Friedrich-Alexander-Universität
Technische Fakultät

Vision Transformer – A Deep Learning model for Image Analysis and Understanding

CML: Control, Machine Learning and Numerics

Group 29

Apurwa Agrawal

Masters in Data Science
Mtrikel Nr. 23161645

Kshitij Dua

Masters in Data Science
Mtrikel Nr. 23032110

Sneha Kumari

Masters in Data Science
Mtrikel Nr. 22970359

Amit Jadhav

Masters in Data Science
Mtrikel Nr. 23036144

Rajesh Madhipati

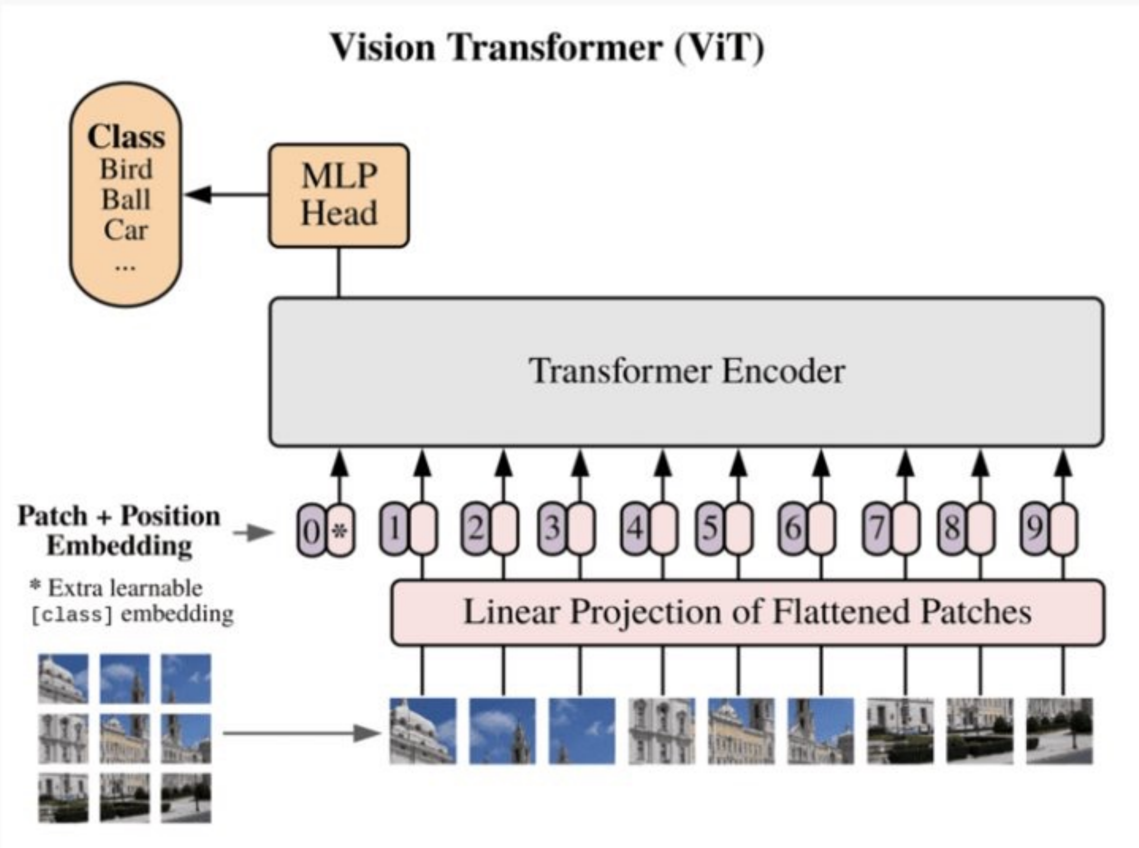
Masters in Data Science
Mtrikel Nr. 23007086

Table of Content

Introduction

Introduction

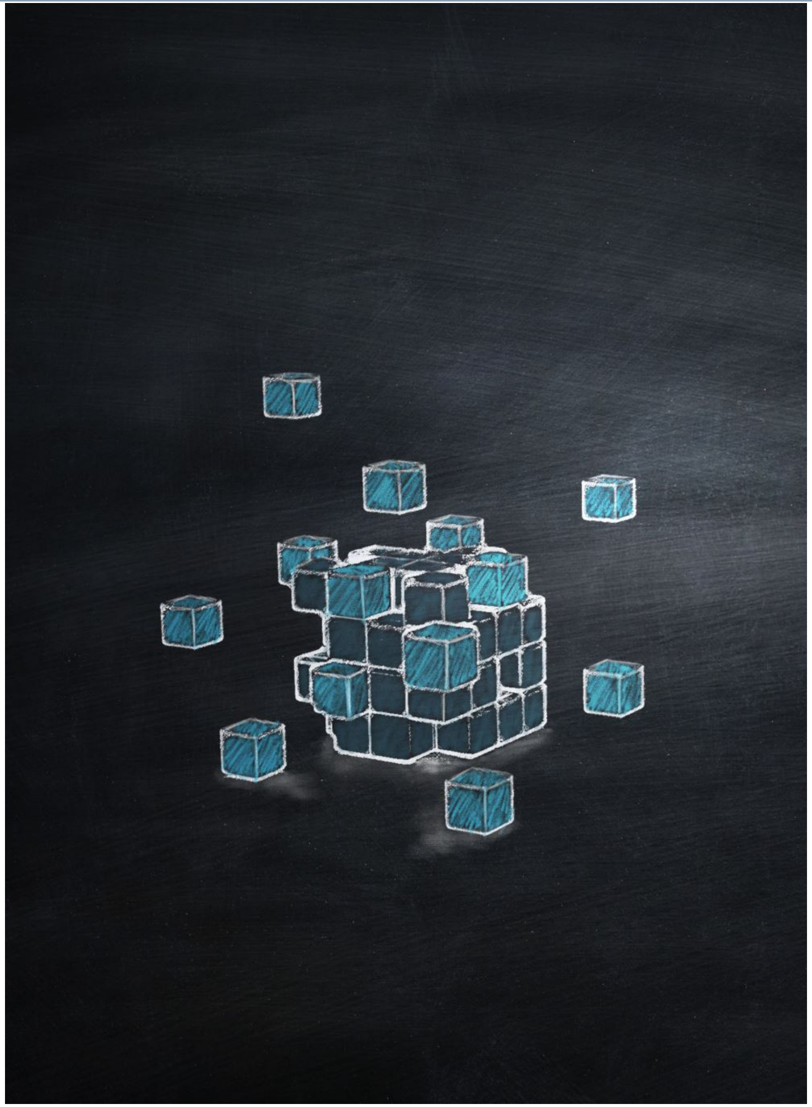
- Ability to extract meaningful insights from images is crucial for driving advancements across various industries and disciplines.
- CNNs faces problem in **capturing long-range dependencies** and modelling global context within images.
- ViT extends the success of the Transformer architecture.
- ViT understand **complex relationships** within images and make informed predictions.



Motivation

Motivation

- Capturing Long-Range Dependencies.
- Flexibility with Variable-Sized Inputs.
- Transferability and Generalization.
- Learning Abstract Representations.



Evolution of Transformers

Evolution of Transformers

2017.6 | Transformer
Solely based on attention mechanism, the Transformer is proposed and shows great performance on NLP tasks.

2018.10 | BERT
Pre-training transformer models begin to be dominated in the field of NLP.

2020.5 | GPT-3
A huge transformer with 170B parameters, takes a big step towards general NLP model.

2020.5 | DETR
A simple yet effective framework for high-level vision by viewing object detection as a direct set prediction problem.

2020.7 | iGPT
The transformer model for NLP can also be used for image pre-training.

2020.10 | ViT
Pure transformer architectures work well for visual recognition.

End of 2020 | IPT/SETR/CLIP
Applications of transformer model on low-level vision, segmentation and multimodality tasks, respectively.

2021 | ViT Variants
Variants of ViT models, e.g., DeiT, PVT, TNT, and Swin.

