HierAttn: Effectively Learn Representations from Stage Attention and Branch Attention for Skin Lesions Diagnosis

Wei Dai\textsuperscript{a}, Rui Liu\textsuperscript{a}, Tianyi Wu\textsuperscript{a}, Min Wang\textsuperscript{a}, Jianqin Yin\textsuperscript{b}, Jun Liu\textsuperscript{a,∗}

\textsuperscript{a}Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China
\textsuperscript{b}School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

A R T I C L E I N F O

Article history:
Received 07 May 2022
Received in final form xx xx 2022
Accepted xx xx 2022
Available online xx xx 2022
Communicated by xxx xxx

Keywords: Skin lesions classification, Same channel attention, Hierarchical pool, Stage attention, Branch attention

A B S T R A C T

An accurate and unbiased examination of skin lesions is critical for the early diagnosis and treatment of skin cancers. The visual feature of the skin lesions varies significantly because skin images are collected from patients with different skin colours by using various devices. Recent studies have developed ensembled convolutional neural networks (CNNs) to classify the images for early diagnosis. However, the practical use of CNNs is limited because their network structures are heavyweight and neglect contextual information. Vision transformers (ViTs) learn the global features by self-attention mechanisms, but they also have comparatively large model sizes (more than 100M). To address these limitations, we introduce HierAttn, a lite and effective neural network with hierarchical and self attention. HierAttn applies a novel strategy based on learning local and global features by a multi-stage and hierarchical network. The efficacy of HierAttn was evaluated by using the dermoscopy images dataset ISIC2019 and smartphone photos dataset PAD-UFES-20. The experimental results show that HierAttn achieves the best top-1 accuracy and AUC among state-of-the-art mobile networks, including MobileNetV3 and MobileViT. The code is available at https://github.com/anthonyweidai/HierAttn. © 2022. All rights reserved.

1. Introduction

Skin conditions and disorders are the most common human diseases to affect millions of people (Guy Jr et al., 2015a,b). Statistical data estimate around 20% of Americans are diagnosed with malignant cutaneous diseases (Esteva et al., 2017). Skin cancer, consisting of non-melanoma and melanoma of the skin, has more than 1.5 million new cases globally in 2020 (Sung et al., 2021). It is also estimated to be the fifth most common detected cancer, with 196,060 new cases, in the U.S. in 2021 (Fund, 2021; Sung et al., 2021). The annual cost of treating new melanoma cases is projected to triple from 2011 to 2030 (Guy Jr et al., 2015b). Proactive detection and early diagnosis are critically important, for a 99% 5-year survival rate for melanoma drop could be achieved by early-stage treatment to only around 27% if detected in its late-stage (Fund, 2021; Sung et al., 2021).

Traditional methods for detecting skin disorders include skin cancer screening, self-examination, and clinical examination. Skin self-examination is the most common method for the early detection of skin diseases. Around 53% of patients with melanomas are self-examined (Avilés-Izquierdo et al., 2016). With limited professional knowledge, numerous patients have doubts in their examination results and are unsure which type of skin disorders they have. Clinical skin examination can provide affirmative screening of skin cancers with a high detection accuracy (Loescher et al., 2013). However, the clinical exami-
nation needs to consume medical professionals’ time to review a large number of dermoscopic images, which is relatively insufficient for all possible patients.

Current technologies make it possible to diagnose skin lesions by analysing skin images. The imaging technologies in skin examination include dermoscopy, reflectance confocal microscopy, total body photography, and teledermatology. Dermoscopy is a non-invasive imaging method without reflecting light to provide 10x magnification of a skin lesion (Loescher et al. 2013). The review of dermoscopic images is time-consuming, and the examination results are subjective to healthcare workers. The diagnostic accuracy of dermoscopy relies on extensive training, practice, professional knowledge and experience to distinguish features and morphologies (Rosendahl et al., 2011).

Due to the sophisticated and unique features of skin lesion images, it is strenuous to detect skin cancer automatically and accurately. Throughout decades of development of technologies and computational resources, deep learning methods have demonstrated great potential to classify multi-class objects and outperform most experts in some cases (Tschandl et al., 2020). A well-trained deep learning model makes it possible to detect skin cancer in an early, treatable stage. Recently, ensemble models with ResNeXt, NASNet, SENet, DenseNet121, EfficientNet performed well, with highest accuracy 94.2% and 92.6% in ISIC2018 and ISIC2019, respectively (Adegun and Viriri, 2021). However, these ensemble models consume substantial computational resources and are highly time-consuming. Furthermore, most of them ignore the global representations from skin lesion images. The lightweight algorithms can assist dermatologists in conducting direct skin cancer screening with limited computational resources from clinical computers. Moreover, even non-professional people, specifically the patients, have a chance to pre-distinguish symptoms of skin cancer at home by such lightweight models.

