Transformer Model

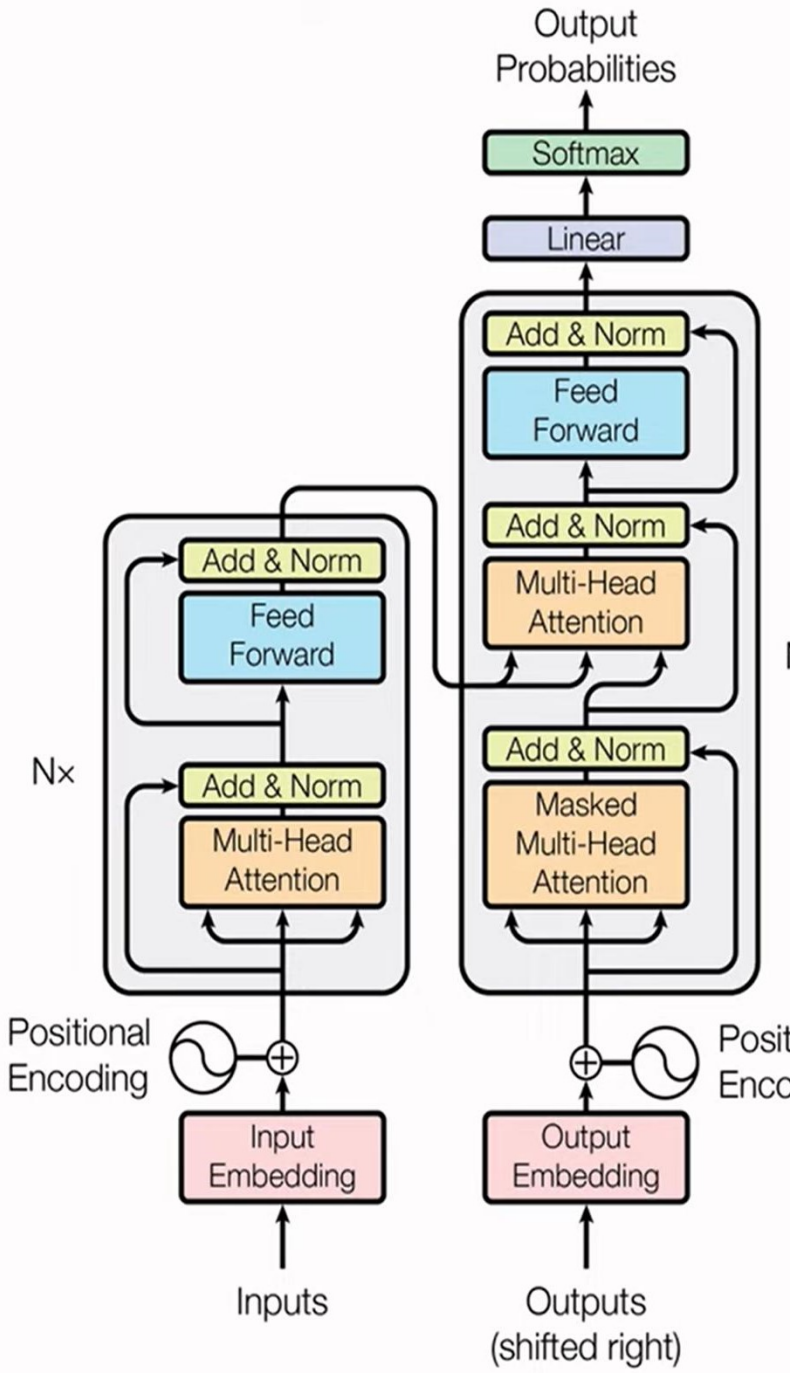
Architecture

Advantages and Disadvantages

Performance Comparison

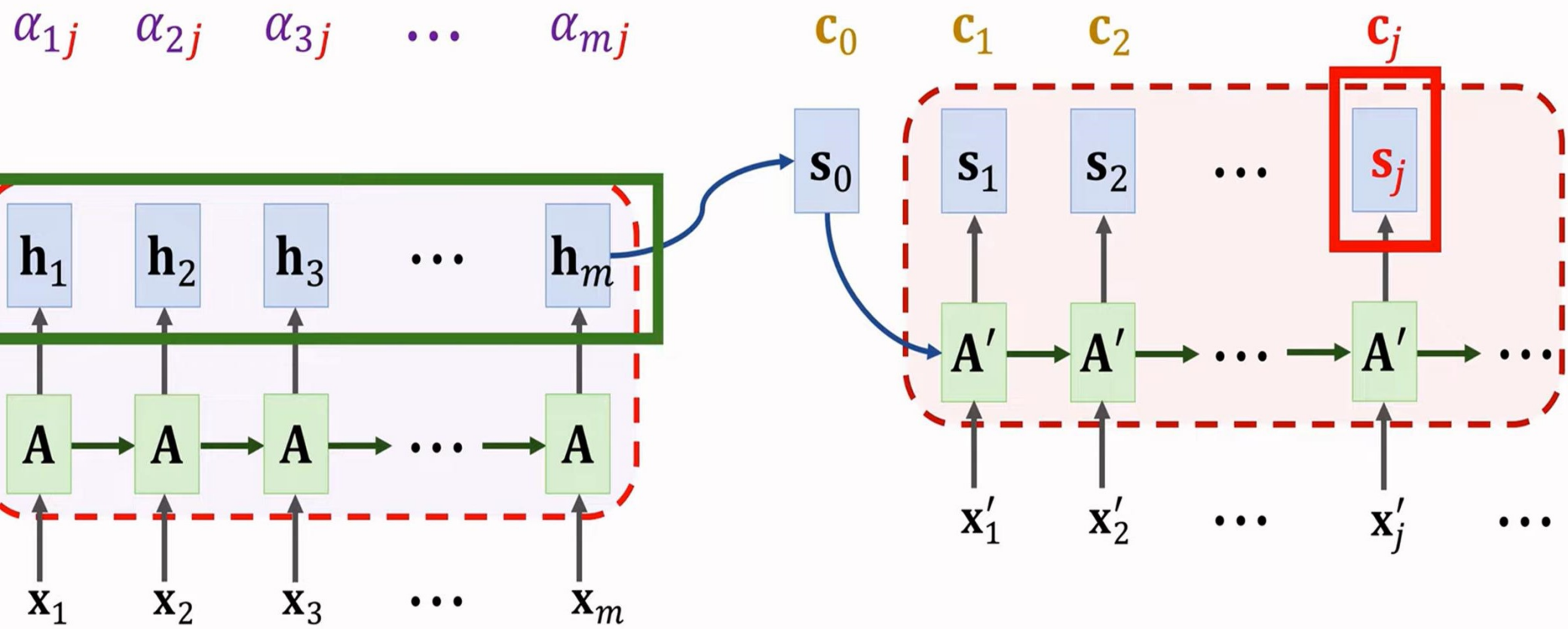
Transformer Model

- Transformer is a Seq2Seq model.
- Transformer is not RNN.
- Purely based on attention and dense layers.
- Higher accuracy than RNNs on large datasets.

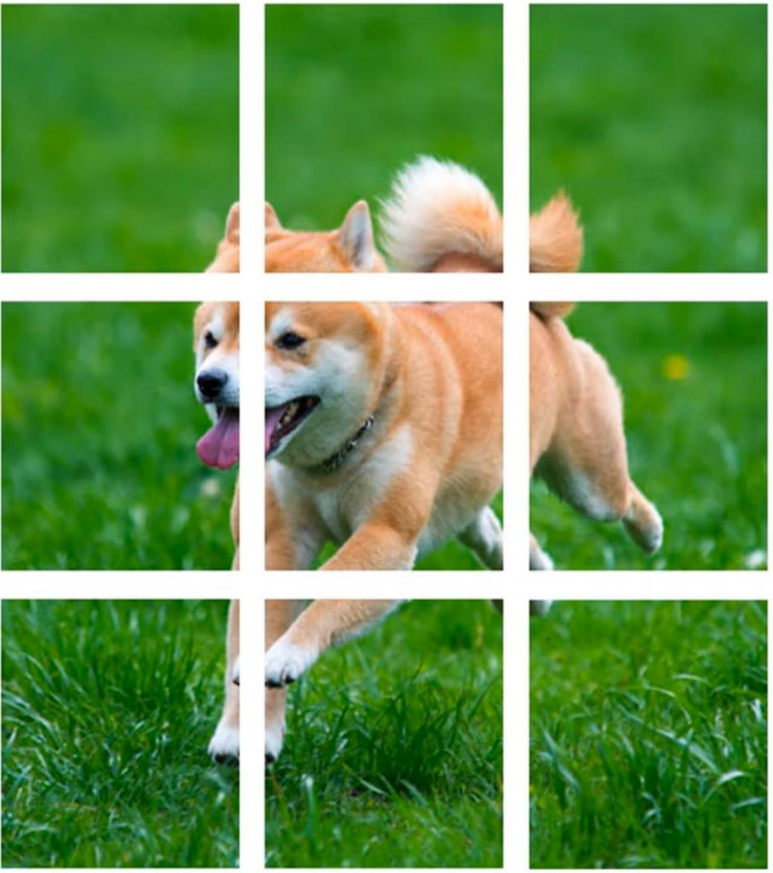


Attention for Seq2Seq Model

Weights: $\alpha_{ij} = \text{align}(\mathbf{h}_i, \mathbf{s}_j)$.



