COMP6211I: Trustworthy Machine Learning

Lecture 3

Minhao Cheng

Exam

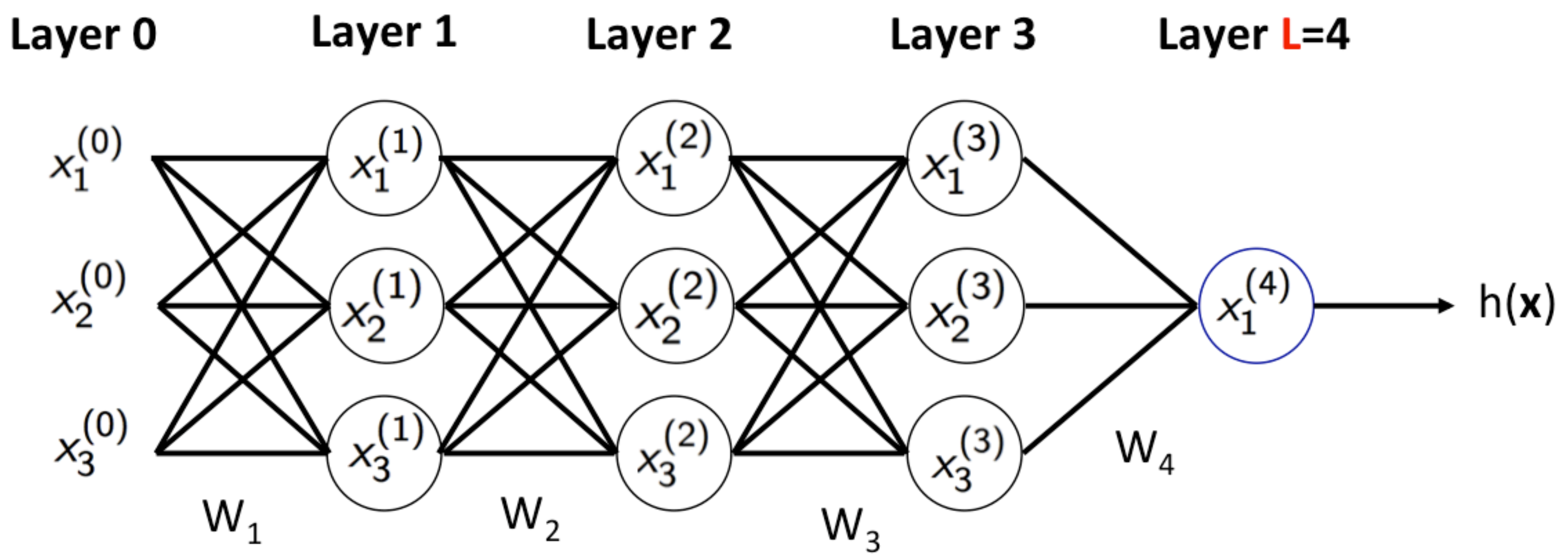
- On next Monday (Feb 20) during the class time
- 80 minutes
- Format:
  - True/False questions with reasons
  - Short answer questions
  - Problems (gradient derivation etc.)

From week 3

- Paper presentation sign-up started today
- Start from Feb 20:
  - Reading summary
  - Paper presentation
  - Class notes & participation
- Project proposal will be due on Feb 24 (1/2 page)
  - Title
  - Proposed problem
  - Proposed methodology (optional)

Convolutional Neural Network

Neural Networks



$$
\begin{aligned}
h(\boldsymbol{x}) &= x_1^{(4)} = \theta(W_4\boldsymbol{x}^{(3)}) = \theta(W_4\theta(W_3\boldsymbol{x}^{(2)})) \\
&= \cdots = \theta(W_4\theta(W_3\theta(W_2\theta(W_1\boldsymbol{x}))))
\end{aligned}
$$

- Fully connected networks ⇒ doesn't work well for computer vision applications

Convolutional Neural Network
Convolution Layer

- Fully connected layers have too many parameters
  - $\Rightarrow$ poor performance
- Example: VGG first layer
  - Input: $224 \times 224 \times 3$
  - Output: $224 \times 224 \times 64$
  - Number of parameters if we use fully connected net:
    - $(224 \times 224 \times 3) \times (224 \times 224 \times 64) =$ 483 billion
  - Convolution layer leads to:
    - Local connectivity
    - Parameter sharing

Convolutional Neural Network

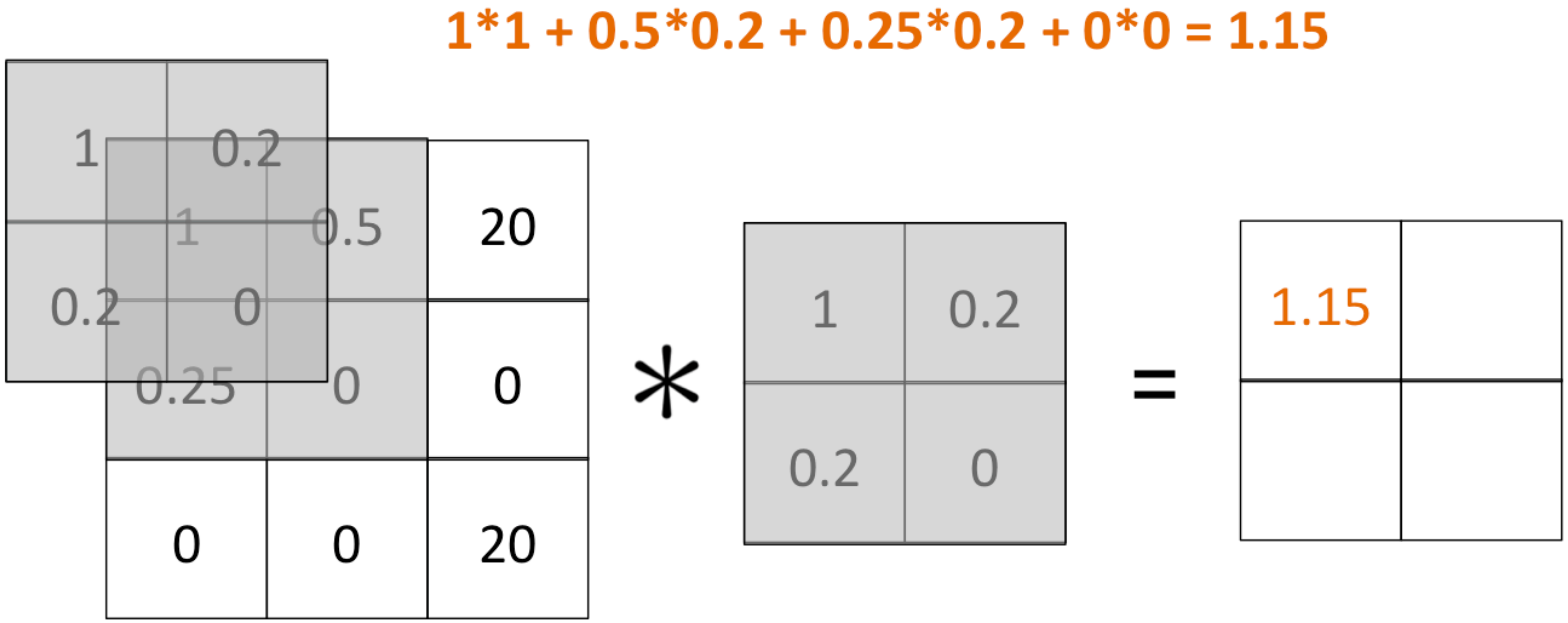
Convolution

- The convolution of an image x with a kernel k is computed as
  - $(x * k)_{ij} = \sum_{pq} x_{i+p,j+q} k_{p,q}$

$$\begin{bmatrix} 1 & 0.5 & 20 \\ 0.25 & 0 & 0 \\ 0 & 0 & 20 \end{bmatrix} * \begin{bmatrix} 1 & 0.5 \\ 0.25 & 0 \end{bmatrix} = \begin{bmatrix} & \\ & \end{bmatrix}$$

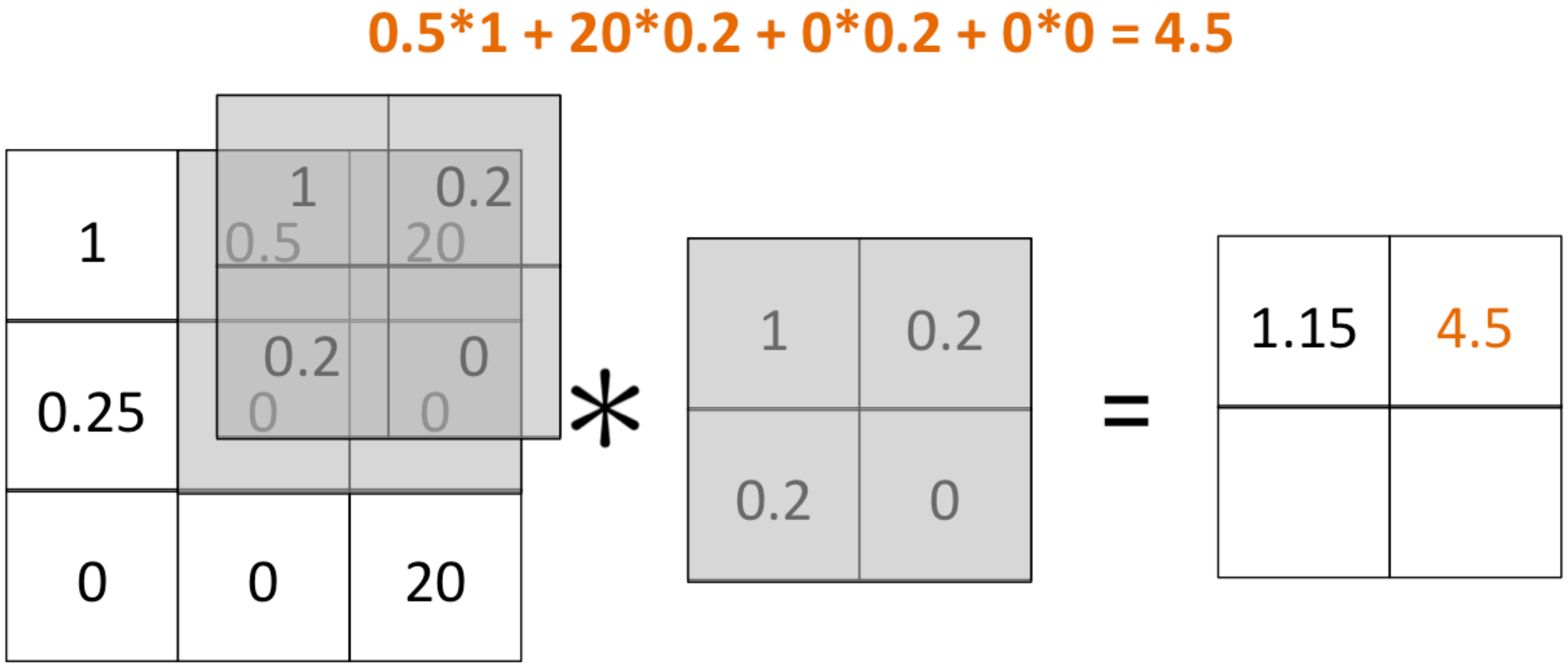
Convolutional Neural Network

Convolution



Convolutional Neural Network

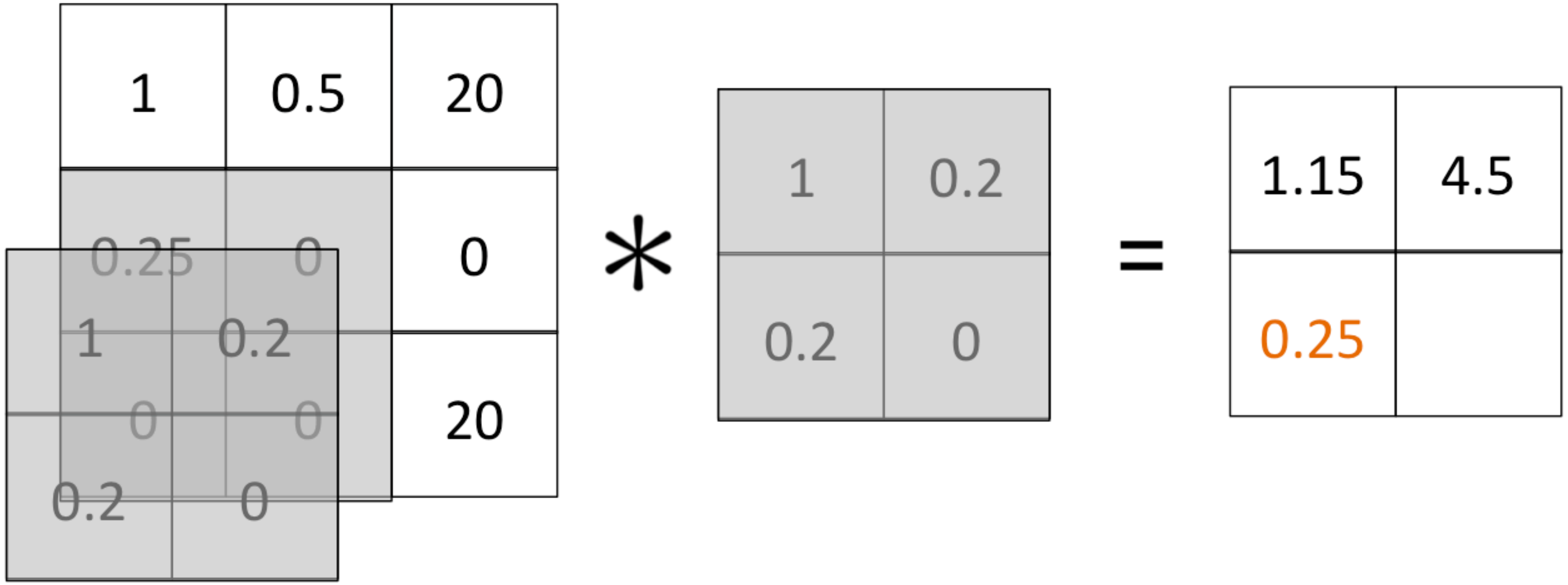
Convolution



Convolutional Neural Network

Convolution

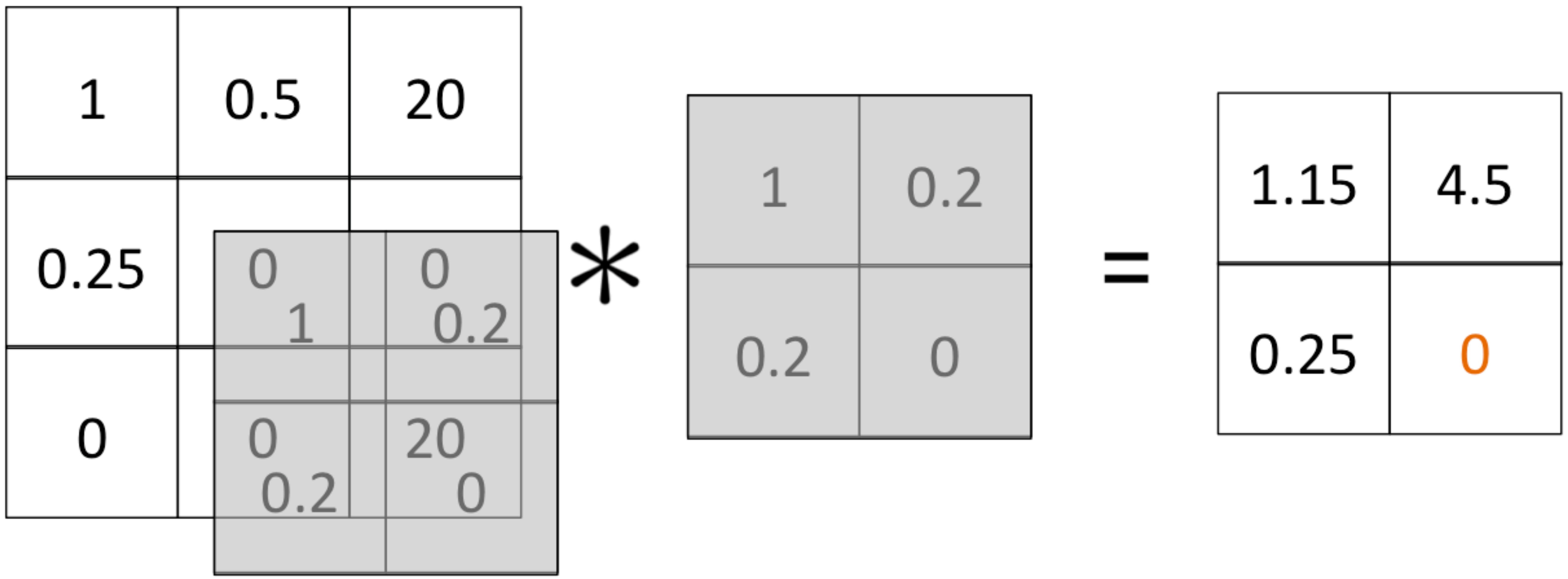
0.25*1 + 0*0.2 + 0*0.2 + 0*0 = 0.25



Convolutional Neural Network

Convolution

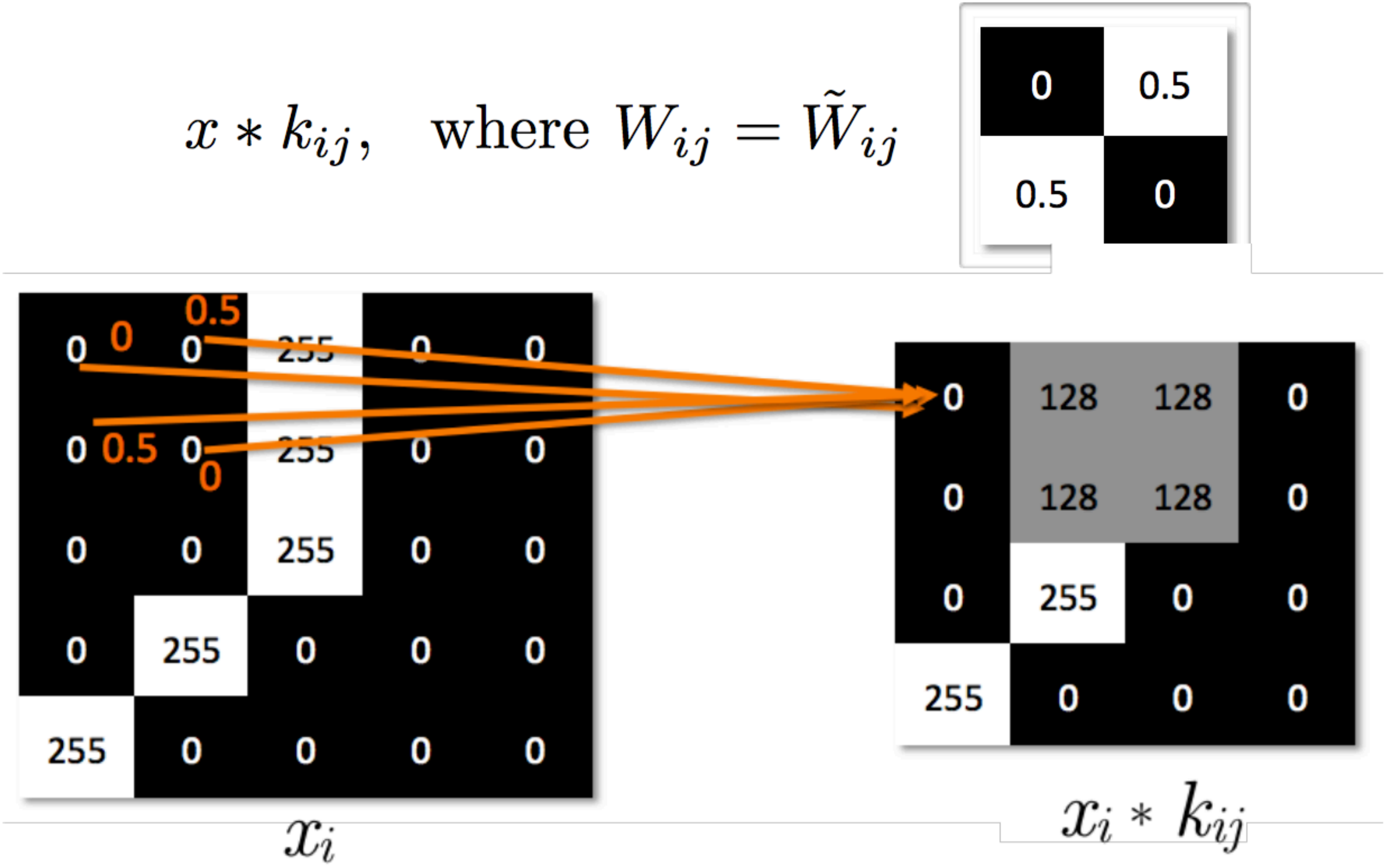
0*1 + 0*0.2 + 0*0.2 + 20*0 = 0



Convolutional Neural Network

Convolution

$$x * k_{ij}, \quad \text{where } W_{ij} = \tilde{W}_{ij}$$



Convolutional Neural Network

Convolution

- Element-wise activation function after convolution
  - $\Rightarrow$ detector of a feature at any position in the image

$$x * k_{ij}, \quad \text{where } W_{ij} = \tilde{W}_{ij}$$

0	0.5
0.5	0

0	0	255	0	0
0	0	255	0	0
0	0	255	0	0
0	255	0	0	0
255	0	0	0	0

x_i

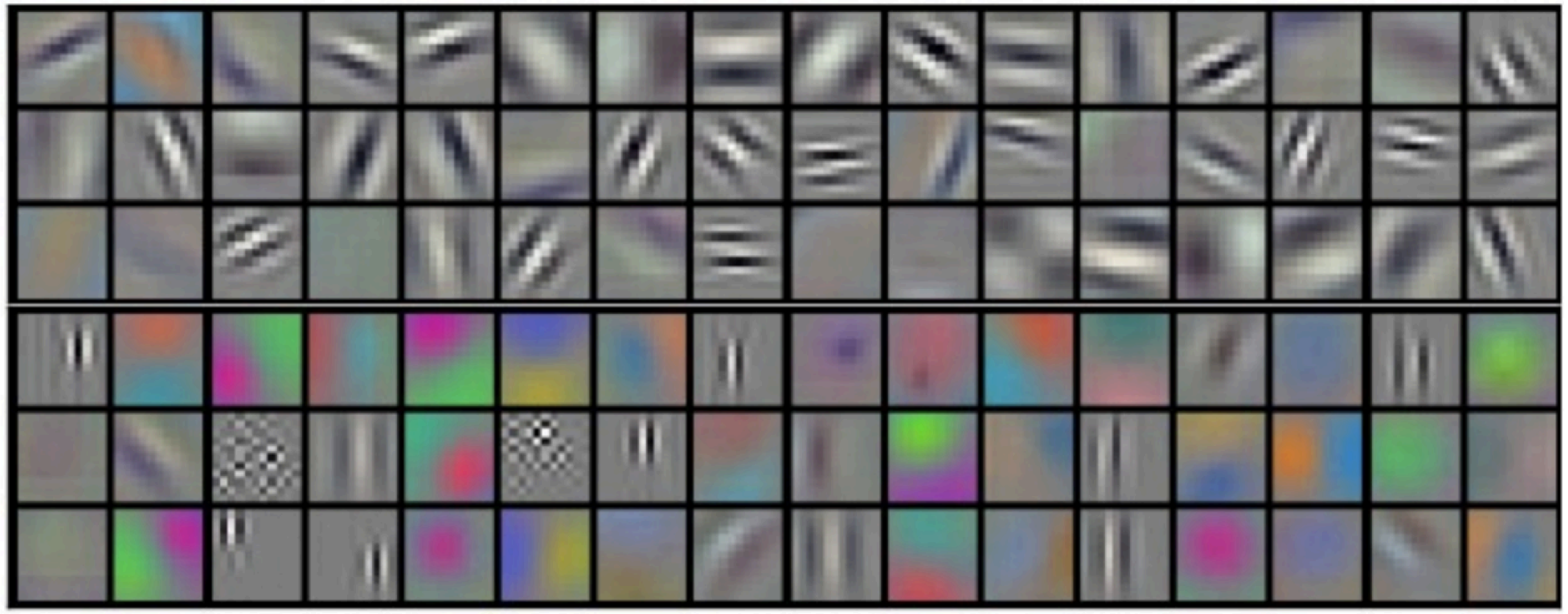
0.02	0.19	0.19	0.02
0.02	0.19	0.19	0.02
0.02	0.75	0.02	0.02
0.75	0.02	0.02	0.02

$\text{sigm}(0.02\ x_i * k_{ij} - 4)$

Convolutional Neural Network

Learned Kernels

- Example kernels learned by AlexNet

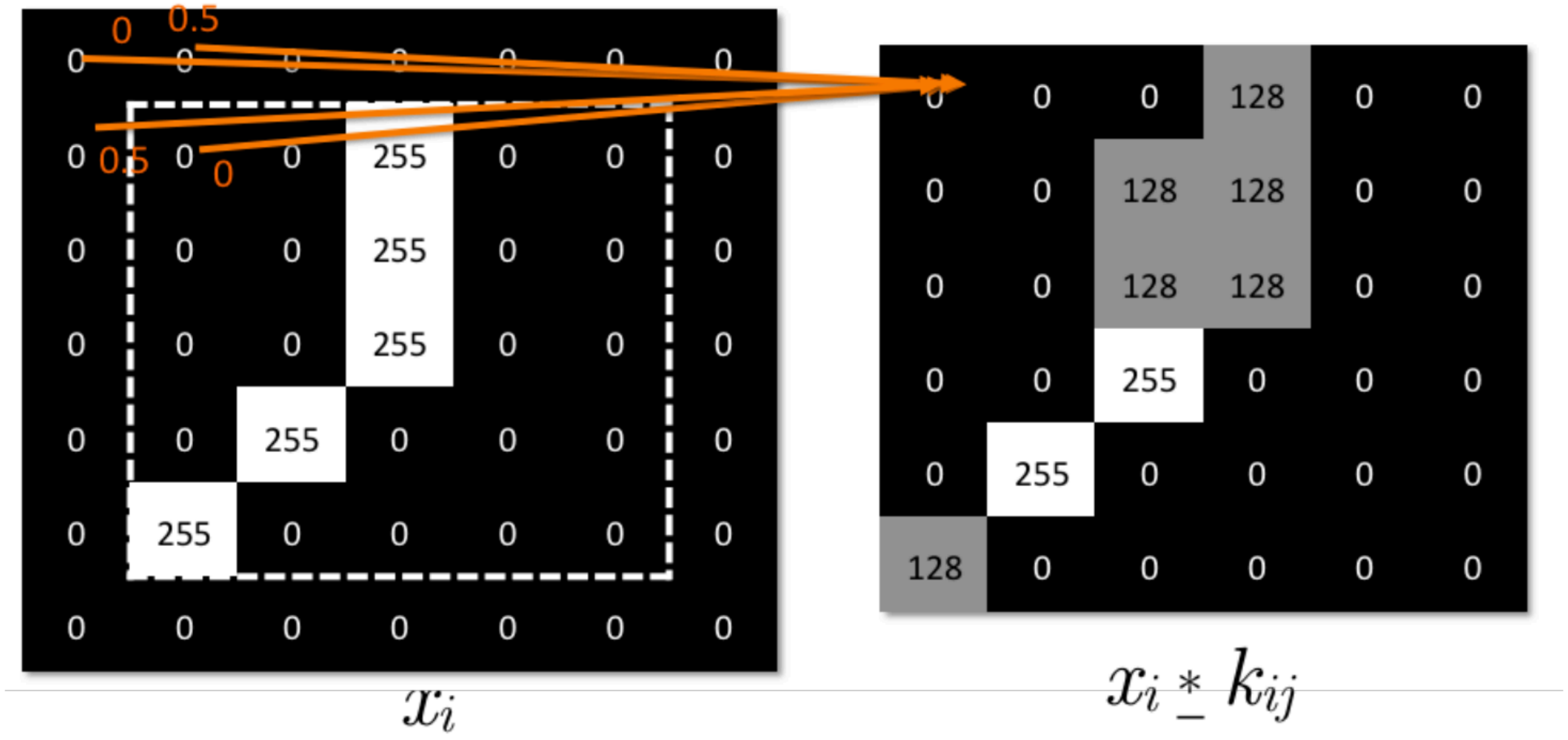


- Number of parameters:
    - Example: 200×200 image, 100 kernels, kernel size 10×10
    - $\Rightarrow 10 \times 10 \times 100$ = 10K parameters

Convolutional Neural Network

Padding

- Use zero padding to allow going over the boundary
  - Easier to control the size of output layer

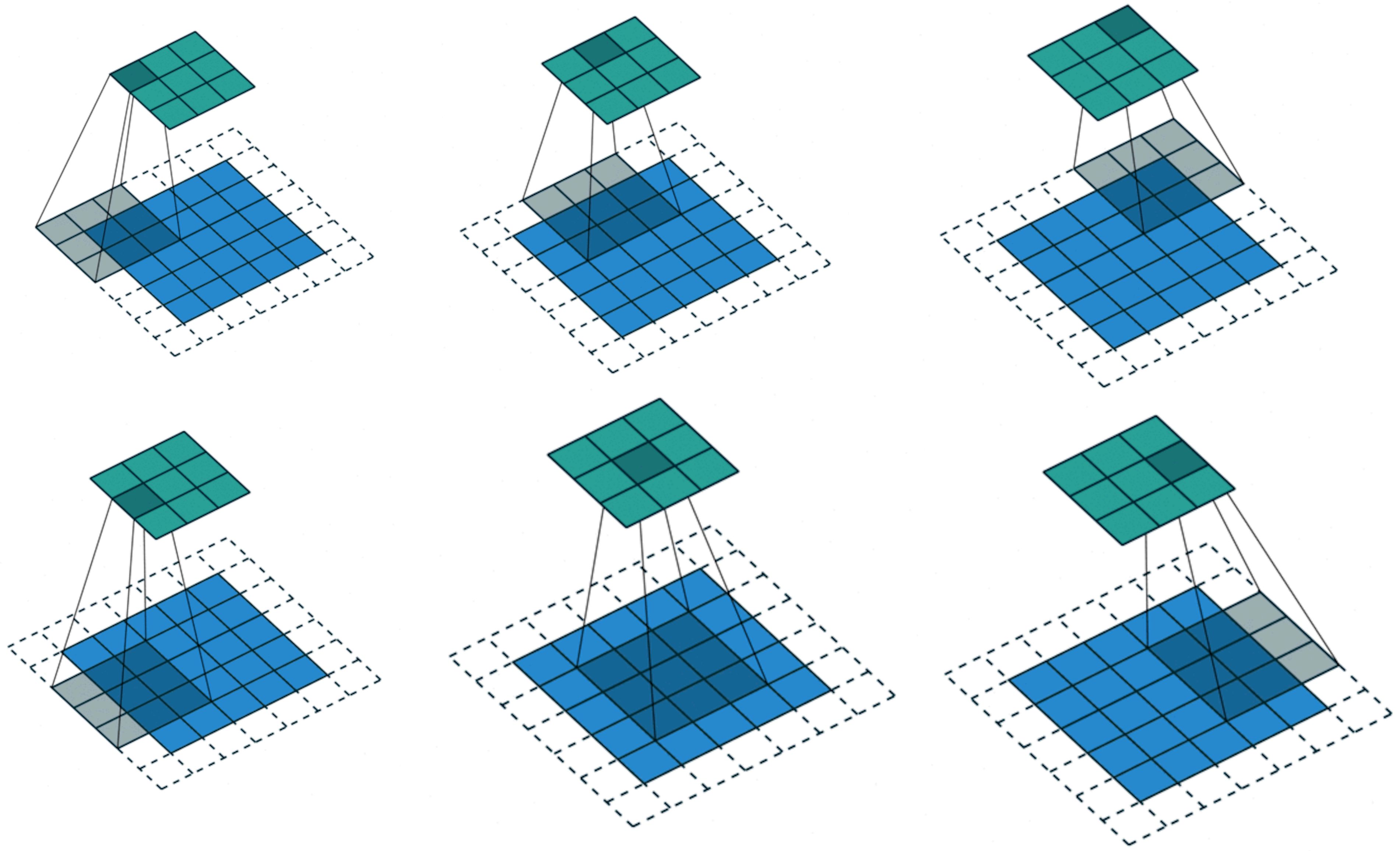


Convolutional Neural Network

Strides

- Stride: The amount of movement between applications of the filter to the input image

