CSCE 638 Natural Language Processing Foundation and Techniques

Lecture 8: Transformers

Kuan-Hao Huang

Spring 2025

(Some slides adapted from Chris Manning, Karthik Narasimhan, Danqi Chen, and Vivian Chen)

Lecture Plan

- Transformers
  - Encoder
  - Decoder
  - Encoder-Decoder
- Transformers Variants
  - Longformer
  - Relative Positional Encoding
  - RoFormer

Recap: Attention Is All You Need

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

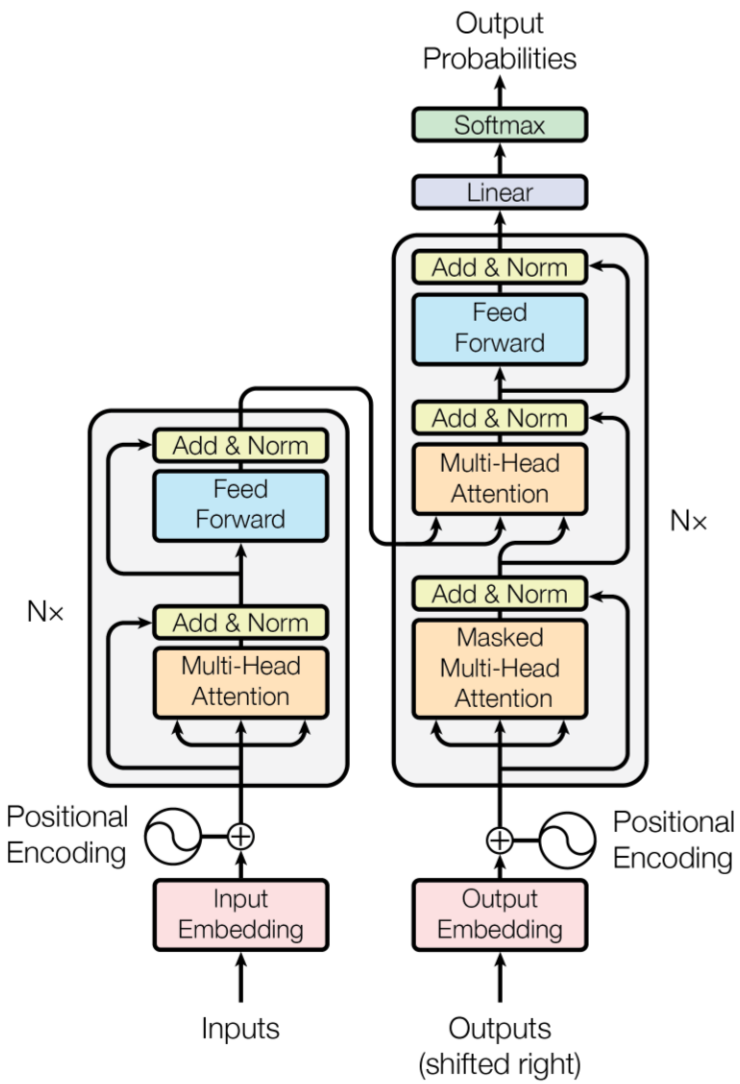
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

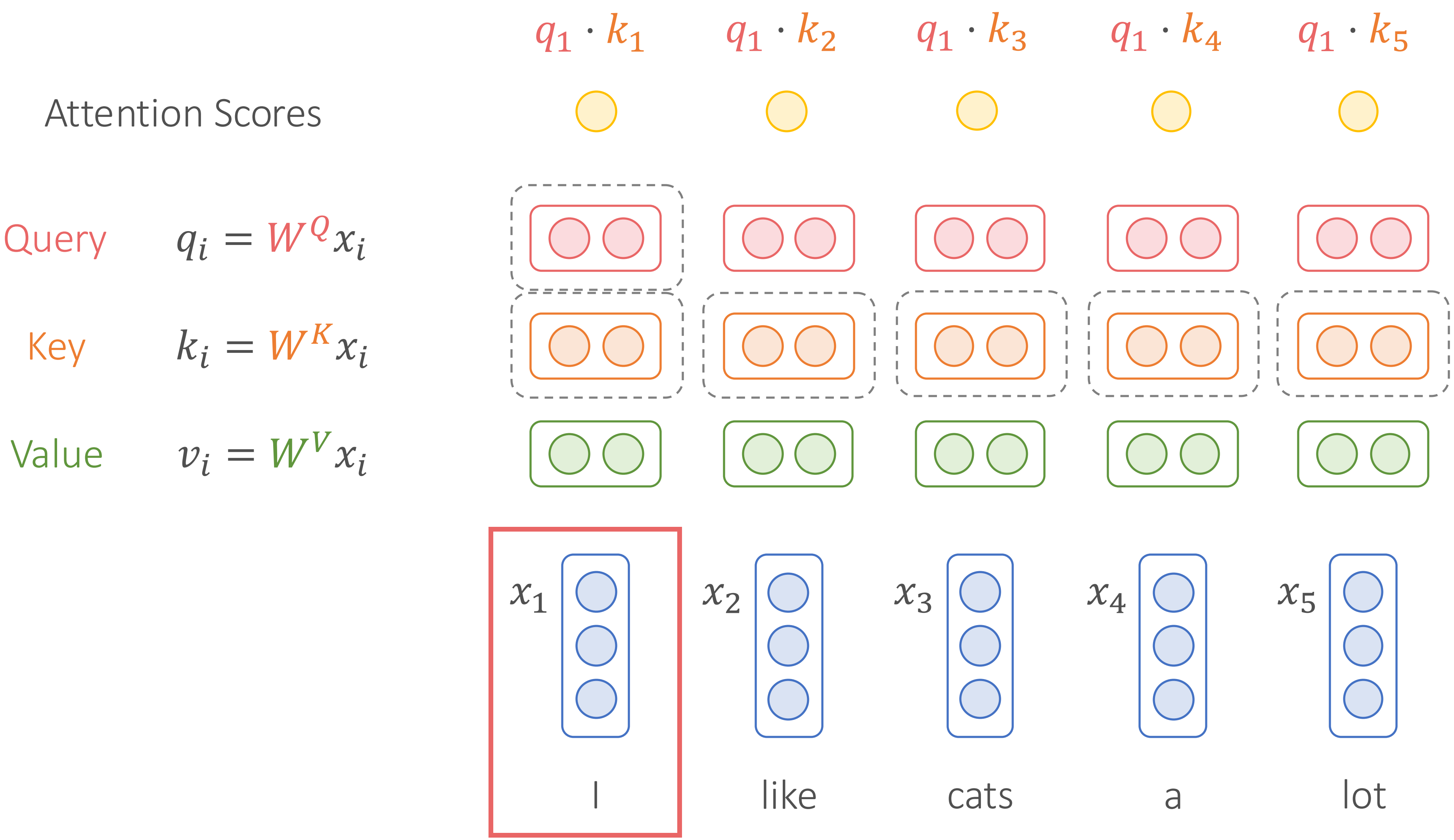
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



Recap: Self-Attention



Recap: Self-Attention

$$\alpha_{1,i} = \text{softmax}\left(\frac{q_1 \cdot k_i}{\sqrt{d}}\right)$$

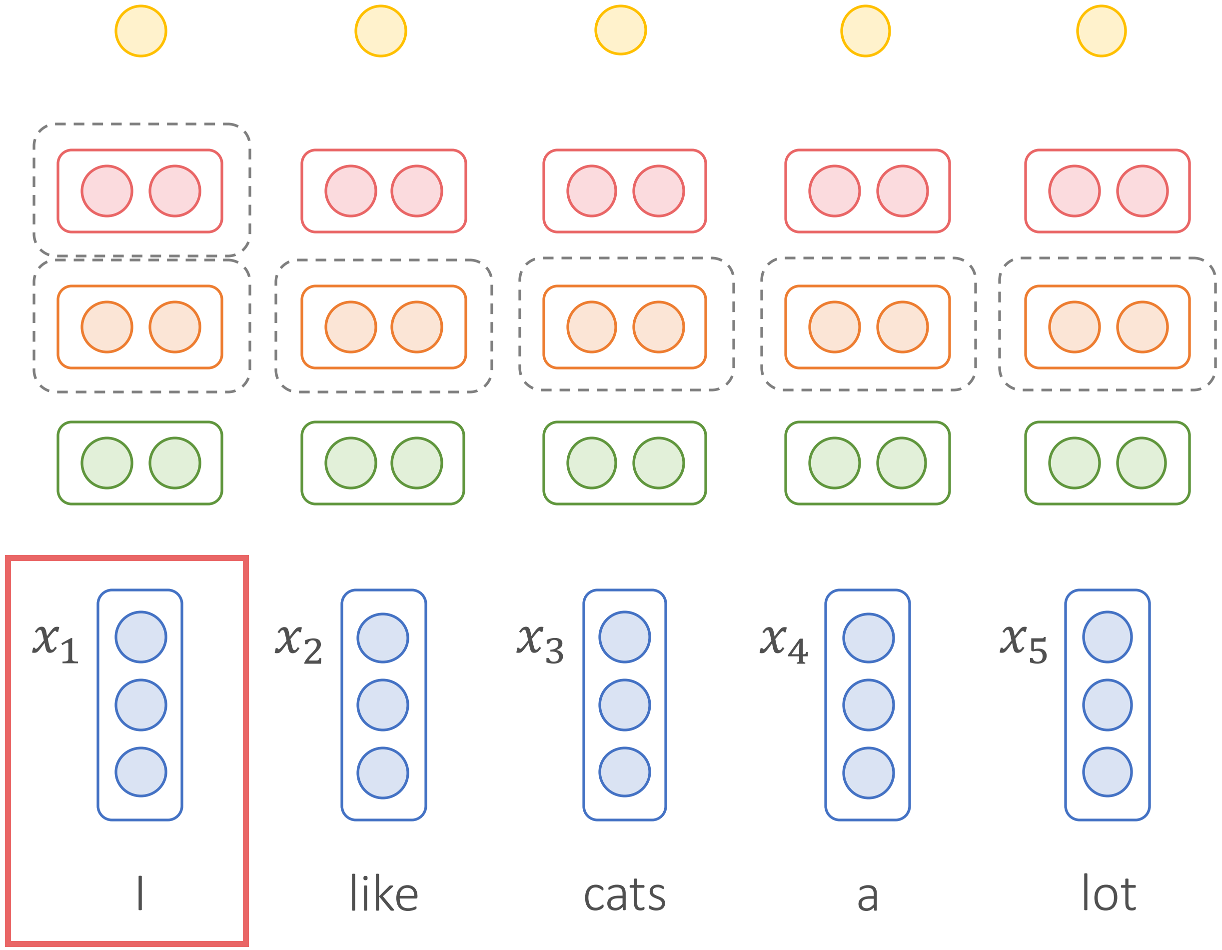
Vector dimension

Normalized Attention Scores

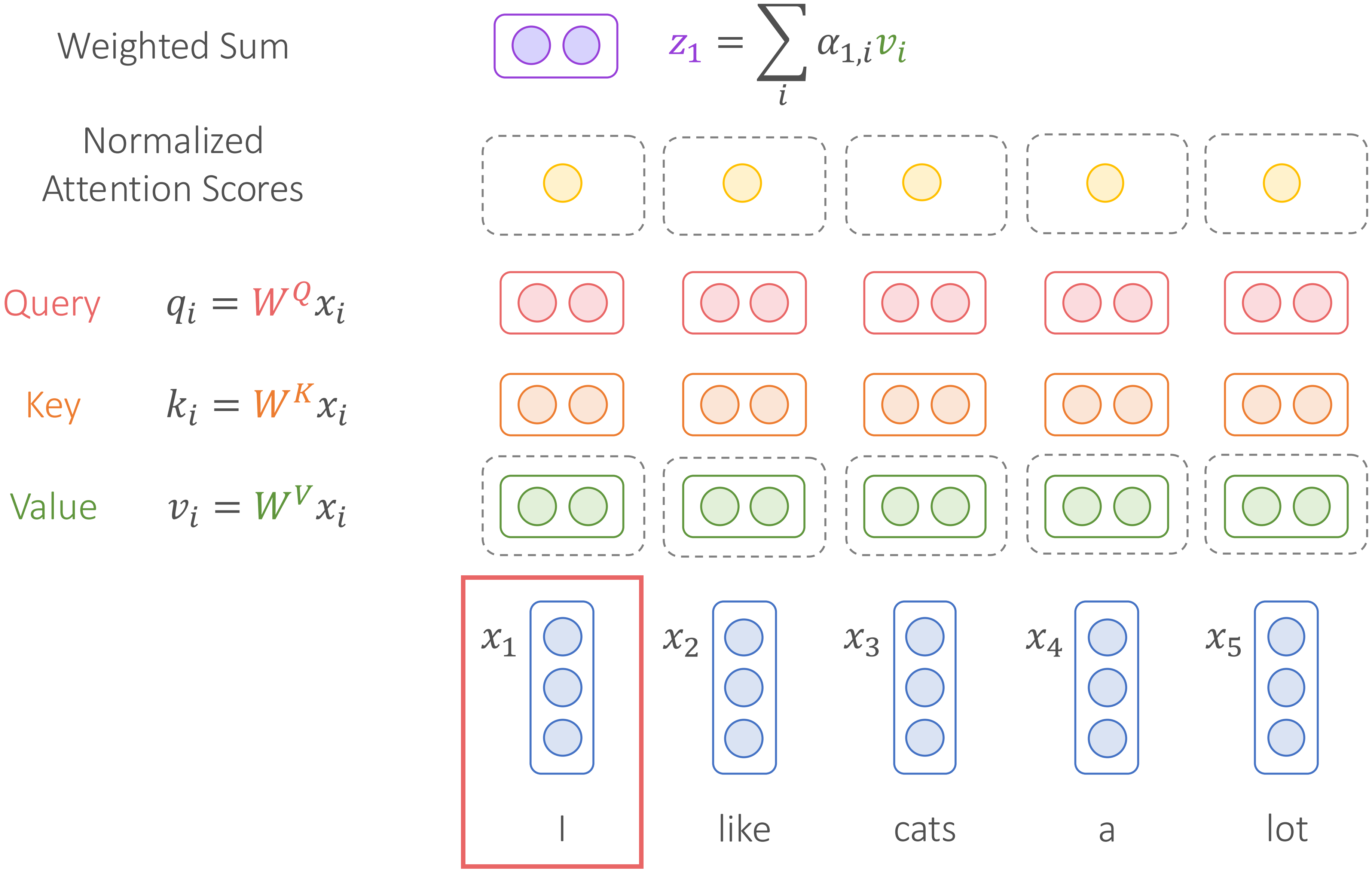
Query $q_i = W^Q x_i$

Key $k_i = W^K x_i$

Value $v_i = W^V x_i$

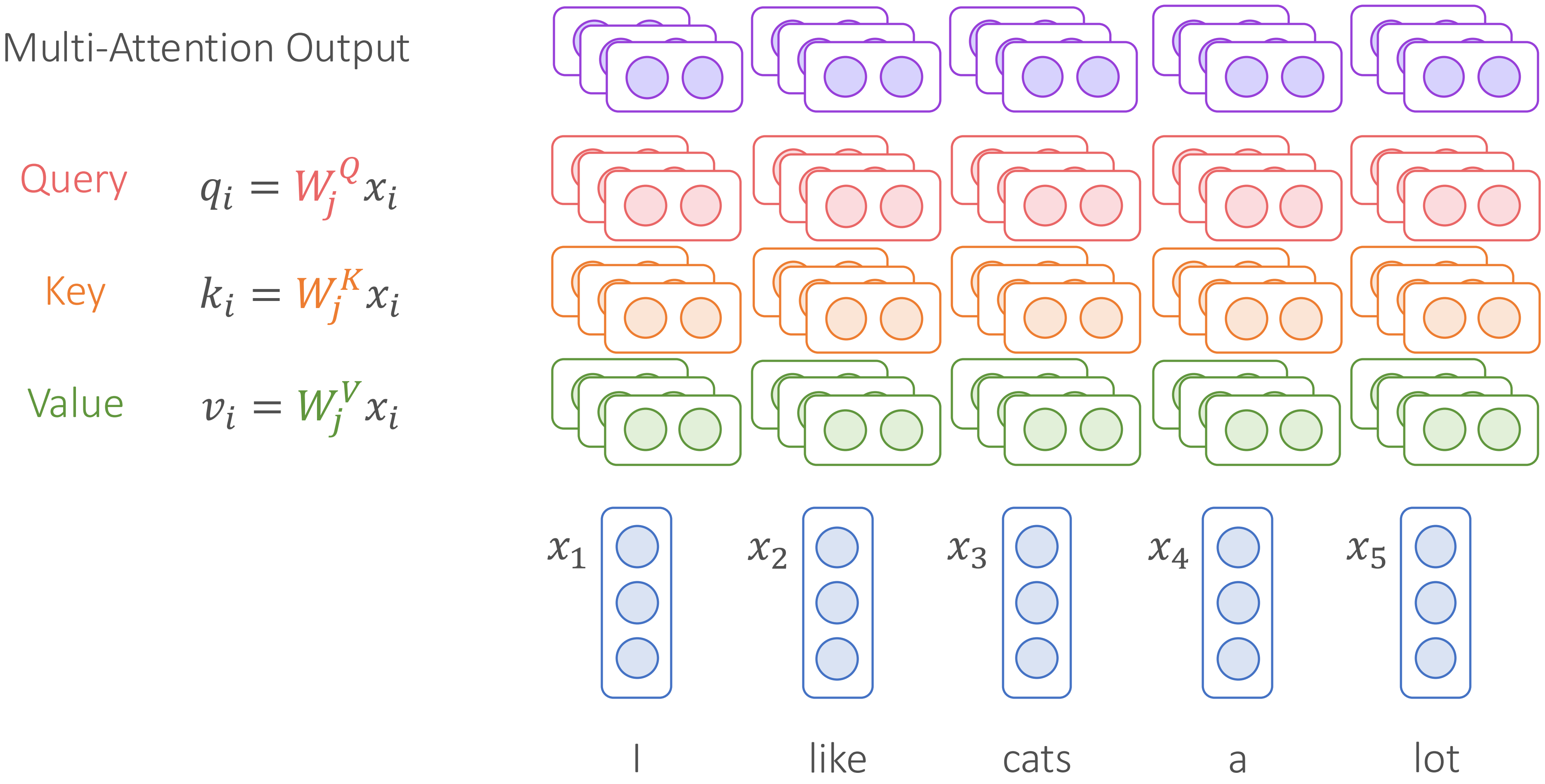


Recap: Self-Attention



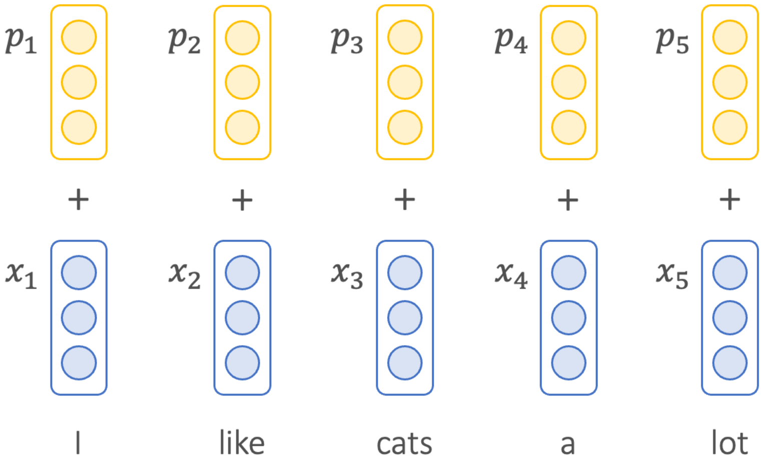
Recap: Multi-Head Attention

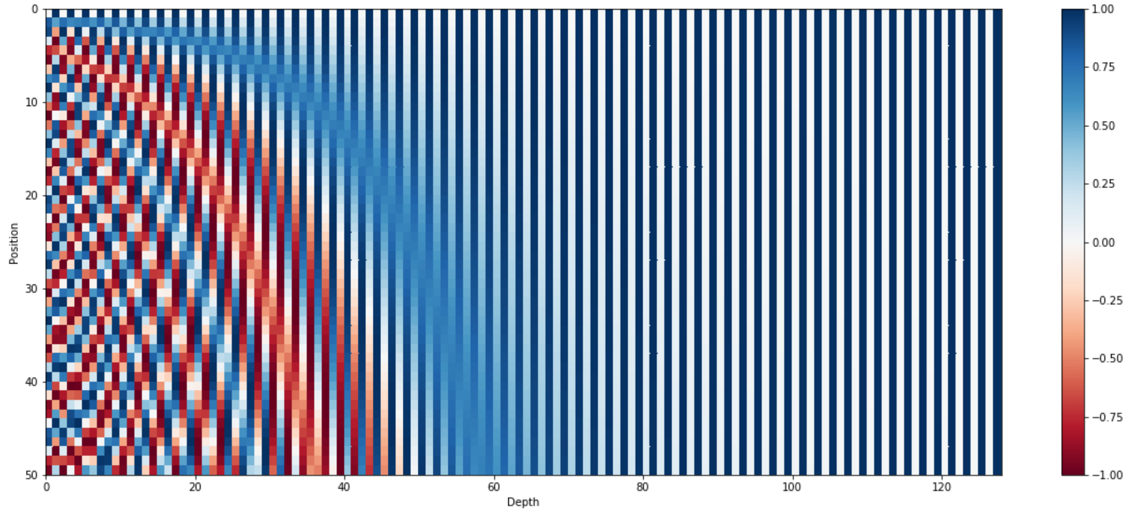
Each attention head focuses on different parts of understanding!

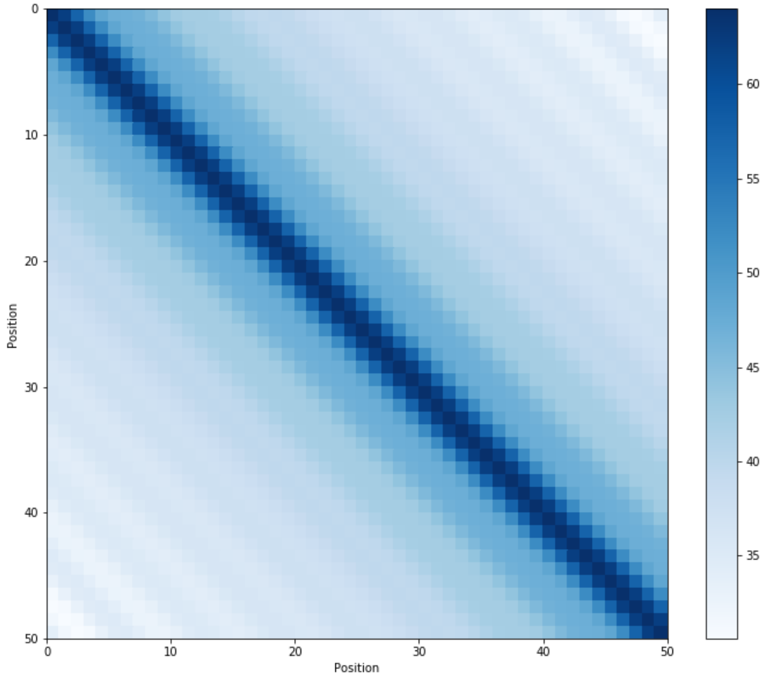


Recap: Positional Encoding

$$x_i \leftarrow x_i + PE_i$$

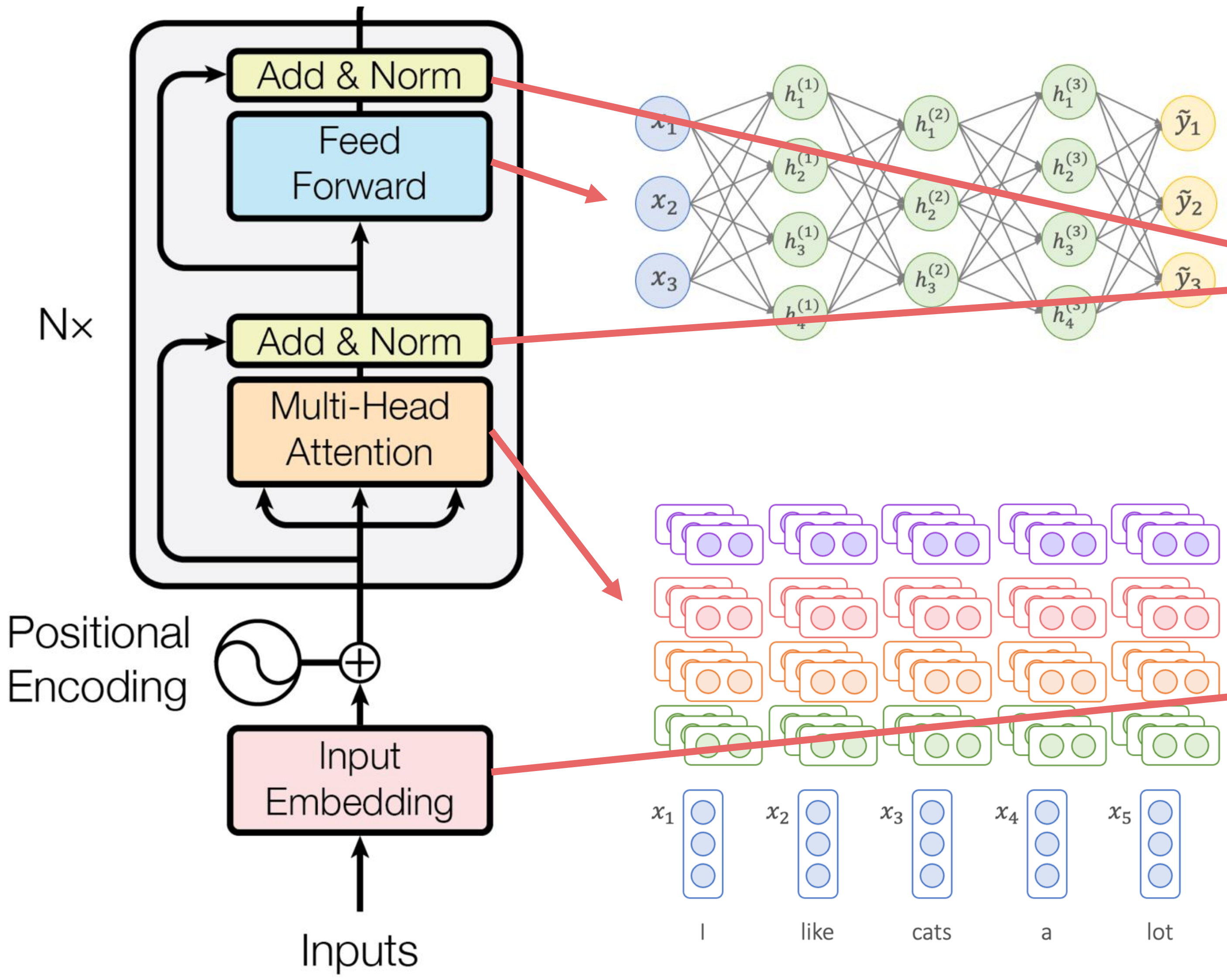






$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Recap: Transformer Encoder



$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

Residual connection (He et al., 2016)
Layer normalization (Ba et al., 2016)

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Transformer

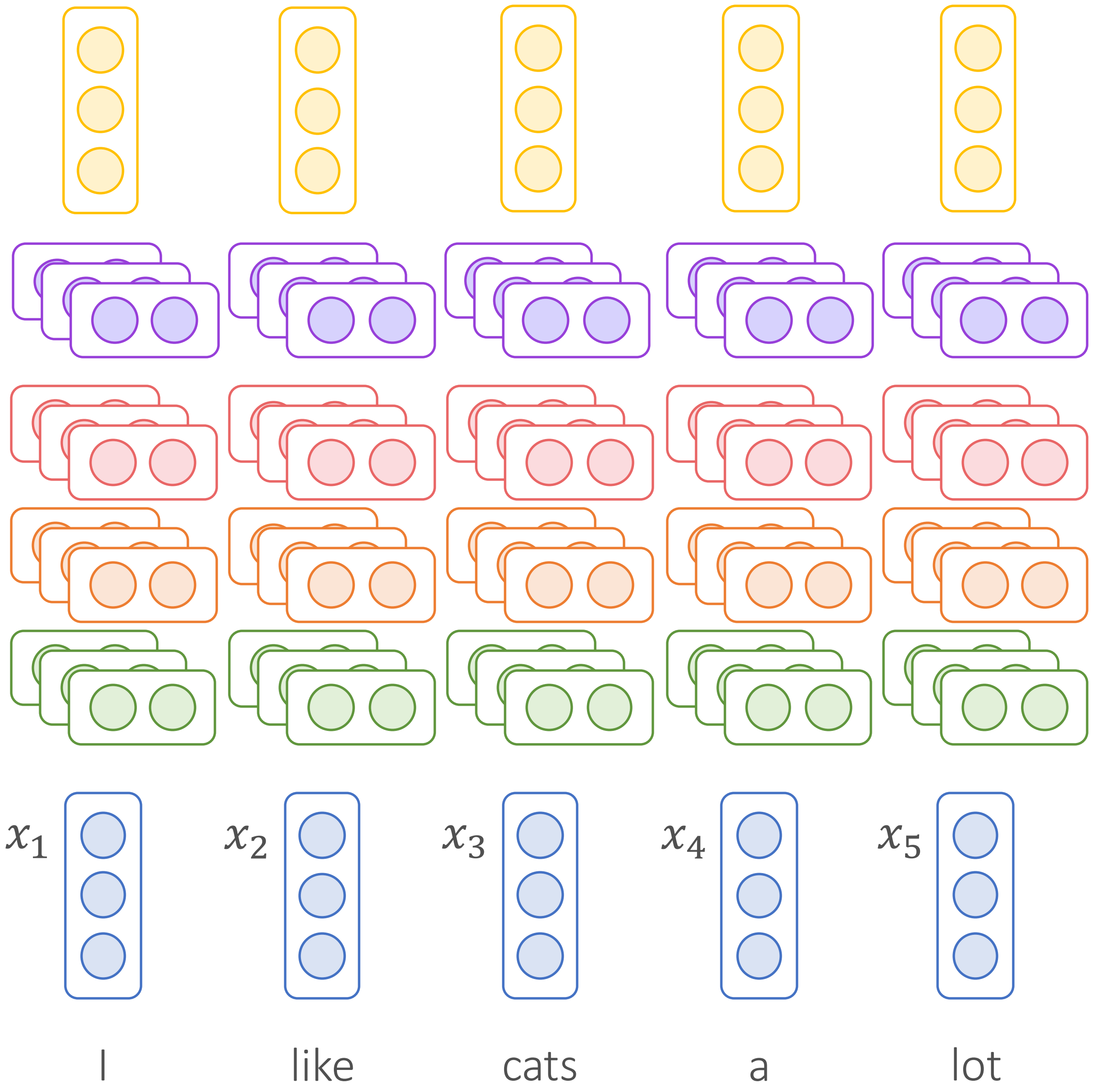
- **Non-recurrence:** easy to parallelize
- **Multi-head attention:** capture different aspects by interacting between words
- **Positional encoding:** capture the order information

