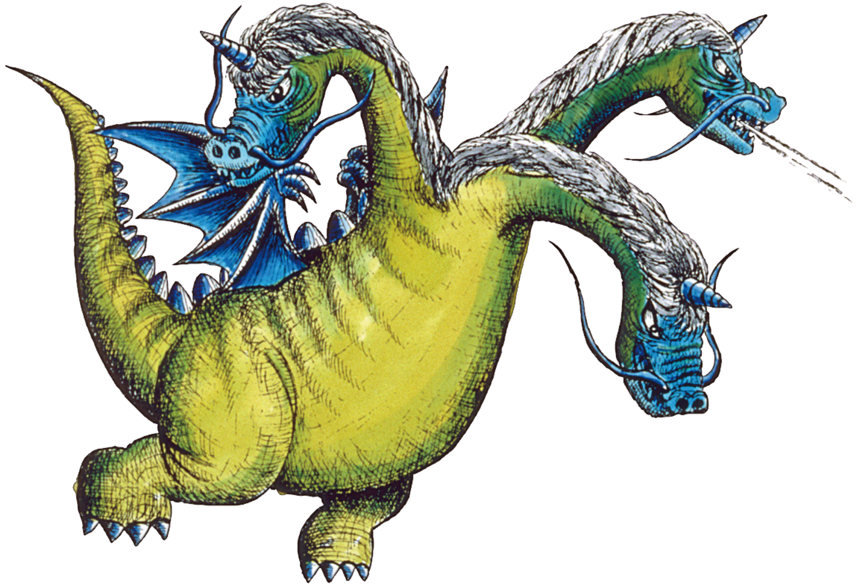
Transformer Model (2/2): From Shallow to Deep

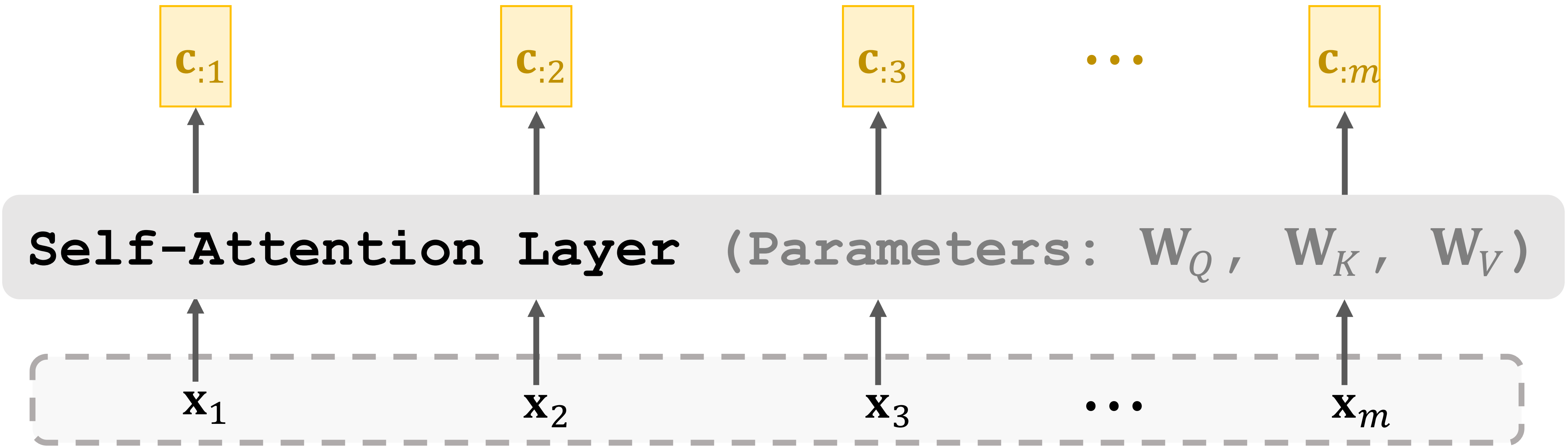
Shusen Wang

Multi-Head Attention



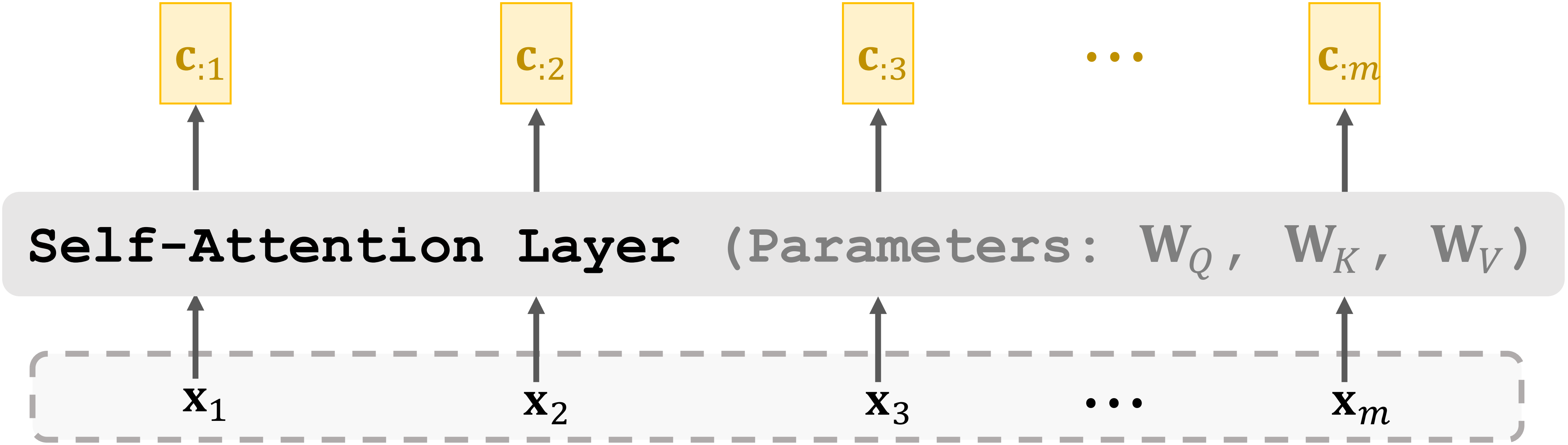
Single-Head Self-Attention

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



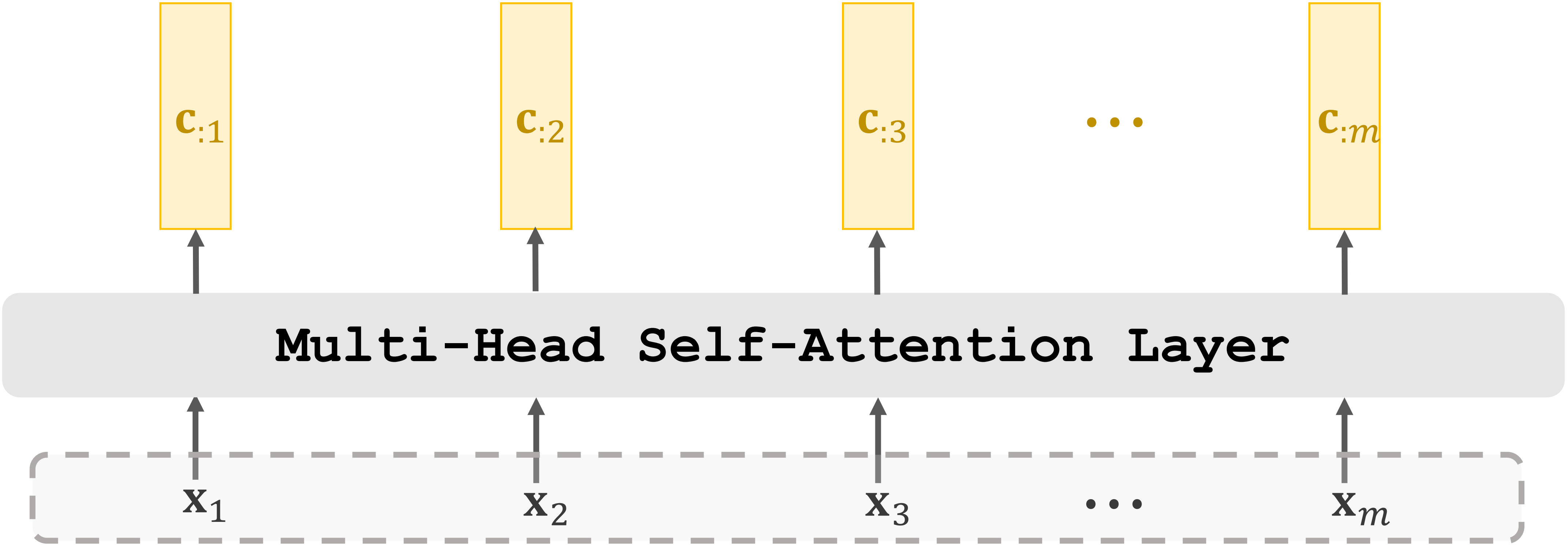
Multi-Head Self-Attention

- 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.



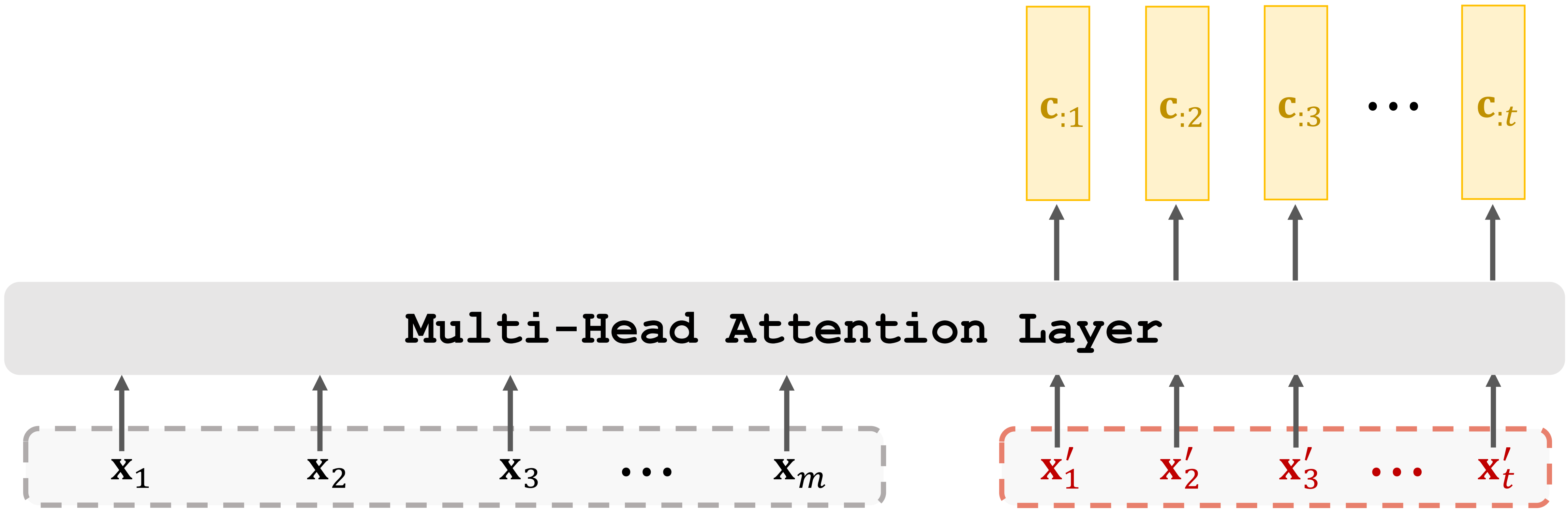
Multi-Head Self-Attention

- 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: $(ld) \times m$.



