Attention Mechanism
and
Transformer

Thanh-Sach LE

LTSACH@hcmut.edu.vn

Faculty of Computer Science and Engineering

Ho Chi Minh City Univ. of Tech., HCM-VNU
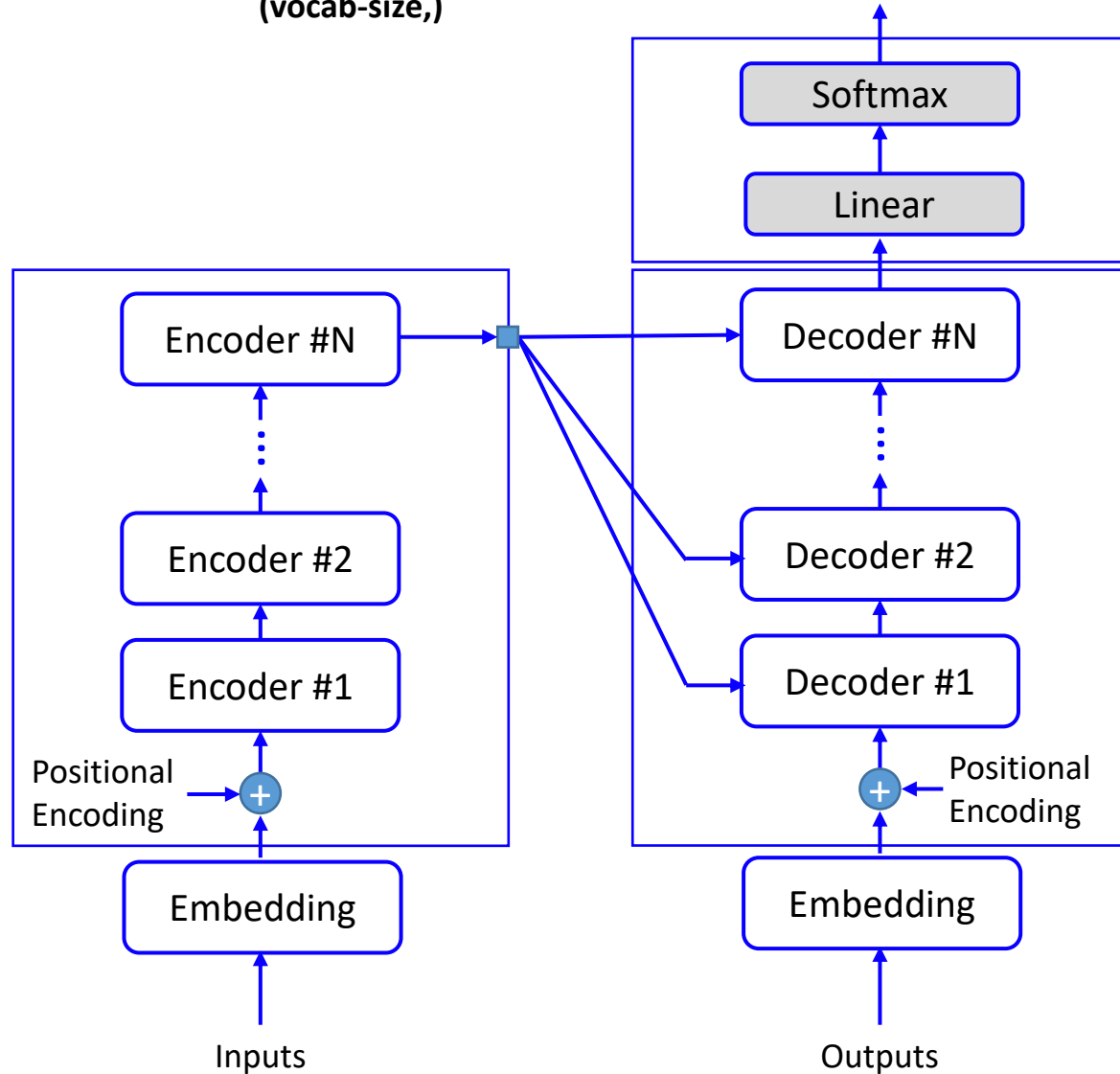
October, 2022