This article aims to address the challenge of reliably and effectively detecting skin cancers with small memory storage and minimal computational cost. Thus, we introduce a new deep learning model to tackle these problems.

The key contributions of this study are:

- We introduce a new lightweight and low latency architecture, HierAttn [see Fig. 1], to distinguish multi-class skin lesions on the ISIC2019 and PAD-UFES-20 datasets. HierAttn is relied on the same channel attention, stage attention, and branch attention and achieves top-level performance while maintaining a small size as classic mobile deep learning models.
- We propose the same channel attention module after depth-wise convolution, which outperforms similar attention methods like squeeze and excitation while does not increase the model’s weight.
- We use branch attention block, based on hierarchical pooling, to progressively learn both local and global representations from censorious learning stages.
- We introduce a new comprehensive data balance method and show that it outperforms random sampling techniques for detecting skin lesions.

We discussed related work in Sec. 2 and presented our proposed methodology in Sec. 3. Finally, we show the experiment results in Sec. 4, discuss the results in Sec. 5 and summarise our findings in Sec. 6.

2. Related work

Data processing. Three commonly used skin lesions datasets (e.g., HAM10000, ISIC2019, PAD20) are utilised in the study for the development and evaluation of deep learning models. The three datasets have the problem of large-scale data imbalance, which significantly decreases the performance of deep learning algorithms. Possible solutions are to perform data balancing for each class or few-shot learning (Prabhu et al., 2018; Weng et al., 2020). Other solutions include the modified loss function that was specifically designed to deal with the heavily imbalanced dataset (Roy et al., 2022). However, the improvement of these methods is relatively limited because they are more likely to misclassify those classes with a rare number of samples. Machine learning approaches are employed to analyse data such as their “hardness” and alleviate the effects of class overlap (Smith et al., 2014). Due to different illumination

![HierAttn architecture](image_url)

Fig. 1: HierAttn architecture. Conv-\(n \times n\) represents a standard convolution and SCADW refers to a depth-wise separable convolution block. Down-sampling blocks are marked with \(\downarrow 2\). Each stage attention has a SCADW block followed by a convolution-transformer hybrid block.
CNNs in skin lesions diagnosis. With the development of deep learning, an increasing number of available architectures were introduced. The deep learning models, including Inception V3, VGG, EfficientNet, ResNet, DenseNet, etc., have been applied to tackle skin lesions classification problems (Esteve et al., 2017; Weng et al., 2020; Gessert et al., 2020). Moreover, the models, assembling ResNetXt, NASNet, SENet, DenseNet121, and EfficientNet in several streams or stages, are often utilised in skin lesions classification (Gessert et al., 2020; Mahbod et al., 2020; Attique Khan et al., 2021). The ISIC2019 challenge winner applied an ensemble EfficientNet to achieve 92.6% average classification accuracy; however, the model has heavy-weight with more than 100 M parameters, making it impractical for clinical and home use (Gessert et al., 2020). To reduce computational cost and model weight, researchers have applied the depth-wise separable convolution method. Based on such convolution method, MobileNetV2, MobileNetV3, MnasNet, ShuffleNet and EfficientNet were constructed and obtained relatively satisfactory performance (Sandler et al., 2018; Zhang et al., 2018; Tan and Le, 2019; Tan et al., 2019; Howard et al., 2019).

