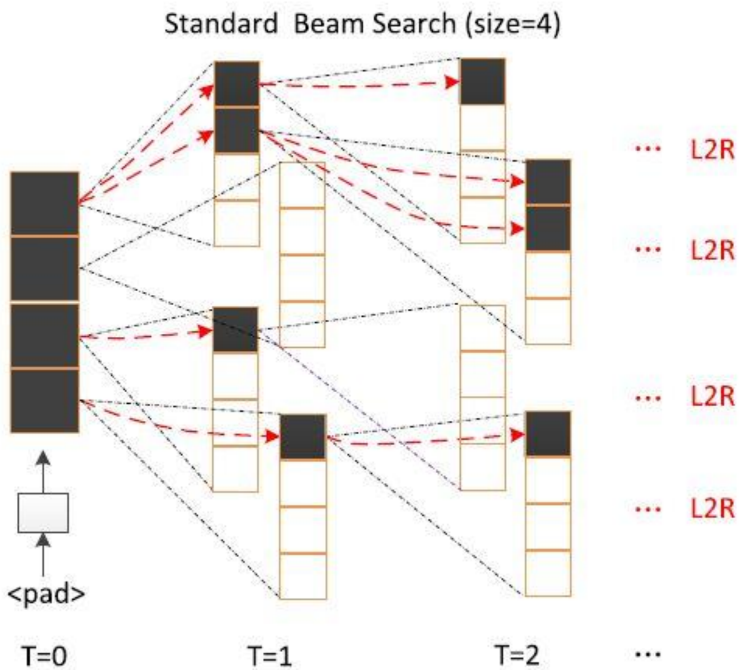


$$p(\overrightarrow{y}_j|\overrightarrow{y}_{<j},\overleftarrow{y}_{<j},x,\theta) = softmax(\overrightarrow{s}^N W)$$
$$p(\overleftarrow{y}_j|\overleftarrow{y}_{<j},\overrightarrow{y}_{<j},x,\theta) = softmax(\overleftarrow{s}^N W)$$

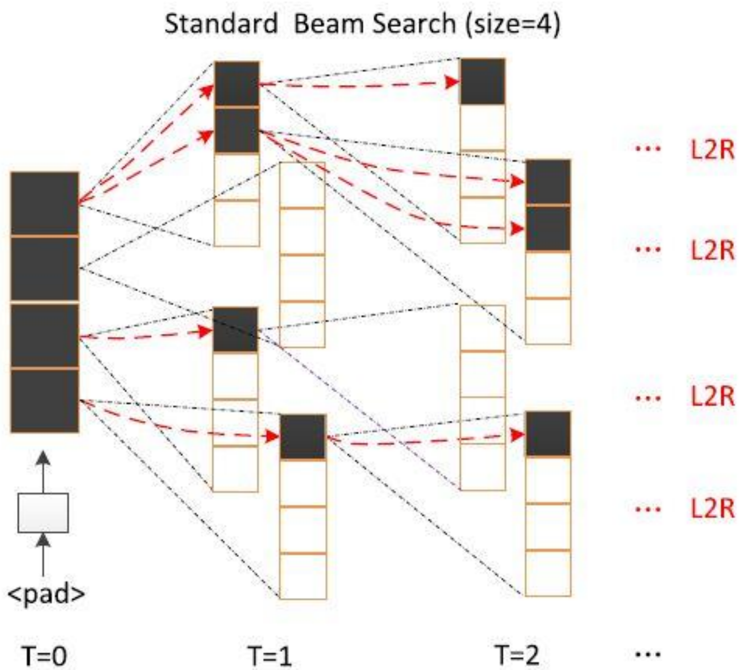
Our proposed bidirectional attention model

Note that all bidirectional information flow in decoder runs in parallel.

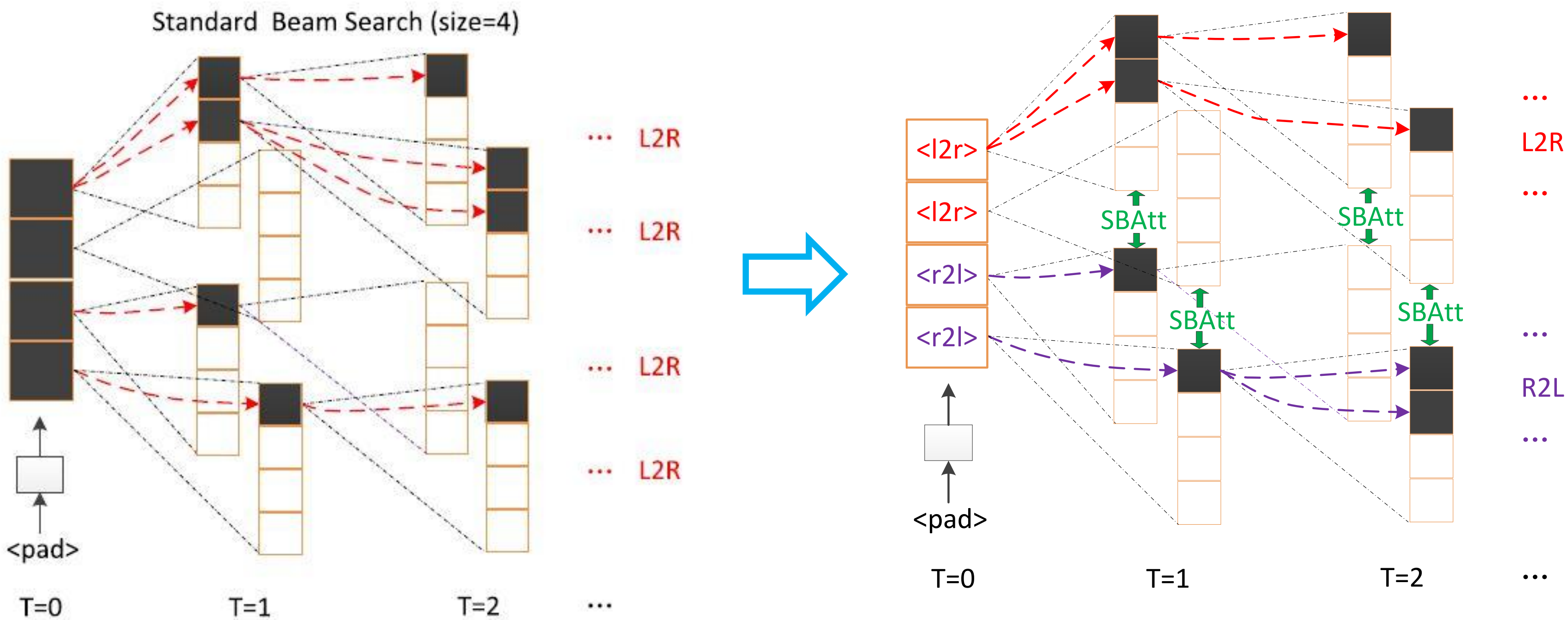
Synchronous Bidirectional Beam Search Algorithm



Synchronous Bidirectional Beam Search Algorithm



Synchronous Bidirectional Beam Search Algorithm



Training

- Training Objective Function

$$J(\theta) = \frac{1}{Z} \sum_{z=1}^{Z} \sum_{j=1}^{M} \{ log\, p(\overrightarrow{y}_j^{(z)} | \overrightarrow{y}_{<j}^{(z)}, \overleftarrow{y}_{<j}^{(z)}, x^{(z)}, \theta)$$

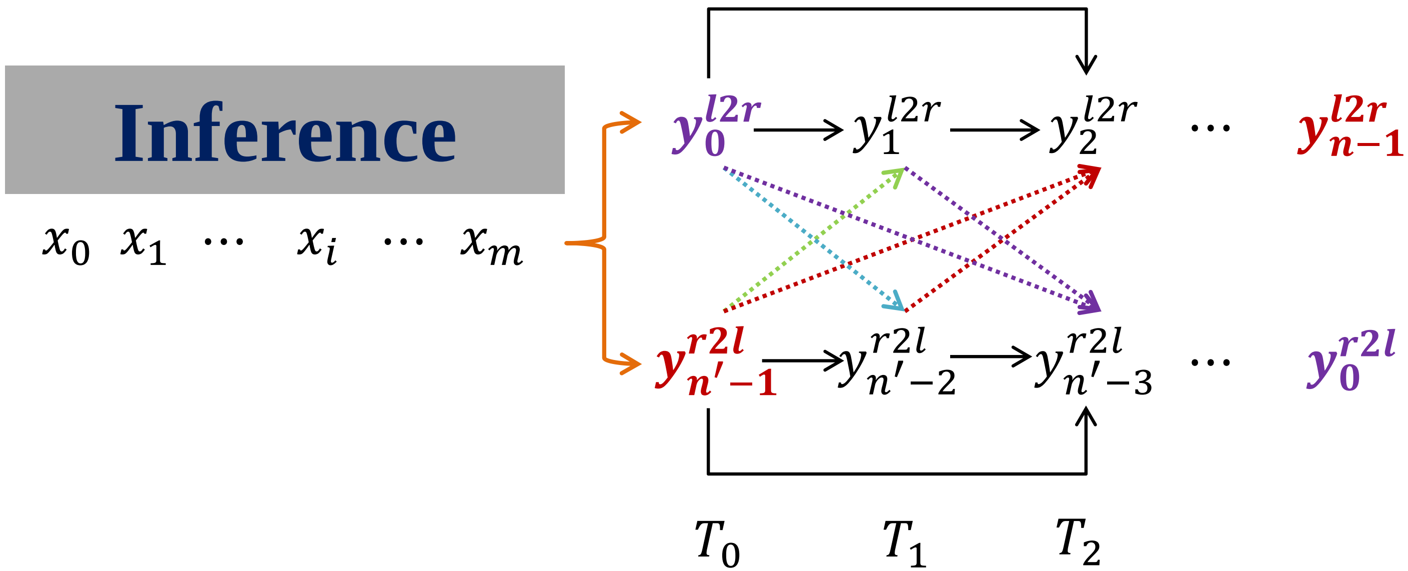
$$+ log\, p(\overleftarrow{y}_j^{(z)} | \overleftarrow{y}_{<j}^{(z)}, \overrightarrow{y}_{<j}^{(z)}, x^{(z)}, \theta) \}$$

Training

Training

Inference

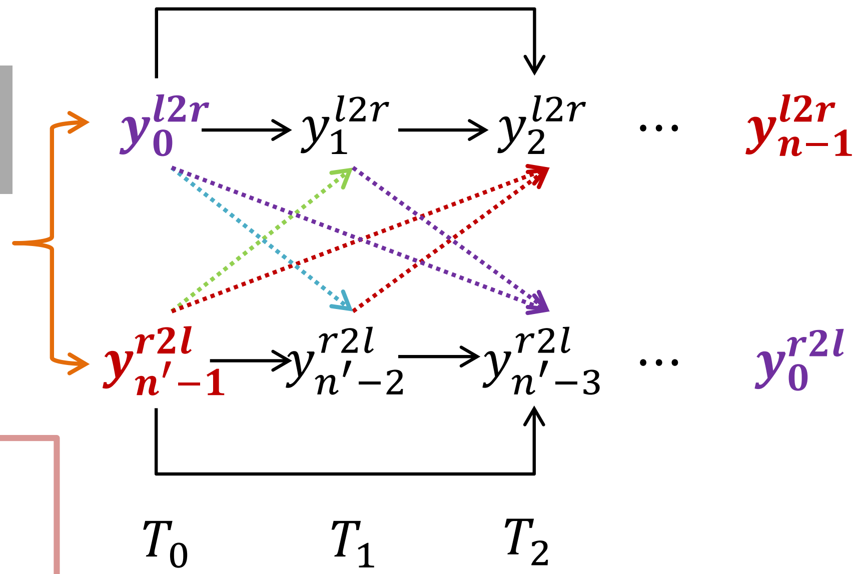
Training


Inference
x_0 x_1 $\cdots$ x_i $\cdots$ x_m
y_0^{l2r} y_1^{l2r} y_2^{l2r} $\cdots$ y_{n-1}^{l2r}
$y_{n'-1}^{r2l}$ $y_{n'-2}^{r2l}$ $y_{n'-3}^{r2l}$ $\cdots$ y_0^{r2l}
T_0 T_1 T_2

Training

Inference

$x_0 \;\; x_1 \;\; \cdots \;\; x_i \;\; \cdots \;\; x_m$

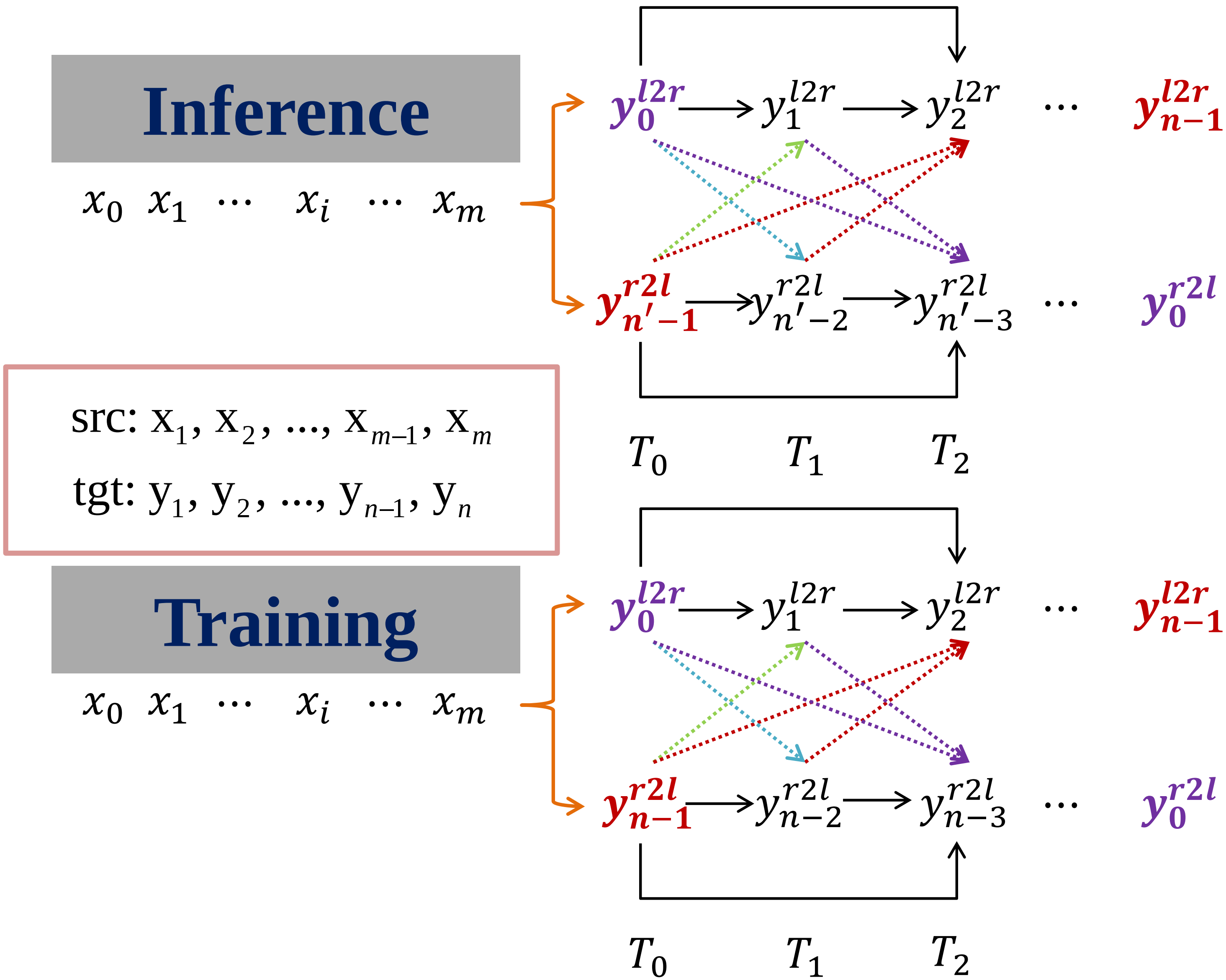


src: $x_1, x_2, \ldots, x_{m-1}, x_m$

tgt: $y_1, y_2, \ldots, y_{n-1}, y_n$

Training

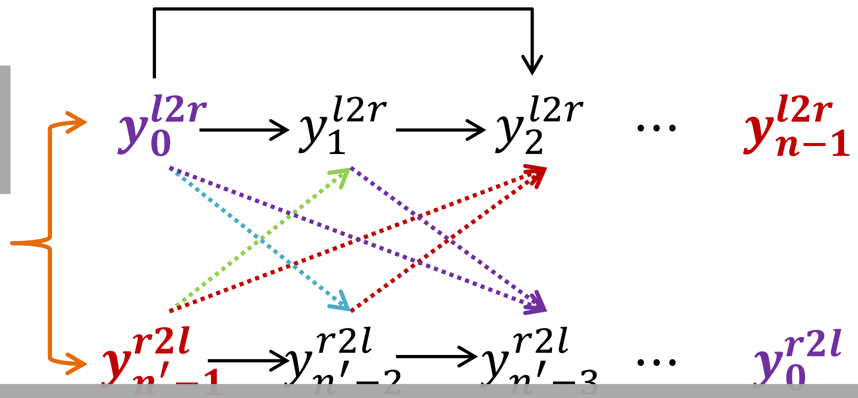
Training


Inference
$x_0 \quad x_1 \quad \cdots \quad x_i \quad \cdots \quad x_m$
$y_0^{l2r} \quad y_1^{l2r} \quad y_2^{l2r} \quad \cdots \quad y_{n-1}^{l2r}$
$y_{n'-1}^{r2l} \quad y_{n'-2}^{r2l} \quad y_{n'-3}^{r2l} \quad \cdots \quad y_0^{r2l}$
$T_0 \quad T_1 \quad T_2$
src: $x_1, x_2, \ldots, x_{m-1}, x_m$
tgt: $y_1, y_2, \ldots, y_{n-1}, y_n$
Training
$x_0 \quad x_1 \quad \cdots \quad x_i \quad \cdots \quad x_m$
$y_0^{l2r} \quad y_1^{l2r} \quad y_2^{l2r} \quad \cdots \quad y_{n-1}^{l2r}$
$y_{n-1}^{r2l} \quad y_{n-2}^{r2l} \quad y_{n-3}^{r2l} \quad \cdots \quad y_0^{r2l}$
$T_0 \quad T_1 \quad T_2$

Training

Inference

$x_0 \; x_1 \; \cdots \; x_i \; \cdots \; x_m$



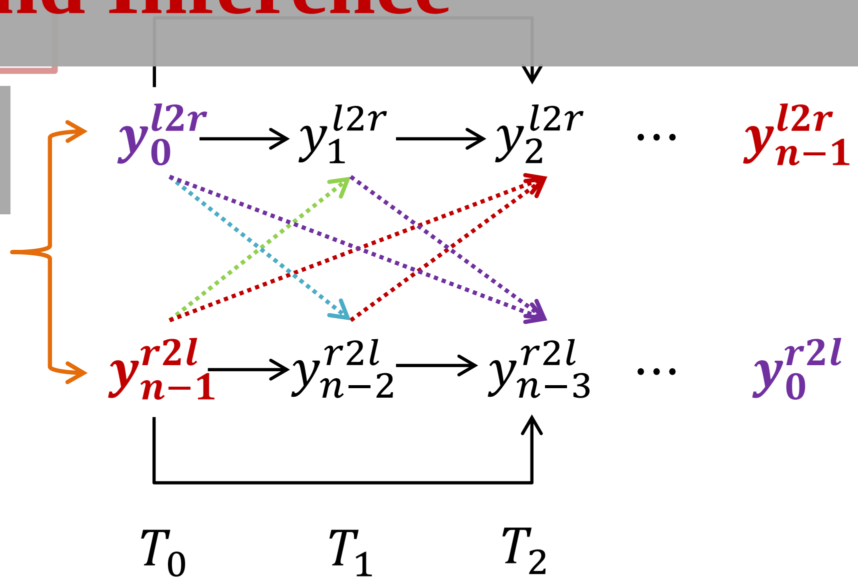
src: $x_1, x_2, \ldots, x_{m-1}, x_m$

tgt: $y_1, y_2, \ldots, y_{n-1}, y_n$

Question: Mismatch between Training and Inference

Training

$x_0 \; x_1 \; \cdots \; x_i \; \cdots \; x_m$



Training Strategy 1

$$J(\theta) = \sum_{t=1}^{T} \left\{ logp\left(\vec{y}^{(t)} | x^{(t)}\right) + logp\left(\overleftarrow{y}^{(t)} | x^{(t)}\right) \right\}$$

- Two-pass method
  - **First-pass**: training $L2R$ and $R2L$, models. Using $L2R$ and $R2L$ to decode the source inputs of bitext, resulting $\left(x^{(t)}, \overrightarrow{y^*}^{(t)}\right)_{t=1}^{T}$ and $\left(x^{(t)}, \overleftarrow{y^*}^{(t)}\right)_{t=1}^{T}$;
  - **Second-pass:** using $\overleftarrow{y^*}_{<i}^{(t)}$ instead of $\overleftarrow{y}_{<i}^{(t)}$ to compute $p\left(\vec{y}_i^{(t)} | \vec{y}_{<i}^{(t)}, x^{(t)}, \overleftarrow{y^*}_{<i}^{(t)}\right)$, similar for $p\left(\overleftarrow{y}_i^{(t)} | \overleftarrow{y}_{<i}^{(t)}, x^{(t)}, \overrightarrow{y^*}_{<i}^{(t)}\right)$.