Content

Machine Translation
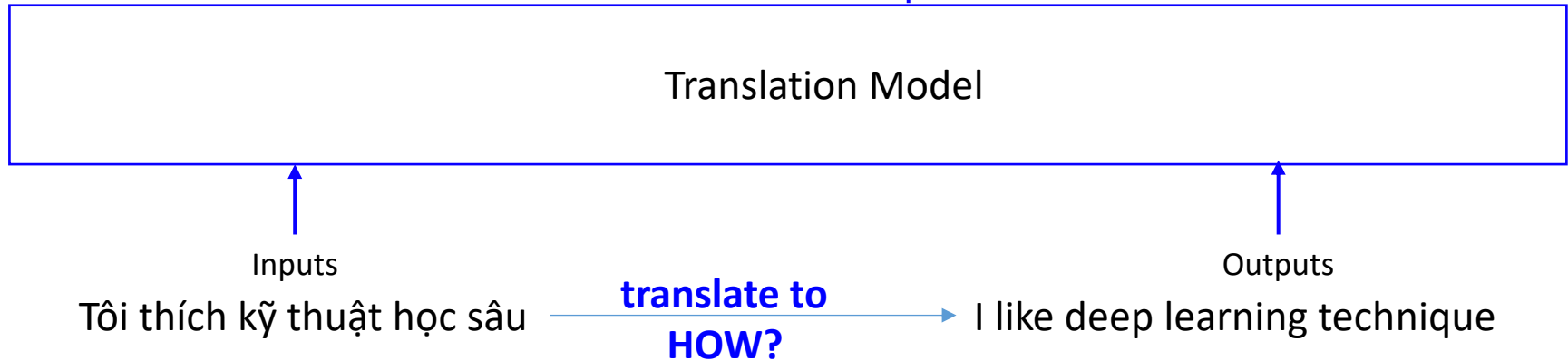
Attention and Transformer

Summary

Machine Translation

Conditional Distribution: P(*next-word* | *Inputs, previous-predicted-words*)
(vocab-size,)

Machine Translation

Conditional Distribution: P(*next-word* | *Inputs, previous-predicted-words*)
(vocab-size,)

Translation Model

Inputs

Tôi thích kỹ thuật học sâu → **translate to HOW?** → I like deep learning technique