Transformer as Token-Level Encoder

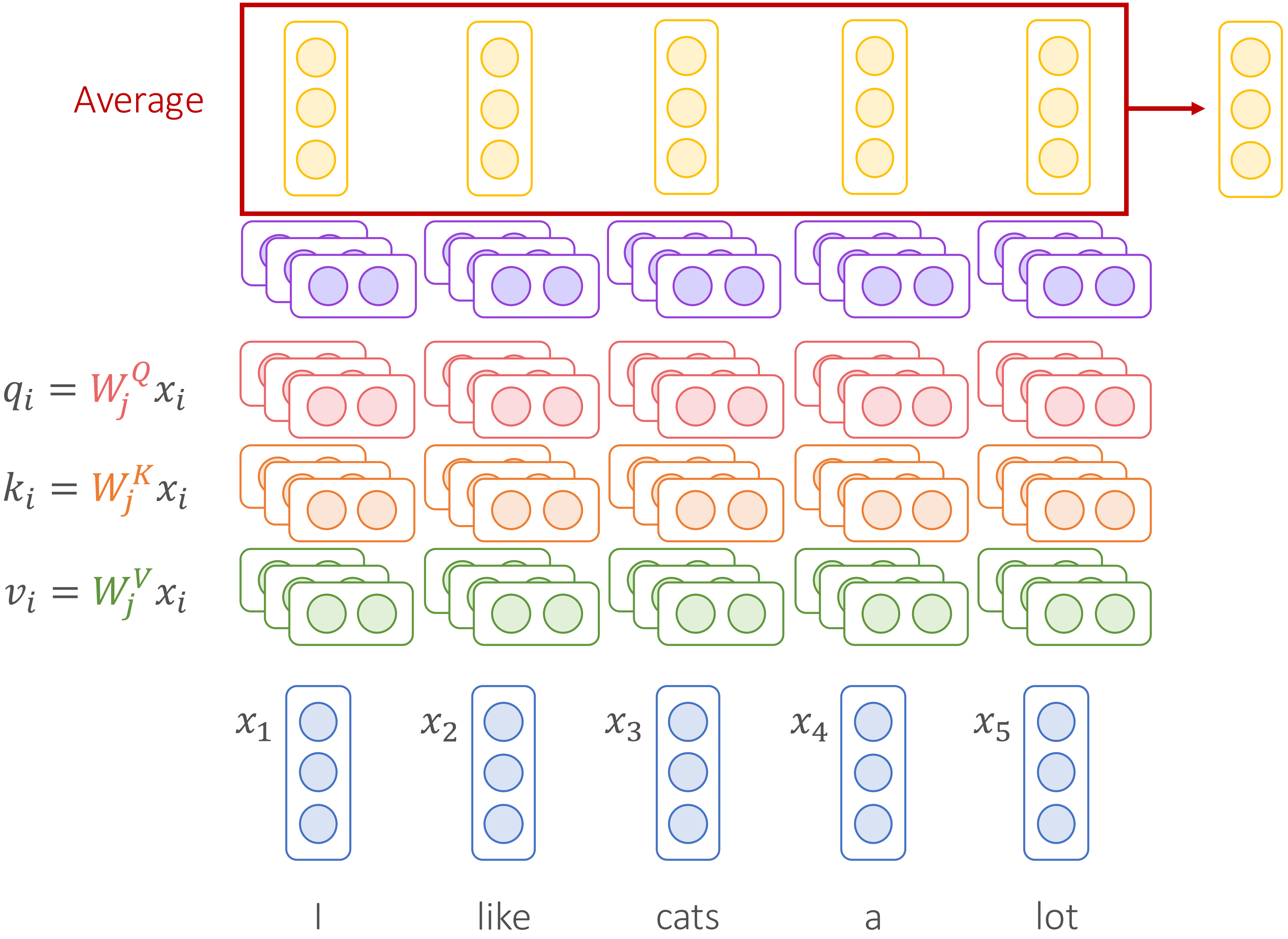
$q_i = W_j^Q x_i$

$k_i = W_j^K x_i$

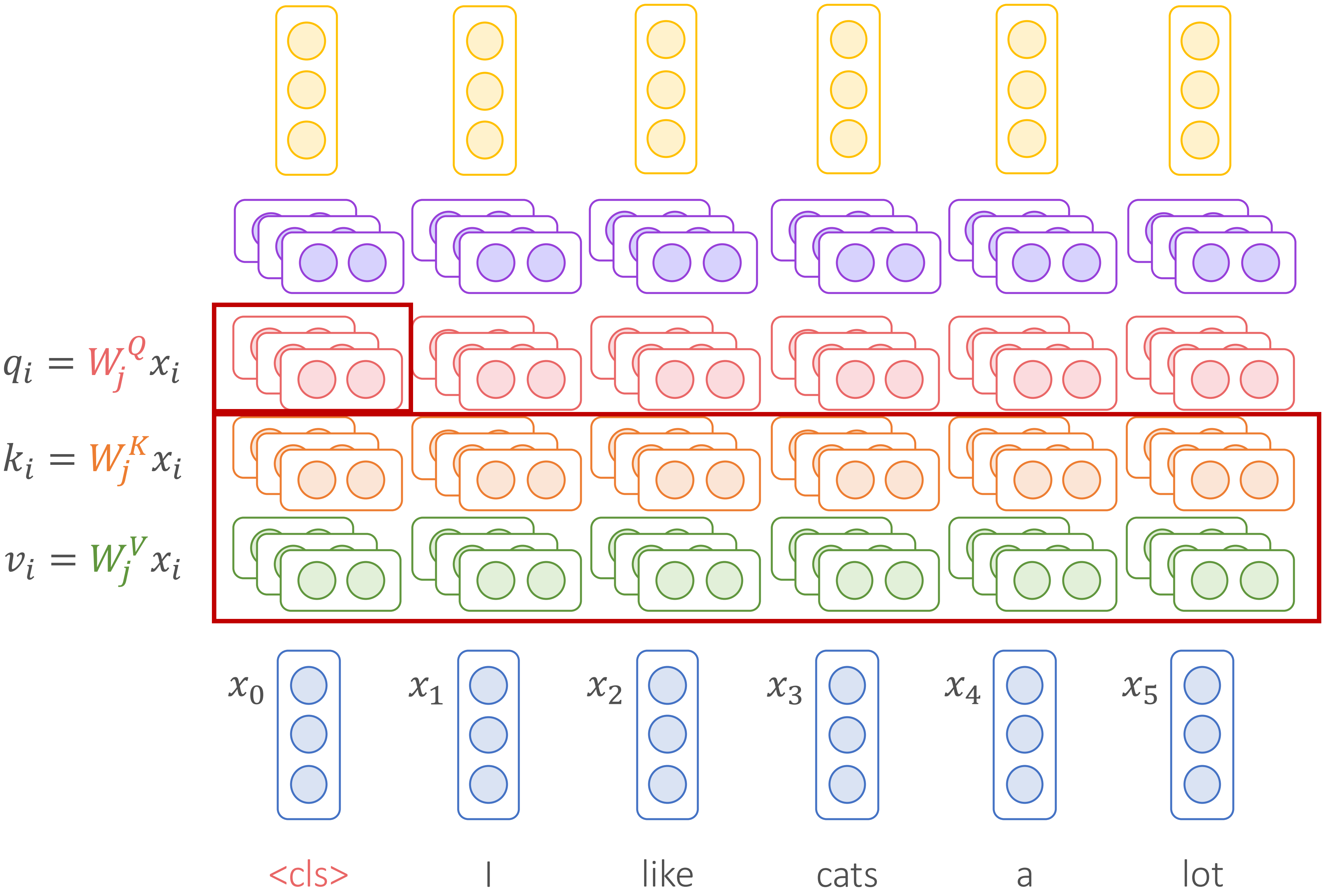
$v_i = W_j^V x_i$



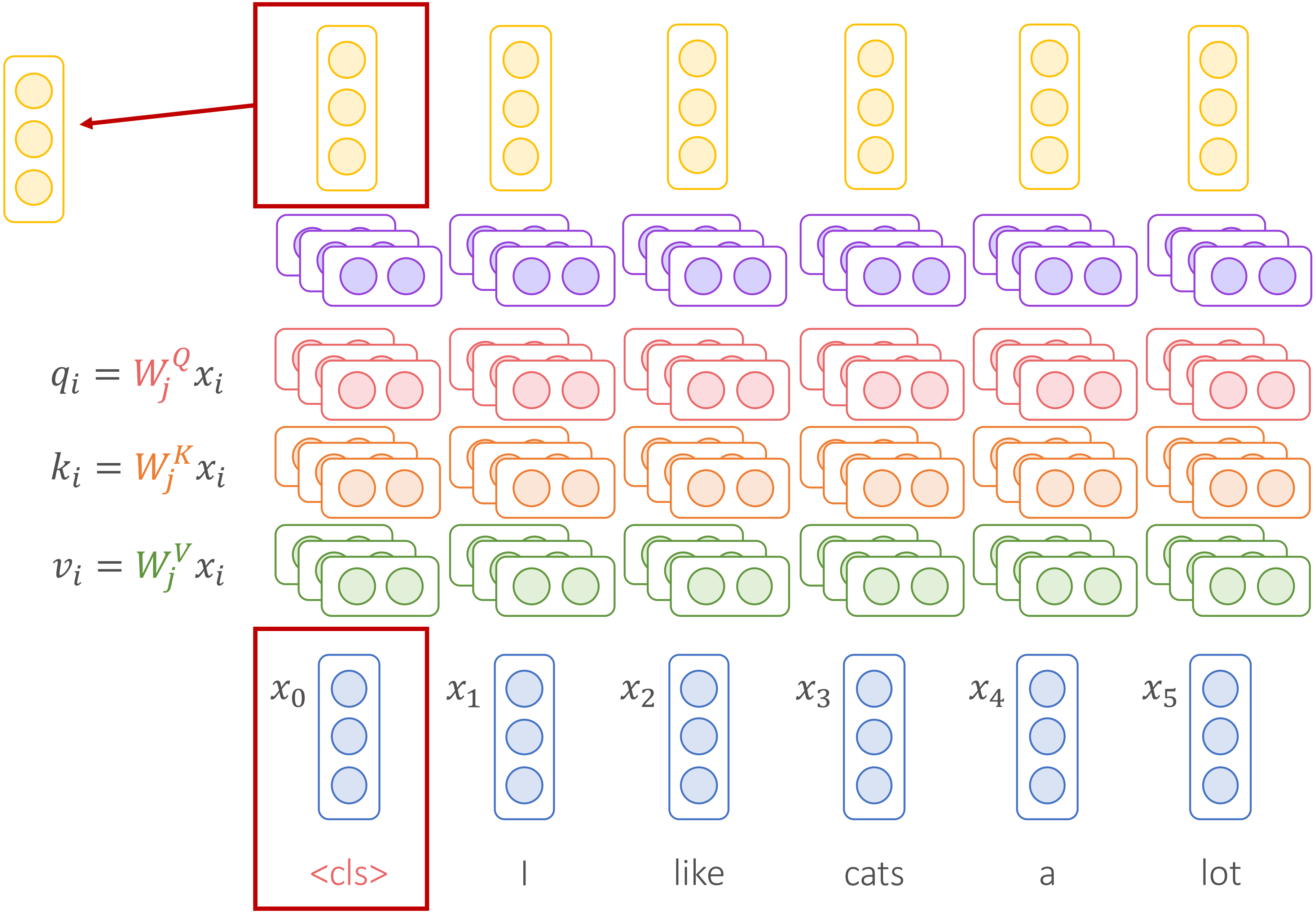
Transformer as Sentence-Level Encoder



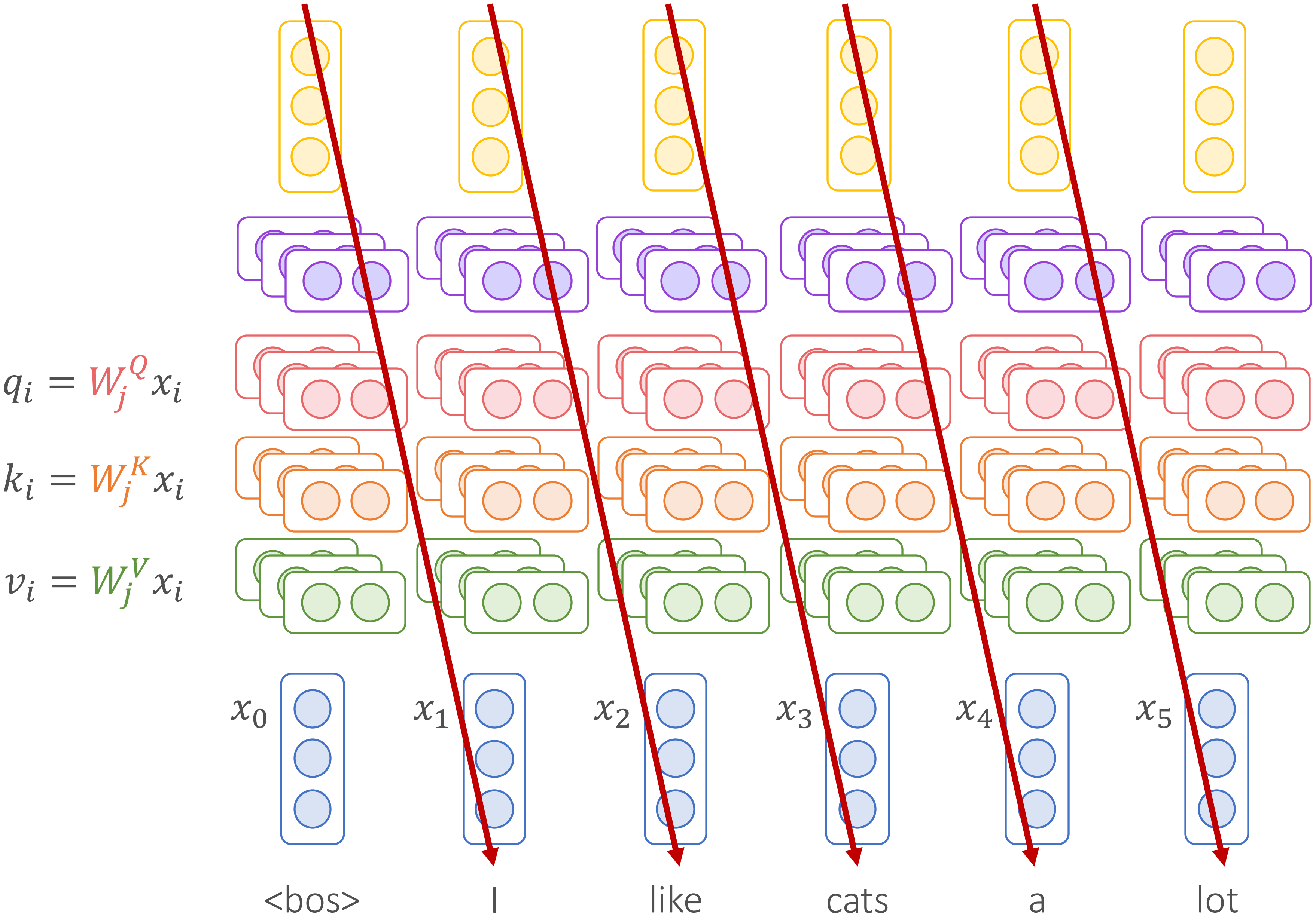
Transformer as Sentence-Level Encoder



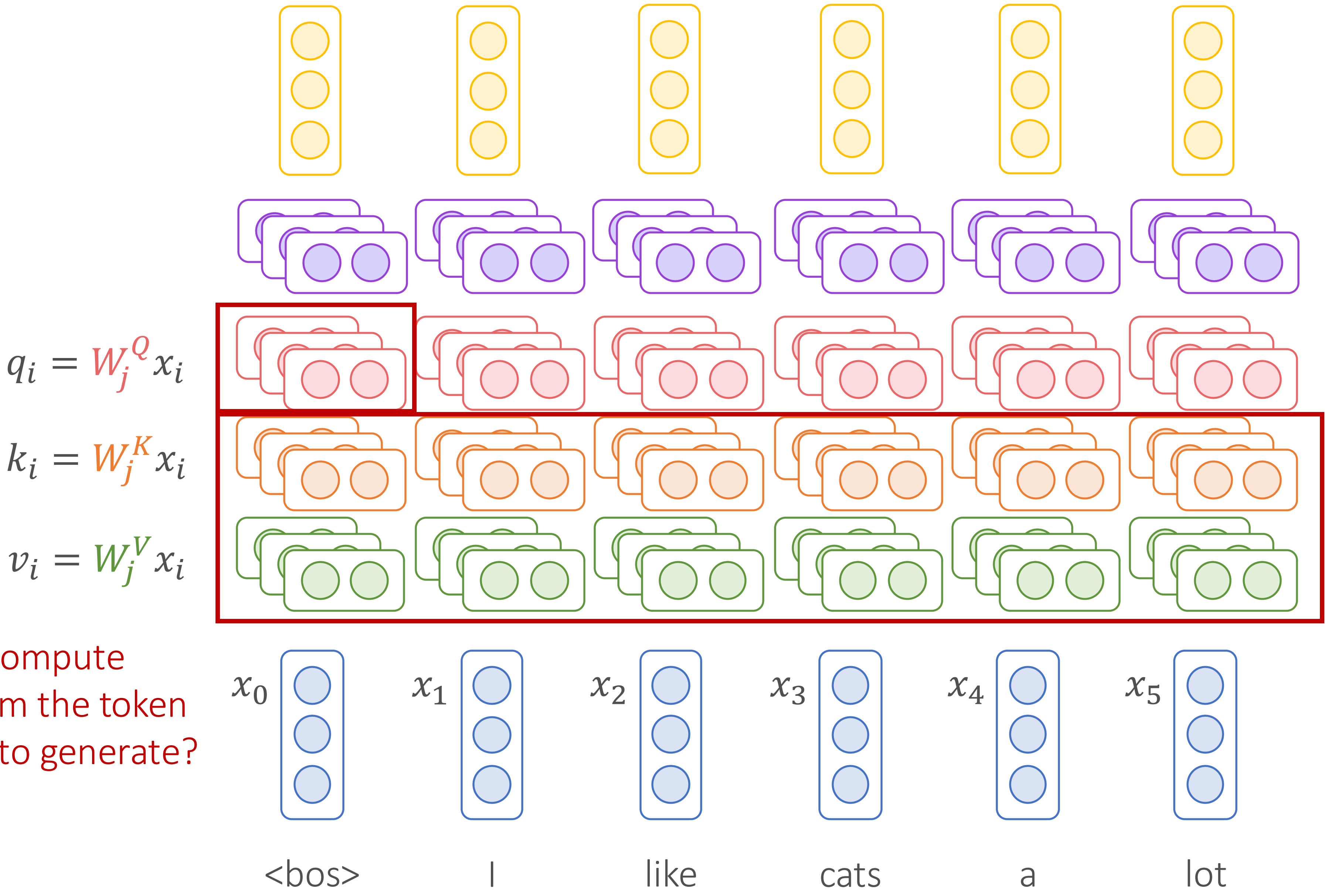
Transformer as Sentence-Level Encoder



Transformer as Decoder?

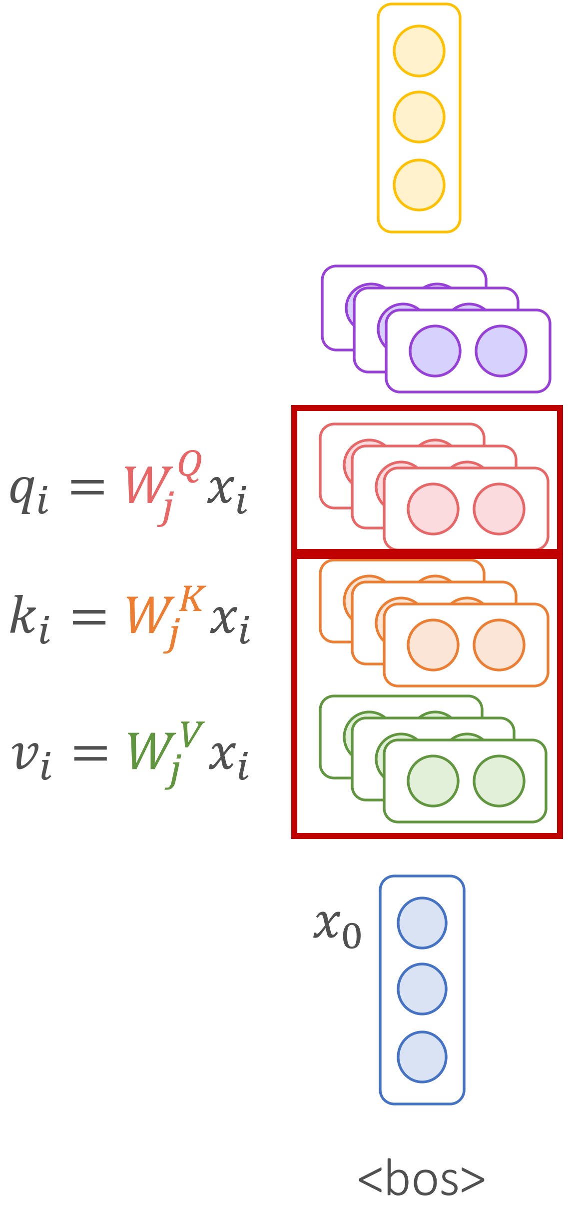


Transformer as Decoder?

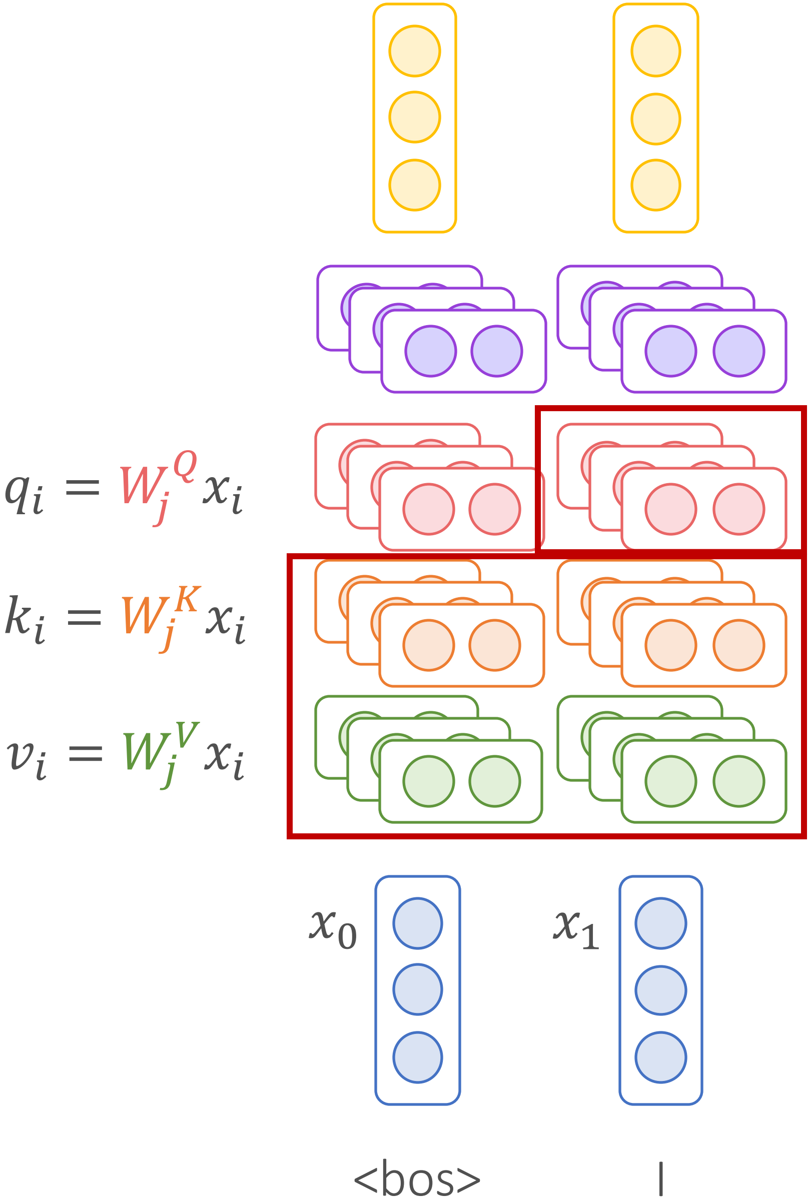


How to compute attention from the token we are going to generate?

