

Neural Language Model

CSE 447 / 517

January 20th, 2022 (Week 3)

Eisenstein (2019) 6 and Appendix A

Logistics

- A3 is due on **Wednesday, 1/26**

Agenda

- Quiz 2 Solutions
- Feedforward Neural Network
- Convolutional Neural network
- Q & A

Quiz 2 - Problem Setup

You wanted to take a class but you were not sure about the workload. You then asked some friends who took the class. They told you the time they spent and the GPA they got in this class, which are in the following table:

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

Based on the table, fill in the blanks. Round answer to 1 decimal place if not specified otherwise.

Quiz 2 - Question 1

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

If you randomly ask one of the friends above, what is the probability that the person got 3.8?

Quiz 2 - Question 1

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

If you randomly ask one of the friends above, what is the probability that the person got 3.8?

$$p(X = 3.8) = \frac{4}{9} \approx 0.4$$

Quiz 2 - Question 2

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

If you randomly ask two of the friends above, what is the probability that they both got 4.0? (Use simplified fraction)

Quiz 2 - Question 2

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

If you randomly ask two of the friends above, what is the probability that they both got 4.0? (Use simplified fraction)

$$\begin{aligned} p(X_1 = 4.0, X_2 = 4.0) &= p(X_1 = 4.0) \cdot p(X_2 = 4.0 \mid X_1 = 4.0) \\ &= \frac{3}{9} \cdot \frac{2}{8} \\ &= \frac{1}{12} \end{aligned}$$

Quiz 2 - Question 3

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

What is the probability of getting 4.0 if you spent 12 hours per week?

Quiz 2 - Question 3

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

What is the probability of getting 4.0 if you spent 12 hours per week?

$$\begin{aligned} p(X = 4.0 \mid H = 12) &= \frac{p(X = 4.0, H = 12)}{p(H = 12)} \\ &= \frac{\frac{2}{9}}{\frac{4}{9}} \\ &= 0.5 \end{aligned}$$

Quiz 2 - Question 4

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

What is the probability of getting 3.9 or above if you spent 10 hours per week?

Quiz 2 - Question 4

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

What is the probability of getting 3.9 or above if you spent 10 hours per week?

$$\begin{aligned} p(X \geq 3.9 | H = 10) &= \frac{p(X \geq 3.9, H = 10)}{p(H = 10)} \\ &= \frac{\frac{2}{9}}{\frac{5}{9}} \\ &= 0.4 \end{aligned}$$

Quiz 2 - Question 5

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

If one of your friends got 3.9 in this class, what is the probability that your friend spent 12 hours in this class?

Quiz 2 - Question 5

3.8	3.8	3.8	3.8	3.9	3.9	4.0	4.0	4.0
10	10	10	12	10	12	10	12	12

If one of your friends got 3.9 in this class, what is the probability that your friend spent 12 hours in this class?

$$\begin{aligned} p(H = 12 \mid X = 3.9) &= \frac{p(H = 12, X = 3.9)}{p(X = 3.9)} \\ &= \frac{\frac{1}{9}}{\frac{2}{9}} \\ &= 0.5 \end{aligned}$$

Quiz 2 - Problem Setup

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

We have trained a Bi-gram model based on some data. We store the frequencies of each pair of words in the following table. Each cell represents the occurrences of the top row word following right after the left column word. For example, “finished” appeared after “I” 40 times.

Quiz 2 - Question 6

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

I ____

Quiz 2 - Question 6

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

I finished ____

Quiz 2 - Question 6

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

I finished work

Quiz 2 - Question 7

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

Is the work ____

Quiz 2 - Question 7

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

Is the work finished

Quiz 2 - Question 8

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

I saw ____ _

Quiz 2 - Question 8

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

I saw the ____ ____

Quiz 2 - Question 8

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

I saw the beautiful ____

Quiz 2 - Question 8

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

I saw the beautiful gift

Quiz 2 - Question 9

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

the ____

Quiz 2 - Question 9

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

the beautiful ____

Quiz 2 - Question 9

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

the beautiful gift ____ ____

Quiz 2 - Question 9

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

the beautiful gift finished ____

Quiz 2 - Question 9

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

Please fill in the most likely word based on the previous word and the frequency table for the following sentences:

the beautiful gift finished work

Quiz 2 - Question 10

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

What is the probability that "**work**" appears after "**finished**"? (Use simplified fraction)

Quiz 2 - Question 10

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

What is the probability that "**work**" appears after "**finished**"? (Use simplified fraction)

Quiz 2 - Question 10

1st/2nd	I	finished	work	saw	the	beautiful	gift
I	0	40	20	30	0	0	0
finished	0	0	10	0	8	5	3
work	0	5	0	0	0	0	0
saw	0	5	5	0	8	5	3
the	0	10	15	0	0	20	10
beautiful	0	0	5	0	0	0	10
gift	0	5	0	0	0	3	0

What is the probability that "**work**" appears after "**finished**"? (Use simplified fraction)

$$p(X_2 = \text{work} \mid X_1 = \text{finished}) = \frac{10}{10+8+5+3} = \frac{5}{13}$$

