Perceiver: General Perception with Iterative Attention

ICML 2021

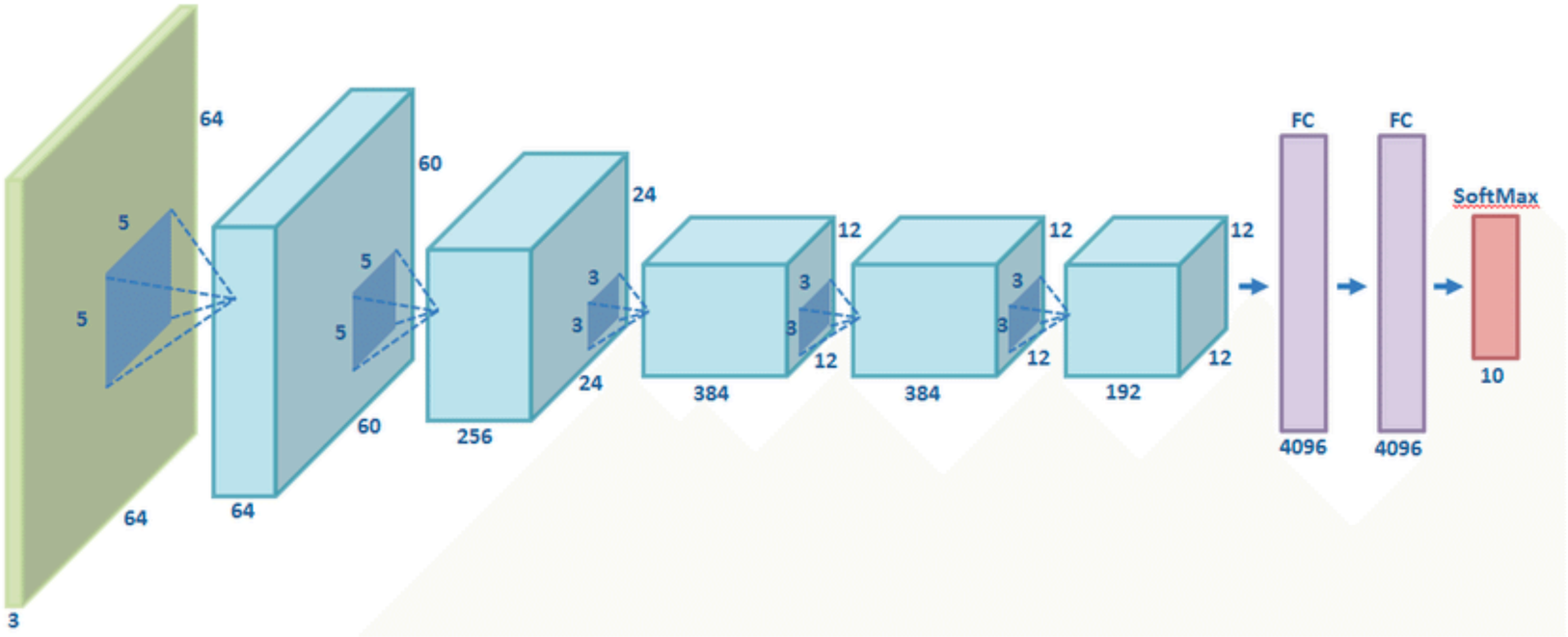
Andrew Jaegle, Felix Gimeno, Andrew Brock,
Andrew Zisserman, Oriol Vinyals, Joao Carreira

Motivation

- We want a general perception model that could process all kinds of modalities (e.g., speech, audio, video, images, etc.).

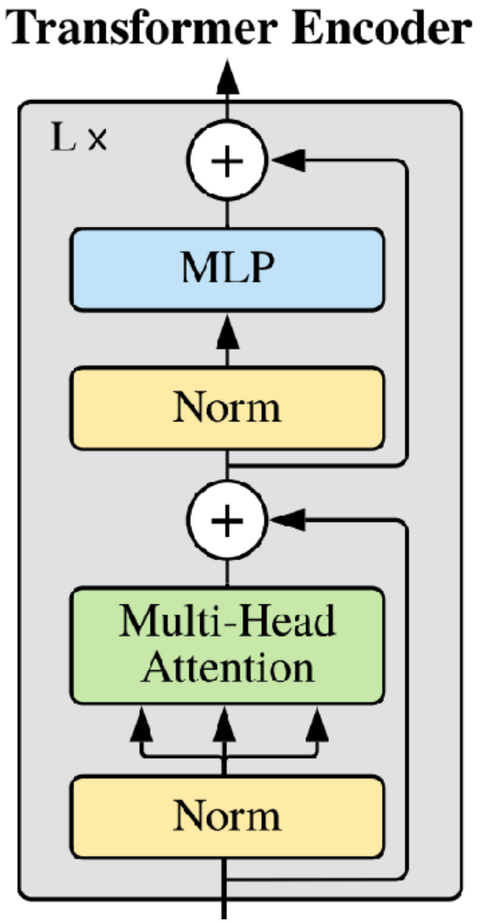
CNNs

- Specifically designed for processing image like data.
- Difficult to apply these models to other types of data.



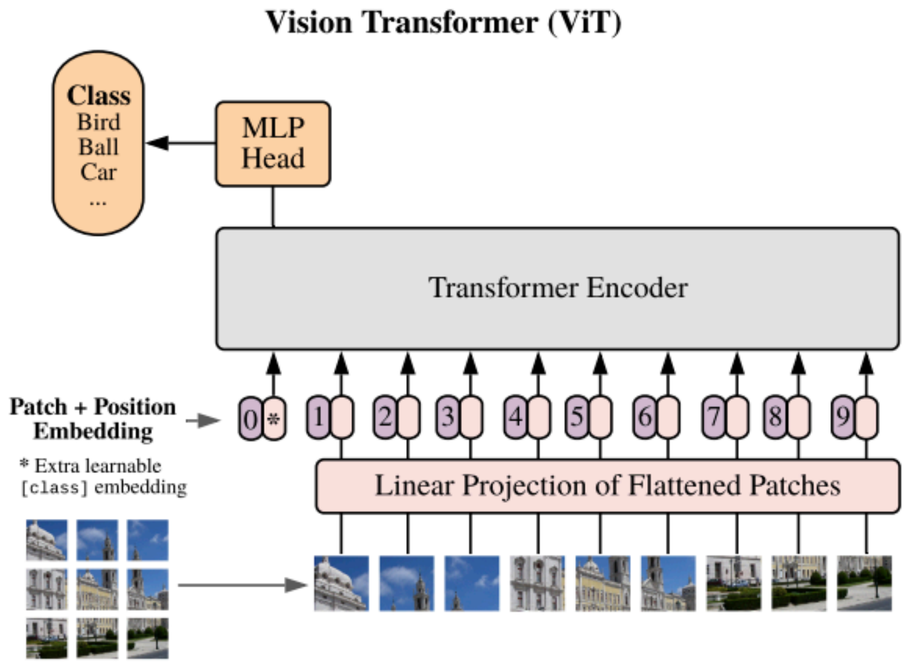
Self-Attention

- Transformers are flexible architectural blocks that make few assumptions about their inputs.



Vision Transformers

- Recent work has using Transformers on images still relies on the pixels' grid structure.



