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

Sheir A. Zaheer SHEIR@KC-ML2.COM
Alexander C. Holston ALEX@KC-ML2.COM
Chan Y. Park CHANYPARK@KC-ML2.COM
KC Machine Learning Lab, Seoul, Rep. of Korea

Editors: List of editors' names

Abstract

We propose discrete roto-translation group equivariant self-attention without position encoding using convolutional patch embedding and convolutional self-attention. We examine the challenges involved in achieving equivariance in vision transformers, and propose a simpler way to implement discretized roto-translation group equivariant vision transformers (ViTs). The experimental results demonstrate the competitive performance of our approach in comparison to the existing approaches for developing roto-translation equivariant ViTs.
Keywords: Roto-translation Equivariance, Group Covolutional Self-Attention, Equivariant transformers

1. Introduction

Equivariant neural networks preserve symmetry between input and output representations (Lim and Nelson, 2022; Wang et al., 2023; Guttenberg et al., 2016) by ensuring that all components transform predictably with input transformations. For instance, in rotation-equivariant models, rotating the input rotates all feature maps and the output accordingly (Bekkers et al., 2018; Cohen and Welling, 2016; Wiersma et al., 2020), a property crucial in molecular analysis (Yi et al., 2023; Liao et al., 2023), medical imaging (Marcos et al., 2017; Veeling et al., 2018), and robotics (Zhao et al., 2023, 2024). Such networks achieve equivariance through architectural choices like rotational convolutions, group convolutions, or equivariant pooling (Marcos et al., 2016; Cohen and Welling, 2016). In 3D tasks, preserving rotational symmetry benefits molecule modeling (Schütt et al., 2021), point-cloud orientation (Dym and Maron, 2020; Chen et al., 2021), and graph-based attention approaches (Liao and Smidt, 2022; Deng et al., 2021). While 2D images lack full 3D positional structure, they still retain orientation information relative to the projection axis, allowing 2D rotation-equivariant models to exploit this symmetry for robust and generalizable vision tasks (Han et al., 2021; Romero and Cordonnier, 2021).

Cohen and Welling (2016) introduced group CNNs to produce rotation-equivariant feature maps, a concept extendable to roto-translation equivariance by leveraging the inherent translation equivariance of CNNs (Romero et al., 2020; Bronstein et al., 2017).

In vision transformers, achieving roto-translation equivariance is challenging due to standard position encodings; however, incorporating equivariance preserving relative position encoding can enable group-equivariant self-attention (Romero and Cordonnier, 2021).

We propose group-equivariant convolutional self-attention (G-CSA) without position encoding for discrete roto-translation equivariance, using convolutional patch embedding

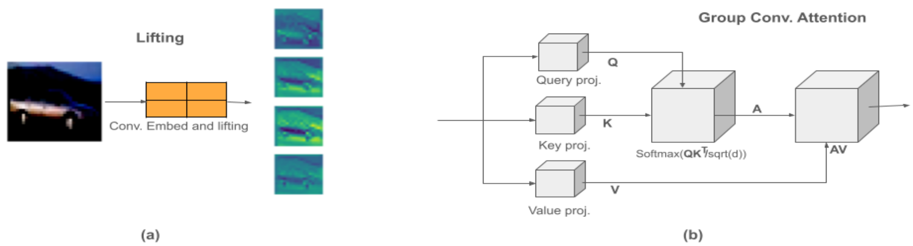

Figure 1: G-CSA transformer: (a) lifting layer for a roto-translation group with 4 elements, and (b) G-CSA with two dimensions for spatial projection and one along the group

and self-attention (Wu et al., 2021). This preserves positional information like in CNNs while retaining transformers' global context capture (Dosovitskiy et al., 2021), eliminating the need for relative position encoding. Experiments with G-CSA ViTs show superior performance to RPE-based approaches with significantly fewer parameters.

