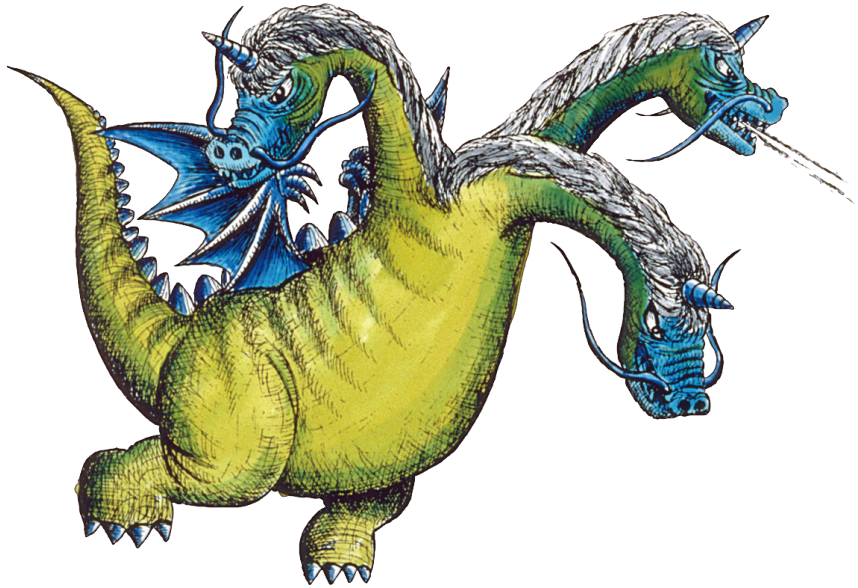
Transformer Model (2/2): From Shallow to Deep

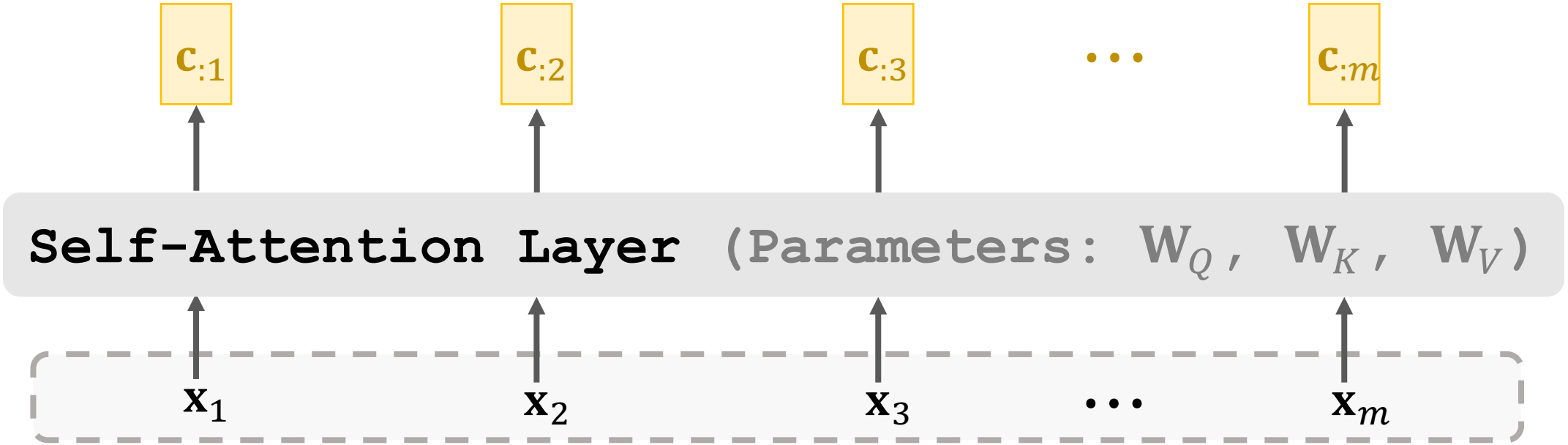
Shusen Wang

Multi-Head Attention



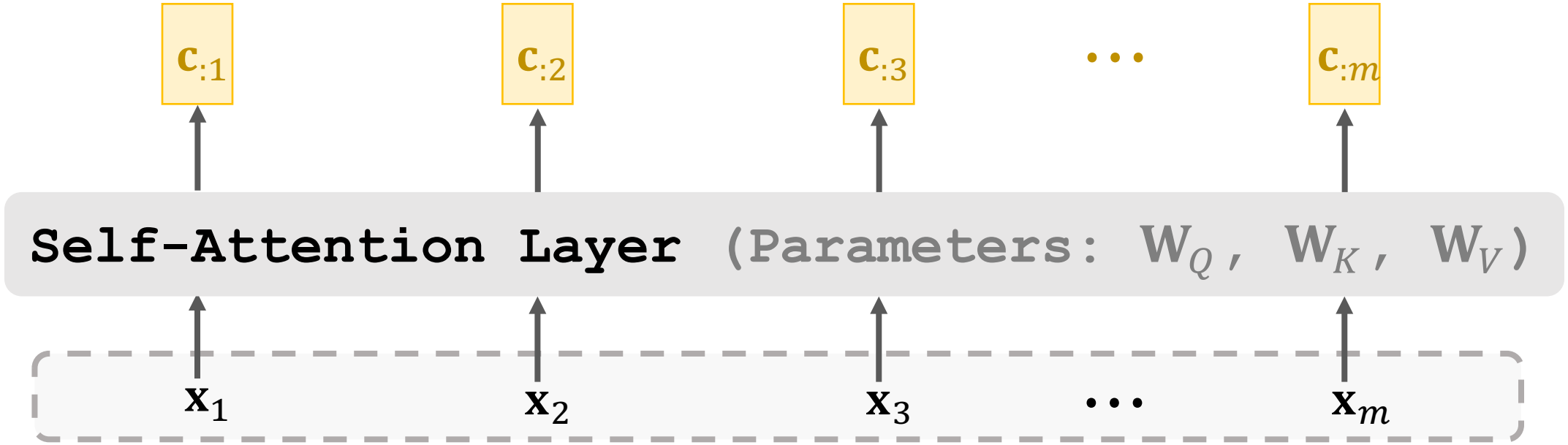
Single-Head Self-Attention

- Self-attention layer: $\mathbf{C} = \mathrm{Attn}(\mathbf{X}, \mathbf{X})$.
- This is called "single-head self-attention".



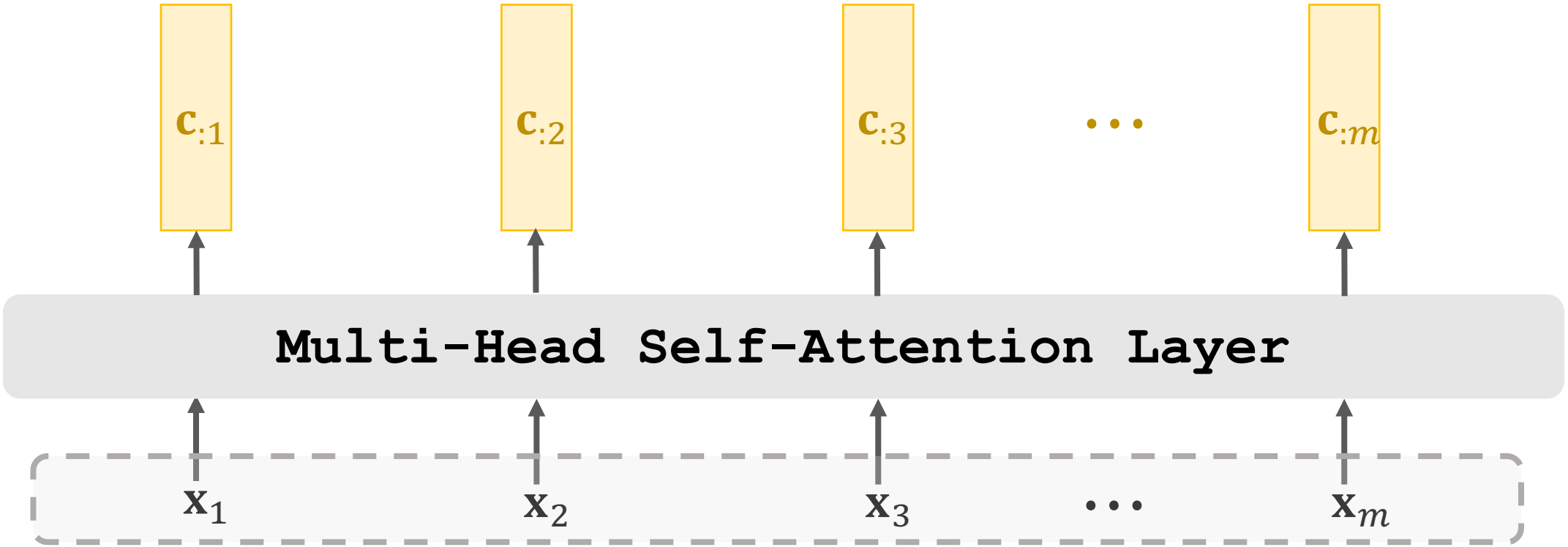
Multi-Head Self-Attention

- 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.



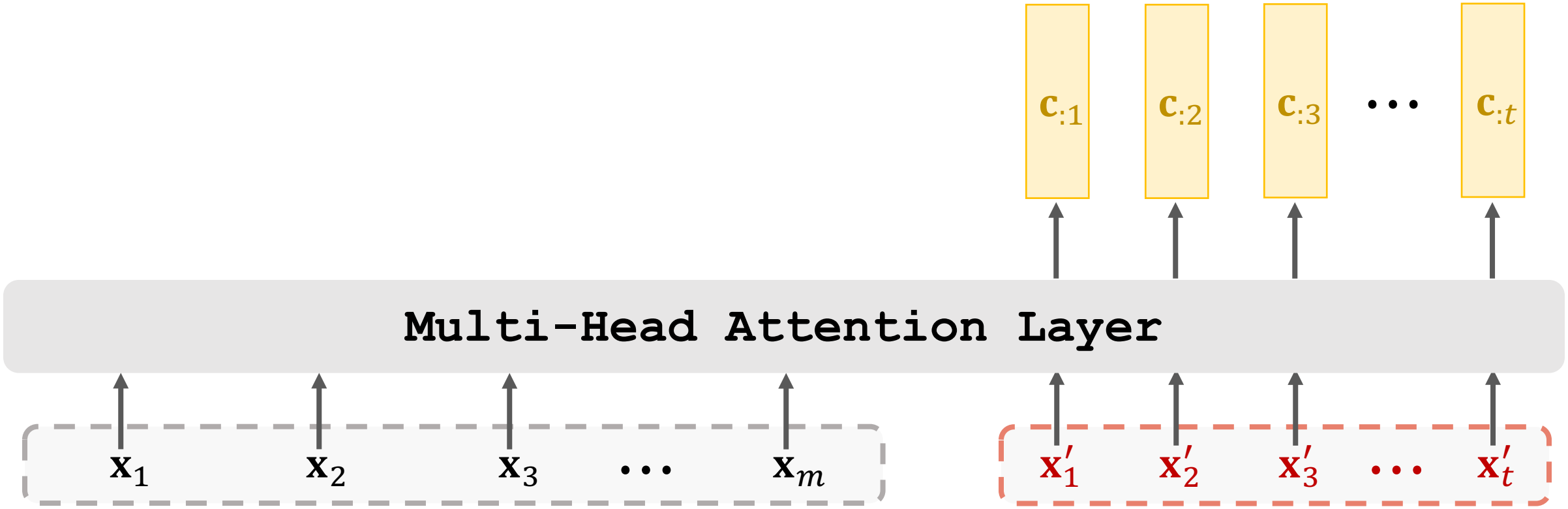
Multi-Head Self-Attention

- 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: $(ld) \times m$.



