Semantic Segmentation - Assignment 3

Syed Saad Ullah Shah - 400202

MSDS SEECS, NUST, H-12, Islamabad, Pakistan.

Corresponding author(s). E-mail(s):
sshah.msds22seecs@seecs.edu.pk;

Abstract

This paper investigates the effectiveness of using different backbone models, including VGG11, VGG13, VGG19, and MobileNetV2, in combination with the U-Net architecture for semantic segmentation on a city dataset. The study compares the performance of these models and evaluates their suitability for urban scene understanding. Experimental results reveal the impact of each backbone model on segmentation accuracy and computational efficiency, providing valuable insights for selecting the most appropriate model in urban computer vision applications. Github code: https://github.com/SaadShah11/CV_MSDS_Assignment3

Keywords: MobileNet V2, UNet, VGG

1 Introduction

Semantic segmentation, a fundamental task in computer vision, involves pixel-level classification and provides detailed understanding of image content. It plays a crucial role in various applications such as autonomous driving, urban planning, and environmental monitoring. One popular architecture for semantic segmentation is U-Net, which exhibits exceptional capabilities in capturing intricate details and localizing objects effectively. However, the choice of backbone model in the U-Net architecture greatly influences its performance.

In this paper, i focus on evaluating the impact of different backbone models, specifically VGG11, VGG13, VGG19, and MobileNetV2, when integrated with the U-Net architecture for semantic segmentation on a city dataset. Each of these backbone models possesses distinct characteristics that contribute to their suitability for urban scene understanding.

The city dataset used in this study comprises diverse urban scenes, encompassing various objects and environmental elements commonly encountered in urban environments, such as roads, buildings, pedestrians, and vegetation. The dataset's diversity enables a comprehensive evaluation of the performance of different backbone models in segmenting urban scenes.

By conducting extensive experiments and evaluations, i aim to provide insights into the strengths and weaknesses of each backbone model, taking into account factors such as segmentation accuracy, computational efficiency, and memory requirements. The performance assessment is carried out using established metrics in the field, including pixel accuracy, mean intersection over union (IoU), and F1 score, which allow for objective and quantitative comparisons.

Through the comparison of VGG11, VGG13, VGG19, and MobileNetV2 as backbone models for the U-Net architecture, i seek to shed light on the trade-offs between segmentation accuracy and computational efficiency. This information is valuable for researchers and practitioners in selecting the most suitable backbone model based on their specific application requirements and constraints.

The findings of this study contribute to the advancement of semantic segmentation in urban environments by providing guidance on selecting the optimal backbone model for the U-Net architecture. By improving the accuracy and efficiency of segmentation algorithms, i can enhance urban scene understanding and facilitate the development of robust computer vision systems for urban applications, ultimately leading to safer and more efficient urban environments.

2 Architectures

UNet:

The U-Net architecture, proposed by Ronneberger et al. in 2015, has become a popular choice for semantic segmentation tasks due to its ability to effectively capture fine-grained details and localize objects. The network architecture is designed to address the challenge of segmenting objects from images while preserving spatial information.

The U-Net architecture follows an encoder-decoder framework, with skip connections that enable the fusion of feature maps from different scales. This helps in capturing both global context and local details, leading to accurate segmentation results. The architecture is named U-Net due to its U-shaped topology, where the encoder and decoder paths are connected by skip connections.

The encoder part of the U-Net consists of a series of convolutional and pooling layers that progressively reduce the spatial dimensions while increasing the number of feature channels. This encoding process helps in capturing abstract and high-level features. The encoder follows a contracting path, reducing the spatial resolution but increasing the receptive field.

The decoder part of the U-Net is responsible for upsampling the feature maps and recovering the spatial information. It consists of a series of upsampling layers, which gradually increase the spatial dimensions, and convolutional layers that decrease the number of feature channels. The decoder follows an expanding path, allowing for precise localization of objects.