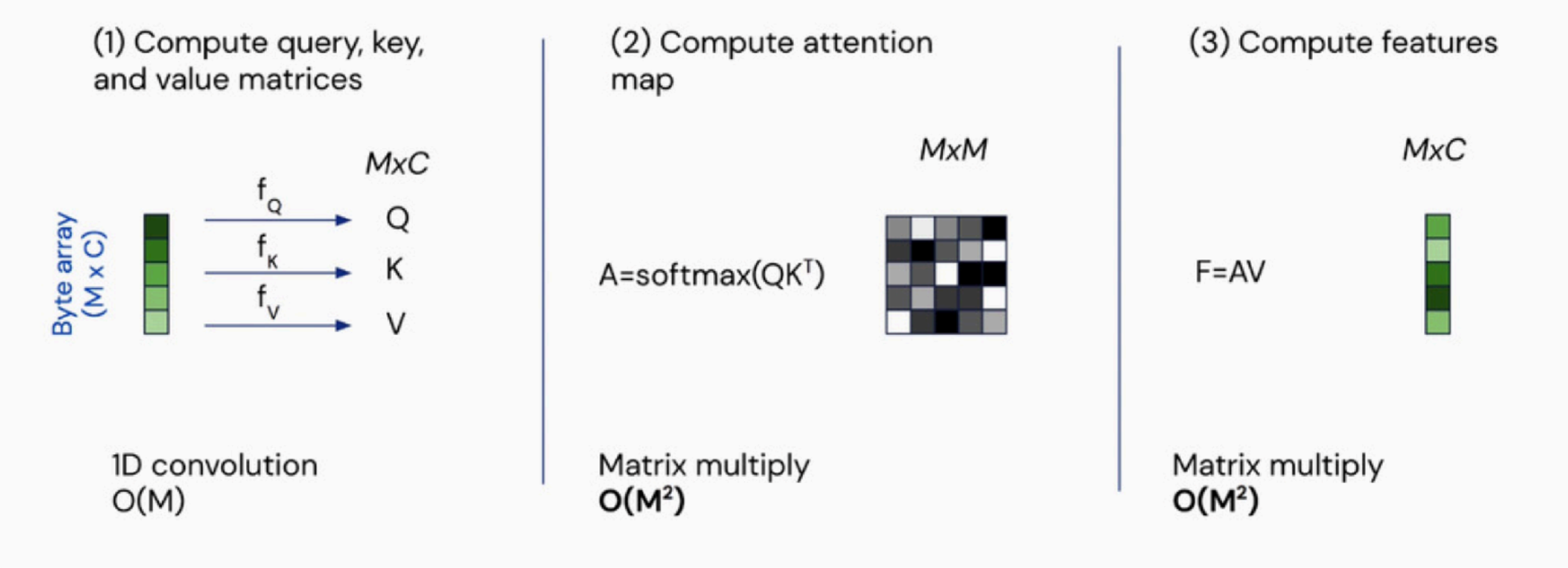
"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Dosovitskiy et al., ICLR 2021

General Perception

- We don’t want to make any assumptions about our input data.

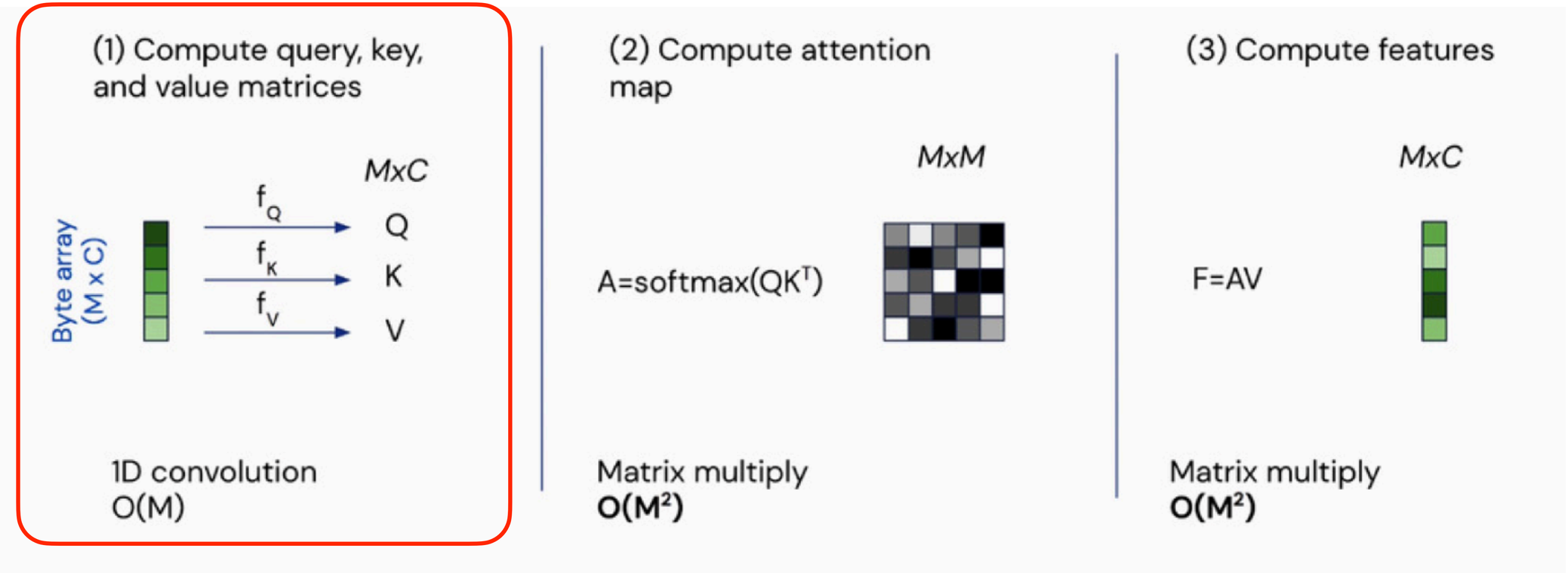
Self-Attention

- M is the number of pixels (e.g., ~50K for a 224x224 image).
- Quadratic cost prevents us from efficiently applying self-attention on images/ videos.



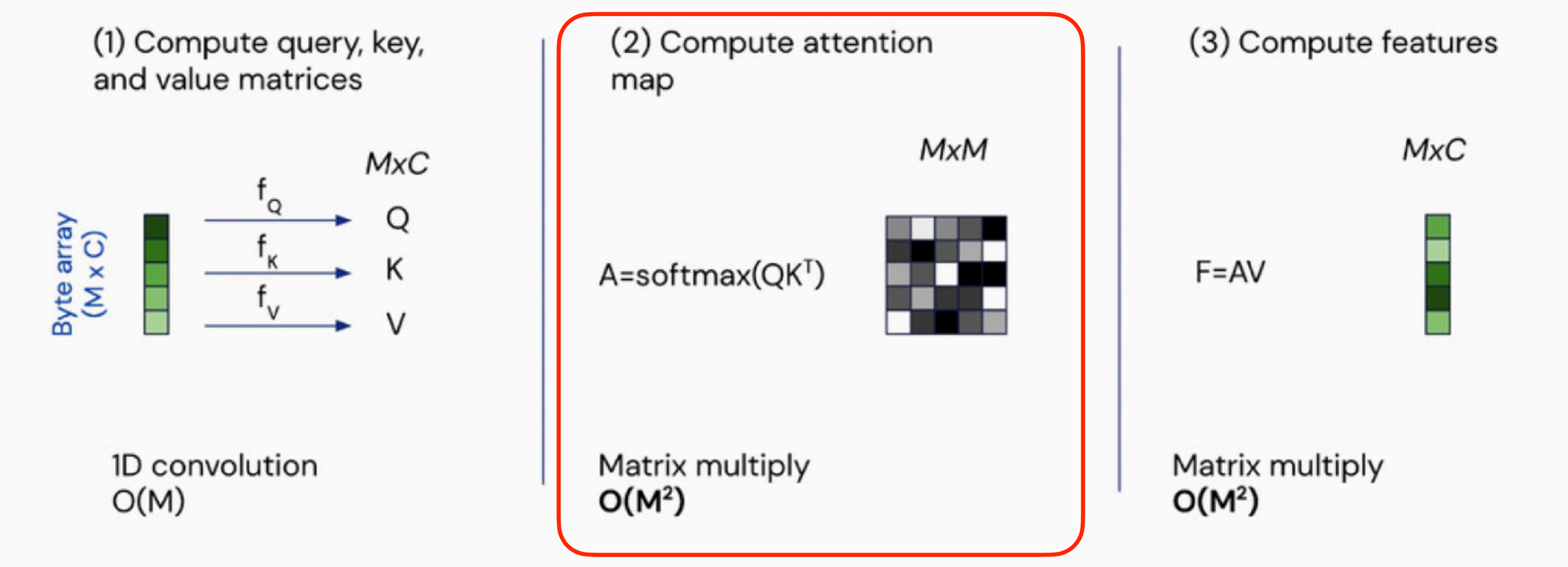
Self-Attention

- M is the number of pixels (e.g., ~50K for a 224x224 image).
- Quadratic cost prevents us from efficiently applying self-attention on images/ videos.



