Perceiver: General Perception with Iterative Attention

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

Presentation By: Alejandro Zamora, Brandon Clark, Joshua Gracie, Dunquan Zheng
Outline

- Introduction
- Architecture
- Experiments
- Ablations
- Conclusion
Introduction and Motivation

- Prior models are not flexible with respect to modality
  - Require specific domain knowledge to function (e.g. locality)
- Multi-Modality
  - Audio
  - Video
  - 3D point Cloud
- Attention Based Architecture
  - Quadratic Time $O(M^2)$
Key Aspects

- Byte Array
- Latent Array
- Cross-Attention
- Latent Transformer
- Reduced Complexity

The puppy was sleeping.
Architecture
Cross Attention

- Latent Array (NxD) is a learned array
  - Queries
- Byte Array (MxC) is taken directly from input
  - Keys and Values
- Choose N << M to help reduce computation
Cross Attention

\[ \text{Layer Norm} \rightarrow NxH \]

\[ \begin{align*} &Q \quad \text{Layer Norm} \quad NxM \\ &QK^T \quad \text{Layer Norm} \quad MxH \\ &\text{Softmax} \quad \text{Layer Norm} \quad MxH \end{align*} \]

\[ \text{Linear} \rightarrow \text{Output} \]

\[ \text{Layer Norm} \rightarrow MxC \]

\[ \begin{align*} &Q \quad \text{Layer Norm} \quad NxH \\ &K \quad \text{Layer Norm} \quad MxH \\ &V \quad \text{Linear} \quad MxC \end{align*} \]
Complexity Analysis

Bottle Neck

Attention = \text{softmax}(QK^T)V

Cross Attention

\(O(NM)\)

1 time per CA-Module

Self Attention

\(O(N^2)\)

L times per Transformer

\(O(NM + LN^2)\)
Byte Array Construction

- Images
  - Individual Pixels (M = 50,176 for a 224x224)
- Videos
  - Tubelets 2x8x8 (M = 12,544 for 32 frames of size 224x224)
- Audio
  - Raw Patches (M = 480 for 1.28s clips)
  - Mel Spectrogram (M = 4800)
- Point Cloud
  - Dependent on # of points
Position Embedding

- Learned Embeddings
- Fourier Feature Embeddings
Fourier Features for Position Embedding

- Increase dimensionality of input channel
- Provide positional information
- Concatenated to input tokens

$$[\sin(f_0 \pi x_d), \cos(f_0 \pi x_d), \ldots, \sin(f_k \pi x_d), \cos(f_k \pi x_d), x_d]$$

- $f_i = 2^i$
- $x_d$ = position value along dimension $d$
- $k$ = # of frequencies/bands
Fourier Feature Example (Image)

Pixel position = (1,0.5)  \[ k = 64 \]

\[ P_0 = [\sin(2^0\pi*1), \cos(2^0\pi*1), \ldots, \sin(2^{64}\pi*1), \cos(2^{64}\pi*1), 1] \]

\[ P_1 = [\sin(2^0\pi*0.5), \cos(2^0\pi*0.5), \ldots, \sin(2^{64}\pi*0.5), \cos(2^{64}\pi*0.5), 0.5] \]

Position Encoding = concat(\(P_0, P_1\))

\[ \text{size}(PE) = d(2k + 1), d = \# \text{ of dimensions} \]
Architecture Details

- Latent Transformers are basic structure
- CA and SA both followed by MLP
- Typically alternate CA and LT
  - Ablation study puts all CA first
## Architecture Variations

<table>
<thead>
<tr>
<th>Modality</th>
<th>Cross-Attention</th>
<th>Latent Transformer Layers</th>
<th>Byte Array size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>8</td>
<td>6</td>
<td>50,176</td>
</tr>
<tr>
<td>Video + Audio</td>
<td>2</td>
<td>8</td>
<td>12,544 +</td>
</tr>
<tr>
<td>Pointcloud</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
LAMB Optimizer

\[ x_{t+1} = x_t + \eta_t u_t \]

\( x_i = \text{weights at time } i \quad \eta_t = \text{learning rate} \quad u_i = \text{update} \)

1.) Normalize the update at every layer
2.) Scale the learning rate at every layer

