Introduction to Computer Vision: Vision Transformers

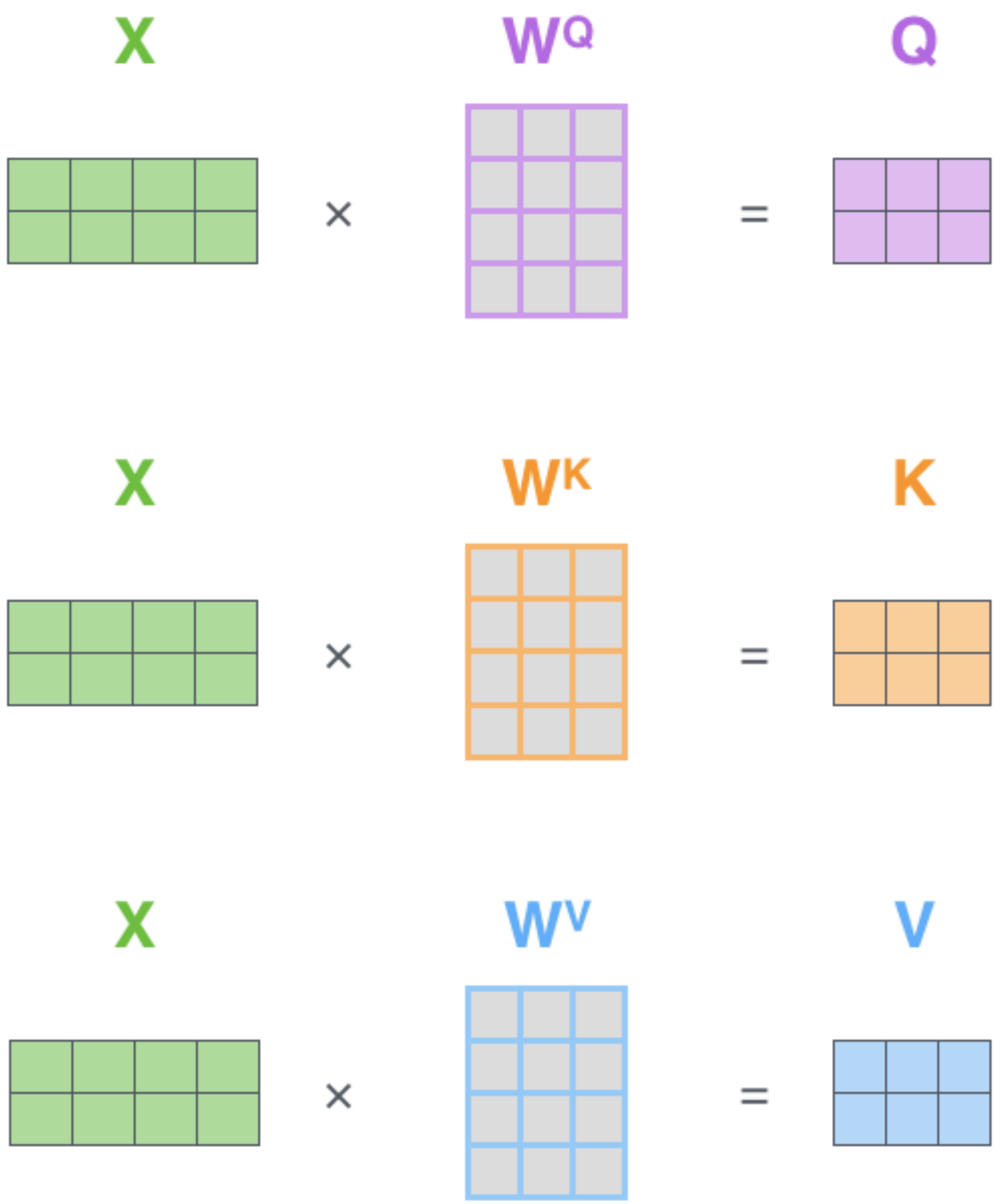
Laurens van der Maaten

Self-attention

- Suppose we receive a set of inputs, for example, image patches
- Each input is represented by an embedding vector
- Self-attention compares all input embeddings to compute a new representation for each input
- To do so, it first creates three vectors per input: *key*, *query*, and *value*

