CSCE 689: Special Topics in Trustworthy NLP

Lecture 6: Natural Language Processing Basics (5)

Kuan-Hao Huang
khhuang@tamu.edu

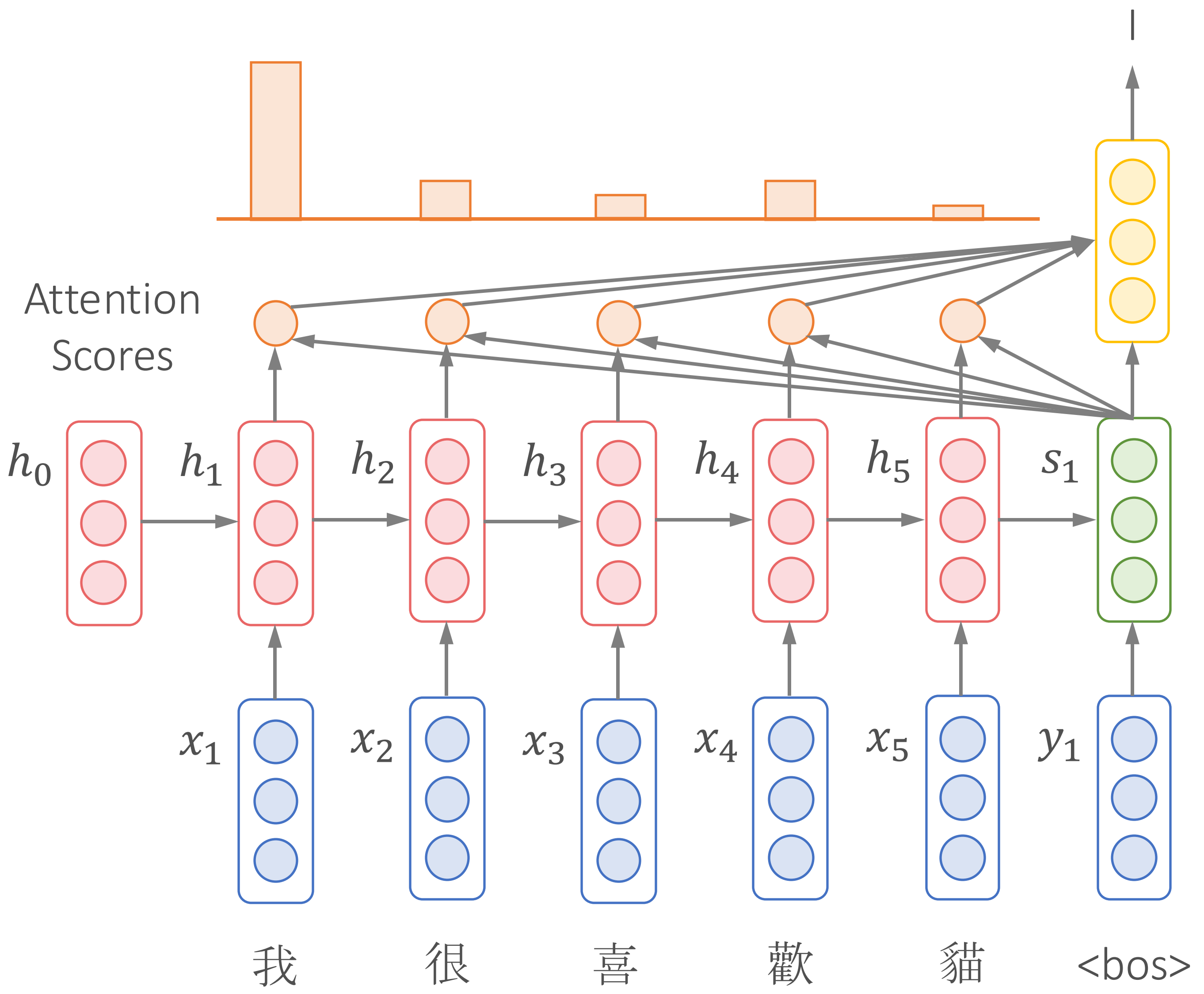
(Some slides adapted from Chris Manning, Karthik Narasimhan)

Presentation Assignment

Lecture Plan

- Natural Language Processing Basics
- Transformers
- Contextualized Representations
- Pre-Training

Recap: LSTM with Attention



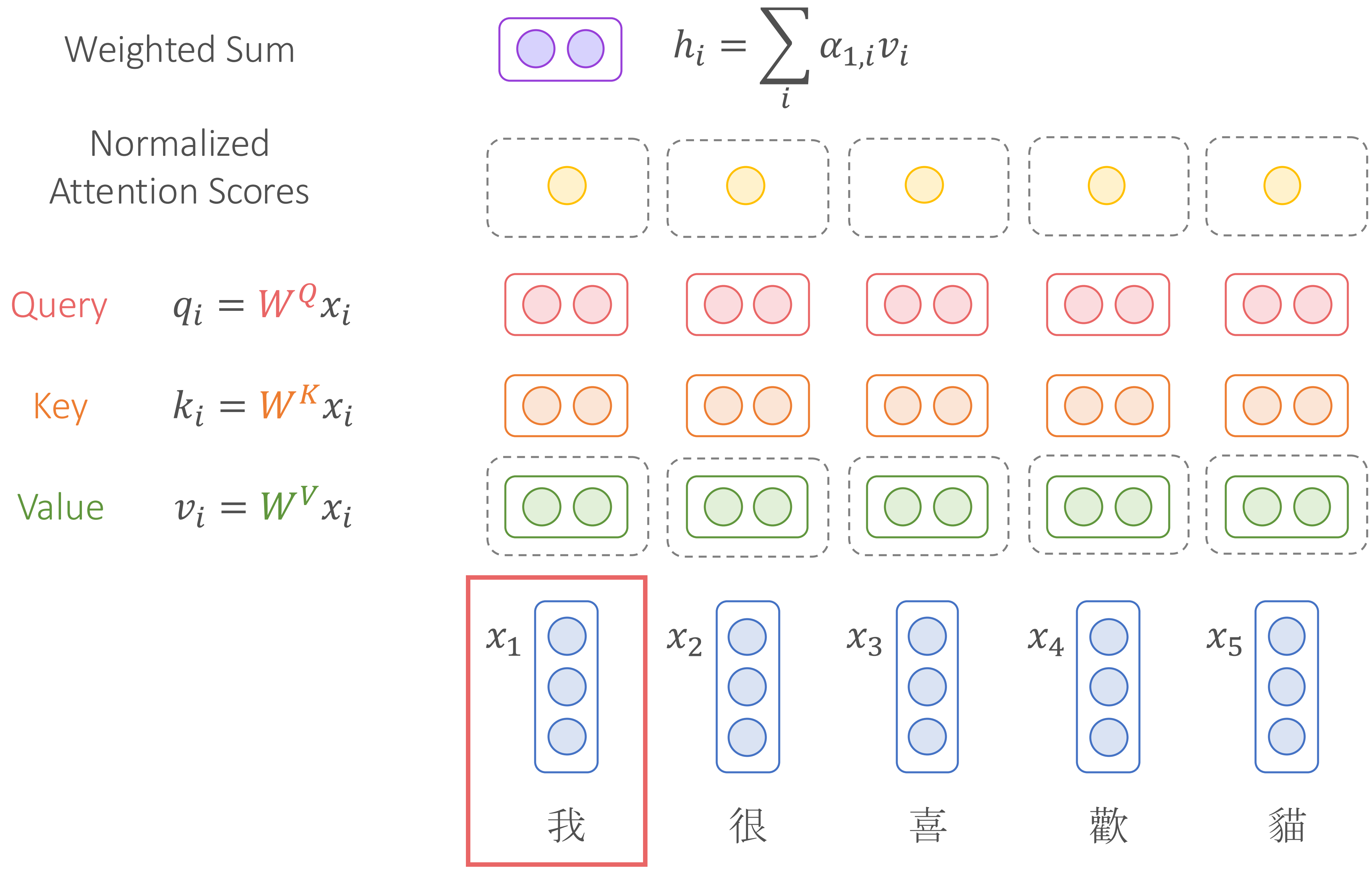
Attention Scores $\alpha_i = h_i^{\mathsf{T}} s_1$

Normalized Attention Scores $\hat{\alpha}_i = \mathrm{softmax}(\alpha_i)$

Weighted Sum $a = \sum_i \hat{\alpha}_i h_i$

Attention Output $\tanh(\mathbf{W}[a; s_1])$

Recap: Self-Attention

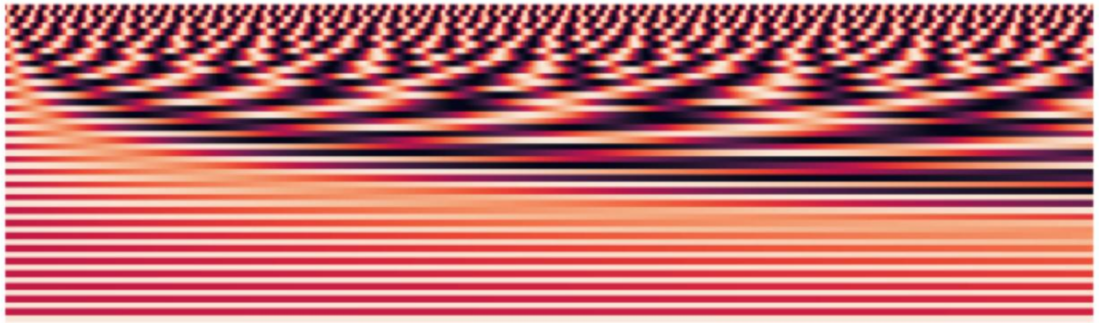


Recap: Positional Encoding

$$x_i \leftarrow x_i + PE_i$$

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$

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

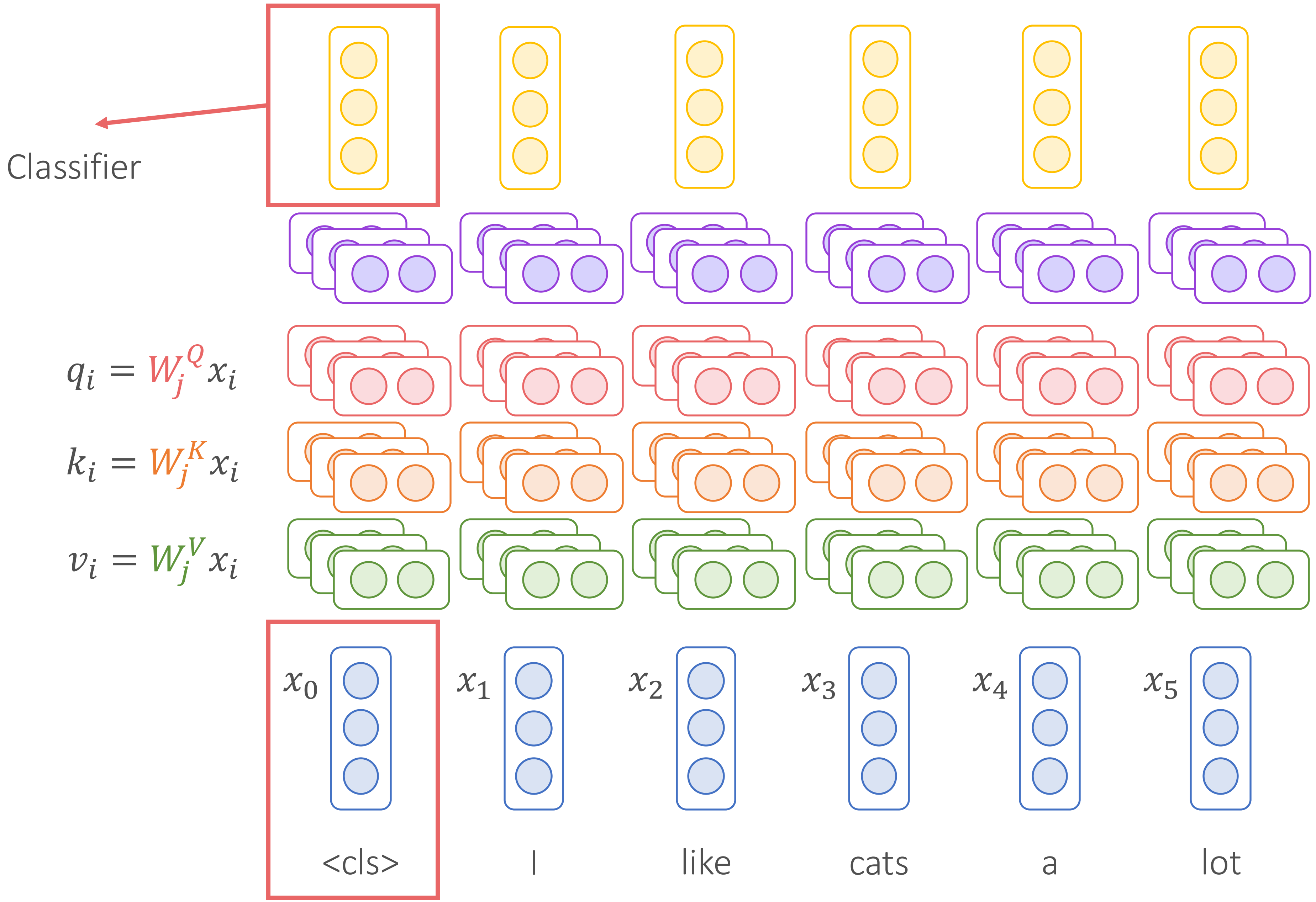


Positional Encoding Matrix with d=4, n=100

Sequence	Index of token, k	i=0	i=0	i=1	i=1
I	0	P_{00}=sin(0) = 0	P_{01}=cos(0) = 1	P_{02}=sin(0) = 0	P_{03}=cos(0) = 1
am	1	P_{10}=sin(1/1) = 0.84	P_{11}=cos(1/1) = 0.54	P_{12}=sin(1/10) = 0.10	P_{13}=cos(1/10) = 1.0
a	2	P_{20}=sin(2/1) = 0.91	P_{21}=cos(2/1) = -0.42	P_{22}=sin(2/10) = 0.20	P_{23}=cos(2/10) = 0.98
Robot	3	P_{30}=sin(3/1) = 0.14	P_{31}=cos(3/1) = -0.99	P_{32}=sin(3/10) = 0.30	P_{33}=cos(3/10) = 0.96

