Neural Language Model

CSE 447 / 517
January 20th, 2022 (Week 3)

Eisenstein (2019) 6 and Appendix A

Logistics

- CSE 517 project proposal is due on **Friday 1/21**

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)

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

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?

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

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:

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

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \ldots, h_{n-1}) =$$

$$\underbrace{\text{softmax}\left(\underbrace{\underset{V}{\mathbf{b}} + \sum_{j=1}^{n-1} \underset{d}{\mathbf{m}_{h_j}} \underset{d \times V}{\mathbf{A}_j} + \underset{V \times H}{\mathbf{W}} \underbrace{\tanh\left(\underbrace{\underset{H}{\mathbf{u}} + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^{\top} \underset{d \times H}{\mathbf{T}_j}}_{\text{affine}} \right)}_{\text{nonlinearity}}}_{\text{affine}} \right)}_{\text{nonlinearity}}$$

Embedding of history token h_j.

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

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

Matrix that transforms the embedding from dimension R^d to R^H.

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \ldots, h_{n-1}) =$$

$$\underbrace{\text{softmax}\left(\underbrace{\underset{V}{\mathbf{b}} + \sum_{j=1}^{n-1} \underset{d}{\mathbf{m}_{h_j}} \underset{d \times V}{\mathbf{A}_j} + \underset{V \times H}{\mathbf{W}} \underbrace{\tanh\left(\underbrace{\underset{H}{\mathbf{u}} + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^\top \underset{d \times H}{\mathbf{T}_j}}_{\text{affine}} \right)}_{\text{nonlinearity}}}_{\text{affine}} \right)}_{\text{nonlinearity}}$$

Sum over all n-1 history tokens.

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

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

Bias term.

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

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

Affine

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

$$p(\cdot \mid h_1, \ldots, h_{n-1}) =$$

$$\underbrace{\text{softmax}\left(\underbrace{\underset{V}{\mathbf{b}} + \sum_{j=1}^{n-1} \underset{d}{\mathbf{m}_{h_j}} \underset{d \times V}{\mathbf{A}_j} + \underset{V \times H}{\mathbf{W}} \underbrace{\tanh\left(\underbrace{\underset{H}{\mathbf{u}} + \sum_{j=1}^{n-1} \mathbf{m}_{h_j}^{\top} \underset{d \times H}{\mathbf{T}_j}}_{\text{affine}} \right)}_{\text{nonlinearity}}}_{\text{affine}} \right)}_{\text{nonlinearity}}$$

Passing the value through a nonlinearity.

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

Nonlinearity

Affine

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

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

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

Nonlinearity

Affine

Repeat!

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

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

Nonlinearity

Affine

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Typical pattern

affine, nonlinear,
affine, nonlinear,
. . .

Feedforward Neural Network

Define the n-gram probability as follows:

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

Softmax to ensure the output sums to 1.

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

Hyperparameters: dimensionalities d and H

Feedforward Neural Network

Define the n-gram probability as follows:

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

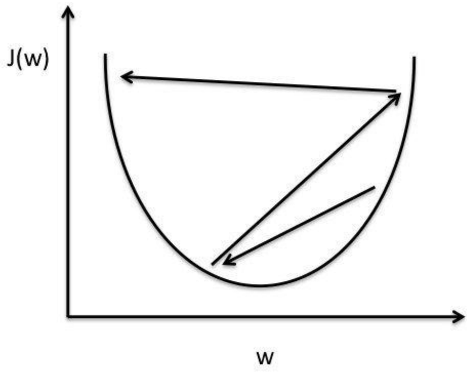
! **Here we also added the affine transformation of embedding of history tokens too.**

Parameters $\boldsymbol{\theta}$ include $\mathbf{M}$ and everything in pink.