Attention Mechanisms. The attention mechanism is a biomimetic cognitive method used in diverse computer vision assignments, such as image classification (Woo et al., 2018; Hu et al., 2018; Dosovitskiy et al., 2020; Hou et al., 2021; Mehta and Rastegari, 2021) and image segmentation (Mehta and Rastegari, 2021; He et al., 2022). An example is SENet, which obtains global representations by squeeze and channel-wise feature response by excitation (Hu et al., 2018). The convolution block attention module (CBAM) improves this idea by using a larger kernel size to encode spatial information (Woo et al., 2018). Coordinated attention further advances this idea by encoding channel relationships and long-range dependencies via average pooling along different axes (Hou et al., 2021). However, they introduce more learnable parameters and consume more computational resources. To improve the computational efficiency and meet the scalability requirement, self-attention mechanisms, particularly transformers, are introduced into computer vision from natural language processing. With self-attention, a vision transformer (ViT) replaces the traditional convolution method with a transformer encoder (Dosovitskiy et al., 2020). Because the input images are directly split into patches and embedded, the ViT models are still quite large, with more than 100 M parameters (Dosovitskiy et al., 2020). To reduce the latency of splitting images, computer scientists from Apple proposed MobileViT by applying the MobileNet V2 block to compose images and then used a transformer to process the information (Mehta and Rastegari, 2021). MobileViT has been reported as an effective network for image classification while maintaining a lightweight.

3. Proposed Methodology

3.1. Dataset

ISIC2019 (Tschandl et al., 2018; Adegun and Viriri, 2021; Combalia et al., 2019) and PAD-UFES-20 (PAD20) (Pacheco et al., 2020), two publicly available skin lesions datasets, are applied in this research for the evaluation of the proposed framework. ISIC2019 dataset consists of 25,331 dermoscopy images with eight categories of skin lesion: actinic keratosis (ACK), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevus (NV), squamous cell carcinoma (SCC), vascular lesion (VASC) (Tschandl et al., 2018; Adegun and Viriri, 2021; Combalia et al., 2019). Three of them belong to non-melanoma skin cancer, ACK, BCC and SCC. Moreover, melanoma, regarded as an uncontrolled growing cell that can produce pigment, is the most severe skin cancer. Furthermore, the vascular lesion is either benign or malignant. Finally, the rest types of skin lesions are benign.

The PAD20 dataset includes 2,298 skin lesion images collected by smartphones with six classes: ACK, BCC, BKL, MEL, NV, and SCC (Pacheco et al., 2020). Compared to ISIC2019, PAD20 has two fewer classes, DF and VASC, because of lacking photos of skin lesions. The data distribution of the two datasets is shown in Fig. 2. Dermoscopy and smartphones are two standard methods to capture skin lesions images. Thus, the two datasets effectively represent current image data on skin lesions.

![Fig. 2: Data distribution on ISIC2019 (left image) and PAD20 (right image)](image)

3.2. Pre-processing

3.2.1. Image pre-processing

The data among different classes from the ISIC2019 dataset has large black areas [see Fig. 3(a)], which damage the evaluation performance of models. Thus, an adaptive cropping strategy is taken to identify and crop these images. Firstly, the input image is turned into greyscale [Fig. 3(b)]. The grey images are then binarized with 50 to 255 thresholds [Fig. 3(c)]. After that, the contour of the binarized images is detected to confirm the circle’s location. When the value of a circle divided by the area of the whole image is between 0.01 and 0.9, the region enclosed by the circle is extracted and saved [Fig. 3(d)]. Post-processing inspection indicates that the method is robust.
3.2.2. Data balance

In this study, an imbalance ratio is defined as the amount of data of the majority class to those of the minority class. The imbalance ratio for ISIC2019 and PAD20 are 53.9 and 16.3, respectively. Such an imbalance ratio can remarkably reduce the metrics of deep learning, according to previous research in (Buda et al., 2018). The data imbalance could lead to low validation metrics for the minority class, though the averaged validation metrics could be high. Data balance by oversampling or undersampling for each class is a practical solution to handle the dataset with a large imbalance ratio. Oversampling and undersampling are simultaneously utilised to balance ISIC2019 and PAD20 data. After sampling, 2500 images and 500 images are collected for each class in the ISIC2019 and PAD20 datasets, respectively. Thus, the amount of data after balancing totals 20000 and 3000 on ISIC2019 and PAD20, respectively, close to the total amount of unbalanced data.

Oversampling uses horizontal flips, random crops, Gaussian blur, linear contrast, random translation, rotation, and shear on a small scale to generate new images from the old images. As for undersampling, a traditional method randomly selects images from the datasets, which has an overlapping class problem. Thus, we adopted an adaptive data analysis method called instance hardness (IH) to alleviate this adverse effect. Instance hardness can be defined as (Smith et al., 2014):

\[
IH_L((x_i, y_i)) = 1 - \frac{1}{|L|} \sum_{j=1}^{|L|} p(y_i|x_i, g(t, \alpha))
\]

where \( L \) is a prior with non-zero probability while treating all other learning algorithms as having zero probability, \( g \) is a machine learning algorithm trained on \( t \) with hyper-parameters \( \alpha \), and \( y_i \) is the label for data \( x_i \).