Outputs

1. Encode(Tôi thích kỹ thuật học sâu)

2. Decode(<START>) → I

3. Decode(<START> I → like)

4. Decode(<START> I like → deep)

5. Decode(<START> I like deep → learning)

6. Decode(<START> I like deep learning → technique)

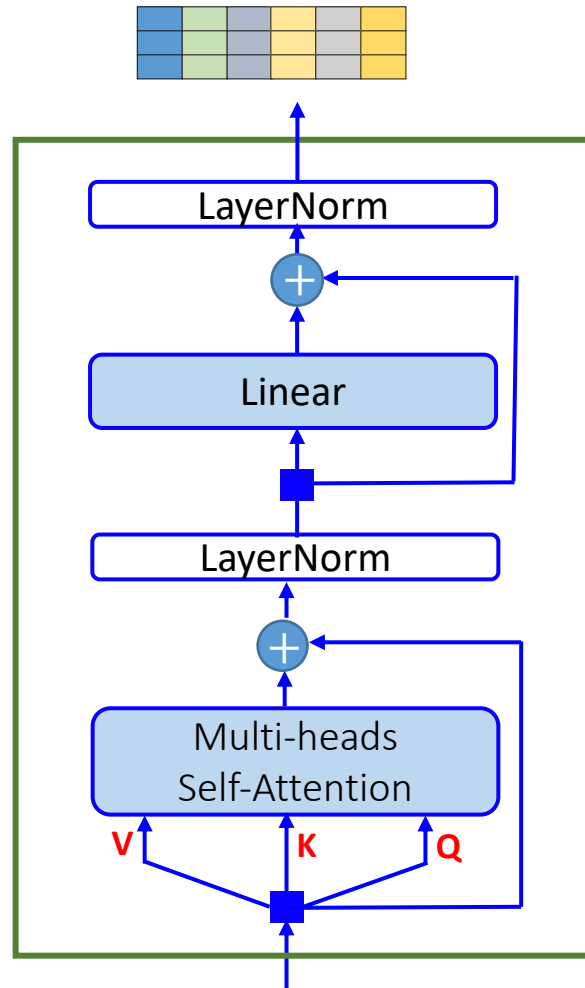7. Decode(<START> I like deep learning technique) <EOS>
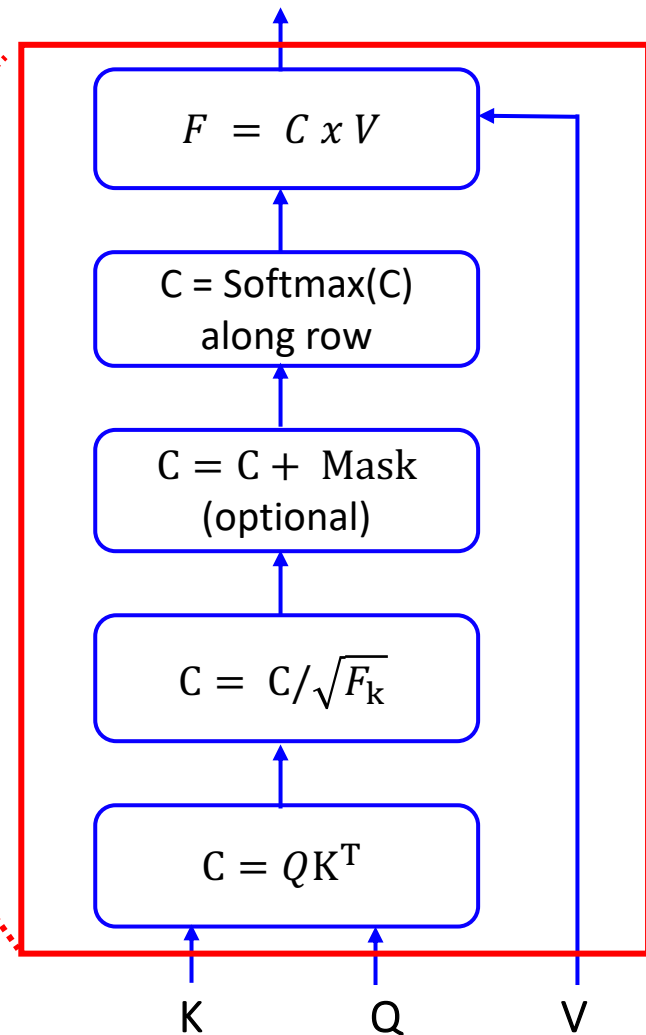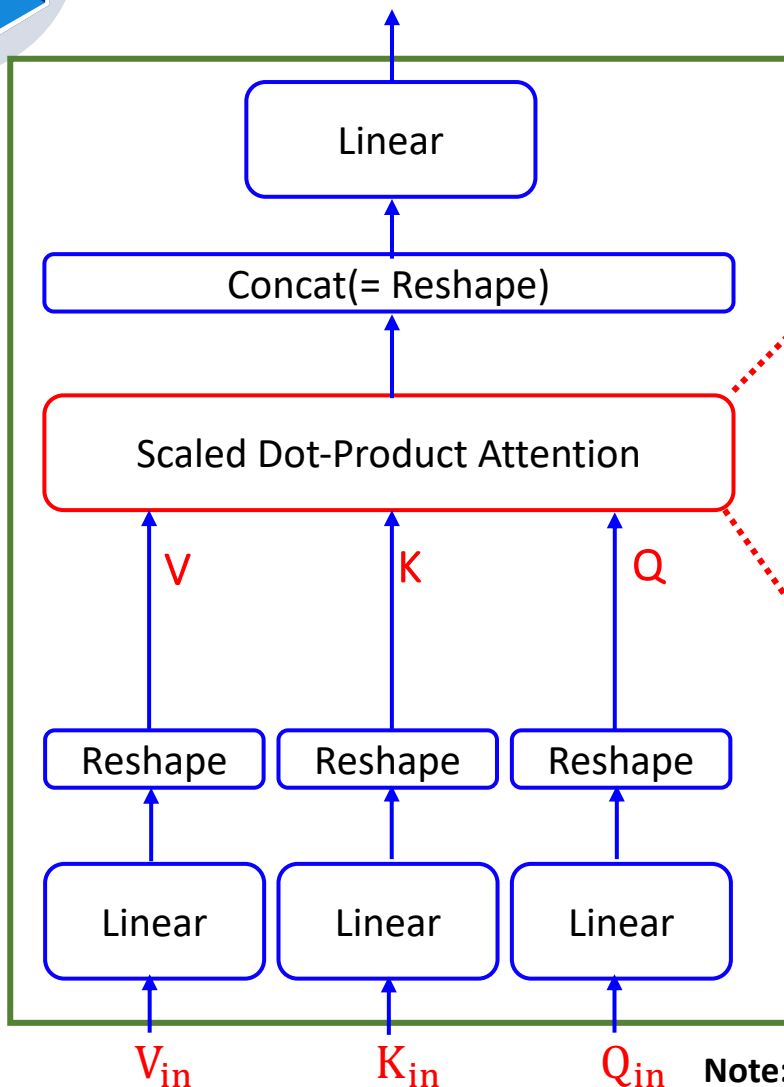
8. Stop, return: I like deep learning technique

