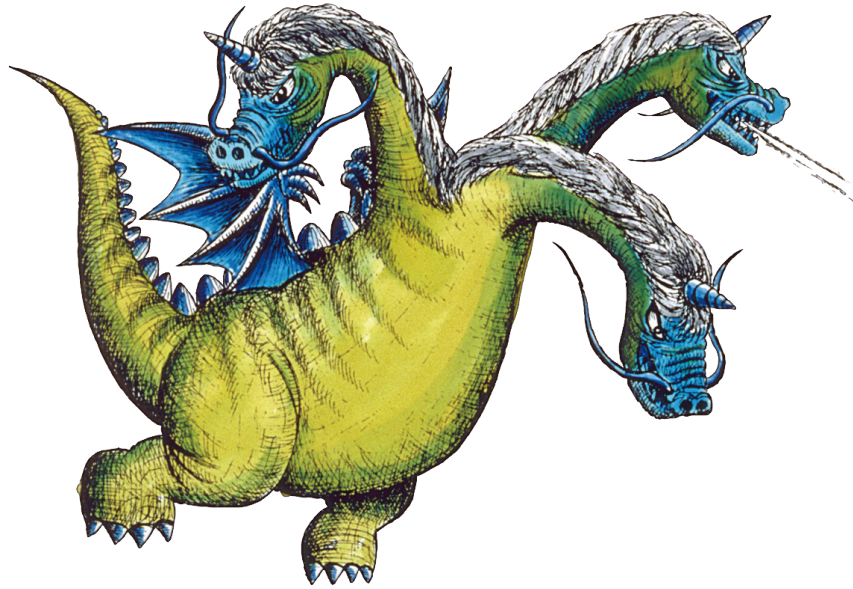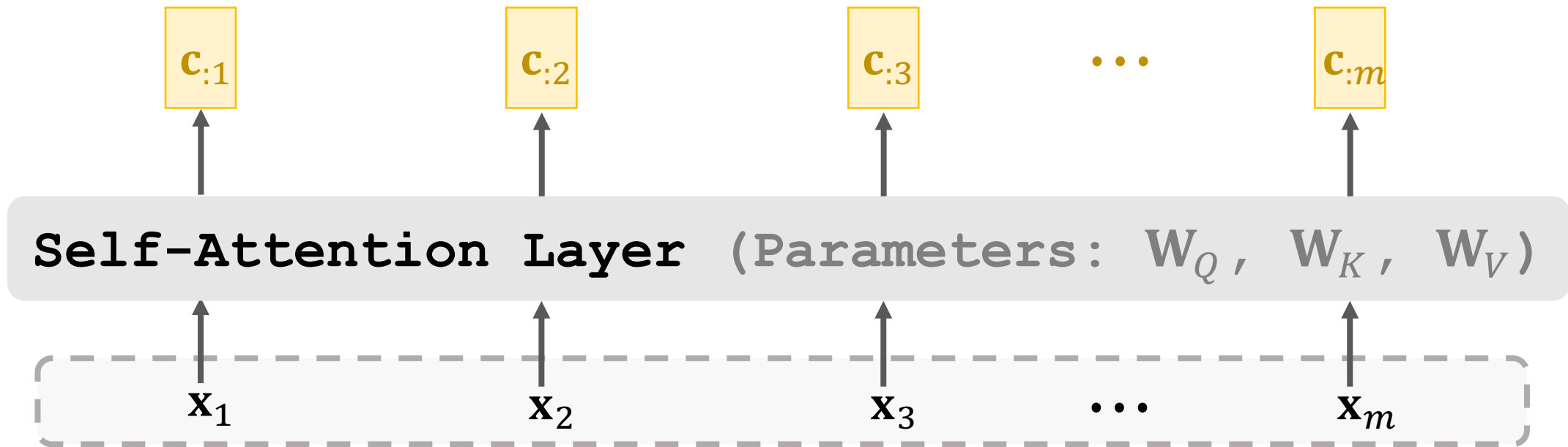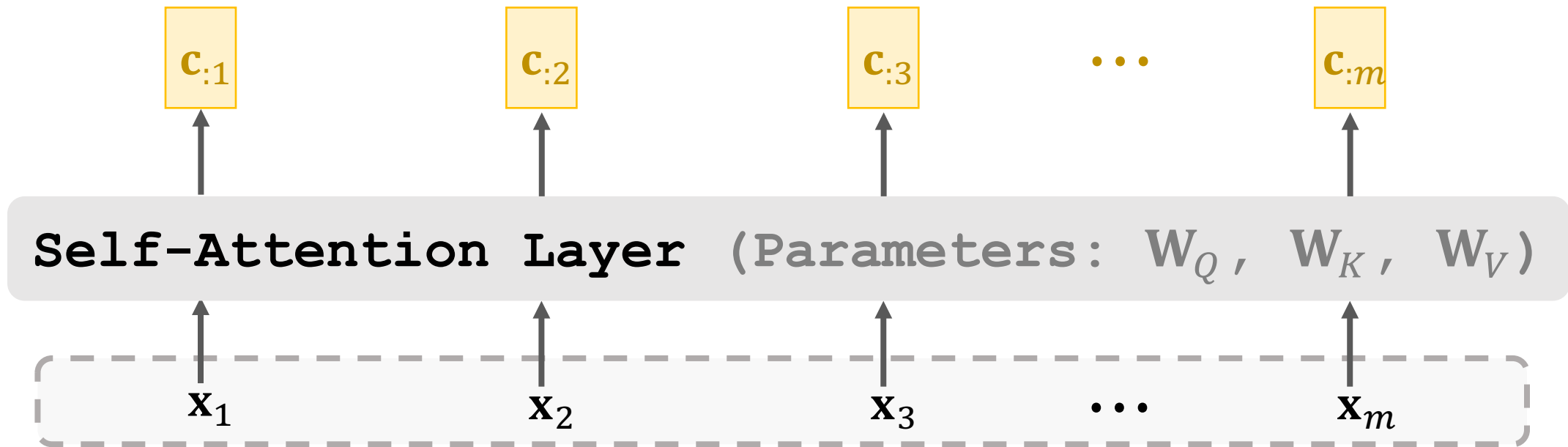# Multi-Head Attention

Single-Head Self-Attention

- Self-attention layer: $\mathbf{C} = \text{Attn}(\mathbf{X}, \mathbf{X})$.

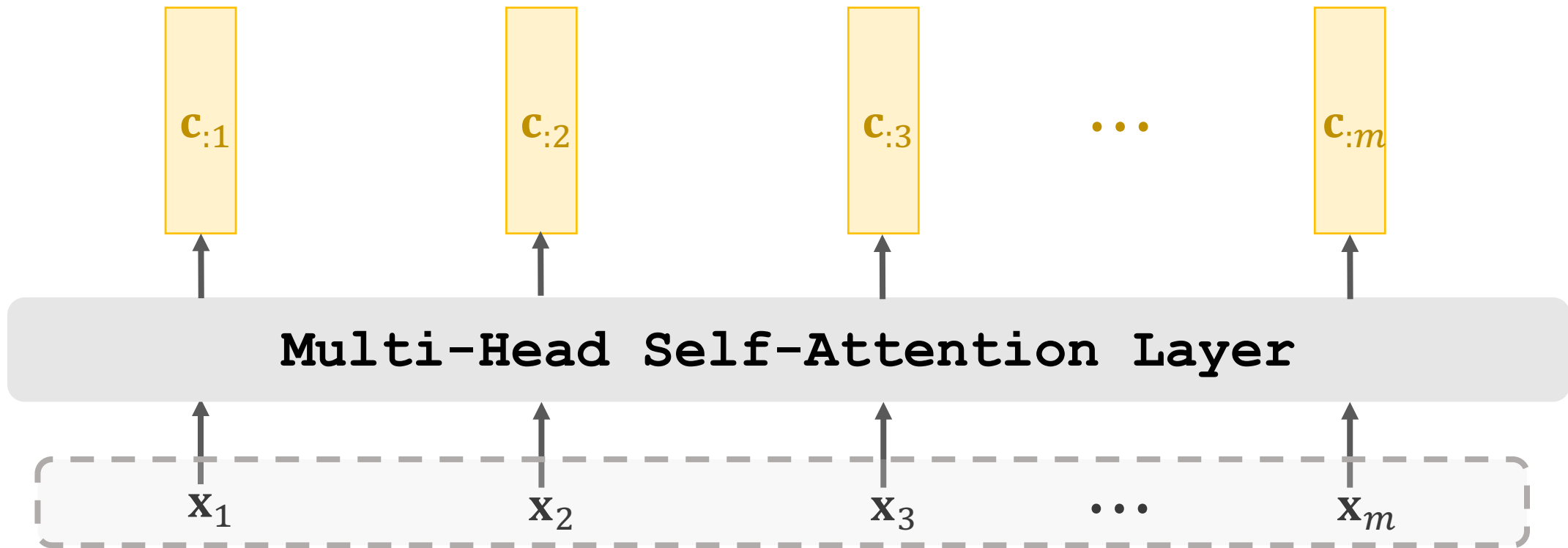- This is called "single-head self-attention".

Multi-Head Self-Attention

- Using l single-head self-attentions (which do not share parameters.)
  - A single-head self-attention has 3 parameter matrices: $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$.
 - Totally $3l$ parameters matrices.

Multi-Head Self-Attention

- Using l single-head self-attentions (which do not share parameters.)

- Concatenating outputs of single-head self-attentions.
  - Suppose single-head self-attentions' outputs are $d{\times}m$ matrices.
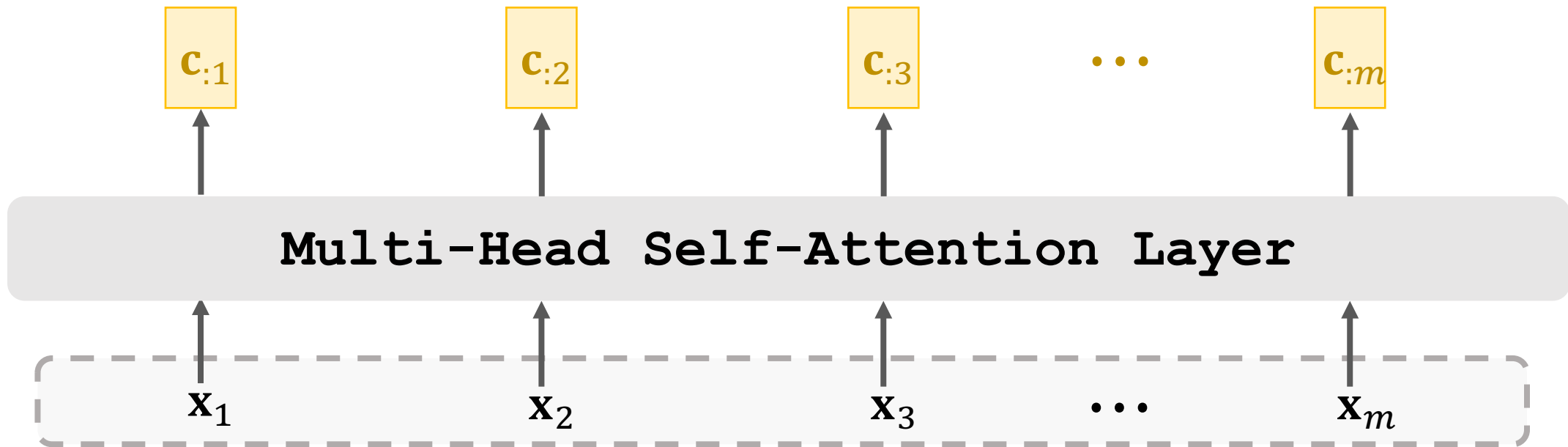 - Multi-head's output shape: $(ld){\times}m$.

Multi-Head Attention

- Using l single-head attentions (which do not share parameters.)
- Concatenating single-head attentions' outputs.
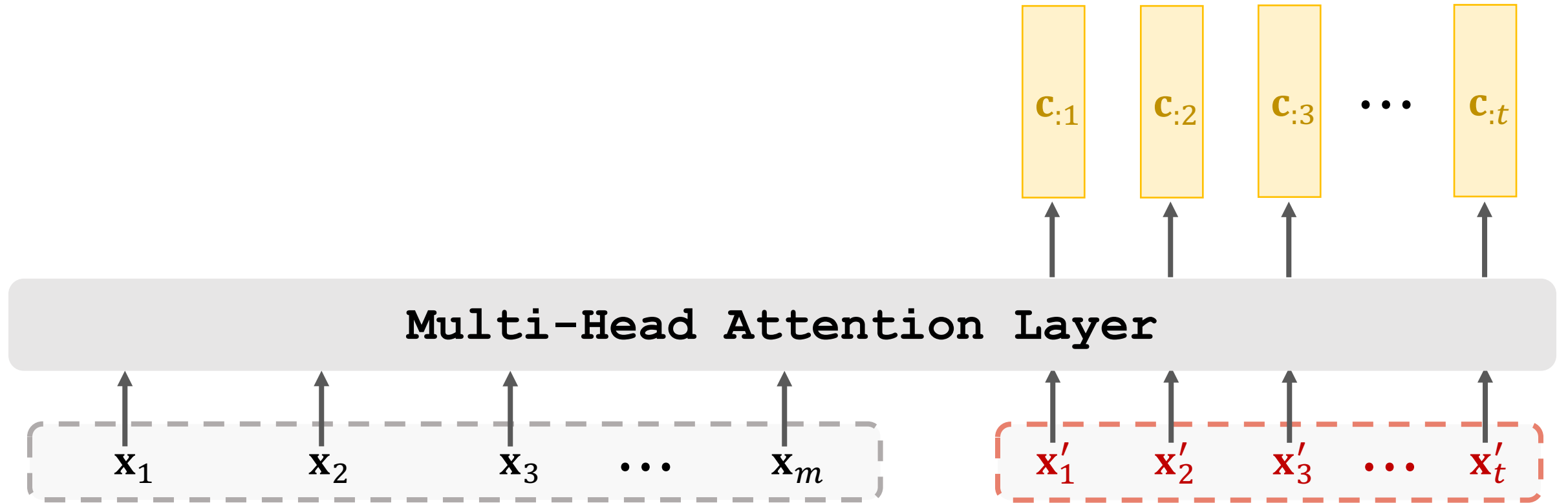
Stacked Self-Attention Layers