\( g_t^{(i)} = \text{gradient at } i^{th} \text{ layer} \quad \phi(||x_t^{(i)}||) = \text{scaling factor} \)

\[ x_{t+1} = x_t - \eta_t \phi(||x_t^{(i)}||) \frac{g_t^{(i)}}{||g_t^{(i)}||} \]
Classifications

- Final output is of size NxD
- Average over N to create a 1xD array of logits
- Logits passed through a linear classification layer
Datasets

- Image net
- AudioSet
- ModelNet40
Experiments: ImageNet Training

- ‘Inception-style’ preprocessing with image size of 224x224
- RandAug
- LAMB Optimizer (instead of SGD)

Parameters:

- 120 epochs
- LR: 0.004 decayed by factor of 10 at epochs [84,102,114]
Experiments: ImageNet Training

- (x,y) coordinates of the cropped image were used to generate positional encodings
- Pre-crop coordinates allow Perceiver to ‘memorize’ the pixels associated with unique features - overfitting
- Cropped coords - augment position and aspect ratio
Experiments: ImageNet Training

1) 8 Cross attends
2) Transformer (6 blocks + 1 self attend with 1 head per block)

- Weights are shared for 2nd cross attend on
  - Weights sharing with all cross attends caused instability
- ~45mil params
Experiments: ImageNet Performance

- ResNet50 and ViT were used for comparison
- Also used a Transformer with pixel inputs - same as transformer layer of Perceiver

<table>
<thead>
<tr>
<th>Model</th>
<th>@2016</th>
<th>@2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50 (He et al., 2016)</td>
<td>77.6</td>
<td>77.9</td>
</tr>
<tr>
<td>ViT-B-16 (Dosovitskiy et al., 2021)</td>
<td>73.5</td>
<td>76.7</td>
</tr>
<tr>
<td>ResNet-50 (FF)</td>
<td>57.0</td>
<td></td>
</tr>
<tr>
<td>ViT-B-16 (FF)</td>
<td>78.0</td>
<td></td>
</tr>
<tr>
<td>Transformer (64x64, FF)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceiver (FF)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments: ImageNet Permutations

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Deterministic Permutation Pattern

<table>
<thead>
<tr>
<th>5</th>
<th>3</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>
Experiments: ImageNet Permutations

Experiments: ImageNet Permutations

- Permutations made to images after position features are generated
- Perceiver is capable of learning the images with no prior knowledge about their spatial features (learned pos.)

<table>
<thead>
<tr>
<th>Model</th>
<th>Raw</th>
<th>Perm.</th>
<th>Input RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50 (FF)</td>
<td>73.5</td>
<td>39.4</td>
<td>49</td>
</tr>
<tr>
<td>ViT-B-16 (FF)</td>
<td>76.7</td>
<td>61.7</td>
<td>256</td>
</tr>
<tr>
<td>Transformer (64x64) (FF)</td>
<td>57.0</td>
<td>57.0</td>
<td>4,096</td>
</tr>
<tr>
<td>Perceiver: (FF)</td>
<td>78.0</td>
<td>78.0</td>
<td>50,176</td>
</tr>
<tr>
<td></td>
<td>70.9</td>
<td>70.9</td>
<td>50,176</td>
</tr>
</tbody>
</table>
Experiments: Images

Original Image  Cross Attn 1  Cross Attn 2  Cross Attn 8
Experiments: Images
Experiments: Images
Experiments: AudioSet Training

- Attends 2 times:
  - Cross attend
  - Transformer (8 self-attention blocks)
- Weights are not shared during AudioSet training because there are only 2 attention iterations
- Trained for 100 Epochs
Experiments: Audio/Video

- Audio input
  - Raw audio file to byte array
  - Spectrogram to byte array
- Video input
  - $2 \times 8 \times 8$ tubelet + 3 Dimensional fourier features, produces spacetime patches
- Audio + Video input
  - Video space time patches + raw audio vectors
  - Video space time patches + spectrogram values
Experiments: Audio/Video Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Audio (mAP)</th>
<th>Video (mAP)</th>
<th>Audio/Video (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-14 (Kong et al., 2020)</td>
<td>43.1</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>Attention AV-fusion (Fayek &amp; Kumar, 2020)</td>
<td>38.4</td>
<td>25.7</td>
<td>46.2</td>
</tr>
<tr>
<td>Perceiver (raw audio)</td>
<td>38.3</td>
<td>25.8</td>
<td>43.5</td>
</tr>
<tr>
<td>Perceiver (mel spectrogram - tuned)</td>
<td>na</td>
<td>na</td>
<td>44.2</td>
</tr>
</tbody>
</table>

