Fine-grained Cross-Layer Attention Framework for Wound Stage Classification

Keval Nagda
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, USA
knagda@ucsc.edu

***Abstract*—Determining progress during wound healing is crucial for effective diagnosis and treatment. Previous works have solved this task using methods paying attention to specific regions of the image. However, we explore an alternative, non-local attention approach and implement a cross-layer attention mechanism that focuses on the areas of interest and considers related spatial regions of the wound. Experimental results and visual representations show that adding cross-layer modules to mid-level and top-level layers enables better classification of wound healing stage and generalization.**

***Index Terms*—Cross-Layer Attention, Wound Healing, Classification**

I. Introduction

Wound healing can be categorized into four distinct phases: hemostasis, inflammatory, proliferation, and maturation [1]. The wound healing process begins with the hemostasis phase, where the wound is quickly closed by clotting. In the second stage of wound healing, inflammation controls swelling and infection, making the wound appear red and reflective. During the proliferative stage, the wound decreases in size as new tissues are built, and the wound's area turns visibly pink. The wound shows the signs of complete healing in the maturation or the remodeling phase as new skin is developed.

Computer vision and machine learning can be used to solve the task of accurate wound healing stage estimation. However, the available data for this task is often scarce and unfit for the desired results. This paper explores different neural network architectures and training methods to overcome this problem. We use a Convolutional Neural Network (CNN) architecture to learn the spatial relationship in the images. CNN's are effective for image classification tasks since different layers in the model capture various features [3]. Moreover, the middle layers learn spatial details while the top layers at the end learn global information from the images. This approach can be an issue while performing Fine-Grained Visual Categorization (FGVC) tasks since the convolution operations concentrate more on the local structure and avoid long-range dependencies.

To overcome the issue of lack of information sharing between the intermediate and top layer and to consider non-local attention, we introduce a Cross-Layer Attention (CLA) on the backbone of a CNN model [4]. The proposed cross-layer mechanism includes two modules: Cross-Layer Context Attention (CLCA) and Cross-Layer Spatial Attention (CLSA). We apply a CLCA module to the intermediate layers. The idea is that the CLCA module improves the global context information on the feature maps from the middle layers. We use a CLSA module on the top layer. The CLSA module enhances the ability to capture spatial attention of the feature maps from the top layer.

We pre-train the CNN using a siamese architecture [5], where we take a stack of two CNN encoders to learn temporal encoding for the wound healing process. We also apply CLA to the CNN encoder and follow the above pre-training steps to explore an alternative approach. We use the respective pre-trained encoder models to perform wound stage classification and to perform a comparative study on the wound stage estimation [6].

II. Description of Data

The dataset used in this work contains 255 cropped images gathered from Yang et al. [7]. The dataset was created by capturing the left and right-wound of 8 mice for the 16 days of the wound closure process, starting from day 0 (the day of surgery) to day 15 (the end of the experimental period). All the images were consistently captured using a mobile device at a distance of 12 cm. Of eight mice, four are young (12-14 weeks old), and four are aged (22-24 months old). The dataset has a missing image of the right wound on Day 9 for the second young mice (Y8-2 Day 9). The original images were cropped around the red circular splint of the wound using an object detection algorithm to avoid the irrelevant parts of the mice, as shown in Fig. 1. This paper uses cropped images only to include the wound to study the difference between the local and non-local attention mechanisms.

A. Pair Dataset Generation

In order to train the siamese network, we create pairs of cropped images to learn the temporal relevance during the wound healing process [8]. The pair dataset contains two sets of pairs - positive pairs (generated in the forward direction from Day 0 to Day 15) and negative pairs (generated in the backward direction from Day 15 to Day 0), resulting in 3,810 image pairs. A pair is positive if the two images in the pair are in ascending temporal order of the day the image was captured. In contrast, the two images in a negative pair appear to be in decreasing temporal order of the day the image was captured.