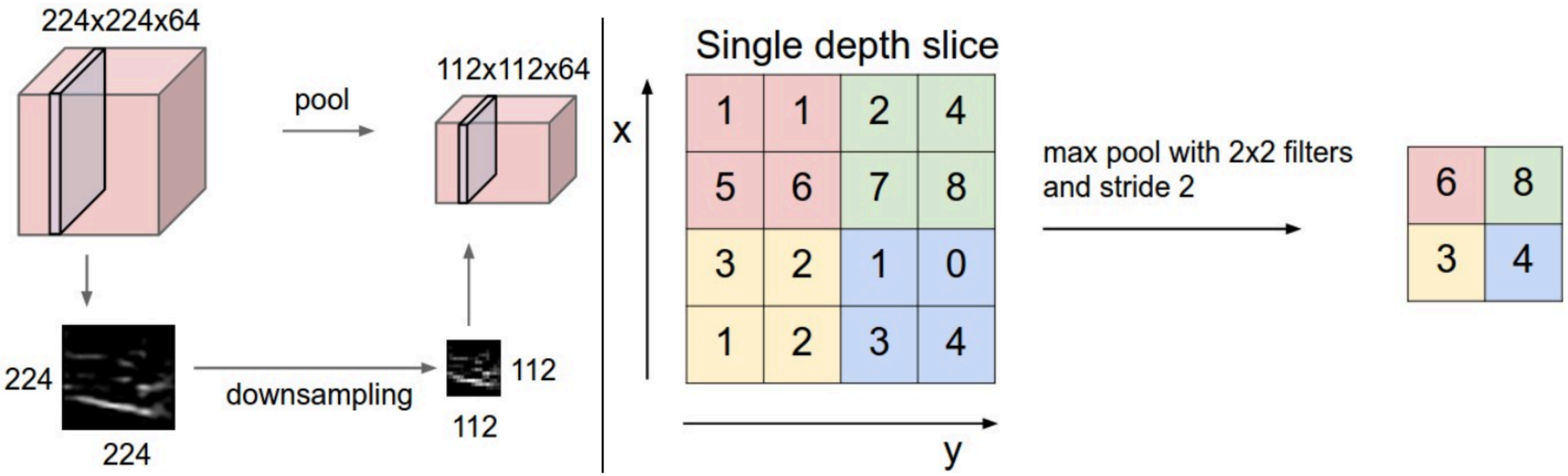
- Stride $(1,1)$: no stride



Convolutional Neural Network

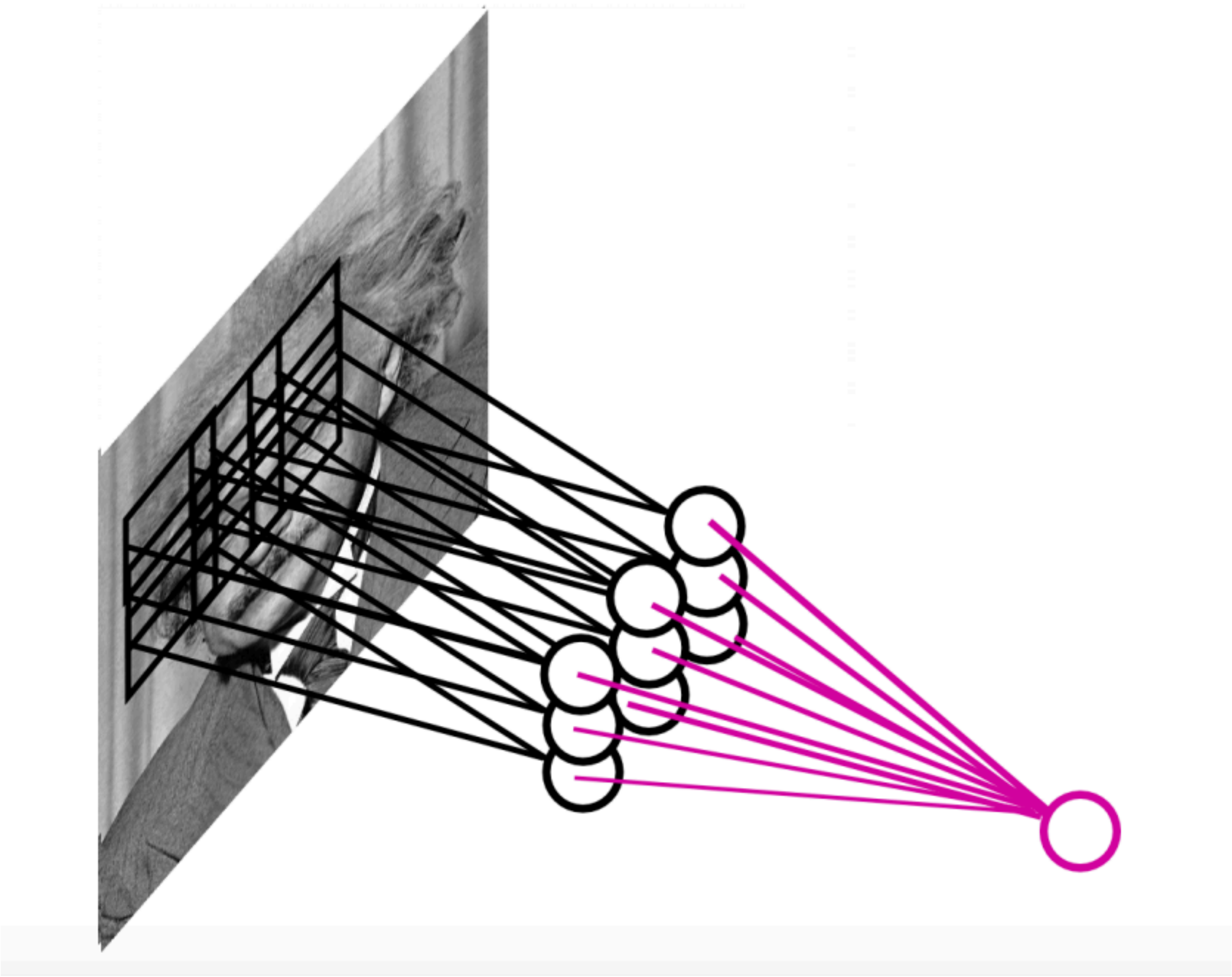
Pooling

- It's common to insert a pooling layer in-between successive convolutional layers
- Reduce the size of presentation, down-sampling
- Example: Max pooling



Convolutional Neural Network
Pooling

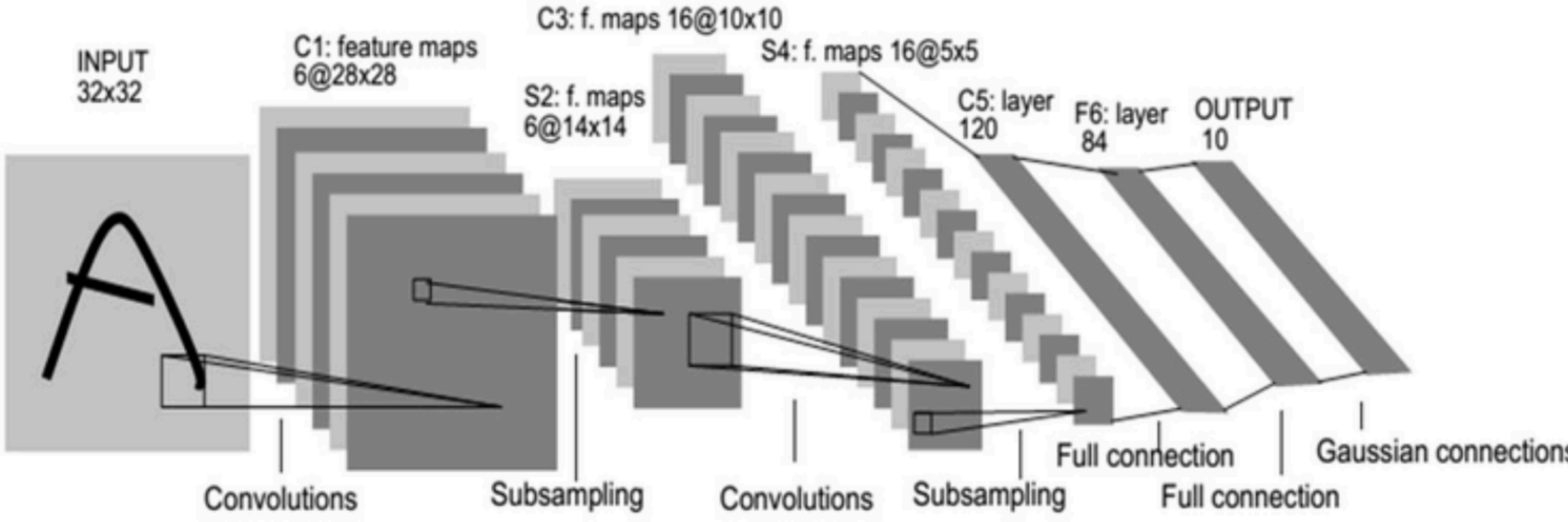
- By pooling, we gain robustness to the exact spatial location of features



Convolutional Neural Network

Example: LeNet5

- Input: 32×32 images (MNIST)
- Convolution 1: 6 5×5filters, stride 1
  - Output: 6 28×28 maps
- Pooling 1: 2×2 max pooling, stride 2
  - Output: 6 14×14 maps
- Convolution 2: 16 5×5 filters, stride 1
  - Output: 16 10×10 maps
- Pooling 2: 2×2 max pooling with stride 2
  - Output: 16 5×5 maps (total 400 values)
- 3 fully connected layers: $120 \Rightarrow 84 \Rightarrow 10$ neurons



Convolutional Neural Network

Training

- Training:
  - Apply SGD to minimize in-sample training error
  - Backpropagation can be extended to convolutional layer and pooling layer to compute gradient!
  - Millions of parameters $\Rightarrow$ easy to overfit

Convolutional Neural Network

Revisit Alexnet

- Dropout: 0.5 (in FC layers)
- A lot of data augmentation
- Momentum SGD with batch size 128, momentum factor 0.9
- L2 weight decay (L2 regularization)
- Learning rate: 0.01, decreased by 10 every time when reaching a stable validation accuracy

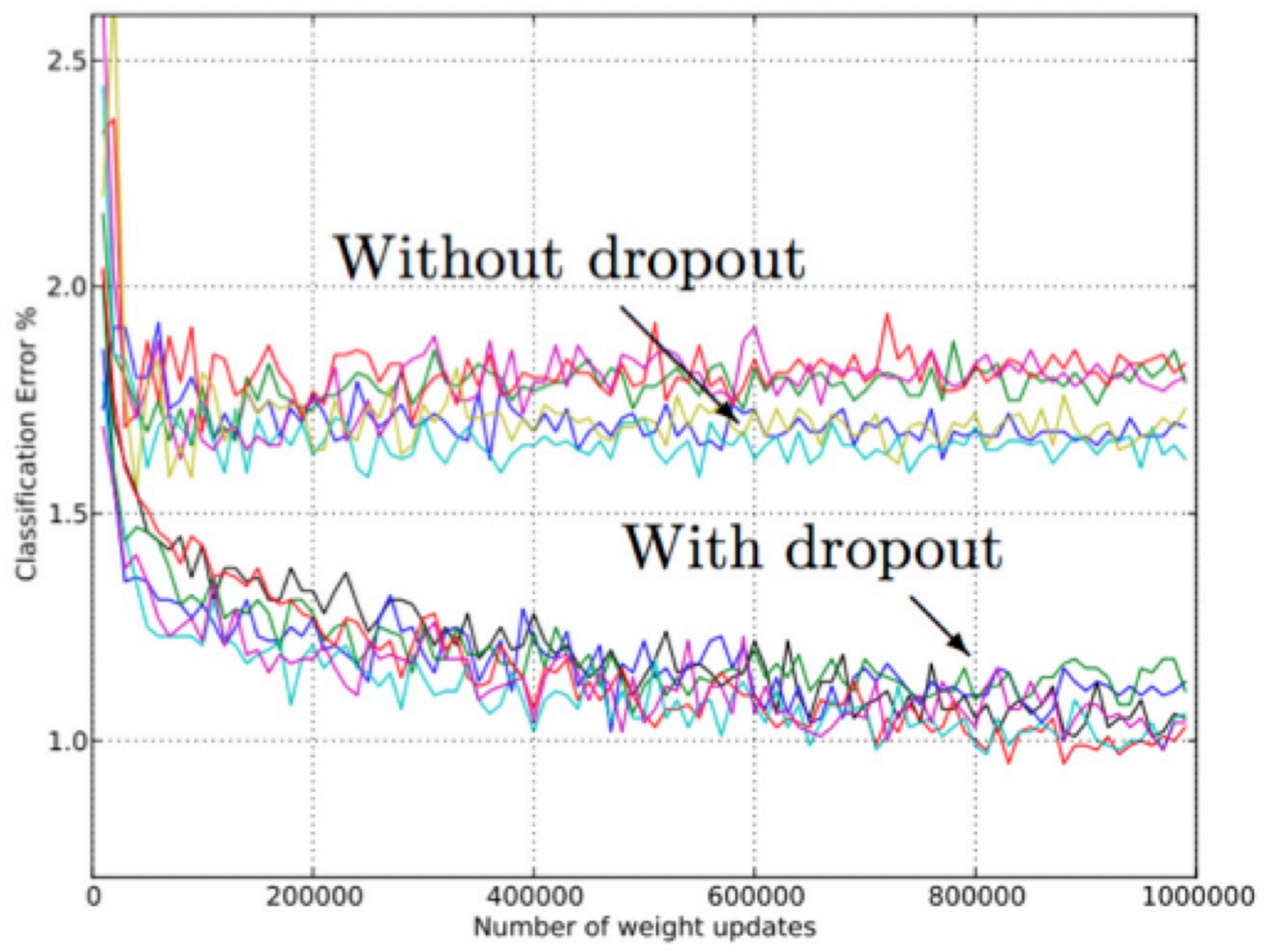
Convolutional Neural Network

Dropout

- One of the most effective regularization for deep neural networks

Method	CIFAR-10	CIFAR-100
Conv Net + max pooling (hand tuned)	15.60	43.48
Conv Net + stochastic pooling (Zeiler and Fergus, 2013)	15.13	42.51
Conv Net + max pooling (Snoek et al., 2012)	14.98	-
Conv Net + max pooling + dropout fully connected layers	14.32	41.26
Conv Net + max pooling + dropout in all layers	12.61	**37.20**
Conv Net + maxout (Goodfellow et al., 2013)	**11.68**	38.57

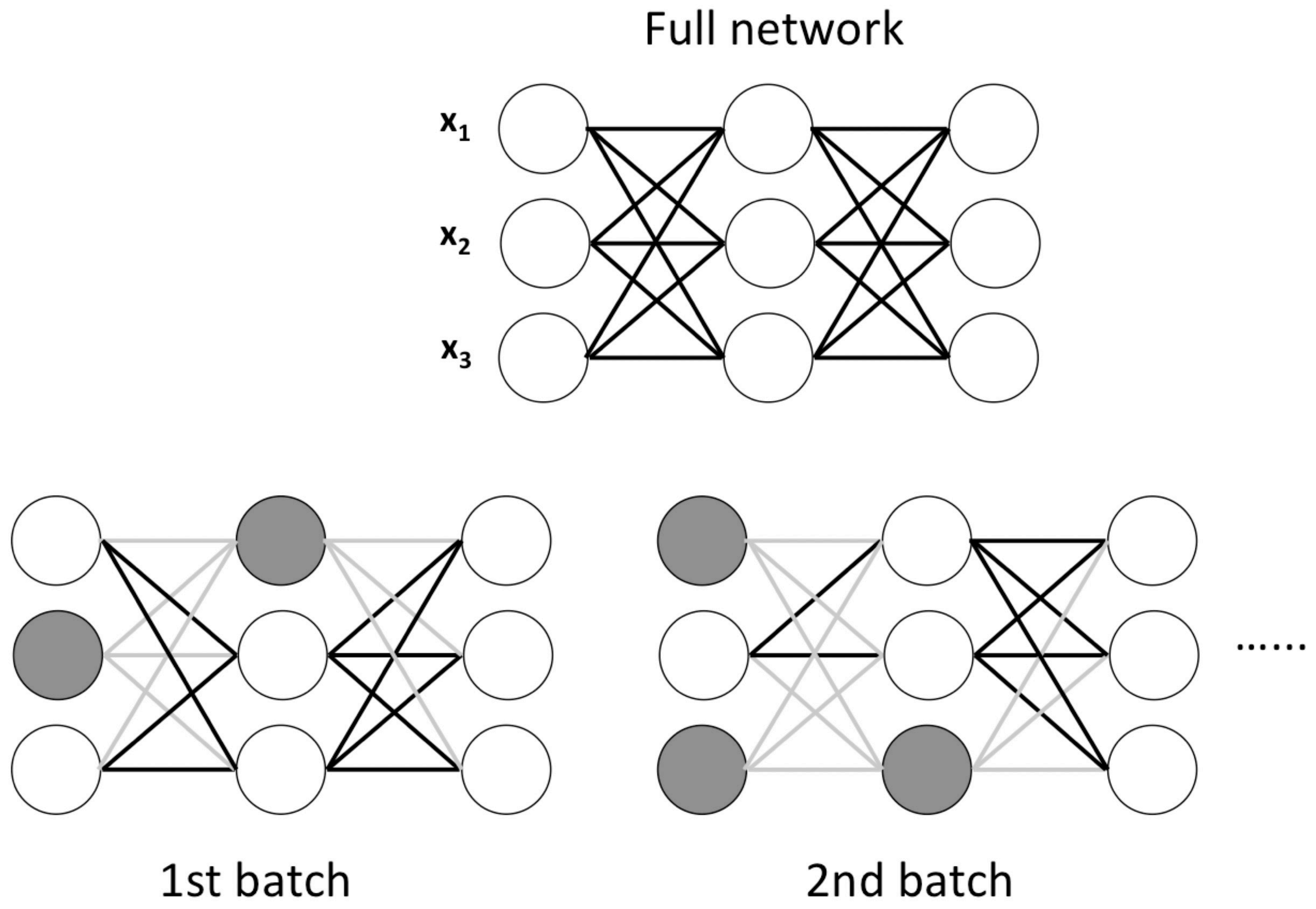
Table 4: Error rates on CIFAR-10 and CIFAR-100.



Srivastava et al, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, 2014.

Convolutional Neural Network
Dropout(training)

- Dropout in the **training** phase:
    - For each batch, turn off each neuron (including inputs) with a probability $1 - \alpha$
    - Zero out the removed nodes/edges and do backpropogation



Convolutional Neural Network

Dropout(test)

- The model is different from the full model:
- Each neuron computes
  - $x_i^{(l)} = B\sigma(\sum_j W_{ij}^{(l)} x_j^{(l-1)} + b_i^{(l)})$
  - Where B is Bernoulli variable that takes 1 with probability α
- The expected output of the neuron:
  - $E[x_i^{(l)}] = \alpha\sigma(\sum_j W_{ij}^{(l)} x_j^{(l-1)} + b_i^{(l)})$
- Use the expected output at test time $\Rightarrow$ multiply all the weights by α

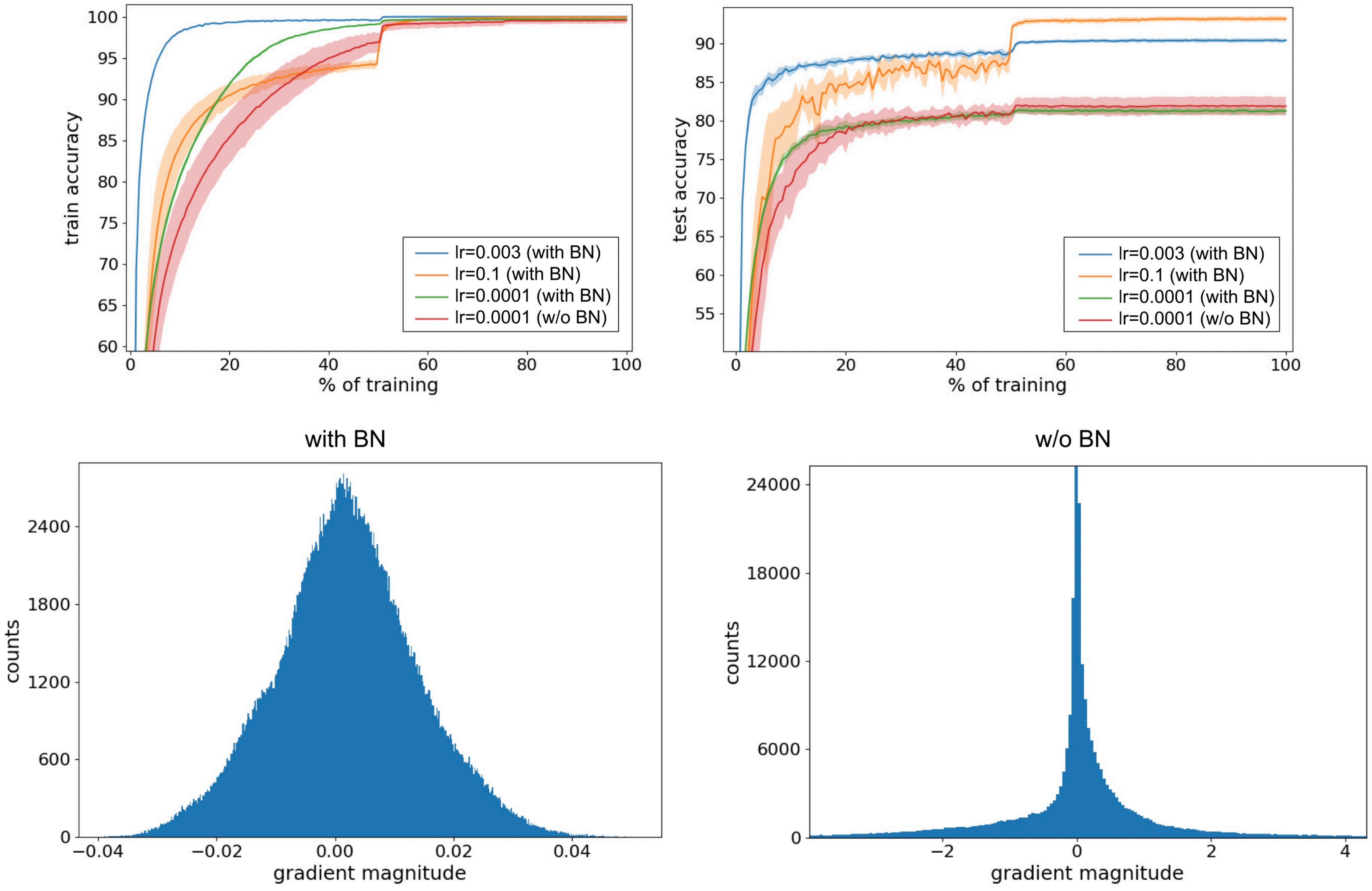
Convolutional Neural Network

Batch Normalization

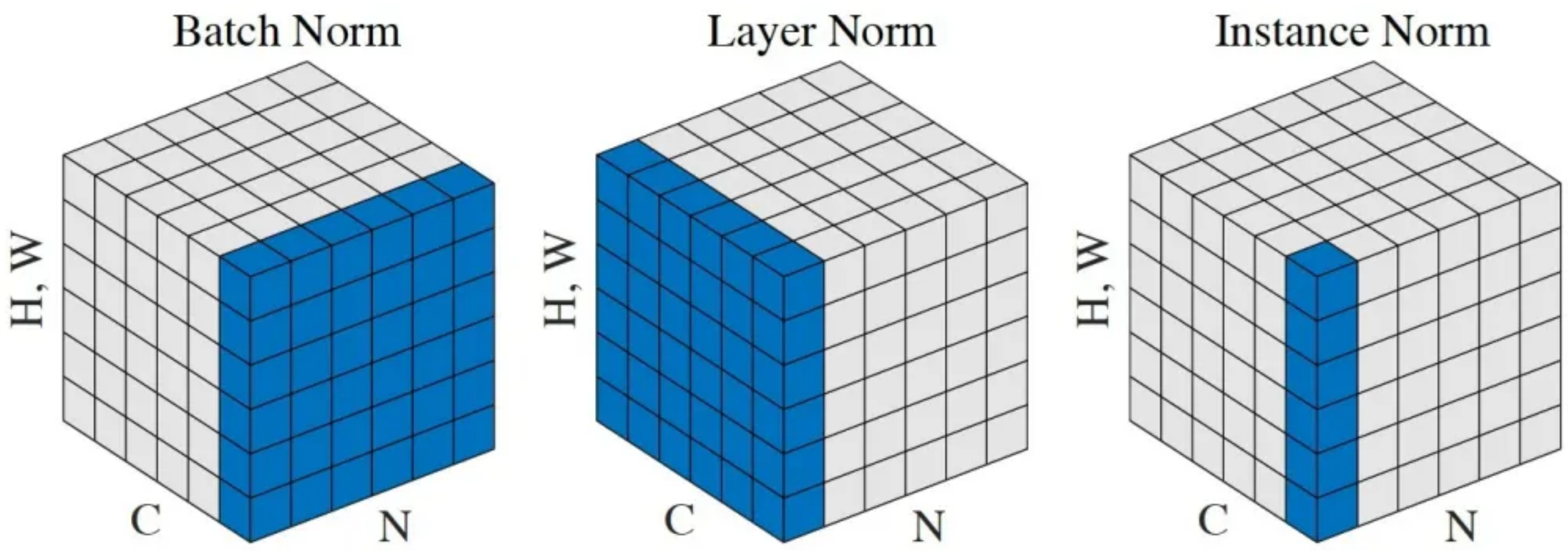
- Initially proposed to reduce co-variate shift
- $$O_{b,c,x,y} \leftarrow \gamma \frac{I_{b,c,x,y} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta \quad \forall b, c, x, y,$$
- $\mu_c = \frac{1}{|B|} \sum_{b,x,y} I_{b,c,x,y}$: the mean for channel c, and σ_c standard deviation.
- γ and β: two learnable parameters

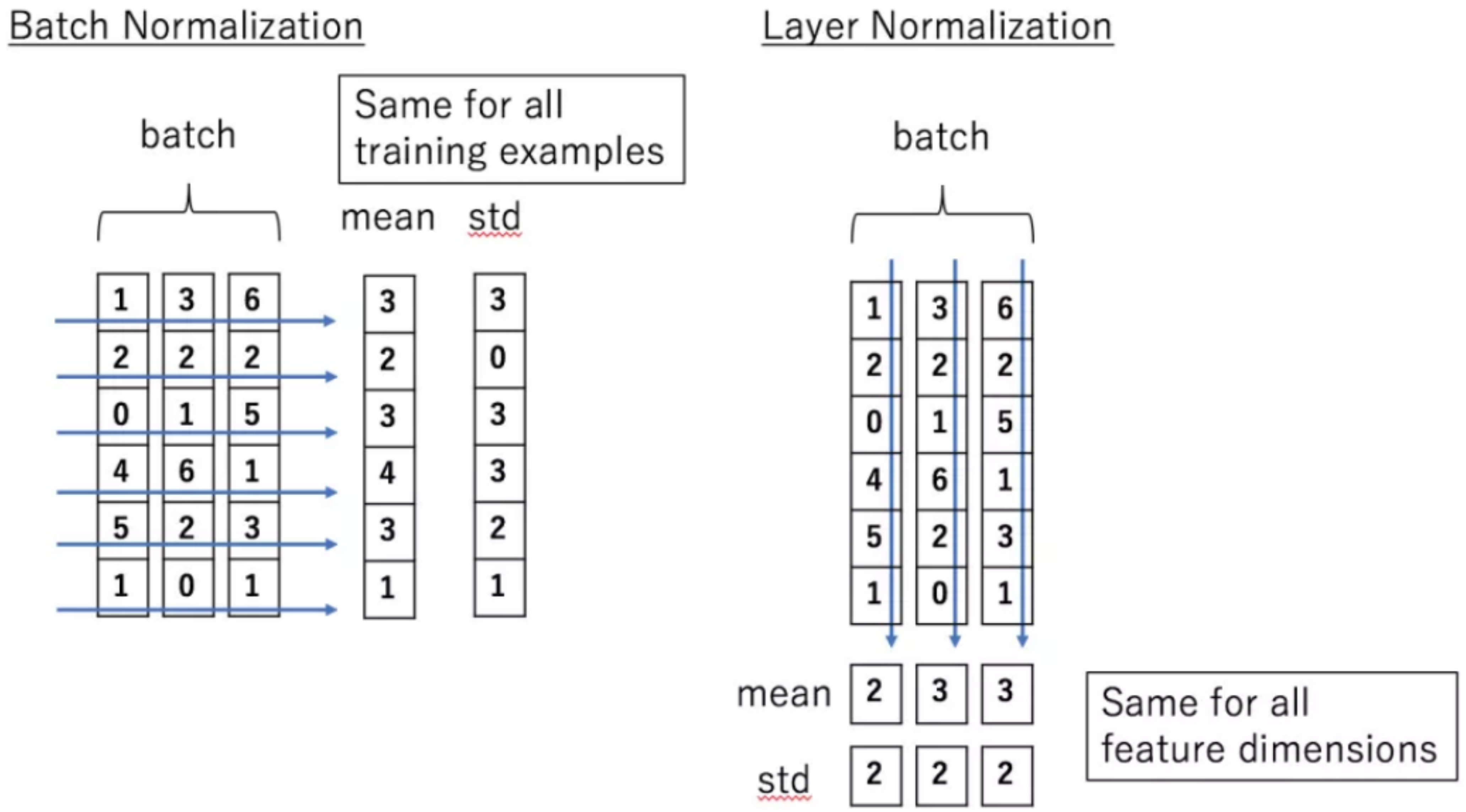
Convolutional Neural Network
Batch Normalization

- Couldn't reduce covariate shift (Ilyas et al 2018)
- Allow larger learning rate
    - Constraint the gradient norm



Convolutional Neural Network
Other normalization

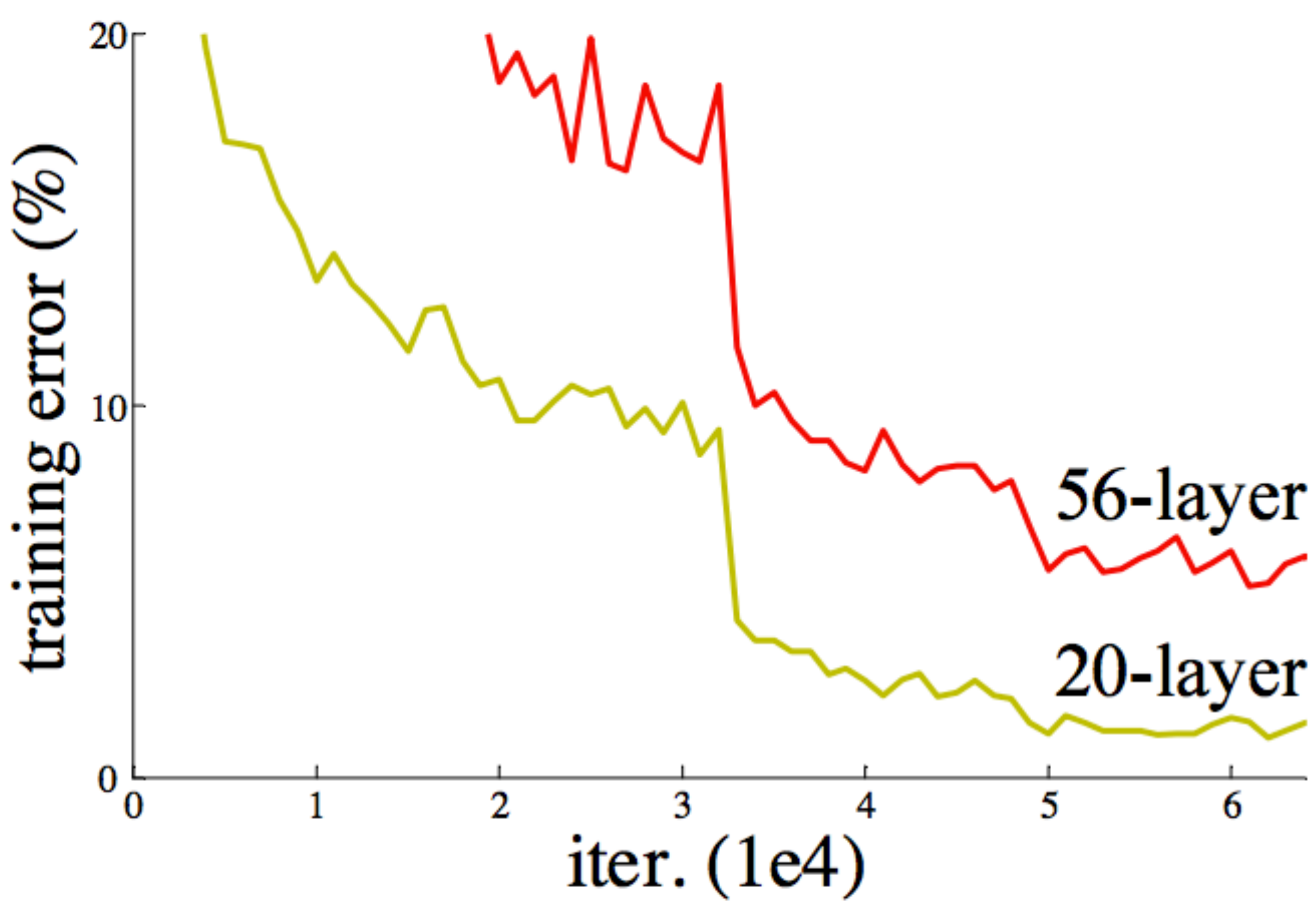


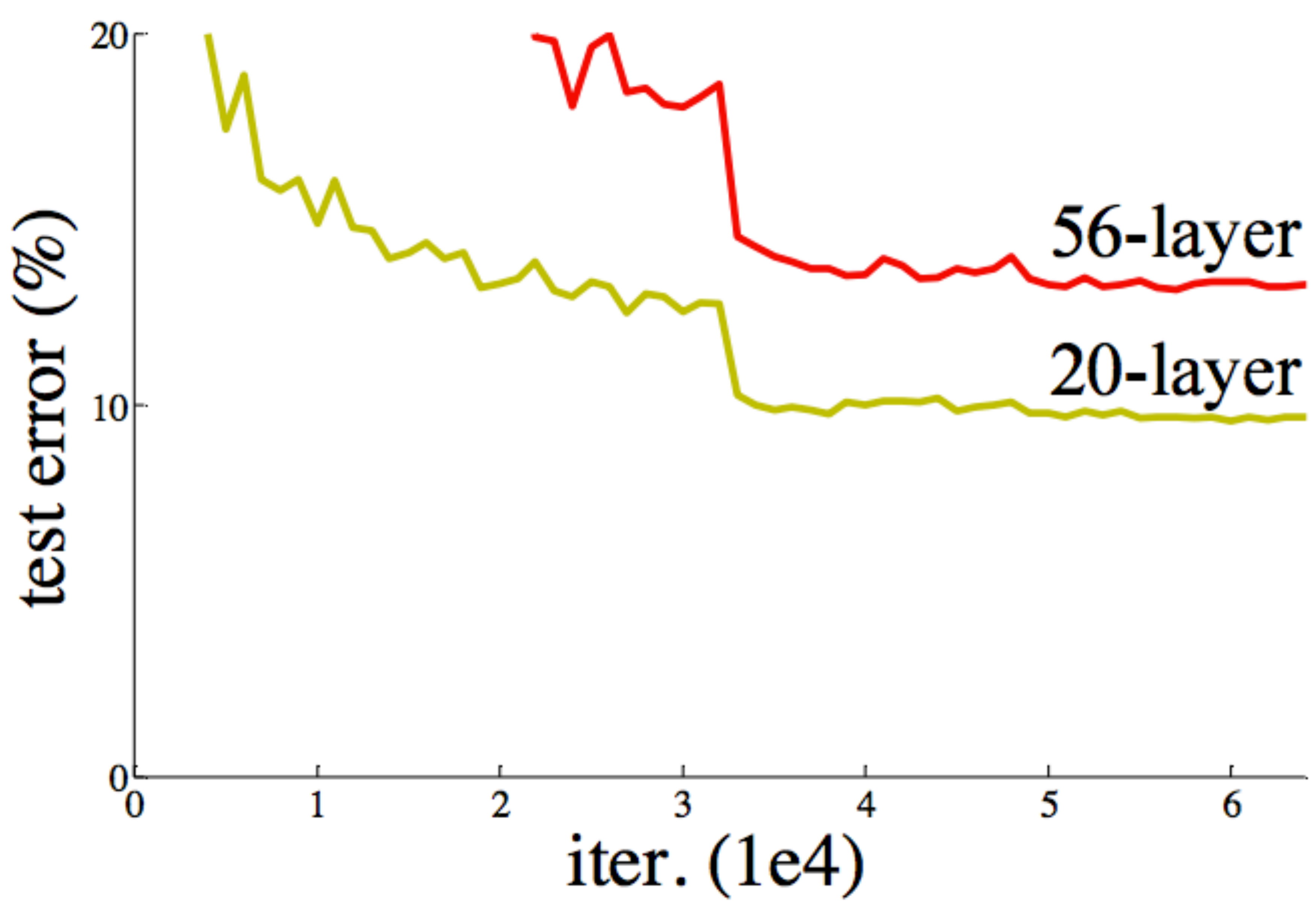


Convolutional Neural Network

Residual Networks

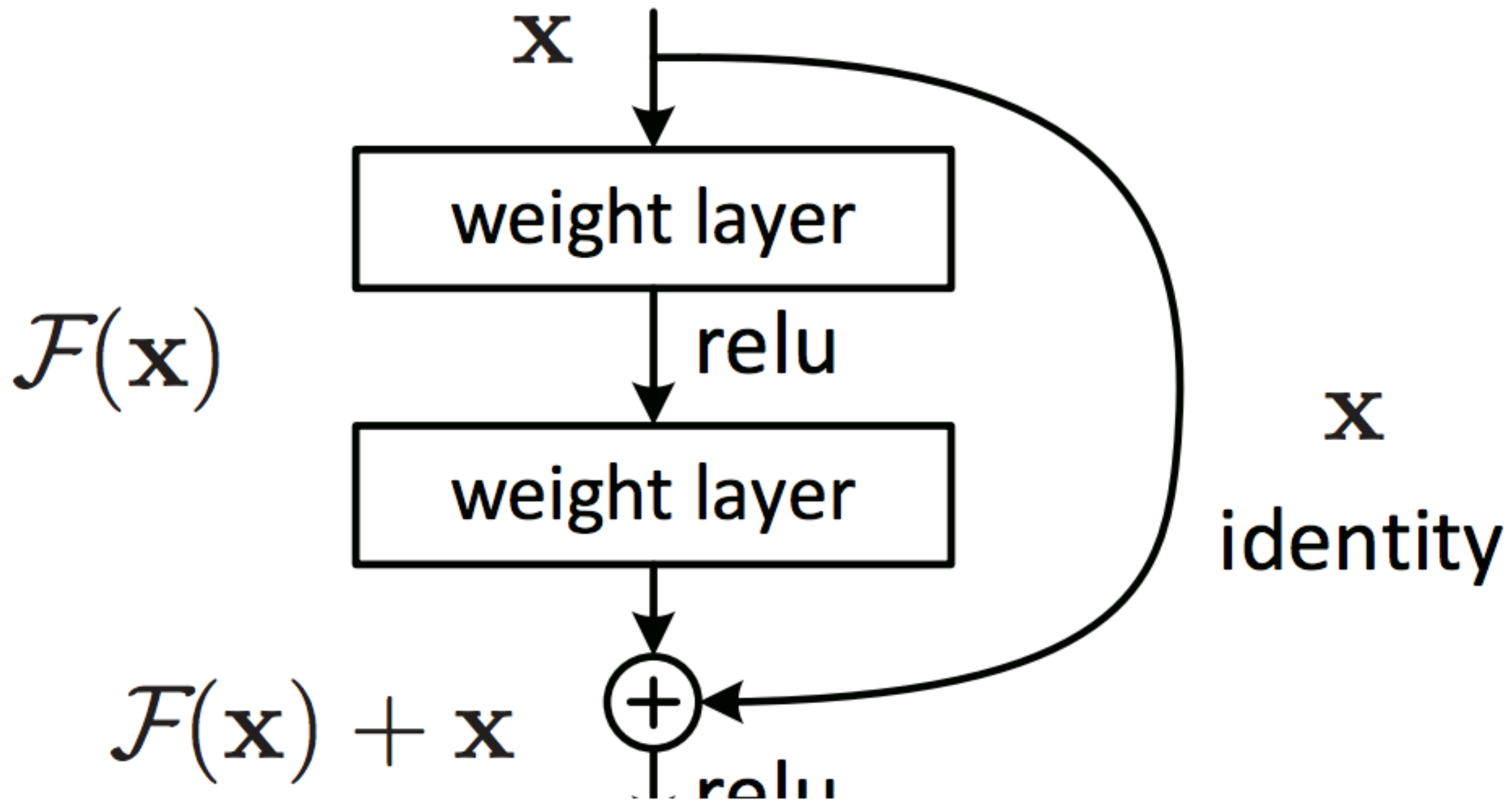
- Very deep convnets do not train well — vanishing gradient problem