Outliers and mislabelled data are expected to have high IH. Thus, IH analysis was applied in undersampling to remove those data with a high IH. This research uses the random forest as the machine learning method \( g \) referenced from the imbalanced-learn study (Lemaitre et al., 2017). Datasets after sampling are renamed as IHISIC20000, RandISIC20000, IHPAD3000, and RandPAD3000. Randomly oversampling is applied to these four datasets. IH and Rand represent the dataset using instance hardness and random sampling methods for undersampling, respectively. Moreover, 20000 and 3000 are the total number of images in the datasets.

3.3. HierAttn Architecture

The standard transformer was applied to process sequences of image patches to learn the inter-patch representations. However, this type of transformer module ignores the inductive biases (e.g., translation equivariance and locality) inherent to CNNs, which leads to poor performance while training with insufficient data (Dosovitskiy et al., 2020). MobileViT was proposed to tackle the loss of inductive biases using the transformer module by taking convolution and transformer to form a hybrid block (Mehta and Rastegari, 2021). However, it cannot thoroughly learn representation due to the information lost. Multilevel loss is widely used to extract feature information from stages to improve the model’s performance. For instance, GoogleNet has three losses and adds them together to form the total loss in training (Szegedy et al., 2015). However, the features in various stages are just directly fed into simple tensor operations. They do not have any influence on each other.

The paper introduces an efficient and lite Convolution-Transformer model, HierAttn [see Fig. 1]. The main idea is to learn local and global representations by utilising attention and self-attention mechanisms effectively. The hierarchical branches allow to keep various information from different depths of layers.

3.3.1. Same Channel Attention

The block uses depth-wise separable convolution (DWSConv) with the same channel attention (SCAttn) is called SCADW block (i.e., green blocks in Fig. 1). Fully convolution has a high computation cost. Google researchers utilised the DWSConv block to reduce the model size (Sandler et al., 2018). Squeeze and excitation attention (SEAttn) blocks were proposed to enhance the expressive power of the learned features after depth-wise convolution in the MobileNetV3 block (Howard et al., 2019). However, the point-wise convolution already extracts the channel-wise information, which suggests that the excitation of SEAttn is more likely a redundant operation. Therefore, the SCAttn is introduced in this study after depth-wise convolution by utilising the attention mechanisms while maintaining the same number of learnable parameters as the DWSConv block. The SCAttn block is illustrated in Fig. 4.

![Fig. 4: Schematic comparison of the original block (without attention mechanism) (a), the SEAttn block (b), and the proposed SCAttn block (c).](image-url)

3.3.2. Stage attention

Each stage attention module (i.e., pink blocks in Fig. 1) has a SCADW block with a stride of 2 followed by a convolution-transformer hybrid block. Inspired by the MobileViT architecture (Mehta and Rastegari, 2021), we simultaneously apply the
convolutions and transformers to learn the local and global representations of an input skin lesion image with fewer parameters. The convolution-transformer hybrid (CTH) block steadily conducts unfolding, transformer encoding and folding like standard convolutions (Mehta and Rastegari 2021). Thus, the CTH block has inductive bias and can process the transformers with convolutions. Recently, vision transformers (ViTs) utilising multi-head self-attention mechanisms have shown great potential in classification tasks (Dosovitskiy et al., 2020). The CTH block also consists of a multi-head self-attention followed by a multilayer perceptron (MLP) layer. Furthermore, we also add a skip connection to link the input and output of the CTH block.

![Convolution-transformer hybrid block.](image)

**Fig. 5:** Convolution-transformer hybrid block.

### 3.3.3. Branch attention

In this study, we propose to use hierarchical pooling and tensor assembling to downsize tensors and learn features from different stages simultaneously. This new technique aims to utilise and improve the interactions of features among different learning stages. Moreover, by keeping the different sizes of pooling results, hierarchical pooling learns the local representation of tensors with large tensor size \( C \times 5 \times 5 \) and attains tensors’ global representation with small tensor size \( C \times 1 \times 1 \). Moreover, the medium tensor size \( C \times 3 \times 3 \) is also designed as a buffer layer to keep both local and global features. The pooled tensors from different branches are then channel-wisely assembled. After that, channel-wise ensembled tensors are depth-wisely assembled. At the end of branch attention, the pixels of assembled tensors are depth-wisely randomised by utilising the Monte Carlo method. The branch attention progress is illustrated in **Fig. 6**. Stage attention thoroughly rearranges feature maps by downsizing feature maps, and the processed features are further extracted by the CTH block as described in the aforementioned section. The branch attention is applied as a particular learning stage after each stage attention block.