Transformer Decoder



Transformer Decoder

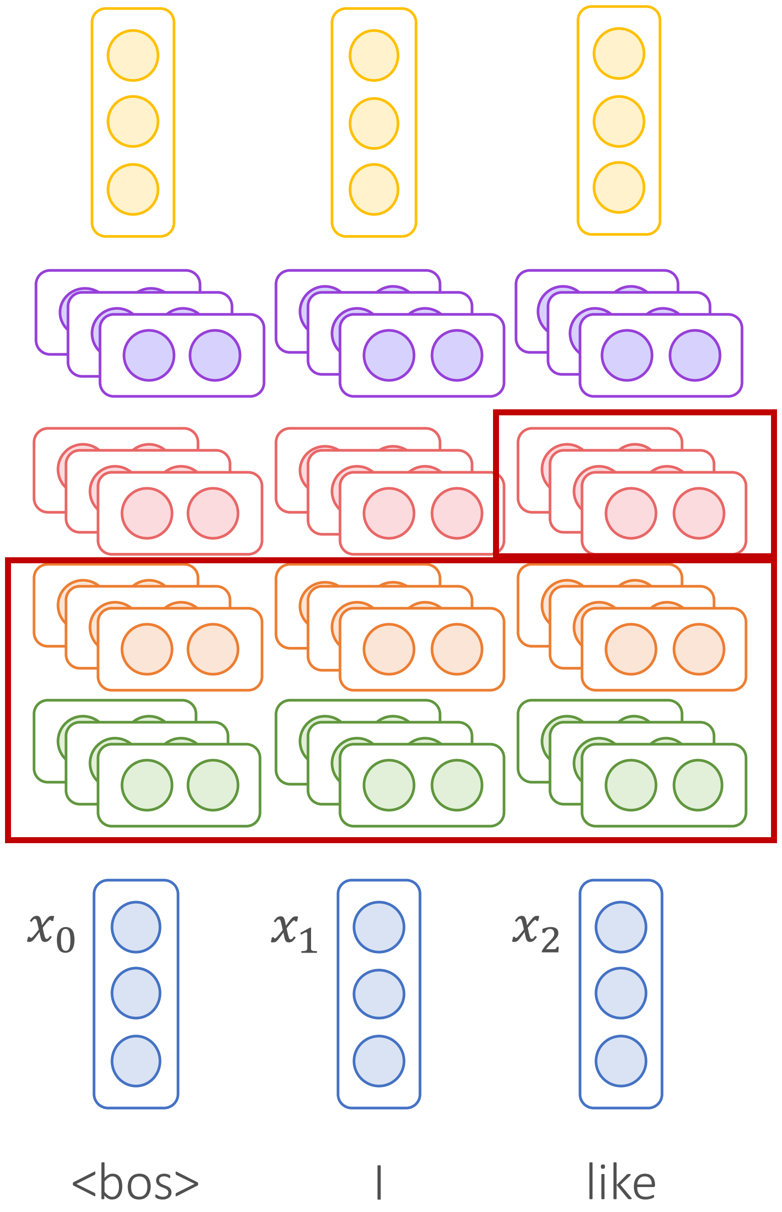


Transformer Decoder

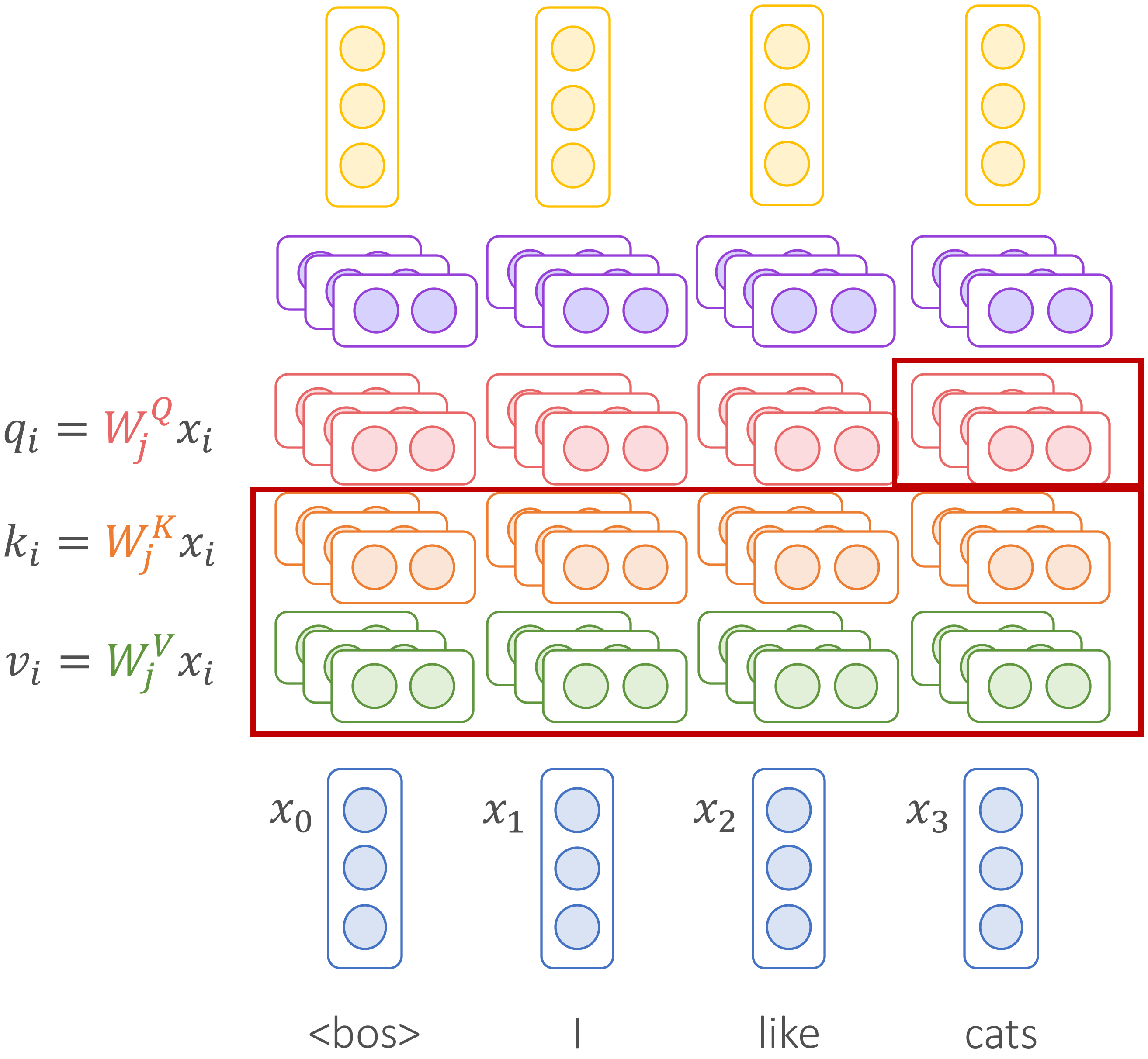
$$q_i = W_j^Q x_i$$

$$k_i = W_j^K x_i$$

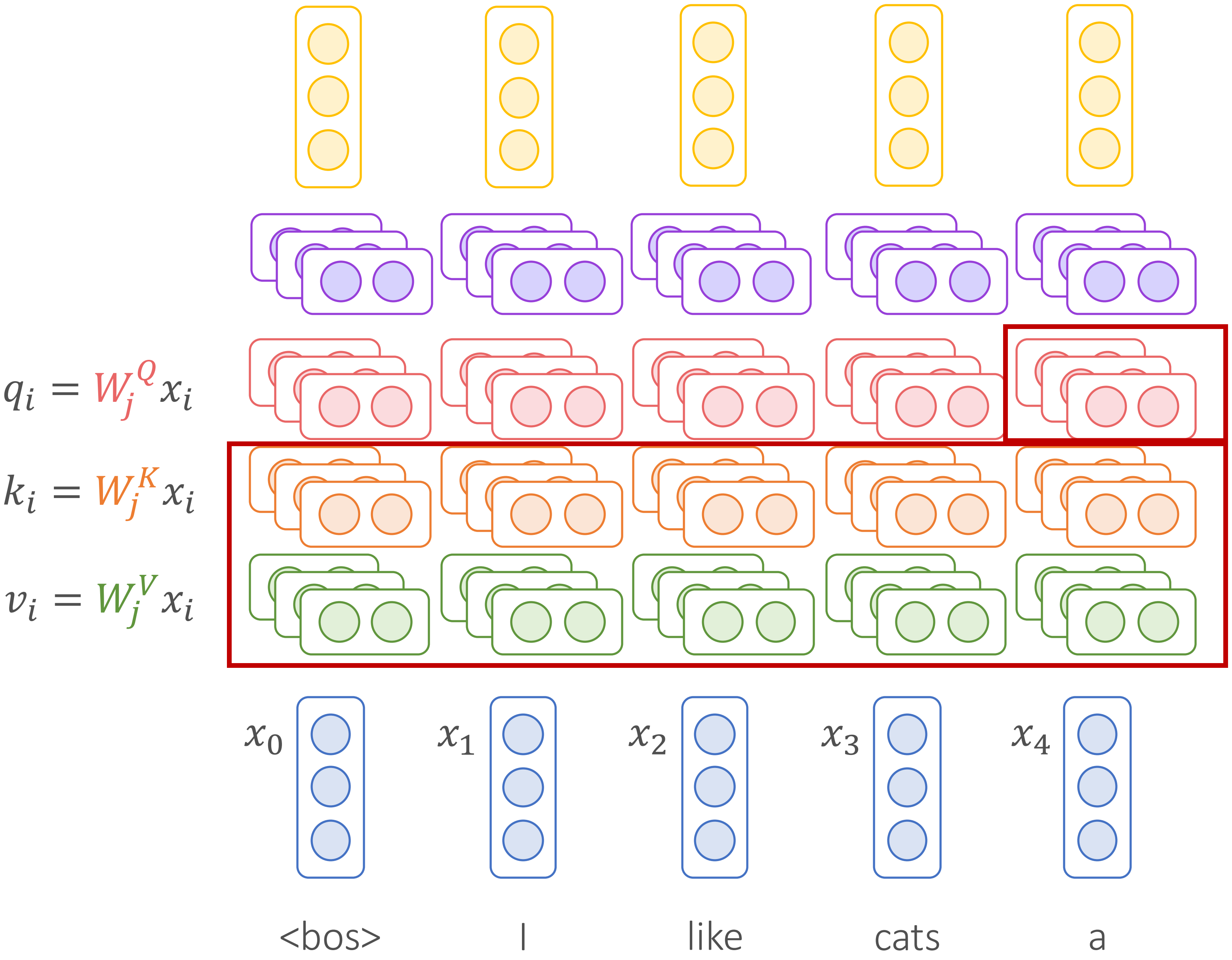
$$v_i = W_j^V x_i$$



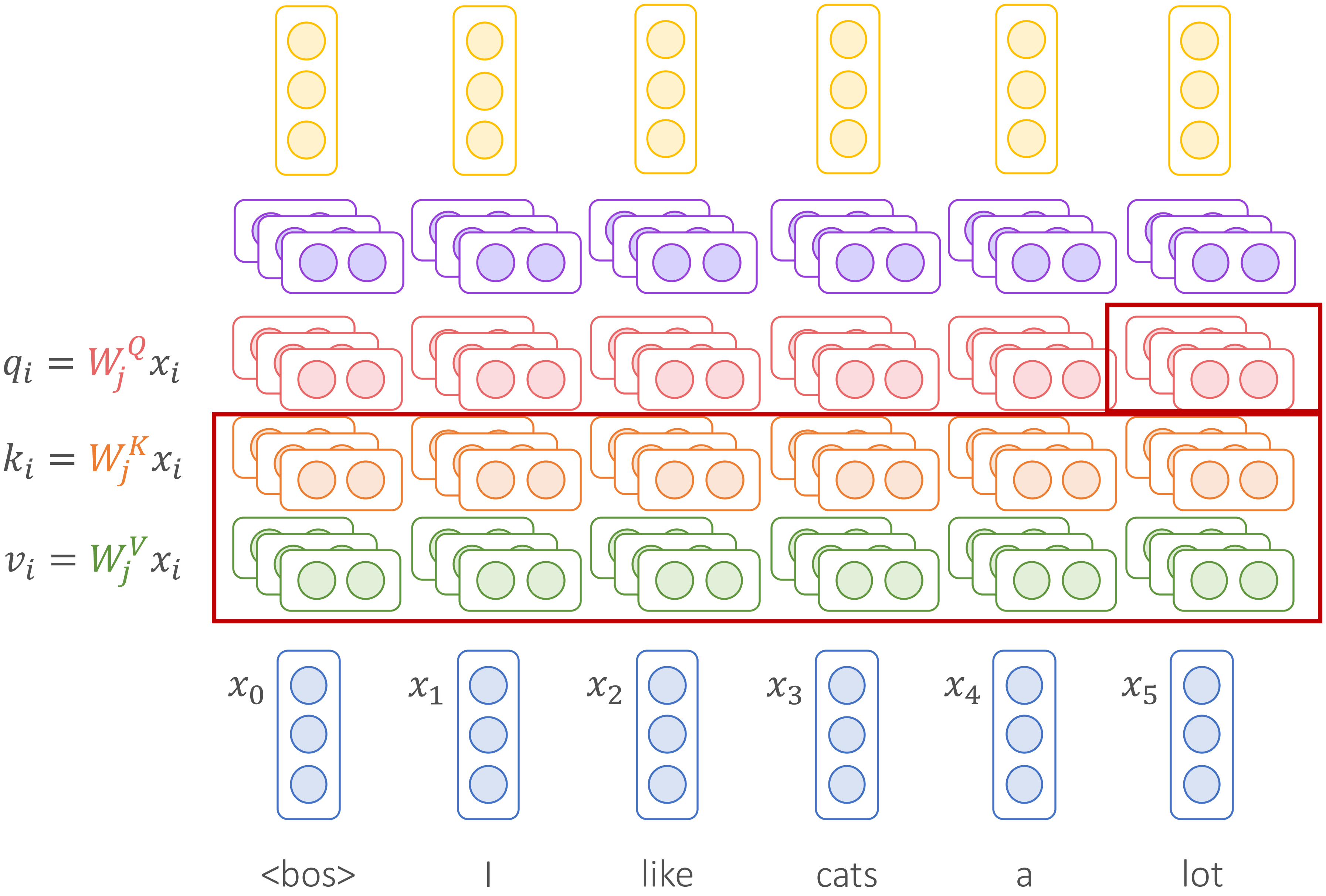
Transformer Decoder



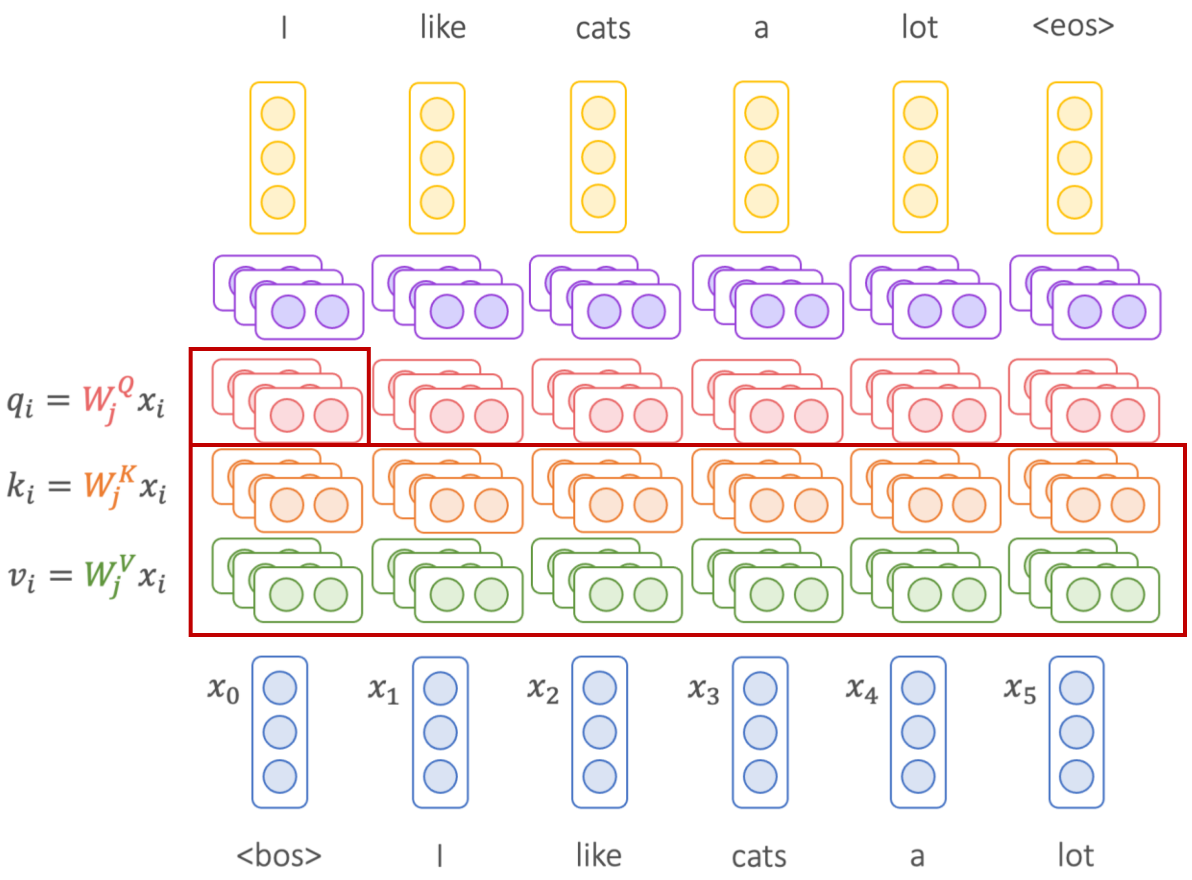
Transformer Decoder



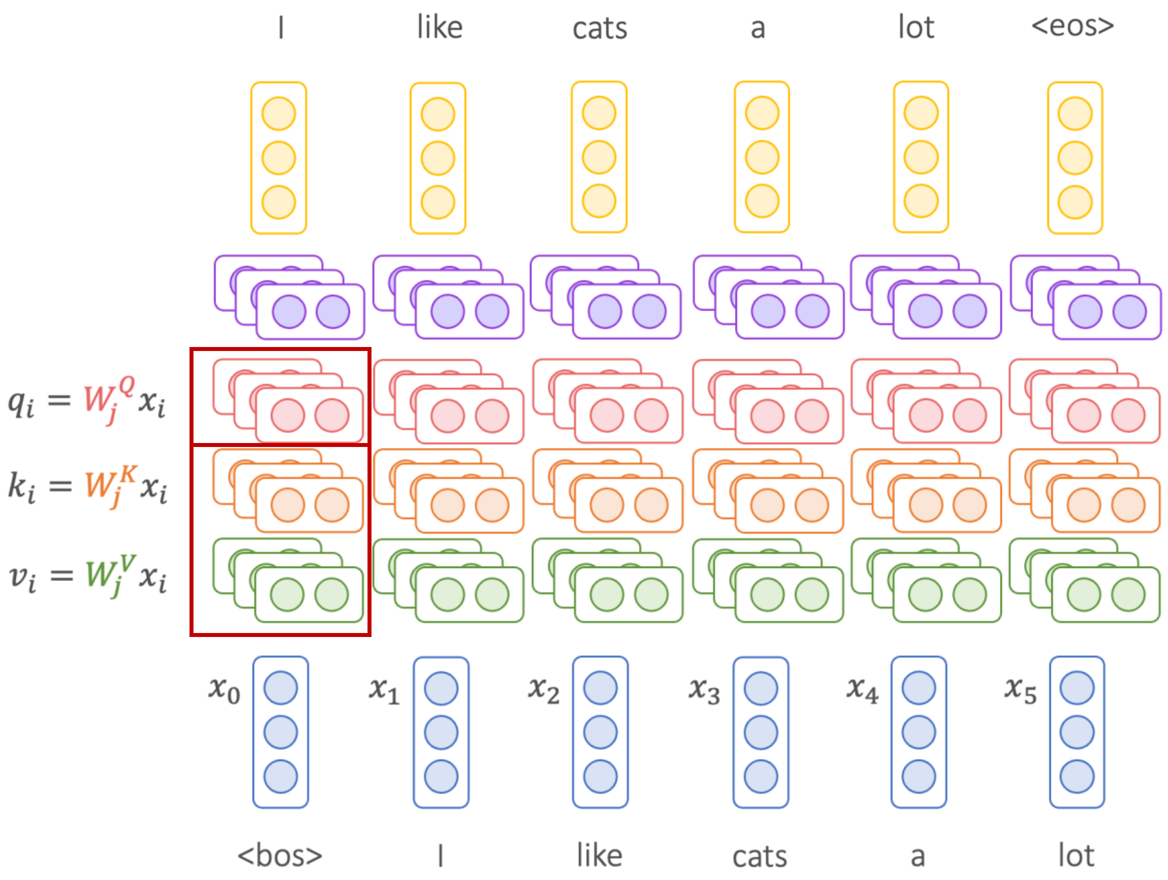
Transformer Decoder



Transformer Encoder vs. Transformer Decoder

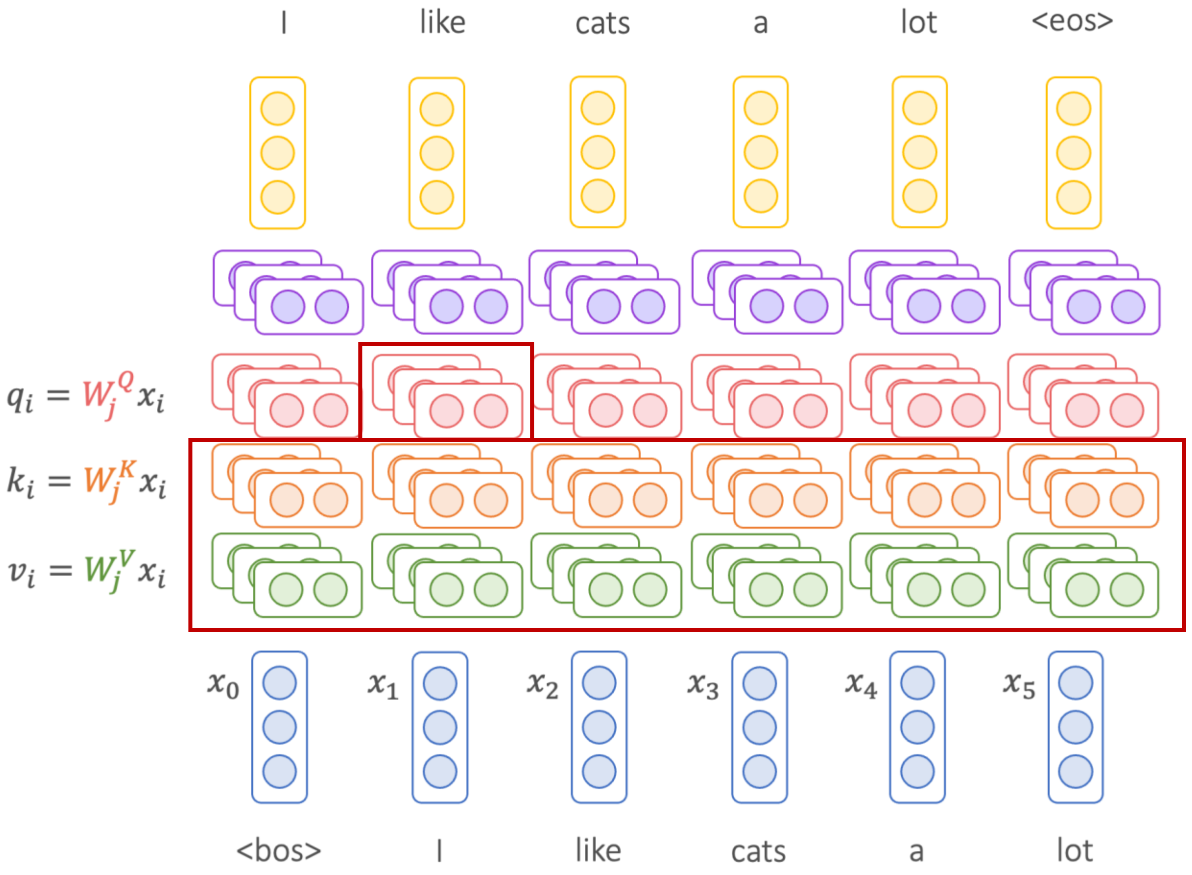


Transformer Encoder

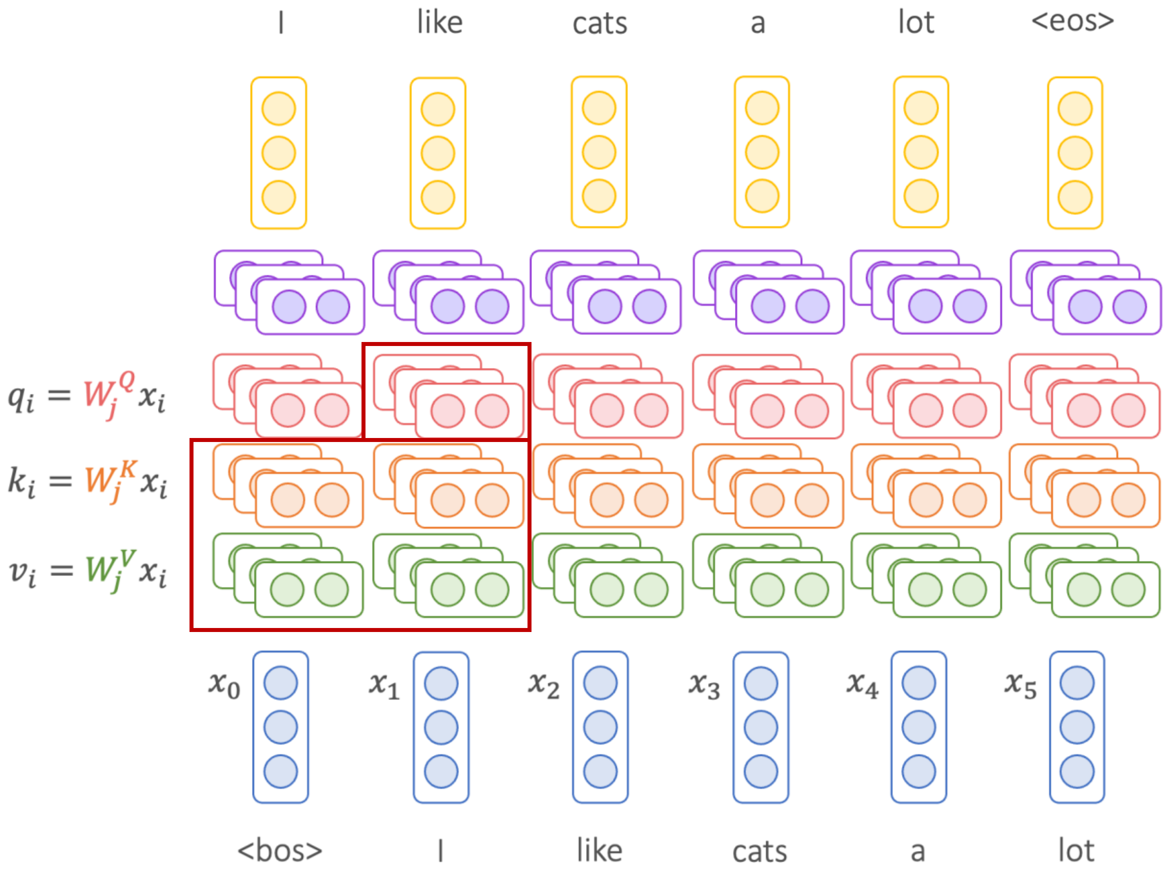


Transformer Decoder

Transformer Encoder vs. Transformer Decoder

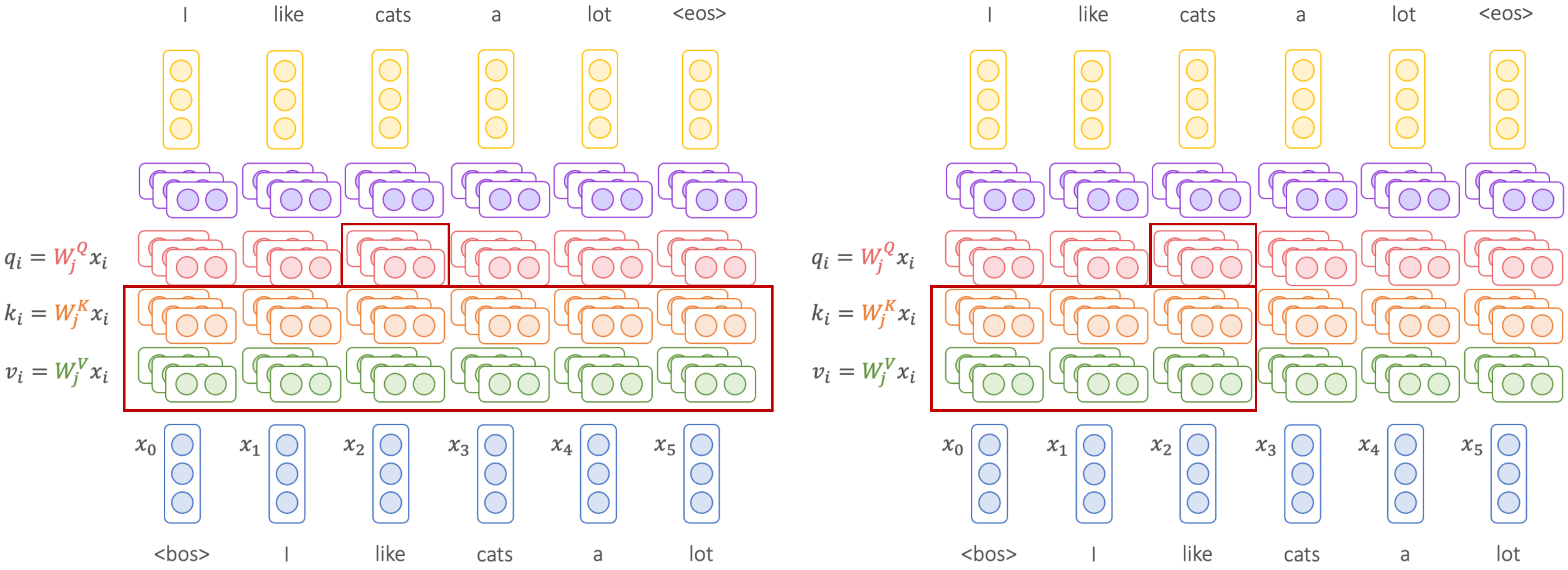


Transformer Encoder



Transformer Decoder

Transformer Encoder vs. Transformer Decoder



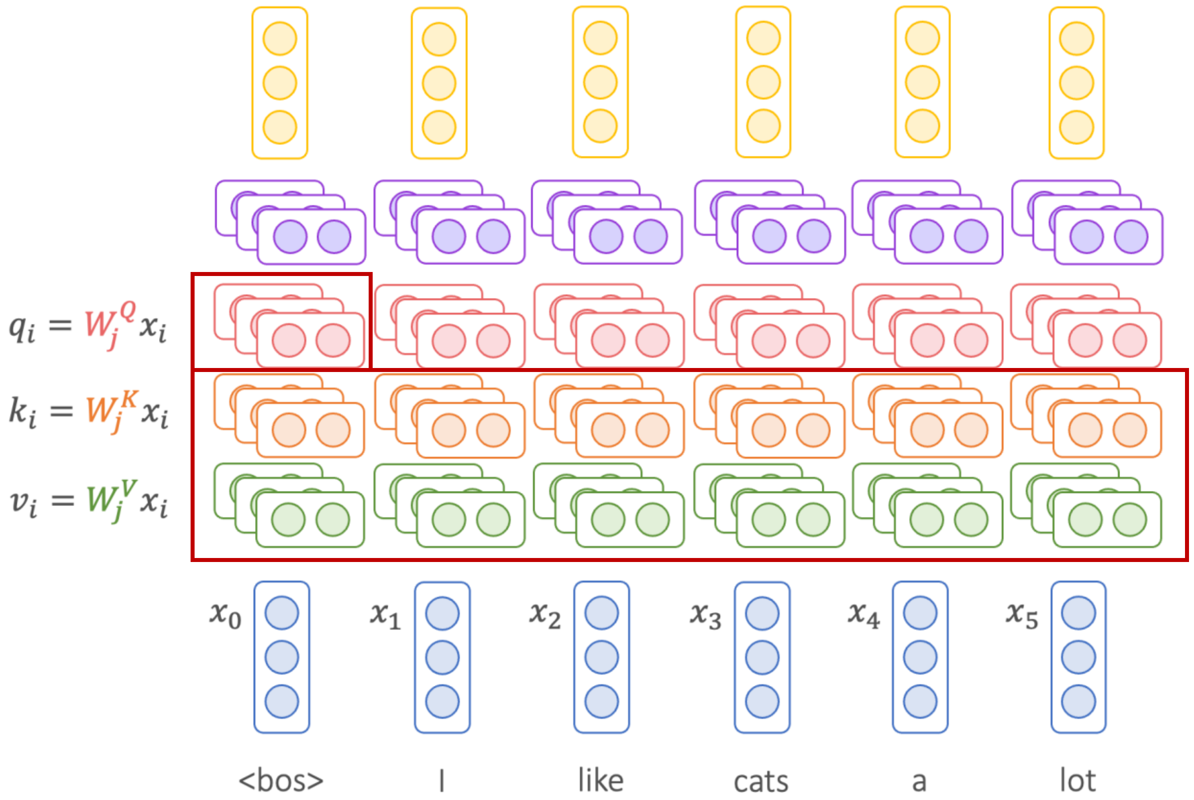
Transformer Encoder

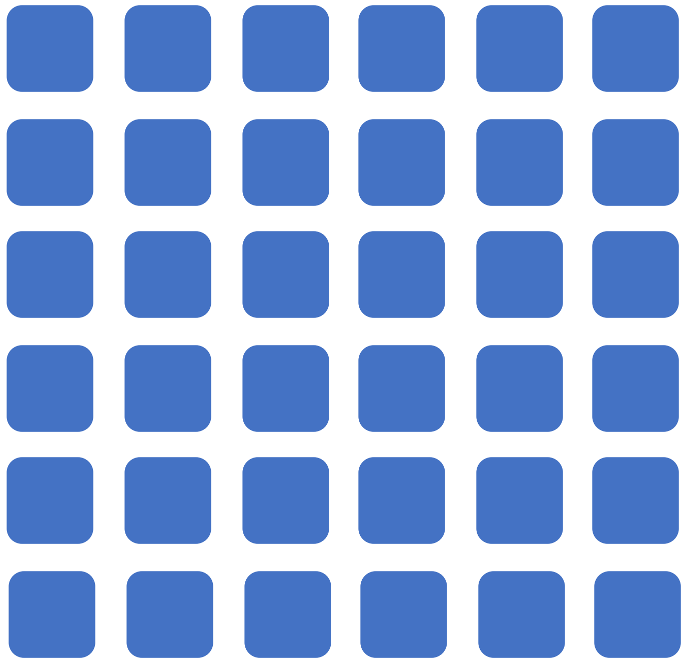
Transformer Decoder

Transformer Encoder vs. Transformer Decoder

- When computing attention for one word
  - Encoder: can see the words before and after this word
  - Decoder: can see the words only before this word

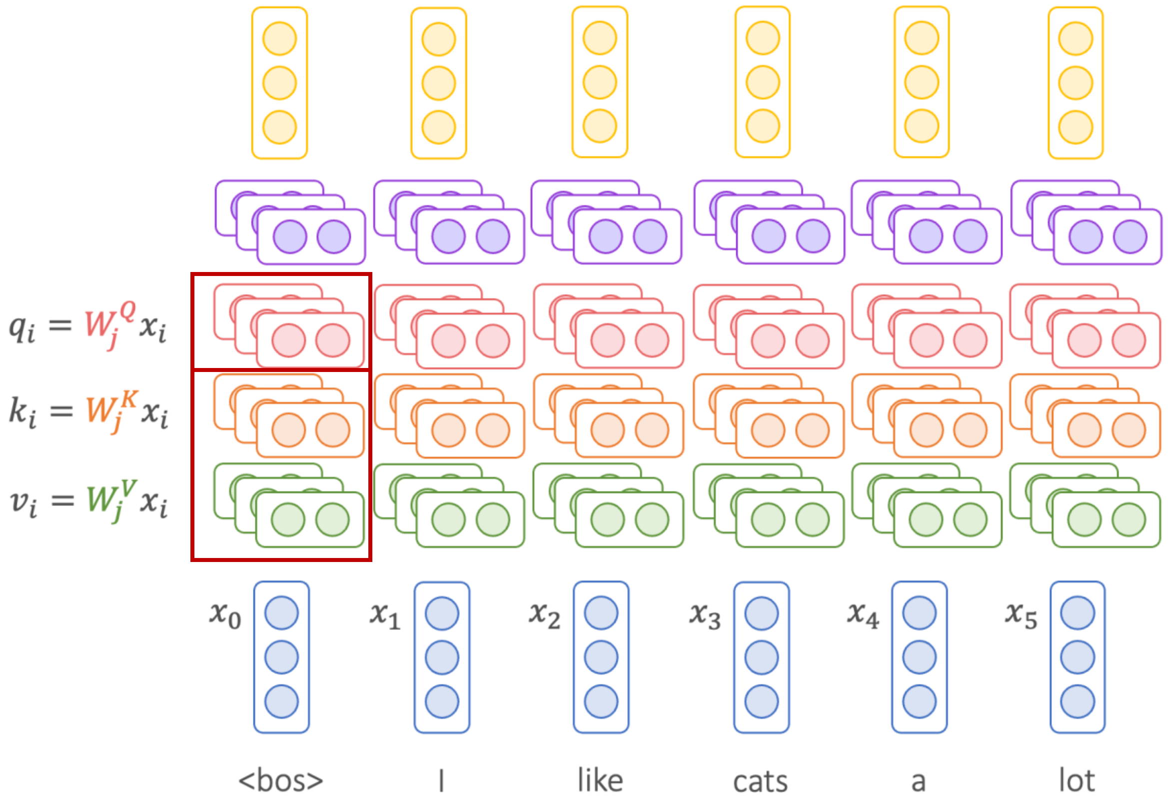
Masked Attention for Transformer Encoder

