ADDIS ABABA UNIVERSITY

Addis Ababa Institute of Technology (AAiT)

Information Technology and Scientific Computing ITSC

Big Data Modelling and Management System

Semantic Urban Scene Segmentation Using VGG-19

Prepared by:

1. *Alefew YimerATR/0973/10*
2. *Yuideg Misganew ATR/2378/10*
3. *Getahun Honelet ATR/3360/10*

Submitted to: - Amanuel Negash.

Jan 2021

Abstract

Image recognition and urban scene analysis are critical in developing smart, sustainable modern cities. In this context, applying the VGG-19 Image Recognition Model has drawn attention as a powerful tool for identifying objects, features, and characteristics in urban scenes. Known for its complexity and high accuracy, the VGG-19 model enables in-depth and precise analysis of urban images. This project applies the VGG-19 model to urban scene analysis, exploring its usage in applications such as traffic monitoring, building and object identification, nature and sky classification, human and vehicle detection, and smart urban planning. Combining artificial intelligence and computer vision, this research aims to understand the potential of the VGG-19 model in enhancing urban environment comprehension and management.

Keywords: VGG-19, image recognition, urban scene analysis.

Chapter I: Introduction

1.1 Background

Image recognition and urban analysis have become rapidly developing research areas in the current digital era. Image segmentation is a suitable choice for analyzing image data. Segmentation is a process of identifying objects and separating specific areas in an image into more distinct parts or classes. This segmentation aims to view and identify relevant areas in the image. Given the ever-evolving role of technology, it can assist in analyzing urban development so that urban aspects and dynamics can be better understood. The analysis of cityscapes has increasing relevance in the context of sustainable urban planning. One of the notable technological breakthroughs is the use of image recognition models using VGG-19, which can recognize visual elements.

Using image recognition technology like VGG-19 allows us to automatically identify and classify the elements of an image with a high degree of accuracy, which can then be used in various applications. The application of technology such as the VGG-19 model in cityscape analysis has great potential to generate valuable data that can be used by stakeholders such as researchers, architects, and others. Developed by the Visual Geometry Group (VGG) team at the University of Oxford, VGG-19 has successfully demonstrated its ability in image recognition and classification with a very high degree of accuracy. This model consists of 19 convolutional layers and is capable of understanding complex features in images, allowing us to perform a more in-depth and accurate analysis of cityscapes. In the research project we conducted, the technical steps required to implement the VGG-19 model according to the dataset used, namely cityscapes, will be discussed, starting with image collection, preprocessing, to feature extraction.

Exploratory case studies show that the VGG-19 model has been used to analyze cityscapes in several major cities. The results of this analysis are expected to provide in-depth insights and understanding of the application of image recognition technology in the context of sustainable urban planning.

1.2 Problem Statement

Problem statement is an important step in research that helps to detail the problems that will be solved through the research. The following are some problem statements that may be relevant to include in the context of our research:

1. How can the VGG-19 image recognition model be efficiently applied to analyze urban scenes?
2. How do variations in light, perspective, and urban structural complexity impact the model's performance?
3. What challenges or limitations exist in applying the VGG-19 model for urban scene analysis, and how can they be addressed?

1.3 Research Objectives

Based on the research problem formulation explained above, the research objectives and benefits are formulated as follows:

1. Implement the VGG-19 Image Recognition Model.
2. Analyze the performance of the VGG-19 model in urban scene contexts.
3. Investigate challenges and limitations in implementing the model.
4. Provide recommendations for future development.

Chapter II: Literature Review

2.1 Transfer Learning

In 2010, Pan et al proposed the concept of learning unknown knowledge through existing knowledge, known as Transfer Learning. The core concept of this learning is to find similarities between existing knowledge and unknown knowledge. Some knowledge domains are too abstract to learn, resulting in high overall learning costs. Therefore, using existing knowledge to aid learning is important [10].

The core concept of transfer learning is how to find relevance between the known and the unknown and learn new knowledge. In transfer learning, existing knowledge is usually called the Source Domain, and unknown knowledge is called the Target Domain. This learning primarily studies how to migrate knowledge from the Source Domain to the Target Domain. In the field of machine learning, transfer learning focuses on applying existing knowledge to the unknown through a defined model.

