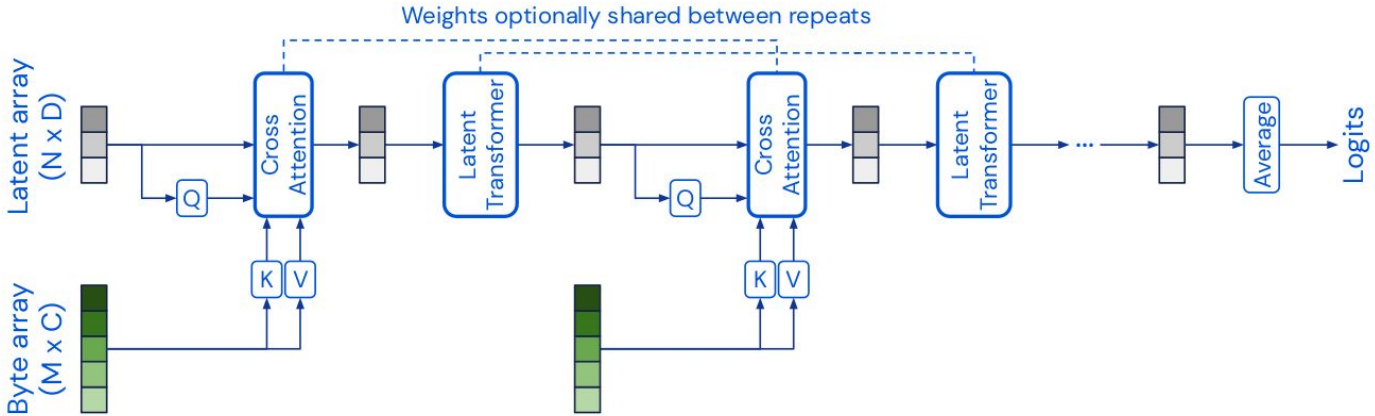
Perceiver

General Perception with Iterative Attention

Why?

- No assumptions on input type.
  - Can handle Images, videos, point clouds, language etc.
  - No inductive bias.
- Efficiency.
  - Attention on images is too expensive.
    - Assuming 224x224 images (50176 pixels)→ Attention complexity O(50176x50176)

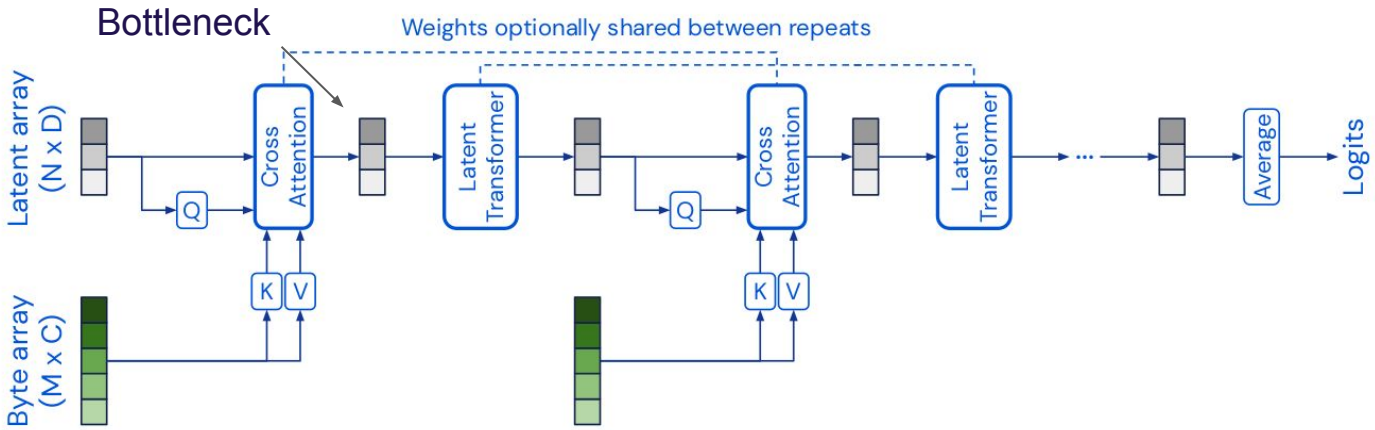
Architecture



Standard Attention: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$ with Q,K MxD → QK^T MxM