Positional Encoding Matrix for the sequence 'I am a robot'

Transformer For Classification – Using Encoder Only

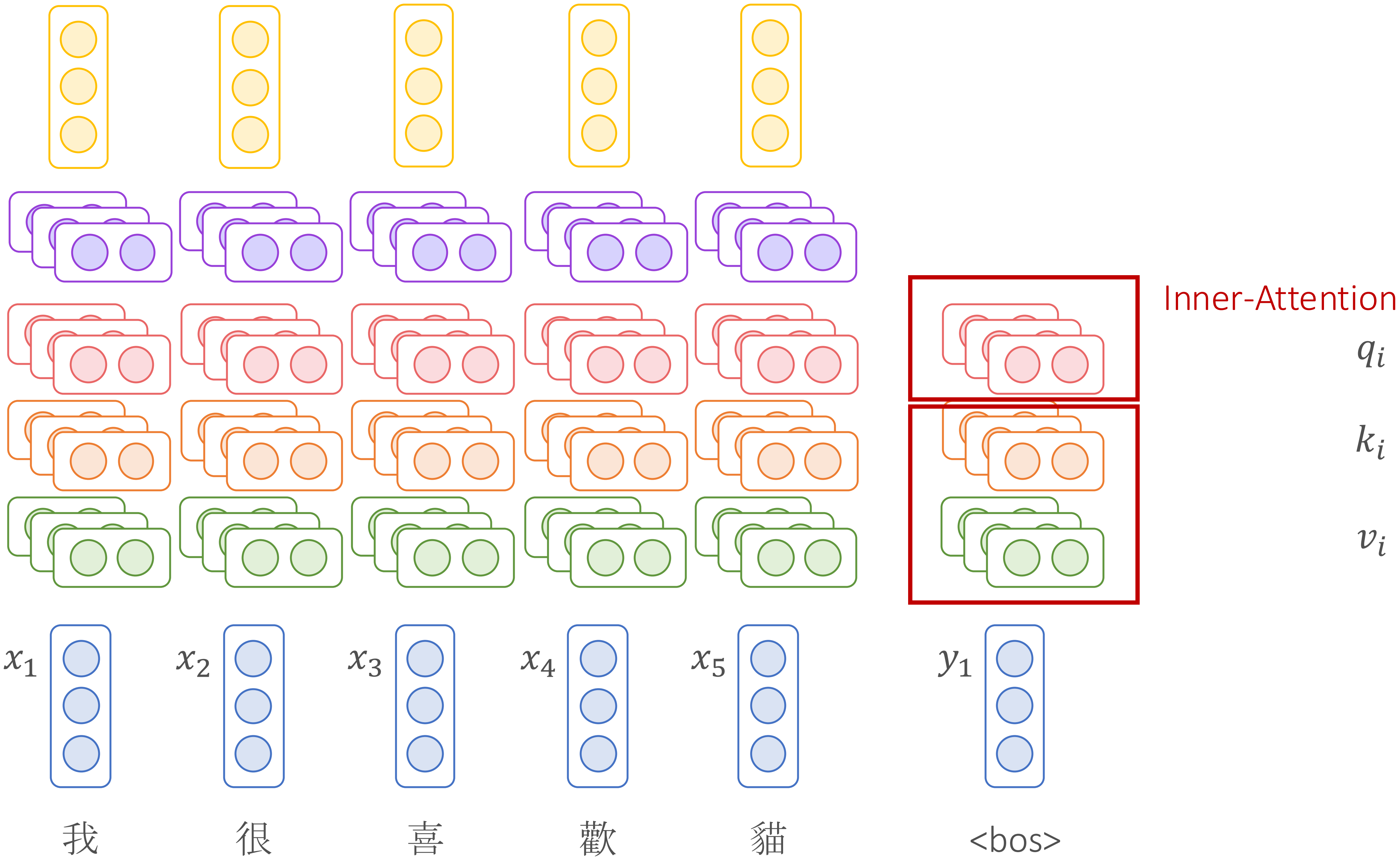


Transformer For Generation

$q_i = W_j^Q x_i$

$k_i = W_j^K x_i$

$v_i = W_j^V x_i$

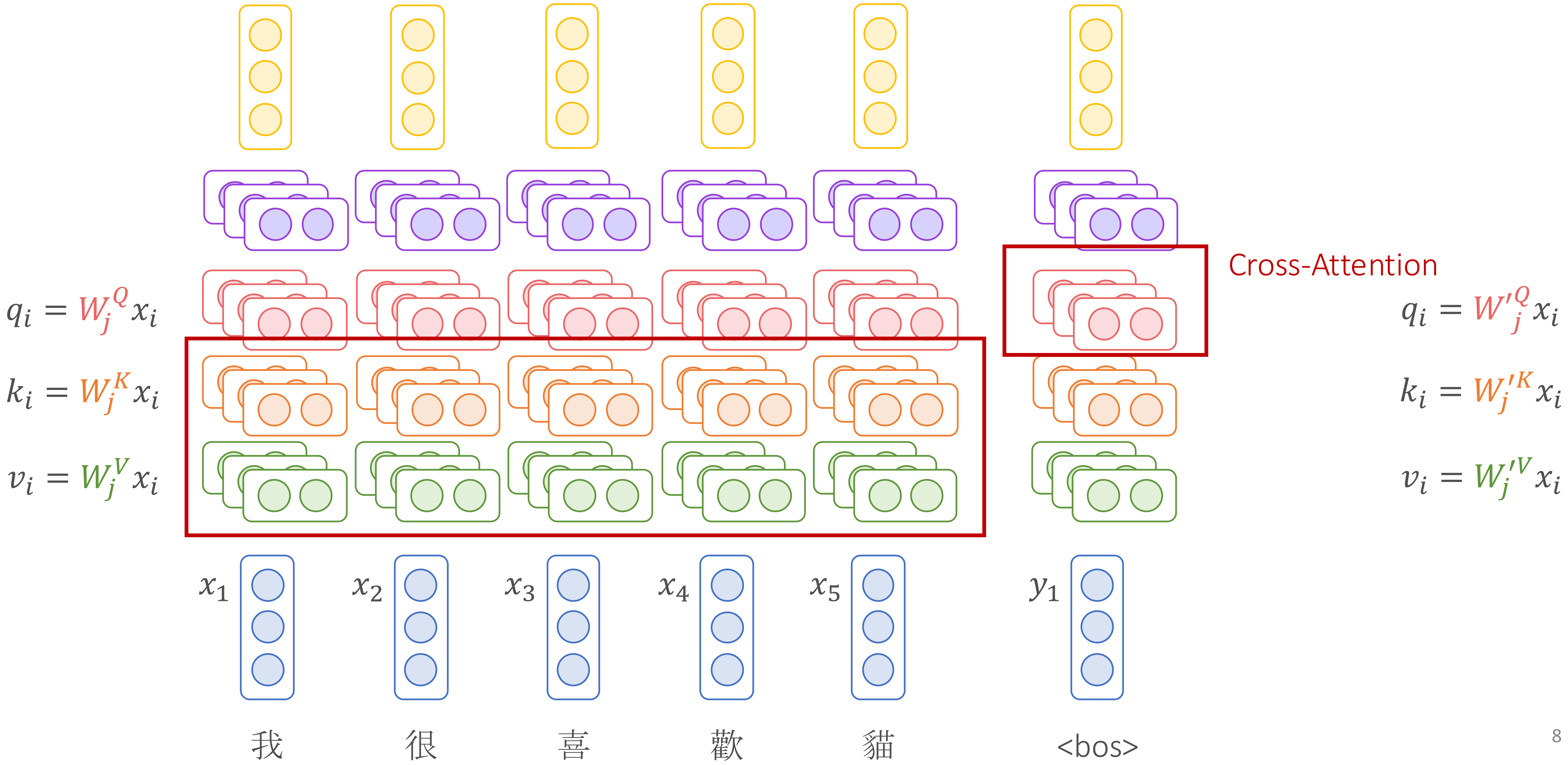


$q_i = W'^Q_j x_i$

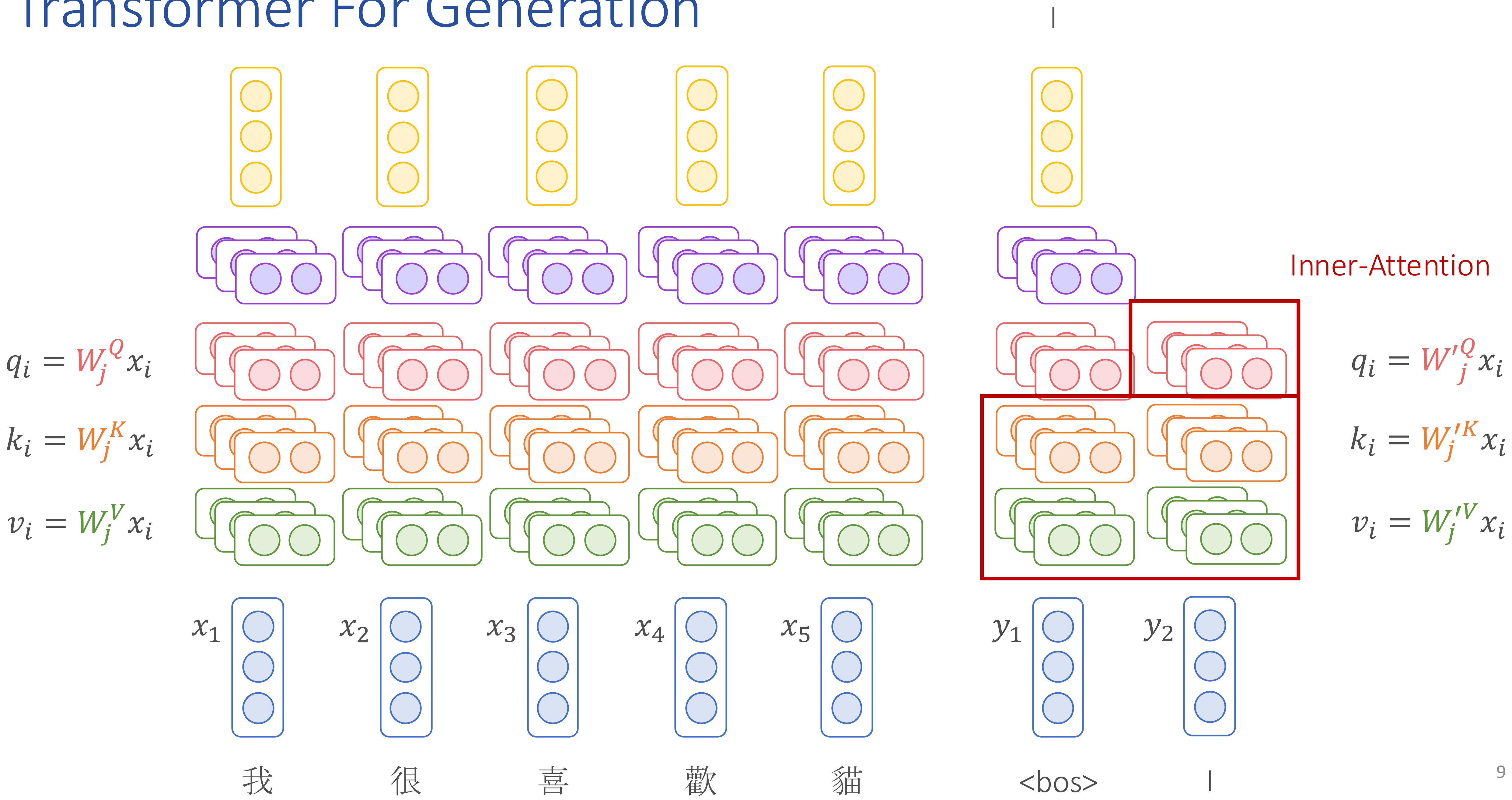