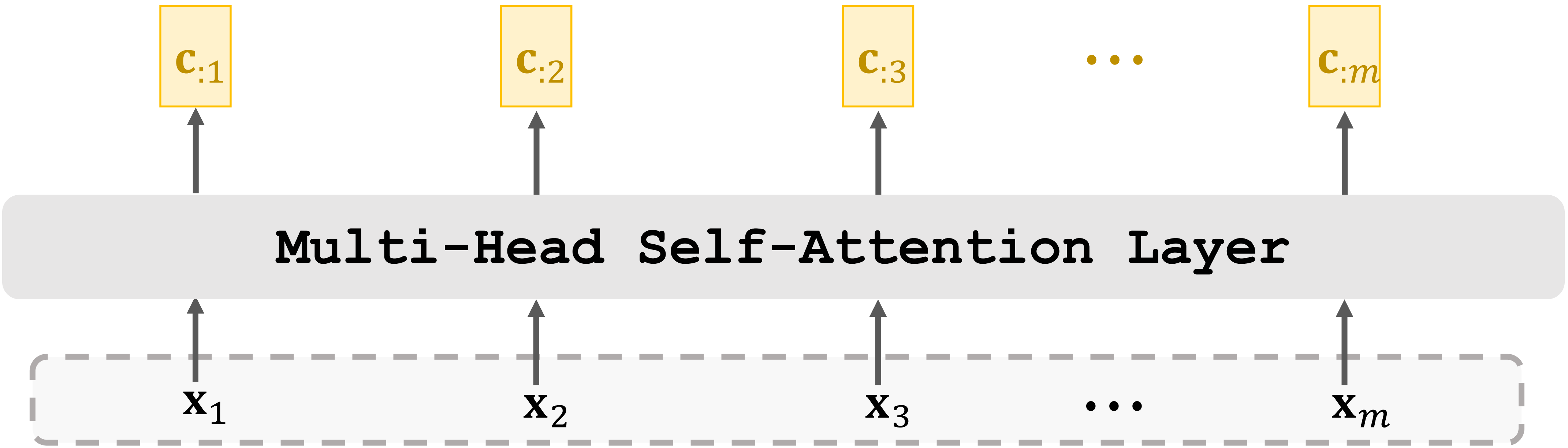
Multi-Head Attention

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

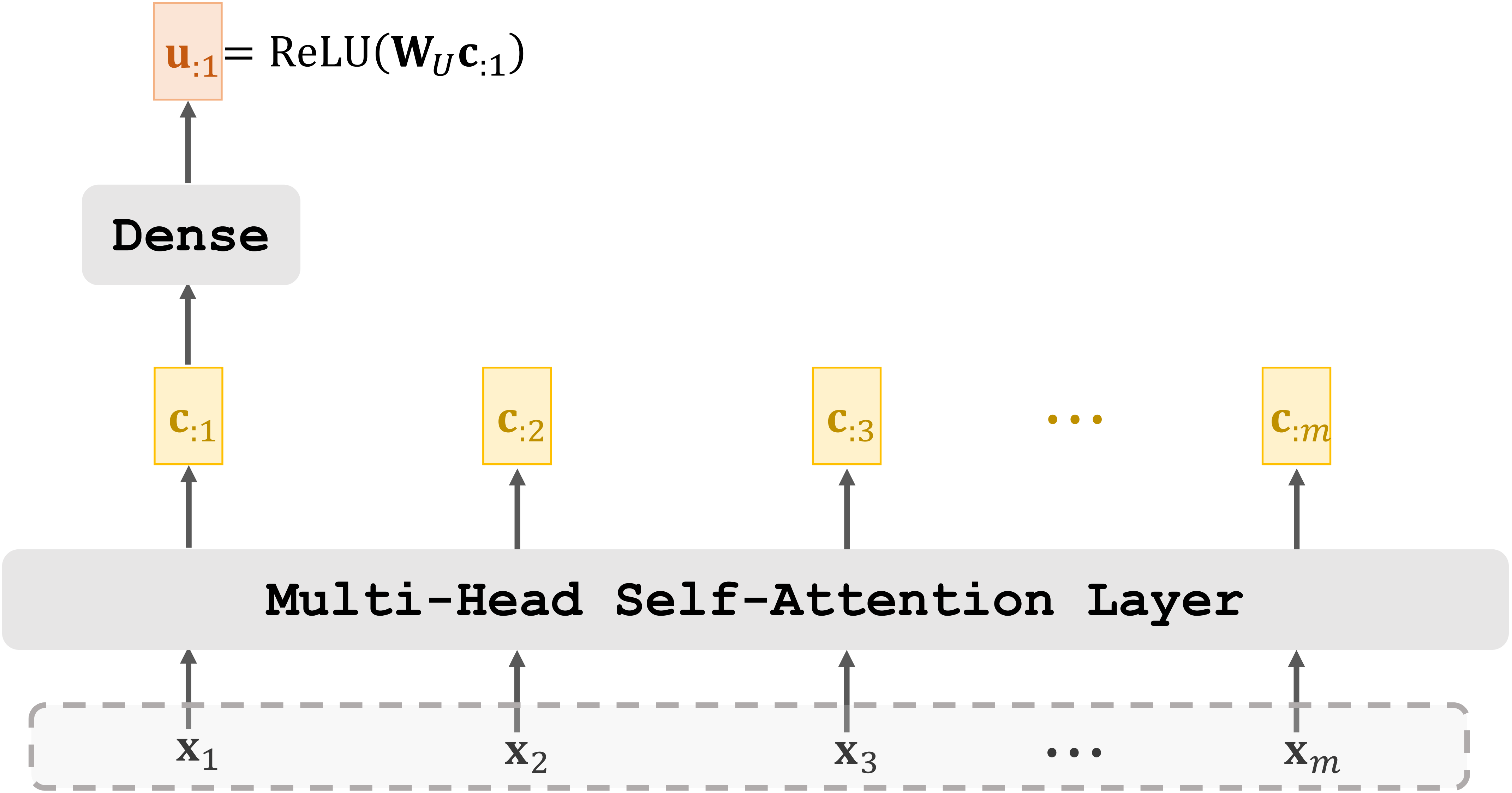


Stacked Self-Attention Layers

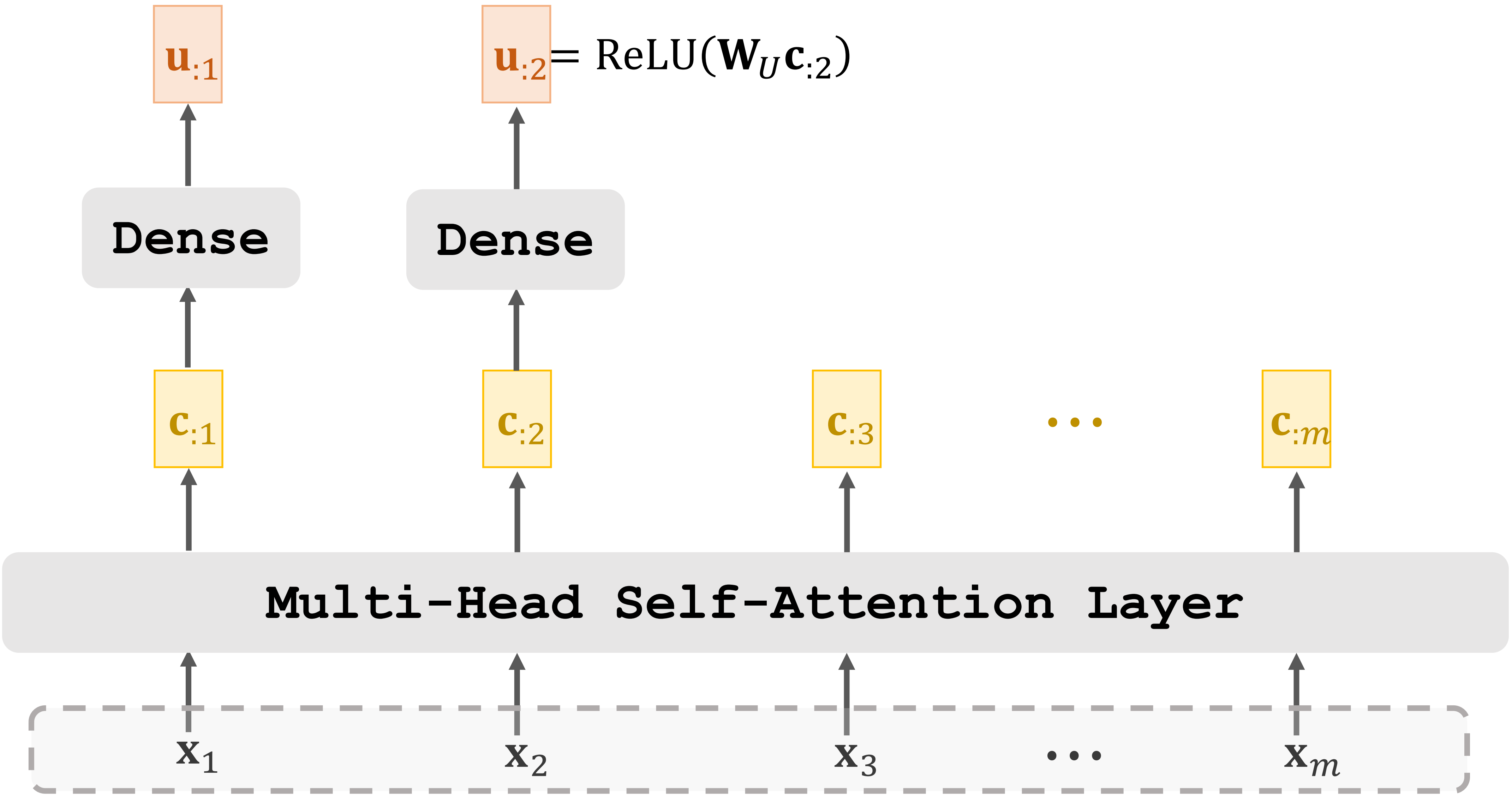
Self-Attention Layer



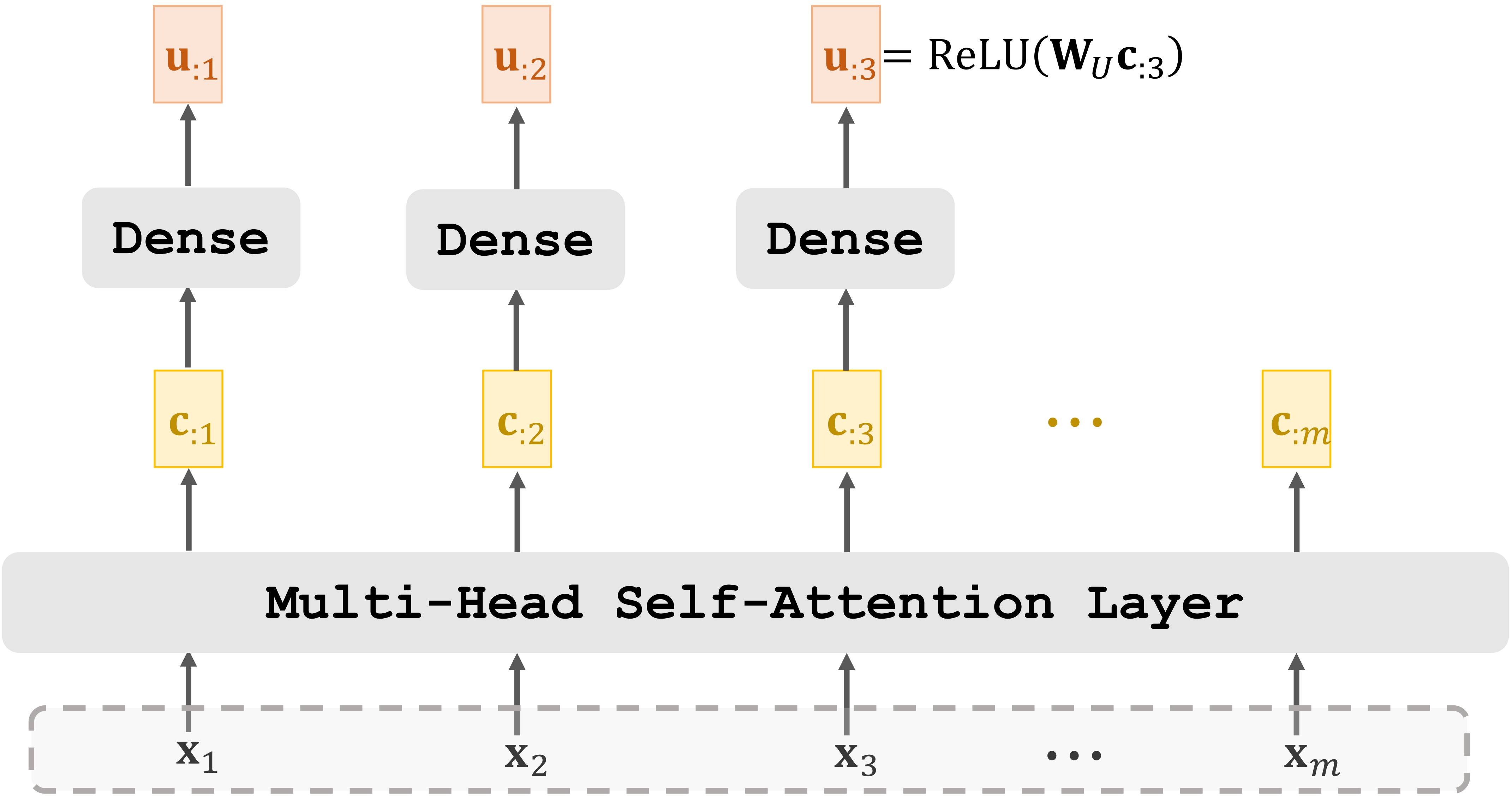
Self-Attention Layer + Dense Layer



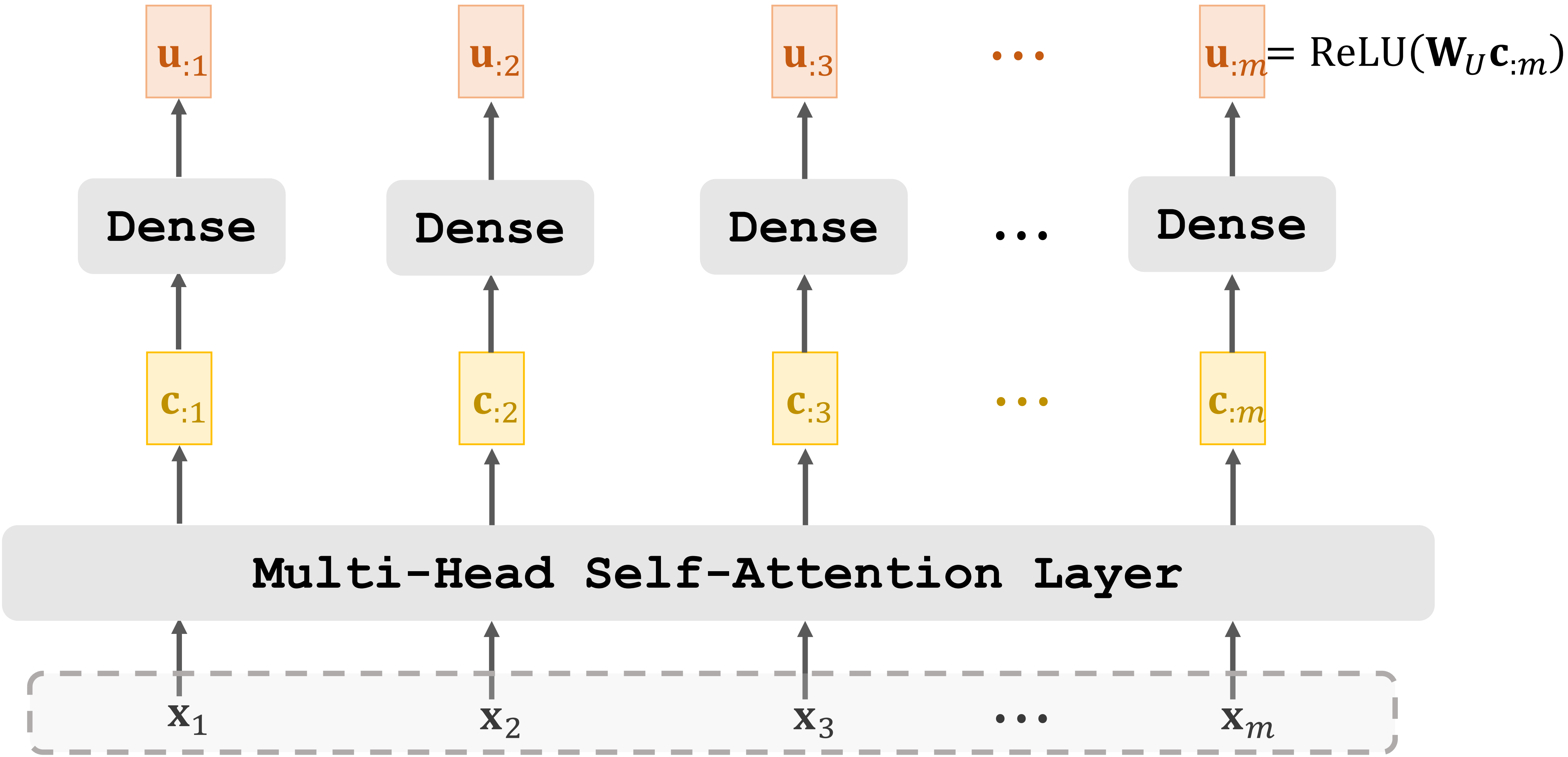
Self-Attention Layer + Dense Layer



