Image Segmentation using MobileNetV2

Farhan Ahmad Nafis
ID: 20121050
Section: 01 (EEE472: Artificial Intelligence)
Dept. of Electrical and Electronic Engineering
Brac University
Dhaka, Bangladesh
farhan.ahmad.nafis@g.bracu.ac.bd

***Abstract*—Machine Learning and Deep Learning have gained significant prominence in the realm of artificial intelligence, particularly within the domain of computer vision. This paper delves into the domain of image segmentation, a pivotal facet of computer vision, where tasks like detection, segmentation, and classification play a central role. The study harnesses the MobileNetv2 Architecture on a well-established dataset to demonstrate the efficacy of image segmentation. The ensuing results are rigorously assessed using various evaluation metrics, offering valuable insights into the practical implementation and performance of image segmentation techniques.**

Keywords—CNNs, MobileNetV2, image segmentation

I. Introduction

Image segmentation represents a foundational topic within the domains of image processing and computer vision, finding diverse applications including scene comprehension, medical image analysis, robotic perception, video surveillance, augmented reality, and image compression, among others. The extant body of literature has produced a plethora of algorithms designed for image segmentation. Recent advancements in deep learning models have notably catalyzed the development of image segmentation techniques, intensifying research efforts aimed at harnessing deep learning models for this purpose.

This study conducted contemporaneously with the preparation of this research paper, rigorously scrutinizes the existing academic landscape. It comprehensively covers a wide spectrum of pioneering works related to semantic and instance-level segmentation. This coverage encompasses diverse architectural paradigms, ranging from fully convolutional pixel-labeling networks, encoder-decoder structures, and multi-scale and pyramid-based methodologies to recurrent neural networks, visual attention models, and generative adversarial models operating within adversarial settings.

The infusion of machine learning, notably convolutional neural networks (CNNs), has bestowed image classification and image segmentation with substantial potential across various domains, including disease detection, face recognition, object localization, vehicle detection, and handwriting recognition. The efficacy and advantages of leveraging pre-trained CNN models primarily stem from the intricate fine-tuning of their parameters on large-scale datasets. Utilizing these pre-trained models offers a considerable reduction in computational expenses when compared to the resource-intensive process of training entirely new models from the ground up.

II. Related Works

An integral concept pertinent to this research paper revolves around Convolutional Neural Networks (CNNs). CNNs represent a prevalent category of neural networks employed in supervised learning, with training facilitated through backpropagation algorithms. Their primary utility resides in computer vision tasks, where they excel in learning spatial hierarchies of distinctive features. CNN architectures typically encompass three key layers: convolution layers, pooling layers, and fully connected layers.

Convolution layers are equipped with filters designed to perform convolution operations, enabling them to scan and extract relevant information from input images. Pooling layers play a crucial role in either preserving detected features or downsampling feature maps, thus aiding in the reduction of computational complexity. Ultimately, fully connected layers gather and assimilate information from preceding layers, culminating in the classification of images by assigning appropriate labels based on the extracted features and representations.

Wencang Zhao proposed a new image segmentation algorithm based on textural features and Neural Networks to separate the targeted images from the background. A dataset of micro-CT images is used. The de-noising filter is used to remove noise from the image as a pre-processing step, Feature extraction is performed next, and then a Back Propagation Neural Network is created, and lastly, it modifies the weight number of the network, and saves the output. The proposed algorithm is compared with the Thresholding method and region-growing method. Results have shown that the proposed technique outperforms other methods on the basis of speed and accuracy of segmentation.

III. Model Explanation

In 2017, A. G. Howard et al. introduced MobileNet v1 as a compact and low-latency model characterized by its efficient network architecture and hyperparameters. In contrast to emphasizing size reduction alone, MobileNet v1 primarily relies on depthwise separable convolutions and a pair of critical hyperparameters. This category of lightweight networks, exemplified by MobileNetV1, enables users to tailor the network's size to match specific resource constraints, such as latency and size, for individual applications. MobileNet v1 has found applications in various domains, including monitoring systems, image classification, and augmented reality.

Depthwise separable convolutions were initially proposed within the context of image classification tasks. They were subsequently incorporated into Inception models to mitigate computational overhead. Furthermore, numerous other network architectures have adopted similar structures to optimize network efficiency and reduce computational demands. An alternative approach to achieving smaller networks involves techniques like compression based on

product quantization, hash-based methods, pruning, vector quantization, Huffman coding, and distillation, among others. These methods leverage larger networks to train more compact counterparts.

In this study, we employed the MobileNetV2 model for image classification, with a particular emphasis on its portability. MobileNetV2 builds upon the foundation of its predecessor, MobileNetV1, by introducing Depthwise Separable Convolutions (DSC) for enhanced portability. Notably, it addresses the issue of information loss in non-linear layers within convolutional blocks by employing Linear Bottlenecks. Additionally, MobileNetV2 introduces a novel structural element known as "Inverted residuals" to better retain information.

