Skin Cancer Classification using Soft Attention and Metadata

Hoang Khoi Do

khoi.dh200322@sis.hust.edu.vn

Ha Noi University of Science and Technology

March 17, 2022

Abstract

Recent estimates are that about 150 million children under five years of age are stunted, with substantial negative consequences for their schooling, cognitive skills, health, and economic productivity. Therefore, understanding what determines such growth retardation is significant for designing public policies that aim to address this issue. We build a model for nutritional choices and health with reference-dependent preferences. Parents care about the health of their children relative to some reference population. In our empirical model, we use height as the health outcome that parents target. Reference height is an equilibrium object determined by earlier cohorts' parents' nutritional choices in the same village. We explore the exogenous variation in reference height produced by a protein-supplementation experiment in Guatemala to estimate our model's parameters. We use our model to decompose the impact of the protein intervention on height into price and reference-point effects. We find that the changes in reference points account for 65% of the height difference between two-year-old children in experimental and control villages in the sixth annual cohort born after the initiation of the intervention.

Keywords: AI-enabled computer-aid diagnosis, Diagnosis, Skin Sancer, Skin Lesion Classification, Artificial Intelligence, Deep Learning, Machine Learning

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my teacher, Dr.Nguyen Viet Dung() who gave me the golden support to do this wonderful project. He gave a chance to use High GPU computing computer for AI Training. Otherwise, he also gave me recommendation on how to implement experiment to make a good conclusion such as what I should focus on, what I need to investigate, which metrics I should consider.

1 Introduction

Skin cancer is one of the most common cancer leading causes of death worldwide. Every day, more than 9500[1] people in the United States are diagnosed with skin cancer. Otherwise, 3.6[1] million people are diagnosed with basal cell skin cancer each year. According to the Skin Cancer Foundation, the global incidence of skin cancer continues to increase[2]. In 2019, it is estimated that 192,310 cases of melanoma will be diagnosed in the United States[2]. On the other hand, if patients are early diagnosed, the survival rate is correlated with 99%. However, once disease progresses beyond the skin, survival is poor[2]. Moreover, with the increasing incidence of skin cancers, low awareness among a growing population, and a lack of adequate clinical expertise and services, there is a need of effective solution.

Recently, deep learning particularly, and machine learning in generally algorithms have emerged to achieve excellent performance on various tasks, especially in skin disease diagnosis tasks. AI-enabled computer-aided diagnostics (CAD) has solutions in three main categories: Diagnosis, Prognosis and Medical Treatment. Medical imaging, including ultrasound, computed tomography, and magnetic resonance imaging, and X-ray image is used extensively in clinical practice. In Diagnosis, Artificial Intelligence (AI) algorithms are applied for disease detection to save progress execution before these diagnosed results are considered by a doctor. In Prognosis, AI algorithms are used to predict the survival rate of a patient based on his/her history medical data. In Medical Treatment, AI models are applied for building solution for a specific disease, medicine revolutionize is an example. In various studies, AI algorithms has provided various end-to-end solutions in the detection of abnormalities such as breast cancer, brain tumors, lung

cancer, esophageal cancer, skin lesions, and foot ulcers across multiple image modalities of medical imaging[2].

In order to adapt the increase in skin cancer case, AI algorithms over the last decade has a great performance. Some typical models that can be mentioned are DenseNet[3], EfficientNet[4], Inception[5], MobileNets[4][6][7], ResNet[8][9], and NasNet[10]. Some of these models have been used as a backbone model in other studies that I will discuss more in the Related Work section.

In this paper, I will analyze the effect of the metadata on classifying skin disease. On the other hand, by analyzing the combination of several backbone models, I will also construct an optimized model that has ability to classify in a balanced way between classes instead of well identifying the majority of classes.