Self-attention

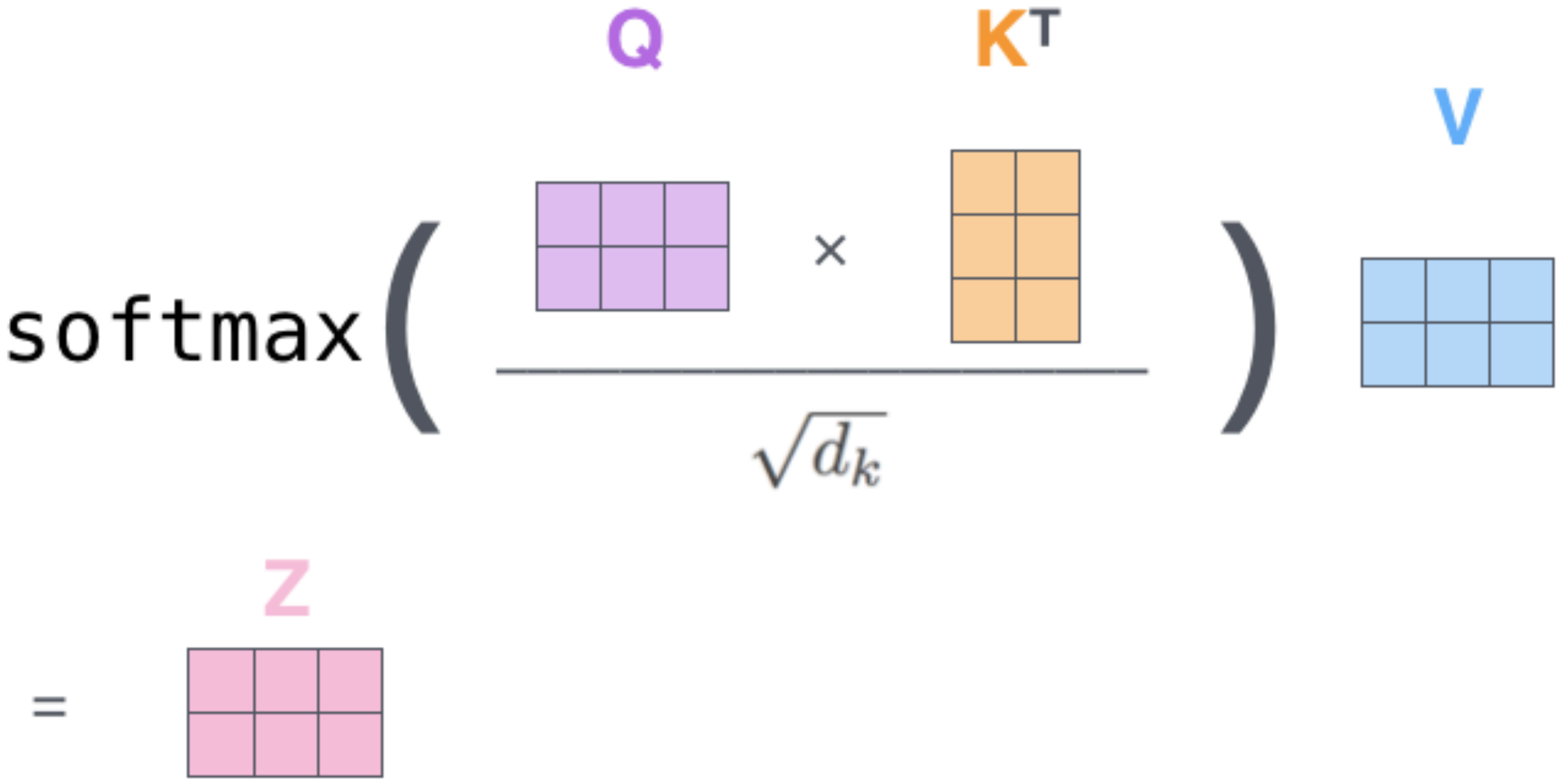
- Computing the *key*, *query*, and *value* for two inputs:



* Figure credit: Jay Allamar

Self-attention

- Use these to compare all inputs:

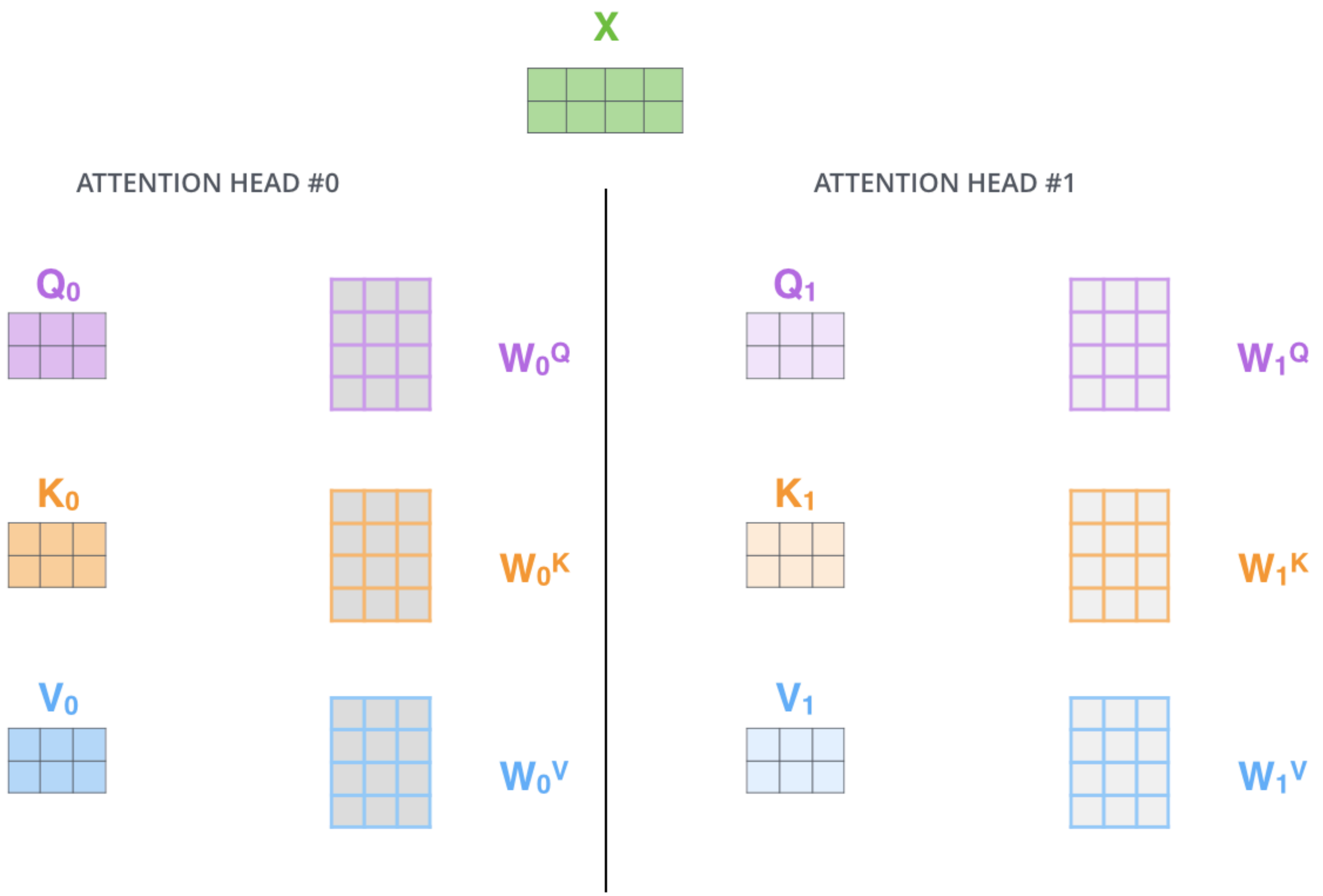


- The output of self-attention is the *expected value* under a probability distribution based on the key-value similarities (dot products)

