



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكرو فيلم

بسم الله الرحمن الرحيم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكرو فيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكرو فيلم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



MONA MAGHRABY



AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering

An Improved Semantic Segmentation for Autonomous Driving

A Thesis submitted in partial fulfilment of the requirements of
M.Sc. in Department of Computer and Systems Engineering
Electrical Engineering

by

Taha Mohamed Ahmed Ibrahim Emara

B.Sc. in Electronics & Communications Department
Electrical Engineering

High Institute of Engineering, El-Shrouk Academy, 2013

Supervised By

Prof. Dr. Hazem Mahmoud Abbas

Prof. Dr. Hossam Eldin Hassan Abdelmunim

Cairo, 2021



AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering

An Improved Semantic Segmentation for Autonomous Driving

by

Taha Mohamed Ahmed Ibrahim Emara

Examiners' Committee

Name and affiliation

Signature

Prof. Dr. Mohsen Abdelrazek Rashwan

Electronics and Communications Engineering

Faculty of Engineering, Cairo university.

.....

Prof. Dr. Hazem Mahmoud Abbas

Computer and Systems Engineering

Faculty of Engineering, Ain Shams University.

.....

Prof. Dr. Mahmoud Ibrahim Khalil

Computer and Systems Engineering

Faculty of Engineering, Ain Shams University.

.....

Prof. Dr. Hossam Eldin Hassan Abdelmunim

Computer and Systems Engineering

Faculty of Engineering, Ain Shams University.

.....

Date: February 2021

Statement

This thesis is submitted as a partial fulfillment of Master degreee in Department of Computer and Systems Engineering, Faculty of Engineering, Ain shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Taha Mohamed Ahmed Ibrahim Emara

Signature

.....

Date: February 2021

Researcher Data

Name: Taha Moahmed Emara

Date of Birth: 01/07/1991

Place of Birth: Damietta, Egypt

Last academic degree: B.Sc.

Field of specialization: Electronics & Communications Engineering

University issued the degree : High Institute of Engineering, El-Shrouk Academy

Date of issued degree : June 2013

Current job : Deep learning Engineer at Cisco

Summary

The First step of autonomous car is based on visual scene understanding of the surrounding environment. This visual understanding entails identification and localization of surrounding objects. Developing a semantic image segmentation architecture, for segmenting the entire view into regions and assigning a semantic label to these regions, lies at the heart of this problem. This thesis proposes an effective and efficient semantic image segmentation model for autonomous driving.

In the last several years, semantic image segmentation likes other computer vision tasks as object detection and image classification, has seen considerable advancements due to the employment of deep learning architectures, especially convolutional neural networks *CNN*. Training such architectures to obtain a high level of accuracy requires a very complex model. Being the autonomous driving a critical real-time application, computationally efficient models are needed. Also, edge devices as mobile phones have a low capacity of computational power. Also, this requires a specific and efficient deep neural networks. Although there are many ways to design deep neural networks and the availability of efficient training hardware, still designing a high accuracy and computationally efficient models is very challenging.

This thesis focuses on providing efficient deep neural networks for semantic image segmentation at two concurrent levels.

Computationally Efficient Model: we designed lightweight neural networks for semantic segmentation by following up the encoder-decoder structure, employing lightweight efficient backbone networks, and designing lightweight efficient decoder module.

High Accuracy Model: while considering the computational cost of the proposed models in our mind, we also consider the performance efficiency in terms of accuracy by employing long and short residual connections and designing efficient module called Deeper Atrous Spatial Pyramid Pooling (DASPP) to capture the extracted features by the encoder section at multi-level context.

We evaluated our model on the standard dataset Cityscapes. Also, in our evaluation procedure, we evaluated our model on severe weather condition on the standard dataset

Foggy Cityscapes. A three variant of semantic segmentations model are proposed to provide multiple trade-offs between accuracy and computational efficiency. Our model LiteSeg-Mobilenet can achieve 161 frame per second (FPS) with mean intersection over union (mIOU) of 67.81% while the previous state-of-the-art ESPNet on the same hardware can achieve 144 FPS with mIOU of 60.3% on the standard Cityscapes test set.

Key words: Semantic Image Segmentation, Convolutional Neural Network, Fully Convolutional Network, Encoder-Decoder Architecture

Acknowledgment

Taha Emara

Computer and Systems Engineering

Faculty of Engineering

Ain Shams University

Cairo, Egypt

February 2021

First, I am so grateful to Allah Almighty for his blessing and grace in giving me the ability to finish this work.