Training Strategy 1

$$J(\theta) = \sum_{t=1}^{T} \left\{ logp(\vec{y}^{(t)} | x^{(t)}) + logp\left(\overleftarrow{y}^{(t)} | x^{(t)}\right) \right\}$$

- Two-pass method
  - **First-pass:** training $L2R$ and $R2L$, models. Using $L2R$ and $R2L$ to decode the source inputs of bitext, resulting $\left(x^{(t)}, \overrightarrow{y^*}^{(t)}\right)_{t=1}^{T}$ and $\left(x^{(t)}, \overleftarrow{y^*}^{(t)}\right)_{t=1}^{T}$;

Problem: Too Time Consuming

  - **Second-pass:** using $\overleftarrow{y^*}^{(t)}_{<i}$ instead of $\overleftarrow{y}^{(t)}_{<i}$ to compute $p\left(\vec{y}_i^{(t)} | \vec{y}_{<i}^{(t)}, x^{(t)}, \overleftarrow{y^*}^{(t)}_{<i}\right)$, similar for $p\left(\overleftarrow{y}_i^{(t)} | \overleftarrow{y}_{<i}^{(t)}, x^{(t)}, \overrightarrow{y^*}^{(t)}_{<i}\right)$.

Training Strategy 2

$$P(y|x) = \begin{cases} \sum_{i=0}^{n-1} p(\vec{y}_i|\vec{y}_0 \cdots \vec{y}_{i-1}, x) & if\ L2R \\ \sum_{i=0}^{n'-1} p(\overleftarrow{y}_i|\overleftarrow{y}_0 \cdots \overleftarrow{y}_{i-1}, x) & if\ R2L \end{cases}$$

- Fine-tuning method
  - **Bidirectional Inference without Interaction**: training $SBNMT$, model with no interaction. The learned $SBNMT$ perfroms $L2R$ and $R2L$ decoding for the source inputs of bitext, resulting $\left(x^{(t)}, \overrightarrow{y^*}^{(t)}\right)_{t=1}^{T}$ and $\left(x^{(t)}, \overleftarrow{y^*}^{(t)}\right)_{t=1}^{T}$;
  - **Fine-tuning with Interaction:** using $\overleftarrow{y^*}_{<i}^{(t)}$ instead of $\overleftarrow{y}_{<i}^{(t)}$ to compute $p\left(\vec{y}_i^{(t)}|\vec{y}_{<i}^{(t)}, x^{(t)}, \overleftarrow{y^*}_{<i}^{(t)}\right)$.

Experiments: Machine Translation

- Setup
  - Dataset:
    (1) NIST Chinese-English translation (2M, 30K tokens, MT03-06 as test set)
    (2) WMT14 English-German translation (4.5M, 37K shared tokens, newstest2014 as test set)
  - Train details:
    (1) *Transformer_big* setting
    (2) Chinese-English: 1 GPUs, single model, case-insensitive BLEU.
    (3) English-German: 3 GPUs, model averaging, case-sensitive BLEU.

Experiments: Machine Translation

- Baselines
  - **Moses:** an Open source phrase-based SMT system.
  - **RNMT:** RNN-based NMT with default setting.
  - **Transformer:** Predict target sentence from left to right.
  - **Transformer(R2L):** Predict sentence from right to left.
  - **Rerank-NMT:** (1) first run beam search to obtain two k-best lists; (2) then re-score and get the best candidate.
  - **ABD-NMT:** (1) use backward decoder to generate reverse sequence states; (2) perform beam search on the forward decoder to find the best translation.

Experiments: Machine Translation

- Results on Chinese-English Translation
  - Translation Quality

Model	DEV	MT03	MT04	M05	MT06	AVE	Δ
Moses	37.85	37.47	41.20	36.41	36.03	37.78	-9.41
RNMT	42.43	42.43	44.56	41.94	40.95	42.47	-4.72
Transformer	48.12	47.63	48.32	47.51	45.31	47.19	-
Transformer(R2L)	47.81	46.79	47.01	46.50	44.13	46.11	-1.08
Rerank-NMT	49.18	48.23	48.91	48.73	46.51	48.10	+0.91
ABD-NMT	48.28	49.47	48.01	48.19	47.09	48.19	+1.00
Our Model	**50.99**	**51.61**	**51.41**	**51.19**	**49.84**	**51.01**	**+3.82**

Table: Evaluation of translation quality for Chinese-English translation tasks with case-insensitive BLEU scores.

Experiments: Machine Translation

- Results on Chinese-English Translation
  - Translation Quality

Model	DEV	MT03	MT04	M05	MT06	AVE	Δ
Moses	37.85	37.47	41.20	36.41	36.03	37.78	-9.41
RNMT	42.43	42.43	44.56	41.94	40.95	42.47	-4.72
Transformer	48.12	47.63	48.32	47.51	45.31	47.19	-
Transformer(R2L)	47.81	46.79	47.01	46.50	44.13	46.11	-1.08
Rerank-NMT	49.18	48.23	48.91	48.73	46.51	48.10	+0.91
ABD-NMT	48.28	49.47	48.01	48.19	47.09	48.19	+1.00
Our Model	**50.99**	**51.61**	**51.41**	**51.19**	**49.84**	**51.01**	**+3.82**

Table: Evaluation of translation quality for Chinese-English translation tasks with case-insensitive BLEU scores.

Experiments: Machine Translation

- Results on Chinese-English Translation
  - Translation Quality

Model	DEV	MT03	MT04	M05	MT06	AVE	Δ
Moses	37.85	37.47	41.20	36.41	36.03	37.78	-9.41
RNMT	42.43	42.43	44.56	41.94	40.95	42.47	-4.72
Transformer	48.12	47.63	48.32	47.51	45.31	47.19	-
Transformer(R2L)	47.81	46.79	47.01	46.50	44.13	46.11	-1.08
Rerank-NMT	49.18	48.23	48.91	48.73	46.51	48.10	+0.91
ABD-NMT	48.28	49.47	48.01	48.19	47.09	48.19	+1.00
Our Model	**50.99**	**51.61**	**51.41**	**51.19**	**49.84**	**51.01**	**+3.82**

Table: Evaluation of translation quality for Chinese-English translation tasks with case-insensitive BLEU scores.

Experiments: Machine Translation

- Results on Chinese-English Translation
  - Translation Quality

Model	DEV	MT03	MT04	M05	MT06	AVE	Δ
Moses	37.85	37.47	41.20	36.41	36.03	37.78	-9.41
RNMT	42.43	42.43	44.56	41.94	40.95	42.47	-4.72
Transformer	48.12	47.63	48.32	47.51	45.31	47.19	-
Transformer(R2L)	47.81	46.79	47.01	46.50	44.13	46.11	-1.08
Rerank-NMT	49.18	48.23	48.91	48.73	46.51	48.10	+0.91
ABD-NMT	48.28	49.47	48.01	48.19	47.09	48.19	+1.00
Our Model	**50.99**	**51.61**	**51.41**	**51.19**	**49.84**	**51.01**	**+3.82**

Table: Evaluation of translation quality for Chinese-English translation tasks with case-insensitive BLEU scores.

Experiments: Machine Translation

- Results on English-German Translation

Model	Test
GNMT (Wu et al., 2016)	24.61
Conv (Gehring et al., 2017)	25.16
AttIsAll (Vaswani et al., 2017)	28.40
Transformer	27.72
Transformer(R2L)	27.13
Rerank-NMT	27.81
ABD-NMT	28.22
Our Model	**29.21**

Table: Results of English-German translation using case-sensitive BLEU.

Experiments: Machine Translation

- Results on English-German Translation

Model	Test
GNMT (Wu et al., 2016)	24.61
Conv (Gehring et al., 2017)	25.16
AttIsAll (Vaswani et al., 2017)	28.40
Transformer	27.72
Transformer(R2L)	27.13
Rerank-NMT	27.81
ABD-NMT	28.22
Our Model	**29.21**

Strong Baselines

Table: Results of English-German translation using case-sensitive BLEU.

Experiments: Machine Translation

- Results on English-German Translation

Model	Test
GNMT (Wu et al., 2016)	24.61
Conv (Gehring et al., 2017)	25.16
AttIsAll (Vaswani et al., 2017)	28.40
Transformer	27.72
Transformer(R2L)	27.13
Rerank-NMT	27.81
ABD-NMT	28.22
Our Model	**29.21**

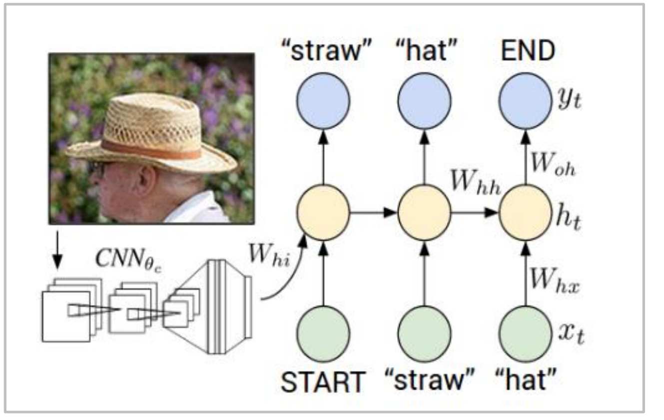
Strong Baselines

(+1.49)

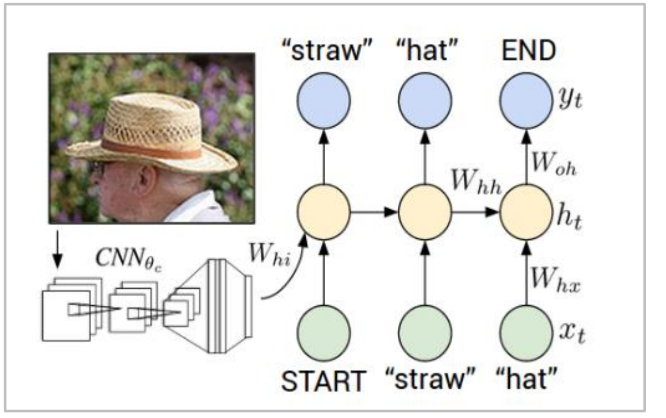
Table: Results of English-German translation using case-sensitive BLEU.

Experiments: Image Caption

Experiments: Image Caption



Experiments: Image Caption



- Setup
  - Dataset:
    (1) Flickr30k (Young et al., 2014)
    (2) 29,000 image-caption for training
    (3) 1014 for validation and 2000 for test
  - Baselines:
    (1) VGGNet encoder + LSMT decoder (Xu et al., 2015)
    (2) Transformer

Experiments: Image Caption

- Results on English Image Caption

  - BLEU score

Method	Validation	Test
Xu et al., (2015)	~	19.90
Transformer	22.11	21.25
Ours	**23.27**	**22.41**

MT Analysis: Unbalanced Outputs

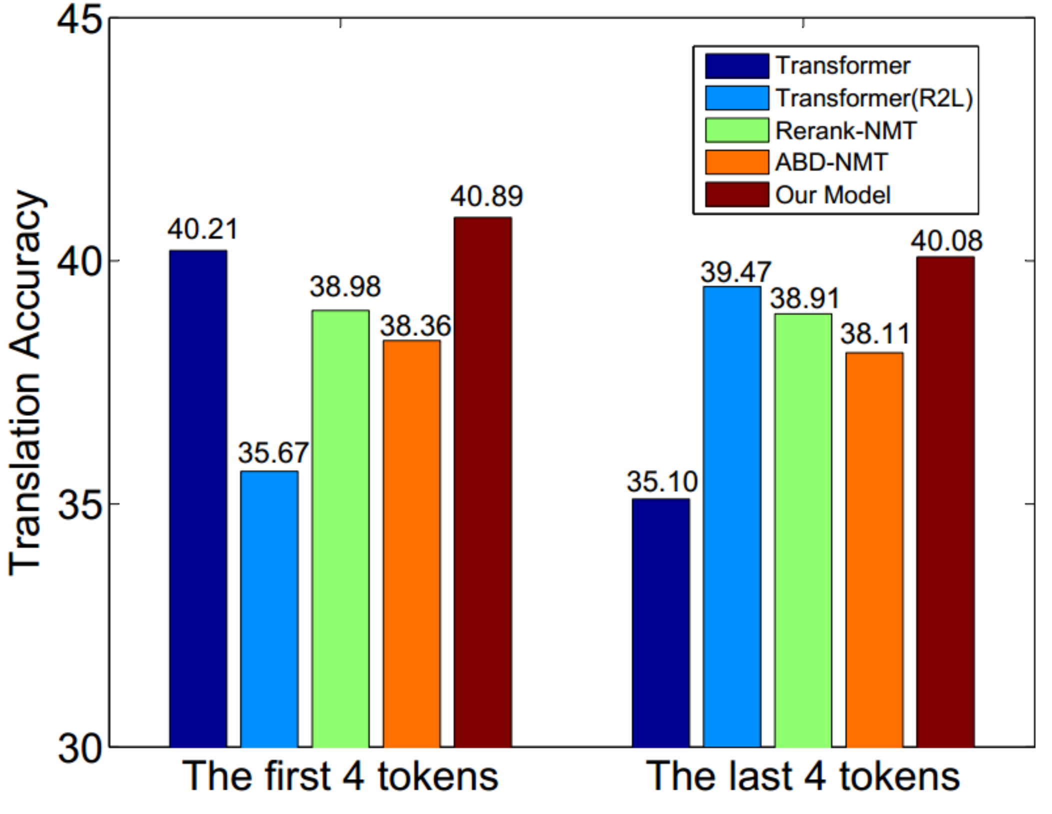
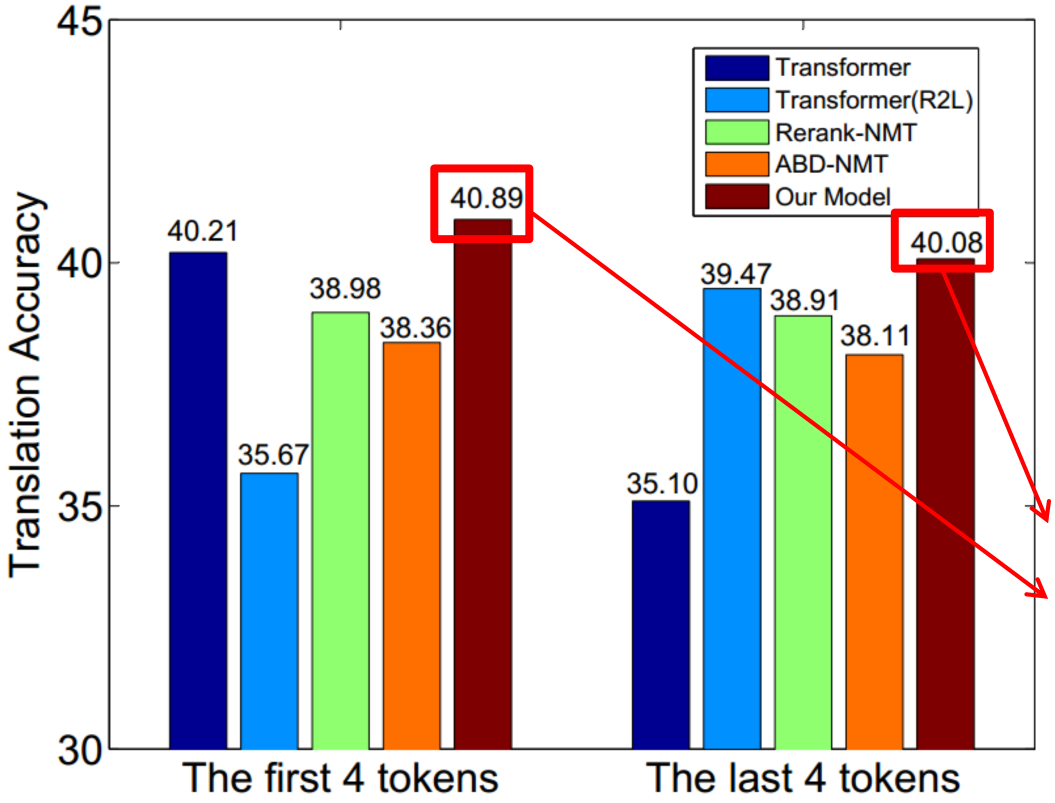


Figure: Translation accuracy of the first 4 tokens and last 4 tokens for L2R, R2L, Rerank-NMT, ABD-NMT and our proposed model.

