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)

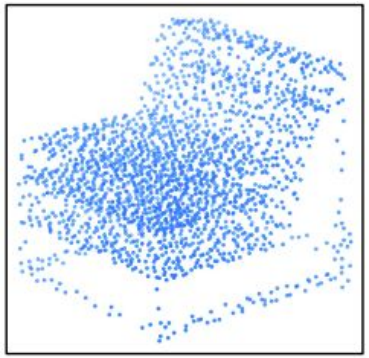


Key Aspects

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

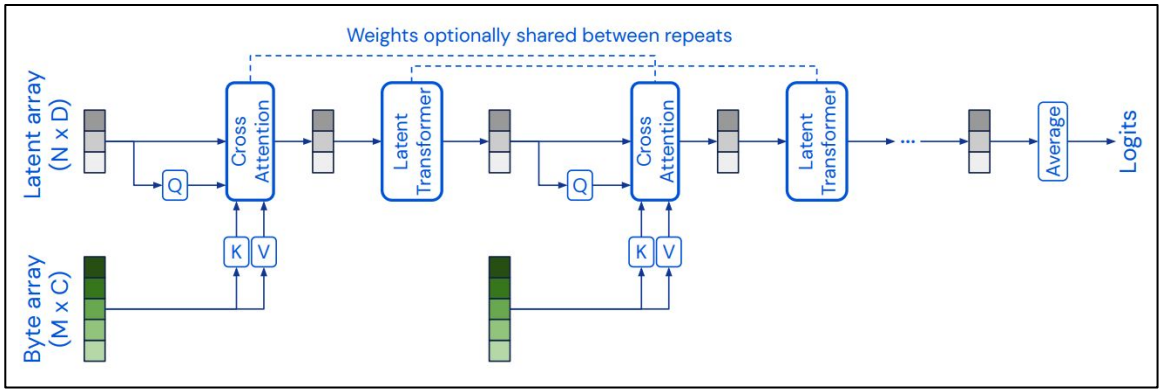








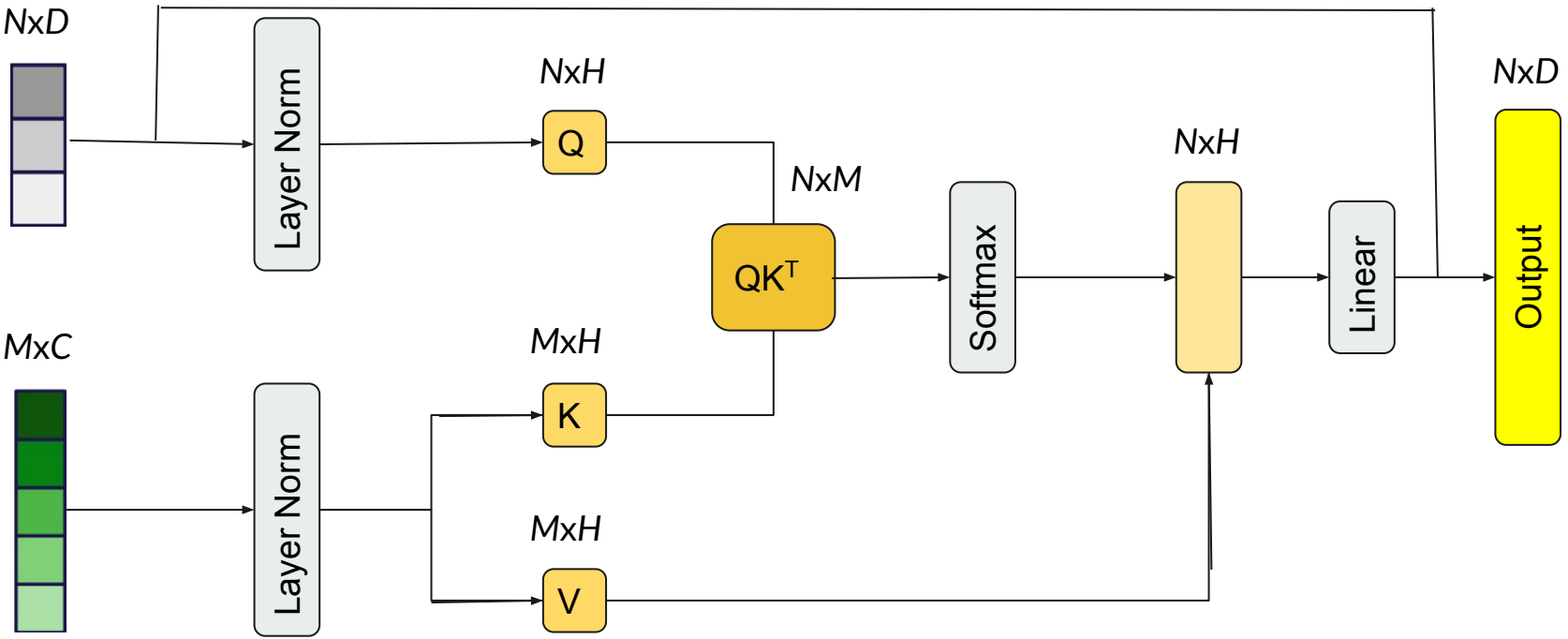
Architecture



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



Complexity Analysis

Bottle Neck

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

Cross Attention

$O(NM)$

1 time per
CA-Module

Self Attention

$O(N^2)$

L times per
Transformer

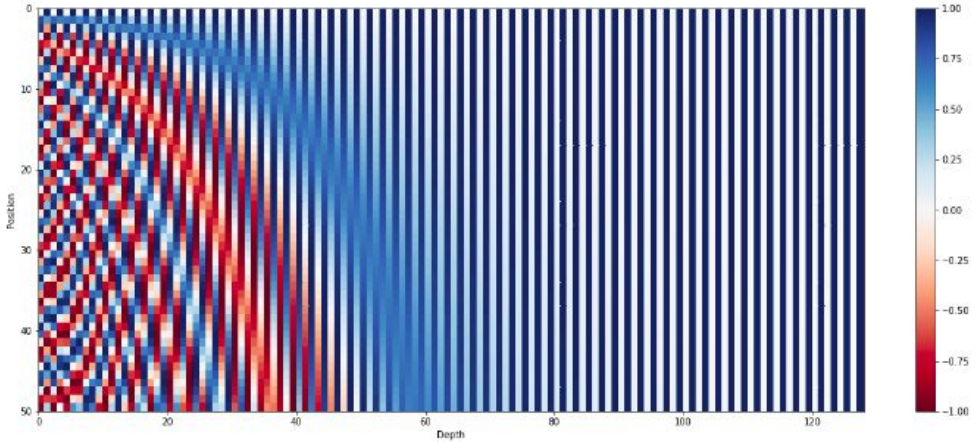
$O(NM + LN^2)$

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

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

$f_i = 2^i$
x_d = position value along dimension d
k = # of frequencies/bands

Fourier Feature Example (Image)

Pixel position = (1,0.5) k = 64