2.2 Deep Learning

It is part of a machine learning method based on artificial neural networks with representation learning. The adjective "deep" refers to the use of many layers in the network. The methods used can be supervised, semi-supervised, or unsupervised [11]. Architectures such as deep neural networks, deep belief networks, recurrent neural networks, convolutional neural networks, and transformers have been applied to various fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection, and board game programs, producing results that are comparable and in some cases surpass human expert performance [12][13][14]. In deep learning, the CNN or Convolutional Neural Network method is very good at finding good features in images to the next layer to form nonlinear hypotheses that can increase the complexity of a model. Complex models will certainly require a long training time, so the use of GPUs is very common in the deep learning world [15].

2.3 VGG Architecture

VGG stands for Visual Geometry Group; it is a standard Convolutional Neural Network (CNN) architecture with many layers. The term "deep" refers to the number of layers, with VGG-16 and VGG-19 consisting of 16 and 19 convolutional layers, respectively. The VGG architecture forms the foundation of innovative object recognition models. Developed as a deep neural network, VGGNet surpasses baselines in many tasks and datasets beyond ImageNet. Additionally, this architecture remains one of the most popular image recognition architectures today.

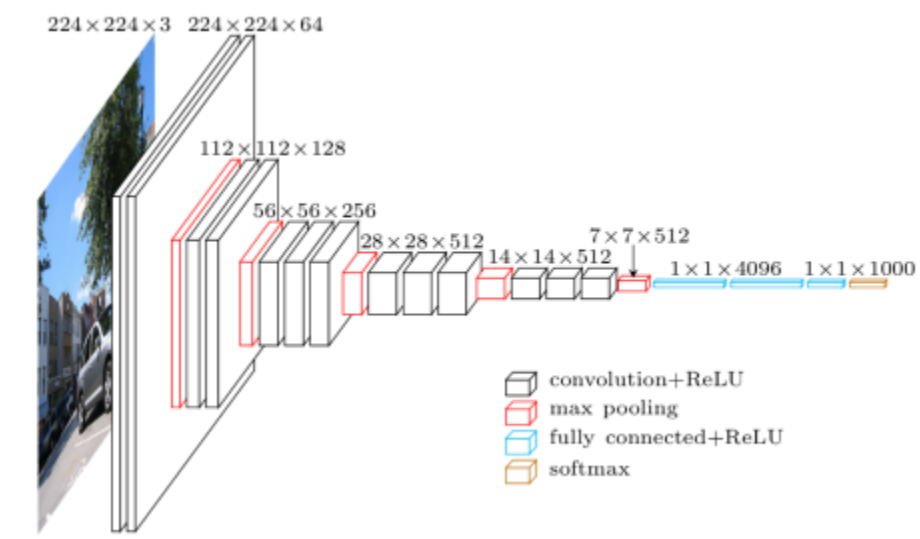

Figure 1: Visual Geometry Group

- **VGG-16**
  The VGG16 model achieved nearly 92.7% top-5 testing accuracy on ImageNet. ImageNet is a dataset comprising over 14 million images across almost 1,000 classes. Additionally, it is one of the most popular models submitted to ILSVRC-2014. It replaces large kernel-sized filters with several 3x3 kernel-sized filters in sequence, resulting in significant improvements over AlexNet. The VGG16 model was trained using an Nvidia Titan Black GPU for several weeks. As mentioned earlier, VGGNet-16 consists of 16 layers and can classify images into 1,000 object categories, such as keyboards, animals, pencils, mice, etc. Moreover, this model has an input image size of 224 x 224.

- **VGG-19**

  The concept of the VGG19 model (also known as VGGNet-19) is similar to VGG16, except that it supports 19 layers. The numbers “16” and “19” indicate the number of weight layers in the model (convolutional layers). This means that VGG19 has three more convolutional layers than VGG16. Further details on the characteristics of the VGG16 and VGG19 networks are discussed later in the article.

