Deep Convolutional Neural Network for Large-Scale Scene Classification

Dave Ojika, Chijua Liu, Rishab Goel, Vivek Viswanath, Arpita Tugave, Shruti Sivakumar

GatorVision @ University of Florida

Introduction

- Humans are extremely proficient at perceiving natural scenes and understanding their contents, but computers are able to do even better
- **Scene classification** is an important problem for computer vision, with the aim of processing the kinds of images we encounter in everyday life
- **ILSVRC** (ImageNet's Large-scale Visual Recognition Challenge) is a challenge to develop algorithm and software for achieving improved performance of vision tasks such as object recognition, localization and scene classification
- **GatorVision**, a team of ECE graduate students at the University of Florida, presents their work on scene classification using convolutional neural networks

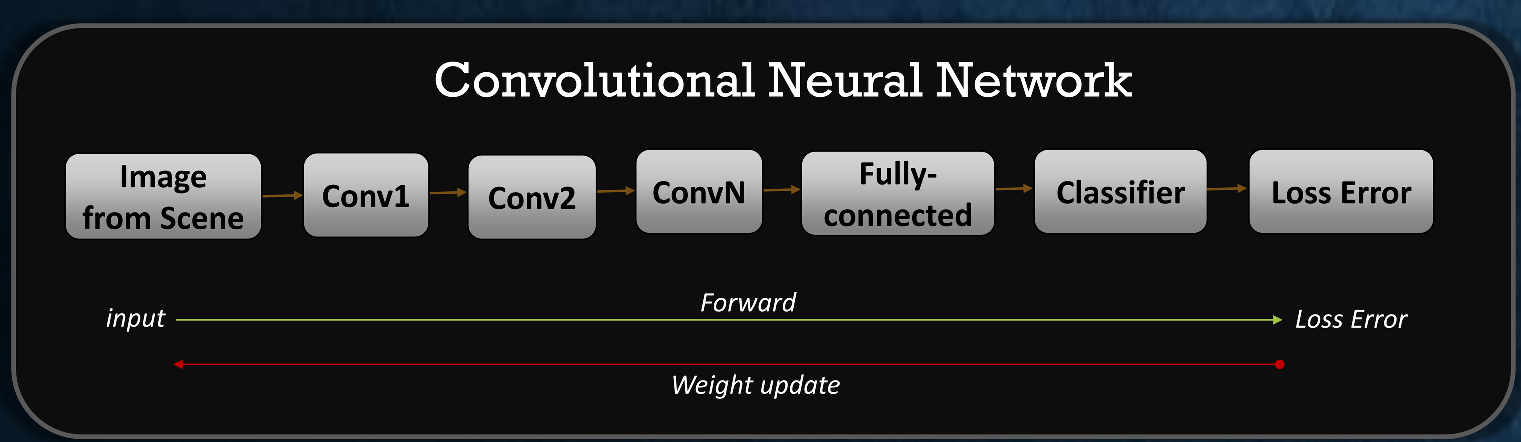
Summary

- We implement a Caffe-based convolutional neural network using the Places2 dataset for a large-scale visual recognition application. We leverage previous research experience by using very small 3 x 3 convolution filters in our architecture
- By varying depth of weight layers, we are able to obtain a suitable parameterization level for training a model towards improved recognition ability compared to configurations used in prior art
- Due to the very large amount of time required to train the model with deeper layers, we deploy Caffe on a multiple GPU environment and leverage optimized libraries from cuDNN to improve training time

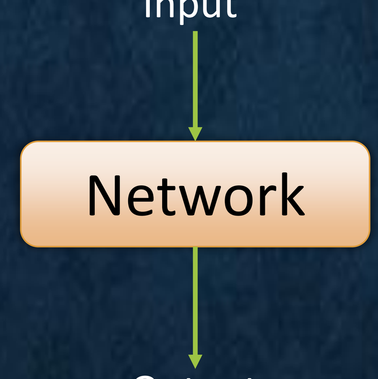
Deep Learning Framework

- **Caffe**: popular open-source software framework for designing and evaluating convolutional networks
  - Supports CPU and GPU for forward and backward passes; compatible with BLAS libraries
  - To accelerate performance, we deploy Caffe on a node with 2 Tesla K80 GPUs, along with cuDNN libraries
- **Network:** Modified VGG
  - 13 convolution layers*
  - 3x3 kernels
  - 3 fully-connected layers