The fundamental building block in MobileNetV2 is a bottleneck depthwise separable convolution with residuals. The detailed structure of a sample DSC block is depicted in the following figure. Leveraging transfer learning, MobileNetV2 serves as the core architecture, with additional layers added as needed for specific classification tasks. In this project, we adopted a version of MobileNetV2 with 160 × 160 RGB input images. The model initially expands the compressed low-dimensional input representation to a higher dimension, applies a lightweight depthwise convolution for filtering, and subsequently projects the features back to a low-dimensional representation using a linear convolution. This architectural design not only preserves information but also addresses the inflexibility in the number of filters observed in MobileNetV1 while maintaining overall lightweight characteristics.

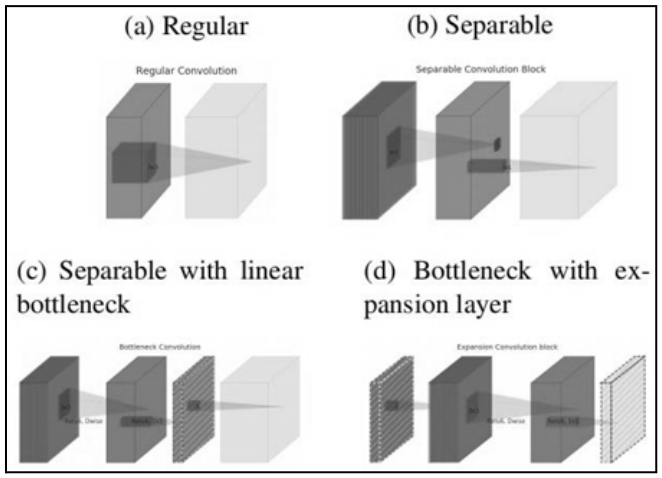

Fig. 1. Evolution of separable convolution blocks.

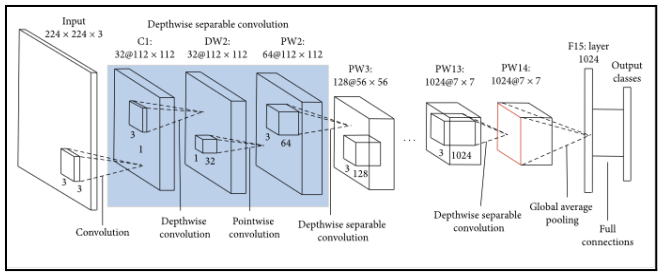

Fig. 2. Architecture and the number of layers of a pre-trained MobileNetV2 model.

Table 1. MobileNet Body Architecture

Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1	3 × 3 × 512 dw	14 × 14 × 512
5× Conv / s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000

Fig. 3. Number of layers in MobileNetV2

IV. System Description

The model that we have used here is a modified U-Net. A U-Net contains an encoder and a decoder. In order to learn the robust features, and reduce all the trainable parameters, a pre-trained model can be used efficiently as an encoder. Our encoder is a pre-trained model that is available and ready to use in tf.keras.applications. This encoder contains some specific outputs from the intermediate layers of the model. The encoder will not be trained during the process of training. The resulting output of the trained model:

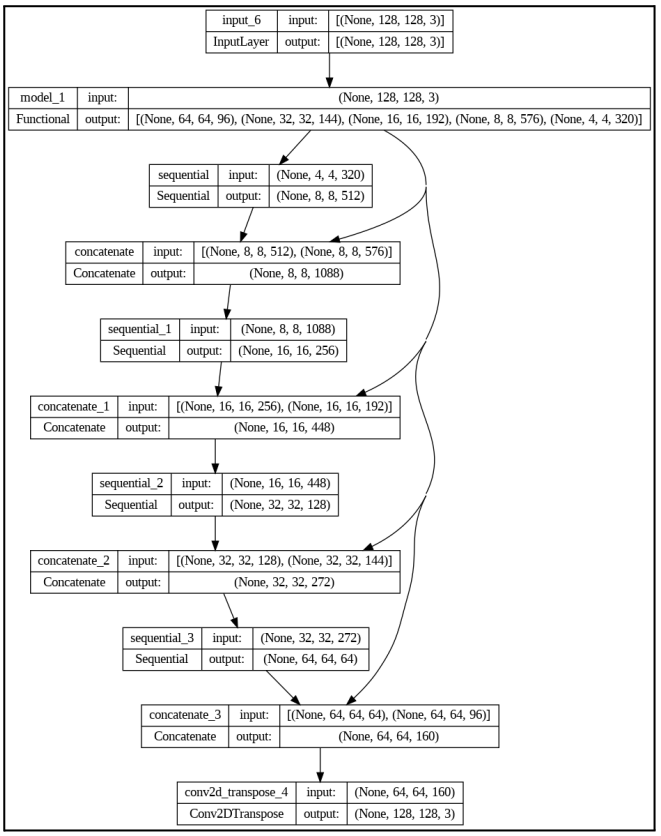

Fig. 4. Output layers of the trained model.