- **AudioSet**
  - Comprised of 2.1 million annotated video
  - 5.8 thousand hours of audio
- CNN-14 specifically trained for audio task and contains pretrained audio from larger dataset
- Perceiver competitive with state of art models
Experiments: Attention maps trained on Audio only

Spectrogram Input

First cross attend

Second cross attend
Experiments: Attention maps trained on video only

First cross attend

Second cross attend
Experiments: Attention maps trained on video and audio

First cross attend

Second cross attend

Spectrogram
Input
Experiments: Attention maps trained on video and audio

Video Input

First cross attend

Second cross attend
Experiments: Point Clouds

- Point cloud input
  - Point cloud is represented by unordered vector of points in space
  - Point cloud input is zero centered
  - Augmentations such as point scaling is implemented
  - Unordered point cloud vector is flatten into byte array
Experiments: Point Clouds Performance

- Predict the class of each object, given the coordinates of ~ 2000 points in 3D space
- PointNet++ specifically trained for point cloud data classification and contains inductive bias pertaining to data

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PointNet++ (Qi et al., 2017)</td>
<td>91.9</td>
</tr>
<tr>
<td>Perceiver</td>
<td>85.7</td>
</tr>
<tr>
<td>Transformer (44x44)</td>
<td>82.1</td>
</tr>
<tr>
<td>ViT-B-2 (FF)</td>
<td>78.9</td>
</tr>
<tr>
<td>ViT-B-4 (FF)</td>
<td>73.4</td>
</tr>
</tbody>
</table>
## Ablation Study

<table>
<thead>
<tr>
<th># cross-attends</th>
<th>Acc.</th>
<th>FLOPs</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (at start)</td>
<td>76.7</td>
<td>404.3B</td>
<td>41.1M</td>
</tr>
<tr>
<td>1 (interleaved)</td>
<td>76.7</td>
<td>404.3B</td>
<td>42.1M</td>
</tr>
<tr>
<td>2 (at start)</td>
<td>76.7</td>
<td>447.6B</td>
<td>44.9M</td>
</tr>
<tr>
<td>2 (interleaved)</td>
<td>76.5</td>
<td>447.6B</td>
<td>44.9M</td>
</tr>
<tr>
<td>4 (at start)</td>
<td>75.9</td>
<td>534.1B</td>
<td>44.9M</td>
</tr>
<tr>
<td>4 (interleaved)</td>
<td>76.5</td>
<td>534.1B</td>
<td>44.9M</td>
</tr>
<tr>
<td>8 (at start)</td>
<td>73.7</td>
<td>707.2B</td>
<td>44.9M</td>
</tr>
<tr>
<td>8 (interleaved)</td>
<td><strong>78.0</strong></td>
<td><strong>707.2B</strong></td>
<td><strong>44.9M</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># cross-attends</th>
<th>Acc.</th>
<th>FLOPs</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>39.4</td>
<td>173.1B</td>
<td>12.7M</td>
</tr>
<tr>
<td>8</td>
<td>45.3</td>
<td>346.1B</td>
<td>23.8M</td>
</tr>
<tr>
<td>12</td>
<td>OOM</td>
<td>519.2B</td>
<td>34.9M</td>
</tr>
</tbody>
</table>
Ablation Study

- Weight is shared between cross attention modules 2-8 in architecture
- Weight is shared across all latent transformers in architecture

<table>
<thead>
<tr>
<th></th>
<th>Valid</th>
<th>Train</th>
<th>Params</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>No weight sharing</td>
<td>72.9</td>
<td>87.7</td>
<td>326.2M</td>
<td>707.2B</td>
</tr>
<tr>
<td>W/ weight sharing</td>
<td>78.0</td>
<td>79.5</td>
<td>44.9M</td>
<td>707.2B</td>
</tr>
</tbody>
</table>
Ablation Study

- Latent index dim
- Latent channel dim
- # self-attends per block
- # cross-attends
Conclusion

- Perceiver model prioritizes flexibility of input over specifically train modalities
- Different Modalities can be combined
- Minimal of inductive biases affects performance in different benchmarks
- Competitive results across different modalities
Sources

- https://gateway.newton.ac.uk/node/10214
- https://kazemnejad.com/blog/transformer_architecture_positional_encoding