Perceiver

General Perception with Iterative Attention

One architecture for all modalities?

- Inductive bias has clearly been valuable when data is small (CNNs, LSTM, etc).
- But every new data modality requires a new architecture.
  - Current approaches use a separate feature extractor per modality
  - Can’t just concatenate audio spectrogram with an image and pass it to a ConvNet
  - → Whenever we fuse new modalities we need to think a lot
- Plus as we scale, inductive bias is not necessary (see eg. ViT) and even may become detrimental

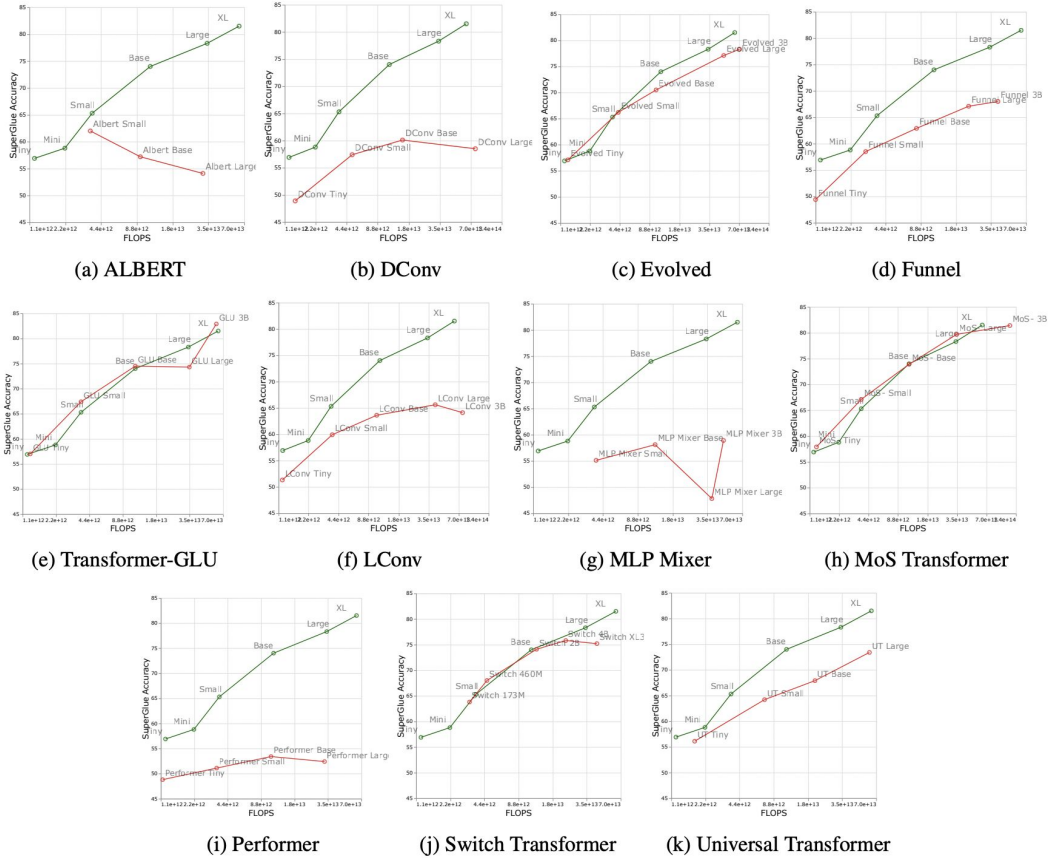
Are we better off building in as much flexibility as possible?

One architecture for arbitrary sequence lengths?

- Image data is processed before attention to account for large # pixels
  - by convolutions (ViT)
  - by dividing images into rows/columns (Sparse Attention)
  - by subsampling (UNITER)
- Why not instead attend to full input image?
  - full attention is quadratic in memory
  - linear attention does not scale