MT Analysis: Unbalanced Outputs



Get the best translation accuracy no matter for the first 4 words and the last 4 words.

Figure: Translation accuracy of the first 4 tokens and last 4 tokens for L2R, R2L, Rerank-NMT, ABD-NMT and our proposed model.

MT Analysis: BLEU along Length

- Analysis
  - Effect of Long Sentence

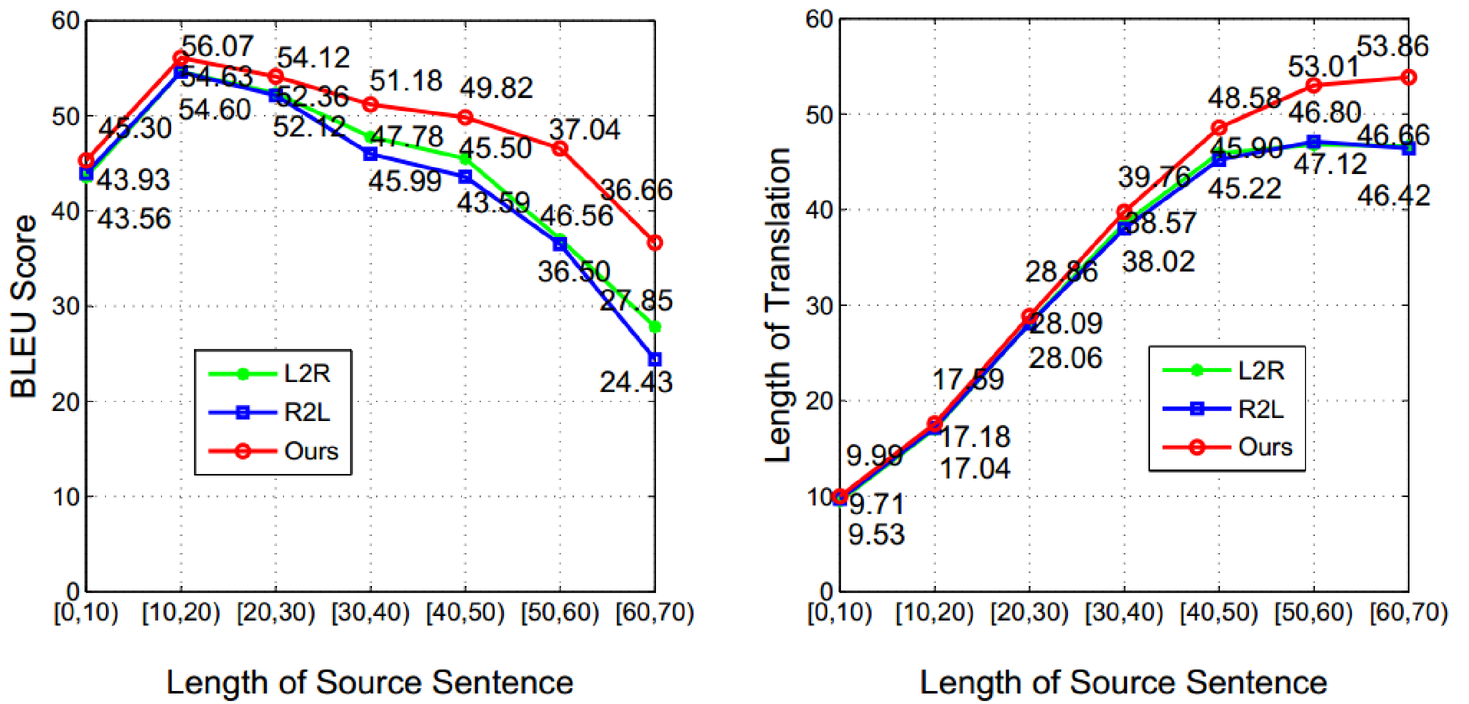


Figure: Performance of translations on the test set with respect to the lengths of the source sentences.

MT Analysis: BLEU along Length

- Analysis
  - Effect of Long Sentence

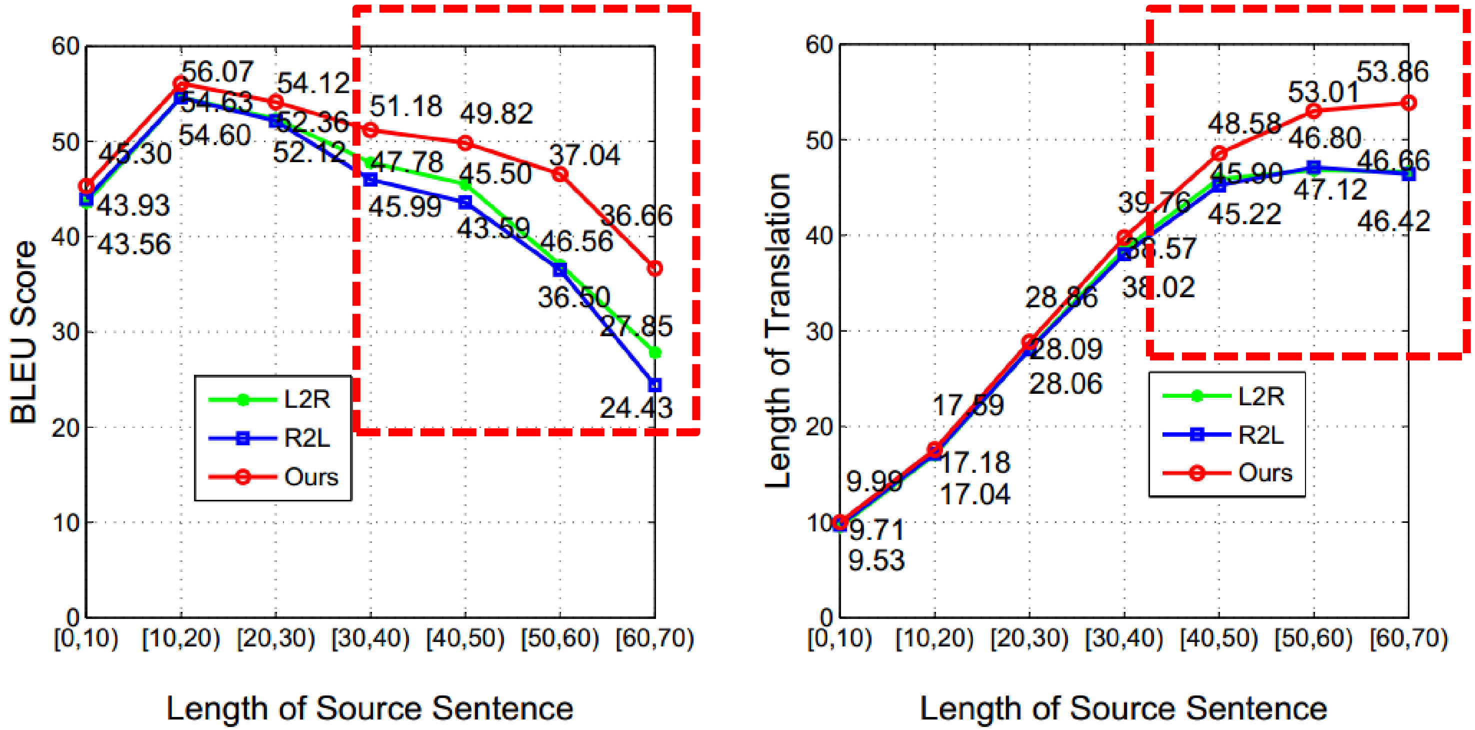


Figure: Performance of translations on the test set with respect to the lengths of the source sentences.

MT Analysis: Case Study

Source	捷克 总统 哈维 卸任 新 总统 仍 未 确定
Reference	czech president havel steps down while new president still not chosen
L2R	czech president leaves office
R2L	the outgoing president of the czech republic is still uncertain
Ours	czech president havel leaves office , new president yet to be determined

Source	他们 正在 研制 一 种 超大型 的 叫做 炸弹 之 母 。
Reference	they are developing a kind of superhuge bomb called the mother of bombs .
L2R	they are developing a super , big , mother , called the bomb .
R2L	they are working on a much larger mother called the mother of a bomb .
Ours	they are developing a super-large scale , called the mother of the bomb .

MT Analysis: Case Study

Source	捷克 总统 哈维 卸任 新 总统 仍 未 确定
Reference	czech president havel steps down while new president still not chosen
L2R	czech president leaves office
R2L	the outgoing president of the czech republic is still uncertain
Ours	czech president havel leaves office , new president yet to be determined

Source	他们 正在 研制 一 种 超大型 的 叫做 炸弹 之 母 。
Refer	
L	
R2	
Ou	

L2R produces good prefix, whereas R2L generates better suffixes.

Our approach can make full use of bidirectional decoding and produce balanced outputs in these cases.

MT Analysis: Parameters and Speeds

MT Analysis: Parameters and Speeds

Model	Param	Speed	
		Train	*Test*
Transformer	207.8M	2.07	19.97
Transformer(R2L)	207.8M	2.07	19.81
Rerank-NMT	415.6M	1.03	6.51
ABD-NMT	333.8M	1.18	7.20
Our Model	207.8M	1.26	17.87

Table: Statistics of parameters, training and testing speeds. Train denotes the number of global training steps processed per second; Test indicates the amount of translated sentences in one second.

MT Analysis: Parameters and Speeds

No additional parameters except for lambda

Model	Param	Speed *Train*	Speed *Test*
Transformer	207.8M	2.07	19.97
Transformer(R2L)	207.8M	2.07	19.81
Rerank-NMT	415.6M	1.03	6.51
ABD-NMT	333.8M	1.18	7.20
Our Model	207.8M	1.26	17.87

Table: Statistics of parameters, training and testing speeds. Train denotes the number of global training steps processed per second; Test indicates the amount of translated sentences in one second.

MT Analysis: Parameters and Speeds

No additional parameters except for lambda

Model	Param	Speed	
		Train	*Test*
Transformer	207.8M	2.07	19.97
Transformer(R2L)	207.8M	2.07	19.81
Rerank-NMT	415.6M	1.03	6.51
ABD-NMT	333.8M	1.18	7.20
Our Model	207.8M	1.26	17.87

Table: Statistics of parameters, training and testing speeds. Train denotes the number of global training steps processed per second; Test indicates the amount of translated sentences in one second.

MT Analysis: Parameters and Speeds

No additional parameters except for lambda

Slightly Slower than baseline Transformer

Model	Param	Speed	
		Train	*Test*
Transformer	207.8M	2.07	19.97
Transformer(R2L)	207.8M	2.07	19.81
Rerank-NMT	415.6M	1.03	6.51
ABD-NMT	333.8M	1.18	7.20
Our Model	207.8M	1.26	17.87

Table: Statistics of parameters, training and testing speeds. Train denotes the number of global training steps processed per second; Test indicates the amount of translated sentences in one second.

Beyond Synchronous Bidirectional Decoding: Improving Efficiency

Beyond Synchronous Bidirectional Decoding: Improving Efficiency

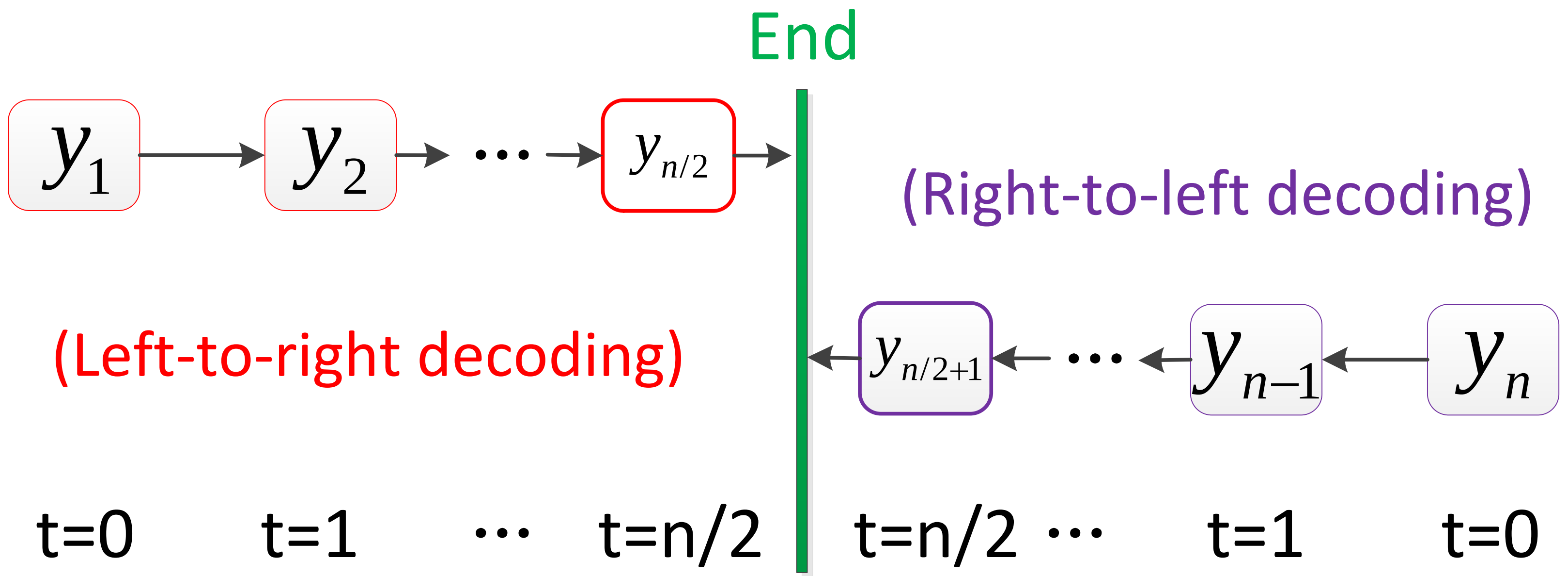
Sequence Generation: from Both Sides to the Middle

Long Zhou, Jiajun Zhang, Chengqing Zong and Heng Yu.

In Proceedings of IJCAI 2019.

Sequence Generation from Both Sides to the Middle

- **SBSG:** Synchronous Bidirectional Sequence Generation
  - Speedup decoding: Generates two tokens at a time
  - Improve quality: Rely on history and future context



Sequence Generation from Both Sides to the Middle

- Training and Inference
  - Following previous work, we also use knowledge distillation techniques to train our model.
  - Training objective:

$$J(\theta) = \frac{1}{Z}\sum_{z=1}^{Z}\sum_{j=1}^{n/2}\{log\, p(\overrightarrow{y}_j^{(z)}|\overrightarrow{y}_{<j}^{(z)}, \overleftarrow{y}_{<j}^{(z)}, x^{(z)}, \theta)$$
$$+ log\, p(\overleftarrow{y}_j^{(z)}|\overleftarrow{y}_{<j}^{(z)}, \overrightarrow{y}_{<j}^{(z)}, x^{(z)}, \theta)\}$$

  - The Smoothing model:

y :	<l2r>	y_1	y_2	$\cdots$	$y_{n/2}$	$y_{n/2+1}$	<eos>
y :	<r2l>	y_n	y_{n-1}	$\cdots$	$y_{n/2+2}$	<null>	<eos>

Experiments on Machine Translation

- Inference speed: ⬆
- Translation quality: ⬆

System	Architecture	English-German Quality	English-German Speed	Chinese-English Quality	Chinese-English Speed	English-Romanian Quality	English-Romanian Speed
Existing NMT systems							
[Gu *et al.*, 2017]	NAT	17.35	N/A	-	-	26.22	15.6×
	NAT (s=100)	19.17	N/A	-	-	29.79	2.36×
[Lee *et al.*, 2018]	D-NAT	12.65	11.71×	-	-	24.45	16.03×
	D-NAT (adaptive)	18.91	1.98×	-	-	29.66	5.23×
[Kaiser *et al.*, 2018]	LT	19.80	3.89×	-	-	-	-
	LT (s=100)	22.50	N/A	-	-	-	-
[Wang *et al.*, 2018] (beam search)	SAT (K=2)	26.90	1.51×	39.57	1.69×	-	-
	SAT (K=6)	24.83	2.98×	35.32	3.18×	-	-
[Wang *et al.*, 2018] (greedy search)	SAT (K=2)	26.09	1.70×	38.37	1.71×	-	-
	SAT (K=6)	23.93	4.57×	33.75	4.70×	-	-
Our NMT systems							
This work (beam search)	Transformer	27.06	1.00×	46.56	1.00×	32.28	1.00×
	Transformer (R2L)	26.71	1.02×	44.63	0.94×	32.29	0.98×
	Our Model	**27.45**	1.38×	**47.82**	1.41×	**33.02**	1.43×
This work (greedy search)	Transformer	26.23	1.00×	44.63	1.00×	31.71	1.00×
	Transformer (R2L)	25.38	0.97×	43.68	0.98×	31.19	1.04×
	Our Model	**27.22**	1.61×	**47.50**	1.51×	**32.82**	1.46×

Experiments on Machine Translation

- Inference speed: ⬆
- Translation quality: ⬆

System	Architecture	English-German		Chinese-English		English-Romanian	
		Quality	Speed	Quality	Speed	Quality	Speed
Existing NMT systems							
[Gu *et al.*, 2017]	NAT	17.35	N/A	-	-	26.22	15.6×
	NAT (s=100)	19.17	N/A	-	-	29.79	2.36×
[Lee *et al.*, 2018]	D-NAT	12.65	11.71×	-	-	24.45	16.03×
	D-NAT (adaptive)	18.91	1.98×	-	-	29.66	5.23×
[Kaiser *et al.*, 2018]	LT	19.80	3.89×	-	-	-	-
	LT (s=100)	22.50	N/A	-	-	-	-
[Wang *et al.*, 2018] (beam search)	SAT (K=2)	26.90	1.51×	39.57	1.69×	-	-
	SAT (K=6)	24.83	2.98×	35.32	3.18×	-	-
[Wang *et al.*, 2018] (greedy search)	SAT (K=2)	26.09	1.70×	38.37	1.71×	-	-
	SAT (K=6)	23.93	4.57×	33.75	4.70×	-	-
Our NMT systems							
This work (beam search)	Transformer	27.06	1.00×	46.56	1.00×	32.28	1.00×
	Transformer (R2L)	26.71	1.02×	44.63	0.94×	32.29	0.98×
	Our Model	**27.45**	1.38×	**47.82**	1.41×	**33.02**	1.43×
This work (greedy search)	Transformer	26.23	1.00×	44.63	1.00×	31.71	1.00×
	Transformer (R2L)	25.38	0.97×	43.68	0.98×	31.19	1.04×
	Our Model	**27.22**	1.61×	**47.50**	1.51×	**32.82**	1.46×