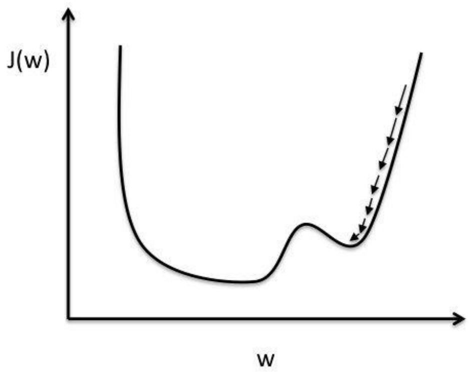
Hyperparameters: dimensionalities d and H

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

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

For Bengio et al. (2003):

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

Feedforward Neural Network: Hyperparameters

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

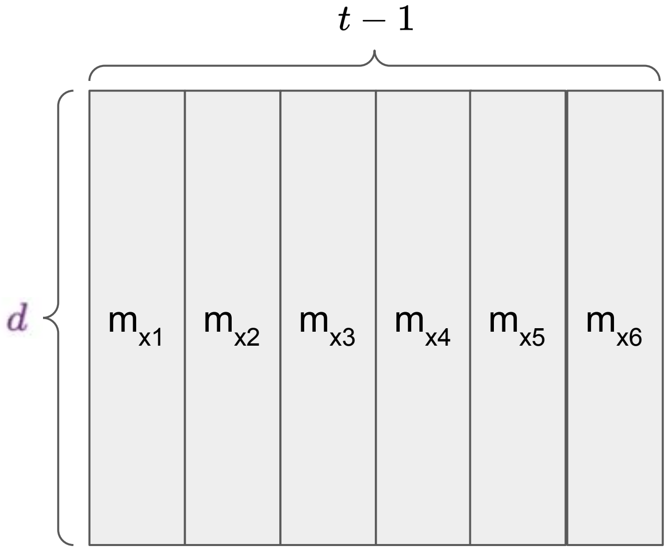
For Bengio et al. (2003):

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

Tune hyperparameters on dev set

Convolutional Neural Network

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



Convolutional Neural Network: Convolution

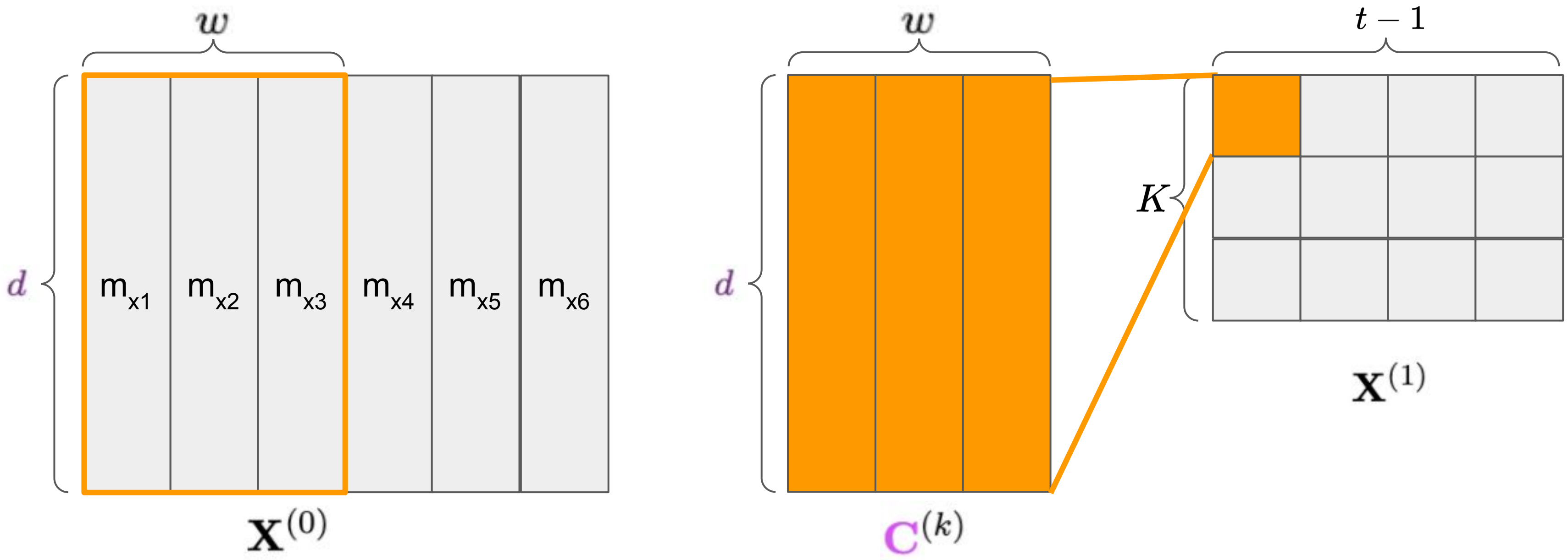
$$\mathbf{X}^{(1)}[k,m] = f\left(b_k + \sum_{i=1}^{d}\sum_{j=1}^{w} \mathbf{C}^{(k)}[i,j] \cdot \mathbf{X}^{(0)}[i, m+j-1]\right)$$

