Recreation and Analysis of Cervical Fuzzy Distance Ensemble

Shiqian Qiu **Jianxiang Ma** **Jiahong Zhai**

Abstract

Affective image recognition from bioimaging images is a challenging task in computer vision, with potential applications in various fields such as cancer diagnosis, fetal examination, and monitoring mental health disorders. The aim of this study is to reproduce the result of paper "A fuzzy distance-based ensemble of deep models for cervical cancer detection"(Pramanik et al., 2022), conduct a sensitivity analysis on the models' hyperparameters and use vision transformer (ViT) (Dosovitskiy et al., 2020) on the same dataset to compare the performance between the new model and the paper's model. The reproduced result verifies that the model yields the best and most consistent validation accuracy with Learning rate = 10^{-4} and Batch size = 16. We also show that vision transformer yields a similar validation accuracy which is slightly lower than that of the proposed model, meaning that vision transformers can also be applied to classifications. Link to our project: `https://github.com/JackQiu09/Recreation-of-CervicalFuzzyDistanceEnsemble`

1 Introduction

In recent years, computer vision and deep learning techniques have shown great potential in clinical medicine and disease diagnosis, the emerging trend these years especially focused on cancer detection (Esteva et al., 2019). Deep learning algorithms are now utilized in cervical cancer detection, which raised mass interest in the industry and academia. The paper "A fuzzy distance-based ensemble of deep models for cervical cancer detection" gave a vanward approach to such a problem. In the paper the combination of three transfer learning models: Inception V3 (Szegedy et al., 2015), MobileNet V2 (Sandler et al., 2018), and Inception ResNet V2 (Szegedy et al. 2016). In this study we used the same dataset and the same learning method from the original work to reproduce and verify the outcome and conduct sensitivity analysis on two hyperparameters: Learning Rate and Batch Size. Furthermore, we used vision transformer to learn from and train on the same dataset in order to compare and analyse the performance of our model and the combination of Inception V3, MobileNet V2, and Inception ResNet V2.

2 Related Works

Inception V3 (Szegedy et al., 2015) is a deep convolutional neural network architecture for image recognition and object detection. This was introduced by GoogLeNet. It is the third edition of Google's Inception Convolutional Neural Network family. Inception V3 made it possible to have a deeper network and maintain the scale of parameters at the same time. This model is also known for factorized convolutions and label smoothing. Inception V3 has been widely used for transfer learning tasks due to its good performance and relatively low computational requirements.

MobileNet V2 (Sandler et al., 2018) is a lightweight convolutional neural network architecture designed for mobile and resource-constrained devices. It's proposed in 2018. This architecture introduces a new inverted residual structure and linear bottlenecks, which reduce the number of

Preprint. Under review.

parameters and computational complexity while maintaining high accuracy.

Inception ResNet V2 (Szegedy et al. 2016) is a deep convolutional neural network architecture that combines the Inception architecture with residual connections. It was proposed in a 2016 paper, "Inception-v4, Inception ResNet and the Impact of Residual Connections on Learning". This architecture builds upon the success of both Inception V3 and Microsoft's ResNet, improving the model's performance and convergence speed. Inception ResNet V2 is known for its high accuracy and has been used extensively in transfer learning tasks, particularly in situations where higher computational resources are available and improved performance is desired.

Vision Transformer (ViT) (Dosovitskiy et al., 2020) is a deep learning architecture for computer vision mainly used in image classification. It was first introduced in the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale".

3 Methods and Algorithms

In this section, we will introduce the the proposed model and vision transformer (ViT) (Dosovitskiy et al., 2020) for cervical cancer detection using Pap smear images. First, the images are resized. Then the images are augmented using data augmentation tools which includes random zooming, shifting, flipping, and rotation. These images are then fed into three CNN models and a vision transformer in which all the models are pre-trained on ImageNet. Additional layers are added to the three CNN models, which includes convolution, max-pooling, and fully connected layers. The overall workflow of the proposed methods is shown in Fig. 1.

