EXERCISE · SLIDE 7 · QUESTION

Tokens and projections

A new token, the same learned map

$x = (2, -1, 1)^T$, $W_Q = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & -1 \end{bmatrix}$. Find $q = W_Q x$.

A $(3, -3)^T$

B $(1, -1)^T$

C A three-coordinate vector

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 7 · SOLUTION 1/2

Tokens and projections

A new token, the same learned map

Solution 1 / 2

Apply the two output rows

x: input column; W_Q: 2 × 3 map; q: output column.

1. $q_1 = 1(2) + 0(-1) + 1(1) = 3.$
2. $q_2 = 0(2) + 2(-1) - 1(1) = -3.$
3. $q = (3, -3)^T$ has two coordinates.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 7 · SOLUTION 2/2

Tokens and projections

A new token, the same learned map

Solution 2 / 2

Check the alternatives

q: projected token.

1. A is correct. B drops coefficients from the map.
2. C keeps the input width, but the two weight rows define two outputs.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 7 · CHECK THE RESULT

Tokens and projections

A new token, the same learned map

$x = (2, -1, 1)^T$, $W_Q = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & -1 \end{bmatrix}$. Find $q = W_Q x$.

- **A** $(3, -3)^T$ ✓
- **B** $(1, -1)^T$
- **C** A three-coordinate vector

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 9 · QUESTION

Scaled dot-product attention

Alignment and length both affect the score

$q = (1, 0)^T$, $k_1 = (1, 0)^T$, $k_2 = (2, 0)^T$, $d_k = 2$. Which key has the larger score?

A Key 1

B Key 2

C They must tie

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 9 · SOLUTION 1/1

Scaled dot-product attention

Alignment and length both affect the score

Solution 1 / 1

Evaluate the dot products

q: query; k_1, k_2: keys; $d_k = 2$.

1. $q^T k_1 = 1$ and $q^T k_2 = 2$.
2. $s_1 = 1/\sqrt{2} < s_2 = 2/\sqrt{2}$.
3. B is correct. Equal directions alone do not imply equal dot products.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 9 · CHECK THE RESULT

Scaled dot-product attention

Alignment and length both affect the score

$q = (1, 0)^T$, $k_1 = (1, 0)^T$, $k_2 = (2, 0)^T$, $d_k = 2$. Which key has the larger score?

A Key 1

B Key 2 ✓

C They must tie

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 13 · QUESTION

Scaled dot-product attention

Unequal scores change the mixture

$s = (0, \log 3)$, $v_1 = (2, 0)^T$, $v_2 = (0, 4)^T$. Find the output.

A $(1, 2)^T$

B $(1/2, 3)^T$

C $(0, 4)^T$

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 13 · SOLUTION 1/2

Scaled dot-product attention

Unequal scores change the mixture

Solution 1 / 2

Convert scores to weights

s: score row; p: its softmax weights.

1. $e^s = (1, 3)$; the sum is 4.
2. $p = (1/4, 3/4)$.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 13 · SOLUTION 2/2

Scaled dot-product attention

Unequal scores change the mixture

Solution 2 / 2

Mix the value vectors

$v_1 = (2, 0)^T$, $v_2 = (0, 4)^T$.

---

1. $y = \frac{1}{4} v_1 + \frac{3}{4} v_2 = (1/2, 3)^T$.
2. B is correct. A treats scores as equal; C replaces soft weighting with a hard maximum.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 13 · CHECK THE RESULT

Scaled dot-product attention

Unequal scores change the mixture

$s = (0, \log 3)$, $v_1 = (2, 0)^T$, $v_2 = (0, 4)^T$. Find the output.

A $(1, 2)^T$

B $(1/2, 3)^T$ ✓

C $(0, 4)^T$

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 17 · QUESTION

Scaled dot-product attention

The scale is determined by the assumptions

Independent zero-mean, unit-variance coordinates; $d_k = 16$. What is the standard deviation of the unscaled dot product?

A 16

B 4

C 1

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 17 · SOLUTION 1/1

Scaled dot-product attention

The scale is determined by the assumptions

Solution 1 / 1

Separate variance from standard deviation

$u = q^T k$; $d_k = 16$.

---

1. $\mathrm{Var}(u) = 16$.
2. $\mathrm{SD}(u) = \sqrt{16} = 4$.
3. B is correct. A confuses variance with standard deviation; C describes the scaled score.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 17 · CHECK THE RESULT

Scaled dot-product attention

The scale is determined by the assumptions

Independent zero-mean, unit-variance coordinates; $d_k = 16$. What is the standard deviation of the unscaled dot product?

A 16

B 4 ✓

C 1

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 20 · QUESTION

Stable attention and masks

A shift or a temperature?