### 3.3.4. Small-scale transformer

Learning global information by transformers utilises fewer parameters than convolutions. However, transformer operations lose spatial bias, increasing computational costs to learn visual representations (Mehta and Rastegari 2021). Thus, they are wide and deep. ViT-Base, ViT-Large, and ViT-Huge models use the number of transformer blocks \( L=12, 24, 32 \) and the number of embedded dimensions \( d = 768, 1024, 1280 \), respectively (Dosovitskiy et al., 2020). However, our new HierAttn only requires \( L = 2, 4, 3 \) and \( d = 96, 120, 144 \), are reference from MobileViT architecture. The number of parameters for the new HierAttn and other state-of-the-art networks is summarised in Table[1]

### 3.3.5. HierAttn architecture

The full name of HierAttn is hierarchical attention. On the one hand, we use SCAttn in each stage attention to process low-level to high-level features. On the other hand, we use hierarchical pooling to obtain hierarchical sizes of tensors, \( C \times 5 \times 5 \), \( C \times 3 \times 3 \) and \( C \times 1 \times 1 \), for learning local and global representations at different degrees. HierAttn is inspired by the philosophy of making full use of local and global information on the tensors and stage conditions of the network. We train HierAttn models at two different network sizes (S: small, XS: extra small). The SCADW blocks in HierAttn are typically responsible for decreasing tensors’ size while alleviating the information loss. The initial layer of stage attention is a \( 3 \times 3 \) standard convolution with a stride of 2, following SCADW blocks and stage attention modules. Moreover, all stage representations are further handled by the branch attention. Furthermore, stochastic depth is applied on every SCADW block and CTH block with a stride of 1 to reduce time cost and validation error.

![Branch attention by hierarchical pooling, assembling branch tensors and randomising tensors.](image)

**Fig. 6:** Branch attention by hierarchical pooling, assembling branch tensors and randomising tensors.
4. Experimental Results

In this section, we first evaluate HierAttn performance on the IHISIC20000 and IHPAD3000 datasets in subsection 4.1. Fig. 7 and Table 1 show that HierAttn delivers significantly better performance than state-of-the-art networks. In subsection 4.2, we conducted ablation studies for data balance methods and HierAttn architecture.

4.1. Image classification on the skin lesions dataset

4.1.1. Implementation details

The HierAttn networks are trained and validated for 500 epochs on one RTX A4000 with a batch size of 64 images using AdamW optimiser (Loshchilov and Hutter, 2017) with 10-fold cross-validation and cross-entropy loss. The learning rate is ceased from 0.002 to 0.0002 during the first 15 epochs and then increased to 0.0002 utilising the cosine scheduler (Loshchilov and Hutter, 2016). L2 weight decay of 0.01 is adopted. Knowledge transfer is applied to reduce the training time and improve model performance. The tunable parameters for initialisation of HierAttn are partly taken from Mobile-Vit, including standard convolution, SCADW blocks, and CTH blocks. All transferred models for training in IHISIC20000 and IHPAD3000 were trained in ImageNet1k (Mehta and Rastegari, 2021) and IHISIC20000, respectively. Thus, the weight of the classifier is ignored while transferring weight to train our models on the skin lesions dataset. Moreover, the transfer learning warm-up is applied to alleviate the negative influence of untransferred layers on transferred layers. On the first training 30 epochs, all transferred layers are frozen. After that, gradient calculation is required for all layers with learnable parameters. In addition, the inference time for each image is calculated by averaging 1000 iterations to demonstrate HierAttn for mobile applications with slower processors.