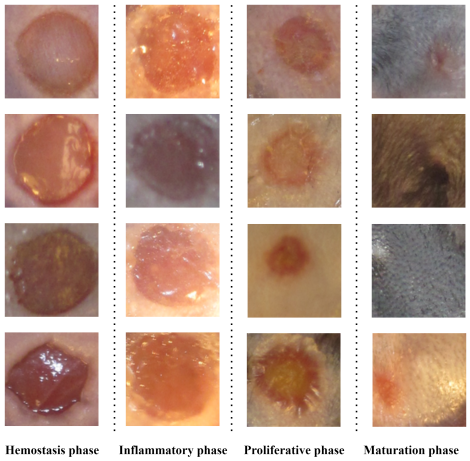

Fig. 1. The representative cropped wound images of all classes categorized into 4 stages of wound healing - hemostasis, inflammatory, proliferative, and maturation.

These pairs are split into 2,850 training, 480 validation, and 480 test samples. The data is split so that all images of the right wound of the aged mouse ID=5 and the left wound of the young mouse ID=4 are in the validation set, and the other wound of the aged mice ID=5 and young mice ID=4 are in the test set.

B. Single Image Dataset

The single image dataset consists of 255 cropped images and is used to train the fine-grained classification models. The dataset is divided into 191 training, 32 validation, and 32 test samples. The method of dataset split is similar to the pair dataset generation. We augment the dataset with linear transformations by performing random rotation, flip, and zoom operations on the images.

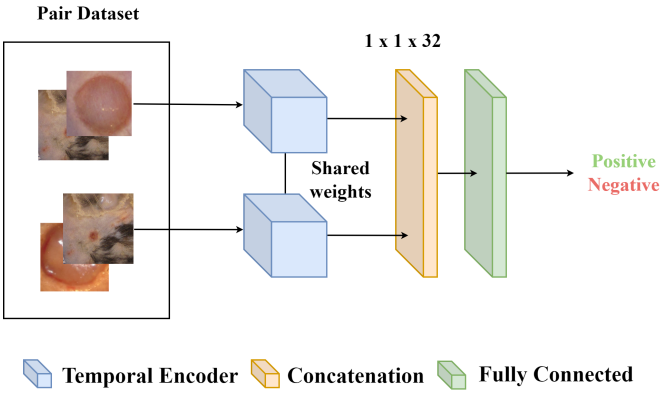

Fig. 2. Pre-training of the temporal encoder on generated pair image dataset. The left (top) encoder and right (bottom) encoder share the weights and are trained on positive and negative pairs. The output of each encoder is a 16-dimensional embedding that is concatenated and fed to a fully-connected layer to perform binary classification [2].

C. Labels

The single-cropped images were manually labeled by ten non-experts, given the guidelines mentioned previously. Each image is labeled with four probabilities representing the distribution of labels given by the annotators. After the pre-training task, we used these labels for the wound stage classification task.

III. Approach

First, self-supervised learning is performed using the convolutional temporal encoder and the pair image dataset using a siamese network, as shown in Fig. 2. The temporal encoder learns temporally-relevant embeddings, which can be helpful for the downstream classification task. To perform the wound stage classification task, we added a fully-connected classification layer to the trained temporal encoder and retrained it on the single image dataset. Finally, cross-layer attention is added to the temporal encoder and trained similarly to compare the results based on the performance of the two approaches.

A. Convolutional Temporal Encoder

In order to learn the temporal context of the wound healing process, we use a simple convolutional temporal encoder, as shown in Fig. 3 [2]. The encoder model is trained using a siamese framework, where a positive or negative pair of images is sent to the encoder. The temporal encoder generates a 16-dimensional embedding for both images, concatenates them, and feeds them combined embedding to a single fully-connected layer for binary classification as a positive or negative pair. We use binary cross-entropy loss function and Adam optimizer to compile the model [10].

The trained temporal encoder is then used for the wound stage estimation. A dropout layer (with dropout rate=0.5) and a fully-connected layer with a softmax activation function are added to the pre-trained temporal encoder for the wound stage classification step. We use categorical cross-entropy as the loss function and the Adam optimizer with the learning rate of 0.01.