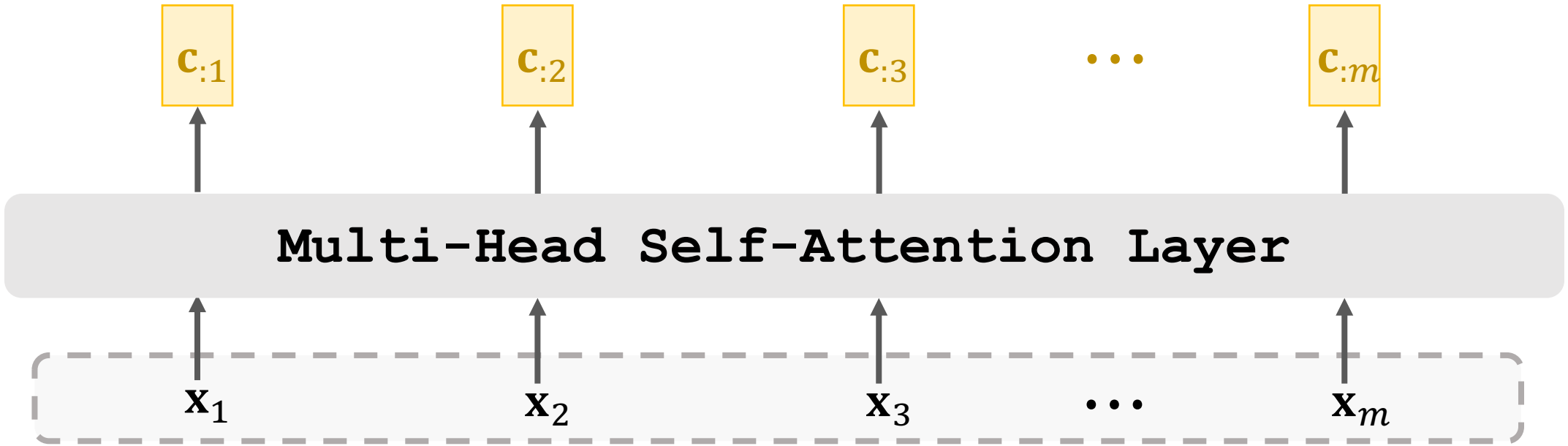
Multi-Head Attention

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

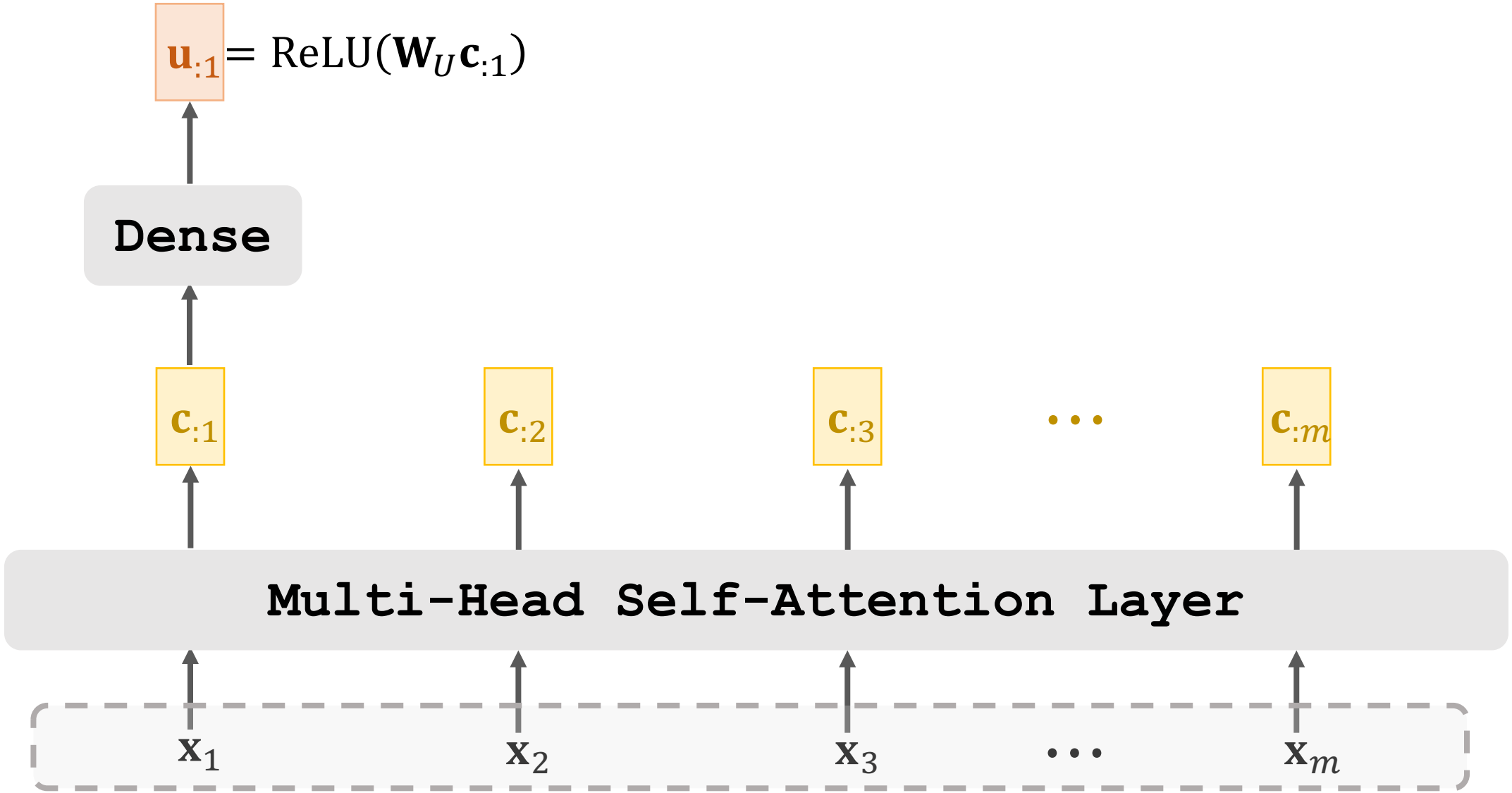


Stacked Self-Attention Layers

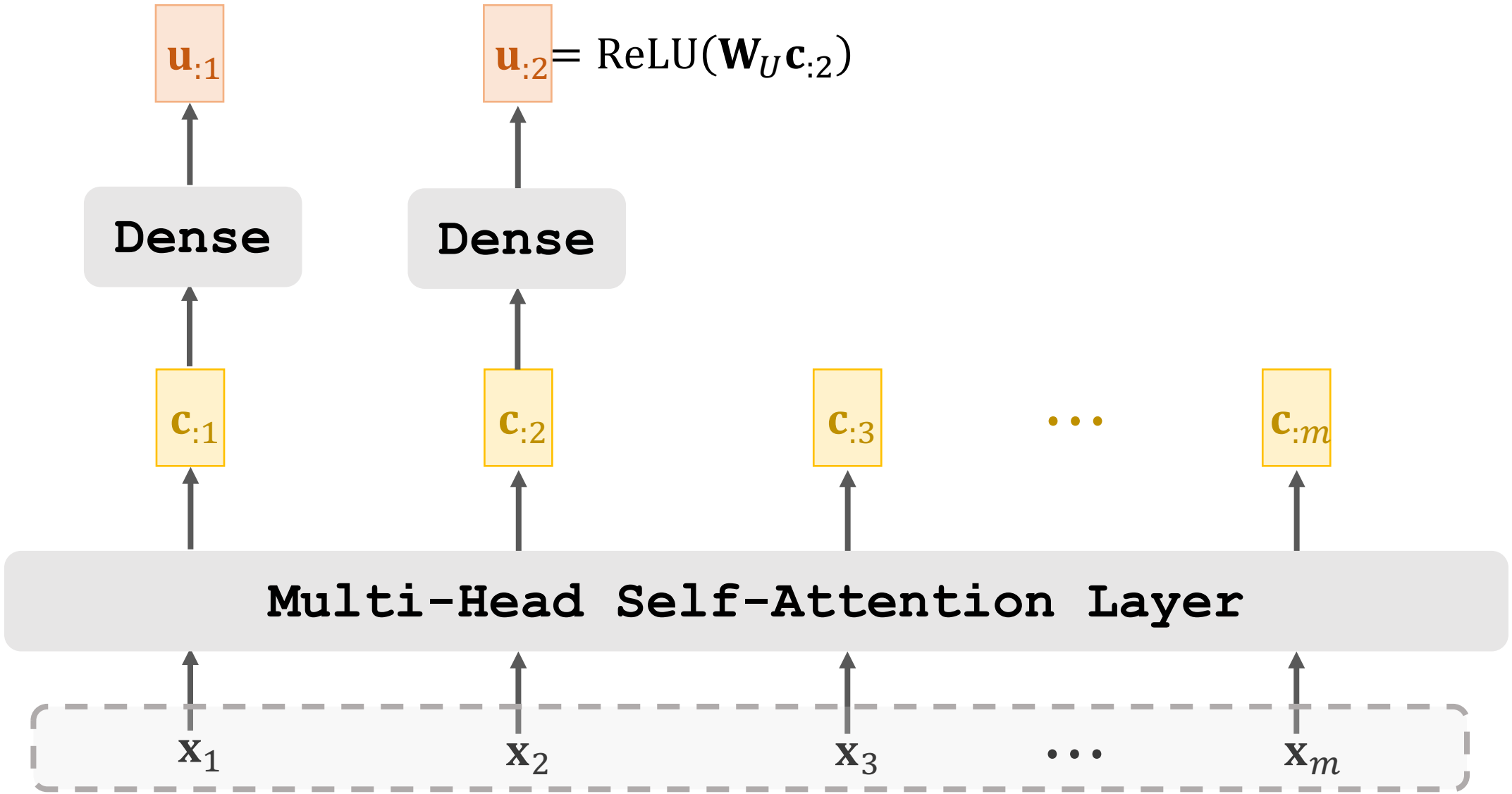
Self-Attention Layer



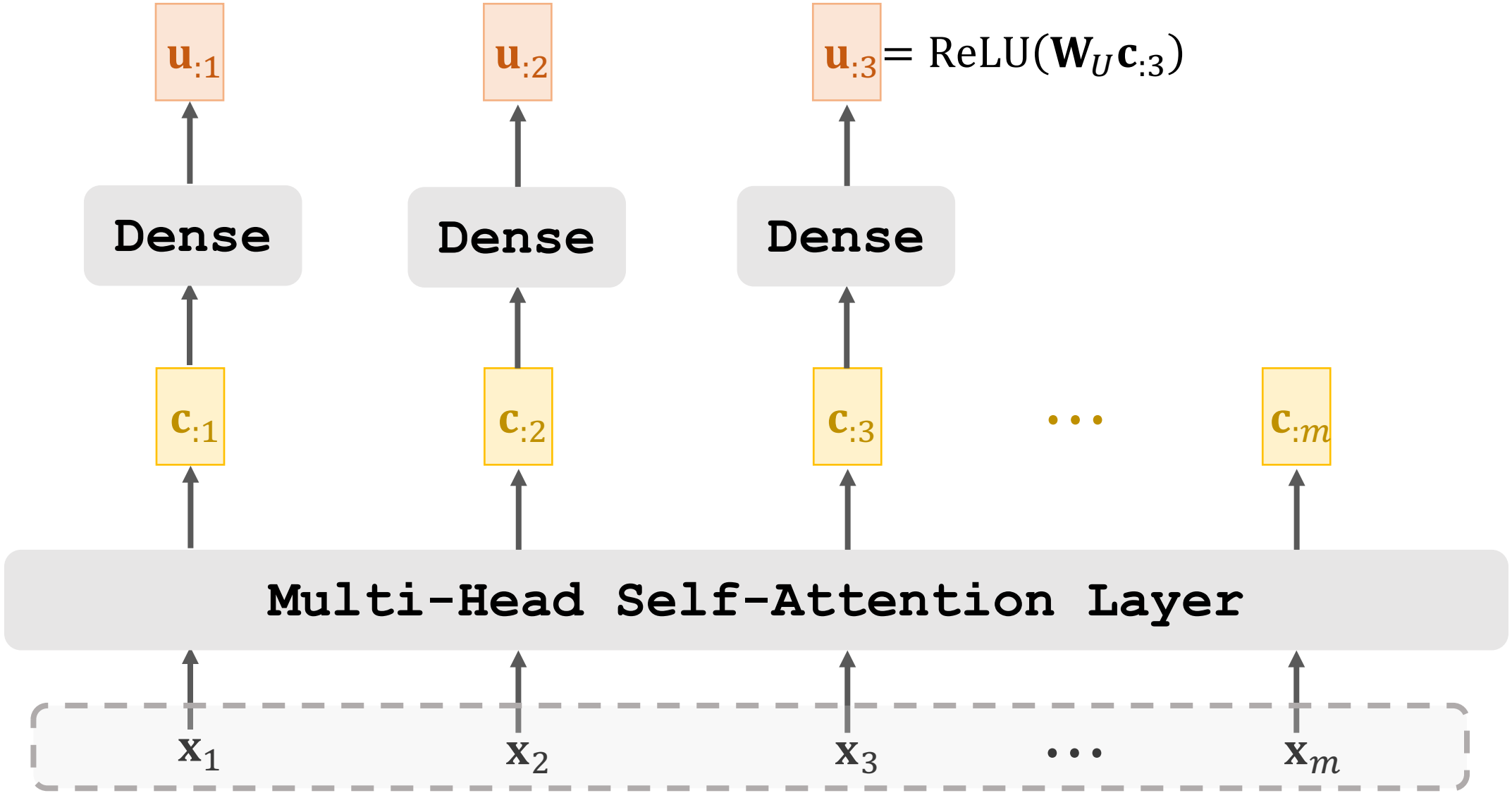
Self-Attention Layer + Dense Layer



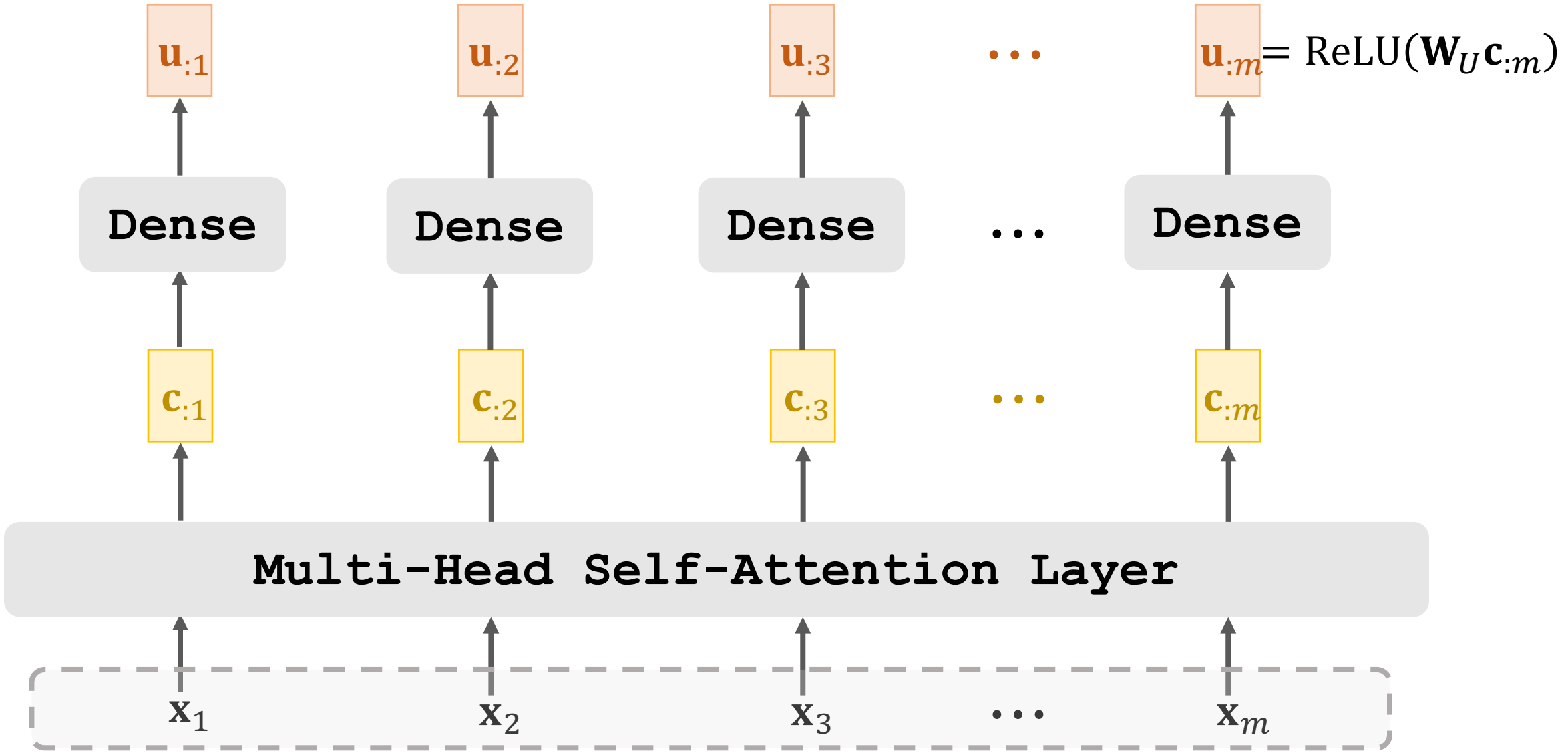