Self-Attention Layer
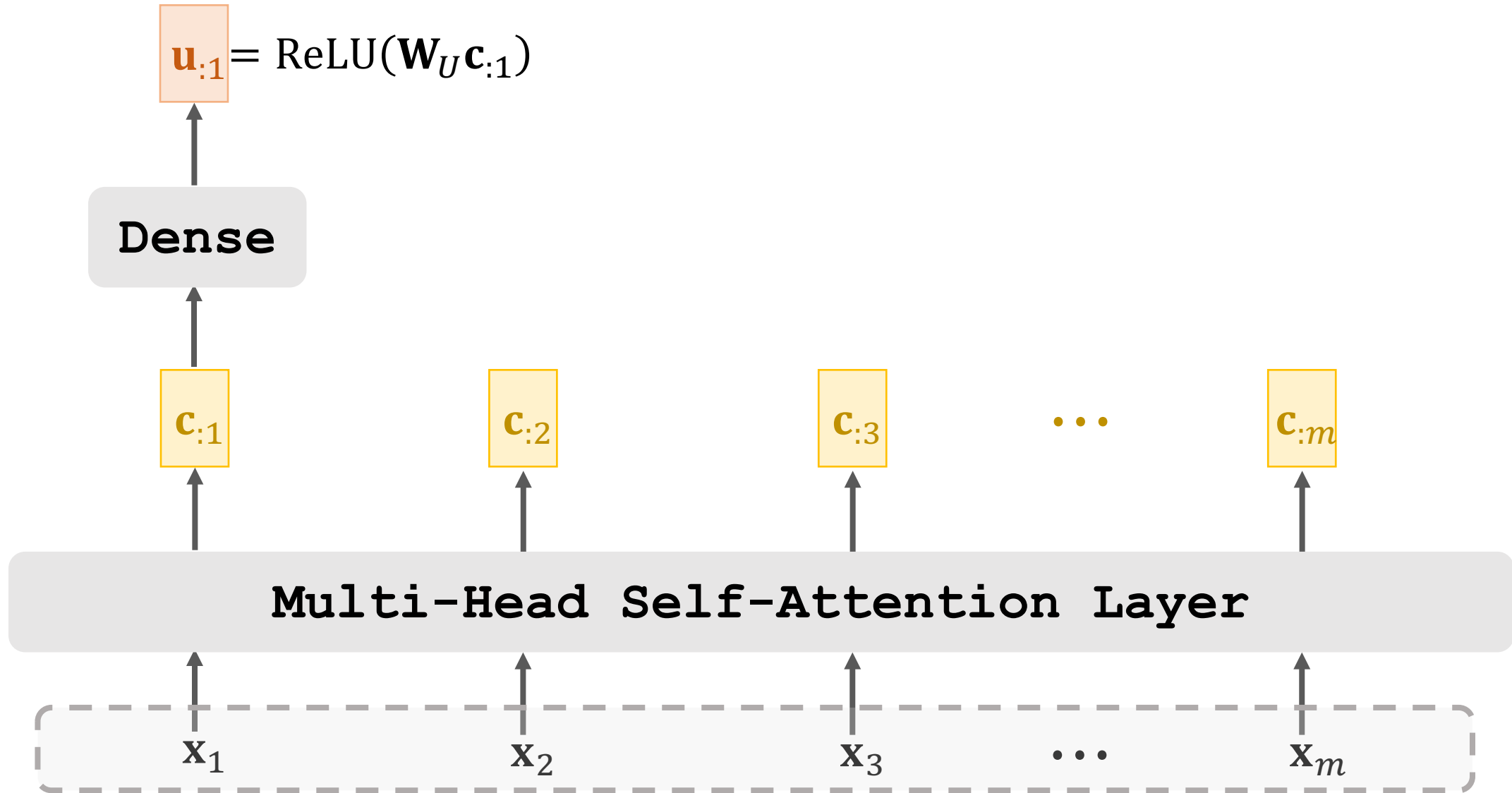
Self-Attention Layer + Dense Layer

$$\mathbf{u}_{:1} = \mathrm{ReLU}(\mathbf{W}_U \mathbf{c}_{:1})$$

Dense

$\mathbf{c}_{:1}$ $\mathbf{c}_{:2}$ $\mathbf{c}_{:3}$ $\cdots$ $\mathbf{c}_{:m}$

Multi-Head Self-Attention Layer

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\cdots$ $\mathbf{x}_m$

Self-Attention Layer + Dense Layer
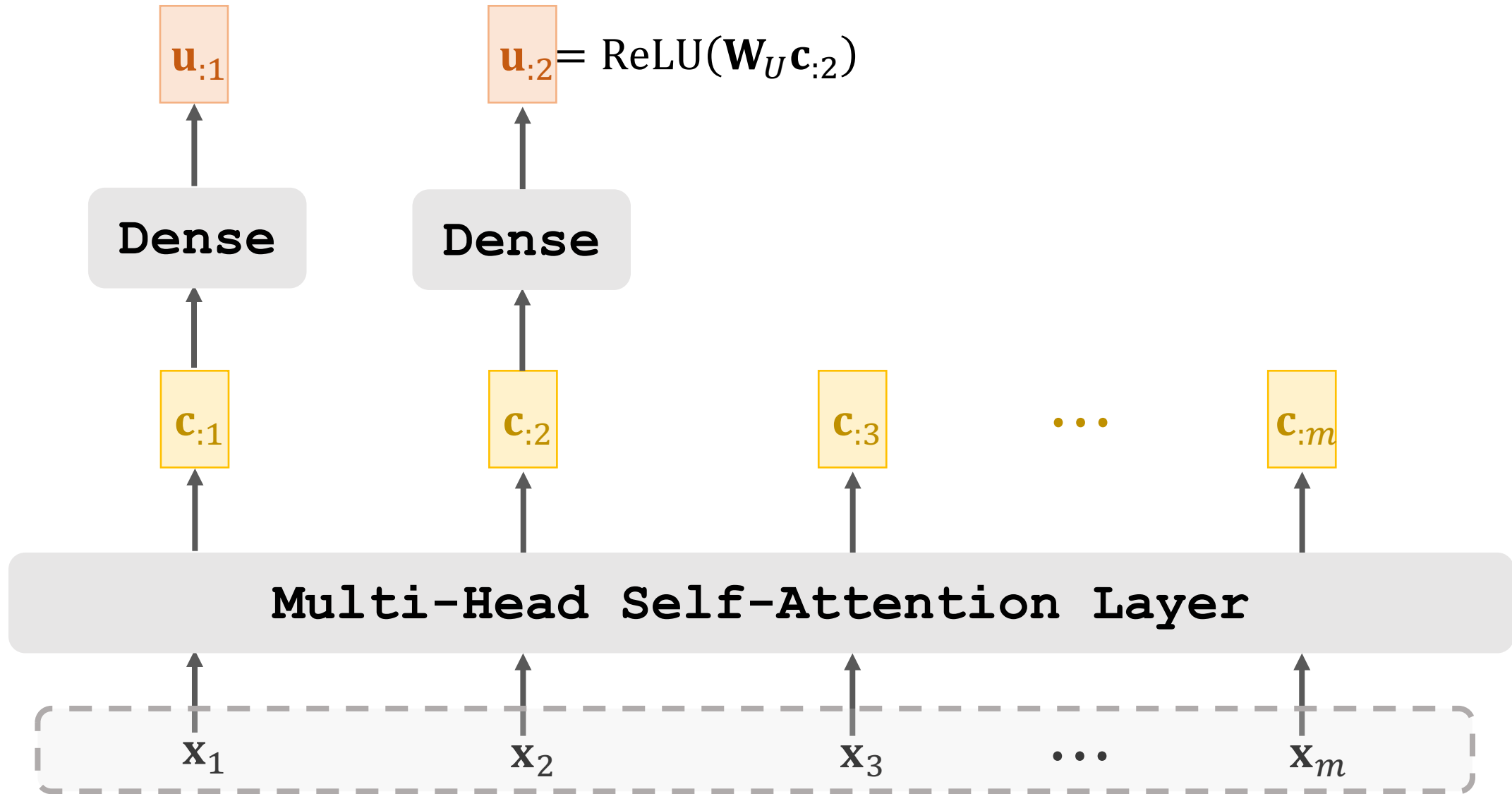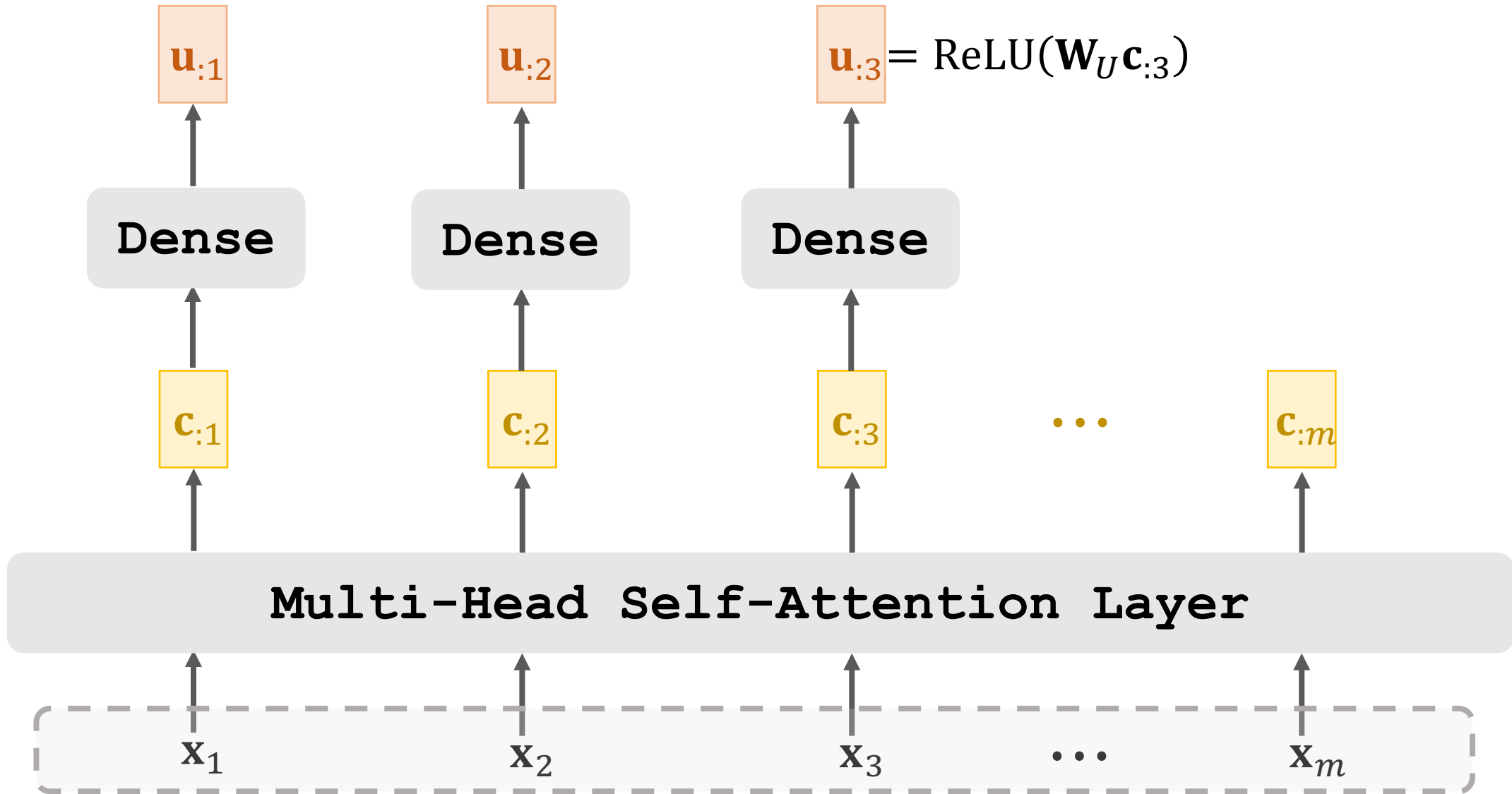
Self-Attention Layer + Dense Layer