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

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

Position Encoding = concat(P_0, P_1)

size(PE) = $d(2k + 1)$, d = # 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

Modality	Cross-Attention	Latent Transformer Layers	Byte Array size

LAMB Optimizer

$x_{t+1} = x_t + \eta_t u_t$ x_i = weights at time i η_t = learning rate u_i = update

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

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

$$x_{t+1} = x_t - \eta_t * \phi(||x_t^{(i)}||) * (g_t^{(i)} / ||g_t^{(i)}||)$$

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

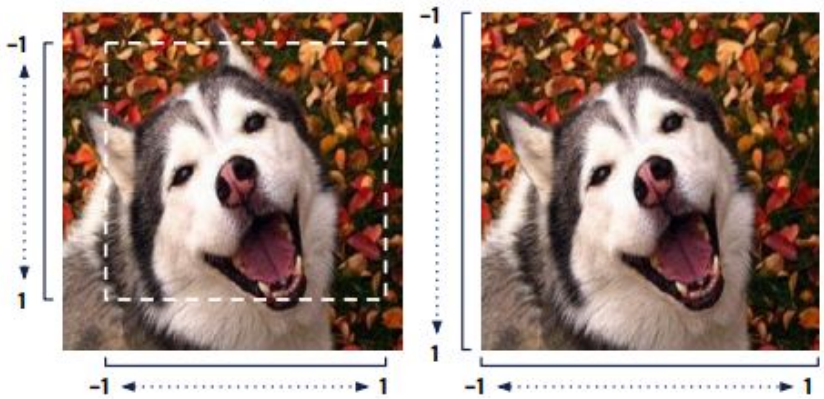
- ‘Inception-style’ preprocessessing with image size of 224x224
- RandAug
- LAMB Optimizer (instead of SGD)