Self-Attention Layer + Dense Layer



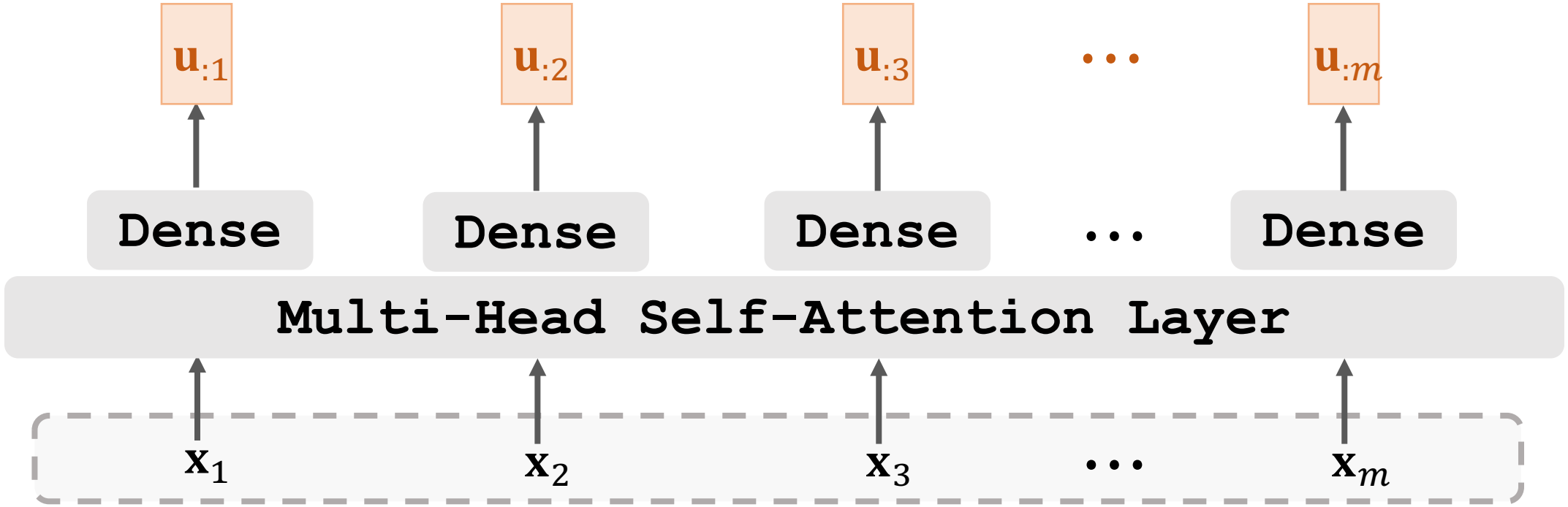
Self-Attention Layer + Dense Layer



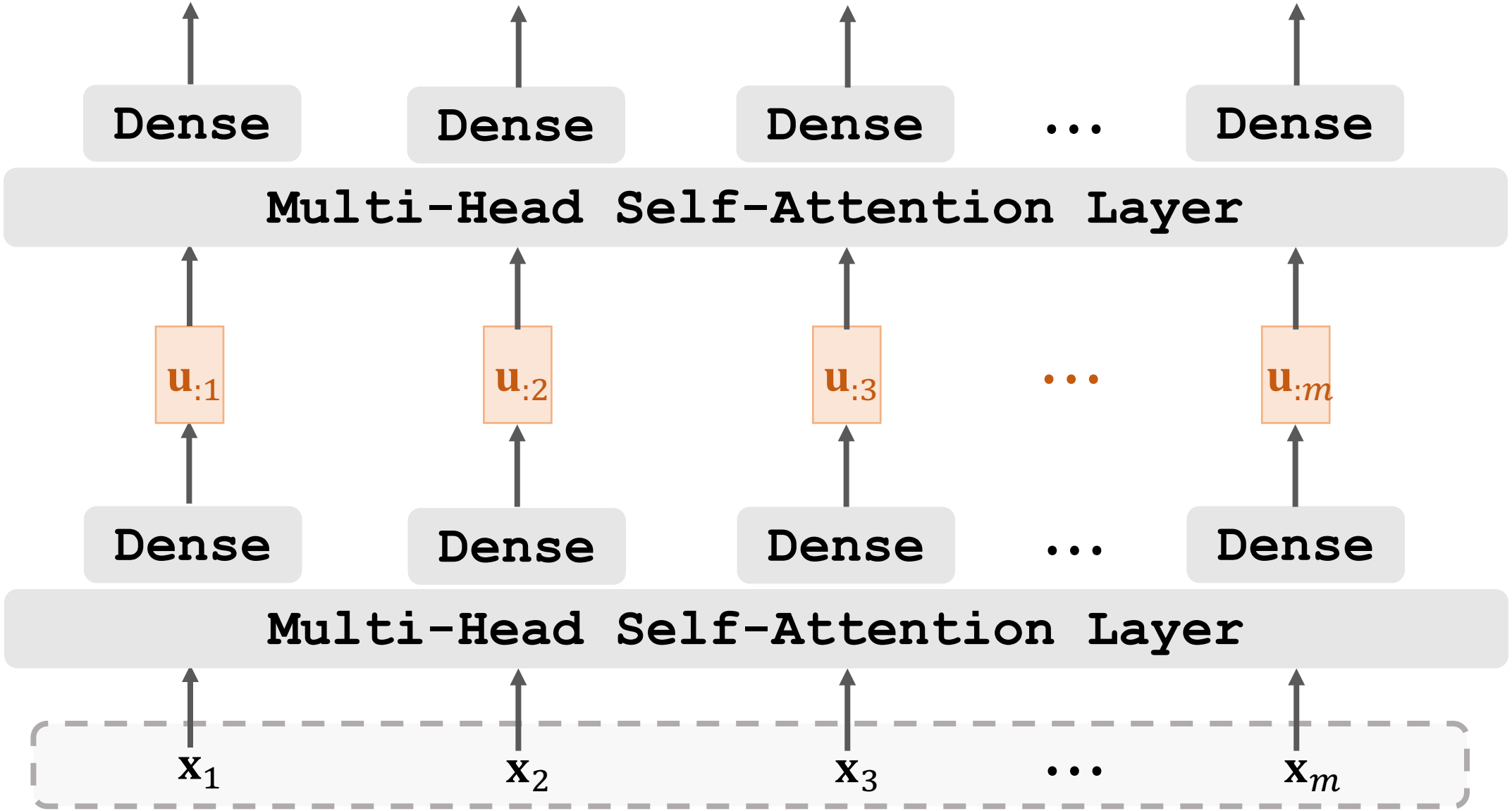
Self-Attention Layer + Dense Layer



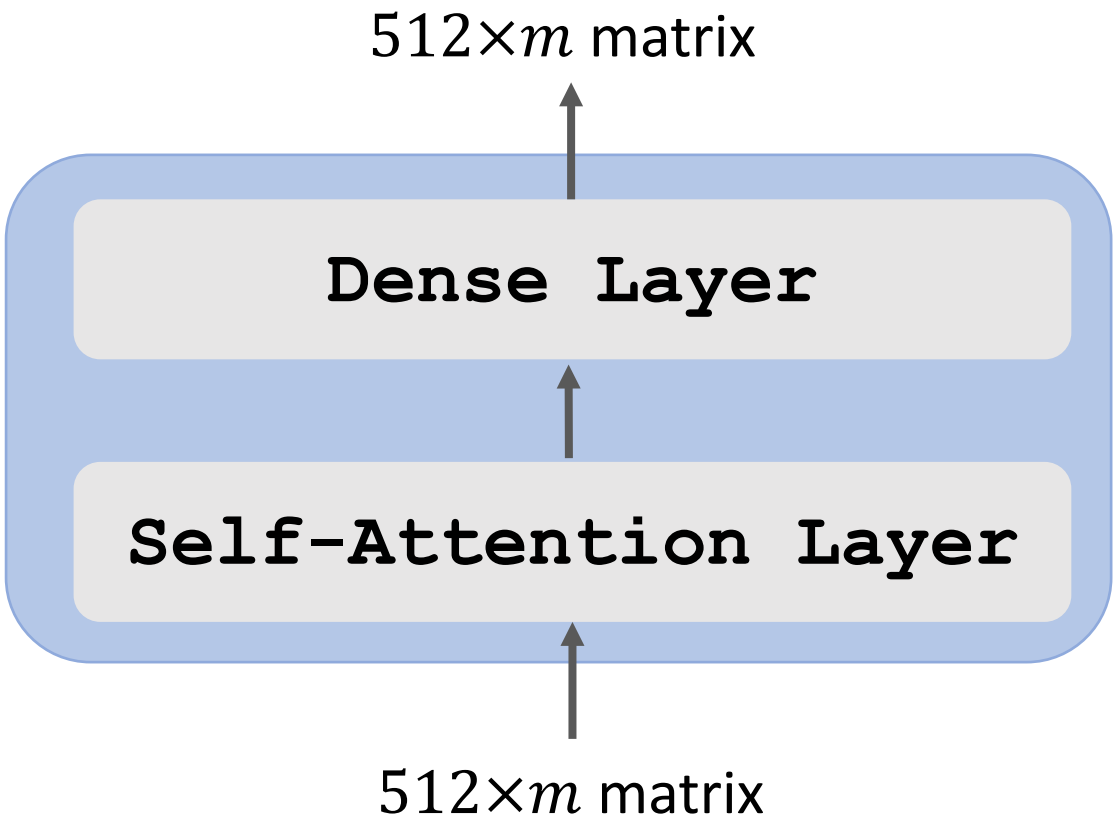
Stacked Self-Attention Layers



Stacked Self-Attention Layers

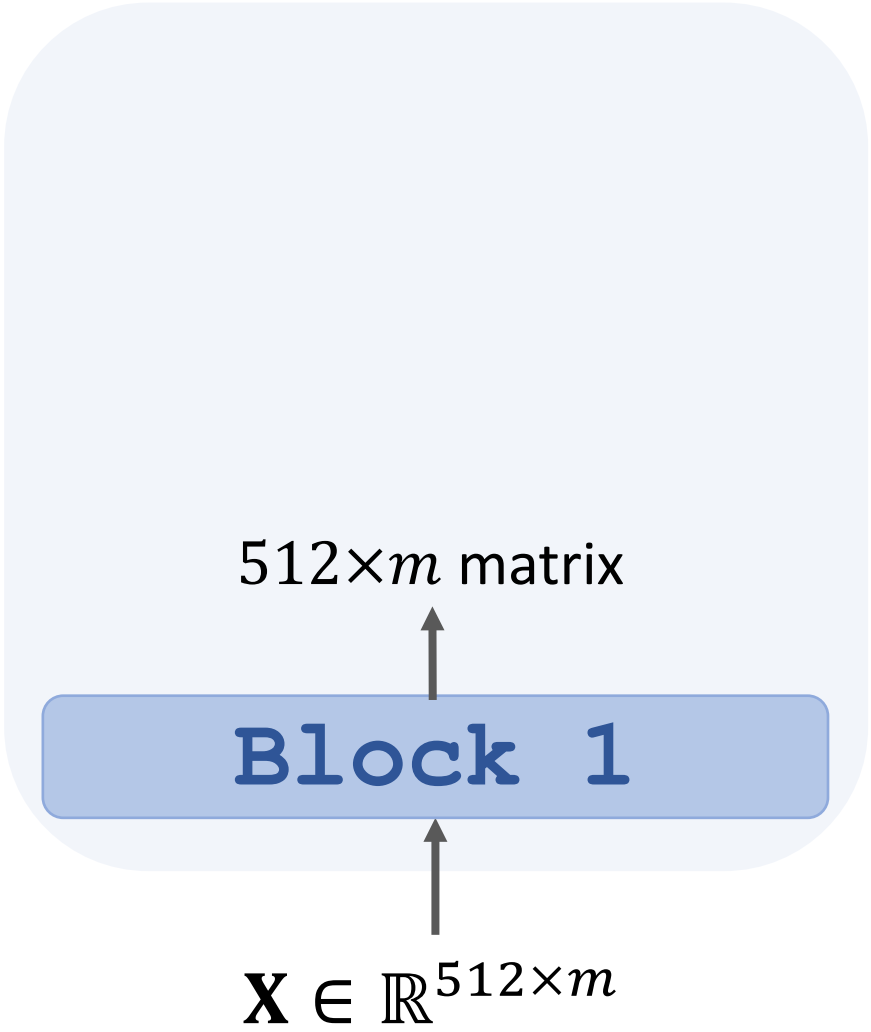


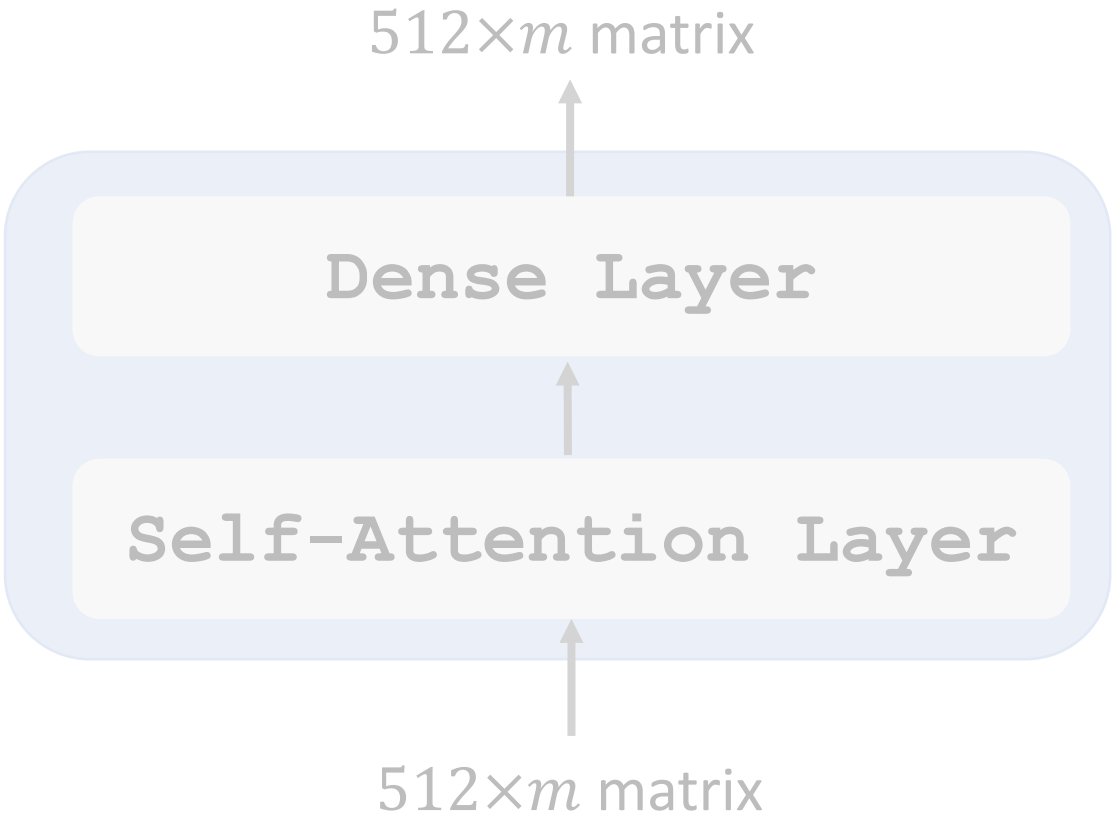
Transformer's Encoder