Convolutional Neural Network
Residual Networks

- Key idea: introduce ``pass through'' into each layer



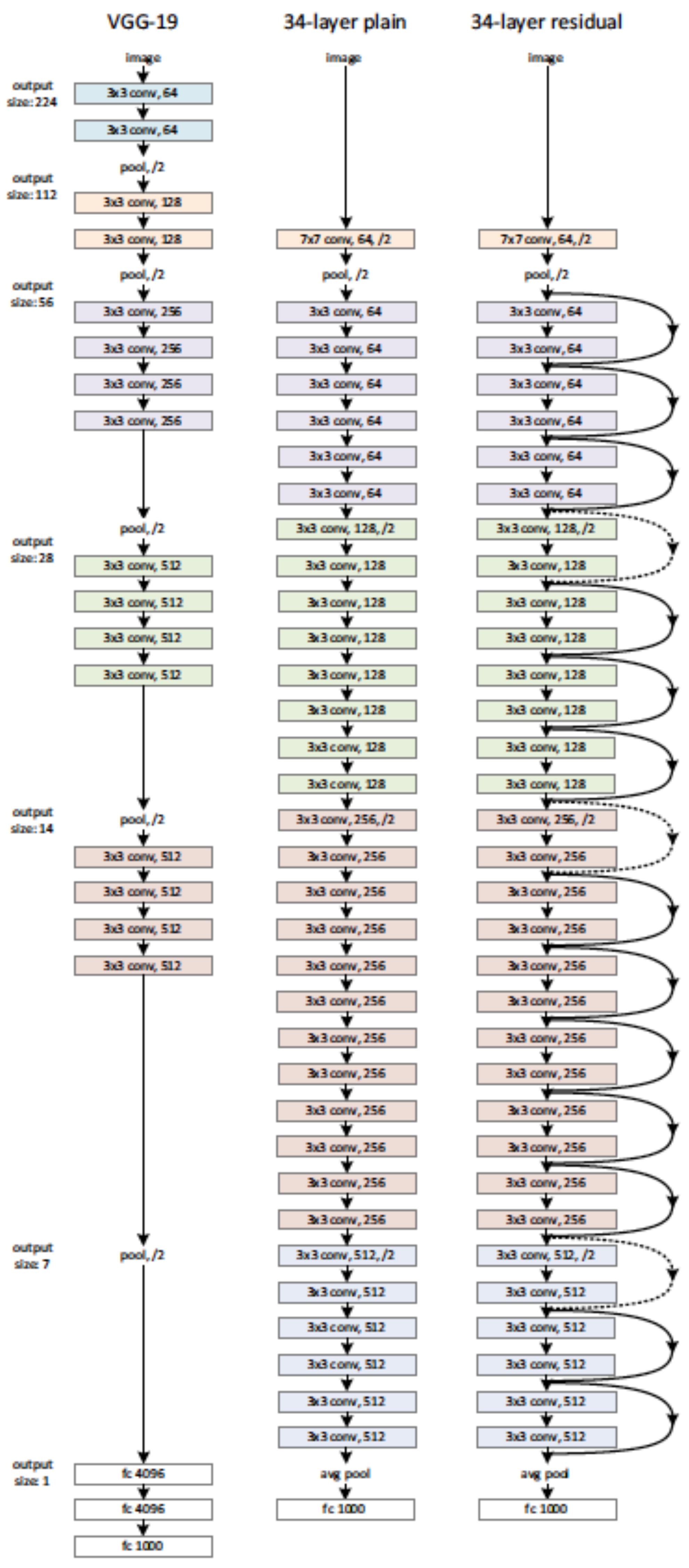
- Thus, only residual needs to be learned

Convolutional Neural Network

Residual Networks

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

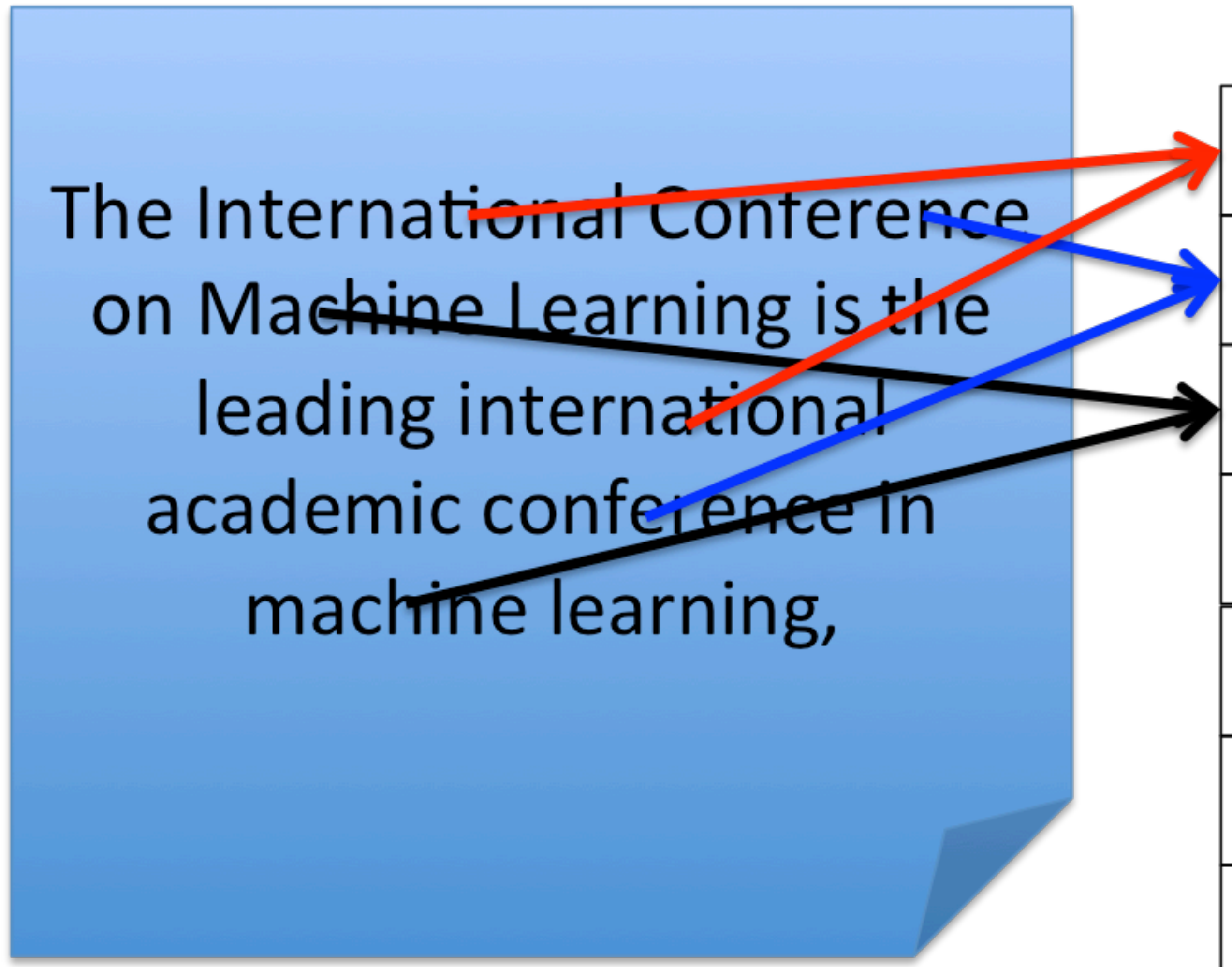
Table 4. Error rates (%) of **single-model** results on the ImageNet validation set (except † reported on the test set).



Representation for sentence/document

Bag of word

- A classical way to represent NLP data
- Each sentence (or document) is represented by a d-dimensional vector $\mathbf{x}$, where x_i is number of occurrences of word i
- number of features = number of potential words (very large)



(international)	2
(conference)	2
(machine)	2
(train)	0
(learning)	2
(leading)	1
(totoro)	0

Representation for sentence/document

Feature generation for documents

- Bag of n-gram features ($n = 2$):
  - 10,000 words $\Rightarrow 10000^2$ potential features

The International Conference on Machine Learning is the leading international academic conference in machine learning,

(international)	2
(conference)	2
(machine)	2
(train)	0
(learning)	2
(leading)	1
(totoro)	0

(international conference)	1
(machine learning)	2
(leading international)	1
(totoro tiger)	0
(tiger woods)	0
(international academic)	1
(international academic)	1

Representation for sentence/document

Bag of word + linear model

- Example: text classification (e.g., sentiment prediction, review score prediction)

- Linear model: $y \approx \text{sign}(w^T x)$ (e.g., by linear SVM/logistic regression)

- w_i: the ``contribution'' of each word

Representation for sentence/document

Bag of word + Fully connected network

- $f(x) = W_L \sigma(W_{L-1} \cdots \sigma(W_0 x))$
- The first layer W_0 is a d_1 by d matrix:
  - Each column w_i is a d_1 dimensional representation of i-th word (word embedding)
  - $W_0 x = x_1 w_1 + x_2 w_2 + \cdots + x_d w_d$ is a linear combination of these vectors
  - W_0 is also called the word embedding matrix
  - Final prediction can be viewed as an $L-1$ layer network on $W_0 x$ (average of word embeddings)
- Not capturing the sequential information

Recurrent Neural Network

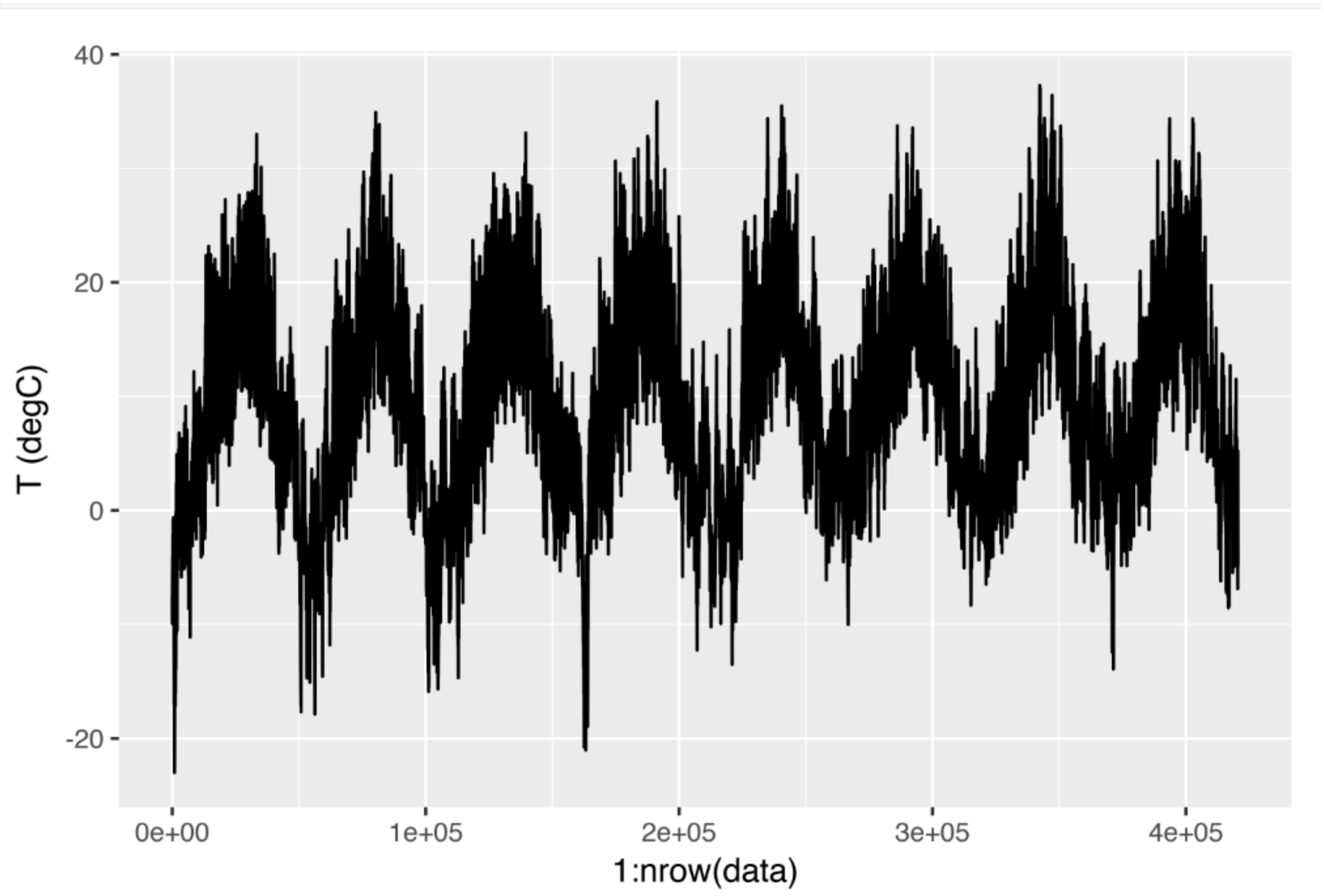
Time series/Sequence data

- Input: $\{x_1, x_2, \cdots, x_T\}$
    - Each x_t is the feature at time step t
    - Each x_t can be a d-dimensional vector
- Output: $\{y_1, y_2, \cdots, y_T\}$
    - Each y_t is the output at step t
    - Multi-class output or Regression output:
        - $y_t \in \{1,2,\cdots, L\}$ or $y_t \in \mathbb{R}$

Recurrent Neural Network

Example: Time Series Prediction

- Climate Data:
  - x_t: temperature at time t
  - y_t: temperature (or temperature change) at time $t + 1$
- Stock Price: Predicting stock price



Recurrent Neural Network

Example: Language Modeling

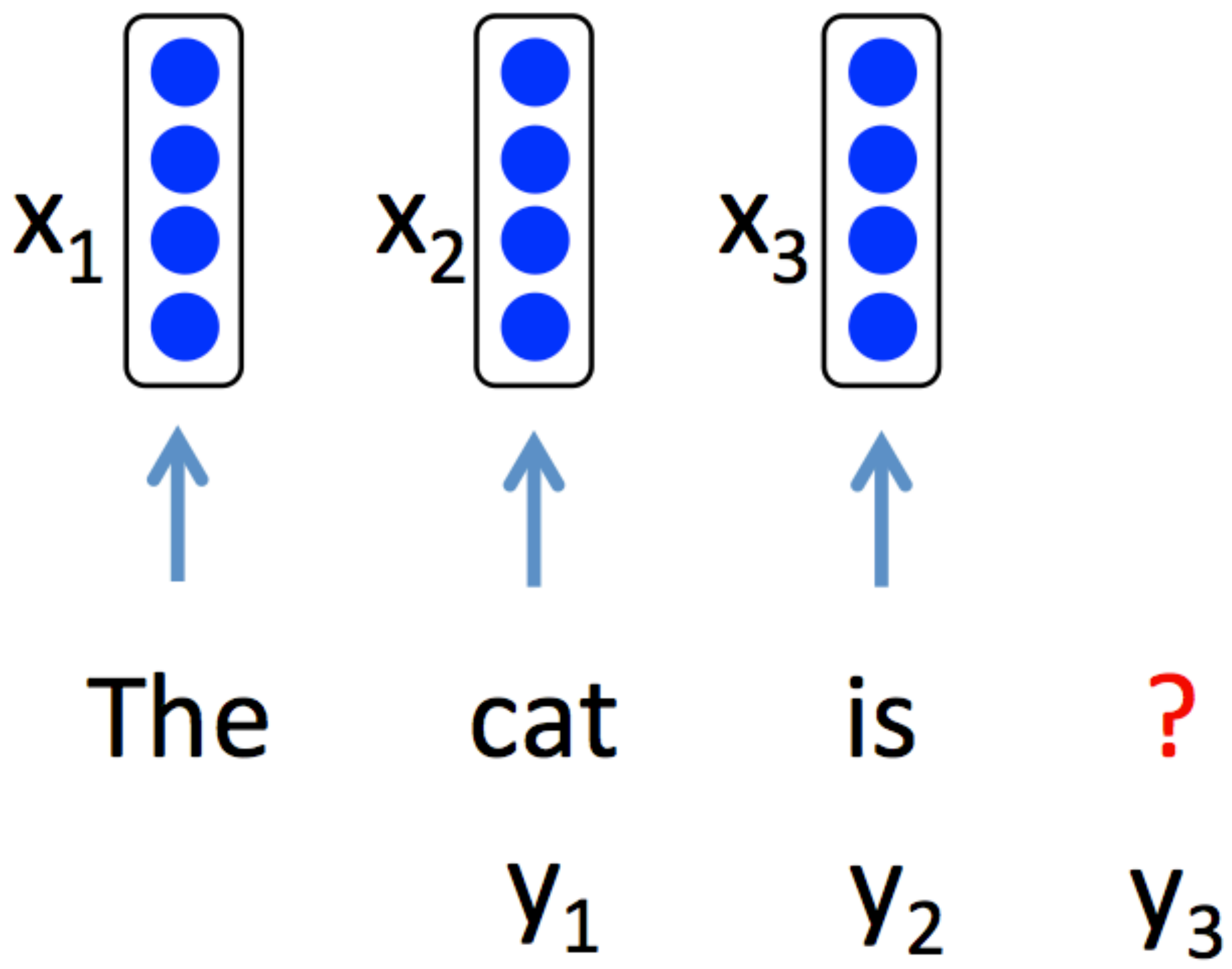
The cat is ?

Recurrent Neural Network

Example: Language Modeling

The cat is ?

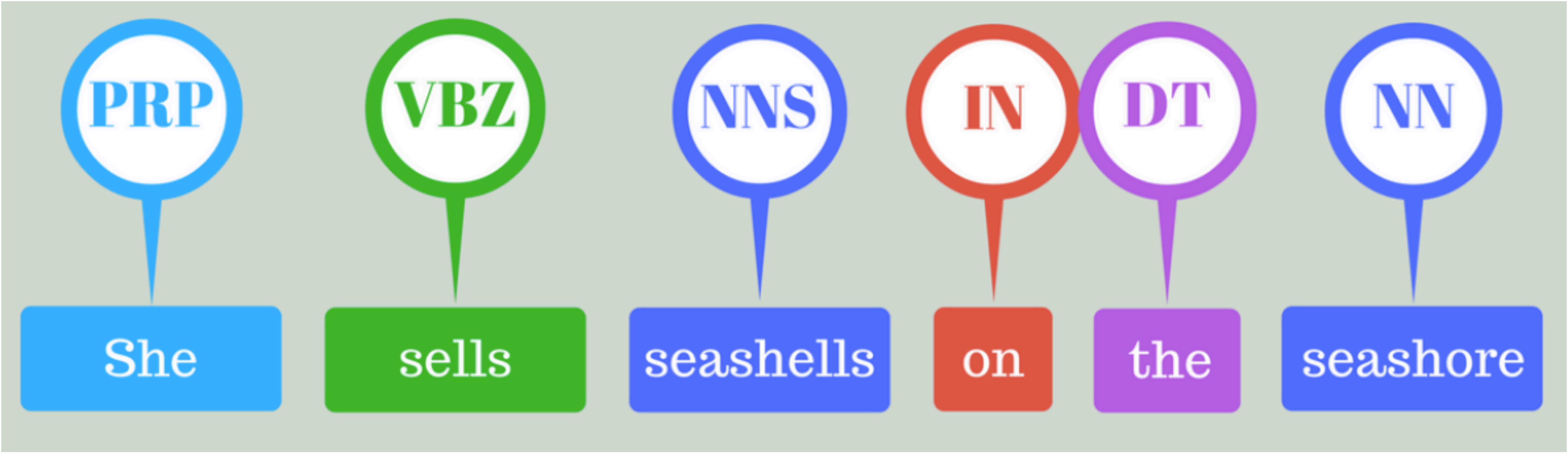
- x_t: one-hot encoding to represent the word at step t ($[0,\ldots,0,1,0,\ldots,0]$)
- y_t: the next word
  - $y_t \in \{1, \cdots, V\}$ V: Vocabulary size



Recurrent Neural Network

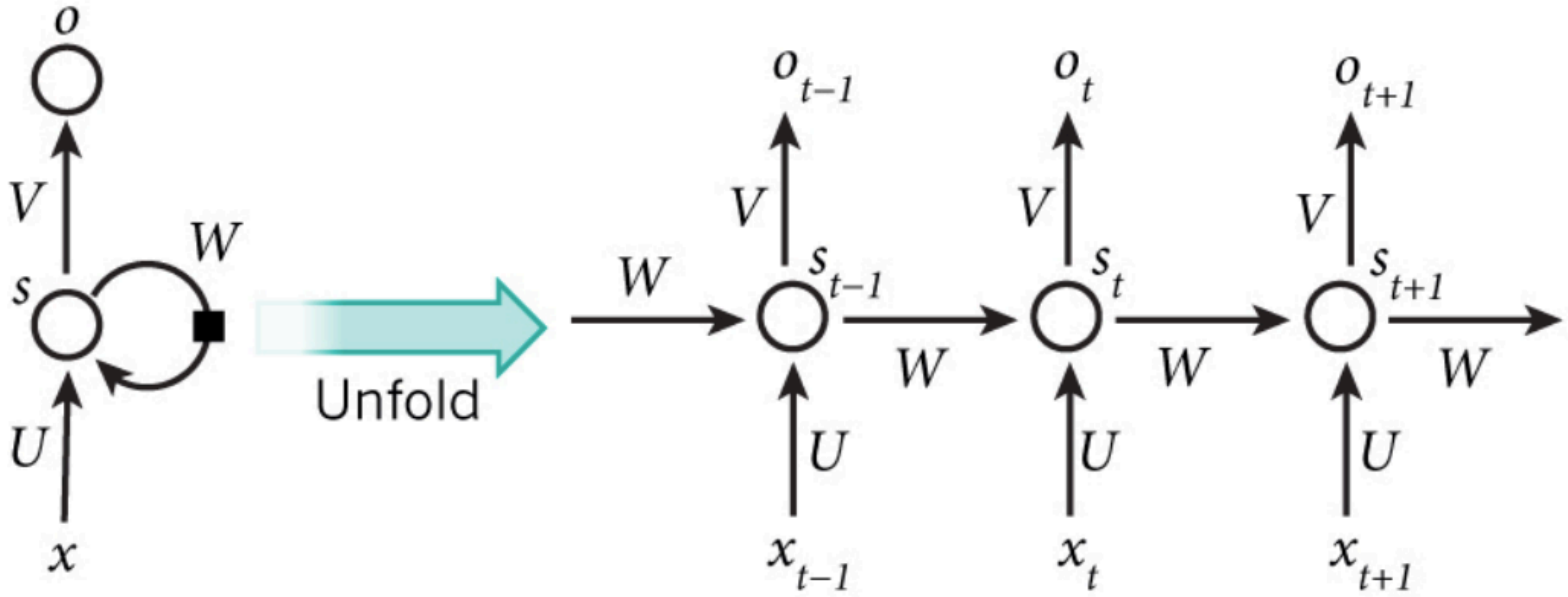
Example: POS Tagging

- Part of Speech Tagging:
    - Labeling words with their Part-Of-Speech (Noun, Verb, Adjective, …)
    - x_t: a vector to represent the word at step t
    - y_t: label of word t



Recurrent Neural Network

Example: POS Tagging



- x_t: t-th input
- s_t: hidden state at time t (`memory'' of the network)
  - $s_t = f(Ux_t + Ws_{t-1})$
  - W: transition matrix, U: word embedding matrix, s_0 usually set to be 0
- Predicted output at time t:
  - $o_t = \arg\max_i (Vs_t)_i$

Recurrent Neural Network

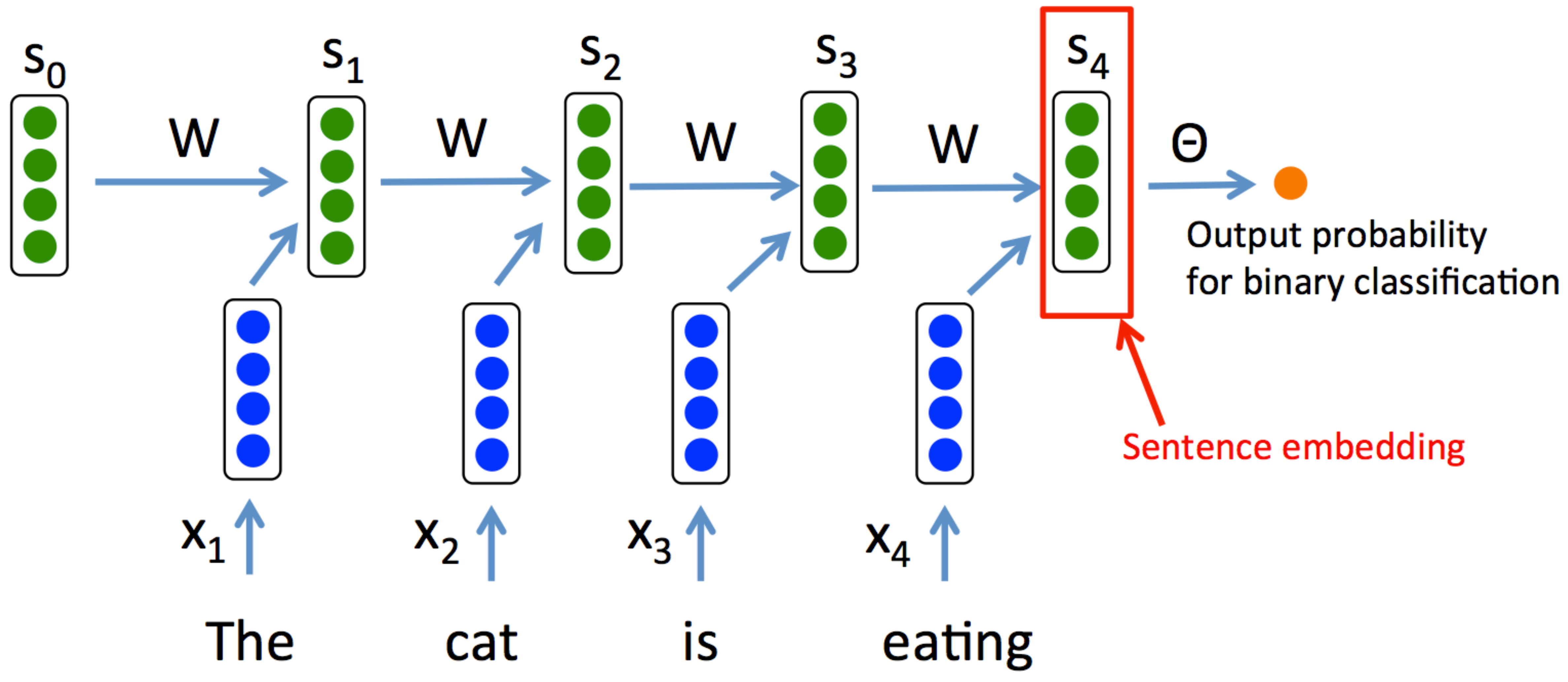
Recurrent Neural Network (RNN)

- Training: Find U, W, V to minimize empirical loss:
- Loss of a sequence:
  - $$\sum_{t=1}^{T} \text{loss}(Vs_t, y_t)$$
  - (s_t is a function of U, W, V)
- Loss on the whole dataset:
  - Average loss over all sequences
- Solved by SGD/Adam

Recurrent Neural Network

RNN: Text Classification

- Not necessary to output at each step
- Text Classification:
  - sentence → category
  - Output only at the final step
- Model: add a fully connected network to the final embedding



Recurrent Neural Network

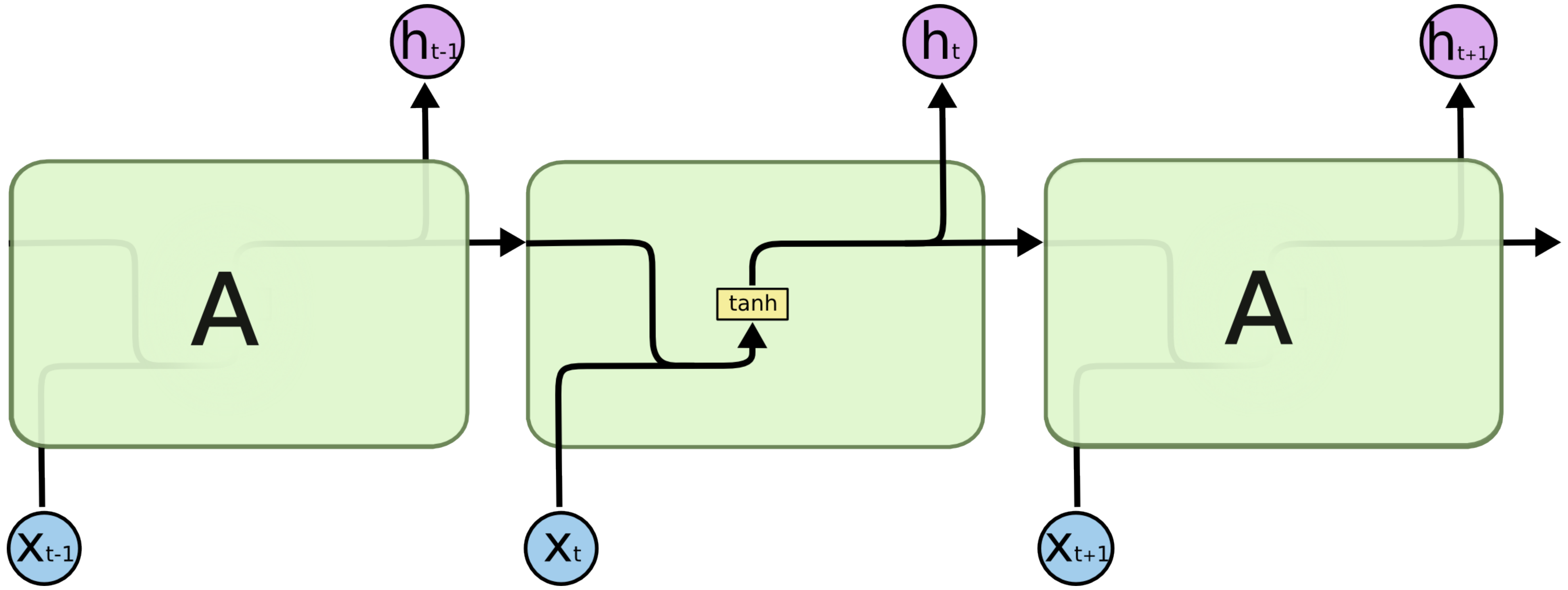
Problems of Classical RNN

- Hard to capture long-term dependencies
- Hard to solve (vanishing gradient problem)
- Solution:
  - LSTM (Long Short Term Memory networks)
  - GRU (Gated Recurrent Unit)
  - …