A score row is $s = (0, \log 3)$. Which change preserves its softmax weights?

- **A** Subtract 5 from both entries
- **B** Multiply both entries by 2
- **C** Add 5 only to the second entry

Blanchard, Higham & Higham (2021), §4 · Course example

EXERCISE · SLIDE 20 · SOLUTION 1/1

Stable attention and masks

A shift or a temperature?

Solution 1 / 1

Identify the invariance

s: logits; softmax normalizes their exponentials.

---

1. $\mathrm{softmax}(s - 51) = \mathrm{softmax}(s)$.
2. $\mathrm{softmax}(2s) = (1/10, 9/10)$.
3. A is correct. B changes the score gap; C changes one exponential relative to the other.

Blanchard, Higham & Higham (2021), §4 · Course example

EXERCISE · SLIDE 20 · CHECK THE RESULT

Stable attention and masks

A shift or a temperature?

A score row is $s = (0, \log 3)$. Which change preserves its softmax weights?

- **A** Subtract 5 from both entries ✓
- **B** Multiply both entries by 2
- **C** Add 5 only to the second entry

Blanchard, Higham & Higham (2021), §4 · Course example

EXERCISE · SLIDE 23 · QUESTION

Stable attention and masks

Mask before normalizing

Scores are $(0, \log 3)$; the second key is forbidden. Which output weights are correct?

A $(1, 0)$

B $(1/4, 0)$

C $(1/4, 3/4)$

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 23 · SOLUTION 1/2

Stable attention and masks

Mask before normalizing

Solution 1 / 2

Normalize only over permitted keys

s: score row; key 2 is unavailable.

---

1. $s + M = (0, -\infty)$.
2. $\exp(s + M) = (1, 0)$ and its sum is 1.
3. $p = (1, 0)$. A is correct.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 23 · SOLUTION 2/2

Stable attention and masks

Mask before normalizing

Solution 2 / 2

Why the other operations differ

p: normalized attention weights.

1. B zeroes a probability after normalization and leaves row sum 1/4.
2. C ignores the mask. Renormalizing allowed probabilities could work, but masking logits is direct.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 23 · CHECK THE RESULT

Stable attention and masks

Mask before normalizing

Scores are $(0, \log 3)$; the second key is forbidden. Which output weights are correct?

- **A** $(1, 0)$ ✓
- **B** $(1/4, 0)$
- **C** $(1/4, 3/4)$

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 25 · QUESTION

Stable attention and masks

Diagnose the NaN without changing the task

A padded query has no allowed key. Stable softmax returns NaN. Which repair preserves the declared information rule?

A. Use uniform weights over all keys

B. Mark the query invalid and output zero

C. Remove all masks for the batch

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 25 · SOLUTION 1/1

Stable attention and masks

Diagnose the NaN without changing the task

Solution 1 / 1

Separate invalid queries from valid predictions

M: mask; padded queries are excluded from the loss.

1. Detect a row with no permitted key before normalization.
2. For that invalid row, use zero weights/output and exclude its loss.
3. B is correct. A and C read forbidden entries.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 25 · CHECK THE RESULT

Stable attention and masks

Diagnose the NaN without changing the task

A padded query has no allowed key. Stable softmax returns NaN. Which repair preserves the declared information rule?

A. Use uniform weights over all keys

B. Mark the query invalid and output zero ✓

C. Remove all masks for the batch

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 28 · QUESTION

Stable attention and masks

Transfer: change the permitted information

$q = (\sqrt{2}, 0)^T$, $k_1 = (0, 1)^T$, $k_2 = (\log 3, 0)^T$; $v_1 = (2, 0)^T$, $v_2 = (0, 4)^T$. Only key 1 is allowed.

A $(1/2, 3)^T$

B $(2, 0)^T$

C $(0, 0)^T$

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 28 · SOLUTION 1/1

Stable attention and masks

Transfer: change the permitted information

Solution 1 / 1

Use the same scores with a different mask

s: scaled scores; p: allowed-key weights.

1. $s = (0, \log 3)$, but $s + M = (0, -\infty)$.
2. $p = (1, 0)$, so $y = 1v_1 + 0v_2 = (2, 0)^T$.
3. B is correct. A ignores masking; C discards a valid key.

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 28 · CHECK THE RESULT

Stable attention and masks

Transfer: change the permitted information

$q = (\sqrt{2}, 0)^T$, $k_1 = (0, 1)^T$, $k_2 = (\log 3, 0)^T$; $v_1 = (2, 0)^T$, $v_2 = (0, 4)^T$. Only key 1 is allowed.