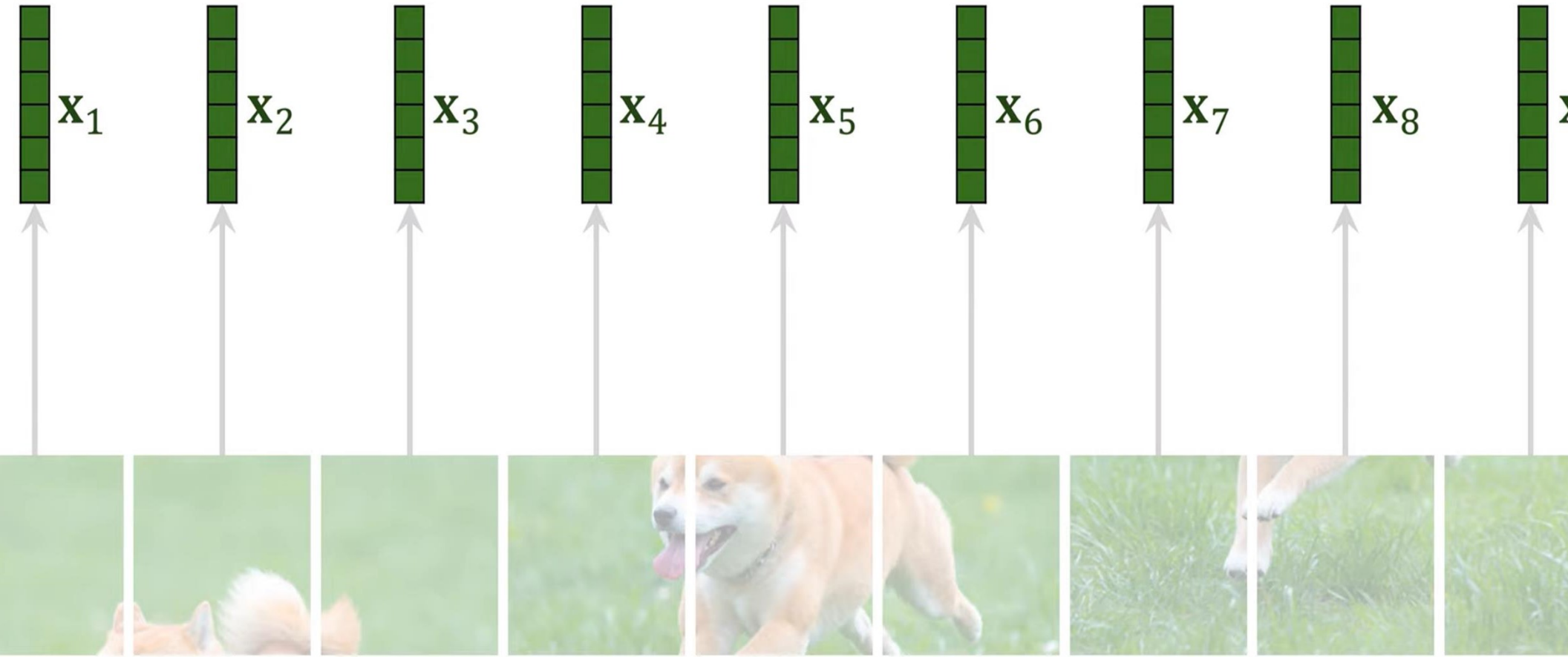
Split Image into Patches

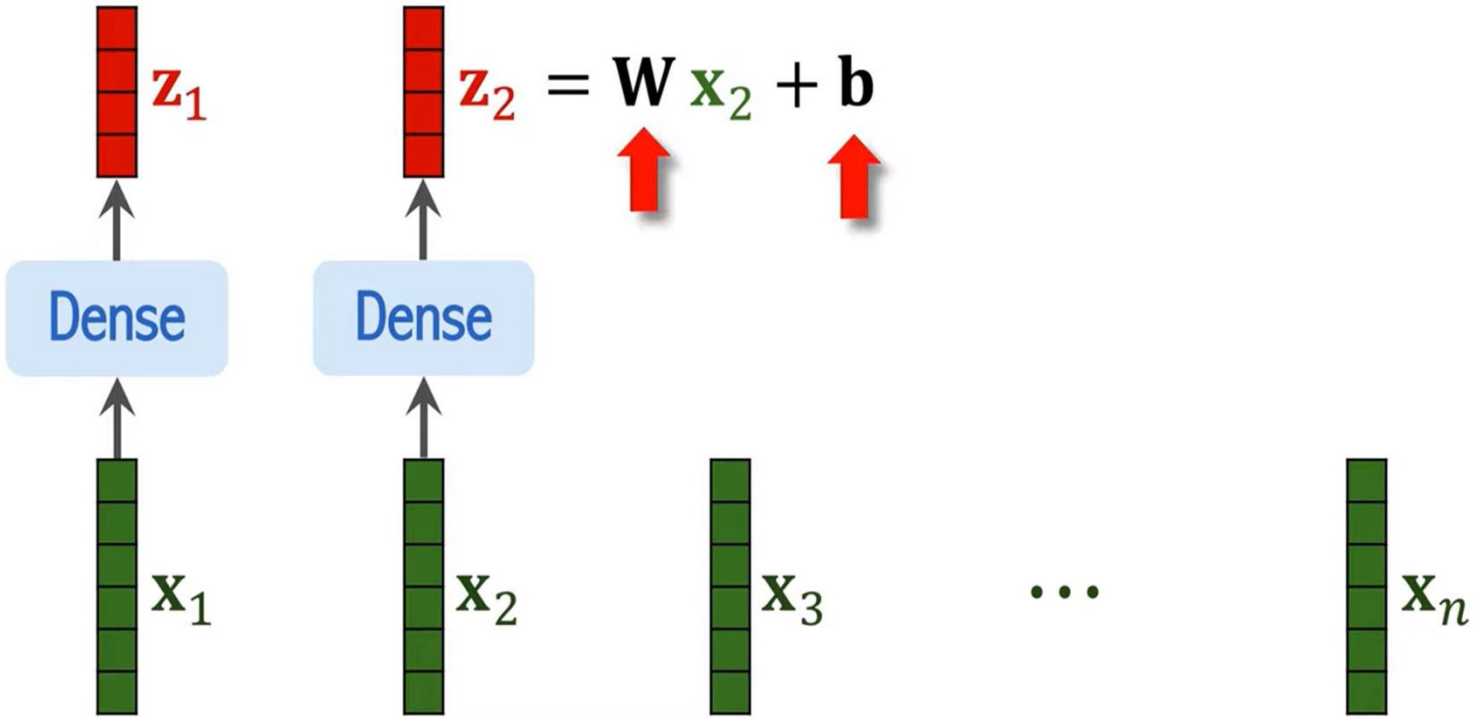


- Here, the patches do not overlap.