2 Related Work

3 Data

3.1 Image Data

The dataset used in this paper is the HAM10000 dataset published by Havard University Dataverse[11]. There are total 7 classes in this dataset containing Actinic keratoses and intraepithelial carcinoma or Bowen's disease (AKIEC), Basal cell Carcinoma (BCC), benign keratosis-like lesions (solar lentigines / seborrheic keratoses andchen-planus like keratoses, BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, VASC). The distribution of the dataset is shown in the table below:

Class	AKIEC	BCC	BKL	DF	MEL	NV	VASC	Total
No. Sample	327	514	1099	115	1113	6705	142	10015

More than 50 percent of lesions are confirmed through histopathology (HISTO), the ground truth for the rest of the cases is either follow-up examination (FOLLOWUP), expert consensus (CONSENSUS), or confirmation by in-vivo confocal microscopy (CONFOCAL). On the other

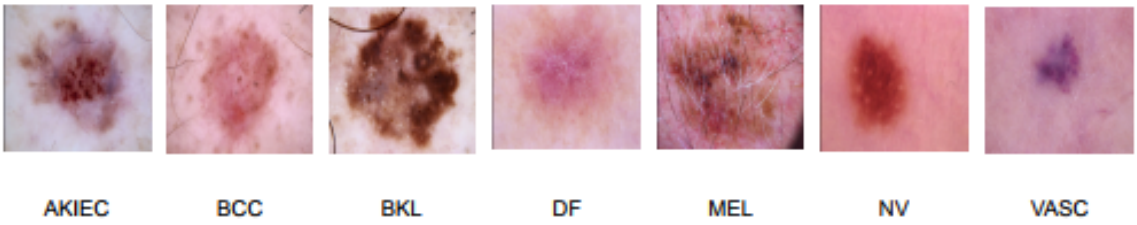

Fig. 1. Example image of each class

hand, before being used for training the whole data is shuffled then split into two part. 90 percent and 10 percent of the data is used for training and validating respectively.

In the previous paper[1], the image data is augmented for all class, the number of image increase to 18015 images. Since, this data is imbalanced, using augmented data may cause the problem of well classify on the majority of class. In this paper, instead of augmenting data, metadata is used. The way of processing metadata is discuss in MetaData section. Images in this dataset has the type of *RGB* and shape of (450, 600). However, Each backbone need the different input size of image as well as the range of pixel value. DenseNet201[3] require the input pixels values are scaled between 0 and 1 and each channel is normalized with respect to the ImageNet dataset. In Resnet50 and Resnet152[8][9], the images are converted from *RGB* to *BGR*, then each color channel is zero-centered with respect to the ImageNet dataset, without scaling. InceptionResNetV2[12], on the other hand, will scale input pixels between -1 and 1. Similarly, three versions of MobileNet[4][6][7], NasNetMobile and NasNetLarge[10] require the input pixel is in range of -1 and 1.

3.2 Metadata

The HAM10000 dataset[11] also contain the metadata of patient including gender, age, and the capturing position. During the data exploration term, I found out that the age category miss 57 data point, then I decided to remove this 57 samples. In the gender and capturing position category contain some samples of unknown. Instead of removing, these unknowns data point is kept and considered as "prefer not to say". Besides, the label of the whole data is preprocessed into one-hot vector.

4 Model Schema

4.1 Input Schema

Using metadata as another input is not new. In paper[13], they decide to keep the missing value and set its value to 0. The sex and anatomical site are categorical encoded. The age, on the other hand is numerical normalized. After processing, the metadata is fed into a two-layer neural network with 256 neurons each. Each layer contains batch normalization, a ReLU[14] activation and dropout with $p = 0.4$. The network's output is concatenated with the CNN's feature vector after global average pooling. Especially, they use a simply data augmentation strategy to address the problem of missing values in metadata. During training, they randomly encode each property as missing with a probability of $p = 0.1$.

In this paper, the unknowns is kept as a type as discussed in Metadata section. Sex, anatomical site and age are also category encoded and numerical normalized, respectively. After processing, the metadata is then concatenated and fed into a dense layer of 4096 neurons. Finally, this this dense layer is then concatenate with the output of Soft-Attention which is then discussed in Soft-Attention section.

The Input schema is described as follow:

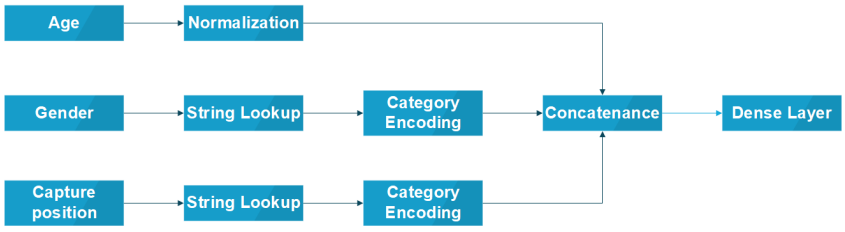

Fig. 2. Input Schema

Image data, on the other hand after being preprocessed, is fed directly into the backbone model.