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



(a) Crop-relative coordinates (b) Image-relative coordinates

Experiments: ImageNet Training

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

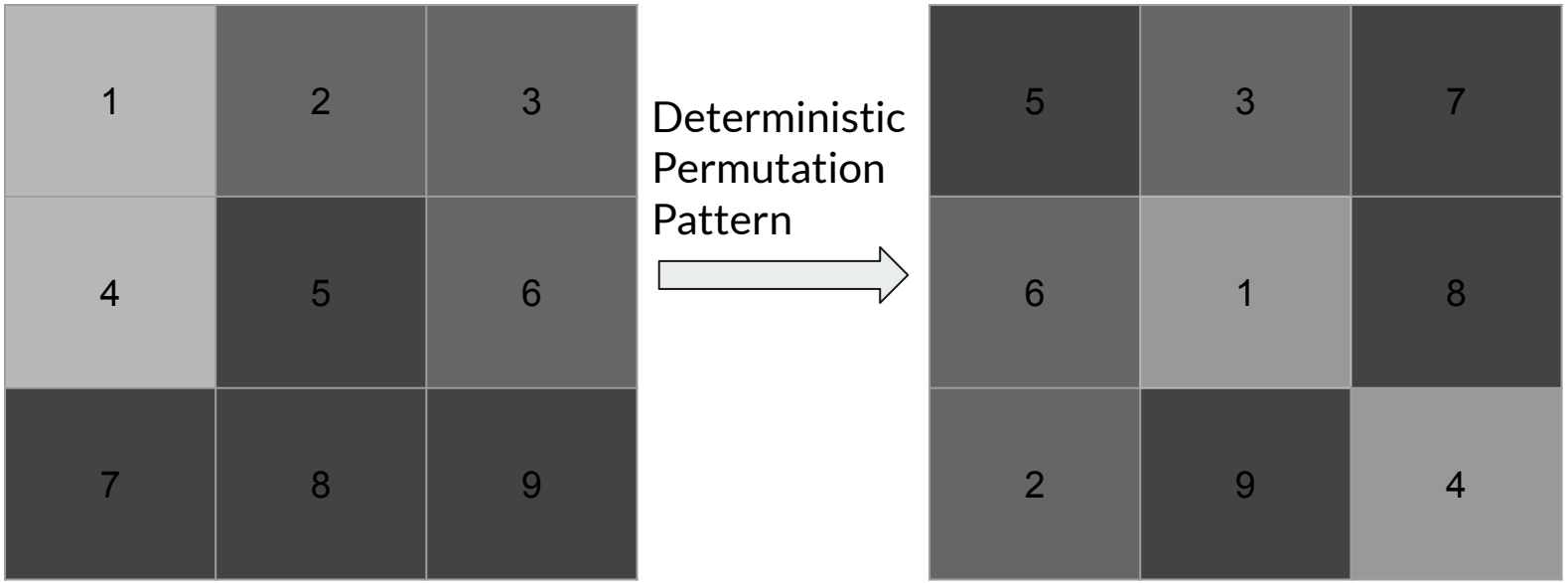
- Weights are shared for 2nd cross attend on
  - Weights sharing with all cross attends caused instability
- ~45mil params

Experiments: ImageNet Performance

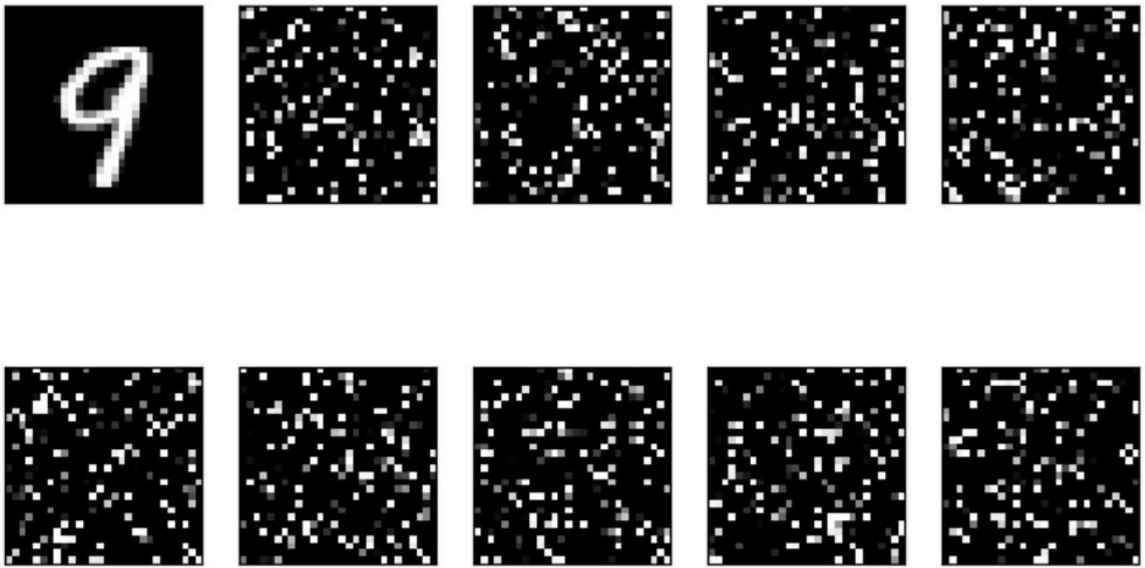
ResNet-50 (He et al., 2016)	77.6
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
ResNet-50 (FF)	73.5
ViT-B-16 (FF)	76.7
Transformer (64x64, FF)	57.0
Perceiver (FF)	78.0

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

Experiments: ImageNet Permutations



Experiments: ImageNet Permutations



Kemker, Ronald, et al. "Measuring catastrophic forgetting in neural networks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1. 2018.

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.)