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

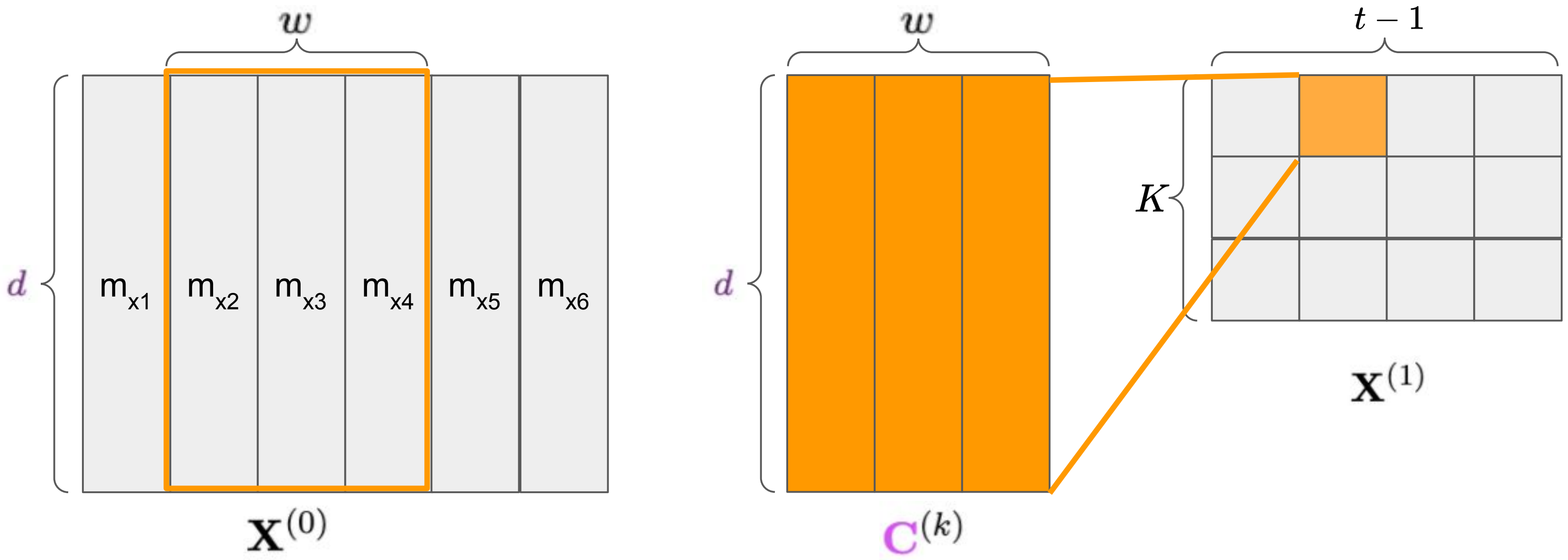
Convolutional Neural Network: Convolution