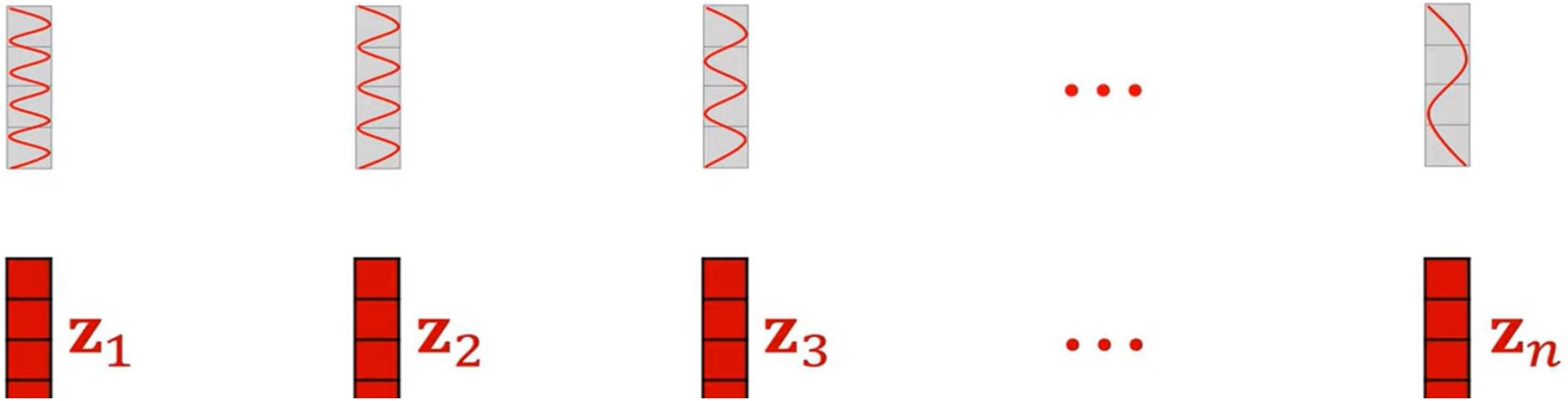
Vectorization

If the patches are $d_1 \times d_2 \times d_3$ tensors, then the vectors are $d_1 d_2 d_3 \times 1$.