I want to express my appreciation and gratitude to Prof. Dr. Hossam Eldin Hassan for his guidance and continuous support over the years during the research and the writing stages. He has been supportive from the first day I began working on my research till the end. His insightful discussions and suggestions on the research helped me finish this work successfully.

I would like to thank Prof. Dr. Hazem Abbas for his guidance, support, insightful discussions, and inspirations to us. He introduced me to the field of machine learning and deep learning during his informative and inspired lectures.

Also, I would like to especially thank my parents for their unconditional love, care, and support. I also would like to thank my wife, for being my pillar of support and her understanding during this work.

Lastly, I would also like to thank the deep learning and the open-source communities for providing us tools, frameworks, and tutorials that made many things easy to understand and implement.

Contents

Contents	vii
List of Figures	xiii
List of Tables	xvii
Abbreviations	xix
1 Introduction	1
1.1 Statement of the problem	1
1.2 History of autonomous driving	3
1.3 Levels of autonomous driving	5
1.4 Rule of AI and Semantic Segmentation models in autonomous driving . .	6
1.5 Goal of The Thesis	8
1.6 Organization of Thesis Chapters	8
2 Artificial Neural Network and Deep Learning	11
2.1 Machine Learning	11
2.1.1 Types Of learning	12
2.1.1.1 Supervised Learning	12
2.1.1.2 Unsupervised Learning	13
2.1.1.3 Semi-Supervised Learning	14
2.1.1.4 Reinforcement Learning	14
2.2 Neural Network	15
2.2.1 Neuron Model	15
2.2.2 Structure of Neural Networks	16
2.2.3 How Neural Networks Make Complex Designs	16
2.2.4 Activation Functions	17
2.2.4.1 Sigmoid	17
2.2.4.2 Hyperbolic Tangent (Tanh)	18
2.2.4.3 Rectified Linear Units (RELU)	19
2.2.4.4 Importance of using non linear activation function.	19
2.2.5 Training of neural networks and Cost Functions	20
2.2.6 Stochastic gradient descent	20
2.3 Convolutional Neural Network (CNN)	22
2.3.1 Convolution Layer	24
2.3.2 Motivation	26
2.3.3 Pooling	26

3	Convolution Neural Networks for Image Classification	29
3.1	Evolution Of Image Classification	29
3.2	Complex Models	30
3.2.1	AlexNet	30
3.2.2	VGG	31
3.2.3	Inception Models	33
3.2.4	ResNet	35
3.3	Lightweight Models	37
3.3.1	MobileNet	37
3.3.2	ShuffleNet	40
3.3.3	Darknet19	42
4	Convolution Neural Networks for Semantic Segmentation	45
4.1	Evolution Of Semantic Image Segmentation	45
4.1.1	Grayscale segmentation	45
4.1.2	Conditional random fields	46
4.2	CNN Based Semantic Segmentation	47
4.2.1	FCN	48
4.2.2	SegNet	50
4.2.3	DeepLabv3+	51
4.3	Real-time Semantic Segmentation	53
4.3.1	ENet	54
4.3.2	ERFNet	55
4.3.3	ESPNet	58
4.3.4	RTSeg	61
5	Proposed Method	63
5.1	Proposed Encoder	63
5.1.1	Backbone Network	63
5.1.2	DASPP	65
5.2	Proposed Decoder	65
5.3	Atrous Convolution	66
5.4	Depthwise Separable Convolution	67
5.5	Long and short residual connection	68
5.6	Upsampling Layers	68
5.6.1	Deconvolution Layers	69
5.6.2	UnPooling Layers	70
5.6.2.1	Interpolation Methods	70
5.7	Output Layer	71
5.7.1	Loss Function Formulation	72
6	Results and Discussion	75
6.1	Evaluation Metrics	75
6.2	Dataset and Computing Environment	76
6.3	Software and Hardware Setup	77
6.4	Training Protocol	78
6.5	Encoder Options	78

6.6	Computational Performance Evaluation	80
6.7	Cityscapes Benchmark Results	80
6.8	Foggy Cityscapes Results	82
7	Conclusion and Future Works	121
7.1	Future Works	121
7.1.1	Knowledge Distillation	122
7.1.2	Model Quantization	122
7.1.3	Winograd fast convolution	123
7.1.4	Augmentation using video data	125
	Bibliography	127