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\underbrace{\mathbf{b}_V + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j} \mathbf{A}_j}_{d \times d \times V} + \underbrace{\mathbf{W}}_{V \times H}}_{\text{affine}} \underbrace{\tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j}^\top \mathbf{T}_j}_{d \times H}}_{\text{affine}} \right)}_{\text{nonlinearity}}}_{\text{nonlinearity}} \right)$$

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\underbrace{\mathbf{b}_v + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j} \mathbf{A}_j}_{d \times d \times V} + \mathbf{W}_{V \times H}}_{\text{affine}} \tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j}^\top \mathbf{T}_j}_{d \times H}}_{\text{affine}} \right)}_{\text{nonlinearity}} \right)_{\text{nonlinearity}}$$

Embedding of
history token h_j .

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_v + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j}^T \mathbf{A}_j}_{\text{affine}} + \mathbf{W}_{v \times H}}_{\text{affine}} \underbrace{\tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^T \mathbf{T}_j}_{\text{affine}} \right)}_{\text{nonlinearity}} \right)_{\text{nonlinearity}}$$

Diagram illustrating the structure of the n-gram probability calculation, showing nested affine and nonlinearity layers. The innermost expression is $\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^T \mathbf{T}_j$, which is labeled as an affine transformation. This is followed by a \tanh nonlinearity. The result is then combined with $\mathbf{b}_v + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^T \mathbf{A}_j + \mathbf{W}_{v \times H}$ in an affine layer, and finally passed through a softmax nonlinearity.

Matrix that transforms the embedding from dimension \mathbb{R}^d to \mathbb{R}^H .

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_v + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j} \mathbf{A}_j}_{\text{affine}} + \mathbf{W}_{v \times H}}_{\text{affine}} \tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_j}_{\text{affine}} \right) \right)$$

nonlinearity

nonlinearity

Sum over all
n-1 history
tokens.

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

Bias term.

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_v + \sum_{j=1}^{n-1} \mathbf{m}_{h_j} \mathbf{A}_j}_{\text{affine}} + \underbrace{\mathbf{W}_{v \times H} \tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_j}_{\text{affine}} \right)}_{\text{nonlinearity}} \right)$$

affine
nonlinearity

affine
nonlinearity

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_V + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j} \mathbf{A}_j}_{\text{affine}} + \mathbf{W}_{V \times H}}_{\text{affine}} \tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_j}_{\text{affine}} \right) \right)$$

Diagram illustrating the structure of the n-gram probability calculation:

- The inner expression $\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_j$ is labeled "affine".
- The \tanh function is labeled "nonlinearity".
- The expression $\mathbf{b}_V + \sum_{j=1}^{n-1} \mathbf{m}_{h_j} \mathbf{A}_j + \mathbf{W}_{V \times H}$ is labeled "affine".
- The entire expression is labeled "nonlinearity".

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_V + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_d}_{d \times V} \mathbf{A}_j + \mathbf{W}_{V \times H}}_{\text{affine}} \underbrace{\tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_j}_{\text{affine}} \right)}_{\text{nonlinearity}} \right)_{\text{nonlinearity}}$$

Passing the value through a nonlinearity.

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_V + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_d^{h_j} \mathbf{A}_{d \times V}}_{\text{affine}} + \underbrace{\mathbf{W}_{V \times H} \tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_d^{h_j} \mathbf{T}_{d \times H}}_{\text{affine}}}_{\text{affine}} \right)}_{\text{nonlinearity}}}_{\text{affine}} \right)$$

Diagram illustrating the structure of the Feedforward Neural Network equation, highlighting the components:

- Nonlinearity**: The \tanh activation function.
- Affine**: The inner affine transformation $\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_d^{h_j} \mathbf{T}_{d \times H}$.

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_V + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_{h_j}_d \mathbf{A}_{d \times V}}_{\text{affine}} + \underbrace{\mathbf{W}_{V \times H} \tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_{d \times H}}_{\text{affine}} \right)}_{\text{nonlinearity}}}_{\text{affine}} \right)_{\text{nonlinearity}}$$

Nonlinearity

Affine

Repeat!

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\mathbf{b}_V + \sum_{j=1}^{n-1} \mathbf{m}_{h_j} \mathbf{A}_j + \mathbf{W}_{V \times H}}_{\text{affine}} \underbrace{\tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_j}_{\text{affine}} \right)}_{\text{nonlinearity}} \right)_{\text{nonlinearity}}$$

The diagram illustrates the structure of the equation for the n-gram probability. It shows a sequence of operations: an affine layer (represented by the sum of bias, input-weight, and hidden-weight terms), followed by a nonlinearity (the tanh activation function), and finally another affine layer (the softmax function). The entire process is labeled as a 'nonlinearity' at the bottom. Brackets and labels identify the 'affine' and 'nonlinearity' components within the equation.

Nonlinearity

Affine

Typical pattern

affine, nonlinear,
affine, nonlinear,
...

