DEEP FEATURES FUSION WITH MUTUAL ATTENTION TRANSFORMER FOR SKIN LESION DIAGNOSIS

Li Zhou *Yan Luo**

Department of Electrical and Computer Engineering
University of Massachusetts Lowell

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506211

ABSTRACT

Early skin lesion diagnosis is crucial to prevent skin cancer, and deep learning (DL) based methods are well exploited to support dermatologists' diagnosis. The data for the diagnosis tasks include dermoscopic lesion images and textual information. It is a challenge to learn features from the multimodal data to improve diagnostic quality. Inspired by the vision and language integration models in Visual Question Answer (VQA), we present an end-to-end neural network model for skin lesion diagnosis using both images and textual information simultaneously. Specifically, we fine-grained features from the two modalities (image and text) of the dataset by the pre-trained DL models. We propose a novel approach named Mutual Attention Transformer (MAT), which consists of self-attention blocks and guided-attention blocks, to enable the interactions between the features from both modalities concurrently. We then develop a fusion mechanism to integrate the represented features before the final classification output layer. The experimental results on the HAM10000 dataset demonstrate that the proposed method outperforms the state-of-art methods for skin lesion diagnosis.

Index Terms— skin lesion classification, attention mechanism, transformer, deep learning

1. INTRODUCTION

The incidence of skin cancer has led to a major public health problem, and both melanoma and non-melanoma skin cancer (NMSC) bear significant morbidity. Early detection and diagnosis of skin cancer are practical ways to increase a survival rate [1]. And they are possible through inspection and analysis of pigmented skin lesions with the help of dermoscopy. Dermoscopy is an imaging technique that eliminates the surface reflection and strength the visualization of deeper skin [2]. Furthermore, computer-aided analyses have shown impressive performance in supporting dermatologist's diagnosis. They mainly make use of dermoscopic lesion images to segment or identify skin lesions. Non-imaging data, such as genetic data and textual data (e.g. sex, age), are usually taken as supplemental information to the image data. It explains the features of images and figure out the relationship among subjects.

Deep Learning (DL) is an efficient assistant for diagnosis and has achieved high performance in practice [3]. Convolutional neural networks (CNNs) are a major type of neural network composed of one or more convolutional layers for local information extraction. It is used heavily in the field of Computer Vision (CV) like image classification and segmentation. Modern very deep CNNs lead to an efficient learning of the input images, such as VGGNet [4], GoogLeNet [5], ResNet [6], etc. They are taken as rich feature extractors to deal with image recognition and other advanced tasks. Recurrent neural networks (RNNs) are another type of neural network with a "memory" to fed information from the previous step to the current. The classic architecture LSTM [7] and its related networks are effectively in Natural Language Processing (NLP) tasks, such as text classification, text generation, semantic representations, and others. Significant progress in DL is about the integration of vision and language, which is applied in the field of Visual Question Answering (VQA). The challenge of the research is about how CV and NLP models interact so that the tasks understand the visual and linguistic information comprehensively. Researchers propose various methods towards jointly learning representations for improving the efficiency of the vision-language tasks. Attention mechanisms [8] are powerful to describe the content of inputs. Moreover, fusion strategies are also critical to the tasks. For example, [9] use a simple element-wise product to merge two vectors in VQA. [10, 11] follow co-attention frameworks to learn visual and textual features simultaneously, and use concatenation and/or sum to fuse multimodal features.

Inspired by the above mentioned model interaction and fusion, this work focuses on classifying skin lesions by using both the dermoscopic images and the corresponding metadata. We propose a novel neural network by introducing a new transformer, termed Mutual Attention Transformer (MAT), to make complementary attention between the imaging content and the textual content. Motivated by the state-of-the-art attention mechanism [8], named the Transformer, which consists of encoder and decoder stacks, we re-design the transformer composed of self-attention (SA) blocks

This work is supported in part by NSF No. 1547428 and 1738965.

and guided-attention (GA) blocks. The SA block depicts self-interactions (i.e., image-to-image or text-to-text) from fine-grained features, while the GA blocks depict guided-interactions (i.e., image-to-text or text-to-image). Pre-trained deep CNN and embedding models are used to represent imaging and textual contents respectively ahead of MAT. Then an adaptive fusion mechanism is introduced, which composite attended features comprehensively, thus the output can be fed into the classifier in the network. The main contributions are summarized as follows: (1) we introduce MAT and an efficient fusion mechanism for feature interactions; (2) we model a novel network that can complementarily learn features from both imaging and textural contents that are applied in the field of dermatoscopic diagnosis; and (3) we conduct extensive experiments on the open benchmark dataset and achieve impressive performance over state-of-art methods.