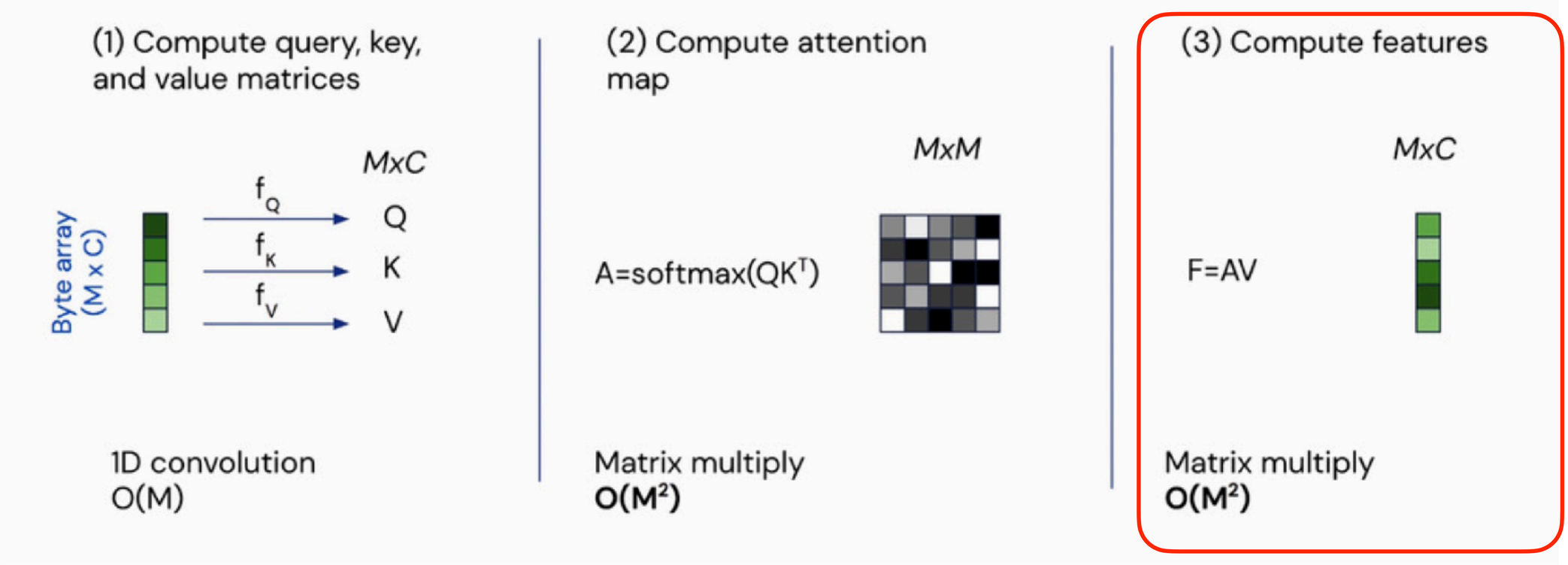
Self-Attention

- M is the number of pixels (e.g., ~50K for a 224x224 image).
- Quadratic cost prevents us from efficiently applying self-attention on images/ videos.



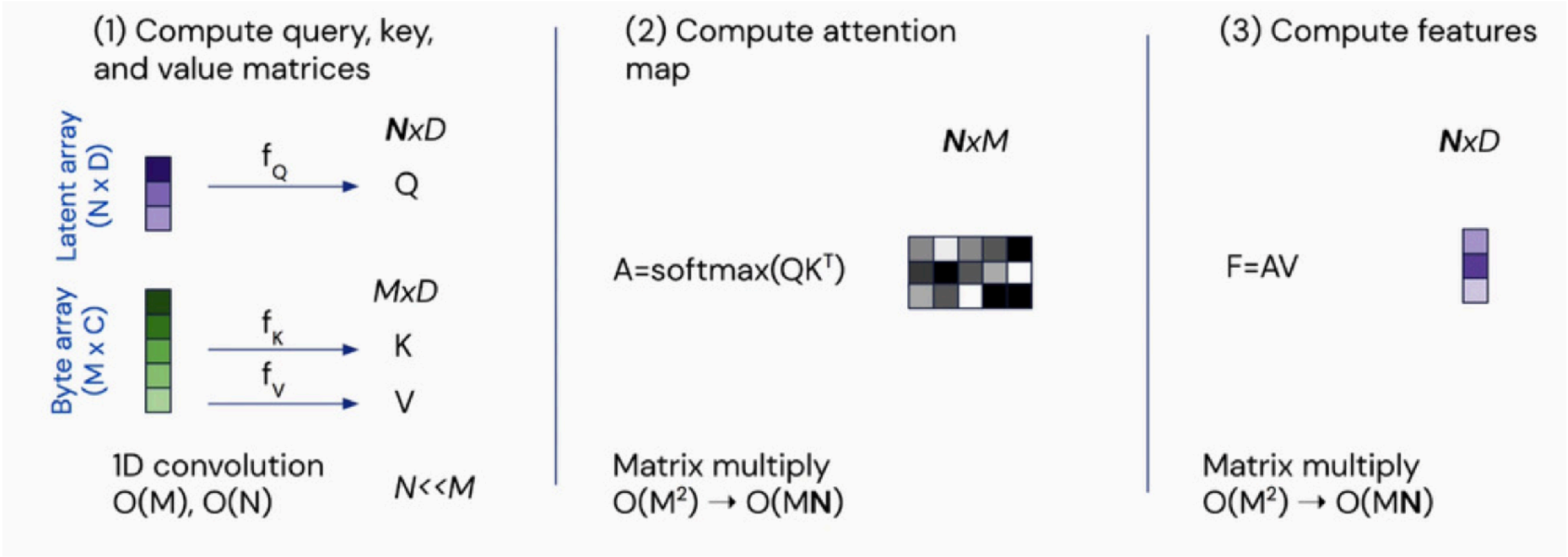
Self-Attention

- M is the number of pixels (e.g., ~50K for a 224x224 image).
- Quadratic cost prevents us from efficiently applying self-attention on images/ videos.



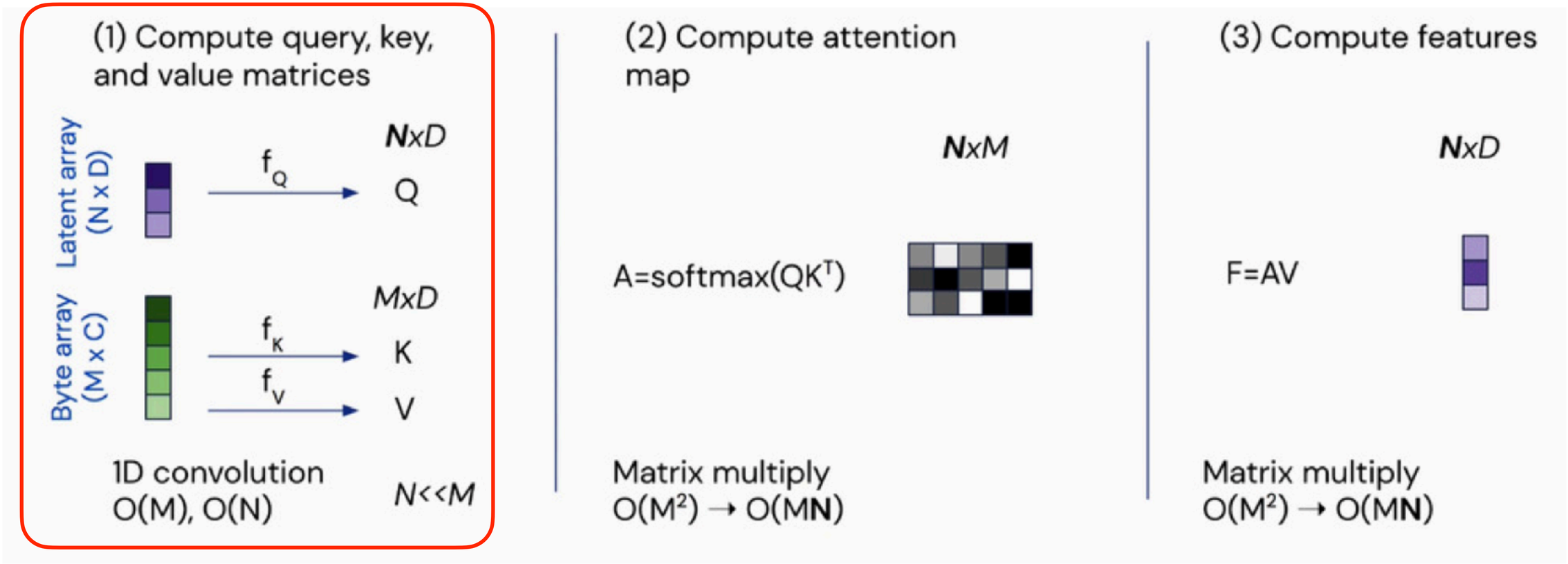
Latent Array

- The core idea is to introduce a small set of latent units that forms an attention bottleneck through which the inputs must pass.



