# Group Convolutional Self-Attention for Roto-Translation Equivariance in ViTs

**Sheir A. Zaheer, Alexander C. Holston, Chan Y. Park**
**KC Machine Learning Lab, *Seoul, Rep. of Korea***

NEURAL INFORMATION PROCESSING SYSTEMS

## Roto-translation equivariant self-attention, w/o position encoding, using convolutional patch embedding and convolutional self-attention.

- G-CSA eliminates the need for relative position encoding
  - by preserving positional information like in CNNs
  - while retaining transformers' global context capture
- ViTs with G-CSA show superior performance to RPE-based approaches with significantly fewer parameters

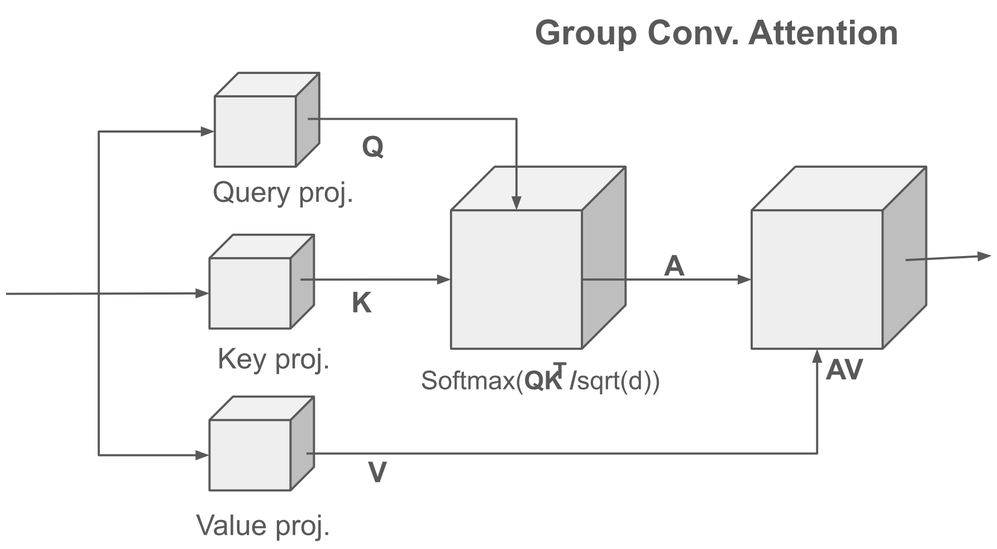
## Lifting Layer

The lifting layer takes an input signal $f : \mathbb{R}^2 \to \mathbb{R}^C$ (e.g., an image with $C$ channels ) and lifts it to a spatial location associated with multiple transformations under group $G$. We define the lifting over position $x$ and the discrete orientations $\theta$ associated with the discrete rotation group, $\mathbb{Z}_N$:

$$F(x, g) = [f * k](x, g) = \sum_c \sum_{x' \in \mathbb{Z}^2} f(x') k_c(g^{-1}(x' - x)), \qquad (1)$$

## Group Convolutional Self-Attention

*G-CSA exploits the inherent discrete translation equivariance of convolutions and doesn't need RPE.*




G-CSA with two dimensions for spatial projection and one along the group.

The *query*, *key*, and *value* mappings in this lifted space are computed using:

$$Q(x, g) = W_Q * F(x, g), \quad K(x, g) = W_K * F(x, g), \quad V(x, g) = W_V * F(x, g), \qquad (2)$$

where $F$ represents the feature at position $x$ and transformations $g$.

To implement G-CSA, we modify typical self-attention by incorporating group structure:

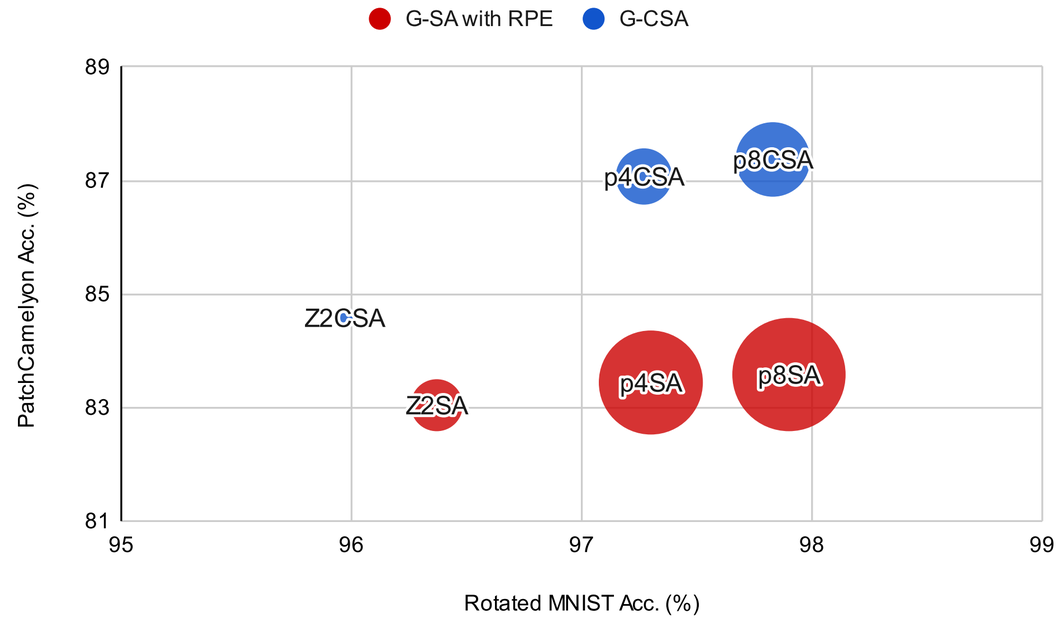
$$G\text{-}CSA(x, g) = \sum_{y \in \mathcal{N}(x)} \sum_{h \in G} A(x, g; y, h)\, V(y, h) \qquad (3)$$

The attention weights $A$ are calculated as:

$$A(x, g; y, h) = \frac{\exp\left(\frac{\langle Q(x,g), K(y,h) \rangle}{\sqrt{d}}\right)}{\sum_{y', h'} \exp\left(\frac{\langle Q(x,g), K(y',h') \rangle}{\sqrt{d}}\right)} \qquad (4)$$

This ensures that attention operates over both spatial and group dimensions while preserving translation equivariance via convolution.

## Performance and Complexity Comparison



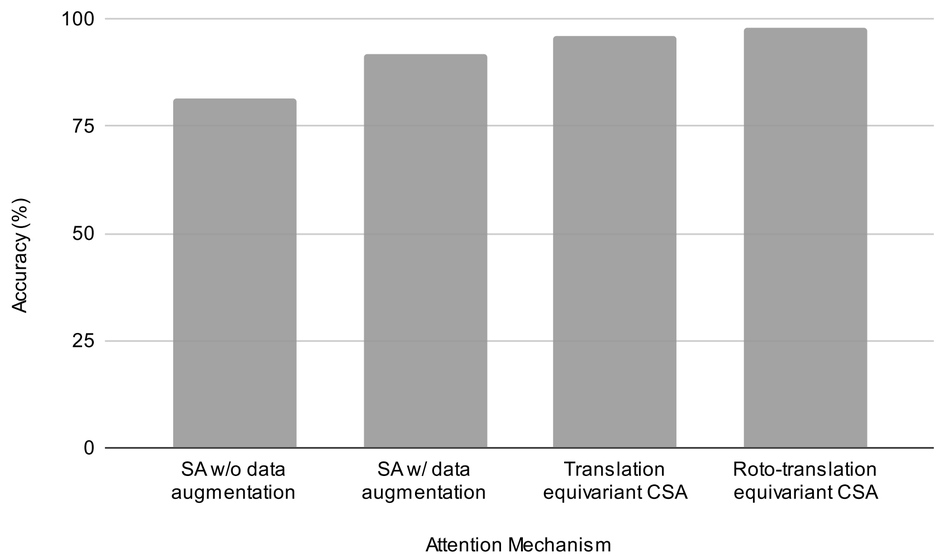

Accuracy comparison of on Rotated MNIST and PatchCamelyon. Sizes of the bubbles are proportional to model complexity (FLOPs)

Average inference runtime†

- *91* ms for G-CSA
- *144* ms for G-SA with RPE.

†RTX3090 GPU, batch of 32 '96 × 96' images

## Equivariance vs Augmentation




## Future Work

- Benchmarking against roto-translation equivariant CNNs.
- Extending G-CSA from discrete $SE(2, N)$-equivariance to full $E(2)$-equivariance.
- Scaling to larger ViTs for larger datasets, *e.g.* Imagenet.
- Application to downstream tasks like object detection and image segmentation.

## References

Alexey Dosovitskiy et al. (2021) "An image is worth 16x16 words: Transformers for image recognition at scale." International Conference on Learning Representations

Taco Cohen, Max Welling (2016),"Group equivariant convolutional networks." *International conference on machine learning*

David W. Romero, Jean-Baptiste Cordonnier (2021), "Group Equivariant Stand-Alone Self-Attention For Vision." *International Conference on Learning Representations*

Haiping Wu et al. (2021), "Cvt: Introducing convolutions to vision transformers." *IEEE/CVF international conference on computer vision*