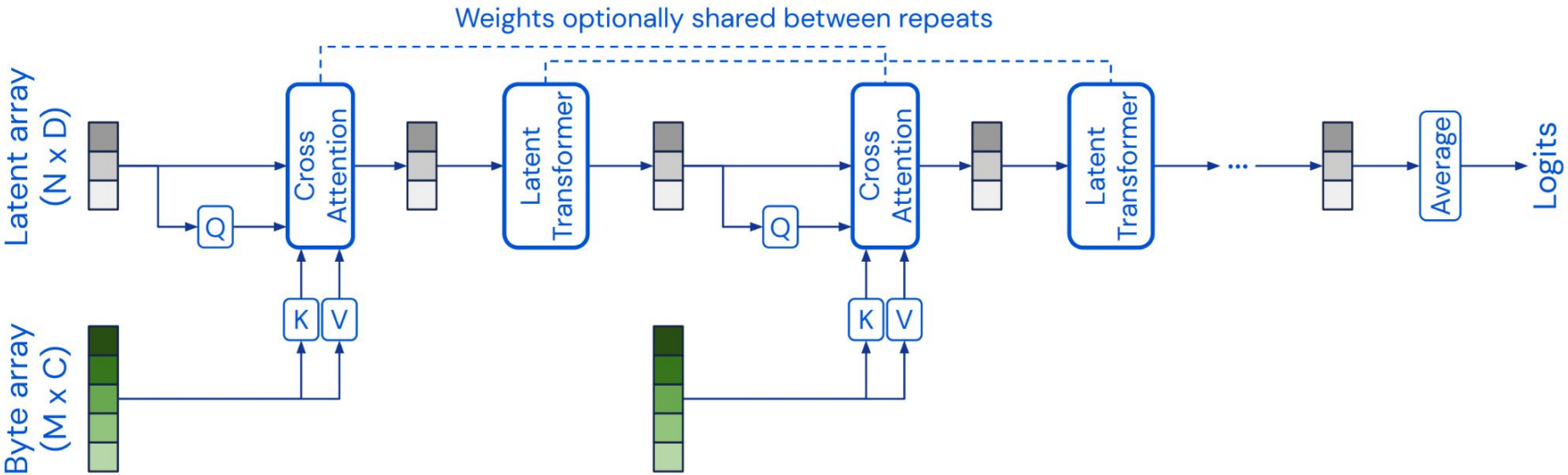
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

One architecture for arbitrary sequence lengths?



downstream accuracy
for vanilla transformer vs
fancy transformers

A Transformer Model for Arbitrary Modalities

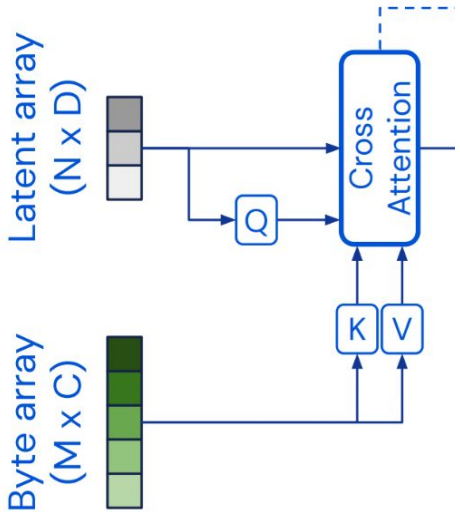


- (1) a cross-attention module mapping (bytes-array, latents) → latents
- (2) a transformer tower mapping latents → latents

Cross-Attention tames quadratic complexity

- Typical attention would be dominated by O(LN^2) term where N is input sequence length, L # layers
- Cross-attention is O(NM) where latent M<<N

- Overall architecture O(NM+ LM^2)
- Note N independent of # layers
- M determined by GPT2-like transformers



Weight sharing + Iterative Attention

- The size of the latents N is a bottleneck
- In practice use multiple cross-attention modules
  - Interleaved residual connections to input
  - But each must be followed by transformer block (adding a huge memory overhead)
- Sharing parameters between transformer blocks improves performance
  - 10x parameter reduction
  - Improves validation performance (overfitting)

	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

# cross-attends	Acc.	FLOPs	Params
1 (at start)	76.7	404.3B	41.1M
1 (interleaved)	76.7	404.3B	42.1M
2 (at start)	76.7	447.6B	44.9M
2 (interleaved)	76.5	447.6B	44.9M
4 (at start)	75.9	534.1B	44.9M
4 (interleaved)	76.5	534.1B	44.9M
8 (at start)	73.7	707.2B	44.9M
8 (interleaved)	**78.0**	707.2B	44.9M

Positional Encodings and Fourier Features

- Attention is permutation invariant, but what if we want to provide structural information to the model?
- Give positional information at multiple frequencies up to Nyquist criterion.
  - Parameterize frequency encoding to take values in

    $$[\sin(f_k \pi x_d), \cos(f_k \pi x_d)]$$

    where f_k is kth band of frequencies spaced equally between 1 and \mu/2 where \mu is the Nyquist frequency.

