SEMANTIC FLOWER SEGMENTATION: PRETRAINED AND CUSTOMISED

P___n4, 20_____5

University of Nottingham

ABSTRACT

Semantic segmentation research plays an important role in image analysis, notably contributing to advancements in medical, botanical fields and agricultural automation. This study evaluates the performance of two deep learning models on a given Oxford flower dataset. The widely adopted DeepLabV3+ model using ResNet50 backbone and an individually developed encoder-decoder architecture inspired by the SegNet architecture. An ablation study performed during the custom model's development demonstrates significant and relevant architectural features for flower semantic segmentation. The pretrained DeepLabV3+ model, fine-tuned with the Adam optimizer achieved a mean accuracy of 99.20% and a mean intersection over union (IoU) of 98.63%, achieving superior performance. In comparison, the custom model reached a mean accuracy of 97.37% and a mean IoU of 94.02%. These results highlight the effectiveness of tailored architectures, approaching the performance of superior and more complex state-of-the-art models for this dataset.

Index Terms— Semantic Segmentation, Oxford Flower Dataset, DeepLabV3+, Resnet50

1. INTRODUCTION

Semantic segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. Advances in deep learning have improved the ability to accurately segment complex images in computer vision. It is important across different fields such as medical imaging, botany, agriculture or autonomous driving. For example, tasks such as flower segmentation can be modified and applied for use in plant health monitoring or detection of invasive weeds.

This study compares and demonstrates superior performance using the existing pretrained DeepLabV3+ model with a ResNet50 backbone. The model utilizes atrous convolutions to capture multi-scale information effectively [1], making it well suited for precise applications like flower segmentation. Alternatives such as ResNet-18 and Inception-ResNet-v2 were considered, however were disregarded due to significantly lower accuracy or larger network/ prediction time [2].

The custom network architecture is shown in Figure 1. An alternative and comparison to existing pretrained nets.

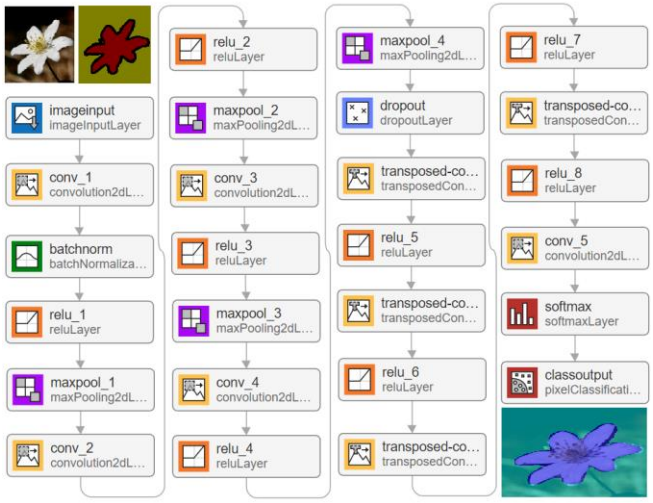

Figure 1. **Proposed custom network architecture.** Images and labels in the top left get taken as input to output at the bottom right.

Designed to address the unique challenges of the Oxford dataset. With the encoder-decoder network resembling SegNet [6], a simpler approach was taken, only containing a single batch normalization layer. Allowing for a well-structured ablation study, demonstrating clear performance changes from model architecture alterations whilst maintaining near superior performance results.

1.1. Dataset

The Oxford Flower dataset is used with the proposed models. Specifically designed for image understanding and computer vision research, it provides a great benchmark for semantic segmentation. This dataset includes several different flower categories commonly found in the United Kingdom providing a diverse range of images to learn from [3, 4]. The images also vary in scale pose and lighting conditions allowing for robust training. Labels are provided for 846/1360 images consisting of 0: null/boundaries, 1: flower, 2: leaves, 3: background and 4: sky. As the models focus on flower segmentation, class 0 was excluded from labelling, 1 was labelled as flower and the remaining labels were set to background. Due to the noise in labels remaining even after resizing images, a median symmetric 5x5 filter was used to reduce label noise on training images as shown in Figure 2. The data is split into 60% training, 20% validation and 20% testing for all models.

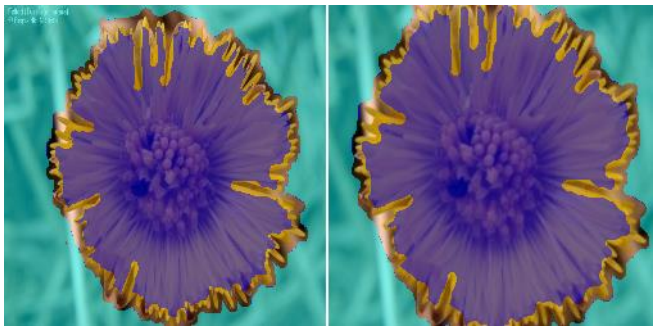

Figure 2. **Images of training labels before and after augmentation.** The left image shows the image label pair as given. The right image shows a random scaling with a factor 1 to 1.2 and a symmetric, label only, median filter of 5x5.