List of Figures

1.1	(a)Stanley car with a laser sensor: the sensor is pointed downwards to scan the ground in front of the vehicle as it is moving. Stanley carries five such sensors, which are mounted at five different angle. (b) Each laser acquires a 3D point cloud over time. Figure from [1].	2
1.2	An example of input RGB image, corresponding semantic segmentation mask , and colour map from Cityscapes dataset [12].	3
1.3	The radio-operated American Wonder. Image From [21]	4
1.4	The bicycle-wheeled Stanford Cart. Image from [23]	4
1.5	Example of conditions when autonomous driving system may fail. Image sources [25–29].	7
2.1	Relationship between artificial intelligence, machine learning, neural networks, and deep learning.	12
2.2	Visualization of the high dimensional MNIST dataset in 2D after applying dimensionality reduction algorithm.	13
2.3	Artificial neuron model.	15
2.4	A fully connected neural network one input layer, one hidden layer, and one output layer.	16
2.5	Plot of Sigmoid function and its derivative.	17
2.6	Plot of Tanh function and its derivative.	18
2.7	Plot of RELU function and its derivative.. . . .	19
2.8	A cost surface $J(\theta)$. The gradient computed at one point in parameter space is indicated by $\partial J(\theta)$. The gradient descent method takes one step in the direction of steepest descent as indicated by $-\mu\partial J(\theta)$	21
2.9	Visualization of the topographical mapping between cortex cells and visual field. Image from [44].	23
2.10	Overview of convolutional and pooling layers.	24
2.11	Numerical example of convolution processing on RGB image.	25
2.12	Example of the computation of the max pooling layer.	27
2.13	Example of the computation of the average pooling layer.	27
3.1	Top-1 Accuracy of different image classification models on ImageNet challenge, from [52].	30
3.2	Diagram of the naive Inception module and the Inception module with dimension reduction.	34
3.3	GoogLeNet architecture diagram. Image from [8].	35
3.4	ResNet Block.	36

3.5	Accuracy versus the number of parameters and the computational cost for different models, from [52]. Note that the accuracy of the model is directly associated with the number of operations needed.	38
3.6	RELU6 activation function.	39
3.7	MobileNet v1 vs MobileNet v2 Convolution Blocks.	40
3.8	(a) is a simple group convolution, the output is determined only by a part of the input, and the effect is naturally poor. While (b) and (c) are the rearrangement of the channel after the first convolution, and then the second convolution, the output is All input decisions. Image from [31] . .	41
4.1	Overview of using CRF in semantic segmentation. When the model predicts the semantic label of each pixel separately, pixels from dog class mixed with pixels from cat class (image c). A more reliable segmentation results is shown in image (d) when we consider the neighbouring relationship between different pixels. Image from [55].	46
4.2	Grid CRF vs Dense CRF. Grid CRF leads to over smoothing around boundaries. Dense CRF is able to obtain fine boundaries. Image from [55].	47
4.3	Fully Convolutional Network by Long et al [2]. transforming a classification model by using convolution layer instead of fully connected layers to create spatial heatmaps. Using a deconvolution layer for upsampling allows for dense inference and pixel labelling learning. Image from [2]. . .	49
4.4	Skip connection approach to feed the low-level features into to the later layers to improve the model accuracy. Image from [2].	49
4.5	The SegNet decoder-encoder architecture. In which there are no fully connected layers and so it's all convolution. A decoder uses transferred pool indices from its encoder to upsample its input to generate a sparse map(s) of features. Image from [56].	50
4.6	Unpooling layer operation, suppose a, b, c, and d are the pixel values of the feature maps in encoder module. In the decoder module, the upsampling layer returns each one of these pixel values only back to the index position that has been recorded in the pooling layer, otherwise Fill in the pixels of with blank (0).	51
4.7	Deeplab-v3+ proposed architecture. Image from [3]	52
4.8	Proposed bottleneck module by ENet [17].	54
4.9	Proposed bottleneck module by ENet [17].	56
4.10	ERFNet proposed architecture by [15].	57
4.11	ESP Block with Hierarchical Feature Fusion (HFF). Image From [16]. . .	59
4.12	Visualization of the effectiveness of using Hierarchical Feature Fusion (HFF) module. Image from [16]	59
4.13	The network diagram of the ESPNet. Image from [16].	60
4.14	Different proposed meta-architectures by RTSeg, a)SkipNet, b)UNet. Image from [20]	61
5.1	General LiteSeg diagram including encoder, decoder and DASPP module.	64
5.2	Illustration of atrous convolution in 2D. (a) 3×3 standard convolution (atrous convolution $rate = 1$); (b) Atrous convolution with $rate r = 2$; (C) Atrous convolution with $rate r = 3$	66