One Block of Encoder

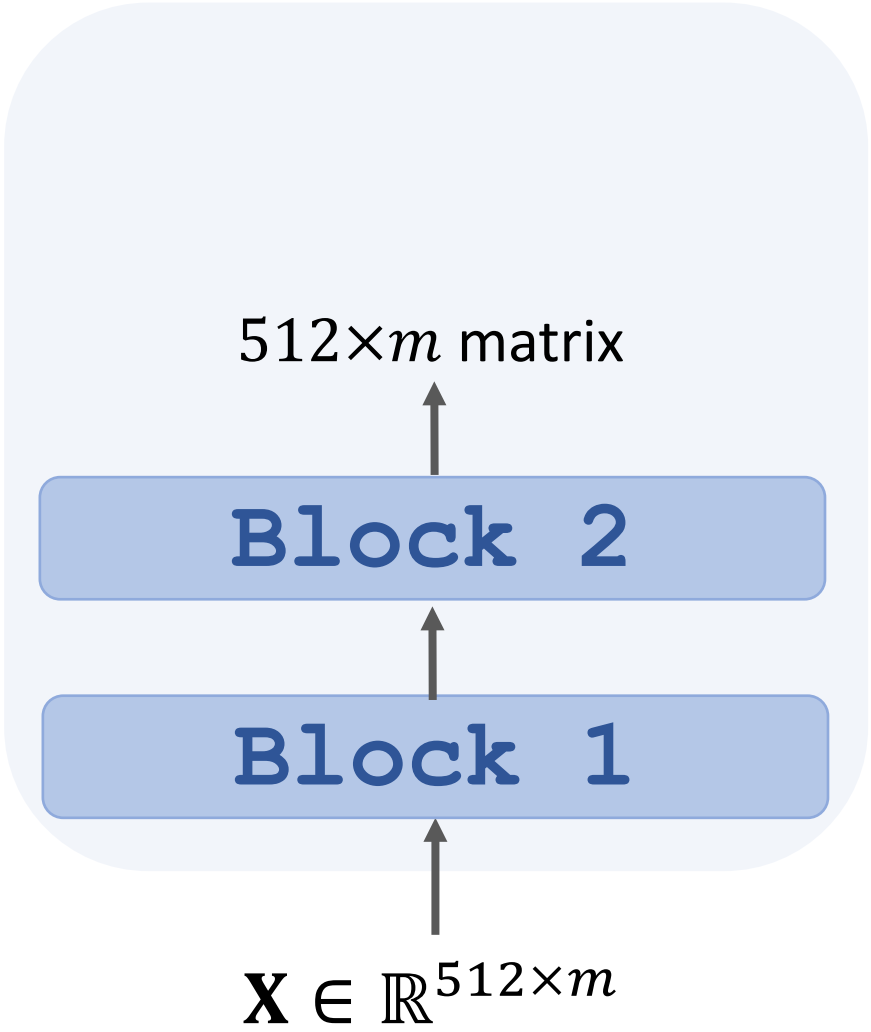
Transformer's Encoder

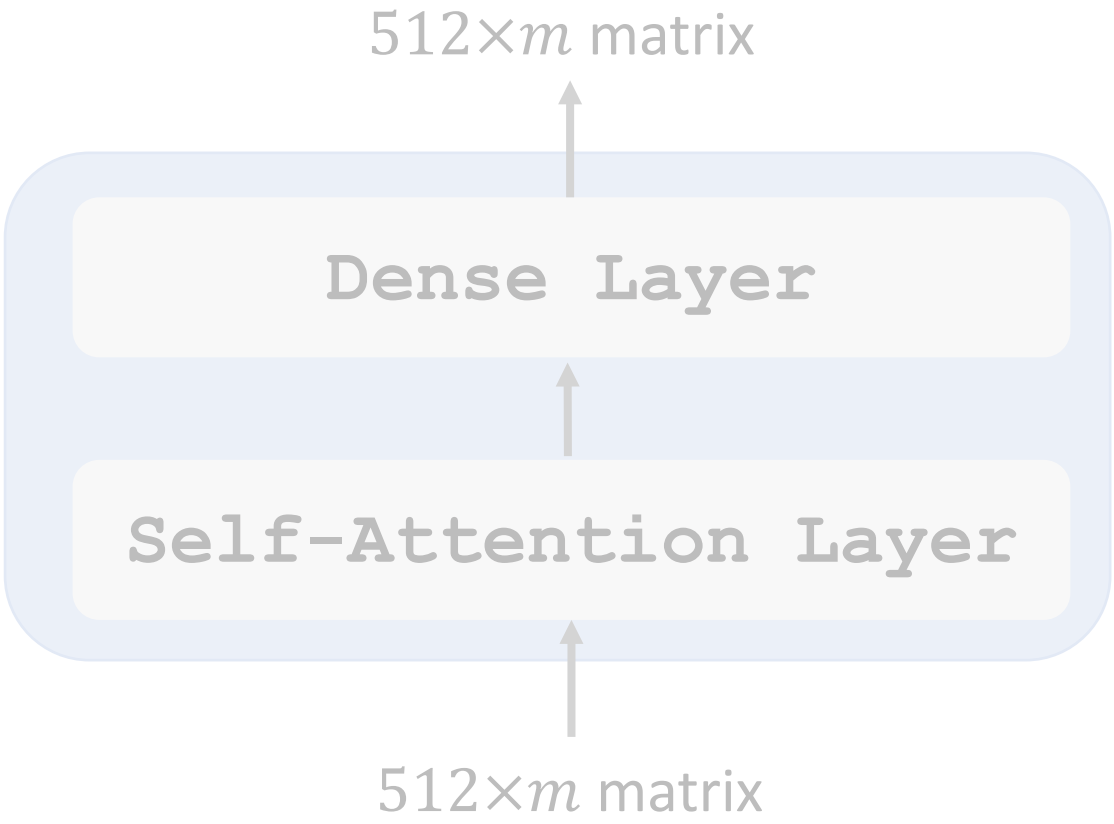




One Block of Encoder

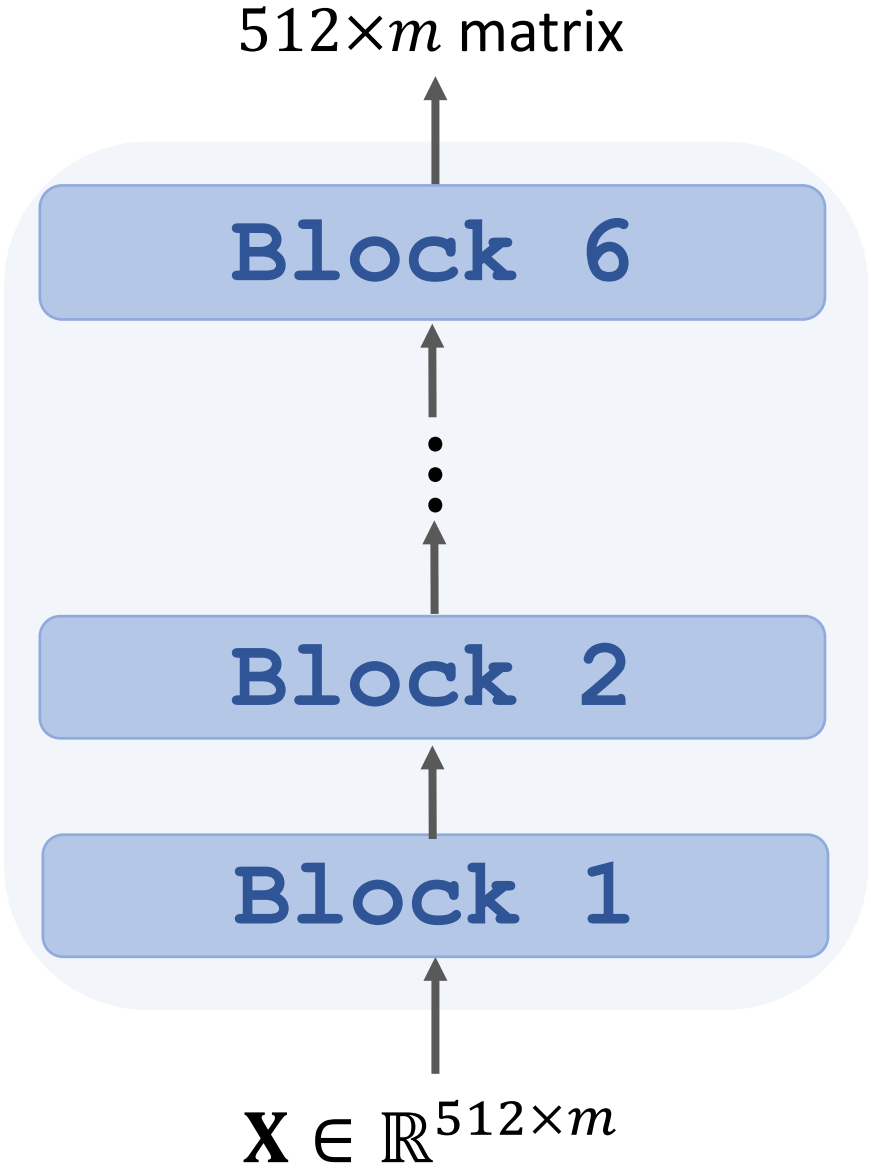
Transformer's Encoder

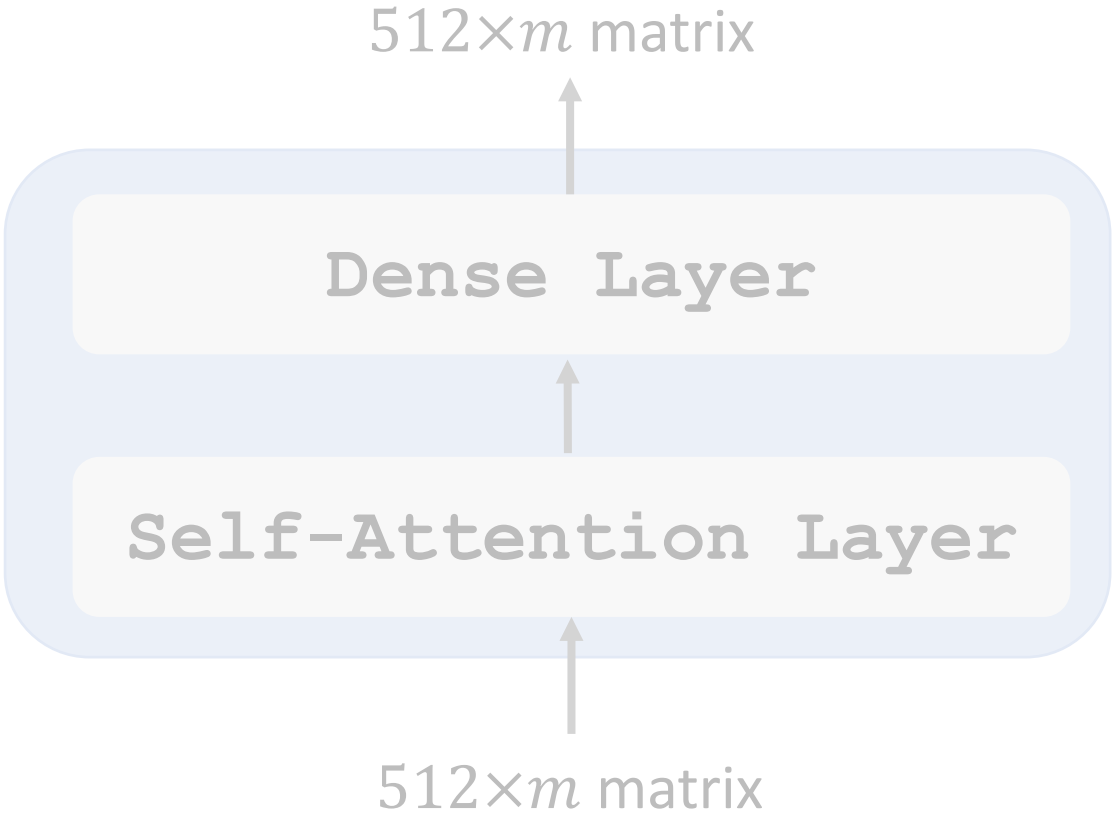




One Block of Encoder

Transformer's Encoder





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:

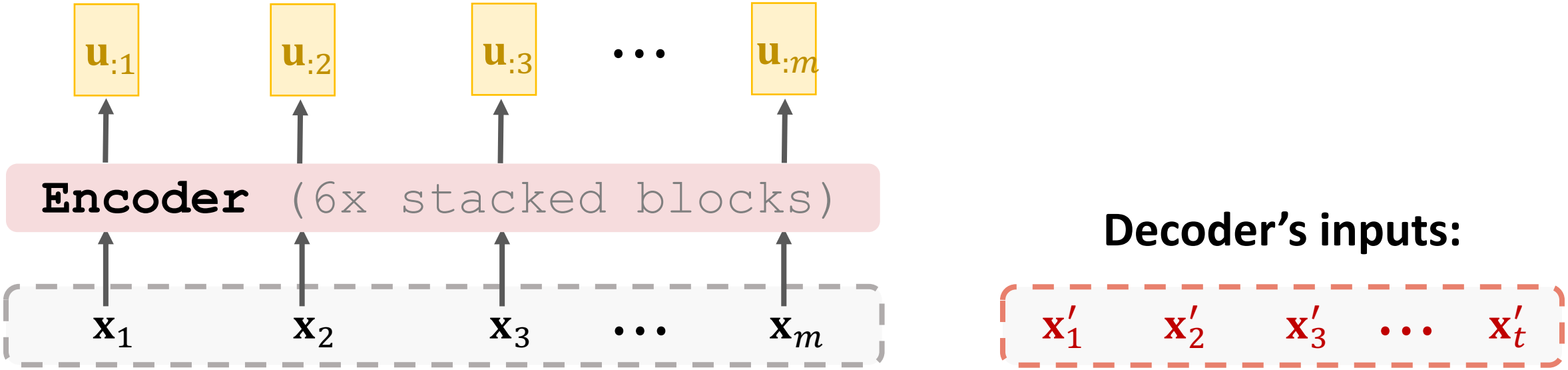
