MorphTransformer : A Progressive Morphological Transformer Approach for Visual Recognition Tasks

Peter Findley

Figure 1. (a) Overview of residual network using 3 × 3 convolution for localization. (b) transformer networks using multi-head self-attention mechanism. (c) MorphTransformer utilizing erosion-dilation based kernels for learning better representations of data along with self-attention mechanism.

Abstract

Vision transformer has demonstrated outstanding results in various vision tasks. The general vision transformer is permutation invariant making it robust for better generalization of the data. To provide local inductive bias and strong representation learning capabilities, in this research proposal, a morphological vision transformer model has been proposed using the potential of progressive kernel receptors with morphological operations. The multi-head self-attention mechanism will be replaced with progressive kernels based attention module in order to cover various frequency information. In practice, MorphTransformer provides a simple but yet powerful representation learning model leveraging potential of morphological operation in progressive receptive areas. Extensive experiments on multiple classification to justify the effectiveness of our proposed solution against state-of-the-art recognition models. Ablative analysis has been conducted to prove the robustness of MorphTransformer in various performance metrics.

1. Introduction

Convolutional Neural Networks (CNNs) have shown significant progress in visual recognition tasks including image classification, object detection. CNNs are better to capture local information, however, a robust network architecture for recognition tasks would need both long-term and short-term dependencies for effective feature learning process. Associated information among neighbourhood pixel is also necessary to perform accurate object recognition task. Residual based neural architecture have proved to learn long-range information providing better accuracy in terms of classification (He et al., 2020). It also uses skip connection for the gradient vanishing problem. However, ResNets can not focus on important parts of receptive regions, thus, attention-based model was introduced (Wu et al., 2019). To use local inductive bias with the robustness of attention-aware representation learning, residual-based attention network has been featured in many studies (Wang et al., 2017; Shi et al., 2018). Self-attention module has become popular in many tasks for context-based addressing mechanism leveraging long-term memory information (Pandey & Wang, 2021). However, these methods do not focus on neighbourhood information.

2. Related Work

Before the Vision Transformer was suggested by Dosoviskiy *et al.* (Dosovitskiy et al., 2020) CNNs dominated the field of visual recognition. CNNs have the strong and effective capacity to extract the feature from the image based on the weight sharing, scale separation, and shift equivariant (Goodfellow et al., 2016). The Transformer network's unique structure endows its permutation equivariance with the ability to obtain inductive bias, even though it didn't exhibit the same strong equivariance representation as CNNs (Dwivedi & Bresson, 2020).

Token Embedding, Positional Embedding, Transformer Encoder, and Classification Head are the specific components of the conventional Visual Transformer's structure (Khan et al., 2022). The research of the Transformer network led to the development of a number of Transformer-based models that can effectively handle vision-related tasks. In order to solve the limitation of ViT, which was that it could only demonstrate excellent performance in large-scale datasets, Touvron *et al.* (Touvron et al., 2021) introduced the DeiT, which makes advantage of distillation learning. The CPVT (Chu et al., 2021) model used various position embedding techniques to boost the ViT model's effectiveness and adaptability.

The CvT (Wu et al., 2021) and CeiT (Yuan et al., 2021) model states that they attempt to combine the Transformer and CNN network in order to extract useful characteristics from each. The Swin Transformer was proposed by Liu *et al.* (Liu et al., 2021) and uses a number of techniques for visual tasks, including patch division, linear embedding, pyramid structure, and window-based MSA.

2.1. Motivation

Transformer-based architecture uses multi-layer perceptron for generalization of image data. These networks are efficient at accuracy metrics for large scale data classification task. However, from the discriminative feature learning perspective, these methods do not work well on for regional neighbourhood information processing. On the other hand, CNN-based models have strong representation learning capabilities. To focus on neighbourhood pixel information as well as the robust generalization of transformer architecture, MorphTransformer has been introduced as a general solution.

2.2. Contribution

The contribution of this research proposal is three folded.