Self-Attention Layer + Dense Layer

$$\mathbf{u}_{:m} = \mathrm{ReLU}(\mathbf{W}_U \mathbf{c}_{:m})$$
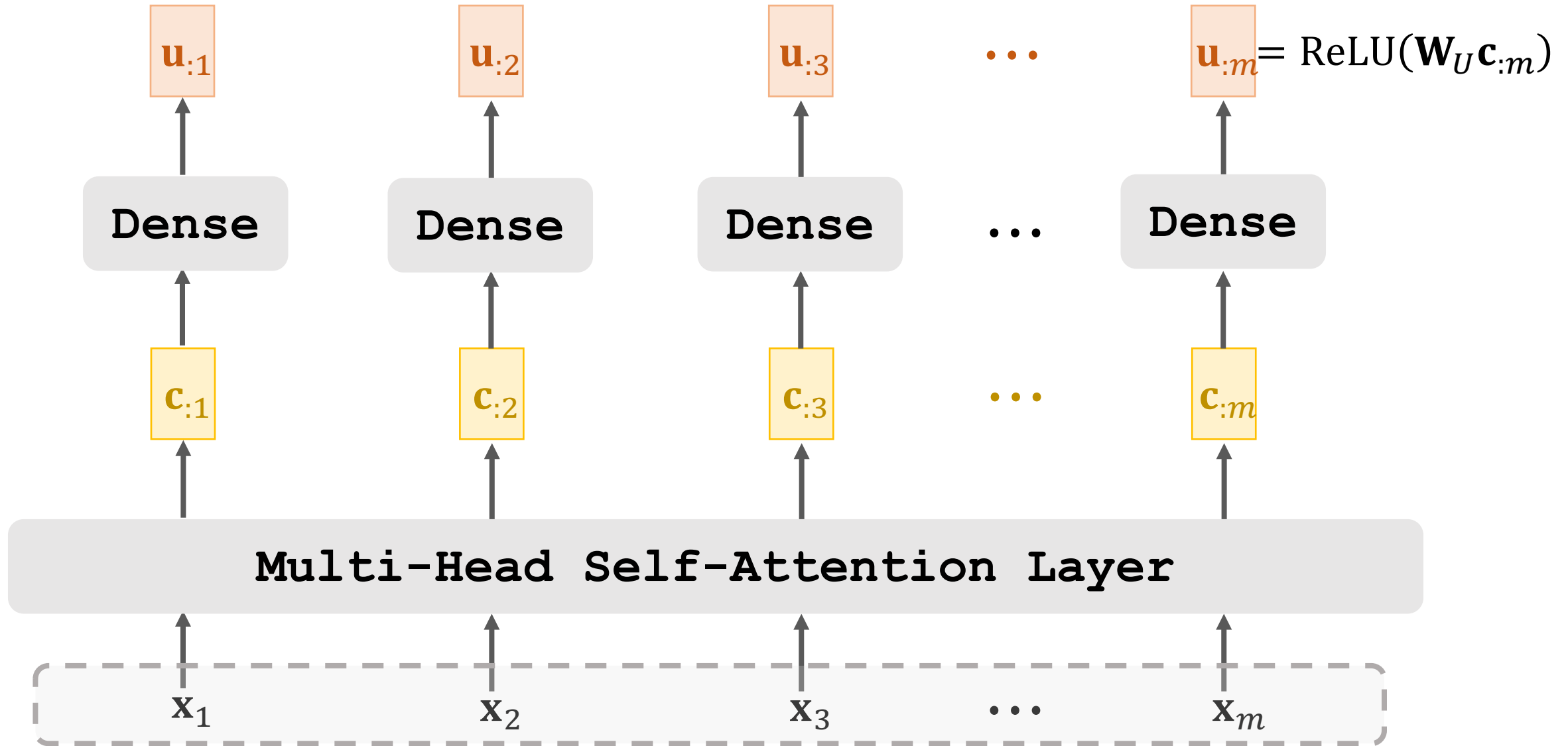
Stacked Self-Attention Layers
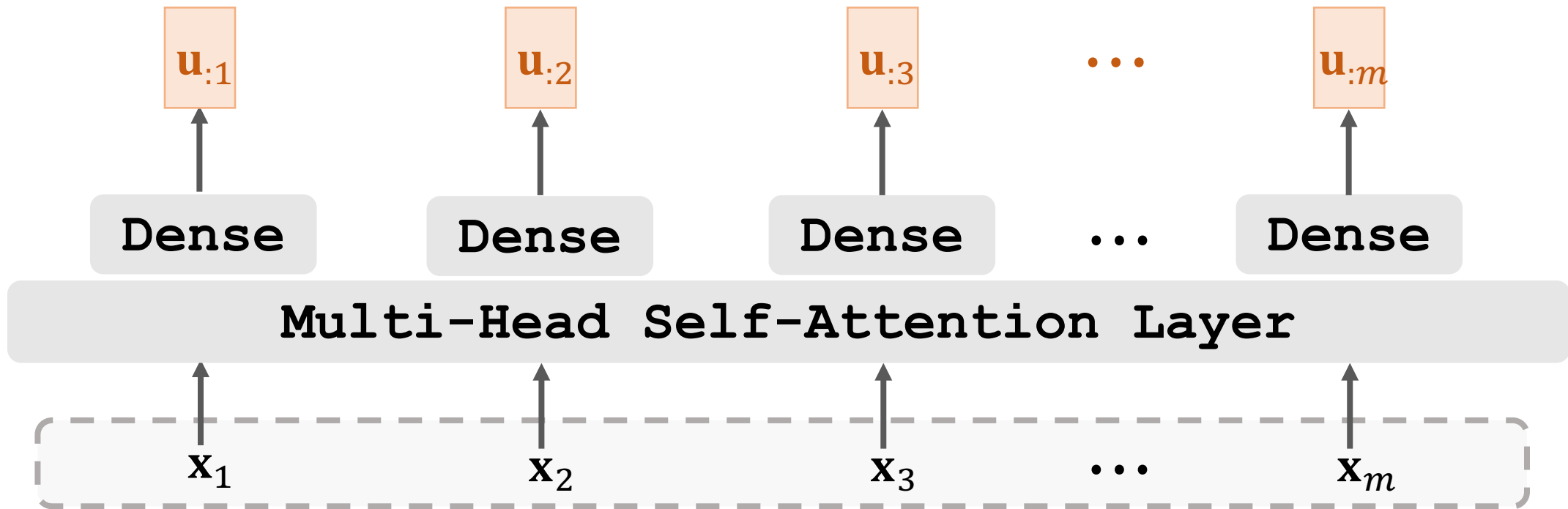
Stacked Self-Attention Layers
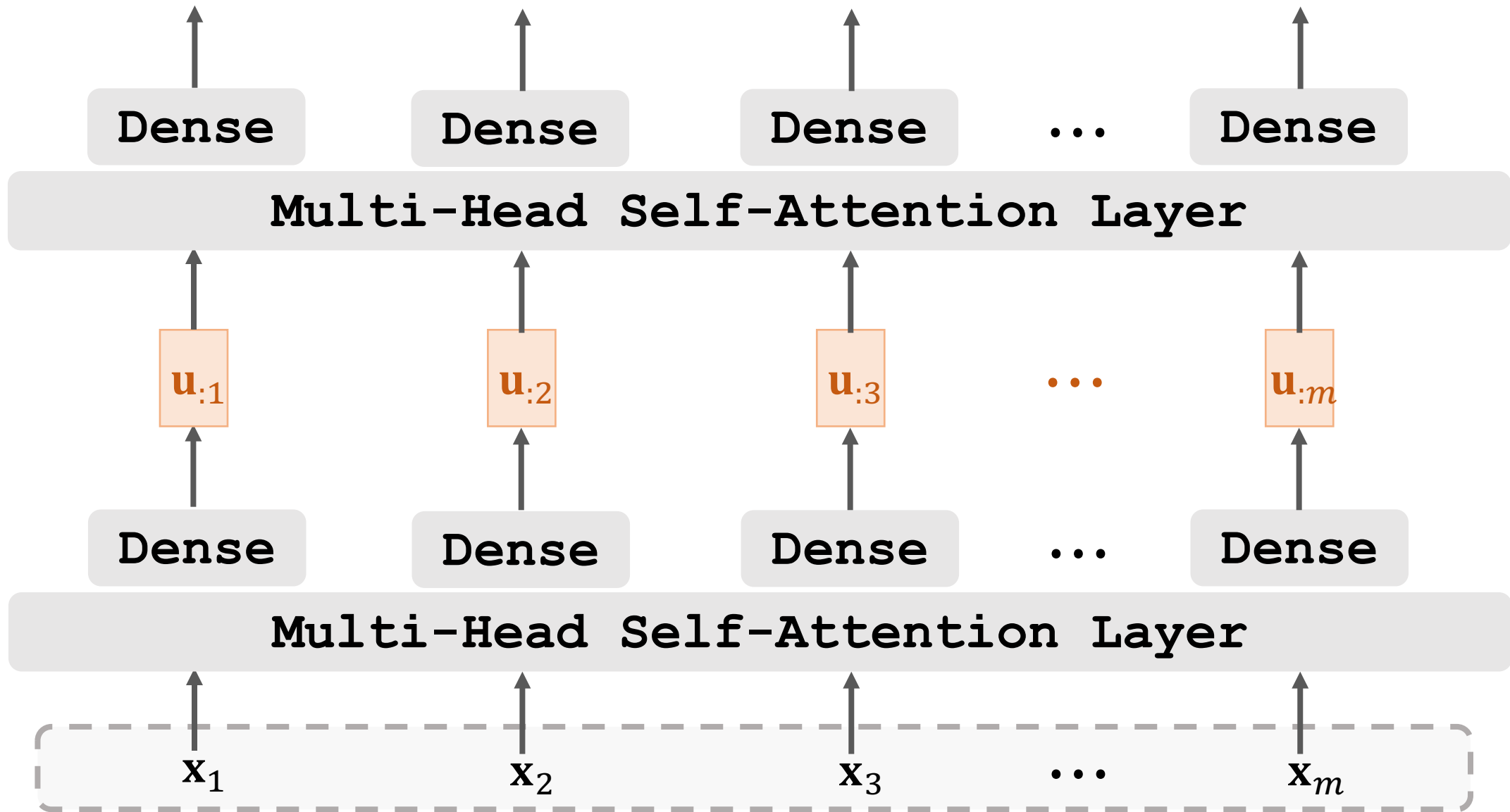
Transformer's Encoder
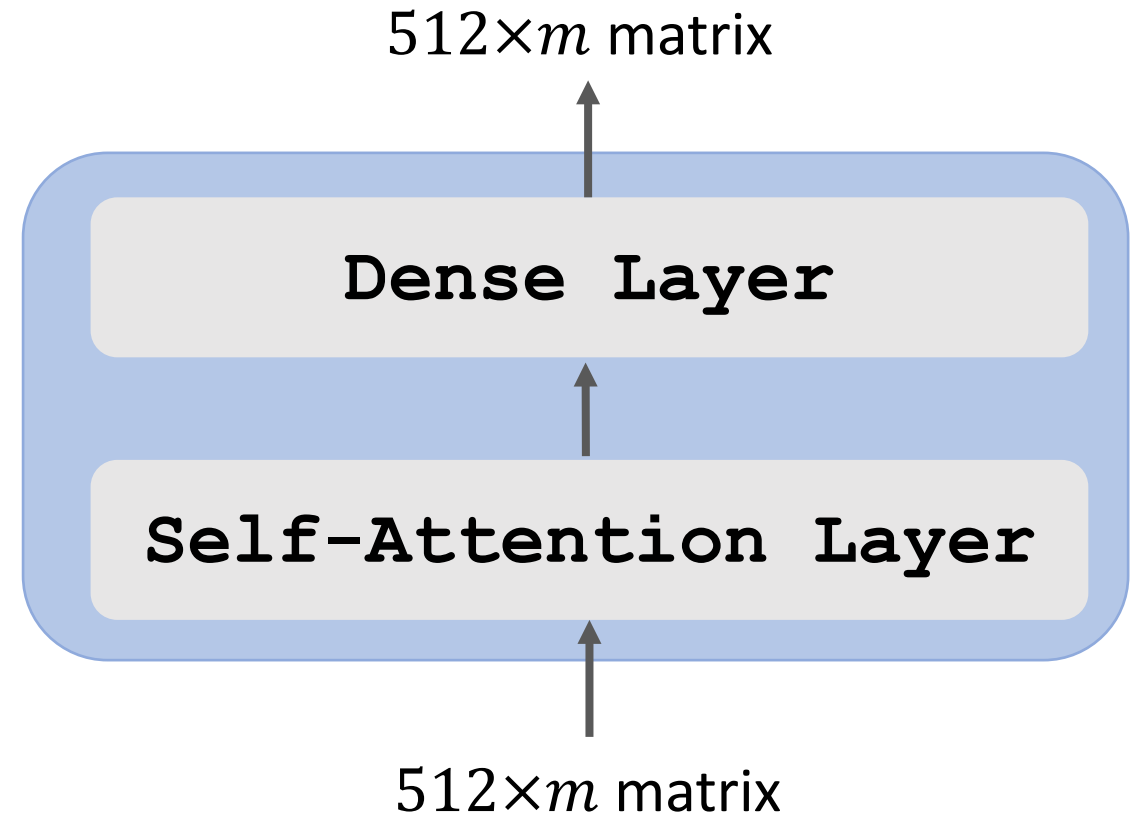
$512 \times m$ matrix