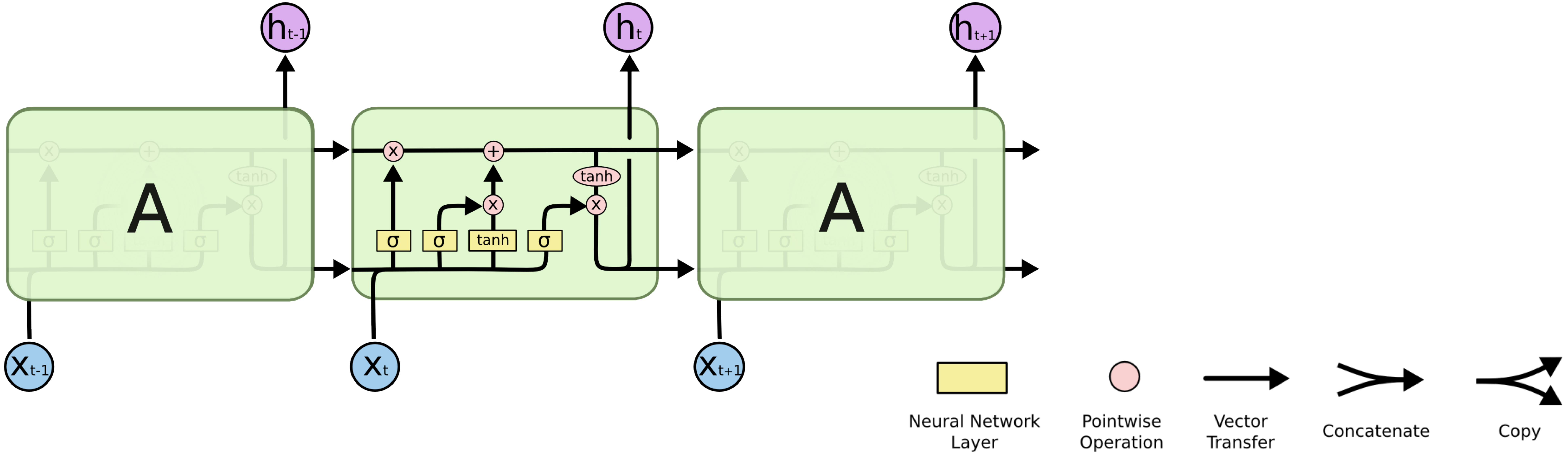
Recurrent Neural Network

LSTM

- RNN:



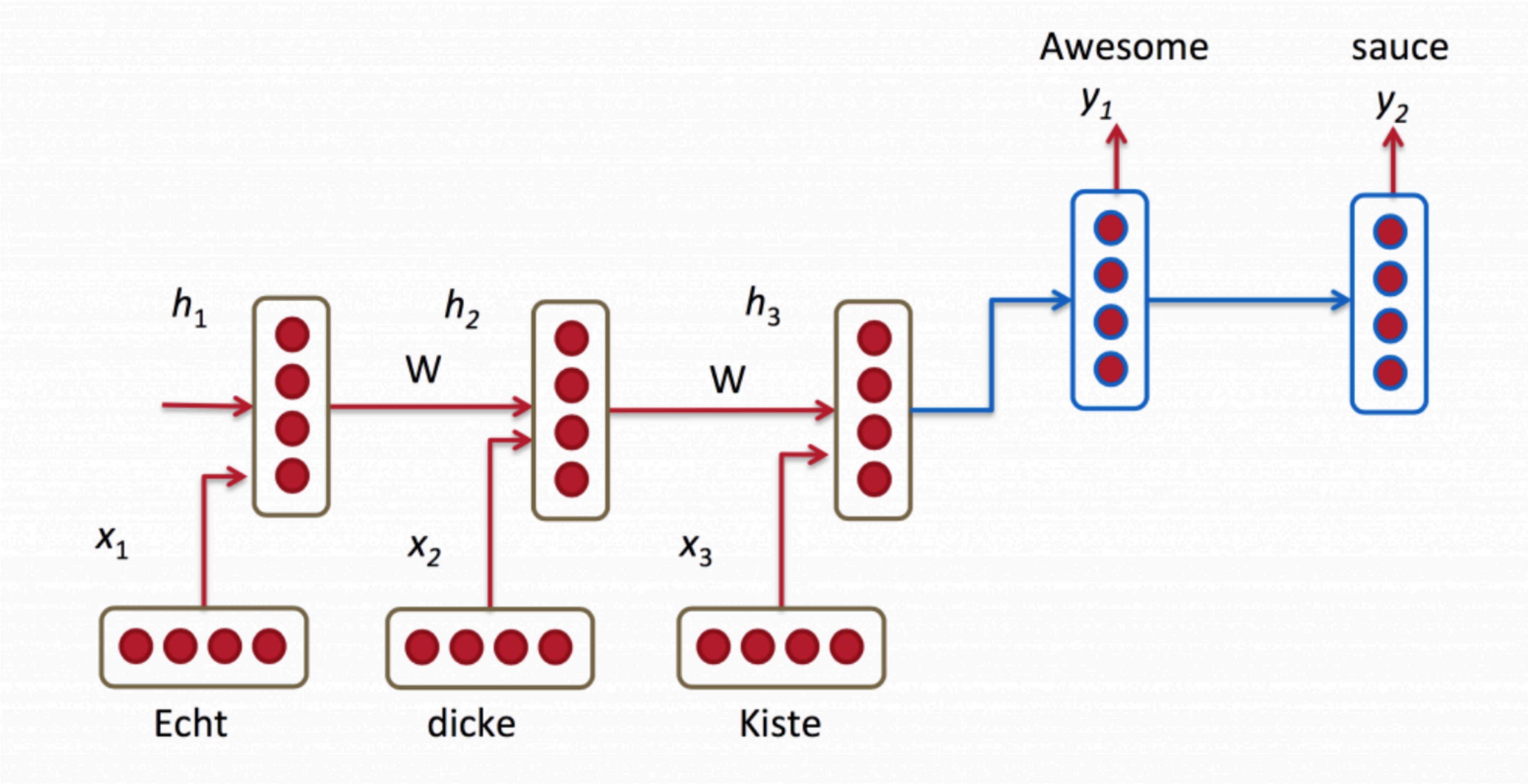
- LSTM:



Recurrent Neural Network

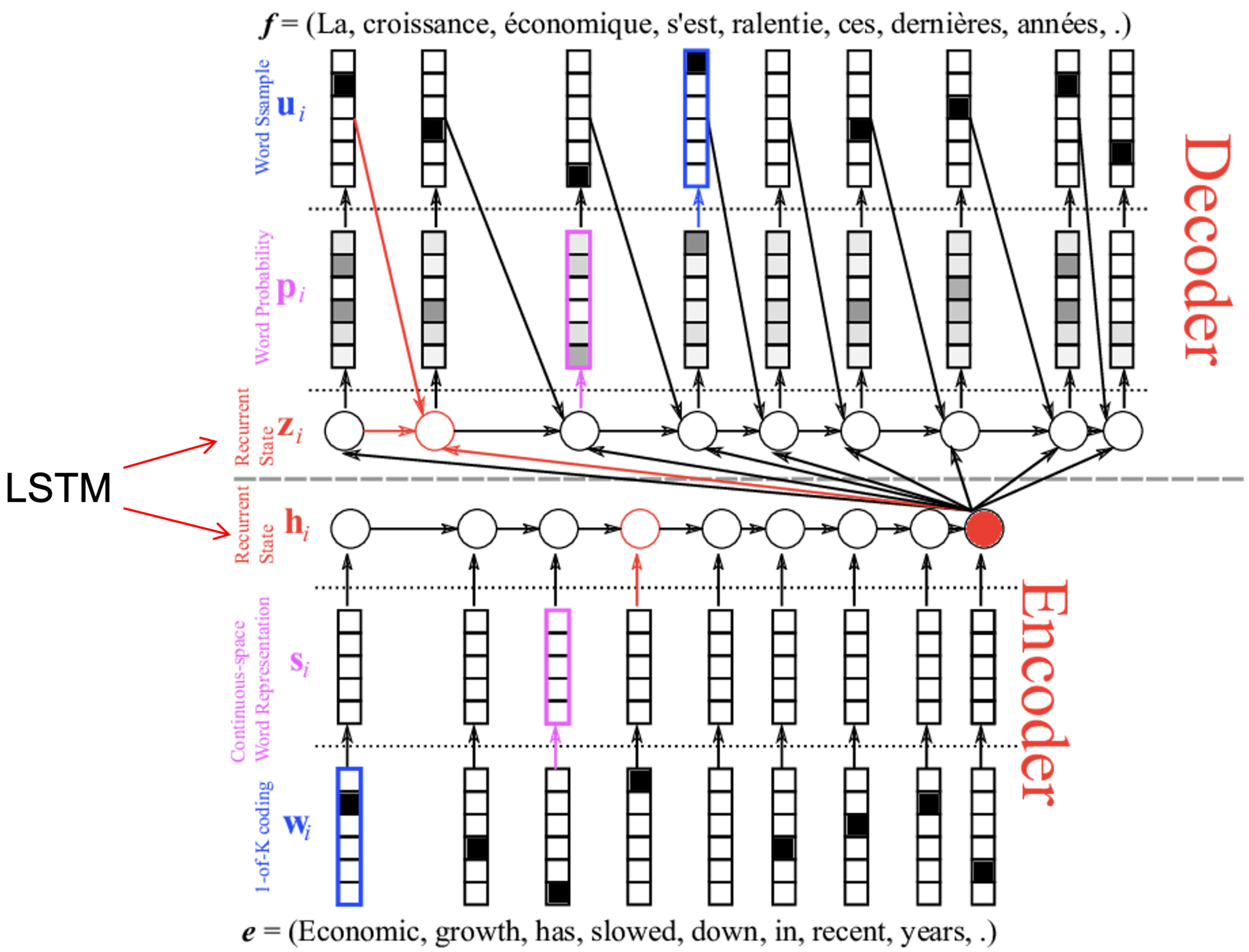
Neural Machine Translation (NMT)

- Out the translated sentence from an input sentence
- Training data: a set of input-output pairs (supervised setting)
- Encoder-decoder approach:
  - Encoder: Use (RNN/LSTM) to encode the input sentence input a latent vector
  - Decoder: Use (RNN/LSTM) to generate a sentence based on the latent vector



Recurrent Neural Network

Neural Machine Translation



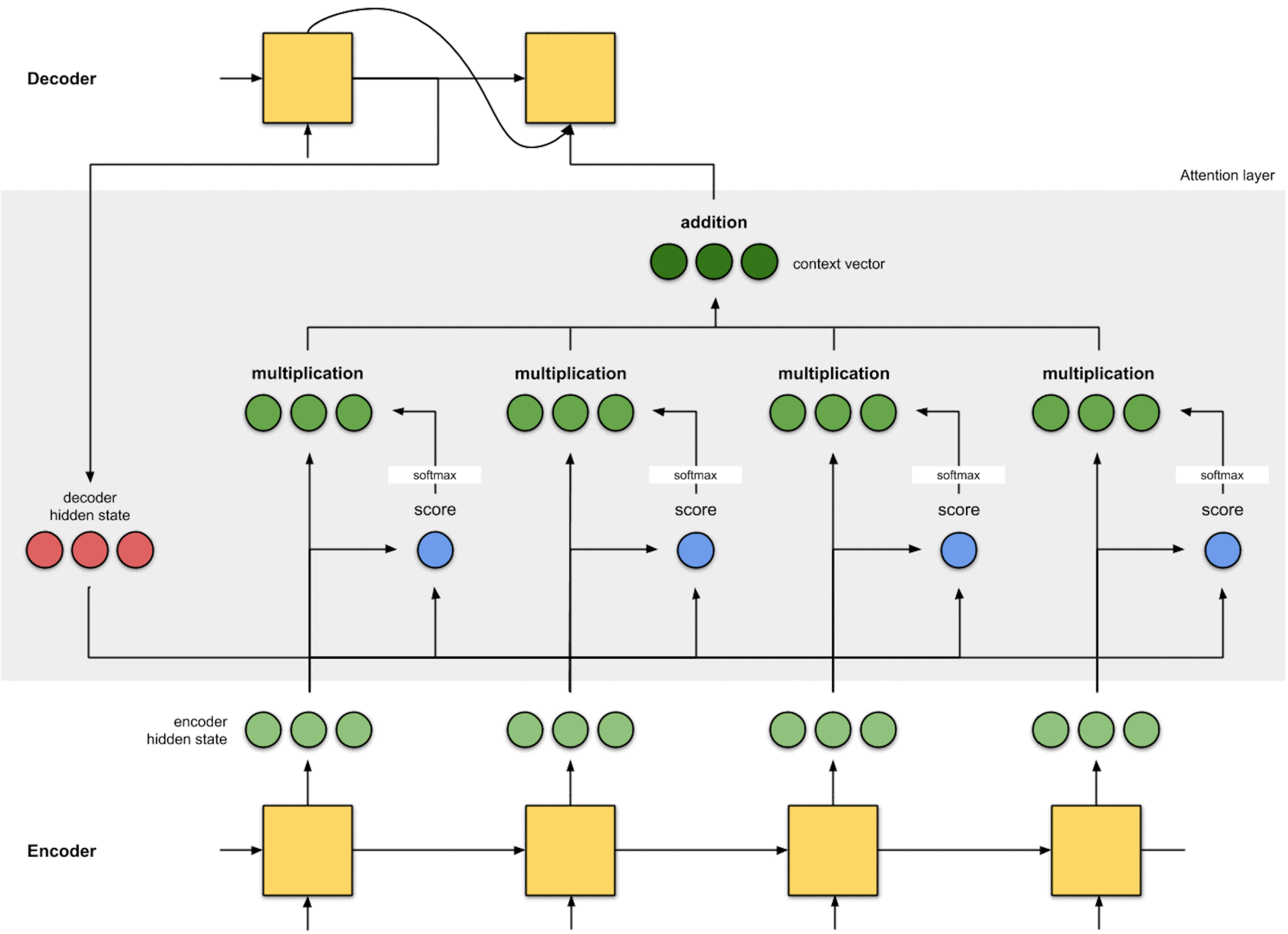
Recurrent Neural Network

Attention in NMT

- Usually, each output word is only related to a subset of input words (e.g., for machine translation)
- Let u be the current decoder latent state, $v_1, \ldots, v_n$ be the latent sate for each input word
- Compute the weight of each state by
  - $p = \text{Softmax}(u^T v_1, \ldots, u^T v_n)$
- Compute the context vector by $Vp = p_1 v_1 + \ldots + p_n v_n$

Recurrent Neural Network

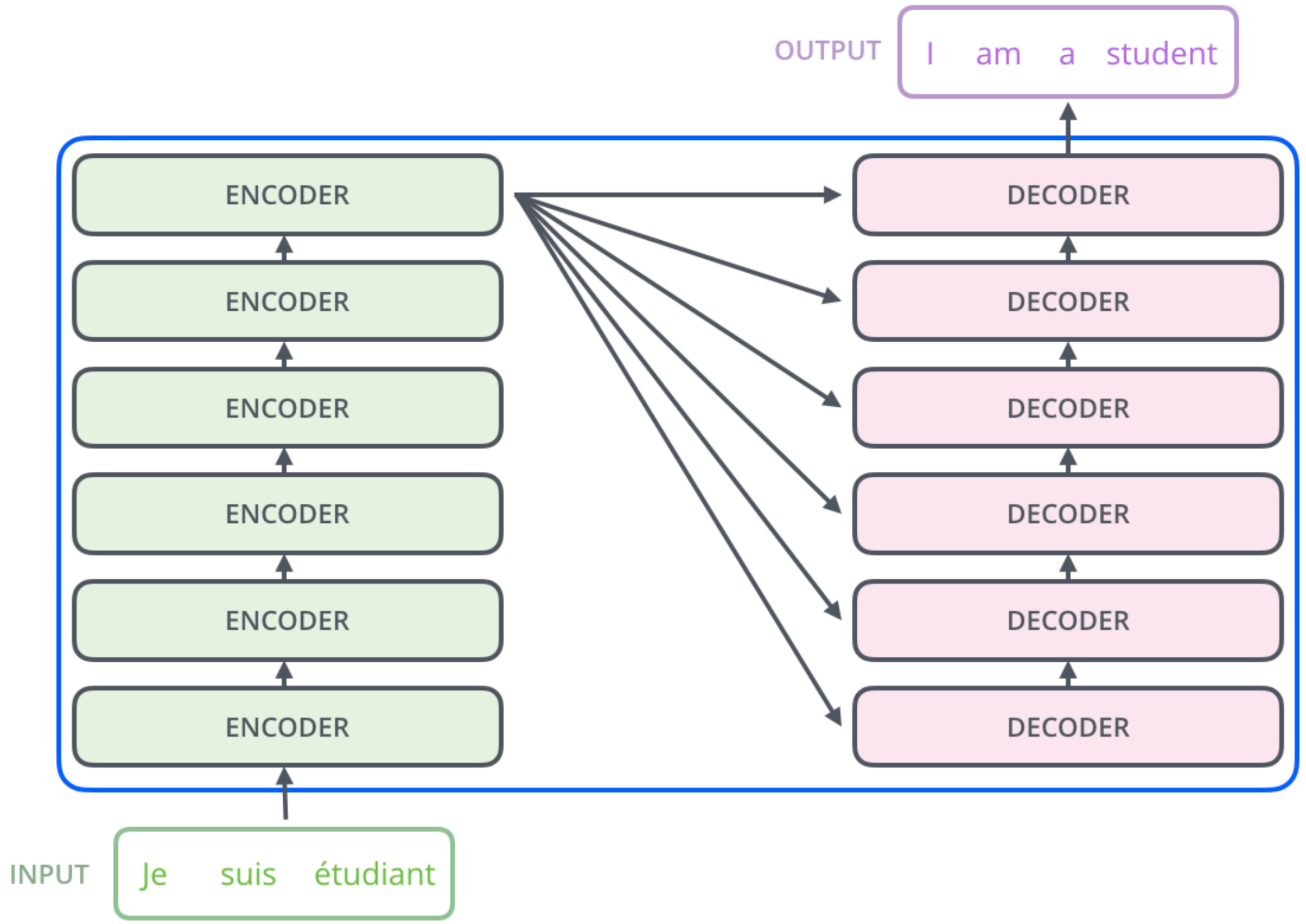
Attention in NMT



Transformer

Transformer

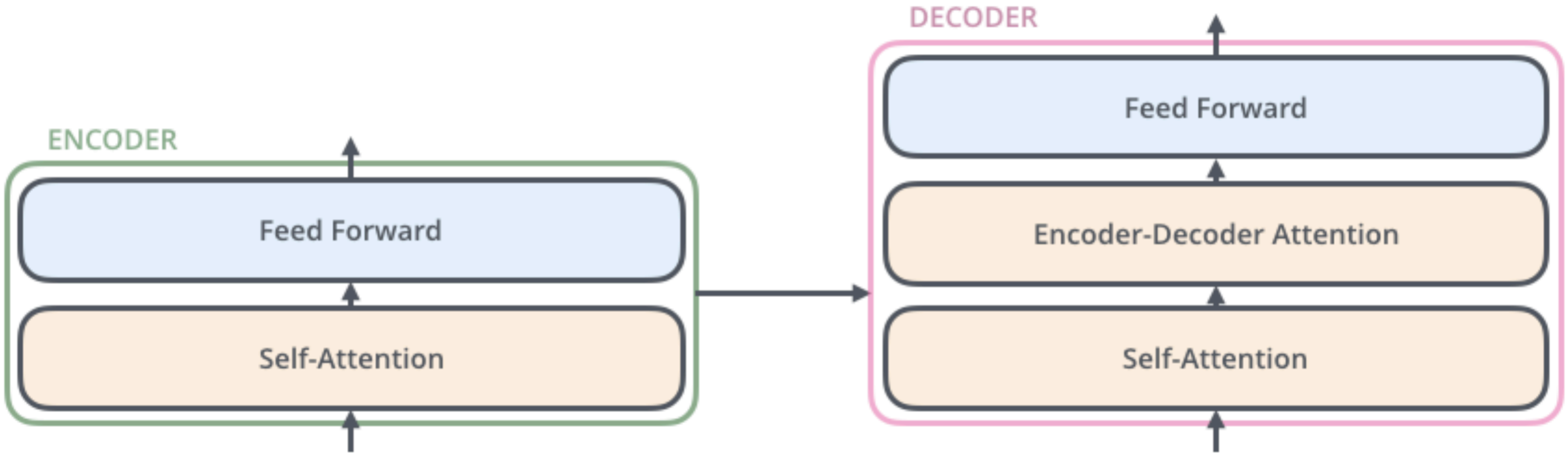
- An architecture that replies entirely on attention without using CNN/RNN
- Proposed in ``Attention Is All You Need'' (Vaswani et al., 2017)
- Initially used for neural machine translation



Transformer

Encoder and Decoder

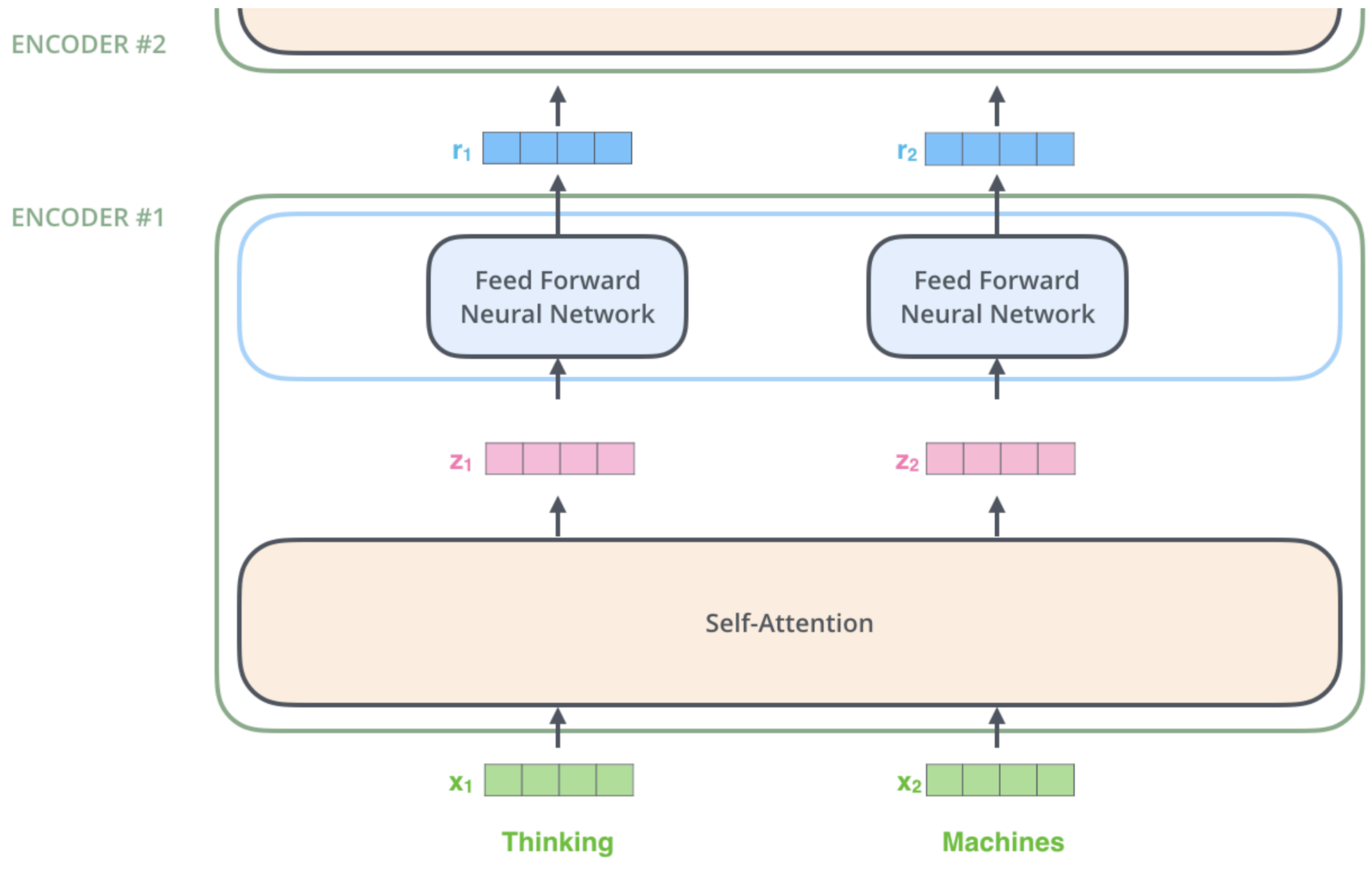
- Self attention layer: the main architecture used in Transformer
- Decoder: will have another attention layer to help it focuses on relevant parts of input sentences.



Transformer

Encoder

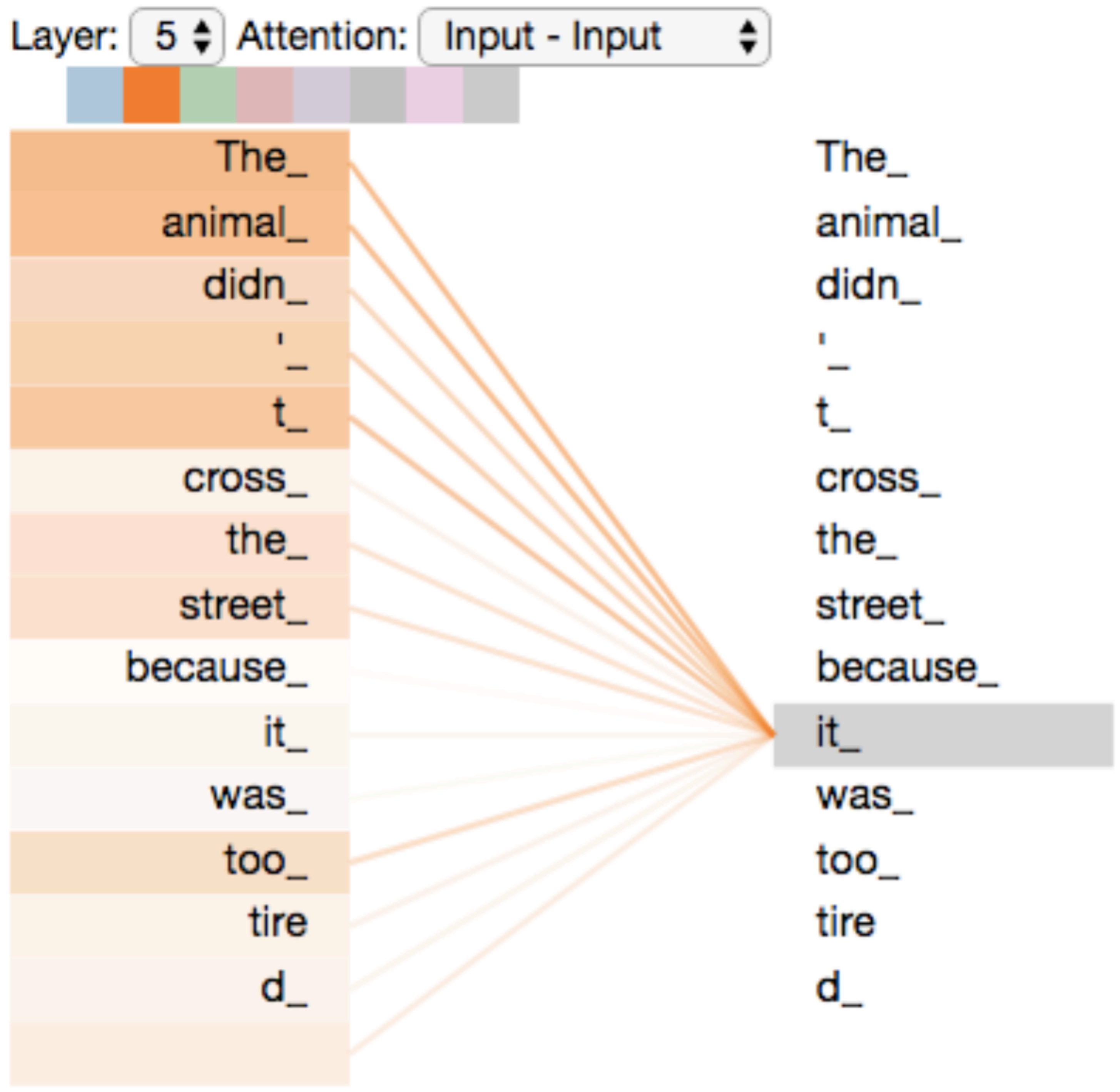
- Each word has a corresponding ``latent vector'' (initially the word embedding for each word)
- Each layer of encoder:
    - Receive a list of vectors as input
    - Passing these vectors to a self-attention layer
    - Then passing them into a feed-foward layer
    - Output a list of vectors



Transformer

Self-attention layer

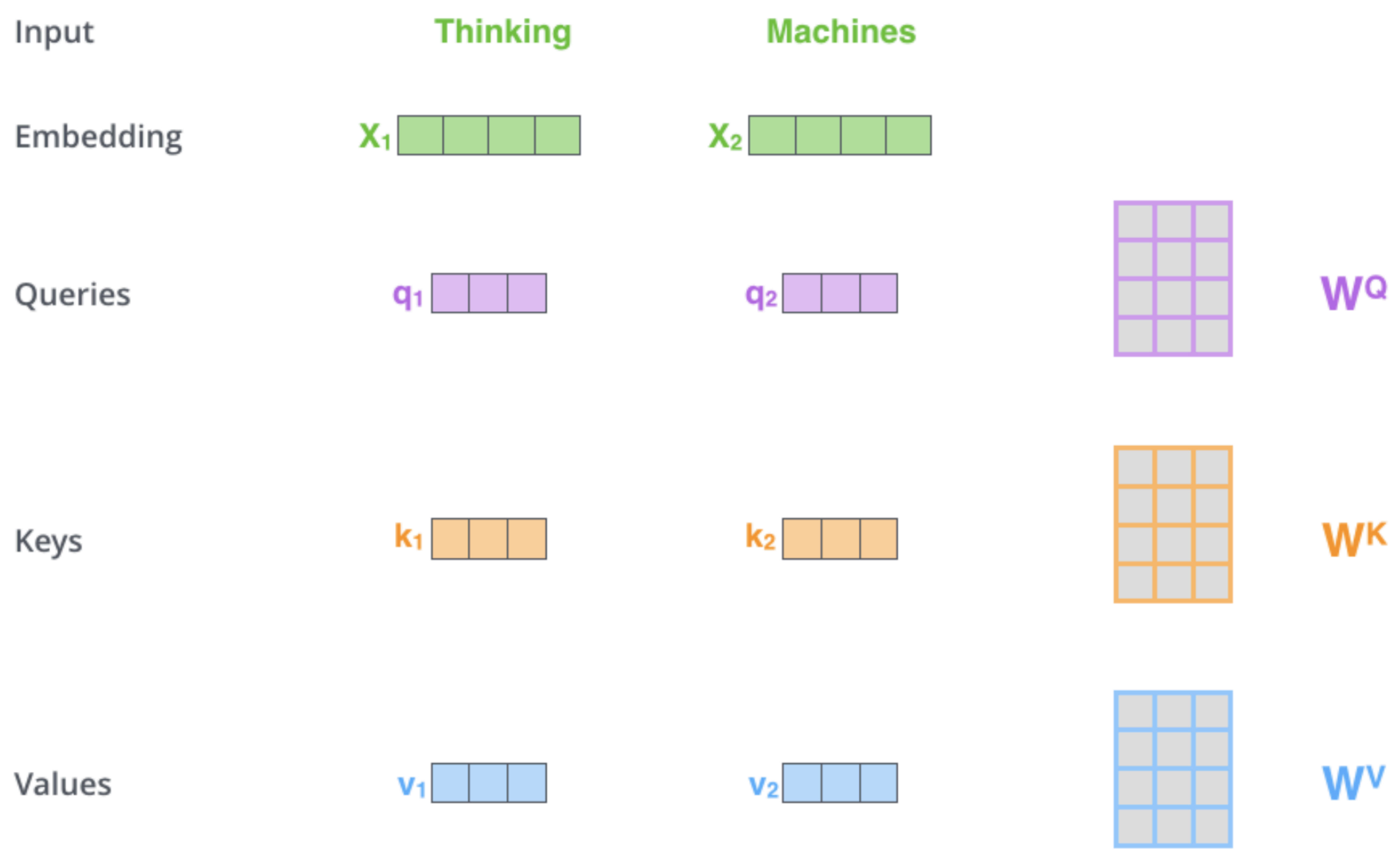
- Main idea: The actual meaning of each word may be related to other words in the sentence
- The actual meaning (latent vector) of each word is a weighted (attention) combination of other words (latent vectors) in the sentences



Transformer

Self-attention layer

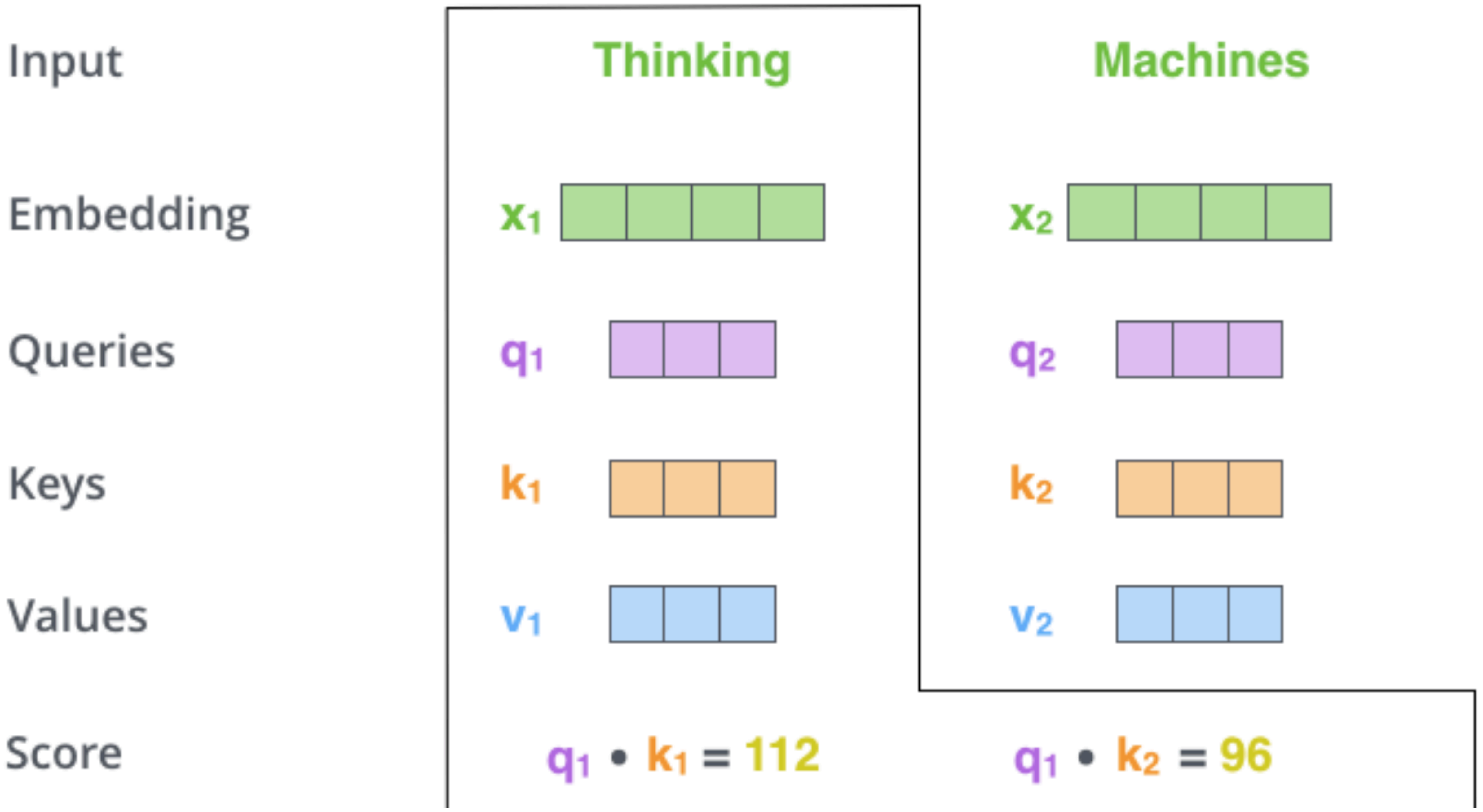
- Input latent vectors: $x_1, \ldots, x_n$
- Self-attention parameters: W^Q, W^K, W^V (weights for query, key, value)
- For each word i, compute
  - Query vector: $q_i = x_i W^Q$
  - Key vector: $k_i = x_i W^K$
  - Value vector: $v_i = x_i W^V$