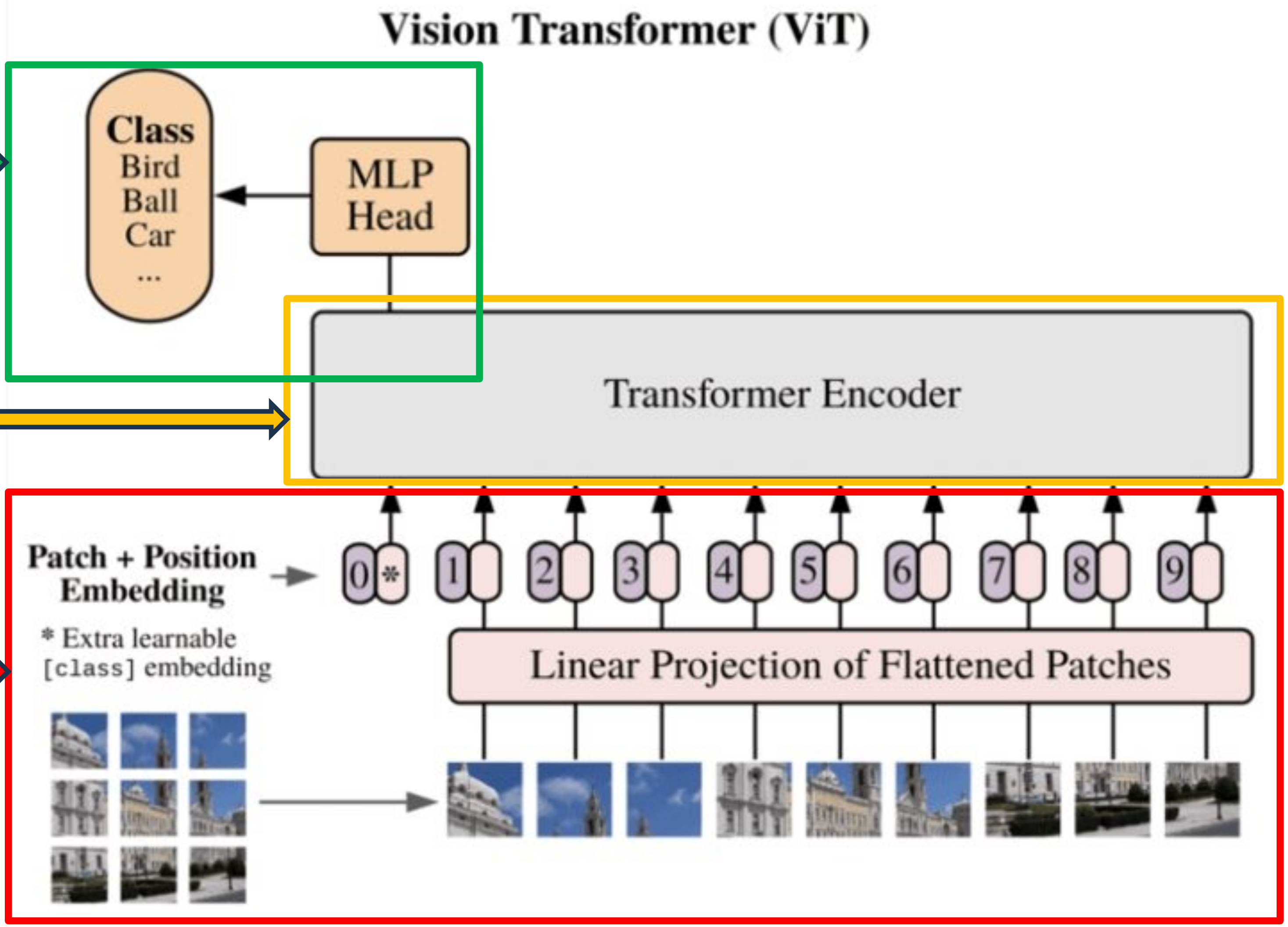
Vision Transformer Architecture

3 Parts of Architecture

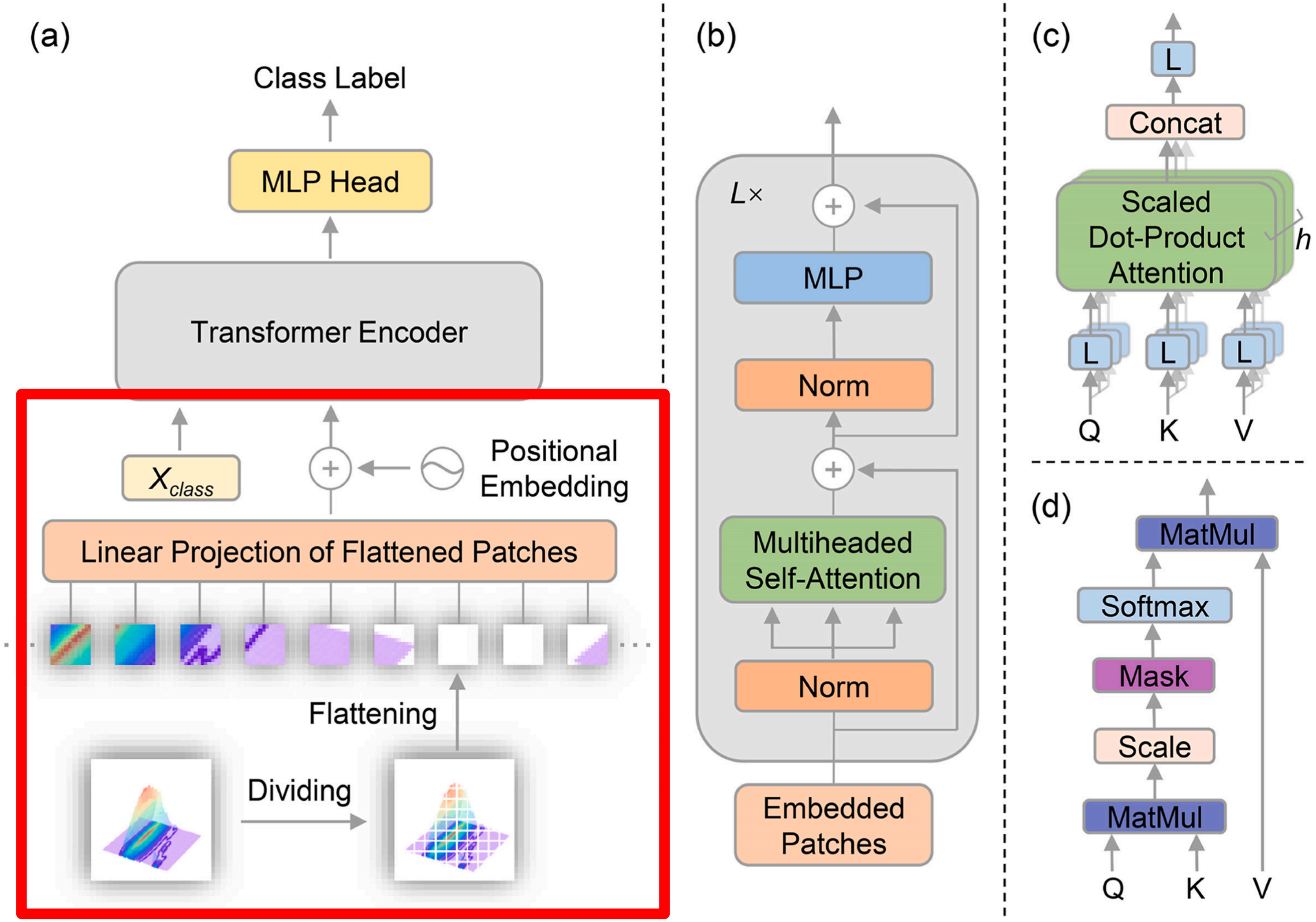
3. The classification head.