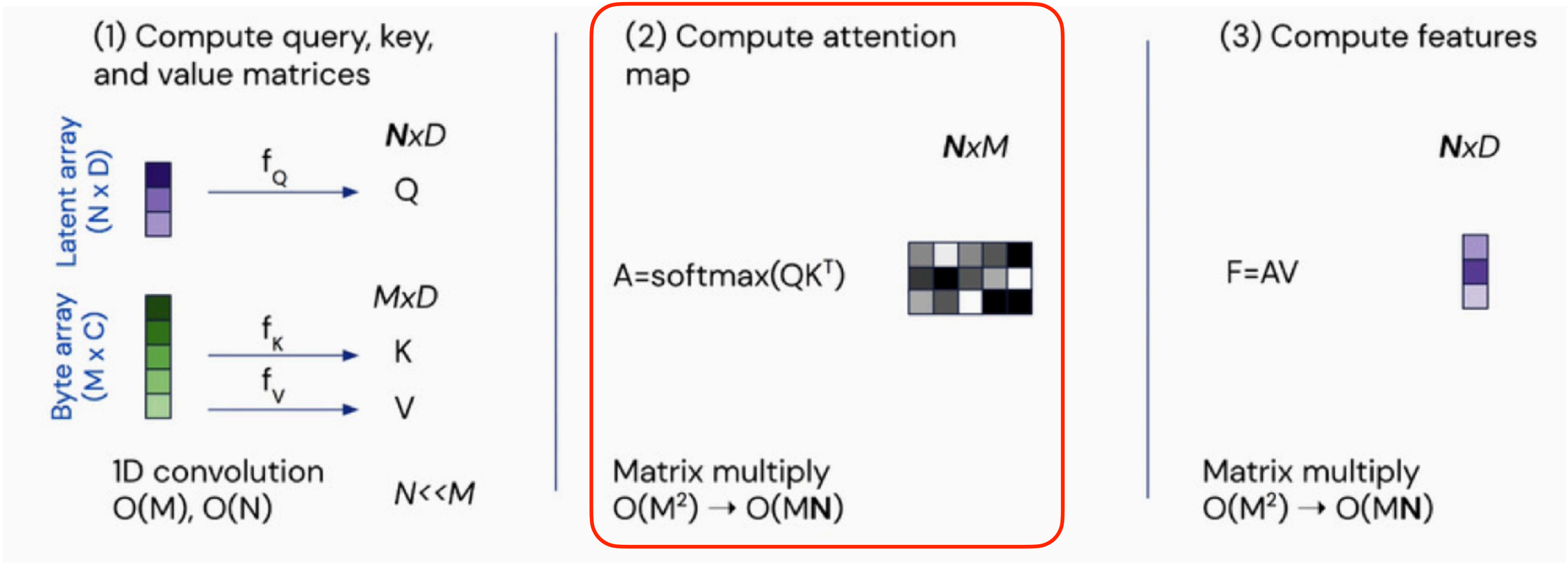
Latent Array

- The core idea is to introduce a small set of latent units that forms an attention bottleneck through which the inputs must pass.



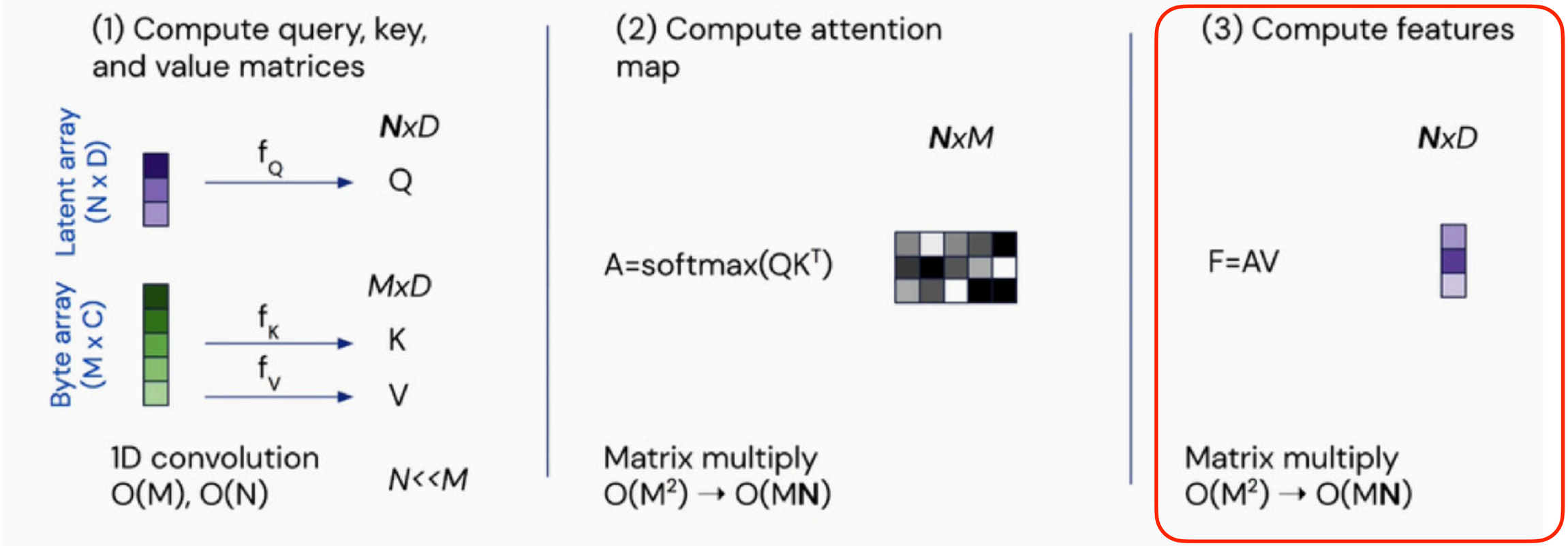
Latent Array

- The core idea is to introduce a small set of latent units that forms an attention bottleneck through which the inputs must pass.