4.1.2. Top-1 Accuracy and Inference Time

Each model with six on IHISIC20000 or eight classes on IHPAD3000 has a similar number of parameters when two deci-

| Model                  | # Parameters/M | Inference time/ms | Top-1 Accuracy/% |
|------------------------|----------------|-------------------|------------------|
| MobileNetV2            | 2.23           | 0.802             | 93.45            | 87.44 |
| MobileViT-s            | 4.94           | 1.528             | 94.72            | 88.22 |
| MobileNetV3_Large      | 4.21           | 0.538             | 94.77            | 88.78 |
| ShuffleNetV2_1×        | 2.28           | 0.307             | 95.23            | 87.89 |
| MnasNet1.0             | 3.11           | 0.997             | 95.45            | 90.33 |
| EfficientNet_b0        | 4.02           | 0.586             | 95.48            | 90.22 |
| HierAttn_xs (ours)     | **1.08**       | 0.878             | **96.15**        | 90.11 |
| HierAttn_s (ours)      | 2.14           | 1.658             | **96.70**        | **91.22** |

Fig. 7: HierAttn vs. classic mobile models on (a) IHISIC20000 and (b) IHPAD3000 validation set.

Table 1: # Parameters, inference time, and top-1 accuracy of different models.

Extra Small Small Medium
Our models
Extra Small Small Medium
(an) (b)
mal places are kept. Thus, the number of parameters of each model is computed with eight classes to simplify the discussion. Fig. 7 compares HierAttn with six other lightweight networks that are also trained on IHISIC20000 and IHPAD3000 datasets. Detailed values are illustrated in Table 1. Fig. 7 demonstrates that HierAttn networks fall in the upper left region, which means they outperform other mobile models with relatively small sizes. For instance, with about 1.08 million parameters, HierAttn_s outperforms MobileNetV2 by 3.25%, MobileNetV3_large by 1.93%, ShuffleNetV2_1x by 1.47%, MnasNet1.0 by 1.25%, and EfficientNet-b0 by 0.55% on IHISIC20000 validation set. Furthermore, HierAttn_s also outperforms MobileViT_s, which is also a convolution-transformer hybrid model, by 1.98% and 3.00% on IHISIC20000 and IHPAD3000 datasets, respectively. Moreover, Table 1 shows the inference time for each image by different models. The inference time for each image by HierAttn_s is lower than 1 ms, 12% faster than MnasNet1.0’s and just 9% slower than MobileNetV2’s.

4.1.3. ROC and AUC

In this paper, the receiver operating characteristic (ROC) curve and its area under the curve (AUC) are applied to demonstrate the general performance of each model. The results can be found in Fig. 8. The figure represents that all models on experiments have more than 0.99 AUC on the IHISIC20000 and 0.98 AUC on the IHPAD3000 validation set, suggesting that they possess satisfactory analytical capacities to conduct multi-classes lesions classification tasks. Furthermore, the ROCs of HierAttn_s and HierAttn_xs are the first and second closest to the point (0, 1) on both IHISIC20000 and IHPAD3000 validation sets. Moreover, HierAttn_s and HierAttn_xs have the first and the second largest AUC on both datasets among all models, for instance, 0.99772 and 0.99558 on the IHISIC20000 validation set, respectively. Thus, the ROCs and AUCs show that HierAttn is the most reliable and superlative model to detect skin lesions among current conventional and advanced mobile models. Meanwhile, the same model’s AUC of the IHPAD3000 validation set is lower than the IHISIC20000 validation set, which means these models perform better on the IHISIC20000 validation set.

In conclusion, these experimental results prove that HierAttn is robust and effective. Thus, it can be practically applied to skin lesions diagnosis.

4.2. Ablation studies

4.2.1. Implementation details

In ablation studies, the model used is HierAttn_s, and the dataset used is IHISIC20000, if not mentioned. Other parameters are the same as Section 4.1.

4.2.2. Data balance methods

Fig. 9 illustrates the top-1 accuracy of different data balance methods on ISIC20000 and PAD3000. Firstly, the top-1 accuracy of IH (instance hardness) is higher than that of Rand (randomised sampling) on both datasets, which means IH can more effectively sample images than Rand. Secondly, IH exceeds 5.9% and 0.7% top-1 accuracy than Rand in ISIC20000 and PAD3000, respectively. The diminishing improvement on the PAD dataset could be caused by the low imbalance ratio of PAD compared to the ISIC dataset (16.3 for PAD vs. 53.9 for ISIC). The effects of alleviating negative influence on classification are more obvious for the dataset with a high imbalance ratio. Thirdly, the error bar of ISIC20000 is shorter than PAD3000, indicating that HierAttn is more robust on ISIC20000 than PAD3000.