4.2 Soft-Attention

Applying Soft-Attention layer in deep learning is not a new approach. Soft-Attention have been used in various application: image caption generation in [15] and handwriting verification in

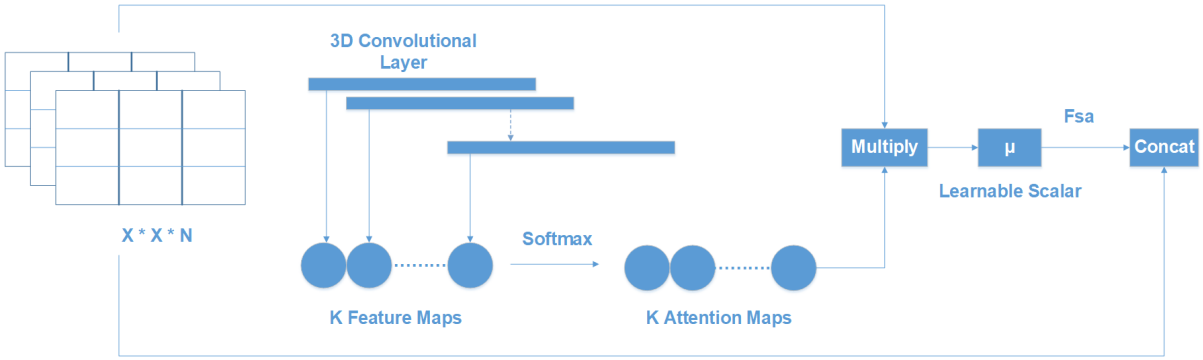

Fig. 3. Soft-Attention Module

[16] respectively. In skin lesion classification, Soft-Attention is used to increase the performance of the model as described in [1]. Soft-Attention has ability to ignore irrelevant areas of the image by multiplying the corresponding feature maps with low weights. The function below describe the flow of Soft-Attention module:

$$f_{sa} = \gamma t \sum_{k=1}^{K} softmax(W_k * t)$$

In order to apply Soft-Attention, there are two main steps. Firstly, the input tensor is put in a grid-based feature extraction from the high-resolution image, where each grid cell is analyzed in the whole slide to generate a feature map[17]. This feature map called $t \in R^{h \times w \times d}$ where $h, w, and d$ is the shape of tensor generate by a Convolution Neural Network(CNN), is then input to a 3D convolution layer whose weights is $W_k \in R^{h \times w \times d \times K}$. The output of this convolution is normalized using softmax function to generate K = 16 attention maps. These 16 attention maps are aggregated to produce a weight function called α. This α function is then multiplied with feature tensor t and scaled by γ, a learnable scalar. Finally, the out of Soft-Attention function f_{sa} is the concatenation of the beginning feature tensor t and the scaled attention maps. In this paper, the Soft-Attention layer is applied in the same way in this paper[1]. The Soft-Attention module is described in the following diagram:

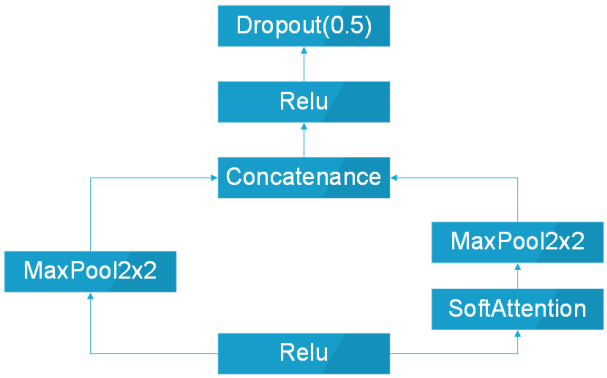

Fig. 4. Soft-Attention Module

4.3 Backbone Model Architecture

In this paper, the backbone models that have been used are DenseNet201[3], Inception[5], MobileNets[4][6][7], ResNet[8][9], and NasNet[10]. The combination of DenseNet201, InceptionResNetV2 and Soft-Attention layer are both test by the previous paper[1] with a great performance. Otherwise, Resnet50 also well classify but with much less number of parameter and depth than based on its f1-score and precision stated. Therefore, in this paper, I will analyze the performance of the model Resnet152 and NasnetLarge which has the larger number of parameter and depth. On the other hand, three version of MobileNet and the NasnetMobile will also be analyzed which has a small number of parameter and depth.