2. Feature extraction via stacked transformer encoders.

1. Patch embedding.



1. Patch Embedding



1. Patch Embedding

Batch of
images **b**

1. Patch Embedding

Batch of
images **b**

One of the images
(b, h, w, c)

b = batch
h = height
w = width
c = channel

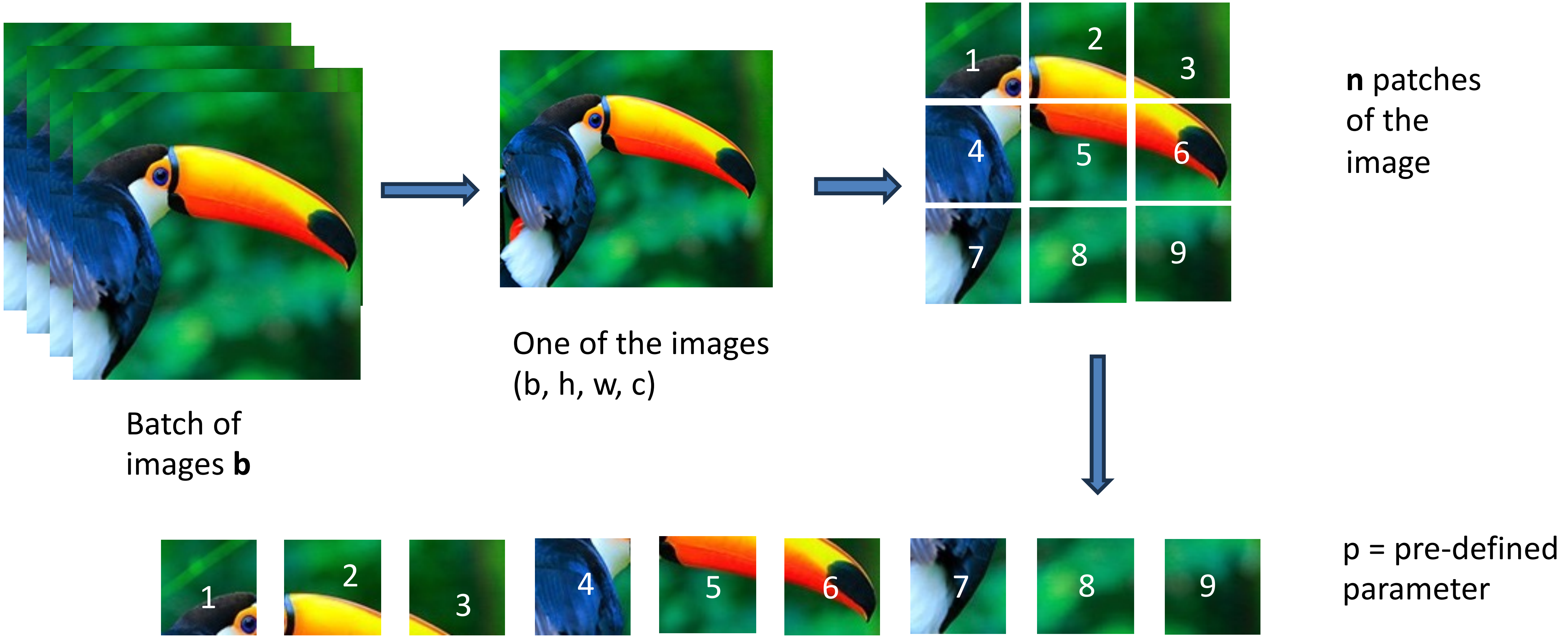
1. Patch Embedding

Batch of
images **b**

One of the images
(b, h, w, c)

n patches
of the
image

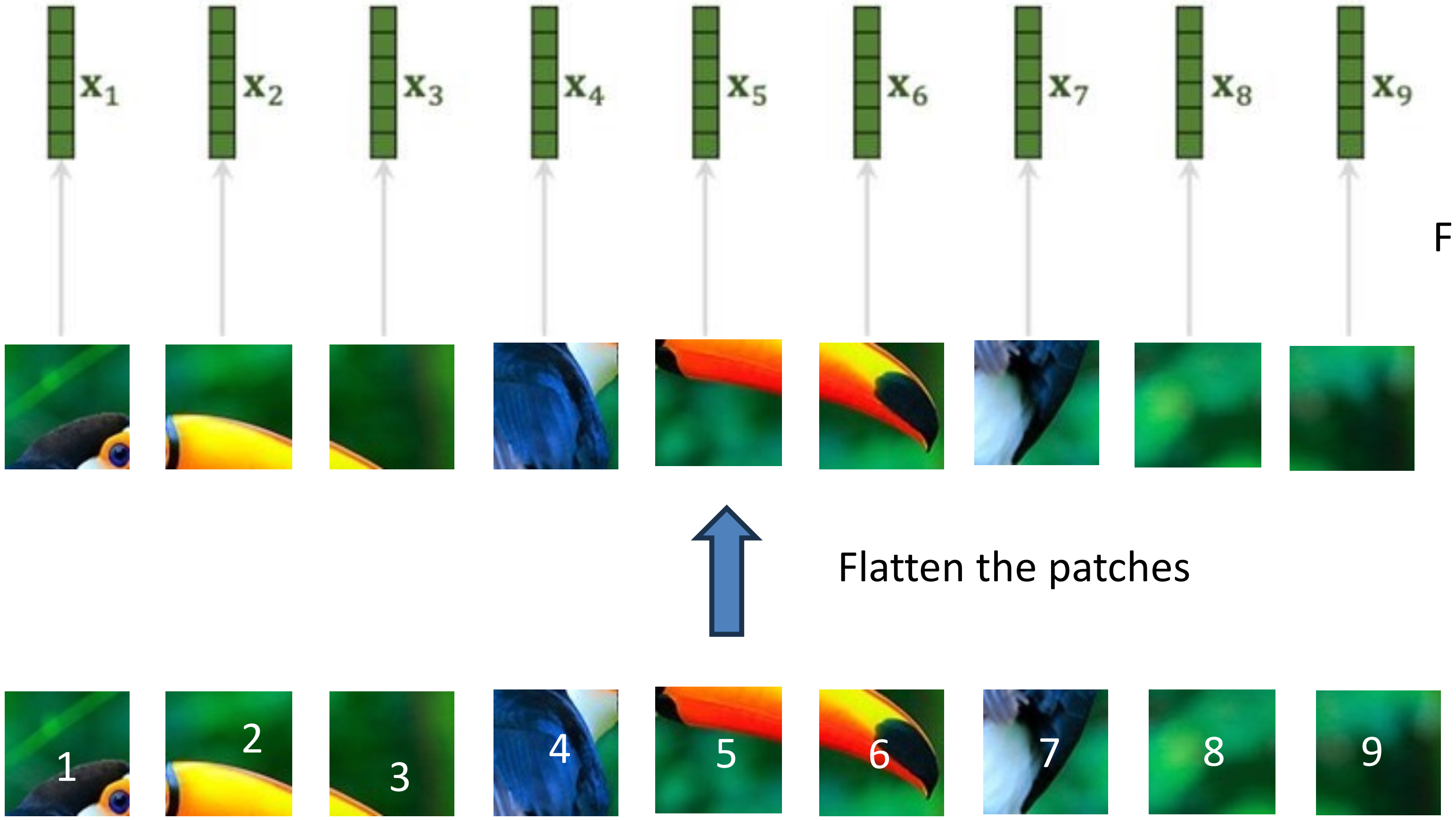
1. Patch Embedding



The image is split into n square patches of shape (p,p,c),

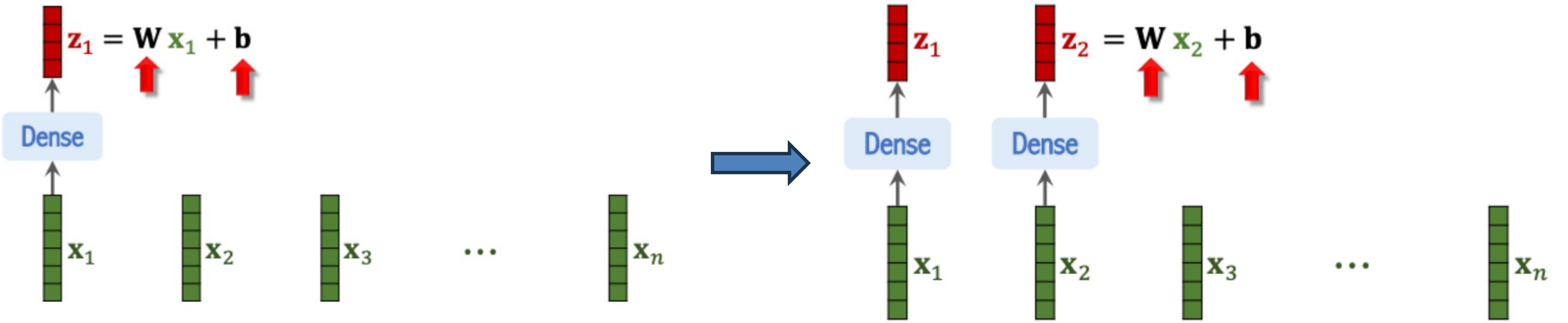
1. Patch Embedding

Flat patches (b, n, (p^2*c))