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\cdots$ $\mathbf{x}_m$

Decoder's inputs:

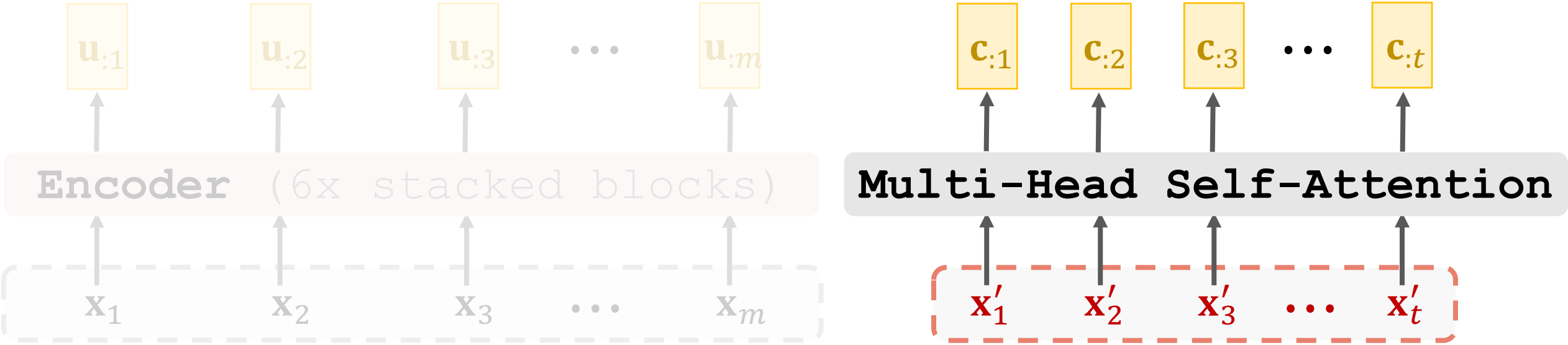
$\mathbf{x}'_1$ $\mathbf{x}'_2$ $\mathbf{x}'_3$ $\cdots$ $\mathbf{x}'_t$

Stacked Attentions

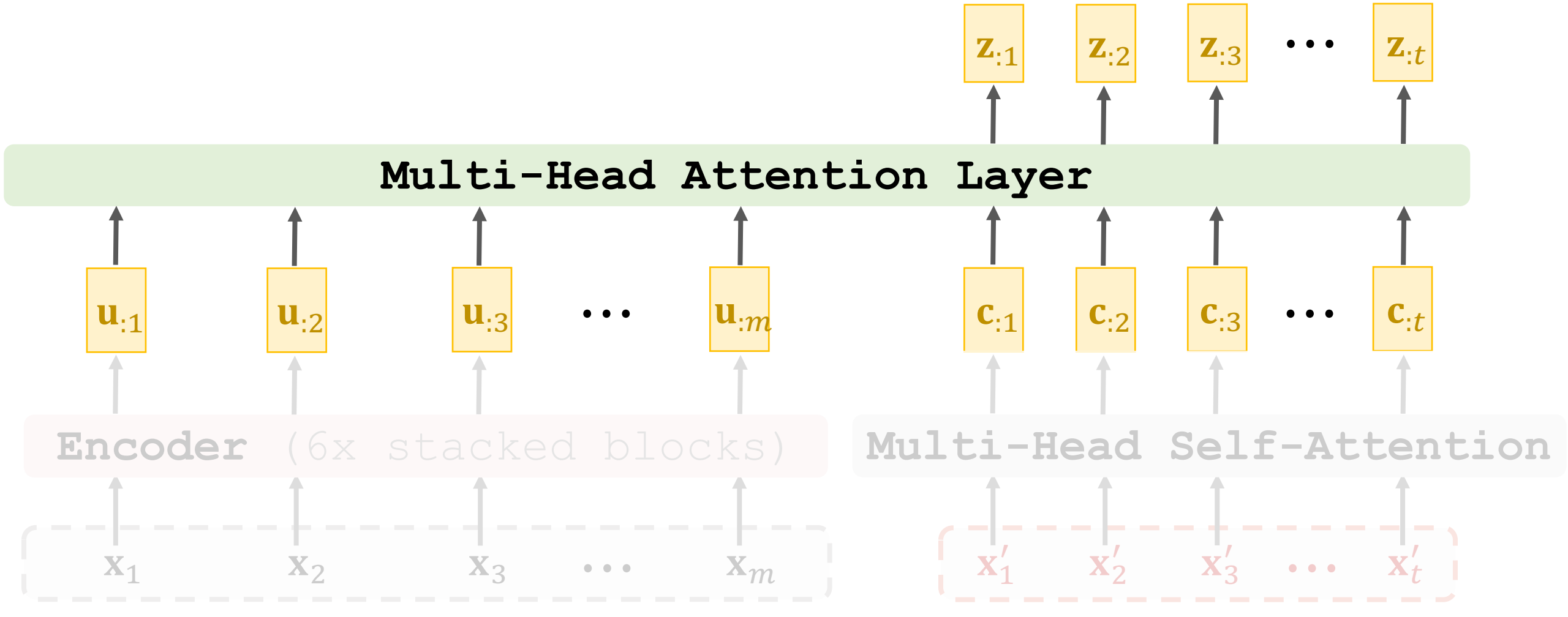
- Transformer's encoder contains 6 stacked blocks.
- 1 block ≈ 1 multi-head attention layer + 1 dense layer.