  - This gives us a size d(2K + 1) positional encoding where d is input data dimension and K is the number of frequencies.
- Concatenate position and input features before passing to Perceiver.

Positional Encodings and Fourier Features

- This approach is general!
  - We can encode whatever structural information is relevant in the input.
  - The model learns whether to use these features instead of imposing them via model architecture.
  - Multimodal data can be distinguished via this encoding.

Evaluations: Vision

- ImageNet (ILSVRC 2012 split)

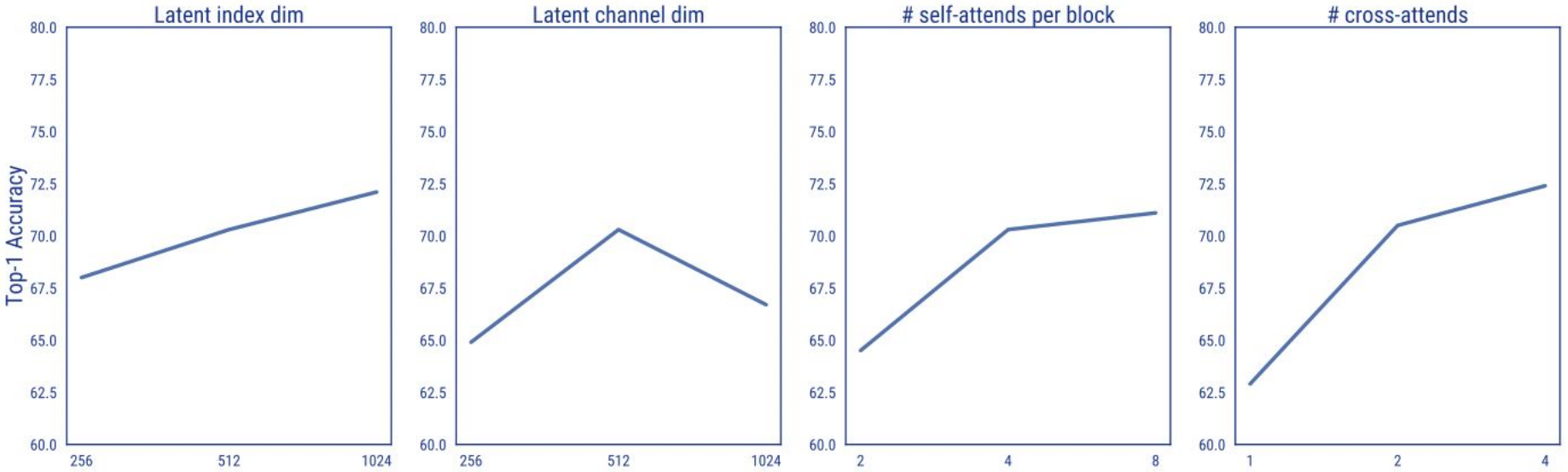
ResNet-50 (He et al., 2016)	77.6
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
ResNet-50 (FF)	73.5
ViT-B-16 (FF)	76.7
Transformer (64x64, FF)	57.0
Perceiver (FF)	78.0

Table 1. Top-1 validation accuracy (in %) on ImageNet. Models that use 2D convolutions exploit domain-specific grid structure architecturally, while models that only use global attention do not. The first block reports standard performance from pixels – these numbers are taken from the literature. The second block shows performance when the inputs are RGB values concatenated with 2D Fourier features (FF) – the same that the Perceiver receives. This block uses our implementation of the baselines. The Perceiver is competitive with standard baselines on ImageNet without relying on domain-specific architectural assumptions.

	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

Table 2. Top-1 validation accuracy (in %) on standard (raw) and **permuted** ImageNet (higher is better). Position encodings (in parentheses) are constructed before permutation, see text for details. While models that only use global attention are stable under permutation, models that use 2D convolutions to process local neighborhoods are not. The size of the local neighborhood at input is given by the input receptive field (RF) size, in pixels.