Shape: (batch-size, ntokens, nfeatures_out)
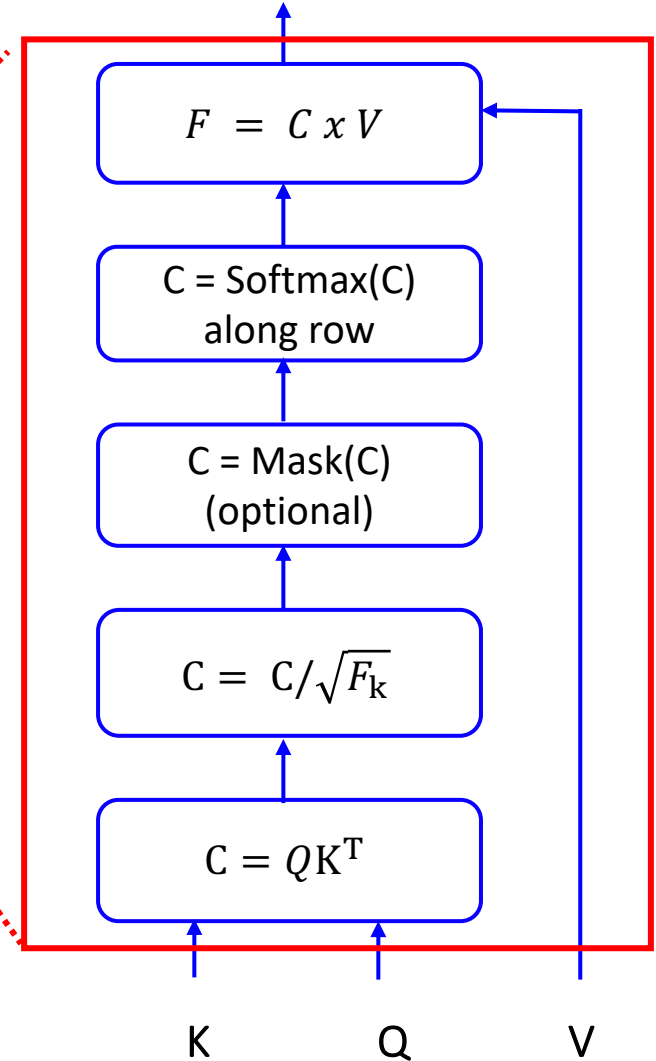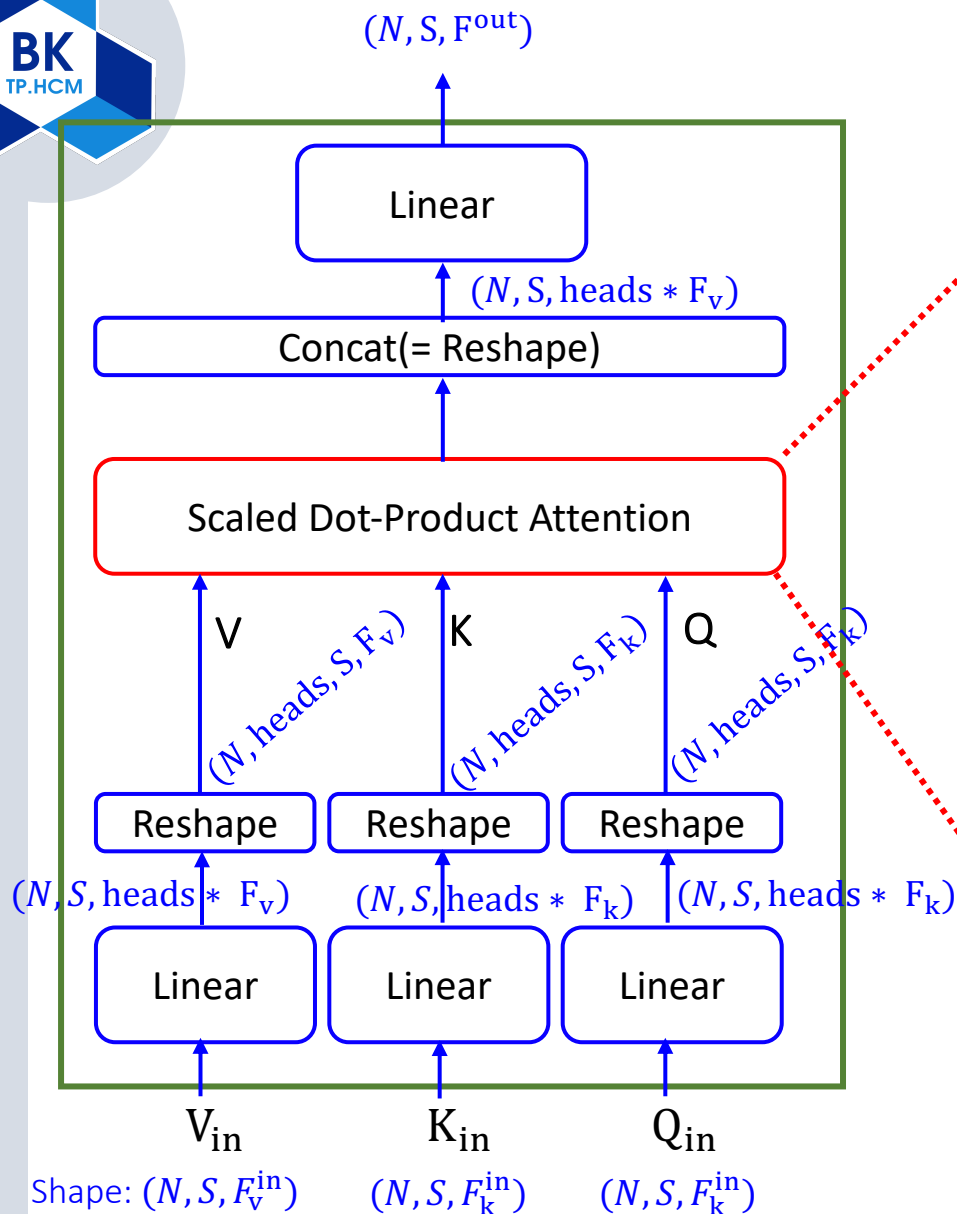