Dense Layer

Self-Attention Layer

$512 \times m$ matrix

One Block of Encoder

Transformer's Encoder

$512 \times m$ matrix

Block 1

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

$512 \times m$ matrix

Dense Layer

Self-Attention Layer

$512 \times m$ matrix

One Block of Encoder

Transformer's Encoder

Block 2

Block 1

$512 \times m$ matrix

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

$512 \times m$ matrix

Dense Layer

Self-Attention Layer

$512 \times m$ matrix

One Block of Encoder

Transformer's Encoder

$512 \times m$ matrix

Block 6

⋮

Block 2

Block 1

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

$512 \times m$ matrix

Dense Layer

Self-Attention Layer

$512 \times m$ matrix
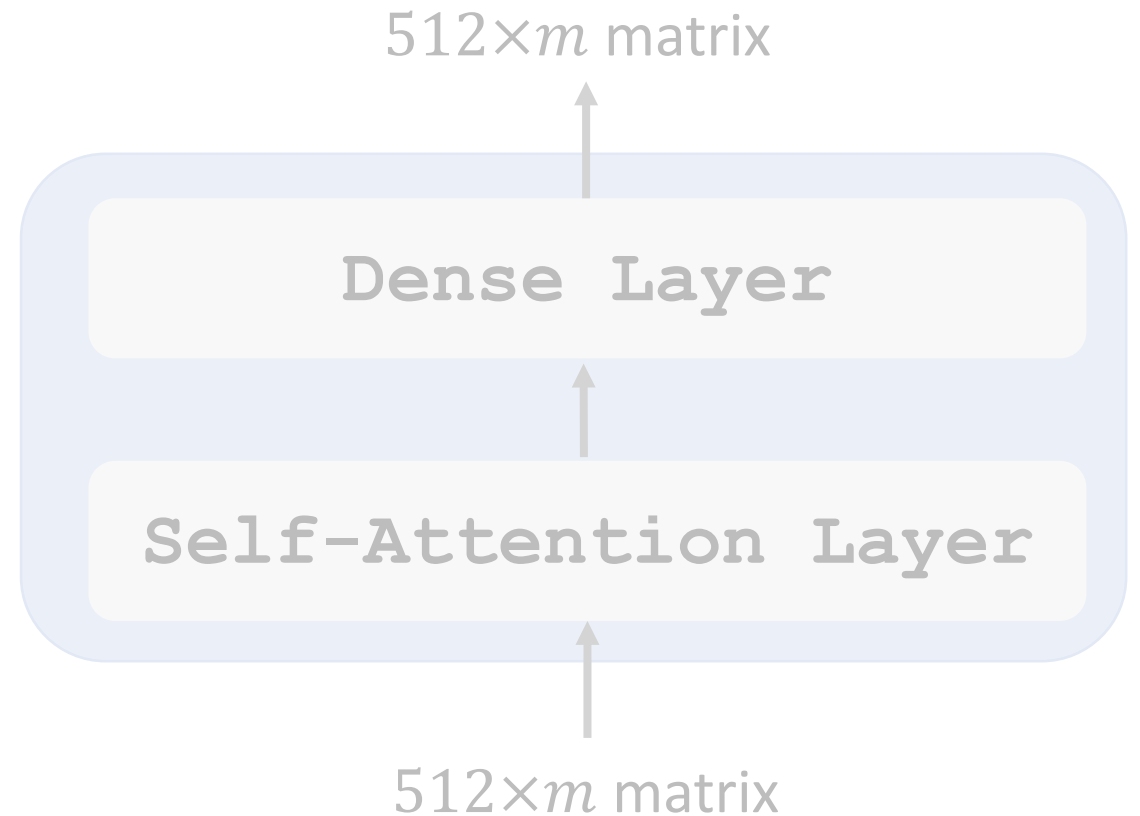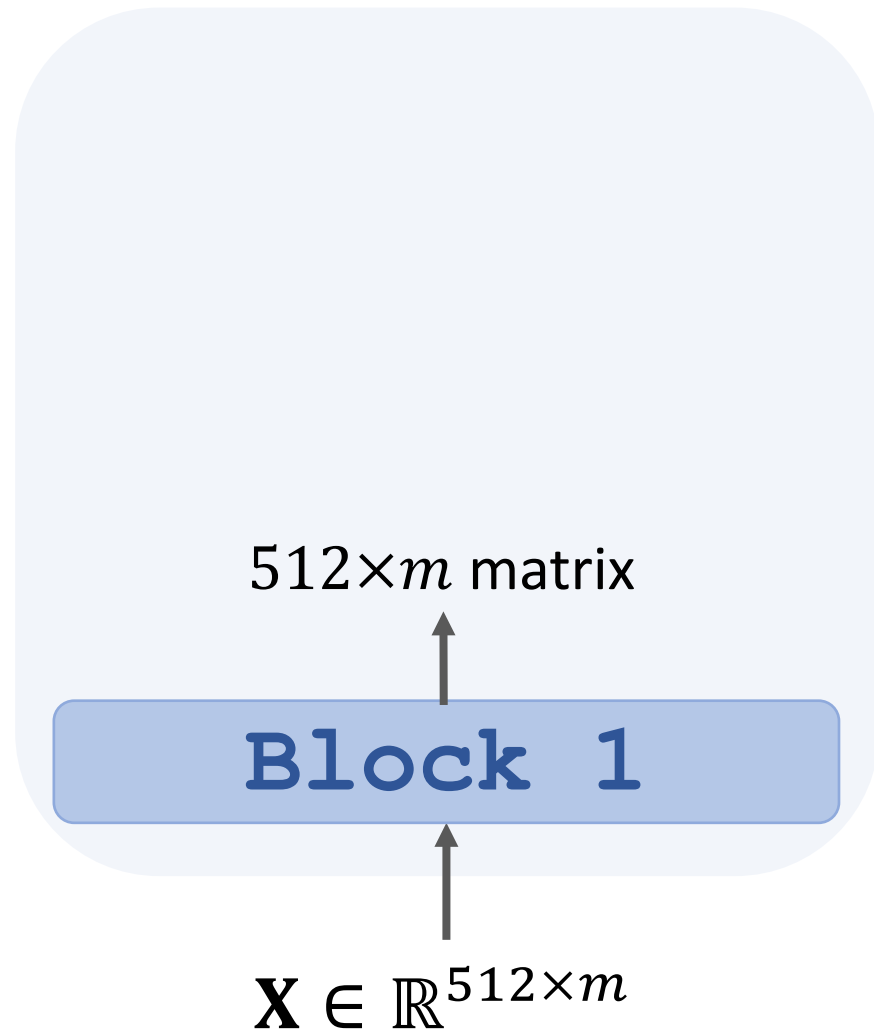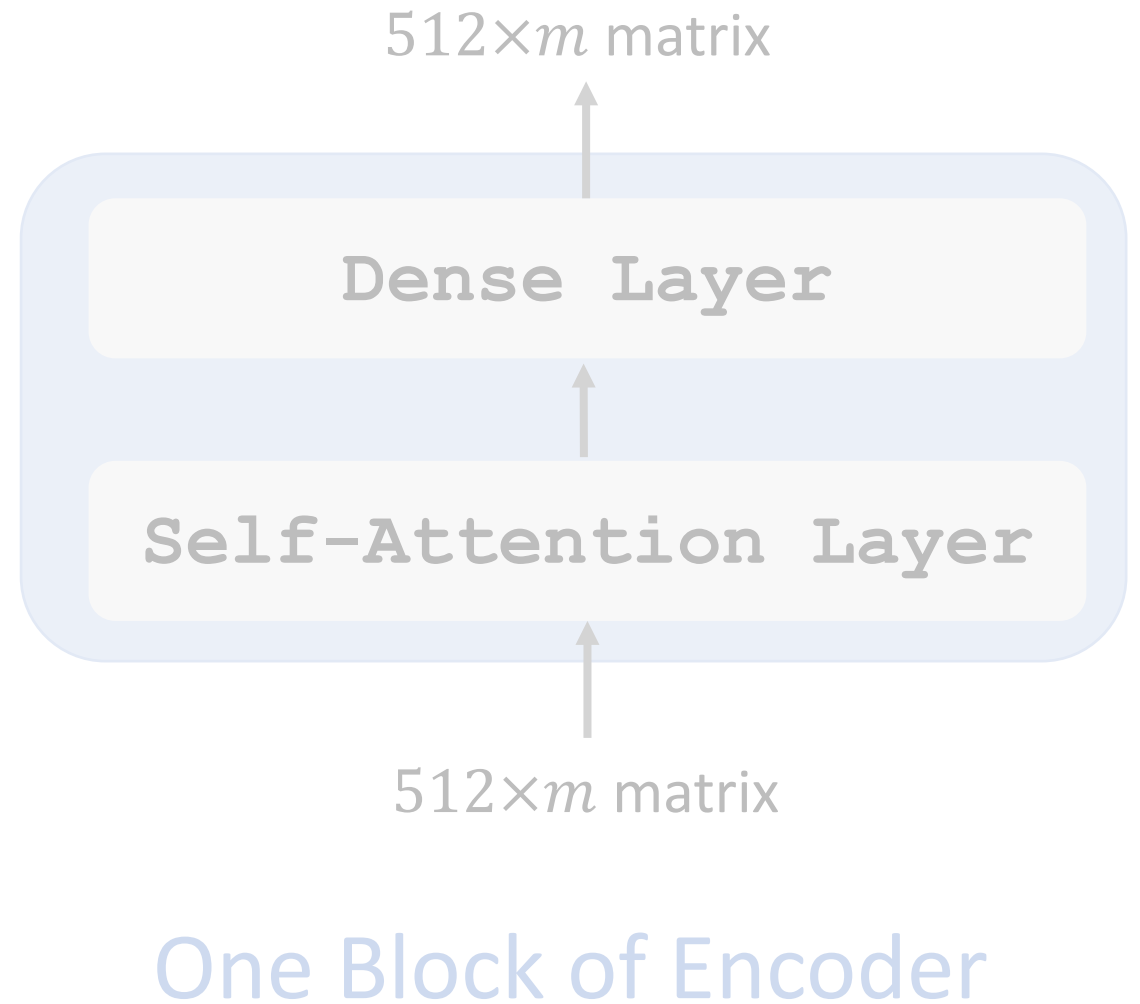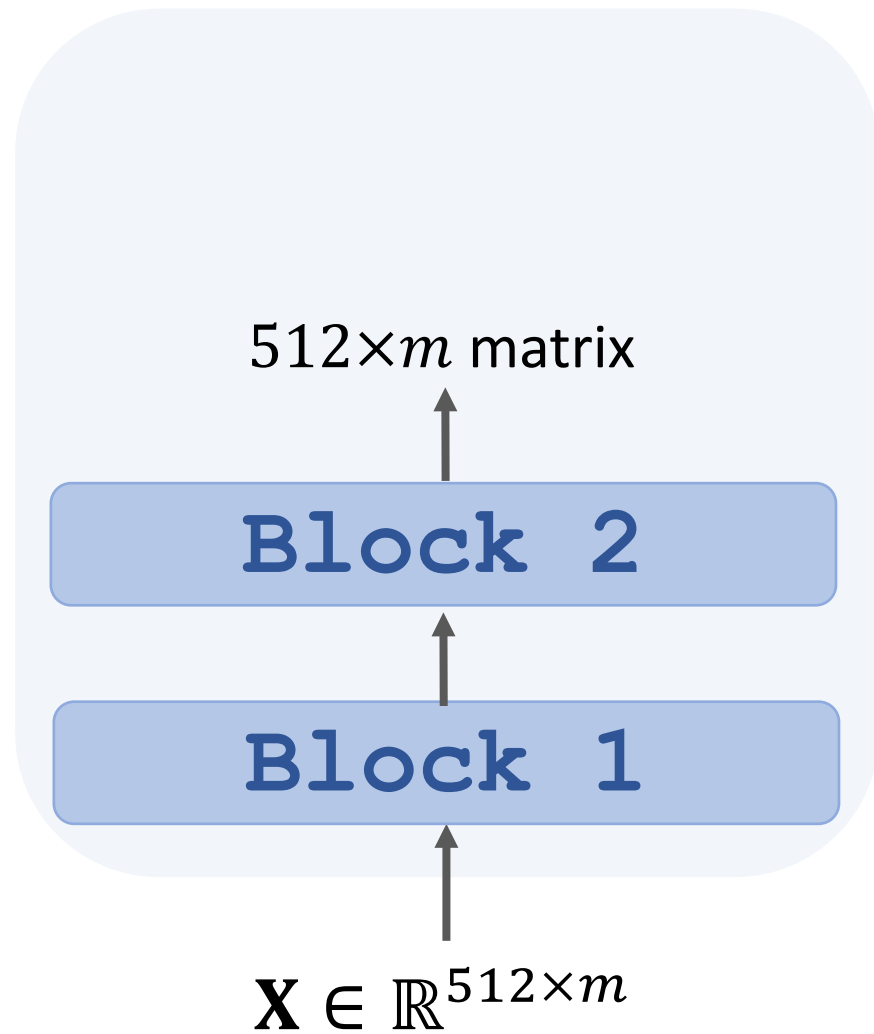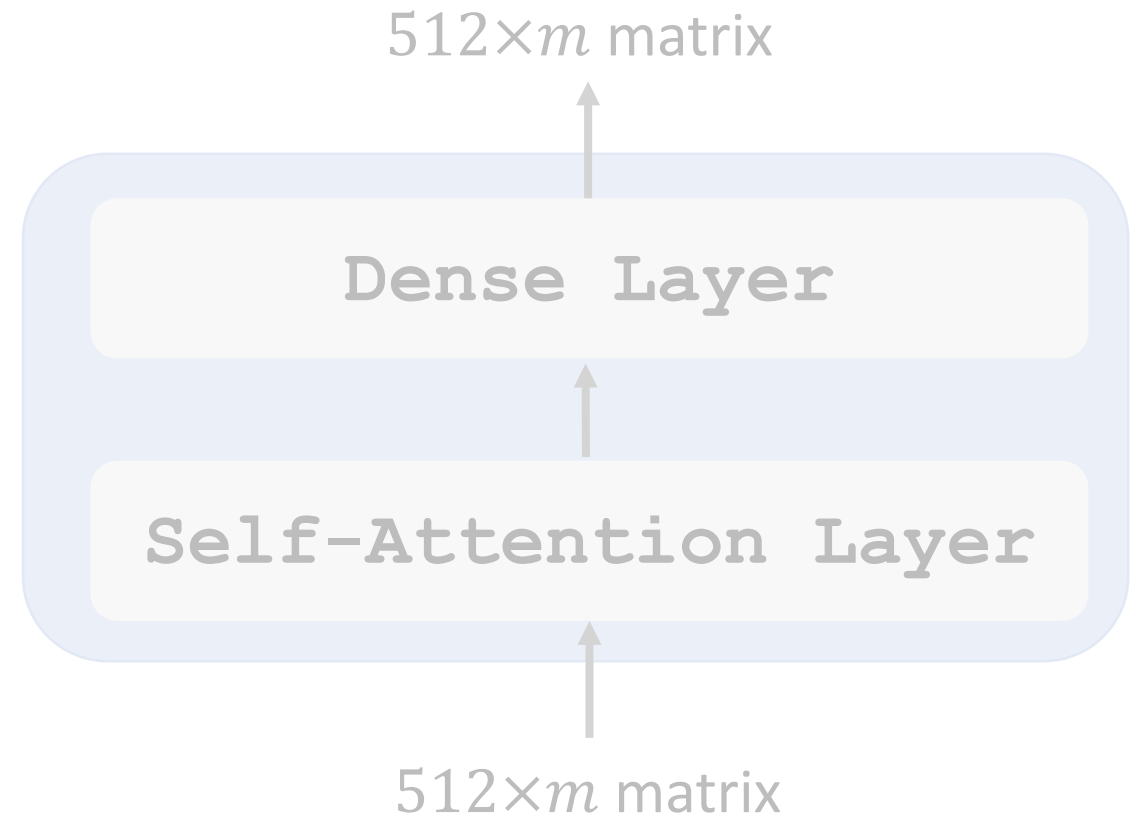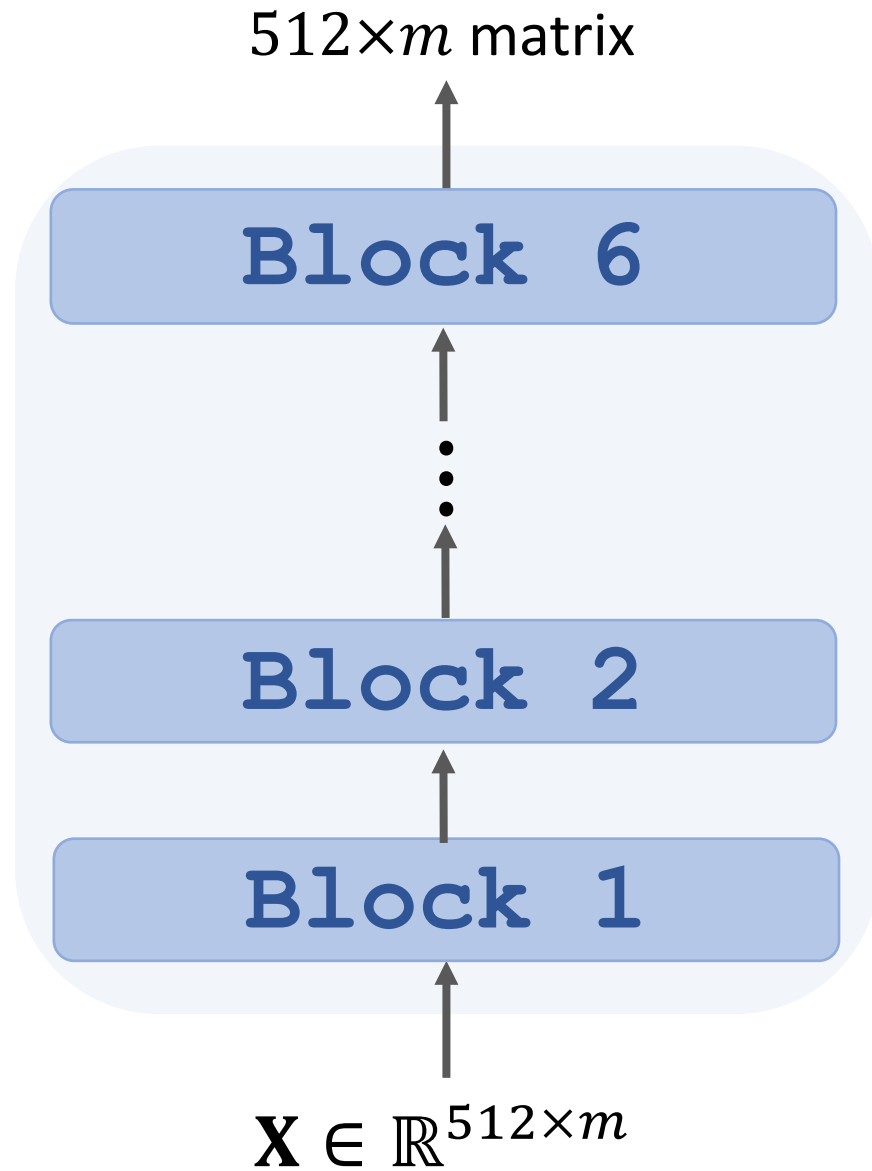
One Block of Encoder

Stacked Attention Layers

Stacked Attentions

- Transformer is a Seq2Seq model (encoder + decoder).

- Encoder's inputs are vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$.

- Decoder's inputs are vectors $\mathbf{x}'_1, \mathbf{x}'_2, \cdots, \mathbf{x}'_t$.

Encoder's inputs:

$$\mathbf{x}_1 \qquad \mathbf{x}_2 \qquad \mathbf{x}_3 \quad \cdots \quad \mathbf{x}_m$$

Decoder's inputs:

$$\mathbf{x}'_1 \quad \mathbf{x}'_2 \quad \mathbf{x}'_3 \quad \cdots \quad \mathbf{x}'_t$$

Stacked Attentions

- Transformer's encoder contains 6 stacked blocks.
- 1 block $\approx$ 1 multi-head attention layer + 1 dense layer.

$$\mathbf{u}_{:1} \quad \mathbf{u}_{:2} \quad \mathbf{u}_{:3} \quad \cdots \quad \mathbf{u}_{:m}$$

Encoder (6x stacked blocks)

$$\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \cdots \quad \mathbf{x}_m$$

Decoder's inputs:

$$\mathbf{x}_1' \quad \mathbf{x}_2' \quad \mathbf{x}_3' \quad \cdots \quad \mathbf{x}_t'$$

Stacked Attentions

$\mathbf{u}_{:1}$ $\mathbf{u}_{:2}$ $\mathbf{u}_{:3}$ $\cdots$ $\mathbf{u}_{:m}$

Encoder (6x stacked blocks)