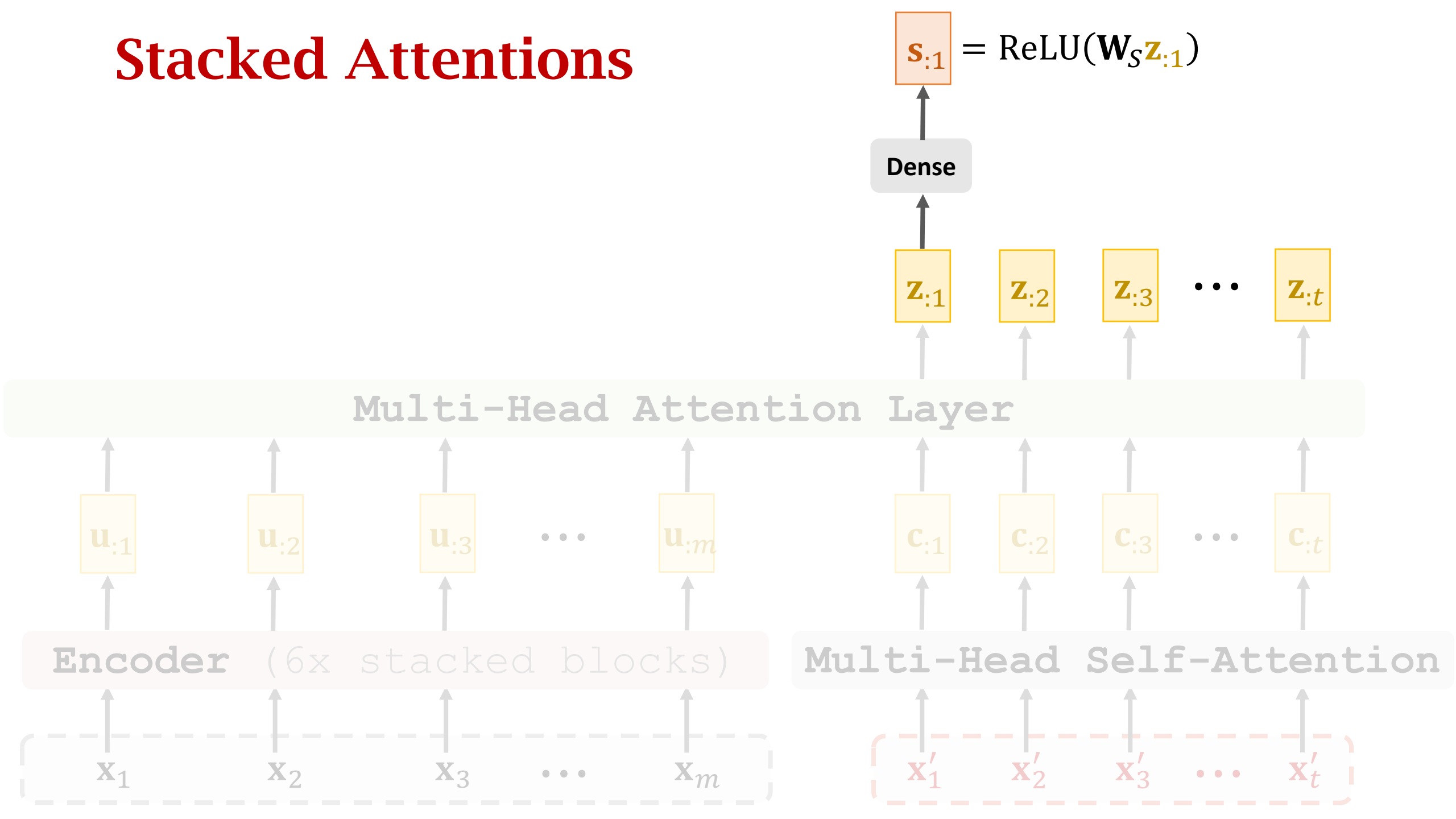
Stacked Attentions



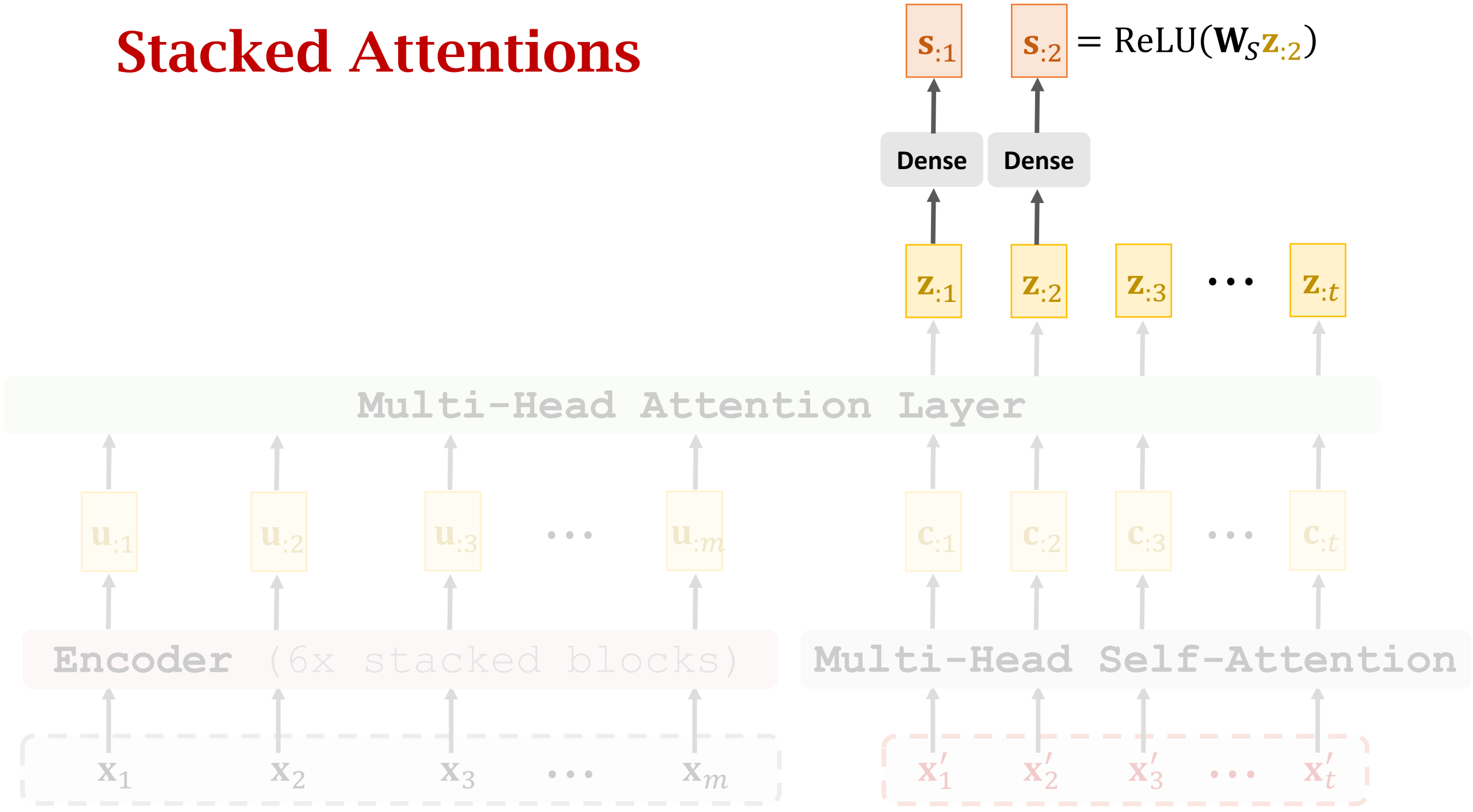
Stacked Attentions



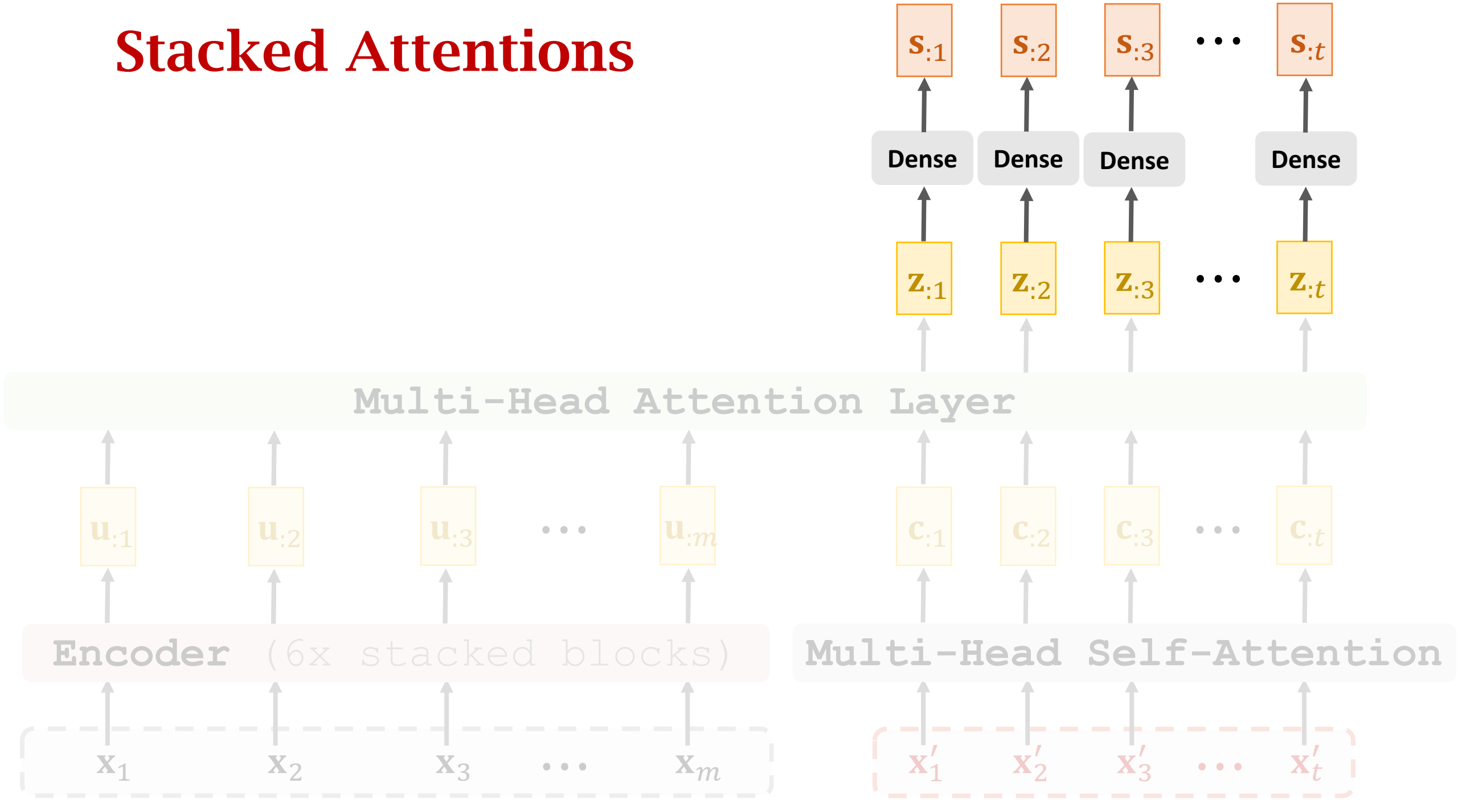
Stacked Attentions



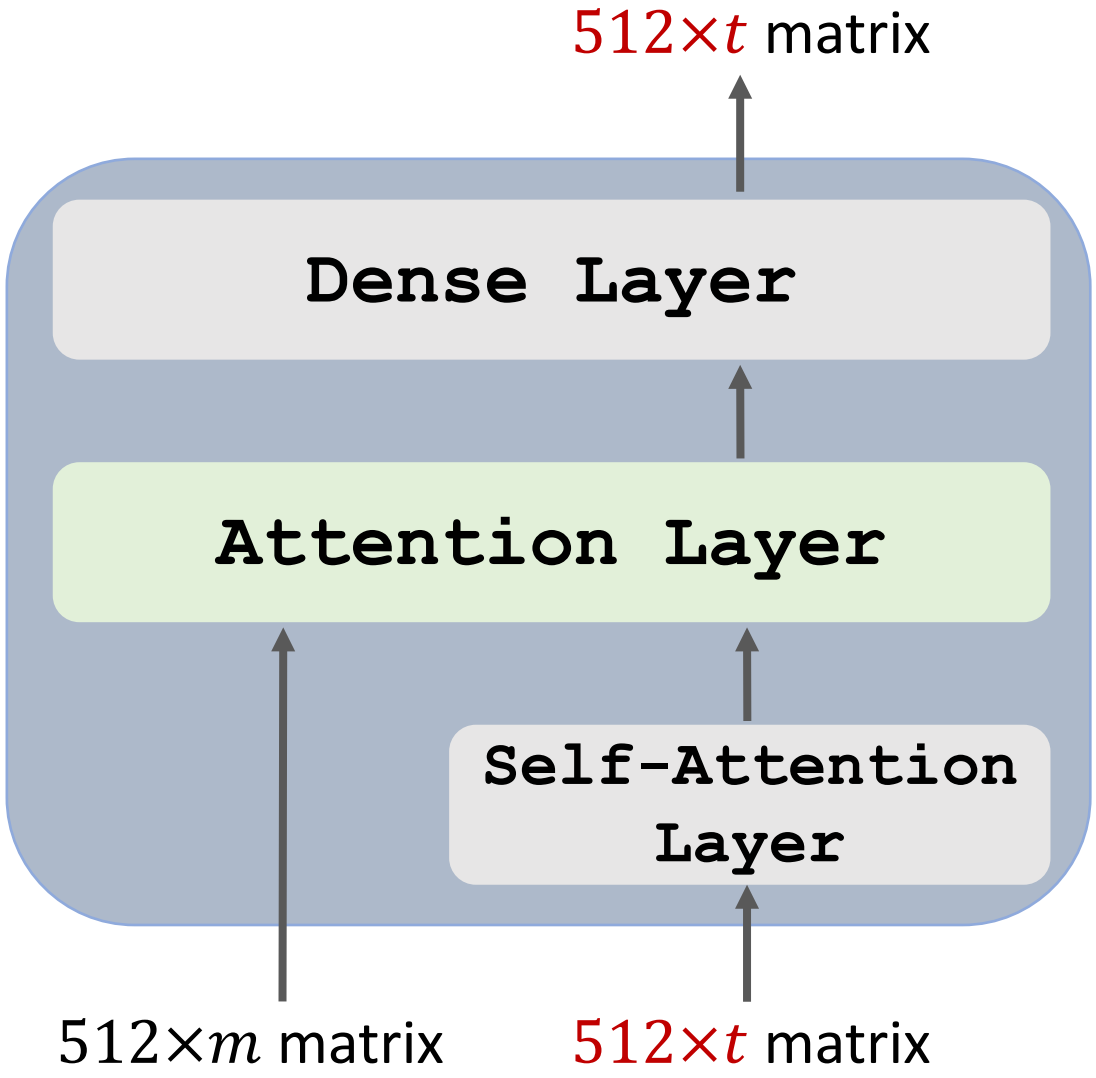
Stacked Attentions



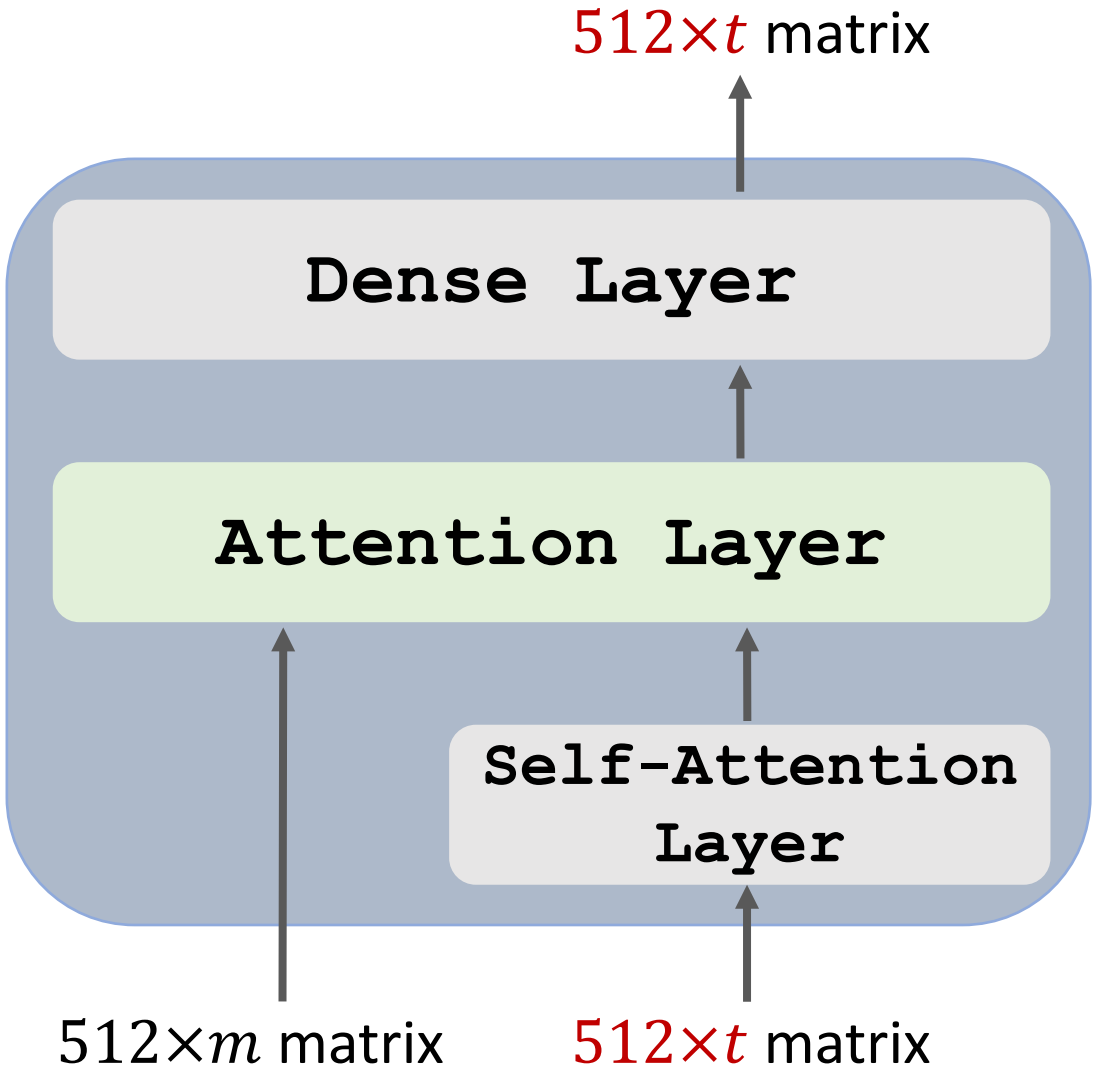
Stacked Attentions

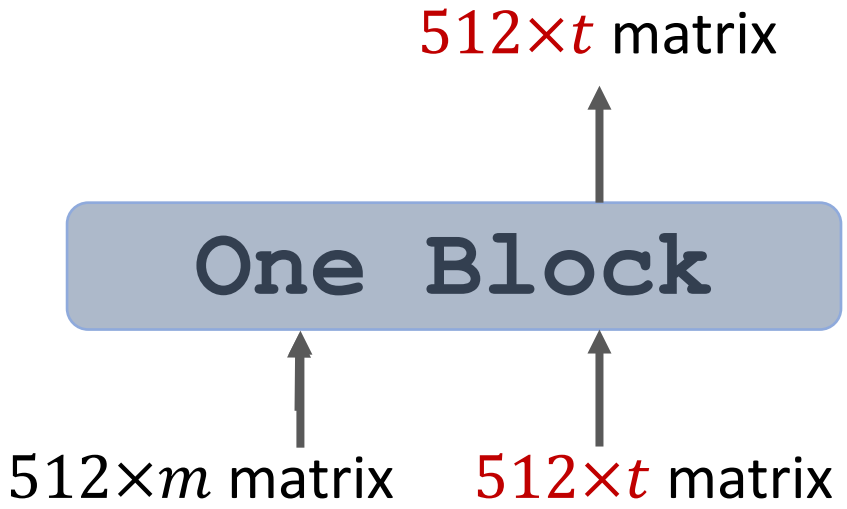