2. Background

2.1. Group Equivariance

Let $\Phi : V_1 \to V_2$ be a map between two spaces V_1 and V_2, and let ρ_1 and ρ_2 be actions of a group G on V_1 and V_2 respectively. Then, Φ is said to be G-equivariant if the following condition holds:

$$\Phi[\rho_1(g)f] = \rho_2(g)[\Phi[f]], \quad \forall g \in G,\ f \in V_1. \tag{1}$$

2.2. Position Encoding and Equivariance in Transformers

Position encoding influences both the equivariance properties and computational cost of self-attention networks. Absolute position encoding (Vaswani et al., 2017) assigns a unique vector to each position, causing the model to learn position-specific patterns and breaking equivariance to transformations such as translations or permutations. In contrast, relative position encoding (RPE) (Shaw et al., 2018) encodes position differences, thus preserving translation equivariance similarly to convolutional networks. This idea can be extended to group-equivariant vision transformers (Romero and Cordonnier, 2021) by incorporating rotation group encodings $G_{e(j)-e(i)}$ alongside horizontal and vertical RPE terms $P_{x(j)-x(i)}$ and $P_{y(j)-y(i)}$, yielding:

$$A := X_i W_Q((X_j + P_{x(j)-x(i)} + P_{y(j)-y(i)} + G_{e(j)-e(i)})W_K)^\top. \tag{2}$$

However, unlike absolute encodings, which are added once to the input, RPE must be computed in every attention step of every layer, leading to additional complexity.

3. ViT with G-CSA

We use the standard ViT architecture with a few modifications. In addition to removing position encoding, we use group convolutional self-attention (Figure 1(b)) and a lifting layer (Figure 1(a)) prior to multi-head attention.

3.1. 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 (Bekkers et al., 2018). Since this work deals with discrete roto-translations, we adapt this lifting operation for discrete rotation groups, where rotations belong to a finite set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ and the input function is $f : \mathbb{Z}^2 \to \mathbb{R}^C$ (an image with discrete pixels with C channels). We define the lifting over position x and the discrete orientations θ associated with the discrete rotation group, $\mathbb{Z}_N$:

$$F(x, g) = [f * k](x, g) = \sum_c \sum_{x' \in \mathbb{Z}^2} f(x') k_c(g^{-1}(x' - x)), \tag{3}$$

where $F : \mathbb{Z}^2 \times \mathbb{Z}_N \to \mathbb{R}^{C'}$ is the lifted feature map, defined over discrete positions and N discrete orientations, and $k_c(g^{-1}(x' - x))$ is a discrete rotation-aware convolutional kernel defined for each $\theta_k = \frac{2\pi k}{N}$, where $k \in \{0, \ldots, N-1\}$ (Figure. 1(a)).

3.2. Group Convolutional Self-attention

G-CSA (Figure. 1(b)) is a mapping from functions defined on an affine group $G = \mathbb{Z}^2 \rtimes \mathbb{Z}_N$ to functions on the same group G modified by the action of group elements. It operates on an input function $F : \mathbb{Z}^2 \times \mathbb{Z}_N \to \mathbb{R}^{C'}$ where F comes from the previous transformer block with G-CSA or the lifting layer (in the case of the first block).

In CVT (Wu et al., 2021), self-attention learns relationships between spatial locations on an input $f : \mathbb{R}^2 \to \mathbb{R}^C$. In G-CSA, we extend this concept to a lifted feature space, where each spatial location is associated with multiple transformations $g \in G$. The *query*, *key*, and *value* mappings in this lifted space are computed using:

$$Q(x, g) = W_Q * F(x, g), \quad K(x, g) = W_K * F(x, g), \quad V(x, g) = W_V * F(x, g), \tag{4}$$

where F represents the feature at position x and transformations g. W_Q, W_K, W_V are group-equivariant convolutional kernels and $*$ denotes convolution.

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

$$G\text{-}CSA(x, g) = \sum_{y \in \mathcal{N}(x)} \sum_{h \in G} A(x, g; y, h) V(y, h) \tag{5}$$

where $\mathcal{N}(x)$ denotes the local neighborhood of x defined by the receptive field of the convolution, and the attention weights A are calculated as:

$$A(x, g; y, h) = \frac{\exp\left(\frac{\langle Q(x,g), K(y,h)\rangle}{\sqrt{d}}\right)}{\sum_{y',h'} \exp\left(\frac{\langle Q(x,g), K(y',h')\rangle}{\sqrt{d}}\right)} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ represents the dot product. This ensures that attention operates over both spatial and group dimensions while preserving translation equivariance via convolution. Appendix A expands on the equivariance of G-CSA

Approach	Model Config	PatchCamelyon		Rotated MNIST			
		Acc. (%)	Params	Acc. (%)	Params	Mul-Add (M)	Total Size (MB)
SA with RPE Romero and Cordonnier (2021)	*Z2SA*	83.04	205.66K	96.37	44.67K	60.16	29.58
	p4SA	83.44		97.30		232.49	161.77
	p8SA	83.58		97.90		462.29	198.05
Ours (CSA without RPE)	*Z2CSA*	84.58	104.96K	95.97	33.35K	29.10	9.09
	p4CSA	87.07		97.27		116.37	35.84
	p8CSA	87.37		97.83		232.98	71.72

Table 1: Classification accuracy and parameters for each model. Model complexity is also provided in terms of total multiplication-addition operations (in millions) and the model memory size (in Megabytes) when trained with batch size of 16.