A $(1/2, 3)^T$

B $(2, 0)^T$ ✓

C $(0, 0)^T$

Vaswani et al. (2017), §3.2.1, Eq. 1 · Course example

EXERCISE · SLIDE 32 · QUESTION

Multi-head attention

Trace a tensor through the heads

$B = 2$, $T = 3$, $D = 8$, $H = 2$. Find query and score shapes after splitting and transposing.

A. Q [2,2,3,4]; S [2,2,3,3]

B. Q [2,3,2,4]; S [2,3,2,2]

C. Q [2,2,3,8]; S [2,2,8,8]

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 32 · SOLUTION 1/2

Multi-head attention

Trace a tensor through the heads

Solution 1 / 2

Name each axis

B: batch; T: tokens; H: heads; $d_h = D/H$.

1. $d_h = 8/2 = 4$.
2. $Q : [B, H, T, d_h] = [2, 2, 3, 4]$.
3. $S : [B, H, T, T] = [2, 2, 3, 3]$.

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 32 · SOLUTION 2/2

Multi-head attention

Trace a tensor through the heads

Solution 2 / 2

Check the alternatives

S: query-by-key scores in each head.

1. A is correct. B omits the token/head transpose.
2. C fails to split the width and confuses features with token counts.

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 32 · CHECK THE RESULT

Multi-head attention

Trace a tensor through the heads

$B = 2$, $T = 3$, $D = 8$, $H = 2$. Find query and score shapes after splitting and transposing.

- **A** Q [2,2,3,4]; S [2,2,3,3] ✓
- **B** Q [2,3,2,4]; S [2,3,2,2]
- **C** Q [2,2,3,8]; S [2,2,8,8]

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 35 · QUESTION

Multi-head attention

Transfer: a different output map

$c = (1, 2, 3, -1)^T$, $W_O = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}$. Find $W_O c$.

A $(3, 2)^T$

B $(4, 1)^T$

C $(1, 2, 3, -1)^T$

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 35 · SOLUTION 1/1

Multi-head attention

Transfer: a different output map

Solution 1 / 1

Read the new map

c: concatenated heads; W_O: two output rows.

1. First output: $0 + 0 + 3 + 0 = 3$.
2. Second output: $1 + 0 + 0 - (-1) = 2$.
3. A is correct. B reuses the earlier map; C stops before projection.

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 35 · CHECK THE RESULT

Multi-head attention

Transfer: a different output map

$c = (1, 2, 3, -1)^T$, $W_O = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}$. Find $W_O c$.

- **A** $(3, 2)^T$ ✓
- **B** $(4, 1)^T$
- **C** $(1, 2, 3, -1)^T$

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 37 · QUESTION

Multi-head attention

Correct shapes can hide a wrong normalization

Scores have shape [B,H,Tq,Tk]. A programmer uses `softmax(S,dim=-2)`. What is normalized?

A. Keys for each query

B. Queries for each key

C. Features of each token

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 37 · SOLUTION 1/1

Multi-head attention

Correct shapes can hide a wrong normalization

Solution 1 / 1

Read the axis labels

$S_{b,h,i,j}$ compares query i with key j.

1. The second-last axis is i, the query index.
2. dim=−2 makes columns sum to one across queries.
3. B describes the bug. Correct attention uses dim=−1 for key normalization.

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 37 · CHECK THE RESULT

Multi-head attention

Correct shapes can hide a wrong normalization

Scores have shape [B,H,Tq,Tk]. A programmer uses `softmax(S,dim=-2)`. What is normalized?

A Keys for each query

B Queries for each key ✓

C Features of each token

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 38 · QUESTION

Multi-head attention

Transfer: broadcast a key-validity mask

Scores have shape [2,3,4,5]: batch, heads, queries, keys. Key validity has shape [2,5]. Which reshape shares the key rule across every head and query?

A [2,1,1,5]

B [2,5,1,1]

C [1,1,2,5]

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 38 · SOLUTION 1/1

Multi-head attention

Transfer: broadcast a key-validity mask

Solution 1 / 1

Align the batch and key axes

$B = 2$, $H = 3$, $T_q = 4$, $T_k = 5$.

1. Use singleton head and query axes: [2,1,1,5].
2. Broadcast those singleton axes to [2,3,4,5].
3. A is correct. B places validity on the head axis; C moves batch into the query axis.

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 38 · CHECK THE RESULT

Multi-head attention

Transfer: broadcast a key-validity mask

Scores have shape [2,3,4,5]: batch, heads, queries, keys. Key validity has shape [2,5]. Which reshape shares the key rule across every head and query?

- **A** [2,1,1,5] ✓
- **B** [2,5,1,1]
- **C** [1,1,2,5]

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 40 · QUESTION

Position and order

What does content alone fail to identify?

The same two feature vectors are presented in opposite orders to unmasked attention with no position features.