Transformer

Self-attention layer

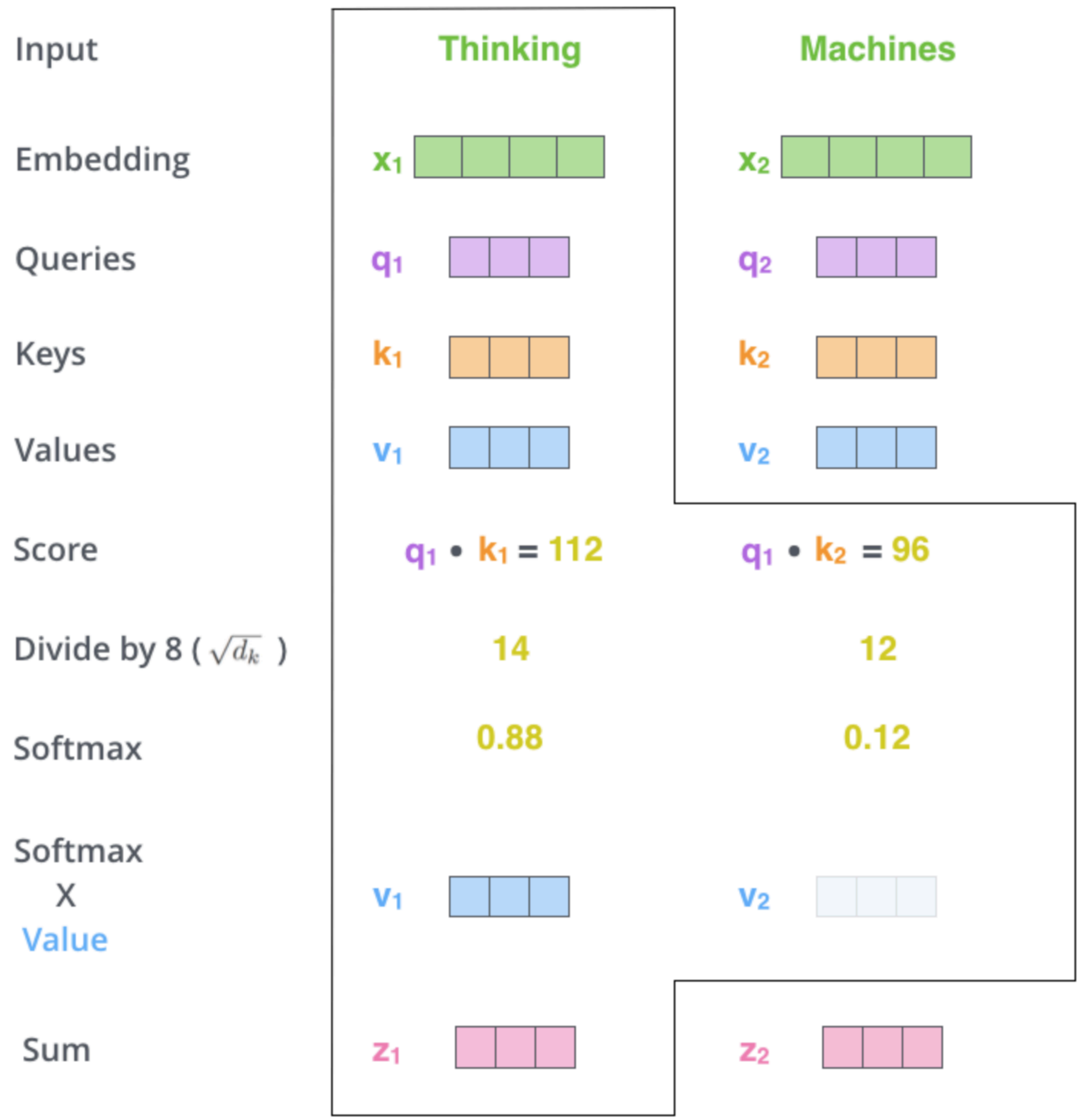
- For each word i, compute the scores to determine how much focus to place on other input words
  - The attention score for word j to word i: $q_i^T k_j$



Transformer

Self-attention layer

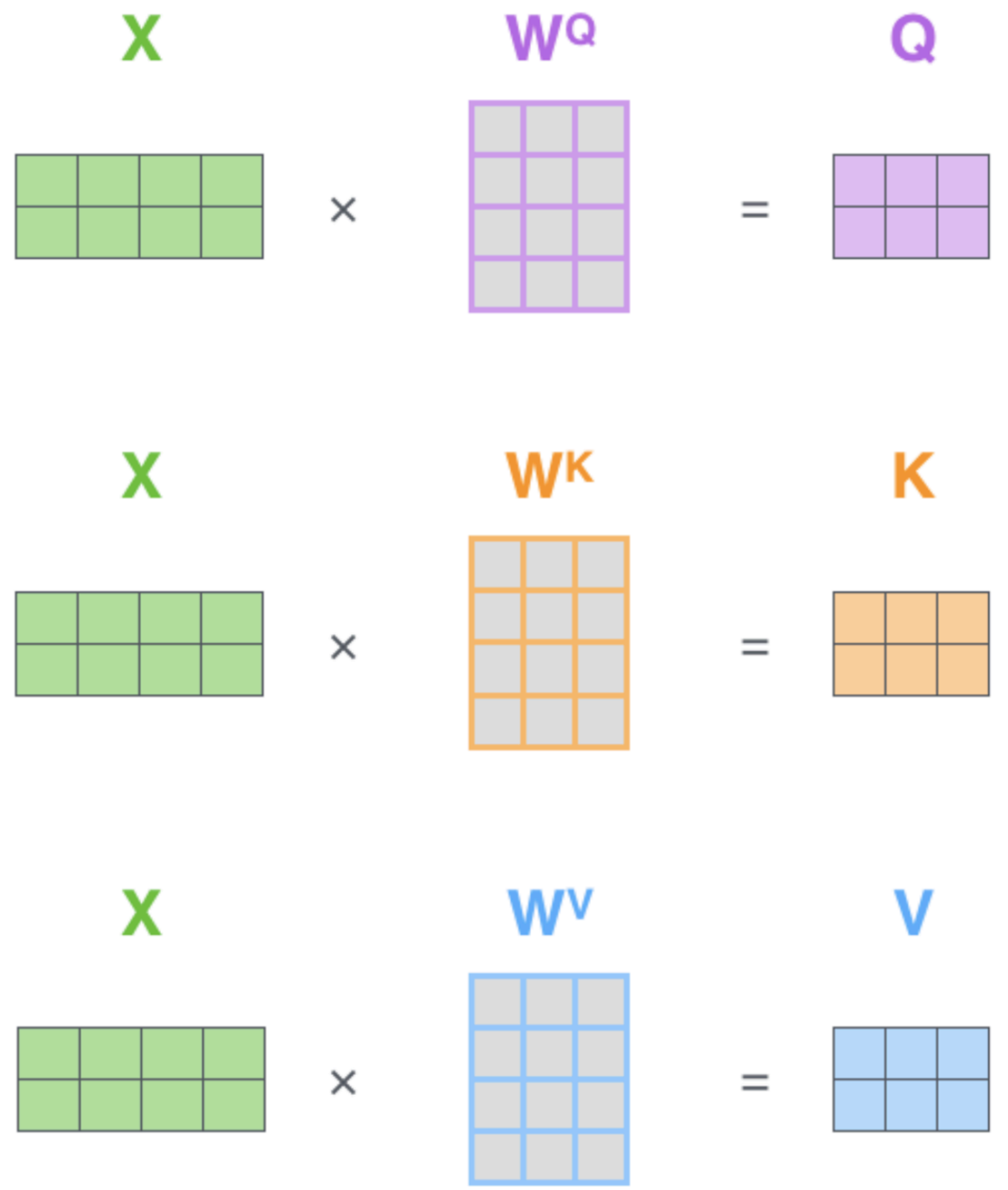
- For each word i, the output vector
  - $\sum_j s_{ij} v_j, \quad s_i = \text{softmax}(q_i^T k_1, \ldots, q_i^T k_n)$

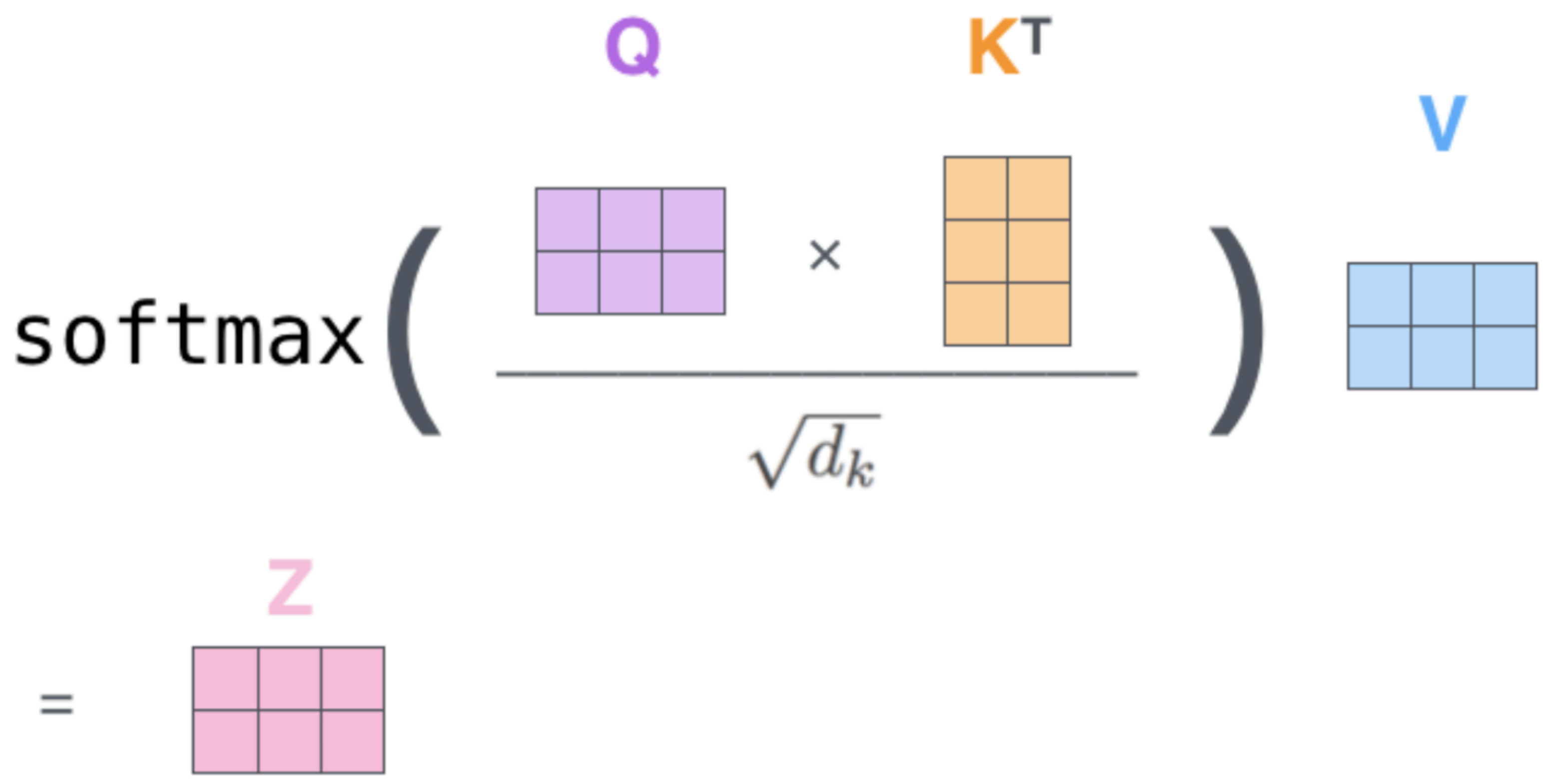


Transformer

Matrix form

- $Q = XW^Q,\ K = XW^K,\ V = XW^V,\quad Z = \text{softmax}(QK^T)V$

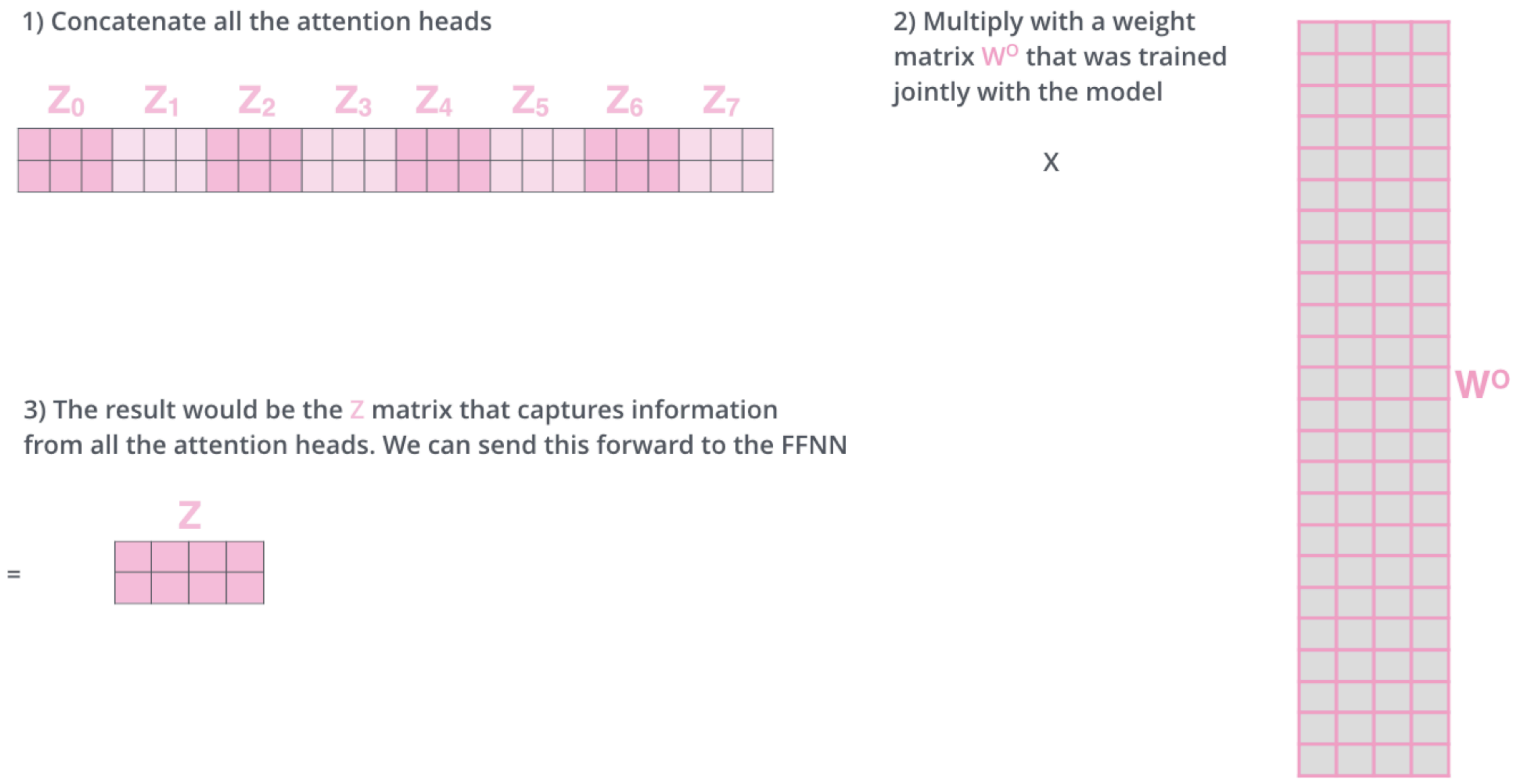




Transformer

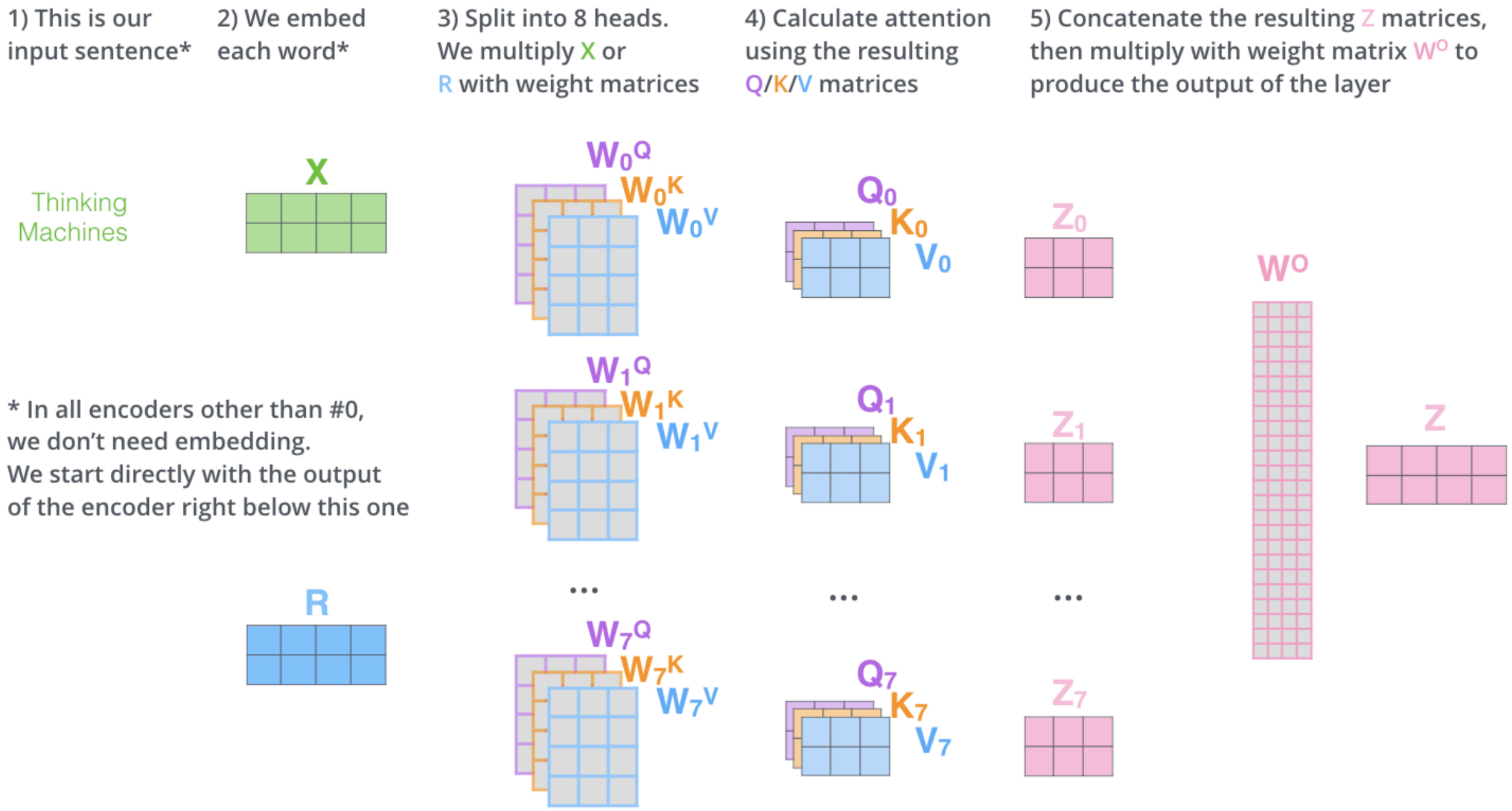
Multiply with weight matrix to reshape

- Gather all the outputs $Z_1, \ldots, Z_k$
- Multiply with a weight matrix to reshape
- Then pass to the next fully connected layer



Transformer

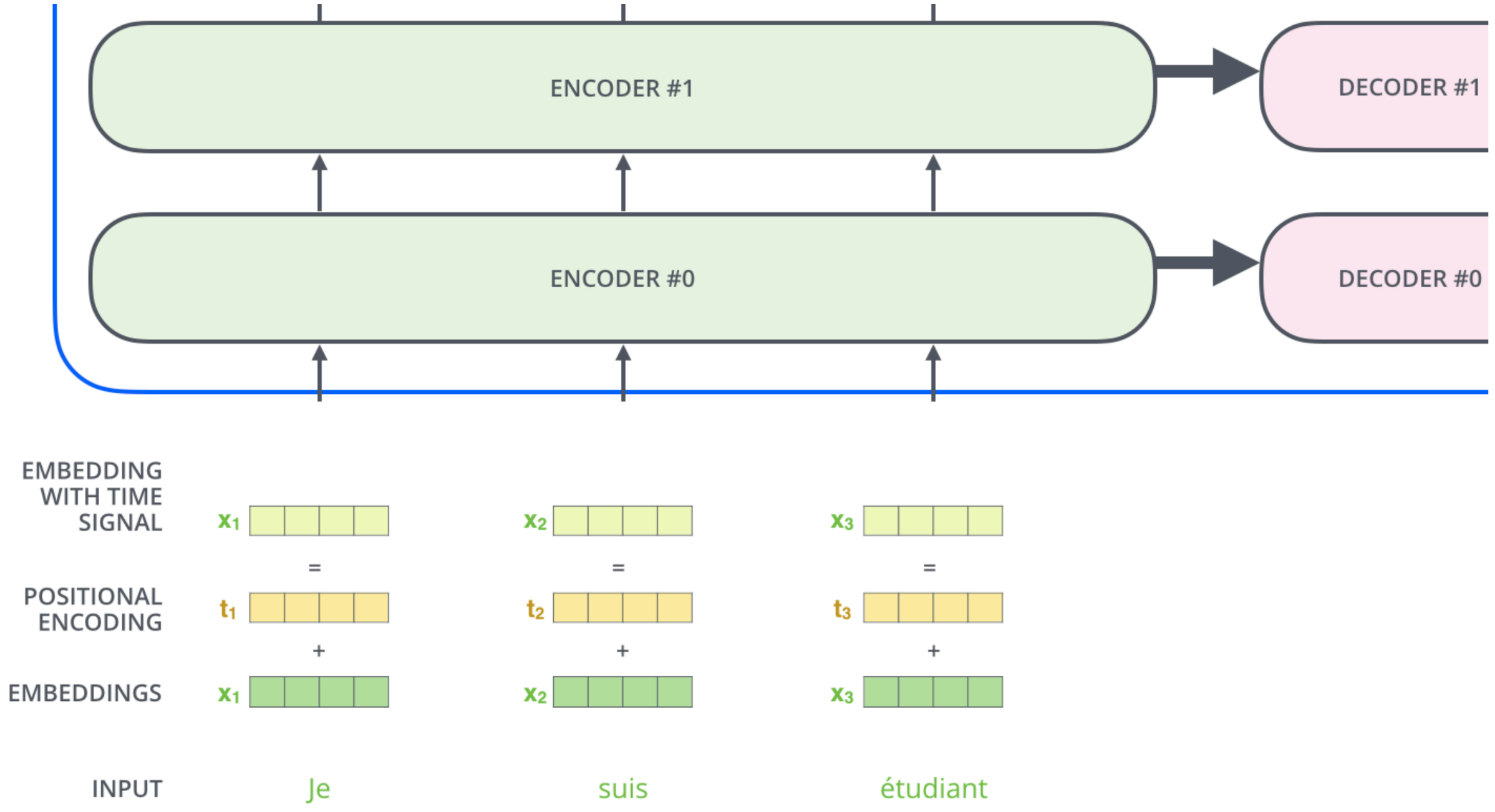
Overall architecture



Transformer

Sinusoidal Position Encoding

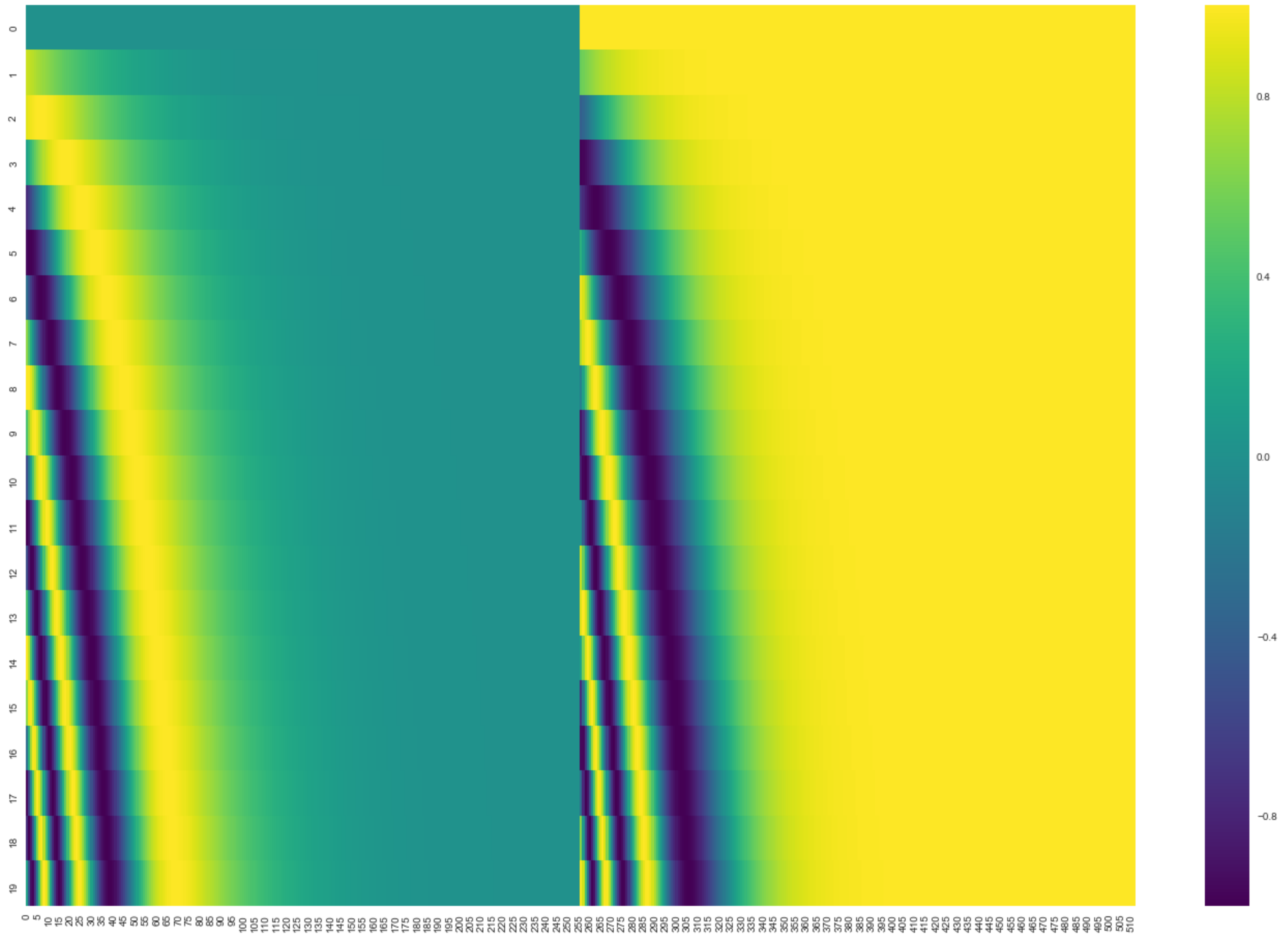
- The above architecture ignores the sequential information
- Add a positional encoding vector to each x_i (according to i)



Transformer

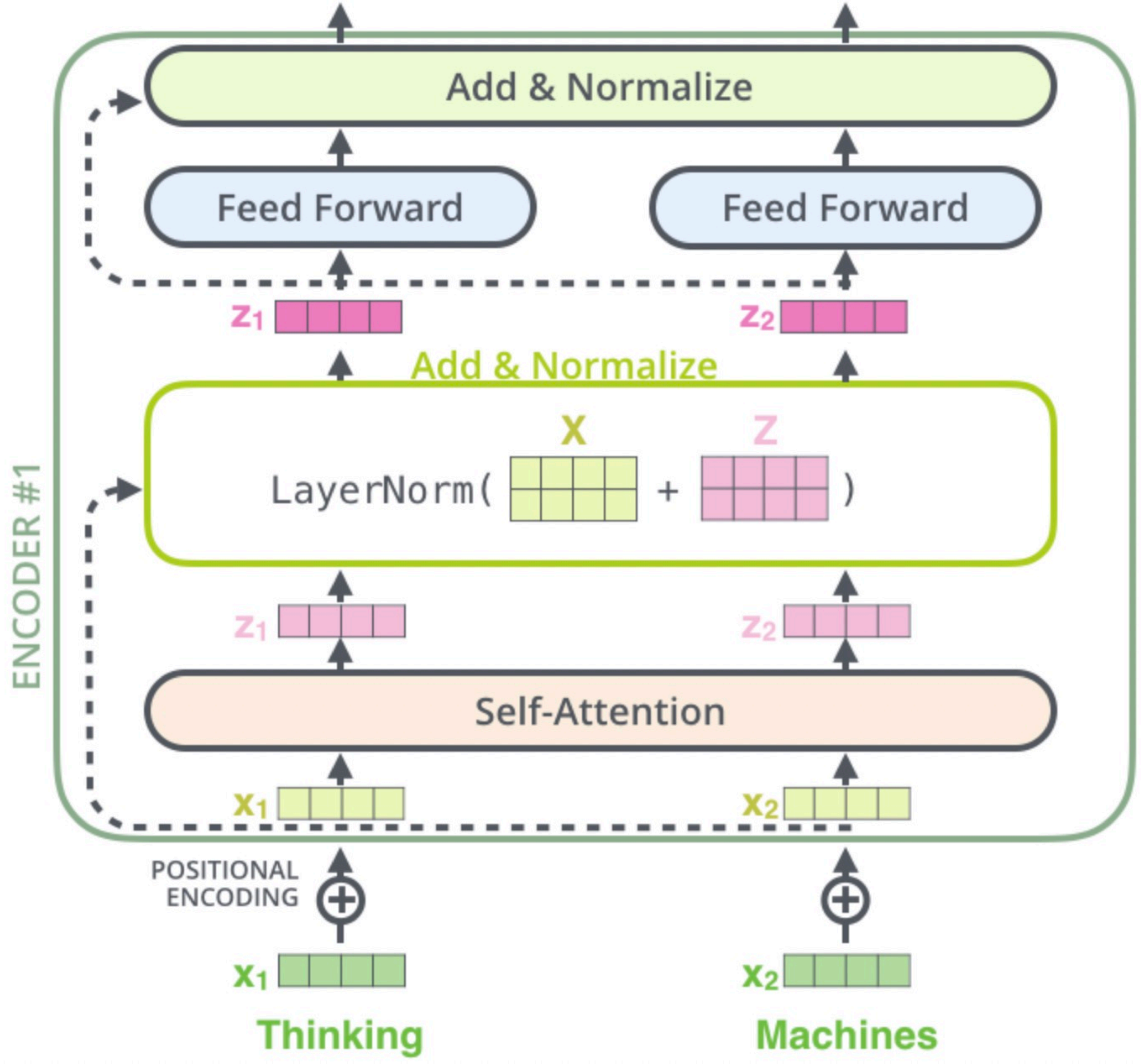
Positional Embedding

- Sin/cosine functions with different wavelengths (used in the original Transformer)
- The jth dimension of ith token $p_i[j] = \begin{cases} \sin(i \cdot c^{\frac{j}{d}}) & \text{if j is even} \\ \cos(i \cdot c^{\frac{j-1}{d}}) & \text{if j is odd} \end{cases}$
- smooth, parameter-free, inductive