Model	Size(MB)	Parameters	Depth
Resnet50	98	25.6M	107
Resnet152	232	60.4M	311
DenseNet201	80	20.2M	402
InceptionResNetV2	215	55.9M	449
MobileNet	16	4.3M	55
MobileNetV2	14	3.5M	105
MobileNetV3Small	Unknown	2.5M	88
MobileNetV3Large	Unknown	5.5M	118
NasnetMobile	23	5.3M	308
NasnetLarge	343	88.9M	533

4.4 Model

The whole architecture of the model used for image feature extraction is applied in the same way in paper [1]. Metadata branch, otherwise is preprocessed before feeding into a dense layer then concatenate with the output of Soft-Attention layer. It is described in the figure below:

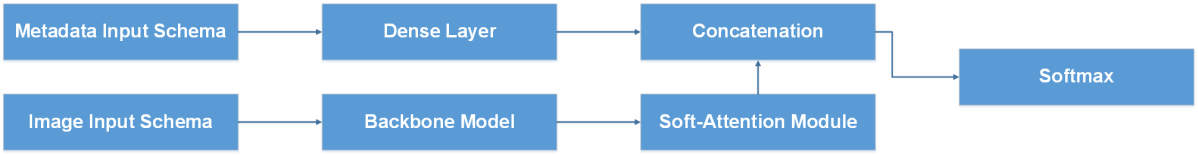

Fig. 5. Overall Model

5 Training

5.1 Loss Function

The loss function used in this paper is categorical cross-entropy. Consider $X = [x_1, x_2, \ldots, x_n]$ as the input feature, $\theta = [\theta_1, \theta_2, \ldots, \theta_n]$. Let N, and C is the number of training examples and number of class respectively. The categorical cross-entropy loss is presented as:

$$L(\theta, x_n) = -\frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N} y_n^c \log(\hat{y}_n^c)$$

where $\hat{y}_i^c$ is the output of model and y_i^c is the target that the model should return.

Since the dataset face the imbalanced problem then I applied the class weight for the loss. This formula below is used to calculate the class weight:

$$W = N \odot D$$

$$D = \left[\frac{1}{C\times N_1} \quad \frac{1}{C\times N_2} \quad \cdots \quad \frac{1}{C\times N_n}\right] = \frac{1}{C} \odot \left[\frac{1}{N_1} \quad \frac{1}{N_2} \quad \cdots \quad \frac{1}{N_n}\right]$$

where N is the number of training sample, C is the number of class, N_i is the number of sample in each class i. D is the matrix contain the inverse of $C \times N_i$. The overall loss function is then

become[18]:

$$L(\theta, x_n) = -\frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N} W_c \times y_n^c \times \log(h_\theta(x_n, c))$$

where W_c is the weight of class c, y_n^c is the expected output of class c at training example n. Otherwise, h_θ is the model with weight θ.

5.2 Evaluation Metrics

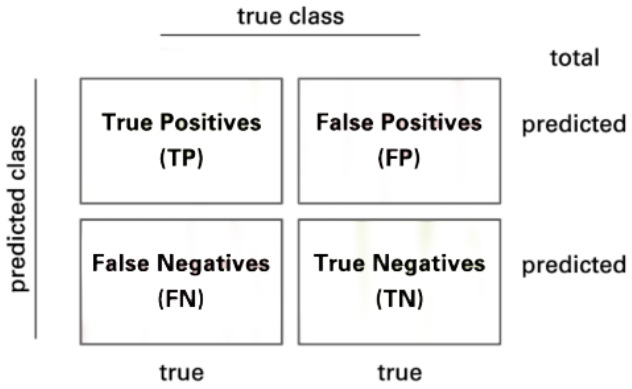

Fig. 6. Confusion Matrix

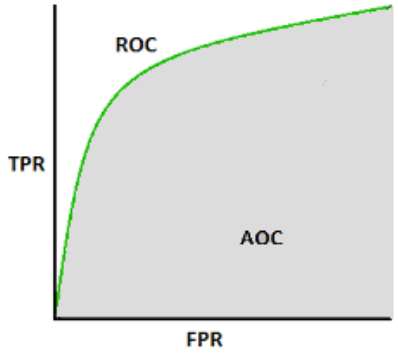

Fig. 7. Area Under the Curve

In this paper, the model is evaluated by using the confusion matrix and related metrics. The figure 4 illustrates the presentation of a 2×2 confusion matrix used for 2 class. Consider a confusion matrix A with C number of class. Let A^i and A^j is the set of A rows and columns respectively.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} \\ a_{21} & a_{22} & \dots & a_{2j} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} \end{bmatrix}$$