4. Experimental Results

We test the proposed group-equivariant convolutional self-attention (G-CSA) for vision transformers by implementing models for *2D Integer Translation* ($\mathbb{Z}^2$) group equivariance, and for *p4* and *p8* roto-translation group equivariance. In the following text, we refer to these models as *Z2CSA*, *p4CSA*, and *p8CSA*, respectively. We compare *G-CSA* models against corresponding models with group equivariant self-attention (G-SA) enriched with relative position encoding (Romero and Cordonnier, 2021).

Table 1 shows the performance comparisons of G-CSA against SA with RPE on rotated MNIST dataset (Larochelle et al., 2007) and PatchCamelyon dataset (RGB images of breast tissue labeled tumorous or non-tumorous) (Veeling et al., 2018). The results show that our models match the performance of the models where group equivariant attention needed to be enriched with RPE in every attention layer. The compared models had the same number of layers and expansions per layer. Table 1 also compares the memory required by the models along with the total number of multiplication and addition operations. The results show a significant reduction in the number of operations in the case of REViTs with G-CSA. Additionally, average inference runtimes on an RTX3090 GPU for a batch of 32 images were 91 ms for G-CSA models and 144 ms for SA with RPE.

5. Discussion and Future Works

Our results show that despite the simpler formulation of ViTs with G-CSA, we were able to achieve competitive results compared to typical group self-attention with RPE. In particular, our results on PatchCamelyon dataset show the effectiveness of our approach on larger image sizes for a real-world application that may benefit from roto-translation equivariant classification. G-CSA not only outperforms, but it also does so with a simpler architecture and smaller roto-translation group sizes.

We proposed G-CSA for roto-translation equivariant transformers. Though more rigorous testing is needed in the future, our results demonstrate that our approach compares well with the existing approaches for roto-translation equivariant image classification. In the future, we also plan to scale up our G-CSA based ViTs to more complex datasets with larger image resolutions, e.g. ImageNet (Deng et al., 2009). Models trained on such datasets also have the potential to be used as roto-translation equivariant backbones for downstream tasks like object detection and image segmentation.

References

Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis, 2018. URL https://arxiv.org/abs/1804.03393.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. ISSN 1558-0792. doi: 10.1109/msp.2017.2693418. URL http://dx.doi.org/10.1109/MSP.2017.2693418.

Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14514–14523, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12200–12209, October 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nadav Dym and Haggai Maron. On the universality of rotation equivariant point cloud networks. *arXiv preprint arXiv:2010.02449*, 2020.

Nicholas Guttenberg, Nathaniel Virgo, Olaf Witkowski, Hidetoshi Aoki, and Ryota Kanai. Permutation-equivariant neural networks applied to dynamics prediction, 2016. URL https://arxiv.org/abs/1612.04530.

Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2786–2795, 2021.

H. Larochelle, D. Erhan, Aaron C. Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *International Conference on Machine Learning*, 2007. URL https://api.semanticscholar.org/CorpusID:14805281.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

Lek-Heng Lim and Bradley J. Nelson. What is an equivariant neural network?, 2022. URL https://arxiv.org/abs/2205.07362.

Diego Marcos, Michele Volpi, and Devis Tuia. Learning rotation invariant convolutional filters for texture classification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, page 2012–2017. IEEE, December 2016. doi: 10.1109/icpr.2016.7899932. URL http://dx.doi.org/10.1109/ICPR.2016.7899932.

Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. doi: 10.1109/iccv.2017.540. URL http://dx.doi.org/10.1109/ICCV.2017.540.

David W. Romero and Jean-Baptiste Cordonnier. Group equivariant stand-alone self-attention for vision. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=JkfYjnOEo6M.

David W. Romero, Erik J. Bekkers, Jakub M. Tomczak, and Mark Hoogendoorn. Attentive group equivariant convolutional networks, 2020. URL https://arxiv.org/abs/2002.03830.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra, 2021. URL https://arxiv.org/abs/2102.03150.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology, 2018. URL https://arxiv.org/abs/1806.03962.

Dian Wang, Jung Yeon Park, Neel Sortur, Lawson L. S. Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry, 2023. URL https://arxiv.org/abs/2211.09231.