- This research proposal introduces a novel morphological operation-driven transformer model for strong localization and generalization of image data.
- The proposed method utilizes different morphological kernels with convolution kernels to cover maximum receptive locations in order to improve the overall feature learning capabilities.
- To the best of our knowledge, this is the first approach that combines transformer network with morphological operation-based kernels.

2.3. Background Studies

The area of mathematical morphology encompasses the image processing operators. These operators are employed in a variety of processes, including edge detection, noise reduction, visual augmentation, and image segmentation. The kernel is a structuring element used by the morphological operators to examine an images. Positioning a little object at every point in the image and comparing it to the surrounding pixels constitutes many procedures. Erosion, dilation, closing, and opening are the primary morphological processes.

The minimum value among all the pixels in its neighborhood is assigned for the pixel in the degraded image. The foreground pixel region boundaries on the image are eroded by the erosion procedure. It also distorts the forms of the items in the image.

Erosion enlarges gaps and spaces between various regions while removing details. The dilation technique increases each pixel in a given area by the maximum number of pixels. The foreground pixels areas' borders are enlarged on the image. By enlarging the bright regions and contracting the dark ones, dilatation increases the size of the items in the image.

Opening is characterized by erosion and then dilatation Minute object protuberances, noise, and small dark fractures are all removed from the image during this operation. Since it tends to remove some of the foreground pixels from the margins of the foreground pixel regions, it has the same effect as erosion. However, it causes less damage than general erosion.

3. Methodology

When it comes to accurately capturing the shape and size of objects in a picture, morphological processes are particularly effective. This study proposes a deep network based on two fundamental morphological procedures. Erosion and morphological dilatation processes are specifically taken into account. In particular, morphological dilation and erosion operations are considered. Let $X \in R^{M \times N \times C}$ be an intermediate feature map extracted from the HSI data, with spatial size $M \times N$ and C channels. The dilation $\oplus$ and erosion $\ominus$ operations over the feature map centered at spatial location (i, j) can be defined as follows:

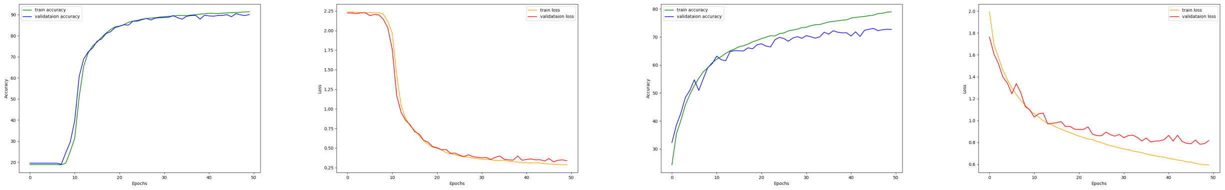
$$(\mathbf{X} \oplus \mathbf{S}^d)(i,j) = \max_{(\hat{i},\hat{j},\hat{k})\in U} (\mathbf{X}_{i+\hat{i},j+\hat{j},\hat{k}} + S^d_{\hat{i},\hat{j},\hat{k}}) \quad (1)$$

$$(\mathbf{X} \ominus \mathbf{S}^e)(i,j) = \min_{(\hat{i},\hat{j},\hat{k})\in U} (\mathbf{X}_{i+\hat{i},j+\hat{j},\hat{k}} - \mathbf{S}^e_{\hat{i},\hat{j},\hat{k}}) \quad (2)$$

where the feature maps of X_1 are combined linearly in order to generate X_2. The linear combination can be viewed as a 1×1 convolution. To generate additional features, we may apply multiple dilation/erosion operations (and generate multiple linear combinations of dilation and erosion). The figure 1 depicts the overall architecture of our proposed MorphTransformer model.

We have taken into account B/4 dilations and B/4 erosions in each block of our studies. Additionally, as dilation and erosion are dependent on min/max operations, several zero gradient values may be produced while performing the backpropagation step. We will employ a convolution layer to precisely improve the intended output of each operation in order to increase the gradient.