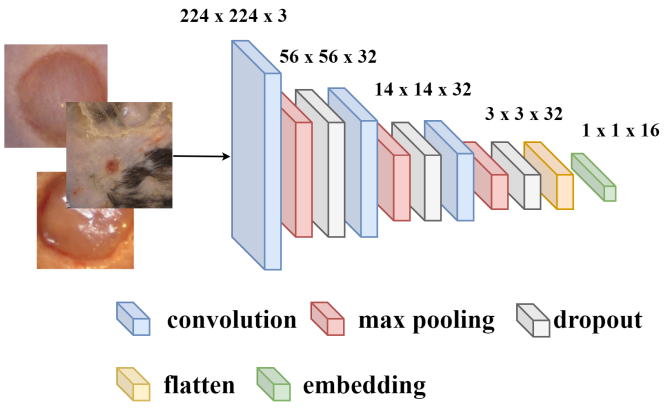

Fig. 3. The convolutional temporal encoder architecture consists of 3 convolutions, pooling, and dropout layers. The output of the temporal encoder is a 16-dimensional temporal embedding learned during the pre-training.

B. Cross-Layer Attention Temporal Encoder

In the CNN model shown in Fig. 3, the first two convolutional layers are considered the middle layers, and the last

convolutional layer is considered the top-level layer. We apply a CLCA module to the second middle and top-level layers. Further, the attention achieved by the CLCA module is fed to a CLSA module along with the attention obtained by the top-level layer, as shown in Fig. 4. The CNN with cross-layer attention is used as the temporal encoder during the pre-training task. The output of each of the branches is a 16-dimensional embedding used to perform binary classification of a pair of images. We use the pre-trained cross-layer attention CNN encoder for the wound stage classification step. The output of each of the branches in the architecture is individually used to predict the wound healing stage. The predictions are combined to give a final output using a softmax activation function.

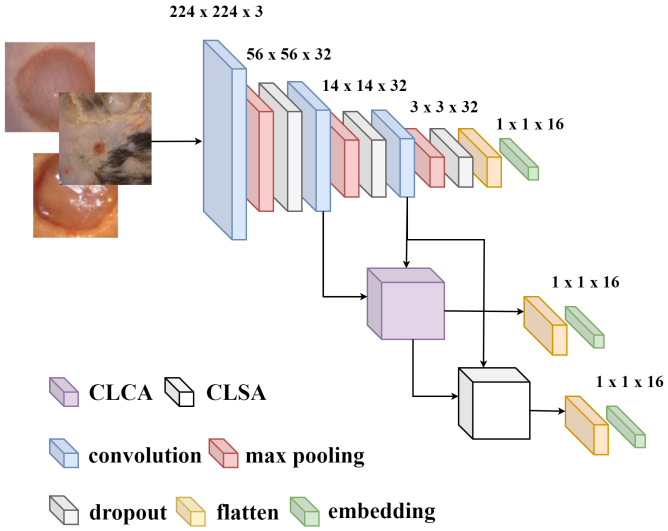

Fig. 4. Cross-layer attention is applied to the convolutional temporal encoder. The cross-layer attention temporal encoder consists of the CLCA and CLSA modules that generates two additional 16-dimensional embeddings.

In the CLCA module (shown in Fig. 5), we first up-sample the top layer feature maps to match them with the respective middle layer feature maps. A self-attention on the middle layer feature maps and the up-sampled top layer feature maps is applied. These self-attention results are combined to obtain enhanced mid-level feature maps with global spatial information. Finally, a weighted sum is performed on the self-attention feature maps. The output of the CLCA module is the refined mid-level feature maps rich in a global spatial context.

The CLSA module is used to improve the top layer's ability to capture local information. We apply average and max pooling simultaneously on the output of the CLCA module as shown in Fig. 6. The output is then processed via a convolutional layer and then downsampled to match the top layer feature map. Finally, attention is applied to the improved downsampled mid-level feature map and the top-level feature map obtained from the top layer.