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\cdots$ $\mathbf{x}_m$

$\mathbf{c}_{:1}$ $\mathbf{c}_{:2}$ $\mathbf{c}_{:3}$ $\cdots$ $\mathbf{c}_{:t}$

Multi-Head Self-Attention

$\mathbf{x}'_1$ $\mathbf{x}'_2$ $\mathbf{x}'_3$ $\cdots$ $\mathbf{x}'_t$
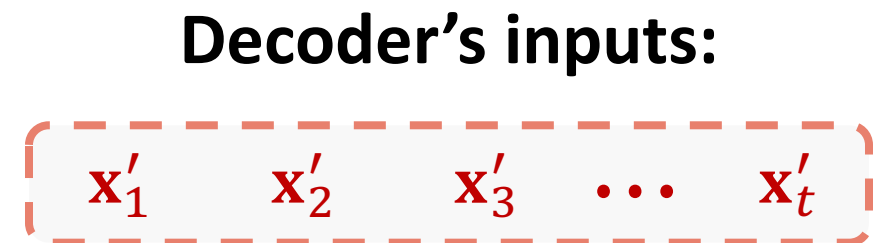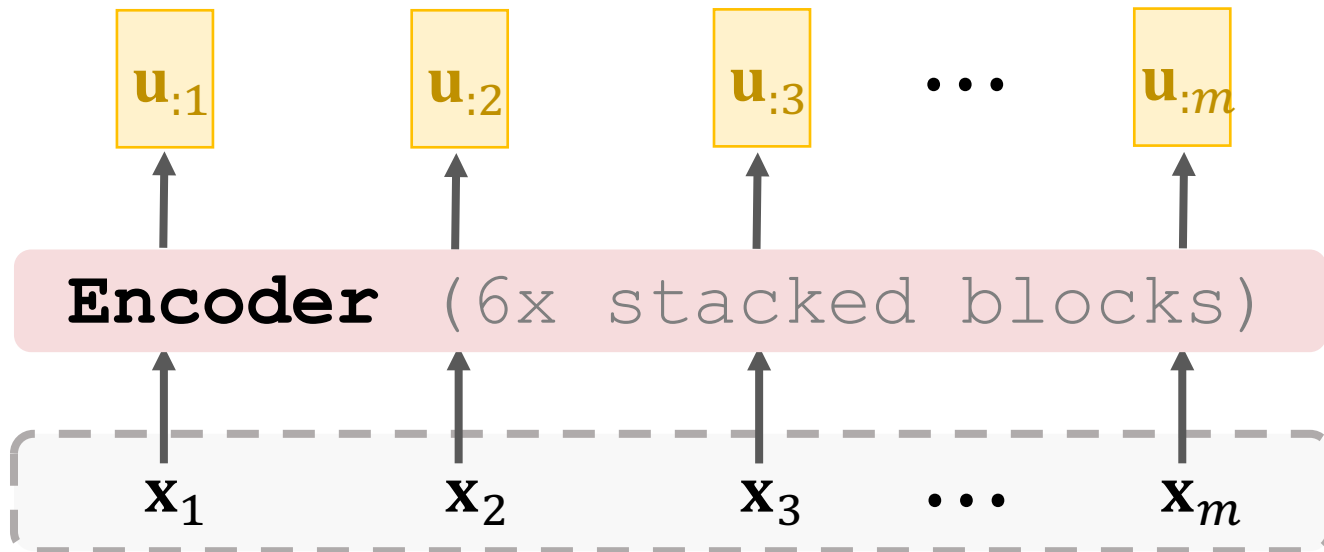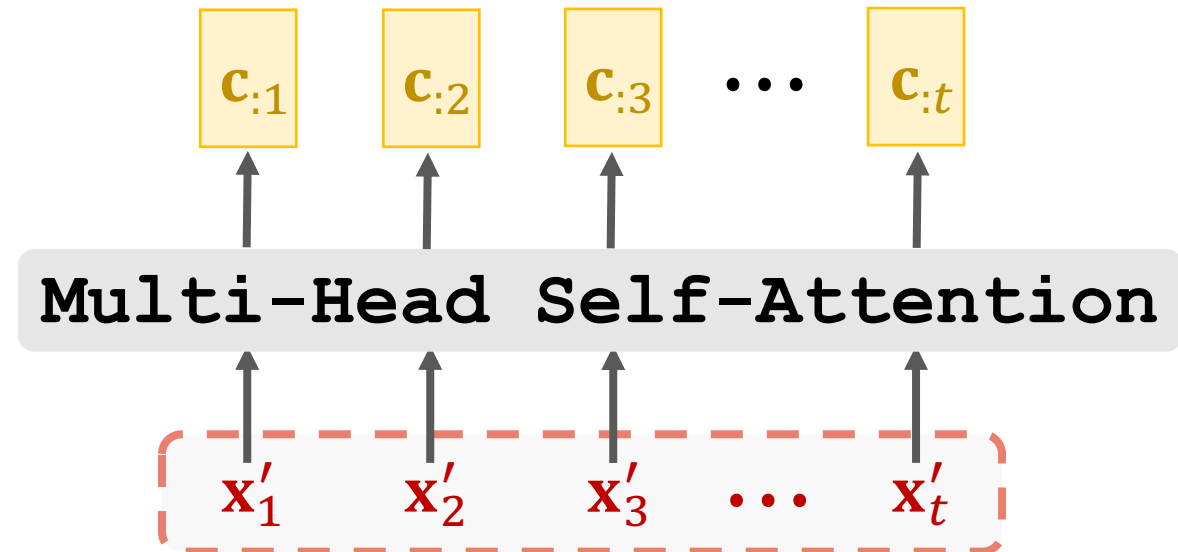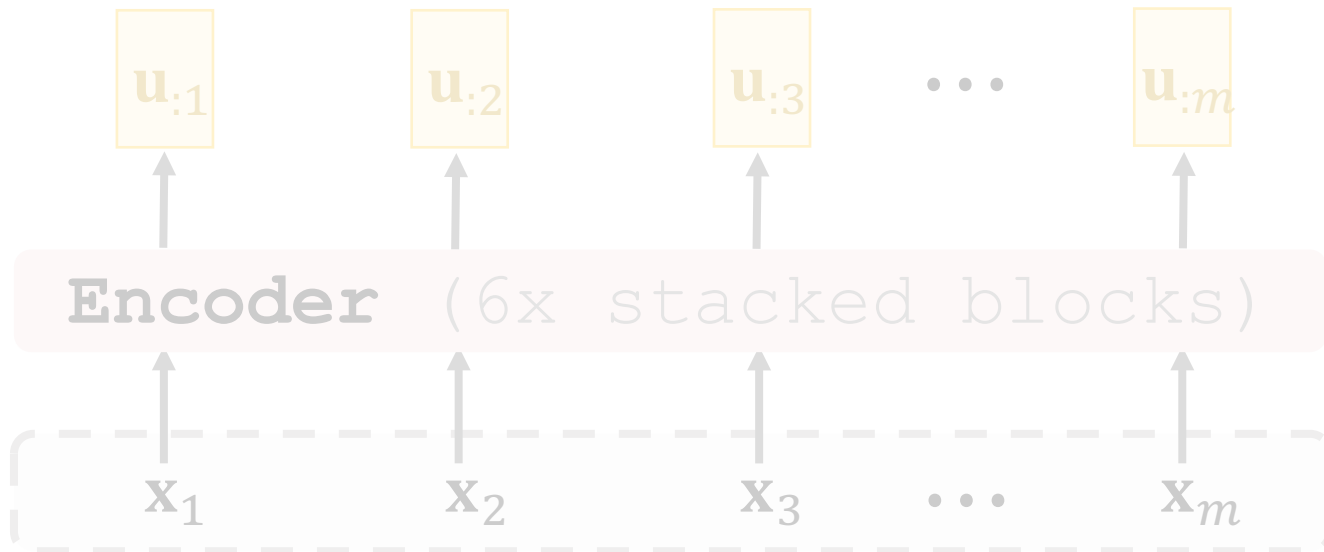
Stacked Attentions

Stacked Attentions

$$\mathbf{s}_{:1} = \text{ReLU}(\mathbf{W}_S \mathbf{z}_{:1})$$

Dense

$\mathbf{z}_{:1}$ $\mathbf{z}_{:2}$ $\mathbf{z}_{:3}$ $\cdots$ $\mathbf{z}_{:t}$

Multi-Head Attention Layer

$\mathbf{u}_{:1}$ $\mathbf{u}_{:2}$ $\mathbf{u}_{:3}$ $\cdots$ $\mathbf{u}_{:m}$ $\mathbf{c}_{:1}$ $\mathbf{c}_{:2}$ $\mathbf{c}_{:3}$ $\cdots$ $\mathbf{c}_{:t}$

Encoder (6x stacked blocks)   Multi-Head Self-Attention

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\cdots$ $\mathbf{x}_m$ $\mathbf{x}'_1$ $\mathbf{x}'_2$ $\mathbf{x}'_3$ $\cdots$ $\mathbf{x}'_t$

Stacked Attentions

$$\mathbf{s}_{:1} \quad \mathbf{s}_{:2} = \mathrm{ReLU}(\mathbf{W}_S \mathbf{z}_{:2})$$

Dense Dense

$$\mathbf{z}_{:1} \quad \mathbf{z}_{:2} \quad \mathbf{z}_{:3} \quad \cdots \quad \mathbf{z}_{:t}$$

Multi-Head Attention Layer

$$\mathbf{u}_{:1} \quad \mathbf{u}_{:2} \quad \mathbf{u}_{:3} \quad \cdots \quad \mathbf{u}_{:m} \qquad \mathbf{c}_{:1} \quad \mathbf{c}_{:2} \quad \mathbf{c}_{:3} \quad \cdots \quad \mathbf{c}_{:t}$$

Encoder (6x stacked blocks)   **Multi-Head Self-Attention**

$$\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \cdots \quad \mathbf{x}_m \qquad \mathbf{x}_1' \quad \mathbf{x}_2' \quad \mathbf{x}_3' \quad \cdots \quad \mathbf{x}_t'$$

Stacked Attentions

Transformer's Decoder : One Block

512×t matrix

Dense Layer

Attention Layer

Self-Attention Layer

512×m matrix   512×t matrix
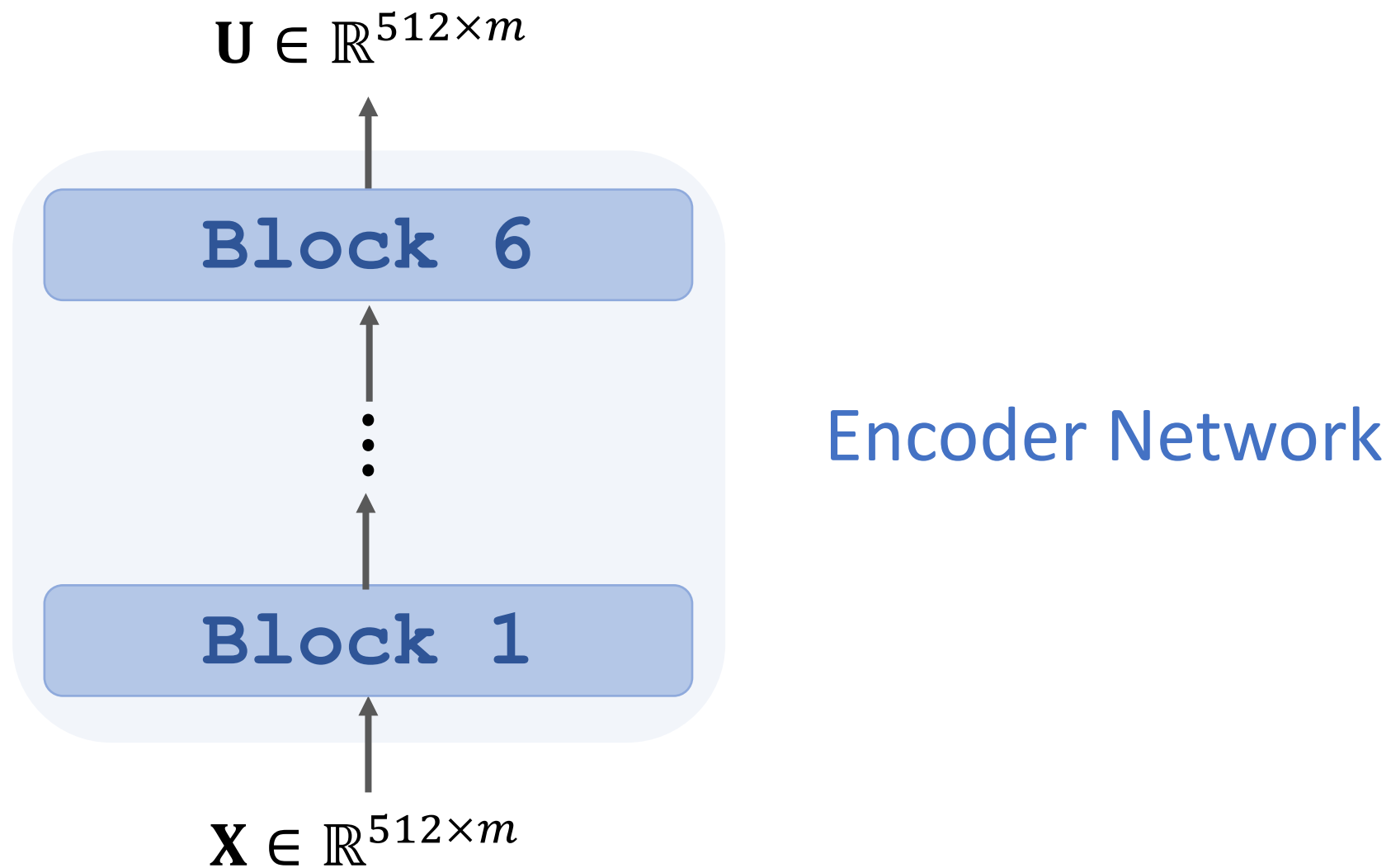
Transformer's Decoder : One Block
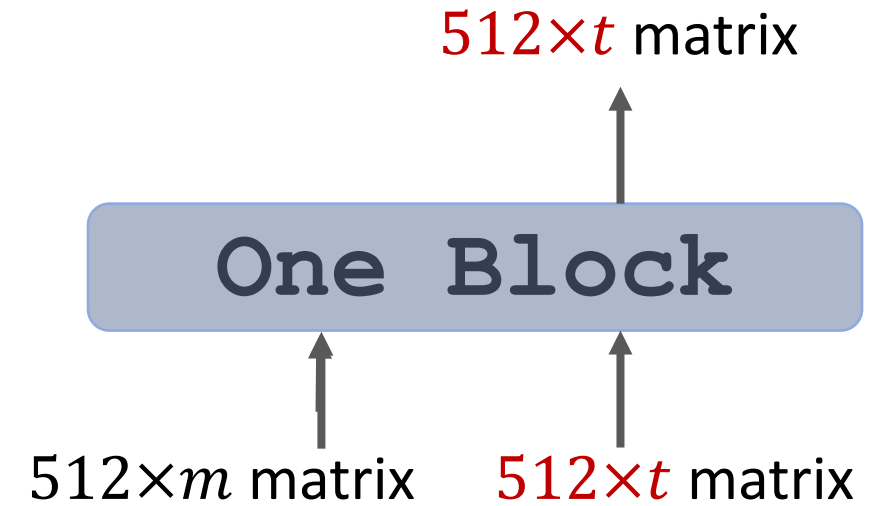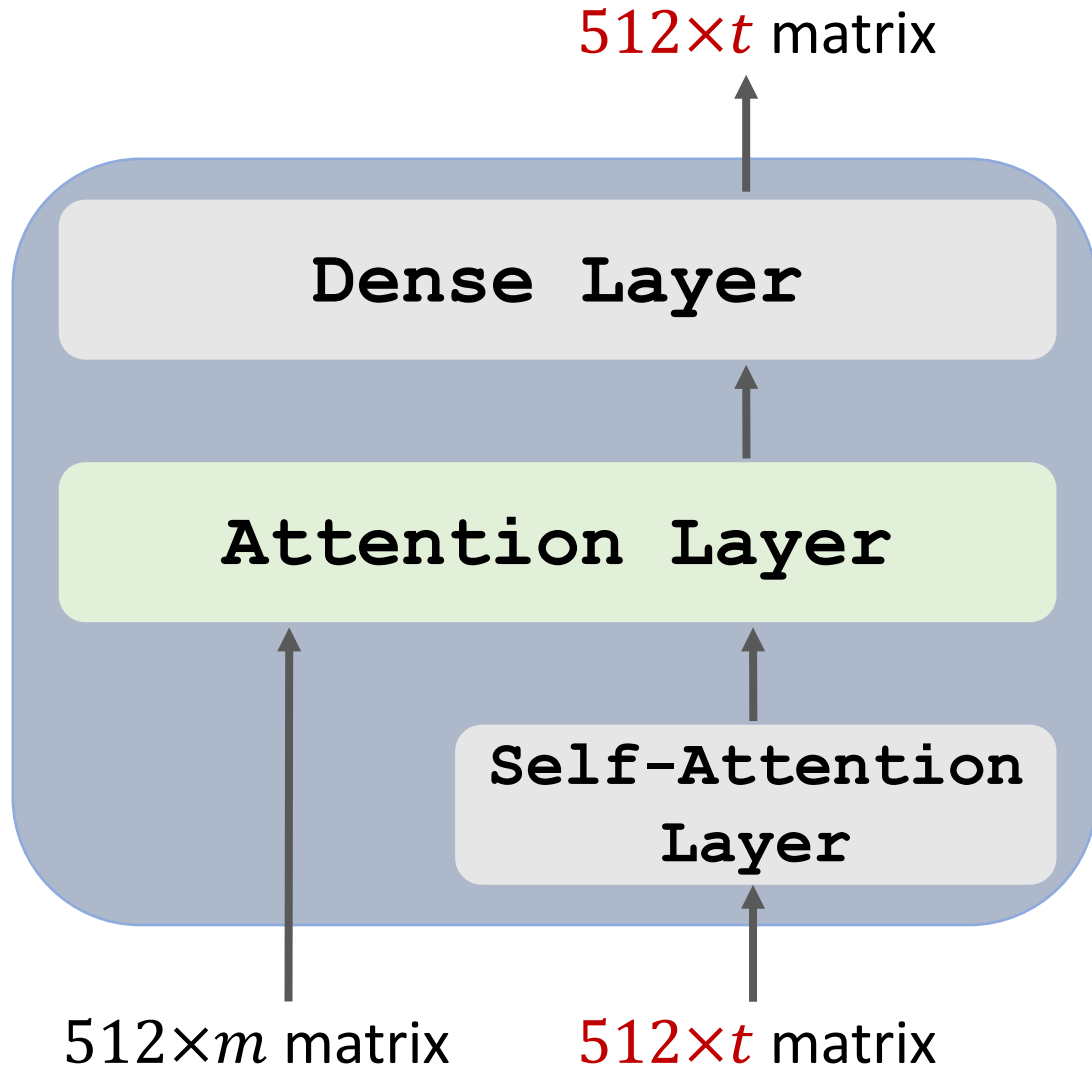
Put Everything Together