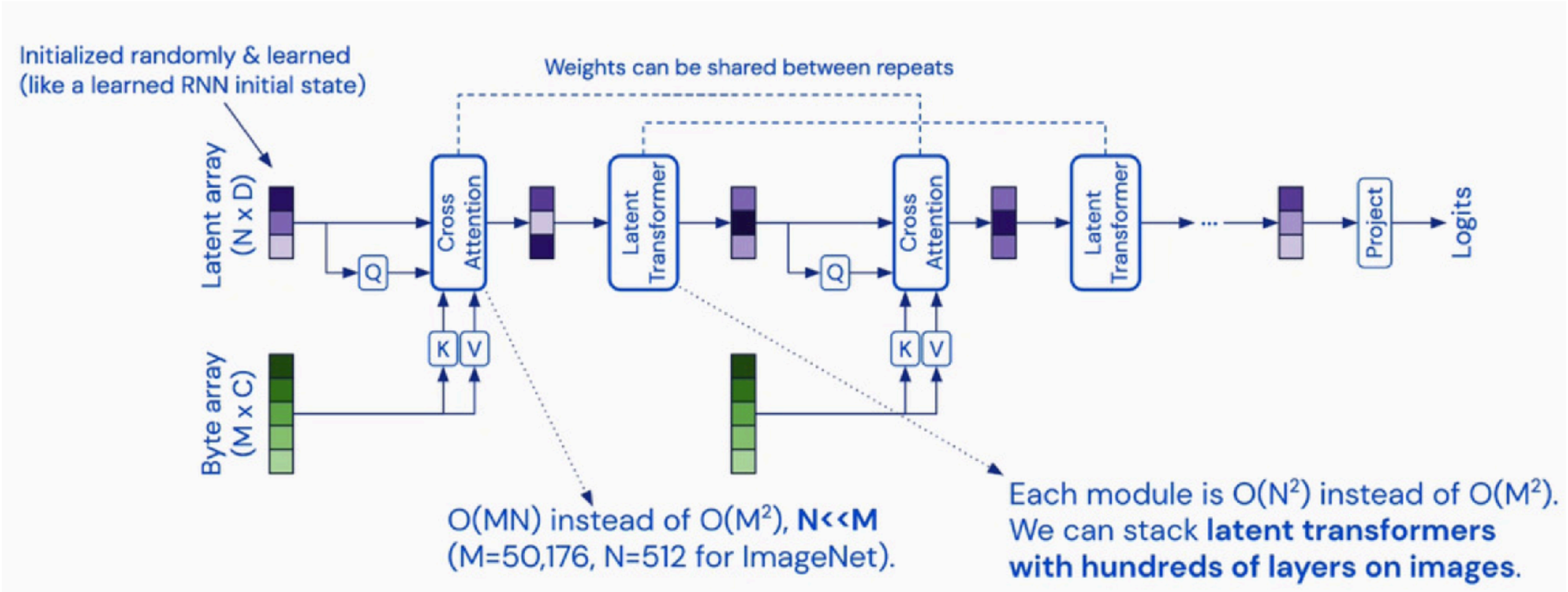


Latent Array

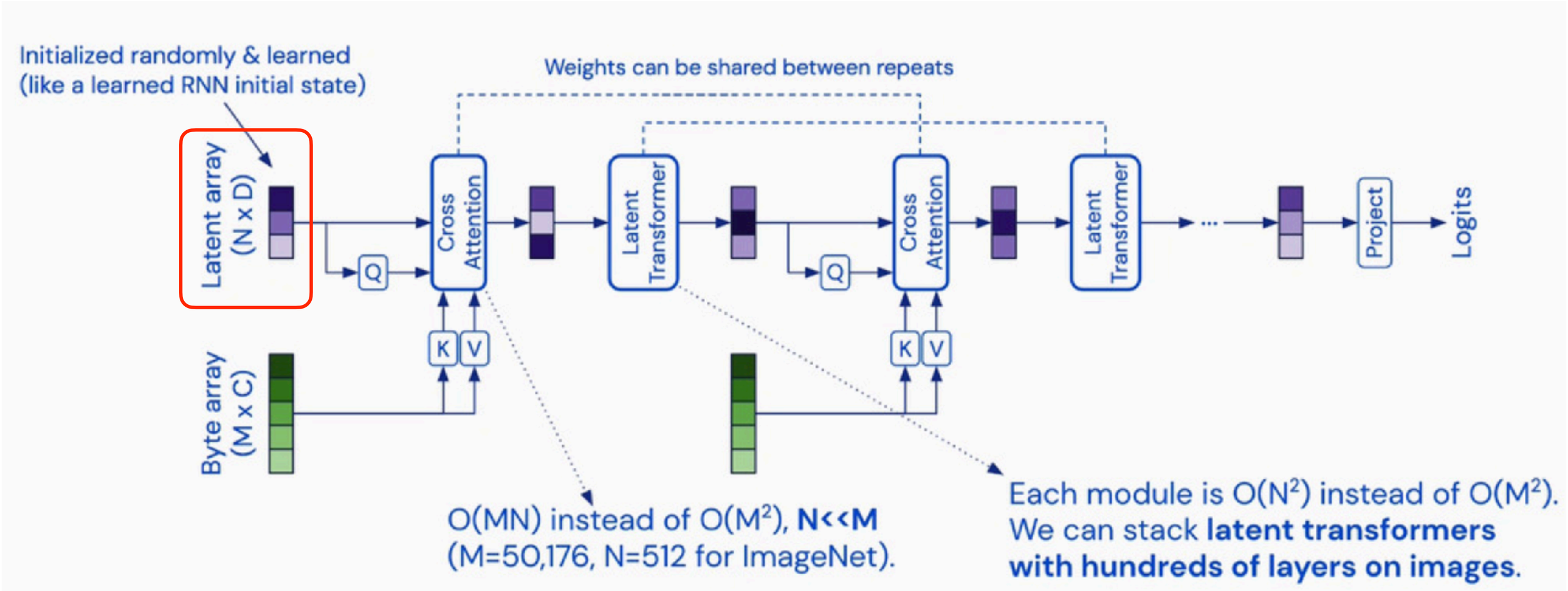
- The core idea is to introduce a small set of latent units that forms an attention bottleneck through which the inputs must pass.