$k_i = W'^K_j x_i$

$v_i = W'^V_j x_i$

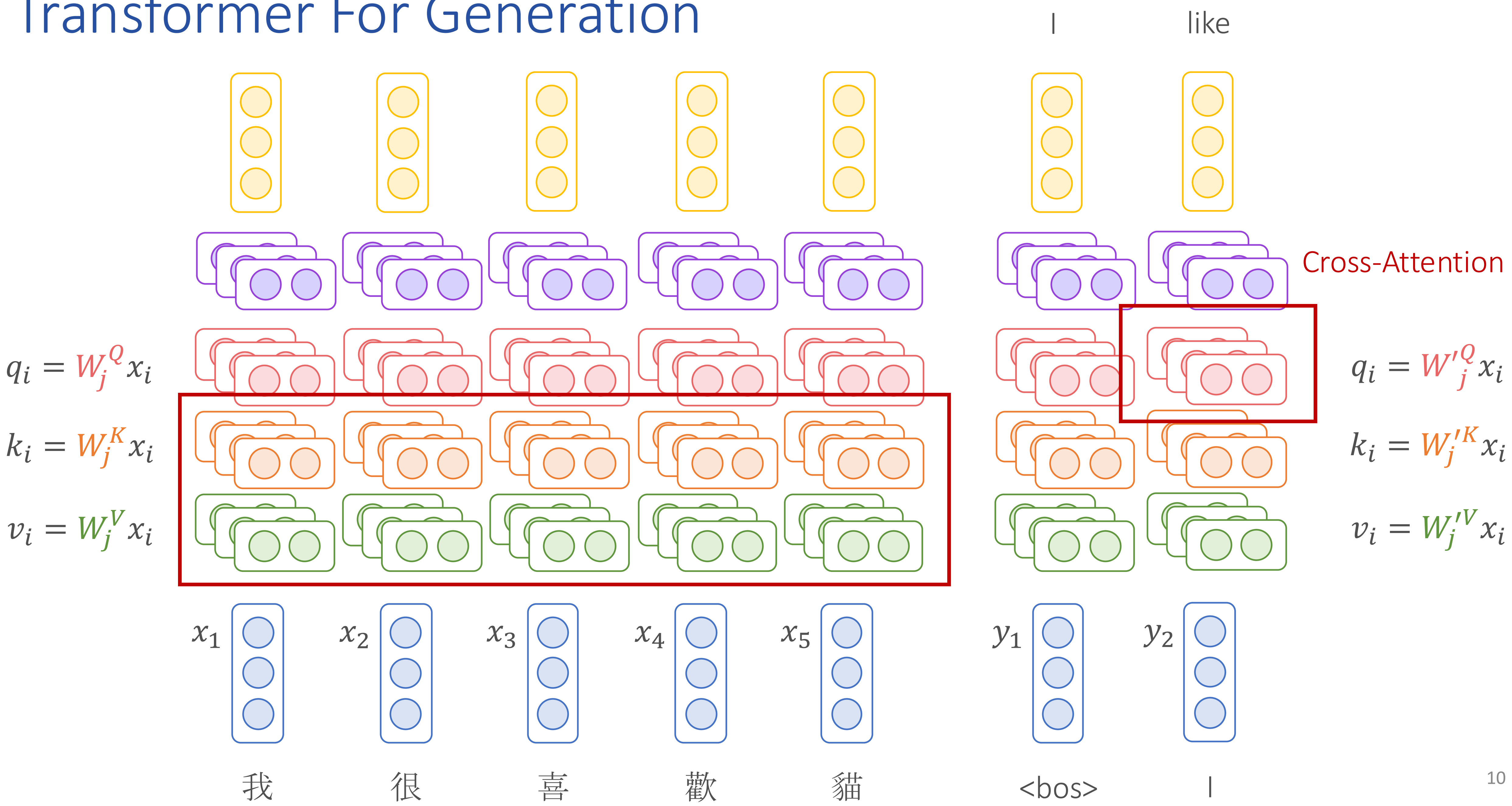
Transformer For Generation



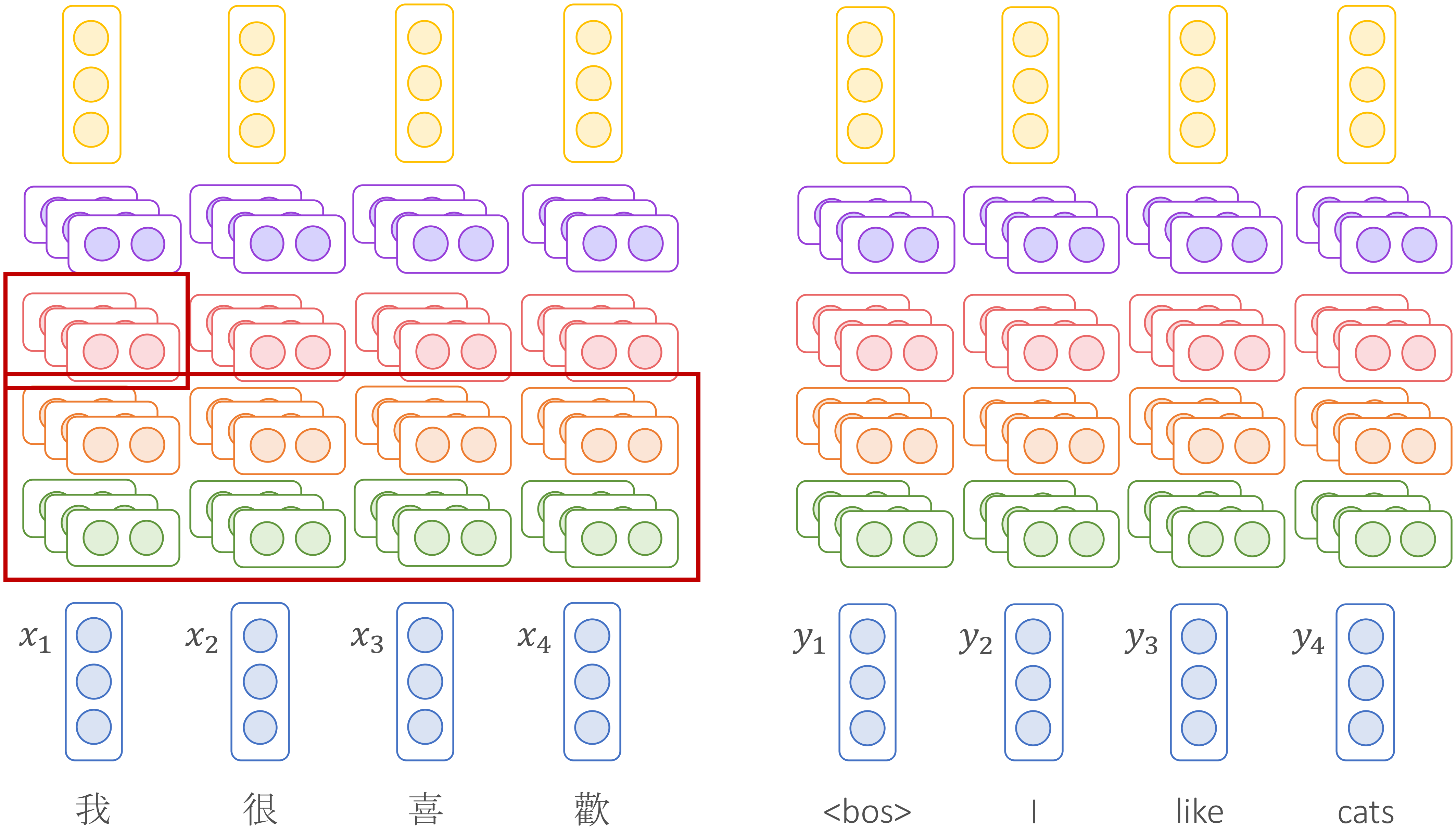
Transformer For Generation



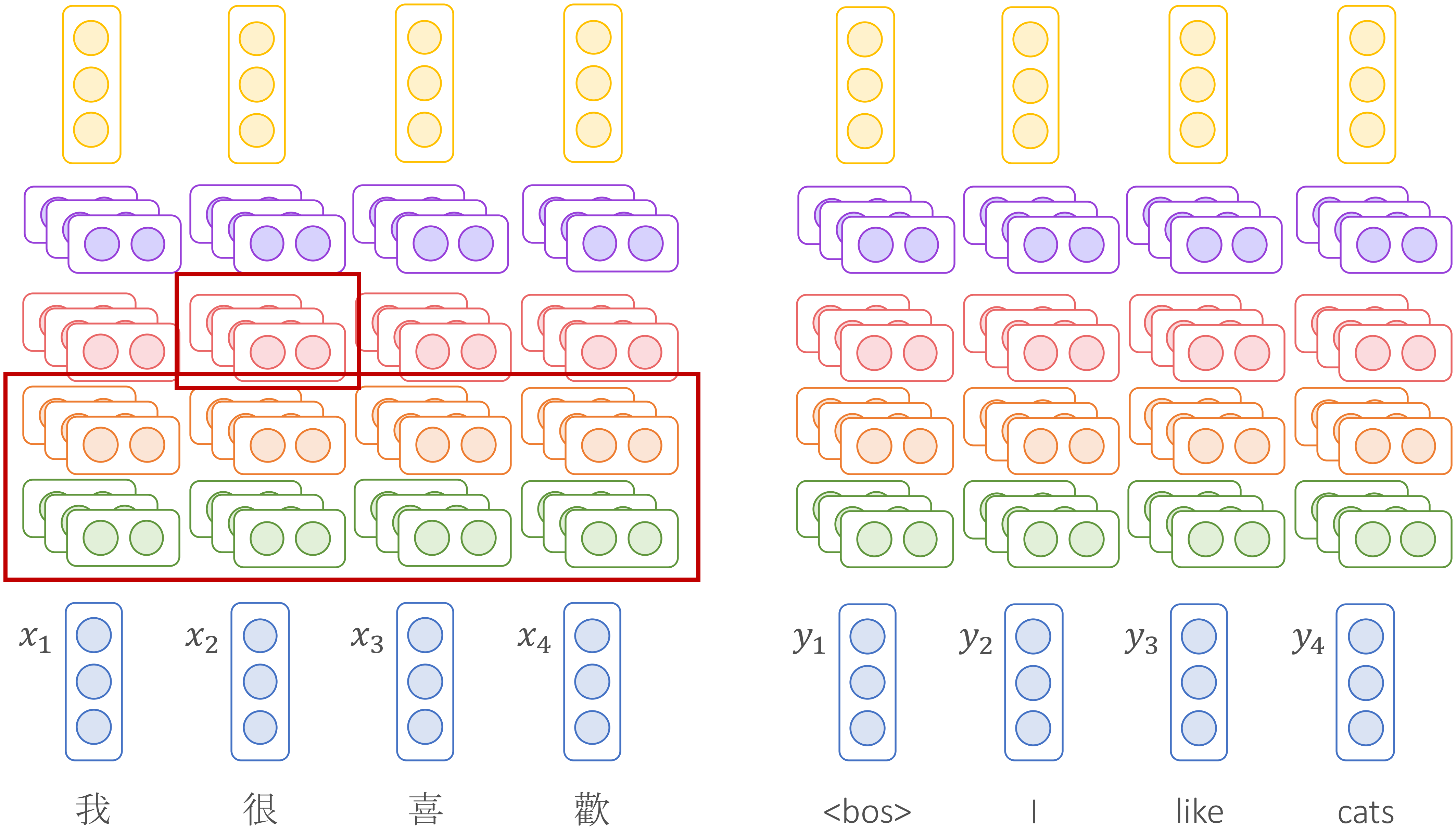
Transformer For Generation



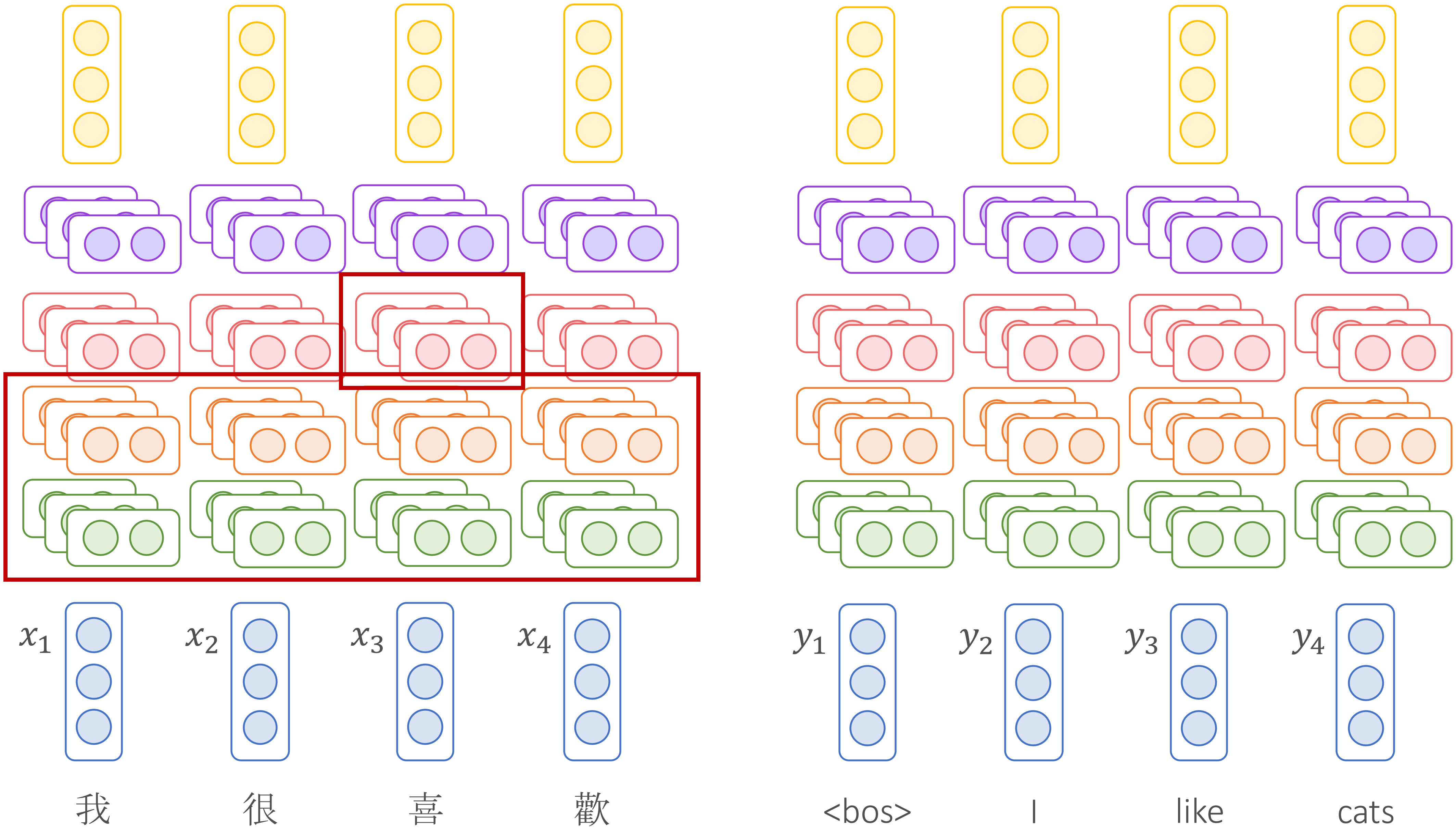
Transformer Encoder vs. Transformer Decoder



