COMP9444 Project Summary

Skin Lesion Classification Using Deep Learning

Group Name: Neural Network Out!

Lihao Wang (z5474242), Haosheng Ren (z5484476), Chaorong Lei (z5405778), Xuefei He (z5504912), Zelong Hu (z5507000)

I. Introduction

The project that we choose is Skin Lesion Classification Using Deep Learning. In this project, 4 different feature extractions and classifiers have been tested to find the best performance model to identify the skin lesion based on the data from ISIC2019.

As melanoma is one of the deadliest types of skin cancer, we chose this topic because of the urgent need to enhance the early detection of skin cancer. Defects on or under the skin are collectively known as skin lesions. We can divide them into two broad categories. The first category is benign skin tumors, these are lesions, such as moles or cysts. The second category is malignant tumor, which refers to malignant skin lesions, such as melanoma, basal cell carcinoma. Although skin lesions are common, they are often difficult to identify by their appearance, and it remains difficult to automatically identify malignant tumors from dermoscopic images. Therefore, we want to do a successful deep learning strategy which can identify dermoscopic lesions.

II. Related Work

- Multi-features extraction based on deep learning for skin lesion classification.

In this paper, different feature extraction and classifiers have been discussed to find the best combination. There are a total of 17 pre-trained CNN architectures used to extract features and 24 different classification methods have been used to classify the skin lesions from two data sets. In the end, the DenseNet 201 combined with Cubic SVM has the best result.

- Skin lesion classification from dermoscopic images using deep learning techniques.

This paper focuses on early melanoma detection. VGGNet CNN architecture and transfer learning paradigm are used to improve the classification process which achieves a higher sensitivity value.

III. Methods

Four models have been used to extract features, including VGG16, ResNet50, ResNet101, DenseNet201. And fully connected layer and Cubic SVM have been used for classification methods. Therefor total 8 training models will be discussed below for both origianl data set and processed data set.

VGG-16 is a deep convolutional neural network with 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. It extracts hierarchical features from images and perform classification based on extracted features.

ResNet50 is an advanced version of CNN based on residual blocks developed by He et al.[3] The main part is using residual blocks to mitigate the gradient degradation problem in the deep networks. He et al mentioned for 50-layer ResNet, each 2-layer block in the 34-layer net was replaced with 3-layer bottleneck block, resulting in a 50-layer Resnet. [4] Bottleneck Block is a bottleneck architecture was

used with three layers 1 x 1 convolution layer, 3 x 3 convolution layer and another 1 x 1 convolution layer. This can help to increase training speed and reduce computational cost.

Resnet101 and resnet50 are basically the same, the only difference is that resnet50 has six blocks with three layers here in conv4_x, and resnet101 has 23 blocks with three layers here.

DenseNet201 is a CNN that consists of a total of 201 layers. Not like the traditional CNN architecture, the DenseNet201 connects each layer to all other layers. This unique feature allows for the reuse of substantial features, which improves the learning process and reduces the risk of overfitting. This architecture is very efficient in the use of parameters and minimizes redundancy. Compared to other similar deep CNNs, the DenseNet201 needs fewer parameters to maintain an even better performance.

To compare the perfermance of each combination, the f1, Accuracy, Average AUC and AUC for each class has been calcualted for the test dataset, refer the section 5 results for details.

The F1 score is the harmonic average of accuracy and recall. Because we do a multi-label classification task, we want to make sure that our model not only correctly identifies relevant instances, but also finds most of them. The F1 Score provides a balance between the two, helping to evaluate the model's performance in correctly identifying each label.

Accuracy measures the proportion of correctly predicted labels among all predictions. It can provide a direct measure of overall correctness. For the multi-label problem, it takes into account the exact match rate, that is, the entire set of predicted labels must match the real label, which is more stringent than the traditional precision metric.

AUROC assesses the model's ability to distinguish categories by plotting true and false positive rates at different threshold Settings. A high AUROC value indicates that the model has a good measure of separability, meaning that it can effectively distinguish between positive and negative classes.

IV. Experimental Setup

Our dataset originates from the ISIC (International Skin Imaging Collaboration), a collaborative effort that focuses on skin imaging. Specifically, the dataset was published in 2019 and includes eight distinct categories of skin diseases. It is a substantial collection, featuring over 20,000 images in the widely used JPG format with total 9 categories. In our dataset, there is a column labelled "UNK" which contains no images. Therefore, we have removed this column, leaving us with a total of eight skin disease categories. The data is split into training, validation, and test sets with a 0.6:0.2:0.2 ratio, designated as the training set, validation set, and test set, to ensure a robust framework for model development and evaluation.

After obtaining the quantity of images for each skin disease, we found that our dataset is highly imbalanced. Therefore, we employed data augmentation and under-sampling to balance it.

The original dataset analysis reveals a significant class imbalance, with the 'NV' (Nevi) category having the highest count at 12,875 instances, which is considerably higher than the other skin disease categories. For instance, 'MEL' (Melanoma) has 4,522 instances, 'BCC' (Basal Cell Carcinoma) has 3,323, and 'SCC' (Squamous Cell Carcinoma) has 628. The least represented categories are 'DF' (Dermatofibroma) with only 239 instances and 'VASC' (Vascular) with 253. To address this imbalance, we will employ undersampling for the 'NV' class to reduce its representation in the training set. This technique will help prevent the model from becoming biased towards the majority class. For the other categories with fewer instances, we will

apply data augmentation to artificially increase their numbers and improve the model's ability to generalize.