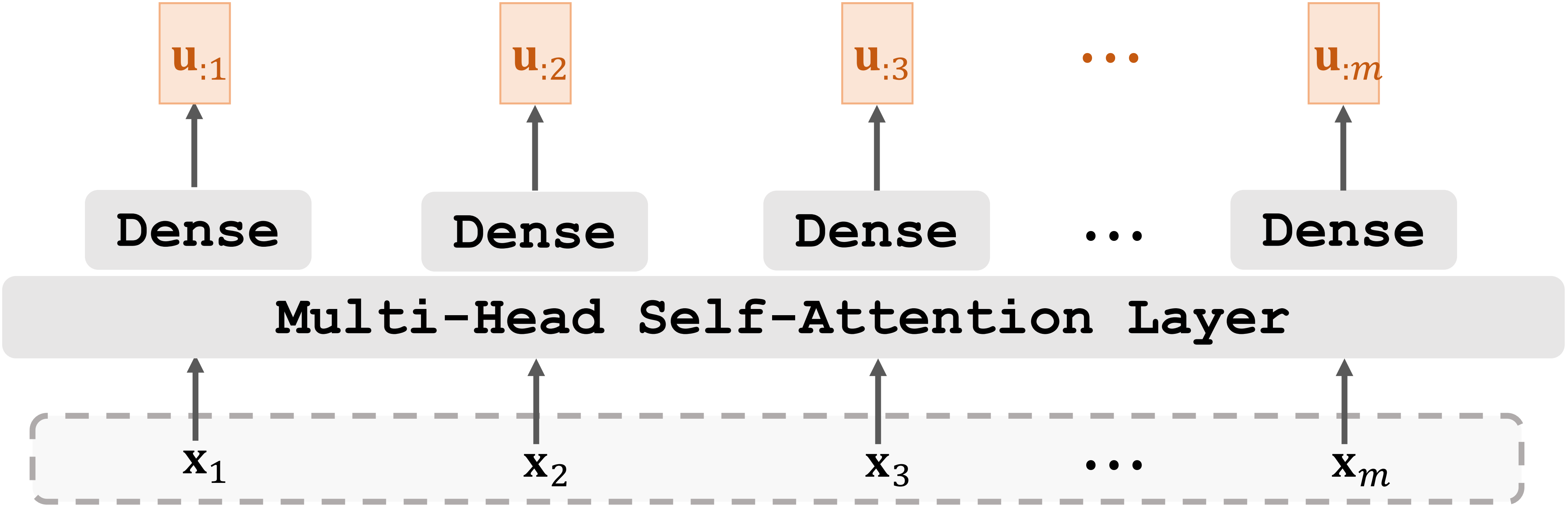
Self-Attention Layer + Dense Layer



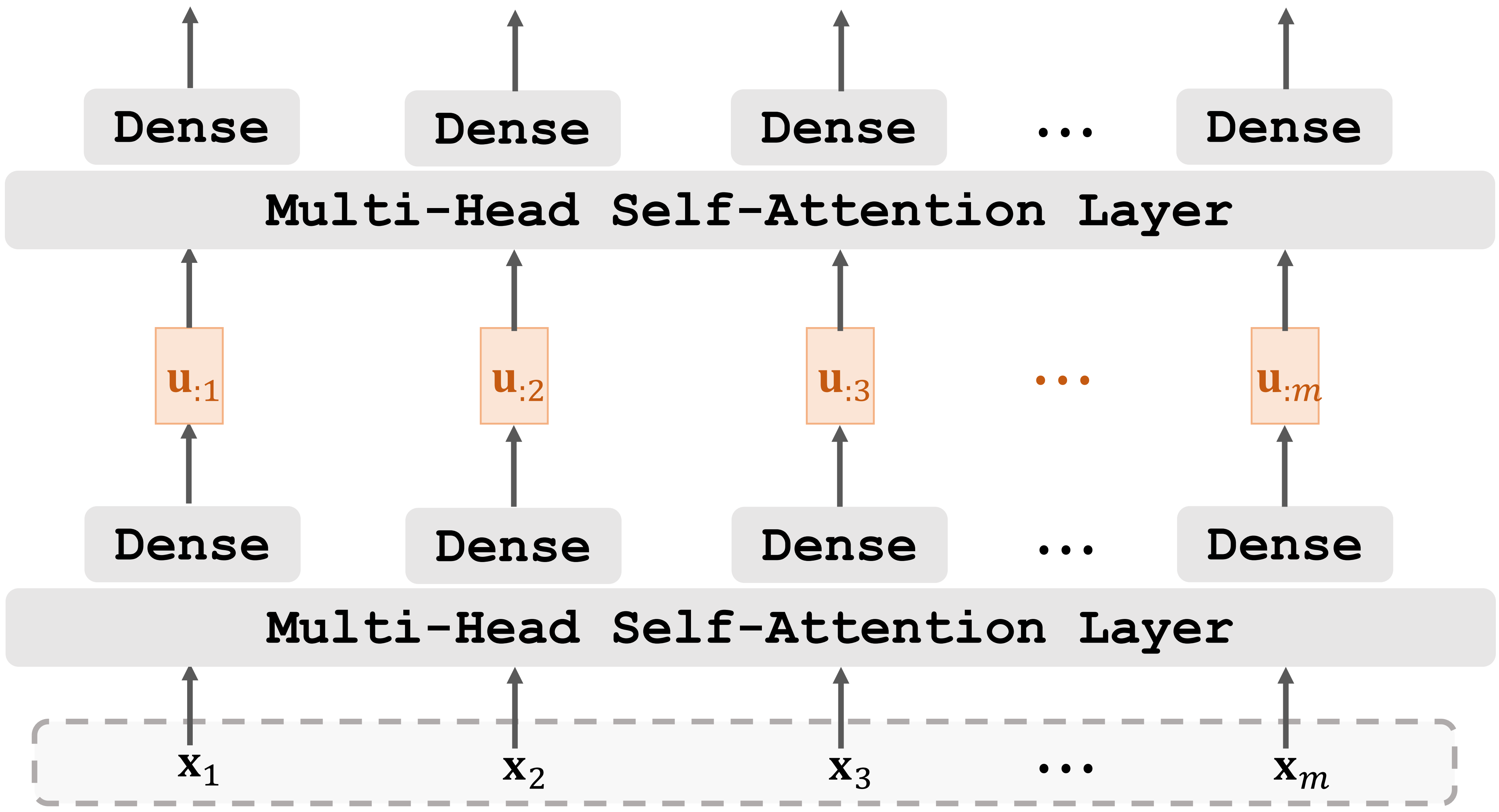
Self-Attention Layer + Dense Layer



Stacked Self-Attention Layers



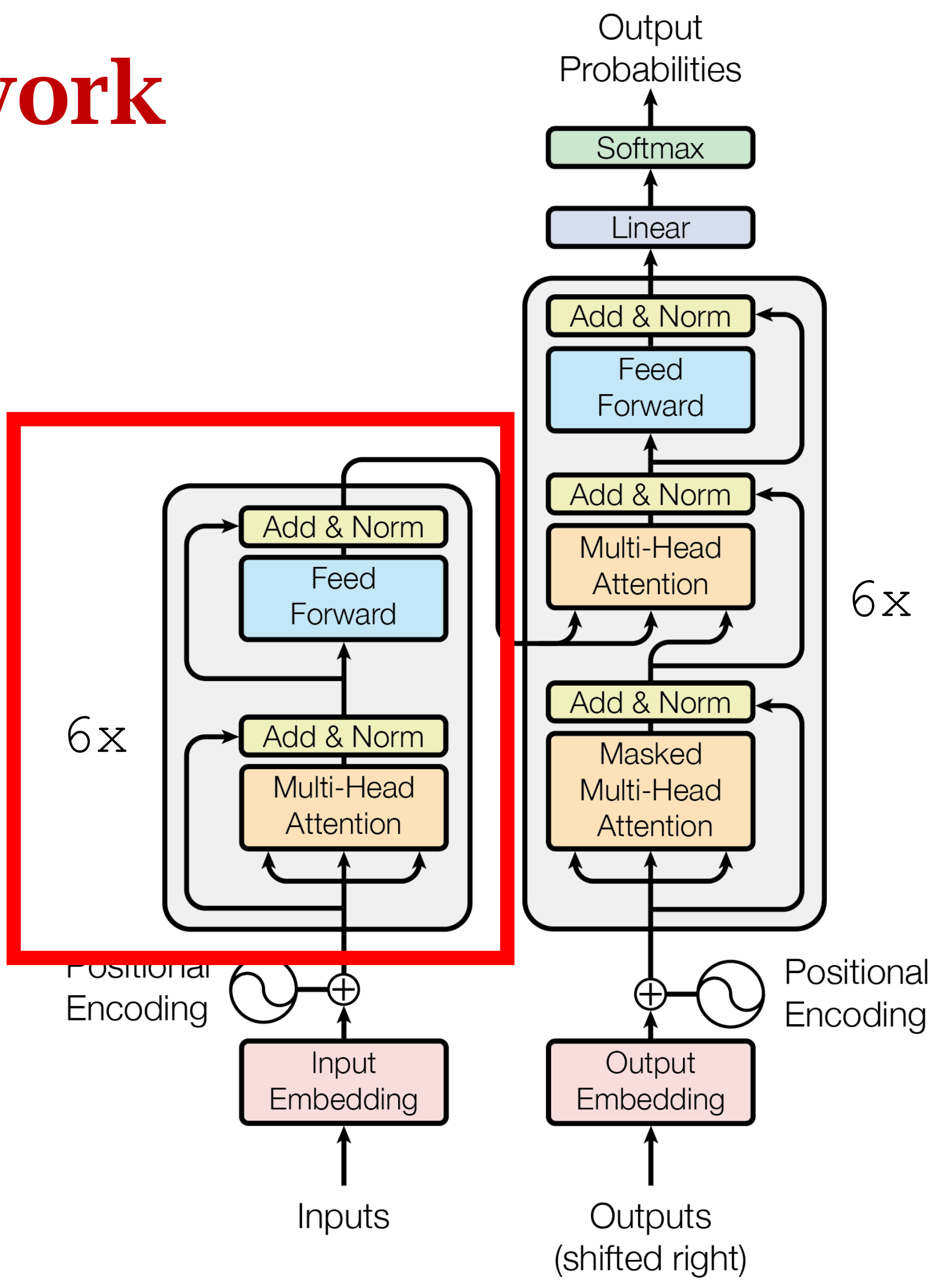
Stacked Self-Attention Layers



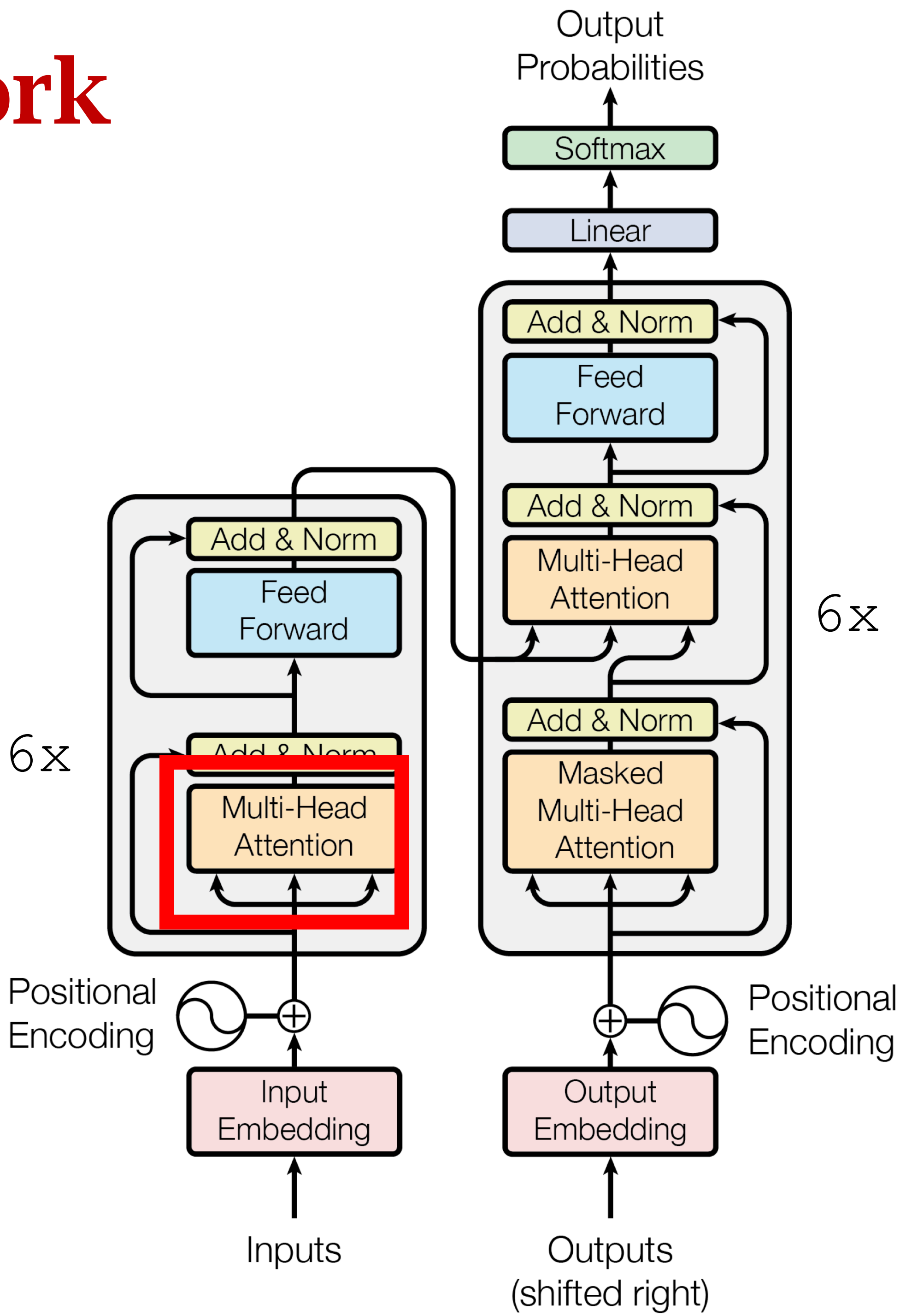
Encoder of Transformer

Encoder Network

- 1 block = self-attention + dense.
- Encoder is a stack of 6 such blocks.