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

Softmax to ensure the output sums to 1.

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\underbrace{\mathbf{b}_V + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_d}_{h_j} \underbrace{\mathbf{A}_{d \times V}}_j}_{\text{affine}} + \underbrace{\underbrace{\mathbf{W}_{V \times H}}_{\text{nonlinearity}} \tanh \left(\underbrace{\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \underbrace{\mathbf{m}_d^\top}_{h_j} \underbrace{\mathbf{T}_{d \times H}}_j}_{\text{affine}} \right)}_{\text{affine}}}_{\text{nonlinearity}} \right)$$

Parameters θ include \mathbf{M} and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \dots, h_{n-1}) = \text{softmax} \left(\underbrace{\underbrace{\mathbf{b}_v + \sum_{j=1}^{n-1} \mathbf{m}_{h_j} \mathbf{A}_j}_{\text{affine}} + \underbrace{\mathbf{W}_{v \times H} \tanh \left(\underbrace{\mathbf{u}_H + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \mathbf{T}_j}_{\text{affine}} \right)}_{\text{nonlinearity}}}_{\text{nonlinearity}} \right)$$

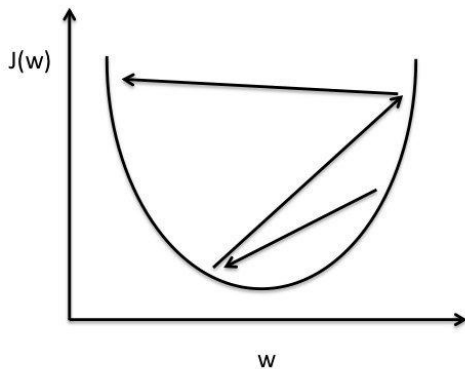
! Here we also added the affine transformation of embedding of history tokens too.

Parameters θ include \mathbf{M} and everything in pink.

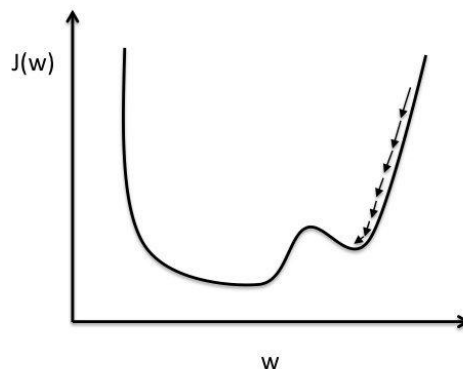
Hyperparameters: dimensionalities d and H

Feedforward Neural Network: Gradient Descent

- Gradient descent is an update on every parameter each iteration
- Does not guarantee to give the optimal solution (gives a local minimum)
- Batch size, epoch, learning rate, various optimizers
 - Stochastic Gradient Descent (SGD), Adam, Adadelta, ...



Large learning rate: Overshooting.



Small learning rate: Many iterations until convergence and trapping in local minima.

Feedforward Neural Network: Hyperparameters

$$D = \underbrace{Vd}_{\mathbf{M}} + \underbrace{V}_{\mathbf{b}} + \underbrace{(n-1)dV}_{\mathbf{A}} + \underbrace{VH}_{\mathbf{W}} + \underbrace{H}_{\mathbf{u}} + \underbrace{(n-1)dH}_{\mathbf{T}}$$

For Bengio et al. (2003):

- ▶ $V \approx 18000$ (after OOV processing)
- ▶ $d \in \{30, 60\}$
- ▶ $H \in \{50, 100\}$
- ▶ $n - 1 = 5$

Feedforward Neural Network: Hyperparameters

$$D = \underbrace{Vd}_{\mathbf{M}} + \underbrace{V}_{\mathbf{b}} + \underbrace{(n-1)dV}_{\mathbf{A}} + \underbrace{VH}_{\mathbf{W}} + \underbrace{H}_{\mathbf{u}} + \underbrace{(n-1)dH}_{\mathbf{T}}$$

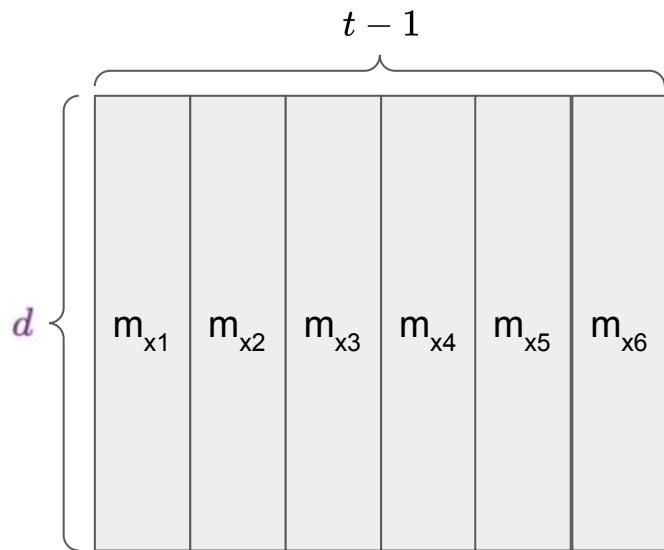
For Bengio et al. (2003):

- ▶ $V \approx 18000$ (after OOV processing)
- ▶ $d \in \{30, 60\}$
- ▶ $H \in \{50, 100\}$
- ▶ $n - 1 = 5$

Tune hyperparameters on dev set

Convolutional Neural Network

Start with $\mathbf{X}^{(0)} = [\mathbf{m}_{x_1}; \mathbf{m}_{x_2}; \dots; \mathbf{m}_{x_{t-1}}]$.