Experiments on Machine Translation

- Inference speed: ⬆
- Translation quality: ⬆

System	Architecture	English-German		Chinese-English		English-Romanian	
		Quality	Speed	Quality	Speed	Quality	Speed
Existing NMT systems							
[Gu *et al.*, 2017]	NAT	17.35	N/A	-	-	26.22	15.6×
	NAT (s=100)	19.17	N/A	-	-	29.79	2.36×
[Lee *et al.*, 2018]	D-NAT	12.65	11.71×	-	-	24.45	16.03×
	D-NAT (adaptive)	18.91	1.98×	-	-	29.66	5.23×
[Kaiser *et al.*, 2018]	LT	19.80	3.89×	-	-	-	-
	LT (s=100)	22.50	N/A	-	-	-	-
[Wang *et al.*, 2018] (beam search)	SAT (K=2)	26.90	1.51×	39.57	1.69×	-	-
	SAT (K=6)	24.83	2.98×	35.32	3.18×	-	-
[Wang *et al.*, 2018] (greedy search)	SAT (K=2)	26.09	1.70×	38.37	1.71×	-	-
	SAT (K=6)	23.93	4.57×	33.75	4.70×	-	-
Our NMT systems							
This work (beam search)	Transformer	27.06	1.00×	46.56	1.00×	32.28	1.00×
	Transformer (R2L)	26.71	1.02×	44.63	0.94×	32.29	0.98×
	Our Model	**27.45**	1.38×	**47.82**	1.41×	**33.02**	1.43×
This work (greedy search)	Transformer	26.23	1.00×	44.63	1.00×	31.71	1.00×
	Transformer (R2L)	25.38	0.97×	43.68	0.98×	31.19	1.04×
	Our Model	**27.22**	1.61×	**47.50**	1.51×	**32.82**	1.46×

Experiments on Machine Translation

- Inference speed: ⬆
- Translation quality: ⬆

System	Architecture	English-German Quality	English-German Speed	Chinese-English Quality	Chinese-English Speed	English-Romanian Quality	English-Romanian Speed
Existing NMT systems							
[Gu *et al.*, 2017]	NAT	17.35	N/A	-	-	26.22	15.6×
	NAT (s=100)	19.17	N/A	-	-	29.79	2.36×
[Lee *et al.*, 2018]	D-NAT	12.65	11.71×	-	-	24.45	16.03×
	D-NAT (adaptive)	18.91	1.98×	-	-	29.66	5.23×
[Kaiser *et al.*, 2018]	LT	19.80	3.89×	-	-	-	-
	LT (s=100)	22.50	N/A	-	-	-	-
[Wang *et al.*, 2018] (beam search)	SAT (K=2)	26.90	1.51×	39.57	1.69×	-	-
	SAT (K=6)	24.83	2.98×	35.32	3.18×	-	-
[Wang *et al.*, 2018] (greedy search)	SAT (K=2)	26.09	1.70×	38.37	1.71×	-	-
	SAT (K=6)	23.93	4.57×	33.75	4.70×	-	-
Our NMT systems							
This work (beam search)	Transformer	27.06	1.00×	46.56	1.00×	32.28	1.00×
	Transformer (R2L)	26.71	1.02×	44.63	0.94×	32.29	0.98×
	Our Model	**27.45**	1.38×	**47.82**	1.41×	**33.02**	1.43×
This work (greedy search)	Transformer	26.23	1.00×	44.63	1.00×	31.71	1.00×
	Transformer (R2L)	25.38	0.97×	43.68	0.98×	31.19	1.04×
	Our Model	**27.22**	1.61×	**47.50**	1.51×	**32.82**	1.46×

Experiments on Text Summarization

- Application to Text Summarization
  - Example:

    the **sri lankan** government on wednesday announced the **closure** of **government schools** with **immediate effect** as a **military campaign** against **tamil separatists escalated** in the north of the country .

    sri lanka closes schools as war escalates

  - Setup

    (1) English Gigaword dataset (3.8M training set, 189K dev set, DUC2004 as our test set)

    (2) shared vocabulary of about 90K word types

    (3) *Transformer_base* setting

    (4) ROUGE-1, ROUGE-2, ROUGE-L

Experiments on Text Summarization

DUC2004	RG-1	RG-2	RG-L	Speed
ABS‡	26.55	7.06	22.05	-
Feats2s‡	28.35	9.46	24.59	-
Selective-Enc‡	29.21	9.56	25.51	-
Transformer	28.09	9.52	24.91	1.00×
SBSG (beam)	28.77	10.11	26.11	1.48×
SBSG (greedy)	28.70	9.88	25.93	2.09×

Experiments on Text Summarization

DUC2004	RG-1	RG-2	RG-L	Speed
ABS‡	26.55	7.06	22.05	-
Feats2s‡	28.35	9.46	24.59	-
Selective-Enc‡	29.21	9.56	25.51	-
Transformer	28.09	9.52	24.91	1.00×
SBSG (beam)	28.77	10.11	26.11	1.48×
SBSG (greedy)	28.70	9.88	25.93	2.09×

Experiments on Text Summarization

DUC2004	RG-1	RG-2	RG-L	Speed
ABS‡	26.55	7.06	22.05	-
Feats2s‡	28.35	9.46	24.59	-
Selective-Enc‡	29.21	9.56	25.51	-
Transformer	28.09	9.52	24.91	1.00×
SBSG (beam)	28.77	10.11	26.11	1.48×
SBSG (greedy)	28.70	9.88	25.93	2.09×

The proposed model significant outperforms the conventional Transformer model in terms of both decoding speed and generation quality.

Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Two Tasks
- Summary and Future Challenges

Interactive Inference for Two Tasks

Interactive Inference for Two Tasks

Synchronously Generating Two Languages with Interactive Decoding

Yining Wang, Jiajun Zhang, Long Zhou, Yuchen Liu and Chengqing Zong.

In Proceedings of EMNLP 2019.

From Generating Two Directions to Generating Two Languages

$y_0^{l2r} \longrightarrow y_1^{l2r} \longrightarrow \cdots \quad y_i^{l2r} \quad \cdots \quad \cdots$

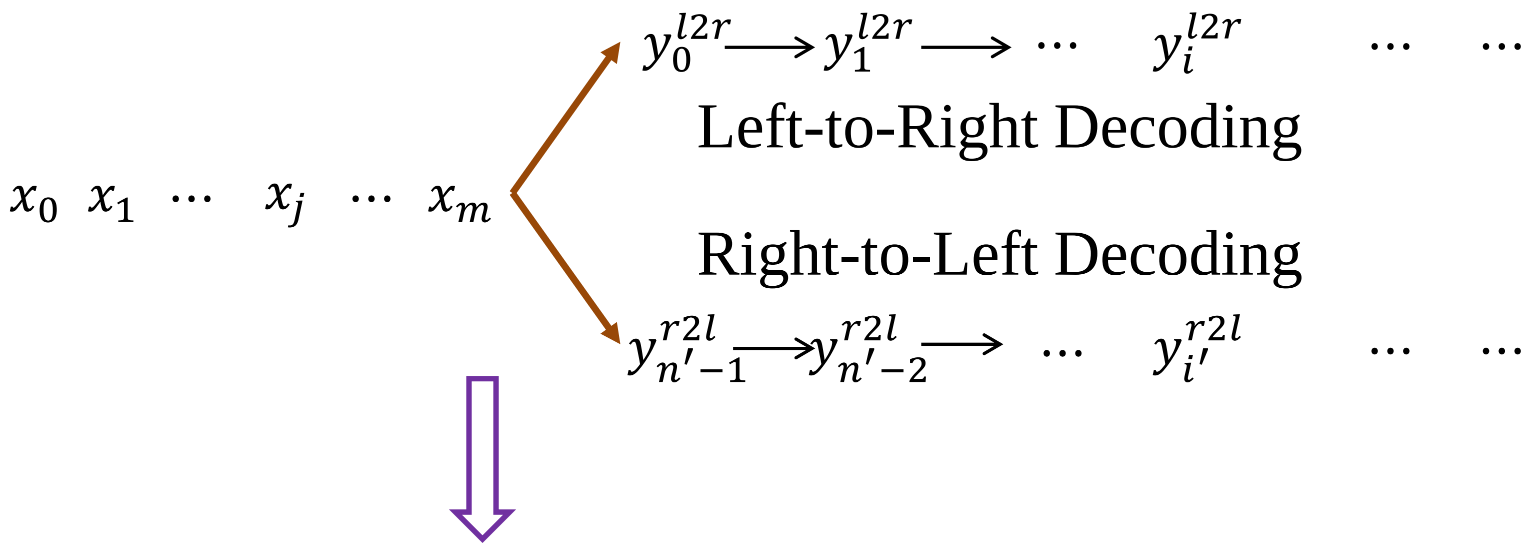
Left-to-Right Decoding

$x_0 \; x_1 \; \cdots \; x_j \; \cdots \; x_m$

Right-to-Left Decoding

$y_{n'-1}^{r2l} \longrightarrow y_{n'-2}^{r2l} \longrightarrow \cdots \quad y_{i'}^{r2l} \quad \cdots \quad \cdots$

From Generating Two Directions to Generating Two Languages



From Generating Two Directions to Generating Two Languages

$x_0 \ x_1 \ \cdots \ x_j \ \cdots \ x_m$

$y_0^{l2r} \longrightarrow y_1^{l2r} \longrightarrow \cdots \quad y_i^{l2r} \quad \cdots \quad \cdots$

Left-to-Right Decoding

Right-to-Left Decoding

$y_{n'-1}^{r2l} \longrightarrow y_{n'-2}^{r2l} \longrightarrow \cdots \quad y_{i'}^{r2l} \quad \cdots \quad \cdots$

$\Downarrow$

$x_0 \ x_1 \ \cdots \ x_j \ \cdots \ x_m$

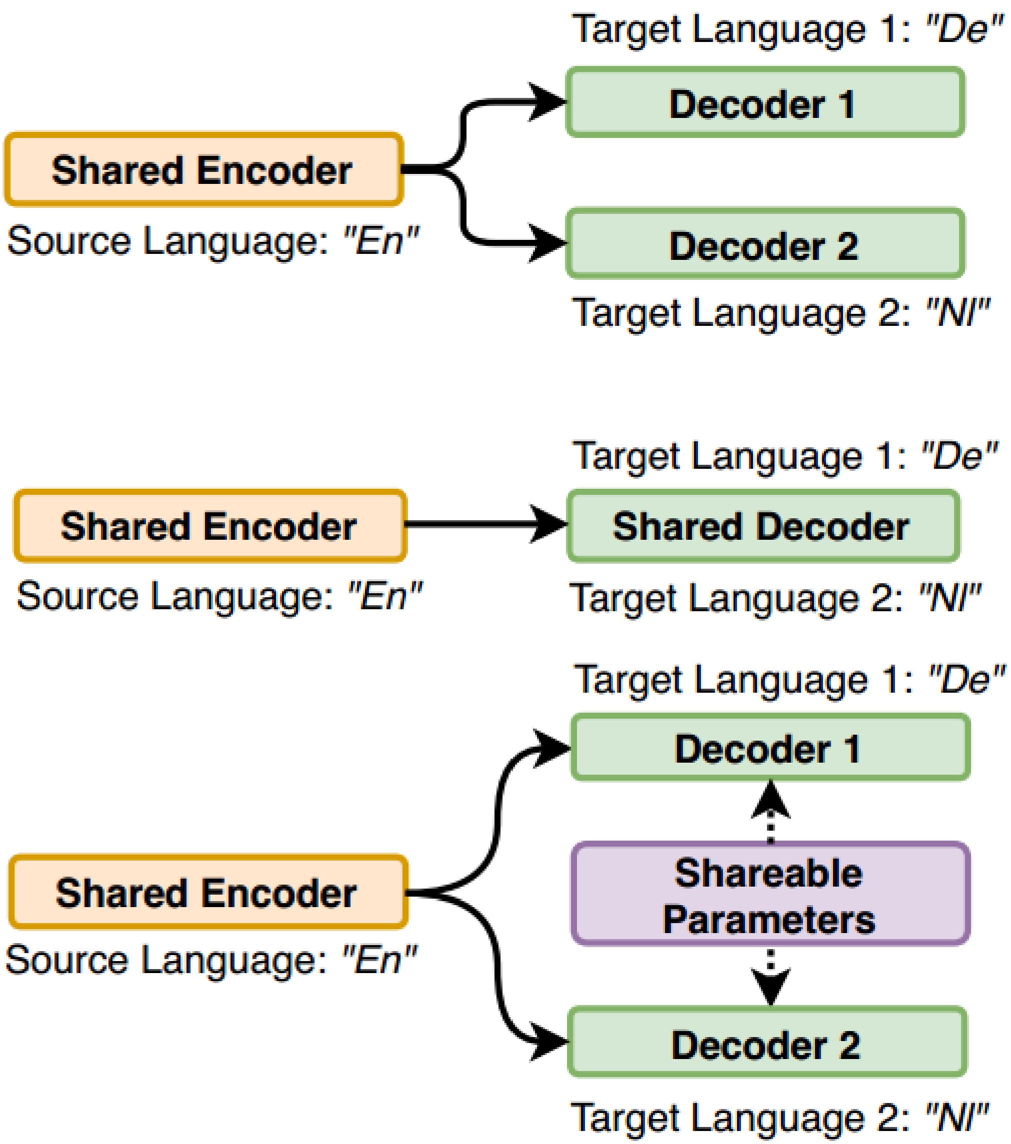
$y_0 \longrightarrow y_1 \longrightarrow \cdots \quad y_i \quad \cdots \quad \cdots$

English-to-Chinese Decoding

English-to-Japanese Decoding

$z_0 \longrightarrow z_1 \longrightarrow \cdots \quad z_j \quad \cdots \quad \cdots$