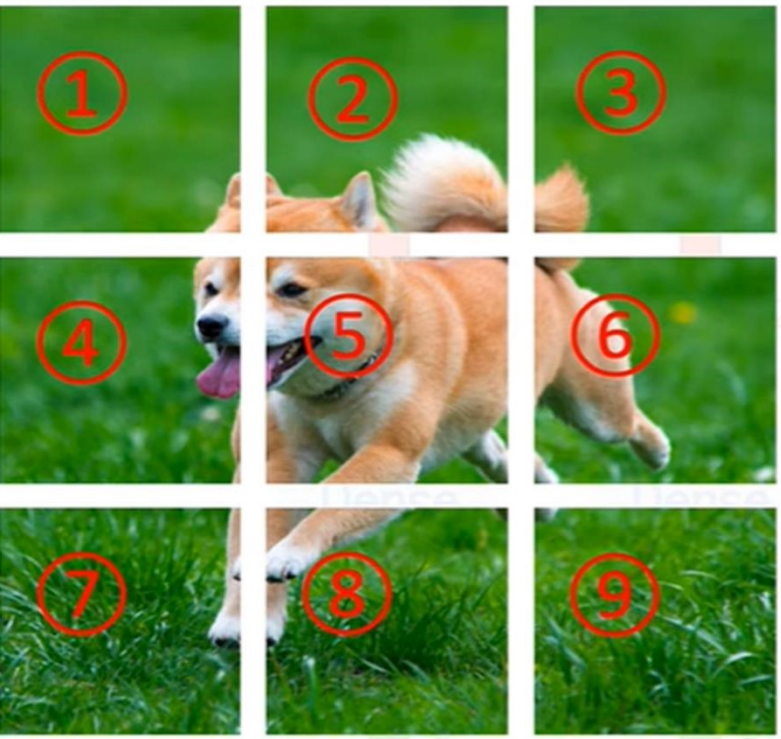


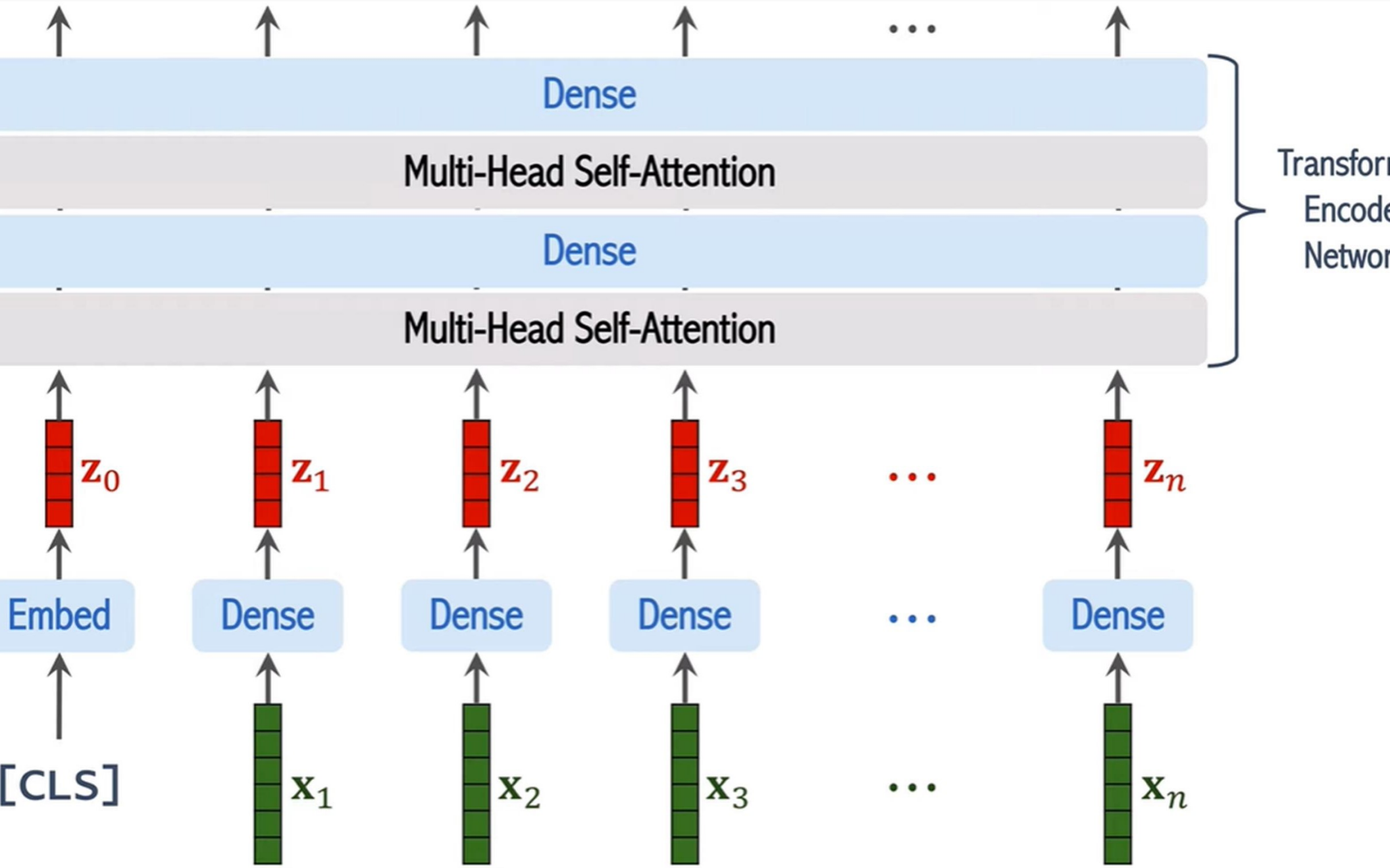


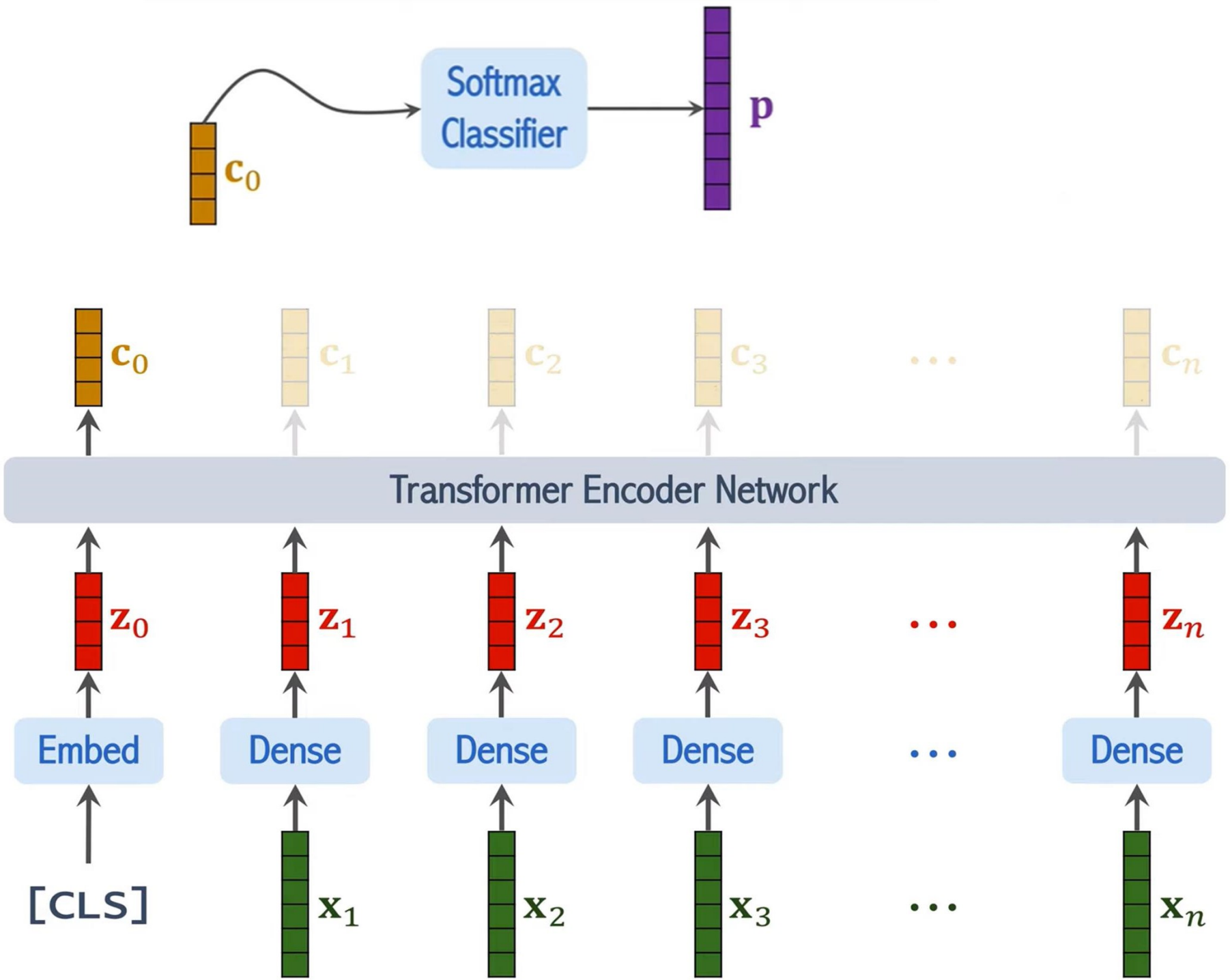
Add positional encoding vectors to $z_1, z_2, \cdots, z_n$.



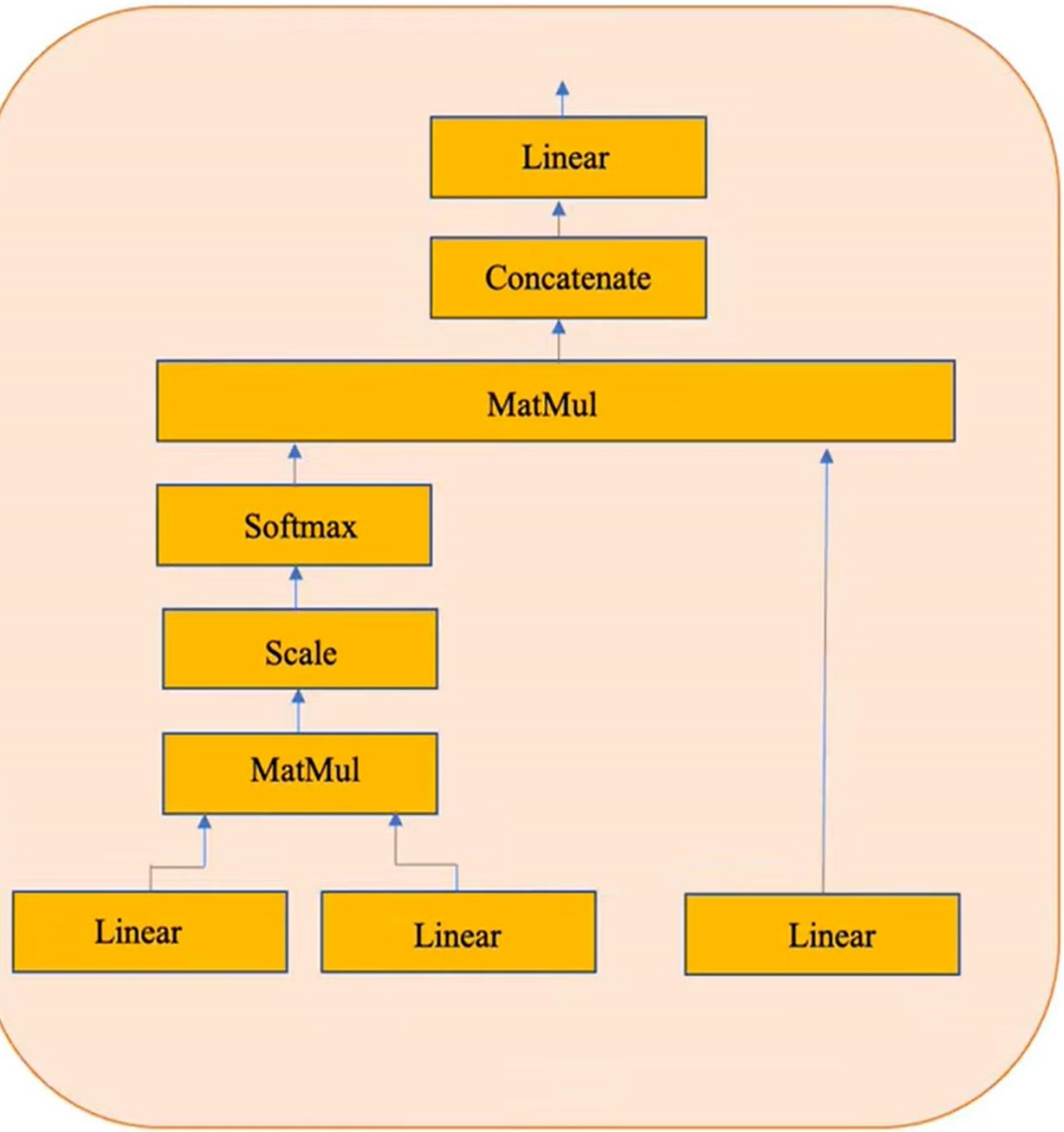
Add positional encoding vectors to $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$. (Why?)