* Figure credit: Jay Allamar

Multi-head self-attention

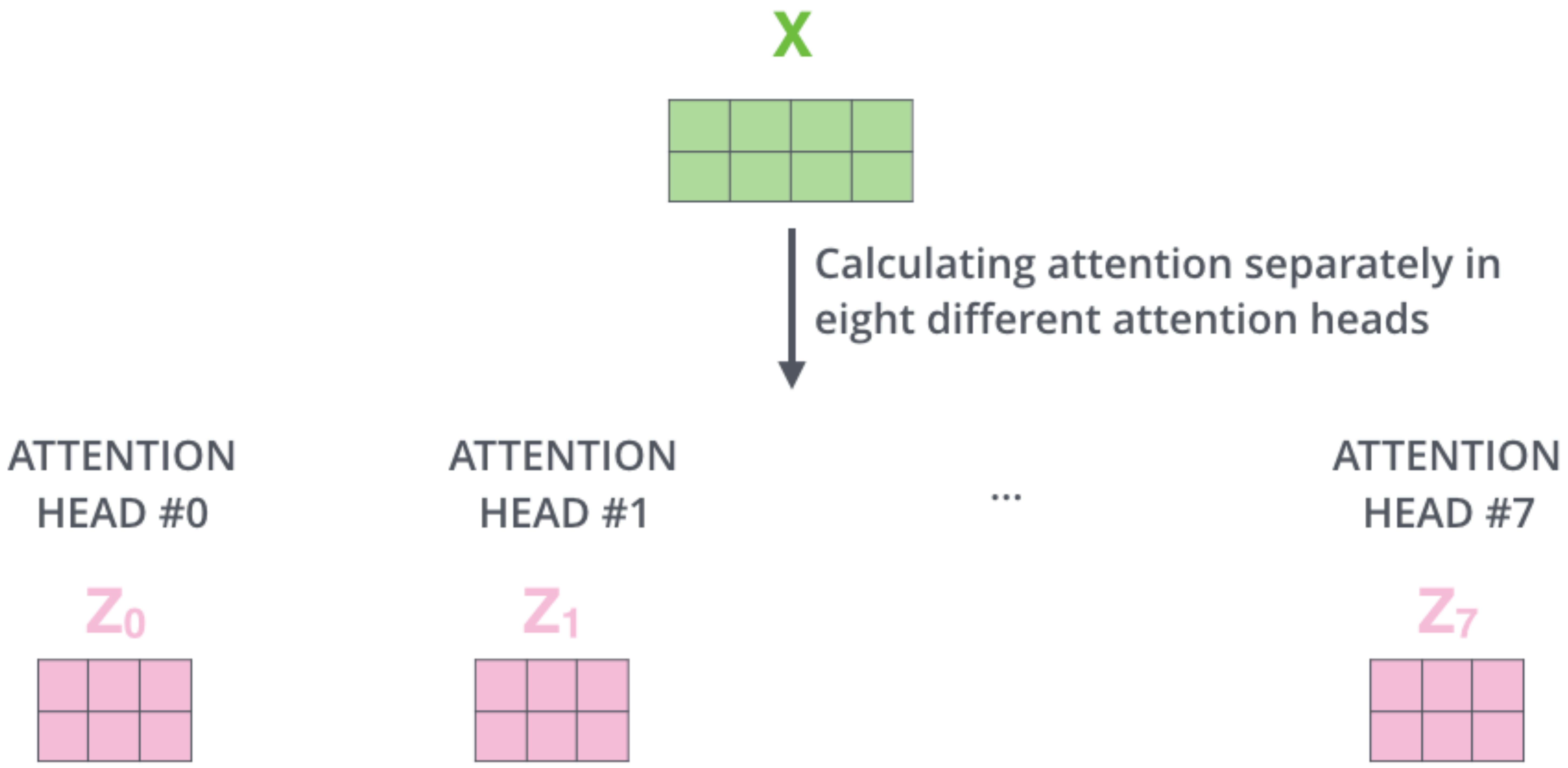
- Repeat process multiple times with different key-query-value triples:



* Figure credit: Jay Allamar

Multi-head self-attention

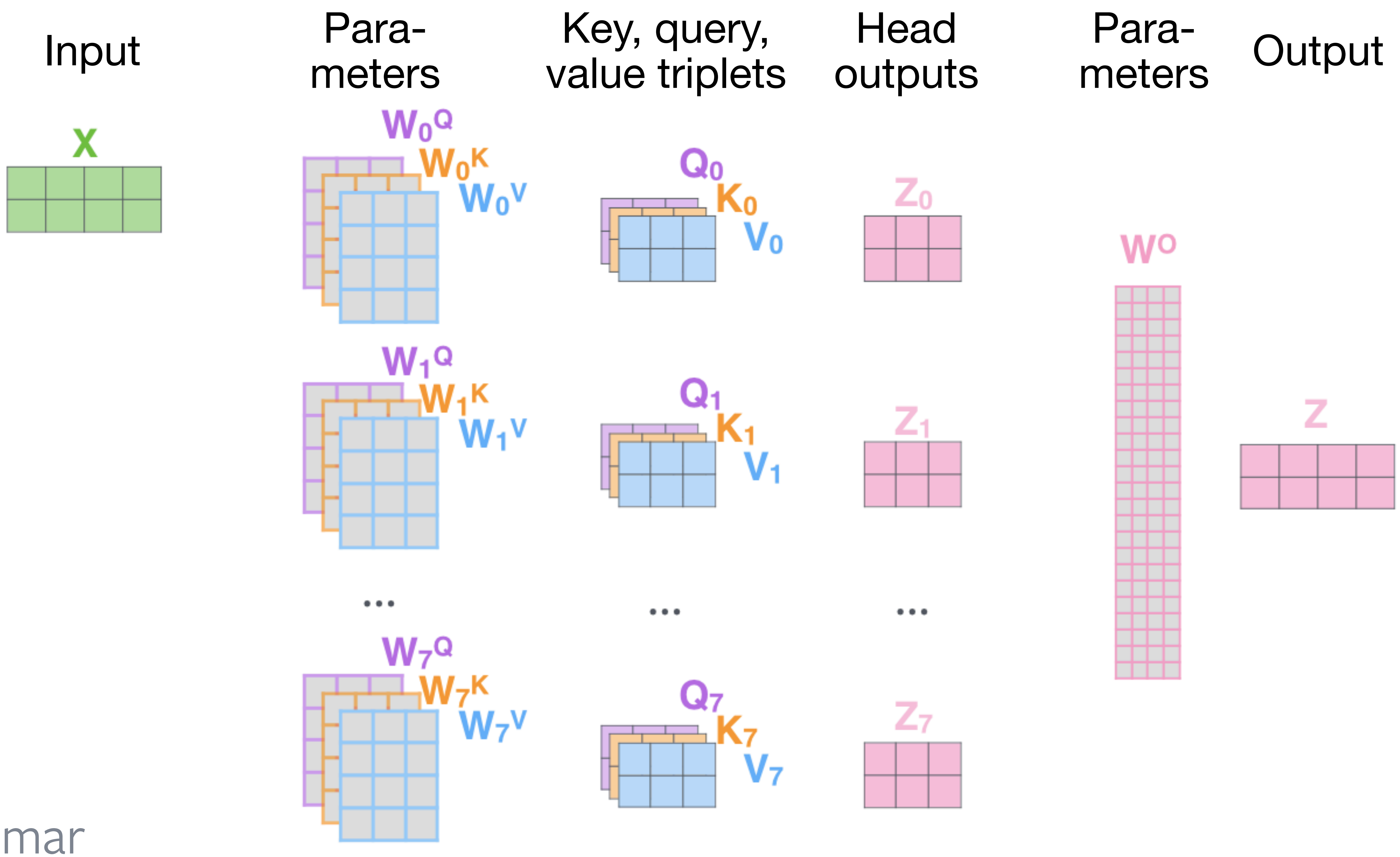
- Repeat process multiple times with different key-query-value triples, and concatenate the resulting outputs:



* Figure credit: Jay Allamar

Multi-head self-attention

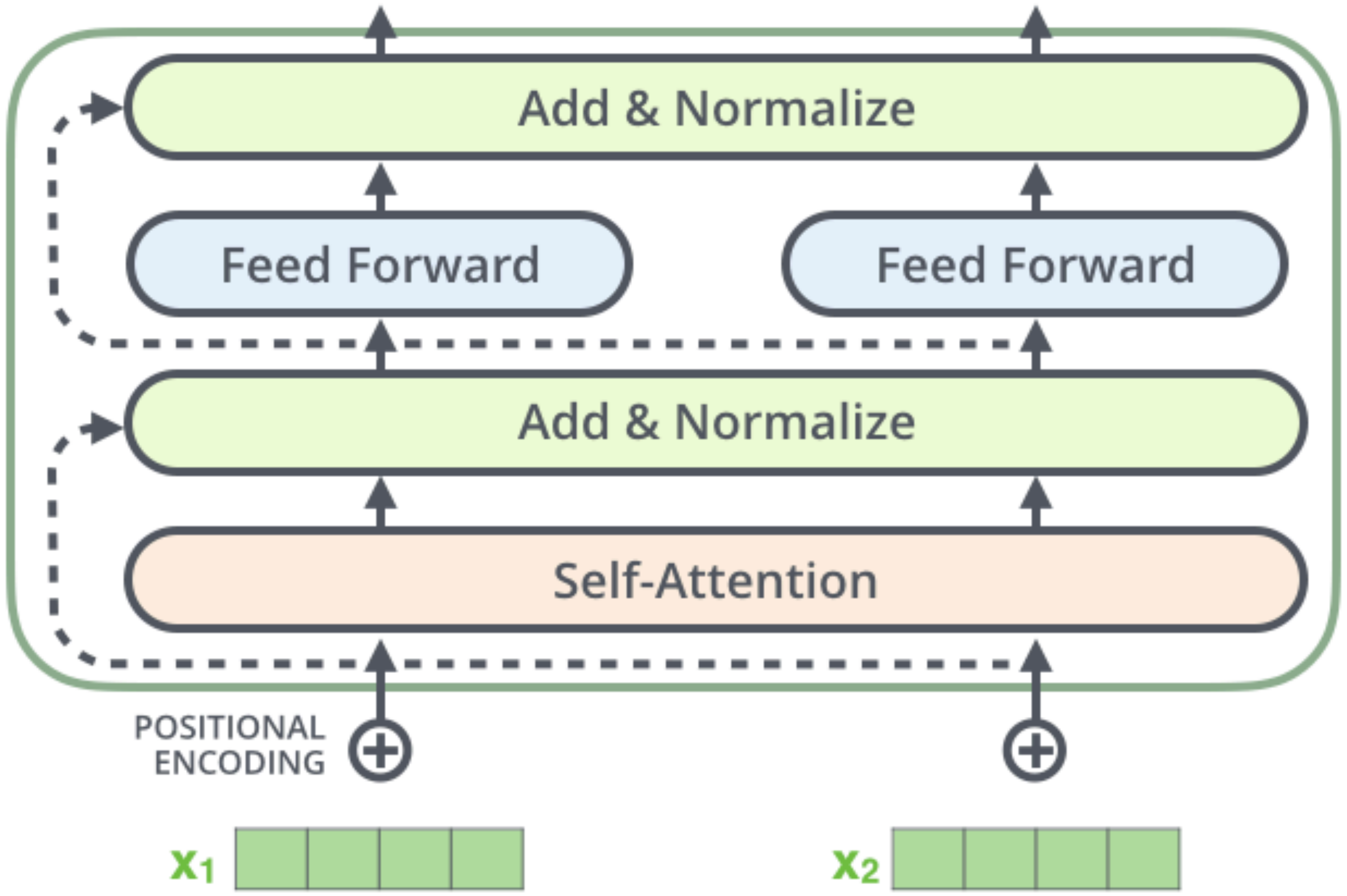
- Project down the result. Full overview of multi-head self-attention:



* Figure credit: Jay Allamar

Transformer block

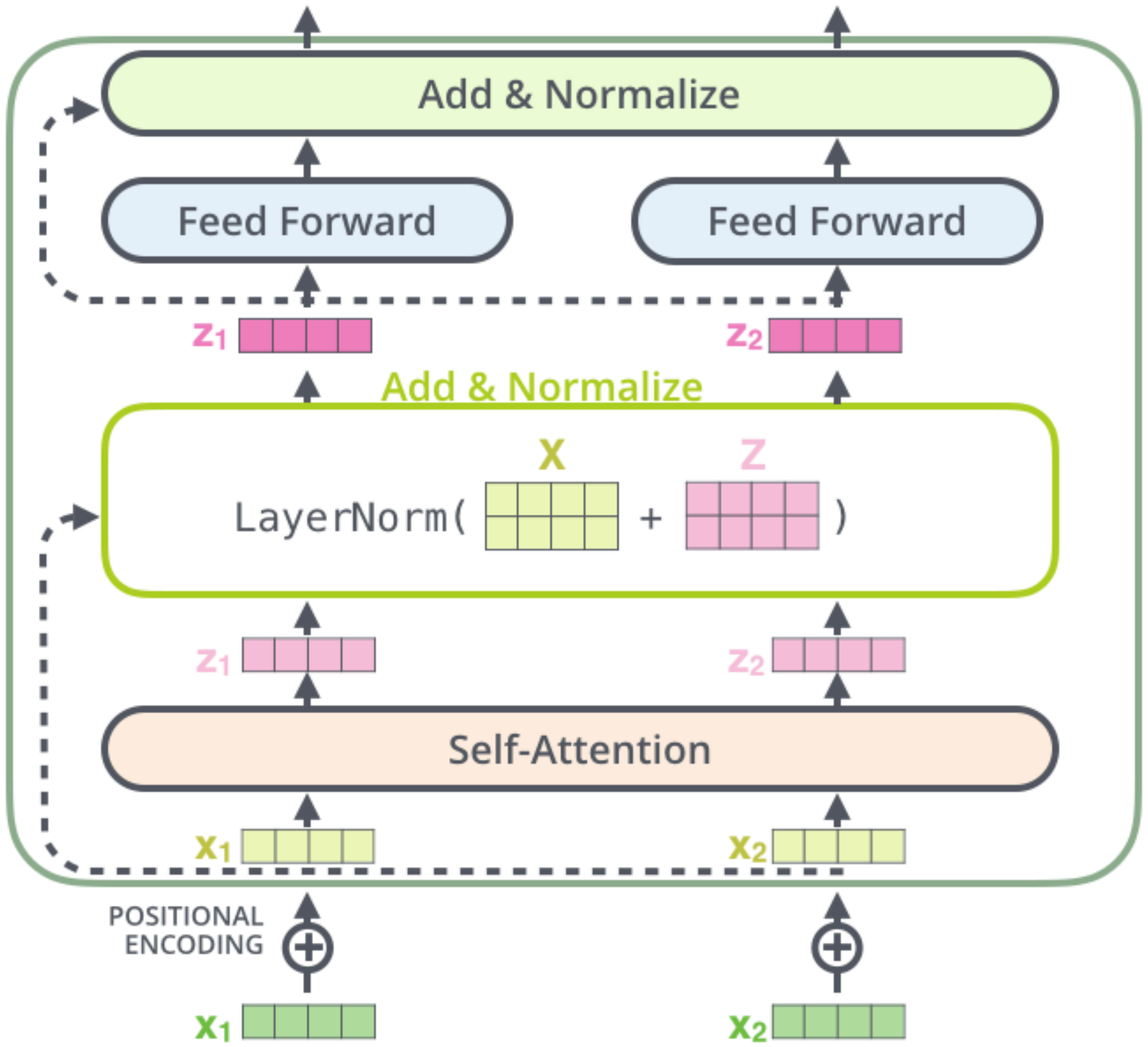
- Transformer encoder blocks combine self-attention and feedforward neural networks via residual connectivity:



* Figure credit: Jay Allamar

Transformer block

- Transformer encoder blocks combine self-attention and feedforward neural networks via residual connectivity:



* Figure credit: Jay Allamar

Classification with Transformers

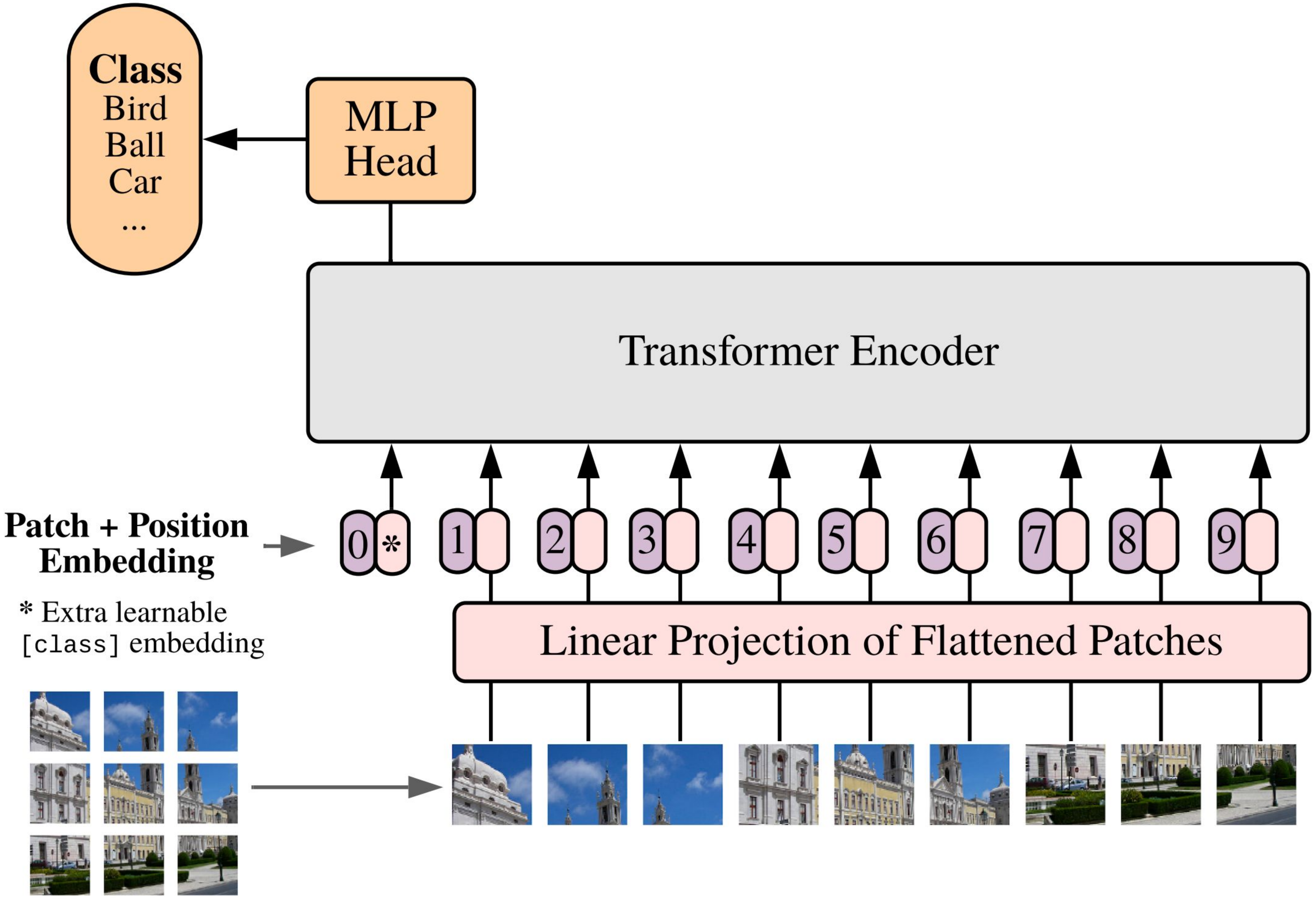
- Add additional `[class]` input to the set of inputs
- Attach a classification model to the corresponding `[class]` output
- Train the entire model to minimize the loss of the classification outputs

Vision transformer

- Flatten patches, and linearly project them
  - Note: This is strided convolution!