However, it is important to maintain the original distribution of data in the validation and test datasets. This approach ensures that the model is evaluated on data that reflects the real-world scenario, including the prevalence of each skin disease category. The balancing efforts are strictly for the training dataset to enhance model performance and fairness in learning from all classes equally. Therefore, we adopted the following balancing scheme for the data: for categories with more than 5000 instances, we used undersampling to reduce their numbers to 80% of their original count; for categories with fewer than 2000 instances, we performed data augmentation to increase their numbers to exceed 2000. This approach allowed us to achieve a more balanced distribution while preserving a sufficient dataset size for effective model training and evaluation.

A custom augmentation transform, RandomRotate180, is defined to randomly rotate images, enhancing the diversity of the training data. Additional augmentations include random horizontal and vertical flips, resized crops, and contrast change.

Data resource: https://challenge.isic-archive.com/data/#2019

V. Results

- Original Data VS Processed Data:

For every combination, it has been demonstrated that the processed data performance is better than the original data set, which means the data processing methods did a good job.

- Different feature extraction:

Based on the same data set, ResNet101 has the best result for classification, followed by DenseNet201, and Resnet50. VGG16 is not good for this dataset because of its poor architecture compared with others.

- Classification Methods:

In this project, the fully connected layer performs better than the Cubic SVM.

The details of each combination show below,

Table 1: Show the results of original dataset.

Original Data	VGG16	VGG16 + Cubic SVM	ResNet50	ResNet50 + Cubic SVM	ResNet101	ResNet101 + Cubic SVM	DenseNet 201	DenseNet201 + Cubic SVM
F1	0.6628	0.6270	0.6590	0.69	0.7524	0.6812	0.9154	0.72
Accuracy	65.56%	66.53%	68.27%	71.98%	94.04%	71.03%	73.86%	74.27%
Average AUC	0.8847	0.8780	0.8985	0.9225	0.9350	0.9124	0.9632	0.9457
AUC PER CLASS								
AUC – Class 1	0.8491	0.8426	0.8050	0.8805	0.8861	0.8788	0.9046	0.9107
AUC – Class 2	0.8942	0.8958	0.9057	0.9242	0.9367	0.9210	0.9416	0.9406
AUC – Class 3	0.9242	0.9163	0.9466	0.9433	0.9604	0.9381	0.9589	0.9617
AUC – Class 4	0.8952	0.8990	0.8905	0.9163	0.9280	0.9204	0.9435	0.9394

AUC – Class 5	0.8370	0.8297	0.8353	0.8860	0.9015	0.8833	0.9014	0.9163
AUC – Class 6	0.8856	0.8636	0.9503	0.9225	0.9578	0.8831	0.9302	0.9441
AUC – Class 7	0.9300	0.8796	0.9589	0.9573	0.9781	0.9716	0.9866	0.9890
AUC – Class 8	0.9114	0.8978	0.8960	0.9501	0.9314	0.9459	0.9521	0.9636

Table 2: Show the results of processed dataset.

Proposed Data	VGG16	VGG16 + Cubic SVM	ResNet50	ResNet50 + Cubic SVM	ResNet101	ResNet101 + Cubic SVM	DenseNet 201	DenseNet201 + Cubic SVM
F1	0.6464	0.6409	0.7851	0.71	0.7524	0.6973	0.9123	0.74
Accuracy	64.42%	66.79%	78.86%	72.19%	95.10%	70.42%	69.03%	74.00%
AUC	0.8879	0.8806	0.9565	0.9192	0.9531	0.9053	0.9505	0.9445
AUC PER CLASS								
AUC – Class 1	0.9053	0.8441	0.9685	0.9299	0.9597	0.9204	0.8822	0.9008
AUC – Class 2	0.9063	0.8985	0.9783	0.9403	0.9735	0.9356	0.9307	0.9332
AUC – Class 3	0.8241	0.9062	0.9290	0.8839	0.9315	0.8707	0.9427	0.9517
AUC – Class 4	0.9003	0.8988	0.9744	0.9184	0.9633	0.9187	0.9455	0.9422
AUC – Class 5	0.8126	0.8038	0.9050	0.8679	0.9049	0.8581	0.8923	0.9109
AUC – Class 6	0.8975	0.8893	0.9398	0.9211	0.9494	0.9185	0.9306	0.9650
AUC – Class 7	0.8962	0.8849	0.9666	0.9336	0.9467	0.9218	0.9922	0.9883
AUC – Class 8	0.9611	0.9189	0.9903	0.9585	0.9960	0.9335	0.9562	0.9635

VI. Conclusions

Because of the limitation of data size and computation resources, the results may change with additional datasets or computation resources. However, it clearly shows the relationship between different feature extraction and classification methods, which provides a method for future classification problems in the medical field, especially for skin lesions.

From the results of 8 different models for both original data and processed data, the best performance combination is ResNet101 and Fully Connected Layer.

VII. Reference

[1]: https://challenge.isic-archive.com/data/#2019
[2]: Youngseok Lee, Jongweon Kim, PSI Analysis of Adversarial-Attacked DCNN Models, Applied Sciences 2023,13,9722, Page 4
[3]: He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770-778
[4]: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", CVPR 2016, Page 775
[5]: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", CVPR 2016, Page 773
[6]:https://www.researchgate.net/figure/Detailed-architecture-of-an-Efficient-DenseNet-201_fig3_360070131