Dense
Multi-Head Self-Attention
Dense
Multi-Head Self-Attention
Transfor
Encode
Networ
z_0
z_1
z_2
z_3
z_n
Embed
Dense
Dense
Dense
Dense
[CLS]
x_1
x_2
x_3
x_n


Softmax
Classifier
p
c_0
c_0
c_1
c_2
c_3
c_n
Transformer Encoder Network
z_0
z_1
z_2
z_3
z_n
Embed
Dense
Dense
Dense
Dense
[CLS]
x_1
x_2
x_3
x_n

Multi-Head Attention



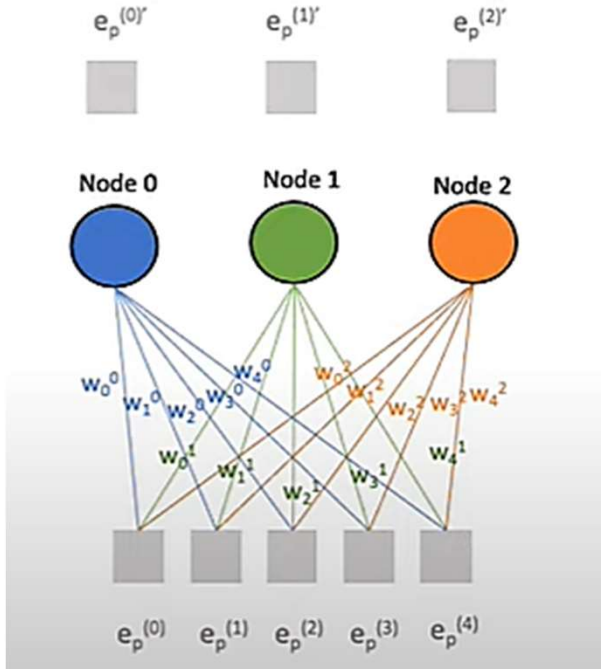
Linear Layer

) Mapping inputs onto the outputs

i) Changing matrix/vector dimensions

Linear Layer

$W_0^0 = 0.6$



Query

Key

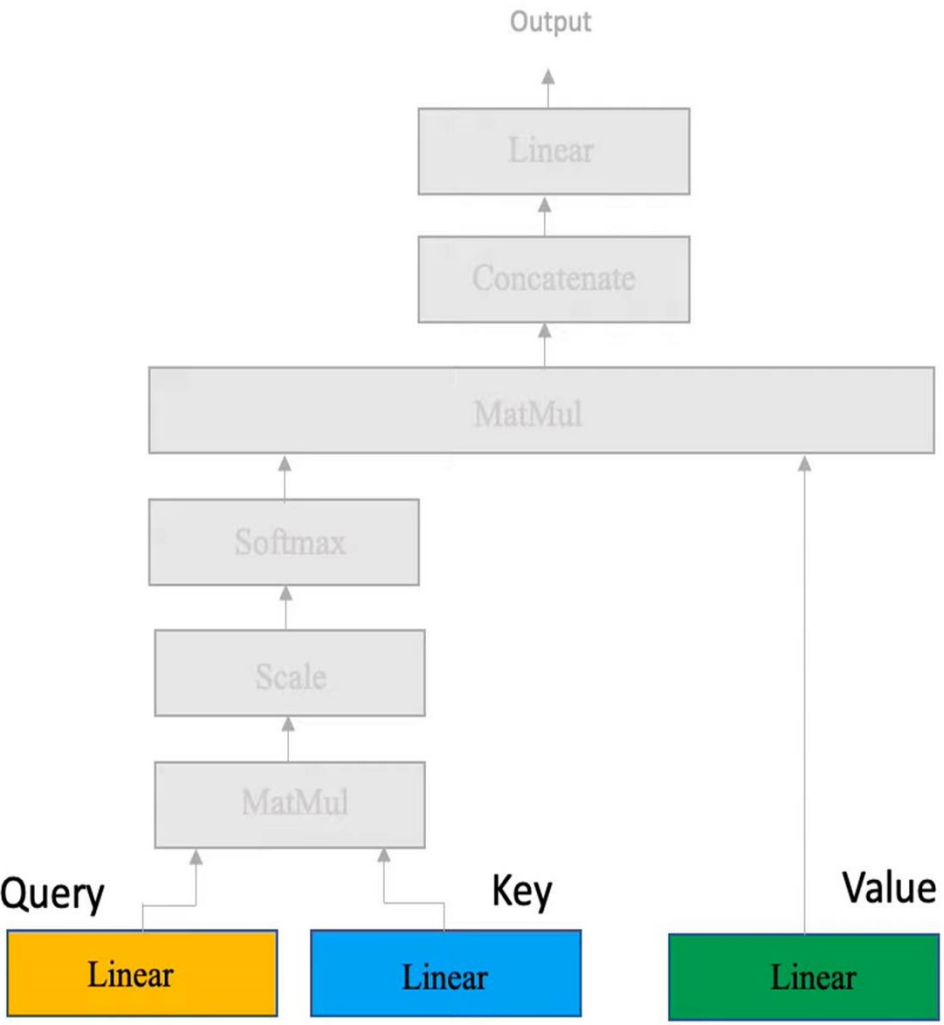
Value

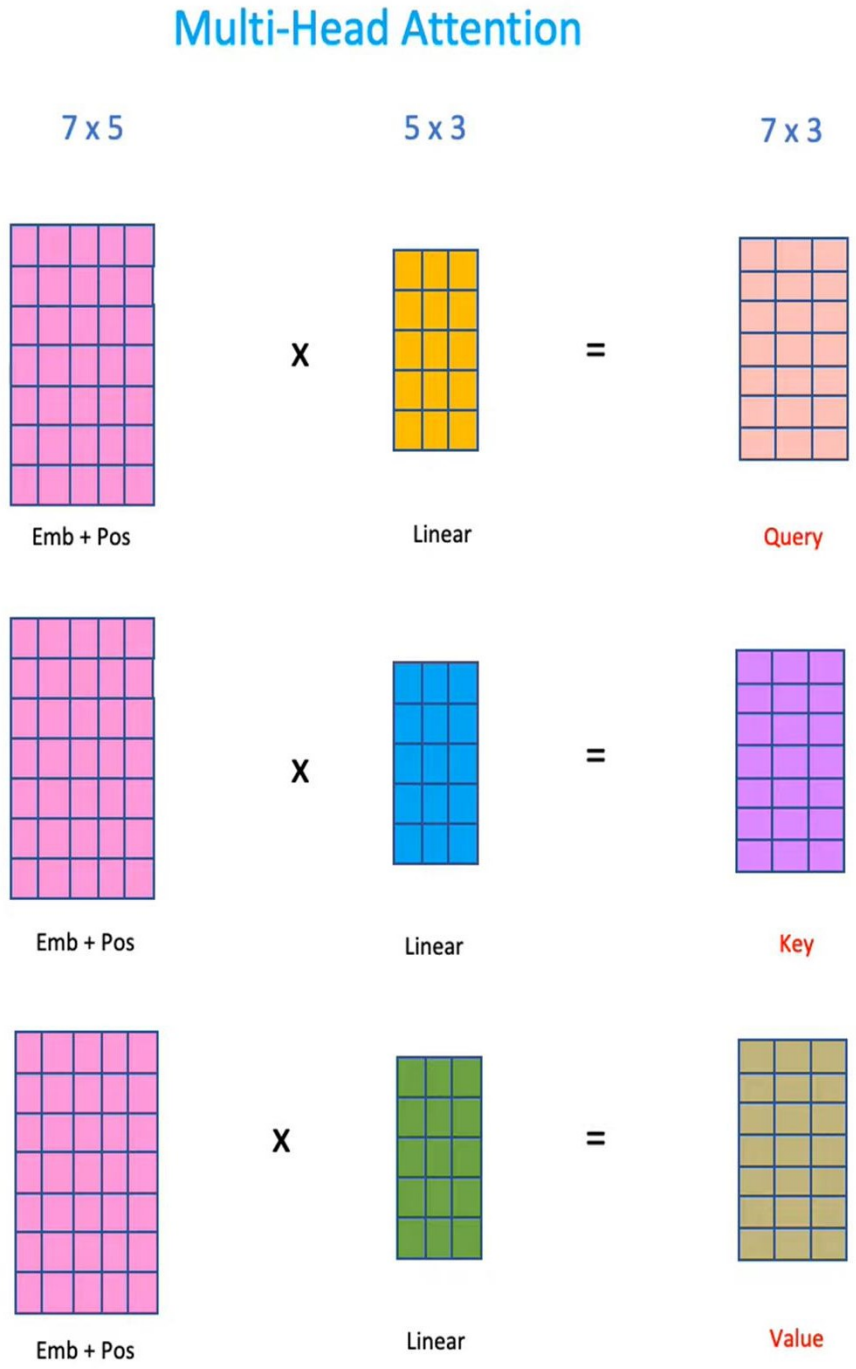
Cosine Similarity

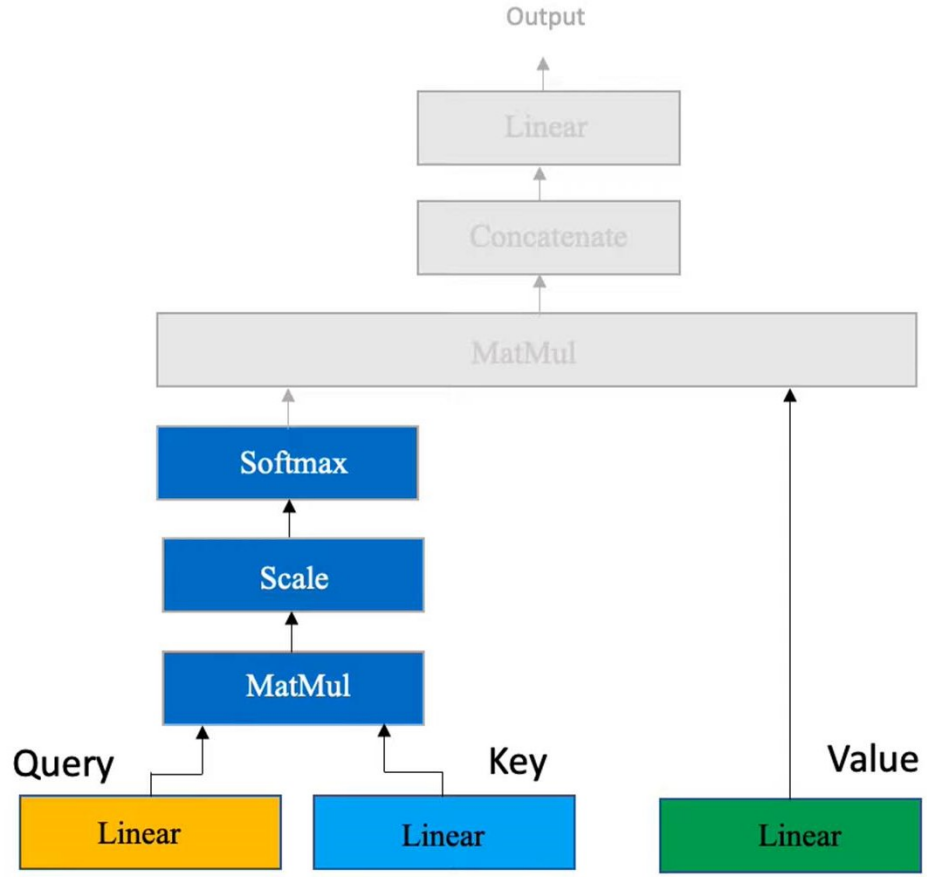
$$\text{Cos}\,(A, B) = \frac{A \,.\, B}{|A|\,|B|}$$