2.4 Image Segmentation

Image segmentation is the process of dividing an image into specific parts or objects based on pixel dimensions. This separation is done by identifying distinct differences or similarities in pixel intensity within the image's dimensions. Classical methods such as thresholding, clustering, and texture-based object separation have laid the foundation for segmentation. More modern approaches, like convolutional neural networks in deep learning, enable image segmentation to become precise and adaptive. Other methods, such as graphical approaches like watershed and region-based techniques, contribute to object separation and a deeper understanding of image structures.

Chapter III: Research Methodology

3.1 Key Concepts

A. VGG-19 Image Recognition Model:
   Convolutional Neural Network (CNN) Architecture: VGG-19 is a deep CNN architecture designed for image recognition tasks. It consists of 19 layers (hence the name "VGG-19") and excels at extracting features from images. The convolutional layers in VGG-19 play a central role in the feature extraction process, generating key features that help the model recognize patterns and objects in images. This capability is particularly useful for urban scene analysis and other image recognition tasks.

B. Transfer Learning:
   Transfer Learning Concept: This section explains the concept of transfer learning and why VGG-19 is often used as a base model for image recognition tasks. It details how a model pre-trained on a large dataset can serve as a foundation for specialized tasks like urban scene analysis.

C. Feature Extraction:
   Feature Extraction in Specific Layers: Explains why certain layers of VGG-19 are frequently utilized for feature extraction in urban scene analysis. These layers effectively capture detailed and meaningful features crucial for segmentation and classification tasks.

D. Evaluation and Performance:
   Performance Metrics: Discusses the performance metrics used to evaluate the model, such as accuracy, F1-score, or other task-specific metrics. These metrics assess the model's effectiveness in meeting the goals of the analysis.

3.2 Methods

This project uses urban scene data focusing on the semantic understanding of urban street environments, which will be processed using a Convolutional Neural Network (CNN). The dataset consists of 30 classes grouped into 8 categories (flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void). Each image has a resolution of 256 x 512 pixels and was captured from 50 cities. Each file contains a combination of the original photo on the left side and the labeled image (semantic segmentation output) on the right side.

A. Data Collection

The data required for this research consists of Cityscapes image pairs captured from a top-down view of the Earth's surface, focusing on flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void. The Cityscapes image pairs are then divided into training and validation datasets, containing 2,780 and 695 images, respectively [7].

B. Data Analysis

The training and testing input data, in the form of satellite-like top-down images, must meet specific labeling requirements:

- Objects labeled as foreground must not contain holes; if the background is visible "through" some foreground objects, it is considered part of the foreground.
- This also applies to regions mixed with two or more classes; such regions are labeled as part of the foreground class.
- Examples include tree leaves in front of a house or sky (all trees) and transparent car windows (all cars)[8].

C. VGG-19 Selection

Why Choose VGG-19?

1. Complexity of Urban Scenes:
   Urban scenes are highly complex, comprising various elements such as buildings, highways, parks, vehicles, and more. A deeper model like VGG-19 is better suited to handle this complexity due to its greater number of layers and capacity to learn intricate image features.
2. Deeper Feature Representation:
   VGG-19, with its additional layers, typically has a superior ability to learn deeper and more complex feature representations. This can result in higher accuracy for tasks like segmentation and classification.
3. Efficiency through Transfer Learning:
   Pre-trained VGG-19 models already have robust feature representations for objects commonly found in urban scenes. This helps save time and computational resources during the training process.

Below is Figure 3, which depicts the plot of the VGG-19 model that has been created. The plot shows several convolutional layers (Conv2D) and pooling layers (MaxPooling2D) used during model training. Each convolutional layer is responsible for extracting complex features from the images, while the pooling layers (MaxPooling2D) aim to reduce the dimensionality of the images and decrease the number of parameters while retaining essential information in the images.

The VGG-19 model architecture comprises multiple convolutional blocks designed to extract features from images that are input into the transfer learning model. Each block consists of several convolutional layers and pooling layers, where each layer has specific parameters that are configured or adjusted during model training. This allows the model to effectively learn the appropriate feature representations from the dataset being used.