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



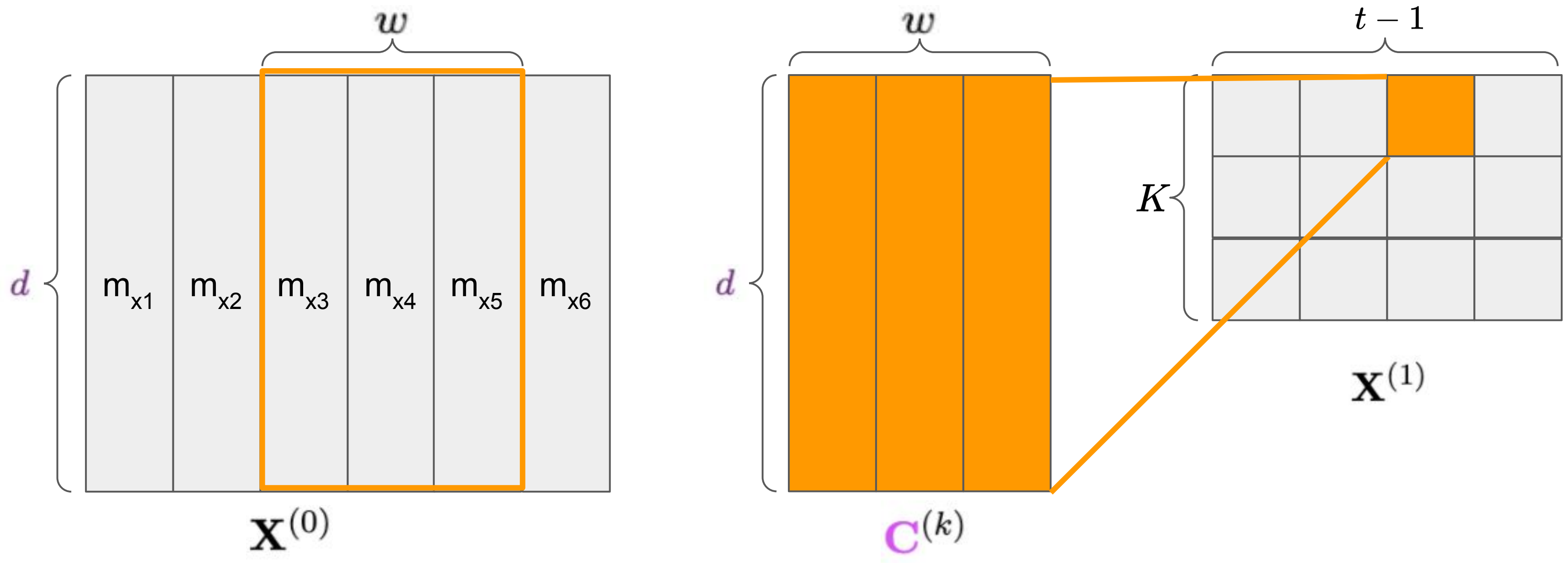
Convolutional Neural Network: Convolution