Convolutional Neural Network: Convolution

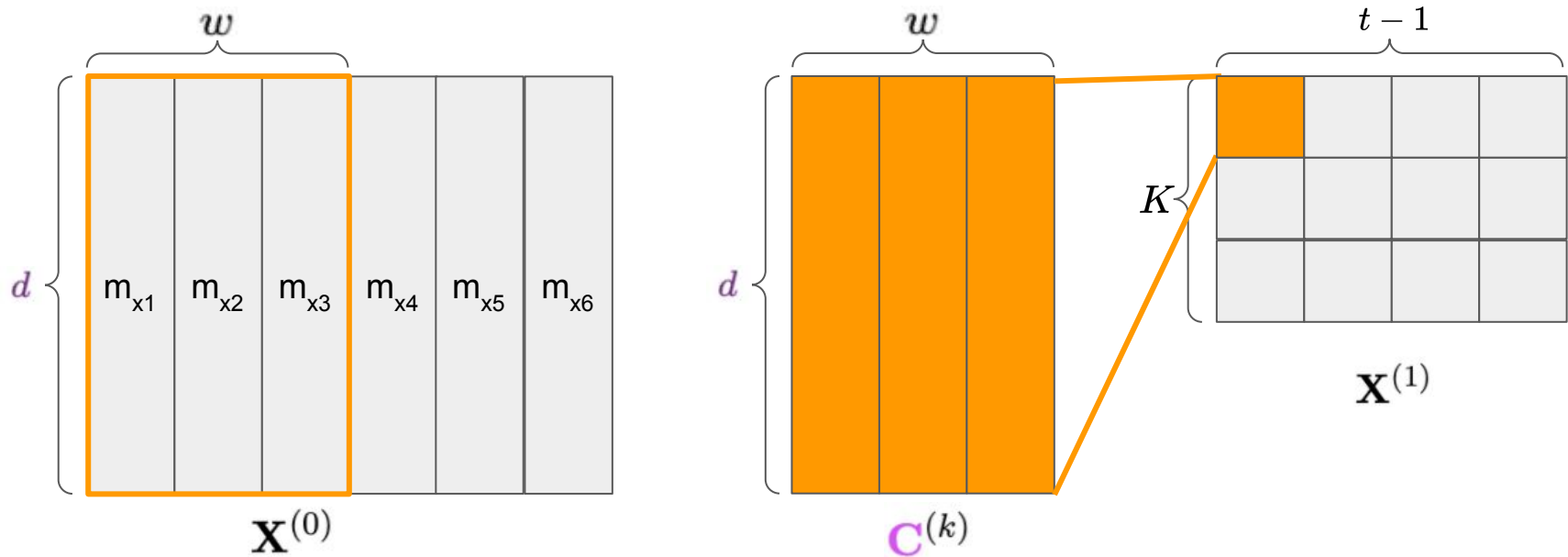
$$\mathbf{X}^{(1)}[k, m] = f \left(\textcolor{violet}{b}_k + \sum_{i=1}^{\textcolor{violet}{d}} \sum_{j=1}^w \textcolor{violet}{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$

f is a nonlinearity (like \tanh). w is the width of the sliding window. Each k is a different “filter” and each m is a word position.

Hyperparameters: number of layers, and, at every layer, f , w , number of filters