Transformer

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

Block 6

$\vdots$

Block 1

Encoder Network

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

Transformer

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

Block 6

$\vdots$

Block 1

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

$512 \times t$ matrix

Block 1

$\mathbf{X}' \in \mathbb{R}^{512 \times t}$

Transformer

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

Block 6

⋮

Block 1

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

$512 \times t$ matrix

Block 2

Block 1

$\mathbf{X}' \in \mathbb{R}^{512 \times t}$

Transformer

$512 \times t$ matrix

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

Block 6

Block 1

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

Block 6

Block 2

Block 1

$\mathbf{X}' \in \mathbb{R}^{512 \times t}$
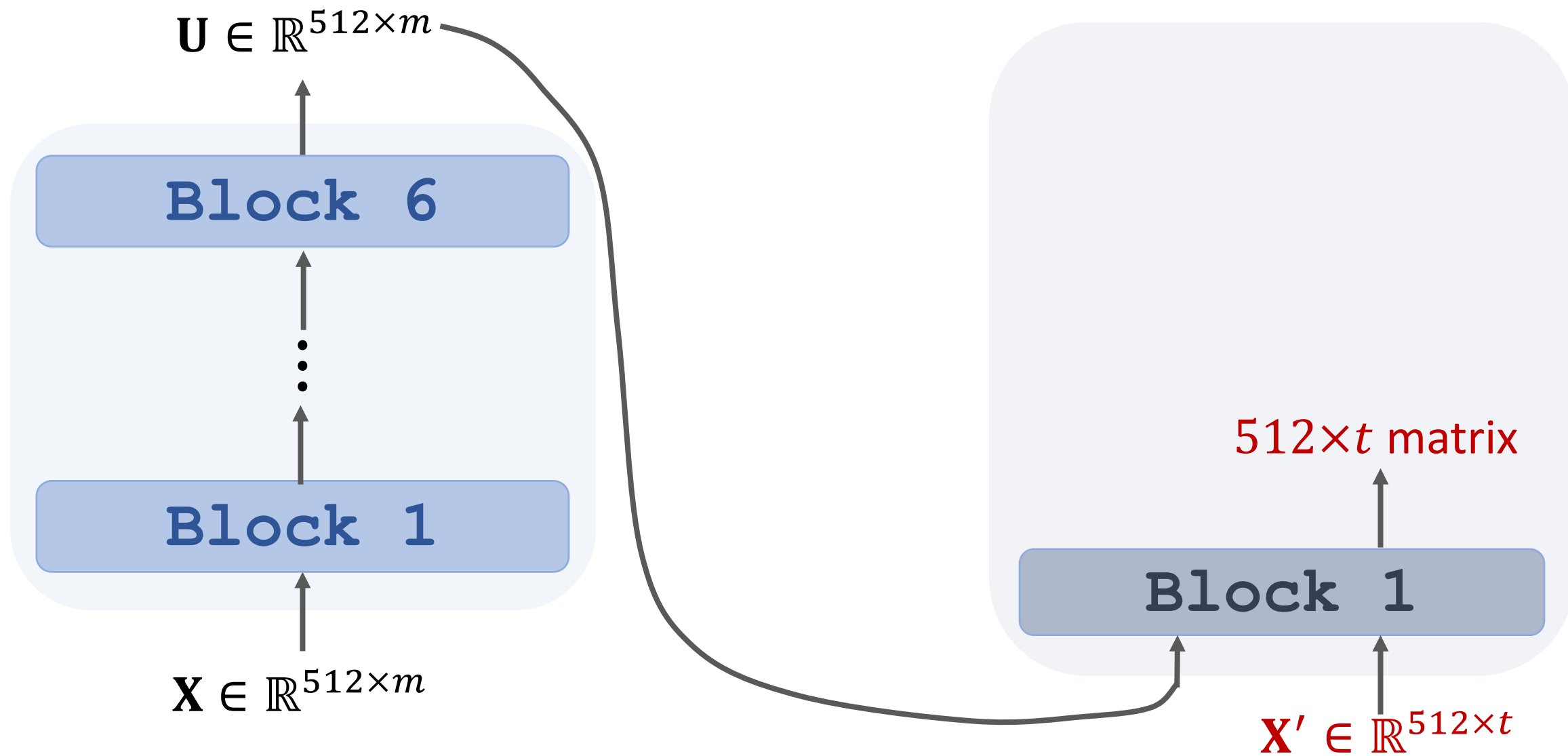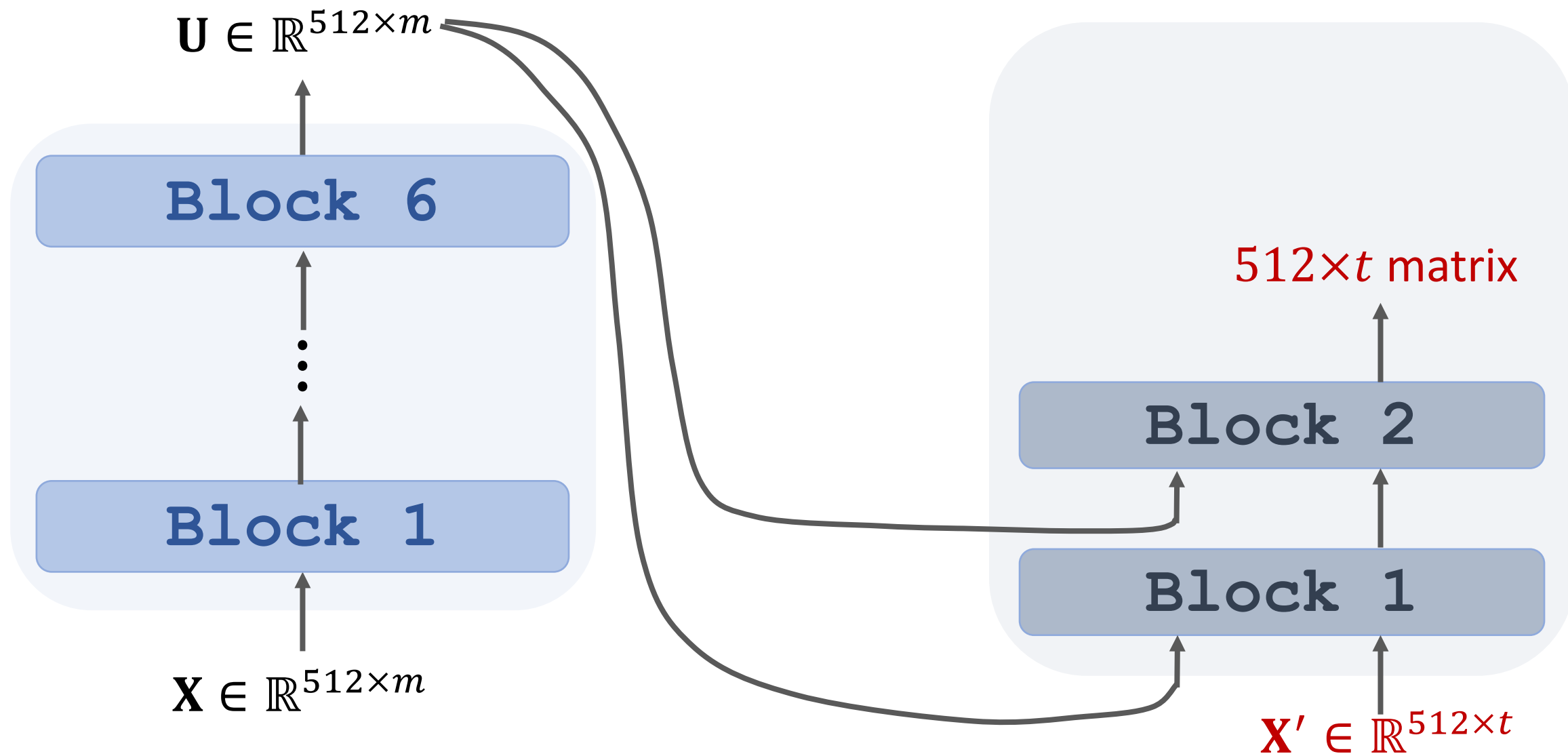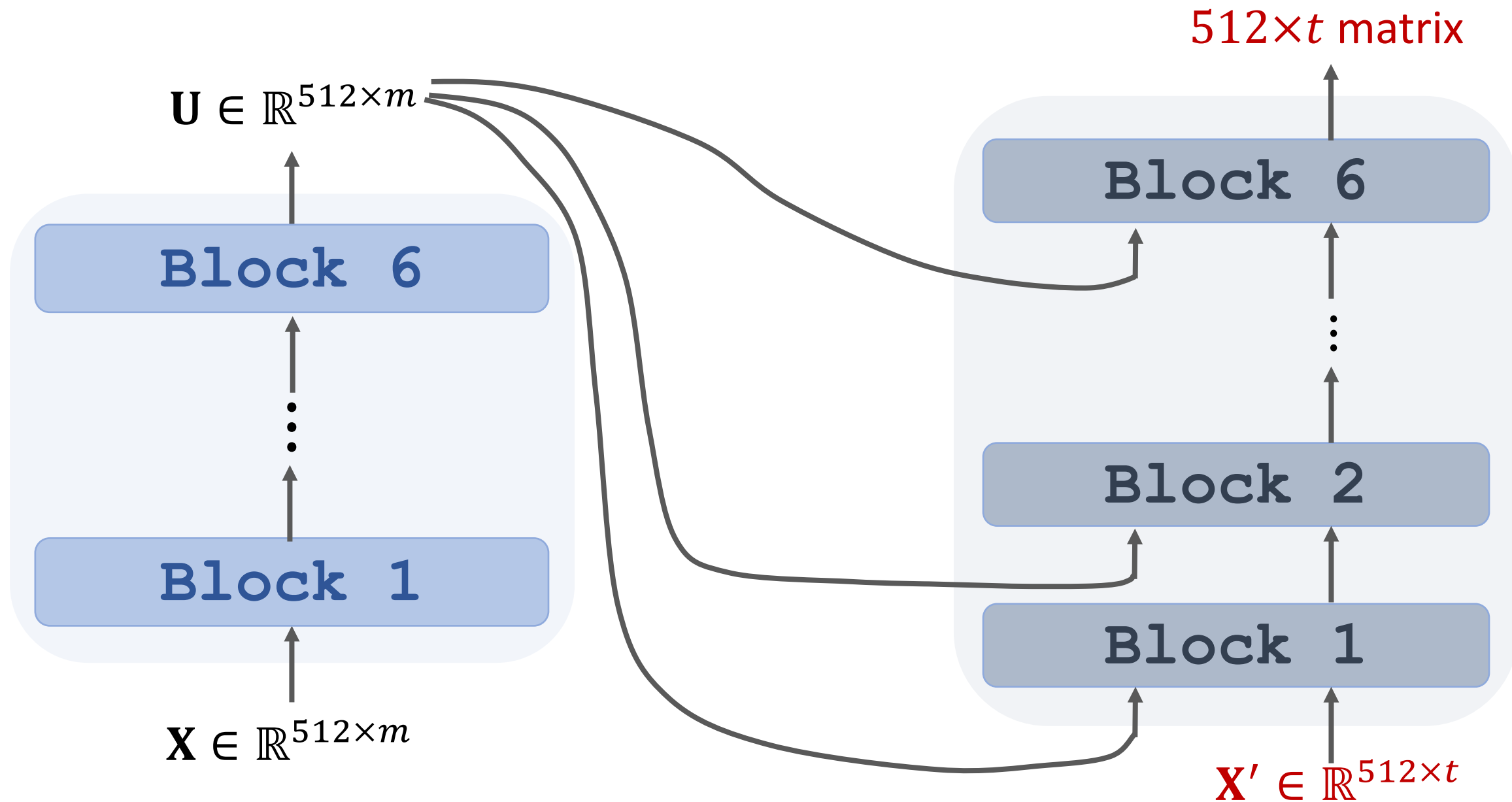
Comparison with RNN Seq2Seq Model
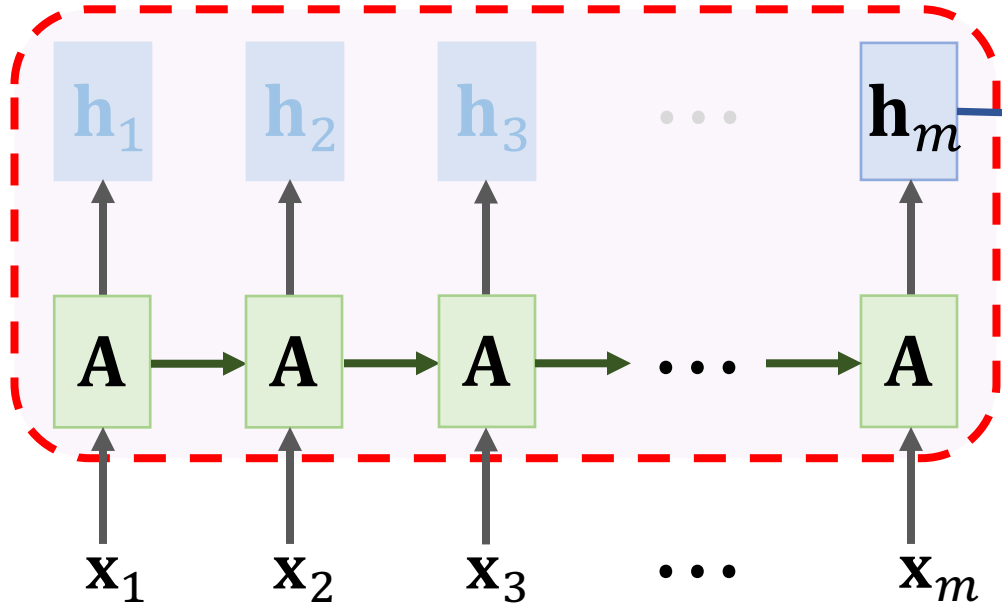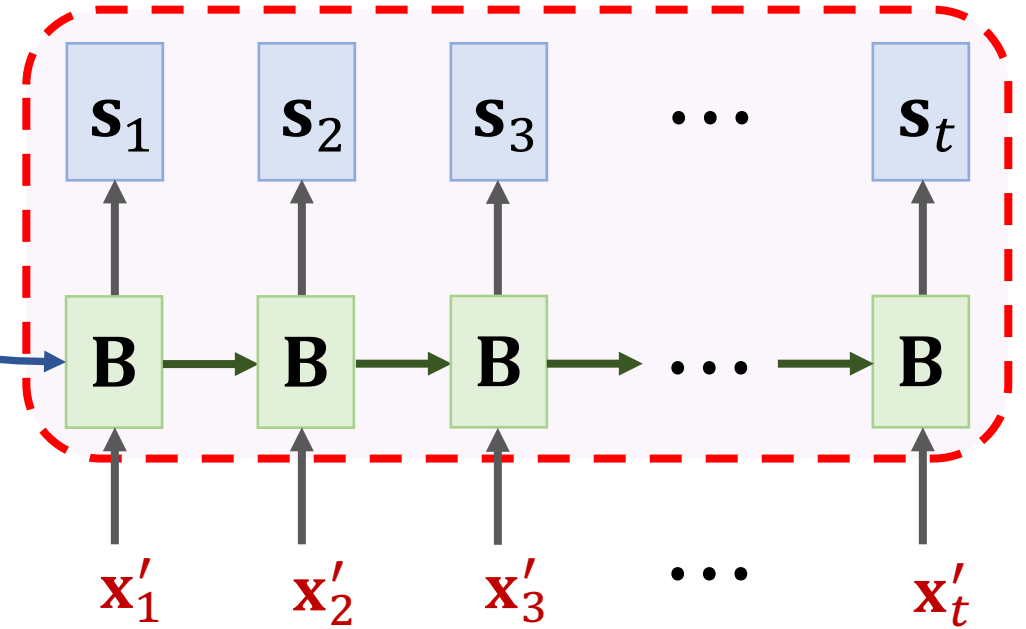
Applied to Machine Translation

Example: English to German

$$\mathbf{U} \in \mathbb{R}^{512 \times m}$$

Encoder

$$\mathbf{X} \in \mathbb{R}^{512 \times m}$$

Example: English to German

$$\mathbf{U} \in \mathbb{R}^{512 \times m}$$

Encoder

$$\mathbf{X} \in \mathbb{R}^{512 \times m}$$

Encoder (6 Blocks)

Example: English to German

$$\mathbf{U} \in \mathbb{R}^{512 \times m}$$

Encoder

$$\mathbf{X} \in \mathbb{R}^{512 \times m}$$

$\mathbf{u}_1$ $\mathbf{u}_2$ $\cdots$ $\mathbf{u}_m$

Encoder (6 Blocks)

$\mathbf{x}_1$ $\mathbf{x}_2$ $\cdots$ $\mathbf{x}_m$

1st English Word   2nd English Word $\cdots$ mth English Word

Example: English to German

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

Encoder

Decoder (6 Blocks)

Example: English to German

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

Encoder

Decoder (6 Blocks)

$\mathbf{y}_1$

$\mathbf{x}'_1$

Start Sign

Example: English to German

Example: English to German

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

Encoder

Decoder (6 Blocks)

$\mathbf{y}_1$

$\mathbf{y}_2$

$\mathbf{x}_1'$

$\mathbf{x}_2'$

Start Sign

1st German Word

Example: English to German

Example: English to German

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

$\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_3$

Encoder

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

Decoder (6 Blocks)

$\mathbf{x}_1'$ $\mathbf{x}_2'$ $\mathbf{x}_3'$

Start Sign

1st German Word

2nd German Word

Example: English to German

$\mathbf{U} \in \mathbb{R}^{512 \times m}$

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

Encoder

Decoder (6 Blocks)

$\mathbf{y}_1$

$\mathbf{y}_2$

$\mathbf{y}_3$

Random Sampling

$\mathbf{x}'_1$

$\mathbf{x}'_2$

$\mathbf{x}'_3$

Start Sign

1st German Word

2nd German Word

. . .