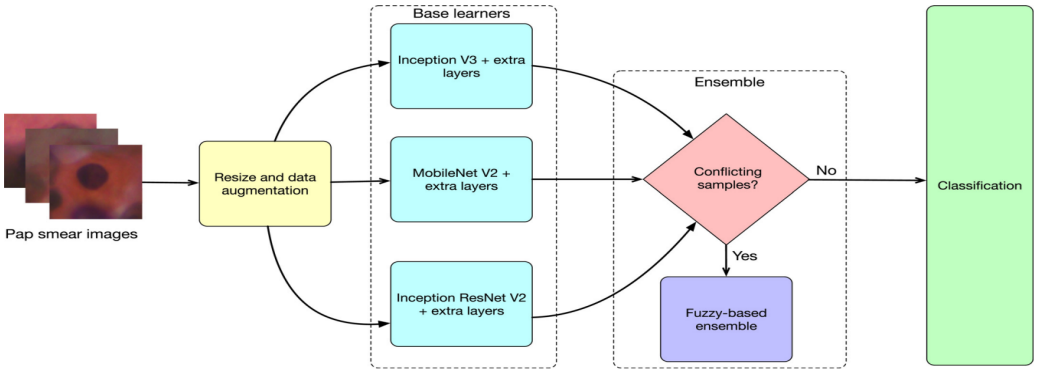

Figure 1: Workflow overview of proposed image classification

3.1 Dataset description

All models are trained and tested on the SIPaKMeD dataset (Plissiti et al., 2018) which is open to public. The dataset includes 4,049 Pap smear images and are classified into five categories: Superficial Intermediate, Parabasal, Koilocytes, Dyskeratotic, and Metaplastic.

3.2 Fuzzy distance-based ensemble

In the original work (Pramanik et al., 2022), the authors proposed a novel approach that try to minimize the difference between the ideal solution and the predicted outcome. We do this by first comparing the most probable class predicted by each base classifier (Inception V3, MobileNet V2 and Inception ResNet V2) to see if they agree. If conflict exits, we calculate the Euclidean, Manhattan, and Cosine distances between the ideal solution (a $\mathbf{1}$ vector) and the confidence score of each classifier for every samples. Then we multiply the result of these three distances and the class with the lowest product is selected.

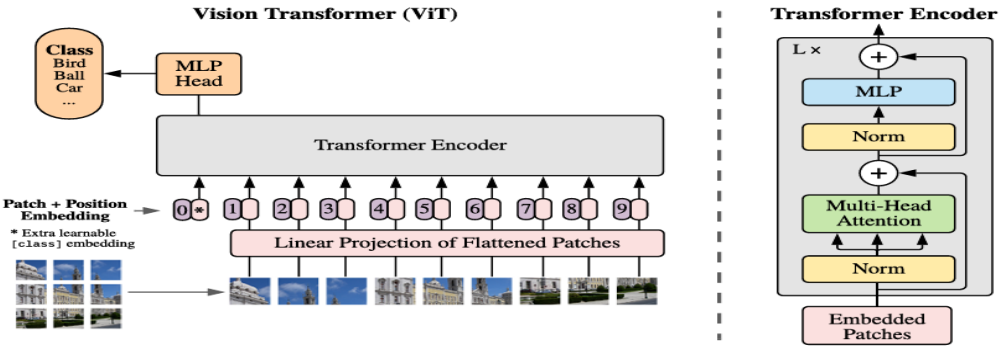

Figure 2: Workflow overview of image classification using Vision Transformer

3.3 Vision Transformer

Vision Transformer was inspired by traditional transformer architecture. Vision Transformer treats images as sequences of visual tokens. Positional embeddings are added to the input tokens to maintain spatial information, and the subsequent layers of the Transformer learn to model the relationships between patches. The output is obtained by applying a classification head to the transformed embeddings of the first token, which is related to a special class token. The overall architecture of the model is shown in Fig. 2.

4 Results

4.1 Reproduction and testing

To conduct sensitivity analysis, we execute the code and train the three models in 70 epoches. The selected Learning Rates are: $10^{-2}, 10^{-4}, 10^{-6}$. The selected Batch Sizes are: 8, 16, 32. Due to resource limitations, the training step will only be one fold instead of 20 like the original work.

For the vision transformer, we used a learning rate of $1e-4$, epoch of 8, and Adam optimizer. The validation accuracy is 95.06% and the validation loss is 0.1569.

Batch size	Inception V3	MobileNet V2	Inception ResNet V2
8	20.49	93.46	94.69
16	19.14	89.26	94.81
32	84.20	31.36	97.78

Table 1: Validation accuracy (in %) with Learning Rate = 10^{-2}

Batch size	Inception V3	MobileNet V2	Inception ResNet V2
8	94.94	93.70	96.42
16	95.43	89.88	95.80
32	94.44	44.44	96.67

Table 2: Validation accuracy (in %) with Learning Rate = 10^{-4}

4.2 Evaluation metrics

We used accuracy, precision, recall, and F-1 score to evaluate the effectiveness of the proposed method. To understand these metrics, we define the following:

- True Positive (TP): an outcome where the model correctly predicts the positive class.

