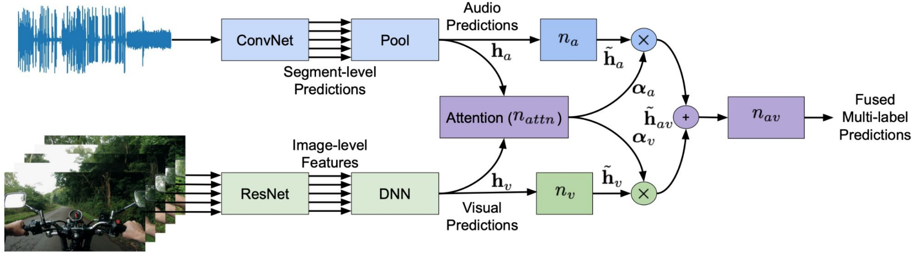
Perceiver: General Perception with Iterative Attention

Junjie Zhao & David Kim

Towards a Unified, Simpler Model

Multimodal Data



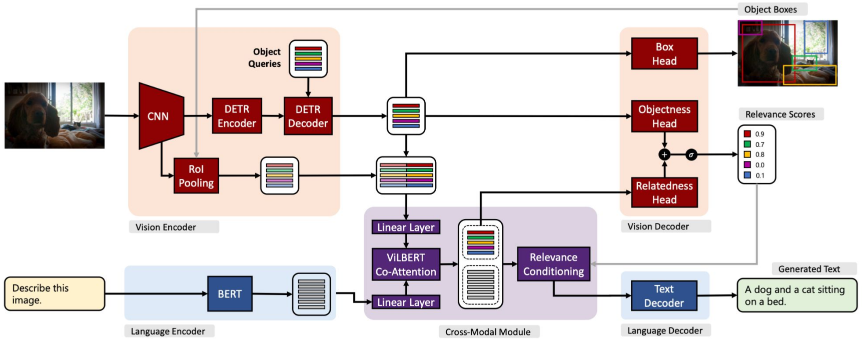


Figure 2. **Architecture of** GPV-1**.** **Vision**, **language**, and **cross-modal** modules are color-coded (see Sec. 3 for details).

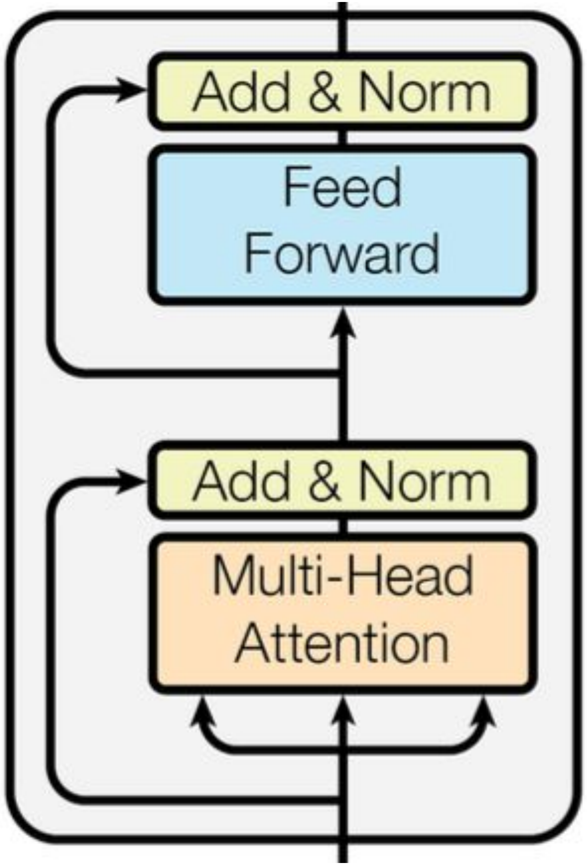
Input Dimensionality

Transformers have been a silver bullet for many tasks but limited by quadratic complexity.

Images: M = 224 x 244 = 50176

1 second of audio consists of 50,000 raw audio samples.

Previous work necessitated modality-specific assumptions about the input data e.g. tokenize images.



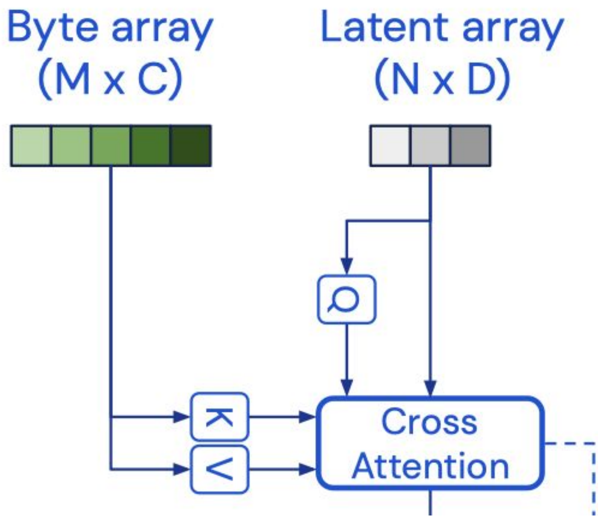
Key Idea: Attention bottleneck to distill Data

Cross-Attention

Input is encoded into a Byte array

Queries come from a much smaller learnable “latent” array initialized by a truncated normal distribution.

