Leveraging CNN and Vision Transformer with Transfer Learning to Diagnose Pigmented Skin Lesions

Jujie Yang *

Department of Computer Science, Durham University, Durham, UK

* Corresponding Author Email: jujie.yang@durham.ac.uk

Abstract. The transformer has emerged in computer vision, but whether it can rival or even replace the status of a convolutional neural network (CNN) still lacks deeper cognition. Therefore, this paper will study the feasibility and effectiveness of transformers in identifying the types of pigmented skin lesions by comparing one of the most widely known CNN architectures with three state-of-the-art Vision Transformer (ViT) models. The research results found that the transformer surprisingly outperforms the CNN. In other words, the overall accuracy and F1-score per class of its classifier are much higher than those of the CNN. In addition, the experimental data results show that if the ViT model has more training data input, its performance seems to have a large room for improvement. Conversely, the CNN model shows no indication that increasing training data will affect its performance. Therefore, if there are more dermoscopy images with corresponding label data to make the dataset larger and balanced, the transformer may completely replace CNN on this task.

Keywords: Transformer; ResNet; Dermoscopy Images; Skin Lesions.

1. Introduction

In recent years, skin cancer has been one of the most common cancers worldwide. Skin cancer often develops from early pigmented lesions. This usually requires many dermatologists to diagnose whether the pigmented lesions are benign or cancerous pathologically. If dermoscopy images can assist dermatologists in understanding the imaging features of skin cancer and quickly gaining experience in differential diagnosis, it will not only improve the efficiency and accuracy of diagnosis but also save a lot of human resources.

Dermoscopy pictures can also be employed to automatically train artificial neural networks to recognize multiple kinds of pigmented skin lesions. Even though a convolutional neural network (CNN) currently seems to be the most famous architecture in deep learning-based image recognition, it still has a less relaxed inductive bias and a smaller receptive field than a transformer. This paper will investigate three state-of-the-art Vision Transformer (ViT) models by adopting transfer learning (TL) and performing corresponding fine-tuning processing. A large-size dataset that will be tested is HAM10000 [3], which collects dermoscopic images of different populations and the corresponding labels. The labels are skin diseases marked by various medical diagnostic methods. The performance of the ViT will be compared with a customized CNN model.

2. Related Work

Convolutional Neural Networks seem to be the mainstream or almost the only algorithm in computer vision for a long time. Since 2020, the transformer has been confirmed to have the potential to perform as much as CNN on image classification tasks. As a result, the attention mechanism is not limited to natural language processing but can also be applied to computer vision [2].

In the field of diagnostic applications of pigmented skin cancer, in the past three years, most researchers still seem to be exploring how CNN can be used to predict better the type of lesions. For instance, they employed various pre-trained deep models, such as VGG19 [3], the Darknet-53 and NasNet-mobile [4], InceptionResNetV2 [5], and also experimented with fine-tuned deep CNN with dedicated preprocessing [6]. Even though there are still a small number of studies that have tried the vision transformer, calculating the accuracy as the only evaluation metric to predict the effect of the

disease cannot fully demonstrate the effect of the transformer [7]. Therefore, the experiments in this paper will assess one more metric, F1-score, to comprehensively compare the performance of two major categories of neural network architectures, CNN and ViT, taking advantage of TL techniques to predict pigmented skin diseases.

3. Method

Two main categories of models are used for dermoscopy image classification, namely CNN and ViT. For the former, this experiment uses a self-defined residual neural network, the ResNet9 Model. For the latter, this experiment employed three types of ViT, namely DeiT [8], Swin Transformer [9], and CrossViT [10], all of which are cutting-edge derivative versions of the fundamental ViT structure [2]. As shown in Figure 1, the Resnet9 model consists of three stages. Each of the first two stages contains two convolutional blocks followed by an identity block, which includes two convolutional layers. And the final stage has max pooling, flattening, and dropout operations followed by a dense layer reducing its input to the number of classes. Three variant models are on top of the first proposed transformer-like architecture employed in image classification. As shown in Figure2, the foundational ViT works in roughly three stages. Stage1 is to split images into patches and flatten them. Stage2 is to produce embedded patches utilizing linear embeddings and positional embeddings. Stage3 is ultimately getting predicted classes by feeding embedded patches into a transformer encoder through an extra multi-layer perceptron (MLP) linear layer.



Fig 1. ResNet9



Fig 2. ViT

Three ViT models investigated in this paper made more or fewer improvements based on the fundamental ViT structure. DeiT has an additional distillation token than ViT. Swin Transformer introduces two new critical concepts: hierarchical feature maps and shifted window attention. CrossViT leverages separate dual branches to extract multi-scale features and a cross-attention module [10] to merge branches. DeiT and Swin Transformer use the transfer learning method to load their respective tiny pre-training models, namely deit_tiny_patch16_224 and swin_tiny _ patch4 _ window7_224. In contrast, CrossViT uses the official defined model.