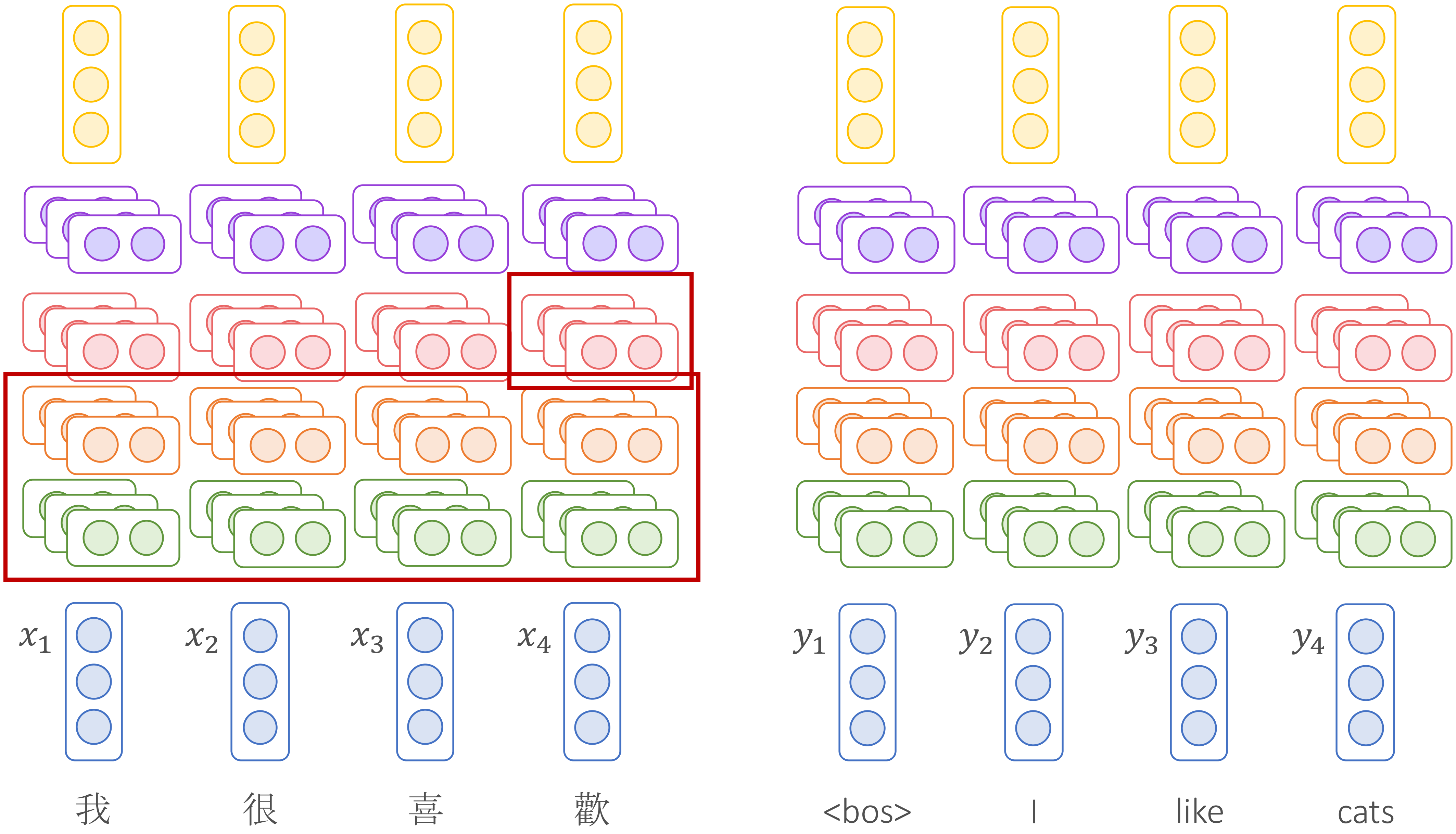
Transformer Encoder vs. Transformer Decoder



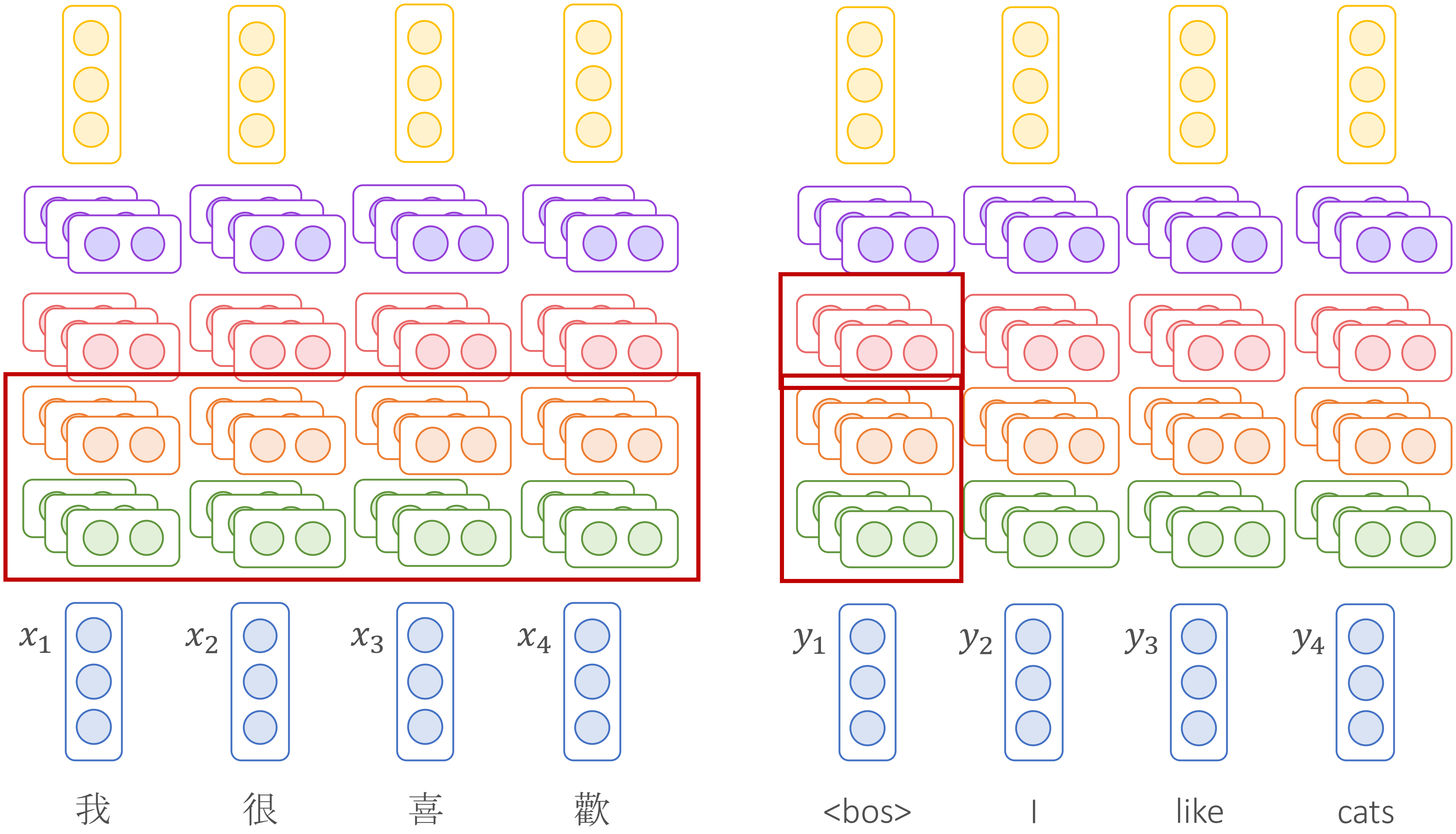
Transformer Encoder vs. Transformer Decoder



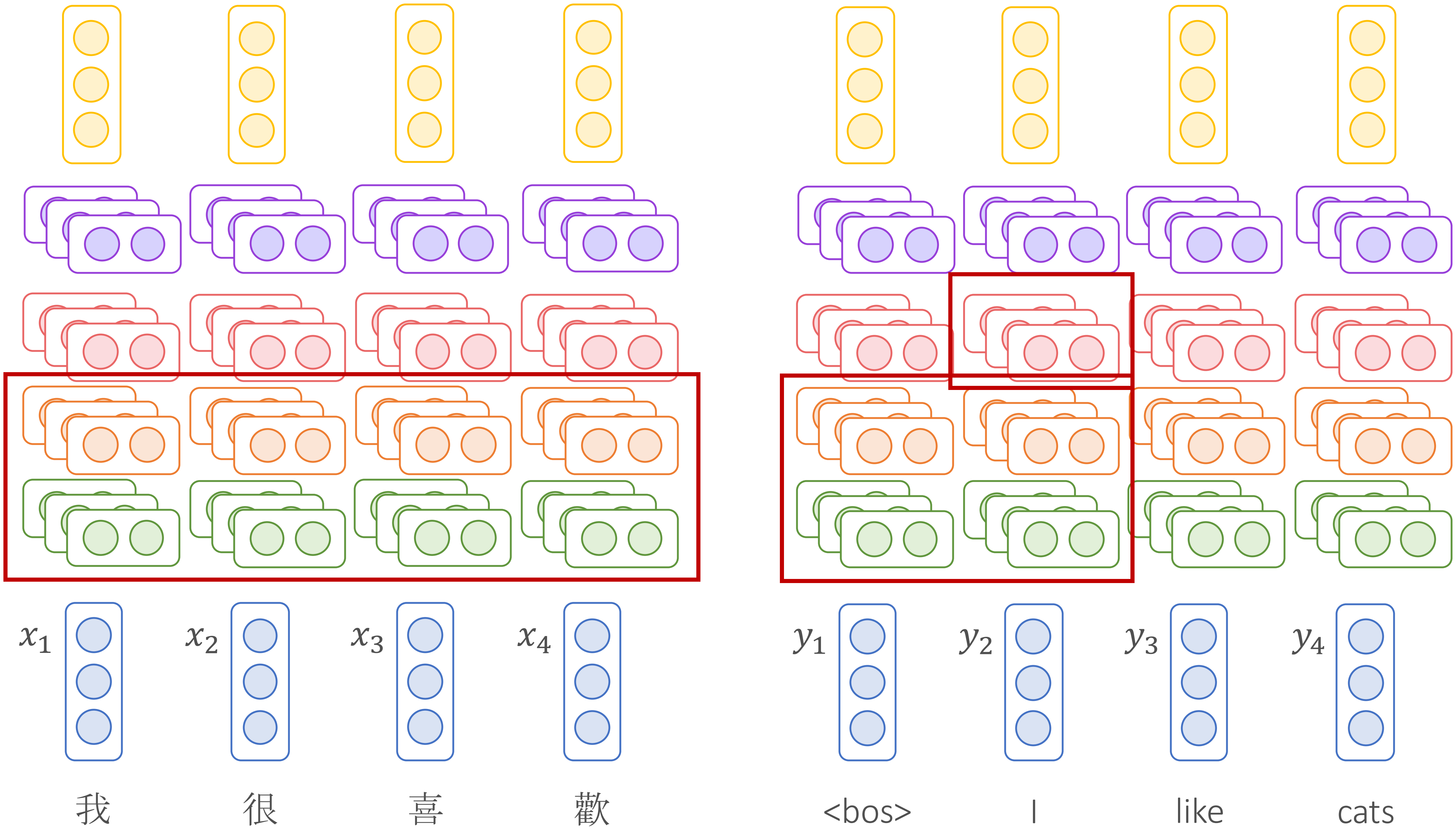
Transformer Encoder vs. Transformer Decoder