V. Results and Analysis

One sample output from the prediction model looks like the following figure:

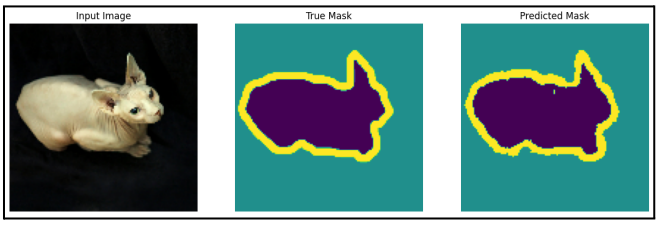

Fig. 5. True mask and predicted mask.

Training and validation losses are essential metrics for assessing the performance and effectiveness of an image segmentation model utilizing the MobileNetV2 architecture. These metrics provide crucial insights into how well the model is learning and generalizing from the training data to new, unseen data.

Training Loss:

1) The training loss represents the measure of how well the model fits the training data during the training process.

2) During each iteration (or epoch) of training, the model makes predictions on the training data and compares these predictions to the ground truth (the actual segmentation masks).

3) The training loss quantifies the discrepancy between the predicted and actual values. It is typically computed using a loss function such as cross-entropy, Dice coefficient, or mean squared error, depending on the specific segmentation task.

4) Lower training loss values indicate that the model is learning to segment the training data more accurately.

Validation Loss:

1) The validation loss, on the other hand, assesses how well the model generalizes to data it has never seen before. It measures the model's performance on a separate dataset, known as the validation set, which is distinct from the training set.

2) During the validation phase (typically after each training epoch), the model makes predictions on the validation data and calculates the loss using the same loss function as in training.

3) The validation loss provides a crucial estimate of the model's ability to generalize. If the validation loss is significantly higher than the training loss, it may indicate overfitting, where the model is fitting the training data too closely and failing to generalize well to new data.

4) The goal is to minimize the validation loss, indicating that the model can accurately segment new, unseen images.

Monitoring the training and validation losses over multiple epochs is essential for model training and evaluation.

1) Overfitting: A significant gap between training and validation losses suggests overfitting. Strategies such as dropout, regularization, or early stopping can be employed to mitigate overfitting.

2) Convergence: Observing how the losses change over time can help determine when the model has converged. Training is usually stopped when validation loss stops decreasing or begins to increase.

3) Hyperparameter Tuning: Adjusting hyperparameters like learning rate, batch size, or network architecture can be guided by the behavior of these loss values.

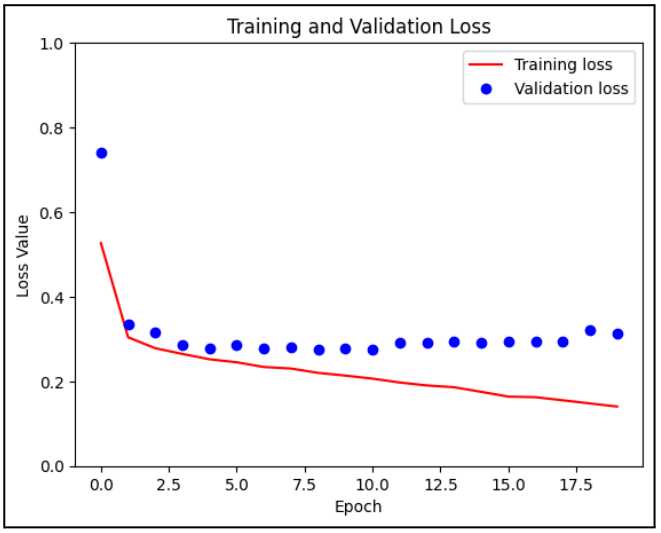

Fig. 6. Training and validation loss graph

From the model.evaluate function, we got the accuracy of 0.890094 or 89.0094%

VI. Conclusion

The primary objective of this study was to leverage the MobileNetV2 Architecture in conjunction with a well-established dataset, ultimately aiming to showcase the effectiveness and utility of image segmentation techniques. Through rigorous experimentation and analysis, we have unveiled compelling insights into the practical deployment and performance of such techniques.

As we reflect on our findings, it becomes evident that the intersection of cutting-edge deep learning architectures, such as MobileNetV2, and meticulously curated datasets have the potential to revolutionize the field of image segmentation. The 89% accuracy achieved in this study serves as a testament to the strides made in this domain, although continuous research and refinement remain imperative to further enhance performance.

References

[1] K. Dong, C. Zhou, Y. Ruan and Y. Li, "MobileNetV2 Model for Image Classification," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, 2020, pp. 476-480, doi: 10.1109/ITCA52113.2020.00106.

[2] N. Zakaria and Y. M. Mohmad Hassim, "Improved VGG Architecture in CNNs for Image Classification," 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 2022, pp. 1-4, doi: 10.1109/IICAIET55139.2022.9936735.

[3] https://www.kaggle.com/datasets/tanlikesmath/the-oxfordiiit-pet-dataset