The True Positive(TP) of all class in this case is the main diagonal of the matrix A. The following method are used to calculate the False Positives(FP), False Negatives(FN), and True Negatives(TN) of all class:

$$FP = -TP + \sum_{k=1}^{i} A_k^i \qquad FN = -TP + \sum_{k=1}^{j} A_k^j$$

$$TN_c = \sum_{i=1}^{C}\sum_{j=1}^{C} a_{ij} - \left[\sum_{k=1}^{i} A^i_{i=ck} + \sum_{k=1}^{j} A^j_{j=ck}\right] + a_{i=c\,j=c} \implies TN = \begin{bmatrix} TN_1 & TN_2 & \dots & TN_c \end{bmatrix}$$

Then, the model is evaluated by the following metrics:

$$\text{Sensitivity(Sens)} = \frac{TP}{TP+FN} \qquad \text{Specificity(Spec)} = \frac{TN}{TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{F1 Score} = \frac{2\times TP}{2\times TP+FP+FN+TN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \qquad \text{Balanced Accuracy} = \frac{\text{Sens}+\text{Spec}}{2}$$

The last metric is the *AUC* score standing for Area Under the Curve which is the Receiver Operating Curve(ROC) that indicate the probability of TP versus the probability of FP.

6 Experiment

All the model in this paper is trained with Adam Optimizer[19]. The initial learning rate is set to 0.001, an learning rate reduction schedule is setup with the minimum learning rate is 0.0000001 with the factor of 0.2. Otherwise, the epsilon argument of the optimizer is set to 0.1. The performance of all model is presented in the figure below:

7 Conclusion

[20]

References

[1] Soumyya Kanti Datta, Mohammad Abuzar Shaikh, Sargur N. Srihari, and Mingchen Gao. Soft-attention improves skin cancer classification performance. Available at https://arxiv.org/abs/2105.03358, 4 Jun 2021.

[2] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. Artificial intelligence-based image classification for diagnosis of skin cancer: Challenges and opportunities. Available at https://arxiv.org/abs/1911.11872, 20 Jun 2020.

[3] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. Available at https://arxiv.org/abs/1608.06993, 28 Jan 2018.

[4] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. Available at https://arxiv.org/abs/1704.04861, 17 Apr 2017.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. Available at https://arxiv.org/abs/1512.00567, 11 Dec 2015.

[6] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. Available at https://arxiv.org/abs/1801.04381, 13 Jan 2018.

[7] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. Available at https://arxiv.org/abs/1905.02244, 20 Nov 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. Available at https://arxiv.org/abs/1512.03385, 10 Dec 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. Available at https://arxiv.org/abs/1603.05027, 25 Jul 2016.

[10] Jonathon Shlens Quoc V. Le Barret Zoph, Vijay Vasudevan. Learning transferable architectures for scalable image recognition. Available at https://arxiv.org/abs/1707.07012, 21 Jul 2017.

[11] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Available at https://arxiv.org/abs/1803.10417, 25 Nov 2018.

[12] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. Available at https://arxiv.org/abs/1905.11946, 11 Sep 2020.

[13] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. Available at https://arxiv.org/abs/1910.03910, 9 Oct 2019.

[14] Abien Fred Agarap. Deep learning using rectified linear units (relu). Available at https://arxiv.org/abs/1803.08375, 7 Feb 2019.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. Available at https://arxiv.org/abs/1502.03044, 19 Apr 2016.

[16] Mohammad Abuzar Shaikh, Tiehang Duan, Mihir Chauhan, and Sargur N. Srihari. 2020 17th international conference on frontiers in handwriting recognition. Sep 2020.

[17] Naofumi Tomita, Behnaz Abdollahi, Jason Wei, Bing Ren, Arief Suriawinata, and Saeed Hassanpour. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. Available at https://arxiv.org/abs/1811.08513, 3 Dec 2019.

[18] YAOSHIANG HO and SAMUEL WOOKEY. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. Available at https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8943952, January 8 2020.

[19] Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. Available at https://arxiv.org/abs/1412.6980, 30 Jan 2017.

[20] Katherine M. Li and Evelyn C. Li. Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks. Available at https://arxiv.org/abs/1807.08332, 22 June 2018.