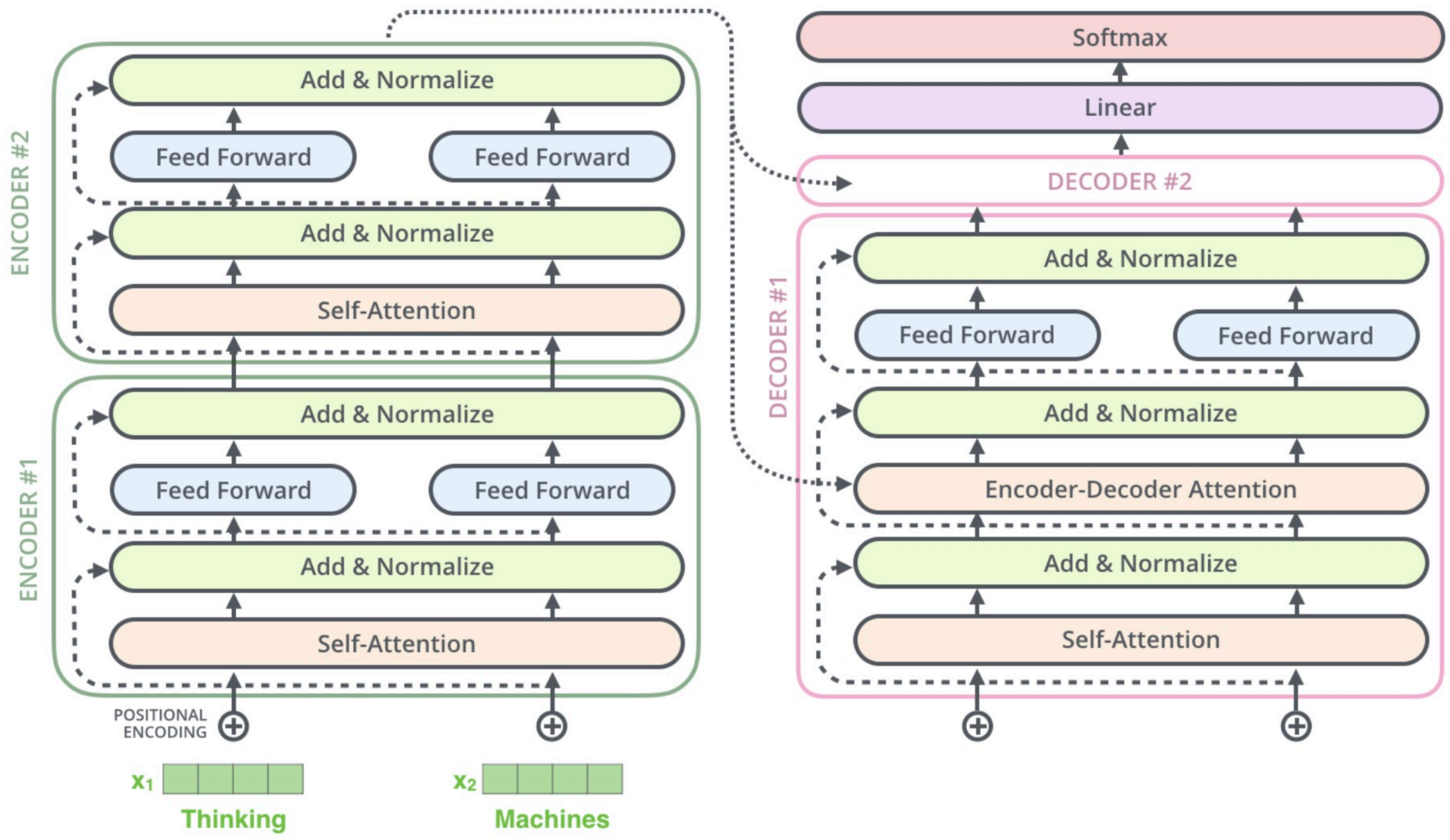
Transformer

Residual



Transformer

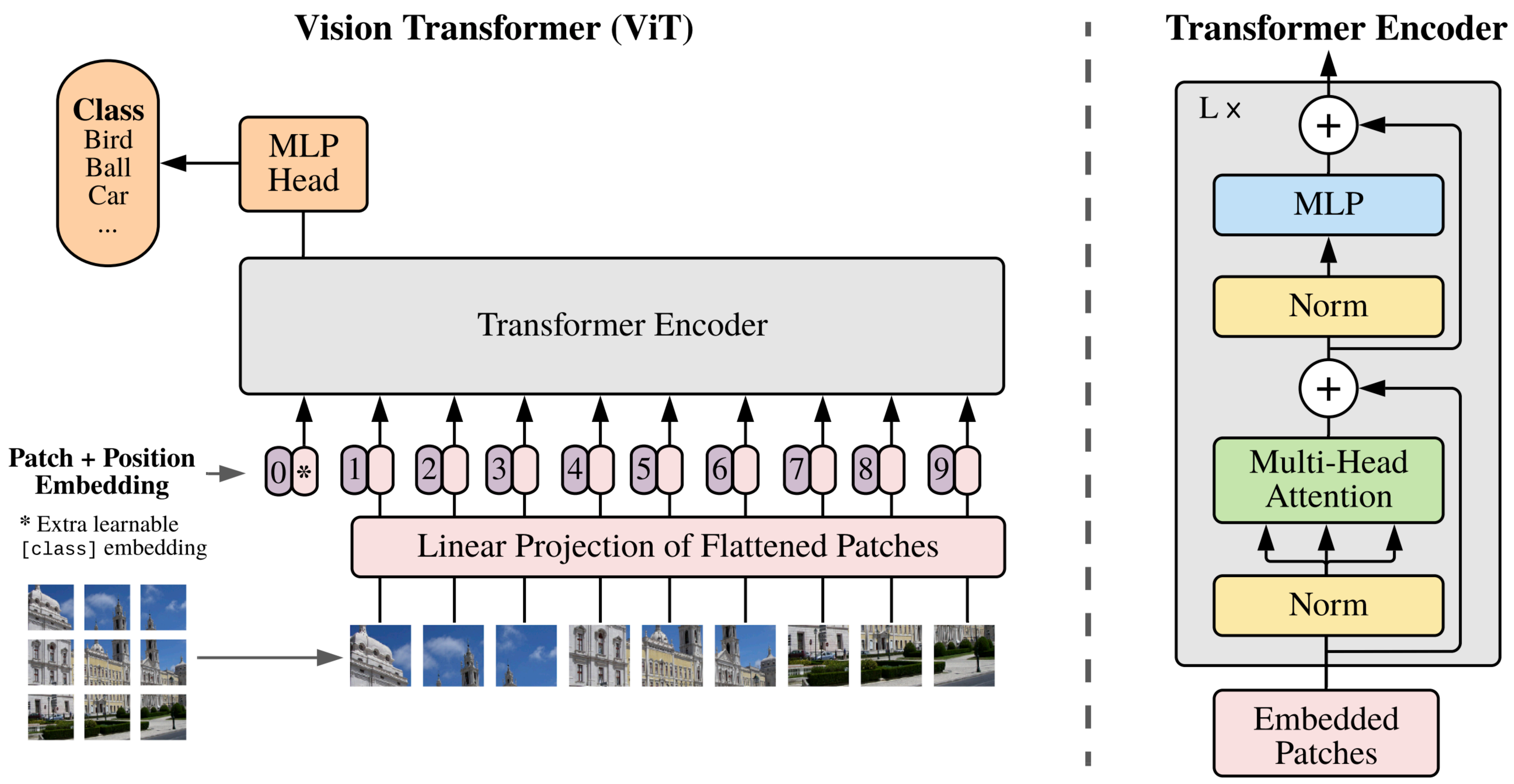
Whole framework



Vision Transformer (ViT)

Vision Transformer (ViT)

- Partition input image into $K \times K$ patches
- A linear projection to transform each patch to feature (no convolution)
- Pass tokens into Transformer



Vision Transformer (ViT)

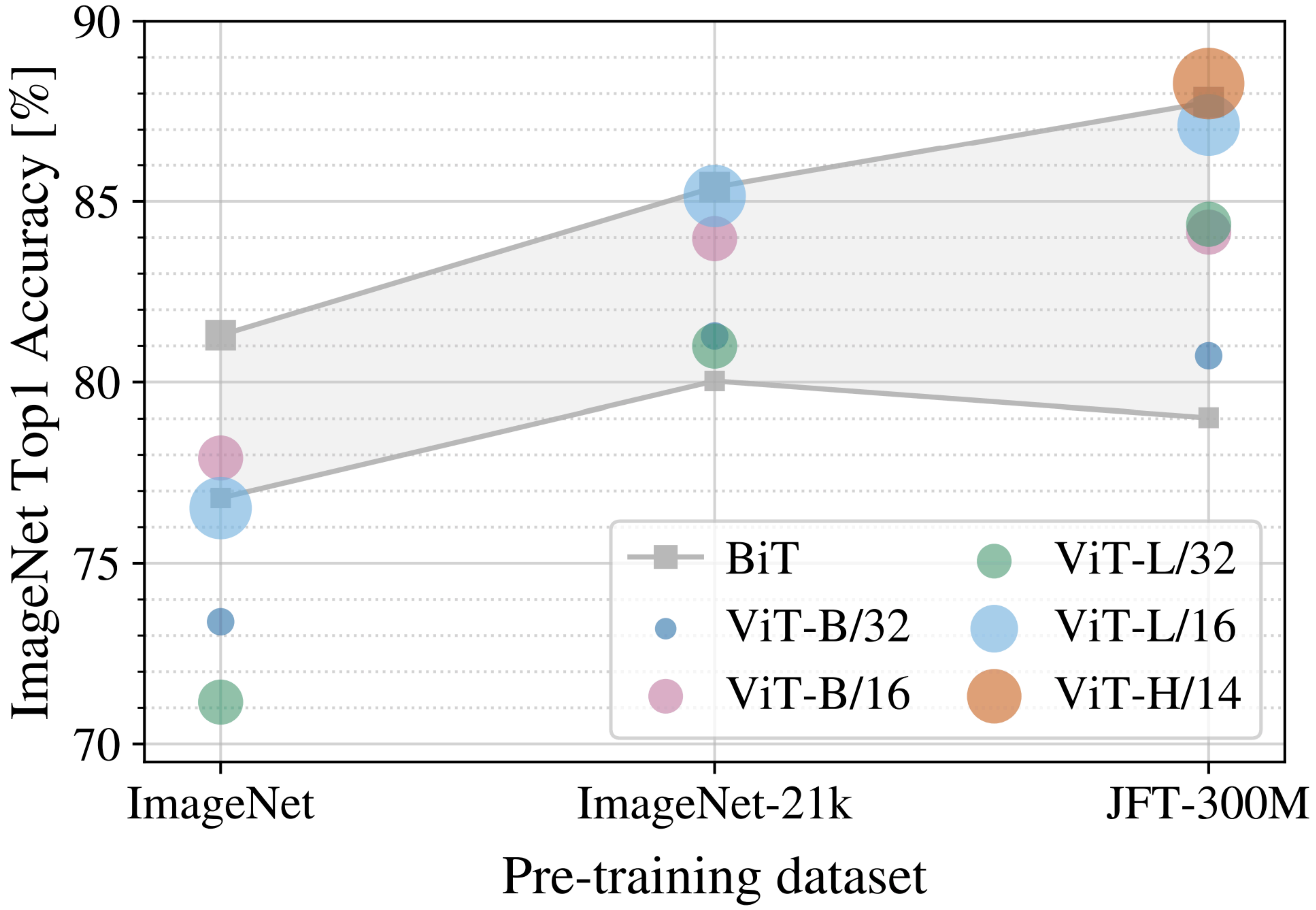
Vision Transformer (ViT)

- Patches are non-overlapping in the original ViT
- $N \times N$ image $\Rightarrow (N/K)^2$ tokens
- Smaller patch size $\Rightarrow$ more input tokens
    - Higher computation (memory) cost, (usually) higher accuracy
- Use 1D (learnable) positional embedding
- Inference with higher resolution:
    - Keep the same patch size, which leads to longer sequence
    - Interpolation for positional embedding

Vision Transformer (ViT)

ViT Performance

- ViT outperforms CNN with large pretraining



Vision Transformer (ViT)

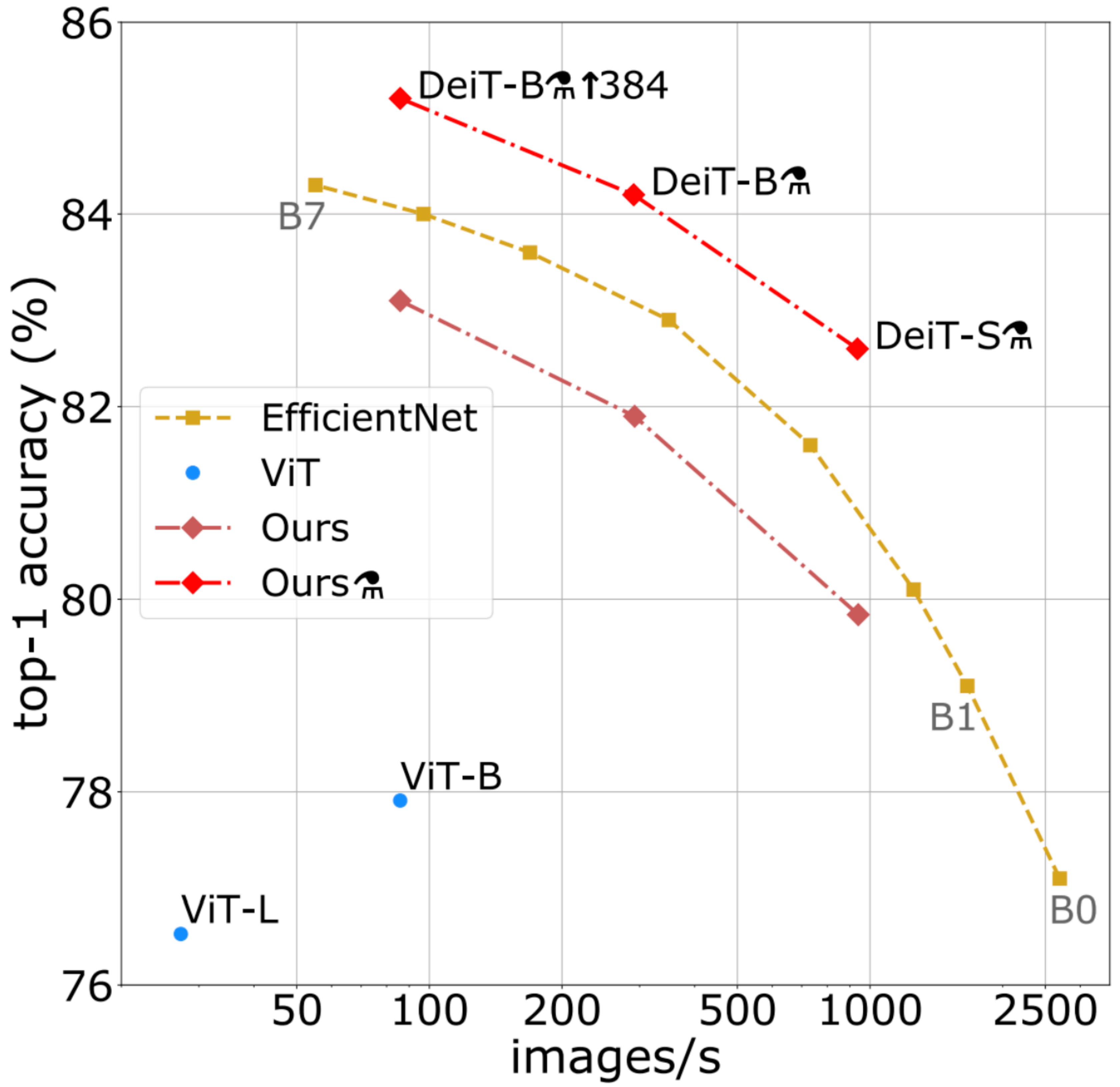
ViT Performance

- Attention maps of ViT (to input)

Vision Transformer (ViT)

ViT v.s. ResNet

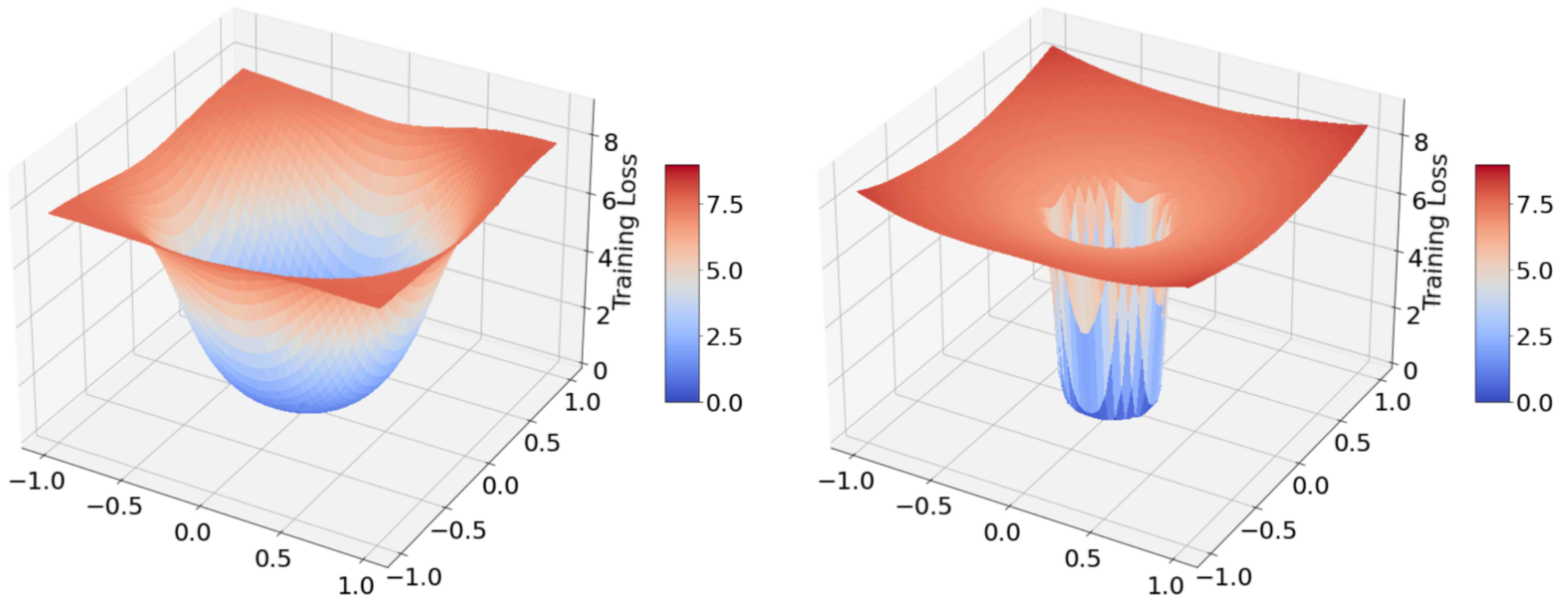
- Can ViT outperform ResNet on ImageNet without pretraining?
- Deit (Touvron et al., 2021):
  - Use very strong data augmentation
  - Use a ResNet teacher and distill to ViT



Vision Transformer (ViT)

ViT v.s. ResNet

- ViT tends to converge to sharper regions than ResNet



(a) ResNet

Leading eigenvalue of Hessian: 179.8

(b) ViT

Leading eigenvalue of Hessian: 738.8

Vision Transformer (ViT)

``Sharpness'' is related to generalization

- Testing can be viewed as a slightly perturbed training distribution
- Sharp minimum $\Rightarrow$ performance degrades significantly from training to testing

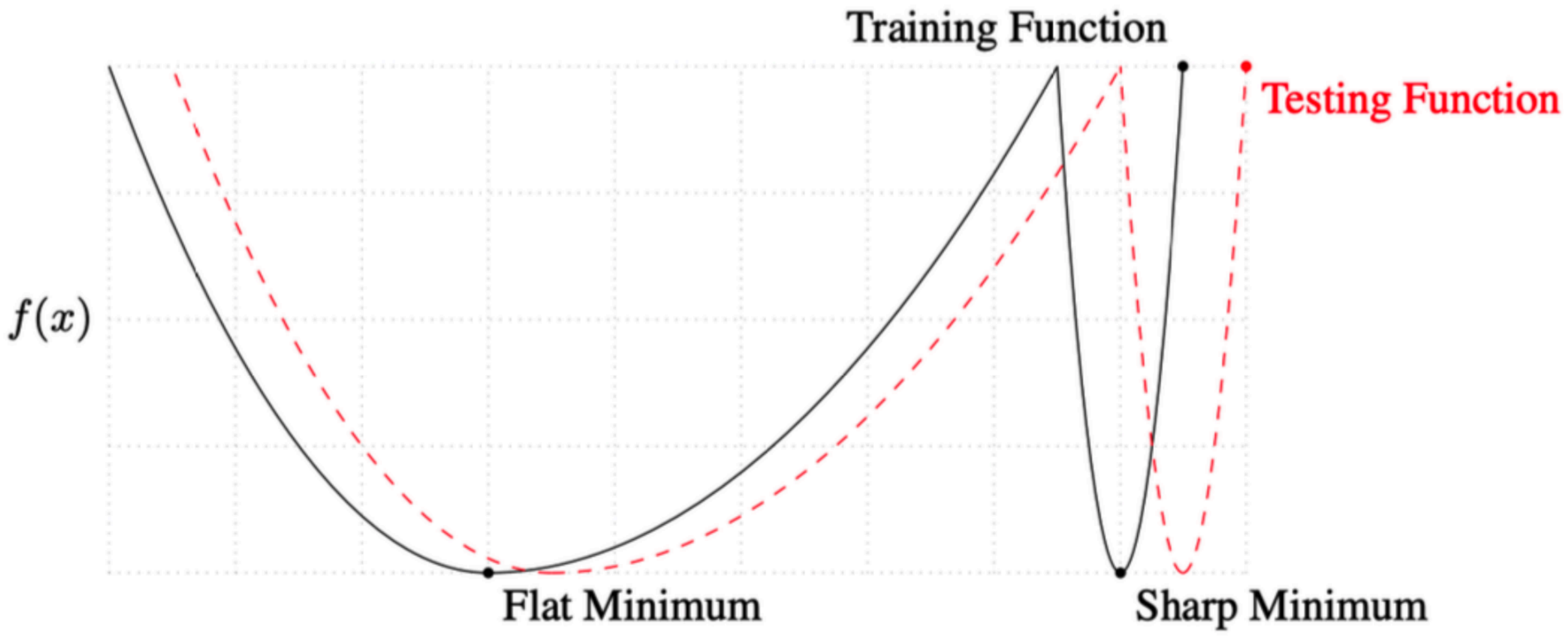


Figure from (Keskar et al., 2017)

Vision Transformer (ViT)

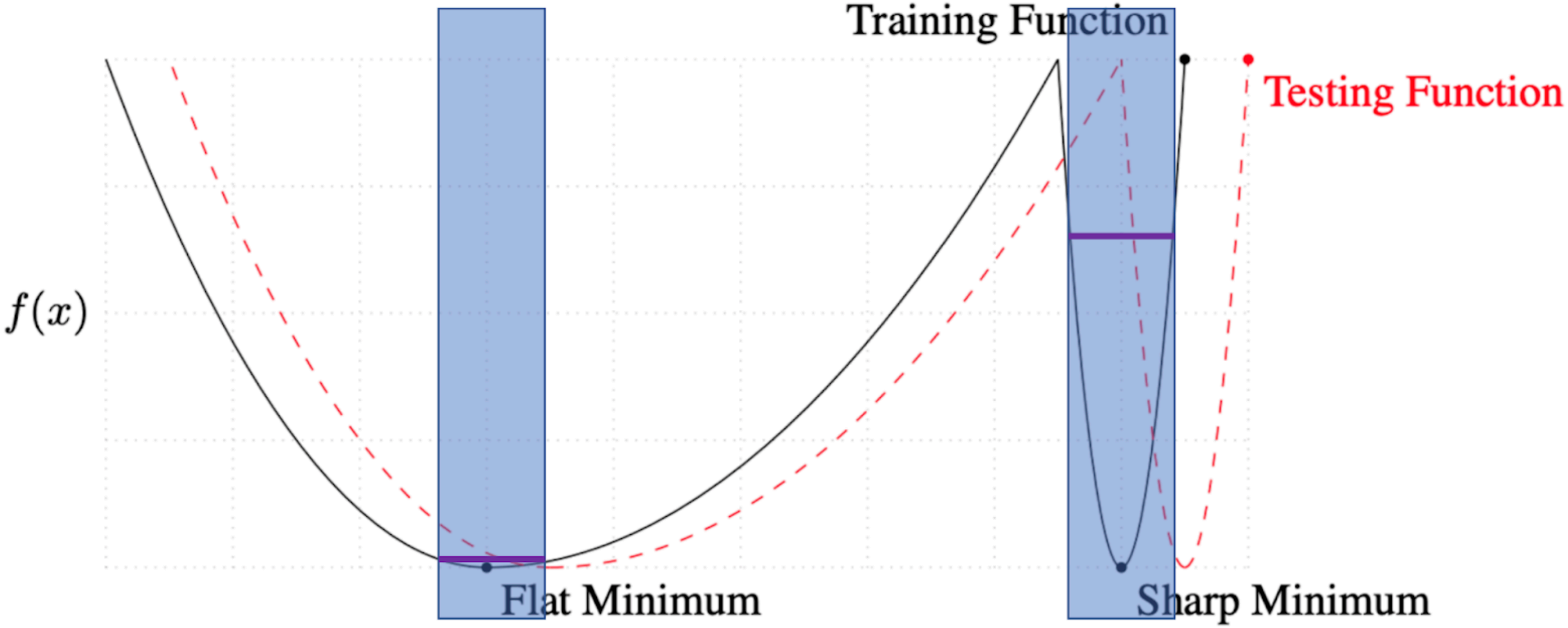
Sharpness Aware Minimization (SAM)

- Optimize the worst-case loss within a small neighborhood
  - $\min_{w} \max_{\|\delta\|_2 \le \epsilon} L(w + \delta)$
  - ϵ is a small constant (hyper-parameter)
- Use 1-step gradient ascent to approximate inner max:
  - $\hat{\delta} = \arg \max_{\|\delta\|_2 \le \epsilon} L(w) + \nabla L(w)^T \delta = \epsilon \frac{\nabla L(w)}{\|\nabla L(w)\|}$
- Conduct the following update for each iteration:
  - $w \leftarrow w - \alpha \nabla L(w + \hat{\delta})$

Vision Transformer (ViT)

Sharpness Aware Minimization (SAM)

- SAM is a natural way to penalize sharpness region (but requires some computational overhead)



Unsupervised pertaining for NLP

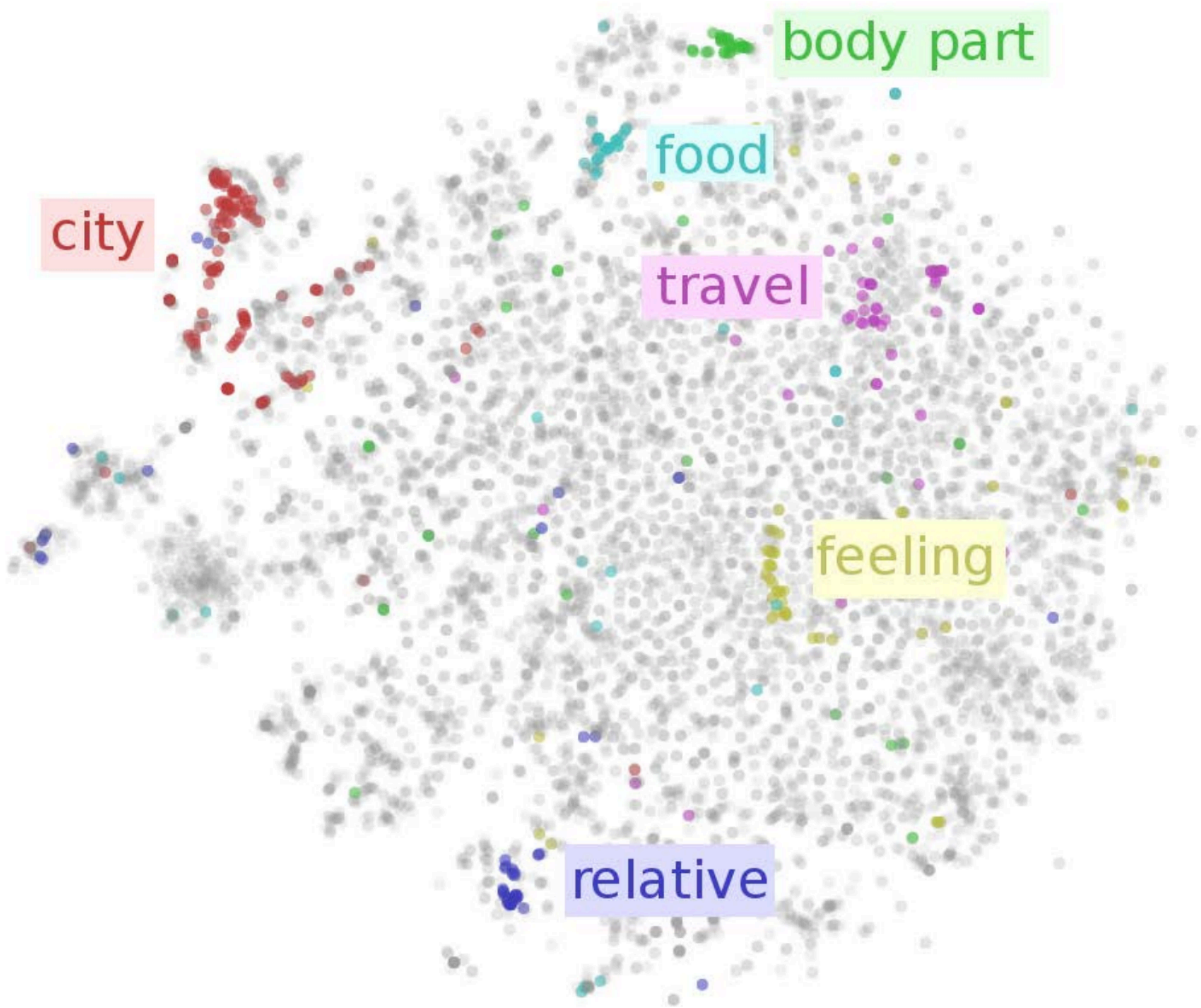
Motivation

- Many unlabeled NLP data but very few labeled data
- Can we use large amount of unlabeled data to obtain meaningful representations of words/sentences?

Unsupervised pertaining for NLP

Learning word embeddings

- Use large (unlabeled) corpus to learn a useful word representation
  - Learn a vector for each word based on the corpus
  - Hopefully the vector represents some semantic meaning
  - Can be used for many tasks
    - Replace the word embedding matrix for DNN models for classification/translation
  - Two different perspectives but led to similar results:
    - Glove (Pennington et al., 2014)
    - Word2vec (Mikolov et al., 2013)



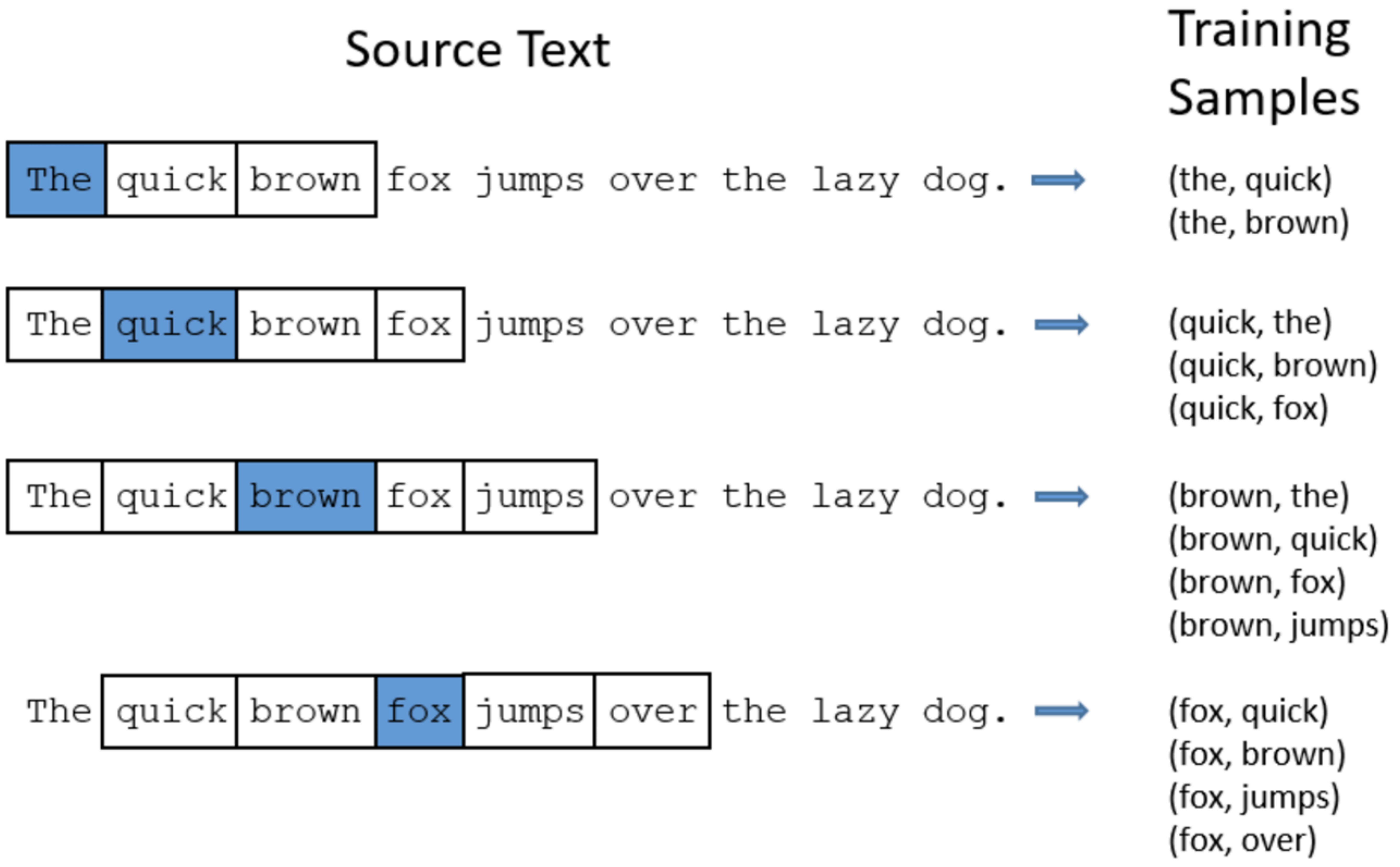
Unsupervised pertaining for NLP

Context information

- Given a large text corpus, how to learn low-dimensional features to represent a word?
- For each word w_i, define the ``contexts'' of the word as the words surrounding it in an L-sized window:
- $$w_{i-L-2}, w_{i-L-1}, \underbrace{w_{i-L}, \cdots, w_{i-1}}_{\text{contexts of } w_i}, w_i, \underbrace{w_{i+1}, \cdots, w_{i+L}}_{\text{contexts of } w_i}, w_{i+L+1}, \cdots$$
- Get a collection of (word, context) pairs, denoted by D.