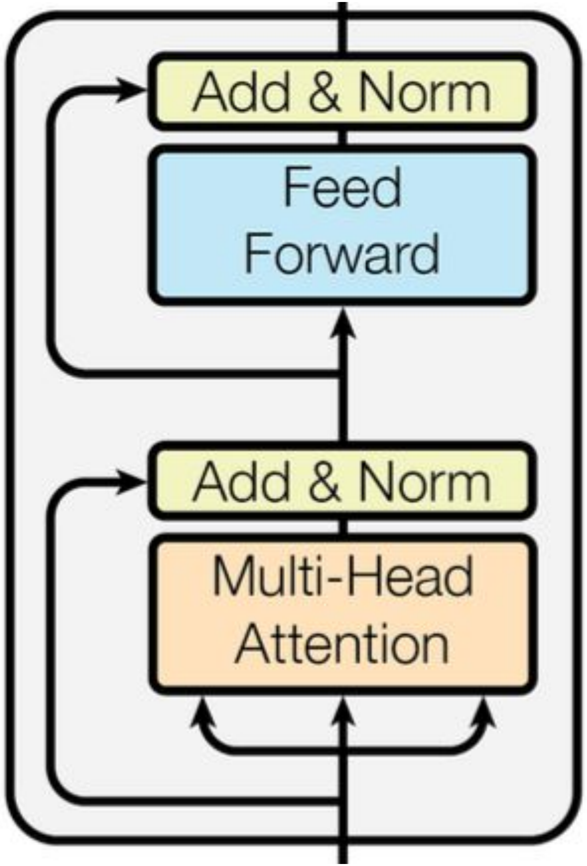
O(M^2) -> O(MN) s.t. N << M



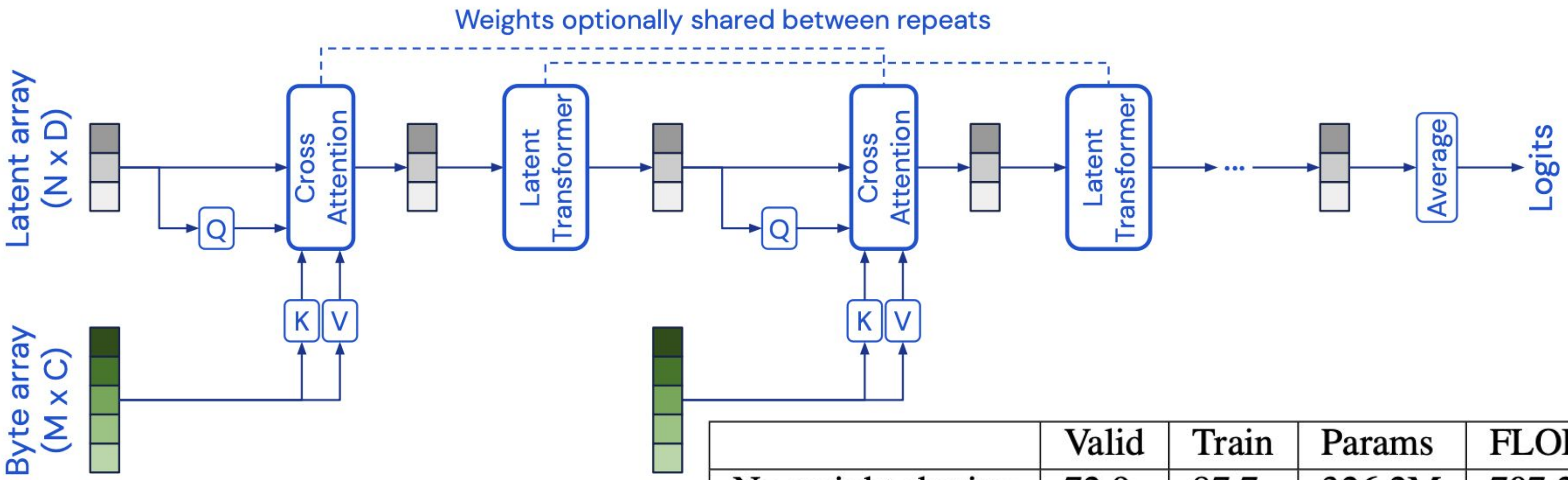
Latent Transformer Block

Just a plain old GPT-2 transformer block which is a modified vanilla transformer. LayerNorm layers are added before and after the Self-Attention blocks

Instead of each block being O(M^2), the latent transformer will be O(N^2) where N << M.