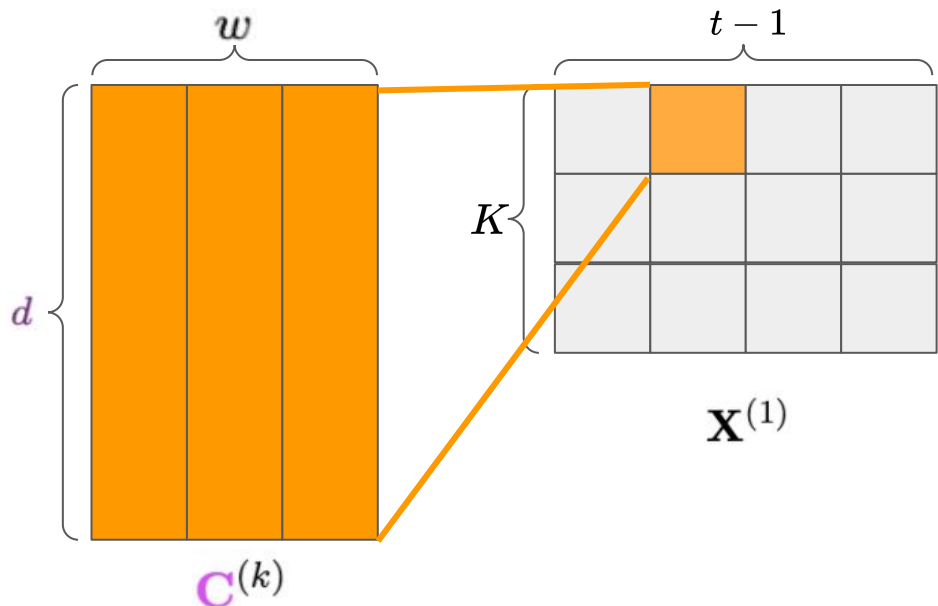
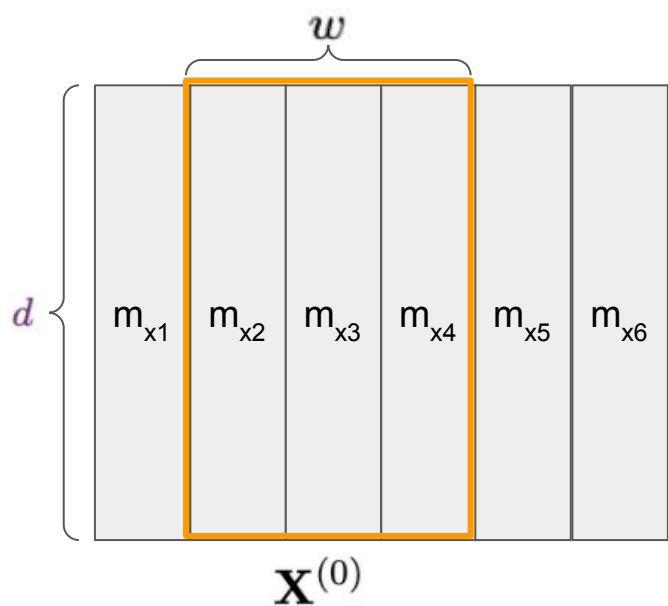
Convolutional Neural Network: Convolution

$$\mathbf{X}^{(1)}[k, m] = f \left(\mathbf{b}_k + \sum_{i=1}^d \sum_{j=1}^w \mathbf{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$



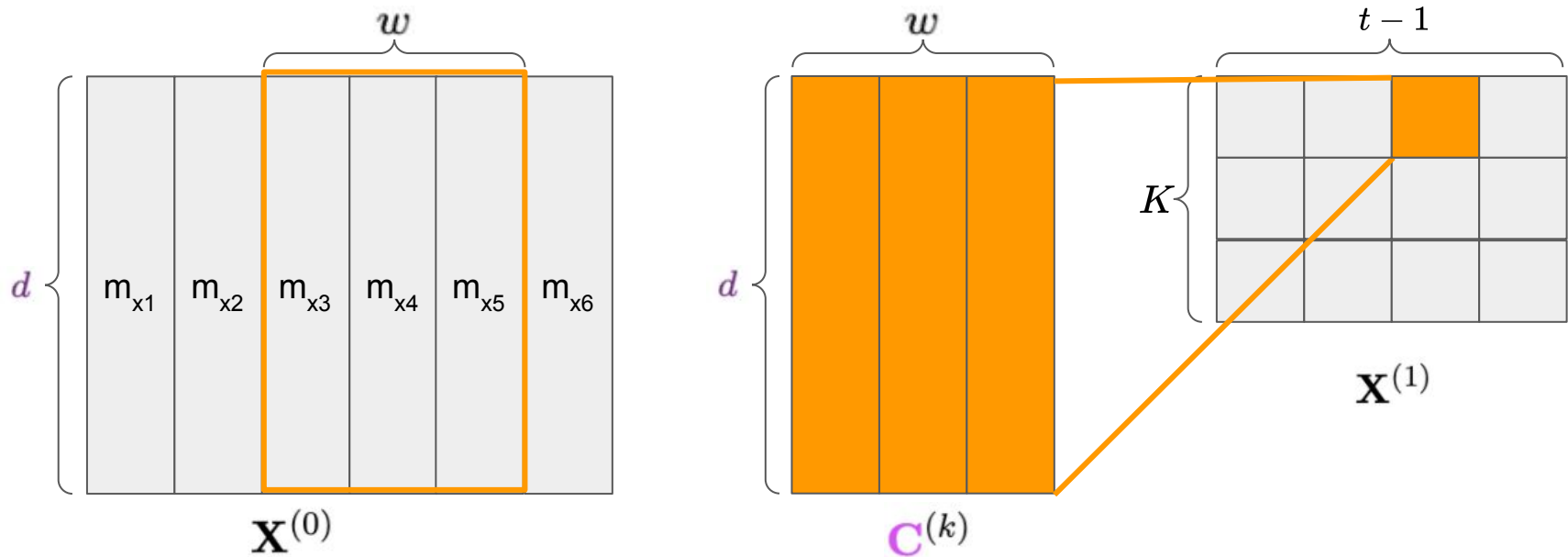
Convolutional Neural Network: Convolution

$$\mathbf{X}^{(1)}[k, m] = f \left(\mathbf{b}_k + \sum_{i=1}^d \sum_{j=1}^w \mathbf{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$