Transformer's Decoder : One Block



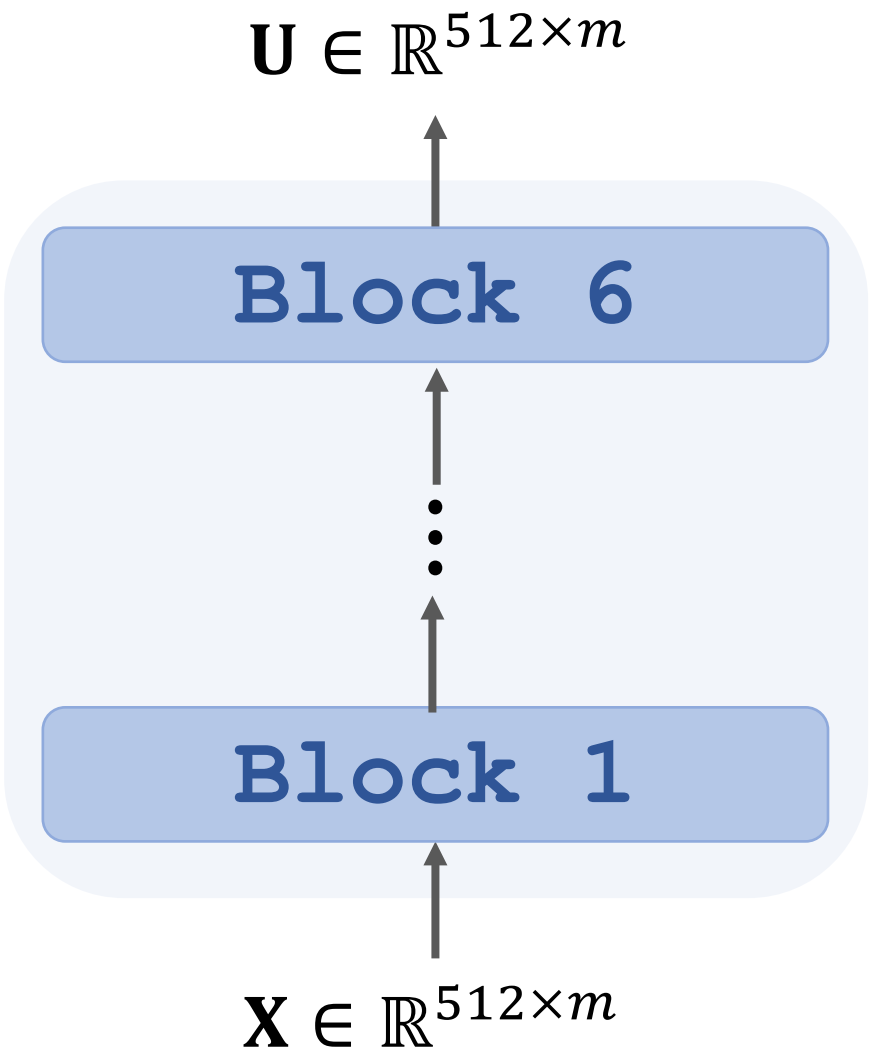
Transformer's Decoder : One Block





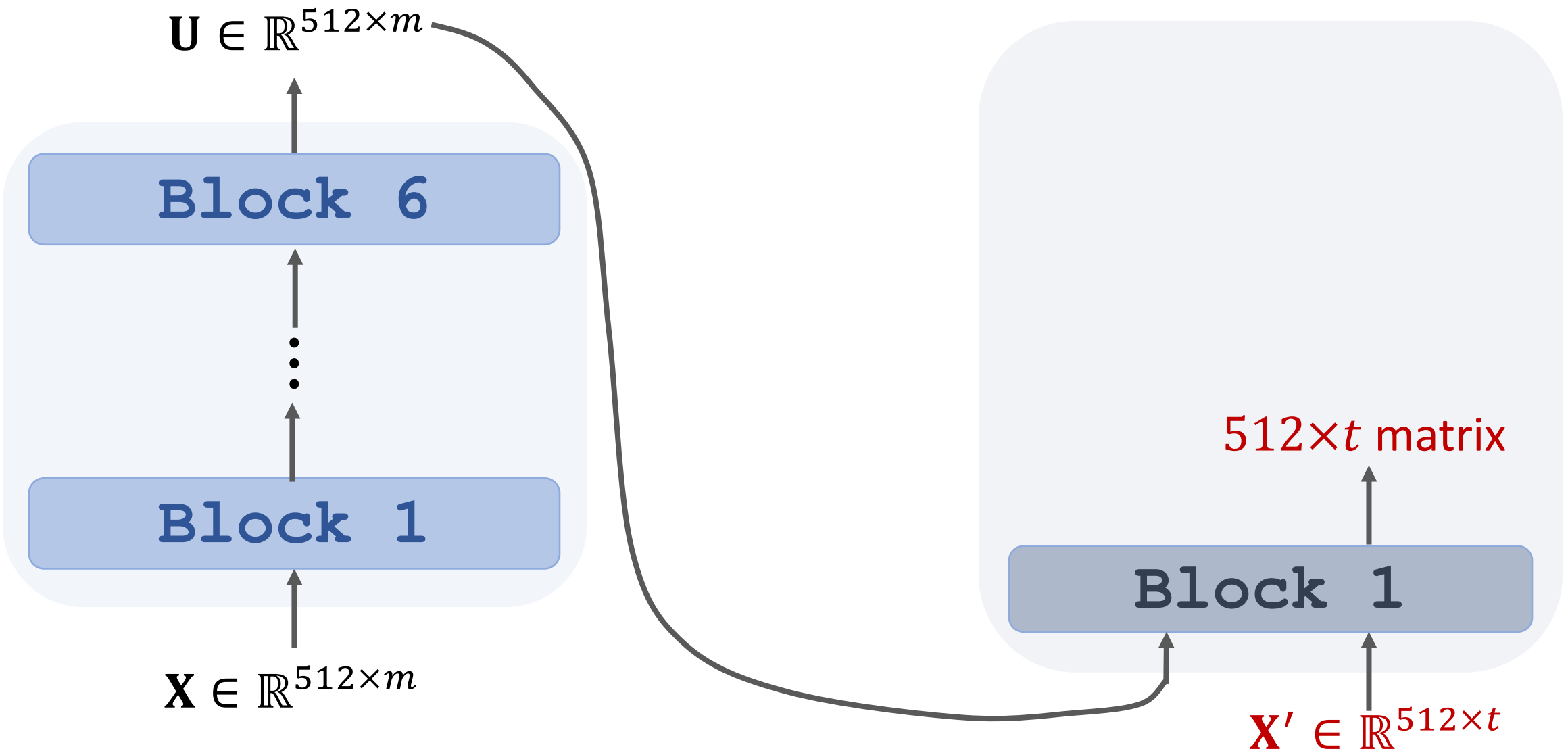
Put Everything Together

Transformer

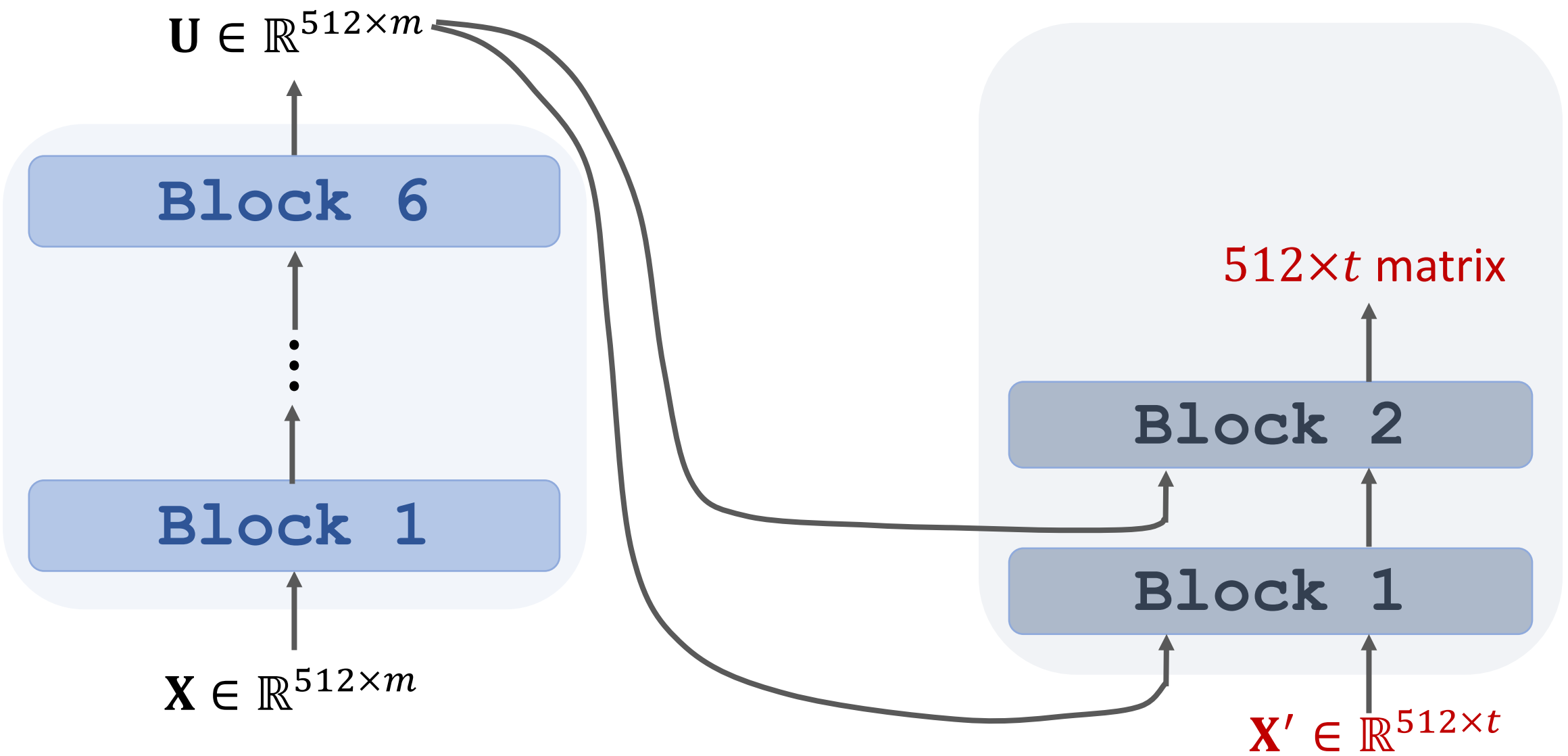


Encoder Network

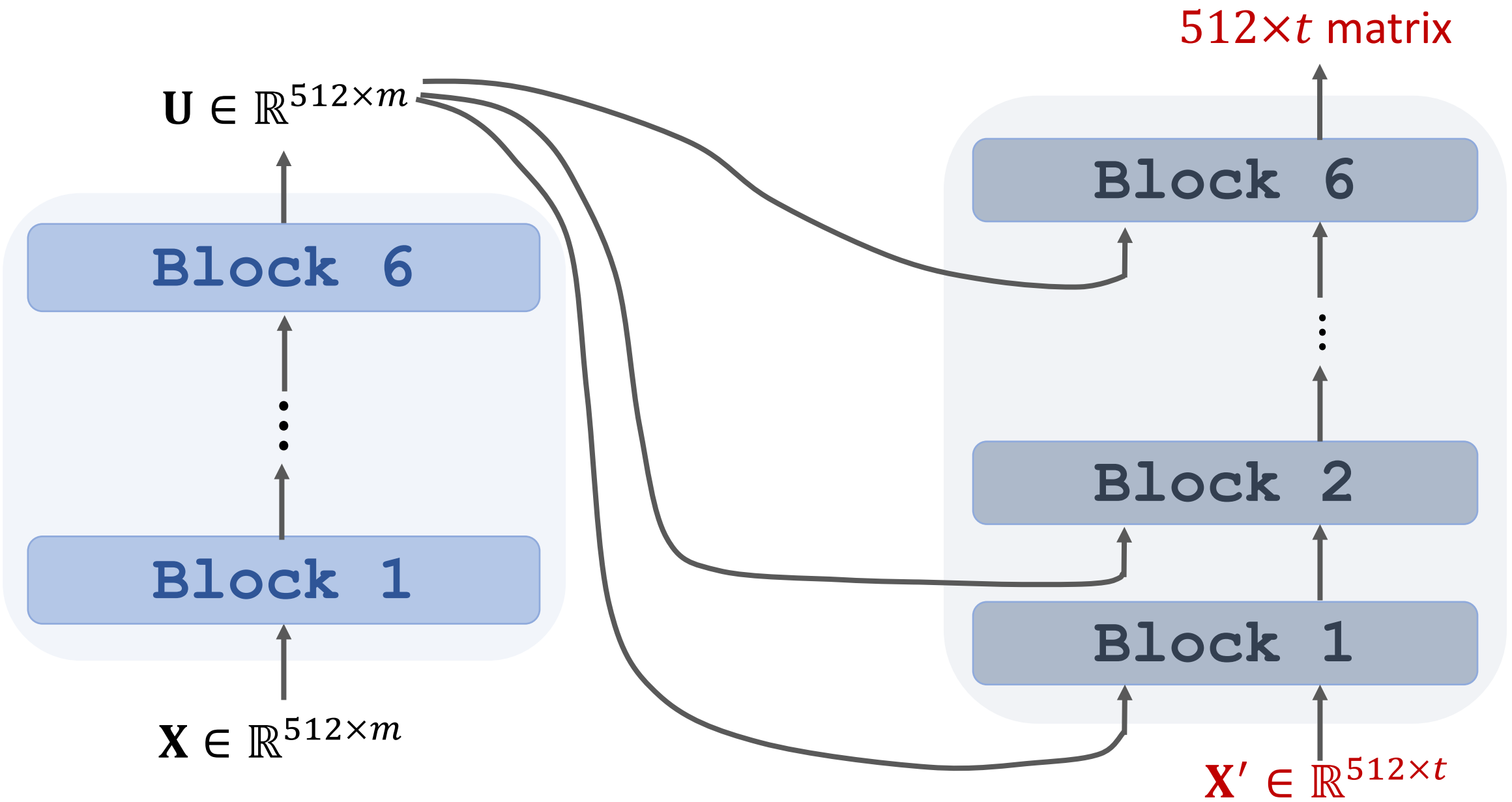
Transformer



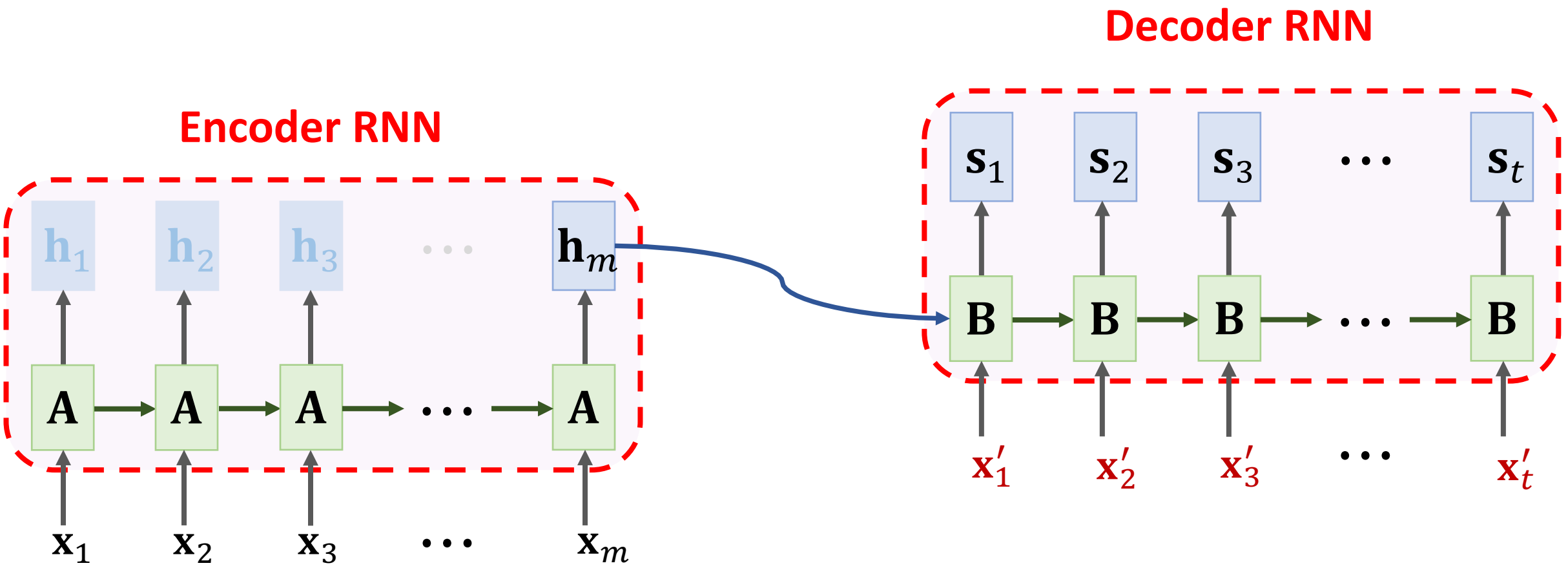
Transformer



