National Tsing Hua University
Fall 2023 11210IPT 553000
Deep Learning in Biomedical Optical Imaging
Report

AUTHOR

柯志明

Student ID: 111003804

Cancer Histology Image Classification

1. Task Introduction

Transfer Learning – We have a collection of 3,750 cancer histology images, each measuring 150x150 pixels, designated for building a 6-class classification model. The dataset comprises 2,550 training examples, 600 validation examples, and 600 testing examples, ensuring a balanced distribution across the six categories. Given the relatively small sample size for training a multi-class image classifier, we intend to utilize transfer learning by selecting an appropriate pre-trained model.

Model Selection – We will test different training modes, specifically comparing the fixed feature extractor approach versus fine-tuning, across various models. Their performances will be evaluated on the validation dataset to determine the most effective approach.

Evaluation Metric – To effectively evaluate the performance of the well-trained model, we will use accuracy, confusion matrix, F1 score, and ROC-AUC as our metrics.

This report will outline the progression of the experiment, covering aspects from data preprocessing and model architecture to training algorithms, evaluation, and conclusion, as follows.

2. Data Preprocessing

We have organized the prepared datasets into pairs for training, validation, and testing as $Data_{train} = \{(X_{train}^{(i)}, Y_{train}^{(i)})\}_{i=1}^{2550}$, $Data_{valid} = \{(X_{valid}^{(i)}, Y_{valid}^{(i)})\}_{i=1}^{600}$, $Data_{test} = \{(X_{test}^{(i)}, Y_{test}^{(i)})\}_{i=1}^{600}$. Additionally, $Data_{tv}$ represents the union of training and validation datasets ($Data_{train} \cup Data_{valid}$), which might be used for final training after the hyperparameters are well tuned and no longer rely on the validation set performance.

Resize and Normalization – Given the requirements of the pre-trained model, all input images in $Data_{train}$, $Data_{valid}$, and $Data_{test}$ are resized from 150×150 pixels to 224×224 pixels. Following this, the input data are normalized using the mean and standard deviation specified by the pre-trained model.

Data Augmentation – Considering the relatively small sample size for training a multi-class image classifier, we are employing data augmentation techniques. We will utilize online augmentation, applying transformations dynamically during the training process rather than pre-processing the data. This approach allows each training iteration to expose the model to slightly varied versions of the data, enhancing generalizability and robustness. Given that histology images are not sensitive to direction, zoom, or absolute color variations, our online augmentation strategy includes various transformations such as rotations, flips, zooms, and color adjustments. These transformations are applied in real-time as the data is loaded for each training epoch, ensuring memory efficiency, and preventing overfitting. The specific types of transformations used are detailed in the following table.

Table 1: Data Transformation Functions and Parameters

Step	Transformation	Parameters
1	RandomHorizontalFlip	p=0.2
2	RandomVerticalFlip	p=0.2
3	RandomRotation	degrees=30
4	ColorJitter	brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1
5	RandomResizedCrop	size=224, scale=(0.9, 1.0), ratio=(0.8, 1.25)

3. Model Architecture

Pre-trained Model – The VGG network is one of the most familiar deep CNN models and is frequently adopted for application in medical image classification [1][2]. Some evidence has shown that VGG16 offers the best performance among the top models in the CNN family, such as VGG19, ResNet50, DenseNet121, InceptionV3, etc. [2][3].

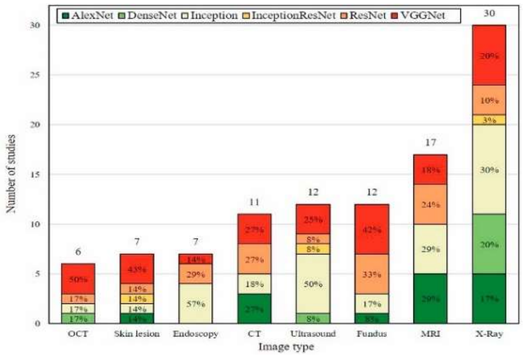

Fig. 1: the frequency of studies using specific types of Transfer Learning CNN models per image type [1]

Thus, we selected VGG16 as the pre-trained model. Considering that histology images generally do not contain large-scale objects specifically, we opted to use only the first three blocks instead of the full five blocks of feature layers in VGG16. The model is adapted by concatenating these first three blocks of VGG16's feature layers with custom fully connected classifier layers, comprising a hidden layer with n_h neurons and an output layer. Here, n_h is set to 128 in Case 1 and 512 in Case 2. The final output layer consists of 6 neurons, corresponding to the number of classes.

4. Training Algorithm and Model Selecting

Fixed Hyperparameters – The common hyperparameters are set as follows: Loss Function – Cross-Entropy, Optimizer – Adam, Learning Rate – 0.01, and Batch Size – 32. The number of epochs is fixed at 100 for the model comparison stage and more for the final training of the leading model.

Compared Cases – We executed the six cases as defined in Table 2. The results of the training and validation performances for these cases are presented in the same table. Case C1.1, distinguished by a training accuracy of 0.9286 and a validation accuracy of 0.8883, emerged as the best performer after 100 epochs. Consequently, Case C1.1 has been selected as the leading model for further training and subsequent evaluation on the testing dataset.

Table 2: Performance Comparison Across Six Cases (100 epochs for each).