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



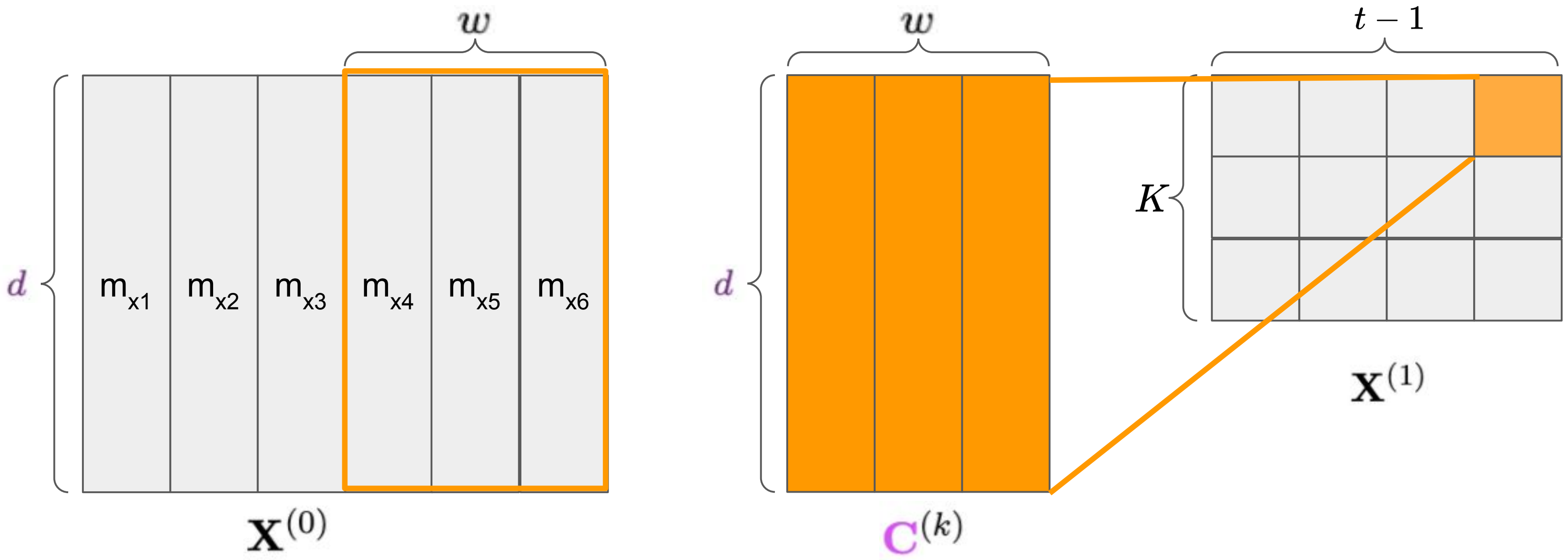
Convolutional Neural Network: Convolution

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



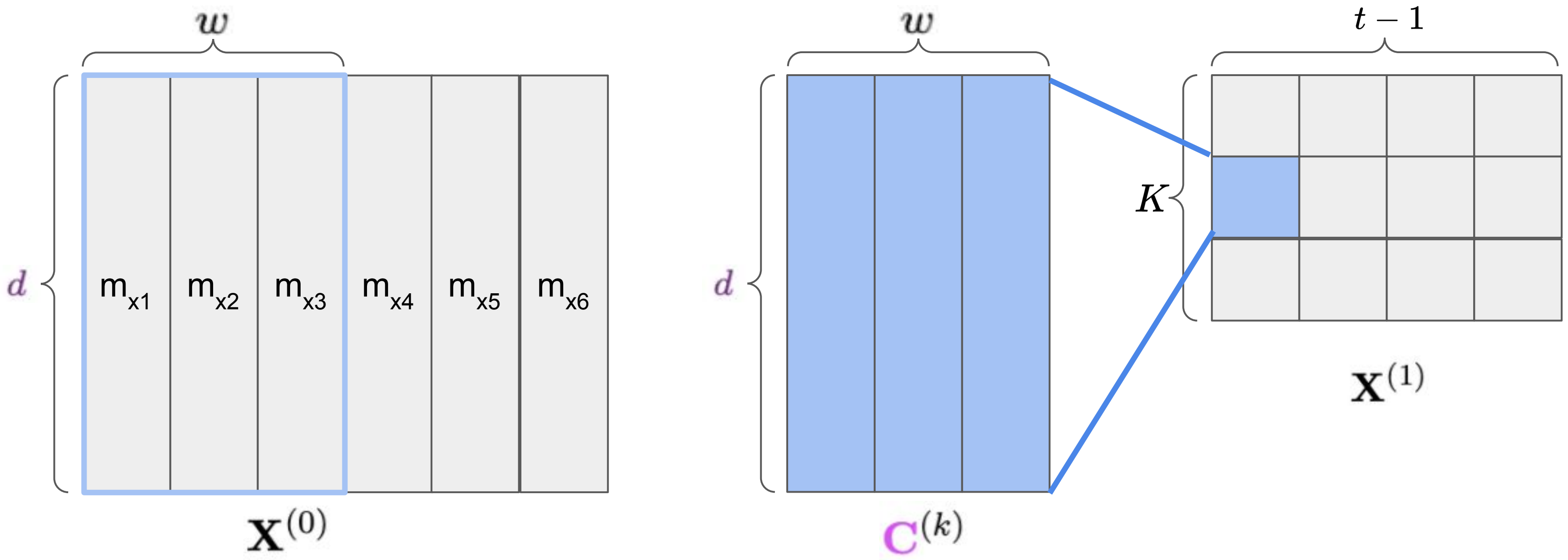
Convolutional Neural Network: Convolution