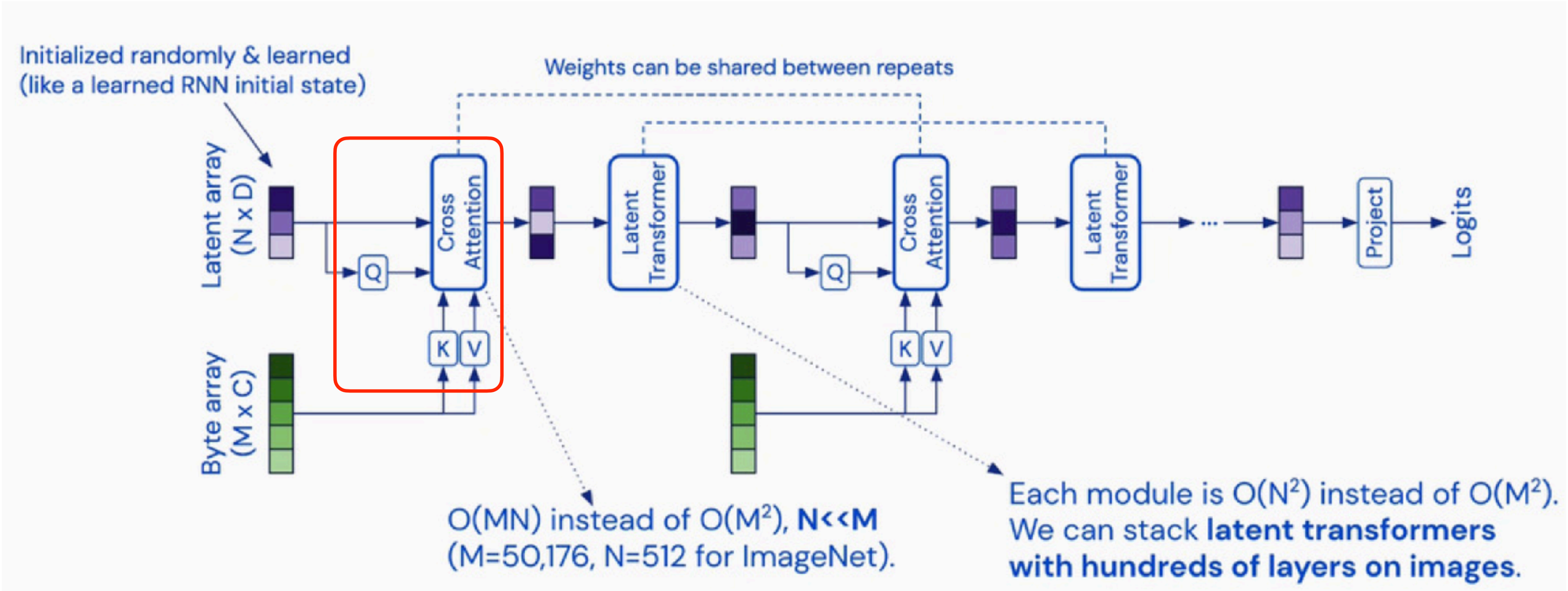
The Perceiver



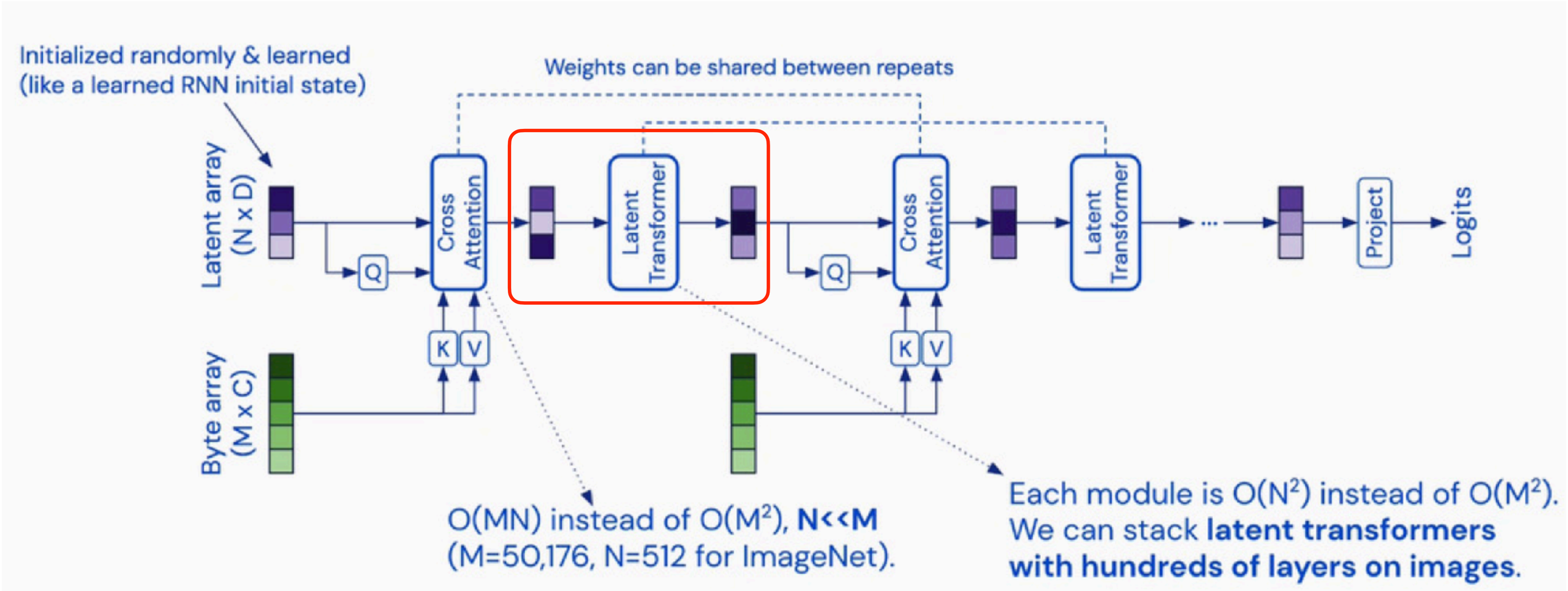
The Perceiver



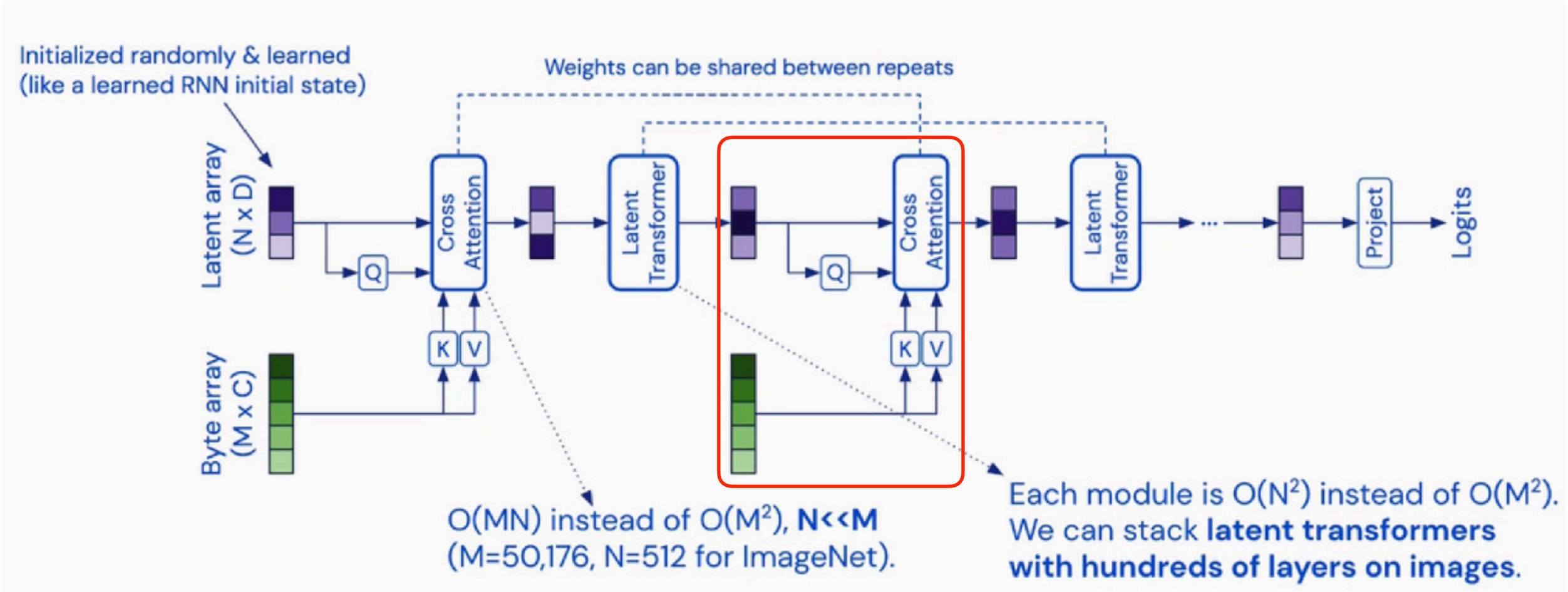
The Perceiver



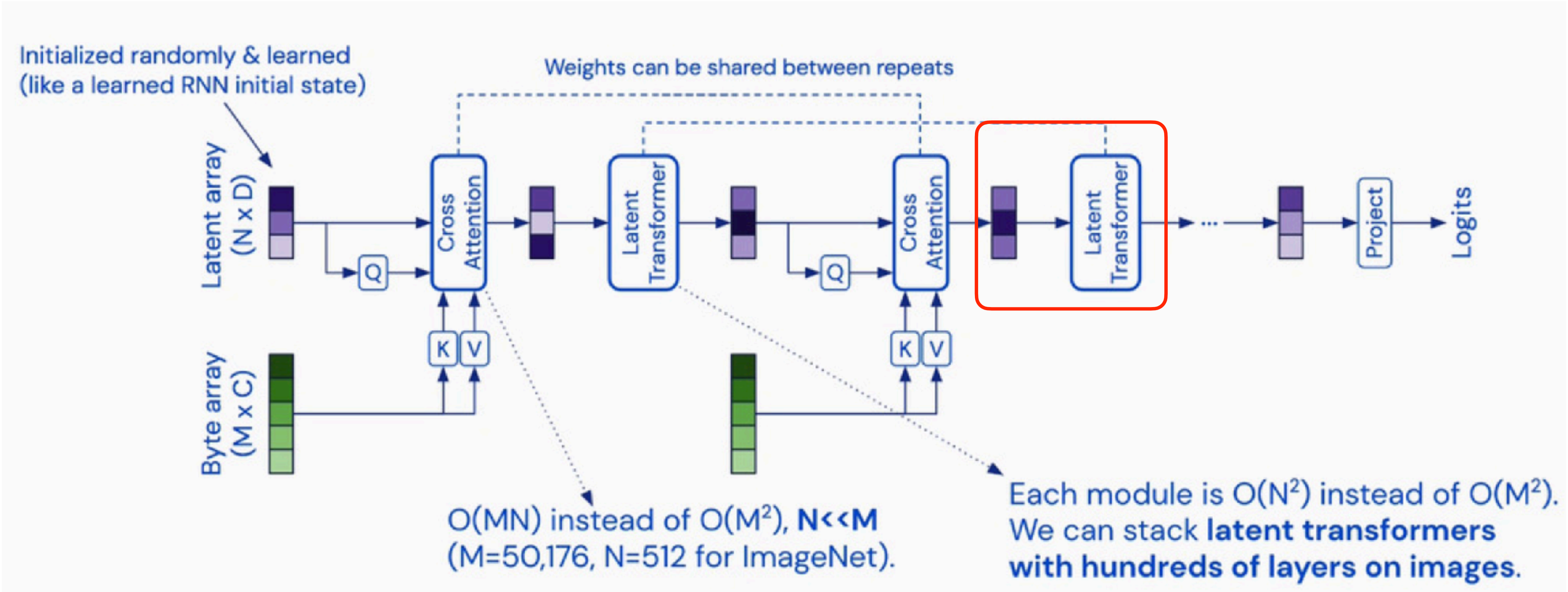
The Perceiver



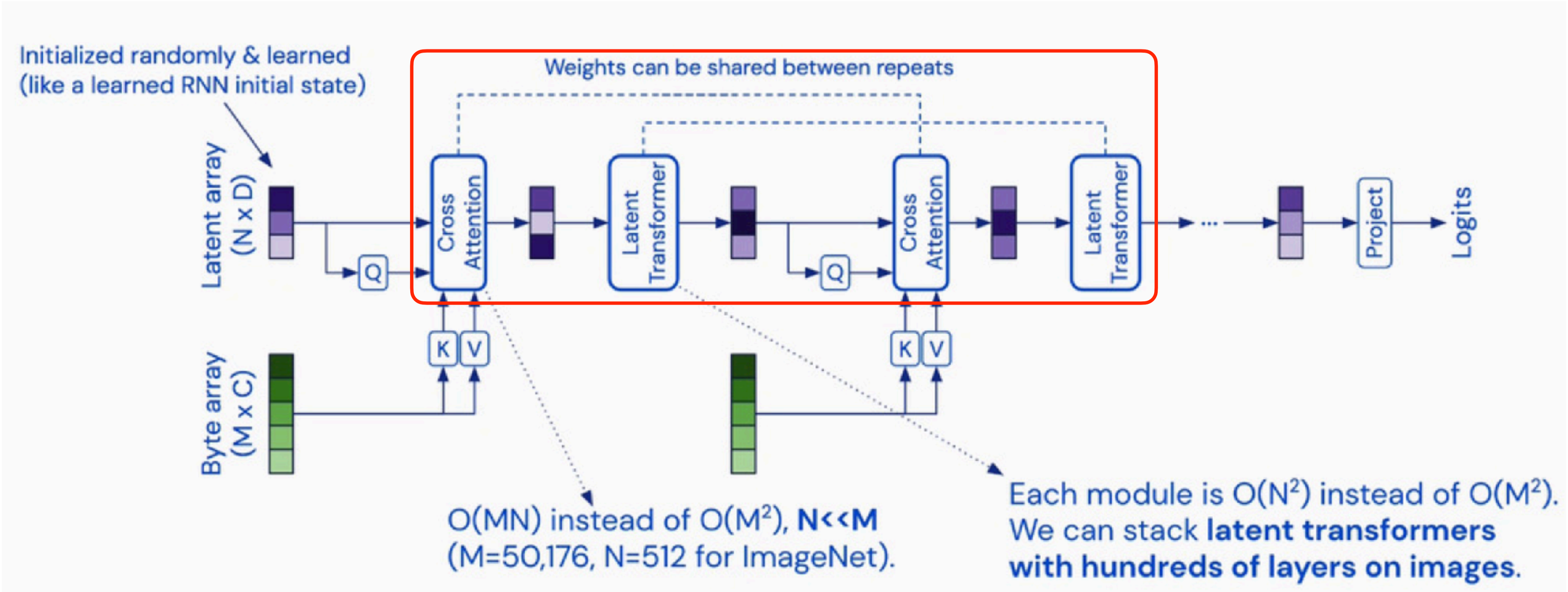
The Perceiver



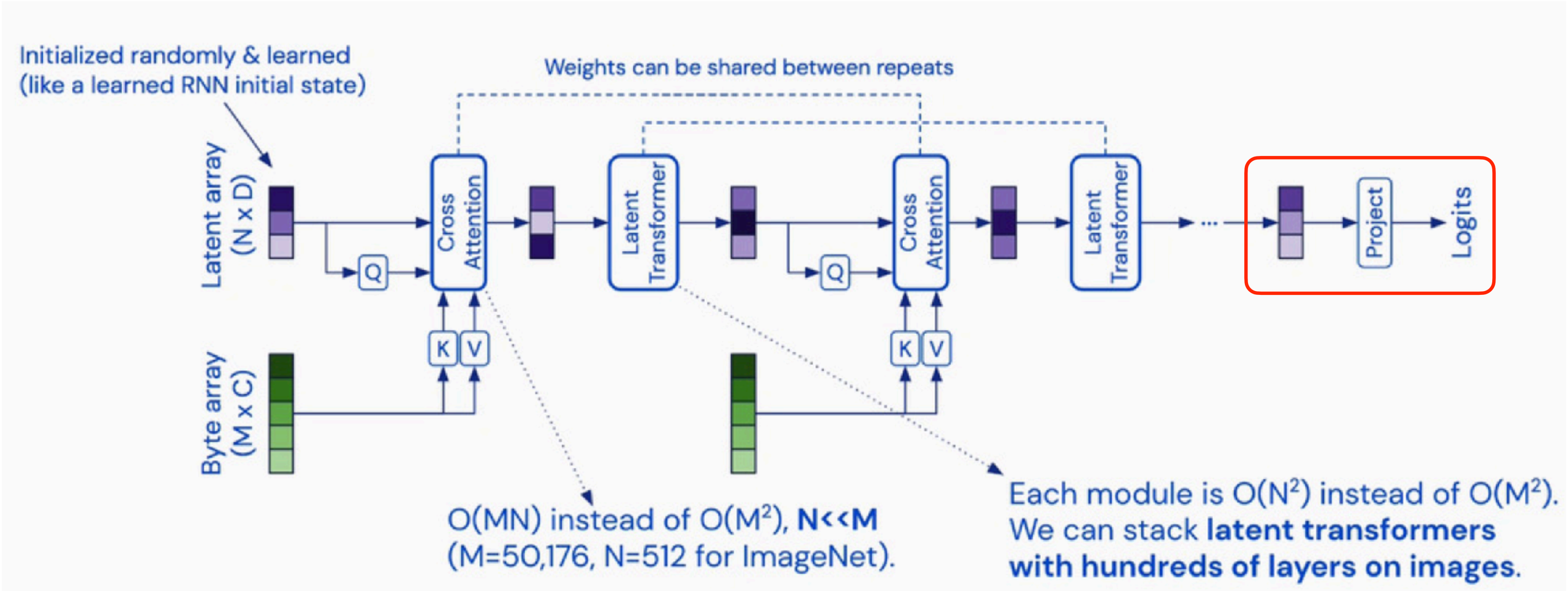
The Perceiver



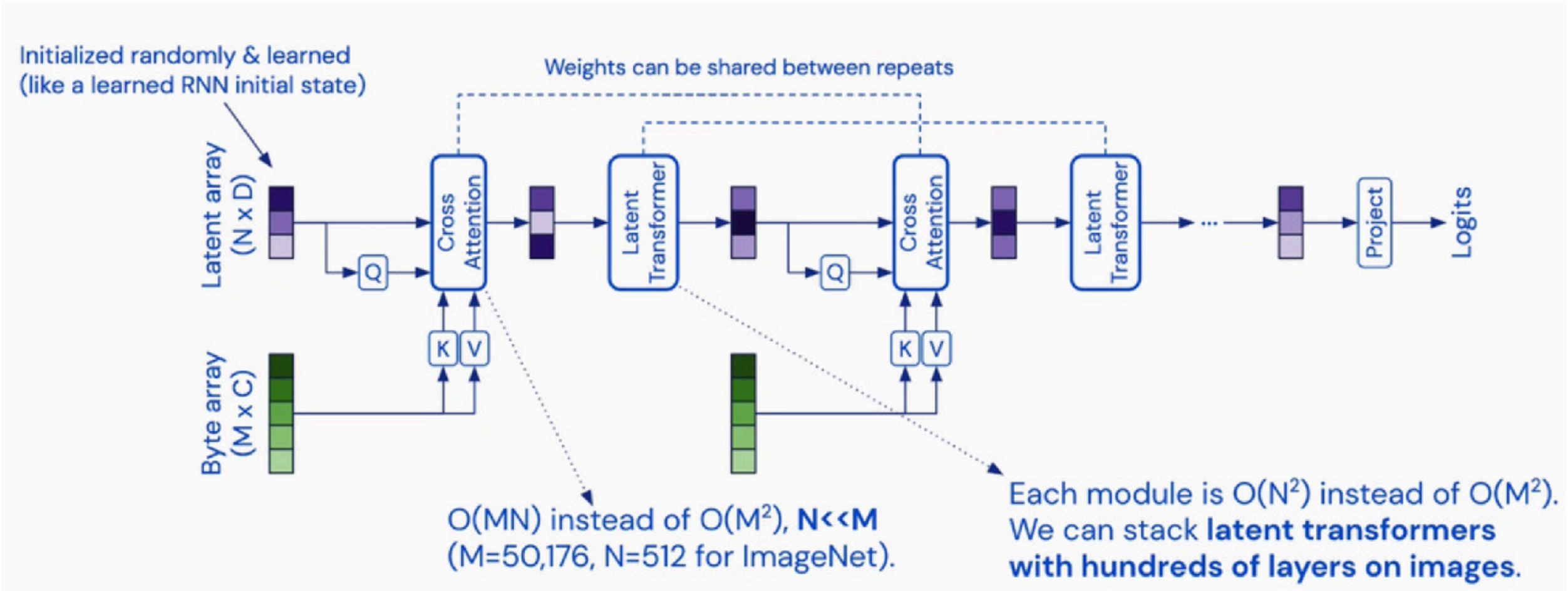
The Perceiver



The Perceiver



The Perceiver



Minimal assumptions about spatial structure (no patches/grids).

Imagenet Classification

- Models that use 2D convolutions (red).
- Models that only use global attention (blue).
- The performance is evaluated using standard top-1 accuracy.

ResNet-50 (He et al., 2016)	77.6
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
ResNet-50 (FF)	73.5
ViT-B-16 (FF)	76.7
Transformer (64x64, FF)	57.0
Perceiver (FF)	78.0