1.2. Evaluation metrics

Global Accuracy: Quantifies the overall population of pixels correctly classified across the entire dataset. It is defined as:

$$Global\ Accuracy = \frac{Total\ Correct\ Predictions}{Total\ Number\ of\ Pixels}$$

Mean Accuracy: Evaluates the average accuracy per class, offering a more balanced view across potentially imbalanced classes. It is defined as:

$$Mean\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{Correct\ Predictions\ for\ Class\ i}{Total\ Predictions\ for\ Class\ i}$$

Mean Intersection over Union (Mean IoU): Measures the average overlap between the predicted and ground truth masks for each class, an essential metric for segmentation. It is defined as:

$$Mean\ IoU = \frac{1}{N}\sum_{i=1}^{N}\frac{Intersection\ of\ Predicted\ and\ True\ Masks\ for\ Class\ i}{Union\ of\ Predicted\ and\ True\ Masks\ for\ Class\ i}$$

Mean IoU in Image Histogram Form: A histogram of IoU scores for each image to visualize the distribution of IoU values across the dataset. Useful for identifying consistency and detecting outliers. It is defined as:

$$Histogram\ of\ IoU\ Scores = Frequency\ of\ each\ IoU\ score\ interval\ across\ all\ images$$

Confusion Matrix: A matrix which displays the prediction results for each class against every other class. It is useful in identifying miscalculations where rows represent true labels and columns represent the predicted labels.

2. RELATED WORK

Semantic segmentation has achieved substantial progress over the last decade, spurred by advances in deep learning and the availability of large and annotated datasets [1, 3, 4, 5, 6, 7]. This evolution of segmentation models from traditional image processing techniques to deep learning networks has significantly improved the accuracy of said models. This section reviews existing research on segmentation and the Oxford Flower Dataset [3, 4].

The concept of fully convolutional networks (FCNs) was first introduced by Long et al. [5], who adapted classification networks into segmentation networks by replacing fully connected layers with convolutional layers. This allowed the network to output spatial maps instead of classification scores making it possible to train a network for pixel-wise prediction. FCNs use a skip architecture which has been influential in the design of many subsequent segmentation models.

Building on the FCN framework, Badrinarayanan et al. developed SegNet [6], a deep convolutional encoder-decoder architecture that uses pooling indices to perform non-linear upsampling in the decoder. This approach reduces the number of parameters increasing accuracy without sacrificing performance.

The U-Net architecture, proposed by Ronneberger et al. [7], further refines the encoder-decoder by introducing skip connections providing a connection between gradients and allowing localization of context information. Originally designed for biomedical image segmentation, U-Net has been widely used and is one of the general purpose pretrained semantic segmentation models in MatLab.

DeepLabV3+ introduced by Chen et al. [1], utilizes atrous convolutions to capture multi-scale information without losing resolution. Their work uses depthwise separable convolutions into the atrous spatial pyramid pooling module, enabling the model to effectively handle segmentation at multiple scales effectively, demonstrating superior performance over alternative models.

In the specific context of flower segmentation, Nilsback and Zisserman have contributed significantly with their works creating a visual vocabulary of flower classification [3] and delving into flower segmentation [4]. Their research has focused on developing robust feature extraction and segmentation techniques focusing on high variability in flower appearance. This is important in advancing the accuracy of classification and segmentation models on the Oxford flower dataset and alternative botanical datasets.

3. EXISTING APPROACH

The existing model of the research uses the DeepLabV3+ model with ResNet50 backbone. Alternative models such as U-Net and SegNet (VGG16 backbone) were also tested, however showed inferior performance. Training data is augmented with a random scaling factor of 1 to 1.2

and a symmetric label only median filter of 5x5. Using the Oxford flower dataset the model is configured to output predictions for two classes: “flower” and “background” using the label separations described previously.

Training the model involves using the Adam optimizer with a learning rate of 1e-4 and batch size of 6 for 40 epochs. The learning rate drop period is enabled and left at the default 10 epochs capturing a wide range of optimization with a default scale factor of 0.1. This allows for quick initial optimization with further precise finetuning without signs of overfitting. With a validation frequency of 50 iterations a frequent best validation loss can be saved, taking 26 minutes to train on a Nvidia 3060TI GPU.