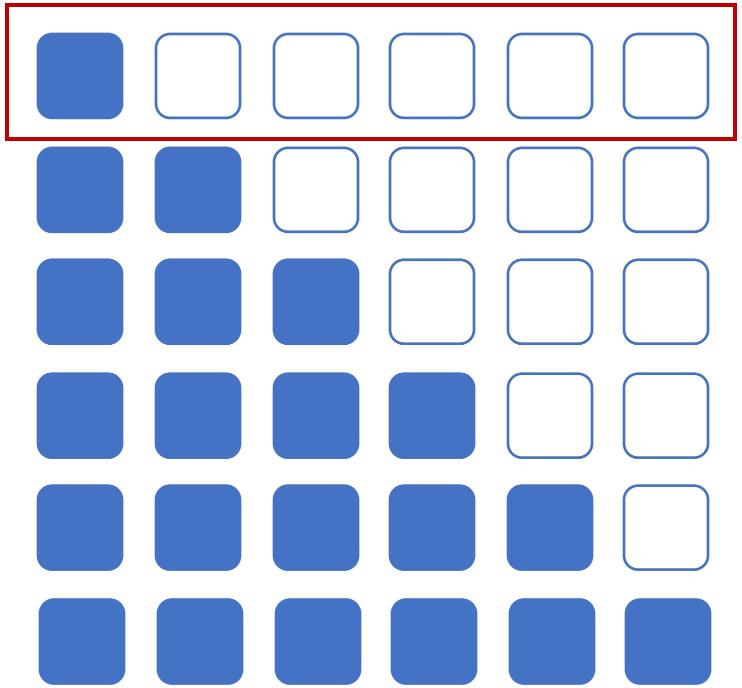




No Masking

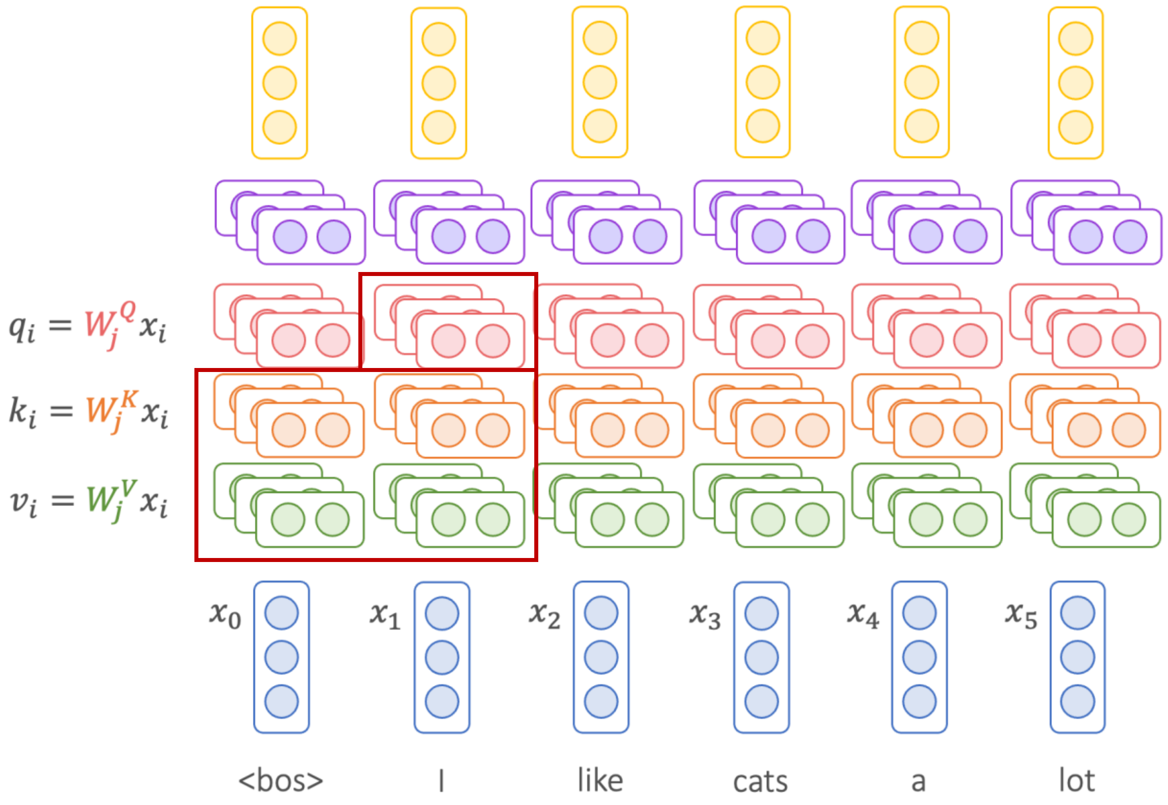
Masked Attention for Transformer Decoder

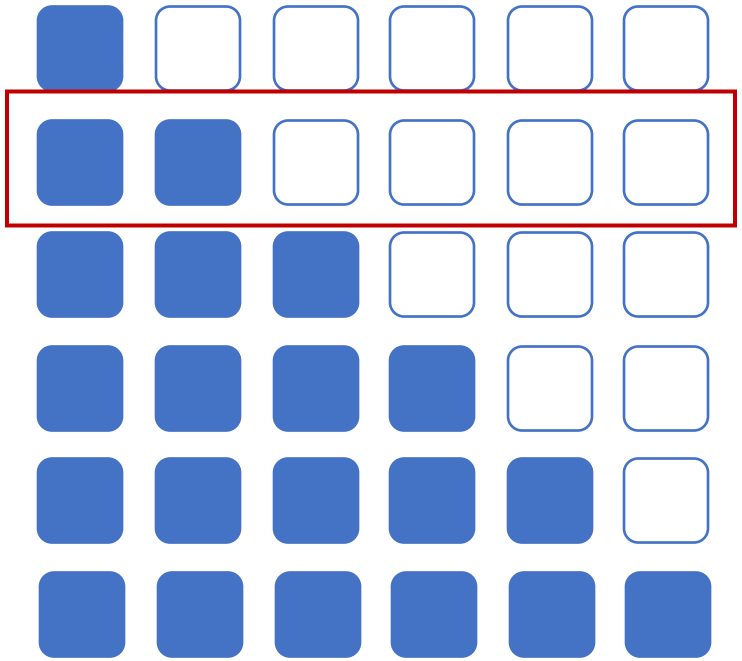




Causal Masking

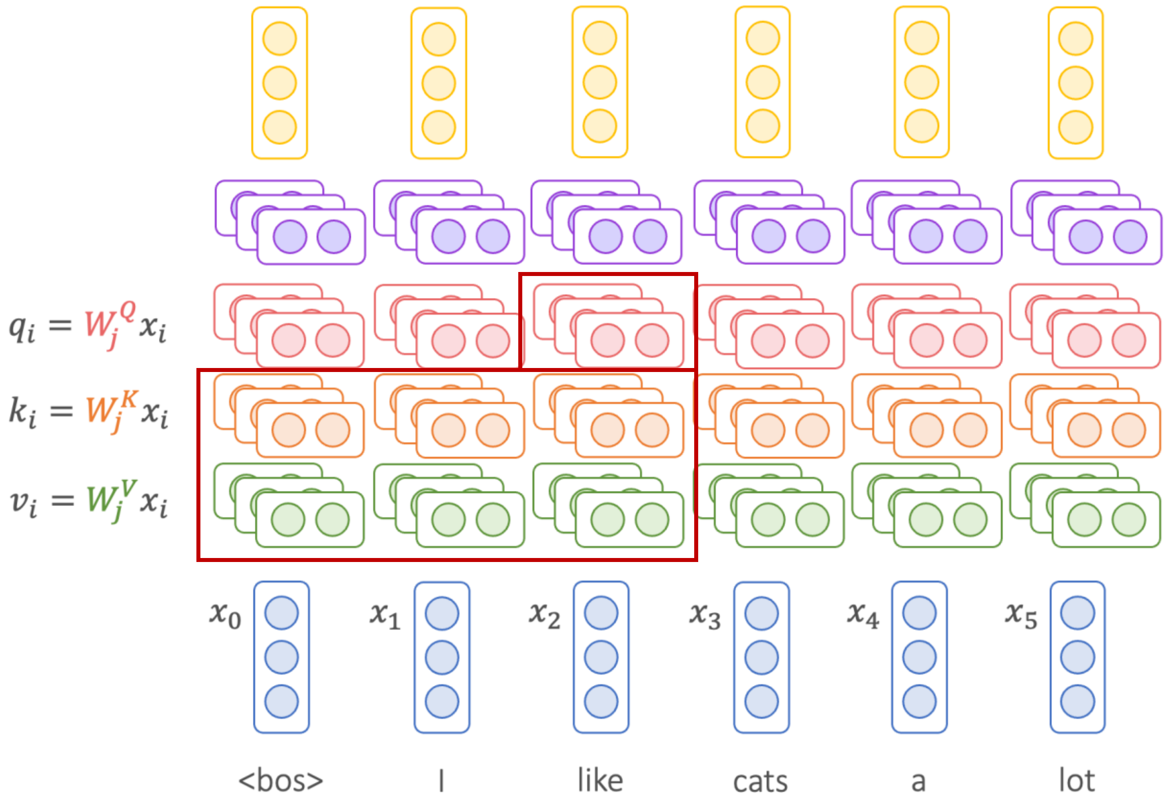
Masked Attention for Transformer Decoder

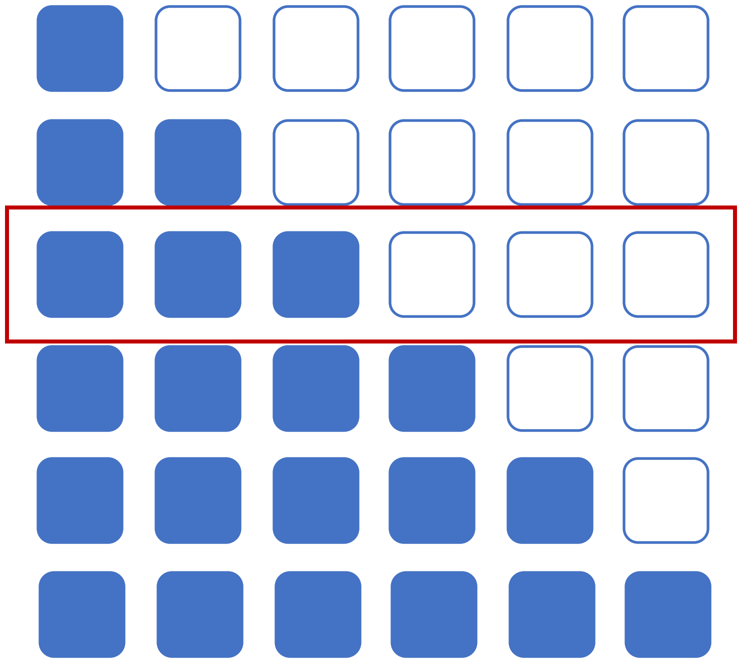




Causal Masking

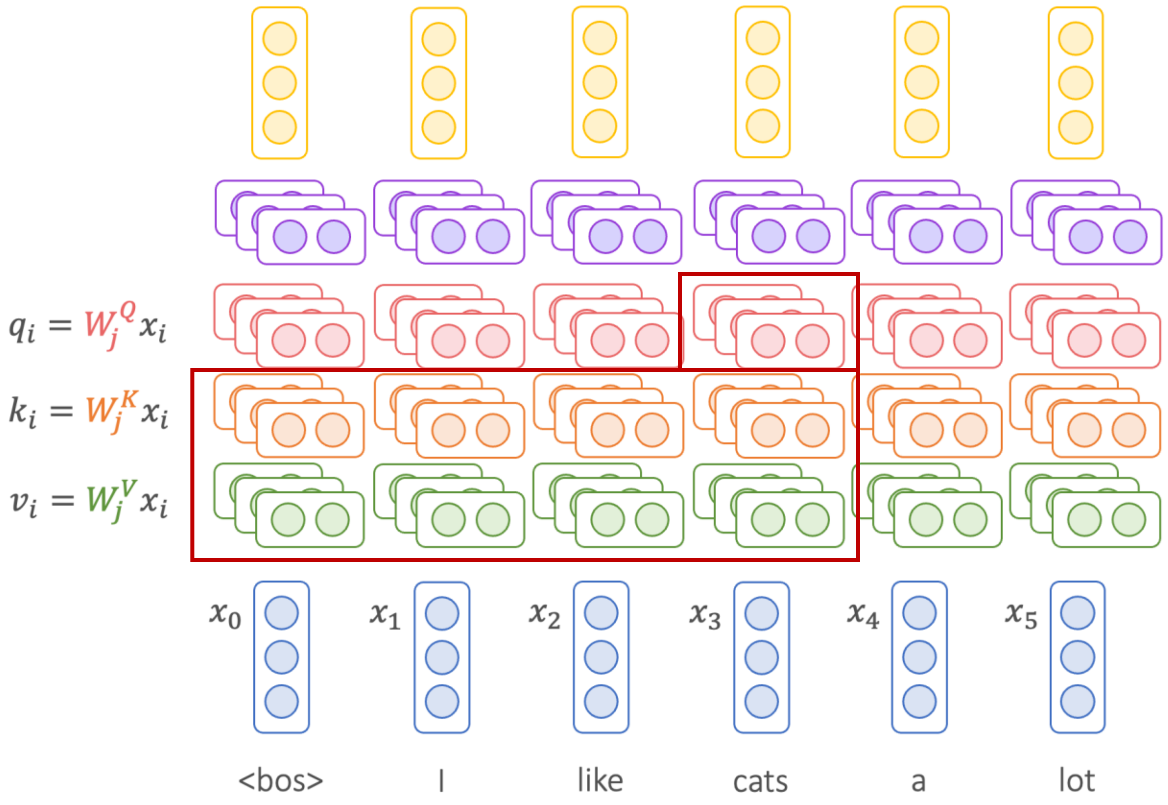
Masked Attention for Transformer Decoder

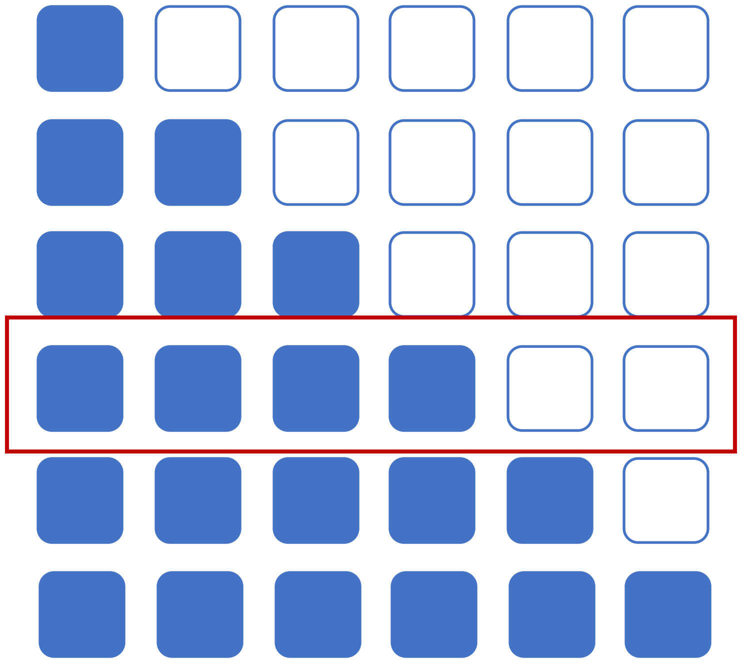




Causal Masking

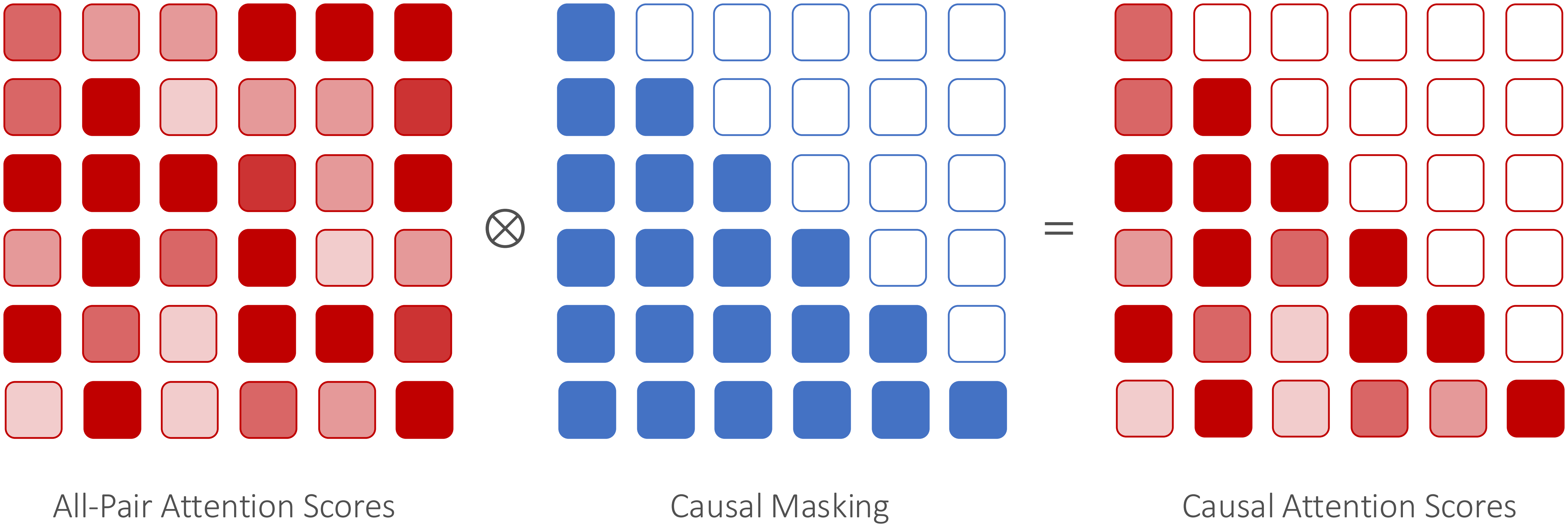
Masked Attention for Transformer Decoder



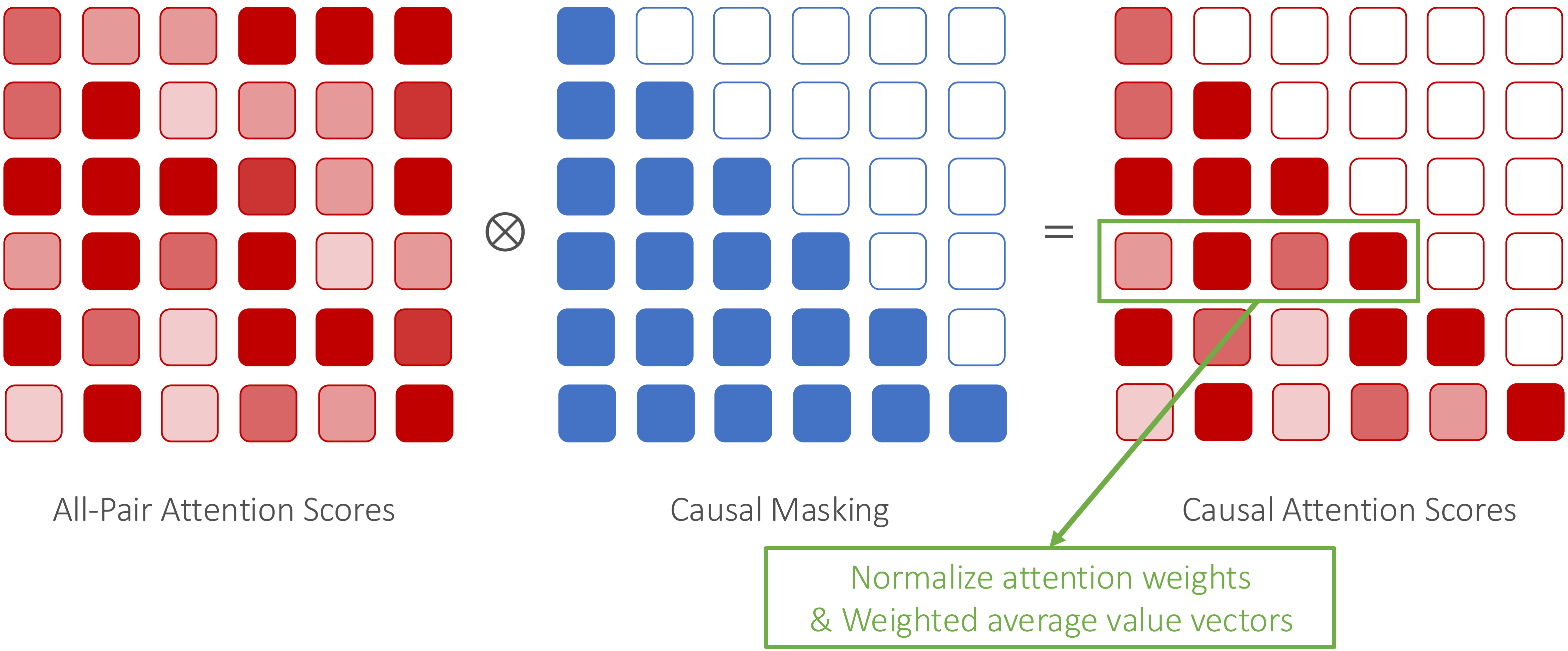


Causal Masking

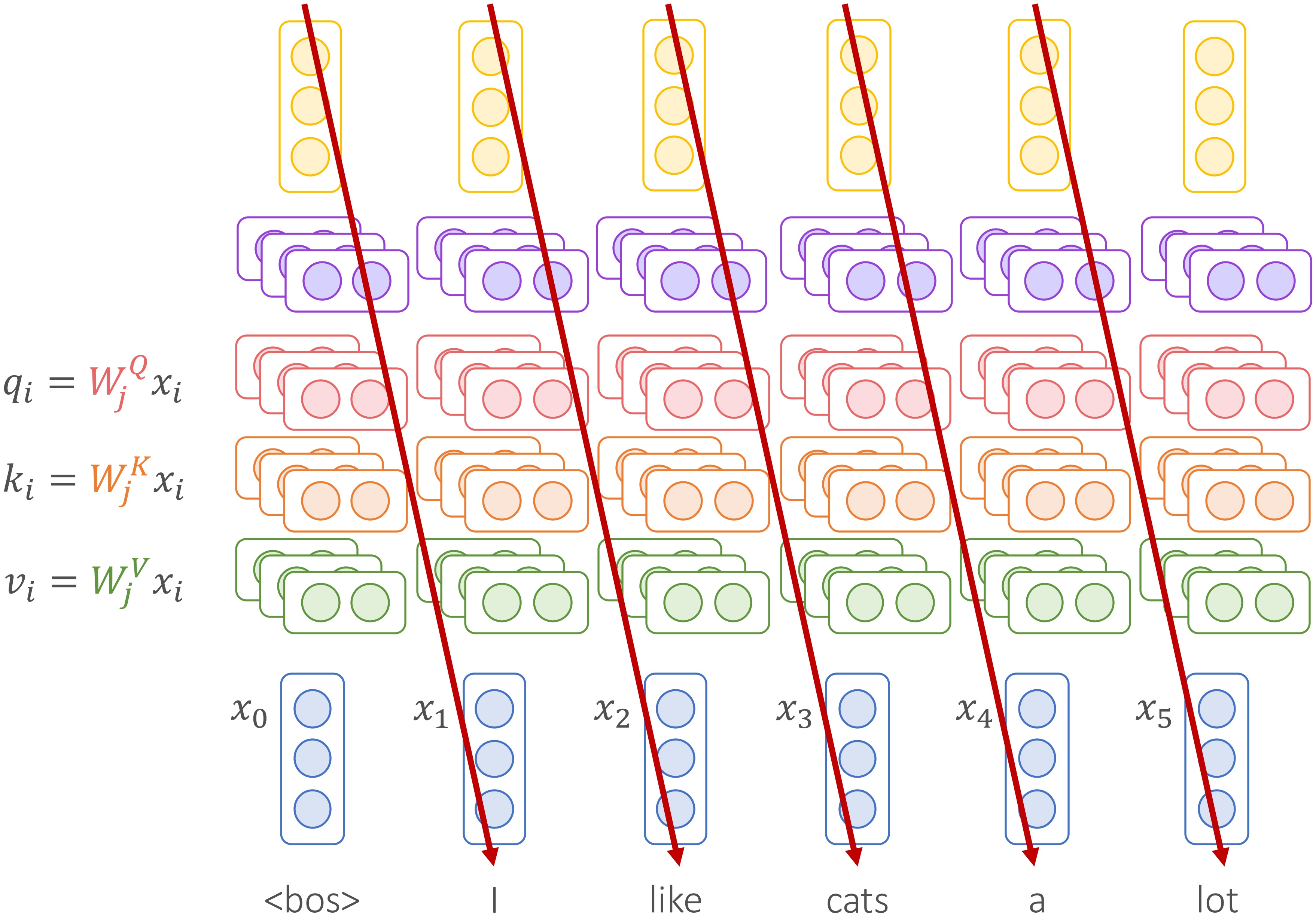
Masked Attention: Implementation



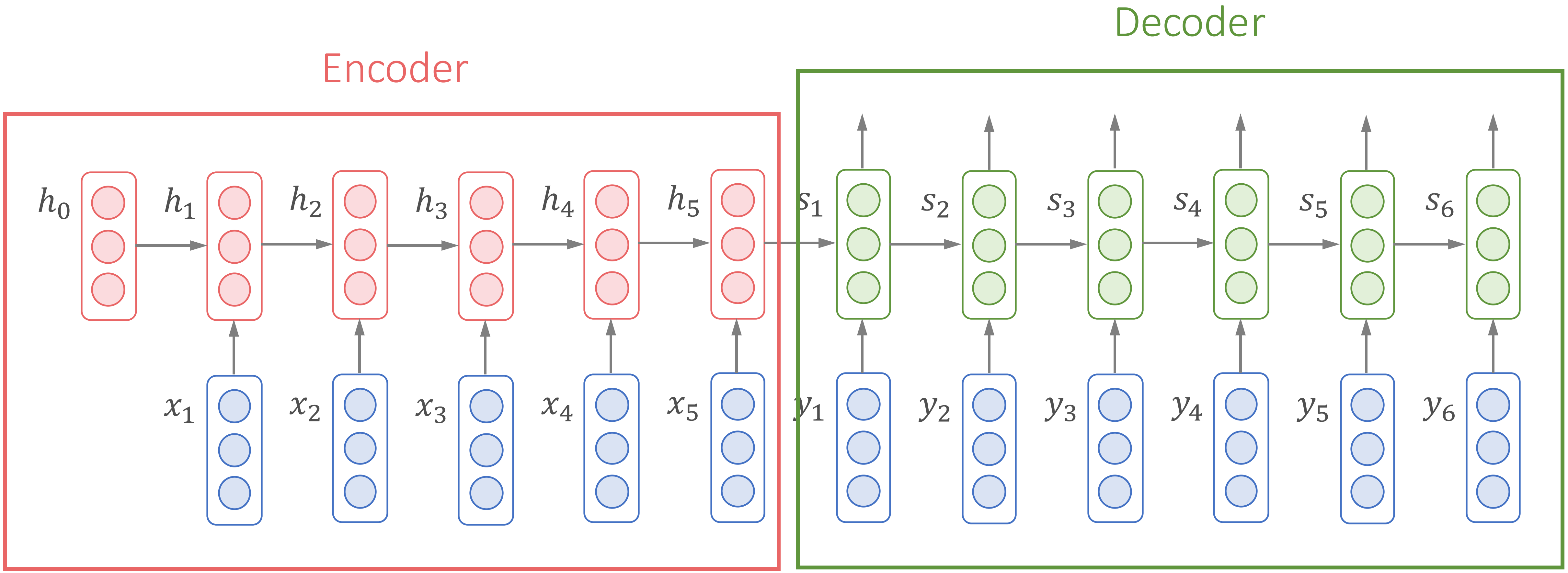
Masked Attention: Implementation



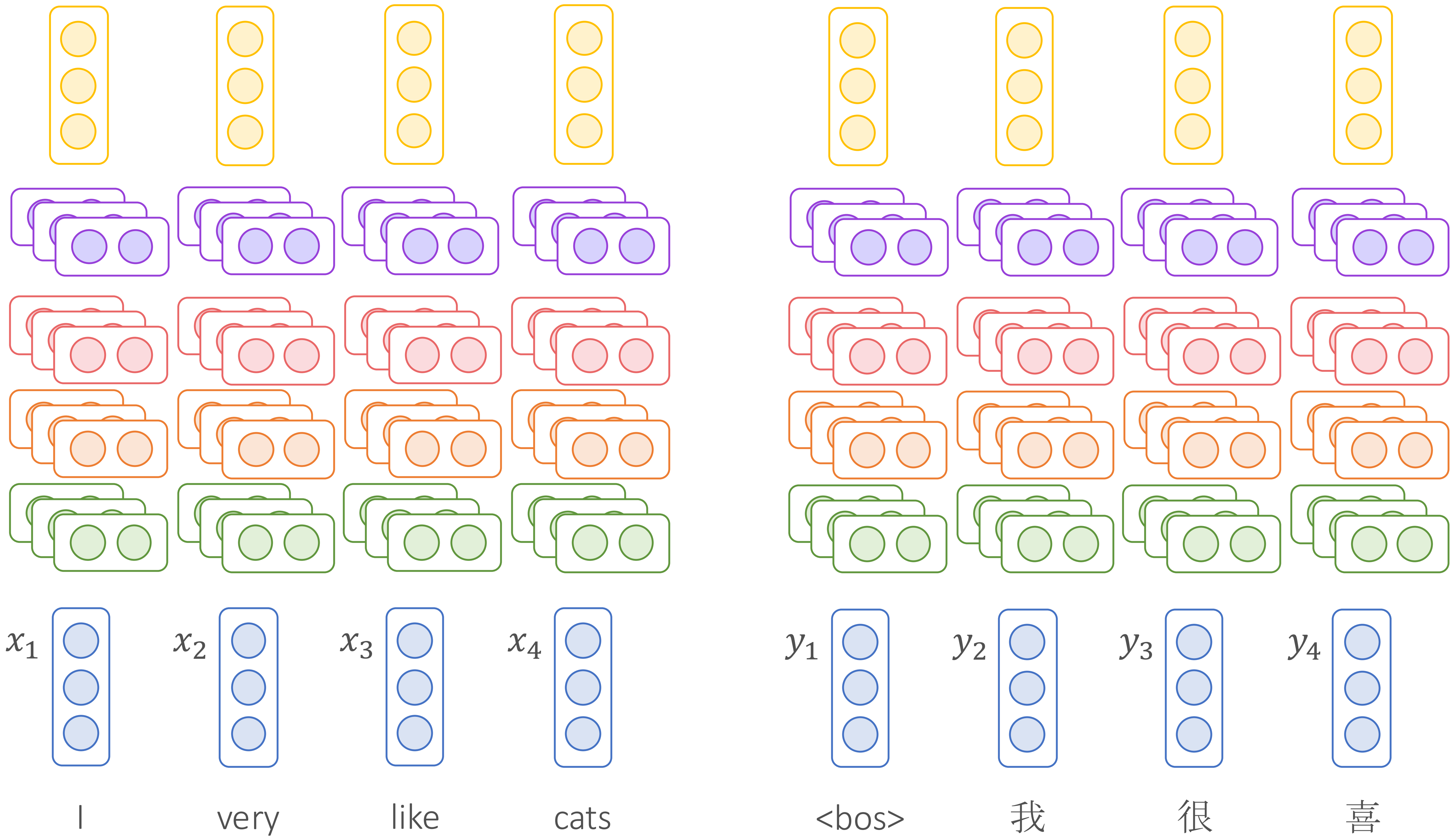
Transformer Decoder



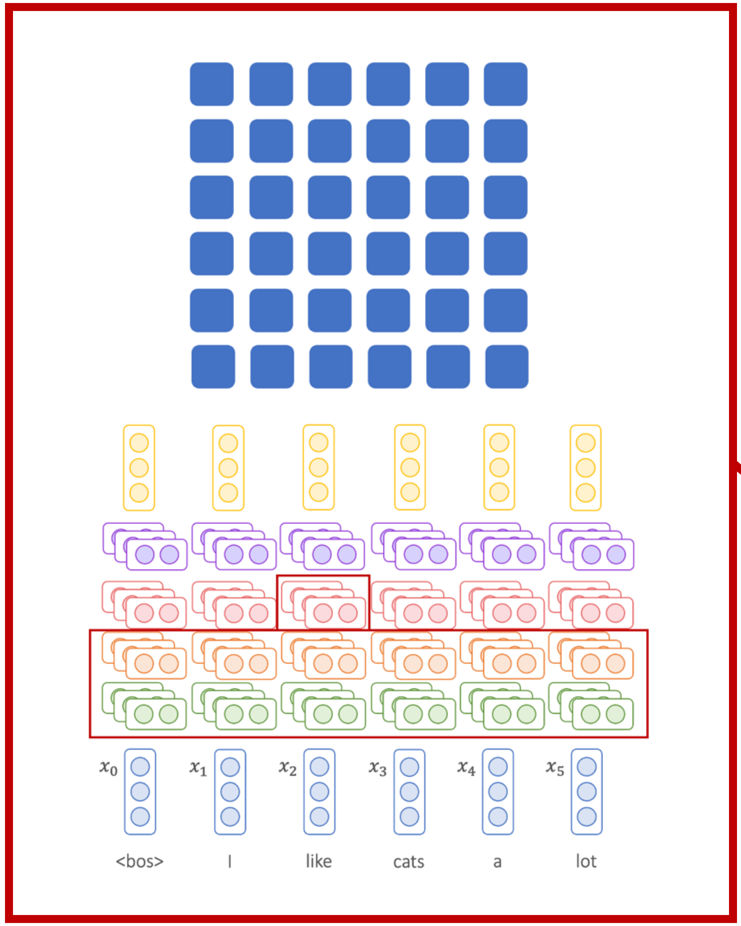
How About Encoder-Decoder (Sequence-to-Sequence)?