	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

Experiments: Images

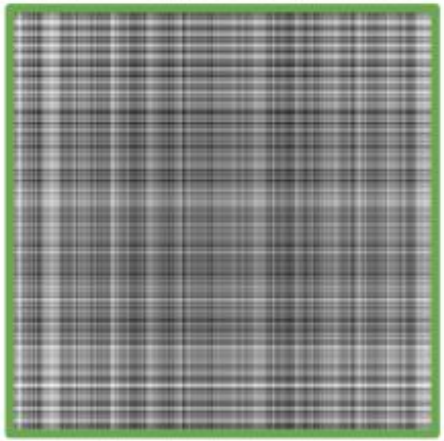




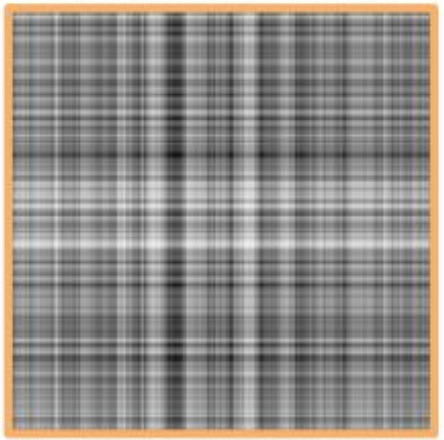
Original Image



Cross Attn 1

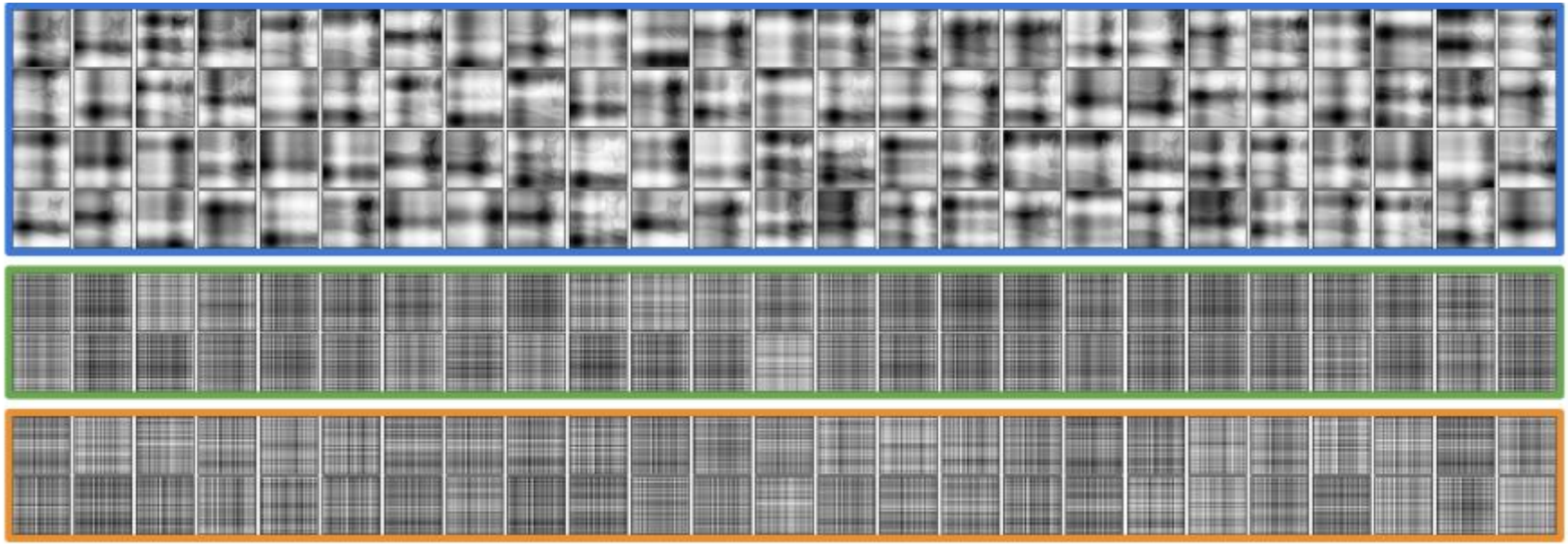


Cross Attn 2

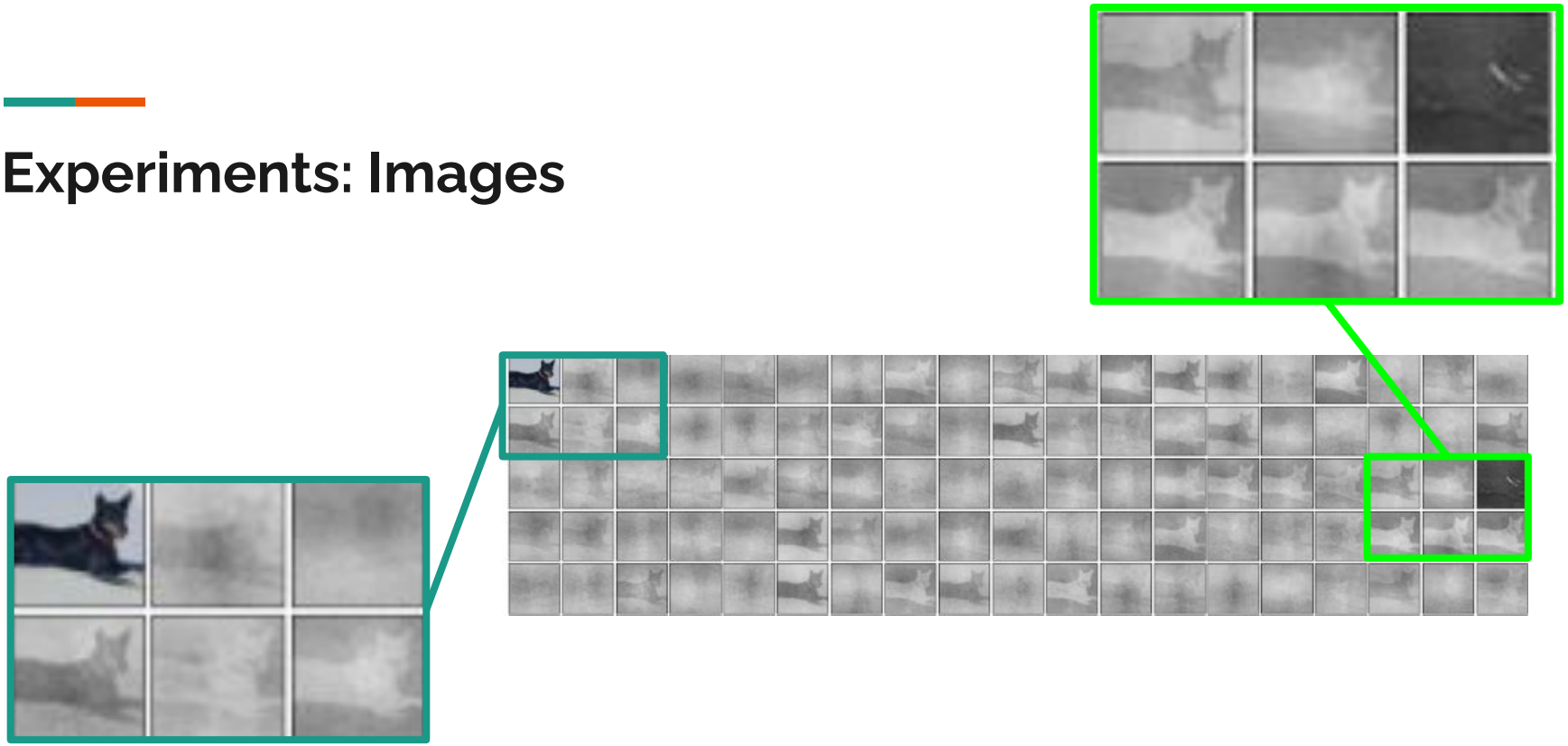


Cross Attn 8

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 x 8 x 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

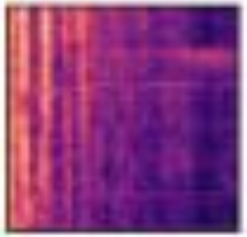