A. Dot products recover their original order

B. Outputs move with the input permutation

C. Every output becomes zero

Vaswani et al. (2017), §3.5 · Course example

EXERCISE · SLIDE 40 · SOLUTION 1/1

Position and order

What does content alone fail to identify?

Solution 1 / 1

Apply the stated symmetry

Π: swap of token positions.

1. $X \mapsto \Pi X$ implies $Y \mapsto \Pi Y$.
2. B is correct: relationships are preserved without an original-position label.
3. A invents missing information; C does not follow from equivariance.

Vaswani et al. (2017), §3.5 · Course example

EXERCISE · SLIDE 40 · CHECK THE RESULT

Position and order

What does content alone fail to identify?

The same two feature vectors are presented in opposite orders to unmasked attention with no position features.

A Dot products recover their original order

B Outputs move with the input permutation ✓

C Every output becomes zero

Vaswani et al. (2017), §3.5 · Course example

EXERCISE · SLIDE 43 · QUESTION

Position and order

Transfer: one sinusoidal pair at a new position

One coordinate pair is $(\sin(p\omega), \cos(p\omega))$. Take $\omega = \pi/2$ and move from p=0 to p=1.

A $(1,0)$

B $(0,1)$

C $(-1,0)$

Vaswani et al. (2017), §3.5 · Course example

EXERCISE · SLIDE 43 · SOLUTION 1/1

Position and order

Transfer: one sinusoidal pair at a new position

Solution 1 / 1

Evaluate in the stated coordinate order

$\omega = \pi/2$; the pair is sine first, cosine second.

1. At $p = 0$: $(\sin 0, \cos 0) = (0, 1)$.
2. At $p = 1$: $(\sin(\pi/2), \cos(\pi/2)) = (1, 0)$.
3. A is correct. B keeps the old position; C uses the wrong sine sign.

Vaswani et al. (2017), §3.5 · Course example

EXERCISE · SLIDE 43 · CHECK THE RESULT

Position and order

Transfer: one sinusoidal pair at a new position

One coordinate pair is $(\sin(p\omega), \cos(p\omega))$. Take $\omega = \pi/2$ and move from p=0 to p=1.

- **A** $(1,0)$ ✓
- B $(0,1)$
- C $(-1,0)$

Vaswani et al. (2017), §3.5 · Course example

EXERCISE · SLIDE 47 · QUESTION

Position and order

Transfer: shift both positions

$q = k = (1, 0)^T$, $\omega = \pi/2$. Change (p, s) from $(0, 1)$ to $(2, 3)$. Find the raw dot product.

A 0

B 1

C -1

Su et al., RoFormer (2021), §3.2 · Course example

EXERCISE · SLIDE 47 · SOLUTION 1/1

Position and order

Transfer: shift both positions

Solution 1 / 1

Use the relative angle

p, s: query/key positions; content unchanged.

---

1. $(s - p)\omega = (3 - 2)\pi/2 = \pi/2.$
2. $q^T R(\pi/2)k = 0.$
3. A is correct. B uses zero offset; C uses a half-turn.

Su et al., RoFormer (2021), §3.2 · Course example

EXERCISE · SLIDE 47 · CHECK THE RESULT

Position and order

Transfer: shift both positions

$q = k = (1, 0)^T$, $\omega = \pi/2$. Change (p, s) from $(0, 1)$ to $(2, 3)$. Find the raw dot product.

- **A** 0 ✓
- **B** 1
- **C** -1

Su et al., RoFormer (2021), §3.2 · Course example

EXERCISE · SLIDE 50 · QUESTION

Transformer blocks

Transfer: normalize one token

One token is $x = (1, 3)$. Use $\gamma = (1, 1)$, $\beta = (0, 0)$ and $\epsilon = 0$ for this nonconstant numerical example.

A $(-1, 1)$

B $(-1/\sqrt{2},\ 1/\sqrt{2})$

C $(1, 3)$

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 50 · SOLUTION 1/2

Transformer blocks

Transfer: normalize one token

Solution 1 / 2

Use the feature mean and mean squared deviation

$D = 2$ features; no batch or token averaging.

---

1. $\mu = (1 + 3)/2 = 2$; $x - \mu = (-1, 1)$.
2. $v = [(-1)^2 + 1^2]/2 = 1$.
3. $\hat{x} = (-1, 1)/\sqrt{1} = (-1, 1)$. A is correct.

Xiong et al. (2020), §2, Fig. 1 · Course example

Transformer blocks

Transfer: normalize one token

Solution 2 / 2

Diagnose the alternative statistics

$\gamma = 1$ and $\beta = 0$ preserve the normalized values.

1. B uses the sample-variance divisor D−1; LayerNorm uses D.
2. C omits centering and scaling. In real implementations keep positive epsilon.