4. CUSTOM APPROACH

The network takes in RGB images of size 256x256 pixels. Both existing and custom models have the same training data augmentation and label classification.

Downsampling consists of several convolutional and max pooling layers with varying filter sizes and depth, the first convolutional layer is followed by batch normalization. All layers use ReLU activations to ensure non-linearity during training. Dropout rate of 40% is introduced at the end to mitigate overfitting and increase performance.

Upsampling layers are constructed using transposed convolutional layers that incrementally increase in spatial dimensions of feature maps.

The network concludes with a 1x1 convolutional layer that maps the deep feature representations to two target classes “flower” and “background”, followed by a softmax layer that calculates the probability distribution of these classes for each pixel. The pixel classification layer uses class weights from the frequency of each class in the entire dataset to address class imbalance.

The model is trained using the Adam optimizer for 60 epochs, with a batch size of 8 and an initial learning rate of 1e-4, adjusted dynamically by a factor of 0.1 every 25 epochs. The model takes 33 minutes to train on a Nvidia 3060TI GPU with the best validation loss model saved at every 25 iterations.

5. RESULTS

5.1. Existing models

Table 1 presents the performance of the pretrained models tested. DeepLabV3+ with a ResNet50 backbone demonstrates best performance achieving a Mean Accuracy of 97.5% and a Mean IoU of 95.3%.

The significant superiority of DeepLabV3+, in equal training conditions, justified focused optimization of the model, discarding alternative existing options of U-Net and SegNet. With further finetuning, notably switching from the SGDM optimizer to Adam as well as adding a learning rate schedule and increasing the number of epochs from 10 to 40, refined the results.

Model	Mean Accuracy	Mean IoU
DeepLabV3+ ResNet50	0.97537	0.95287
U-Net	0.89504	0.78434
SegNet VGG16	0.88014	0.78292

Table 1. Performance results of different existing models with equivalent settings, 10 epochs, all reached plateau RNG controlled.

The final pretrained existing finetuned model metrics are as follows: global accuracy of **99.419%**; mean accuracy of **99.202%**; mean IoU of **98.629%**; weighted IoU of **98.844%**; and a mean BF score of **93.86%**.

Given the provided labels frequently have noise, fail to classify border/null values correctly as well as any flowers which might be in the background, it is proposed that the model achieves superior performance for the dataset.

5.2. Custom model

An ablation study was carried out to understand the impact of different architectural changes and training strategies on a SegNet like architectures performance [6]. Table 2 showcases these variations, notably highlighting the importance of dropout, encoder-decoder layer count and training data. Please note that the baseline model was modified with listed methods to achieve varying results. Notably the baseline model used the entire image dataset provided, substituting the missing labels with previously described existing model generated ones. However, due to lower performance the proposed new dataset was abandoned for the original flower dataset provided.

Method	Mean Accuracy	Mean IoU
Without pooling 1st layer	0.9316	0.83593
With 3 poolings (baseline)	0.93388	0.84144
400 feature layers	0.9081	0.77804
40% dropout	0.93654	0.85315
60% dropout	0.93114	0.85705
Original label ds	0.95498	0.89314
Filtered training images	0.89762	0.79658
2 convolutional layers	0.92852	0.82416
4 convolutional layers	0.95307	0.88478
6 convolutional layers	0.93251	0.85378
Original label ds + 4 layers	0.97099	0.93146
Original label ds + 4 layers + 40% dropout	0.96934	0.93135
Original label ds + 4 layers + 40% dropout, adjusted lr	**0.97386**	**0.94017**
Original labels + 40% dropout	0.95718	0.9001

Table 2. Effects on performance results given architectural changes to a baseline model, RNG set for a controlled dataset.

Importantly the incorporation of 40% dropout improved the Mean IoU to 85.3%, whilst lower than 60% dropout it was chosen for mean accuracy. The optimal configuration involved using the original label dataset with four convolutional layers, 40% dropout and adjusted learning rate. This model achieved the following metrics: global accuracy of **97.362%**; mean accuracy of **97.386%**; mean IoU of **94.017%**; weighted IoU of **94.906%**; and a mean BF score of **87.45%**. A small even split of misclassifications as can be seen in the confusion matrix Table 3.

True Class	**Normalized Confusion Matrix (%)**	
Background	97.3%	2.7%
Flower	2.6%	97.4%
Predicted Class	Background	Flower

Table 3. Confusion matrix for the custom model architecture.

Similarly to the existing network, given flawed provided labels, the performance of a model that lacks features such as U-Net skip connections [7] or pretraining, is very good.