Batch size	Inception V3	MobileNet V2	Inception ResNet V2
8	76.54	92.72	95.80
16	77.65	90.12	95.19
32	84.69	42.72	96.30

Table 3: Validation accuracy (in %) with Learning Rate = 10^{-6}

- True Negative (TN): an outcome where the model correctly predicts the negative class.
- False Positive (FP): an outcome where the model incorrectly predicts the positive class.
- False Negative (FN): an outcome where the model incorrectly predicts the negative class.
- accuracy is calculated as such $\frac{TP+TN}{TP+TN+FP+FN}$
- precision is calculated as such $\frac{TP}{TP+FP}$
- recall is calculated as such $\frac{TP}{TP+FN}$
- F1-score is calculated as such $\frac{2\times Precision\times Recall}{Precision+Recall}$

Metric	Inception V3	MobileNet V2	Inception ResNet V2	Proposed Method
Accuracy	95.43	89.88	95.80	95.92
Precision	95.46	90.12	95.89	95.97
Recall	95.67	90.07	95.98	96.09
F1 score	95.52	90.02	95.92	96.01

Table 4: Result (in %) by performing one run with Learning Rate = 10^{-4} and Batch Size = 16

5 Discussion and conclusion

The recreated data is very similar to the original paper. And with the result of vision transformer as a comparison, we can see that transformers can in fact perform as good as CNNs in image classification tasks.

There are advantages and disadvantages to both the proposed method and vision transformer. The advantages of the proposed method is that it is non-trainable and does not require additional training data to achieve optimal results. It uses diversified distance metrics to create a fuzzy distance based ensemble, which results in a good decision-making process. The ensemble model also includes three diverse deep learning architectures as base learners to improve overall performance. The advantages of using vision transformers for this task include their ability to handle global dependencies in images, their capacity for handling large datasets, and their flexibility in terms of adapting to different image resolutions.

The proposed ensemble method has limitations as it is static and may not be effective if the feature representation is biased towards a specific class. In the case of an equal distribution of classes, the proposed method may be erroneous as cosine distance measures an angle between two vectors, which could result in a null aperture and a cosine distance of zero. Additionally, deep learners must be carefully trained with the appropriate hyperparameters to avoid overfitting, and the proposed ensemble method may not always provide the desired results, like any other ensemble technique.

On the other hand, some disadvantages of using vision transformers include their high computational and memory requirements, which can make training and inference slower and more resource-intensive.

Due to limited time and resources, we were only able to run the training for once instead of running k-fold cross validation like the original paper. In conclusion, our recreation of the paper was successful, and with the additonal vision transformer for comparison, we can see the potential of transformers in the field of image classification.

Author Contributions

Shiqian Qiu:

Executed the original code with learning rate 10^{-4} on all three batch sizes. Created and executed the code for the vision transformer. Partitioned and organized image data. Edited and revised the Method, Results, and Discussion section.

Jianxiang Ma:

Executed the original code with learning rate 10^{-2} and on three batch sizes. Edited and revised the Introduction, Abstract, and Results section.

Jiahong Zhai:

Executed the original code with learning rate 10^{-6} on all three batch sizes to recur the result of the original paper. Wrote the proposal of the project. Helped analyse the result of Vision Transformer. Wrote Introduction, Related Work and Methods and Algorithms of final report.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021*, 2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[2] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 24-29. https://doi.org/10.1038/s41591-018-0300-7

[3] Plissiti, M. E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., & Charchanti, A. (2018). SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images. *25th IEEE International Conference on Image Processing (ICIP)* (pp. 3144-3148). IEEE.

[4] Pramanik, R., Biswas, M., Sen, S., de Souza Júnior, L. A., Papa, J. P., & Sarkar, R. (2022). A fuzzy distance-based ensemble of deep models for cervical cancer detection. *Computer Methods and Programs in Biomedicine*, 219, 106776. https://doi.org/10.1016/j.cmpb.2022.106776

[5] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv preprint arXiv:1801.04381. https://arxiv.org/abs/1801.04381

[6] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 31* (1). https://doi.org/10.1609/aaai.v31i1.11231

[7] Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (2015). Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818-2826.