Unsupervised pertaining for NLP

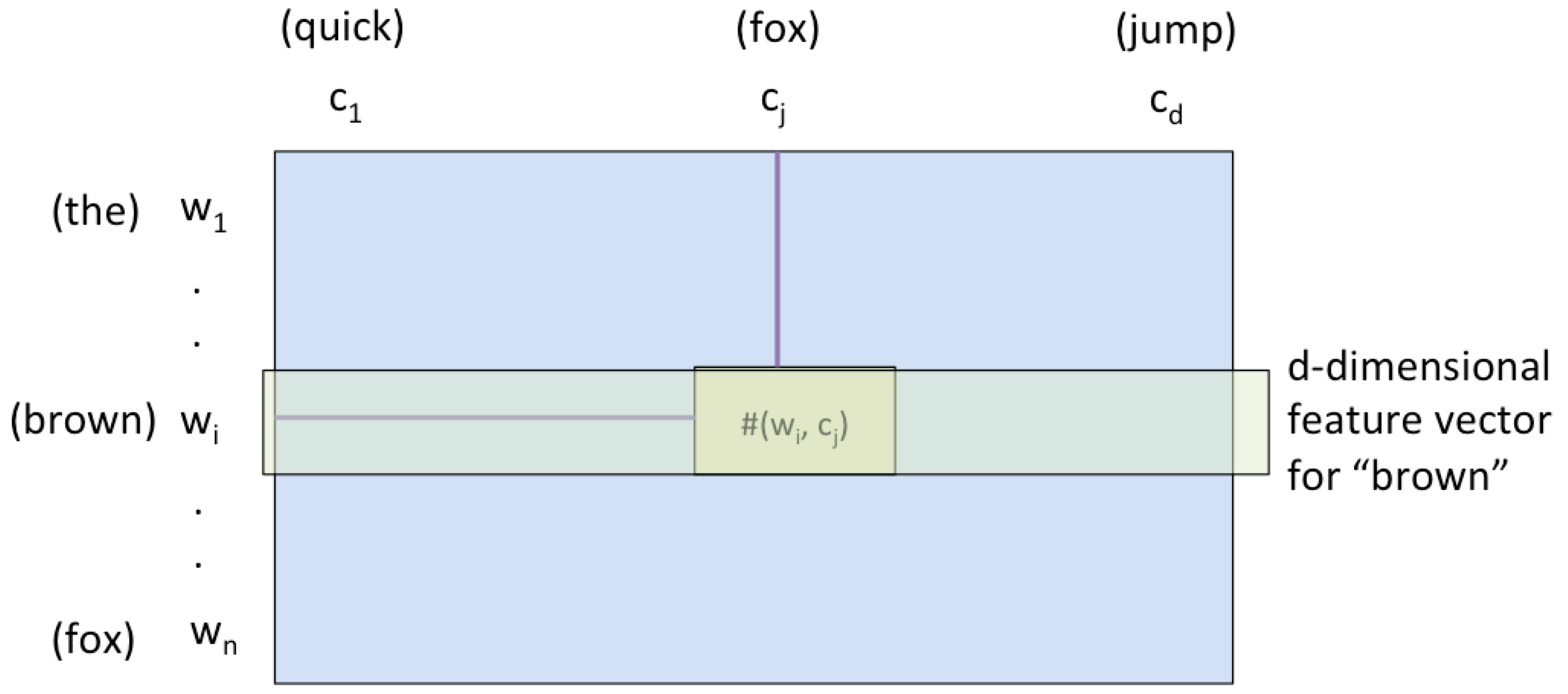
Examples



Unsupervised pertaining for NLP

Use bag-of-word model

- Idea 1: Use the bag-of-word model to ``describe'' each word
- Assume we have context words $c_1, \cdots, c_d$ in the corpus, compute
  - $\#(w, c_i) :=$ number of times the pair(w, c_i) appears in D
- For each word w, form a d-dimensional (sparse) vector to describe w
  - $\#(w, c_1), \cdots, \#(w, c_d),$



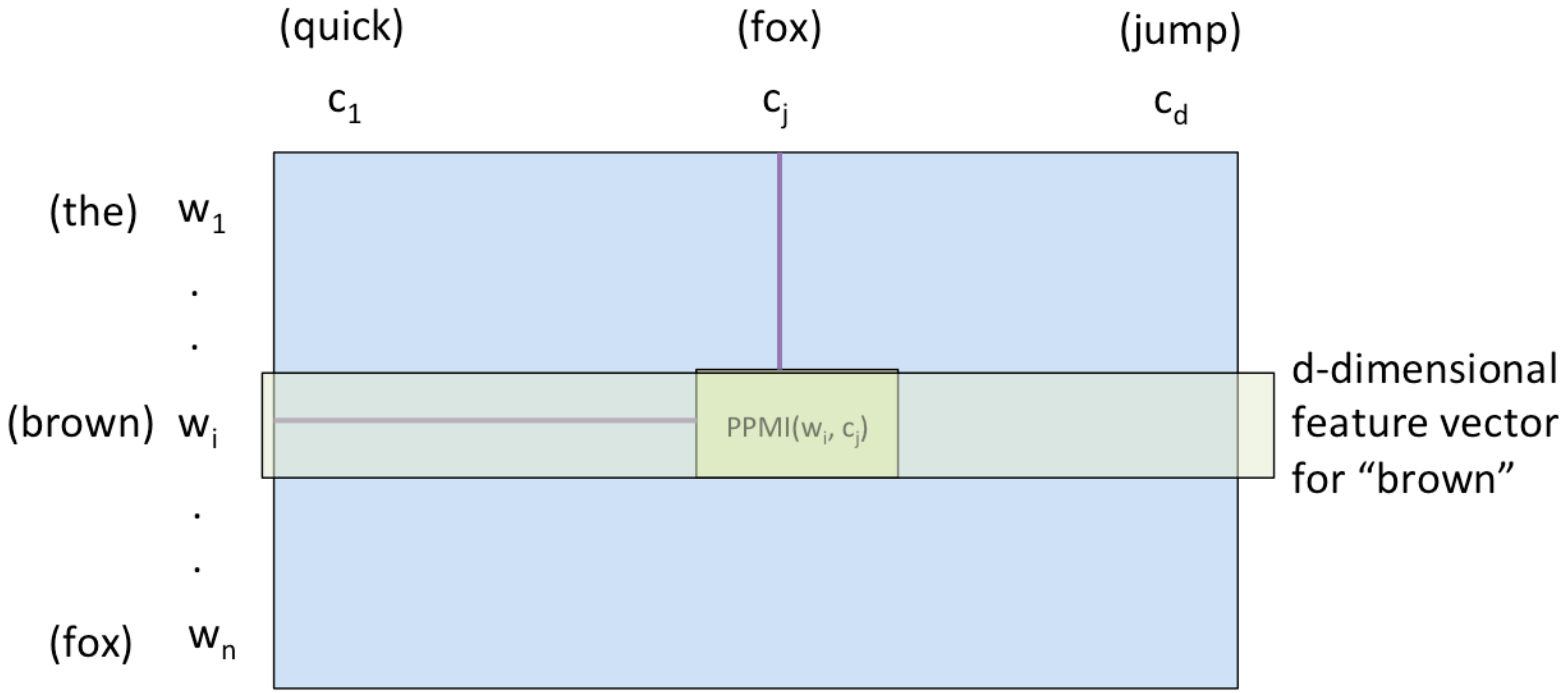
Unsupervised pertaining for NLP

PMI/PPMI Representation

- Similar to TF-IDF: Need to consider the frequency of each word and each context
- Instead of using co-ocurrent count $\#(w, c)$, we can define pointwise mutual information:
  - PMI$(w, c) = \log(\frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}(c)}) = \log \frac{\#(w, c)|D|}{\#(w)\#(c)}$,
  - $\#(w) = \sum_c \#(w, c)$: number of times word w occurred in D
  - $\#(c) = \sum_w \#(w, c)$: number of times context c occurred
  - $|D|$: number of pairs in D
- Positive PMI (PPMI) usually achieves better performance:
  - PPMI$(w, c) = \max($PMI$(w, c), 0)$
- M^{PPMI}: a n by d word feature matrix, each row is a word and each column is a context

Unsupervised pertaining for NLP

PPMI Matrix



Unsupervised pertaining for NLP

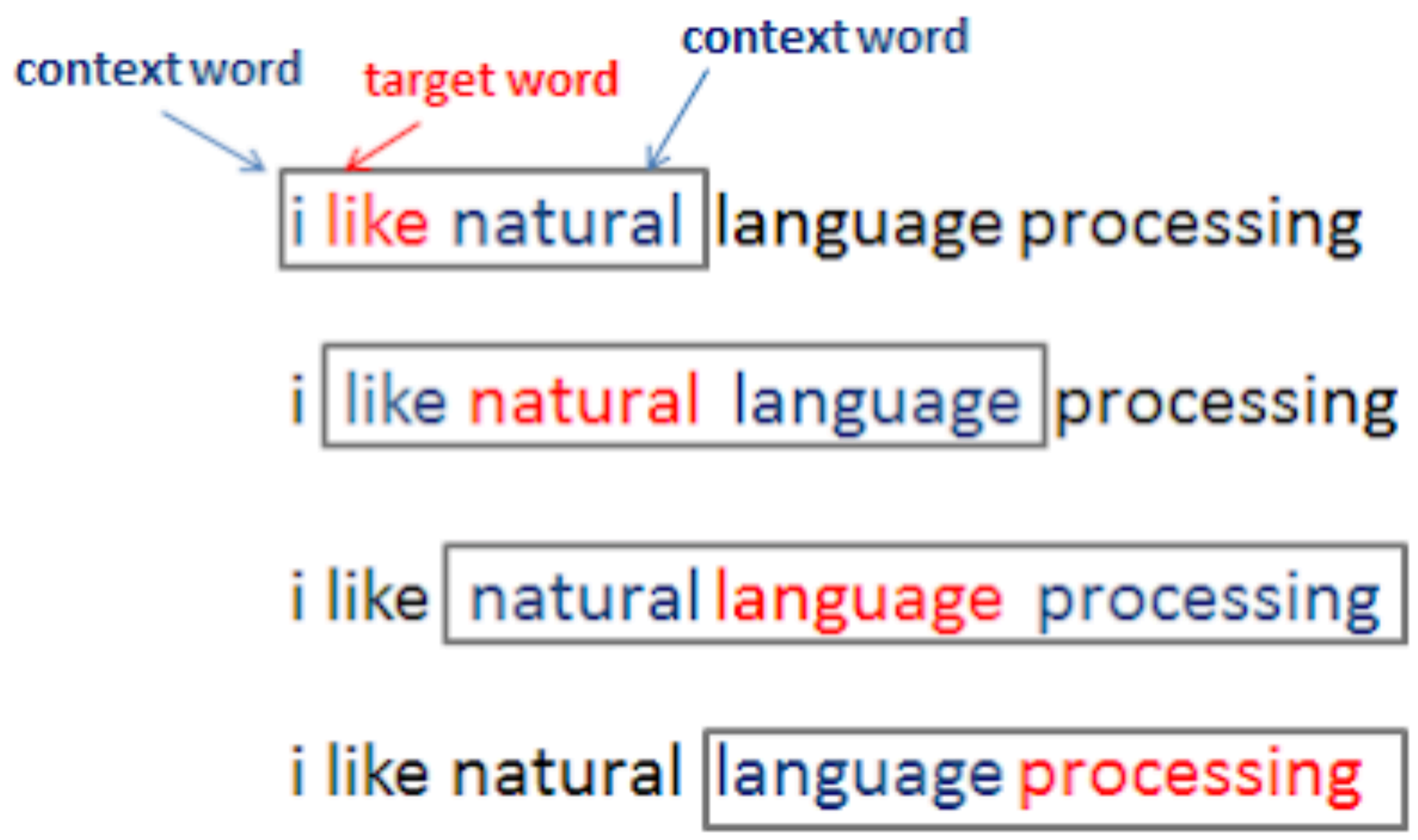
Generalized Low-rank Embedding

- SVD basis will minimize
  - $\min_{W,V} \|M^{\text{PPMI}} - WV^T\|_F^2$
- Glove (Pennington et al., 2014)
  - Negative sampling (less weights to 0s in M^{PPMI})
  - Adding bias term:
    - $M^{\text{PPMI}} \approx WV^T + b_w e^T + e b_c^T$
- Use W or V as the word embedding matrix

Unsupervised pertaining for NLP

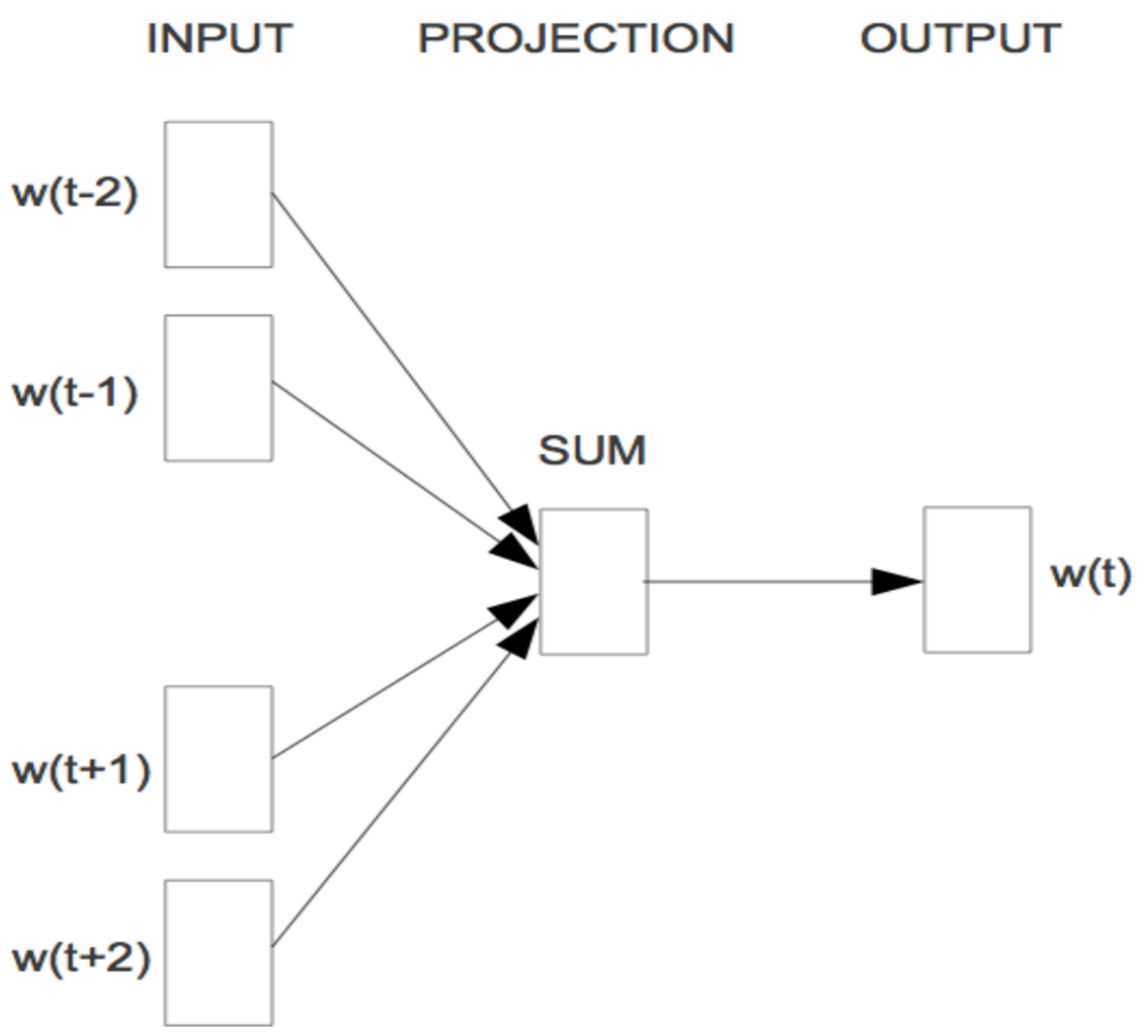
Word2vec (Mikolov et al., 2013)

- A neural network model for learning word embeddings
- Main idea:
  - Predict the target words based on the neighbors (CBOW)
  - Predict neighbors given the target words (Skip-gram)



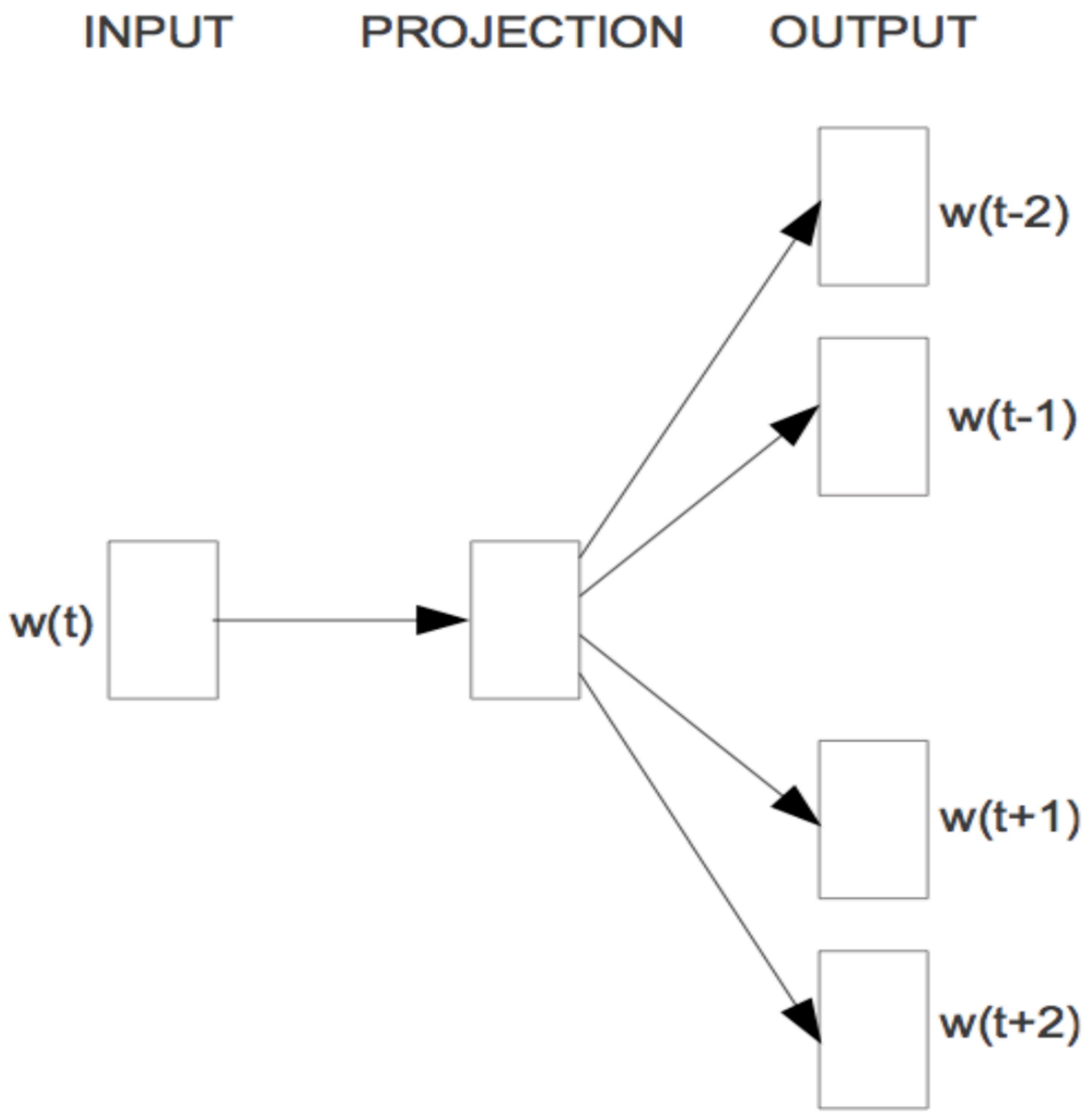
Unsupervised pertaining for NLP
CBOW (Continuous Bag-of-Word model)

- Predict the target words based on the neighbors



Unsupervised pertaining for NLP
Skip-gram

- Predict neighbors using target word



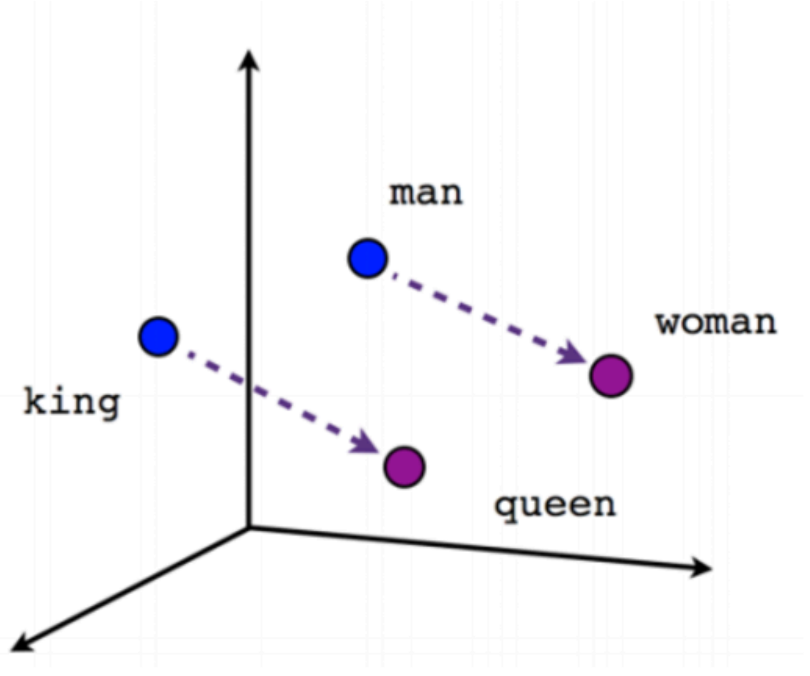
Unsupervised pertaining for NLP

More on skip-gram

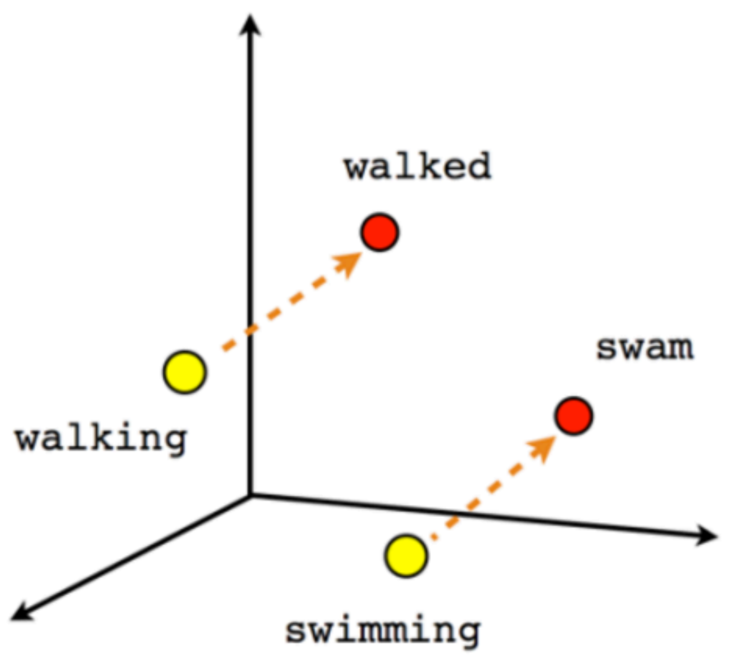
- Learn the probability $P(w_{t+j} | w_t)$: the probability to see w_{t+j} in target word w_t's neighborhood
- Every word has two embeddings:
  - v_i serves as the role of target
  - u_i serves as the role of context
- Model probability as softmax:
  - $P(o | c) = \dfrac{e^{u_o^T v_c}}{\sum_{w=1}^{W} e^{u_w^T v_c}}$

Unsupervised pertaining for NLP
Results

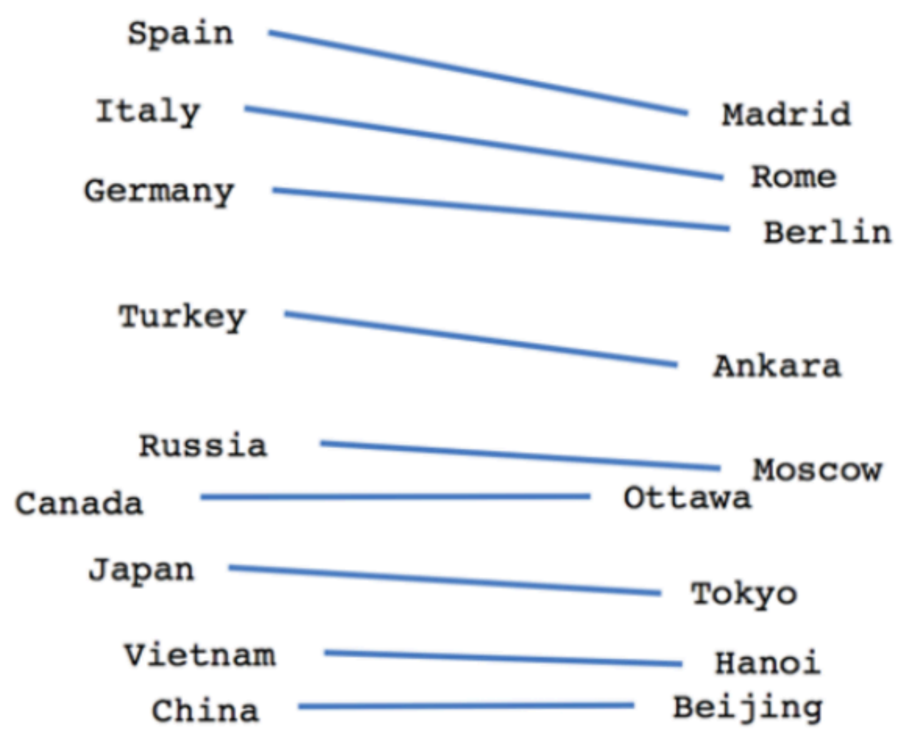
- The low-dimensional embeddings are (often) meaningful:



Male-Female



Verb tense



Country-Capital

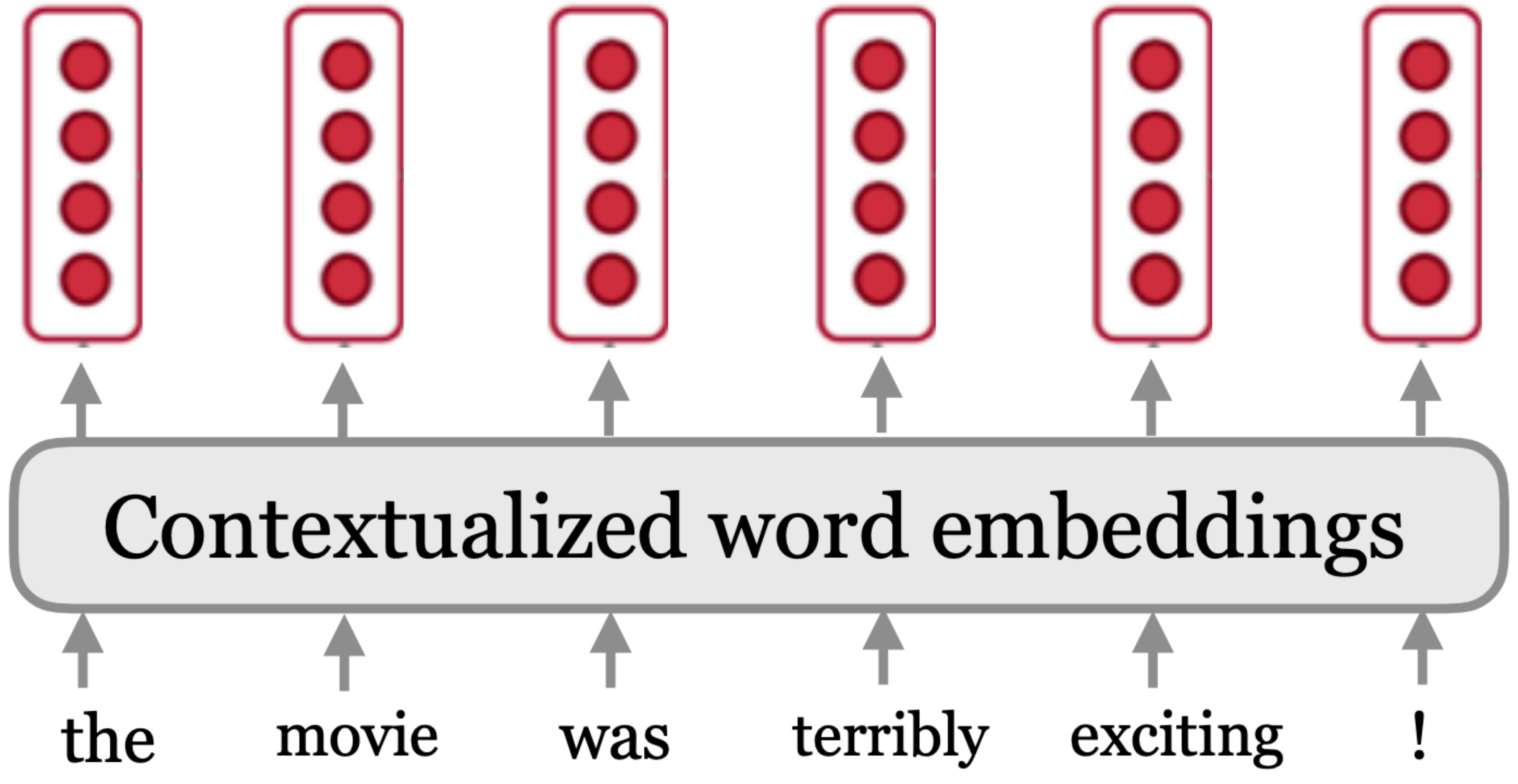
Contextual embedding

Contextual world representation

- The semantic meaning of a word should depend on its context

```
   [0.9, -0.2, 1.6, …]                  [-1.9, -0.4, 0.1, …]
            ↑                                      ↑
  open a bank account               on the river bank
```

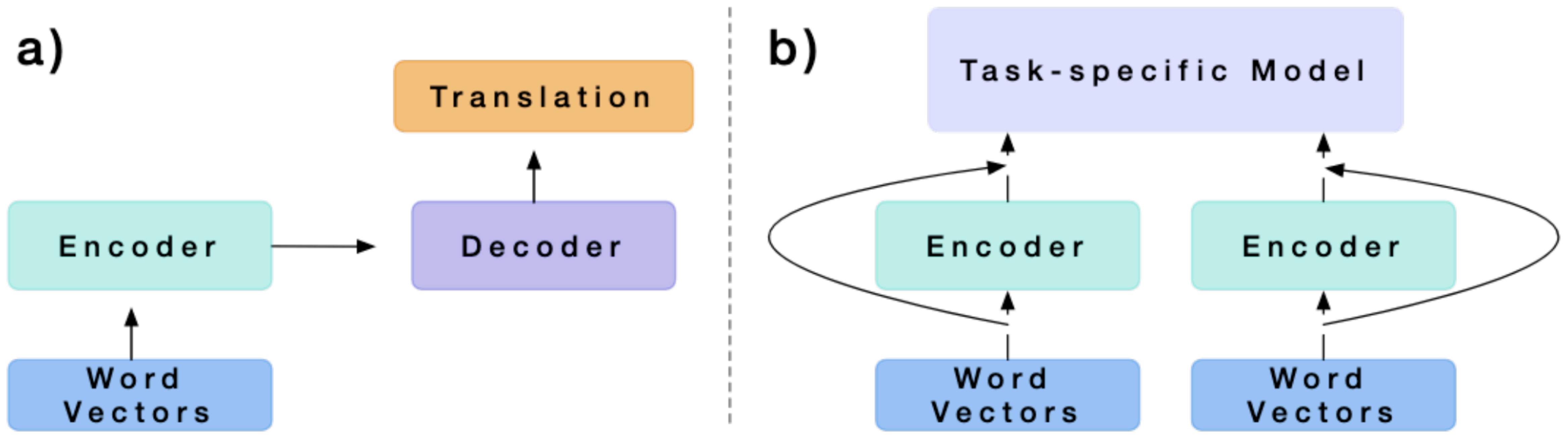
- Solution: Train a model to extract contextual representations on text corpus



Contextual embedding

CoVe (McCann et al., 2017)

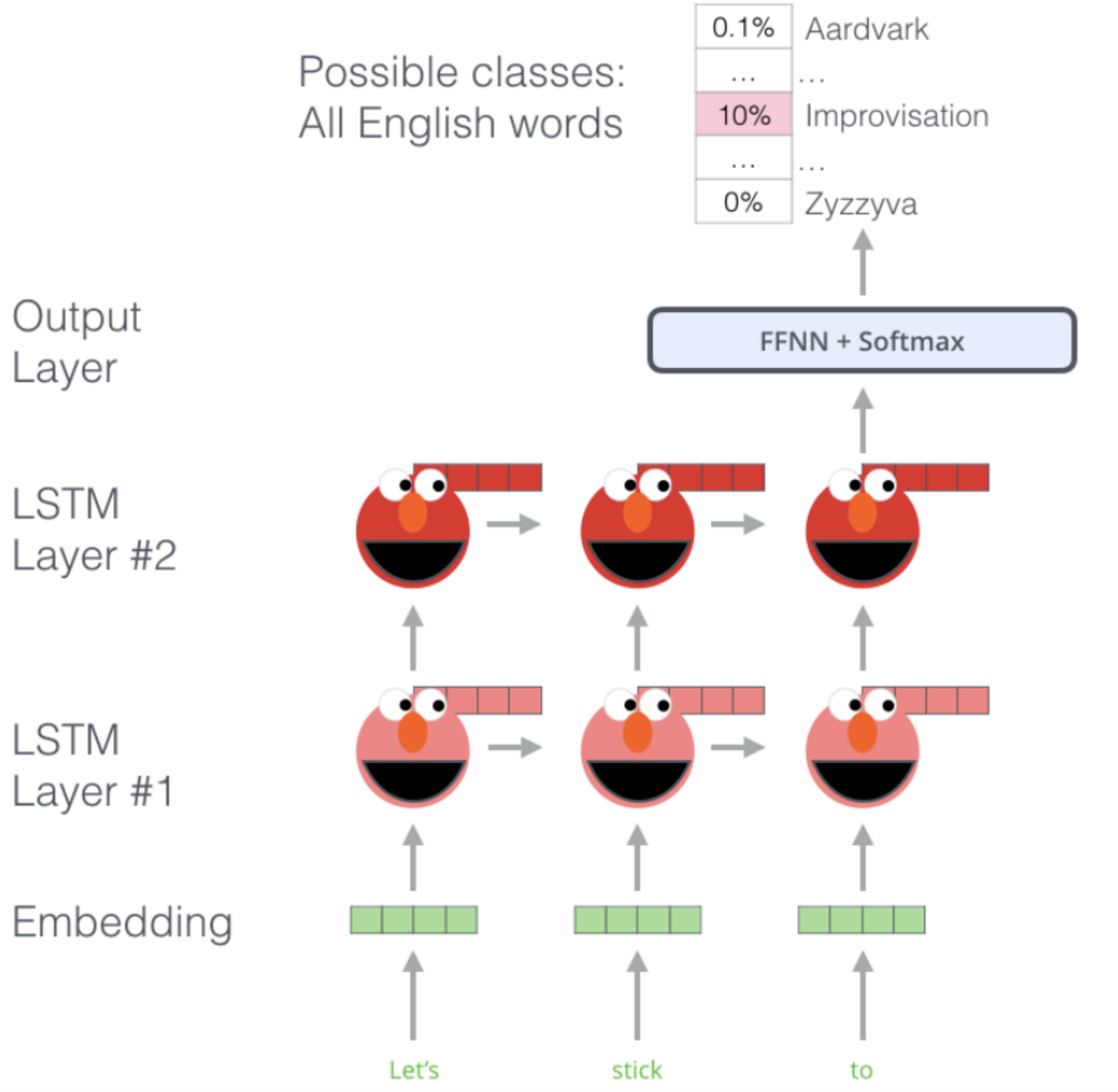
- Key idea: Train a standard neural machine translation model
- Take the encoder directly as contextualized word embeddings
- Problems:
    - Translation requires paired (labeled) data
    - The embeddings are tailored to particular translation corpuses