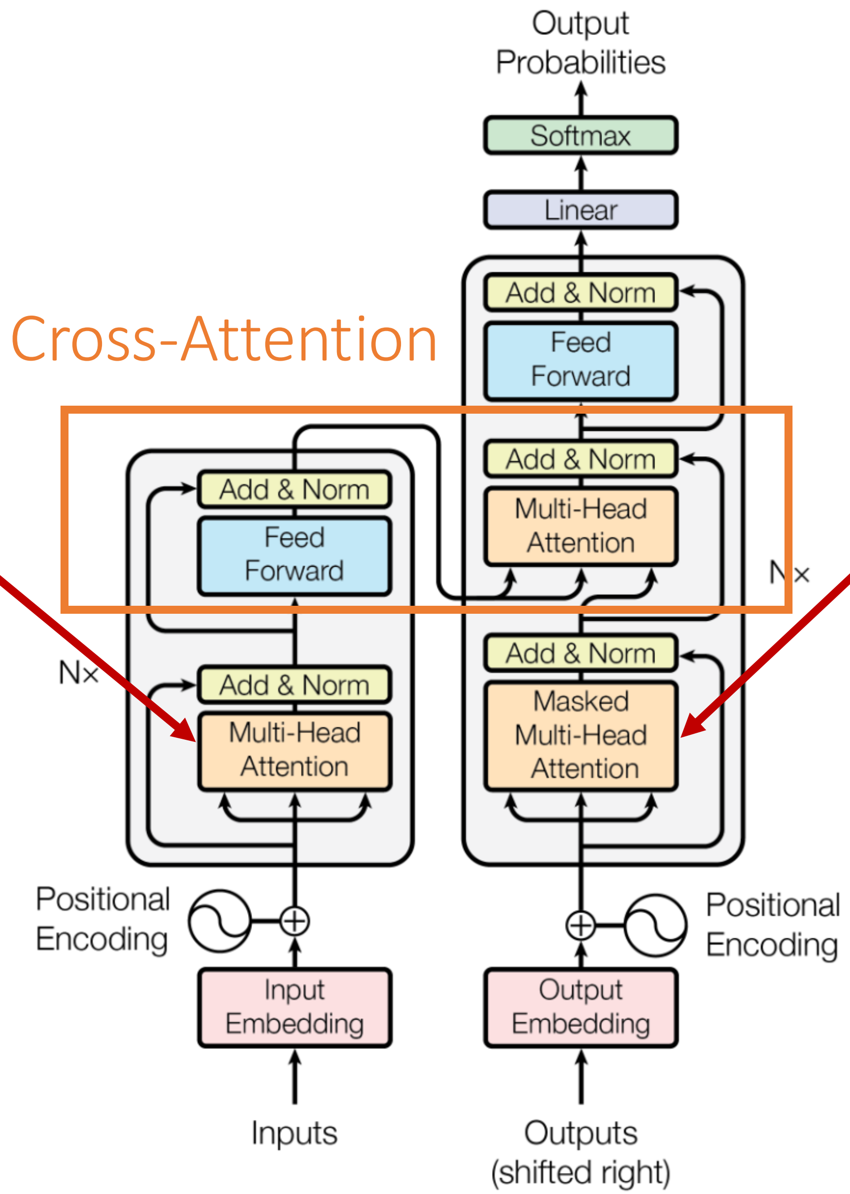
Transformer Encoder-Decoder (Sequence-to-Sequence)

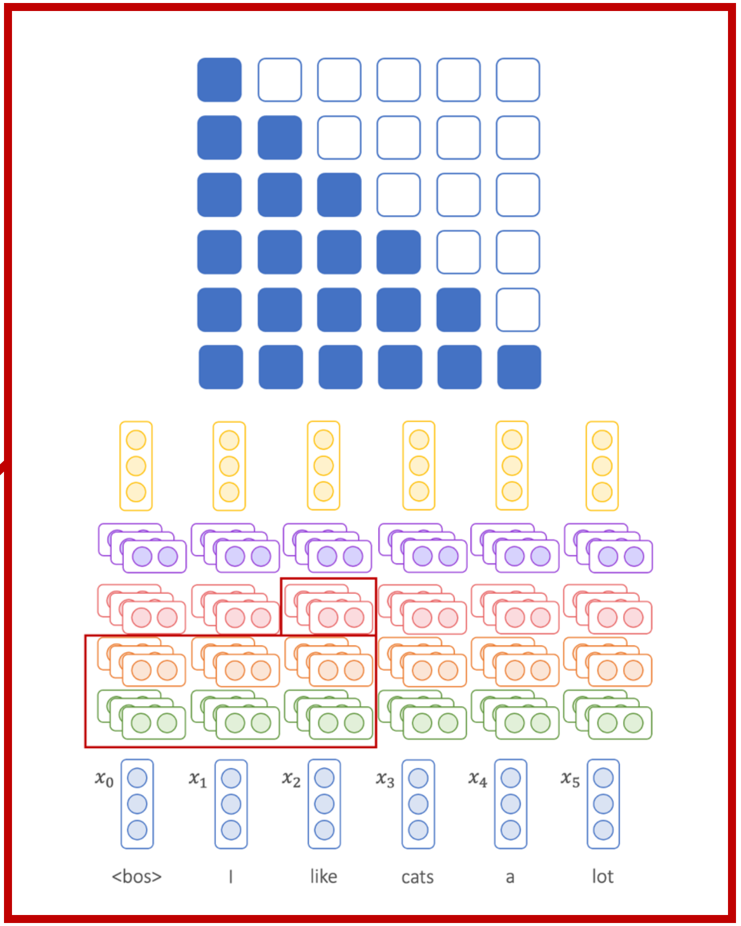


Transformer Encoder-Decoder (Sequence-to-Sequence)



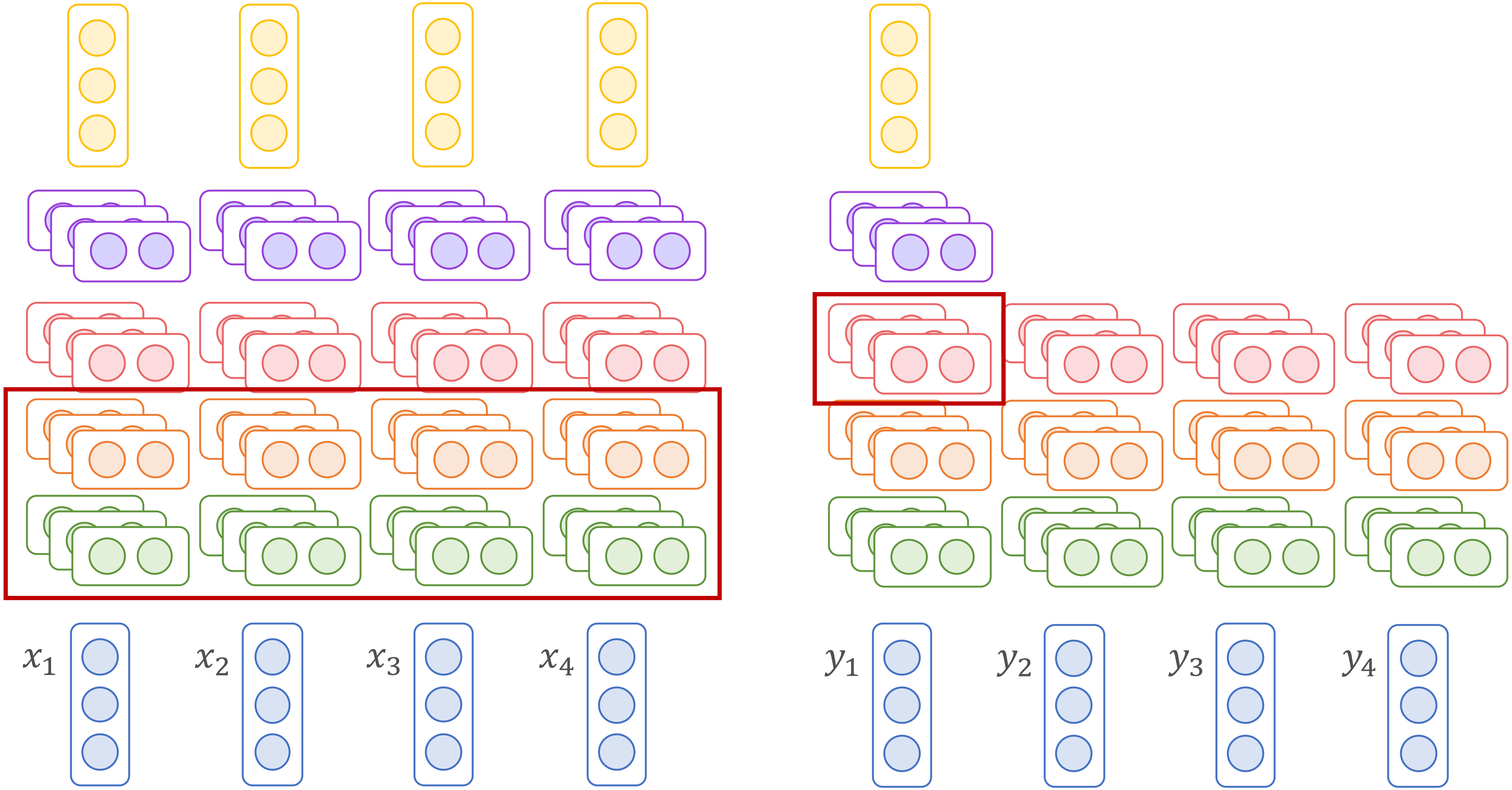
Transformer Encoder



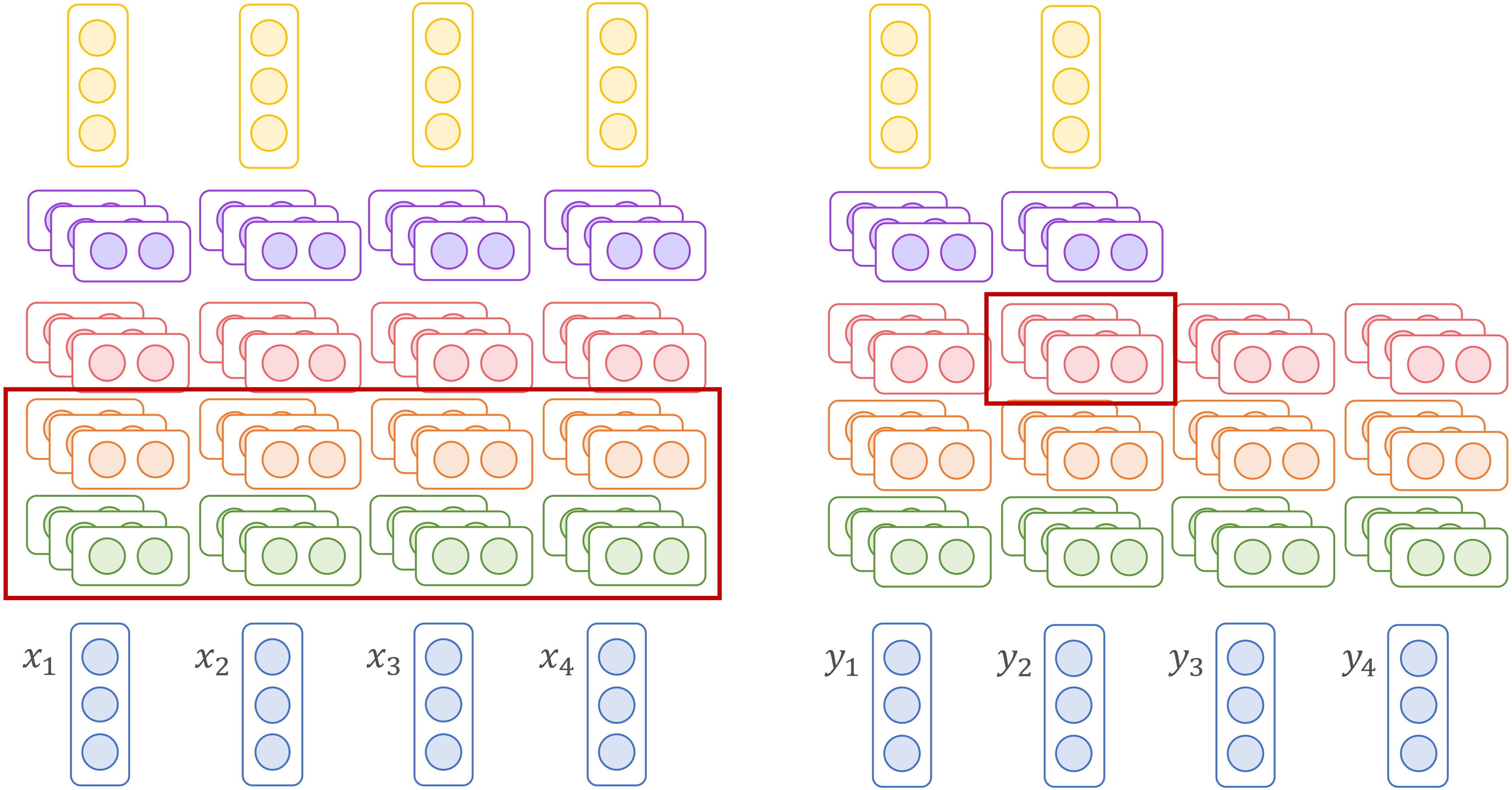


Transformer Decoder

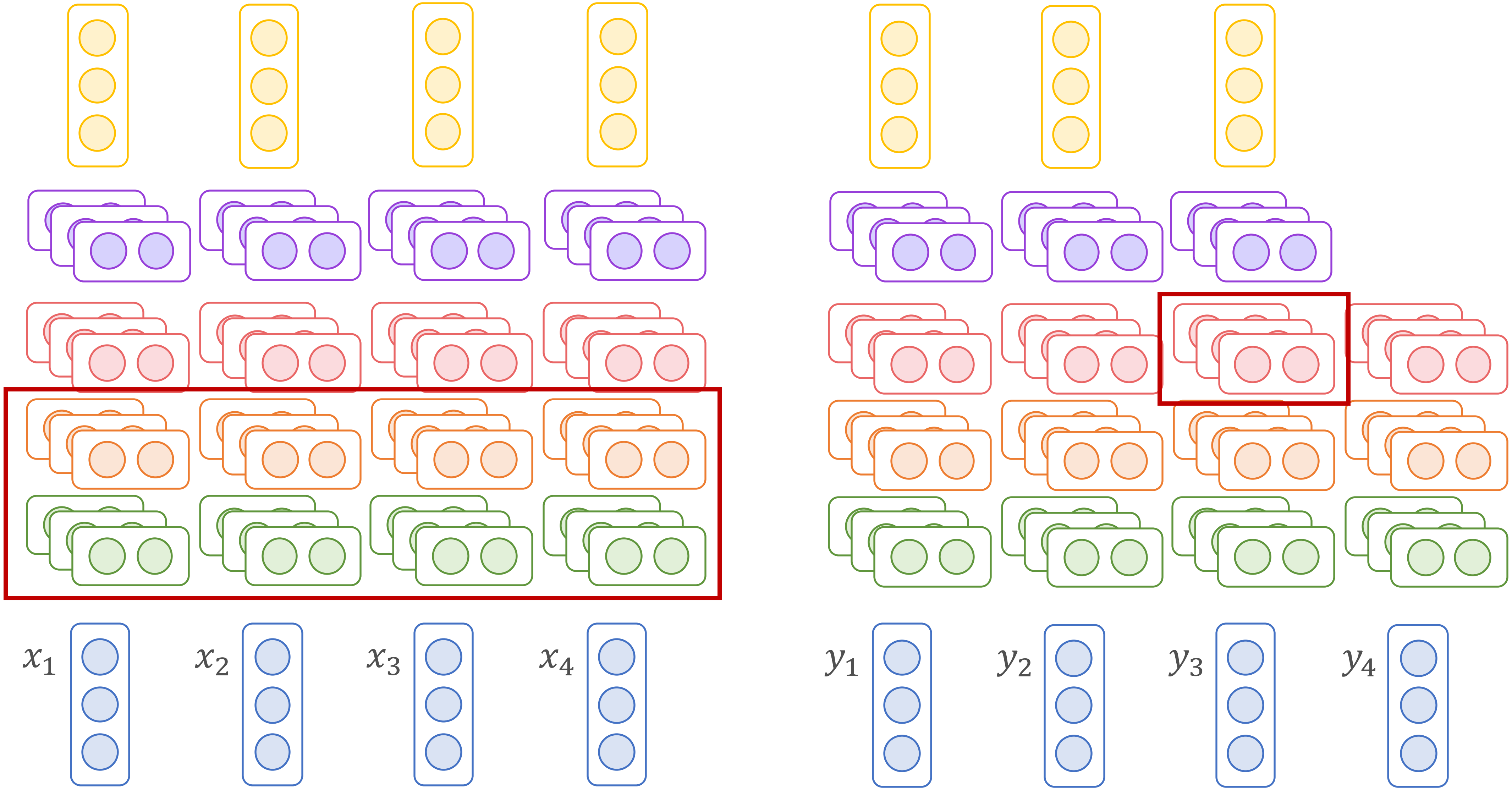
Cross-Attention



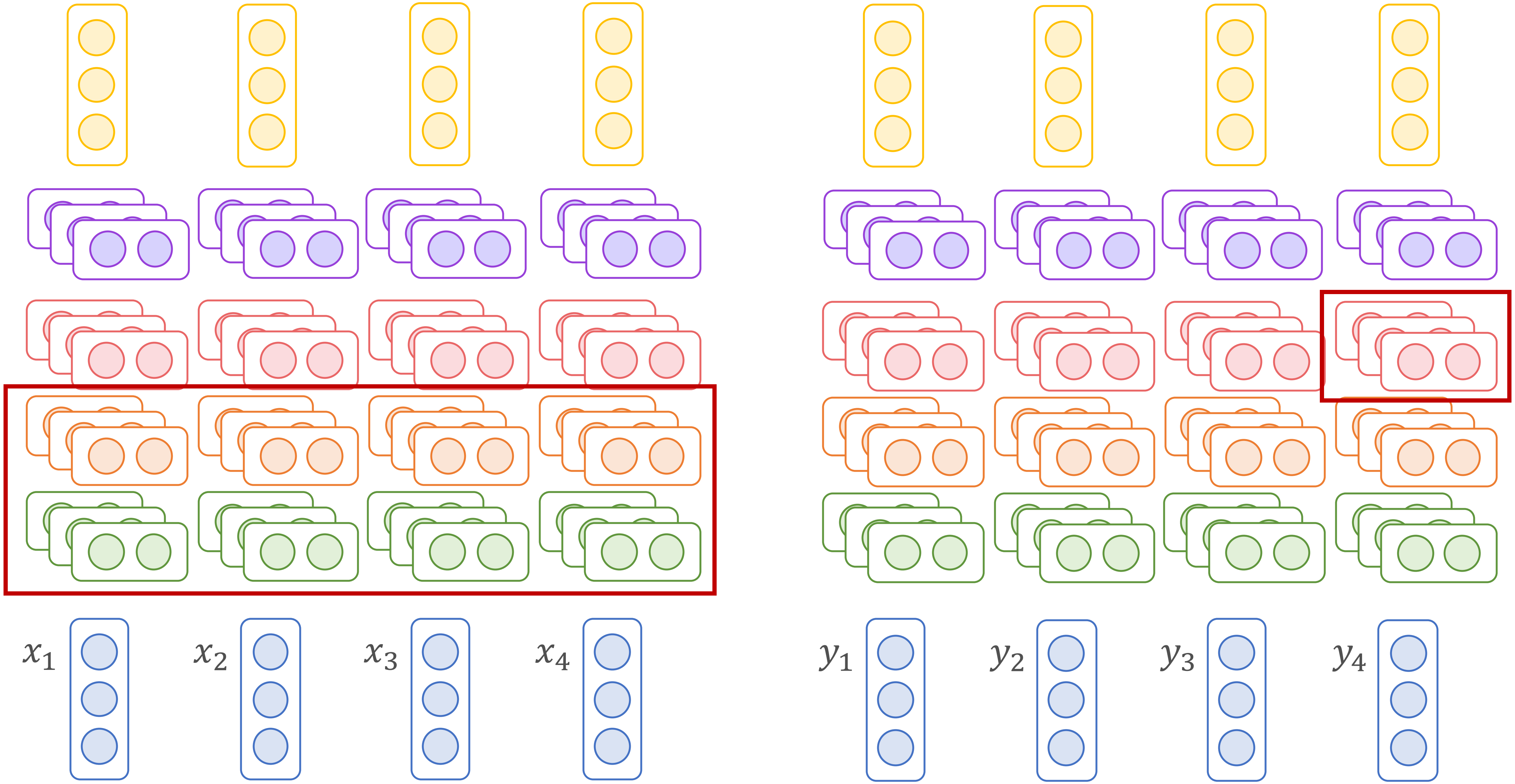
Cross-Attention



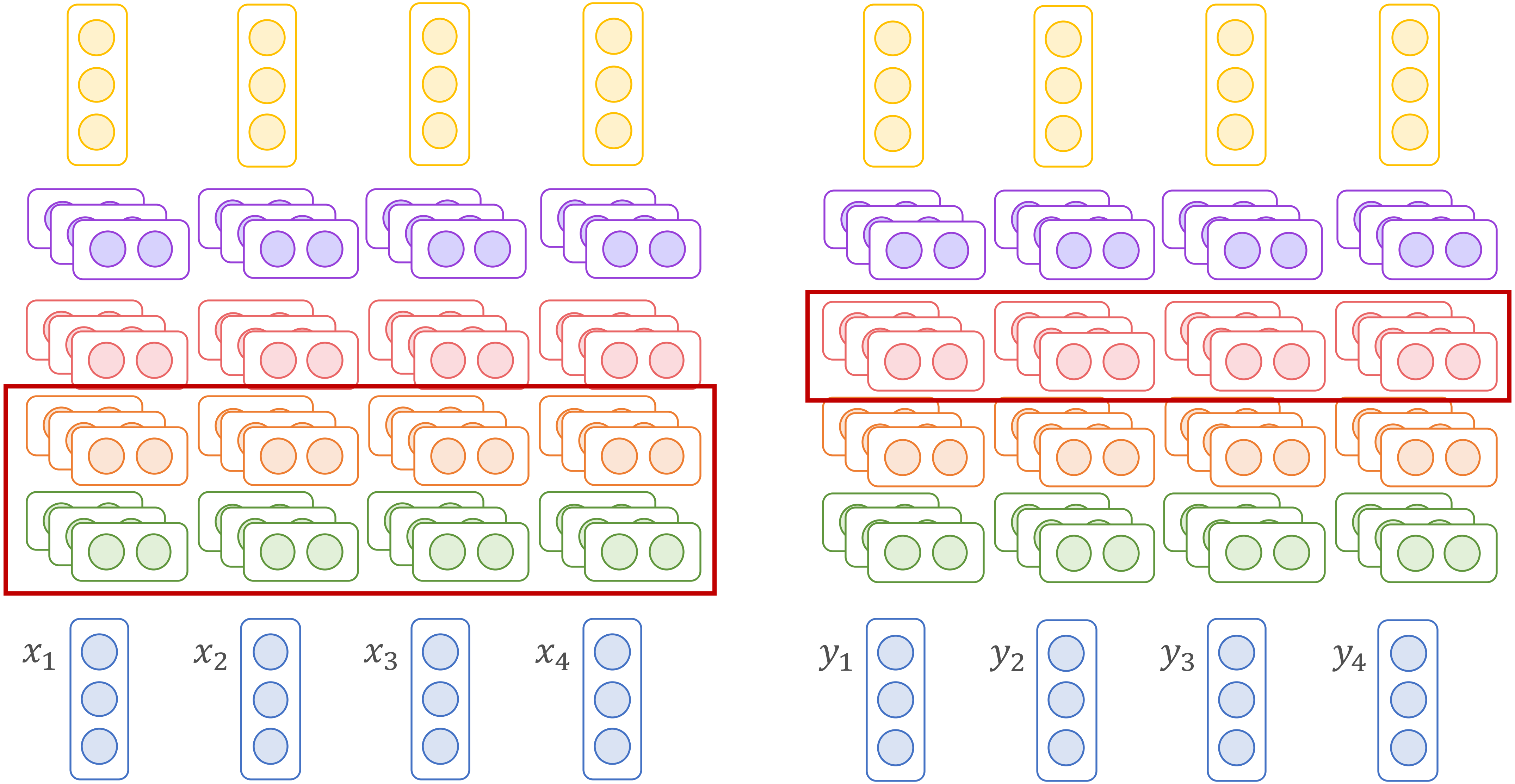
Cross-Attention



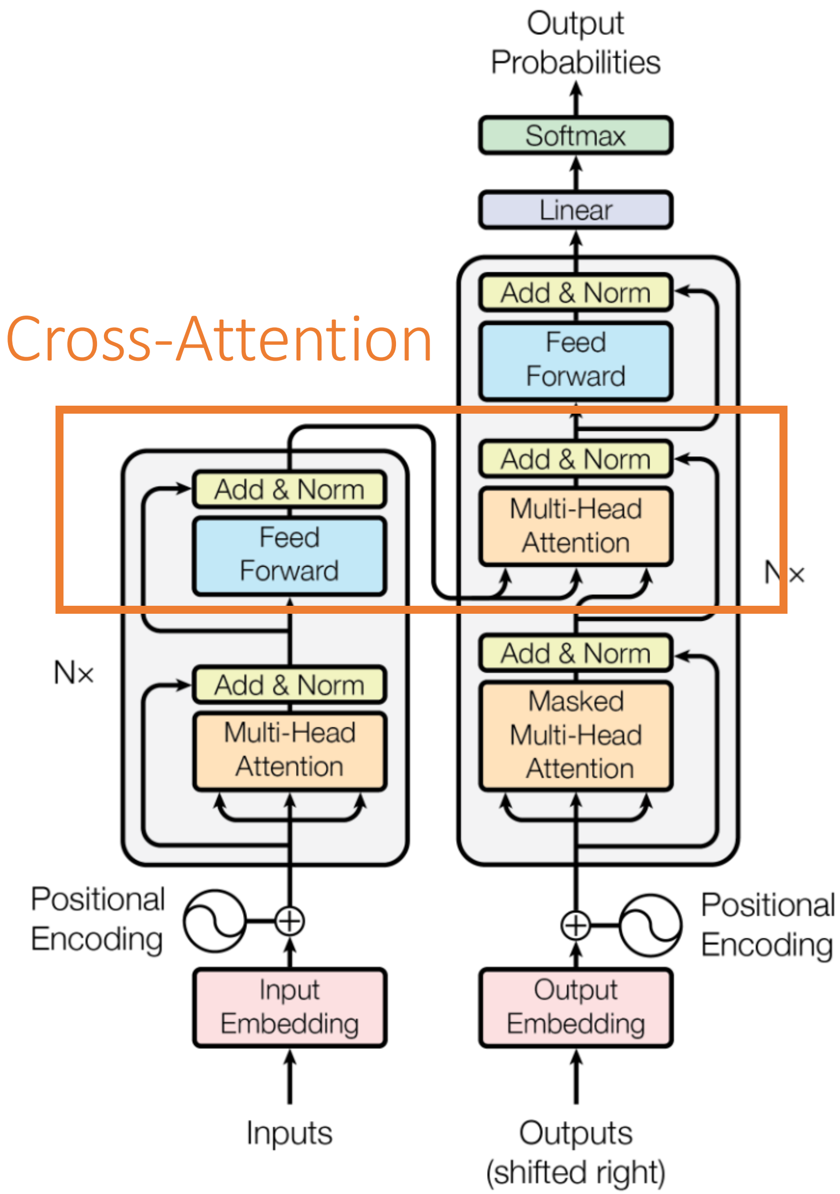
Cross-Attention



Cross-Attention



Transformer



Transformer on Machine Translation

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

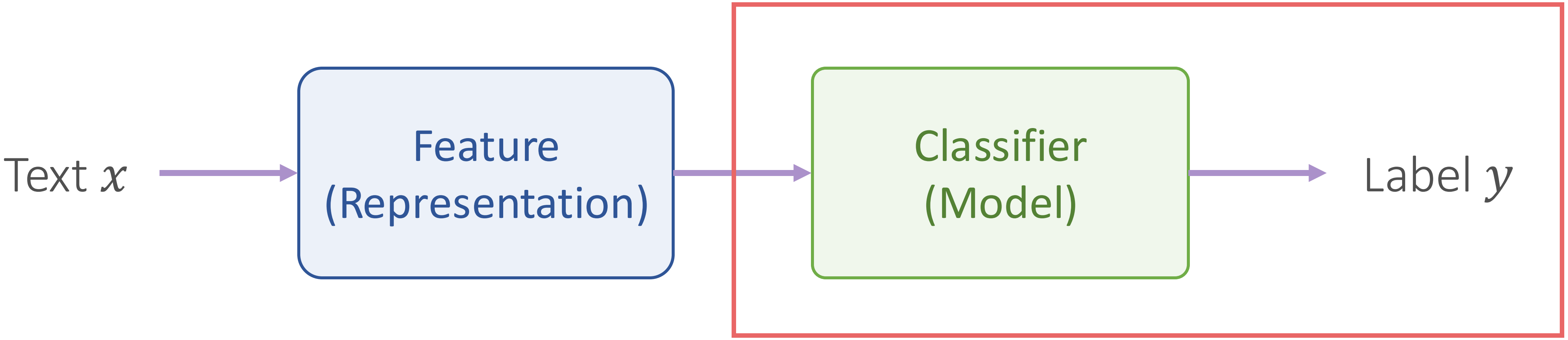
Attention Is All You Need, 2017

Transformer on Document Generation

Model	Test perplexity	ROUGE-L
seq2seq-attention, $L = 500$	5.04952	12.7
Transformer-ED, $L = 500$	2.46645	34.2
Transformer-D, $L = 4000$	2.22216	33.6
Transformer-DMCA, no MoE-layer, $L = 11000$	2.05159	36.2
Transformer-DMCA, MoE-128, $L = 11000$	1.92871	37.9
Transformer-DMCA, MoE-256, $L = 7500$	1.90325	38.8