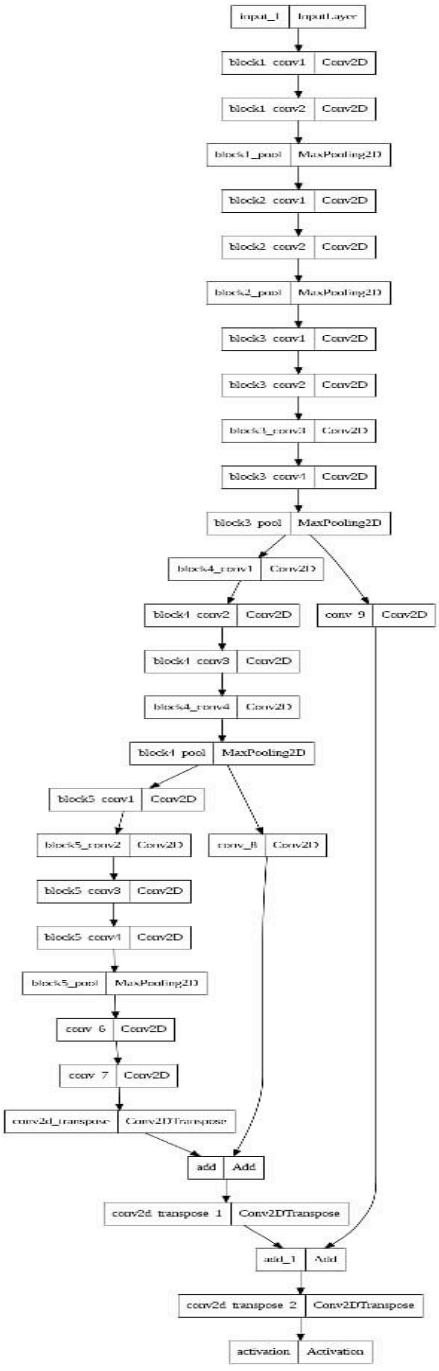

Figure 3: Visualization of Model Architecture

In preprocessing, scaling adjustments are made to facilitate neural network software in processing the data. Additionally, image rotation is altered to enable the model to understand objects from various orientations. A data generator is created to train and test the neural network model. The generator dynamically retrieves images from the designated directory during training. Using a generator is beneficial to avoid loading the entire image dataset into memory at once.

Chapter IV: Discussion

4.1 Results and Discussion

The Cityscape dataset is a collection of images extracted from a video taken from a moving vehicle in Germany. The images in the dataset include original images and semantic segmentation labels. Semantic segmentation labels are generated through image processing and computer vision techniques, which categorize the pixels in an image into specific objects or areas. Semantic segmentation produces pixel mappings that provide detailed information about the objects and elements in the image.

Figure 2. Original Image

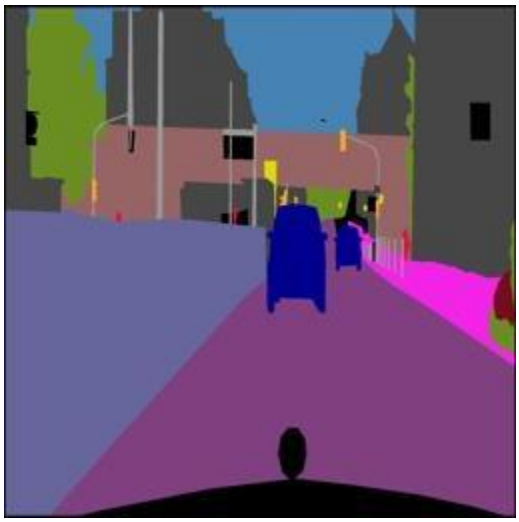

Figure 3. Semantic Segmentation

In Figure 2, the image shows the original data from the Cityscape dataset, which is a still frame extracted from a video captured by a vehicle-mounted camera. Figure 3, on the other hand, represents the semantic segmentation of the original image, where various objects are differentiated by distinct colors. These data are further processed for model analysis and accuracy predictions of the VGG19 model on the Cityscape dataset.