The input/output dimension is tentatively selected to 2048. However, this may be changed according to the future set up. Each block is subjected to concatenation in order to avoid feature loss due to morphological operations. We will



(a)

(b)

(c)

(d)

Figure 2. Training and validation accuracy and loss on SVHN dataset, (b) test accuracy and loss on CIFAR10 dataset

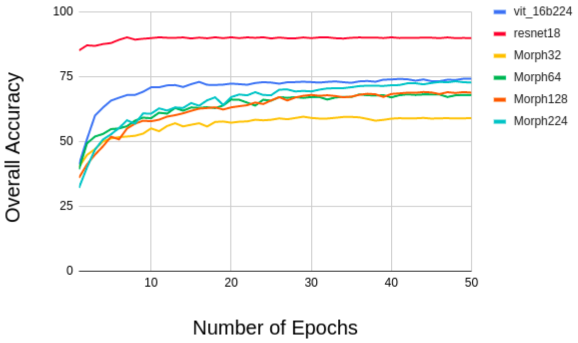


Figure 3. Comparative analysis of MorphTransformer against various deep learning solutions for CIFAR10 datasets.

use sparse categorical crossentropy with AdamW optimizer to handle the zero gradient caused by the morphological operations.

4. Comparative Discussions

To compare our results with state-of-the-art deep learning solutions, we have adopted vision transformer and residual networks. We have used Cifar10 and SVHN to demonstrate the superiority of our results. From the figure 5, it can be noticed that MORPHTRANSFORMER achieved equivalent results compared to vision transformer model. ResNet demonstrated good results due to local inductive bias.

From the figure 4, we can notice that we have more VRAM than other deep learning settings due to morphological operation in the transformer. This also significantly increased the training time.

4.1. Experimental Settings

For the experiment, pytorch framework will be used with Nvidia GPU 3090 TX, Intel Xeon Processor with 256 GB RAM. The experiment will be repeated multiple times in order to ensure reproducibility. We have experimented with different images sizes including 32, 64, 128 and 224. We have found the best results for 224 × 224 images. Images at 32x32 were too small and had declining accuracy after enough epochs, declining from 60 to 55, while the others increased We have adopted Adam optimizer with categorical cross-entropy loss function. The learning rate has been set to 0.0005. Stdev of weights was set to 1

5. Concluding Remarks

In this research project, a novel morphological operation-based transformer architecture has been proposed for strong generalization and representation learning from data. The featured model is be able to extract robust discriminative features from image data improving the image classification performance. The proposed MORPHTRANSFORMER model showed better generalization on various datasets including Cifar10 and SVHN. The main goal of this project is to build a robust network that can perform generalization and localization from image data in a sufficient way. Our future work is to improve the classification result and focus on attention-free transformer.

References

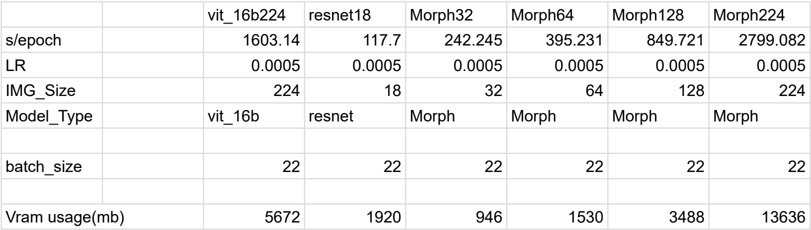
Chu, X., Zhang, B., Tian, Z., Wei, X., and Xia, H. Do we really need explicit position encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 3(8), 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dwivedi, V. P. and Bresson, X. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

He, F., Liu, T., and Tao, D. Why resnet works? residuals generalize. *IEEE transactions on neural networks and learning systems*, 31(12):5349–5362, 2020.



		vit_16b224	resnet18	Morph32	Morph64	Morph128	Morph224
s/epoch		1603.14	117.7	242.245	395.231	849.721	2799.082
LR		0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
IMG_Size		224	18	32	64	128	224
Model_Type		vit_16b	resnet	Morph	Morph	Morph	Morph
batch_size		22	22	22	22	22	22
Vram usage(mb)		5672	1920	946	1530	3488	13636

Figure 4. Experimental settings for MorphTransformer

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Pandey, A. and Wang, D. Dense cnn with self-attention for time-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1270–1279, 2021.