Some hyperparameter sweeps for ImageNet



Some weirdness: use LAMB optimizer, use RandAugment, strange custom LR

Evaluations: Audio and Video

- AudioSet
  - 1.7M 10s long training videos and 527 classes
  - Sample 32-frame clips for training
  - Eval by avg prediction from 16 overlapping 32-frame clips

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

Evaluations: Point Clouds

- ModelNet40
  - Point clouds from 3D triangular meshes
  - Predict object from 2000 points in 3D space

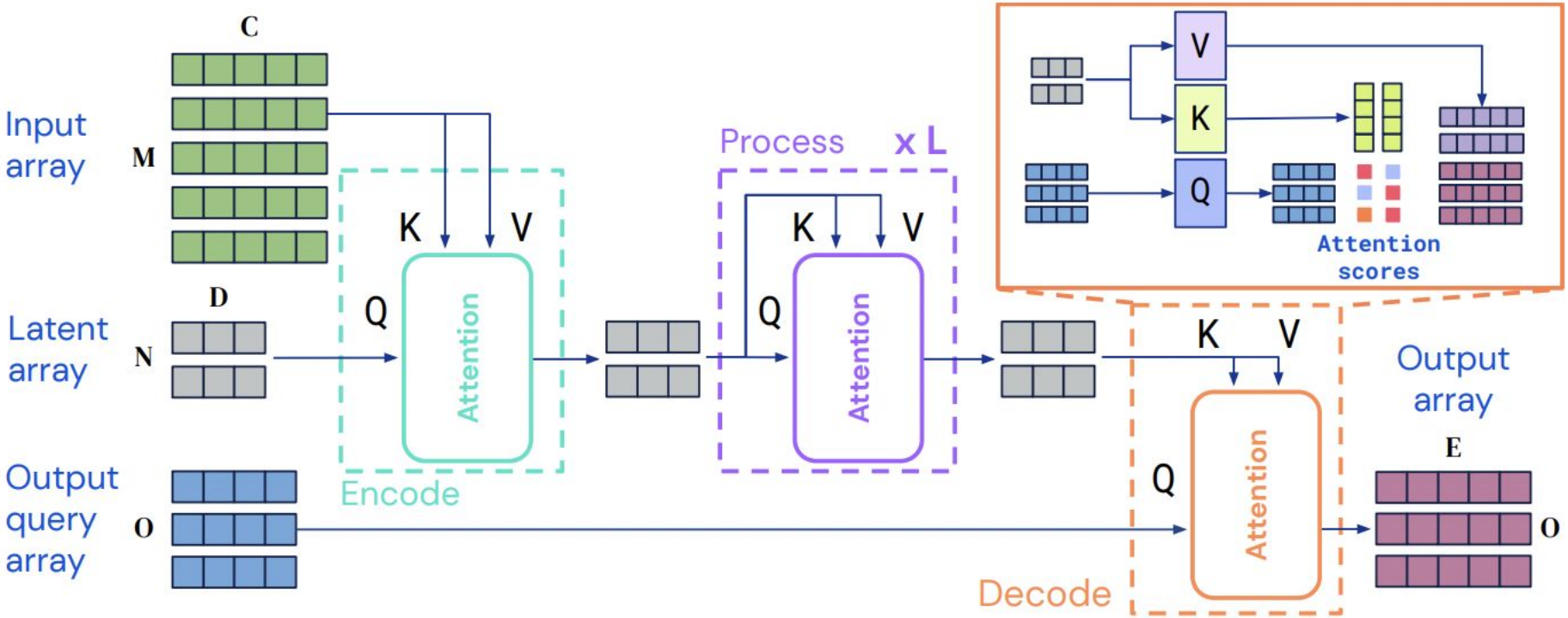
	Accuracy
PointNet++ (Qi et al., 2017)	**91.9**
ResNet-50 (FF)	66.3
ViT-B-2 (FF)	78.9
ViT-B-4 (FF)	73.4
ViT-B-8 (FF)	65.3
ViT-B-16 (FF)	59.6
Transformer (44x44)	82.1
Perceiver	**85.7**