Conventional Multilingual Translation

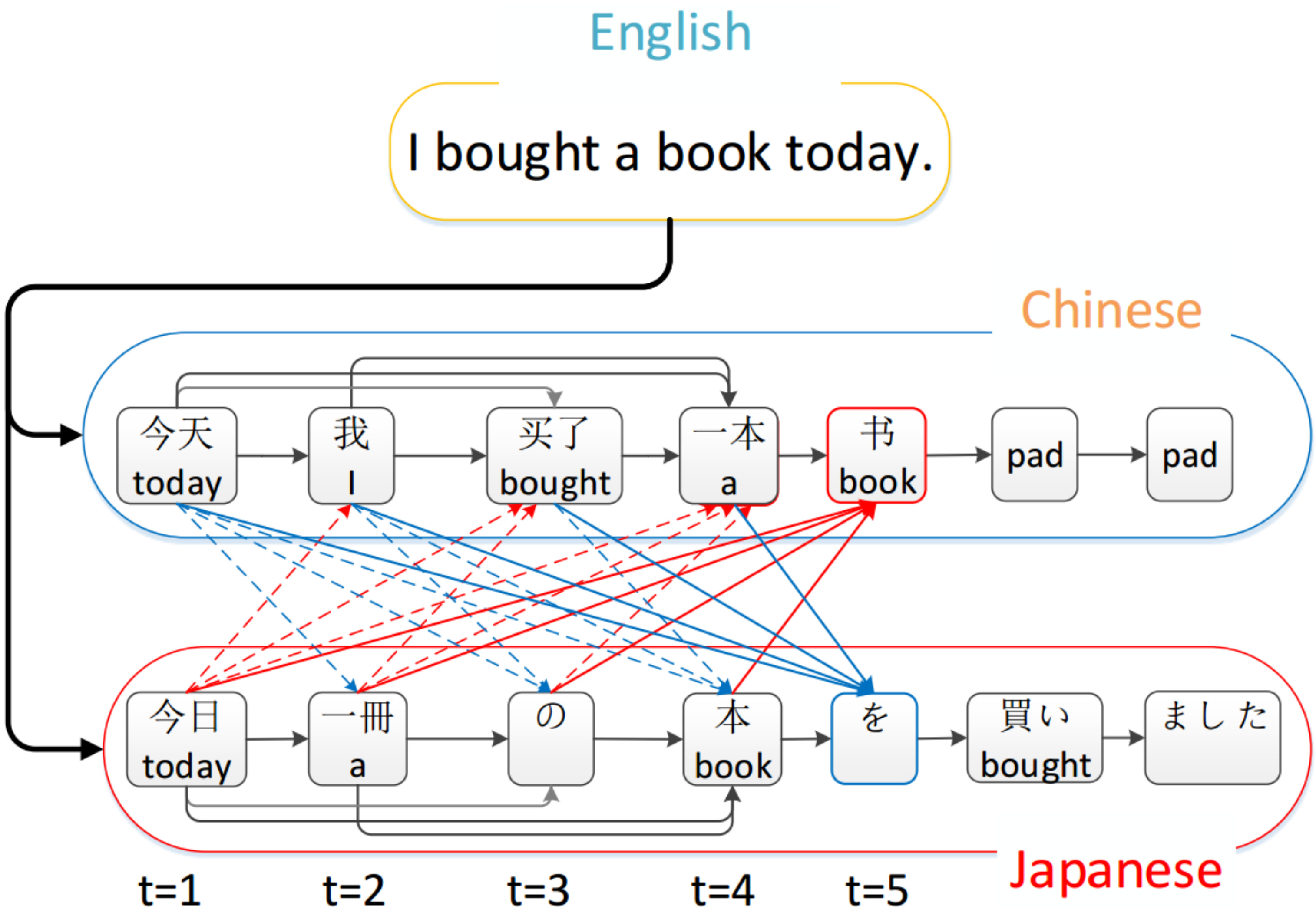


Separate Encoder or Decoder network

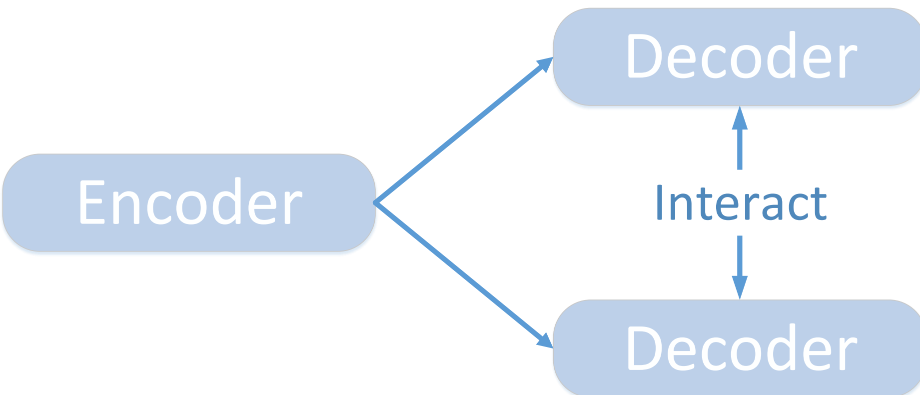
Shared Encoder or Decoder network

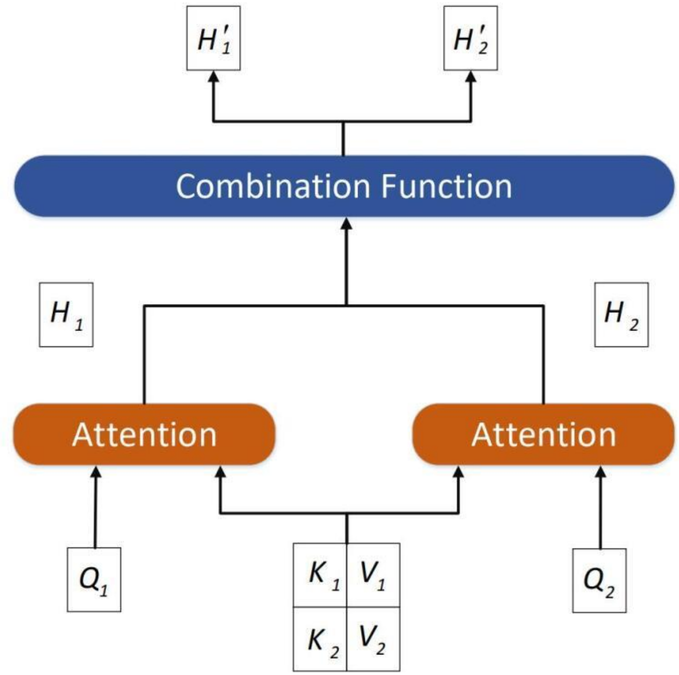
Shared with partial parameter

Synchronously Generating Two Languages with Interactive Decoding



Synchronously Generating Two Languages with Interactive Decoding

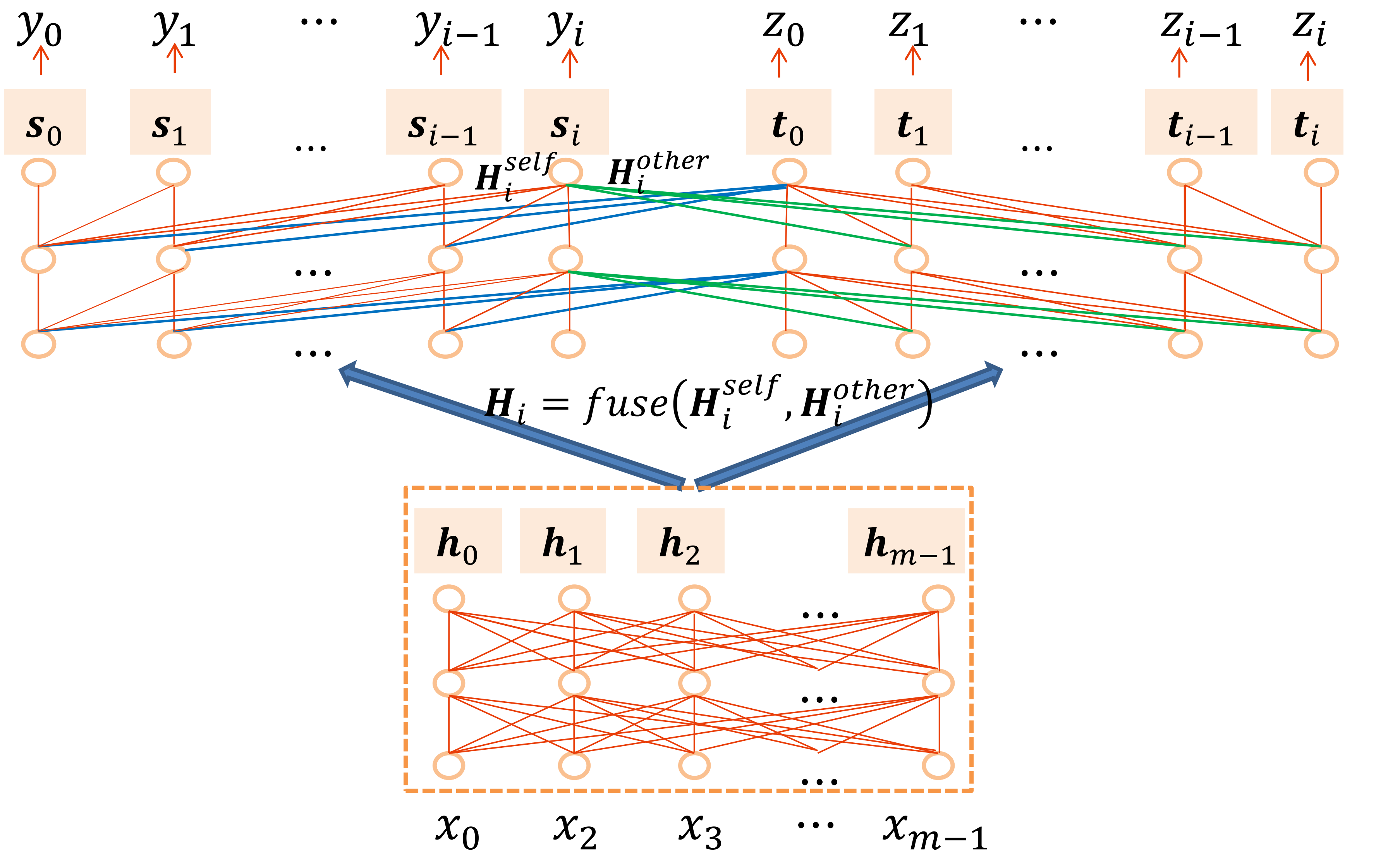




Synchronous Self-Attention Model:

$$H_1' = \mathrm{SyncAtt}(Q_1, [K_1; K_2], [V_1; V_2])$$
$$H_2' = \mathrm{SyncAtt}(Q_2, [K_1; K_2], [V_1; V_2])$$

Synchronous Bi-language Attention



Some Experiments

1. Large Scale

	WMT14 subset	
	En-De	En-Fr
Train	2.43M	2.43M
Test	3003	3003

2. Small Scale

	IWSLT			
	En-Ja	En-Zh	En-De	En-Fr
Train	223K	231K	206K	233K
Test	3003	3003	1305	1306

Training Data Construction

$$y_0 \longrightarrow y_1 \longrightarrow \cdots \quad y_i \quad \cdots \quad \cdots$$

English-Chinese Decoding

$$x_0 \ x_1 \ \cdots \ x_j \ \cdots \ x_m$$

English-to-Japanese Decoding

$$z_0 \longrightarrow z_1 \longrightarrow \cdots \quad z_j \quad \cdots \quad \cdots$$

Training Data Construction

$$y_0 \longrightarrow y_1 \longrightarrow \cdots \quad y_i \quad \cdots \quad \cdots$$

English-Chinese Decoding

$$x_0 \;\; x_1 \;\; \cdots \;\; x_j \;\; \cdots \;\; x_m$$

English-to-Japanese Decoding

$$z_0 \longrightarrow z_1 \longrightarrow \cdots \quad z_j \quad \cdots \quad \cdots$$

Training Instance Format Requirement: trilingual translation example (x, y, z) in which (x, y) and (x, z) are parallel sentence

Training Data Construction

Step1(train): $(x_1, y_1) \xrightarrow{\text{Training}}$ Model: M_1

$(x_2, y_2) \xrightarrow{\text{Training}}$ Model: M_2

Step2 (decode): $M1 \xrightarrow[\text{Inference}]{x_2} (x_2, y_2^*)$

$M2 \xrightarrow[x_1]{\text{Inference}} (x_1, y_1^*)$

Step3 (combination):

$$(x_1, y_1, y_1^*) \cup (x_2, y_2^*, y_2)$$

Main Results

- **English-Chinese/Japanese and English-German/French**

Method	En-Zh/Ja		En-De/Fr	
	En-Zh	En-Ja	En-De	En-Fr
Indiv	15.68	16.56	27.11	40.62
Indiv + pseudo	16.72	18.02	28.47	40.39
Multi	17.06	18.31	27.79	40.97
Multi + pseudo	17.10	18.40	28.56	40.62
Sync-Trans	**17.97**	**19.31**	**29.16**	**41.53**

- Indiv: System learned with bilingual training
- Multi: shared encoder-decoder networks
- Sync-Trans significantly outperforms Indiv and Multi

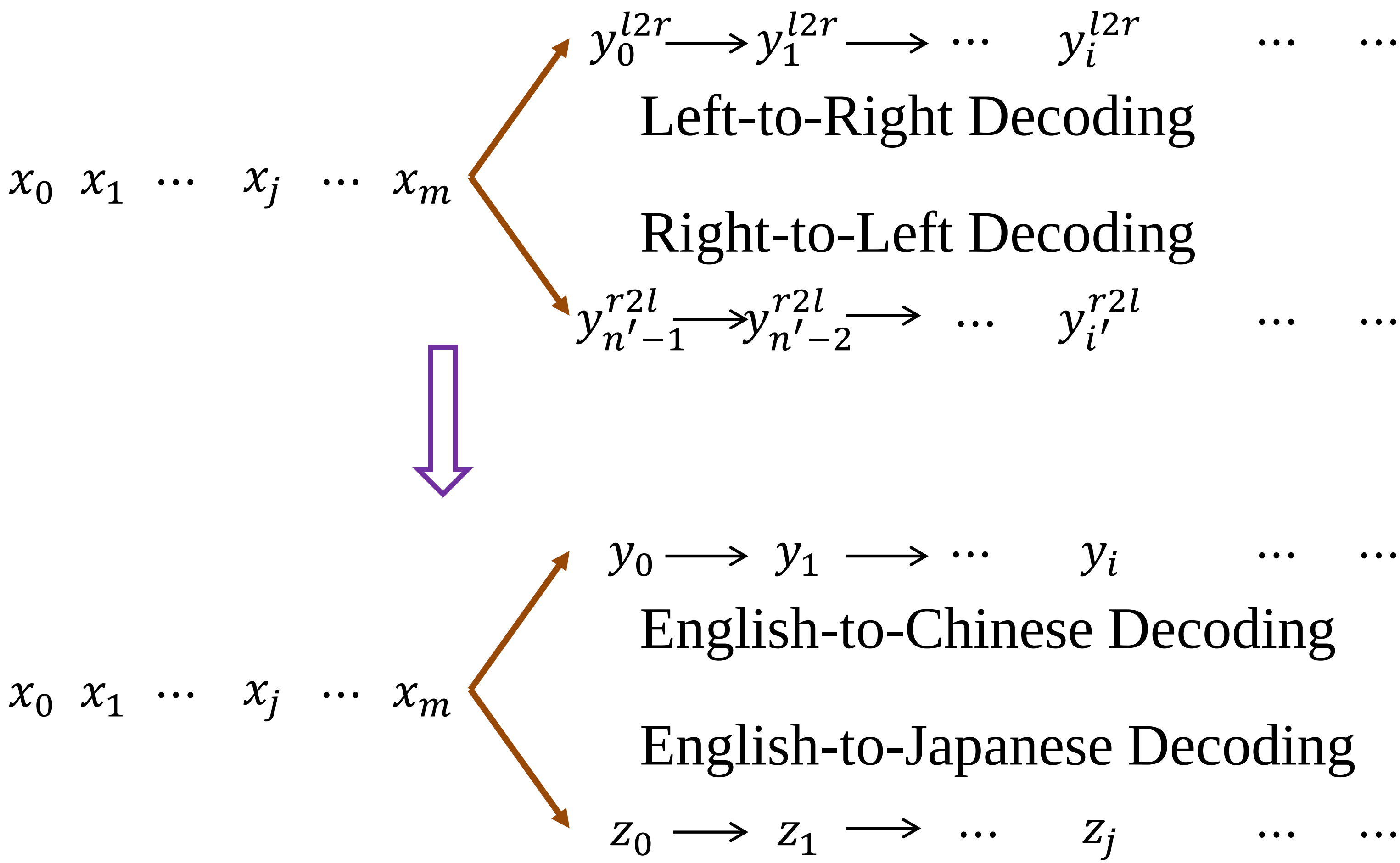
Main Results

- **Large-scale WMT Dataset**

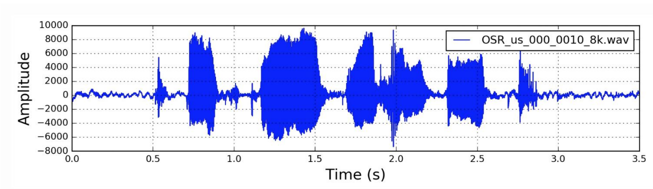
Method	WMT14 subset		WMT14
	En-De	En-Fr	En-De
Indiv	24.33	37.12	26.53
Multi	23.46	36.33	25.81
Sync-Trans	**24.84**$^{\dagger *}$	**37.66**$^{\dagger *}$	**27.01**$^{\dagger *}$

- Sync-Trans significantly outperforms Indiv and Multi

From Bidirection to Two Tasks



Interactive Inference for other Two Tasks



$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_i \quad \cdots \quad \cdots$

Speech Recognition

Speech-to-Text Translation

$y_0 \longrightarrow y_1 \longrightarrow \cdots \longrightarrow y_j \quad \cdots \quad \cdots$

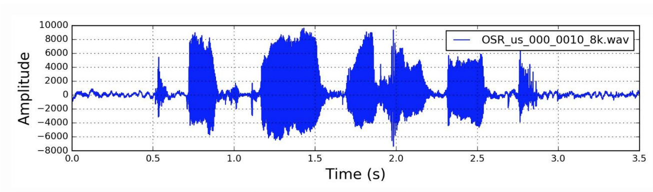
$y_0 \longrightarrow y_1 \longrightarrow \cdots \longrightarrow y_i \quad \cdots \quad \cdots$

To English Caption

To Chinese Caption

$z_0 \longrightarrow z_1 \longrightarrow \cdots \longrightarrow z_j \quad \cdots \quad \cdots$

Interactive Inference for other Two Tasks



$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_i \quad \cdots \quad \cdots$

Speech Recognition

Speech-to-Text Translation

$y_0 \longrightarrow y_1 \longrightarrow \cdots \longrightarrow y_j \quad \cdots \quad \cdots$

$y_0 \longrightarrow y_1 \longrightarrow \cdots \longrightarrow y_i \quad \cdots \quad \cdots$

To English Caption

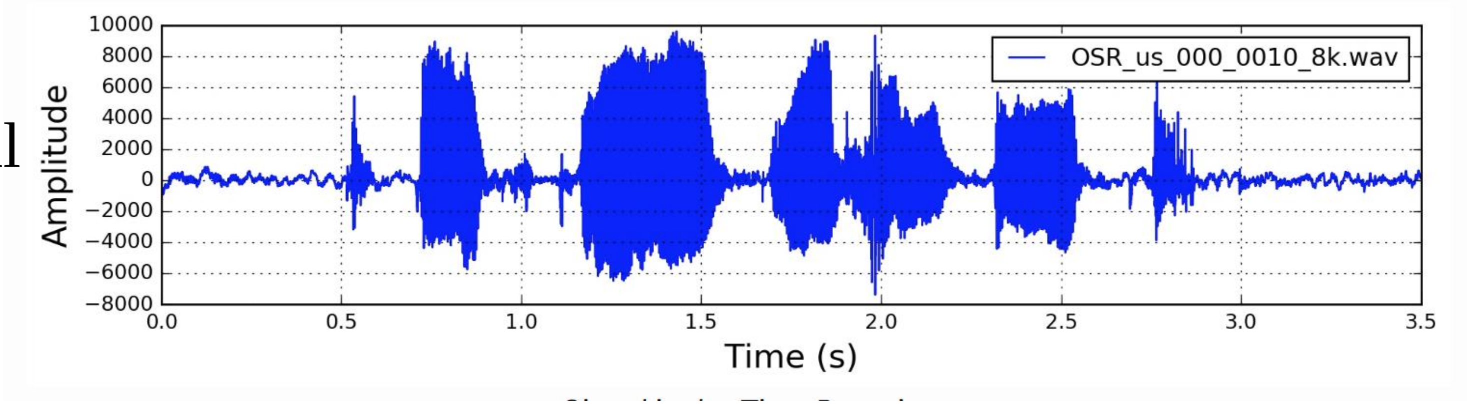
To Chinese Caption

$z_0 \longrightarrow z_1 \longrightarrow \cdots \longrightarrow z_j \quad \cdots \quad \cdots$

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Speech Features**

Original signal



Mel滤波

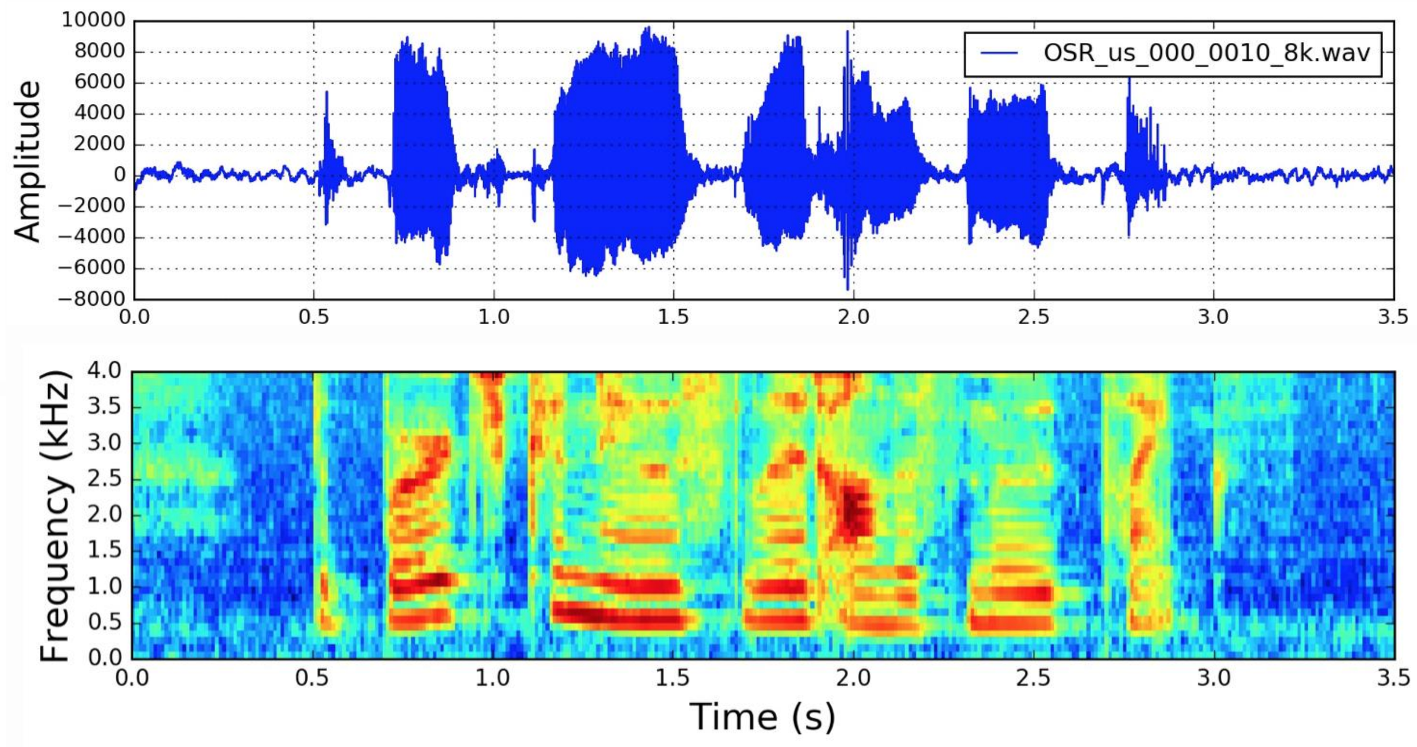
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Speech Features**