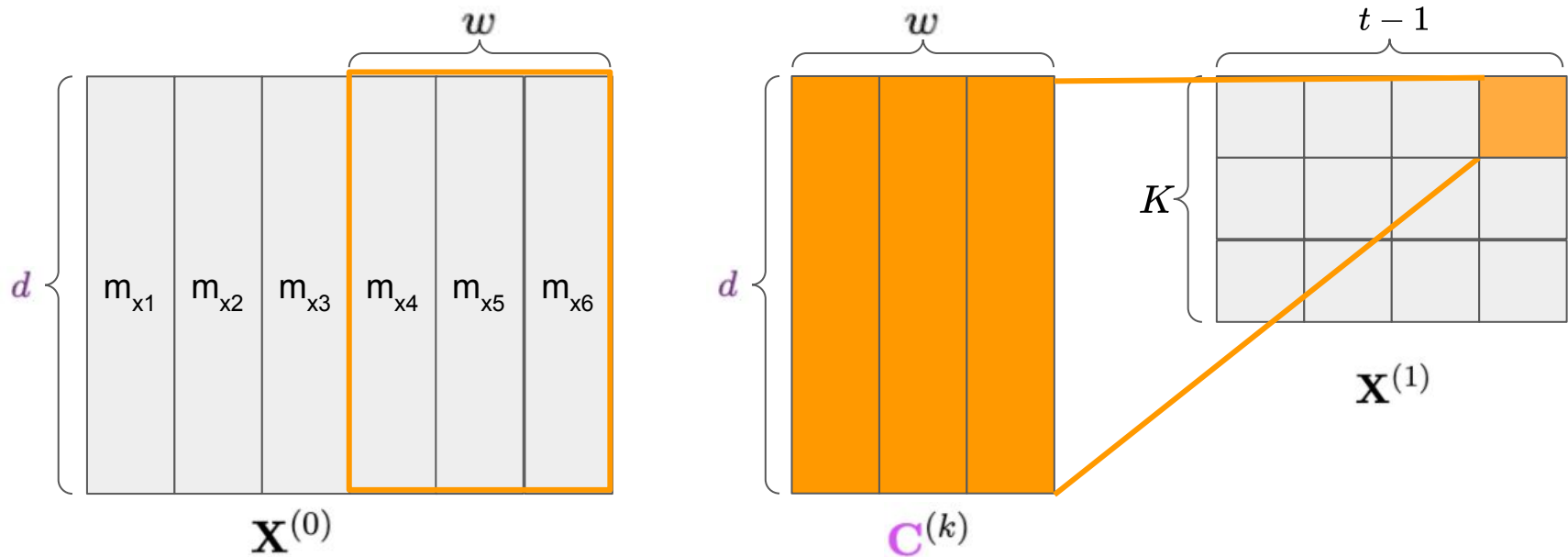
Convolutional Neural Network: Convolution

$$\mathbf{X}^{(1)}[k, m] = f \left(\mathbf{b}_k + \sum_{i=1}^d \sum_{j=1}^w \mathbf{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$



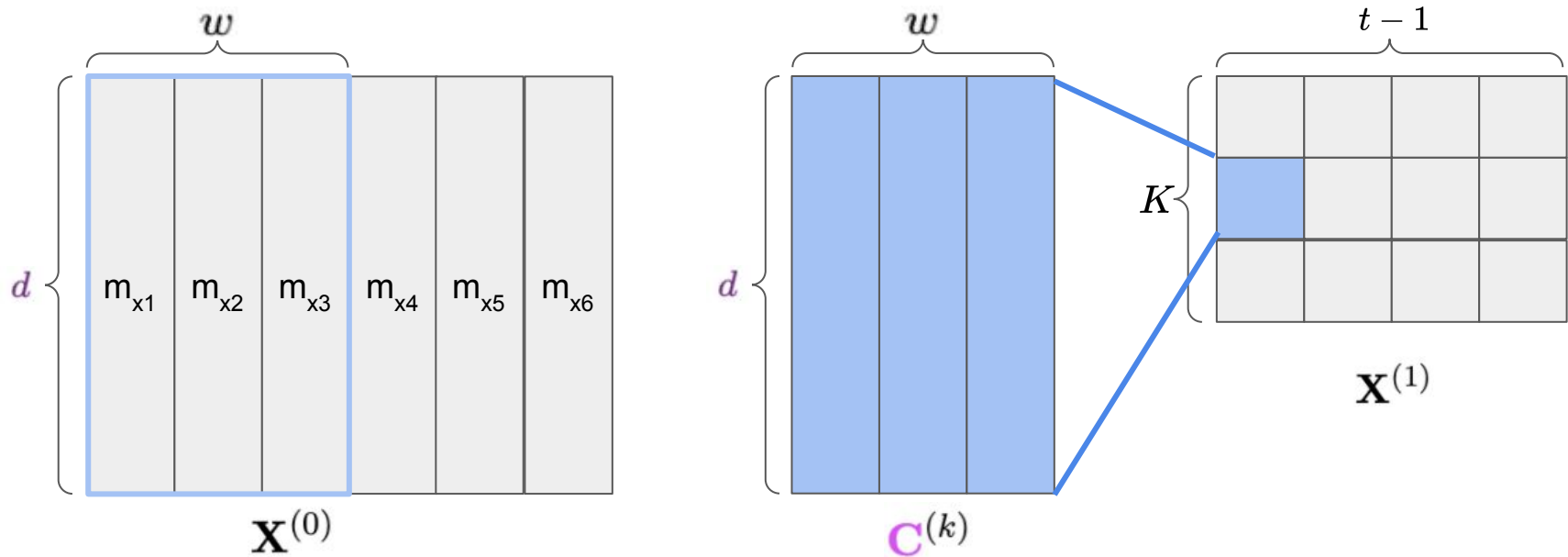
Convolutional Neural Network: Convolution