Perceiver IO

General Architecture for Structured Inputs and Outputs

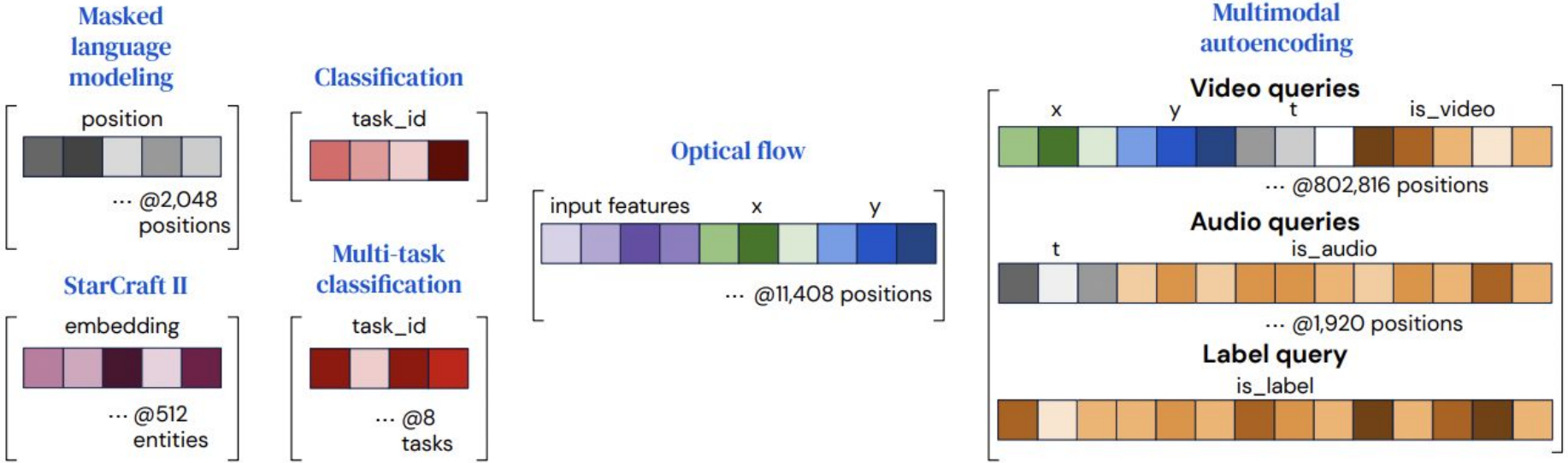
One architecture for all tasks?

- Perceiver was designed only for classification.
- Can we do multitask learning with a single architecture?



Output Queries

- Classification: fixed or learned vector of length # categories
- Spatial Structure (semantic segmentation): vector of positional encodings
- Task/Modality Specific: learn a single query for each task or modality
- Starcraft: One vector per unit



Evaluations (Overview)

Modalities	Tasks	Preprocessing	Postprocessing	# Inputs	# Outputs
Text	Token-level pred.	Tokenization + Embed.	Linear projection	512×768	512×768
Text	Byte-level pred.	Embed.	None	$2,048 \times 768$	$2,048 \times 768$
Text	Multi-task (8 tasks)	Embed.	None	$2,048 \times 768$	8×768
Video	Flow prediction	None	None	$365,056 \times 64$	$182,528 \times 64$
Video	Flow prediction	Concat	None	$182,528 \times 64$	$182,528 \times 64$
Video	Flow prediction	Conv+maxpool	RAFT upsampling	$22,816 \times 64$	$11,408 \times 64$
Video	Flow prediction	Conv+maxpool+concat	RAFT upsampling	$11,408 \times 64$	$11,408 \times 64$
Video+Audio+Label	Autoencoding	Patch: 1x4x4 Vid, 16 Aud	None	$50,657 \times 704$	$803,297 \times 512$
Image	Classification	None	None	$50,176 \times 3$	$1 \times 1,000$
Image	Classification	Linear projection	None	$50,176 \times 256$	$1 \times 1,000$
Image	Classification	Conv+maxpool	None	$3,136 \times 64$	$1 \times 1,000$
StarCraft Unit Set	Encoding and Classification	Tokenization	Pointer network	512×256	512×128
Video+Audio	Classification	Patch: $2 \times 8 \times 8$ Vid, 128 Aud	None	$13,024 \times 487$	1×527
Video+Audio	Classification	Patch: $2 \times 8 \times 8$ Vid. Aud $\rightarrow$ mel-spectrogram	None	$17,344 \times 487$	1×527

