Skysong Innovations
Advancing the Arizona State University
Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M23-162P^

Published: 2/13/2024

## Inventors

Baoxin Li

Sachin Chhabra

## Contact

Physical Sciences Team

# Self-Supervised Technique for Training Vision Transformers

Convolution neural networks (CNNs) have made tremendous progress in various image processing fields like object recognition, segmentation, etc. This progress is primarily a result of supervised training on labeled data. The features learned by such a network are highly transferable and can be used for similar tasks. However, labeling a dataset is generally expensive and time-consuming. Self-supervised learning alleviates this problem by learning rich features without the need for manual annotation of labeling the dataset. Self-supervised learning techniques were designed to train CNNs. In recent years, Vision Transformers (ViTs) have surpassed CNNs. ViTs are considered data-hungry models as they outperform CNNs when huge labeled training data is available. Existing self-supervised techniques can be applied to ViTs but ViTs process images differently than CNNs. CNNs process the image like a grid and learn shared kernels to extract features whereas ViTs divide the images into patches and apply self-attention to the embeddings of the patches. There is a need for a self-supervised training technique for ViTs that considers how ViTs process images.

Researchers at Arizona State University have developed a self-supervised training technique for Vision Transformers (ViTs). The technique trains a ViT to predict the rotation of patches as well as that of the image. Given an image, the image or its patches are rotated at a time. The classification head of the ViT contains the global information of the image and the patch heads contain the local information. The classification head is used to predict the rotation angle of the image and the other heads are used to predict the rotation angle for their respective patches. This way, the training technique learns to extract both global and local features.

Related Publication: PatchRot: A Self-Supervised Technique for Training Vision Transformers

Potential Applications:

- For training Vision Transformers (ViTs)

Benefits and Advantages:

- Self-supervised technique for ViTs to predict rotation angles of images and image patch
- Easy-to-understand and implement
- Trains a ViT to predict rotation angles (0°, 90°, 180°, 270°) of images and image patches
- ViT pre-trained with this technique achieves superior results on downstream supervised learning

- Works for rotation invariant objects as well