- Other tricks:
  - Skip connection.
  - Normalization.



Encoder Network



Multi-head self-attention:

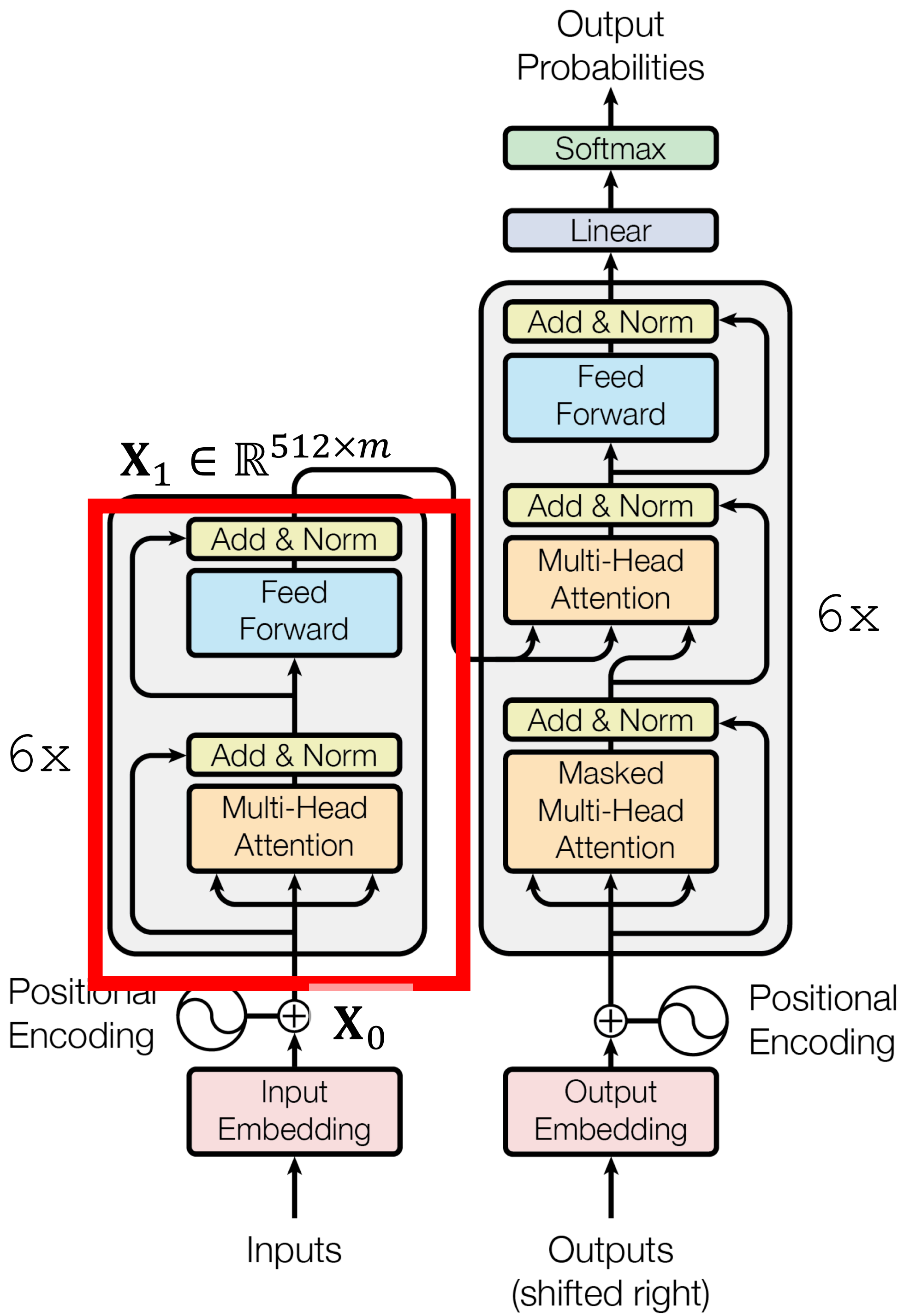
- Input shape: $512 \times m$.
- Use 8 single-head attentions.
- Every single-head attention outputs a $64 \times m$ matrix.
- Thus the output shape is $512 \times m$.

Encoder Network

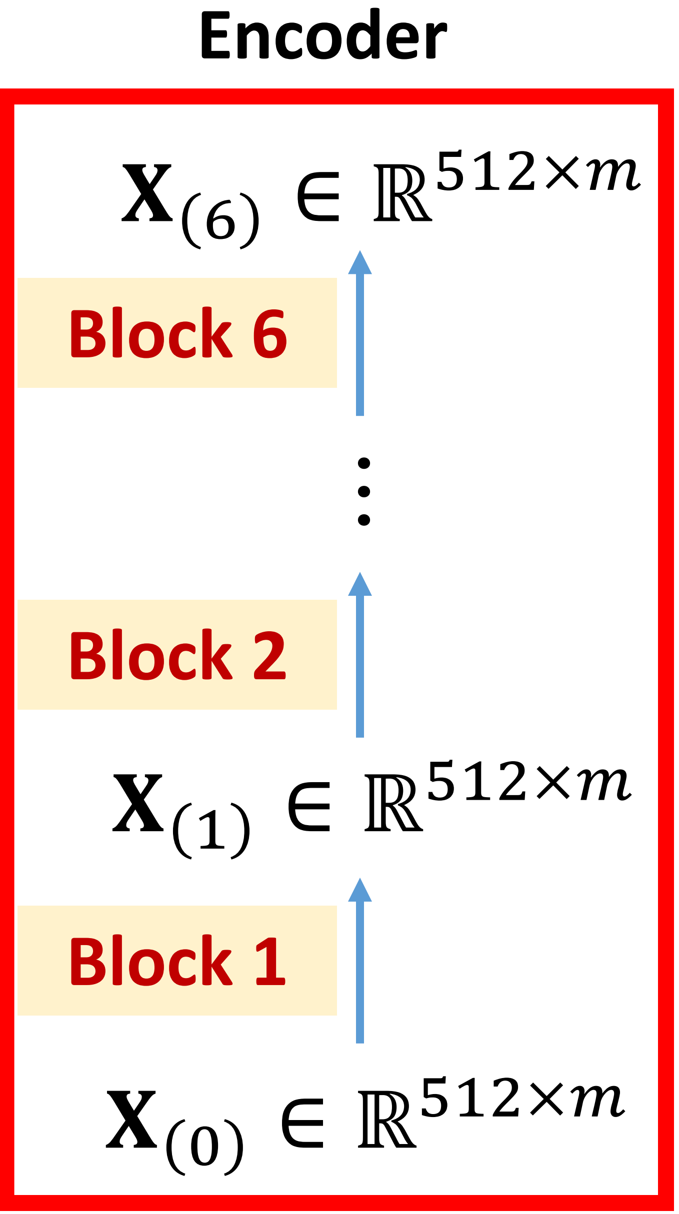
$$\mathbf{X}_{(1)} \in \mathbb{R}^{512\times m}$$