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



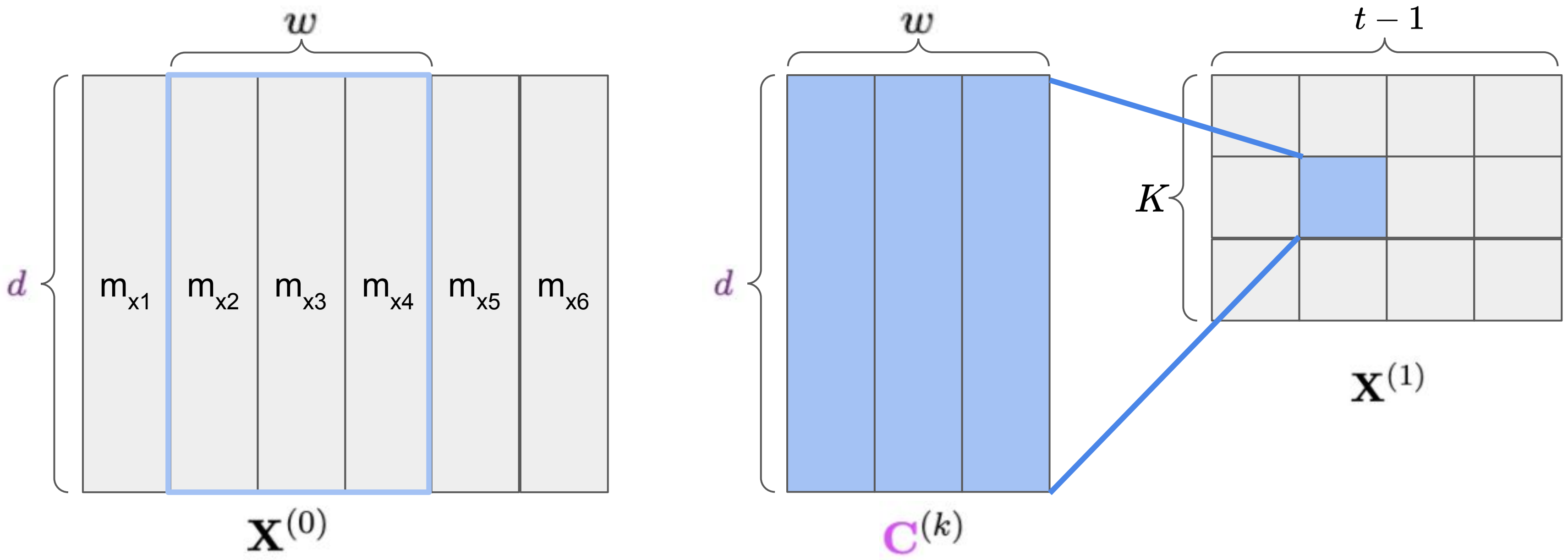
Convolutional Neural Network: Convolution