Every patch has a feature vector of dimesion (1, p^2*c)

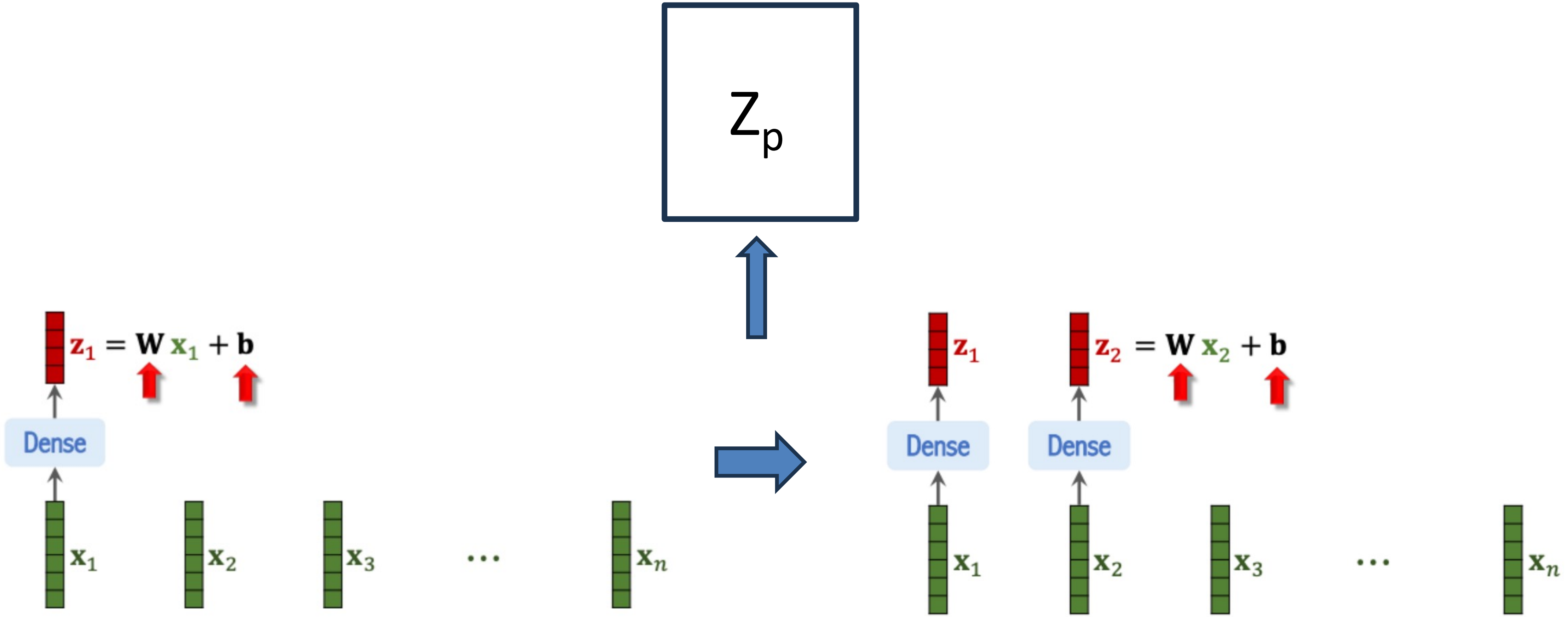
1. Patch Embedding



Embeddings ((p^2 * c), d)

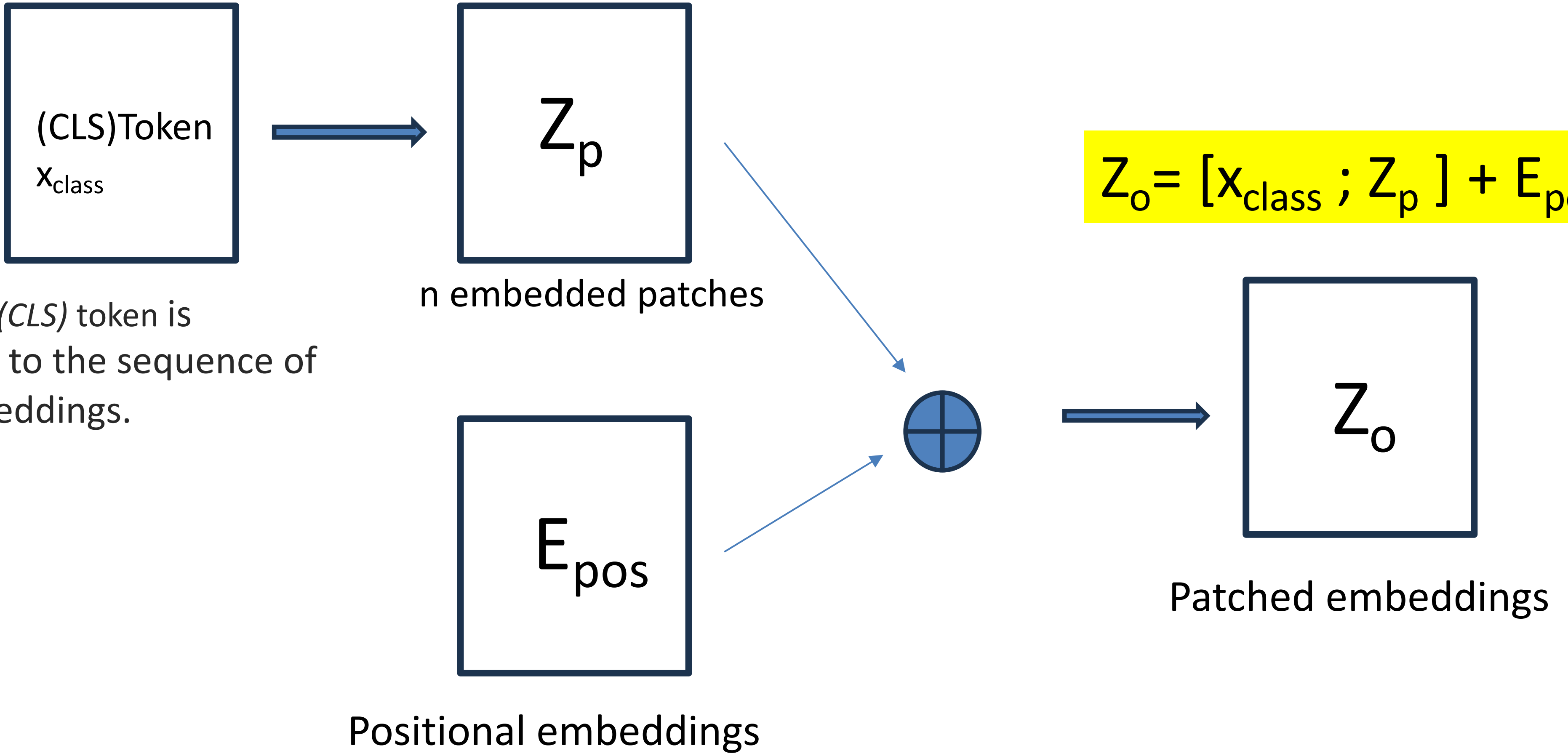
The flattened patches are multiplied with a **trainable** embedding tensor, which learns to linearly project each flat patch

