DEEP NEURAL NETWORKS ON EMBEDDED SYSTEMS

Rishav Sharma (_TEQIP Intern_)

ABOUTS:

Education:

Rishav Sharma is an undergraduate in Computer Science and Engineering
Jorhat Engineering College, Jorhat, India.

Working under Faculty Supervisor:

Dr. Sparsh Mittal.
Assistant Professor, Department of Computer Science and Engineering, IIT Hyderabad
Research Interests:
Computer architecture (CPUs and GPUs), processor architectures for machine learning, neural network accelerators,VLSI, approximate computing.

Guided by :

Nandan Kumar Jha
M.Tech in Computer Science and Engineering, IIT Hyderabad

Internship Provided by:

Technical Education Quality Improvement Programme (TEQIP) Cell, IIT Hyderabad.

Internship Duration :

One Month (15th June'18 - 15th July'18)

Abstract:

DNNs (Deep Neural Networks) are very popular and have become mainstay in computer vision and natural language processing. DNN processing requires huge computations and power. High resource requirements (in terms of memory, power and compute) of DNN prohibits it's wide deployment across edge devices. Several networks (MobileNet, SqueezeNet, SqueezeNext etc.) have been proposed for resource constraint platforms. These proposed nets have less parameters and hence reduce the storage requirements on embedded platform. We have used Raspberry Pi 3 Model B, which has quad core Cortex A53 @1.2GHz with 1GB RAM, as embedded platform and evaluated wide range of DNNs (size varying from 2.3MB to 243MB)

Acknowledgments:

I am very much thankful to TEQIP team and also to MHRD, Govt of India for such an excellent internship programme. I would like to thank my guide Dr.Sparsh Mittal and mentor Mr. Nandan Kumar Jha at IIT Hyderabad.

My Contribution:

Till date Deep Neural Networks have not been extensively studied and evaluated on embedded platforms like Raspberry Pi. There are claims for some models to have run efficiently on embedded platforms. But from our experiments we observed that very few models run with efficient consumption of resources.

Conclusion:

Not only model size matters: AlexNet having size 243.9MB completed running successfully but Xception Net having 91MB size does not run.

RAM consumption and model size are not proportional: AlexNet having size 243.9MB consumes 497 MB of RAM but ResNet-101 having size 178.8 consumes 834 MB of RAM even if the model size is lesser than the former.

Inference time and model size are not proportional: SqueezeNet V1.1 and SqueezeNet V1.0 both have size 5MB but inference time is twice for the second.

Learning Accomplishments:

I started the internship without any idea of Machine Learning, Deep Neural Networks and various frameworks used. I have learnt a lot about these fields and have worked in Caffe on my system(debian 9) and Raspberry Pi 3(ubuntu 16.04) for experiments. To understand the tools I have learnt Core Python and Python visualization, for source code version control, I have learnt git and bitbucket. From TEQIP seminars by Mr. Nandan K. Jha, Ms. Poonam Rajpoot, Mr. Maruthi S. Inukonda, I have learnt Convolution Neural Network, Autonomous Driving Systems, Containerization of GPUs.

Results:

Sl No.	Model Name	Model Size (MB)	Inference time (Batch Size=1) ms	RAM (Batch Size=1) MB
1	AlexNet	243.9	1713	497
2	GoogleNet	53.5	2868	135
3	SqueezeNet V1.0	5	1397	70
4	SqueezeNet V1.1	5	674	46
5	1.0-G-SqNxt-23	2.3	829	126
6	1.0-SqNxt-23	3	888	127
7	1.0-SqNxt-23v5	3.8	706	101
8	Zynqnet	10.1	898	74
9	MobileNet	17	1536	222
10	NIN	30	1485	84
11	DenseNet-121	32.3	6436	658
12	DenseNet-169	57.3	7793	816
13	Inception-V3	95.5	5031	410
14	ResNet-101	178.7	16145	834
15	ResNext50_32x4d	191	8042	690

A vote of thanks to TEQIP for adding in some knowledge as well as memories.