Rotation Equivariant Convolutional Vision Transformer

Kartik Sachdev
RWTH Aachen University

Introduction

A **Vision Transformer architecture** is introduced to induce **rotational equivariance to a cyclic subgroup** C_4 in a transformer architecture. Achieved by combining:

- $E(2)$**-steerable convolutions** [6]
- **Convolutional Vision Transformers (CvT)** [7]

The architecture is trained in a supervised learning fashion on a **highly symmetrical dataset** of **simulated strong gravitational lensing images** for binary classification.

The model was trained with only **65k parameters** and achieves a test accuracy of over **97.1%**. Notably, the proposed architecture has approximately **one ninth of parameters** compared to CvT, while achieving a comparable test accuracy.

Problem Statement

Strong gravitational lensing is a phenomenon where the light of quasar or other distant, luminous object is bent and distorted by the gravity of massive foreground galaxy, which contain **dark matter**.

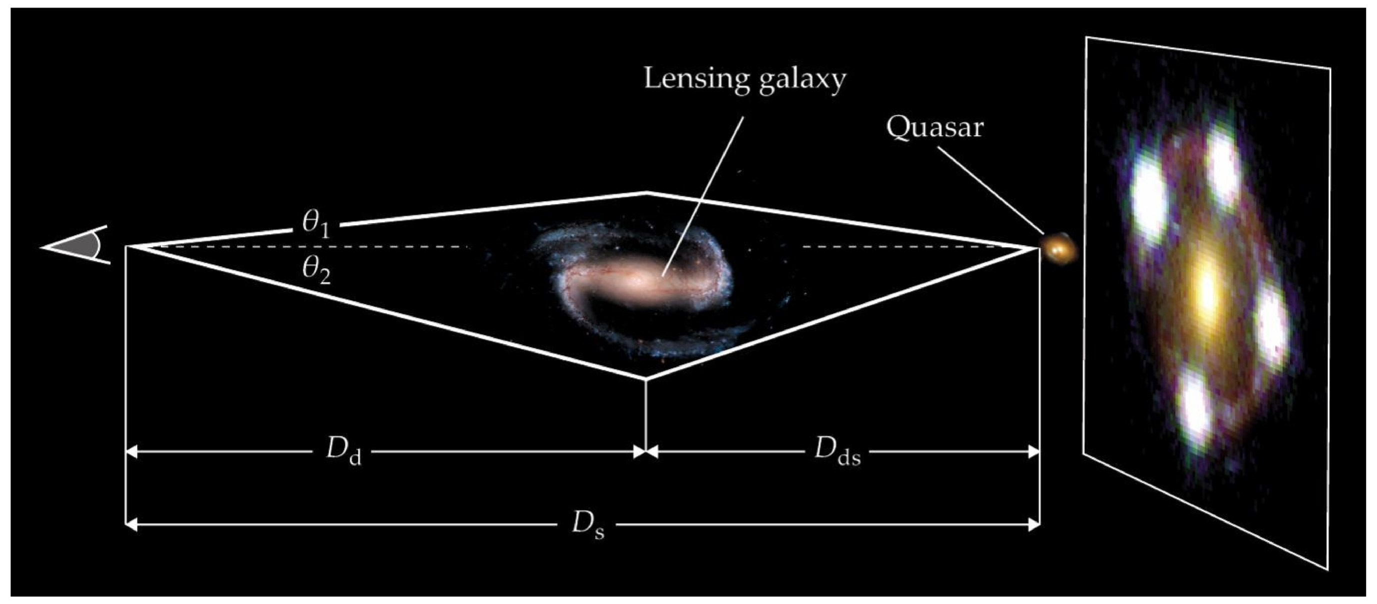

Figure 1. Strong gravitational lensing phenomenon [2]

- Strong gravitational lensing is an effective method for **detecting** the **substructures of dark matter halos**
- The substructure provides essential information for the identification of the true nature of dark matter [3]
- **Dataset** consists of **simulated strong lensing images** to determine if the image has substructure or not
- Dataset was generated using the package **PyAutoLens** and consists of 5000 images each of simulated strong lensing images with and without dark matter substructure