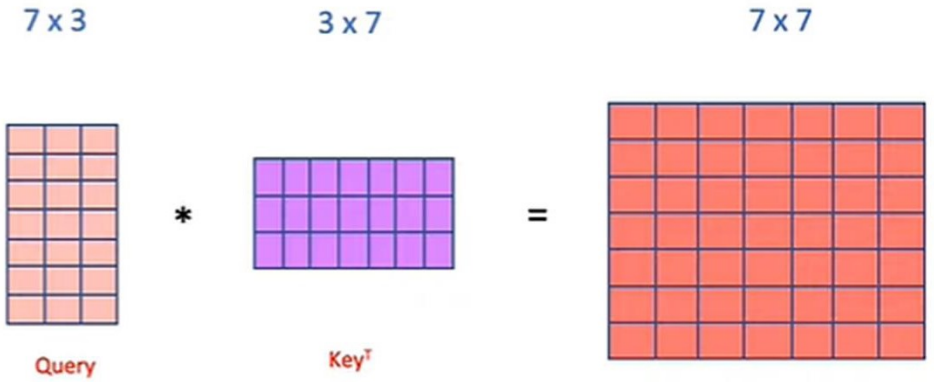
Similarity b/w Matrices

$$\text{similarity}\,(A, B) = \frac{A \,.\, B^{T}}{scaling}$$

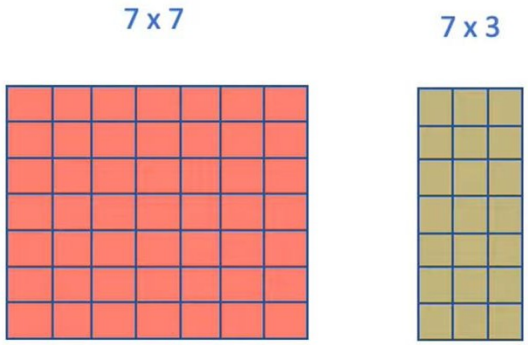

Output
Linear
Concatenate
MatMul
Softmax
Scale
MatMul
Query
Key
Value
Linear
Linear
Linear


Multi-Head Attention
7 x 5
5 x 3
7 x 3
Emb + Pos
Linear
Query
Emb + Pos
Linear
Key
Emb + Pos
Linear
Value





Attention Filter



$$\text{softmax}\left(\frac{Q.K^T}{\sqrt{d_k}}\right) \quad V$$

Intuition

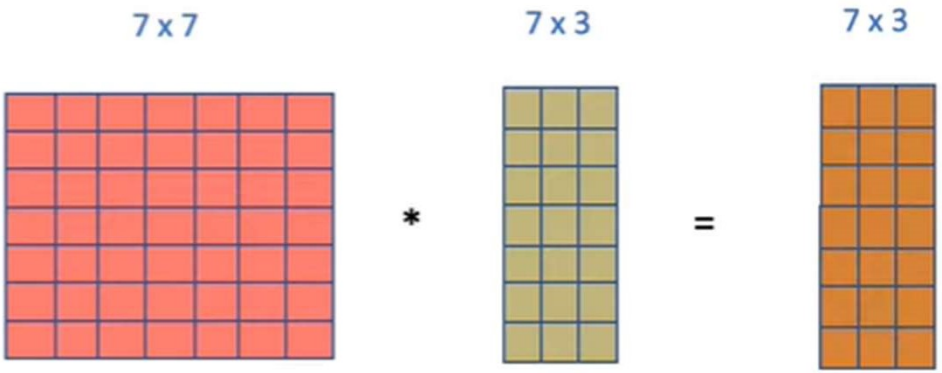
Attention Filter

Original Image

Filtered Image

*

=

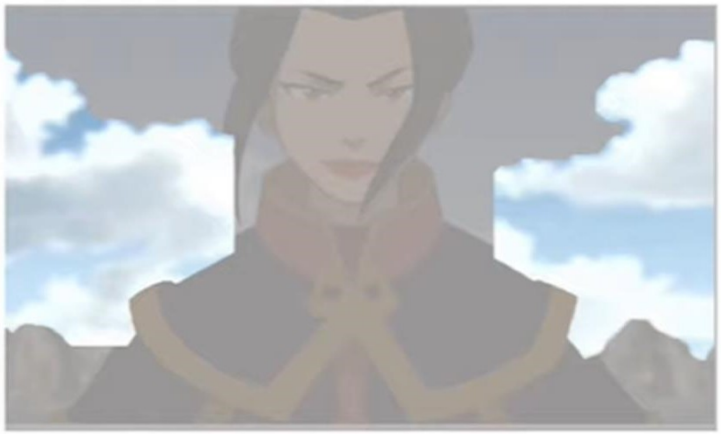


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q.K^T}{\sqrt{d_k}}\right).V$$