Case	n_h	Training Mode	Training Accuracy	Validation Accuracy	Time Consumption
C1.1	**128**	**C1.1 Fixed Features (FF)**	**0.9286**	**0.8883**	1:25:29
C1.2	128	C1.2 Fine Tuning (FT)	0.8012	0.8033	1:30:11
C1.3	128	C1.3 FF (50) + FT (50)	0.7106	0.7683	1:28:53
C2.1	512	C2.1 Fixed Features (FF)	0.9110	0.8583	**1:09:06**
C2.2	512	C2.2 Fine Tuning (FT)	0.7961	0.8033	1:34:20
C2.3	512	C2.3 FF (50) + FT (50)	0.9188	0.8867	1:11:49

5. Performance

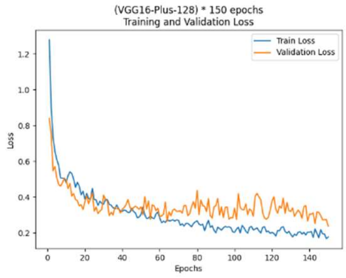

Fig. 2.1: Loss curves for Case C1.1

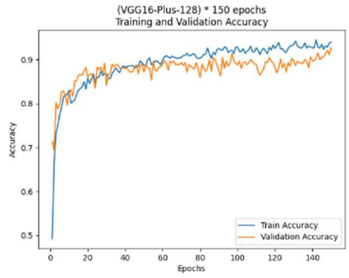

Fig. 2.2: Accuracy curves for Case C1.1

VGG16-Plus-128 – Let’s name the model from Case C1.1 as VGG16-Plus-128. We continued training VGG16-Plus-128 and ultimately stopped at epoch 150, where we achieved the highest validation accuracy across all epochs. The learning curves of VGG16-Plus-128 are shown in Figure 2. We can observe that the curves converge with a small gap between training and validation losses/accuracies, indicating that this model does not suffer from overfitting.

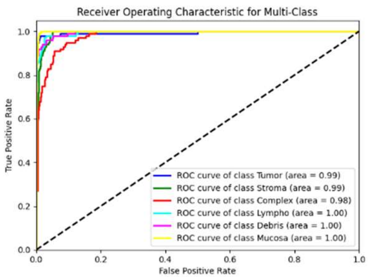

Fig. 3.1: ROC curves

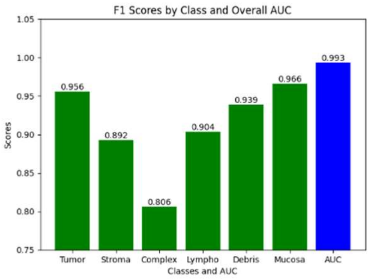

Fig3.2: F1-scores and average AUC

Accuracies – The well-trained model achieved a training accuracy of 93.96%, a validation accuracy of 92.67%, and a testing accuracy of 91.00%.

F1-Score and ROC-AUC – As depicted in Fig. 3, the model achieved a macro-averaged AUC of 0.993 using the One-vs-Rest method. This score indicates an excellent level of separability and a robust capability to distinguish between different classes. Regarding the F1-scores, most are above 0.9, with the notable exceptions being the 'Stroma' class at 0.892 and the 'Complex' class at 0.806.

Confusion Matrix – The multi-class confusion matrix for VGG16-Plus-128 is depicted in Fig. 4. A detailed examination of this matrix reveals that the 'Complex' class demonstrates the lowest distinctiveness. This is highlighted by a True Positive count of 83, alongside 23 False Positives and 17 False Negatives. The 'Complex' class likely represents a variety of cellular features that are challenging to categorize into simpler, more definitive classes. Consequently, a notably higher proportion of examples from the 'Complex' class are misclassified, with frequent confusion occurring particularly with the 'Lympho' and 'Stroma' classes.

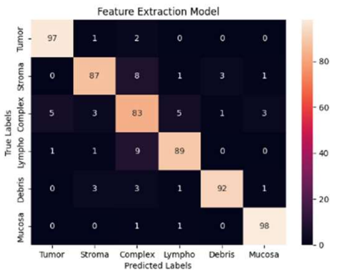

Fig. 4: The confusion matrix for VGG16-Plus-128

6. Conclusion

This task is interesting and challenging. Due to space limitations, many initial attempts were not included in the report. Moreover, the best model presented in the report is only a relatively better choice among a few models and experiments, and not necessarily the best among all possibilities. With more time for experimentation, it would certainly be possible to find a stronger model and conduct more training for improvement.

Furthermore, for certain classes with relatively weaker distinction abilities, it's crucial to go beyond simply increasing sample sizes for model training. Leveraging existing medical knowledge is essential to understand potential challenges in identifying these classes. Take, for instance, the 'Complex' class of tissues, which may be so named due to its composite and multiple tissue characteristics. This complexity makes it especially susceptible to having overlapping features with other categories, leading to confusion. Given this context, the poorer score in terms of distinctiveness for the 'Complex' class is quite understandable.

7. References

[1] M.A. Morid, A. Borjali, G. Del Fiol, “A scoping review of transfer learning research on medical image analysis using ImageNet”, *Computers in Biology and Medicine Volume 128, January 2021, 104115*.

[2] N.D. Kathamuthu, S. Subramaniam, Q.H. Le, S. Muthusamy, H. Panchal, S.C.M. Sundararajan, A.J. Alrubaie, M.M.A. Zahra, “A deep transfer learning-based convolution neural network model for COVID-19 detection using computed tomography scan images for medical applications”, *Advances in Engineering Software Volume 175, January 2023, 103317*.

[3] K. Kamal, H. EZ-ZAHRAOUY, “A comparison between the VGG16, VGG19 and ResNet50 architecture frameworks for classification of normal and CLAHE processed medical images”, *Research Square, 28 Apr 2023*.