2. FEATURE FUSION WITH MAT

In this section, we introduce the MAT based neural network for complementary learning of the multimodal data for the classification task. The overview of the proposed architecture is illustrated in Fig. 1. The MAT network accepts the imaging and textual data and represents them by pre-trained models as fine-grained features respectively, followed by the transformer unit with the MAT to obtain the attended multimodal features simultaneously and then the fusion unit to composite the features, yielding integrated features projected into the classifier in the output unit.

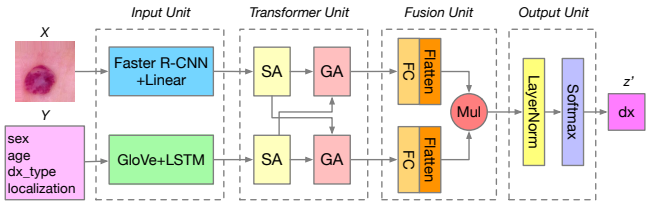


Fig. 1. The architecture of the mutual attention transformer (MAT) based neural network for skin lesions diagnosis. X and Y denote the image and metadata inputs respectively; z' denotes the predicted diagnostic category (dx) of the image.

2.1. Multimodal Data Representation

In the input unit, we use two pre-trained deep models to extract imaging and textual features in parallel. For the input images, unlike the general processing that segments the lesion from each dermoscopic image at the beginning, we intend to keep the information of skin surrounding a lesion. Thus, we apply a pre-trained model named Faster R-CNN (use ResNet-101 as its backbone) [12] to extract features from the images. It is the bottom-up mechanism that proposes a set of regional features. In detail, we truncate the pre-trained model with the first 8 layers and keep the top-n ($n = 100$) regional features of dimension d, further use a linear layer as a pre-processing, and result in a feature matrix $X \in \mathbb{R}^{n\times d}$ of each image. For the pairwise metadata with multiple words, we first pad with a maximum of 10 words to deal with the missing features. Then, we use the pre-trained GloVe [13] weights to transform the textual data with embedding dimension d', then the word embeddings are fed into a 1-layer LSTM with d hidden size, and result in a unified textual feature matrix $Y \in \mathbb{R}^{n\times d}$ of the corresponding image.

2.2. Mutual Attention Transformer

2.2.1. Attention Blocks

Scaled Dot-Product Attention [8] is the core component of the attention mechanism. It maps a query and a key-value pair to the attended features of the query. In practice, a set of queries and key-value pairs are packed into matrices as Q, K, V with the same dimensions $\mathbb{R}^{n\times d}$ by padding respectively, where n is the number of inputs and d is the dimensionality of the input features. The attended features are computed as:

$$Att(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where $\frac{1}{\sqrt{d}}$ is the scaling factor. Note that, we mask out padding values by setting them to $-\infty$ followed behind the scaling step to overcome the underflow problem [11] and implement dropout [14] after the softmax step to avoid overfitting.

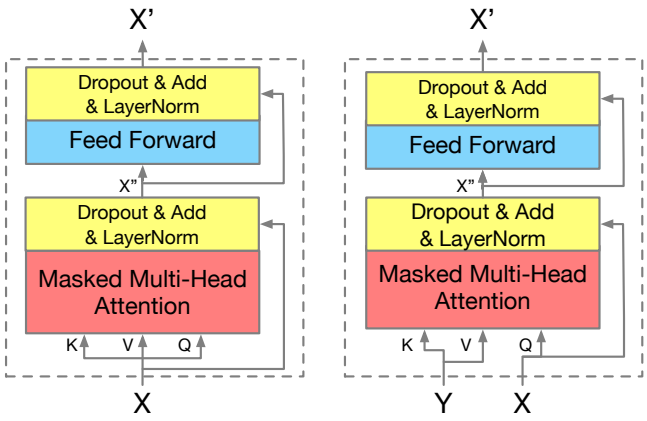
Further, based on the core attention, Multi-Head Attention (MHA) [8] is introduced to improve the attended features. In detail, it uses h scaled dot-product attention layers (denote as 'heads') running with different linear projections for inputs, yielding the attended features of each layer, and concatenates them by weighted function to result in the jointly features. Functions are listed as follows:

$$\begin{aligned} MultiAtt(Q,K,V) &= Concat\left(H_1,...,H_h\right)W^o \\ where\ H_i &= Att\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_o}$ are the weighted matrices of the i-th head, $d_o = d/h$ is the dimension of the attended features of each head, and $W_i^Q \in \mathbb{R}^{hd_o\times d}$ is the weight matrix for concatenate h heads.

Inspired by the encoder-decoder strategy of the transformer [8], we derive two basic attention blocks as shown in Fig. 2 to resolve the attended features for skin lesions diagnosis, i.e. self-attention (SA) block and guided-attention (GA) block. Uniformly, both SA and GA are two layers, namely masked MHA and the feed-forward network (FFN). FFN contains a linear transformation (FC($4d$)) accompany by a rectified linear unit function (ReLU) and a Dropout, followed by another FC(d). The output of each layer is processed in sequence of a Dropout, a shortcut connection [6] and a layer

normalization [15]. The difference between the two blocks is that SA takes only one modal of input features while GA takes multimodal features. Note that, the dimensionality of the input features (X) of an attention block is equal to the dimensionality of the output features (X').



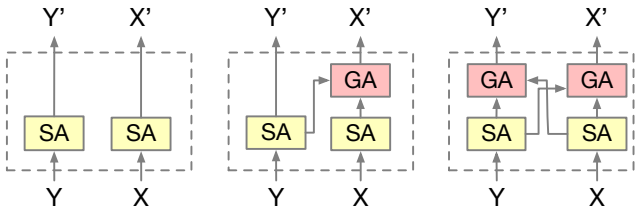
(a) Self-Attention (SA) (b) Guided-Attention (GA)

Fig. 2. Two basic attention blocks with masked multi-head attention. X and Y denote the different modalities of input features; X'' denotes the intermediate features outputted from the MHA; X' denotes the attended features.

2.2.2. *Structured Transformer*

Following the architecture shown in Fig. 1, we first take the network without the transformer unit as a baseline, i.e. input features are directly passed through the transformer unit with the identity mapping. Then we structure three types of transformer units to deal with the multimodal problem.

The proposed transformers, as depicted in Fig. 3, are built on the composition of the SA and/or GA. Intuitively, the attended imaging and textual features are concurrently obtained by an attention block separately. Thus, we introduce the SA-SA transformer as shown in Fig. 3 (a). The extracted imaging features ($X \in \mathbb{R}^{n\times d}$) are fed into a SA block result in a group of inner-attended features ($X' \in \mathbb{R}^{n\times d}$), and the same workflow for the textual features. Compared with the SA-SA transformer, the SA-SGA in Fig. 3 (b) places a GA block on top of a SA block. The transformer can learn the guided attention over the inner-attended features. In detail, both the principal features outputted from a SA block and the guided features ($X'' \in \mathbb{R}^{n\times d}$) outputted from the first layer (MHA) of another SA block are passed through a GA block to model guided-attended features (X'). There are two versions of the transformer, i.e. SA(textual)-SGA(imaging) and SA(imaging)-SGA(textual). The SGA-SGA transformer in Fig. 3 (c) is the symmetric module designed for the mutual attention between the two inputs. The transformer loads a GA block on top of each SA block. Both of the original features (X and Y) are self-attended concurrently through SA blocks and the outputs are guided with each other across GA blocks in the meantime. Note that, all the transformers work without feature dimension reduction because they are cascaded stacked with attention blocks.



(a) SA-SA (b) SA-SGA (c) SGA-SGA

Fig. 3. Three transformer units for feature representation.