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 50 · CHECK THE RESULT

Transformer blocks

Transfer: normalize one token

One token is $x = (1, 3)$. Use $\gamma = (1, 1)$, $\beta = (0, 0)$ and $\epsilon = 0$ for this nonconstant numerical example.

- **A** $(-1, 1)$ ✓
- **B** $(-1/\sqrt{2},\ 1/\sqrt{2})$
- **C** $(1, 3)$

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 52 · QUESTION

Transformer blocks

A residual connection constrains the shape

X has shape [2,5,12], while the attention branch returns [2,5,8]. Can they be directly added?

A. Yes, ordinary broadcasting repairs it

B. No, the branch must return width 12

C. Yes, because both have five tokens

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 52 · SOLUTION 1/1

Transformer blocks

A residual connection constrains the shape

Solution 1 / 1

Compare corresponding coordinates

X: residual input; branch: learned update.

1. Feature widths 12 and 8 differ and neither is one.
2. Ordinary broadcasting cannot align these feature coordinates.
3. B is correct: use the output projection to return width 12.

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 52 · CHECK THE RESULT

Transformer blocks

A residual connection constrains the shape

X has shape [2,5,12], while the attention branch returns [2,5,8]. Can they be directly added?

A Yes, ordinary broadcasting repairs it

B No, the branch must return width 12 ✓

C Yes, because both have five tokens

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 53 · QUESTION

Transformer blocks

Transfer: a missing path changes the architecture

$Y = X + FFN(MHA(X))$ is described as the pre-LN block. What is missing?

A. Only position information

B. Both normalizations and the intermediate residual

C. A final softmax only

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 53 · SOLUTION 1/1

Transformer blocks

Transfer: a missing path changes the architecture

Solution 1 / 1

Compare the computation graph

U: intermediate residual representation.

1. $U = X + MHA(LN(X))$.
2. $Y = U + FFN(LN(U))$.
3. B is correct. Position features or final softmax do not restore these missing paths.

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 53 · CHECK THE RESULT

Transformer blocks

Transfer: a missing path changes the architecture

$Y = X + FFN(MHA(X))$ is described as the pre-LN block. What is missing?

A Only position information

B Both normalizations and the intermediate residual ✓

C A final softmax only

Xiong et al. (2020), §2, Fig. 1 · Course example

EXERCISE · SLIDE 56 · QUESTION

Cost and the KV cache

Transfer: a longer sequence

Keep B, H and datatype fixed. Double T. How does the materialized score-tensor memory change?

A ×2

B ×4

C Unchanged

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 56 · SOLUTION 1/1

Cost and the KV cache

Transfer: a longer sequence

Solution 1 / 1

Count both token axes

S has shape [B,H,T,T].

1. $BH(2T)^2 = 4BHT^2$.
2. B is correct. A counts only the extra queries and forgets the extra keys.
3. C confuses fixed feature width with a fixed pair count.

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 56 · CHECK THE RESULT

Cost and the KV cache

Transfer: a longer sequence

Keep B, H and datatype fixed. Double T. How does the materialized score-tensor memory change?

A ×2

B ×4 ✓

C Unchanged

Vaswani et al. (2017), §3.2.2 · Course example

EXERCISE · SLIDE 59 · QUESTION

Cost and the KV cache

Transfer: can this memory be reused?

Camera tokens are replaced by a new frame. Can their old cross-attention keys and values serve as exact representations of the new frame?

A Yes, when shapes match

B No, recompute the changed memory

C Yes, if action horizon is fixed

Shazeer (2019), §2, incremental attention · Course example

EXERCISE · SLIDE 59 · SOLUTION 1/1

Cost and the KV cache

Transfer: can this memory be reused?

Solution 1 / 1

Trace projections back to their inputs

E: encoded camera tokens; K,V are their projections.

---

1. $K = EW_K^T$, $V = EW_V^T$.
2. A new frame can change E and both projections.
3. B is correct. Matching shape or horizon does not imply matching values.

Shazeer (2019), §2, incremental attention · Course example

EXERCISE · SLIDE 59 · CHECK THE RESULT

Cost and the KV cache

Transfer: can this memory be reused?

Camera tokens are replaced by a new frame. Can their old cross-attention keys and values serve as exact representations of the new frame?

A Yes, when shapes match

B No, recompute the changed memory ✓

C Yes, if action horizon is fixed

Shazeer (2019), §2, incremental attention · Course example

EXERCISE · SLIDE 61 · QUESTION

Cost and the KV cache

Transfer: count incremental comparisons

Generate four tokens from an empty prefix. Count query-key comparisons for one head, including each current token. Compare cached decoding with full-prefix recomputation.