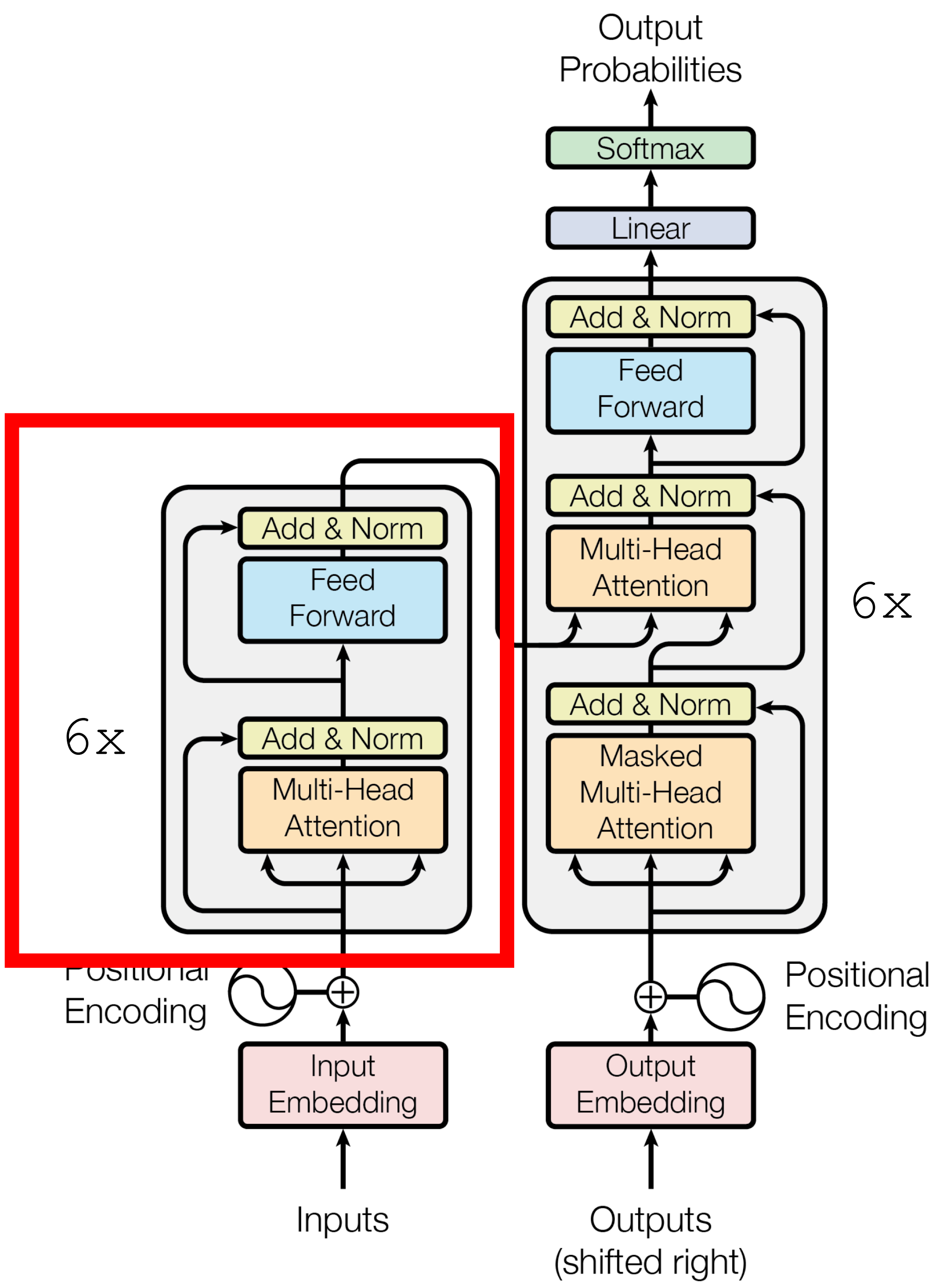
Block 1

$$\mathbf{X}_{(0)} \in \mathbb{R}^{512\times m}$$



Encoder Network





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:

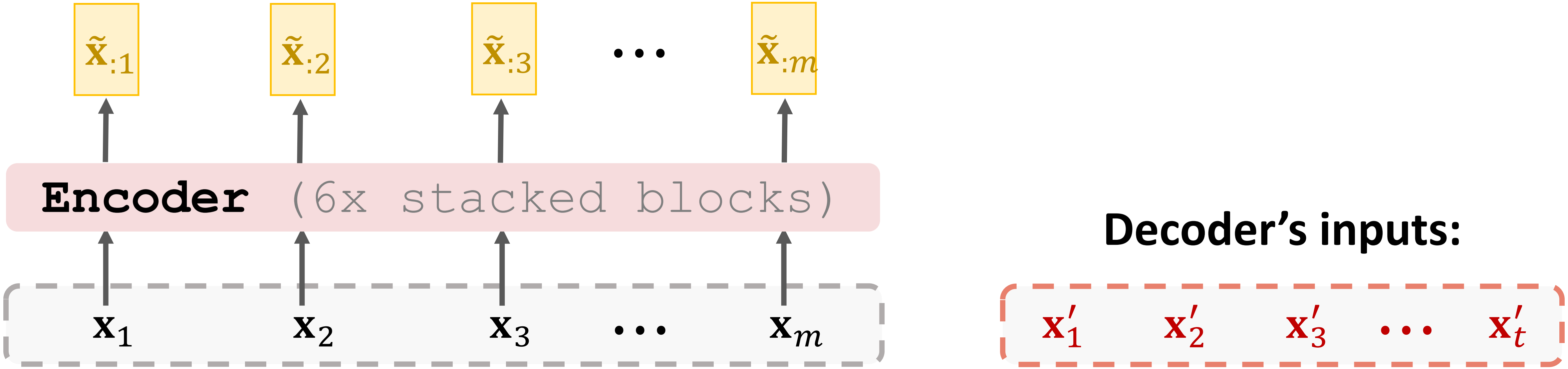
$\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \cdots \quad \mathbf{x}_m$

Decoder's inputs:

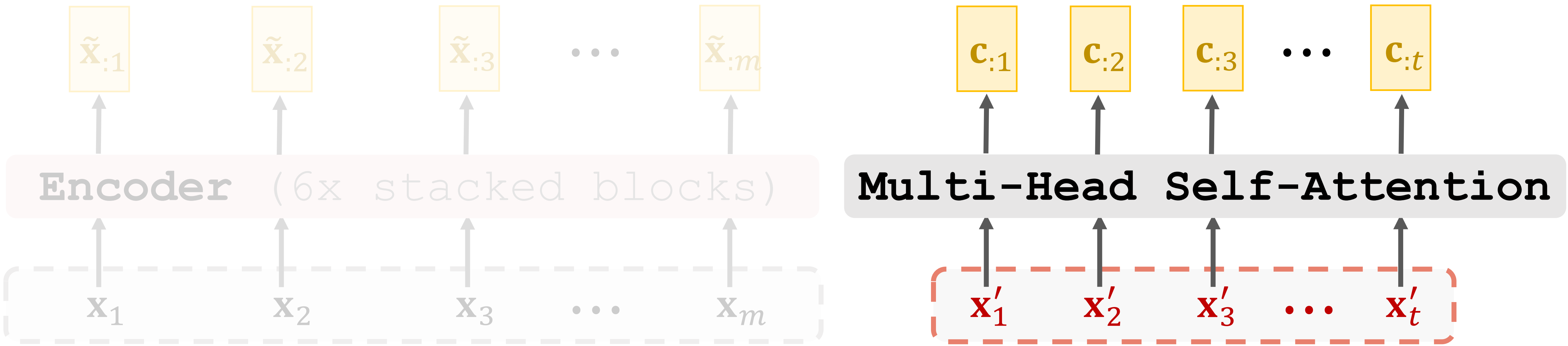
$\mathbf{x}'_1 \quad \mathbf{x}'_2 \quad \mathbf{x}'_3 \quad \cdots \quad \mathbf{x}'_t$

Stacked Attentions

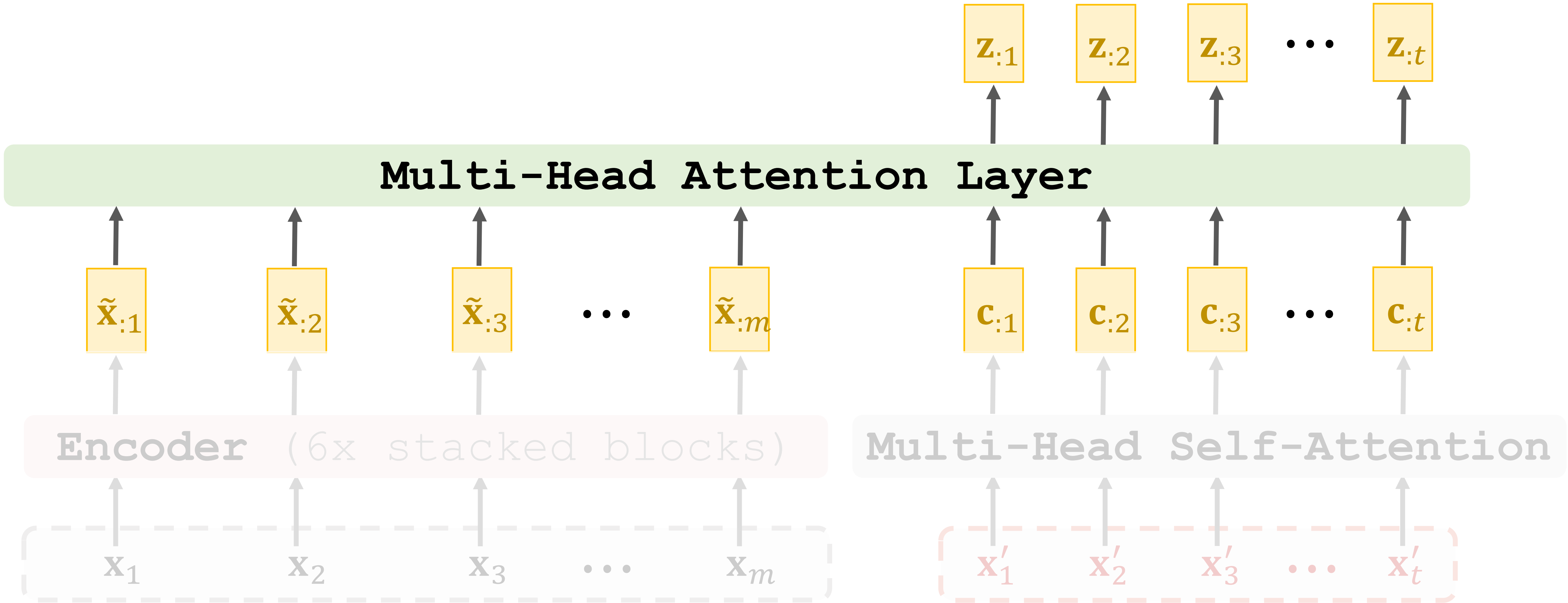
- Transformer's encoder contains 6 stacked blocks.
- 1 block $\approx$ 1 multi-head attention layer + 1 dense layer.