Based on the method employed with the VGG19 model, changes in **val_loss** and **val_accuracy** were observed across training epochs. To produce accurate and reliable predictions, variations in learning rate, batch size, and epoch number were experimented with. The predicted results from the variations are shown below.

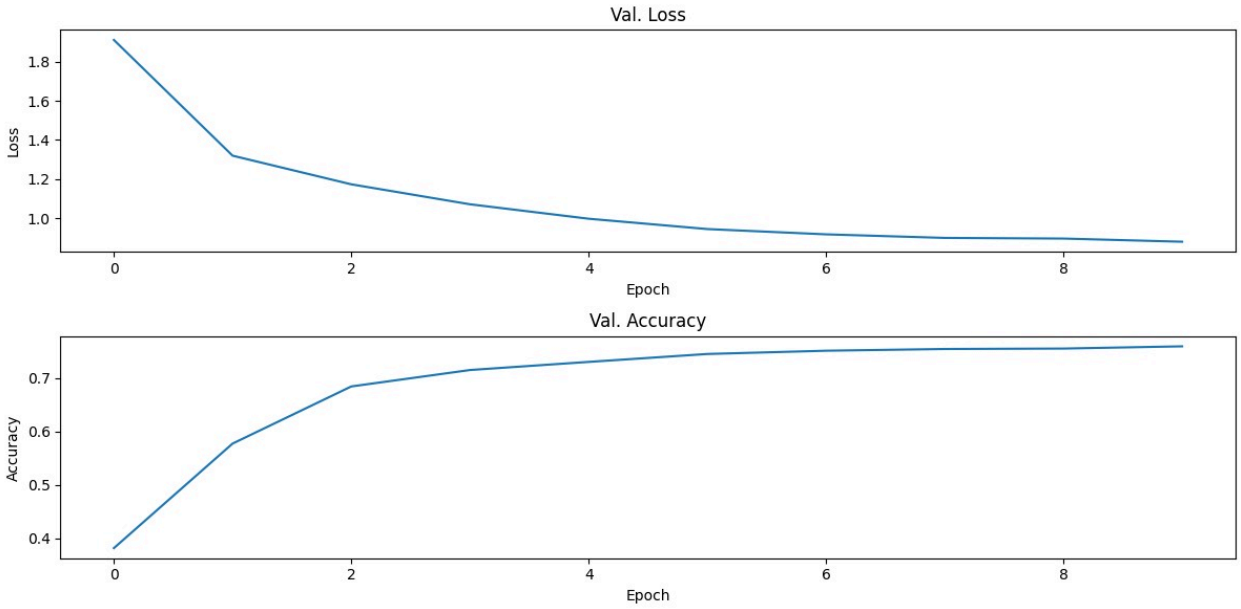

Figure 4. Visualization of Predicted Results

Based on Figure 4, the model achieved an accuracy of 78% at epoch 10. As accuracy increases with additional epochs and loss decreases, it indicates that the model is performing quite well. The segmentation prediction results are visualized to demonstrate the quality of the segmentation, where the accuracy graph steadily increases with more epochs, and the loss graph correspondingly decreases. The visualization highlights the predictions made on the Cityscape dataset using the model and its parameters.