A Cached 10; recomputed 30

B Cached 4; recomputed 16

C Cached 16; recomputed 16

Shazeer (2019), §2, incremental attention · Course example

EXERCISE · SLIDE 61 · SOLUTION 1/1

Cost and the KV cache

Transfer: count incremental comparisons

Solution 1 / 1

Add work over all four decoding steps

$t = 1, 2, 3, 4$; one new query uses t keys.

---

1. Cached: $1 + 2 + 3 + 4 = 10$.
2. Full-prefix reruns: $1^2 + 2^2 + 3^2 + 4^2 = 30$.
3. A is correct. B drops past keys and repeated passes; C counts only one final full-sequence pass.

Shazeer (2019), §2, incremental attention · Course example

EXERCISE · SLIDE 61 · CHECK THE RESULT

Cost and the KV cache

Transfer: count incremental comparisons

Generate four tokens from an empty prefix. Count query-key comparisons for one head, including each current token. Compare cached decoding with full-prefix recomputation.

- **A** Cached 10; recomputed 30 ✓
- **B** Cached 4; recomputed 16
- **C** Cached 16; recomputed 16

Shazeer (2019), §2, incremental attention · Course example

EXERCISE · SLIDE 64 · QUESTION

Action chunking

Chunk length is not execution length

A policy predicts K=6 future actions and executes E=2 before reading a new observation. Which targets from the current chunk are executed?

A. Times t through t+5

B. Times t and t+1

C. Only time t+5

Zhao et al. (2023), §IV-A · Course example

EXERCISE · SLIDE 64 · SOLUTION 1/1

Action chunking

Chunk length is not execution length

Solution 1 / 1

Separate prediction and execution

$K = 6$: predicted rows; $E = 2$: executed prefix length.

1. The chunk covers target times t,…,t+5.
2. Its first two rows target t and t+1.
3. B is correct. A confuses K with E; C chooses the last row rather than the prefix.

Zhao et al. (2023), §IV-A · Course example

EXERCISE · SLIDE 64 · CHECK THE RESULT

Action chunking

Chunk length is not execution length

A policy predicts K=6 future actions and executes E=2 before reading a new observation. Which targets from the current chunk are executed?

A. Times t through t+5

B. Times t and t+1 ✓

C. Only time t+5

Zhao et al. (2023), §IV-A · Course example

EXERCISE · SLIDE 67 · QUESTION

Temporal ensembling

Transfer: three estimates for one target

Oldest-to-newest predictions for target s: 1, 3, 5. Let $m = \log 2$. Find the ensemble.

A 3

B 15/7

C 27/7

Zhao et al. (2023), §IV-A · Course example

EXERCISE · SLIDE 67 · SOLUTION 1/2

Temporal ensembling

Transfer: three estimates for one target

Solution 1 / 2

Weight, sum, normalize

$j = 0, 1, 2$: oldest-to-newest; all predictions target s.

1. $w = (1, 1/2, 1/4)$, with sum $7/4$.
2. Weighted sum $= 1 + 3/2 + 5/4 = 15/4$.
3. $\bar{a}_s = (15/4)/(7/4) = 15/7$.

Zhao et al. (2023), §IV-A · Course example

EXERCISE · SLIDE 67 · SOLUTION 2/2

Temporal ensembling

Transfer: three estimates for one target

Solution 2 / 2

Interpret the result

$\bar{a}_s$: ensemble action.

1. B is correct. A uses equal weights.
2. C reverses the weights and favors the newest prediction.

Zhao et al. (2023), §IV-A · Course example

EXERCISE · SLIDE 67 · CHECK THE RESULT

Temporal ensembling

Transfer: three estimates for one target

Oldest-to-newest predictions for target s: 1, 3, 5. Let $m = \log 2$. Find the ensemble.

A 3

B 15/7 ✓

C 27/7

Zhao et al. (2023), §IV-A · Course example

EXERCISE · SLIDE 72 · QUESTION

ACT: observations to actions

Transfer: encoder tokens and action slots

A model has 20 observation tokens and predicts a chunk of 6 actions. Each action query reads every observation token. What is one head’s cross-attention score shape?

A. 20×20

B. 6×20

C. 6×6

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 72 · SOLUTION 1/1

ACT: observations to actions

Transfer: encoder tokens and action slots

Solution 1 / 1

Identify query and key collections

$K = 6$: action queries; $T_o = 20$: observation keys.

1. $Q_a : [6, d_k]$, $K_o^T : [d_k, 20]$.
2. $Q_a K_o^T : [6, 20]$.
3. B is correct. A and C replace cross-attention with self-attention on one collection.

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 72 · CHECK THE RESULT

ACT: observations to actions

Transfer: encoder tokens and action slots

A model has 20 observation tokens and predicts a chunk of 6 actions. Each action query reads every observation token. What is one head’s cross-attention score shape?