Ruben Wiersma, Elmar Eisemann, and Klaus Hildebrandt. Cnns on surfaces using rotation-equivariant features. *ACM Transactions on Graphics*, 39(4), August 2020. ISSN 1557-7368. doi: 10.1145/3386569.3392437. URL http://dx.doi.org/10.1145/3386569.3392437.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

Yiqiang Yi, Xu Wan, Yatao Bian, Le Ou-Yang, and Peilin Zhao. Etdock: A novel equivariant transformer for protein-ligand docking. *arXiv preprint arXiv:2310.08061*, 2023.

Linfeng Zhao, Xupeng Zhu, Lingzhi Kong, Robin Walters, and Lawson L. S. Wong. Integrating symmetry into differentiable planning with steerable convolutions, 2023. URL https://arxiv.org/abs/2206.03674.

Linfeng Zhao, Hongyu Li, Takn Padr, Huaizu Jiang, and Lawson L.S. Wong. E(2)-equivariant graph planning for navigation. *IEEE Robotics and Automation Letters*, 9 (4):3371–3378, April 2024. ISSN 2377-3774. doi: 10.1109/lra.2024.3360011. URL http://dx.doi.org/10.1109/LRA.2024.3360011.

Appendix A. Equivariance of G-CSA

Here, we show that G-CSA is group equivariant. Let T_g be a group transformation acting on a feature function $F : X \times G \to \mathbb{R}^d$. A transformation $g \in G$ acts as:

$$(T_g F)(x, h) = F(g^{-1}x, g^{-1}h) \tag{7}$$

where $g^{-1}x$ is undoing the transformation g on the spatial point x, and $g^{-1}h$ refers to inverting the transformation g before applying the transformation h.

Given the transformed feature T_gF, we compute the *query, key, and value* mappings from (4) using group-equivariant convolutions:

$$\begin{aligned} &Q_g(x, h) = W_Q * (T_g F)(x, h),\, K_g(x, h) = W_K * (T_g F)(x, h), \\ &V_g(x, h) = W_V * (T_g F)(x, h) \end{aligned} \tag{8}$$

Since these are implemented via equivariant convolutions:

$$Q_g(x, h) = Q(g^{-1}x, g^{-1}h), \quad K_g(x, h) = K(g^{-1}x, g^{-1}h), \quad V_g(x, h) = V(g^{-1}x, g^{-1}h) \tag{9}$$

Then, we compute the attention weights for the transformed feature using (6):

$$A_g(x, h; y, h') = \frac{\exp\left(\frac{\langle Q_g(x,h), K_g(y,h')\rangle}{\sqrt{d}}\right)}{\sum_{y',h''} \exp\left(\frac{\langle Q_g(x,h), K_g(y',h'')\rangle}{\sqrt{d}}\right)} \tag{10}$$

Using the equivariance of Q and K from (9), we substitute:

$$\langle Q_g(x, h), K_g(y, h')\rangle = \langle Q(g^{-1}x, g^{-1}h), K(g^{-1}y, g^{-1}h')\rangle. \tag{11}$$

Since the dot product is invariant to transformations applied to both vectors, we rewrite:

$$\langle Q(g^{-1}x, g^{-1}h), K(g^{-1}y, g^{-1}h')\rangle = \langle Q(x', h), K(y', h')\rangle \tag{12}$$

where $x' = g^{-1}x$ and $y' = g^{-1}y$. This implies:

$$A_g(x, h; y, h') = A(g^{-1}x, g^{-1}h; g^{-1}y, g^{-1}h') \tag{13}$$

Using (5), G-CSA output is:

$$G\text{-}CSA_G(x, h) = \sum_{y,h'} A_g(x, h; y, h')V_g(y, h') \tag{14}$$

Substituting the equivariance property of V_g from (9):

$$V_g(y, h') = V(g^{-1}y, g^{-1}h') \tag{15}$$

we obtain:

$$G\text{-}CSA_G(x, h) = \sum_{y,h'} A(g^{-1}x, g^{-1}h; g^{-1}y, g^{-1}h')V(g^{-1}y, g^{-1}h'). \tag{16}$$

Rewriting with $y' = g^{-1}y$, $h' = g^{-1}h'$:

$$G\text{-}CSA_G(x, h) = T_g\left(G\text{-}CSA(x, h)\right) \tag{17}$$

which shows *equivariance*:

$$G\text{-}CSA(T_gF) = T_g\left(G\text{-}CSA(F)\right) \tag{18}$$