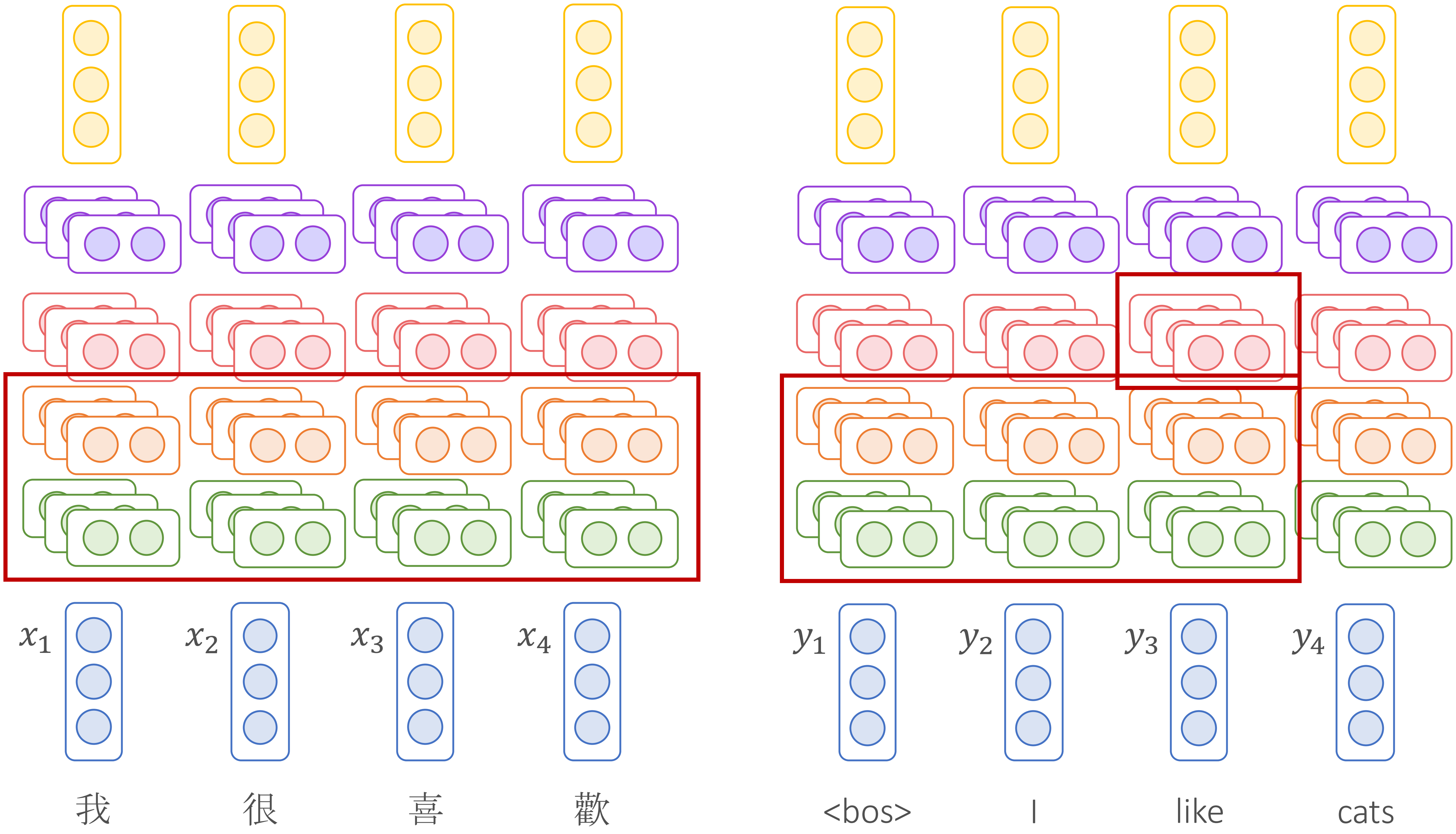
Transformer Encoder vs. Transformer Decoder



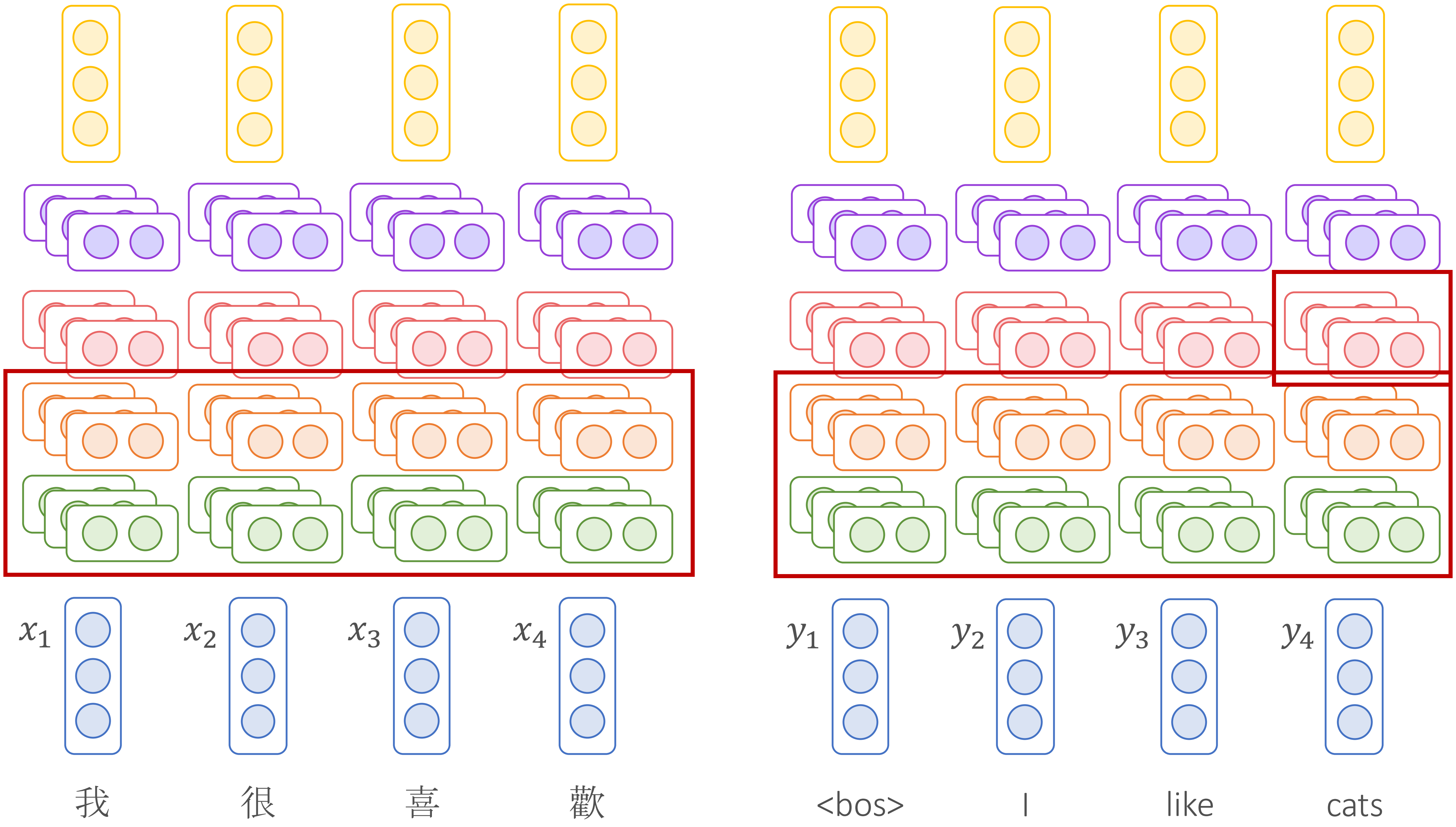
Transformer Encoder vs. Transformer Decoder



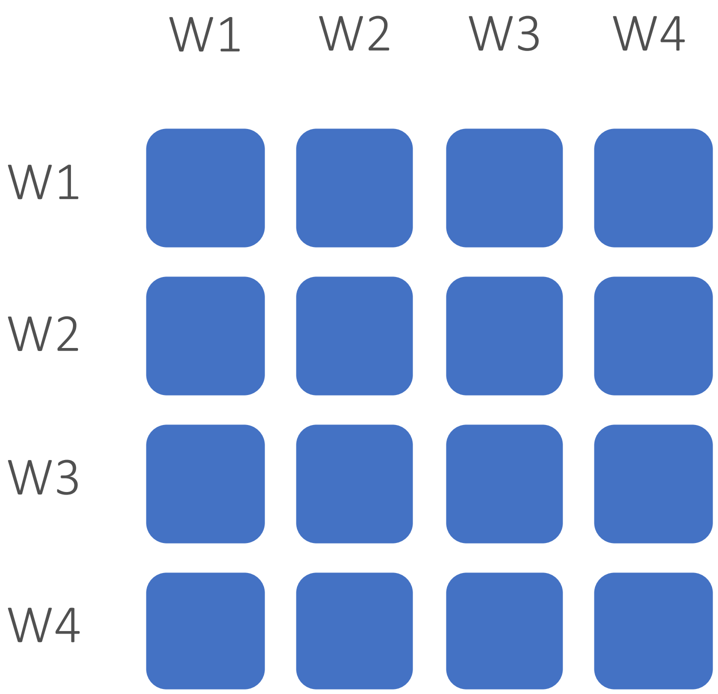
Transformer Encoder vs. Transformer Decoder



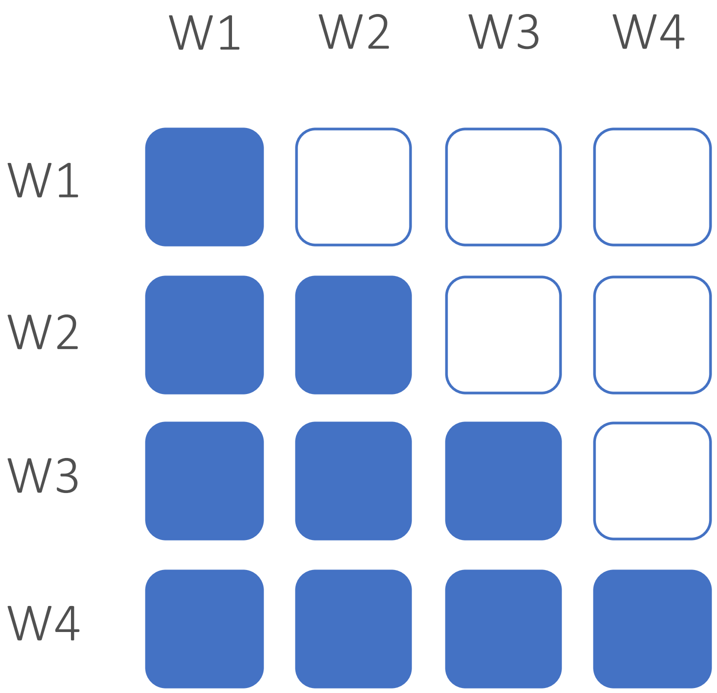
Transformer Encoder vs. Transformer Decoder



Transformer Encoder vs. Transformer Decoder



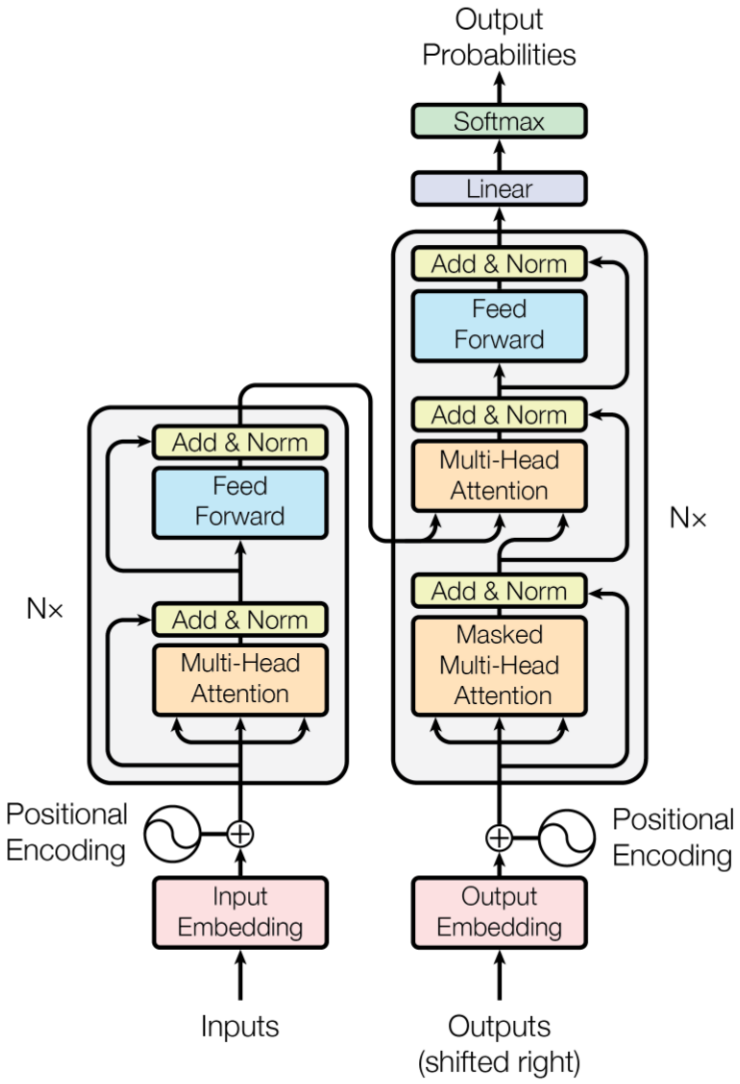
Encoder
Full Attention



Decoder
Causal Attention

Transformers

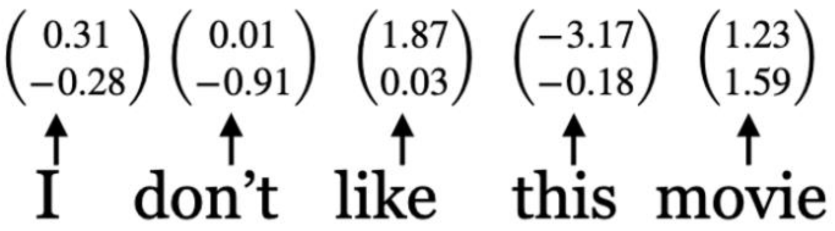
- Main architectures
  - Self-attention
  - Feed forward
  - Positional encoding
- Transformer encoder = a stack of encoder layers
- Transformer decoder = a stack of decoder layers