Contextual embedding

Language model pretraining task

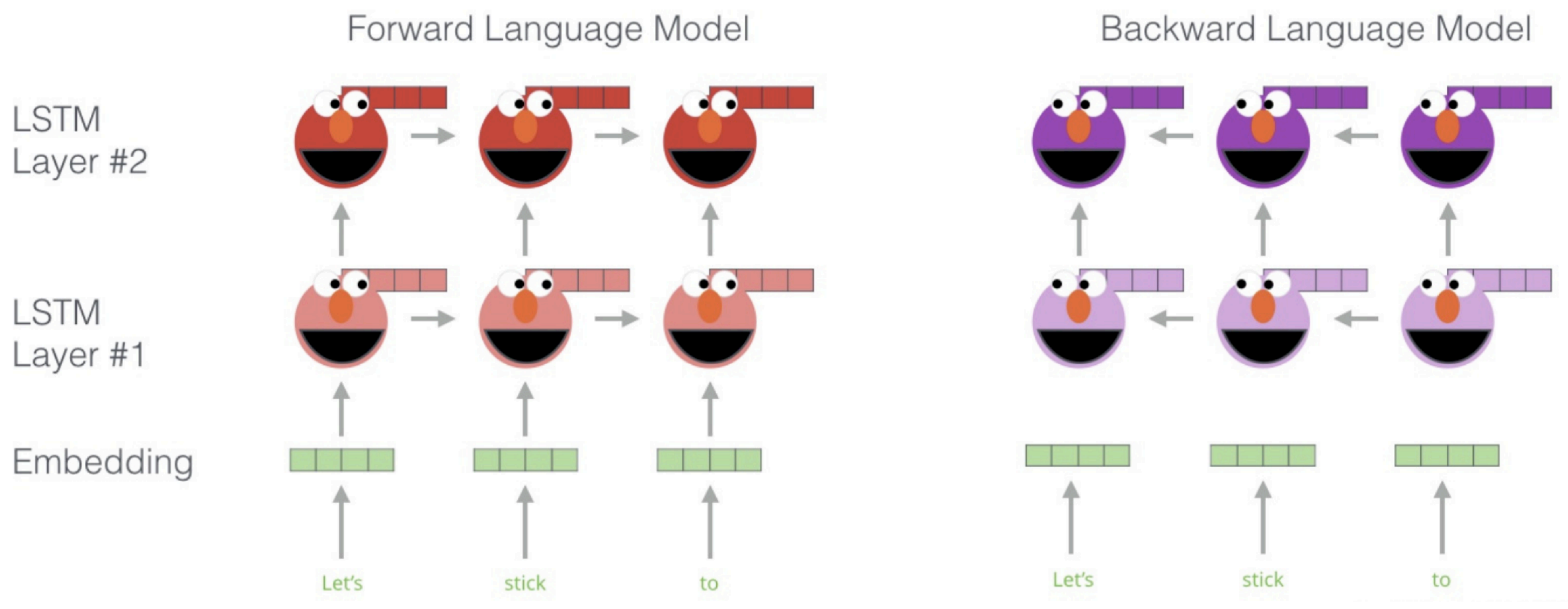
- Predict the next word given the prefix
- Can be defined on any unlabeled document



Contextual embedding

ELMo (Peter et al., 2018)

- Key ideas:
  - Train a foward and backward LSTM language model on large corpus
  - Use the hidden states for each token to compute a vector representation of each word
  - Replace the word embedding by Elmo's embedding (with fixed Elmo's LSTM weights)



Contextual embedding

ELMo results

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Contextual embedding

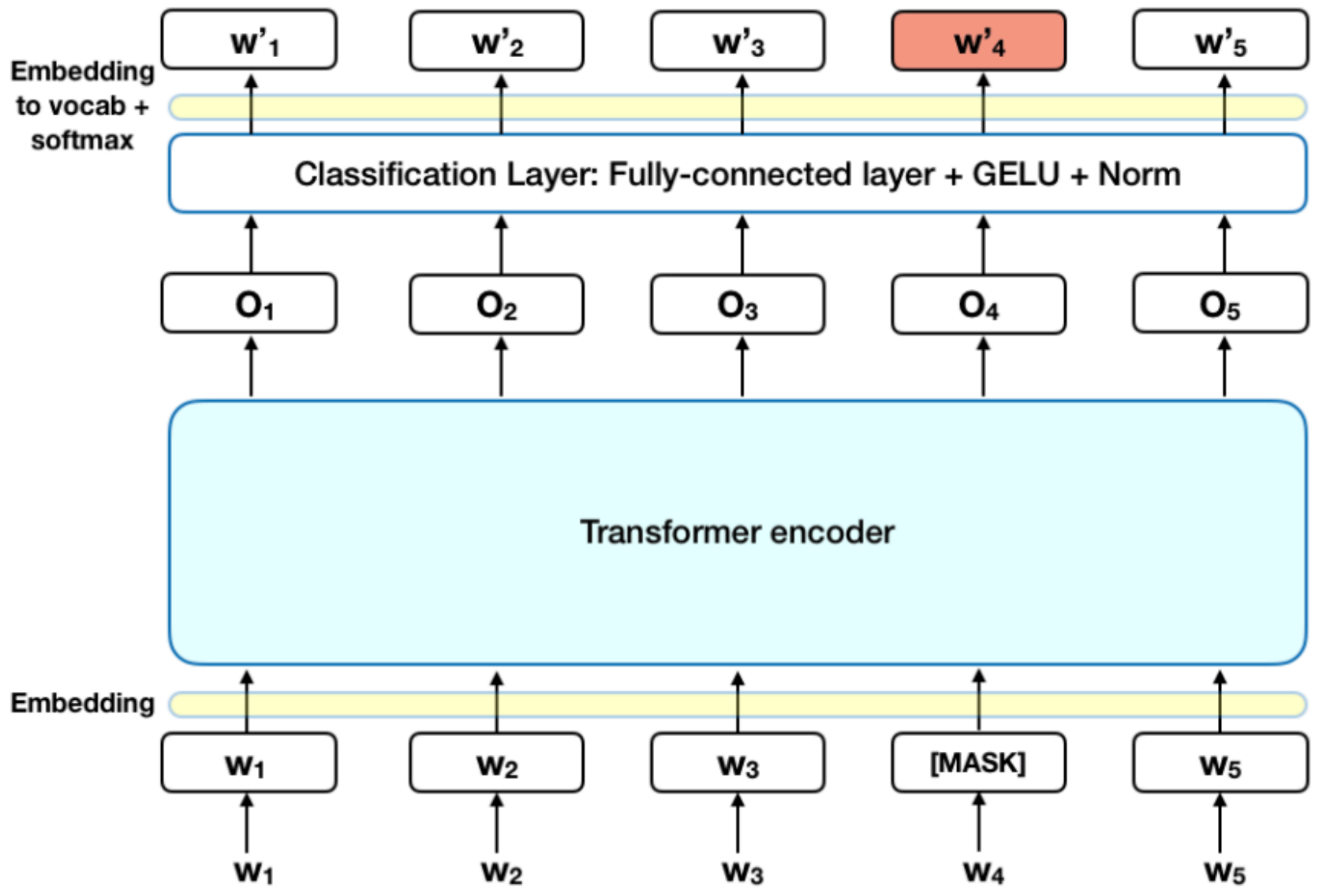
BERT

- Key idea: replace LSTM by Transformer
- Define the generated pretraining task by masked language model
- Two pretraining tasks
- Finetune both BERT weights and task-dependent model weights for each task

Contextual embedding

BERT pretraining loss

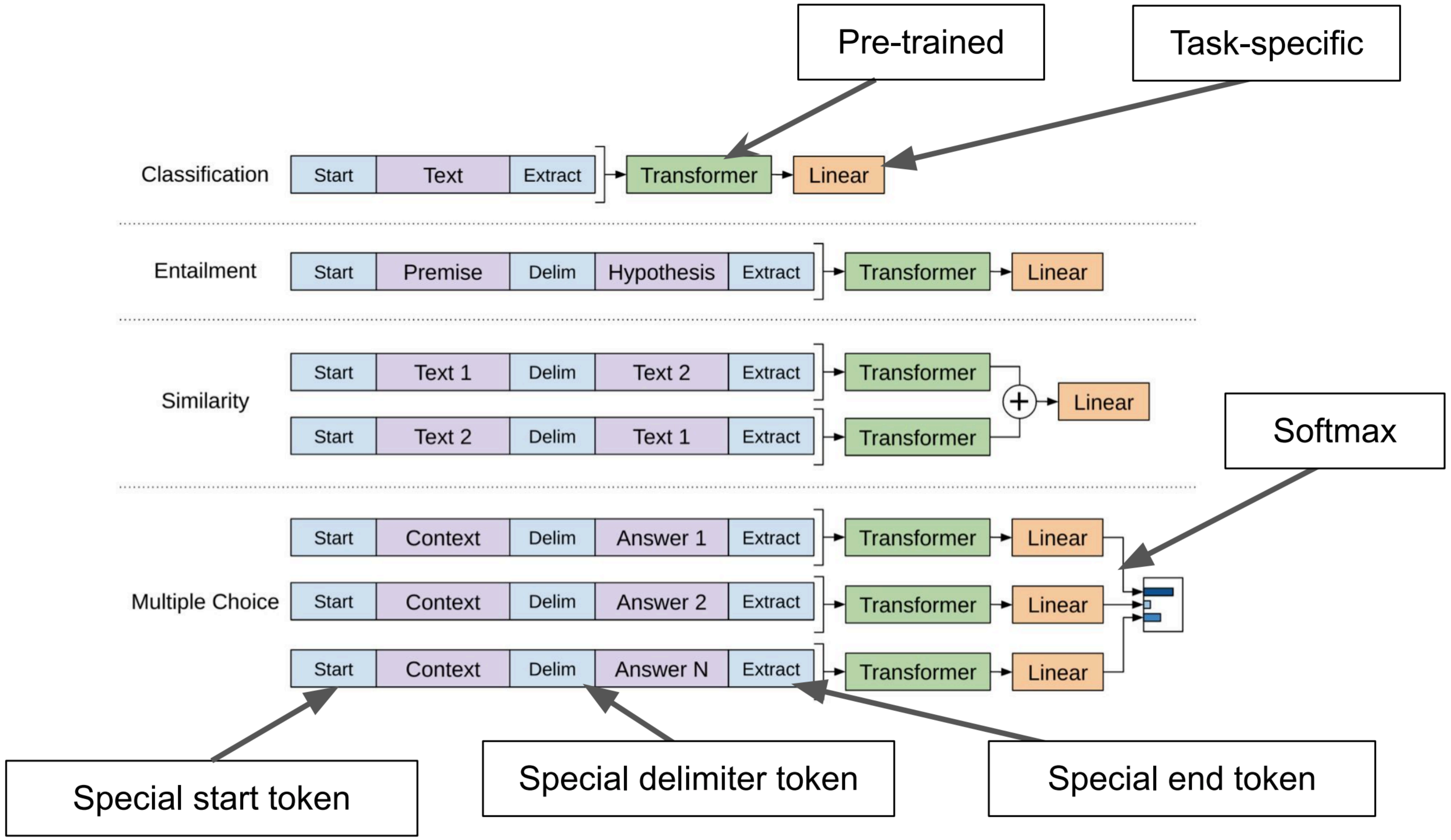
- Masked language model: predicting each word by the rest of sentence
- Next sentence prediction: the model receives pairs of sentences as input and learns to predict if the second sentence is the subsequent sentence in the original document.



Contextual embedding

BERT finetuning

- Keep the pretrained Transformers
- Replace or append a layer for the final task
- Train the whole model based on the task-dependent loss



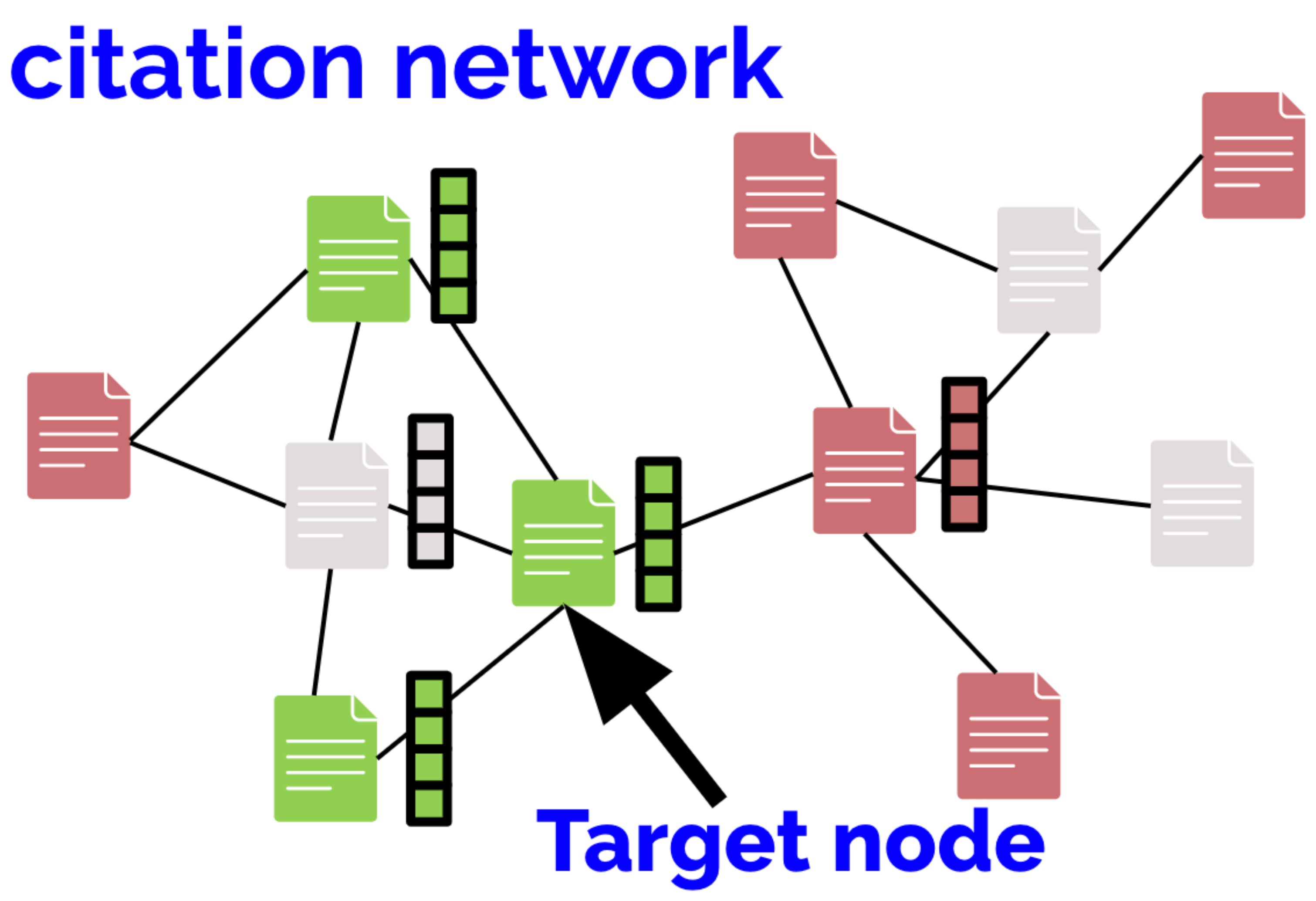
Contextual embedding
BERT results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Graph Convolutional Neural Network

Node classification problem

- Given a graph of N nodes, with adjacency matrix $A \in \mathbb{R}^{N \times N}$
- Each node is associated with a D-dimensional feature vector.
- $X \in \mathbb{R}^{N \times D}$: each row corresponds to the feature vector of a node
- Observe labels for a subset of nodes: $Y \in \mathbb{R}^{N \times L}$, only observe a subset of rows, denoted by Y_S
- Goal: Predict labels for unlabeled nodes (transductive setting) or
- test nodes (inductive setting) or test graphs (inductive setting)



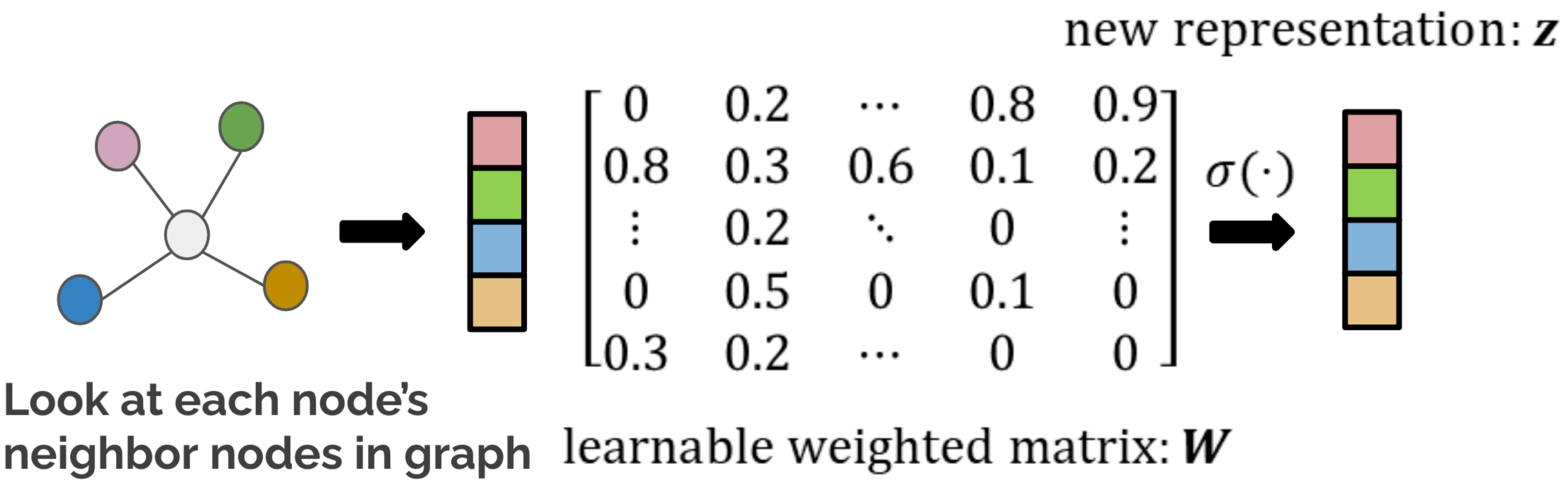
Graph Convolutional Neural Network

Graph Convolution Layer

- GCN: multiple graph convolution layers
- $\hat{A}$: normalized version of A:
  - $\tilde{A} = A + I, \quad \tilde{D}_{uv} = \sum_{v} \tilde{A}_{uv}, \quad P = \tilde{D}^{-1}\hat{A}$
- Graph convolution:
  - Input: features for each node $H^{(l)} \in \mathbb{R}^{n \times D}$
  - Output: features for each node $H^{(l+1)}$ after gathering neighborhood information
  - Convolution: $PH^{(l)}$: Aggregate features from neighbors
  - Convolution + fully-connected layer + nonlinear activation:
    - $H^{(l+1)} = \sigma(PH^{(l)}W^{(l)})$,
    - $W^{(l)}$ is the weights for the linear layer
    - $\sigma(\cdot)$: usually ReLU function

Graph Convolutional Neural Network

Graph convolutional network



Graph Convolutional Neural Network

Graph convolutional network

- Initial features $H^{(0)} := X$
- For layer $l = 0, \ldots, L$
  - $Z^{(l+1)} = PH^{(l)}W^{(l)}, \quad H^{(l+1)} = \sigma(Z^{(l+1)}),$
- Use final layer feature $H^{(L)} \in \mathbb{R}^{N \times K}$ for classification:
  - $\text{Loss} = \dfrac{1}{|S|} \sum_{s \in S} \text{loss}(y_s, Z_s^{(L)})$
  - Each row of $Z_s^{(L)}$ corresponds to the output score for each label
  - Cross-entropy loss for classification

Graph Convolutional Neural Network

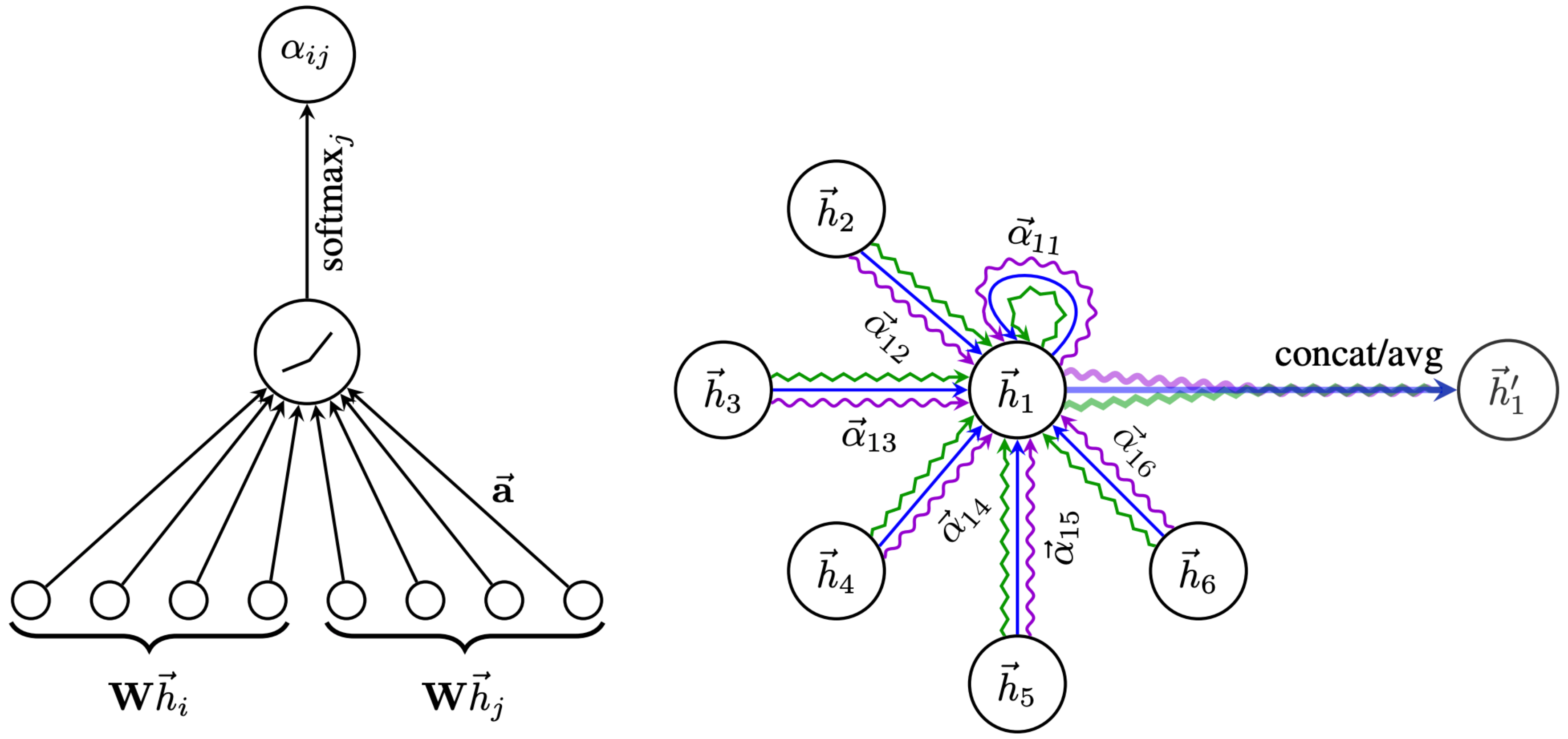
Graph convolutional network

- Model parameters: $W^{(1)}, \cdots, W^{(L)}$
- Can be used to
  - Predict unlabeled nodes in the training set
  - Predict testing nodes (not in the training set)
  - Predict labels for a new graph
- Also, features extracted by GCN $H^{(L)}$ is usually very useful for other tasks

Graph Convolutional Neural Network

Graph Attention Networks

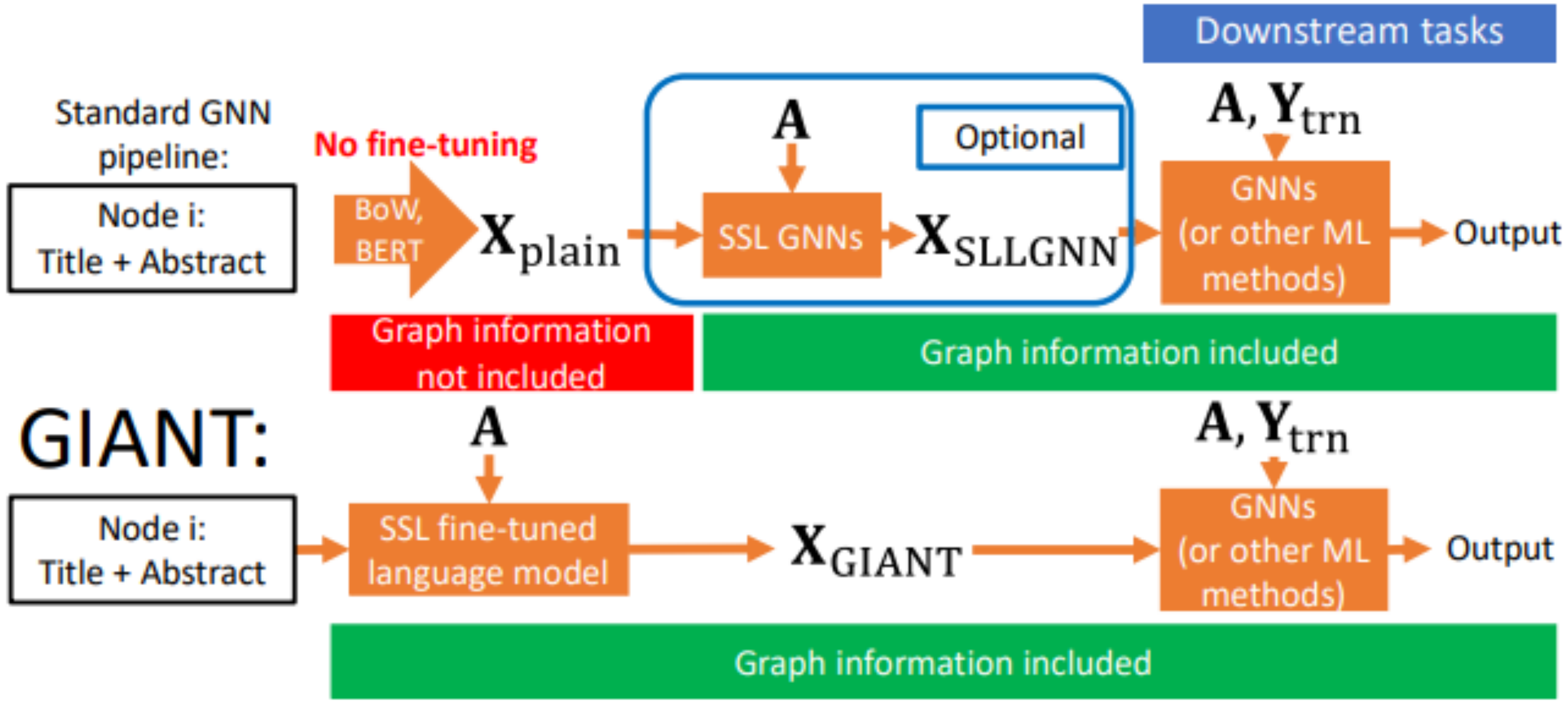
- Each edge may not contribute equally
- Using attention mechanism to automatically assign weights to each edge:
  - $\alpha_{i,j} = \dfrac{\exp(\text{LeakyReLU}(a^T[Wh_i \mid Wh_j]))}{\sum_{k \in N_i} \exp(\text{LeakyReLU}(a^T[Wh_i \mid Wh_k]))}$
  - where h_i, h_j are the features for node i and j at previous layer, W is the GNN weight, a is the additional learnable parameter for attention



Graph Convolutional Neural Network

GNN Pretraining

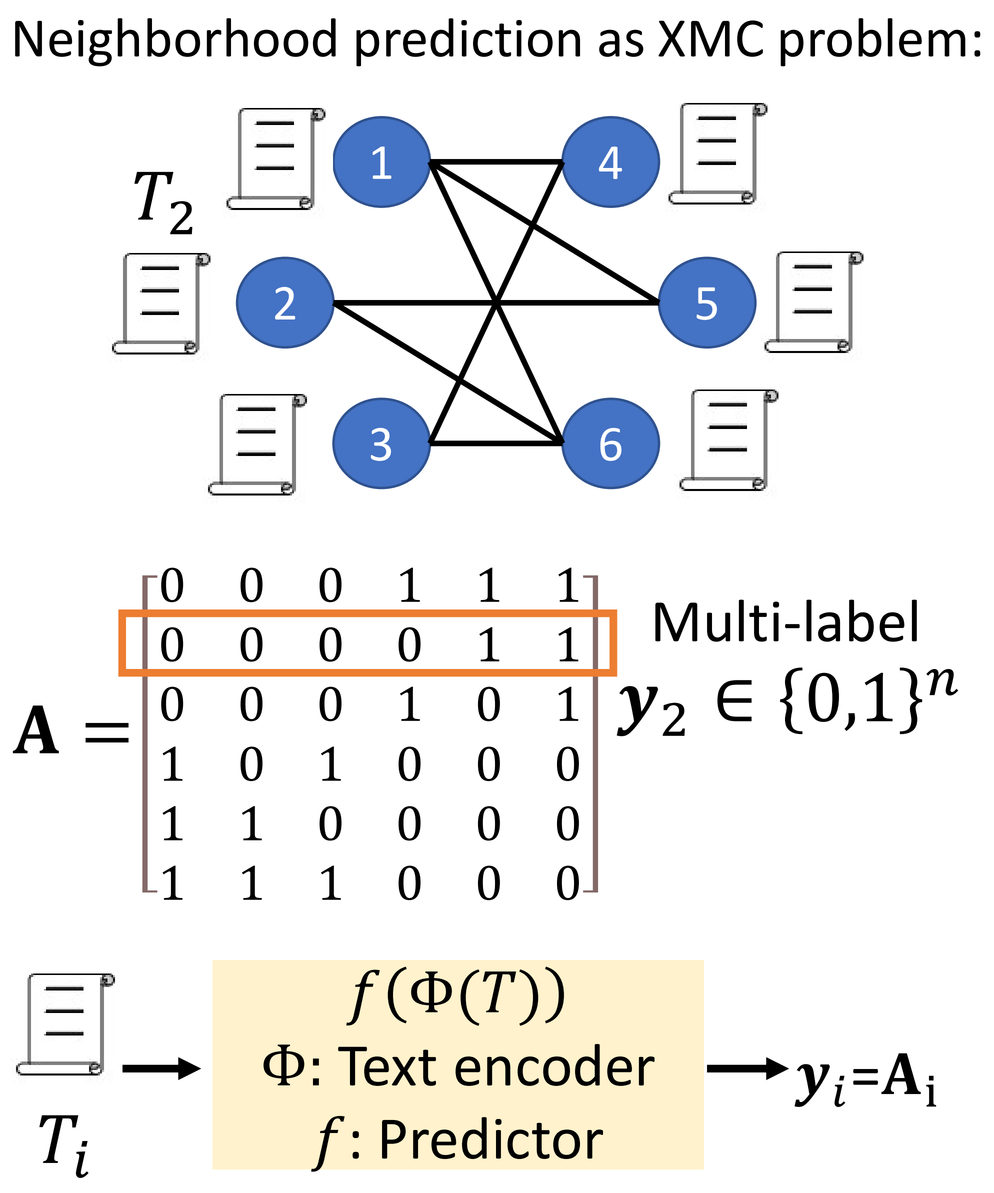
- Standard GNN pipeline:
  - Text features ⇒ BERT/Word2vec ⇒ GNN
- GIANT-XRT: pretrain the feature extractors (e.g., BERT) based on the graph information.



Graph Convolutional Neural Network

GIANT-XRT

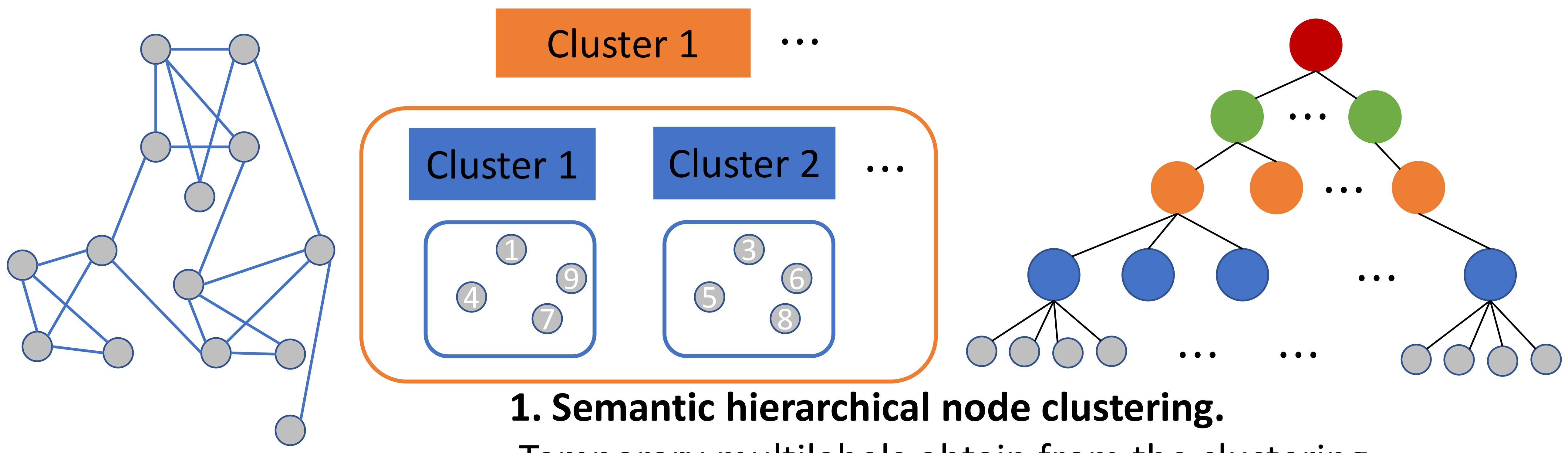
- Pretraining task: Predicting the Neighbors of each node
- Train BERT encoder to predict each row of adjacency matrix ⇒Multilabel classification with huge number of labels



Graph Convolutional Neural Network

GIANT-XRT

- State-of-the-art eXtreme Multilabel Classification (XMC) usually conducts multi-layer predictions.
- Example: PECOS, Parabel, …



1. Semantic hierarchical node clustering.
Temporary multilabels obtain from the clustering.