4.2.3. Attention blocks in DWSConv

Table 2 shows the number of parameters, inference time and top-1 accuracy of models using different attention mechanisms in DWSConv. It can be discovered that without the attention mechanism [see Fig. 4 (a)], the network only achieved only 96.20% top-1 accuracy. However, HierAttn with SEAttn or SCAttn attains at least 4.5% more top-1 accuracy than without using the attention mechanism in DWSConv. Although the top-1 accuracy of HierAttn with SCAttn is only 0.5% more than HierAttn with SEAttn, HierAttn with SCAttn is faster for inference with smaller the number of parameters than HierAttn with SEAttn.
Mobile application for skin cancer diagnosis allows dermatologists to perform point of care testing. Moreover, possible patients can carry out further detection by utilizing mobile applications while doing regular self-exam. HierAttn has a statistically close speed to classic mobile networks, which shows great potential to be developed on mobile devices. If the skin lesion is recognised as MEL, BCC, ACK, SCC and VASC with more than 50% possibility, users are suggested to go to a clinic or hospital to perform further diagnosis. Otherwise, it is more likely that the detected area of the skin is healthy. We expect that HierAttn can be deployed on the mobile phone to assist ordinary people in performing regular self-check in the future.

6. Conclusion

In this paper, we propose a HierAttn network consisting of SCAttn, stage attention, and branch attention for skin lesions diagnosis. SCAttn directly extracts global features by only global average pooling without operating channel-wise information abundantly. Stage attention consists of a SCADW block to downsize feature maps and a CTH block to learn local and global representations effectively. Branch attention applies hierarchical pooling after each stage attention to learn local and global representations while improving the interaction of feature relationships. Additionally, we propose a comprehensive data balance method based on instance hardness analysis undersampling and random oversampling. With these novel modules, HierAttn can achieve better skin lesions classification results, 96.70% top-1 accuracy and 0.9972 AUC on IHISIC20000 and 91.22% top-1 accuracy and 0.98816 AUC on IHPAD3000 validation set, than other state-of-the-art mobile networks.

Acknowledgments

Acknowledgments should be inserted at the end of the paper, before the references, not as a footnote to the title. Use the unnumbered Acknowledgements Head style for the Acknowledgments heading.

References

Adegun, A., Viriri, S., 2021. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artificial Intelligence Review 54, 811–841.
Adegun, A., Viriri, S., 2021. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artificial Intelligence Review 54, 811–841.
Aguirre Khan, M., Sharif, M., Akram, T., Kadry, S., Hsu, C., 2021. A two-stream deep neural network-based intelligent system for complex skin cancer types classification. International Journal of Intelligent Systems.
Avilés-Izquierdo, J.A., Molina-López, I., Rodríguez-Lomba, E., Marquez-Rodas, I., Suarez-Fernandez, R., Lazaro-Ochaita, P., 2016. Who detects melanoma? impact of detection patterns on characteristics and prognosis of patients with melanoma. Journal of the American Academy of Dermatology 75, 967–974.
Barata, C., Marques, J.S., Celebi, M.E., 2014. Improving dermoscopy image analysis using color constancy, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 3527–3531.
Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks 106, 249–259.
Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al., 2019. Bcn2000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288.
Supplementary Material

HierAttn architecture

The HierAttn architecture is motivated by MobileNetV3, EfficientNet, MobileViT, and GoogleNet. The first layer of the HierAttn is a standard $3 \times 3$ convolution. Learning from MobileNetV3, we employ SCADW blocks to extract features with depth-wise separable convolution and attention mechanisms followed by three-stage attentions. Referencing MobileViT, we utilise the CTH block in stage attention to learn local and global representations and Swish as the activation function (Mehta and Rastegari, 2021). EfficientNet shows insights on using stochastic depth to stabilise the training process (Tan and Le, 2019). We adopt this method in every skip connection of the HierAttn. Finally, GoogleNet has a perception that uses stage information by multi-loss during training (Szegedy et al., 2015). Based on this, we further combine stage information by hierarchical pooling using hierarchical attention mechanisms. The architecture parameters are shown in Table 3.

Pre-processing

The source code for skin lesion image pre-processing like Fig. 2 is publicly available on https://github.com/anthonyweidai/circle_extractor.