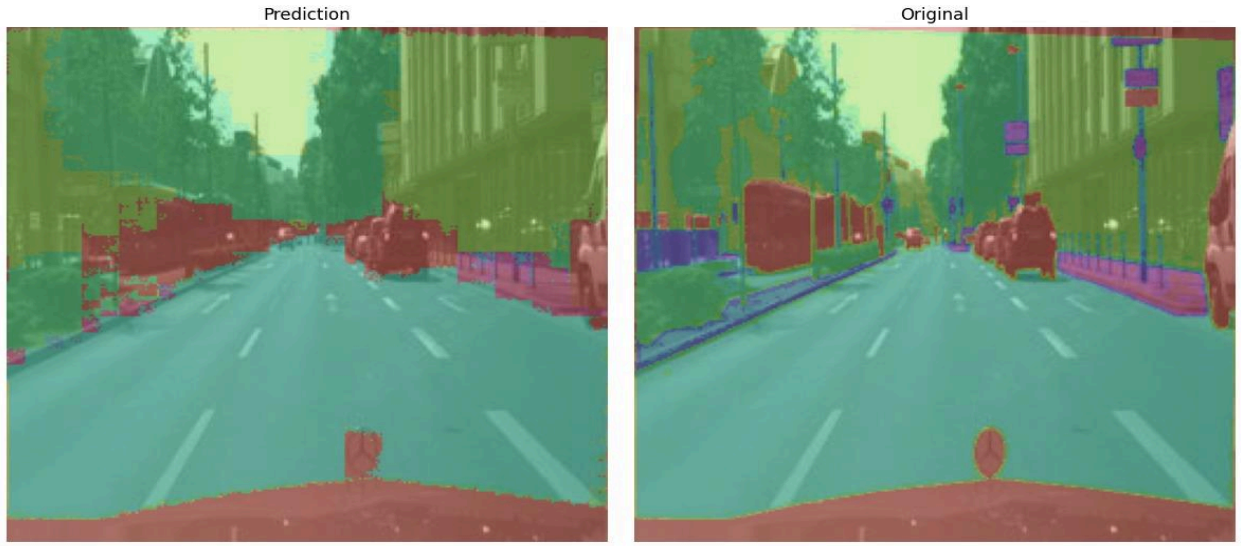

Figure 5. Segmentation Prediction Results

In Figure 5, the visualization of the model's predictions shows an accuracy of 78% on the dataset. This visualization indicates that certain objects are not fully covered by the segmentation colors, which aligns with the model's segmentation accuracy score.

Chapter V: Conclusion and Recommendations

5.1 Conclusion

In the experiment we conducted, we used the Cityscapes dataset, which contains images from videos captured by vehicles in 50 cities, including original images and semantic segmentation. Through preprocessing that involved pixel intensity normalization, image rotation, and the use of a data generator, we successfully trained and tested the neural network model. However, the training results obtained were less than optimal as we only inputted 2 epochs. From the 2 input epochs, an accuracy of 78% was obtained. The provided architecture visualization gives the reader an overview of how the model architecture works.

5.2 Suggestions

The trained model has shown promising results for use in subsequent image segmentation tasks. Considering the training results and segmentation visualization, the VGG19 model can be effectively used for image segmentation tasks with the Cityscapes image pairs dataset. Further evaluation is needed to assess the model's performance in depth and for broader implementation.

References

[1] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

[2] Yang, X., Luo, P., Loy, C. C., & Tang, X. (2015). WIDER FACE: A Face Detection Benchmark. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* (pp. 3320–3328).

[5] Arsanjani, J. J., Sayyad, G., Sarai, A., & Wakefield, S. (2013). Integrating space syntax into GIS for evaluating the impact of urban form on pedestrians' walking behavior: The case of Siena, Italy. *Environment and Planning B: Planning and Design, 40*(3), 505–527.

[6] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the ACM International Conference on Multimedia (MM)* (pp. 675–678).

[7] Cityscape dataset overview

[8] Cityscapes dataset official site

[10] J. C. Hung, K. C. Lin, & N. X. Lai, “Recognizing Learning Emotion Based on Convolutional Neural Networks and Transfer Learning,” *Applied Soft Computing Journal, vol. 84*, p. 105724, 2019, doi: 10.1016/j.asoc.2019.105724.

[11] LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). "Deep Learning." *Nature, 521*(7553), 436–444.

[12] Ciresan, D.; Meier, U.; Schmidhuber, J. (2012). "Multi-column Deep Neural Networks for Image Classification." *2012 IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3642–3649.

[13] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey (2012).

[14] *TechCrunch.* May 25, 2017. Archived on June 17, 2018. Retrieved June 17, 2018.

[15] Danukusumo, Kefin Pudi. (2017). *Implementation of Deep Learning Using Convolutional Neural Networks for Image Classification of Temples Based on GPU.*