Project Proposal

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” and conduct a sensitivity analysis on the hyperparameters of the models. Then use vision transformer (Dosovitskiy et al. 2020) on the same dataset to compare the performance between the new model and the paper's model.

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. In this area mass image recognition learning methods have been implied by researchers to learn and recognize medical images like MRI image and X-ray plates in order to help doctors to diagnose whether the patient has cancer, and furthermore to predict the category and stage of such cancer. Cervical cancer is one of the leading causes of women’s death.

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” gives 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. 2017), with additional layers are used for training the model.

In this project we will use the same dataset and the same learning method from the original paper to reproduce and verify the outcome. And conduct a sensitivity analysis on the hyperparameters of the models. Furthermore, we will use our own vision transformer algorithm 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.

Related Works:

Inception V3 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 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 parameters and computational complexity while maintaining high accuracy.

Inception-ResNet V2 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 is a deep learning architecture for computer vision mainly used in image classification. It’s first introduced in the paper “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”.

Method / Algorithm:

Inception V3 is a CNN architecture made for image recognition and classification. After processing images through inception modules with parallel convolutional branches, auxiliary classifiers in intermediate layers provide additional supervision, solving the vanishing gradient problem and improving regularization. Then the global average pooling layer reduces the scale of parameters and the chance of overfitting. Then a fully connected layer and softmax activation function will give final class probabilities. The model is trained using stochastic gradient descent.

MobileNet V2 will first process and normalize input images. Then use depthwise separable convolutions to depthwise convolutions and pointwise convolutions. The core of MobileNet V2 is inverting residual blocks with linear bottlenecks. The output feature from the final convolutional layer will pass through the global average pooling layer, which reduces the spatial dimensions to generate a fixed-length feature vector. The feature vectors are then processed by a fully connected layer and output final class probabilities. MobileNet V2 is also trained using stochastic gradient descent.

Inception-ResNet V2 is the CNN model that combines Inception architecture and residual connections. The model accepts images with 299x299 pixels and 3 color channels as input. Stem model reduces the spatial dimensions and increases the depth of feature maps. Then Inception-ResNet blocks will learn features at different scales. The outputs are concatenated and linked with input by residual connections. Reduction-A and Reduction-B blocks reduce spatial dimensions of feature maps and deepen the depth. Then the global average pooling layer will generate a feature vector. After that dropout layer will prevent overfitting. The feature vectors then go through a fully connected layer followed by the softmax activation which outputs the probabilities for each class.

Vision Transformer is 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 the 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.

Reference:

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, Oct. 2020.

[Online]. Available: https://arxiv.org/abs/2010.11929

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," arXiv preprint arXiv:1512.00567, Dec. 2015.

[Online]. Available: https://arxiv.org/abs/1512.00567

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2018, pp. 4510-4520.

[Online]. Available: https://arxiv.org/abs/1801.04381

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in Proc. Thirty-First AAAI Conference on Artificial Intelligence, Feb. 2017.

[Online]. Available: https://arxiv.org/abs/1602.07261