$$\mathbf{X}^{(1)}[k, m] = f \left(\textcolor{violet}{b}_k + \sum_{i=1}^{\textcolor{violet}{d}} \sum_{j=1}^w \textcolor{violet}{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$



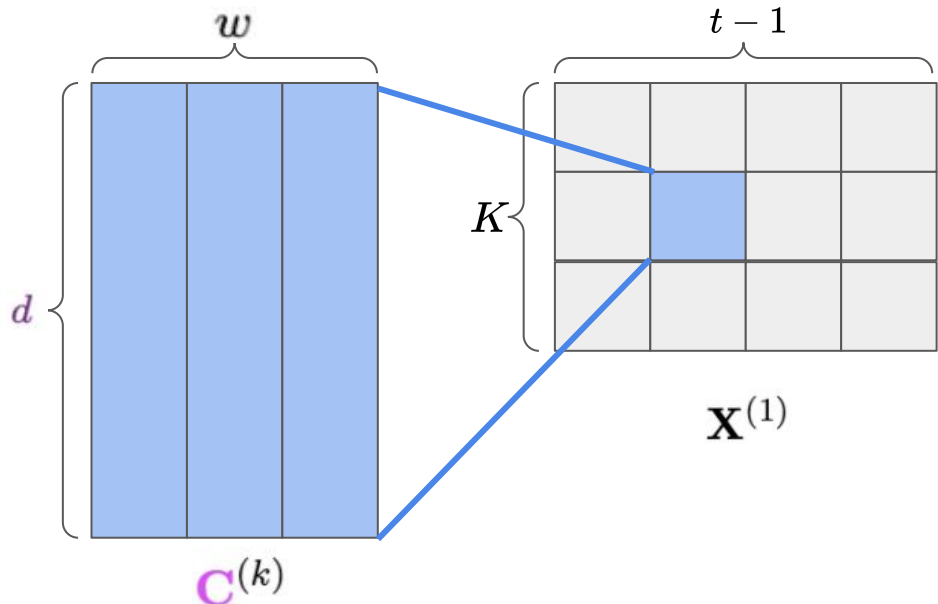
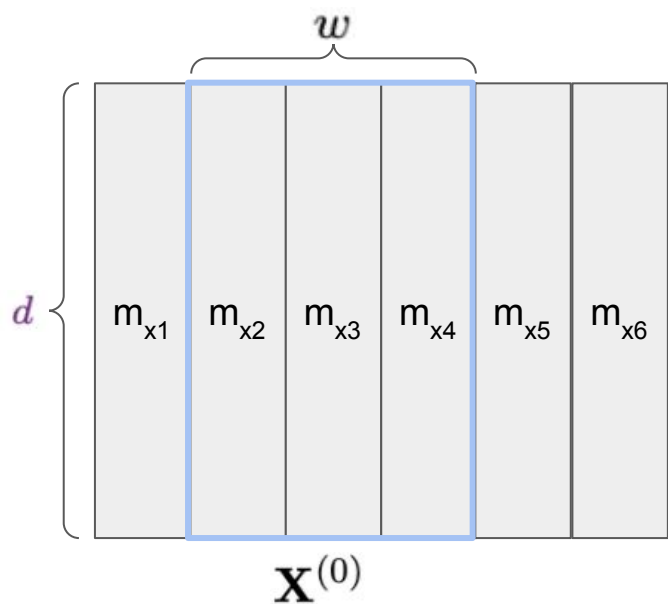
Convolutional Neural Network: Convolution

$$\mathbf{X}^{(1)}[k, m] = f \left(\mathbf{b}_k + \sum_{i=1}^d \sum_{j=1}^w \mathbf{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$