The skip connections play a crucial role in the U-Net architecture. They connect corresponding encoder and decoder layers to preserve fine-grained details and provide spatial information. These connections help in combining both low-level and high-level features, allowing the network to leverage both local and global information during segmentation. The skip connections effectively bridge the gap between the encoder and decoder paths, enabling accurate localization.

During the training phase, the U-Net architecture is trained using a pixel-wise cross-entropy loss function. The loss function compares the predicted segmentation map with the ground truth map and penalizes the differences between them. The network parameters are optimized using gradient-based optimization methods, such as stochastic gradient descent (SGD) or Adam, to minimize the loss function.

In the inference phase, given an input image, the U-Net generates a segmentation map by propagating the image forward through the network. The final output is a pixel-level classification map, where each pixel is assigned a label corresponding to the object or background class.

The U-Net architecture has been widely adopted and extended for various segmentation tasks, including biomedical image segmentation, autonomous driving, and scene understanding. Its modular design and skip connections allow for flexibility and customization, enabling researchers to incorporate different backbone models and adapt the network architecture to specific requirements.

In summary, the U-Net architecture is a powerful tool for semantic segmentation, providing a balance between capturing global context and preserving spatial details. Its encoder-decoder framework with skip connections enables accurate object localization and semantic segmentation. By leveraging the strengths of U-Net and incorporating suitable backbone models, researchers can further enhance the performance of semantic segmentation systems.

3 Results

Table 1 Semantic Setgmentation Using UNet with VGG11 Backbone

Architecture	Accuracy	F1-Score	mIoU	Sensitivity	Specificity
UNet - VGG-11	0.90	0.68	0.73	0.68	0.69

Table 2 Semantic Setgmentation Using UNet with VGG13 Backbone

Architecture	Accuracy	F1-Score	mIoU	Sensitivity	Specificity
UNet - VGG-13	0.90	0.68	0.71	0.69	0.68

Table 3 Semantic Setgmentation Using UNet with VGG19 Backbone

Architecture	Accuracy	F1-Score	mIoU	Sensitivity	Specificity
UNet - VGG-19	0.91	0.69	0.71	0.69	0.68

Table 4 Semantic Setgmentation Using UNet with MobileNet-V2 Backbone

Architecture	Accuracy	F1-Score	mIoU	Sensitivity	Specificity
UNet - MobiuleNet-V2	0.67	0.91	0.67	0.73	0.69

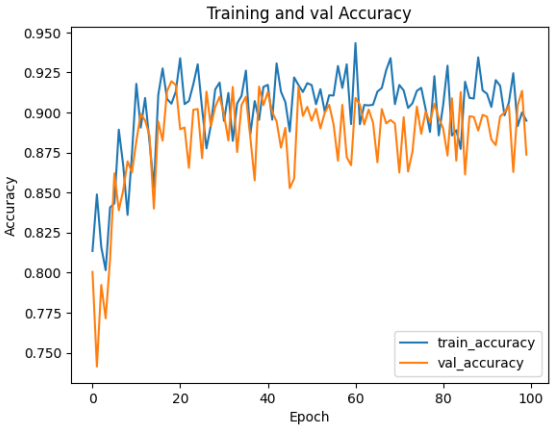
4 Methods

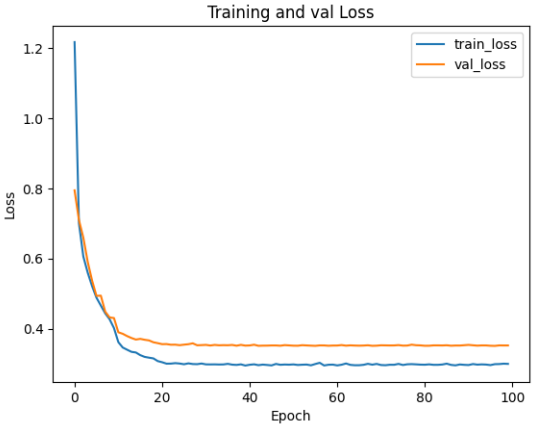
The proposed method involves the following steps:

1. **Dataset:** I downloaded and read the city dataset, then i converted it into dataloader for use by pytorch. Batch size was kept at 8.
2. **Preprocessing:** Before training the models, we performed data preprocessing, including cropping the faces from the images, resizing them to 256x256 pixels, and normalizing the pixel values to have zero mean and unit variance. We also performed data augmentation by randomly flipping the images horizontally and applying random rotations, translations, and zooms.
3. **Model Architecture:** For semantic segmentation, i used 4 CNN architectures backbones with UNet. I used the pre-trained models as feature extractors and fine-tuned them on the City dataset by adding three fully connected layers with 512, 256, and 128 neurons, respectively. The output of the last layer was a softmax function for semantic segmentation into 12 classes.

4. **Training:** I trained the models using the Adam optimizer with a learning rate of 0.001 for 100 epochs. I used the miou loss function for semantic segmentation.
5. **Evaluation:** To evaluate our approach, we used the City dataset validation set, which contains images.
6. **Implementation Details:** We implemented our approach using the PyTorch deep learning framework, which provides efficient tools for building and training deep neural networks. We used a single NVIDIA Tesla V100 GPU for training and evaluation.

   Overall, our approach involved preprocessing the data, using pre-trained UNet model with different backbones, fine-tuning the models for semantic segmentation, training the models using Adam with miou, and evaluating the models using the City dataset validation set.


Training and val Accuracy
Accuracy
Epoch
train_accuracy
val_accuracy


Training and val Loss
Loss
Epoch
train_loss
val_loss


Training and val mIoU
mIoU
Epoch
train_mIoU
val_mIoU

5 Conclusion

In this paper, i investigated the impact of different backbone models, namely VGG11, VGG13, VGG19, and MobileNetV2, when integrated with the U-Net architecture for semantic segmentation on a city dataset. Through extensive experiments and evaluations, i gained valuable insights into the performance of these models and their suitability for urban scene understanding.

My results revealed the strengths and weaknesses of each backbone model in terms of segmentation accuracy and computational efficiency. The VGG models, with their deep architectures, demonstrated superior performance in capturing fine-grained details and producing accurate segmentation results. However, they also exhibited higher computational and memory requirements. On the other hand, MobileNetV2, with its lightweight design, showcased efficiency in terms of computational resources but had a slightly lower accuracy compared to the VGG models.

By comparing and evaluating these backbone models, i provided guidance for practitioners in selecting the most appropriate model based on their specific application requirements. For scenarios where high accuracy is paramount and computational resources are not a major constraint, VGG models such as VGG19 can be a suitable choice. On the other hand, if computational efficiency is crucial, MobileNetV2 offers a good trade-off between accuracy and efficiency.

My study contributes to the field of semantic segmentation in urban environments by providing insights into the performance of U-Net with different backbone models. By selecting an optimal backbone model, researchers and practitioners can develop more robust computer vision systems for urban applications, such as autonomous driving, urban planning, and environmental monitoring.

In conclusion, the evaluation and comparison of VGG11, VGG13, VGG19, and MobileNetV2 as backbone models for the U-Net architecture in semantic segmentation on a city dataset highlighted the trade-offs between accuracy and computational efficiency. This research facilitates the advancement of urban scene understanding and assists in the selection of suitable backbone models for improved computer vision systems in urban environments.

Links for Code and Report

Github Repository for Code:
https://github.com/SaadShah11/CV_MSDS_Assignment3

Overleaf Link for Report: https://www.overleaf.com/read/nfnrnmfyjzhm

References

UNet: https://paperswithcode.com/method/u-net

8 *Semantic Segmentation*

Segmentation$_m odels$:

MobileNet V2: https://pytorch.org/vision/main/models/mobilenetv2.html

UNet using Pytorch: https://github.com/milesial/Pytorch-UNet