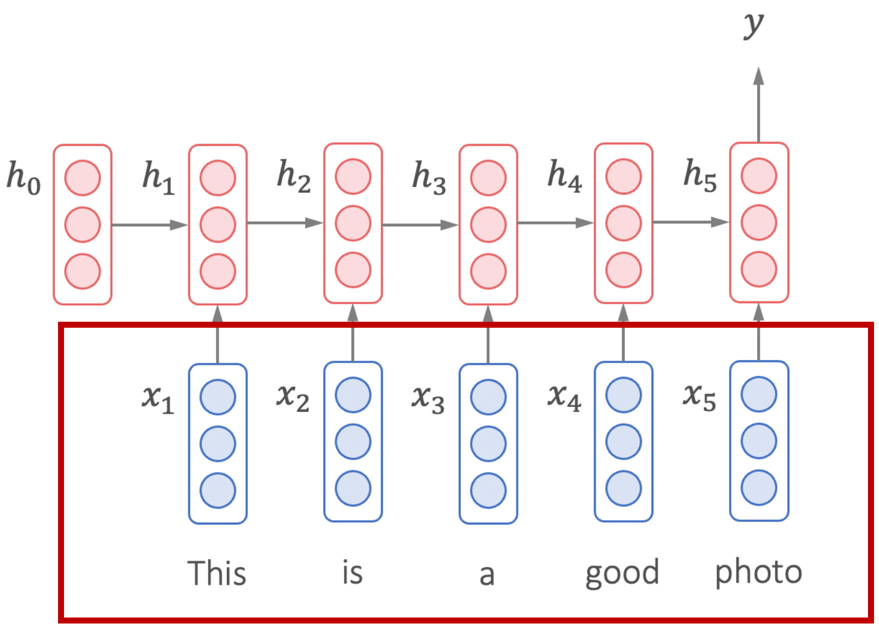


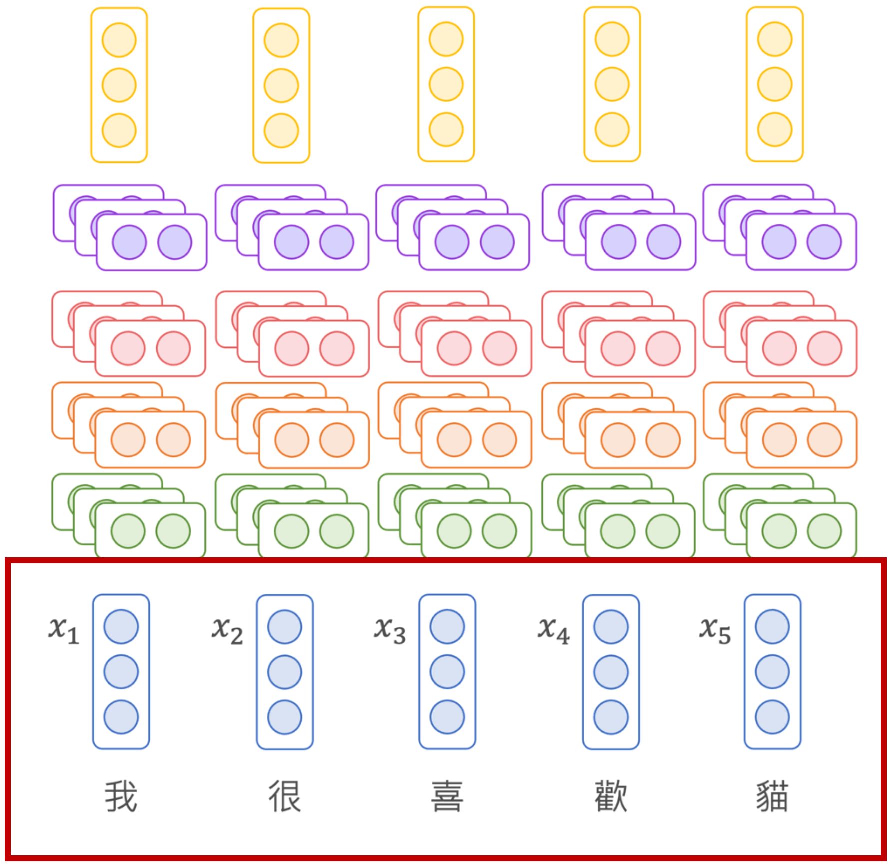
Lecture Plan

- Natural Language Processing Basics
- Transformers
- Contextualized Representations
- Pre-Training and Fine-Tuning

Static Word Embeddings







Static Word Embeddings

- One vector for each word type
- How about words with multiple meanings?

mouse^1 : .... a *mouse* controlling a computer system in 1968.
mouse^2 : .... a quiet animal like a *mouse*
bank^1 : ...a *bank* can hold the investments in a custodial account ...
bank^2 : ...as agriculture burgeons on the east *bank*, the river ...

Contextualized Word Embeddings

- The embeddings of a word should be conditioned on its context

Distributional hypothesis: words that occur in similar contexts tend to have similar meanings

J.R.Firth 1957

- "You shall know a word by the company it keeps"
- One of the most successful ideas of modern statistical NLP!

...government debt problems turning into	***banking***	*crises as happened in 2009...*
...saying that Europe needs unified	***banking***	*regulation to replace the hodgepodge...*
...India has just given its	***banking***	*system a shot in the arm...*

Contextualized Word Embeddings

- Chico Ruiz made a spectacular play on Alusik's grounder …
- Olivia De Havilland signed to do a Broadway play for Garson …
- Kieffer was commended for his ability to hit in the clutch , as well as his all-round excellent play …
- … they were actors who had been handed fat roles in a successful play …
- Concepts play an important role in all aspects of cognition …

Contextualized Word Embeddings

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

ELMo: Embeddings from Language Models

- Deep contextualized word representations, NAACL 2018
  - 15K+ citations
- Key ideas
  - Learning contextualized embeddings with LSTM-based language models on a large corpus
  - Use the hidden states of the LSTMs for each token to compute a vector representation of each word

Deep contextualized word representations

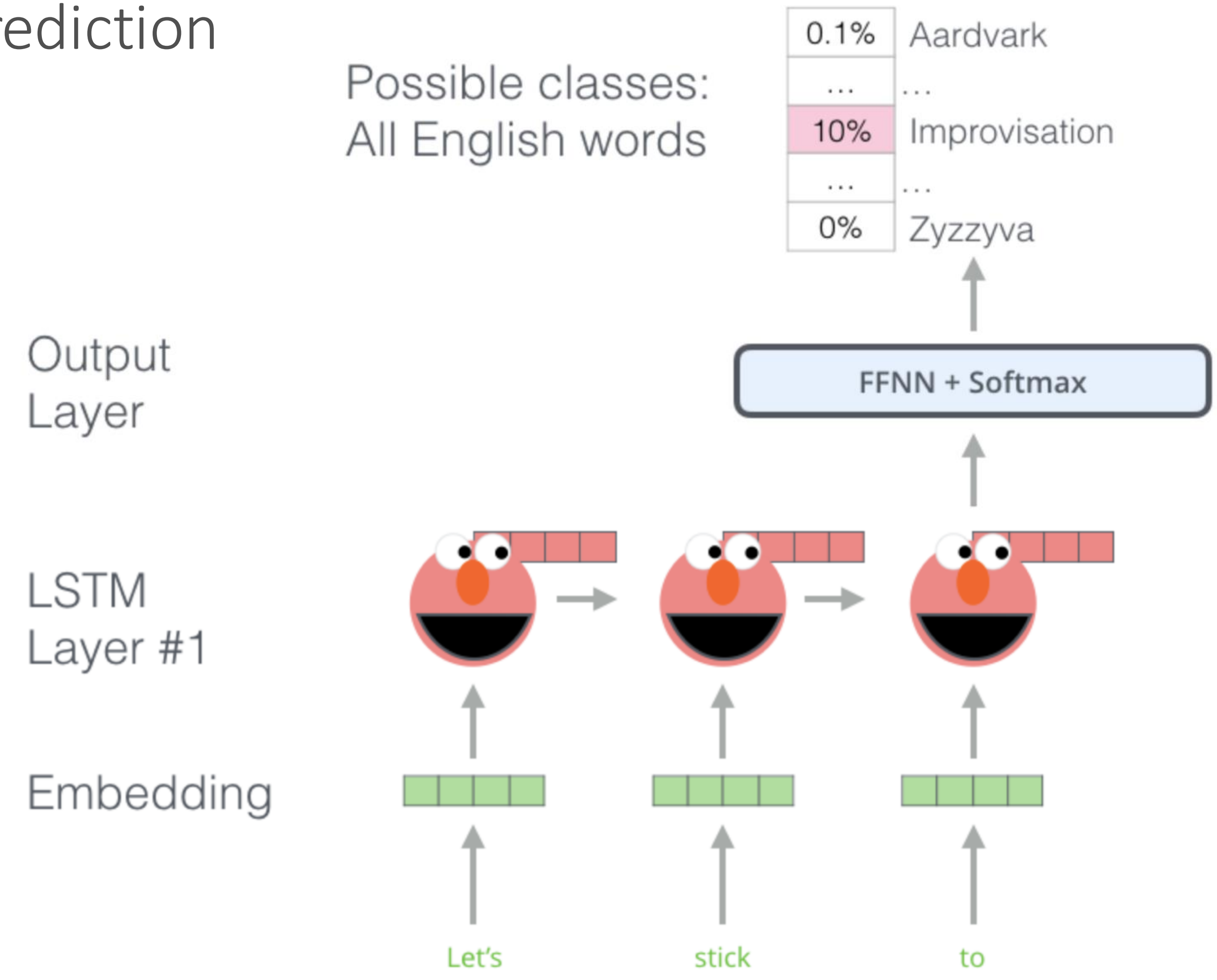
Matthew E. Peters[†]**, Mark Neumann**[†]**, Mohit Iyyer**[†]**, Matt Gardner**[†]**,**
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*]**, Kenton Lee**[*]**, Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