2.3. Multimodal Fusion and Output

Followed by the transformer unit, we propose a multimodal fusion algorithm to incorporate relations between the two attended features ($X', Y' \in \mathbb{R}^{n\times d}$). At first, they are linearly and parallelly projected into a Fully Connected layer (FC(d)) accompanied by a ReLU and a Dropout to get their transformed representation. Next to FC a sum function used to flat the dimensionality of both modalities ($\tilde{X}, \tilde{Y} \in \mathbb{R}^{d}$). We opt for an element-wise multiplication of both modalities as the last step of the feature integration, result in the fused feature $z \in \mathbb{R}^{d}$. The output unit is a chain of operations following a layer normalization, a Dropout, a FC and a Softmax layer. The fused features are transformed into a probability vector $z' \in \mathbb{R}^{C}$, where C is the number of the classes.

3. EXPERIMENTS

3.1. Dataset and Setup

We evaluate our method on the benchmark dataset HAM10000 [16], which consists of 10015 dermatoscopic images of pigmented lesions. There are 7 diagnostic classes (0:bcc, 1:df, 2:mel, 3:bkl, 4:nv, 5:akiec, 6:vasc) and 4 attributes (sex, age, diagnostic type and localization). We do a stratified split on the non-duplicated lesions of HAM1000 and split the dataset in an 80%-10%-10% fashion of training/validation/test datasets. Since the datasets are very uneven (0.67 of nv vs. 0.01 of df), we augment data by rotation, shifting, flipping, and resizing to make the datasets balanced.

3.2. Implementation Details

Following the flowchart of the MAT network in Fig. 1, the hyper-parameters are listed as follows. Over the MAT network, the number of heads h is 8 in each attention block, the hidden size d is 512, and the dropout rate is 0.1. For training, the batch size is 64, the base learning rate is 10^{-4} with decay ratio equals to 0.2, and the optimizer is Adam algorithm

[17] with $\beta = (0.9, 0.98)$ and $\epsilon = 10^{-9}$. Also, early stopping is employed to avoid the overfitting while training the network. It records the validation loss with the maximum of 100 epochs, and the patience is set to 20. Losses are measured by the Cross Entropy function because of the single-label categorical problem. Predicted results are scored using the normalized accuracy classification score (ACC), the label ranking average precision (LRAP) and the macro area under the receiver operating characteristic curve (AUC) metrics. Our framework was implemented in Python3 with Pytorch1.7.1 and Keras2.24 library.

3.3. Experimental Results

MAT Network For the network in Fig. 1, we obtain our best results for the diagnostic task. The training process returned the early stopping checkpoint on epoch 11. which is shown in Fig. 4 with 91.64% validation accuracy and 1.25 validation mean loss. Results on the test dataset are 92.55% ACC, 95.63% LRAP, and 98.28% AUC. Fig. 5 shows the macro-average ROC curve and AUC values for each class. Each curve represents the performance in distinguishing classes of lesions. It indicates that the model has a balanced and competitive capacity for the multi-class classification.

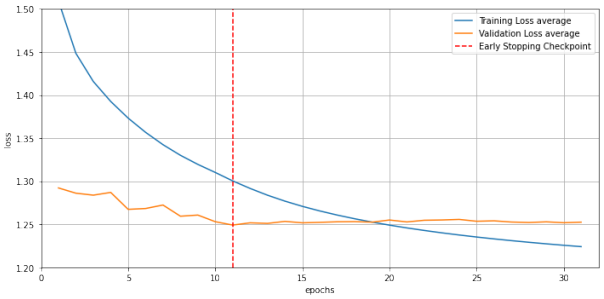


Fig. 4. Model loss.

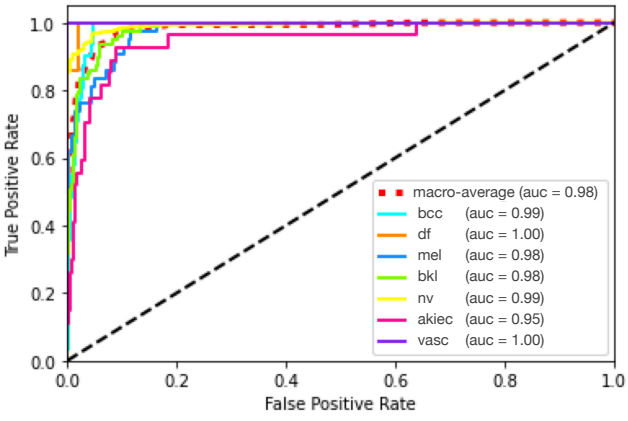


Fig. 5. Receiver Operating Characteristic (ROC) curve.