1. Patch Embedding



Embeddings

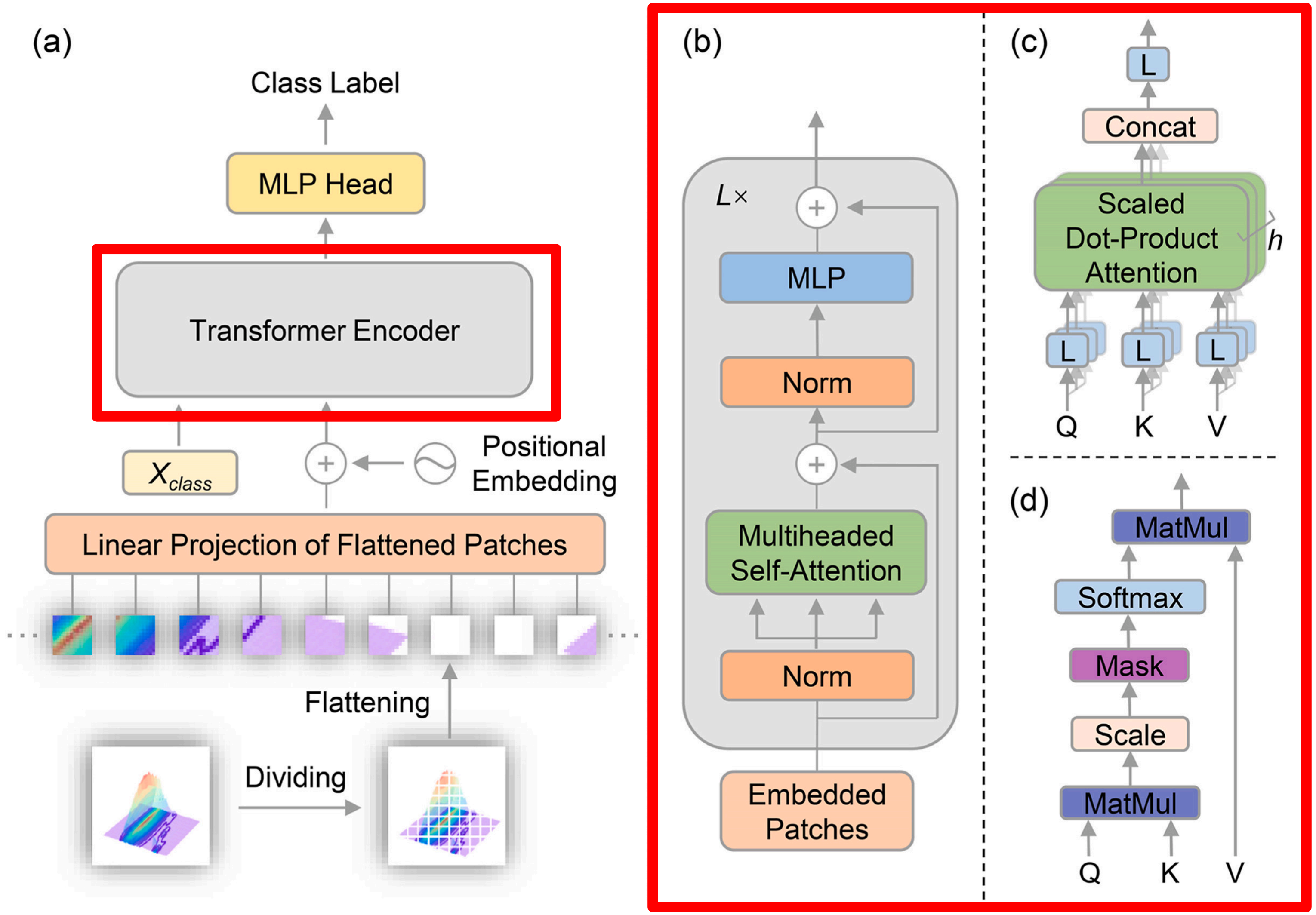
1. Patch Embedding



A learnable *(CLS)* token is prepended to the sequence of patch embeddings.

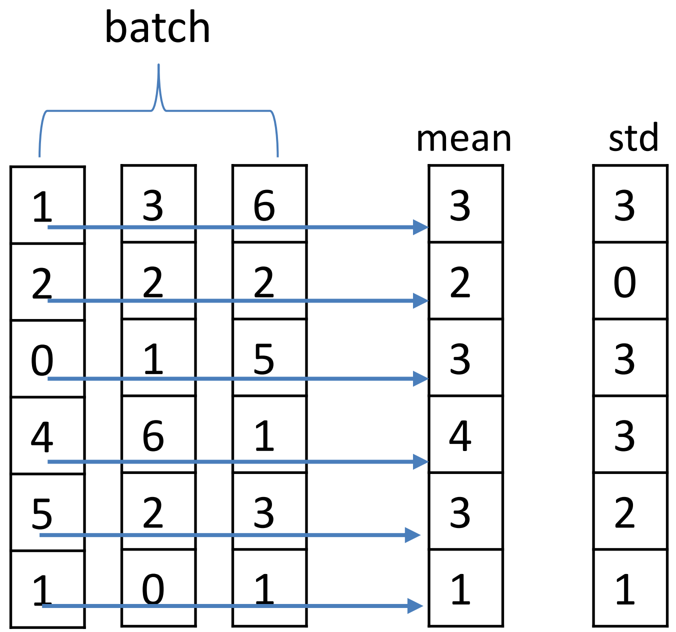
$$Z_o = [x_{class} ; Z_p] + E_{pos}$$

2. Transformer Encoding



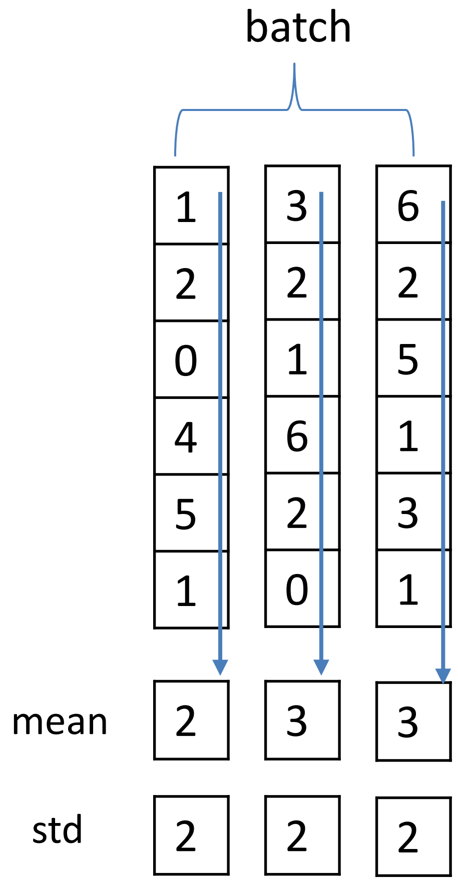
2.1 Layer Normalization

Batch Normalization



Sample for all training examples

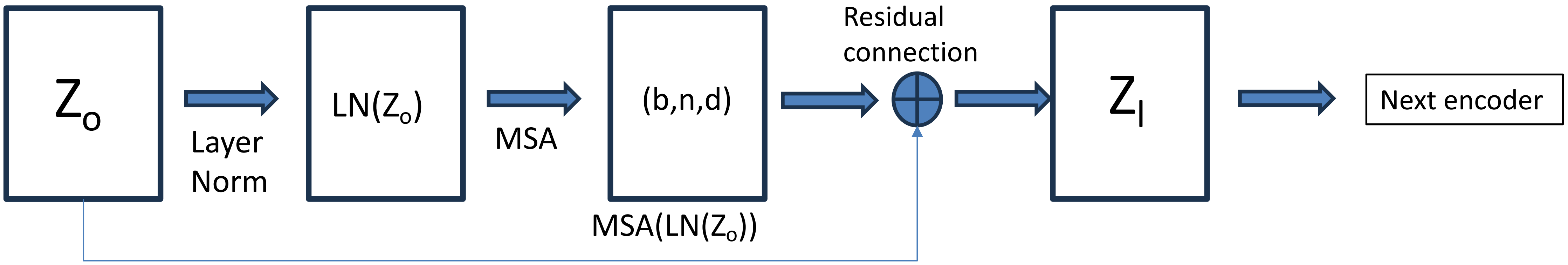
Layer Normalization



Same for all features dimensions

2.1 Layer Normalization

2. Transformer Encoding



Layer Normalization:

$$\mu_i = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$$

$$\sigma_i^2 = \frac{1}{n}\sum_{i=1}^{n}(zij_{-}\mu_i)^2$$

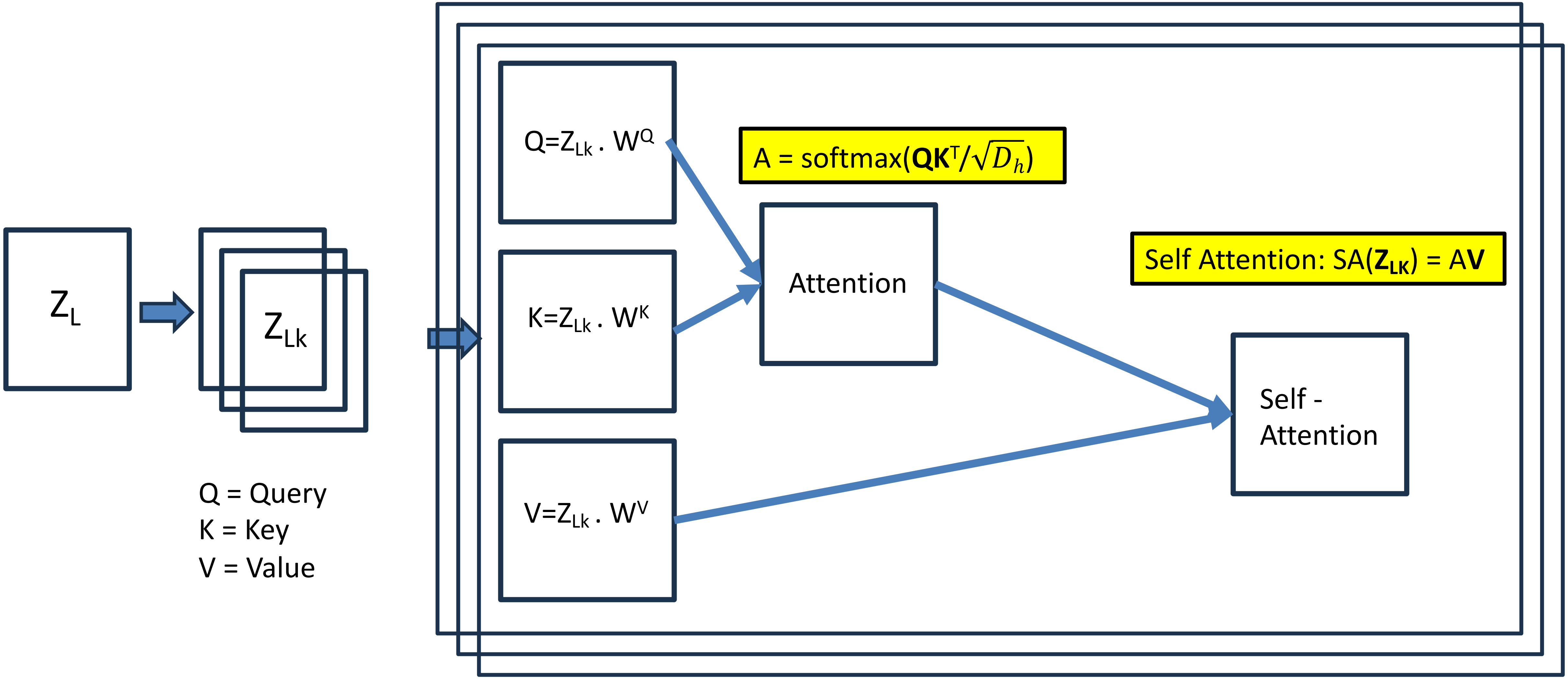
$$\hat{X}_{ij} = \frac{z_{ij} - \mu_i}{\sqrt{\sigma^2_i + \varepsilon}}$$

$$LN(Z_o) = \hat{X}$$

Z'_e = MSA(LN(Z_o)) + Z_l , where e = 1 … L (2)