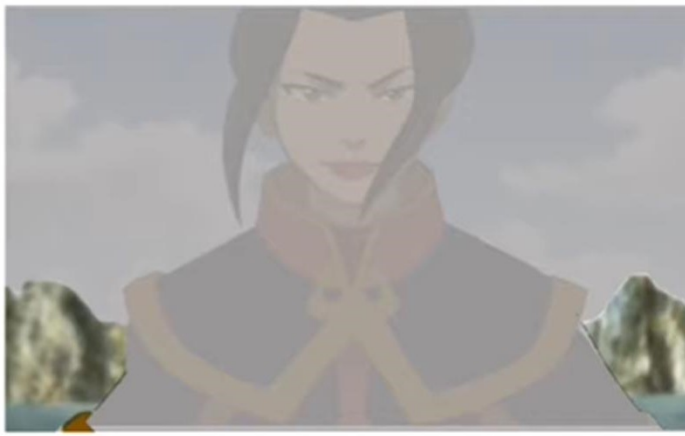
Intuition for Multi-head Attention

Attention Filter 1

Attention Filter 2

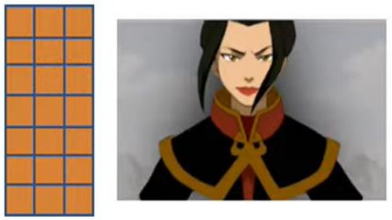


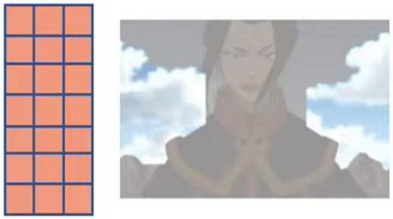
Attention Filter 3



Multi-Head Attention

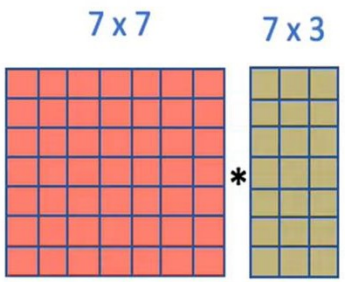
Filtered Value

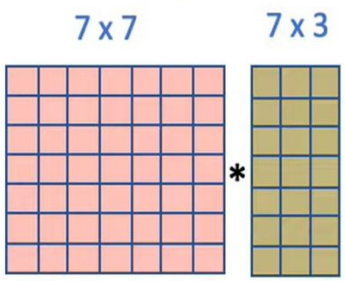


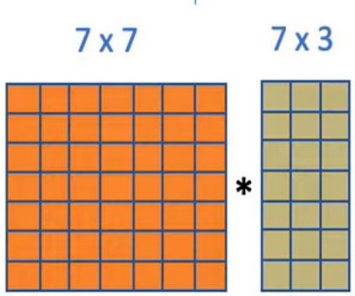


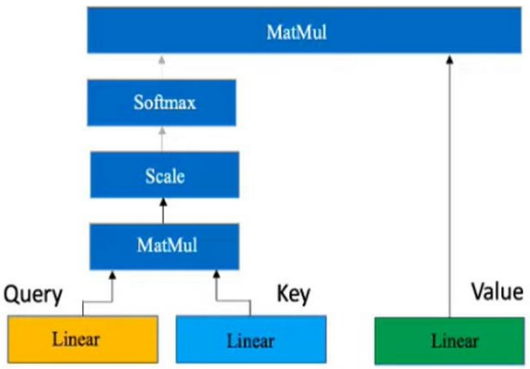
7 x 7 * 7 x 3 Value

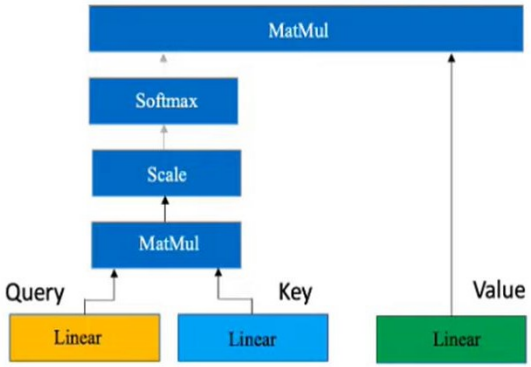
Attention Filter

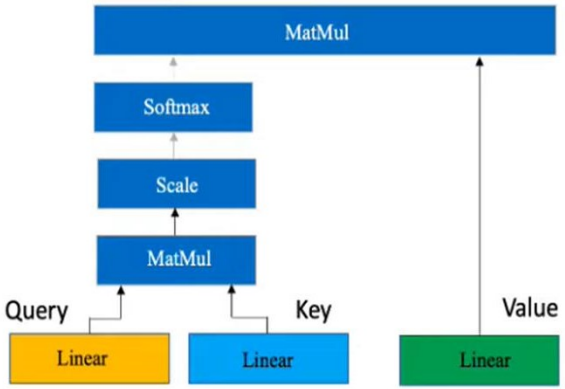


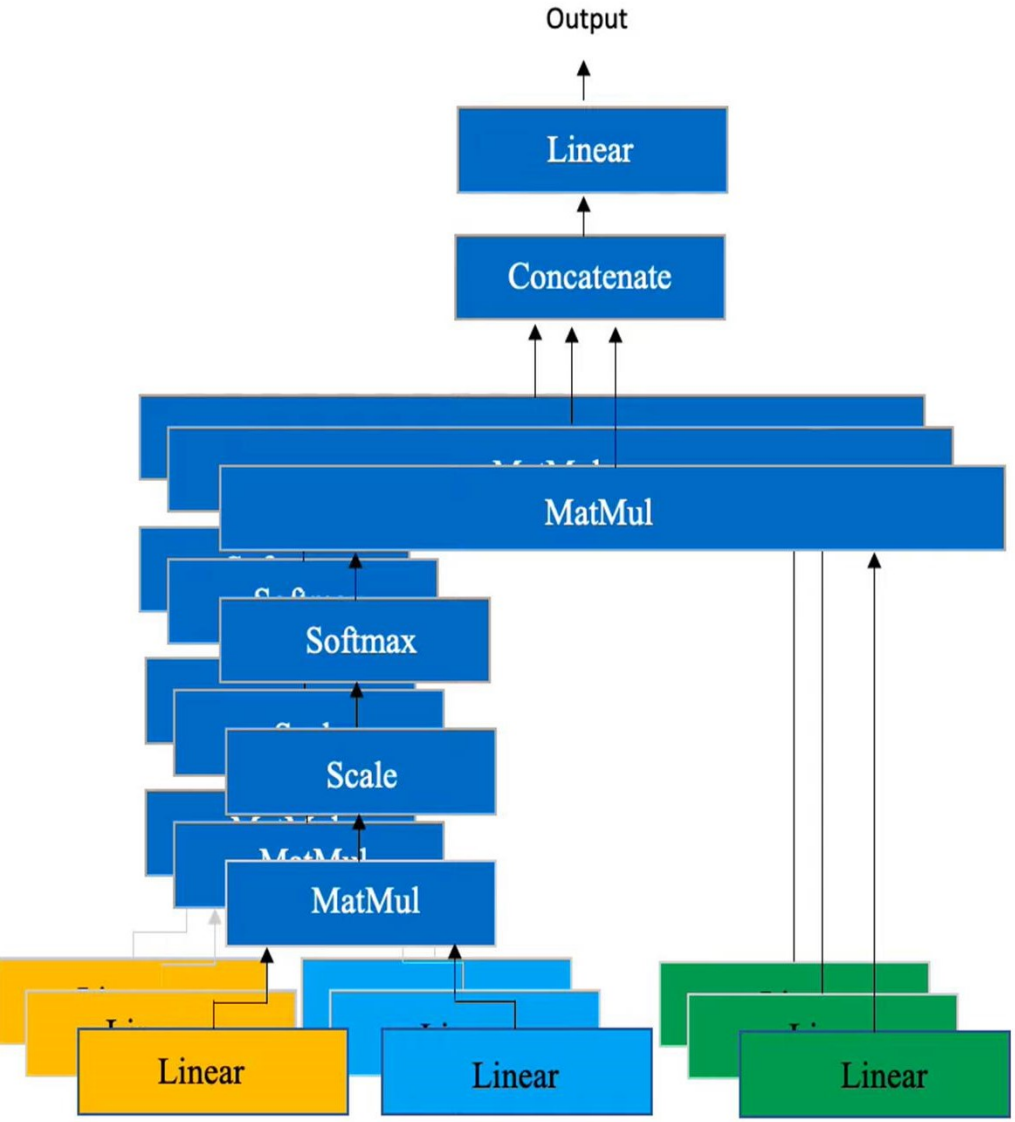