Shape: (batch-size, ntokens, nfeatures_in)

Multi-heads Self-Attention

Note:
(1) **Reshape**: actually = reshape, then transpose
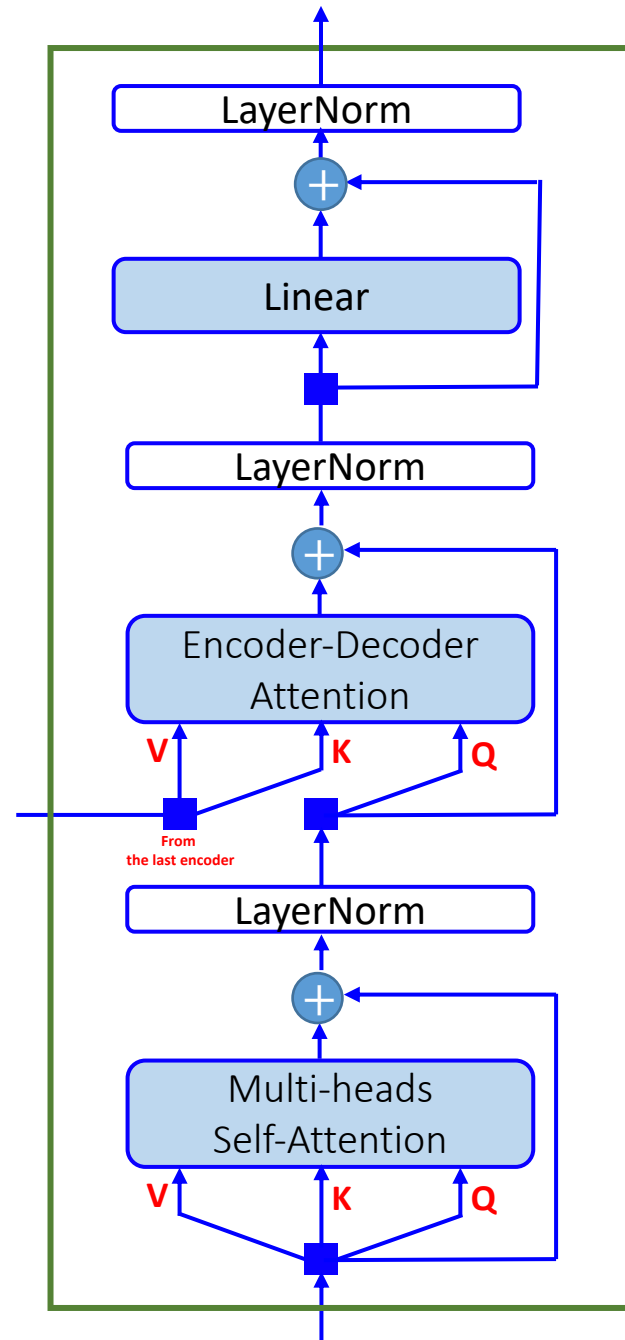(2) **Concate**: actually = transpose, then reshape
(3) **Mask**: contain 0 or –inf

Shape: (N, S, F_v^{in}) (N, S, F_k^{in}) (N, S, F_k^{in})

$(N, S, heads * F_v)$ $(N, S, heads * F_k)$ $(N, S, heads * F_k)$

$(N, heads, S, F_v)$ $(N, heads, S, F_k)$ $(N, heads, S, F_k)$

$(N, S, heads * F_v)$

(N, S, F^{out})

V_{in} K_{in} Q_{in}

Linear Linear Linear

Reshape Reshape Reshape

V K Q

Scaled Dot-Product Attention

Concat(= Reshape)

Linear

$C = QK^T$

$C = C/\sqrt{F_k}$

C = Mask(C) (optional)