2.2 Multi-head Attention

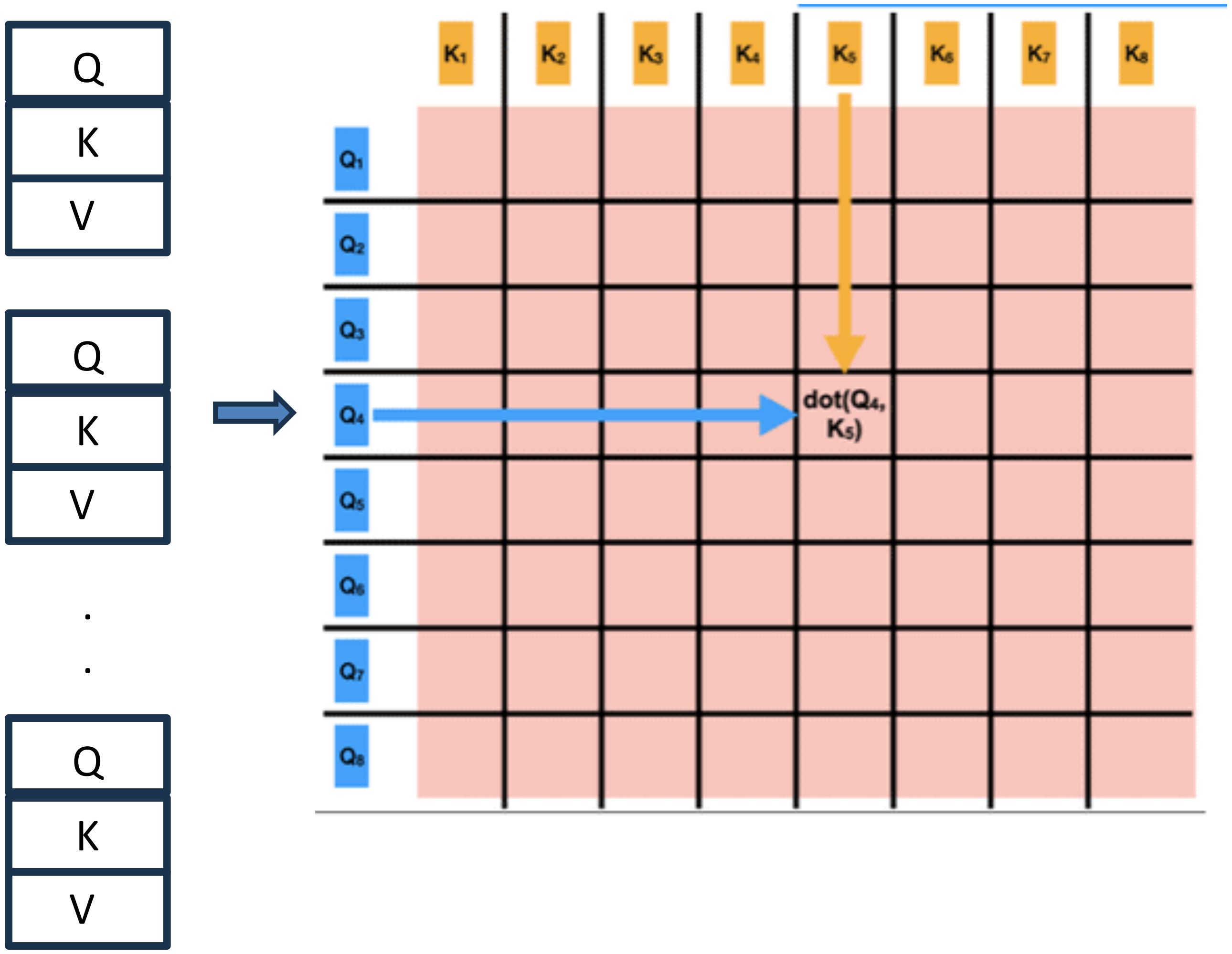
2. Transformer Encoding



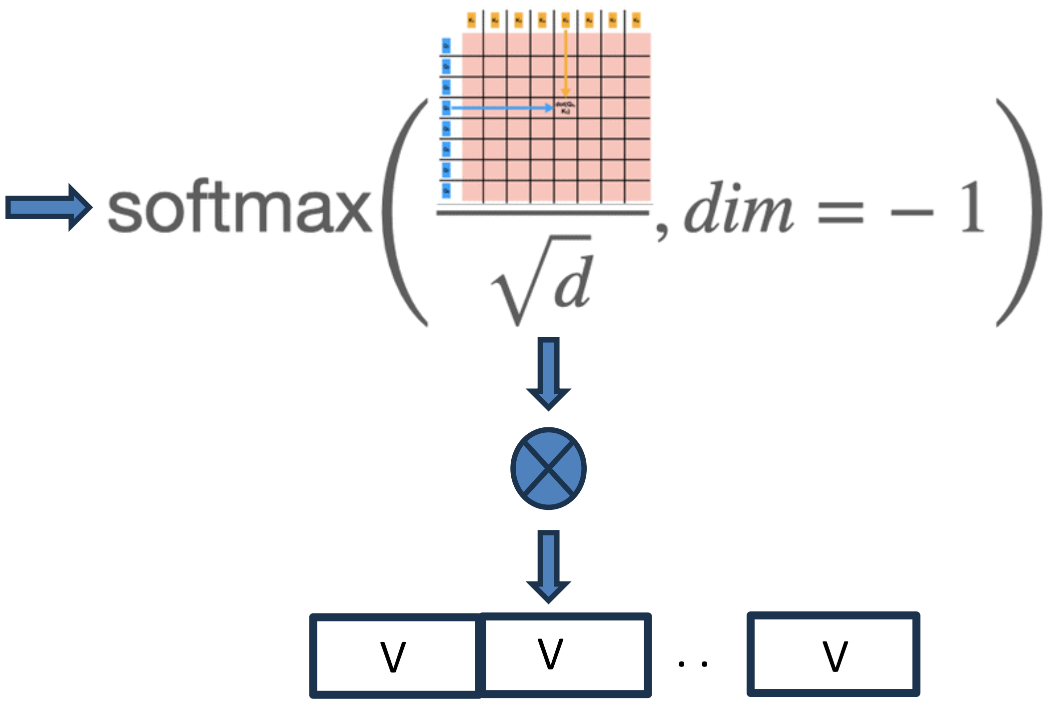
Q = Query
K = Key
V = Value

2.2 Multi-head Attention

2. Transformer Encoding



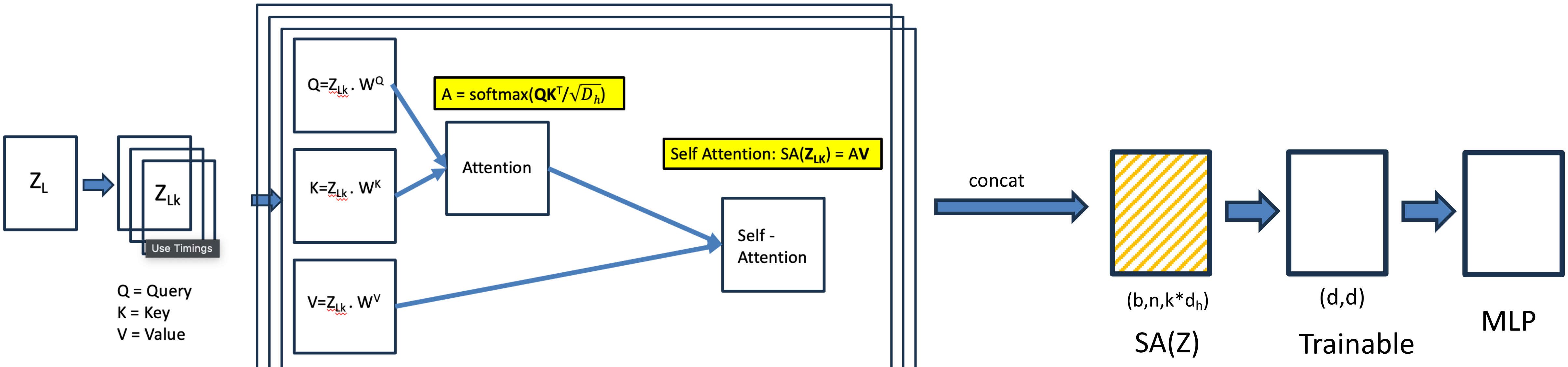
$A = \text{softmax}(\mathbf{QK}^T/\sqrt{D_h})$



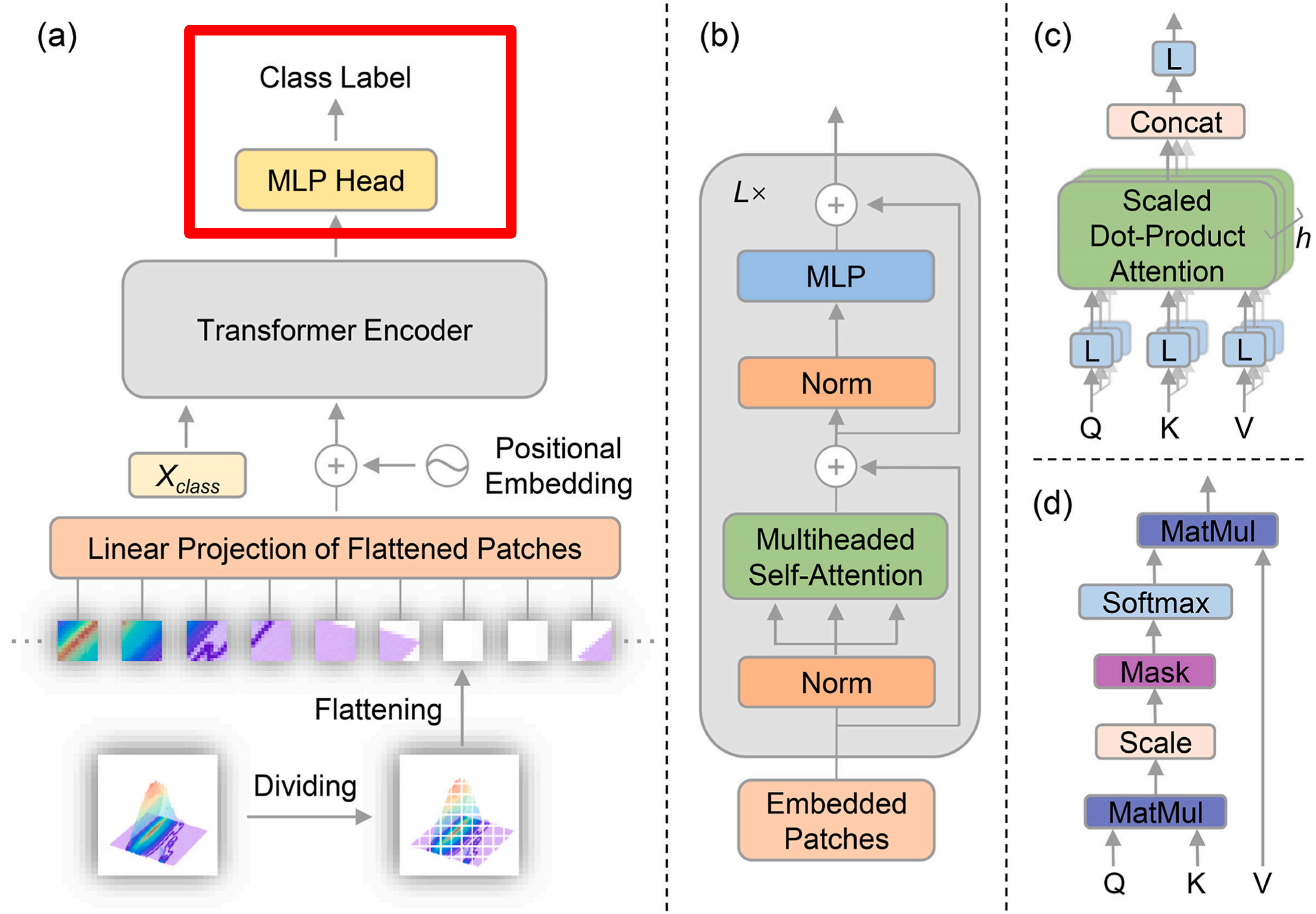
Self Attention: $SA(\mathbf{Z}_{LK}) = A\mathbf{V}$