Generating Wikipedia by Summarizing Long Sequences, 2018

A General Framework for Text Classification



- Teach the model how to make prediction y
- Logistic regression, neural networks, CNN, RNN, LSTM, Transformers

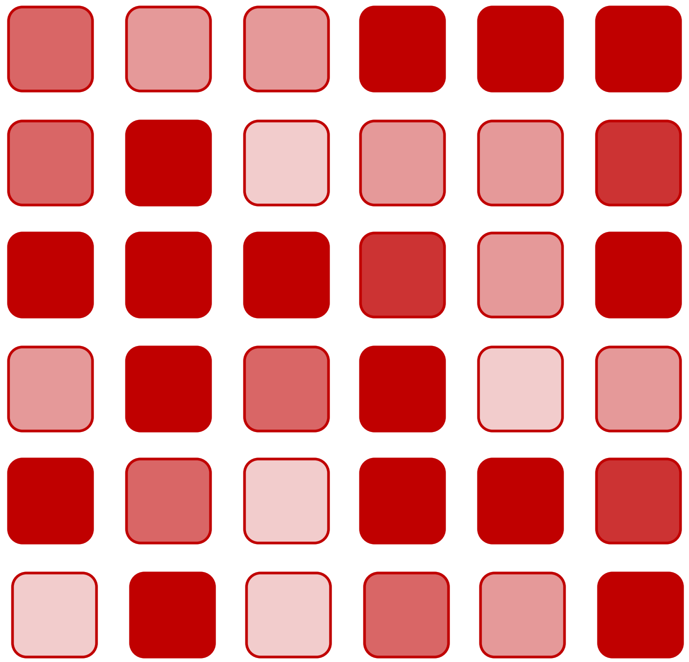
Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$

Lecture Plan

- Transformers
  - Encoder
  - Decoder
  - Encoder-Decoder
- Transformers Variants
  - Longformer
  - Relative Positional Encoding
  - RoFormer

Computation in Transformer

- All-pair attention scores
  - Complexity $O(length^2)$
- When the input is long → slow

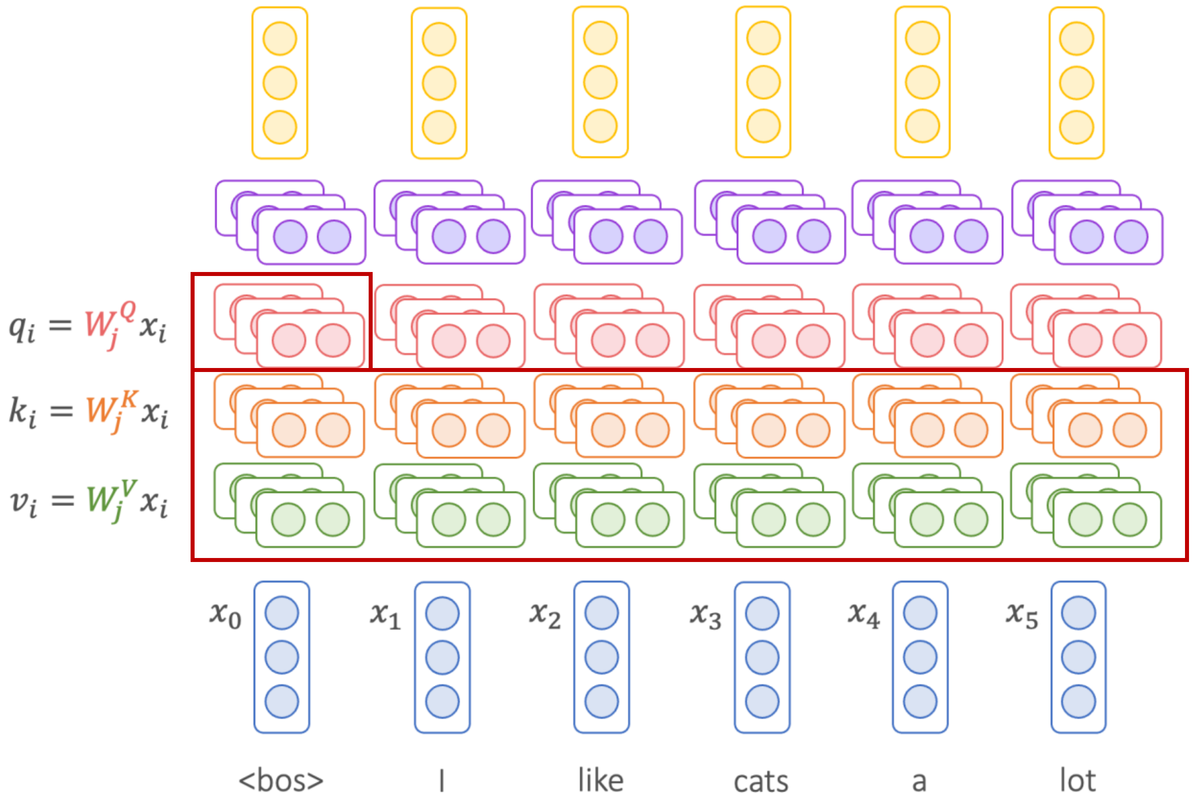


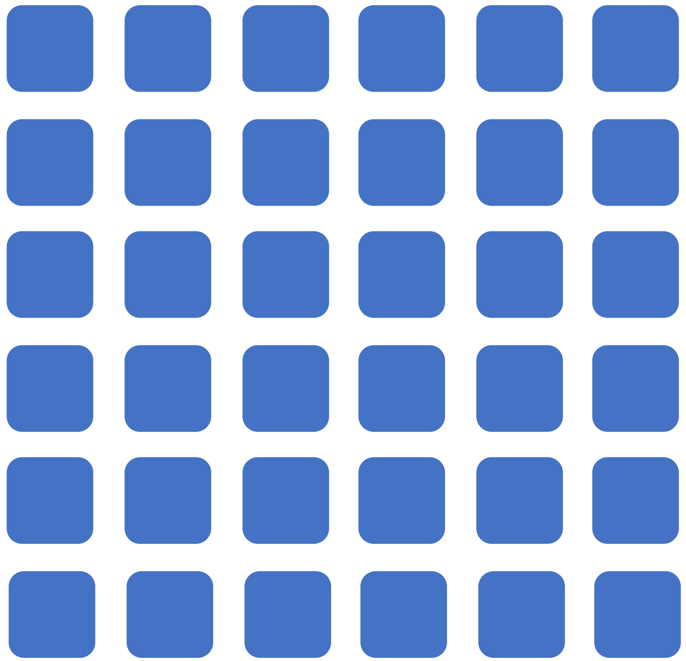
LongFormer

- Don't compute all-pair attention score
  - Manipulate attention mask
- Capture local information to reduce computational load
  - Idea is similar to convolutional neural network