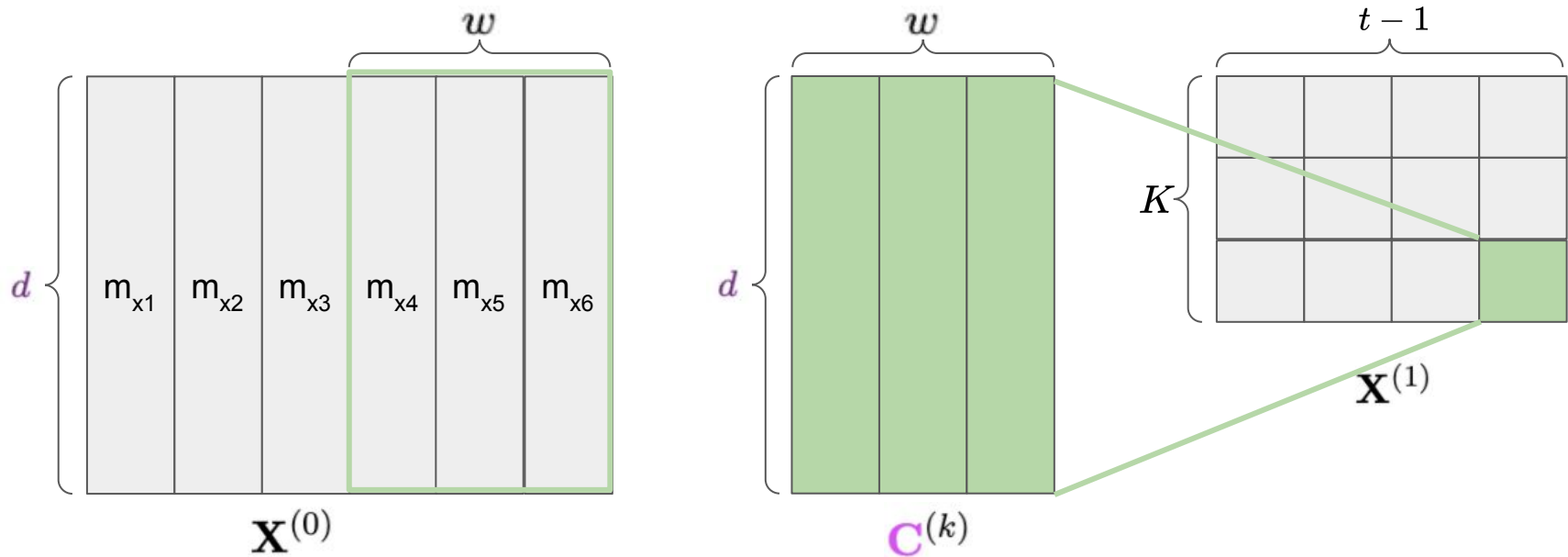
Convolutional Neural Network: Convolution

$$\mathbf{X}^{(1)}[k, m] = f \left(\textcolor{violet}{b}_k + \sum_{i=1}^{\textcolor{violet}{d}} \sum_{j=1}^w \textcolor{violet}{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$

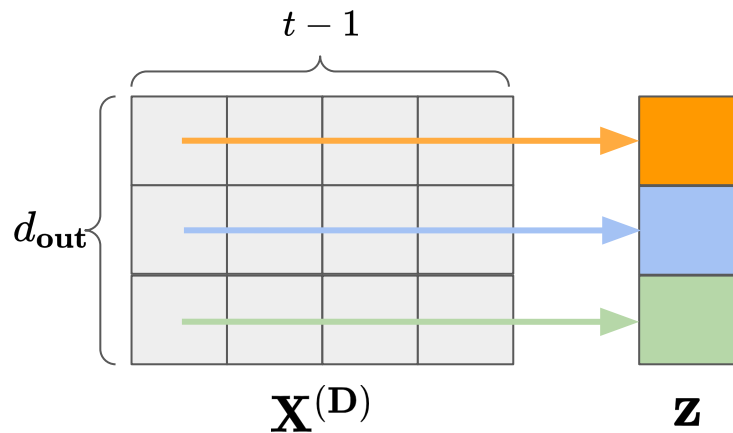


Convolutional Neural Network: Convolution

$$\mathbf{X}^{(1)}[k, m] = f \left(\mathbf{b}_k + \sum_{i=1}^d \sum_{j=1}^w \mathbf{C}^{(k)}[i, j] \cdot \mathbf{X}^{(0)}[i, m + j - 1] \right)$$



Convolutional Neural Network: Pooling



Pooling takes $\mathbf{X}^{(D)} \in \mathbb{R}^{d_{\text{out}} \times (t-1)}$ and maps it into $\mathbb{R}^{d_{\text{out}}}$.

Two standard options (with no additional parameters) are max pooling,

$$z_k = \max_j \mathbf{X}^{(D)}[k, j];$$

and average pooling,

$$z_k = \frac{1}{t-1} \sum_{j=1}^{t-1} \mathbf{X}^{(D)}[k, j].$$

Finally, $\text{softmax}(\mathbf{z})$ gives a probability distribution over outputs.

Q & A