Shi, Z., Chen, C., Xiong, Z., Liu, D., Zha, Z.-J., and Wu, F. Deep residual attention network for spectral image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

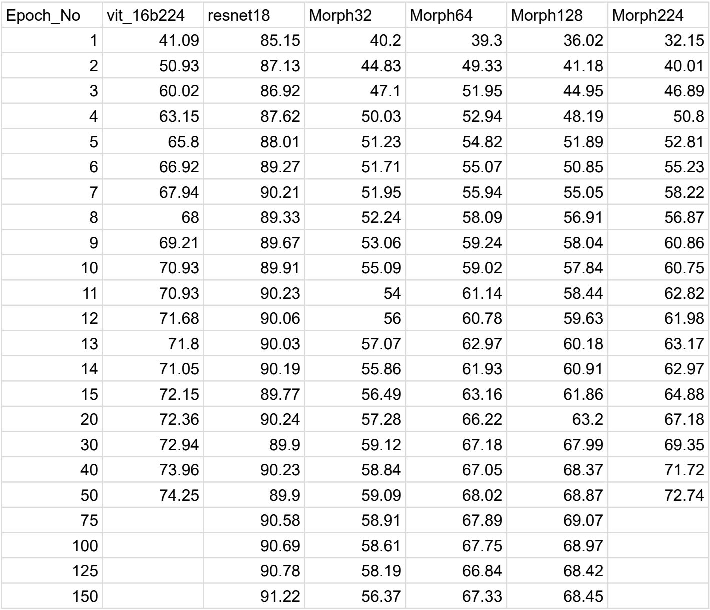
Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2017.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., and Zhang, L. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.

Wu, Z., Shen, C., and Van Den Hengel, A. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., and Wu, W. Incorporating convolution designs into visual transformers.



Epoch_No	vit_16b224	resnet18	Morph32	Morph64	Morph128	Morph224
1	41.09	85.15	40.2	39.3	36.02	32.15
2	50.93	87.13	44.83	49.33	41.18	40.01
3	60.02	86.92	47.1	51.95	44.95	46.89
4	63.15	87.62	50.03	52.94	48.19	50.8
5	65.8	88.01	51.23	54.82	51.89	52.81
6	66.92	89.27	51.71	55.07	50.85	55.23
7	67.94	90.21	51.95	55.94	55.05	58.22
8	68	89.33	52.24	58.09	56.91	56.87
9	69.21	89.67	53.06	59.24	58.04	60.86
10	70.93	89.91	55.09	59.02	57.84	60.75
11	70.93	90.23	54	61.14	58.44	62.82
12	71.68	90.06	56	60.78	59.63	61.98
13	71.8	90.03	57.07	62.97	60.18	63.17
14	71.05	90.19	55.86	61.93	60.91	62.97
15	72.15	89.77	56.49	63.16	61.86	64.88
20	72.36	90.24	57.28	66.22	63.2	67.18
30	72.94	89.9	59.12	67.18	67.99	69.35
40	73.96	90.23	58.84	67.05	68.37	71.72
50	74.25	89.9	59.09	68.02	68.87	72.74
75		90.58	58.91	67.89	69.07	
100		90.69	58.61	67.75	68.97	
125		90.78	58.19	66.84	68.42	
150		91.22	56.37	67.33	68.45	

Figure 5. End-to-end training results using our method and other state-of-the-art deep learning models

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 579–588, 2021.