C = Softmax(C) along row

$F = C \ x \ V$

K Q V

DsciLab
LTSACH@hcmut.edu.vn

7

Note:
(1) **Reshape**: actually = reshape, then transpose
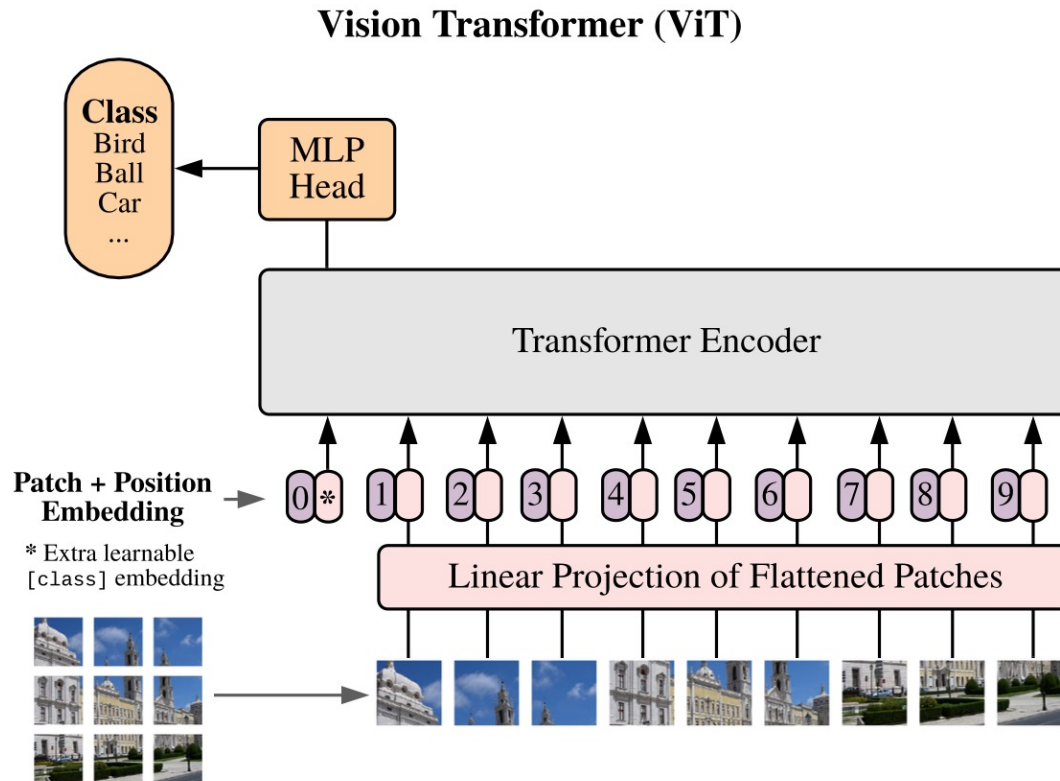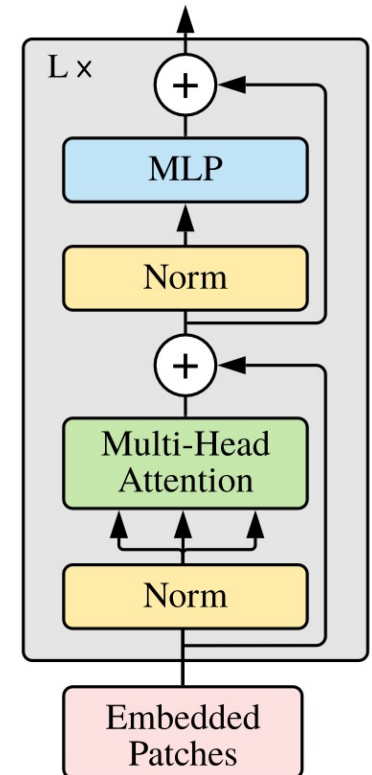**(2)** **Concate**: actually = transpose, then reshape

Decoder >>

Vision Transformer (ViT)

References

❖Attention Mechanisms — Multi-Head attention (with code)
> ➢"https://medium.com/@smitasasindran/12-attention-mechanisms-multihead-attention-958041a35553

❖Transformers Explained Visually (Part 2): How it works, step-by-step
> ➢https://towardsdatascience.com/transformers-explained-visually-part-2-how-it-works-step-by-step-b49fa4a64f34

❖How Transformers work in deep learning and NLP: an intuitive introduction
> ➢https://theaisummer.com/transformer/

❖Attention Mechanisms and Transformers (with code)
> ➢https://d2l.ai/chapter_attention-mechanisms-and-transformers/multihead-attention.html#sec-multihead-attention

❖Keras, ViT
> ➢https://keras.io/examples/vision/image_classification_with_vision_transformer/

Thank you for your listening

Welcome to your sharing and comments!

Lê Thành Sách

LTSACH@hcmut.edu.vn