In preprocessing the loaded data, all models perform data augmentation on the trained data, including randomly flipping, changing the image's brightness, contrast, saturation, and hue, and randomly erasing small patches of pixels, followed by normalizing.

In the model training process, CNN and ViT use their respective two sets of optimizers, schedulers, and criteria. ResNet9 uses an Adam optimizer for gradient descent, a one-cycle learning rate (LR) scheduler, and a cross-entropy loss function. In contrast, the three ViT models use an AdamW optimizer, a cosine LR scheduler, and a label smoothing cross-entropy loss function, for which the label smoothing is widely used in advanced image multiclass classification models to improve the ability of generalization and model calibration.

Finally, this research will apply the F1-score measurement in addition to overall accuracy to assess each classifier's quality because the HAM10000 dataset is unbalanced. F1-score is usually leveraged to evaluate classifiers when the dataset is unbalanced, or the prediction errors are expensive. The skin cancer classifier in this paper is appropriate for assessing both accuracy and F1-score.

4. Experiment

The experiment dataset, HAM10000, has a total of 10,015 dermoscopy images of seven types of pigmented skin lesions. The seven types are Melanocytic nevi (nv), Melanoma (mel), Benign keratosis-like lesions (bkl), Basal cell carcinoma (bcc), Actinic keratoses (akiec), Vascular lesions (vasc), and Dermatofibroma (df). Since the training of transformers tends to crave more training data than the training of CNNs, three sets of comparative experiments will be designed to explore how transformers perform on dermoscopic images. Each set of experiments compares the three ViT models with the customized ResNet9 model. The difference between the three sets of experiments is the division of the whole data set. In other words, the size of the test set and the training set are different. The size ratios of the test and training sets for the three groups are approximately 7:3, 8:2 and 9:1, respectively. For example, a percentage of nearly 7:3 means that out of 10,015 total data, there are 3,000 test data, and all the remaining data (7,015) are training data for the model. For the remaining two groups of experiments, the proportions of the data were divided in the same way.

Three ViT models receive square inputs with a size of 224×224 after the same operations are performed on the input dermoscopy images of the ResNet9, which are resized and center cropped into a square with a side length of 32. During the training phase of ResNet9, it clips gradients in the range [-0.1, 0.1] and sets up a customized Adam optimizer with weight decay (WD) and a one-cycle LR scheduler. The optimizer sets 0.0004 and 0.0005 WD. The scheduler corresponding to these two values has tried four different upper LR boundaries in order, 0.01, 0.005, 0.001, and 0.00001, in the cycle for each parameter group to achieve a convergent accuracy. In comparison, during the training step of three ViT models, it sets up an AdamW optimizer with 5e-4 LR and 0.05 WD and a cosine annealing scheduler with a minimum LR of 5e-6.

Table 1. Overall accuracy and F1-score per class for each model with different dataset partitions

Model	Accuracy (%)	F1-score for akiec	F1-score for bcc	F1-score for bkl	F1-score for df	F1-score for mel	F1-score for nv	F1-score for vasc
ResNet9 (7:3)	78.28	0.43	0.54	0.50	0.10	0.38	0.90	0.76
DeiT (7:3)	82.43	0.48	0.66	0.56	0.22	0.54	0.92	0.66
Swin (7:3)	83.72	0.57	0.72	0.61	0.16	0.54	0.92	0.75
CrossViT (7:3)	76.62	0.32	0.44	0.51	0	0.31	0.90	0.68
ResNet9 (8:2)	78.12	0.42	0.54	0.51	0	0.33	0.90	0.79
DeiT (8:2)	81.12	0.41	0.66	0.51	0.34	0.46	0.92	0.76
Swin (8:2)	84.35	0.49	0.73	0.67	0.43	0.50	0.93	0.75
CrossViT (8:2)	77.86	0.11	0.40	0.48	0.07	0.32	0.90	0.66
ResNet9 (9:1)	78.82	0.39	0.54	0.52	0	0.41	0.90	0.89
DeiT (9:1)	85.07	0.40	0.75	0.67	0.42	0.62	0.93	0.73
Swin (9:1)	84.92	0.36	0.70	0.69	0.32	0.63	0.93	0.66
CrossViT (9:1)	79.02	0.31	0.52	0.51	0	0.42	0.91	0.55

Table 1 shows the outcomes for the three partitioned sets of experiments in terms of overall accuracy and F1-scores for each class. First, comparing the accuracy of ResNet9 and the three ViT models, the accuracy of ResNet9 in the three sets of experiments was always maintained at about 78%. However, the accuracy of DeiT and Swin was always much higher than ResNet9, which was held at more than 80%. Although the result of CrossViT is consistently below 80%, it is surprising that with the increase in training data, the accuracy of CrossViT is constantly improving. It eventually surpassed ResNet9's overall accuracy when the training data amount reached 9015, which corresponds to the size ratio of roughly 9:1 for the training dataset to the validation set. The above

observations have already demonstrated that the ViT's performance on the dermoscopy image classification task is not inferior to CNN.