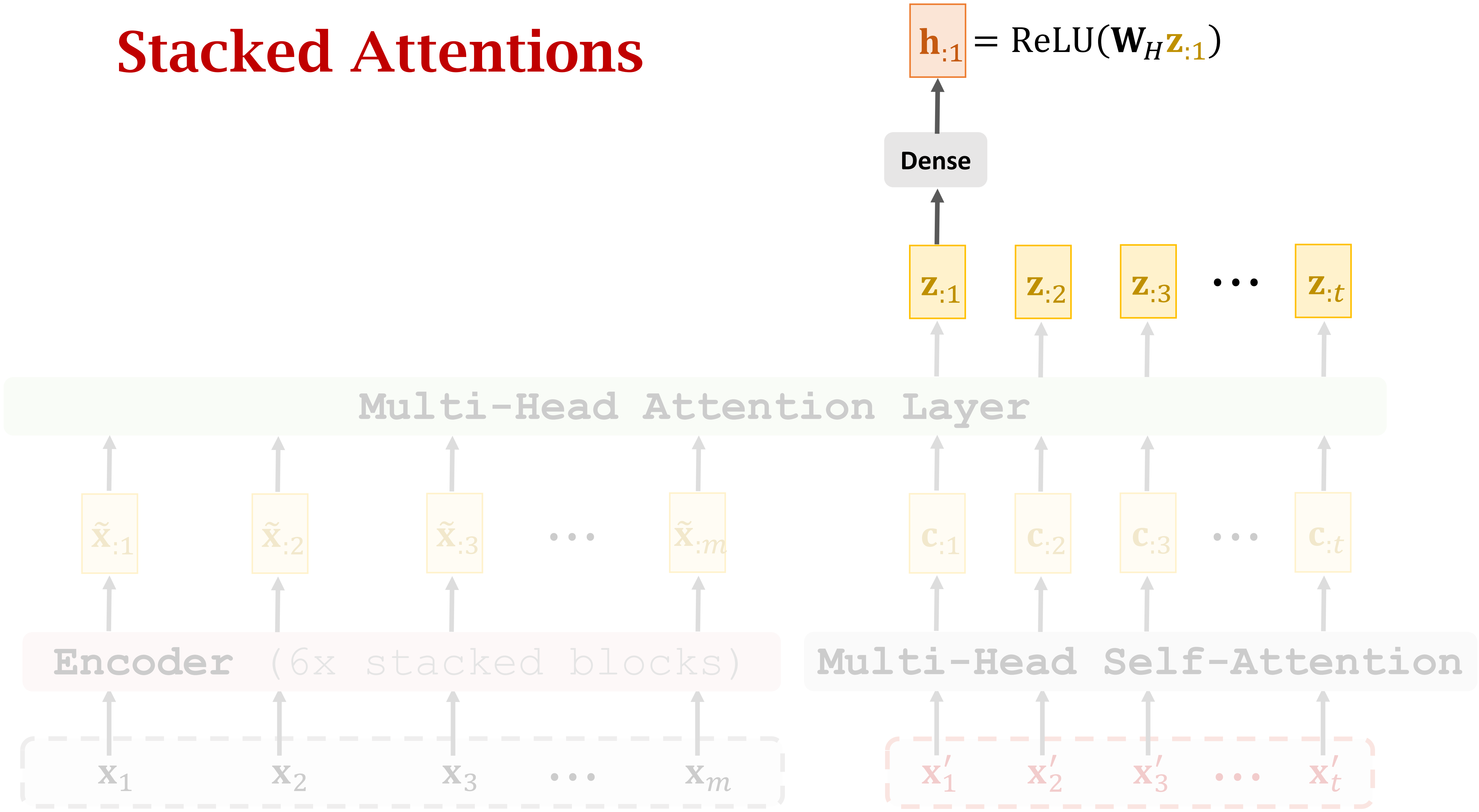
Stacked Attentions



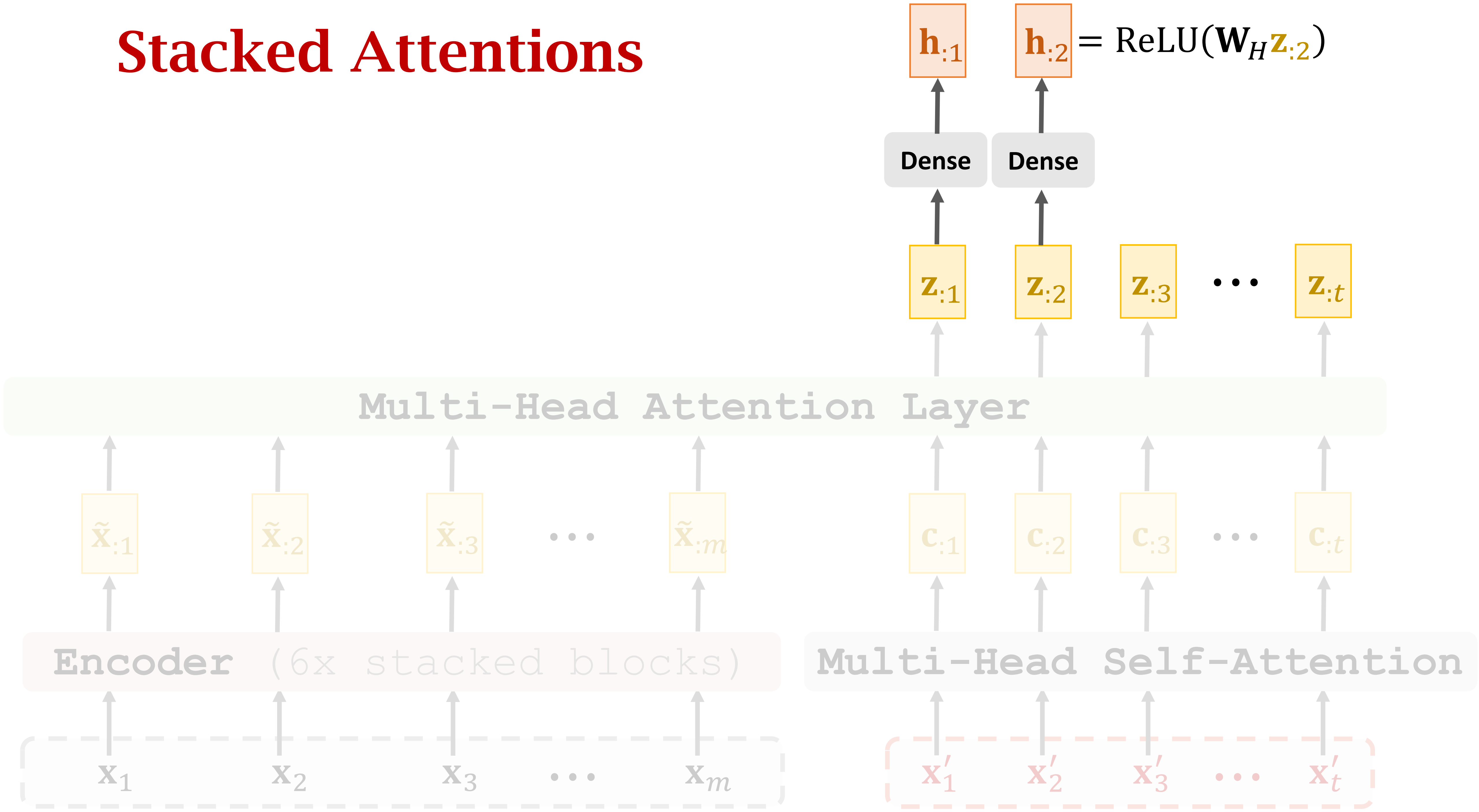
Stacked Attentions



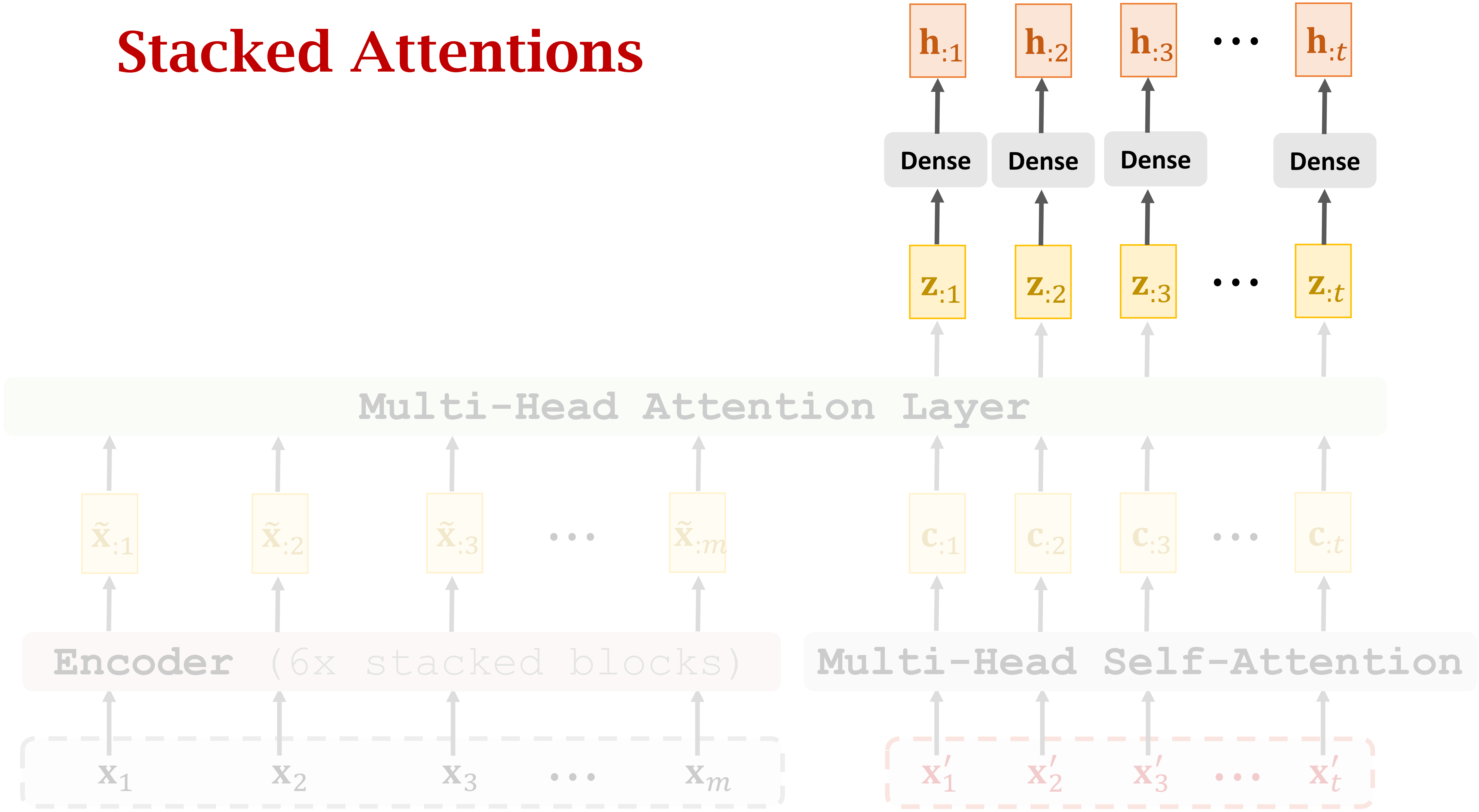
Stacked Attentions



Stacked Attentions

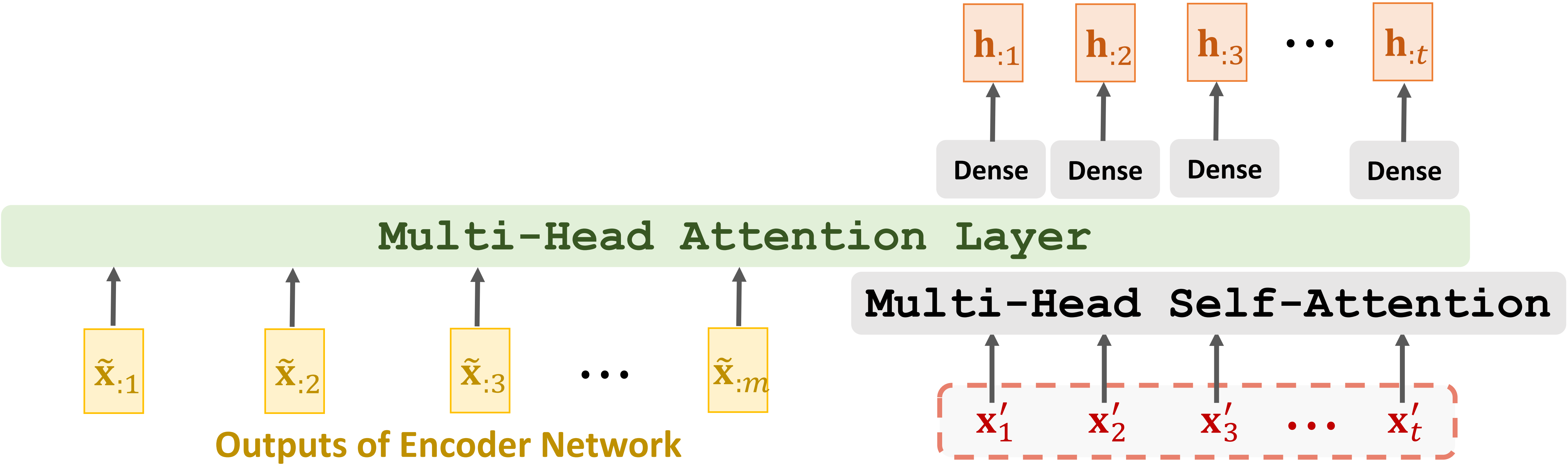


Stacked Attentions

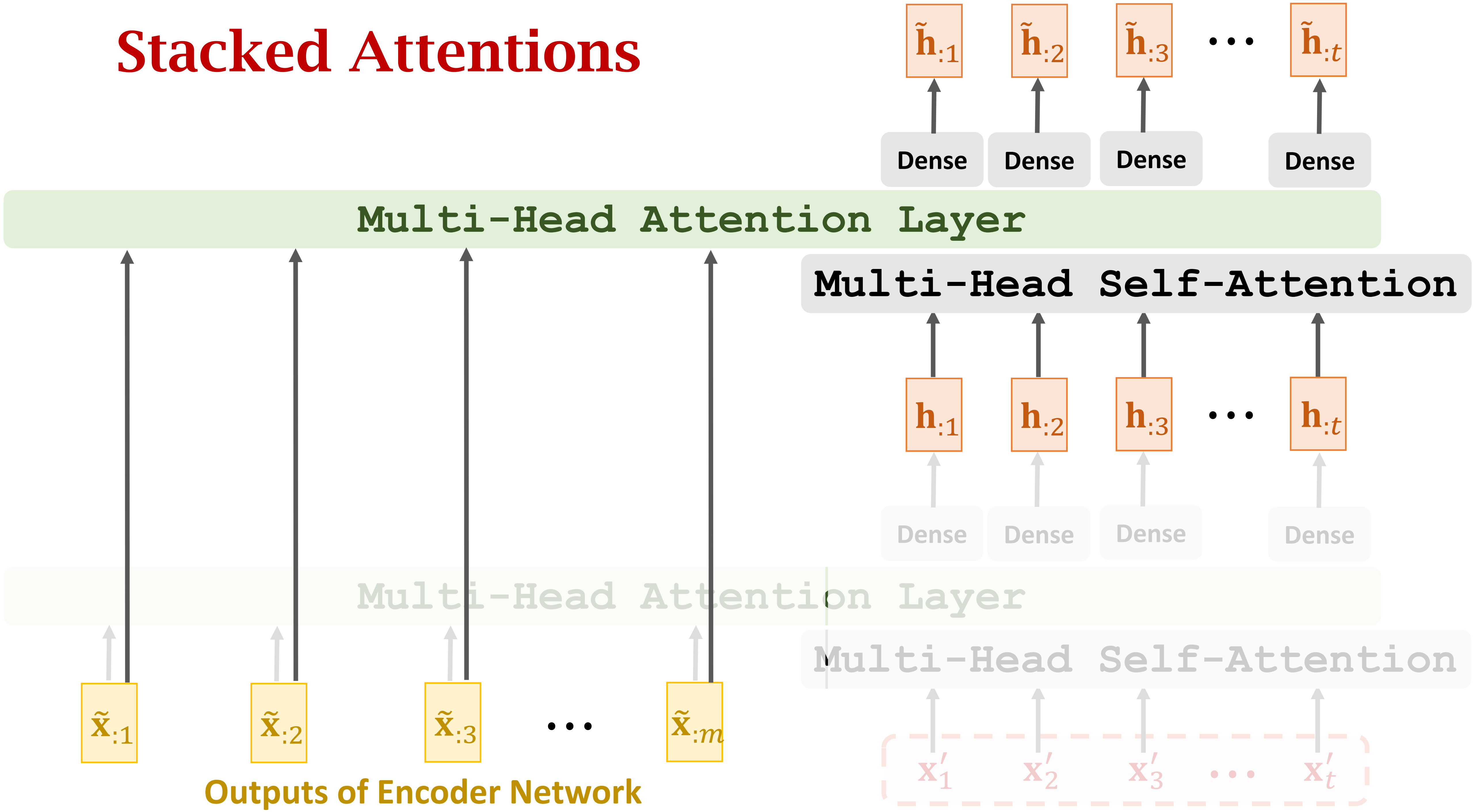


Stacked Attentions

- We have stacked 3 layers: self-attention + attention + dense.
- They together map $(\tilde{\mathbf{X}}, \mathbf{X}')$ to $\mathbf{H}$.
- One block of Transformer's decoder is the stack of the 3 layer.



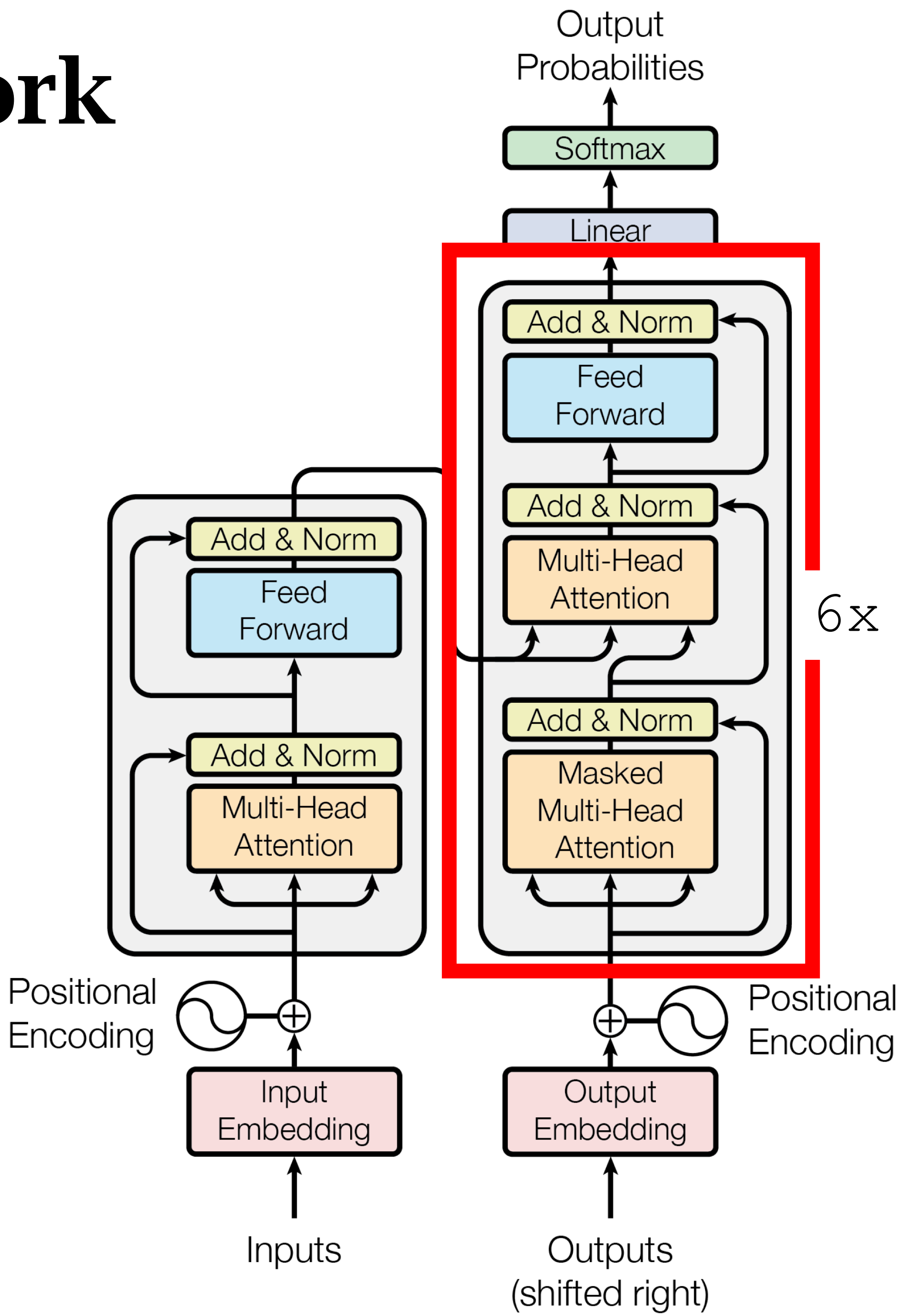