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



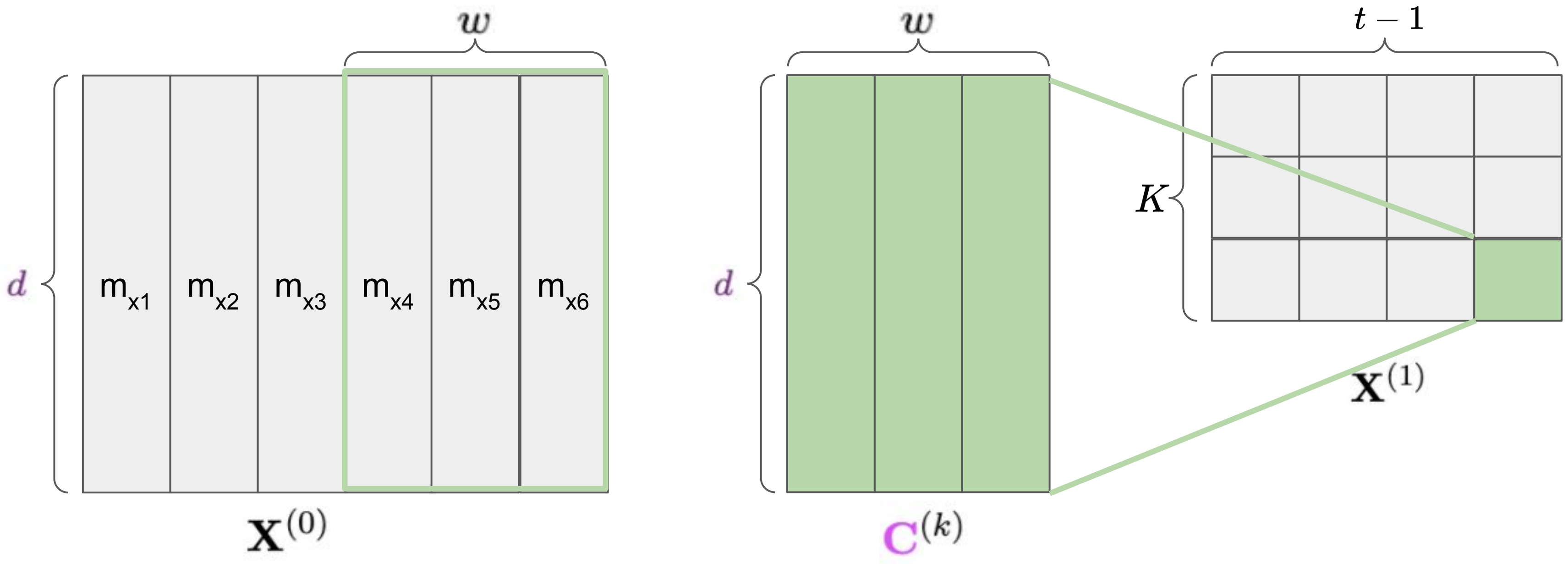
Convolutional Neural Network: Convolution

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

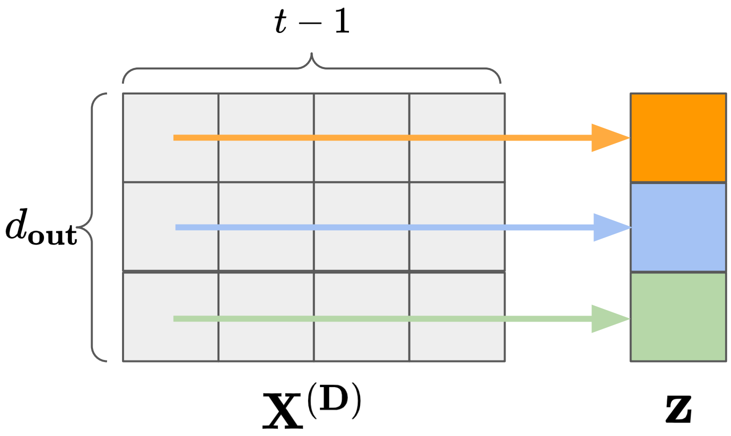


Convolutional Neural Network: Convolution

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



Convolutional Neural Network: Pooling



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

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

Q & A