A. 20×20

B. 6×20 ✓

C. 6×6

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 76 · QUESTION

ACT: latent-variable training

Log-variance is not the standard deviation

One latent coordinate has $\mu = 1$, $\ell = \log \sigma^2 = \log 4$, and $\epsilon = -1$. Find z.

A -3

B -1

C 1

Kingma & Welling (2014), §2, Appendix B · Course example

EXERCISE · SLIDE 76 · SOLUTION 1/1

ACT: latent-variable training

Log-variance is not the standard deviation

Solution 1 / 1

Convert the parameter before sampling

ℓ: log-variance; σ: standard deviation.

1. $\sigma^2 = e^{\ell} = 4$, so $\sigma = e^{\ell/2} = 2$.
2. $z = \mu + \sigma\epsilon = 1 + 2(-1) = -1$.
3. B is correct. A uses variance as scale; C ignores the sampled noise.

Kingma & Welling (2014), §2, Appendix B · Course example

EXERCISE · SLIDE 76 · CHECK THE RESULT

ACT: latent-variable training

Log-variance is not the standard deviation

One latent coordinate has $\mu = 1$, $\ell = \log \sigma^2 = \log 4$, and $\epsilon = -1$. Find z.

A -3

B -1 ✓

C 1

Kingma & Welling (2014), §2, Appendix B · Course example

EXERCISE · SLIDE 79 · QUESTION

ACT: latent-variable training

Transfer: which divergence creates the gap?

$\log p_\theta(A \mid o) = \mathcal{L}_{ELBO} + G$. Which G makes this an exact identity?

A $D_{KL}(q \| p_\theta(z \mid A, o))$

B $-D_{KL}(q \| p(z))$

C The action prediction error alone

Kingma & Welling (2014), §2, Appendix B · Course example

EXERCISE · SLIDE 79 · SOLUTION 1/1

ACT: latent-variable training

Transfer: which divergence creates the gap?

Solution 1 / 1

Distinguish two different distributions

q: encoder; posterior: conditioned on the observed target.

1. $G = D_{KL}(q \| \text{true posterior}) \geq 0$.
2. The prior KL is already inside $\mathcal{L}_{ELBO}$ with a minus sign.
3. A is correct. B confuses regularization with the posterior approximation gap.

Kingma & Welling (2014), §2, Appendix B · Course example

EXERCISE · SLIDE 79 · CHECK THE RESULT

ACT: latent-variable training

Transfer: which divergence creates the gap?

$\log p_\theta(A \mid o) = \mathcal{L}_{ELBO} + G$. Which G makes this an exact identity?

- **A** $D_{KL}(q \| p_\theta(z \mid A, o))$ ✓
- **B** $-D_{KL}(q \| p(z))$
- **C** The action prediction error alone

Kingma & Welling (2014), §2, Appendix B · Course example

EXERCISE · SLIDE 84 · QUESTION

ACT: latent-variable training

Transfer: combine the two loss terms

Four action errors are $1, -1, 0, 2$. A one-dimensional encoder has $\mu = 1$, $\sigma^2 = 1$. Use mean L1 and $\beta = 2$. Find total loss.

A 1

B 2

C 5

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 84 · SOLUTION 1/2

ACT: latent-variable training

Transfer: combine the two loss terms

Solution 1 / 2

Reconstruct with the stated reduction

$Kd_a = 4$ action coordinates; errors are prediction minus target.

---

1. $\mathcal{L}_{rec} = (|1| + |-1| + |0| + |2|)/4 = 1.$
2. $D_{KL} = \frac{1}{2}(1 + 1 - 1 - \log 1) = 1/2.$

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 84 · SOLUTION 2/2

ACT: latent-variable training

Transfer: combine the two loss terms

Solution 2 / 2

Combine without changing the reduction

$\beta = 2$; mean reconstruction error equals 1.

1. $\mathcal{L} = 1 + 2(1/2) = 2.$
2. B is correct. A drops the KL term; C uses summed rather than mean L1.

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 84 · CHECK THE RESULT

ACT: latent-variable training

Transfer: combine the two loss terms

Four action errors are $1, -1, 0, 2$. A one-dimensional encoder has $\mu = 1$, $\sigma^2 = 1$. Use mean L1 and $\beta = 2$. Find total loss.

A 1

B 2 ✓

C 5

[Zhao et al. (2023), §IV-B–C, Appendix C](#) · Course example

EXERCISE · SLIDE 86 · QUESTION

ACT: deployment

Training and inference use different information

Which input route matches the reported ACT inference procedure?