Then observing the changes in the accuracy of the same model as the training data increases, ResNet9 does not seem to be related to the training data size. In contrast, the accuracy of the other three ViT models has not only significantly improved, for which the improvement rate is at least 1%, but also tends to be positively correlated to the training data size. Among the three ViT models, DeiT has significantly increased from 81.12 to 85.07, nearly 4%. This observation implies that with more training data, the accuracy of the CNN models may not improve much, but the improvement effect of the ViT models is obvious.

Table 2. The amount of data per class for the whole dataset and training datasets of three experimental groups

Group	nv	mel	bkl	bcc	akiec	vasc	df
Dataset	6705	1113	1099	514	327	142	115
7:3	4656	847	760	363	213	97	79
8:2	5334	941	870	416	252	110	92
9:1	6020	1022	988	462	291	128	104

Finally, the experiment calculated the F1-score, which is the higher and closer to 1 for each class, the better the classifier may be. As shown in Table 2, the whole dataset is quite unbalanced. For example, the data volume of nv accounts for approximately two-thirds of the entire dataset; in contrast, the data volume of vasc and df each only accounts for about one percent of the whole dataset. However, the training dataset in three different data partition ways does not change the size relationship of the seven classes. In other words, according to the order from left to right in Table 2, the amount of data for the seven labels always decreases monotonically. The nv class has the largest amount of data, and its F1-score is the highest no matter which model or data division it is, almost close to 1. And the df with the least amount of data, its scores are overwhelmingly the worst. These two observations imply that the amount of data positively correlates with the F1-score. However, the amount of vasc data is the second smallest, but the f1-score is not necessarily lower than other classes except nv, such as mel, whose data amount is nearly ten times that of vasc. It leads to a conjecture that features for some types of skin diseases may be so inconspicuous or confusing that the classifier cannot effectively judge them even after receiving a lot of training data.

5. Conclusion

In conclusion, transformers have achieved ideal results comparable to or even better than CNN in exploring the viability and effectiveness of the state-of-the-art ViT on dermoscopy image classification task. Although the CNN model selected in the experiment for comparison with ViT is not designed to be complex or diverse, the effect achieved by ViT is undeniable. In the future, if the transformer wants to achieve better results in recognizing pigmented skin lesions, the critical point is to increase the training data. One way is to input more manually labeled images, and the other is to consider using generative adversarial neural networks to produce more data to make the dataset more balanced.

References

[1] Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-sources dermatoscopic images of common pigmented skin lesions, Scientific Data, vol. 5, no. 1, pp. 1-9. doi: 10.1038/sdata.2018.161.

[2] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[3] Abuared, Nour, Alavikunhu Panthakkan, Mina Al-Saad, Saad Ali Amin, and Wathiq Mansoor. 2020. Skin Cancer Classification Model Based on VGG 19 and Transfer Learning. In 2020 3rd International Conference on Signal Processing and Information Security (ICSPIS), DUBAI, United Arab Emirates, pp. 1–4. doi: 10.1109/ICSPIS51252.2020.9340143.

[4] Khan, Muhammad Attique, Tallha Akram, Muhammad Sharif, Seifedine Kadry, and Yunyoung Nam. 2021. Computer Decision Support System for Skin Cancer Localization and Classification. Computers, Materials & Continua, vol. 68, no. 1, pp. 1041–1064. doi: 10.32604/cmc.2021.016307.

[5] Chang, Chung Hsing, Wei En Wang, Feng Yi Hsu, Rou Jhen Chen, and Han Chao Chang. 2022. AI HAM 10000 Database to Assist Residents in Learning Differential Diagnosis of Skin Cancer. In 2022 IEEE 5th Eurasian Conference on Educational Innovation (ECEI), pp. 1–3. doi: 10.1109/ECEI53102.2022.9829465.

[6] Ali, Md Shahin, Md Sipon Miah, Jahurul Haque, Md Mahbubur Rahman, and Md Khairul Islam. 2021. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. Machine Learning with Applications, vol. 5, p. 100036. doi: 10.1016/j. mlwa. 2021. 100036.

[7] Pedro, Rafael, and Arlindo L. Oliveira. 2021. Assessing the Impact of Attention and Self-Attention Mechanisms on the Classification of Skin Lesions. arXiv preprint arXiv:2112.12748.

[8] Touvron, Hugo, Matthieu Cord, and Hervé Jégou. 2022. DeiT III: Revenge of the ViT. arXiv preprint arXiv: 2204.07118.

[9] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 9992–10002. doi: 10. 1109/ ICCV48922.2021.00986.

[10] Chen, Chun-Fu Richard, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 347–356. doi: 10.1109/ICCV48922.2021.00041.