  **All convolutional layers are followed by ReLU layer*



- **Task:** Train, validate and test
  - **Input**: 380,950 images from test-set, with 401 possible classes
  - **Output**: 5-top class predictions for every input image



Training and Testing

- **Dataset**

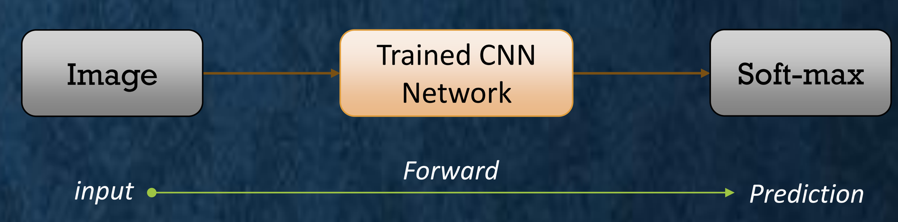
Dataset	Resolution	Train	Validation	Test
Places2	256 x 256	8,097,967	20,500	380,950

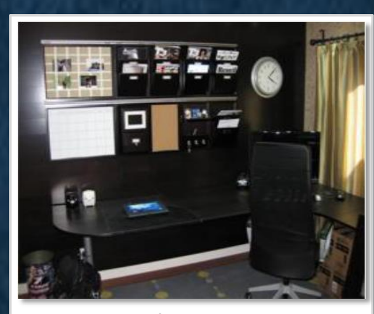
- **Training**
  - Start loss error: 5.83
  - Final loss error: 1.22
  - Iterations: > 160,000
  - *Learning rate manually adjusted during training*
  - *Best accuracy on validation-set: %55.4*

1. Convolution
2. Rectified Linear Unit
3. Fully-connected
4. Pooling
5. Soft-max (classification)

Performance-critical Layers

- **Testing**

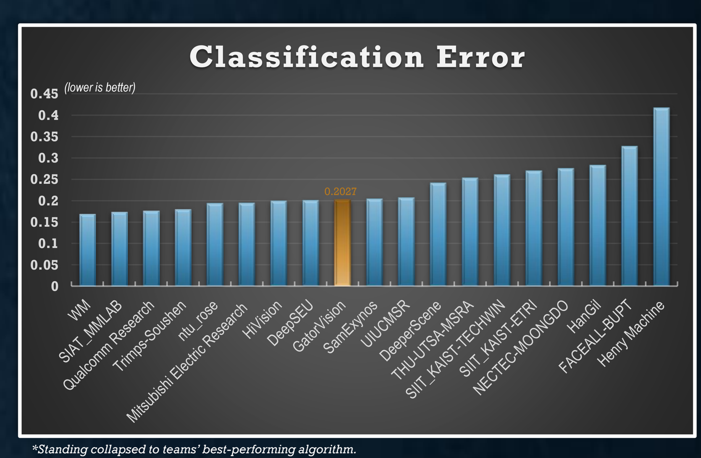




Image

Class #	Class	Probability
189	Home Office	0.365
263	Office	0.246
257	Music Studio	0.117
324	Server Room	0.087
356	Server Room	0.053

Prediction



**Standing collapsed to teams' best-performing algorithm.*

ImageNet Official Ranking

Conclusions & Feature Work

- Although, there is room for improvement, the prediction performance of our model is close to current state of the art techniques
- It is necessary to develop heterogeneous programming model to leverage new and emerging architectures such as FPGAs
- Low-power, mobile platforms can also benefit from CNN-based vision tasks by optimizing the network configuration for a more efficient performance-per-watt

Acknowledgments

Thanks to **NVIDIA** for provision of computing infrastructure with GPUs, and to **University of Florida** for support and supercomputing access to HiperGator. Finally, thanks to ImageNet for the opportunity to participate in ILSVRC 2015.

Contact: davido@ufl.edu

ImageNet and MS COCO Workshop

ICCV 2015
Santiago, Chile | 17th Dec