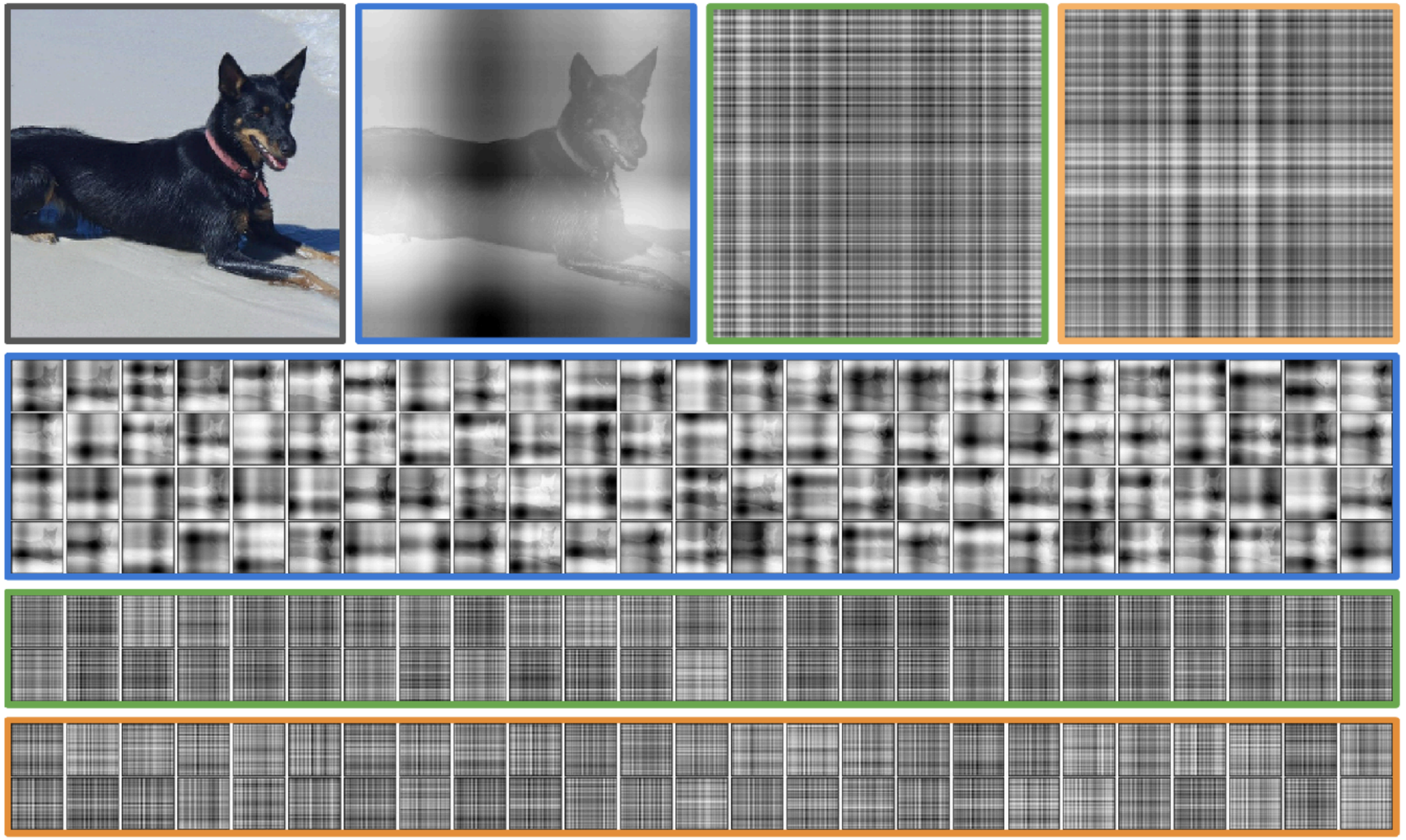
Imagenet Classification

- Models that use 2D convolutions (red).
- Models that only use global attention (blue).
- The performance is evaluated using standard top-1 accuracy.

ResNet-50 (He et al., 2016)	77.6
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
ResNet-50 (FF)	73.5
ViT-B-16 (FF)	76.7
Transformer (64x64, FF)	57.0
Perceiver (FF)	78.0

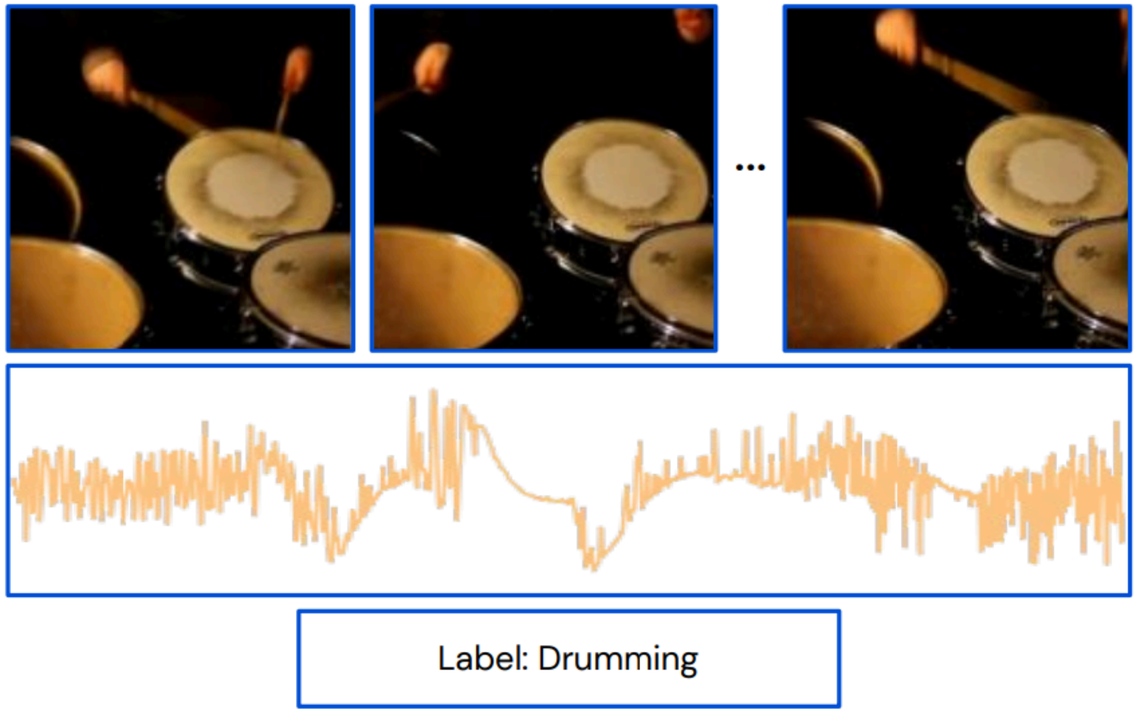
The Perceiver is competitive with standard baselines without relying on domain-specific architectural assumptions

Imagenet Classification



Audio Event Classification

- The authors experiment with audio event classification in video using the AudioSet dataset.
- AudioSet consists of ~1.7M 10s long training videos and 527 classes.



Audio Event Classification

- The authors experiment with audio event classification in video using the AudioSet dataset.
- AudioSet consists of ~1.7M 10s long training videos and 527 classes.

Model / Inputs	Audio	Video	A+V
Benchmark (Gemmeke et al., 2017)	31.4	-	-
Attention (Kong et al., 2018)	32.7	-	-
Multi-level Attention (Yu et al., 2018)	36.0	-	-
ResNet-50 (Ford et al., 2019)	38.0	-	-
CNN-14 (Kong et al., 2020)	43.1	-	-
CNN-14 (no balancing & no mixup) (Kong et al., 2020)	37.5	-	-
G-blend (Wang et al., 2020c)	32.4	18.8	41.8
Attention AV-fusion (Fayek & Kumar, 2020)	38.4	25.7	46.2
Perceiver (raw audio)	38.3	25.8	43.5
Perceiver (mel spectrogram)	38.4	25.8	43.2

3D Object Recognition

- The task is to predict the class of each object, given the coordinates of ~ 2000 points in 3D space.
- ModelNet dataset consists of 9,843 training examples spanning 40 object categories.

3D Object Recognition

- Models that use geometric features (red).
- 3D agnostic approaches (blue).
- The performance is evaluated using standard top-1 accuracy.

	Accuracy
PointNet++ (Qi et al., 2017)	**91.9**
ResNet-50 (FF)	66.3
ViT-B-2 (FF)	78.9
ViT-B-4 (FF)	73.4
ViT-B-8 (FF)	65.3
ViT-B-16 (FF)	59.6
Transformer (44x44)	82.1
Perceiver	**85.7**

Optical Flow Prediction

- The method is evaluated on Sintel and KITTI datasets.
- The performance is evaluated using average end-point error (EPE) (lower is better).

Optical Flow Prediction

- The Perceiver model outperforms highly specialized optical flow architectures on several popular optical flow benchmarks.

Network	Sintel.clean	Sintel.final	KITTI
PWCNet [75]	2.17	2.91	5.76
RAFT [84]	1.95	2.57	**4.23**
Perceiver IO	**1.81**	**2.42**	4.98

Prediction Ground Truth

Contributions

- A new interesting approach that provides a general architecture for processing many different modalities/tasks.
- Proposes how to scale the approach to large input sizes by introducing the latent array and cross-attention concepts.
- Competitive results on a variety of very different benchmarks.

Discussion Questions

- What’s the benefit of having a single architecture that can process multiple modalities simultaneously?

Discussion Questions

- What’s the benefit of having a single architecture that can process multiple modalities simultaneously?

- Is the proposed approach scalable?

Discussion Questions

- What's the benefit of having a single architecture that can process multiple modalities simultaneously?
- Is the proposed approach scalable?
- Multi-task multi-modal learning?