Experiments: Audio/Video Performance

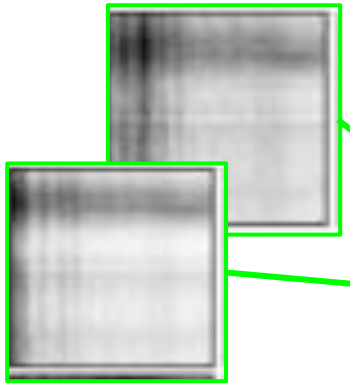
Model	Audio (mAP)	Video (mAP)	Audio/ Video (mAP)
CNN-14 (Kong et al., 2020)	**43.1**	na	na
Attention AV-fusion (Fayek & Kumar, 2020)	38.4	25.7	**46.2**
Perceiver (raw audio)	38.3	**25.8**	43.5
Perceiver (mel spectrogram - tuned)	na	na	44.2

- 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

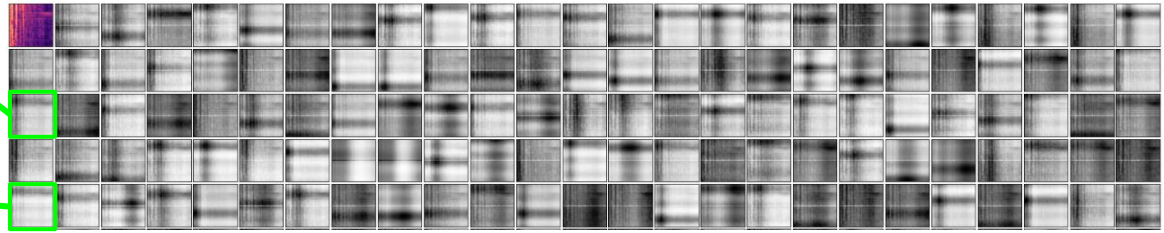
Experiments:Attention maps trained on Audio only