Language Modeling

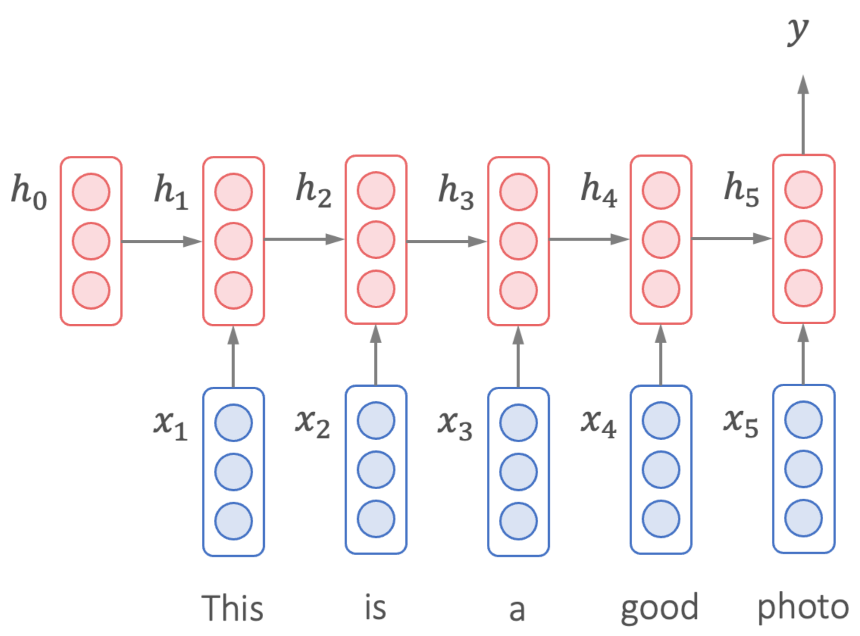
- Next word prediction

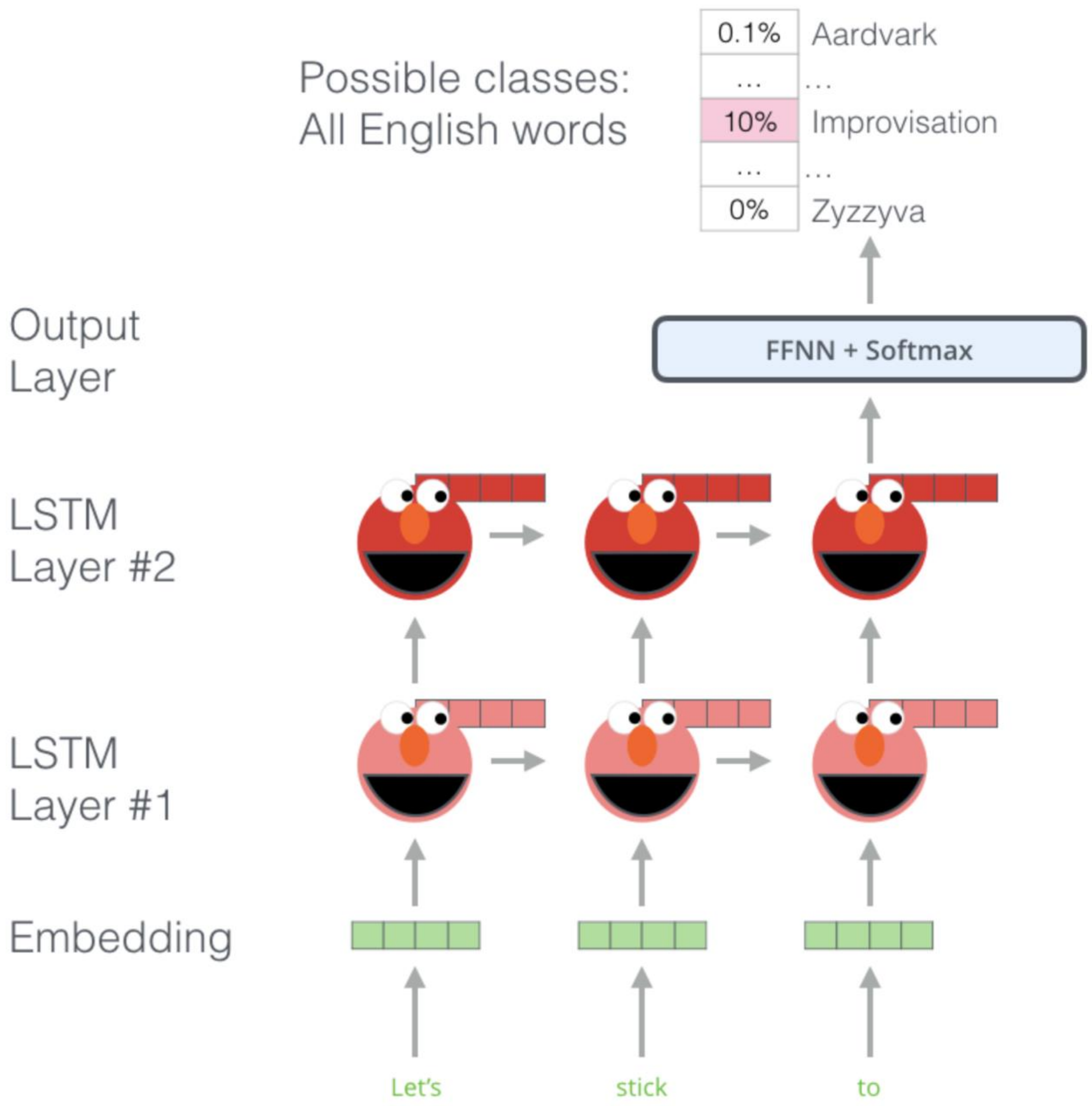


Source: http://jalammar.github.io/illustrated-bert/

Language Modeling

- Stacked LSTM

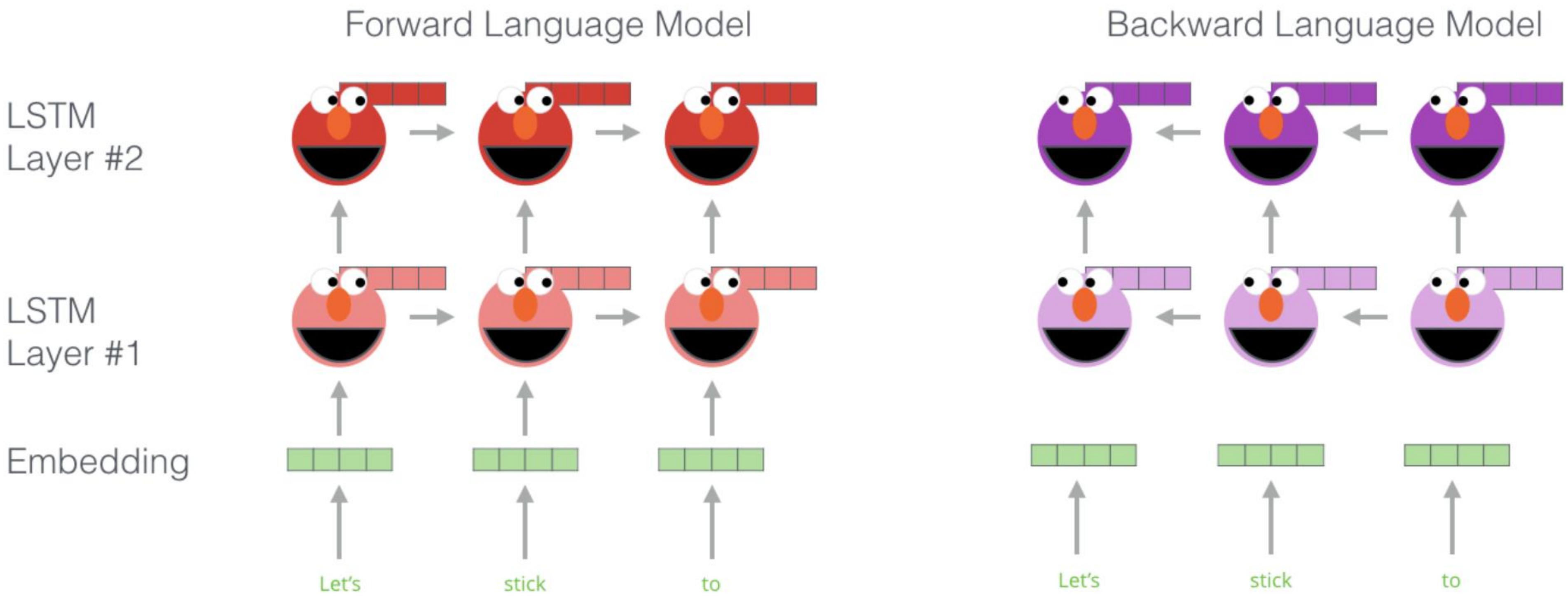




Source: http://jalammar.github.io/illustrated-bert/

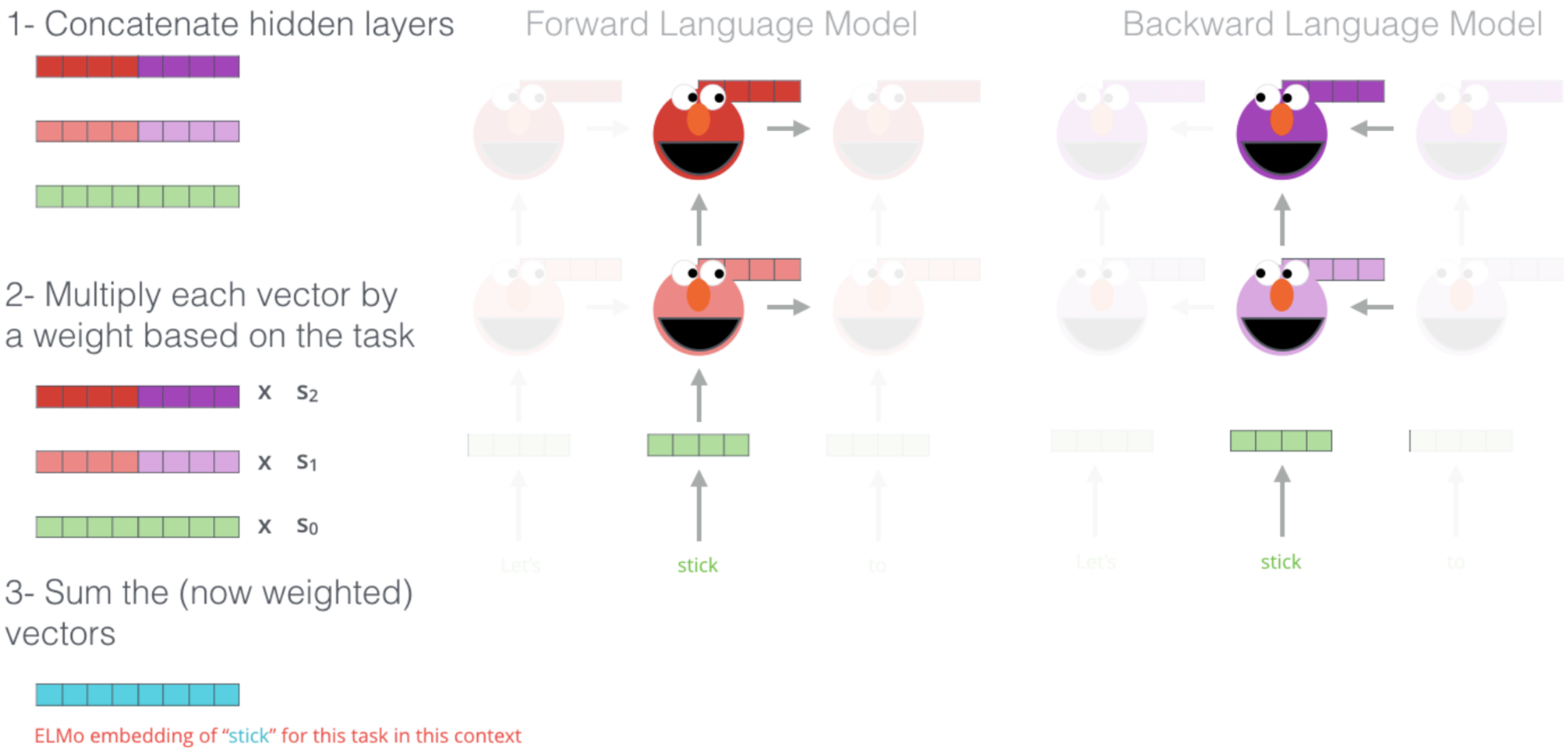
Language Modeling

- Bi-directional language modeling



Source: http://jalammar.github.io/illustrated-bert/

Contextualized Word Embeddings

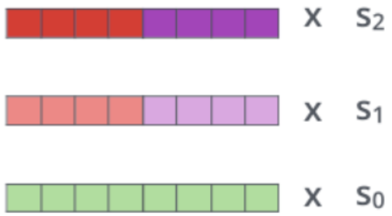


Source: http://jalammar.github.io/illustrated-bert/

Task-Specific Weights

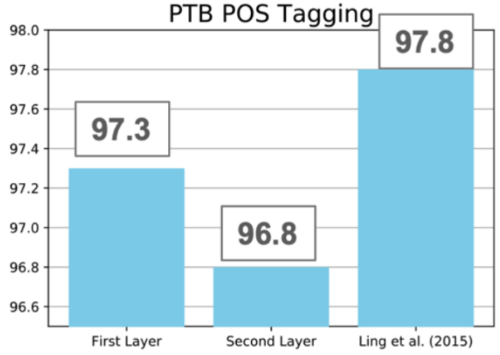
1- Concatenate hidden layers

2- Multiply each vector by a weight based on the task

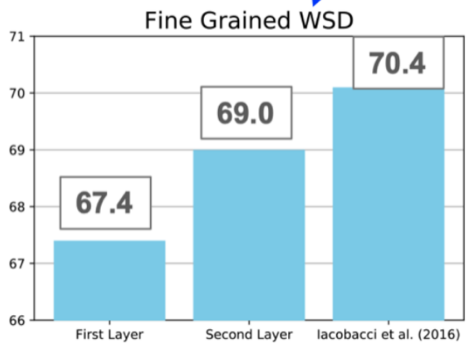


3- Sum the (now weighted) vectors

WSD = word sense disambiguation



first layer > second layer



second layer > first layer

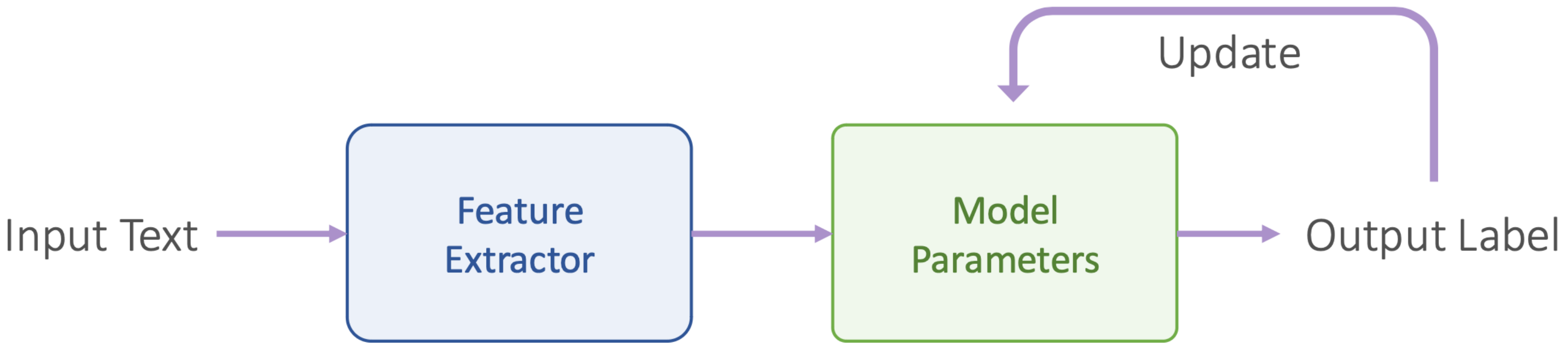
Lecture Plan

- Natural Language Processing Basics
- Transformers
- Contextualized Representations
- Pre-Training

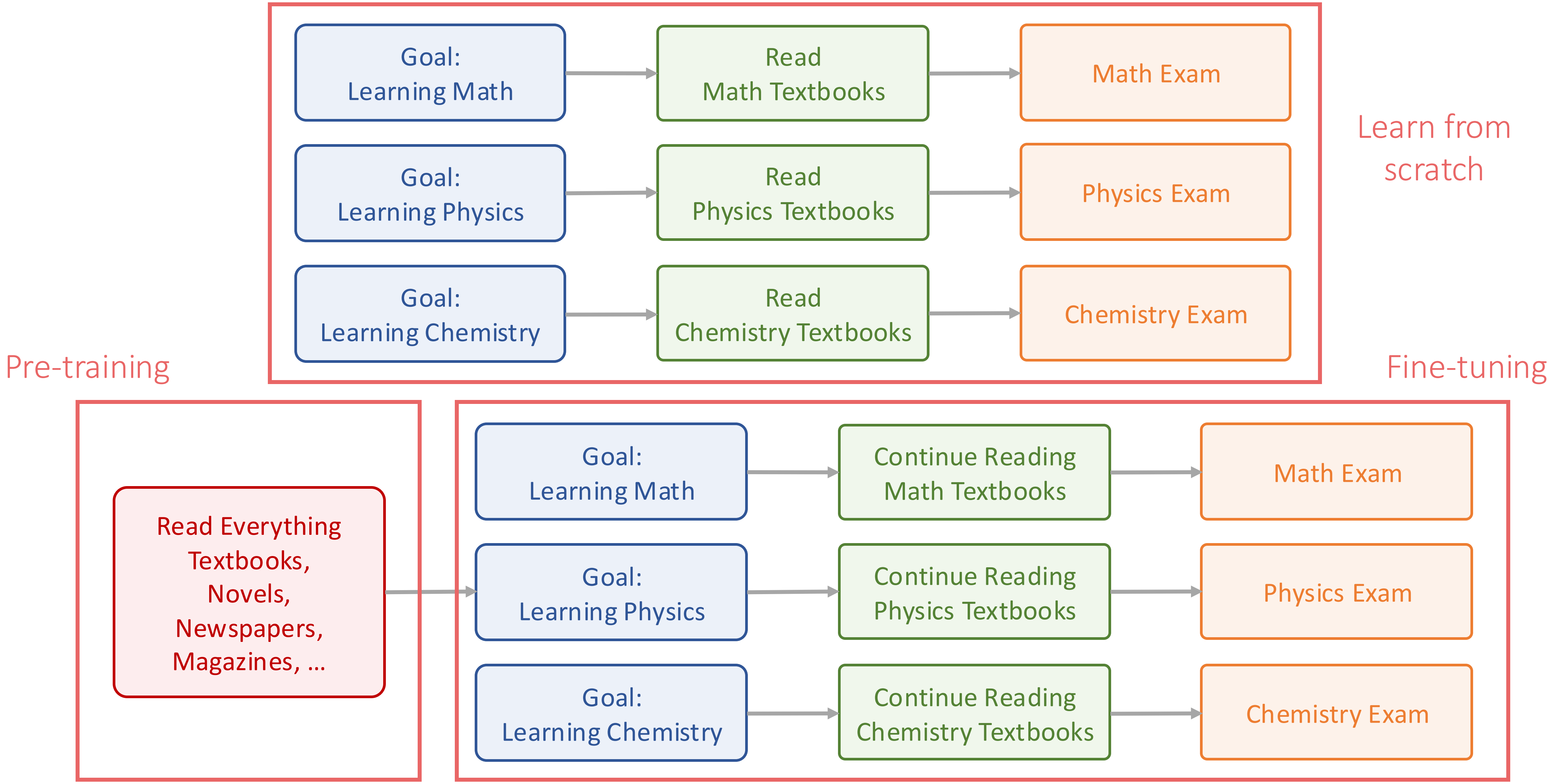
Feature-Based vs. Fine-Tuning Approaches