Oversampling. One of the oversampling results is illustrated in Fig. 10.

IH analysis. The demonstration of IH analysis and random undersampling methods can be found in Fig. 11. The four colours represent the four undersampled classes in ISIC2019. The ratio of data points is the same as the classes in ISIC2019. This figure shows that IH analysis can efficiently remove outliers from the datasets, while the random undersampling method cannot, and there is a more considerable amount of overlap sampled data. Fig. 9 depicts that the performance of the HierAttn improves by 5.9% by using IH on ISIC20000.

Additional ablations

Transfer learning warm-up. We consider that the models are unstable after partially transferring tunable parameters, and statistical methods randomly initialise those layers without transferring. We apply a warm-up technique to alleviate the influence of random weight initialisation for transferred layers. We freeze the transferred layers, locking gradient descent, in the first 30 epochs. In these 30 epochs, only those layers without transferring can update their parameters. Table 4 shows that the performance of the HierAttn improves by 0.42% on the IHISIC2019 dataset.

Stochastic depth. Stochastic depth, also regarded as “layer dropout”, is implemented in each layer with a skip connection in HierAttn. Typically, it is in all SCADW blocks with a stride of 2.

| Layer          | Output size | Repeat | Output Channels | Output size | Repeat |
|----------------|-------------|--------|-----------------|-------------|--------|
| Image          | $256 \times 256$ |        |                 | XS          | S      |
| $3 \times 3, \downarrow 2$ SCADW | $128 \times 128$ | 1      | 16              | 16          | S      |
| $3 \times 3, \downarrow 2$ SCADW | $64 \times 64$ | 1      | 24              | 24          | S      |
| Stage1 Conv3 $3 \times 3, \downarrow 2$ CTH block | $32 \times 32$ | 1      | 48              | 48(d = 64)  | S      |
| Stage1 Conv3 $3 \times 3, \downarrow 2$ CTH block | $16 \times 16$ | 1      | 64              | 64(d = 64)  | S      |
| Stage1 Conv3 $3 \times 3, \downarrow 2$ CTH block | $8 \times 8$ | 1      | 80              | 80(d = 60)  | S      |
| Branch attention Conv1 $1$ | $8 \times 8$ | 1      | 192             | 192         | S      |
| Branch attention Linear | $1 \times 1$ | 1      | 768             | 768         | S      |
| # Parameters   |             |        |                 | 1.08 M      | 2.14 M |

Table 3: HierAttn architecture. Here, $d$ means dimensionality of the input to the transformer layer in the CTH block. In the CTH block, kernel size is set as three and patch height and width are set as two.
Fig. 11: Data undersampling before and after processing by (a) random sampling (b) instance hardness threshold

Table 4: Impact of transfer learning warm-up. Here, results are demonstrated for the HierAttn_s model on the IHISIC2019 dataset.

| Warm up | Top-1 Accuracy/\% |
|---------|-------------------|
| Without | 96.28             |
| With    | 96.70             |

of 1 and all CTH blocks. Table 5 demonstrates that the stochastic depth effectively enhances the performance of HierAttn_s by 0.65%. Note that even without this stochastic depth, the performance of HierAttn_s delivers similar or better results than SOTA mobile models (Fig. 7).

Table 5: Impact of stochastic depth. Here, results are demonstrated for the HierAttn_s model on the IHISIC2019 dataset.

| Stochastic depth | Top-1 Accuracy/\% |
|------------------|-------------------|
| Without          | 96.05             |
| With             | 96.70             |

Skip connection of CTH block. We add a skip connection link in the CTH (convolution-transformer hybrid) block with a stride of 1. Moreover, we also use stochastic depth with those modules with skip connections. Thus, we can not only reuse the lower lever feature but alleviate gradient descent. Table 6 shows a 0.23% improvement in the performance of HierAttn_s with skip connection. Furthermore, MobileViT achieved 0.5% less performance with skip connections on CTH blocks but without stochastic depth (Mehta and Rastegari 2021). It demonstrates that stochastic depth can reduce the negative influence of adding two distinctive features in the skip connection of the CTH block.

Table 6: Impact of skip connection. Here, results are demonstrated for the HierAttn_s model on the IHISIC2019 dataset.

| Skip connection | Top-1 Accuracy/\% |
|-----------------|-------------------|
| Without         | 96.47             |
| With            | 96.70             |