Transformer



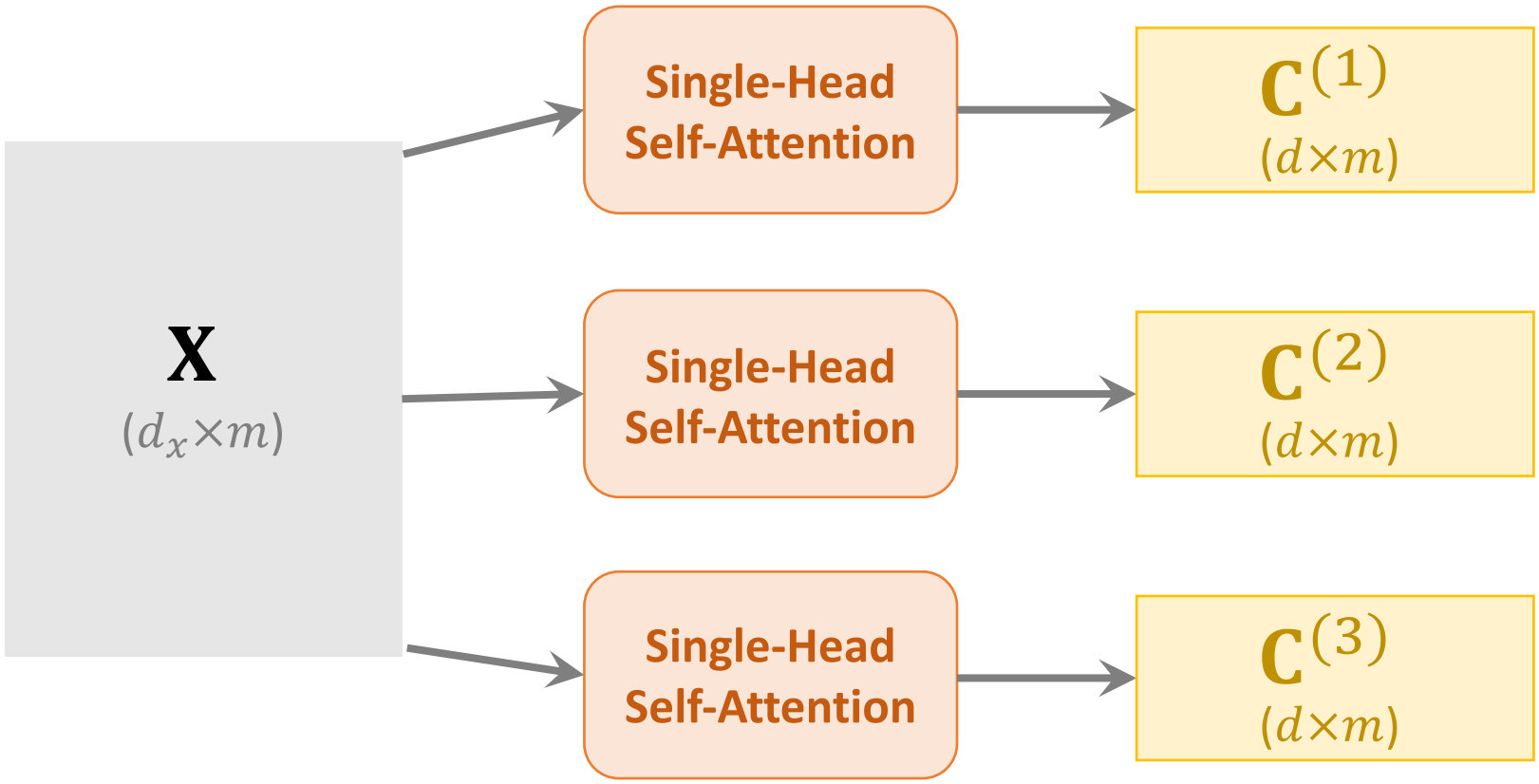
Comparison with RNN Seq2Seq Model



Summary

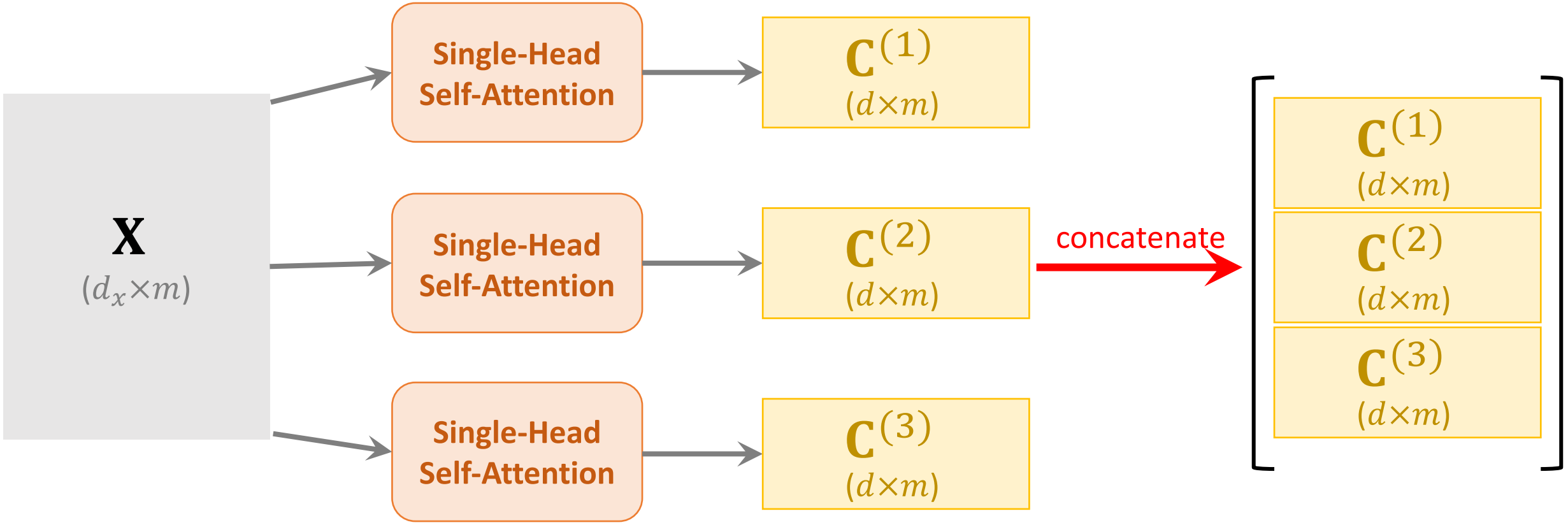
From Single-Head to Multi-Head

- 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 $\approx$ 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 $\approx$ 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.

Thank you!