Example: English to German
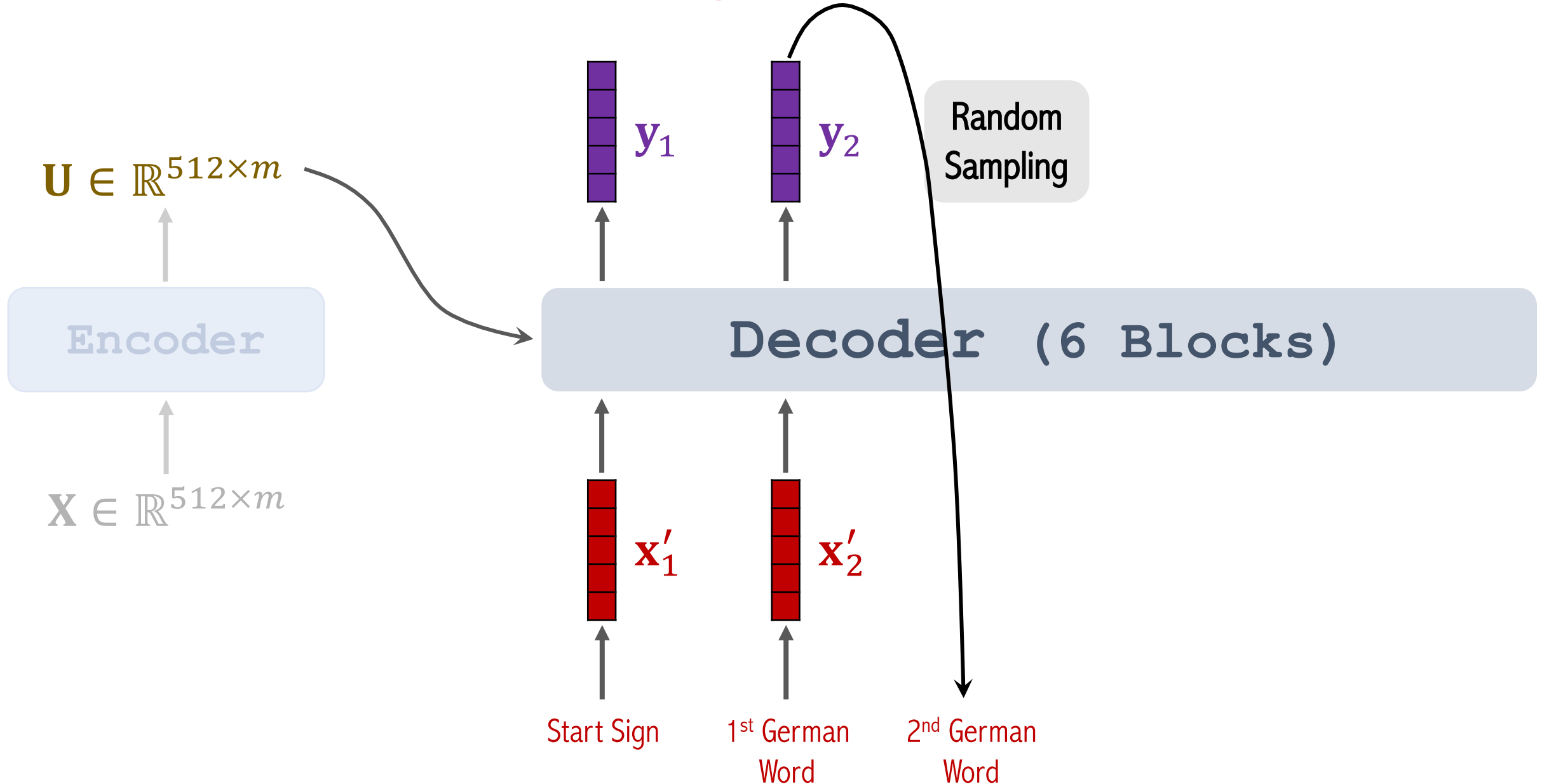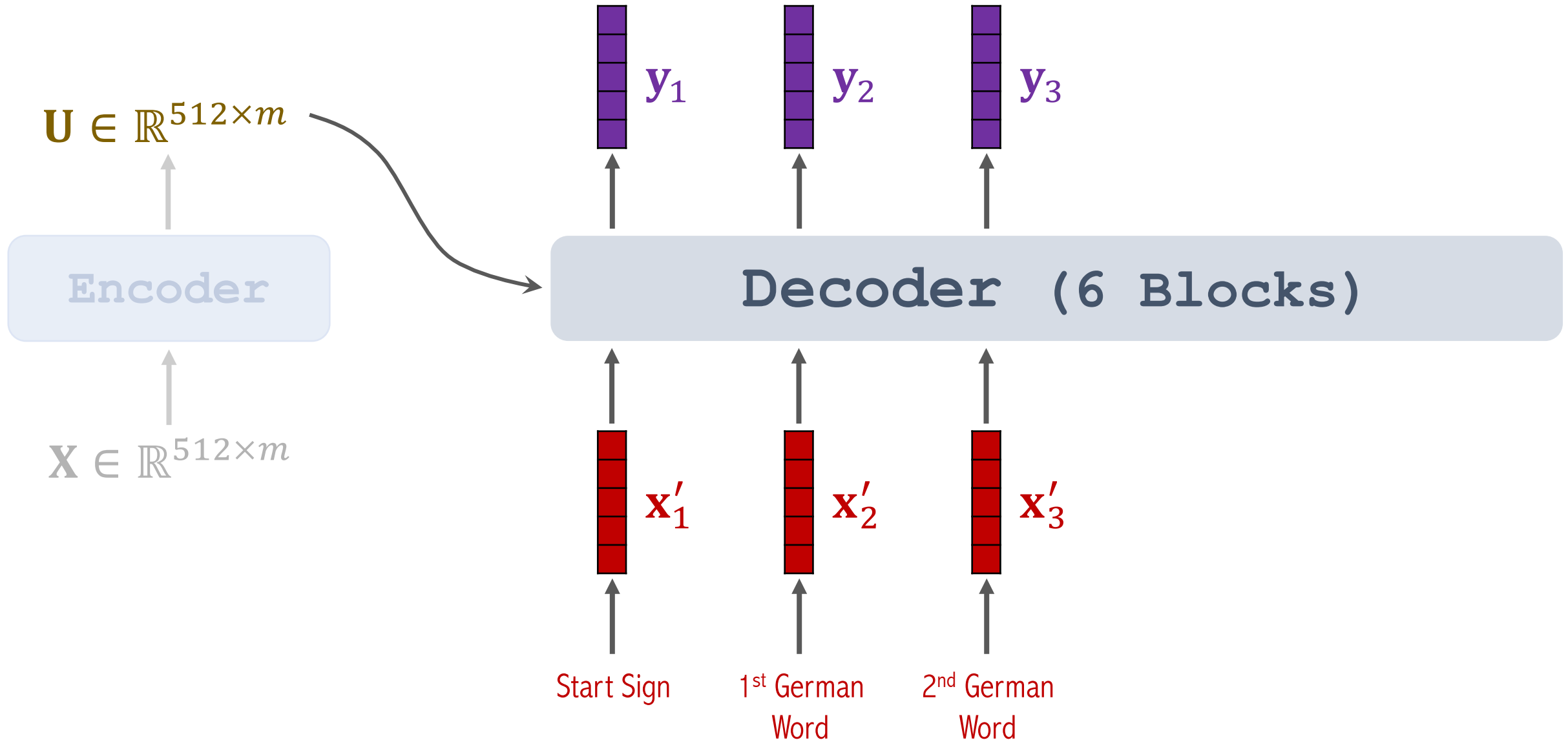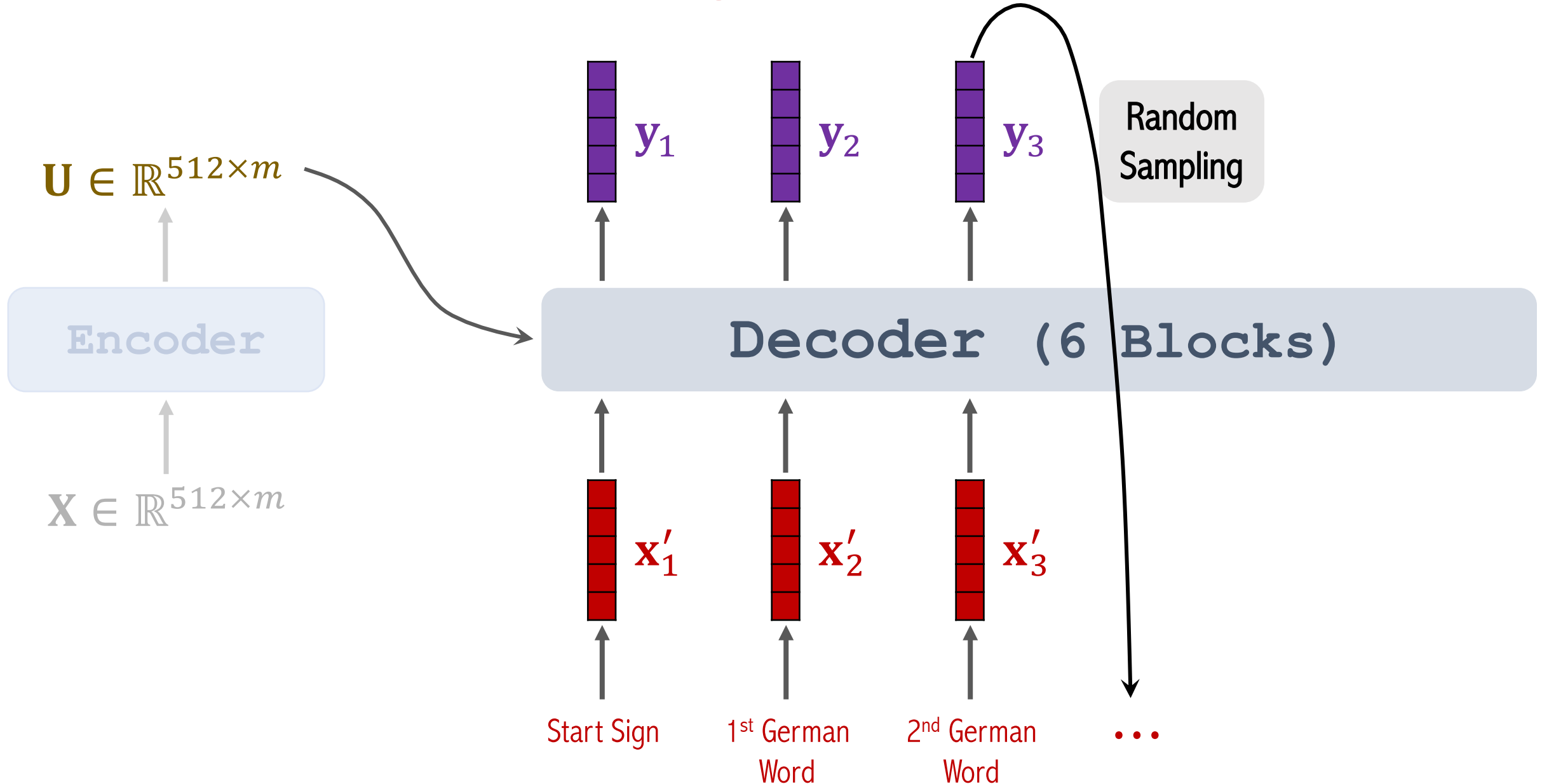
Example: English to German