Evaluations (Queries)

Domain	Input Modality	Encoder KV input	Encoder KV channels	Decoder query input	Decoder query channels
Language (MLM)	Text	byte/token encoding + learned pos	768	learned pos	1280
Language (Perceiver IO++ MLM)	Text	byte/token encoding + learned pos	768	learned pos	1536
Language (GLUE)	Text	byte/token encoding + learned pos	768	Class query (per-task)	1280
Language (Perceiver IO++ GLUE)	Text	byte/token encoding + learned pos	768	Class query (per-task)	1536
Optical Flow	Video (concat. frames)	[conv or Linear(concat RGB), 2D FFs]	322	[Linear(RGB), 2D FFs]	322
Optical Flow	Video	[conv or Linear(RGB), 3D FFs]	451	[conv features, 3D FFs]	451
Kinetics	Video, Audio, Label	[patched RGB, 3D FFs, learned modality feat.] [patched sound pressure, 1D FF, learned modality feat.] [one-hot label, learned modality feat.]	704 704 704	[3D FFs, learned modality feat.] [1D FF, learned modality feat.] [learned modality feat.]	1026 1026 1026
ImageNet (2D FFs)	Image	[RGB, 2D FFs]	261	Class query (single)	1024
ImageNet (learned pos)	Image	[Linear(RGB), learned pos]	512	Class query (single)	1024
ImageNet (conv)	Image	[Conv features, 2D FFs]	322	Class query (single)	1024
StarCraft II	SC2 entities	Entity features	128	Entity features	128
AudioSet	Video, Audio	[patched RGB, 3D FFs, learned modality feature] [patched sound pressure, 1D FFs, learned modality feature]	487 487	Class query (single)	1024
AudioSet	Video, Mel-spectrogram	[patched RGB, 3D FFs, learned modality feature] [mel-spectrogram features, 1D FFs, learned modality feature]	487 487	Class query (single)	1024

Evaluations (Masked Language Modeling, GLUE)

Model	Tokenization	M	N	Depth	Params	FLOPs	SPS	Avg.
BERT Base (test)	SentencePiece	512	512	12	110M	109B	-	81.0
BERT Base (ours)	SentencePiece	512	512	12	110M	109B	7.3	81.1
Perceiver IO Base	SentencePiece	512	256	26	223M	119B	7.4	**81.2**
BERT (matching FLOPs)	UTF-8 bytes	2048	2048	6	20M	130B	2.9	71.5
Perceiver IO	UTF-8 bytes	2048	256	26	201M	113B	7.6	81.0
Perceiver IO++	UTF-8 bytes	2048	256	40	425M	241B	4.2	**81.8**

Table 1: **Perceiver IO on language**: results on the GLUE benchmark (Avg. = average performance, higher is better). Following Devlin et al. (2019) we exclude the WNLI task. We use Pearson correlation on STS-B, Matthews correlation on CoLa and accuracy on the remaining tasks. BERT Base (test) performance is reported from Devlin et al. (2019). SPS = train-time steps per second. M = # inputs and N = # latents.

Bytes-level Perceiver outperforms sentencepiece tokenized BERT model

Evaluations (Optical Flow)

- Given 2 consecutive frames of video, estimate the displacement for each pixel in the first image
- SOTA methods decompose this into (1) find correspondence between points, (2) compute relative offsets, (3) propagate result to regions of image that aren't anchored by an object e.g. empty space between objects
- Perceiver just extracts 3x3 patches and adds positional encoding, then passes to PerceiverIO

Network	Sintel.clean	Sintel.final	KITTI
PWCNet (Sun et al., 2018)	2.17	2.91	5.76
RAFT (Teed & Deng, 2020)	1.95	2.57	**4.23**
Perceiver IO	**1.81**	**2.42**	4.98

Evaluations (Multimodal Encoding)

- Kinetics-700-2020 dataset
  - Video, audeo, class labels
  - Encode and reconstruct all modalities simultaneously
  - Train on 16 224x224 frames with 30k raw audio samples → 1920 16d vectors and one 700d one-hot class label per sample

Figure 4: Multimodal audio-video-label autoencoding with 88x compression. Side-by-side: inputs on left, reconstructions right. See the supplemental material for example output video and audio.