IV. Results

The convolutional temporal encoder achieves a validation accuracy of 90.4% and a test accuracy of 91.5% when trained

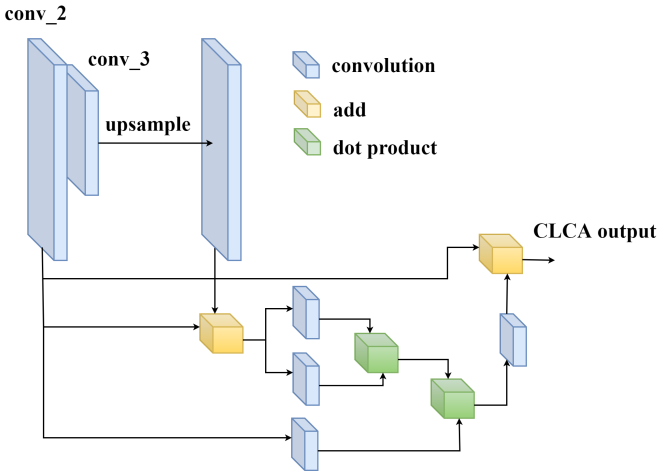

Fig. 5. The structure of the CLCA module used to improve the global context at the intermediate layer using the feature information at the top layer.

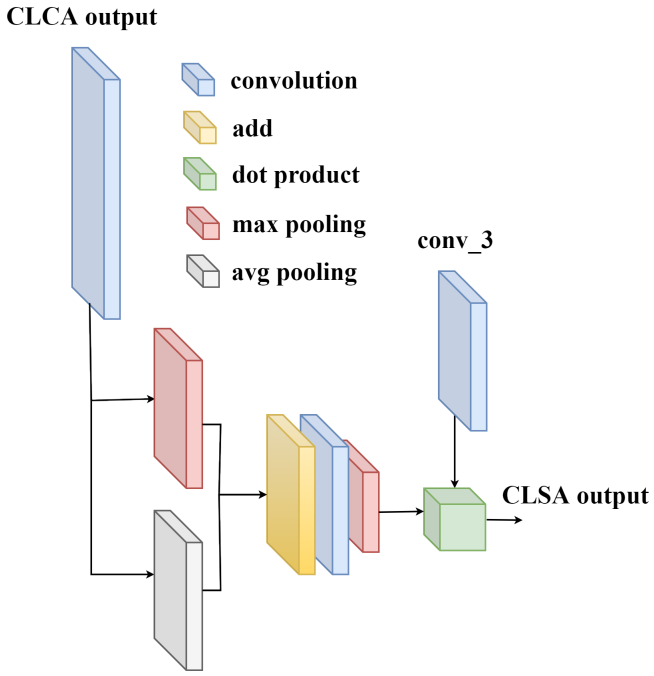

Fig. 6. Illustration of the CLSA module architecture utilizing the CLCA output to enhance the spatial attention maps from the final layer.

on the pair image dataset. The test accuracy for the wound stage classification task performed using the CNN encoder pre-trained on a single image dataset achieves a validation accuracy of 78.1% and a test accuracy of 81.3%. The network uses the Adam optimizer with the learning rate of 0.01 for the pre-training task. We pre-train the temporal encoder for 25 epochs and then retrain it for 100 epochs at a lower learning rate of 0.0001 for the wound stage classification task.

The cross-layer attention encoder with CNN as the backbone model on evaluation performs well with a validation accuracy of 91.3% and a test accuracy of 87.3%. However, the classification performed using cross-layer attention achieves a

validation accuracy of 82.4% and a test accuracy of 75.0% on the stage estimation task. To closely evaluate the performance of the two methods, the hyper-parameters used for the convolutional and cross-layer attention temporal encoder remain the same. Other related metrics are presented in Table I.

TABLE I
PERFORMANCE METRICS

Model Name	**Evaluation**		
	Dataset	***Accuracy***	***Loss***
CNN Temporal Encoder	Training	89.9%	0.22
	Validation	90.4%	0.23
	Test	91.5%	0.26
Wound Stage Classifier (pretrained CNN Encoder)	Training	82.2%	0.51
	Validation	78.1%	0.89
	Test	81.3%	0.56
Cross-Layer Attention Temporal Encoder	Training	96.8%	0.08
	Validation	91.3%	0.42
	Test	87.3%	0.79
Wound Stage Classifier (pre-trained Cross-Layer Attention Encoder)	Training	89.5%	0.36
	Validation	82.4%	0.53
	Test	75.0%	0.16