Iterative Cross-Attention & Weight Sharing



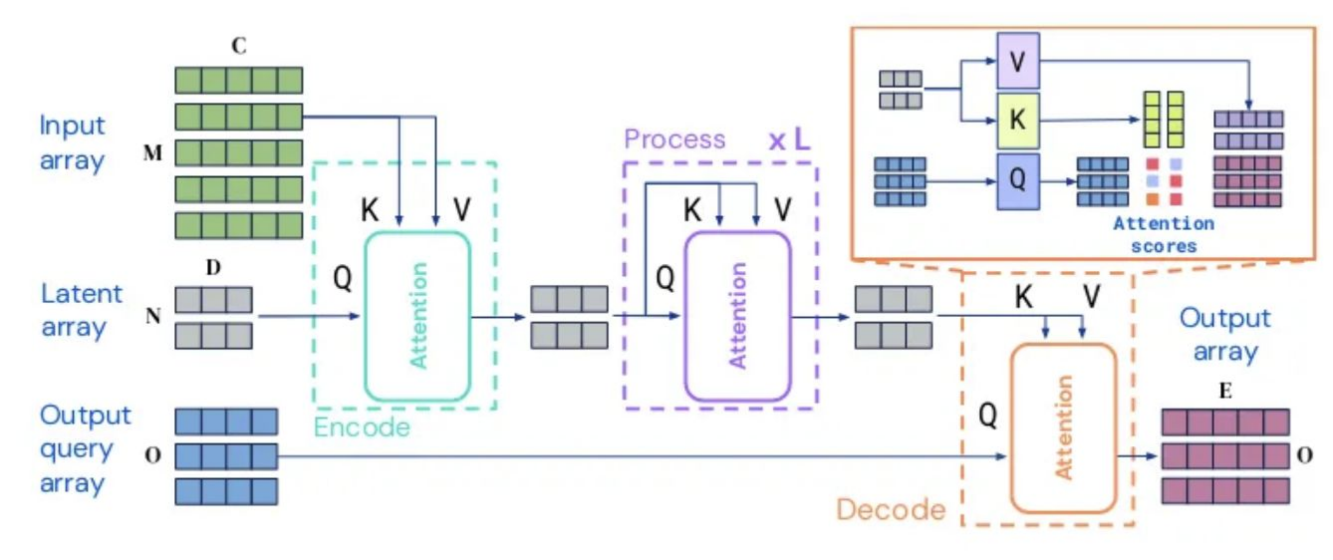
	Valid	Train	Params	FLOPs
No weight sharing	72.9	87.7	326.2M	707.2B
W/ weight sharing	**78.0**	79.5	44.9M	707.2B

Perceiver Architecture achieves generalizability with minimal assumptions about input data structure

Fourier Positional Encodings

	Raw	Perm.	Input RF
ResNet-50 (FF)	73.5	39.4	49
ViT-B-16 (FF)	76.7	61.7	256
Transformer (64x64) (FF)	57.0	57.0	4,096
Perceiver:			
(FF)	78.0	78.0	50,176
(Learned pos.)	70.9	70.9	50,176

Perceiver IO



Experiments

Single-image classification on ImageNet

Audio event classification on AudioSet (1.7M 10s long training videos and 527 classes)

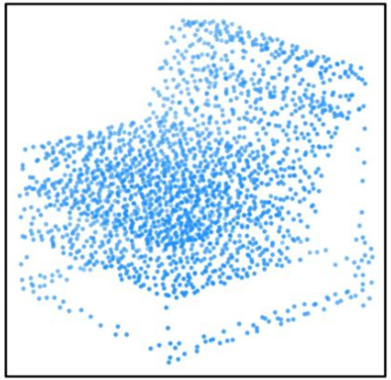
- Audio only
- Video
- Audio + video

Classification on ModelNet40 (Point clouds derived from 3D meshes)

Perceiver trained on diverse range of input data

Sample from ImageNet

Video and audio from AudioSet



Point clouds from ModelNet40

Experiments on ImageNet

Compared with ResNet-50 and ViT that use 2D convolutions; And Perceiver, Transformer that only use global attention

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 validation accuracy (in %) on ImageNet

Experiments on Permuted ImageNet

Evaluate how important domain-specific assumptions about grid structure are to the performance.

While models that only use global attention are stable under permutation, models that use 2D convolutions to process local neighborhoods are not

	Raw	Perm.	Input RF
ResNet-50 (FF)	73.5	39.4	49
ViT-B-16 (FF)	76.7	61.7	256
Transformer (64x64) (FF)	57.0	57.0	4,096
Perceiver:			
(FF)	78.0	78.0	50,176
(Learned pos.)	70.9	70.9	50,176

Top-1 validation accuracy (in %) on standard and **permuted** ImageNet

Experiments on AudioSet

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

Mean average precision (mAP) on audio, video and audio+video inputs

Experiments on ModelNet

	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**

Ablation Study