Example: English to German

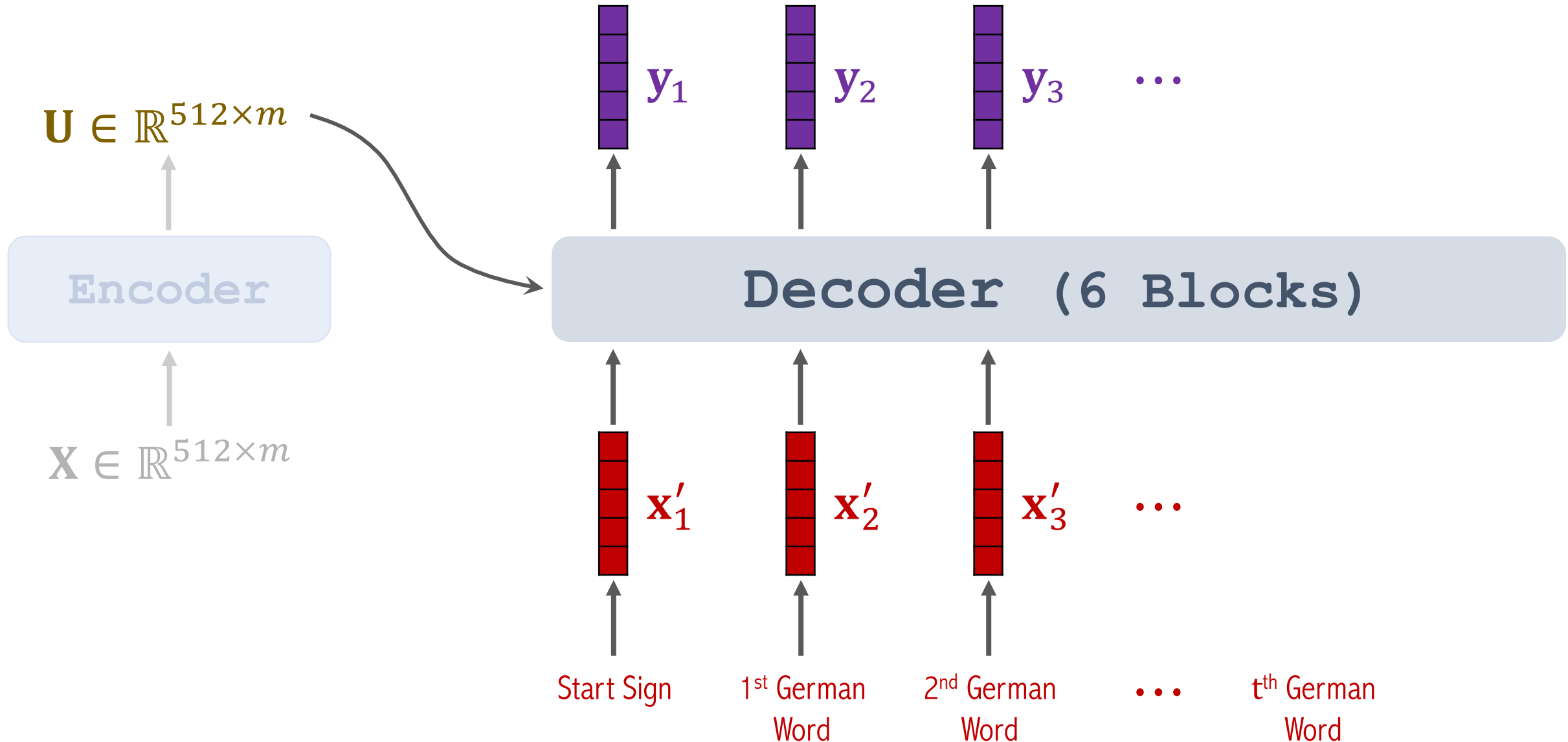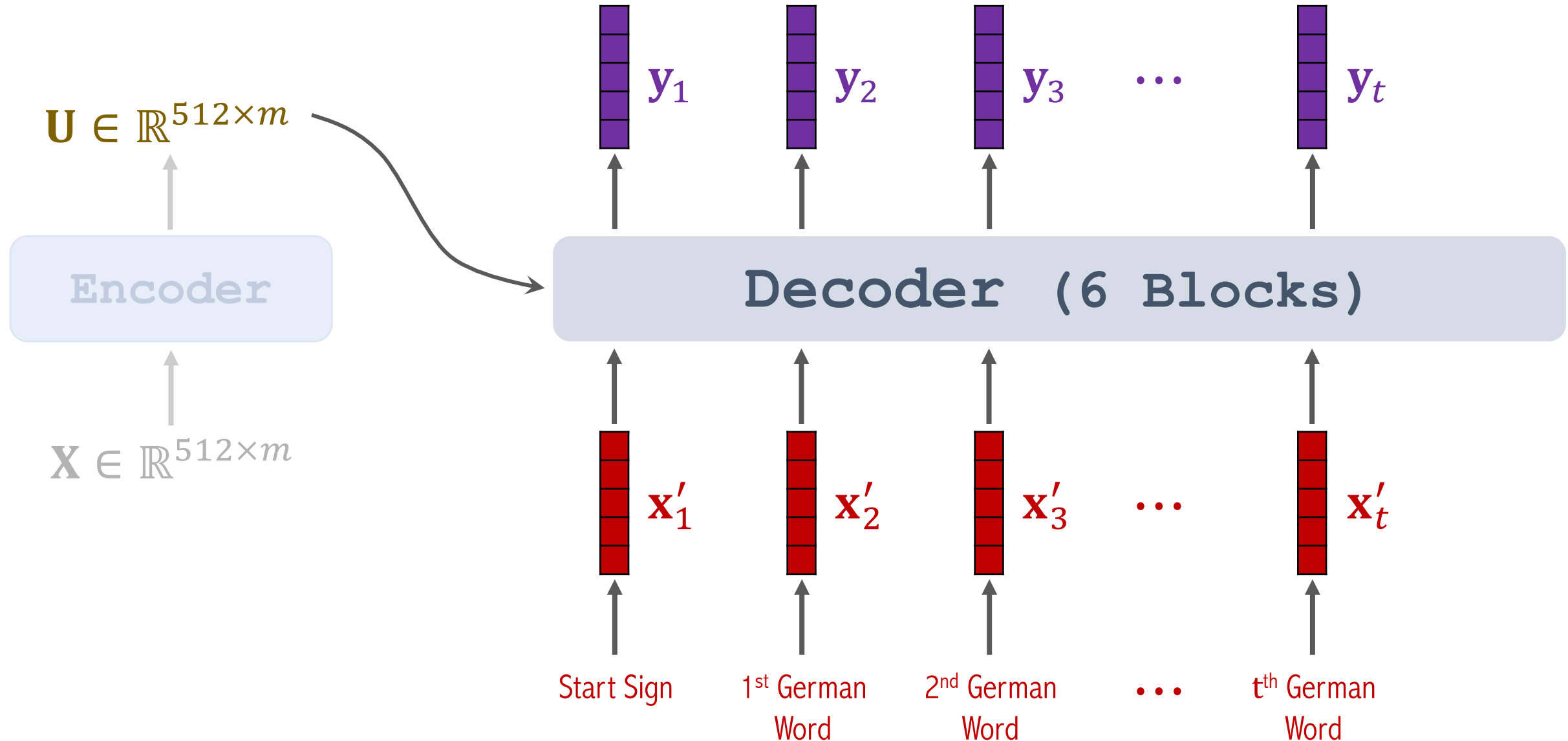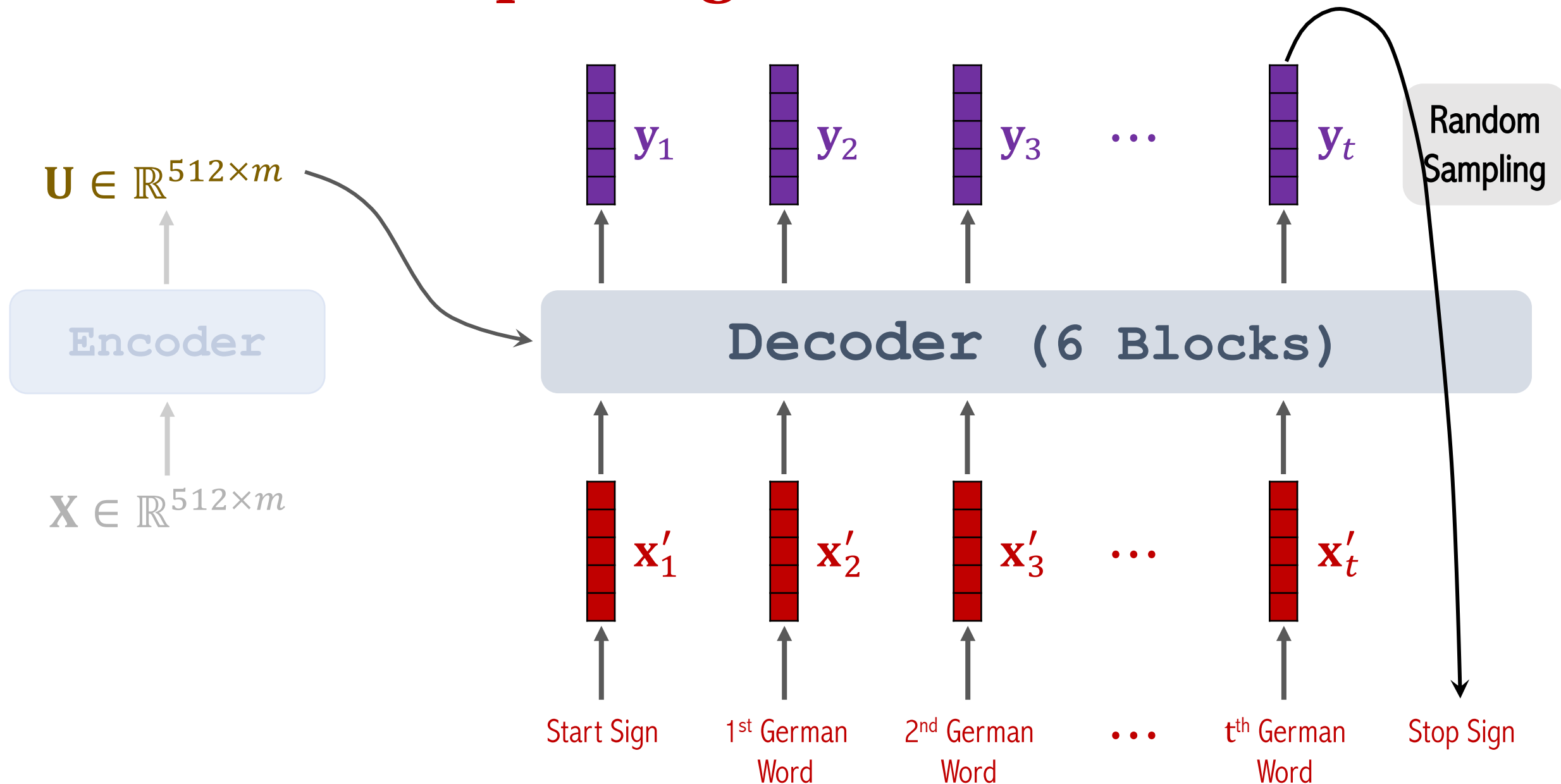$\mathbf{U} \in \mathbb{R}^{512 \times m}$

Encoder

$\mathbf{X} \in \mathbb{R}^{512 \times m}$

$\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_3$ $\cdots$ $\mathbf{y}_t$

Random Sampling

Decoder (6 Blocks)

$\mathbf{x}'_1$ $\mathbf{x}'_2$ $\mathbf{x}'_3$ $\cdots$ $\mathbf{x}'_t$

Start Sign   1st German Word   2nd German Word $\cdots$ tth German Word   Stop Sign
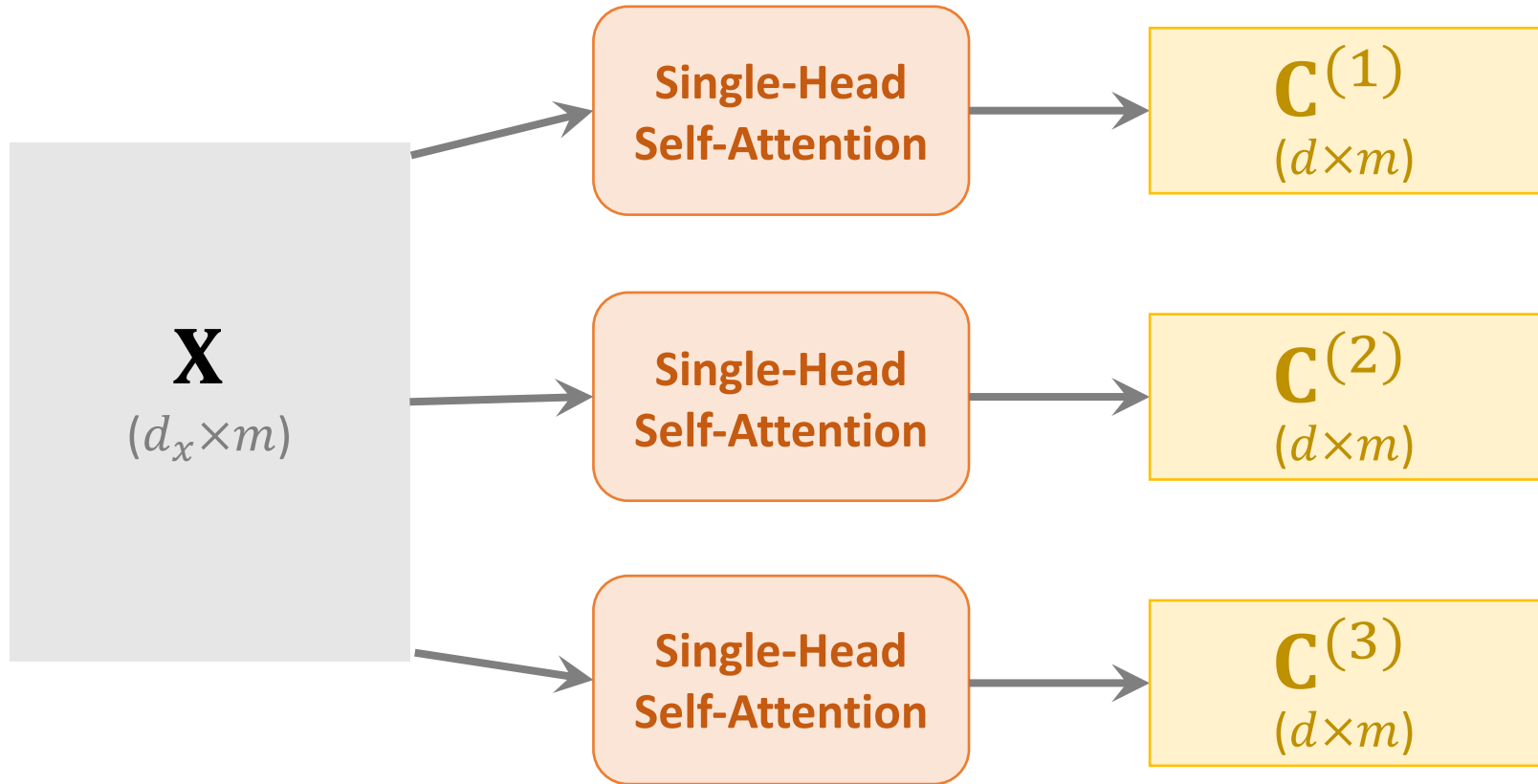
Summary

From **Single-Head** to **Multi-Head**

- Single-head self-attention can be combined to form a multi-head self-attention.

From **Single-Head** to **Multi-Head**

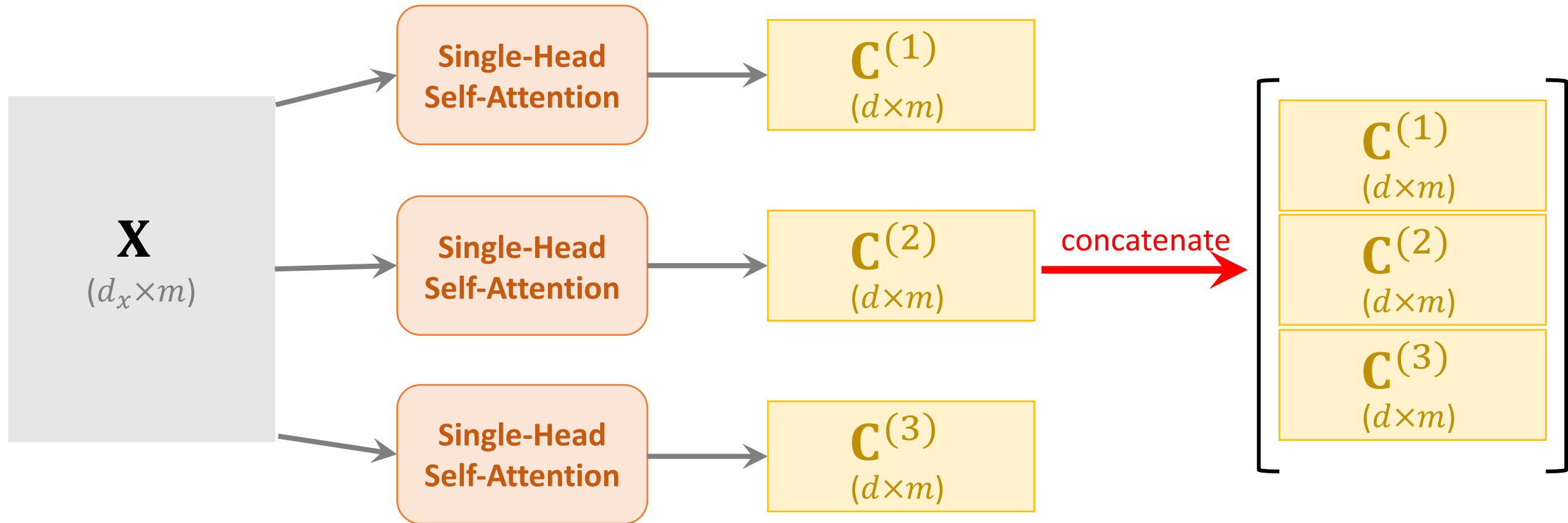- Single-head self-attention can be combined to form a multi-head self-attention.

From **Single-Head** to **Multi-Head**

- Single-head self-attention can be combined to form a multi-head self-attention.

- Single-head attention can be combined to form a multi-head attention.

Encoder Network of Transformer

- 1 encoder block ≈ multi-head self-attention + dense.

- Input shape: $512{\times}m$.

- Output shape: $512{\times}m$.

- Encoder network is a stack of 6 such blocks.

Decoder Network of Transformer

- 1 decoder block ≈ multi-head self-attention + multi-head attention + dense.

- Input shape: $(512 \times m \ , \ 512 \times t)$.

- Output shape: $512 \times t$.

- Decoder network is a stack of 6 such blocks.

Transformer Model

- Transformer is Seq2Seq model; it has an encoder and a decoder.

- Transformer model is not RNN.

- Transformer is based on attention and self-attention.

- Transformer outperforms all the state-of-the-art RNN models.