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:M23-163P^
Published: 2/13/2024

Inventors

Baoxin Li

Sachin Chhabra

Contact

Physical Sciences Team

Regularization Technique for Training Vision Transformers

Transformers were originally designed for natural language processing but their application to other domains is rapidly gaining traction. In computer vision, convolution neural networks (CNNs) are the traditional choice of deep learning framework for image recognition tasks. Recently, Vision Transformers (ViTs) have been created a new benchmark by outperforming CNNs. Both CNNs and ViTs take image input and predict its label but differ in the way they process the images. CNNs process an image spatially like a grid and use kernels to extract features whereas ViTs divide the image into fixed size patches and use self-attention mechanism.

However, ViTs require a huge amount of labeled data (labeled training data) to outperform CNNs. This presents a major challenge when dealing with small datasets in that ViTs often overfit and result in poor generalization. To combat overfitting, many commonly used regularization solutions are incorporated into training CNNs as well as ViTs. Although, these solutions were originally designed for CNNs not ViTs. What is needed is a data augmentation technique specialized for ViTs to boost their performance.

Researchers at Arizona State University have developed a data augmentation technique that increases the amount of training data and thereby regularizes the performance of Vision Transformers (ViTs). The technique (1) prevents overfitting by regularizing the network by mixing images and labels; and (2) linearly interpolates an image consistently within the label space. The technique is an effective regularization technique for ViTs and outperforms state-of-the-art methods for datasets like CIFAR-10 and CIFAR-100. Experimentation has also shown the technique can be used with unlabeled data (extending to a semi-supervised learning setting).

Related Publication: PatchSwap: A Regularization Technique for Vision Transformers

Potential Applications:

- For training Vision Transformers (ViTs)

Benefits and Advantages:

- Effective in supervised and semi-supervised learning settings
- Ensures ViT can be trained effectively with small training dataset
- Regularizes ViTs by swapping image patches between two images to create a regularized input for training