- Add position embeddings to encode spatial location

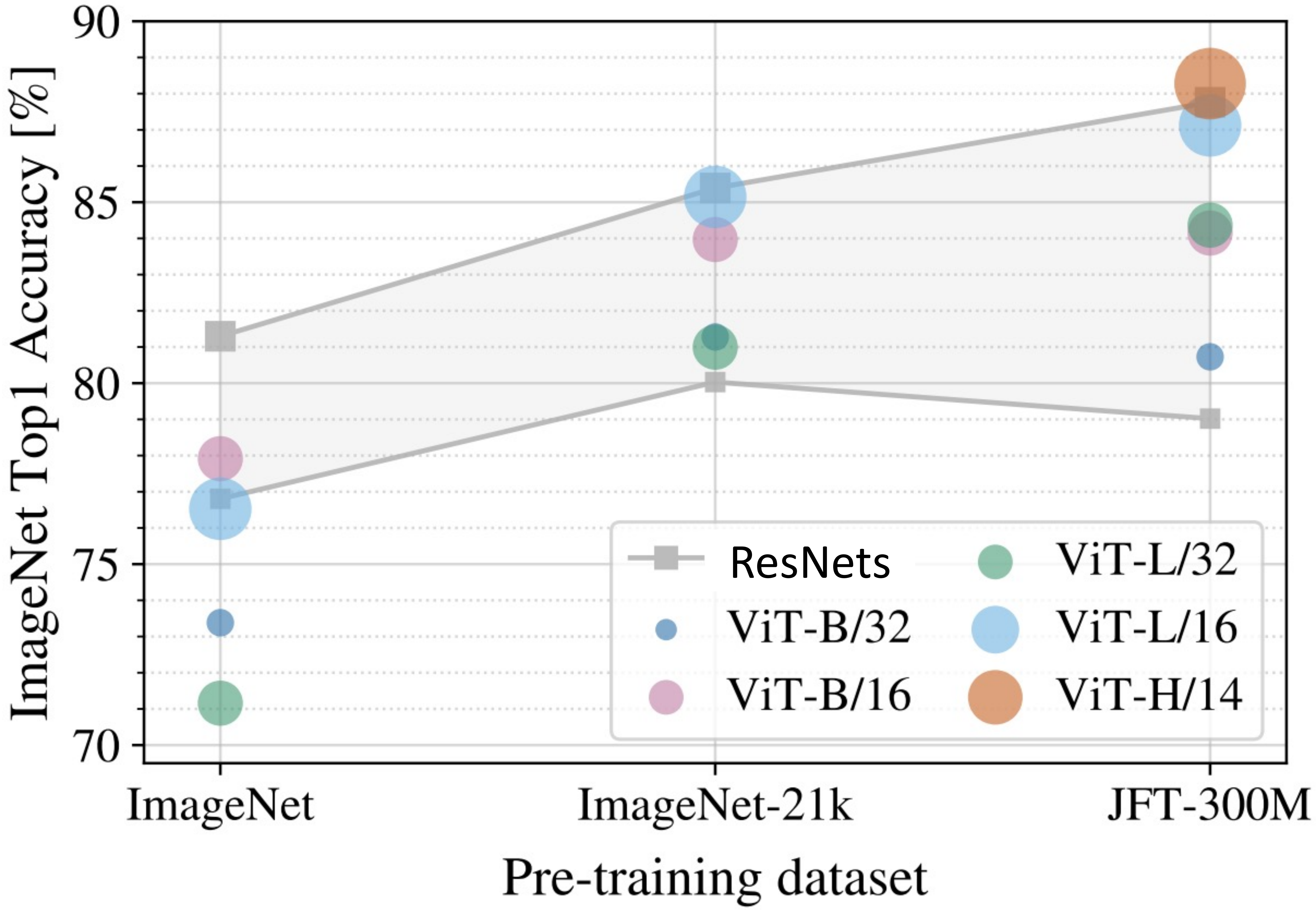
- Minimize multi-class logistic loss



* Figure credit: Lucas Beyer

ViTs versus ResNets

- ViTs outperform ResNets if trained on <u>very large</u> image datasets:



* Figure credit: Lucas Beyer

Summary

- Self-attention computes new representations for a set inputs by comparing all of those inputs
- Transformers combine multi-head self-attention layers and feed-forward models with residual connectivity
- Vision transformers are Transformers that treat images as a set of patches
- Vision transformers can outperform ResNets when trained on very large datasets
- Transformers allow one to combine different modalities very naturally

Reading material

- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin. **Attention is All You Need**. In *Advances in Neural Information Processing Systems* (*NeurIPS*), 2017.

- A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. In *International Conference on Learning Representations* (*ICLR*), 2021.

Questions?