A. Encode the unknown future target chunk

B. Current observations plus z=0

C. Draw z from the demonstration posterior

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 86 · SOLUTION 1/1

ACT: deployment

Training and inference use different information

Solution 1 / 1

Identify what is available now

o_t: current observations; the future demonstration is unavailable.

1. Discard the demonstration-conditioned latent encoder.
2. Set z to the standard-prior mean, zero, and run the decoder.
3. B is correct. A and C require future target information.

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 86 · CHECK THE RESULT

ACT: deployment

Training and inference use different information

Which input route matches the reported ACT inference procedure?

A. Encode the unknown future target chunk

B. Current observations plus z=0 ✓

C. Draw z from the demonstration posterior

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 88 · QUESTION

ACT: deployment

Transfer: trace the live policy loop

At target time s, two chunks predict actions 2 and 4, oldest first. Use $m = \log 2$. A fresh image arrives after execution. What is the executed target and the next computation?

A. Execute 8/3; encode the fresh observation

B. Execute 3; change β

C. Execute 4; reuse all old image features

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 88 · SOLUTION 1/2

ACT: deployment

Transfer: trace the live policy loop

Solution 1 / 2

Compute the same-time ensemble

$r_0 < r_1$; both predictions target s.

1. $w = (1, 1/2)$; normalize to $(2/3, 1/3)$.
2. $\bar{a}_s = (2/3)2 + (1/3)4 = 8/3$.

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 88 · SOLUTION 2/2

ACT: deployment

Transfer: trace the live policy loop

Solution 2 / 2

Close the feedback loop

o_{s+1}: fresh observations; inference latent remains zero.

1. Encode o_{s+1} and predict $\hat{A}_{s+1} = f_\theta(o_{s+1}, 0)$.
2. A is correct. B changes the ensemble and a training coefficient. C discards the stated weighting and fresh evidence.

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 88 · CHECK THE RESULT

ACT: deployment

Transfer: trace the live policy loop

At target time s, two chunks predict actions 2 and 4, oldest first. Use $m = \log 2$. A fresh image arrives after execution. What is the executed target and the next computation?

- **A** Execute 8/3; encode the fresh observation ✓
- **B** Execute 3; change β
- **C** Execute 4; reuse all old image features

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 89 · QUESTION

ACT: deployment

Transfer: a changed scene after prediction

A chunk predicts six actions. After one executed action, the object moves. Which response uses the new observation?

A. Keep executing all old targets without recomputing

B. Encode the new observation and predict a new chunk

C. Change the KL coefficient at inference

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 89 · SOLUTION 1/1

ACT: deployment

Transfer: a changed scene after prediction

Solution 1 / 1

Locate the input that changed

o: observation; learned model parameters remain fixed.

1. Object motion changes the current observation.
2. Re-encoding it can change attention memory and future predicted actions.
3. B uses that new evidence. A is open-loop; C changes a training parameter rather than reading the scene.

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 89 · CHECK THE RESULT

ACT: deployment

Transfer: a changed scene after prediction

A chunk predicts six actions. After one executed action, the object moves. Which response uses the new observation?

A. Keep executing all old targets without recomputing

B. Encode the new observation and predict a new chunk ✓

C. Change the KL coefficient at inference

Zhao et al. (2023), §IV-B–C, Appendix C · Course example

EXERCISE · SLIDE 92 · QUESTION

Discrete tokens and continuous actions

Transfer: the retained set depends on mass

Sorted token probabilities are $(0.5, 0.3, 0.2)$ and top-p threshold $\rho = 0.7$. Which distribution is sampled?

A $(0.5, 0.3, 0.2)$

B $(0.625, 0.375, 0)$

C $(1, 0, 0)$

Holtzman et al. (2020), §3.1 · Course example

EXERCISE · SLIDE 92 · SOLUTION 1/1

Discrete tokens and continuous actions

Transfer: the retained set depends on mass

Solution 1 / 1

Find the smallest sufficient prefix

$\rho = 0.7$: cumulative-mass threshold.

1. $0.5 < 0.7$, while $0.5 + 0.3 = 0.8 \geq 0.7$.
2. Retain two tokens and divide by 0.8: $(5/8, 3/8, 0)$.
3. B is correct. A omits truncation; C stops before reaching the threshold.

Holtzman et al. (2020), §3.1 · Course example

EXERCISE · SLIDE 92 · CHECK THE RESULT

Discrete tokens and continuous actions

Transfer: the retained set depends on mass

Sorted token probabilities are $(0.5, 0.3, 0.2)$ and top-p threshold $\rho = 0.7$. Which distribution is sampled?

A $(0.5, 0.3, 0.2)$

B $(0.625, 0.375, 0)$ ✓

C $(1, 0, 0)$

Holtzman et al. (2020), §3.1 · Course example