Visualizations from the conv_2 layer of the CNN-based encoder show that the model focuses more on specific wound regions by applying local attention. It is observed that the attention maps from the conv_2 focus on the skin region around the wound (at the corners) along with some parts of the wound. In such a case, the area around the wound may or may not be relevant and may not contain the essential information required for accurate wound stage classification, as shown in the maturation phase in Fig. 7. On the other hand, the attention map obtained from the cross-layer attention model shows that the model learns spatial and context information about the image and focuses on the healed wound area. A similar determination can be made from the hemostasis phase from the activation map shown in Fig. 7. In the second inflammatory stage, the wound appears saturated with a glossy region, and the convolutional model is able to determine the circular area of the wound. However, it misses other regions which are a part of the wound. The cross-layer spatial module is shown to capture the features on the circumference of the wound, which indicates the wound healing. The saliency map from the conv_9 layer shows that almost the entire portion of the wound is highlighted as expected.

V. CONCLUSIONS

We compare local and non-local attention methods for fine-grained visual classification of wound healing stages. Experimental results obtained from both techniques show that cross-layer methods enable more attention to the exposed wound region. It is evident that the cross-layer context attention module enhances the global context for the intermediate layer by utilizing the top-level feature maps and the cross-layer spatial attention module further improves the spatial feature extraction at the top layer. Saliency map visualizations show the effectiveness of adding cross-layer attention to the convolutional neural network in classifying the wound healing stage.

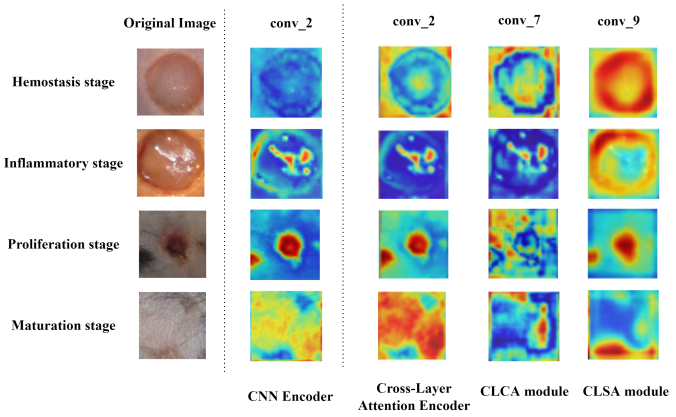

Fig. 7. Visualizations of attention maps for the wound stage classification task of the 4 stages of wound.

REFERENCES

[1] S. a. Guo and L. A. DiPietro, "Factors affecting wound healing," Journal of dental research, vol. 89, no. 3, pp. 219–229, 2010.

[2] Héctor Carrión et al., "HealNet - self-supervised acute wound heal-stage classification," Medical image computing and computer assisted intervention–-MICCAI 2022: 25th International Conference, Singapore, September 18– 22, 2022.

[3] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[4] R. Huang, Y. Wang, H. Yang, "Cross-layer attention network for fine-grained visual categorization", The 8th Workshop on Fine-Grained Visual Categorization (FGVC8) .

[5] J. Bromley, I. Guyon, Y. LeCun, E. S¨ackinger, and R. Shah, "Signature verification using a siamese time delay neural network," Advances in neural information processing systems, vol. 6, 1993.

[6] R. Behrouz, et al. "Multiclass wound image classification using an ensemble deep CNN-based classifier." Computers in Biology and Medicine 134 (2021): 104536.

[7] Y. Hsin-ya, B. Michelle, C. Hector, I. Rivkah (2022), Photographs of 15-day wound closure progress in C57BL/6J mice, Dryad, Dataset, https://doi.org/10.25338/B84W8Q.

[8] A. K. Mokin, A. V. Gayer, A. V. Sheshkus, V. L. Arlazarov, "Auto-clustering pairs generation method for Siamese neural networks training," Proc. SPIE 12084, Fourteenth International Conference on Machine Vision (ICMV 2021), 120841A (4 March 2022); https://doi.org/10.1117/12.2623139.

[9] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, and D. Mahajan, "Clusterfit: Improving generalization of visual representations," in Proceedings of the IEEE/CVF.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.