Longformer: The Long-Document Transformer, 2020

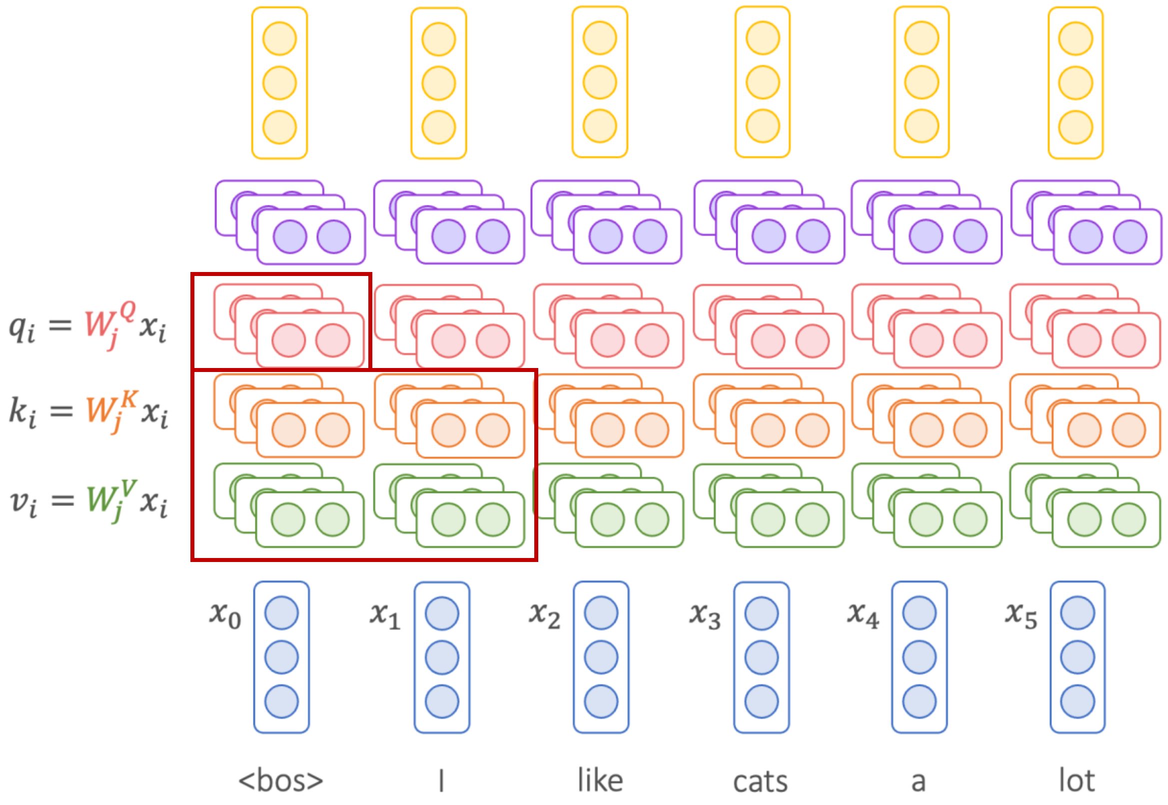
Transformer Encoder

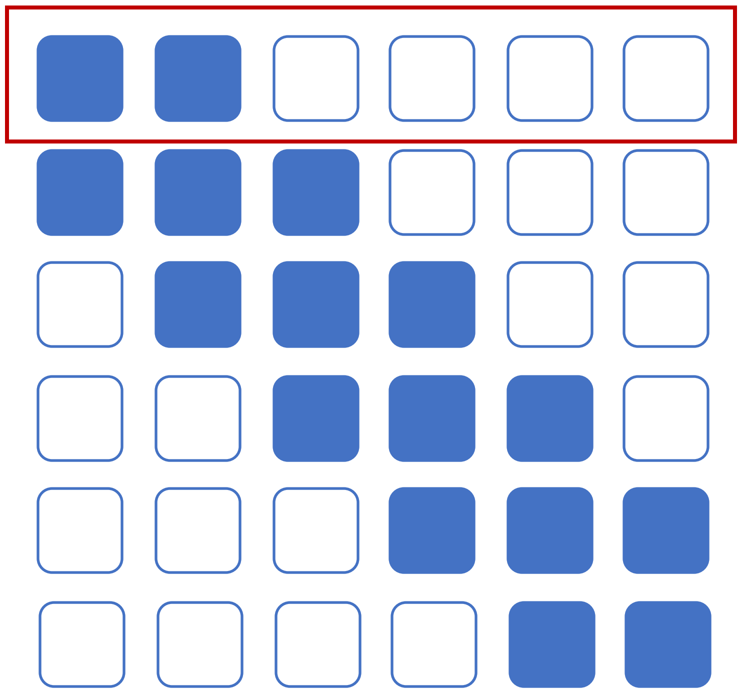




No Masking

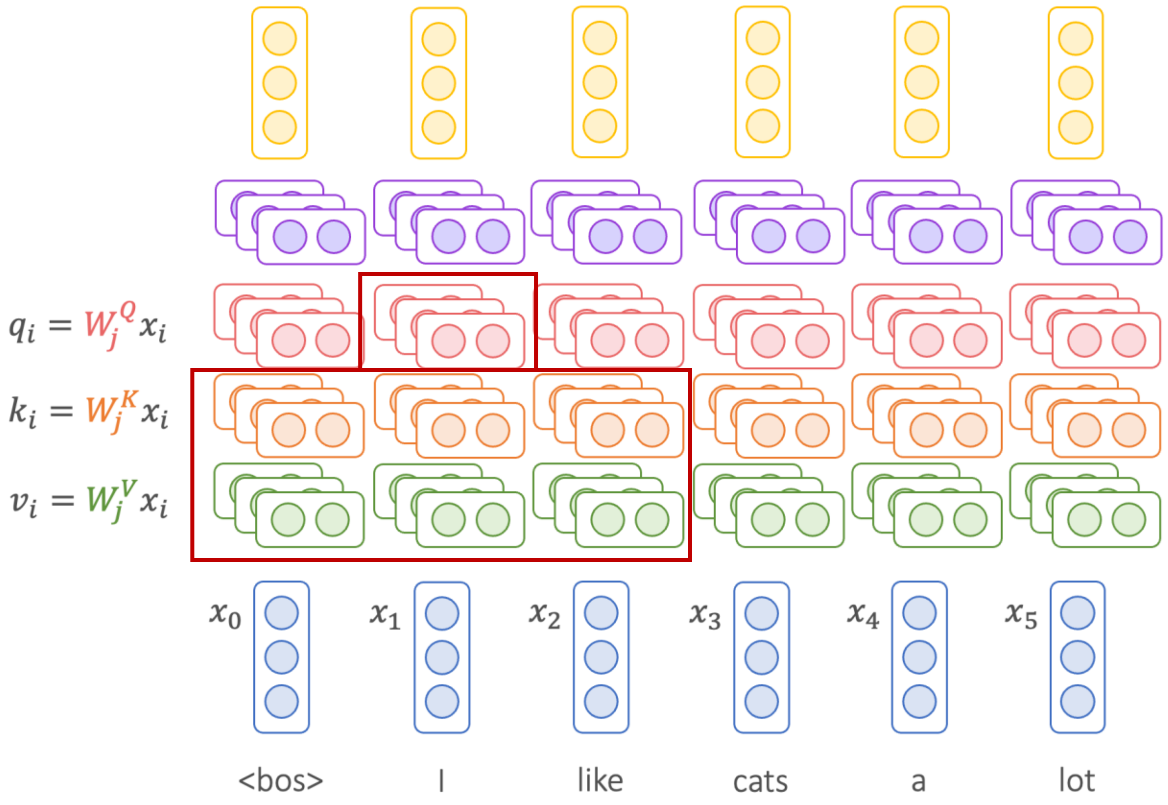
Sliding Window Attention Masking

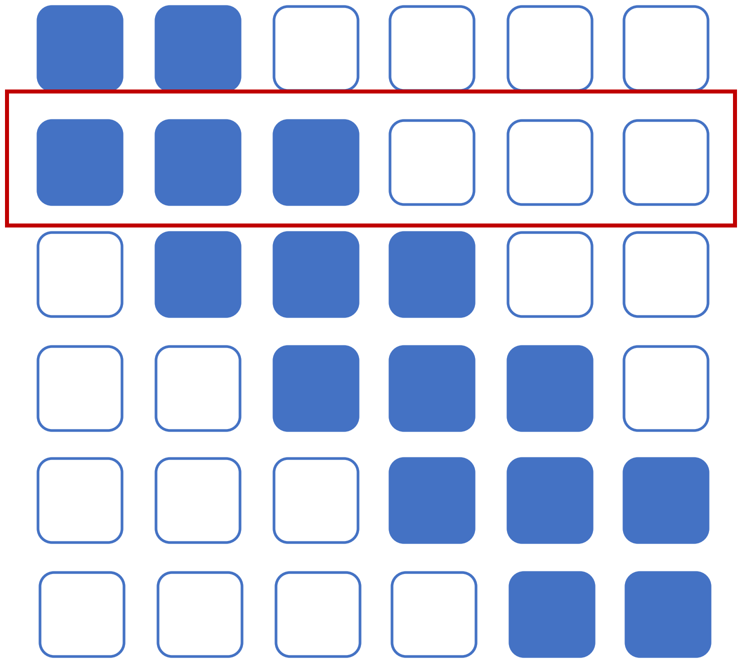




Sliding Window Attention Masking

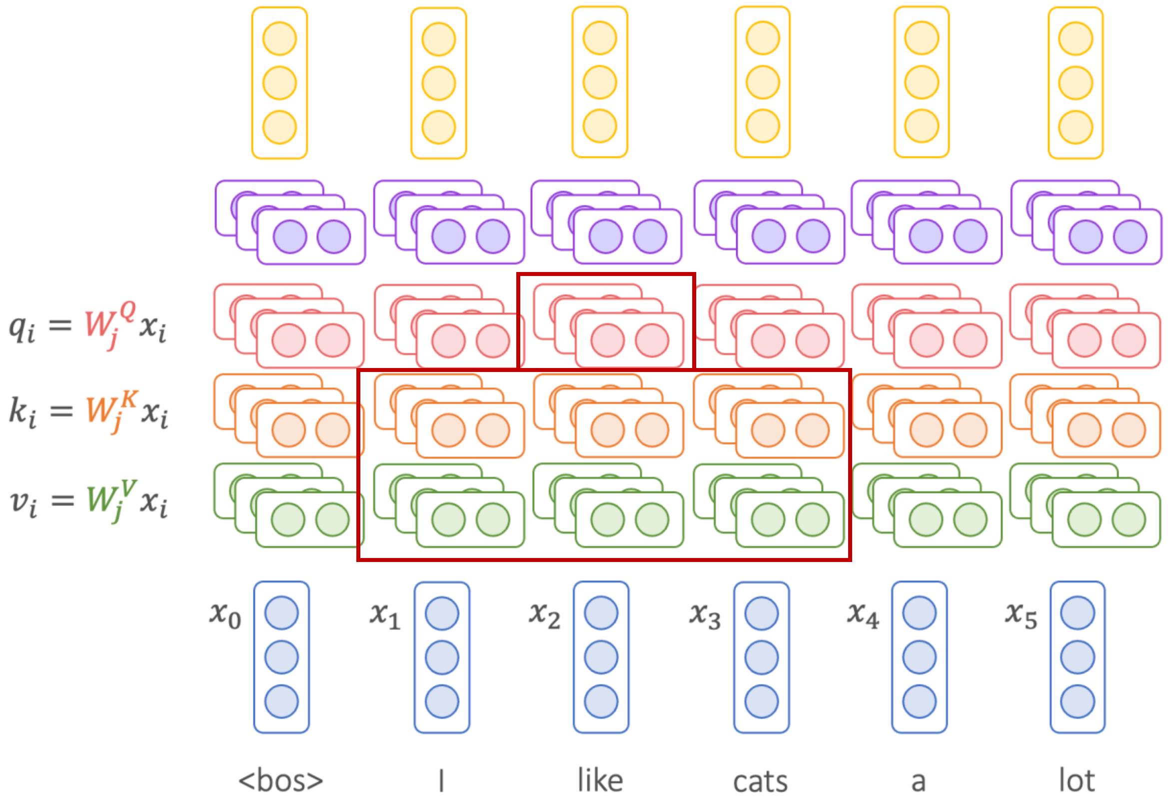
Sliding Window Attention Masking

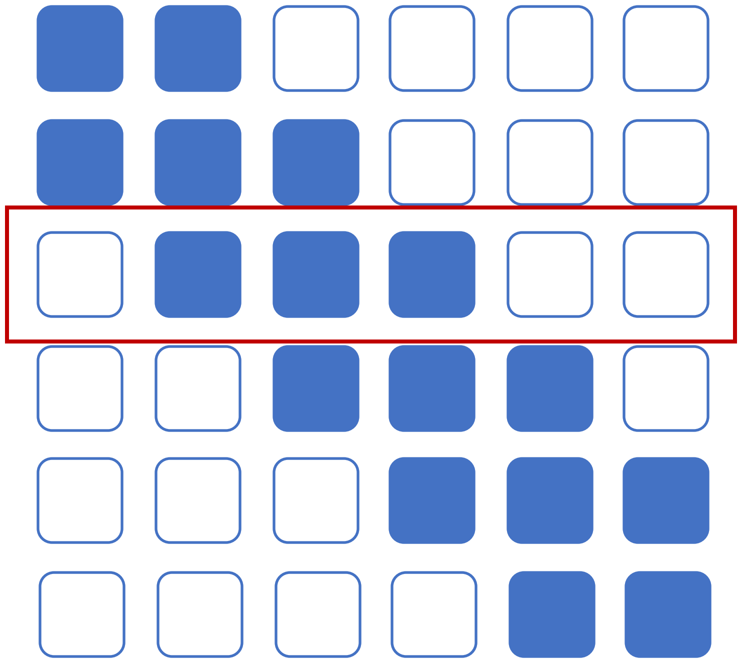




Sliding Window Attention Masking

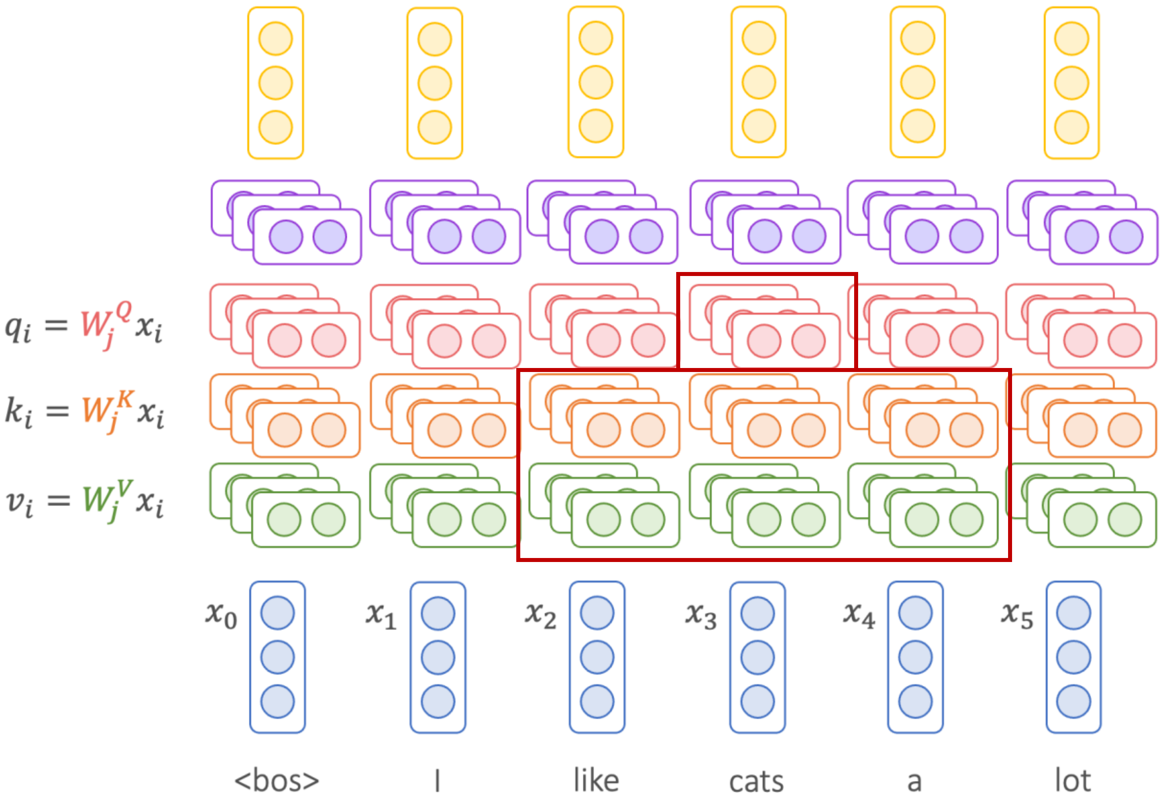
Sliding Window Attention Masking

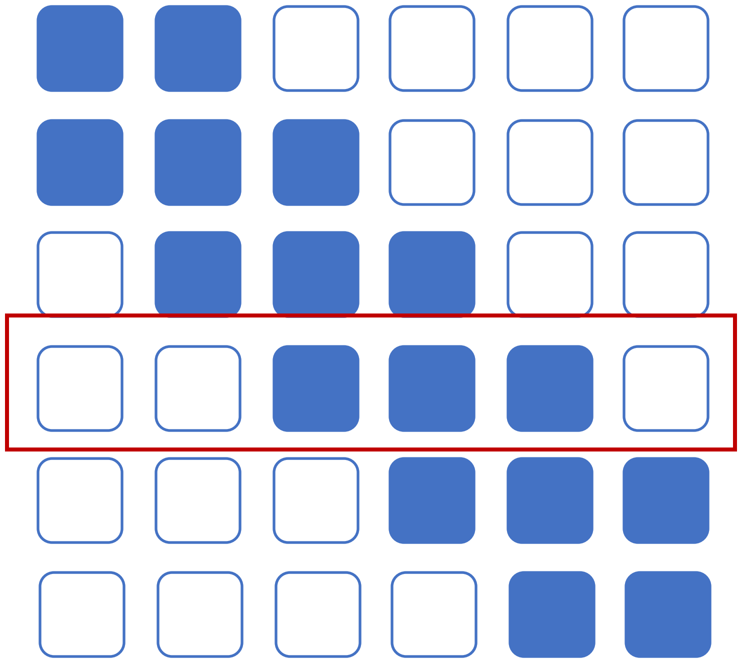




Sliding Window Attention Masking

Sliding Window Attention Masking





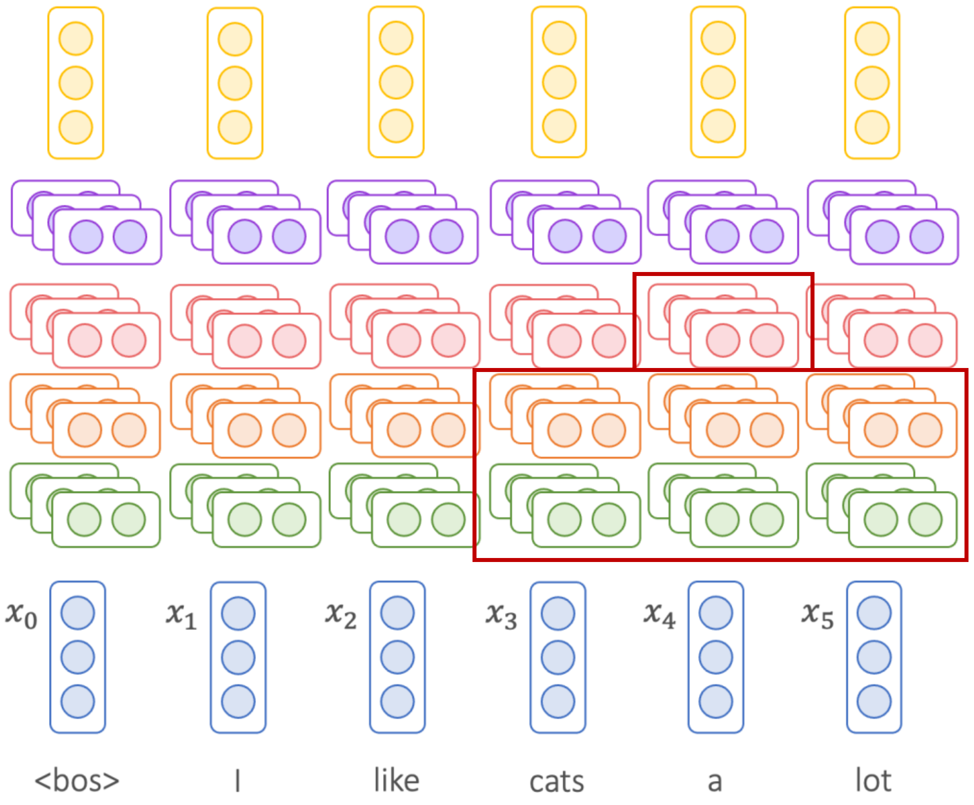
Sliding Window Attention Masking

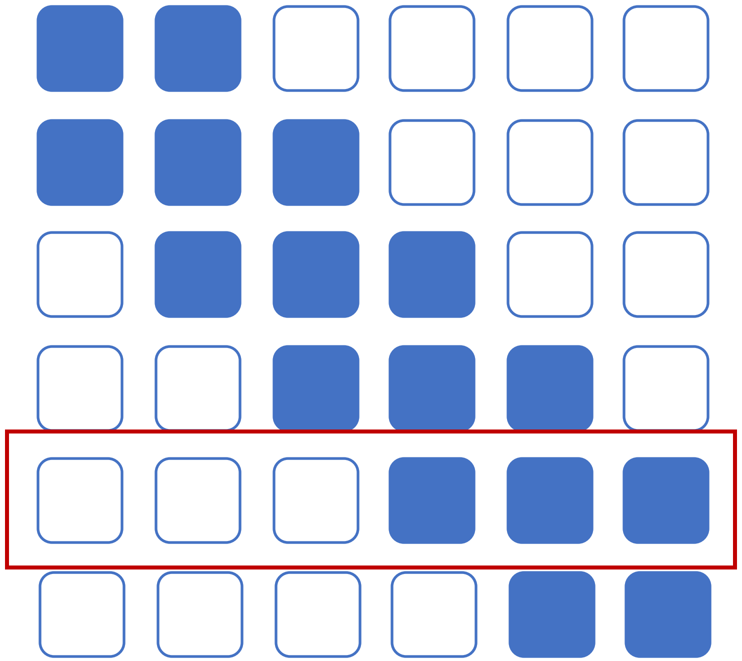
Sliding Window Attention Masking

$q_i = W_j^Q x_i$

$k_i = W_j^K x_i$

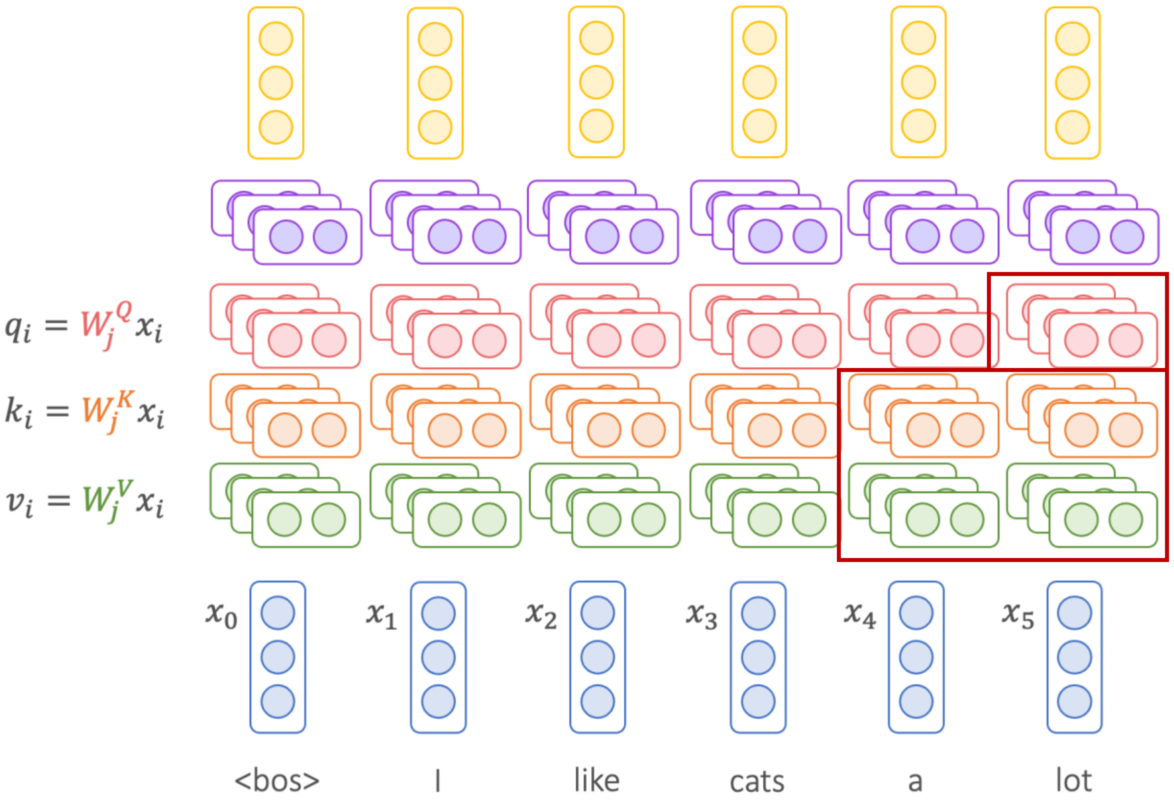
$v_i = W_j^V x_i$

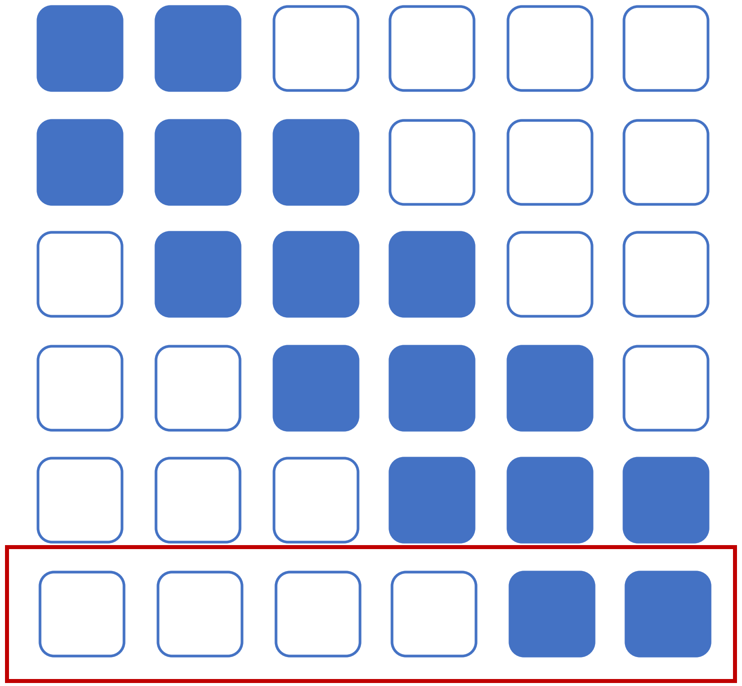




Sliding Window Attention Masking

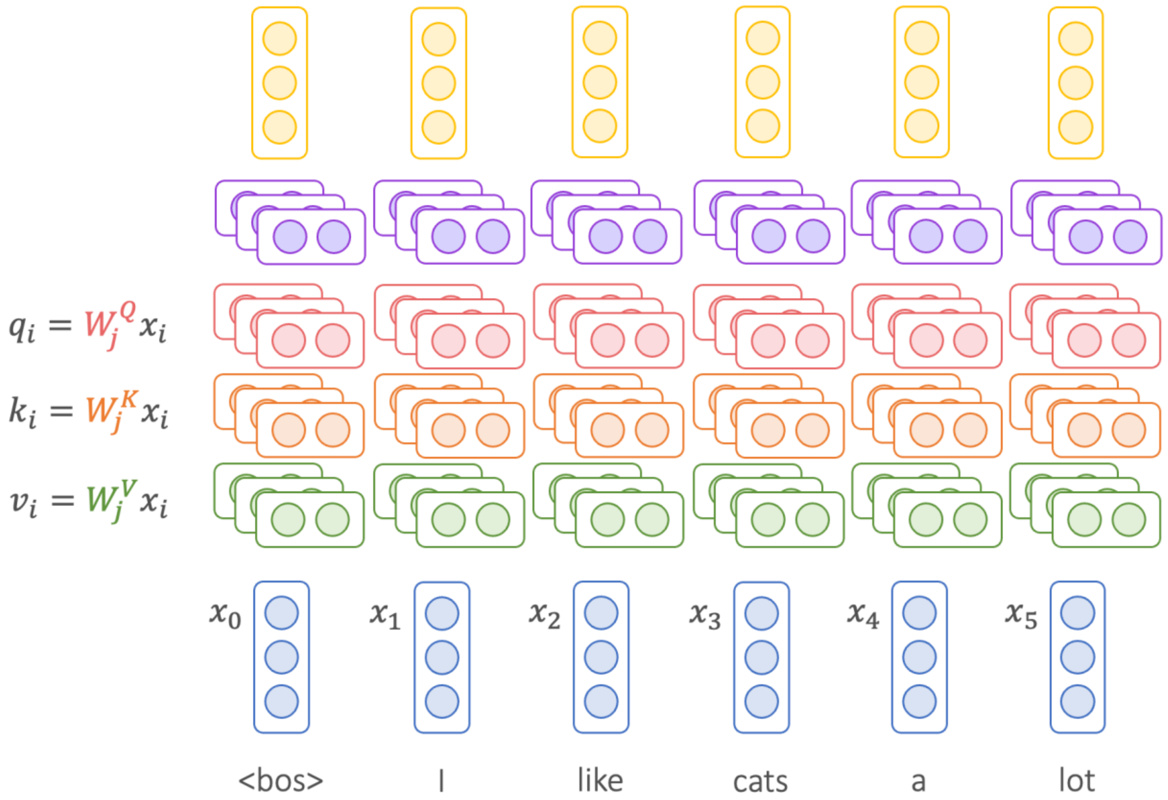
Sliding Window Attention Masking

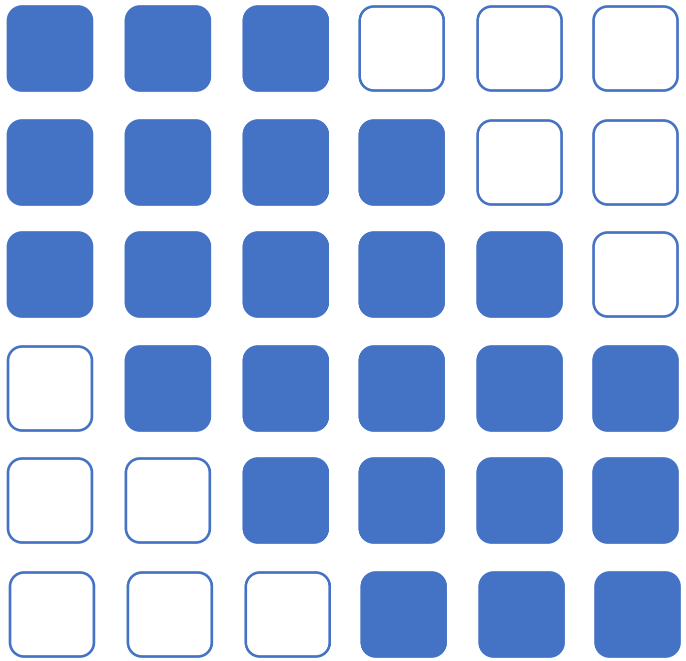




Sliding Window Attention Masking

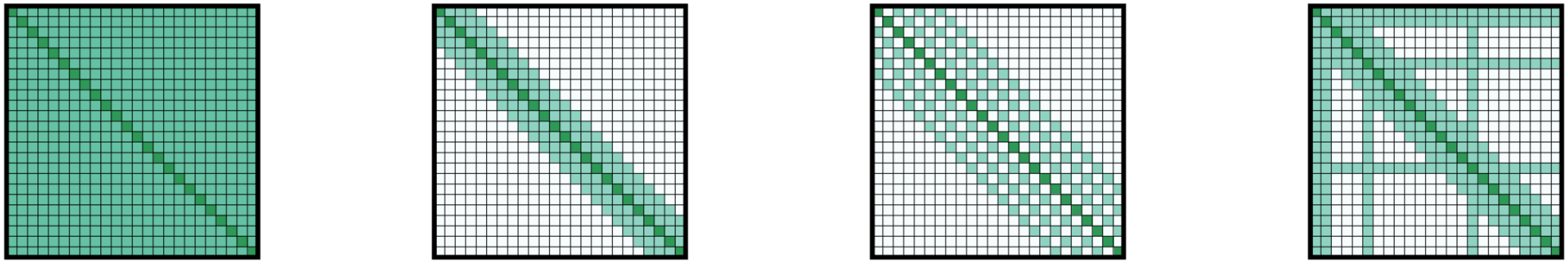
Sliding Window Attention Masking





Sliding Window Attention Masking

Different Types of Attention Masks



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

LongFormer Results on Language Modeling

Model	#Param	Dev	Test
Dataset `text8`			
T12 (Al-Rfou et al., 2018)	44M	-	1.18
Adaptive (Sukhbaatar et al., 2019)	38M	1.05	1.11
BP-Transformer (Ye et al., 2019)	39M	-	1.11
Our Longformer	41M	1.04	**1.10**
Dataset `enwik8`			
T12 (Al-Rfou et al., 2018)	44M	-	1.11
Transformer-XL (Dai et al., 2019)	41M	-	1.06
Reformer (Kitaev et al., 2020)	-	-	1.05
Adaptive (Sukhbaatar et al., 2019)	39M	1.04	1.02
BP-Transformer (Ye et al., 2019)	38M	-	1.02
Our Longformer	41M	1.02	**1.00**

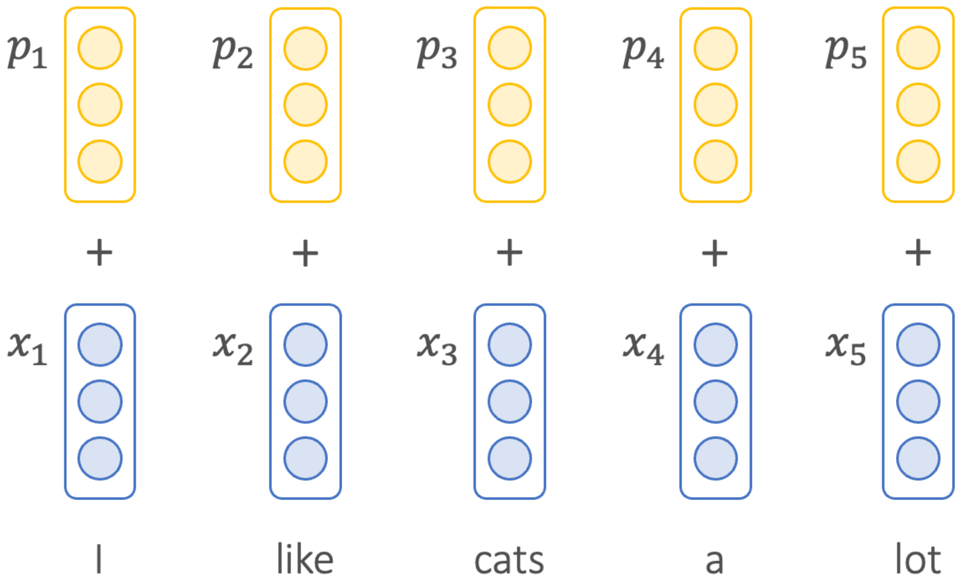
Longformer: The Long-Document Transformer, 2020

Lecture Plan

- Transformers
  - Encoder
  - Decoder
  - Encoder-Decoder
- Transformers Variants
  - Longformer
  - Relative Positional Encoding
  - RoFormer

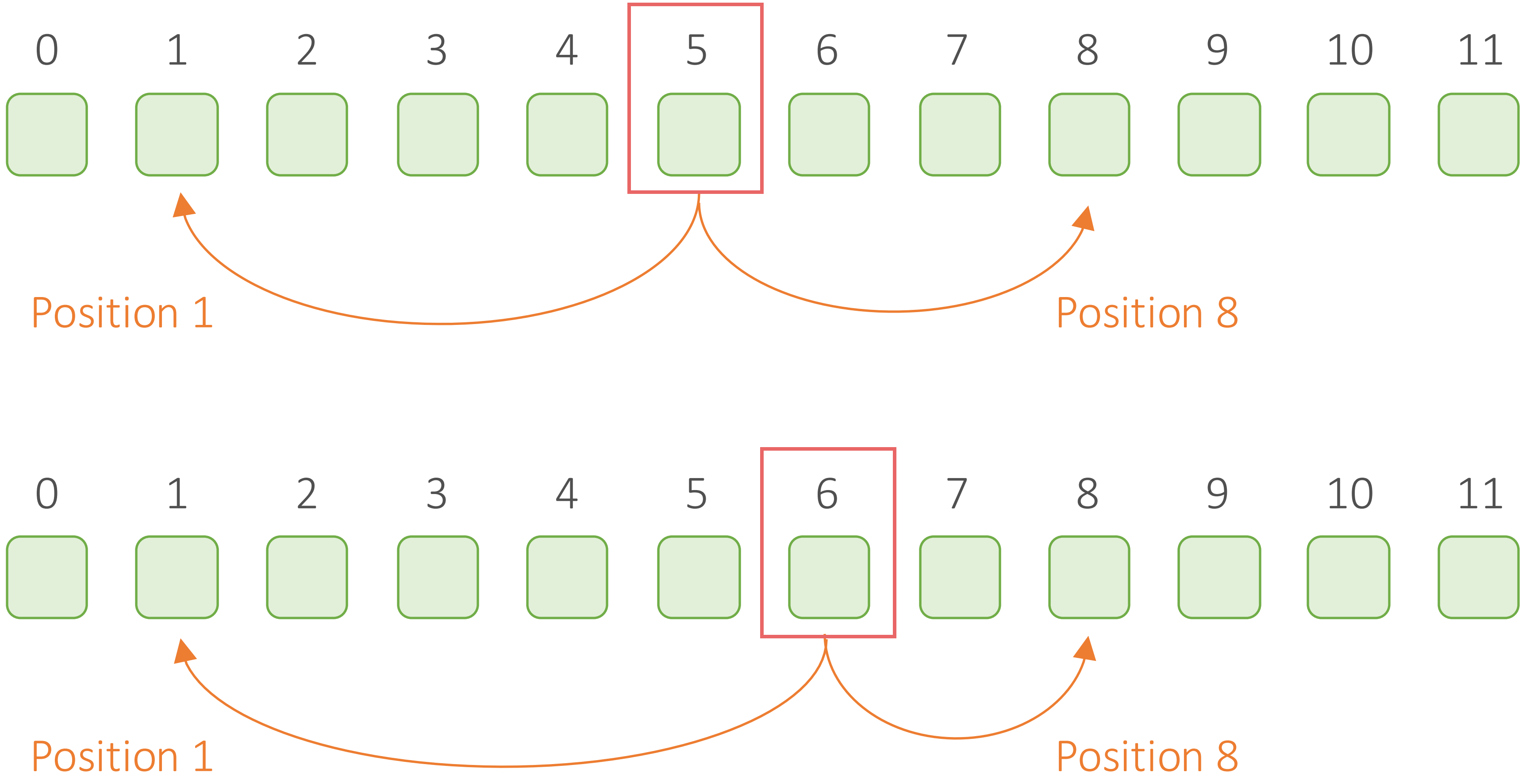
Absolute Positional Encoding

$$x_i \leftarrow x_i + PE_i$$

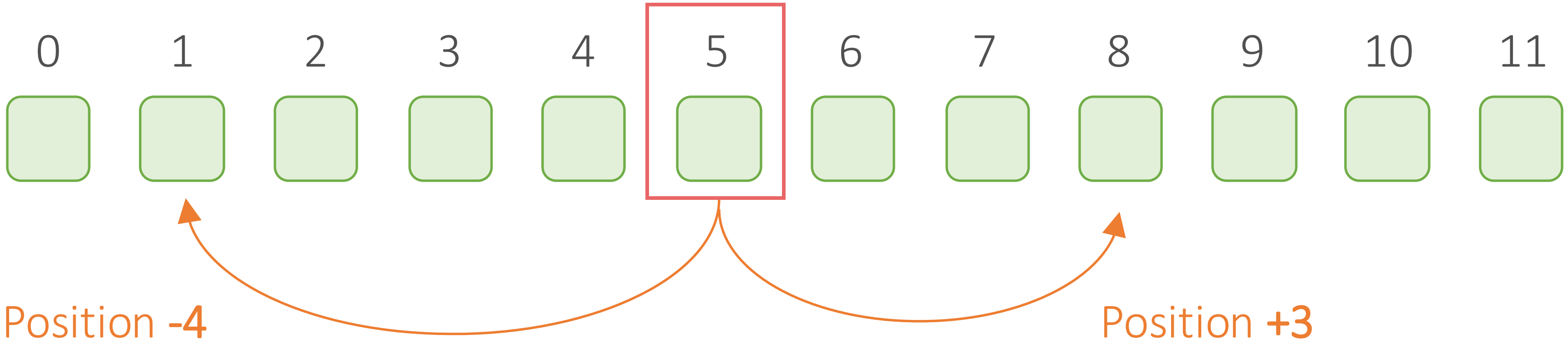


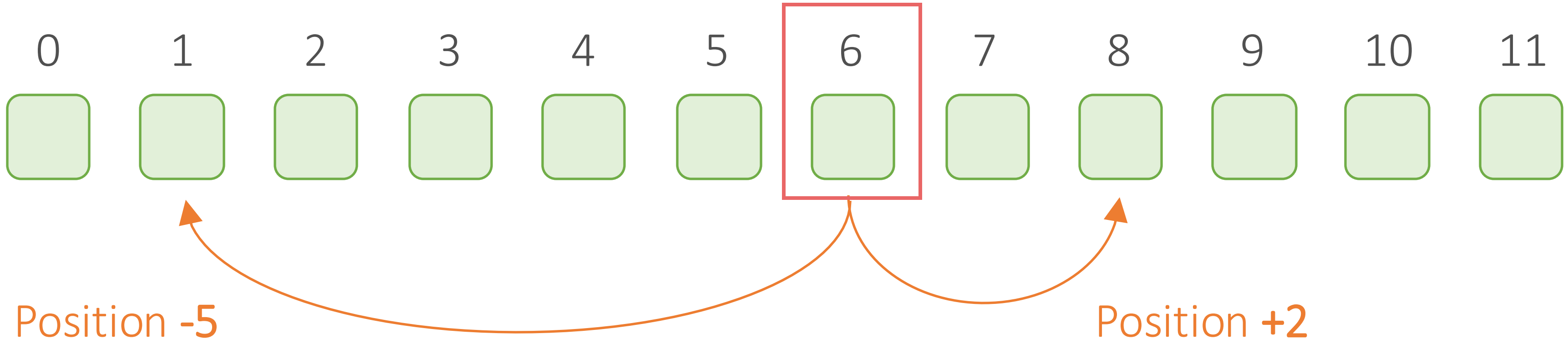
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Absolute Position



Relative Position





Why Relative Position?

- More contextual awareness
  - Position -4: 4 position before this word
  - Position +3: 4 position after this word
- Generalization to longer sequences

Relative Position

$r_{i,j}$ j

i

0	+1	+2	+3	+4	+5	+6	+7	+8	+9
-1	0	+1	+2	+3	+4	+5	+6	+7	+8
-2	-1	0	+1	+2	+3	+4	+5	+6	+7
-3	-2	-1	0	+1	+2	+3	+4	+5	+6
-4	-3	-2	-1	0	+1	+2	+3	+4	+5
-5	-4	-3	-2	-1	0	+1	+2	+3	+4
-6	-5	-4	-3	-2	-1	0	+1	+2	+3
-7	-6	-5	-4	-3	-2	-1	0	+1	+2
-8	-7	-6	-5	-4	-3	-2	-1	0	+1
-9	-8	-7	-6	-5	-4	-3	-2	-1	0

Relative Position with Clipping

$r_{i,j}$

j

i

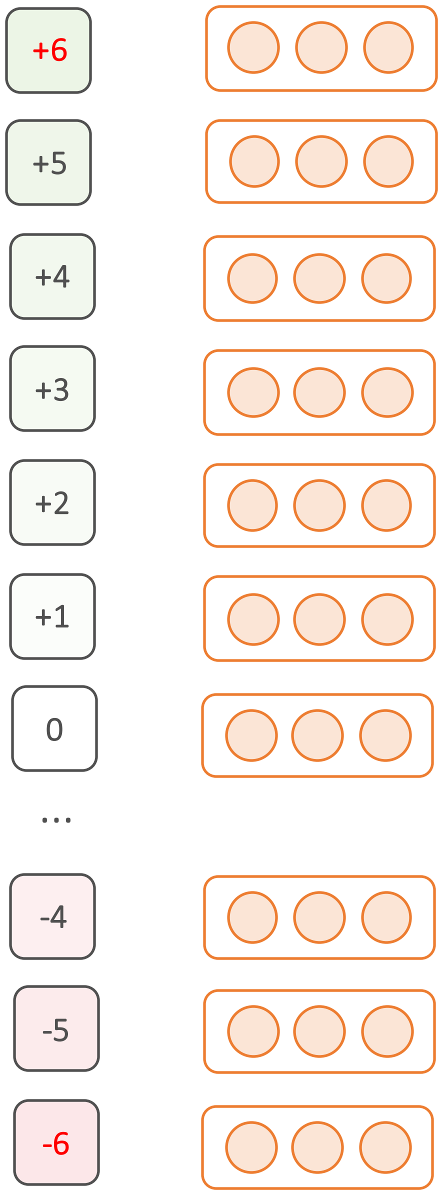
0	+1	+2	+3	+4	+5	+6	+6	+6	+6
-1	0	+1	+2	+3	+4	+5	+6	+6	+6
-2	-1	0	+1	+2	+3	+4	+5	+6	+6
-3	-2	-1	0	+1	+2	+3	+4	+5	+6
-4	-3	-2	-1	0	+1	+2	+3	+4	+5
-5	-4	-3	-2	-1	0	+1	+2	+3	+4
-6	-5	-4	-3	-2	-1	0	+1	+2	+3
-6	-6	-5	-4	-3	-2	-1	0	+1	+2
-6	-6	-6	-5	-4	-3	-2	-1	0	+1
-6	-6	-6	-6	-5	-4	-3	-2	-1	0

Limited relative positions

Map Relative Positions to Embeddings

$r_{i,j}$ (j →, i ↓)

0	+1	+2	+3	+4	+5	+6	+6	+6	+6
-1	0	+1	+2	+3	+4	+5	+6	+6	+6
-2	-1	0	+1	+2	+3	+4	+5	+6	+6
-3	-2	-1	0	+1	+2	+3	+4	+5	+6
-4	-3	-2	-1	0	+1	+2	+3	+4	+5
-5	-4	-3	-2	-1	0	+1	+2	+3	+4
-6	-5	-4	-3	-2	-1	0	+1	+2	+3
-6	-6	-5	-4	-3	-2	-1	0	+1	+2
-6	-6	-6	-5	-4	-3	-2	-1	0	+1
-6	-6	-6	-6	-5	-4	-3	-2	-1	0



Self-Attention

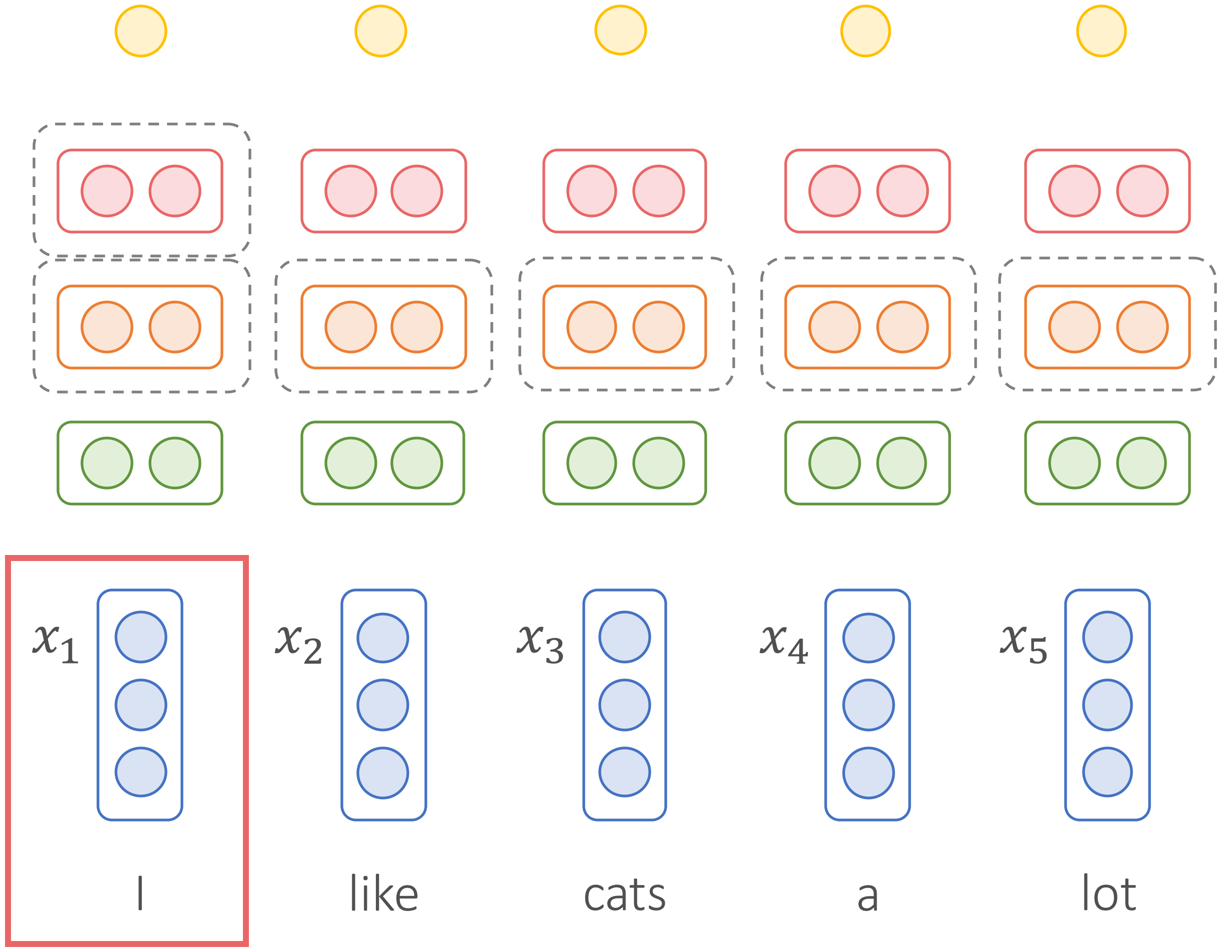
$$\alpha_{1,i} = \text{softmax}\left(\frac{q_1 \cdot k_i}{\sqrt{d}}\right)$$

Normalized
Attention Scores

Query $q_i = W^Q x_i$

Key $k_i = W^K x_i$

Value $v_i = W^V x_i$



Self-Attention

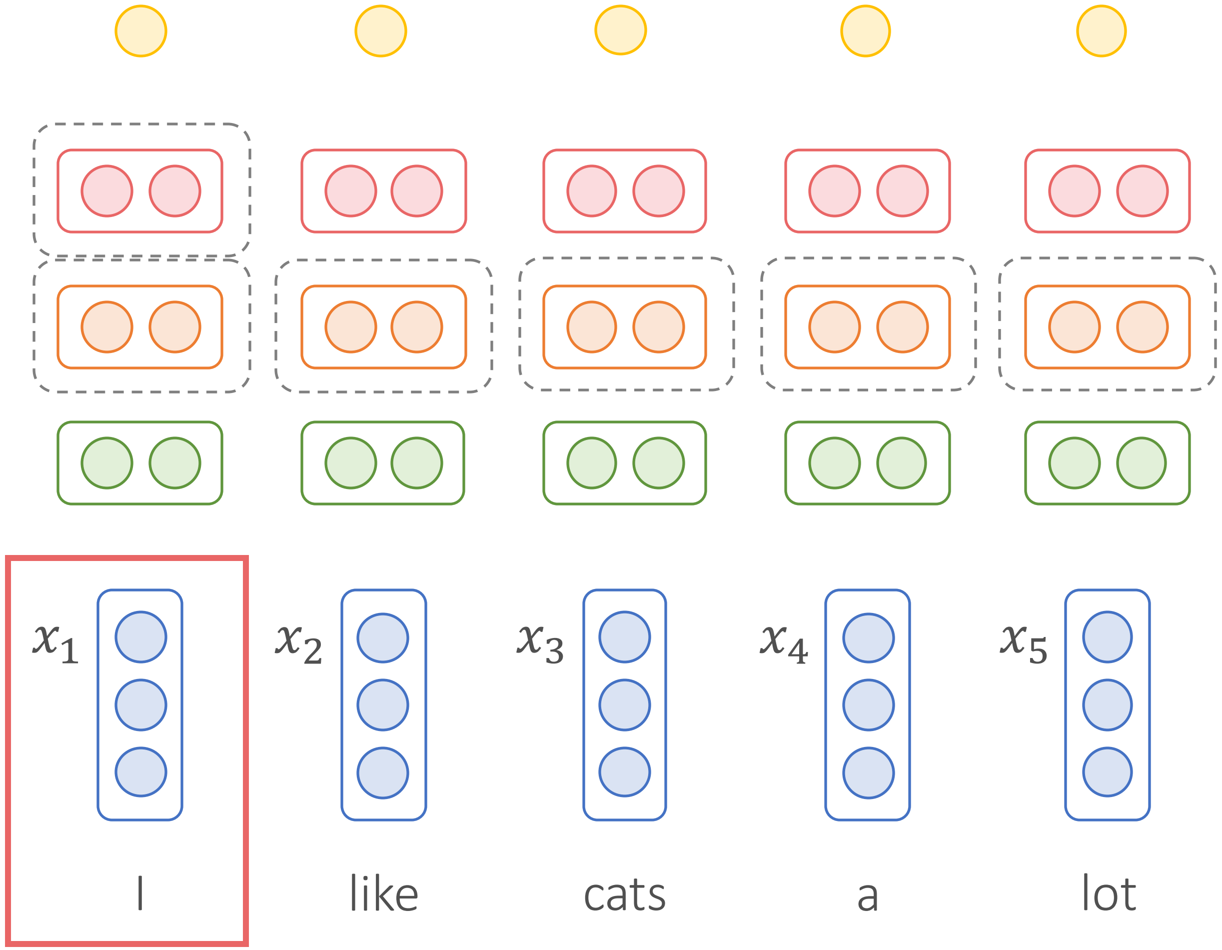
$$\alpha_{1,i} = \mathrm{softmax}\left(\frac{W^Q x_1 \cdot W^K x_i}{\sqrt{d}}\right)$$