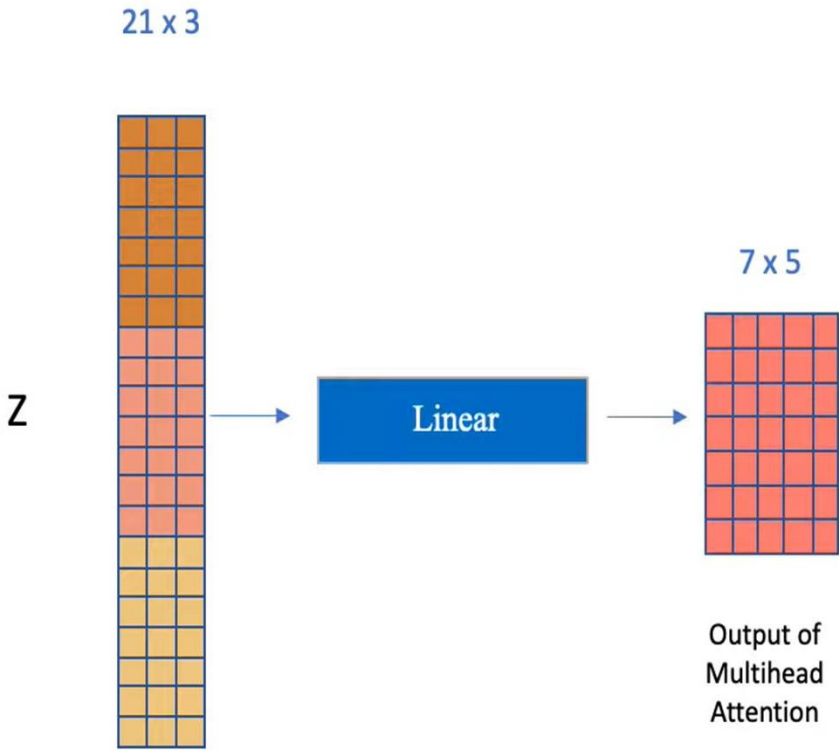








Multi-Head Attention



Training of Transformer Model

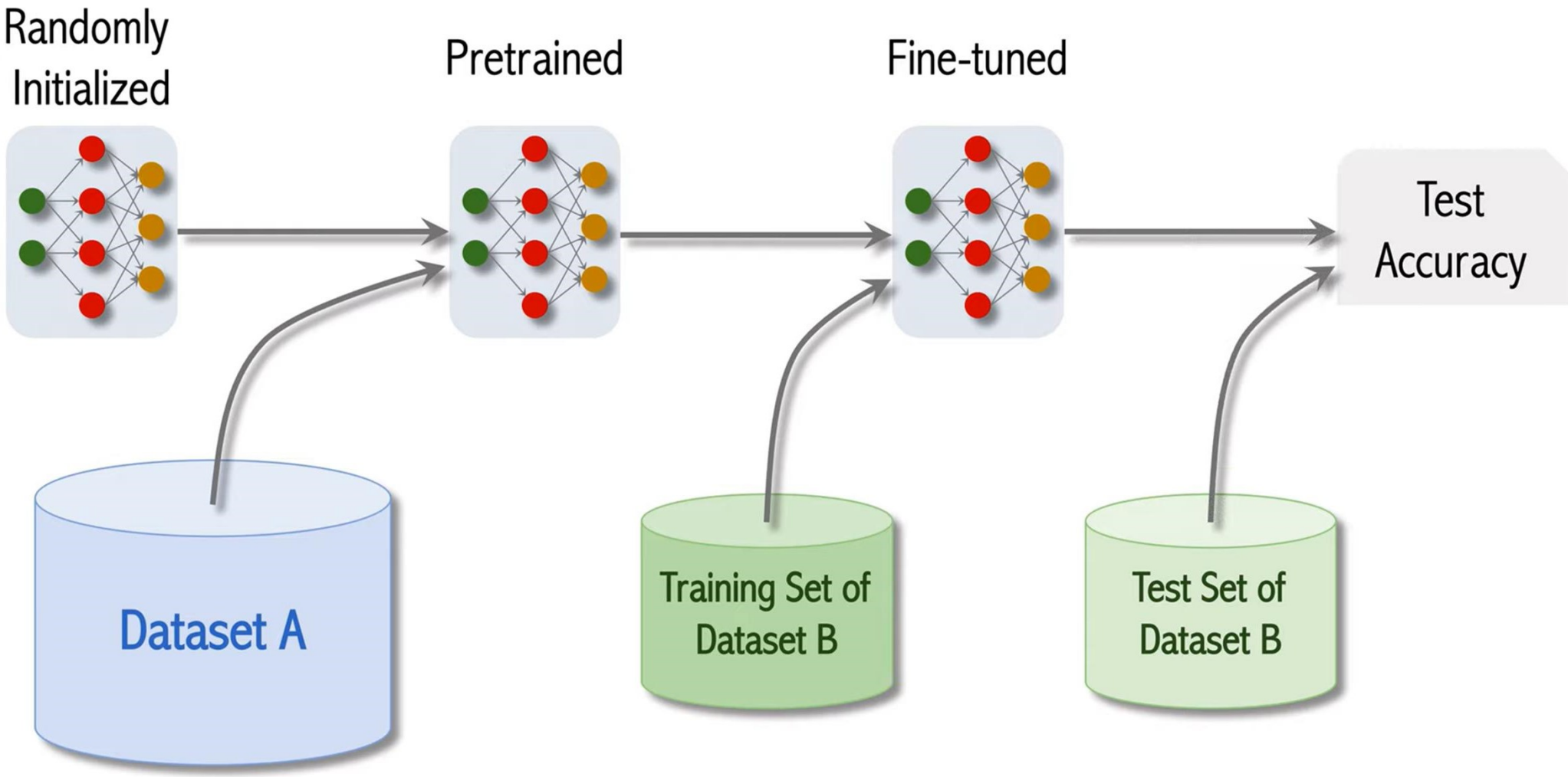
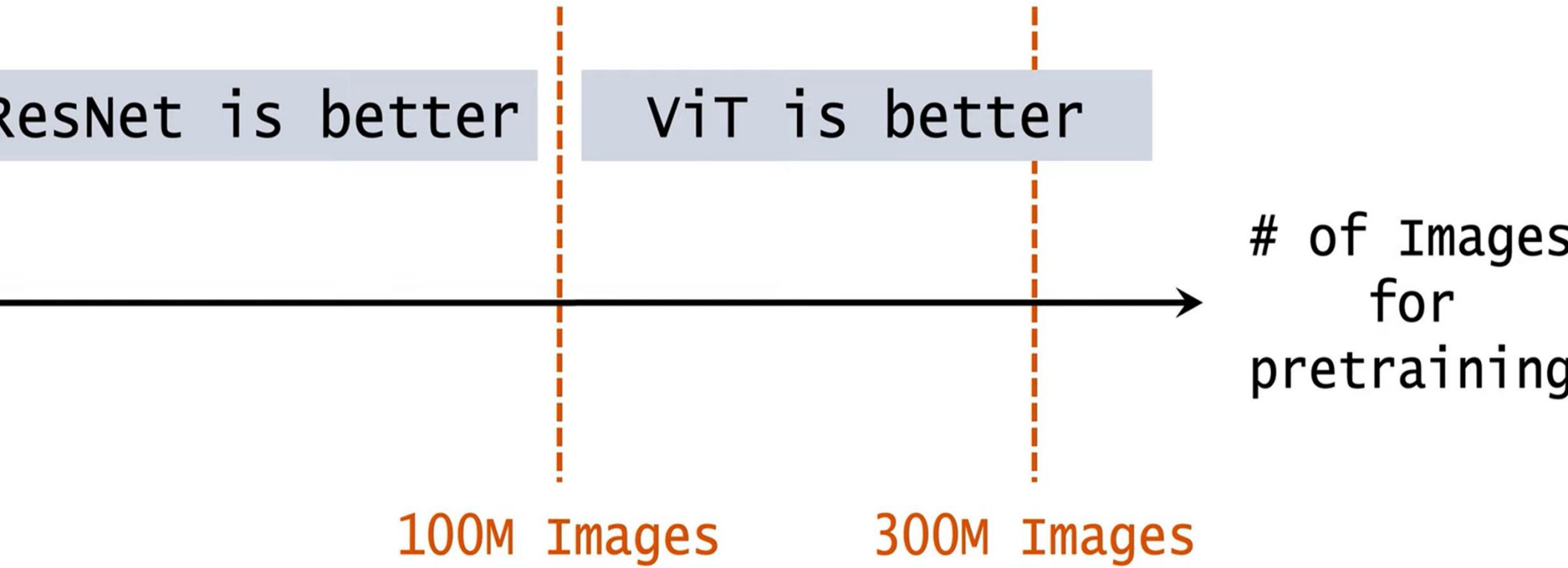


Image Classification Accuracies

- Pretrain the model on Dataset A, fine-tune the model on Dataset B, and evaluate the model on Dataset B.

- Pretrained on ImageNet (small), ViT is slightly worse than ResNet.
- Pretrained on ImageNet-21K (medium), ViT is comparable to ResNet.
- Pretrained on JFT (large), ViT is slightly better than ResNet.

Image Classification Accuracies



THANK YOU