Architecture

If you share the weights (more efficient) its basically an RNN



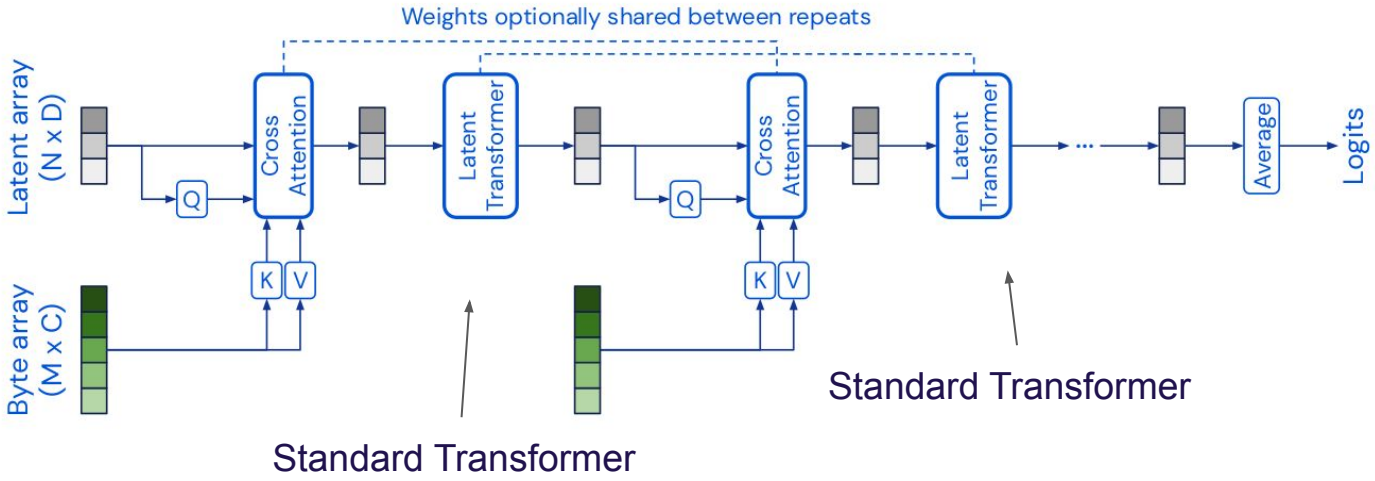
Cross Attention: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$

where Q deviates and comes from a latent array NxD where N<<M.

QK^T → MxN

Architecture

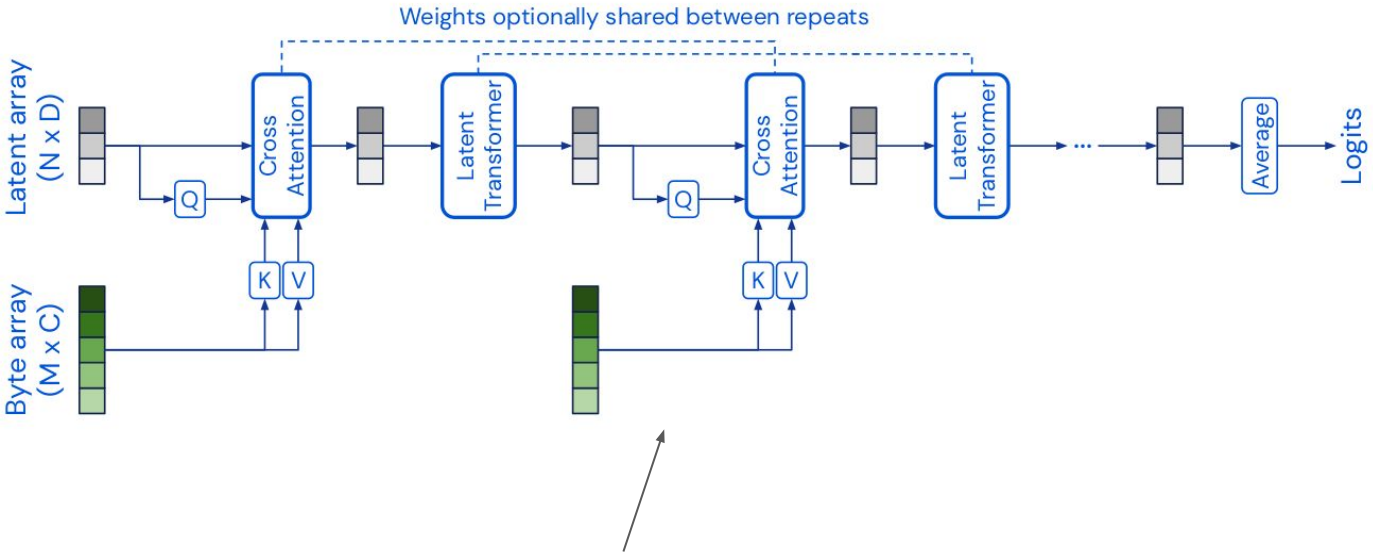
If you share the weights (more efficient) its basically an RNN



Transformer block is cheap now → We can go deep regardless of input size.

Architecture

If you share the weights (more efficient) its basically an RNN



- The output of cross attention (bottleneck) is now less expressive.
  - We can repeatedly compute cross attention with input as we go deeper.

Experiments

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

Top-1 Acc ImageNet

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
Perceiver (mel spectrogram - tuned)	-	-	44.2

mAP in AudioSet

	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**

Top-1 Acc on ModelNet-50 (Point Clouds)

Negatives?

- As with most transformers the input is complemented by positional encodings
- Is this compatible with all modalities?
  - Typically yes, but ok.

Video. A full 32 frame clip at 224x224 resolution has more than 2 million pixels. We experimented using tiny space-time patches with dimensions 2x8x8, resulting in a total of 12,544 inputs to the Perceiver. We compute Fourier features for horizontal, vertical and time coordinates (scaled to [-1, 1]), and concatenated them with the RGB values.

Hmm

With great flexibility comes great overfitting, and many of our design decisions were made to mitigate this. In future work, we would like to pre-train our image classification model on very large scale data (Dosovitskiy et al., 2021). We