2.2 Multi-head Attention

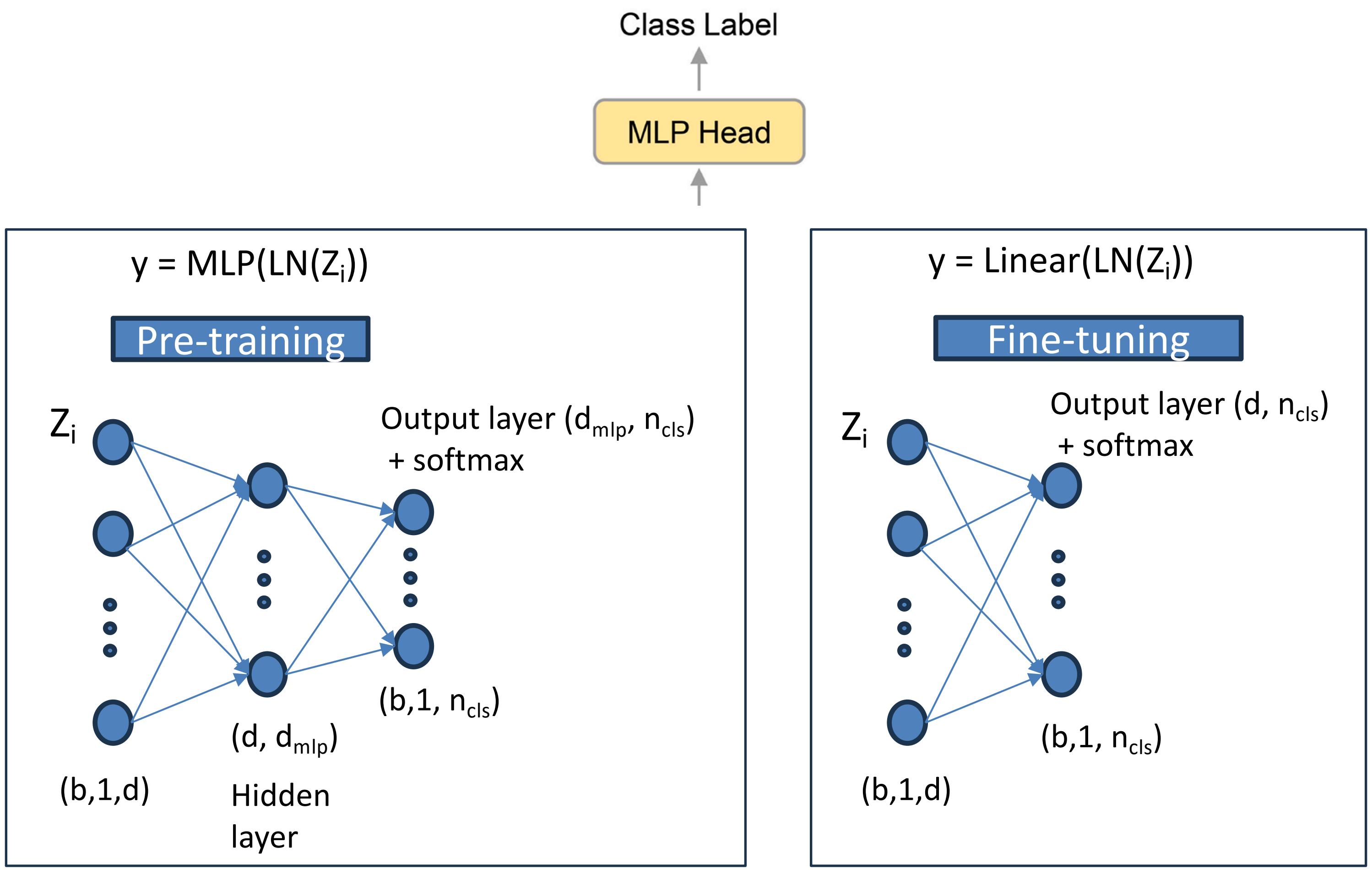
2. Transformer Encoding



3. Classification Head



3. Classification Head



- [cls] token is used in the classification head.
- Pre-training - 2 layer of MLP used, hence 2 weight matrices
  - W_h [d, d_{mlp}]
  - W_o [d_{mlp}, d]
- Fine-tuning – single layer used, hence only 1 tensor [d, n_cls]
- Output: Probability associated with each of n_{cls} classes

Training Vision Transformer

Training Vision Transformers

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Applications of ViT

- Image Classification
- Image captioning
- Object Detection
- Semantics Segmentation
- Video Understanding
- Image Generation

Performance, Limitations and Future Improvements

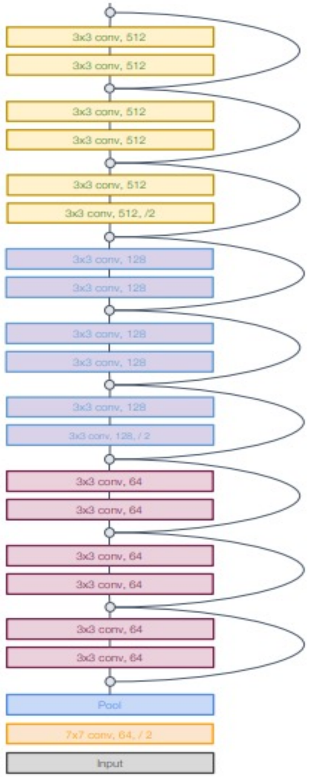
VIT vs CNN

Stage 3:
256 x 14 x 14

Stage 2:
128 x 28 x 28

Stage 1:
64 x 56 x 56

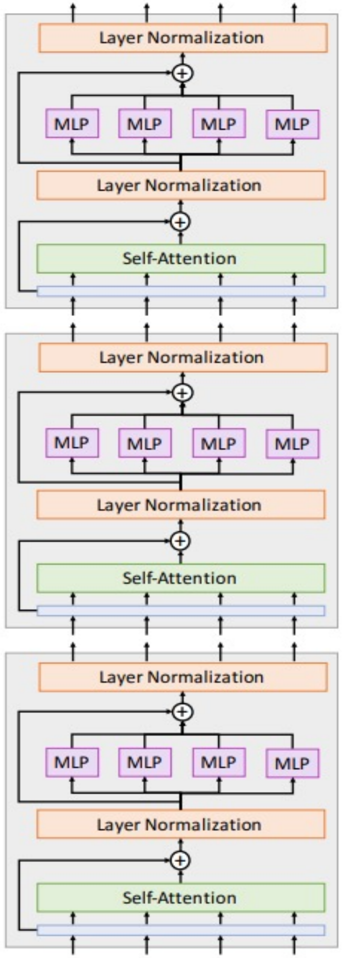
Input:
3 x 224 x 224



In most CNNs (including ResNets), **decrease** resolution and **increase** channels as you go deeper in the network (Hierarchical architecture)

Useful since objects in images can occur at various scales

In a ViT, all blocks have same resolution and number of channels (Isotropic architecture)



3rd block:
768 x 14 x 14

2nd block:
768 x 14 x 14

1st block:
768 x 14 x 14

Input:
3 x 224 x 224

Limitations of ViT

- Limited spatial information
- Computational Complexity
- Generalization to Diverse Data
- Memory and Computational Efficiency
- Continual Learning and Adaptability

Discussion and Conclusion

- **Superior Performance:**
  - Demonstrated exceptional performance, surpassing traditional CNNs in various image analysis tasks..
- **Global Context Modelling:**
  - capture global contextual information, enabling modeling of long-range dependencies and interactions between image elements.
- **Scalability and Adaptability:**
  - Scalability and Adaptability: ViTs are highly scalable and adaptable, making them suitable for diverse applications and datasets.

Future Improvements

- Enhanced attention mechanism for better long-range dependency modeling.
- Improved model efficiency with reduced computational complexity.
- Advancements in interpretability for better understanding of attention maps.
- Integration of multimodal information for enhanced performance.

References

- Han, Kai & Wang, Yunhe & Chen, Hanting & Chen, Xinghao & Guo, Jianyuan & Liu, Zhenhua & Tang, Yehui & Xiao, An & Xu, Chunjing & Xu, Yixing & Yang, Zhaohui & Zhang, Yiman & Tao, Dacheng. (2022). **A Survey on Vision Transformer.** IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 1-1. 10.1109/TPAMI.2022.3152247.

- C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi and J. Zhong, "**Attention Is All You Need In Speech Separation**," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 21-25, doi: 10.1109/ICASSP39728.2021.9413901.

- Dosovitskiy, Alexey & Beyer, Lucas & Kolesnikov, Alexander & Weissenborn, Dirk & Zhai, Xiaohua & Unterthiner, Thomas & Dehghani, Mostafa & Minderer, Matthias & Heigold, Georg & Gelly, Sylvain & Uszkoreit, Jakob & Houlsby, Neil. (2020). **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.**

Thankyou