Stacked Attentions



Decoder of Transformer

Decoder Network

- 1 block = self-attention + attention layer + dense.
- Decoder is a stack of 6 such blocks.



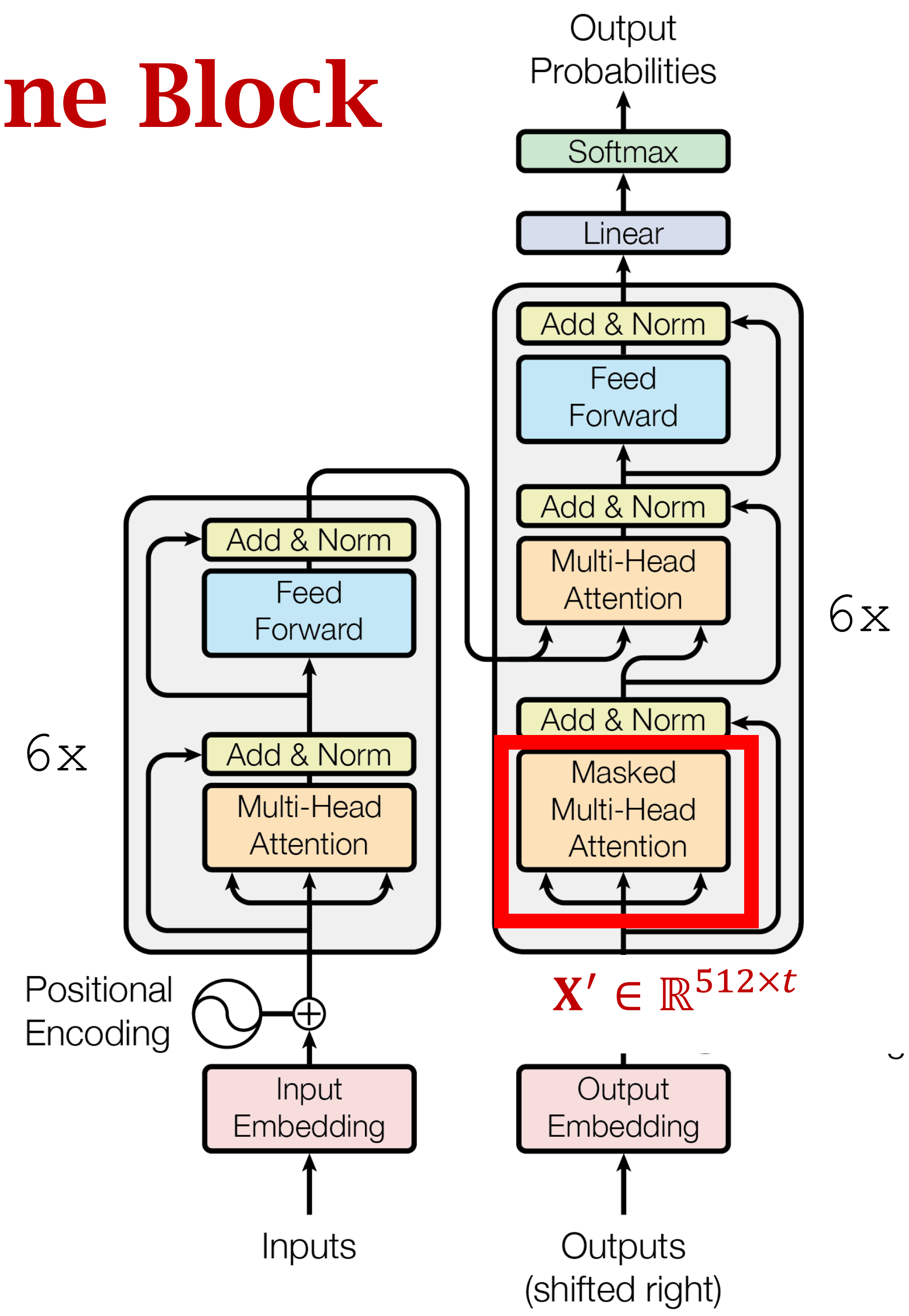
Decoder Network: One Block

Multi-head self-attention.

- Input shape: $512 \times t$.
- Use 8 single-head self-attentions:

$$\mathbf{C} = \mathrm{Attn}(\mathbf{X}', \mathbf{X}').$$

- Each outputs $64 \times t$ matrix.
- Output shape: $512 \times t$.



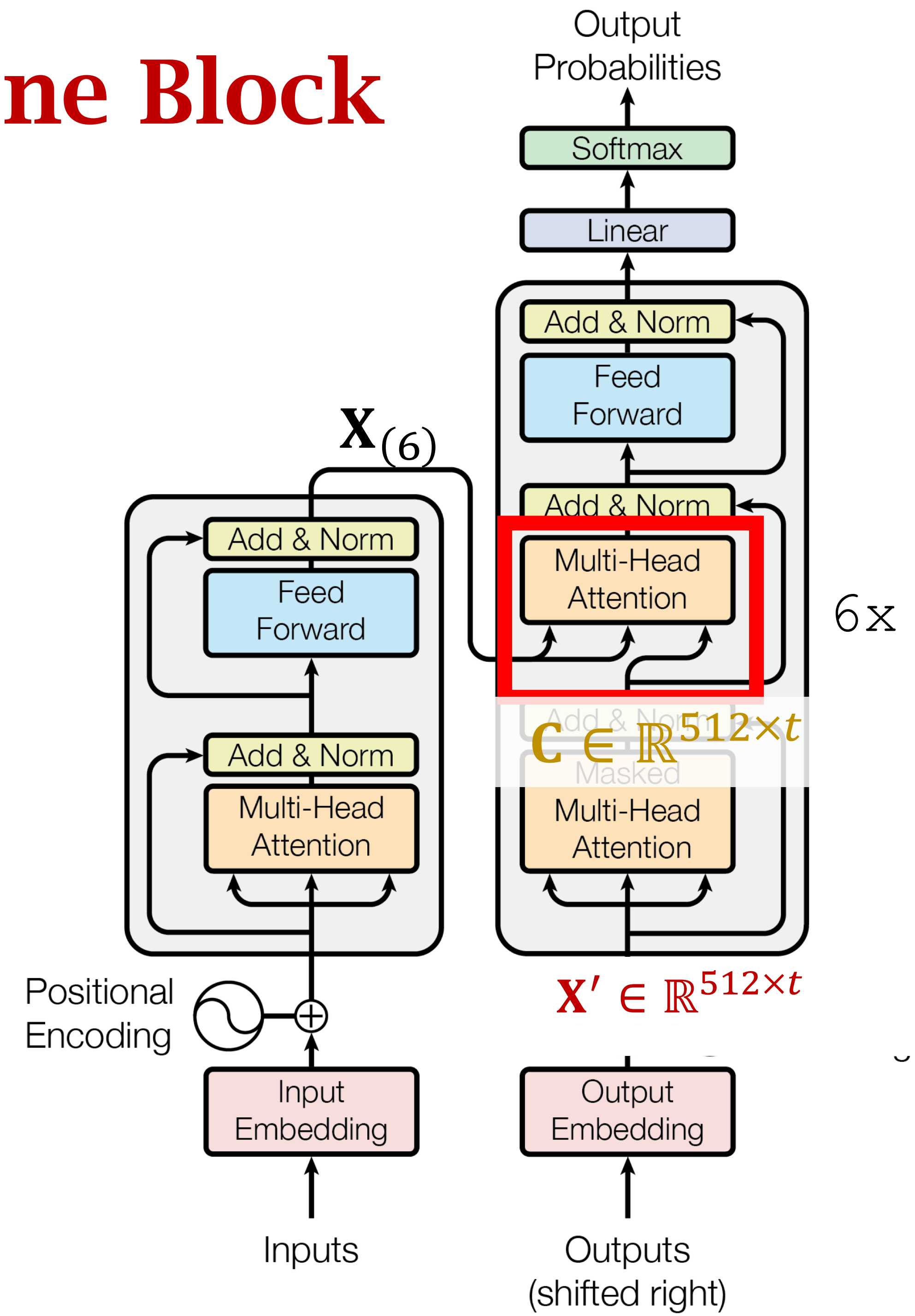
Decoder Network: One Block

Multi-head attention.