Figure 2. Actual Gravitational Lensing Image [1]

Implementation Details

A single convolution block consisting of modules - $E(2)$**-steerable convolution, batch-normalization and ReLU** is added before the **single stage of CvT**. Details:

- The input and output field types of convolution block are composed of **3 trivial and 10 regular representations** with rotational action of C_4
- The output of the convolution block is passed through 2 pooling layers namely, **anti-aliased channel-wise average pooling** and **group pooling**
- The output is then, fed to the single stage of CvT. The stage consists of **Convolutional Token Embedding layer** to generate tokens for the **Convolutional Transformer Blocks** which performs the **Multi-Head Self-Attention**
- Finally, an **MLP layer** is used for binary classification

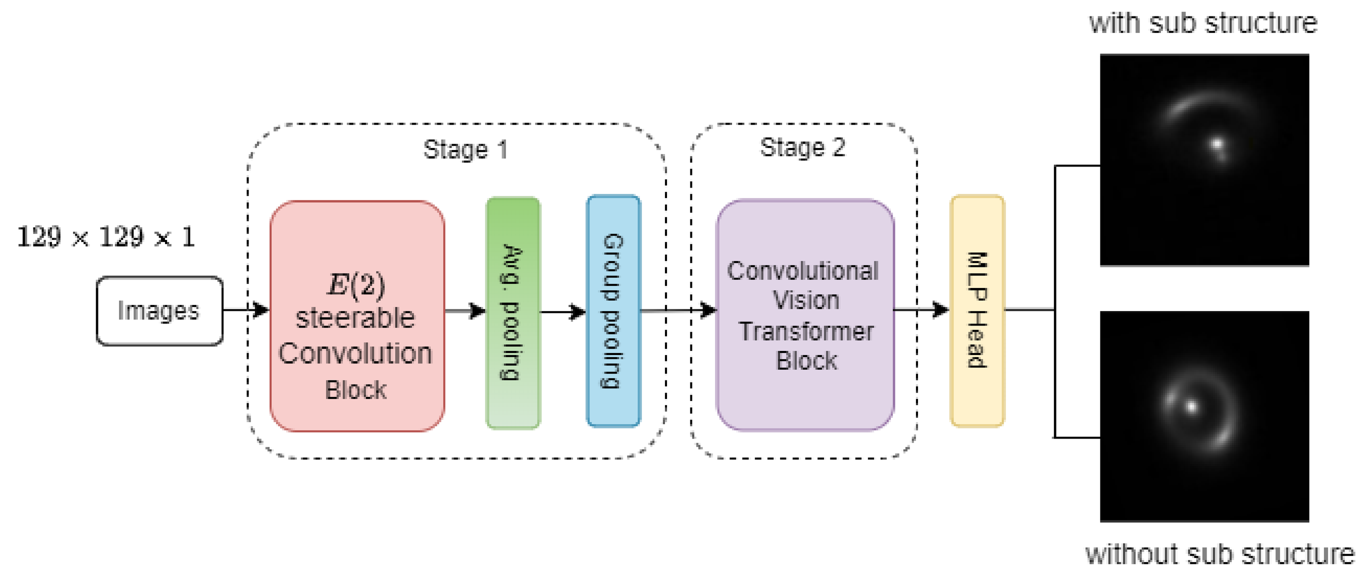

Figure 3. Simple illustration of the proposed model architecture

Intuition

To strategically place $E(2)$ **steerable convolutional layers before Convolutional Vision Transformer** block and **empirically verify** a hypothesis of making Vision Transformers equivariant under rotation. Thus, it would combine the **advantages** of both the types of networks:

Equivariant Network

Exploit the known inherent rotational symmetries present in the dataset

Convolutional Vision Transformer Network

Leverage its following advantages:

- Shift, scale invariance
- Global context
- Train on small dataset

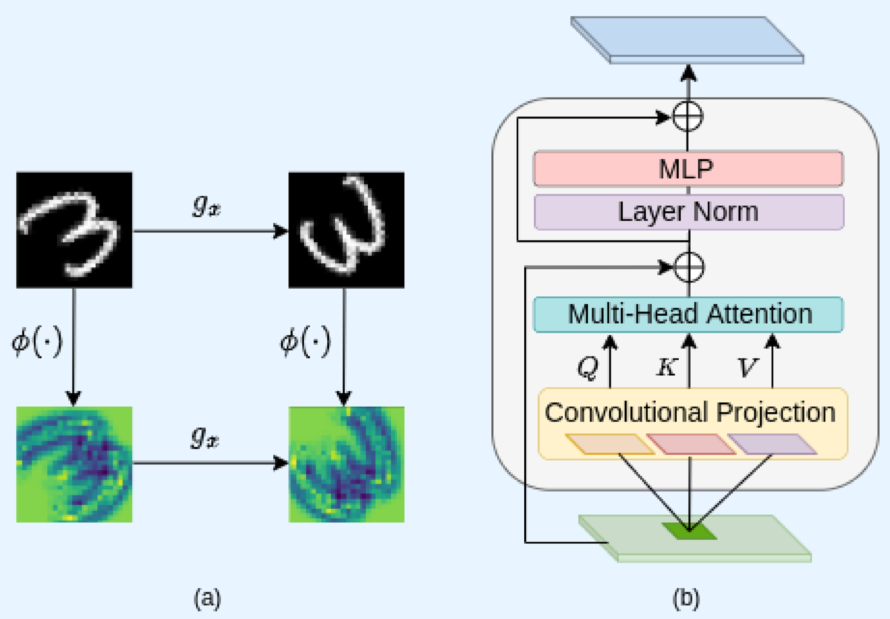

Figure 4. Illustration of (a) Rotation Equivariance [4] and (b) Single CvT block [7]

Architecture

	EqCvT	
	Layer Name	Details
Stage 1	Conv. Point Avg. Pool Group Pool	$\begin{bmatrix} 7 \times 7, \text{Stride}=1 \\ \text{Field Type}=\text{Regular} \\ \text{Fields}=10 \end{bmatrix} \times 1$
Stage 2	Conv. Embed.	$3 \times 3, 32, \text{Stride}=2$
	Conv. Proj. MHSA MLP	$\begin{bmatrix} 3 \times 3, 32 \\ \text{H}_1=3, \text{D}_1=32 \\ \text{R}_1=2 \end{bmatrix} \times 2$
Head	Linear	64
Params	65k	
Accuracy	97.1%	

Figure 5. Architecture of the proposed model

Results and Future Work

The proposed model obtains an **accuracy of 97.1%** that is 1% less than 2-stage CvT but with a significant **reduction of 86% in parameter count**. To check for the **rotational equivariance**, the **standard deviation of the output logits** were compared for an input with 8 rotations in multiples of $\pi/4$. **EqCvT** had a standard deviation of **0.48** while **CvT** had **0.96**, empirically showing **more robustness to rotation**.

This work is a part of an **ongoing project with Google Summer of Code (GSoC)** and provides **preliminary result**s for combining equivariant networks with transformers. **Future work** includes:

- Making the model **equivariant under all** $E(2)$ **isometries** of the image plane - **translations, rotations and reflections**
- Testing the architecture on **different** and **bigger datasets**
- Using **other versions of Vision Transformers**

Acknowledgments

I would like to thank Machine Learning for Science (ML4Sci) with the participating organizations University of Alabama, Brown University and BITS Pilani Hyderabad for providing the dataset.

References

[1] Gravitational lensing. `https://hubblesite.org/contents/articles/gravitational-lensing`. Accessed: 2022-06-17.

[2] Gravitational-lensing measurements push hubble-constant discrepancy past 5σ. `https://physicstoday.scitation.org/doi/10.1063/PT.3.4424`. Accessed: 2022-06-17.

[3] Stephon Alexander, Sergei Gleyzer, Evan McDonough, Michael W Toomey, and Emanuele Usai. Deep learning the morphology of dark matter substructure. *The Astrophysical Journal*, 893(1):15, 2020.

[4] Naman Khetan, Tushar Arora, Samee Ur Rehman, and Deepak K. Gupta. Implicit equivariance in convolutional networks. *CoRR*, abs/2111.14157, 2021.

[5] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[6] Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[7] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.