Leveraging Locality Inductive Bias for Automated Medical Imaging Diagnosis

Jungmin Ha[1], Jinkyu Kim[2], Jaekoo Lee[1]
[1]College of Computer Science, Kookmin University
[2]Department of Computer Science and Engineering, Korea University
Contact: jinkyukim@korea.ac.kr, jaekoo@kookmin.ac.kr

Introduction

1) Importance of diagnosis skin cancer

- Recently, the incidence of skin cancer has been increasing globally

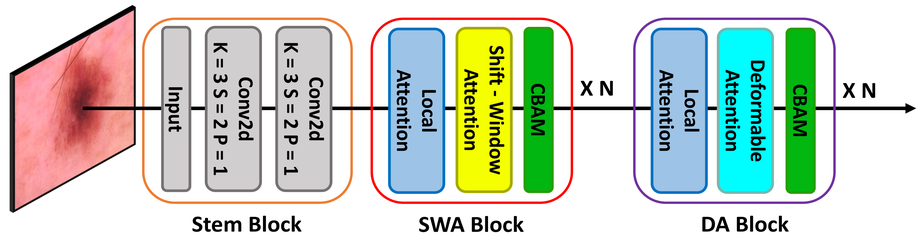
2) Insufficient medical data

- Until recently, CNN has been overwhelmingly prevalent
- Vision Transformer necessitates extensive amounts of data for training

3) Vision Transformer with a lack of inductive bias

- Slowing training convergence speed

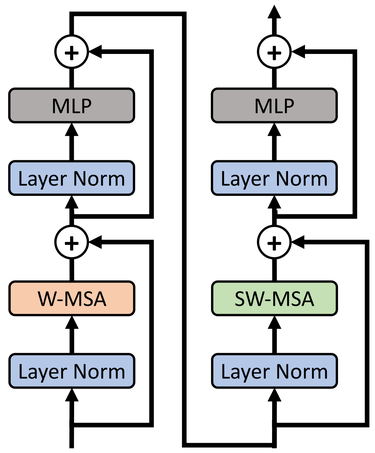
Model Overview



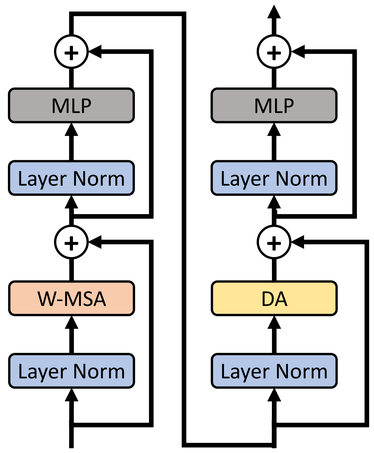
- **Vision Transformer based model**
- **To address locality, we added SWA blocks and DA blocks.**
- **To enhance clarity, CBAM separates object characteristics and location information.**

Transformer Blocks

- **SWA Block [2]**



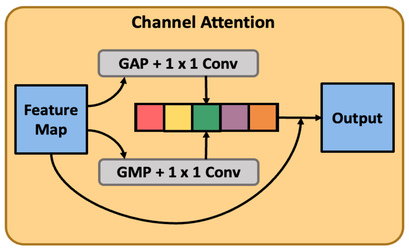
- **DA Block**



W-MSA	**Extracts critical information within the window.**
SW-MSA	**Extracts important information between windows.**
DA	**Extracts important information from the overall feature map.**

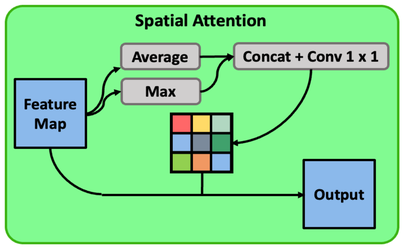
CBAM

- **Channel Attention [3]**



- Extracts vital channel information.

- **Spatial Attention [3]**



- Extracts vital Spatial information.

Experiments

Classification performance comparison on HAM10000 dataset [1]

Networks	Resolution	Params(M)	MACs(G)	Accuracy(↑)
ResNet50	224×224	25.6	4.1	97.34 (97.71)
GoogLeNet	224×224	13.0	1.5	93.96 (96.38)
Inception V3	299×299	27.2	2.9	97.34 (95.17)
MobileNet V3	224×224	5.5	0.2	97.10 (97.34)
FixCaps	299×299	0.8	1.4	96.14 (96.62)
ViT-B/32	224×224	88.2	4.4	96.01 (95.53)
Swin-B	224×224	87.8	10.2	95.41 (93.60)
Ours	224×224	59.7	15.8	**98.19 (97.83)**

- Scores in parenthesis represent results with the preprocessing
- Applying black hat transform to preprocessing method

Example of preprocessing image

Conclusion & Discussion

- **Achieved the highest performance on HAM10000**
- **Limited to simple classification task based on images**
- **Further research is needed to validate in diverse dense prediction tasks**

References

1. schandl, Philipp and Rosendahl, Cliff and Kittler, Harald: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5(1), pp. 1-9 (2018)

2. Liu, Ze and Lin, Yutong and Cao, Yue and Hu, Han and Wei, Yixuan and Zhang, Zheng and Lin, Stephen and Guo, Baining: Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012-10022. (2021).

3. Woo, Sanghyun and Park, Jongchan and Lee, Joon-Young and Kweon, In So: Cbam: Convolutional block attention module, Proceedings of the European conference on computer vision (ECCV), pp. 3-19. (2018).