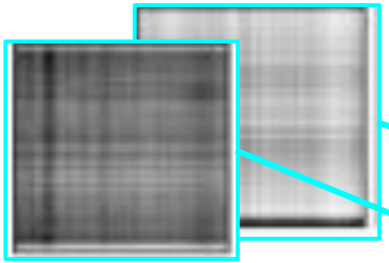
Spectrogram Input



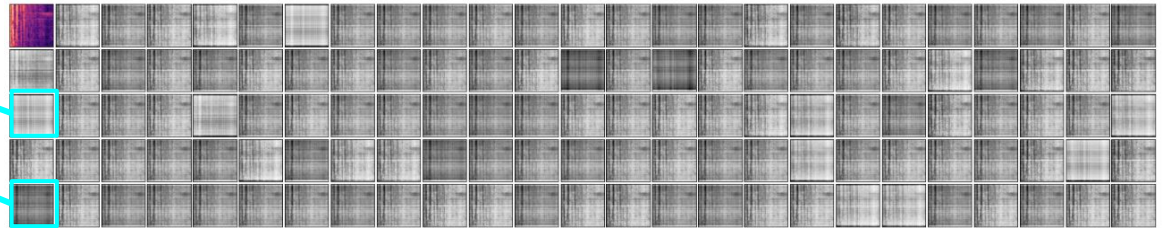


First cross attend





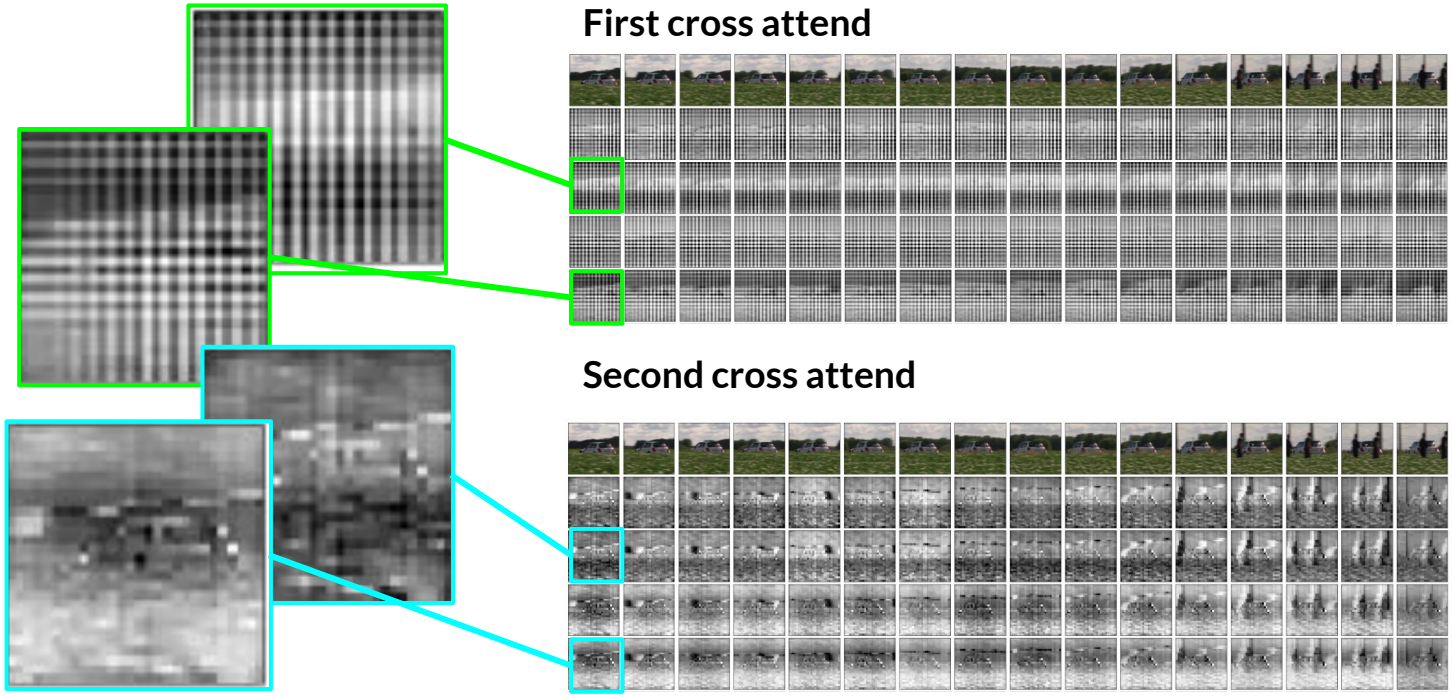
Second cross attend



Experiments: Attention maps trained on video only

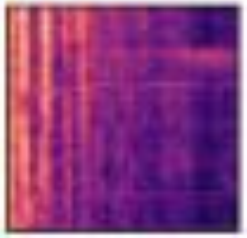
Video Input



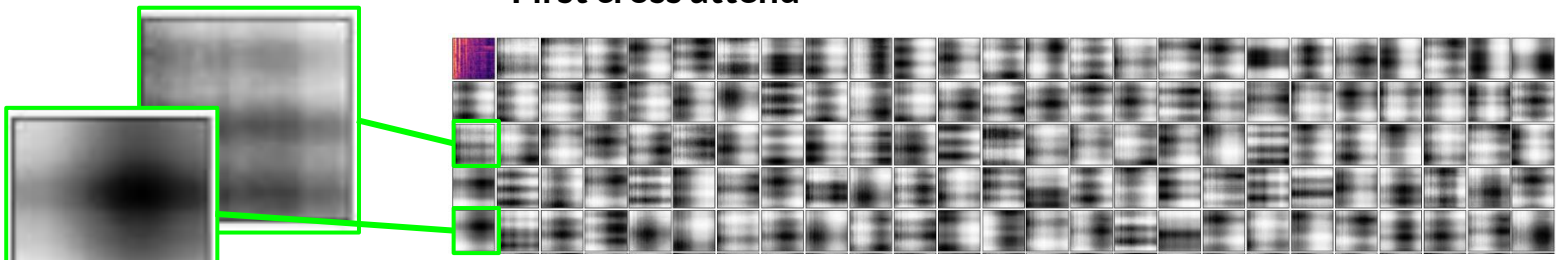


Experiments: Attention maps trained on video and audio

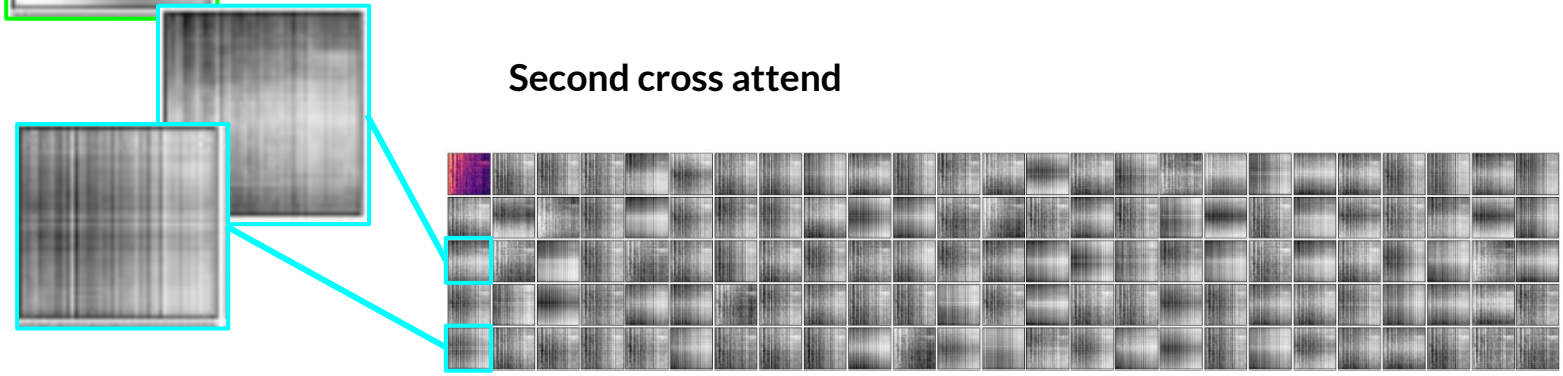
Spectrogram Input



First cross attend



Second cross attend

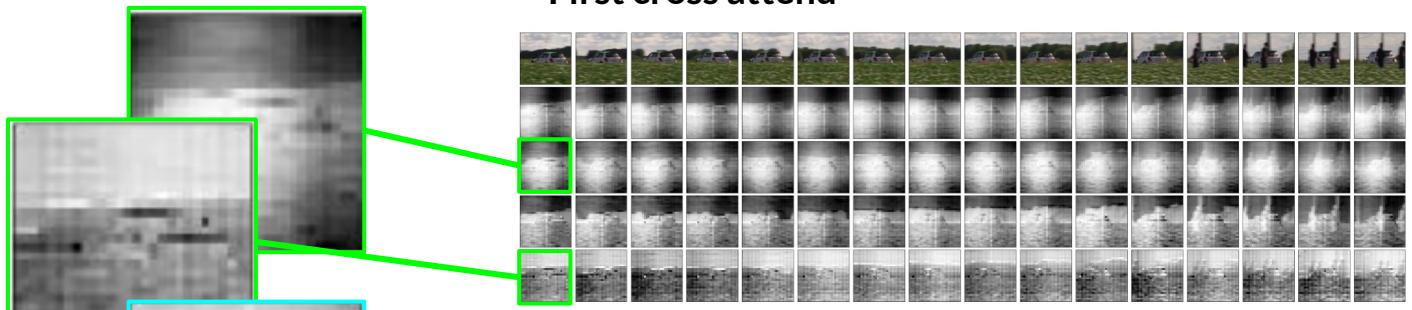


Experiments: Attention maps trained on video and audio

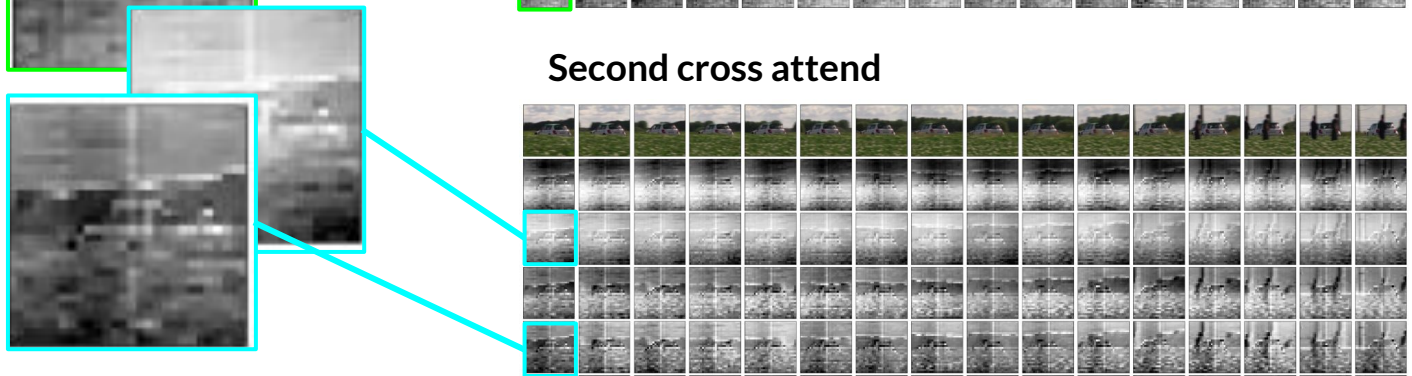
Video Input



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

Model	Accuracy(mAP)
PointNet++ (Qi et al., 2017)	**91.9**
Perceiver	85.7