- Use 8 single-head attentions:

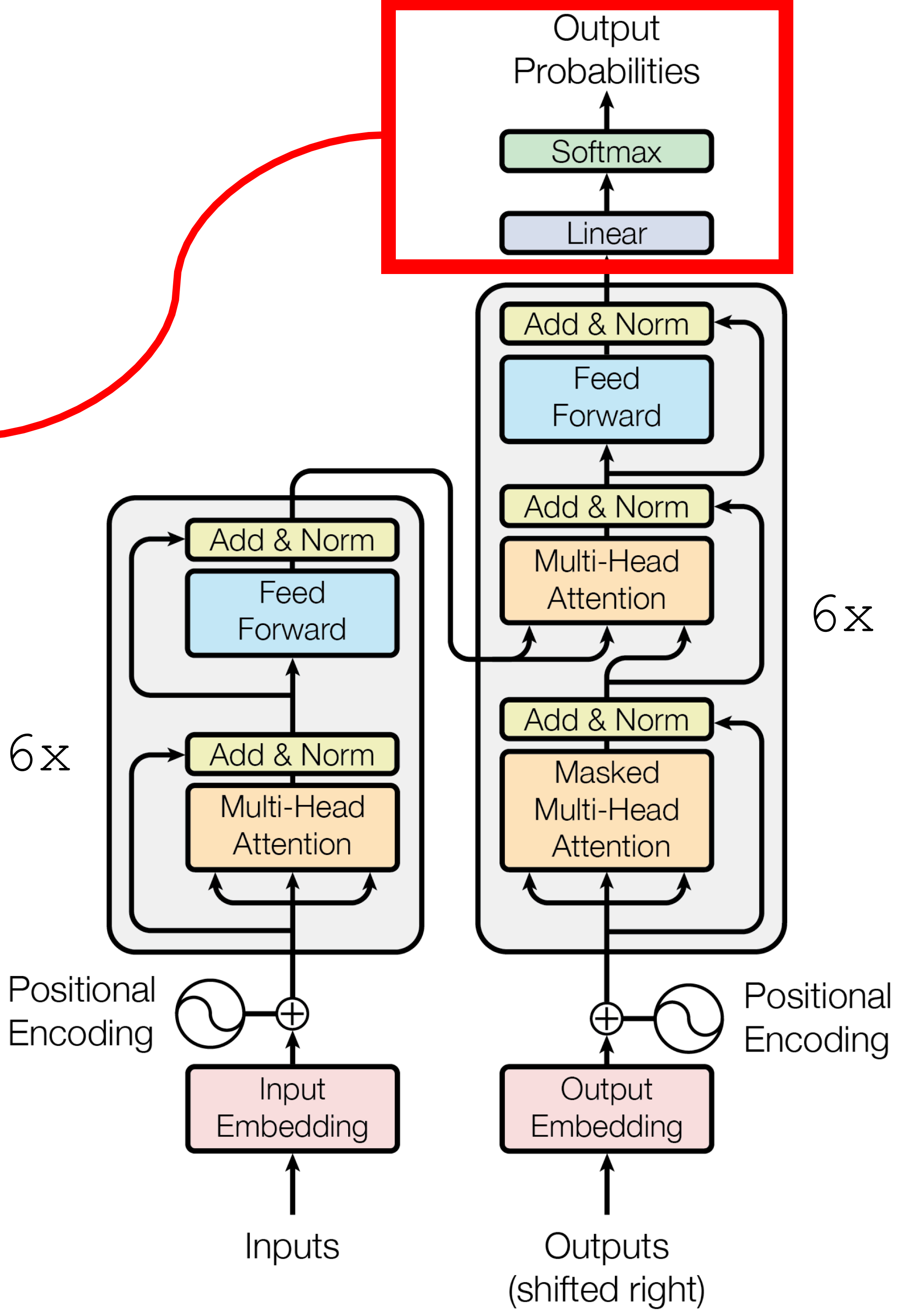
$$\text{Attn}\left(\mathbf{X}_{(6)}, \mathbf{C}\right).$$

- Each outputs $64 \times t$ matrix.
- Output shape: $512 \times t$.



Decoder Network

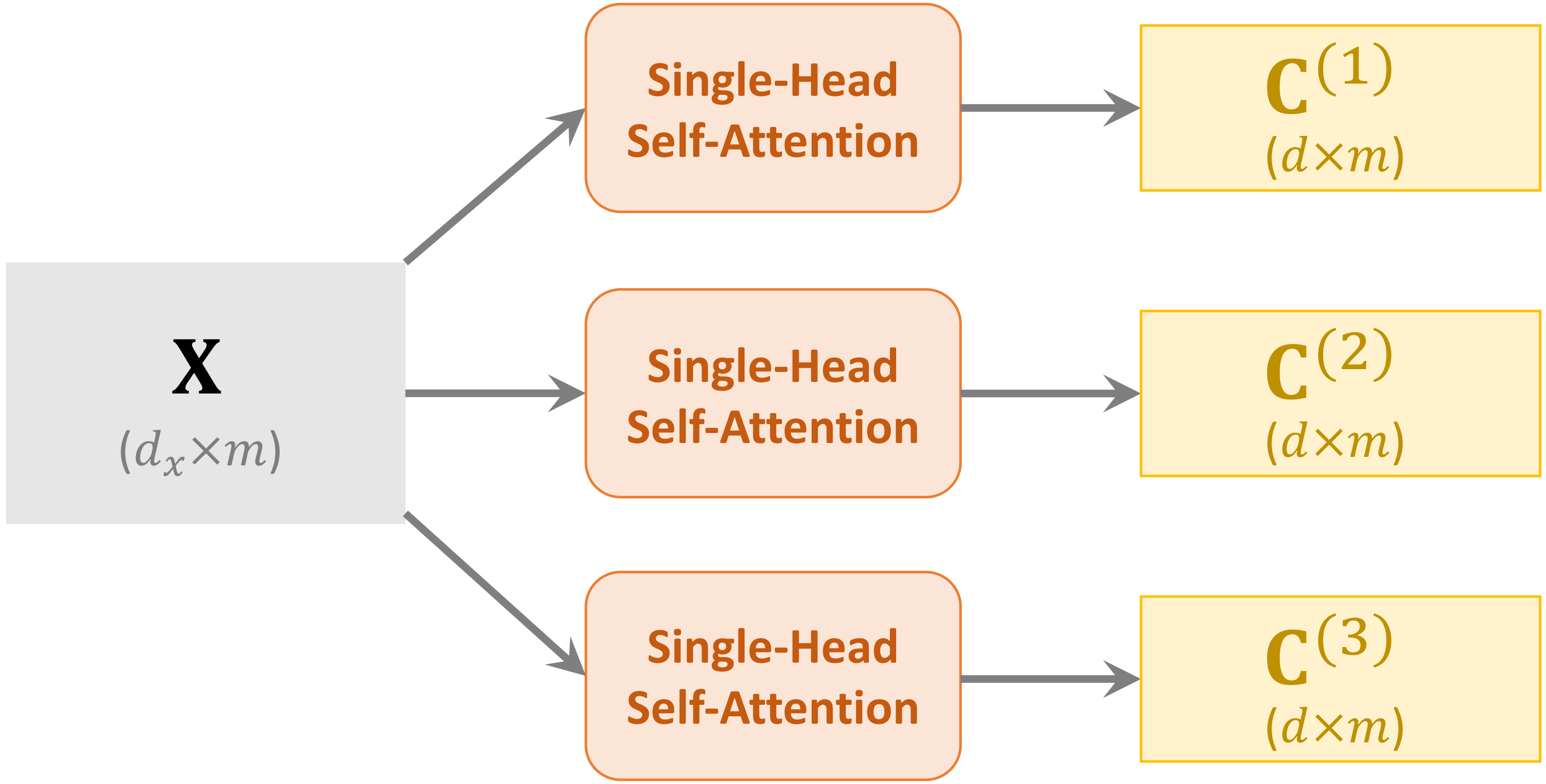
- Output a distribution over the vocabulary.
- Compare the distribution with the one-hot encode of the label.
- ➔ Loss, e.g., cross-entropy.
- ➔ Gradient.
- ➔ Update model parameters.



Summary

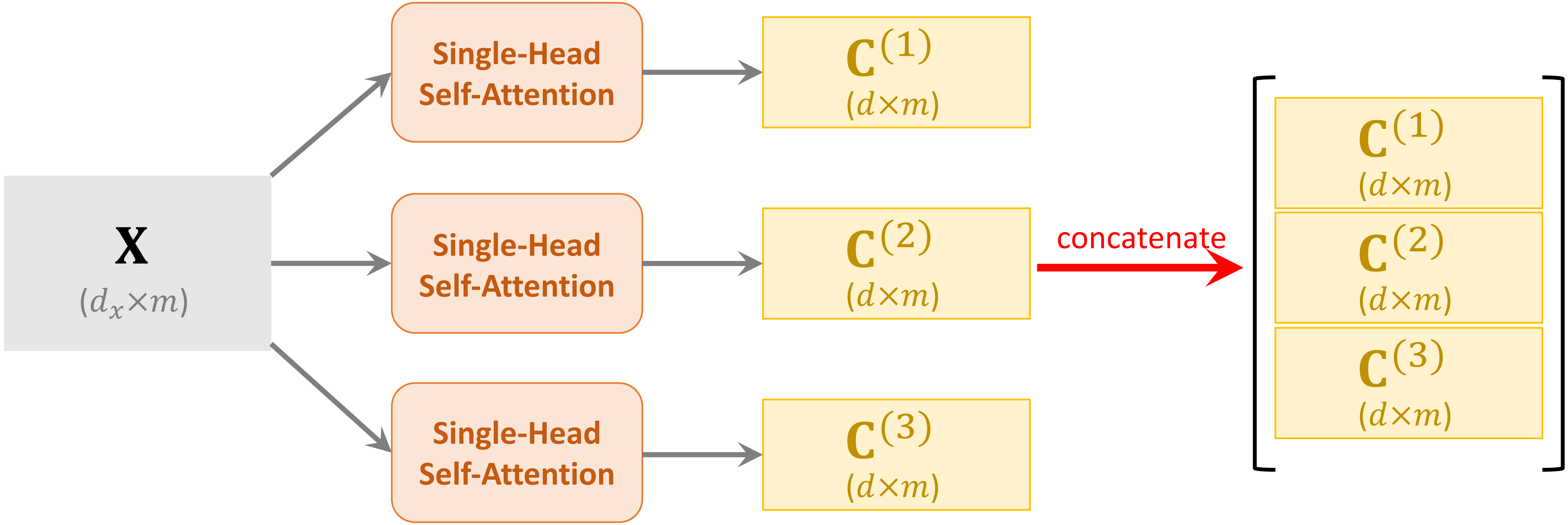
From Single-Head to Multi-Head

- 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.



Encoder Network of Transformer

- 1 encoder block $\approx$ 8-head self-attention + dense.
- Encoder network is a stack of 6 such blocks.
- Input shape: $512 \times m$.
- Output shape: $512 \times m$.

Decoder Network of Transformer

- 1 decoder block $\approx$ 8-head self-attention + 8-head attention + dense.
- Encoder network is a stack of 6 such blocks.
- Input shape: $512 \times t$.
- Output shape: $512 \times t$.

Summary

- Transformer model is not RNN.
- Transformer is based on attention and self-attention.
- Upside: Outperform all the state-of-the-art RNN models.
- Downside: Much more expensive than RNN models.

Thank you!