Original signal

Mel滤波

Filter bank



Interactive Inference for Speech Recognition and Speech-to-Text Translation

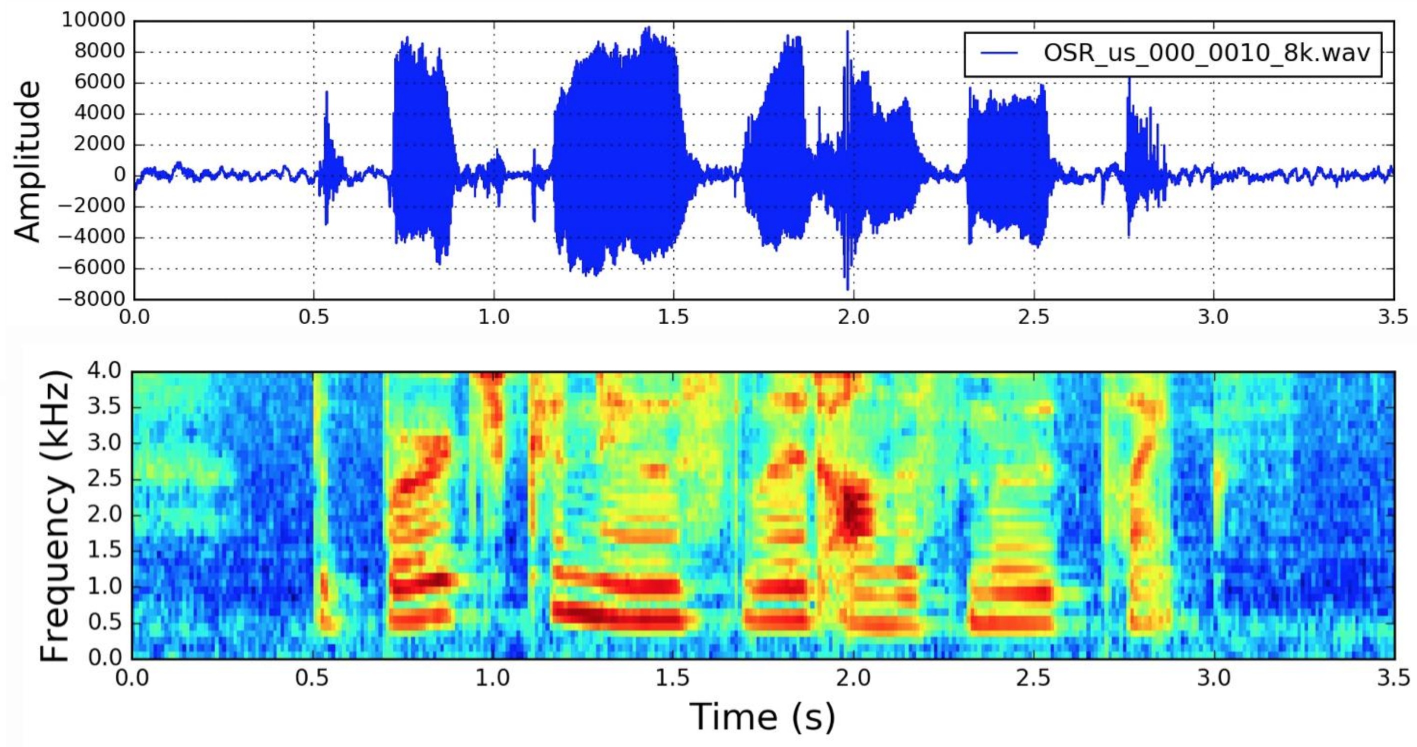
- **Speech Features**

Original signal

Mel滤波

Filter bank

离散余弦变换



Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Speech Features**

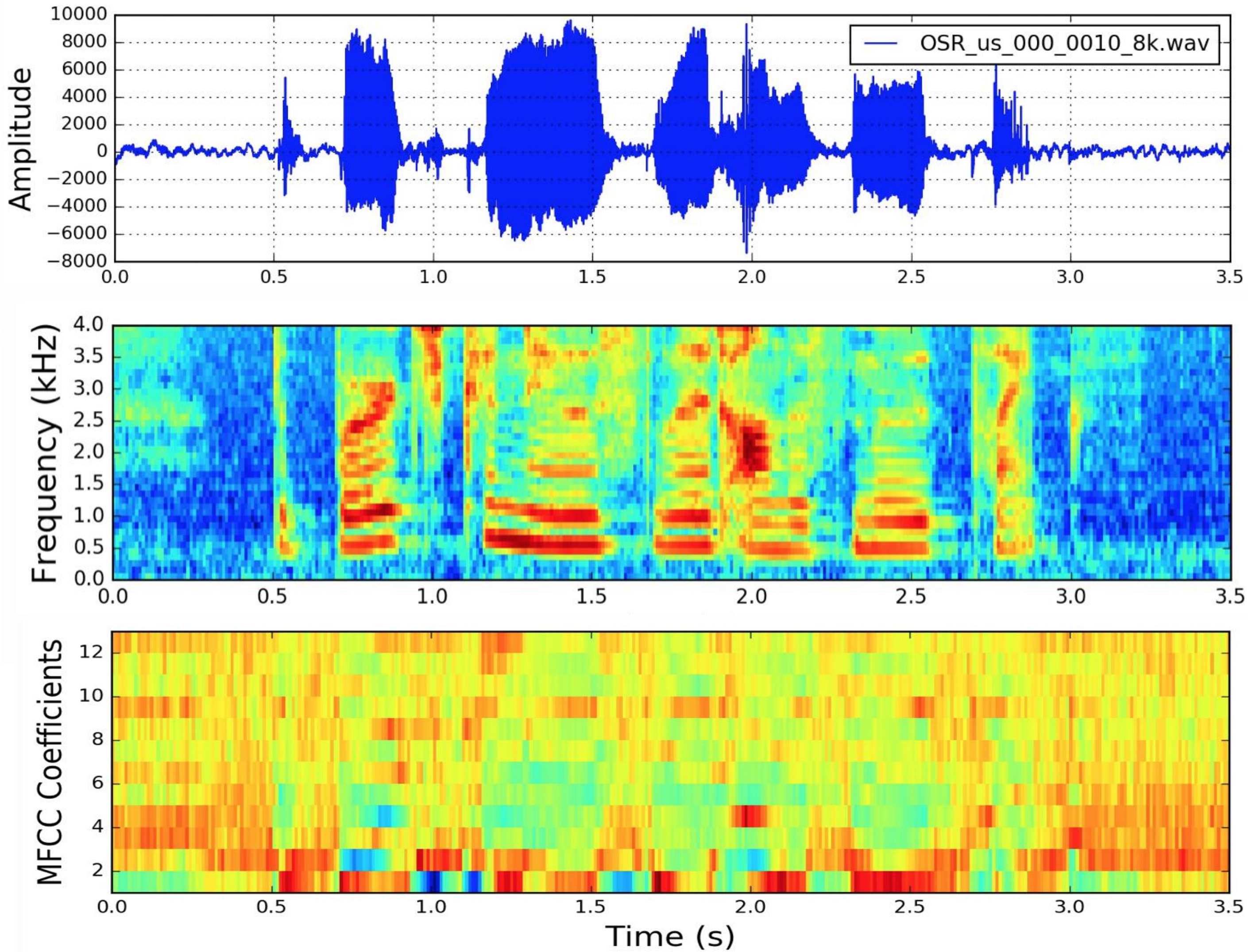
Original signal

Mel滤波

Filter bank

离散余弦变换

MFCC



Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Speech Features**

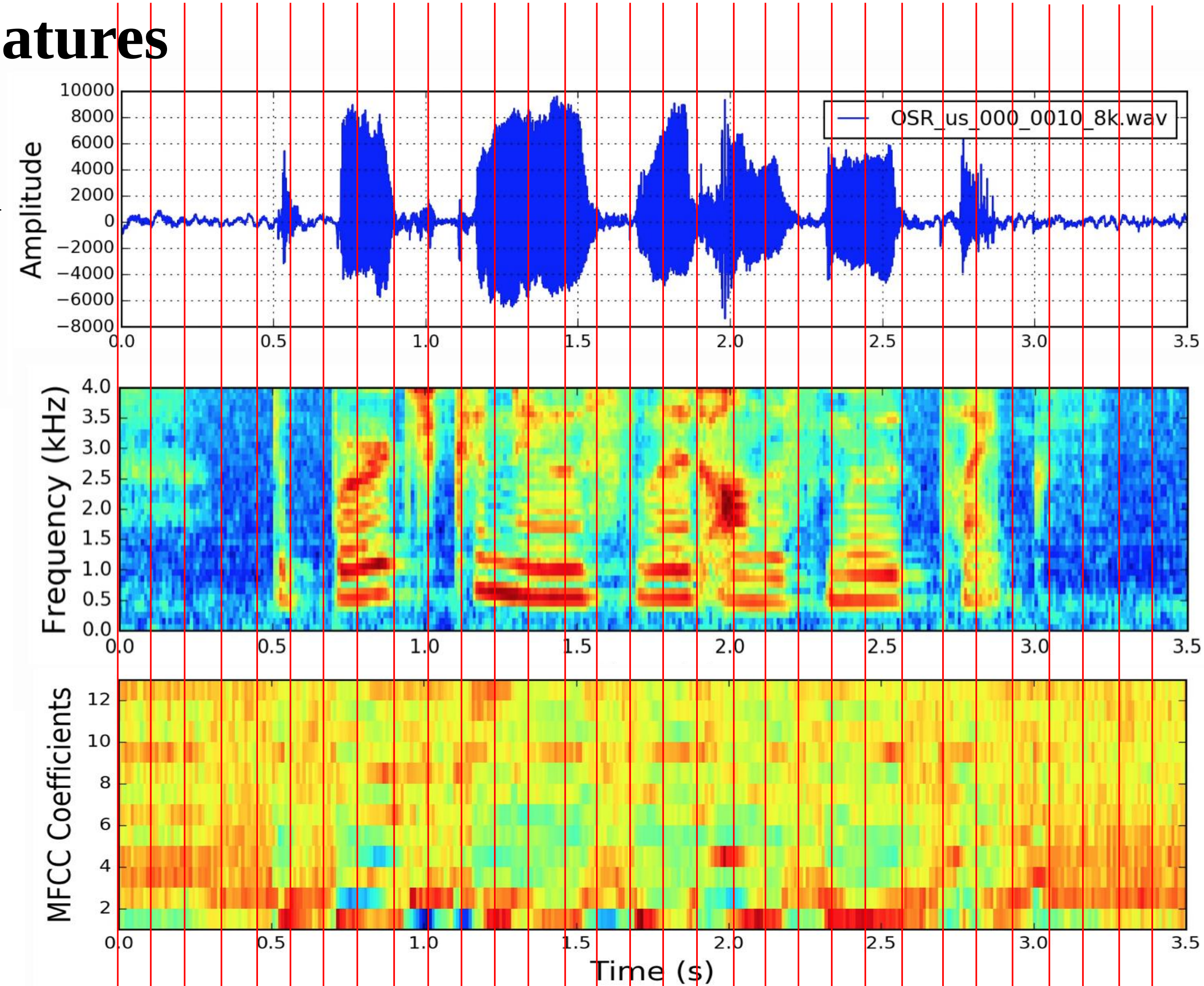
Original signal

Mel滤波

Filter bank

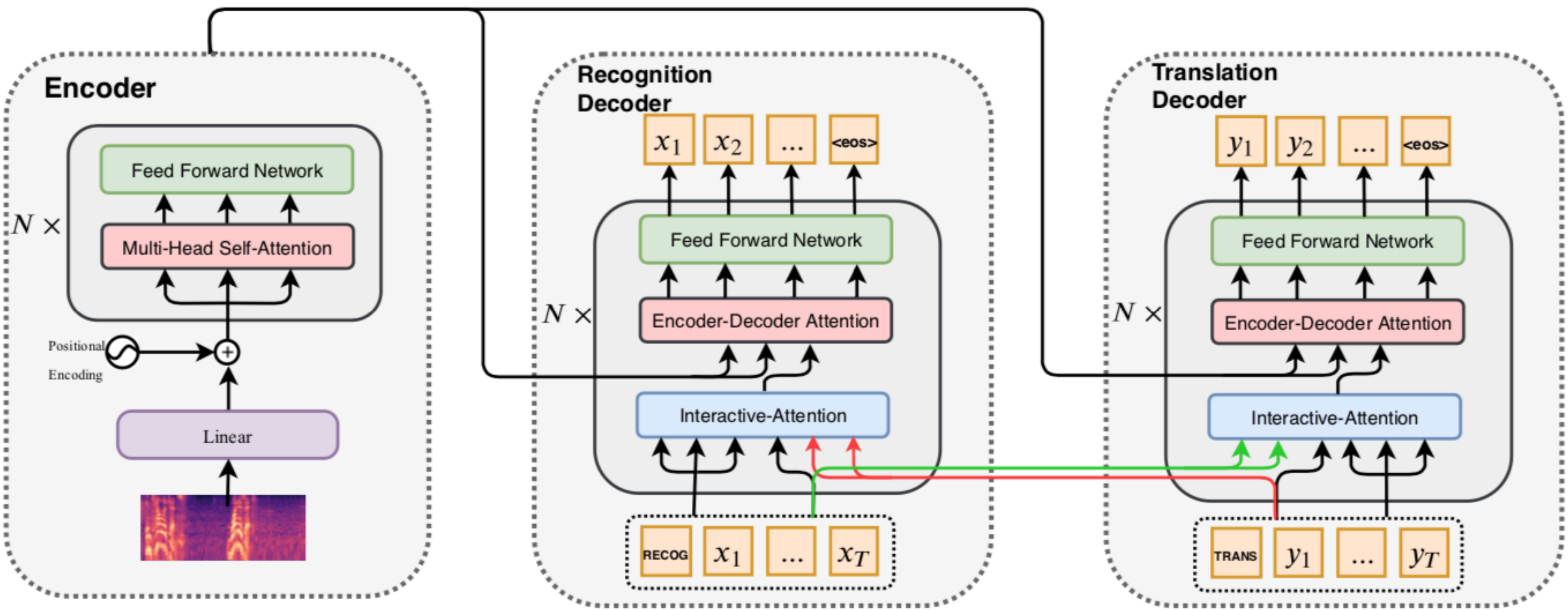
离散余弦变换

MFCC



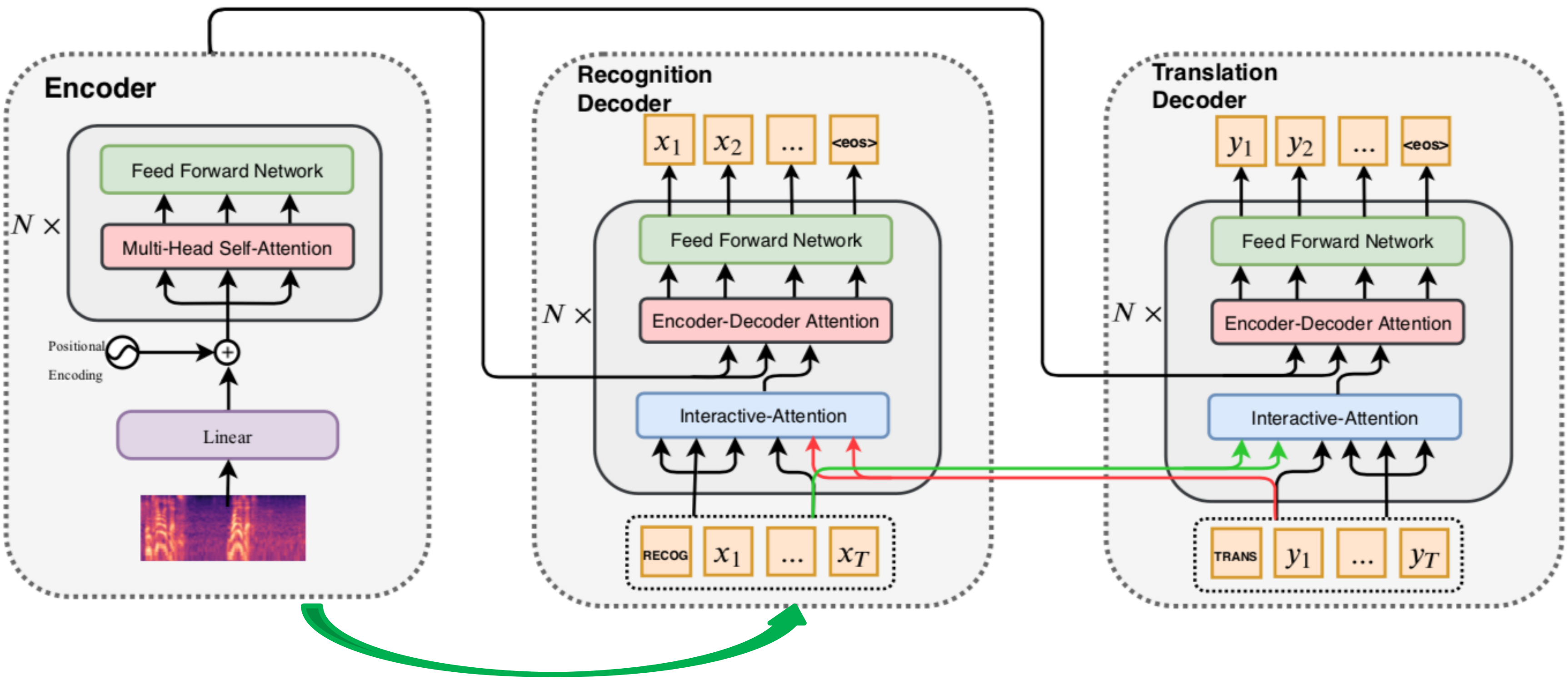
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Overall Architecture**