- Task-specific features + task-specific model
- General embeddings + task-specific model
- General embeddings + general model + task-specific fine-tuning

Pre-Training



Pre-Training



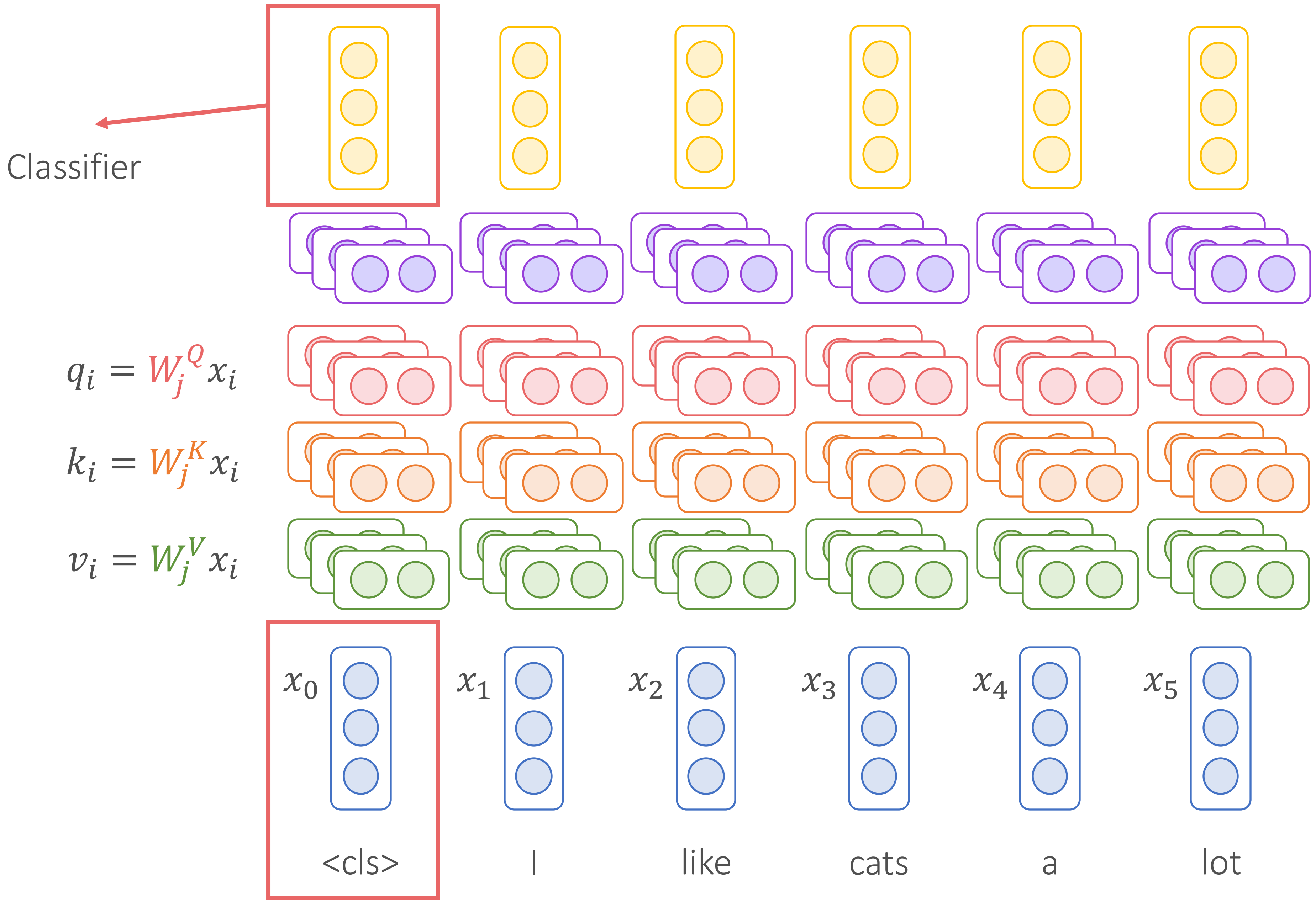
Bidirectional Encoder Representations from Transformers

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019
  - 110K+ citations
- Learn general knowledge with a large corpus
- Re-use model weights for fine-tuning

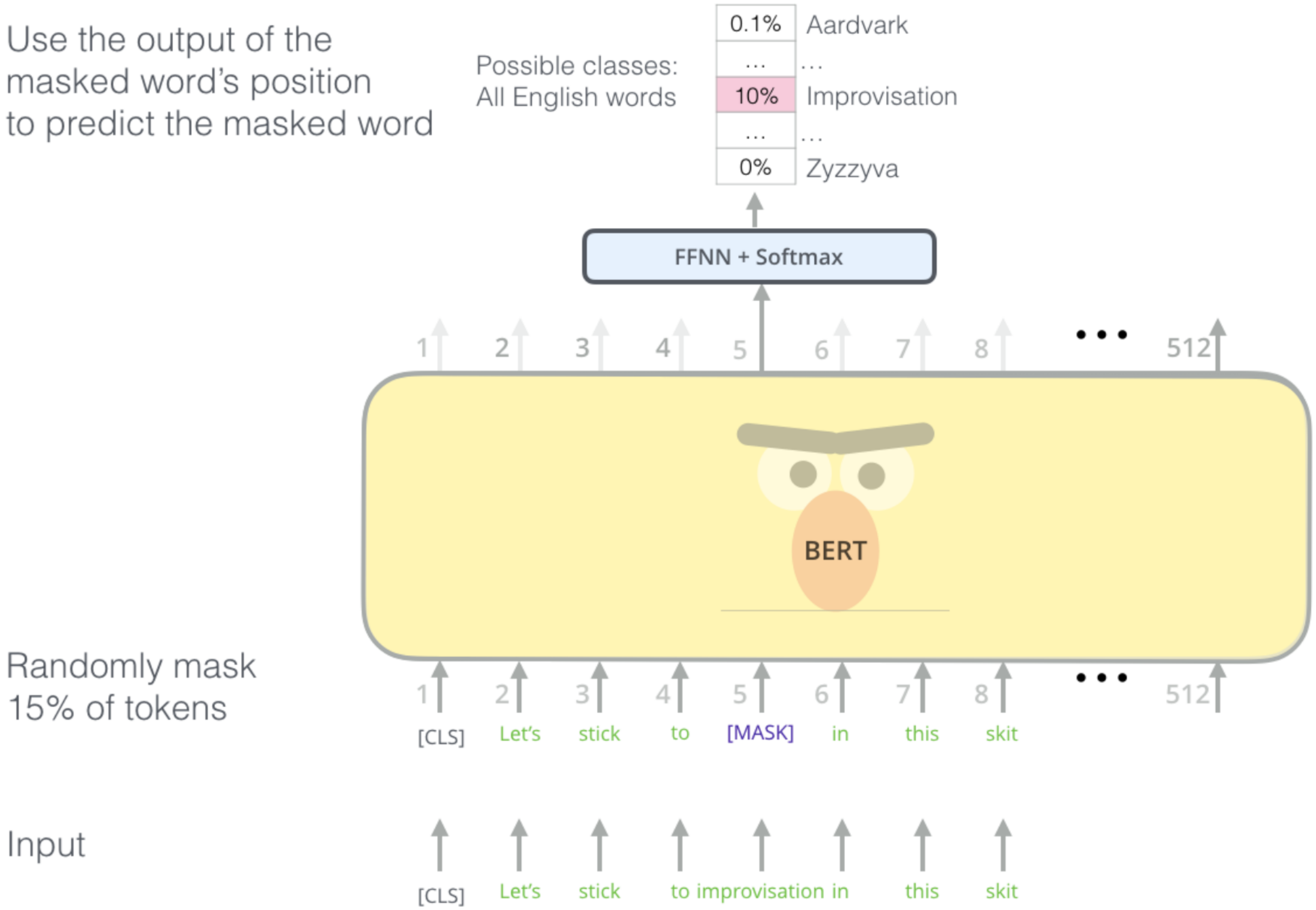
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Transformer Encoder Only

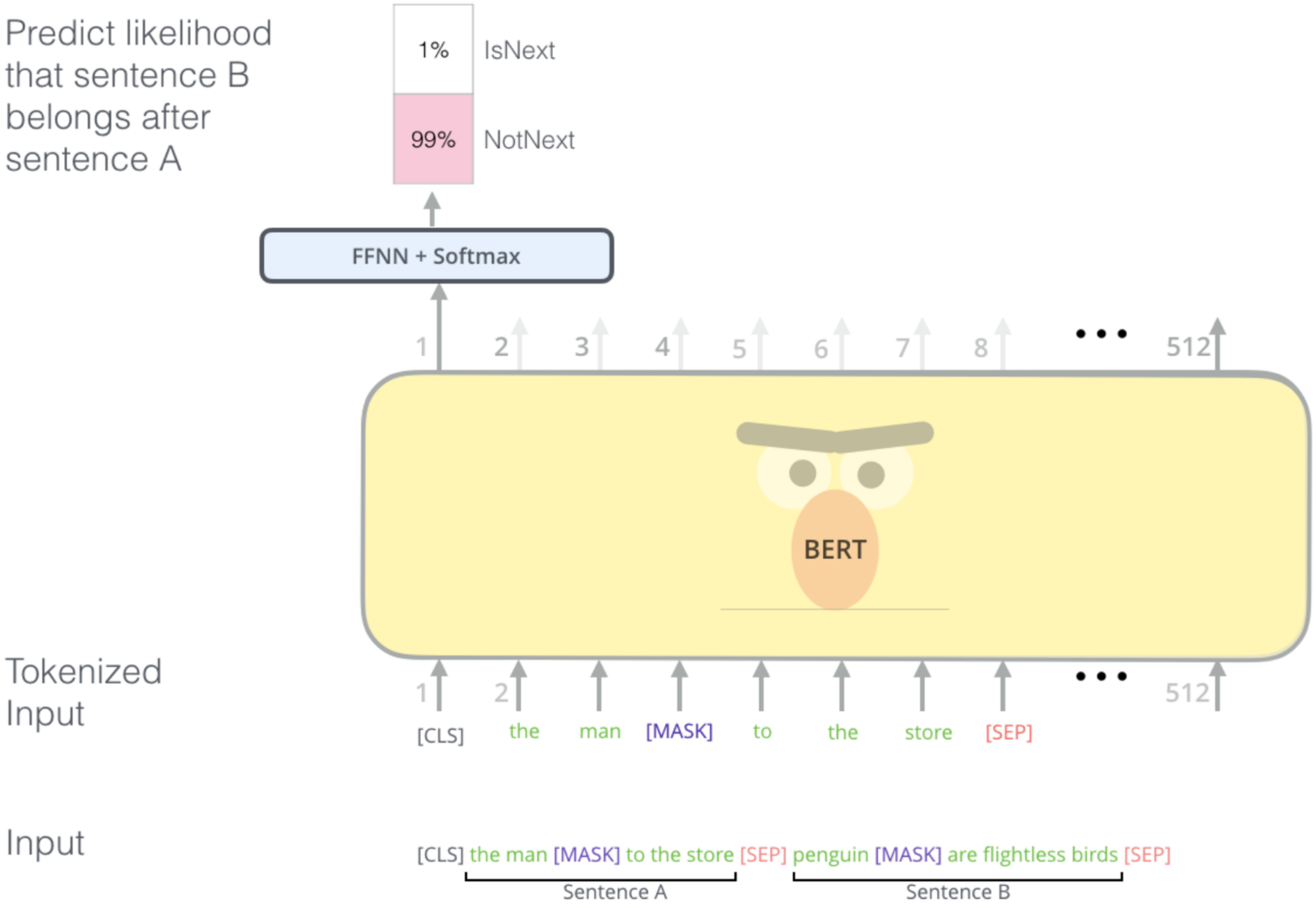


Pre-Training Task: Masked Language Modeling



Source: http://jalammar.github.io/illustrated-bert/

Pre-Training Task: Next Sentence Classification



Source: http://jalammar.github.io/illustrated-bert/

Next Lecture

- Natural Language Processing Basics
- Pre-Training
- Generative Pre-Training
- Language Models