Normalized
Attention Scores

Query $q_i = W^Q x_i$

Key $k_i = W^K x_i$

Value $v_i = W^V x_i$



Self-Attention with Relative Position Embeddings

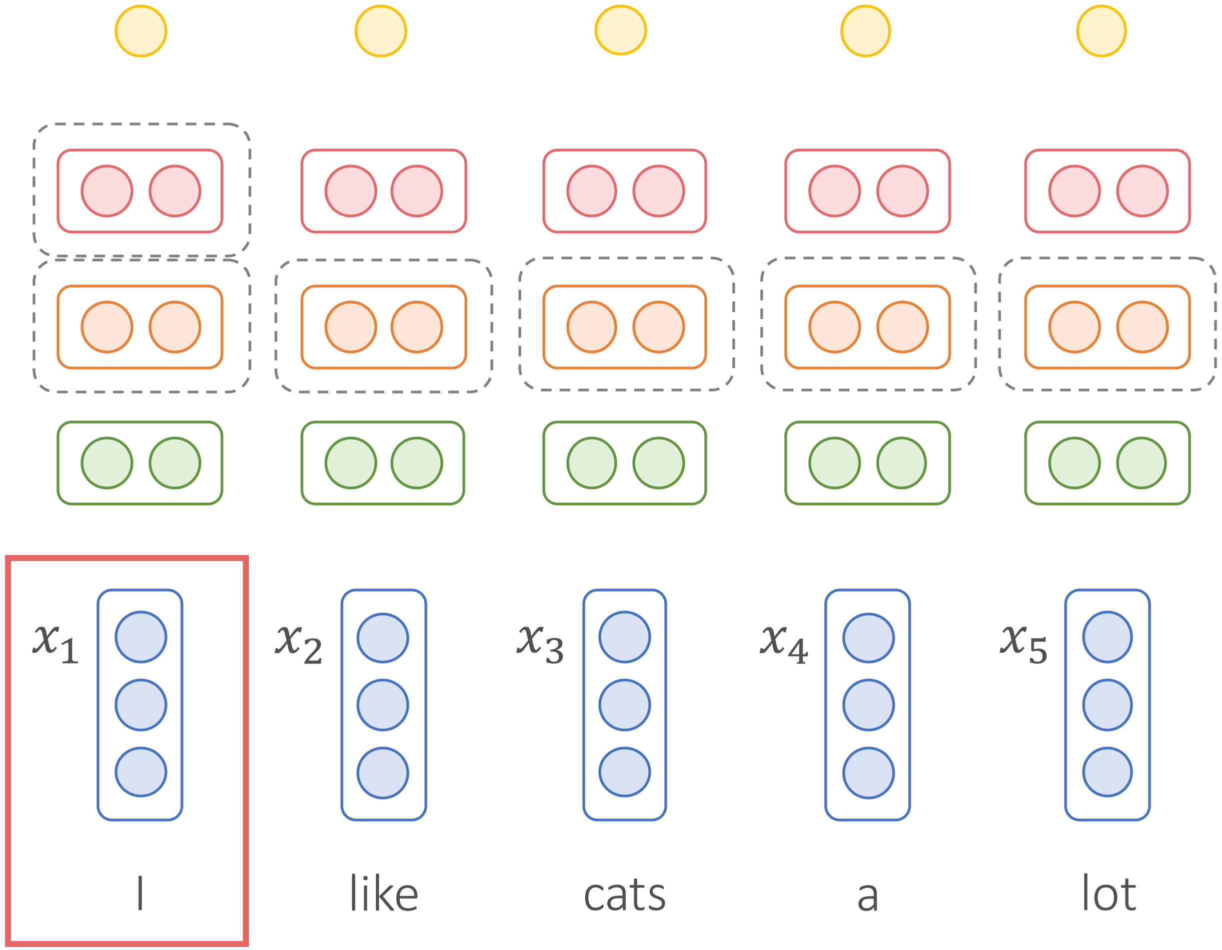
$$\alpha_{1,i} = \text{softmax}\left(\frac{W^Q x_1 \cdot W^K\left(x_i + RE(r_{1,i})\right)}{\sqrt{d}}\right)$$

Normalized Attention Scores

Query $q_i = W^Q x_i$

Key $k_i = W^K x_i$

Value $v_i = W^V x_i$



Self-Attention with Relative Position Embeddings

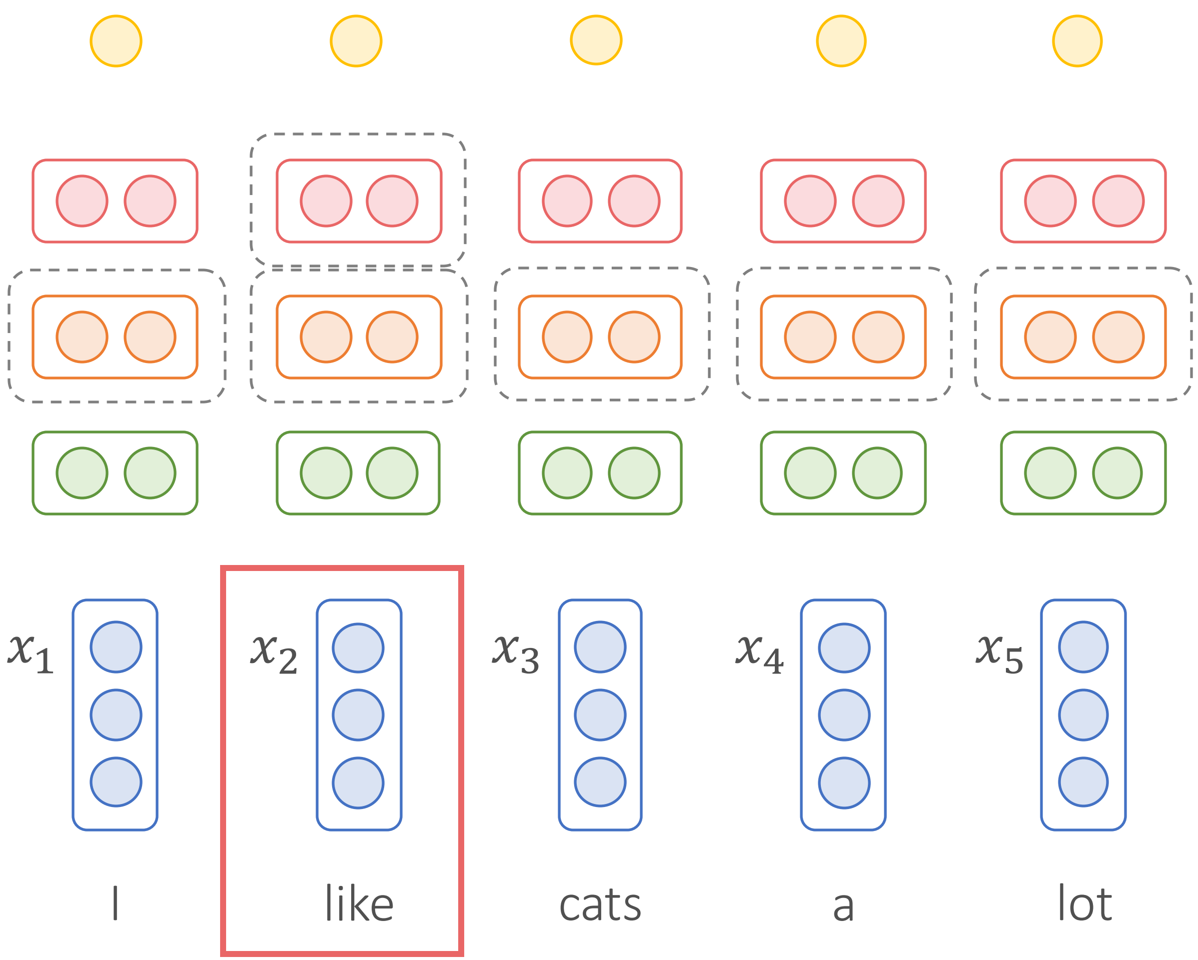
$$\alpha_{2,i} = \text{softmax}\left(\frac{W^Q x_2 \cdot W^K\left(x_i + RE(r_{2,i})\right)}{\sqrt{d}}\right)$$

Normalized Attention Scores

Query $q_i = W^Q x_i$

Key $k_i = W^K x_i$

Value $v_i = W^V x_i$



Relative Positions for Machine Translation

Model	Position Information	EN-DE BLEU	EN-FR BLEU
Transformer (base)	Absolute Position Representations	26.5	38.2
Transformer (base)	Relative Position Representations	**26.8**	**38.7**
Transformer (big)	Absolute Position Representations	27.9	41.2
Transformer (big)	Relative Position Representations	**29.2**	**41.5**

Lecture Plan

- Transformers
  - Encoder
  - Decoder
  - Encoder-Decoder
- Transformers Variants
  - Longformer
  - Relative Positional Encoding
  - RoFormer

RoFormer

- Improved version of relative positional encoding
  - Rotary Position Embedding (RoPE)
- Most advanced large language models use RoPE

Self-Attention with Relative Position Embeddings

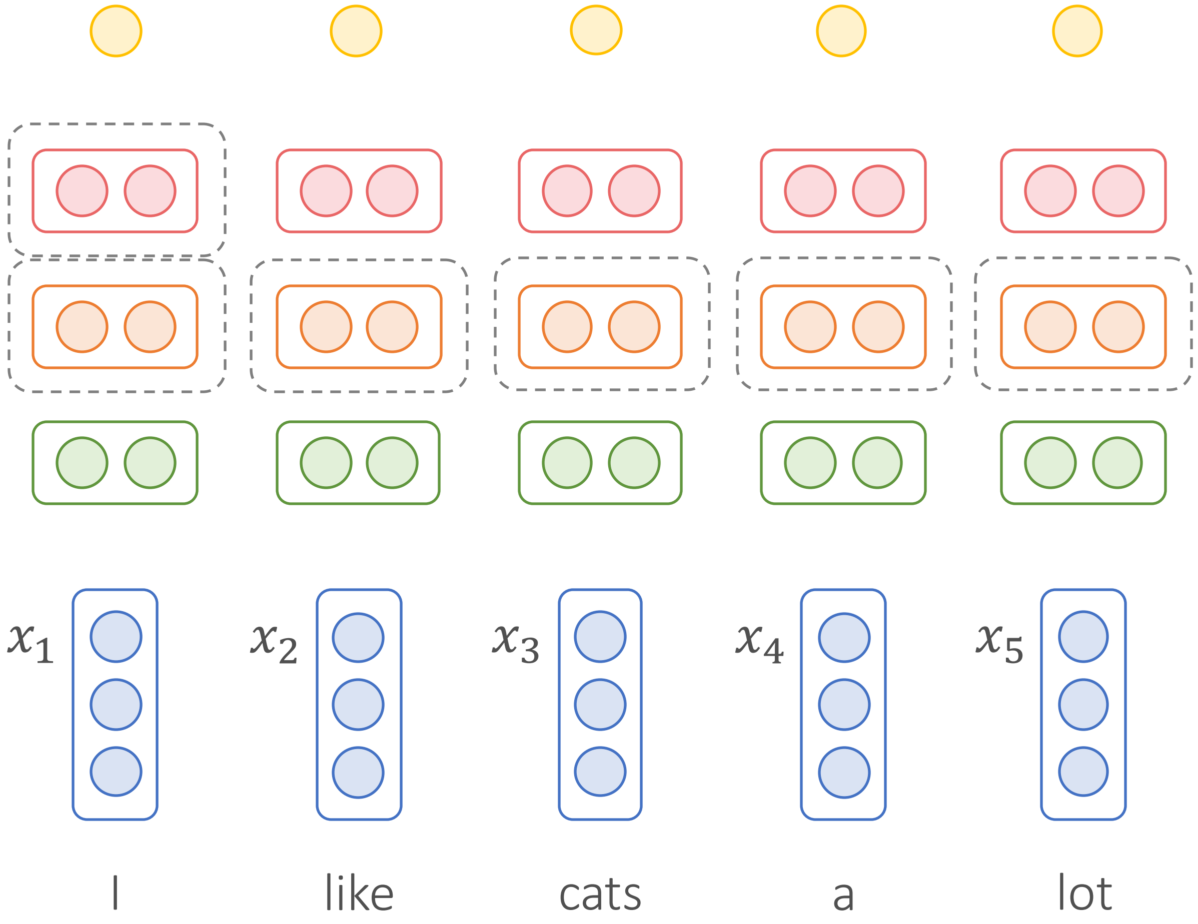
$$\alpha_{m,n} = \text{softmax}\left(\frac{W^Q x_m \cdot W^K (x_n + RE(r_{m,n}))}{\sqrt{d}}\right)$$

Normalized Attention Scores

Query $q_i = W^Q x_i$

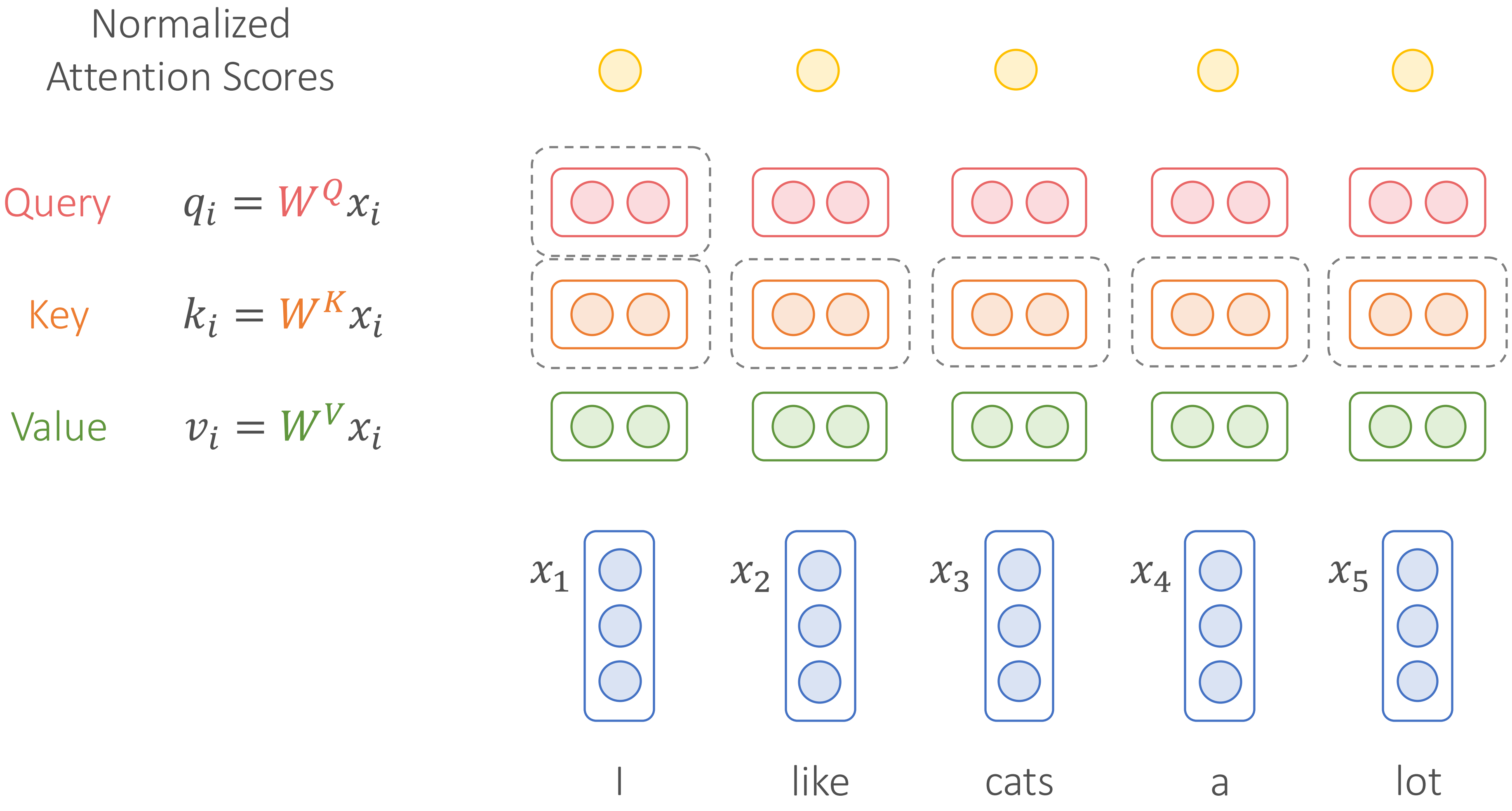
Key $k_i = W^K x_i$

Value $v_i = W^V x_i$



Self-Attention with RoPE (In 2D Case)

$$\alpha_{m,n} = \text{softmax}\left(\frac{\langle (W^Q x_m) e^{im\theta} \cdot (W^K x_n) e^{in\theta} \rangle}{\sqrt{d}}\right)$$



Self-Attention with RoPE (In 2D Case)

Equivalent to rotate $W^Q x_m$ with angle $m\theta$

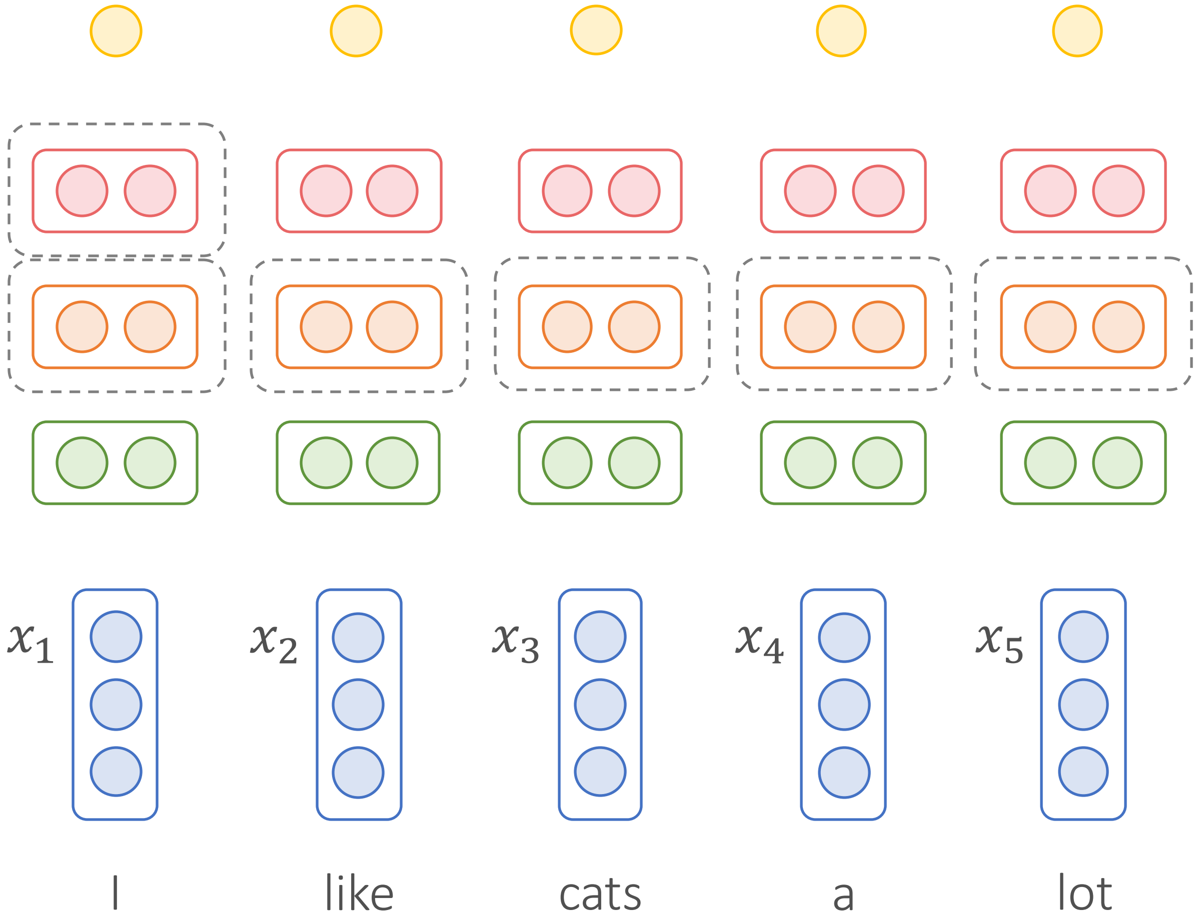
$$\alpha_{m,n} = \mathrm{softmax}\left(\frac{\langle \boxed{(W^Q x_m)e^{im\theta}} \cdot (W^K x_n)e^{in\theta}\rangle}{\sqrt{d}}\right)$$

Normalized Attention Scores

Query $q_i = W^Q x_i$

Key $k_i = W^K x_i$

Value $v_i = W^V x_i$



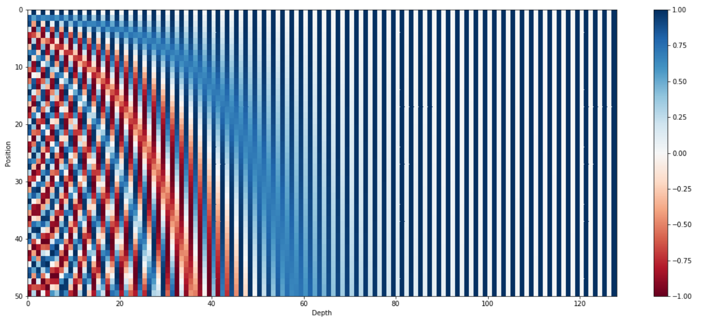
General Form of RoPE

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_{\{q,k\}} \boldsymbol{x}_m$$

Different base angle $\theta_1, \theta_2, \ldots, \theta_{d/2}$

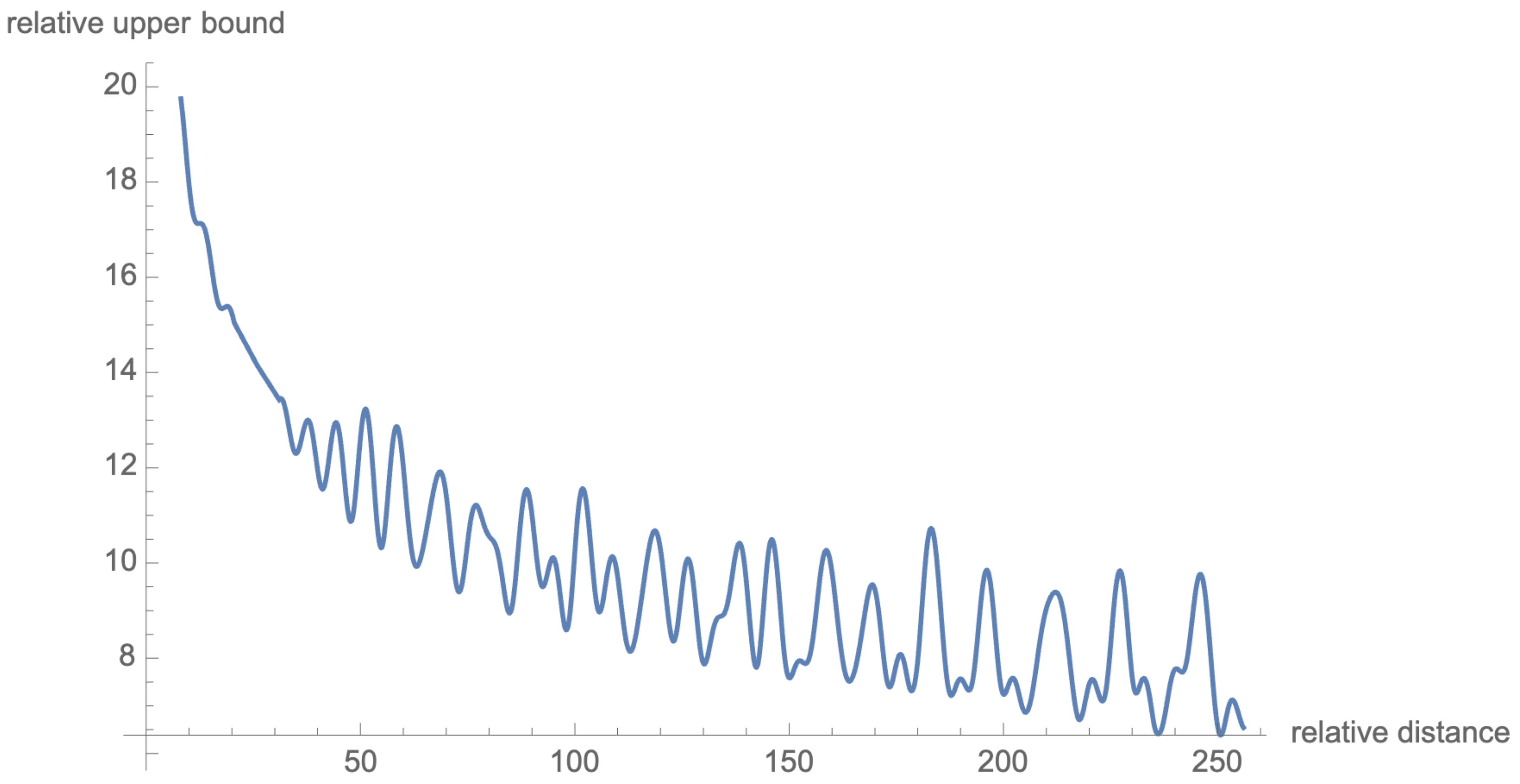
$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = (\boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_q \boldsymbol{x}_m)^\intercal (\boldsymbol{R}^d_{\Theta,n} \boldsymbol{W}_k \boldsymbol{x}_n) = \boldsymbol{x}^\intercal \boldsymbol{W}_q R^d_{\Theta,n-m} \boldsymbol{W}_k \boldsymbol{x}_n$$

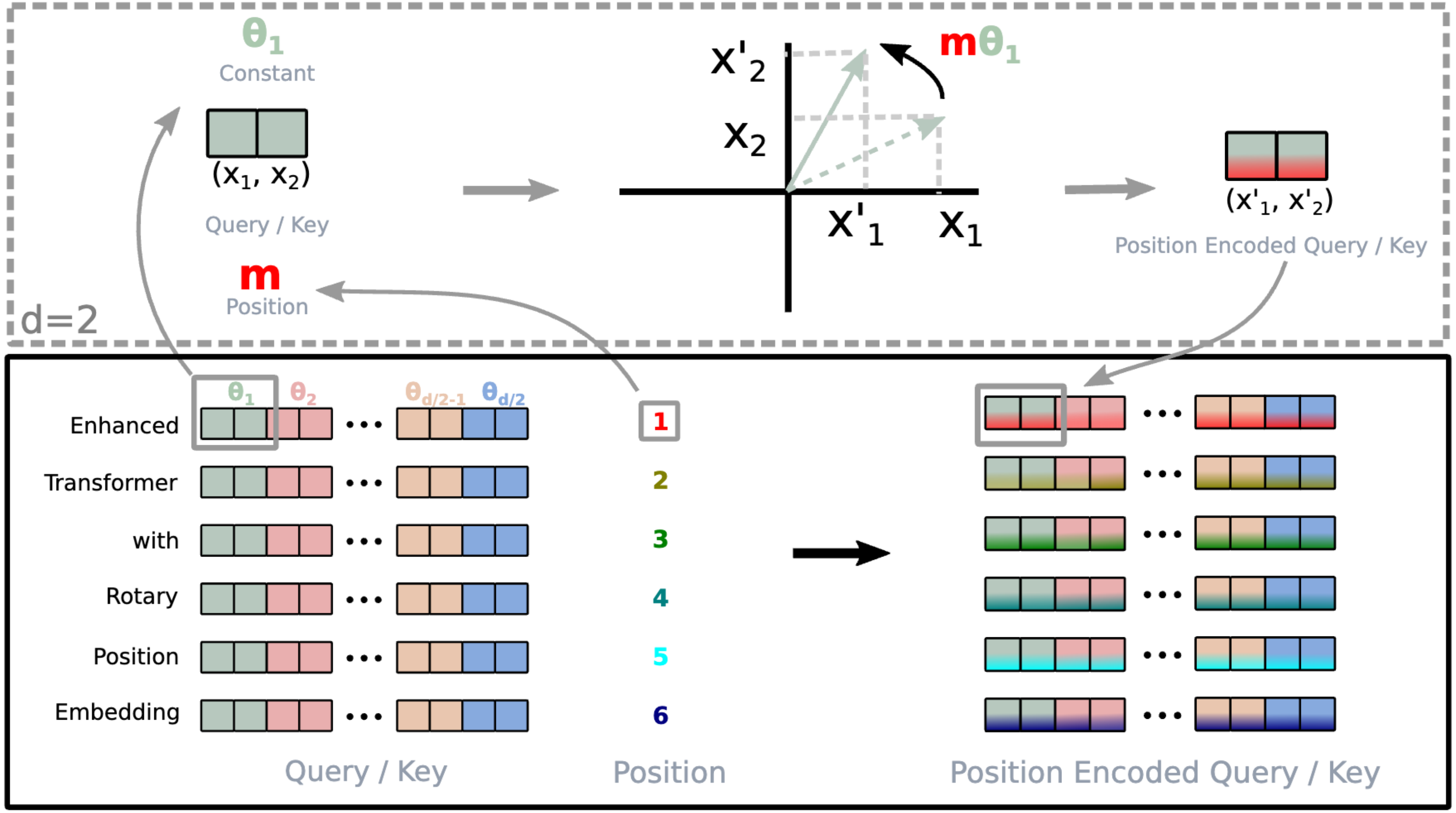


Similar to the idea of using different flipping frequency
for Sinusoidal positional encoding

RoPE Similarity over Position Difference



RoPE Implementation



RoPE Performance

Model	MRPC	SST-2	QNLI	STS-B	QQP	MNLI(m/mm)
BERTDevlin et al. [2019]	88.9	93.5	90.5	85.8	71.2	84.6/83.4
RoFormer	**89.5**	90.7	88.0	**87.0**	**86.4**	80.2/79.8

Lecture Plan

- Transformers
  - Encoder
  - Decoder
  - Encoder-Decoder
- Transformers Variants
  - Longformer
  - Relative Positional Encoding
  - RoFormer