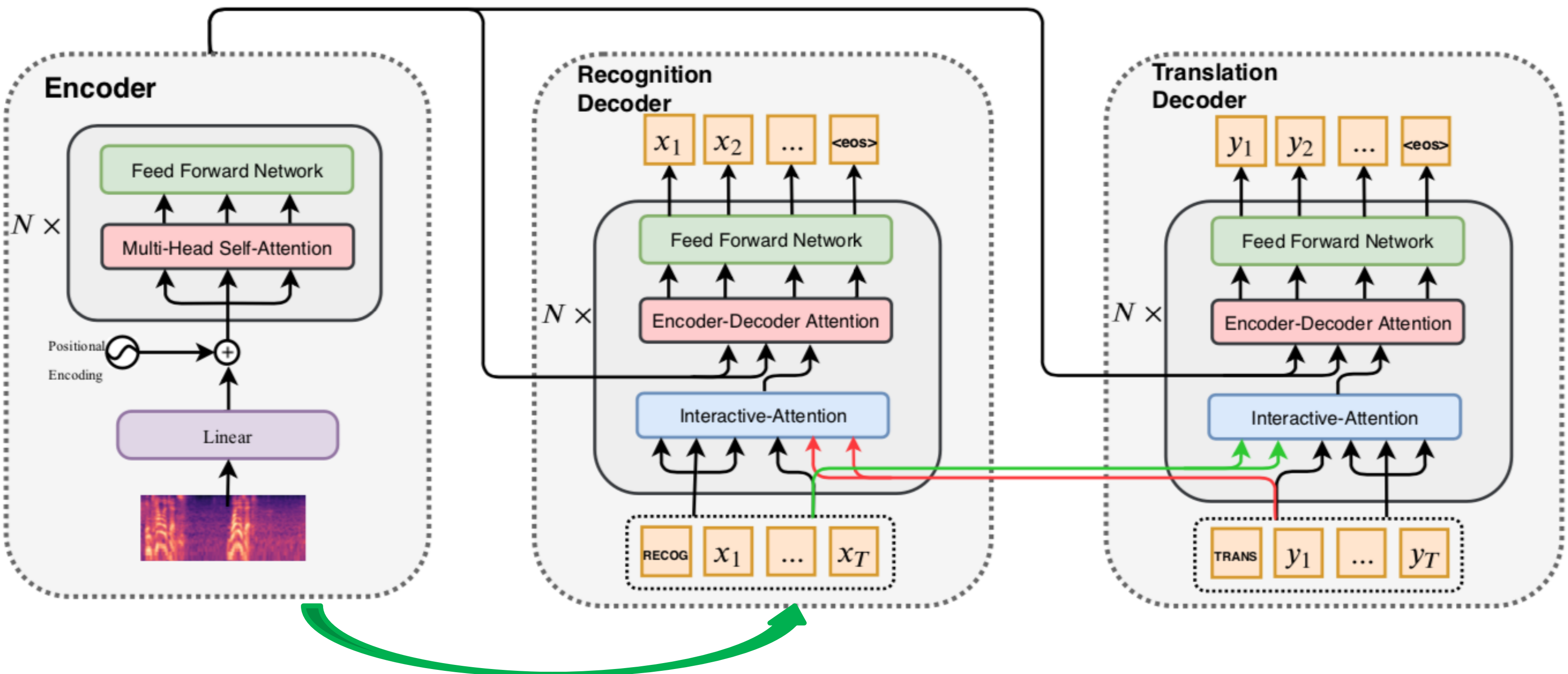
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Overall Architecture**



Interactive Inference for Speech Recognition and Speech-to-Text Translation

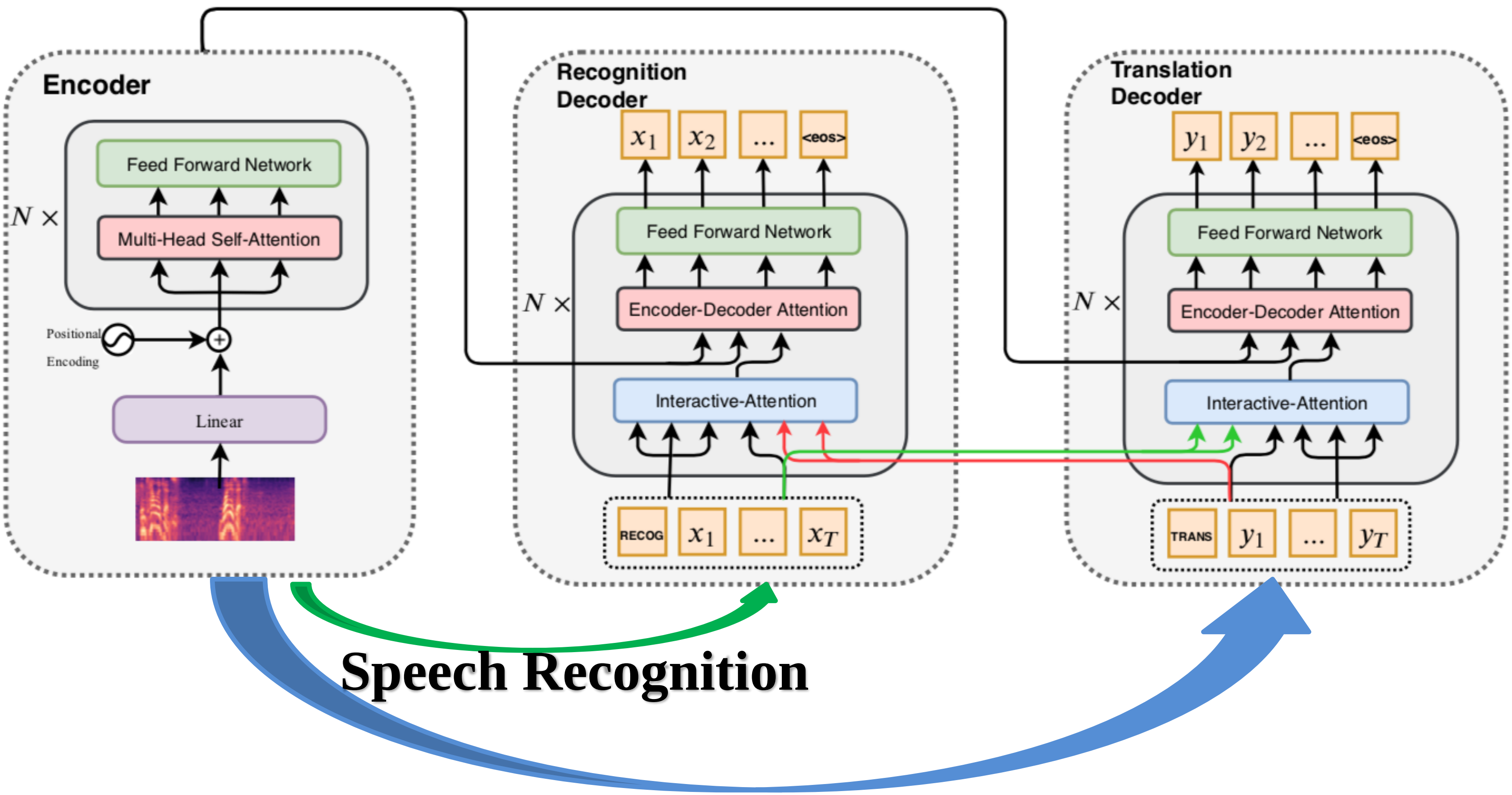
- **Overall Architecture**



Speech Recognition

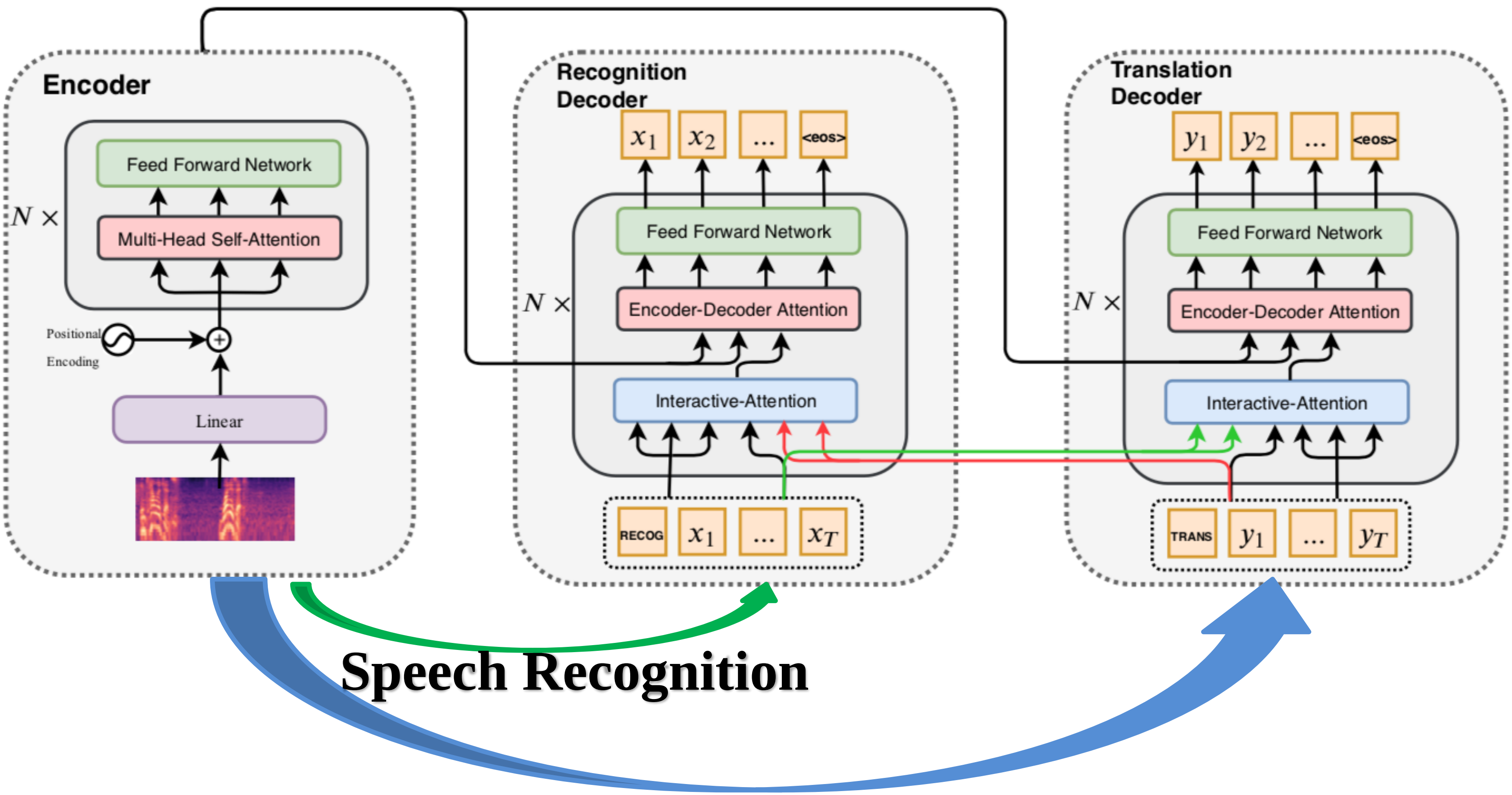
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Overall Architecture**



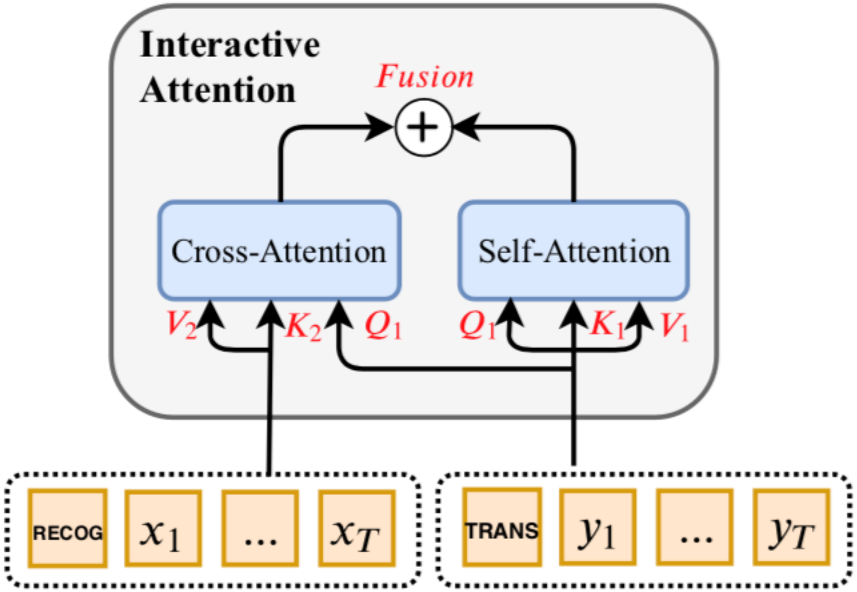
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Overall Architecture**



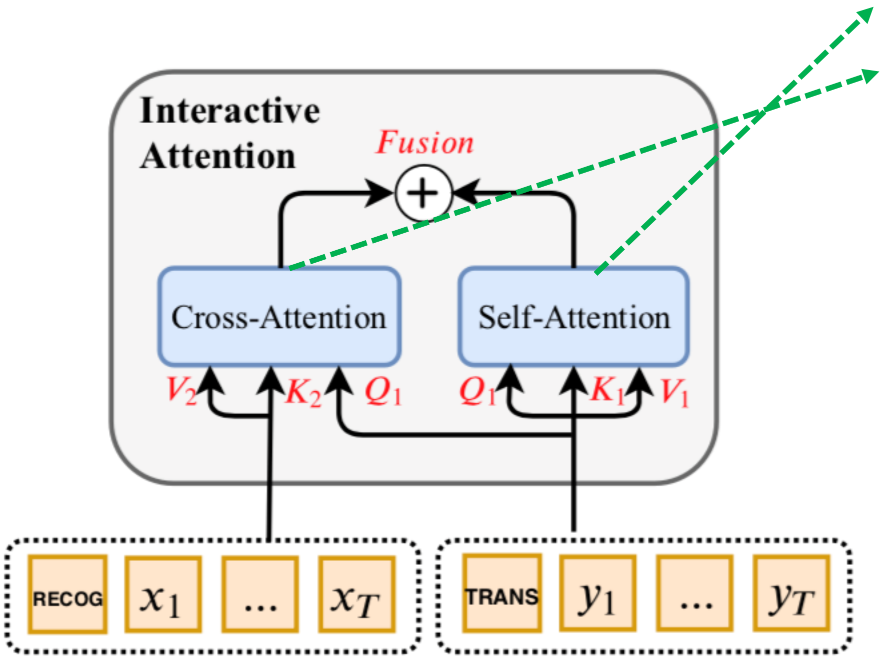
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Interactive Attention**



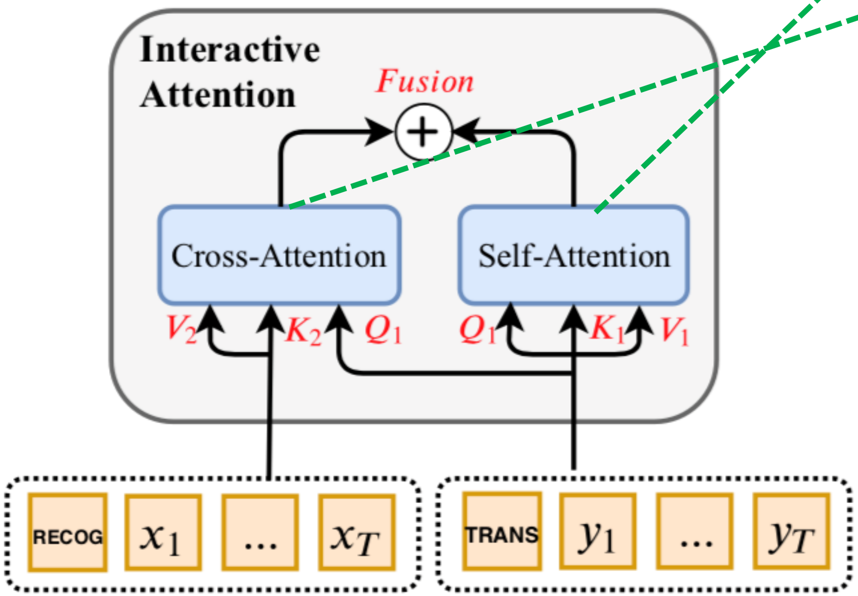
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Interactive Attention**



Interactive Inference for Speech Recognition and Speech-to-Text Translation

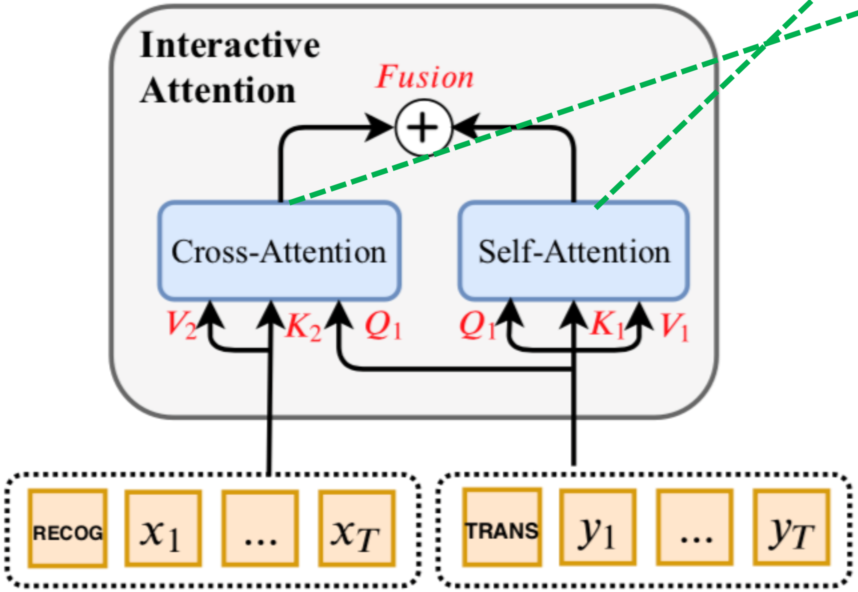
- **Interactive Attention**



$$H_1 = Attention(Q_1, K_1, V_1)$$
$$H_2 = Attention(Q_1, K_2, V_2)$$
$$H_{final} = Fusion(H_1, H_2)$$

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Interactive Attention**



$$H_1 = Attention(Q_1, K_1, V_1)$$
$$H_2 = Attention(Q_1, K_2, V_2)$$
$$H_{final} = Fusion(H_1, H_2)$$

- **Fusion**
  - Linear Interpolation

    $$H_{final} = \lambda_1 * H_1 + \lambda_2 * H_2$$
  - Nonlinear Interpolation

    $$H_{final} = \lambda_1 * H_1 + \lambda_2 * \tanh(H_2)$$
  - Gate Interpolation

    $$r, z = \sigma(W[H_1; H_2])$$
    $$H_{final} = r \odot H_1 + z \odot H_2$$

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- Experimental Setup
  - Dataset:

    TED En-Fr，En-Zh
  - Train details:

    (1) *Transformer_big* setting

    (2) English-Chinese: 2 GPUs, character BLEU

    (3) English-French: 2 GPUs, tokenizer BLEU

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- Data Size

Corpus		Total		Source (per segment)		Target (per segment)
		segments	hours	frames	words	words
Fisher/Callh-ome (En-Es)	train	138,819	138:00	762	20.7	20.3
	dev	3,961	2:00	673	17.9	17.9
	test	3,641	3:44	657	18.3	18.3
TED (En-Zh/ En-Fr)	train	305,971	527:00	662	17.9	20.3
	dev	1,148	2:23	659	18.2	17.9
	test	1,223	2:37	624	18.3	18.1

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- Data Size

Corpus		Total		Source (per segment)		Target (per segment)
		segments	hours	frames	words	words
Fisher/Callh -ome (En-Es)	train	138,819	138:00	762	20.7	20.3
	dev	3,961	2:00	673	17.9	17.9
	test	3,641	3:44	657	18.3	18.3
TED (En-Zh/ En-Fr)	train	305,971	527:00	662	17.9	20.3
	dev	1,148	2:23	659	18.2	17.9
	test	1,223	2:37	624	18.3	18.1

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- Baselines
  - **Pipeline:** Transformer ASR + Transformer MT
  - **Pre-trained E2E:** Pretrain on ASR, fintune on ST
  - **Multi-task:** ASR + ST with encoder shared
  - **Two-stage:** (1) use the first decoder to generate transcription sequence; (2) use the output of first decoder on the second decoder

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Evaluation Metrics**
  - **ASR Metric**

REF:	各位	来宾	*	各位	合作	伙伴	媒体界	的	朋友们	下午	好
ASR:	各位	来宾	个	各位	合作	伙伴	媒体界	*	朋友们	刚	好
			I					D		S	

$$WER = 100 \cdot \frac{S + D + I}{N} \%$$

  - **MT Metric**

$BLEU$

Interactive Inference for Speech Recognition and Speech-to-Text Translation

- Overall Results

Model	**En-De**		**En-Fr**		**En-Zh**		**En-Ja**	
	WER(↓)	BLEU(↑)	WER(↓)	BLEU(↑)	WER(↓)	BLEU(↑)	WER(↓)	BLEU(↑)
MT	/	22.19	/	30.68	/	25.01	/	22.93
Pipeline	14.29	19.50	14.20	26.62	14.20	21.52	14.21	**20.87**
E2E	14.29	16.07	14.20	27.63	14.20	19.15	14.21	16.59
Multi-task	14.20	19.08	13.04	28.71	13.43	20.60	14.01	18.73
Two-stage	14.27	20.08	13.34	**30.08**	13.55	20.29	13.85	19.32
Interactive	**14.16**	**21.11**	**12.58**	29.79	**13.38**	**21.68**	**13.52**	20.06

Interactive Inference for Image Caption in Two Languages

$y_0 \longrightarrow y_1 \longrightarrow \cdots \longrightarrow y_i \quad \cdots \quad \cdots$

To English Caption

To German Caption

$z_0 \longrightarrow z_1 \longrightarrow \cdots \longrightarrow z_j \quad \cdots \quad \cdots$

- Dataset:
  (1)Multi30k **(**Elliott et al., 2016**):** English and German Captions
  (2) 29,000 image-caption for training
  (3) 1014 for validation and 2000 for test

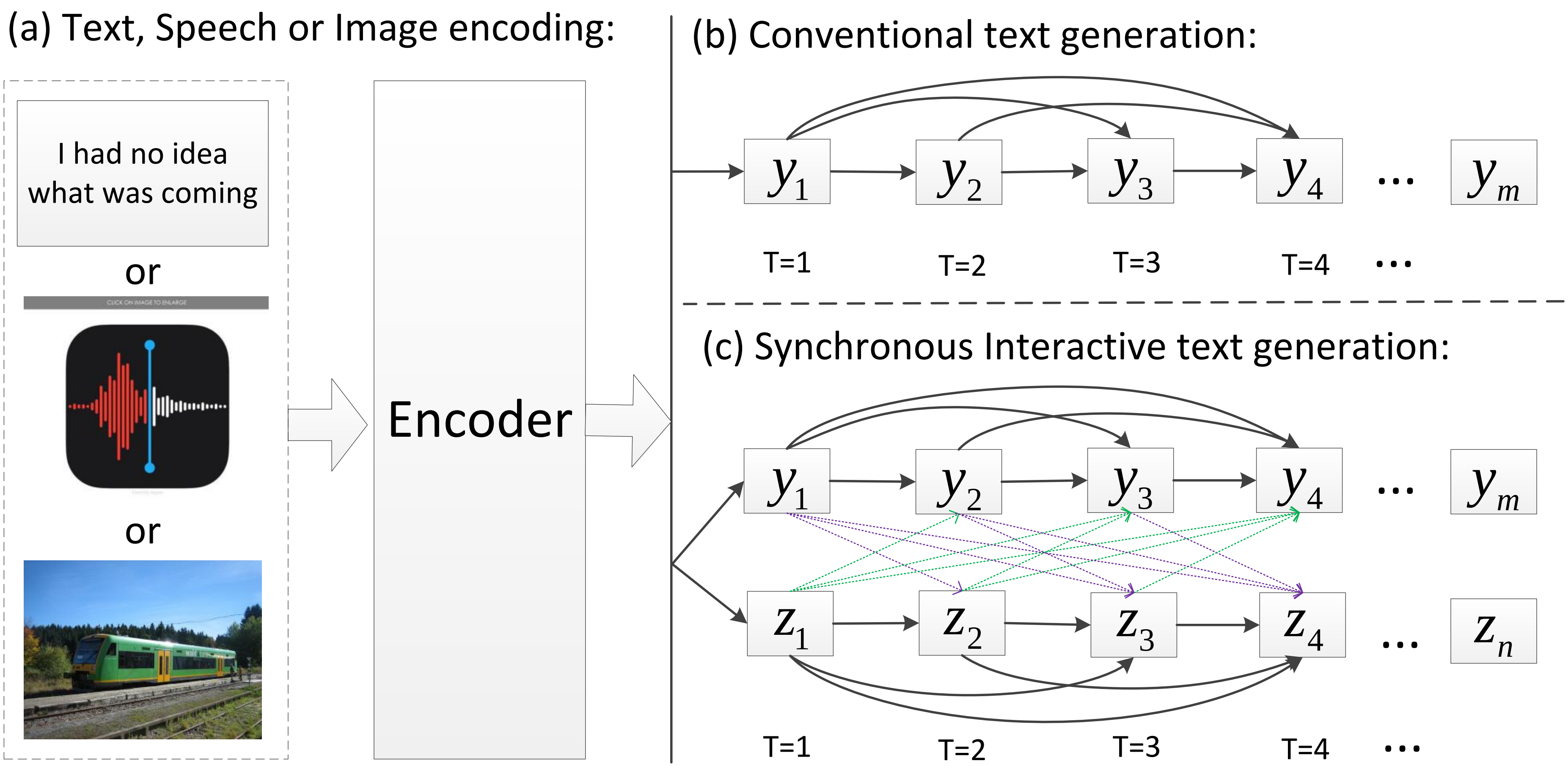
- Baselines:
  (1) VGGNet encoder + LSMT decoder (Xu et al., 2015)
  (2) Transformer

Experiments: Image Caption in Two Languages

- Results on English and German Image Captions
  - BLEU score

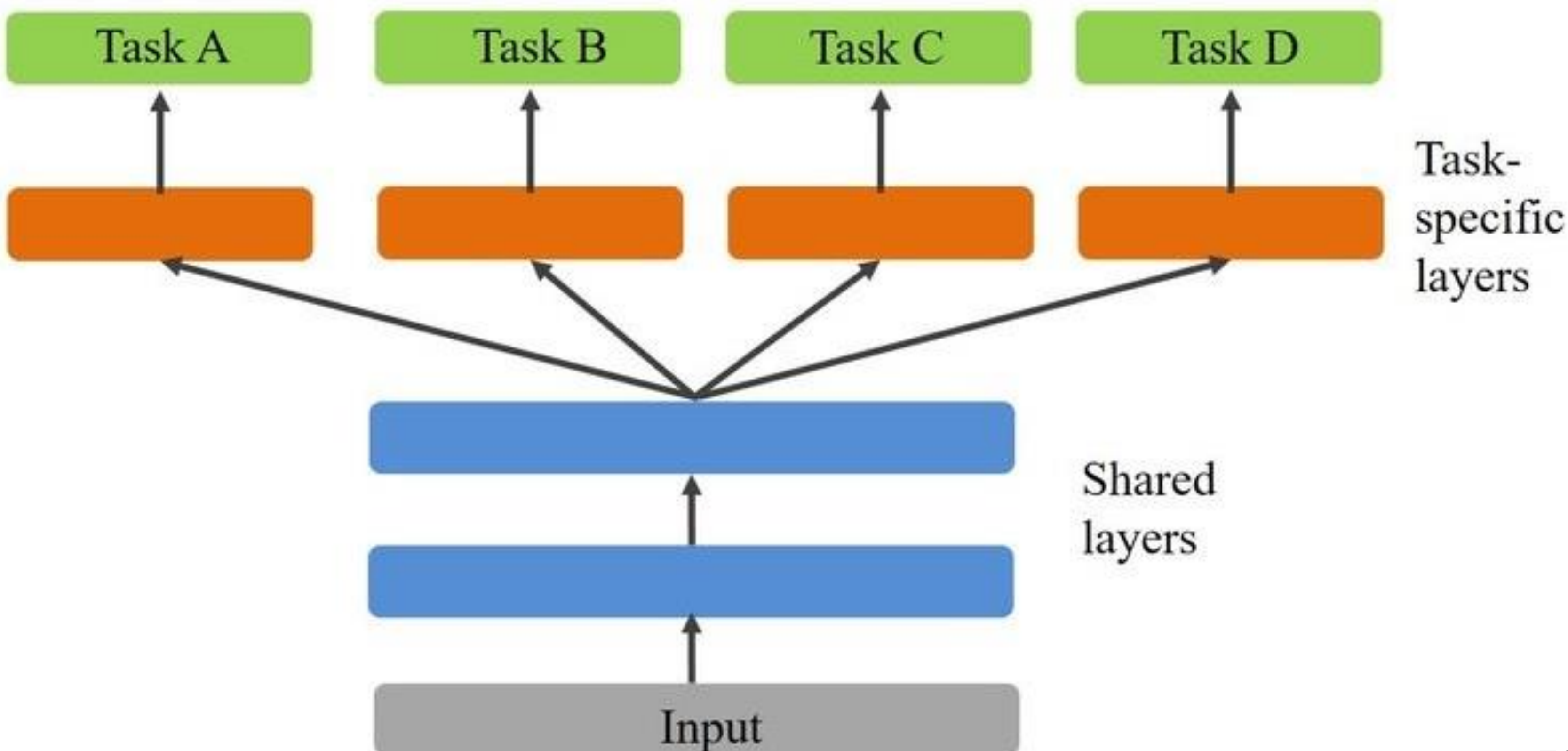
Method	English	German
Xu et al., (2015)	19.90	~
Jaffe (2017)	~	11.84
Transformer	21.25	13.55
Ours	**22.54**	**15.49**

Unified Text Generation from Text, Speech and Image



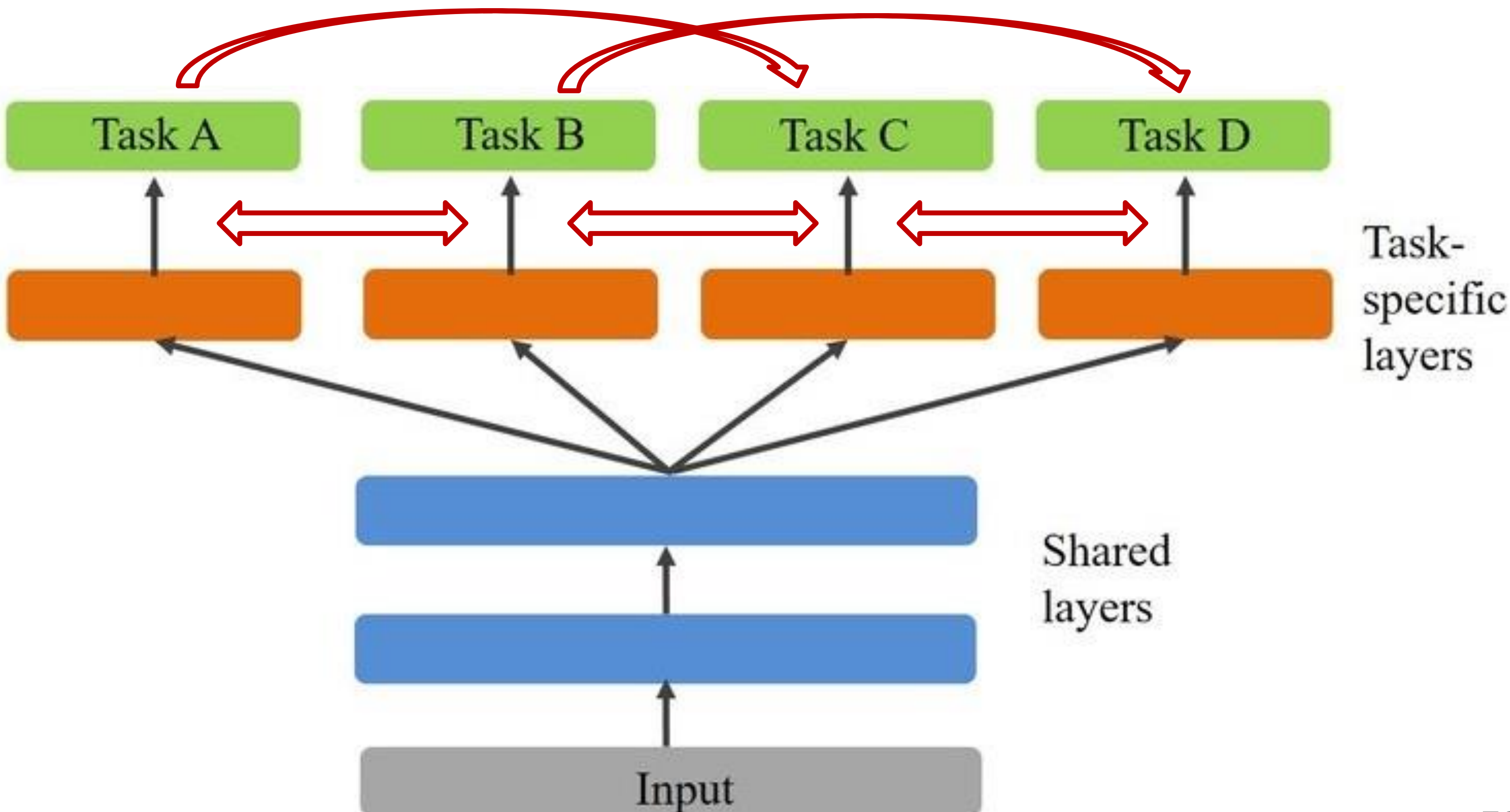
And Beyond …

- **Why not interactive inference for multi-task learning?**



And Beyond …

- **Why not interactive inference for multi-task learning?**



Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Two Tasks
- Summary and Future Challenges

Summary

- The synchronous bidirectional Inference model that can take full advantage of both history and future information provided by bidirectional decoding states, achieving promising results.

Summary

- The synchronous bidirectional Inference model that can take full advantage of both history and future information provided by bidirectional decoding states, achieving promising results.

•The bidirectional inference model can be further extended to inference from both sides to the middle to improve the efficiency.

Summary

- The synchronous bidirectional Inference model that can take full advantage of both history and future information provided by bidirectional decoding states, achieving promising results.

•The bidirectional inference model can be further extended to inference from both sides to the middle to improve the efficiency.

•The bidirectional inference model can be generalized to generate two languages synchronously and interactively.

Summary

- The synchronous bidirectional Inference model that can take full advantage of both history and future information provided by bidirectional decoding states, achieving promising results.

•The bidirectional inference model can be further extended to inference from both sides to the middle to improve the efficiency.

•The bidirectional inference model can be generalized to generate two languages synchronously and interactively.

•Our main code is available at https://github.com/ZNLP/sb-nmt. Feel free to have a try！

Summary

- The synchronous bidirectional Inference model that can take full advantage of both history and future information provided by bidirectional decoding states, achieving promising results.

•The bidirectional inference model can be further extended to inference from both sides to the middle to improve the efficiency.

•The bidirectional inference model can be generalized to generate two languages synchronously and interactively.

Message: Synchronous Interactive Inference May Reshape Multi-task Generation

•Our main code is available at https://github.com/ZNLP/sb-nmt. Feel free to have a try！

Future Challenges

- How to generalize the interactive inference idea into multi-task problems in which three or more tasks are concerned?
- How to perform efficient training without generating pseudo parallel instances?
- How to effectively combine bidirectional inference in the multi-task interactive inference problem?

Reference

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *In NAACL-HLT* 2019 (**Best Paper**).
2. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by Generative Pre-Training. Technical report, OpenAI.
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NIPS-2017.
4. Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Agreement on target-bidirectional neural machine translation. NAACL-2016.
5. Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. Regularizing neural machine translation by target-bidirectional agreement. AAAI-2019.
6. Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. Neural system combination for machine translation. ACL-2017.
7. Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Rongrong Ji, and Hongji Wang. Asynchronous bidirectional decoding for neural machine translation. AAAI-2018.
8. Long Zhou, Jiajun Zhang, and Chengqing Zong. Synchronous Bidirectional Neural Machine Translation. TACL-2019.

Reference

9. Long Zhou, Jiajun Zhang, Chengqing Zong and Heng Yu. Sequence Generation: From Both Sides to the Middle. IJCAI-2019.

10. Yining Wang, Jiajun Zhang, Long Zhou, Yuchen Liu and Chengqing Zong. Synchronously Generating Two Languages with Interactive Decoding. EMNLP-2019

11. Jiajun Zhang, Long Zhou, Yang Zhao and Chengqing Zong. Synchronous Bidirectional Inference for Neural Sequence Generation. arXiv preprint arXiv:1902.08955.

12. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. ICML-2015.

13. Desmond Elliott, Stella Frank, Khalil Sima'an, Lucia Specia. Multi30K: Multilingual English-German Image Descriptions. Proceedings of the 5th Workshop on Vision and Language. 2016.

14. Alan Jaffe. Generating Image Descriptions using Multilingual Data. Proceedings of the Second Conference on Machine Translation 2017.

15. Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL-2014.

中国科学院自动化研究所

谢谢！

Thanks!