Transformer (44x44)	82.1
ViT-B-2 (FF)	78.9
ViT-B-4 (FF)	73.4

Ablation Study

# 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

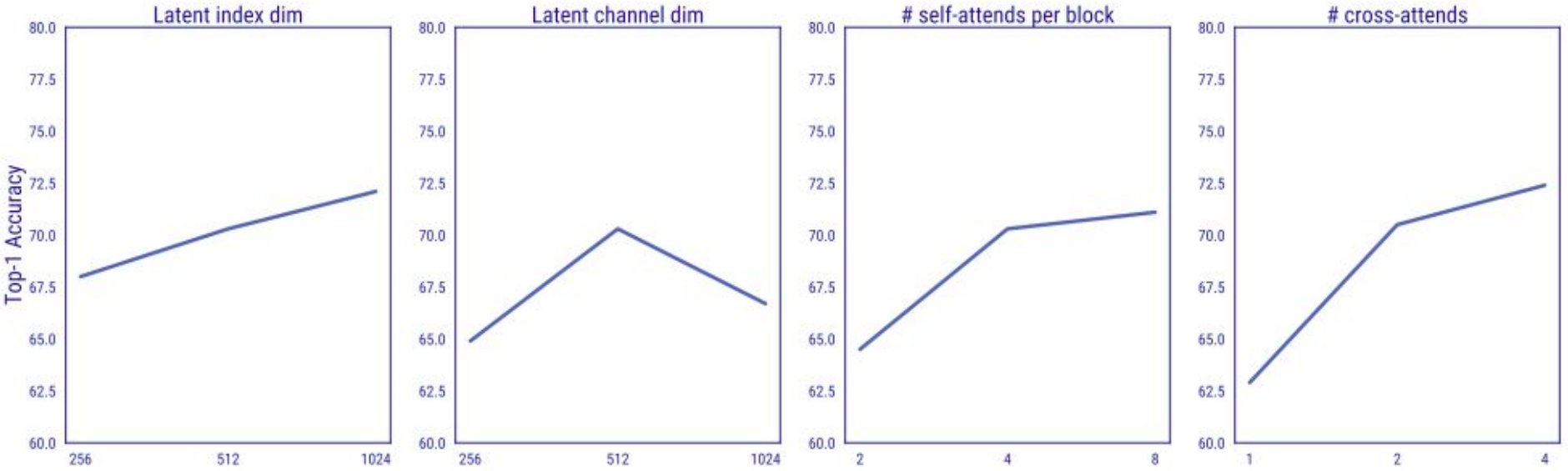
# cross-attends	Acc.	FLOPs	Params
4	39.4	173.1B	12.7M
8	45.3	346.1B	23.8M
12	OOM	519.2B	34.9M

Ablation Study

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

	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

Ablation Study



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

- Jaegel, et al. “Perceiver: General Perception with Iterative Attention”. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021
- You, Y., et al. “Large batch optimization for deep learning: Training bert in 76 minutes”. In Proceedings of International Conference on Learning Representations (ICLR), 2020.
- Gemmeke, J. F., et al. “Audio Set: An ontology and human-labeled dataset for audio events”. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.
- Deng, et al. “ImageNet A large-scale hierarchical image database”. IEEE 2009
- Wu, Z., et al. “3D shapenets: A deep representation for volumetric shapes”. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- Qi, C. R., et al. “PointNet++: Deep hierarchical feature learning on point sets in a metric space”. In Proceedings of Neural Information Processing Systems (NeurIPS), 2017.
- Kemker, Ronald, et al. "Measuring catastrophic forgetting in neural networks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1. 2018.
- https://gateway.newton.ac.uk/node/10214
- https://kazemnejad.com/blog/transformer_architecture_positional_encoding