5.3. Visualization

The performance of the custom model Mean IoU in Figure 3 demonstrates most of the image segmentations clustering around the 0.9 to 1.0 mark. While a few outliers could indicate a challenging flower class.

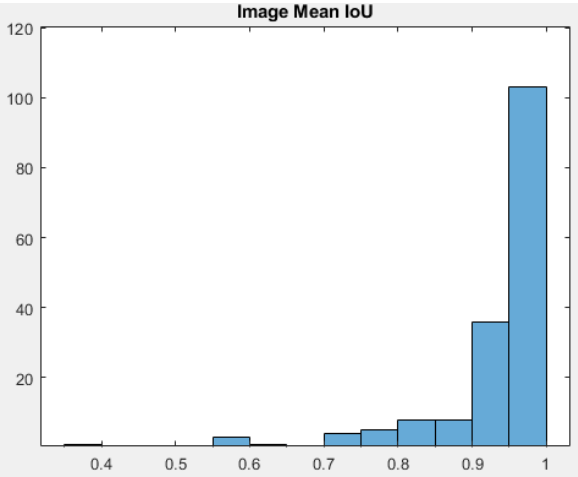

Figure 3. **Mean IoU histogram of the custom model.**

The practical functionality for existing and custom models is illustrated through the output segmentations in Figures 4 and 5 respectively. Both figures reveal a high level of accuracy in segmenting a wide range of flowers from the background. The pretrained model exhibits nearly flawless segmentation with sharp clean boundaries. In contrast while the custom model achieves similar high-quality segmentation, there are minor instances of label spillage, where the segmentation slightly exceeds the actual boundaries of the flowers, indicating areas for refinement.

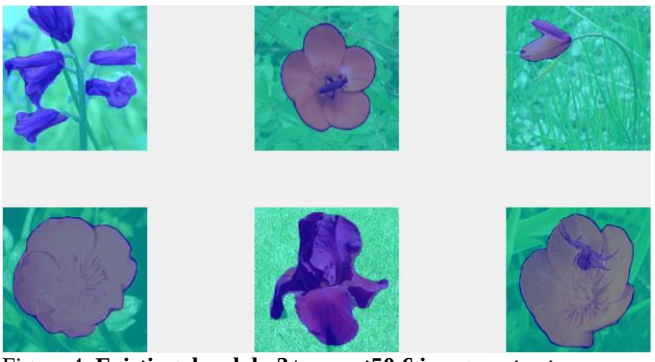

Figure 4. **Existing deeplabv3+ resnet50 6 image output.**

Figure 5. **Proposed custom network 6 image output.**

6. ETHICS

Whilst this research uses publicly available data, ethics must be considered in the use of developed models. Care must be taken to avoid misuse of the technology in applications where automated segmentation could lead to ecological disturbances, such as falsely labeled agriculture targeting specific plant species.

7. CONCLUSION

This study demonstrated and analyzed the effectiveness of a pretrained DeepLabV3+ model and a custom build encoder-decoder model on the Oxford flower dataset. The DeepLabV3+ model achieved near perfect segmentation metrics on a noisy and poorly labeled dataset, showcasing the benefits of using pretrained architectures for complex semantic segmentation tasks. Likewise, a custom model, tailored specifically for the dataset also showed promising results, highlighting the potential advantages of simple yet effective specialized architectures. These findings show the significance of choosing appropriate model architectures based on the dataset and task requirements. Future work will focus on further optimizing existing pretrained models and designing custom models with latest architectural features, to improve performance and integrate into the real world.

8. REFERENCES

[1] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (n.d.). *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*. [online] Available at: https://arxiv.org/pdf/1802.02611.

[2] uk.mathworks.com. (n.d.). *Pretrained Deep Neural Networks - MATLAB & Simulink - MathWorks United Kingdom*. [online] Available at: https://uk.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html. [Accessed 11 May. 2024].

[3] Nilsback, M.-E. . and Zisserman, A. (2006). *A Visual Vocabulary for Flower Classification*. [online] IEEE Xplore. doi:https://doi.org/10.1109/CVPR.2006.42.

[4] M.-E. Nilsback and A. Zisserman (2007). Delving into the whorl of flower segmentation. *CiteSeer X (The Pennsylvania State University)*. doi:https://doi.org/10.5244/c.21.54.

[5] Long, J., Shelhamer, E. and Darrell, T. (n.d.). *Fully Convolutional Networks for Semantic Segmentation*. [online] Available at: https://arxiv.org/pdf/1411.4038.

[6] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), pp.2481–2495.

[7] Ronneberger, O., Fischer, P. and Brox, T. (n.d.). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. [online] Available at: https://arxiv.org/pdf/1505.04597.