Transformer Units We compare the transformer unit within the MAT network with other proposed transformers.

Table 1. Ablation study for the proposed MAT network with different transformer units. CKPT: early stopping checkpoint; metrics ACC, LRAP and AUC are evaluated on the test dataset.

Model	CKPT	ACC	LRAP	AUC
Base	13	0.9154	0.9514	0.9823
SA-SA	31	0.9194	0.9520	0.9808
SA-SGA	32	0.9104	0.9434	0.9673
SGA-SGA	11	0.9255	0.9563	0.9828

Table 1 outlines the ablation experiments validating the attentional mechanism and the choice of the transformers. The base model indicates the network trained without the transformer unit and yields 91.54% ACC. From the table, we can see that introducing SGA-SGA outperforms the base model on all metrics. It shows the effectiveness of the mutual attention transformer. The MAT network also outperforms the other models that replace the transformer unit with SA-SA or SA-SGA. All the models have a high chance ($> 95\%$) to distinguish positive and negative classes based on AUCs. It is interesting to note that SA-SGA downgrades the performance of the model compared with the base one, so the transformer technique should be carefully applied to a multimodal problem. Also, the network on skin image only yields a lower ACC (92.47%) compared with the two modalities task.

Sate-of-the-art Comparison In table 2, we compare the MAT network against the state-of-the-art methods on the HAM10000 dataset. Our solutions comprehensively outperform the others in terms of the list metrics. We do not list all methods because of space limitations.

Table 2. Comparisons on HAM10000.

Methods	Year	ACC	AUC
Multi-model[18]	2019	89.80%	0.98
MobileNet[19]	2019	92.70%	0.96
DenseNet[20]	2020	85.80%	0.88
Semi-supervised[21]	2020	92.54%	0.94
Ours	2021	92.55%	0.98

4. CONCLUSIONS

In this paper, we proposed a novel MAT neural network to comprehensively learn features from the multimodal data for skin lesions diagnosis. Specifically, we design a transformer unit composed of SA blocks and GA blocks, which depicts self-attended features and guided-attended features concurrently. With the fusion unit and the output unit, multimodal features are integrated and result in a predicted vector. Results validate the improved performance for the diagnosis.

5. REFERENCES

[1] Goutam Kumar Jana, Anshita Gupta, Arpita Das, Ramasish Tripathy, and Prasenjit Sahoo, "Herbal treatment to skin diseases: A global approach.," *Drug Invention Today*, vol. 2, no. 8, 2010.

[2] Sameena Pathan, K Gopalakrishna Prabhu, and PC Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—a review," *Biomedical Signal Processing and Control*, vol. 39, pp. 237–262, 2018.

[3] Ramsha Baig, Maryam Bibi, Anmol Hamid, Sumaira Kausar, and Shahzad Khalid, "Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images-a review," *Current Medical Imaging*, vol. 16, no. 5, pp. 513–533, 2020.

[4] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[9] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[10] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Hierarchical question-image co-attention for visual question answering," *arXiv preprint:1606.00061*, 2016.

[11] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 6281–6290.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[13] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[16] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, pp. 180161, 2018.

[17] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] Muhammad Attique Khan, Muhammad Younus Javed, Muhammad Sharif, Tanzila Saba, and Amjad Rehman, "Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification," in *2019 international conference on computer and information sciences (ICCIS)*. IEEE, 2019, pp. 1–7.

[19] Ensaf Hussein Mohamed and Wessam H El-Behaidy, "Enhanced skin lesions classification using deep convolutional networks," in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2019, pp. 180–188.

[20] Hsin-Wei Huang, Benny Wei-Yun Hsu, Chih-Hung Lee, and Vincent S Tseng, "Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers," *The Journal of Dermatology*, 2020.

[21] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Transactions on Medical Imaging*, 2020.