EFFICIENT AND VERSATILE DATA ANALYTICS FOR DEEP NETWORKS

Javier Espinosa
Postdoctoral Research Fellow, BSC-LAFMIA
javier.espinosa@bsc.es

and D. Garcia-Gausulla, J. Moreno-Vázquez, J. Conejero, G. Vargas-Solar,
R. Badia, U. Cortés, T. Suzummura

Work partially funded by CONACyT & CNRS
“The brain performs its canonical task (learning) by tweaking its myriad connections according to a secret set of rules”
- Biologically-inspired programming paradigm which enables a computer to learn from observational data
- Learn by repeatedly making tiny tweaks to neurons and synapse
- Used for classifying and clustering patterns (**images**, **sound**, **text**)
DEEP CONVOLUTIONAL NEURONAL NETWORK

- Performs object recognition with images (ex. faces, individuals, street signs)
- Each layer trains on a distinct set of features based on the previous layer’s output
DEEP CONVOLUTIONAL NEURONAL NETWORK

- Performs object recognition with images (ex. faces, individuals, street signs)

How to extract knowledge from Deep Networks?
Tiramisu

Knowledge Extraction Environment

Tiramisu-Engine

Storage

SPARK
Map-Reduce
PyCOMPSs
Parallel SDK

Caffe
TensorFlow
DL4J
Theano
Deep Learning Development Toolkit

Tiramisu
PRINCIPLE: BACKWARD ENGINEERING

“Seeking Activation Patterns”

LAYERS

L1  L2  L3  L4  L5  L6  L7  L8  L9
OBJECTIVE: PATTERN ANALYSIS

- church
- mosque
- bell cot
- mosque
- church
- stupa
- horse cart
- horse
- rickshaw
KNOWLEDGE EXTRACTION WORKFLOW

1. "IMAGE → VECTOR"

Images → HPC nodes → ImageVectors

Logs of activated neurons

Wordnet Ontology

1. "IMAGE → VECTOR"
 KNOWLEDGE EXTRACTION WORKFLOW

2. \( \{ \text{VECTOR}_N \} \rightarrow \text{CLASS}_V\text{ECTOR} \)

Buildings Vectors

Algebraic operations

HPC nodes

ClassVector

Buildings Features Vector
HPC nodes

KNOWLEDGE EXTRACTION WORKFLOW

3 “DATA ANALYTICS”

Data Analytics operation

HPC nodes

Similarity Matrixes

Vectors

Equations

PyComs Task

Data analytics operation

PyComs Task

Data analytics operation

PyComs Task

Data analytics operation


PROBLEM:
Simple I/O access
Data is the bottleneck
**Conjecture**: Optimizing in any two RUM areas negatively impacts the third

**Strategy**: Adaptive RUM methods based on application and hardware requirements
Distributed File System

Data transformation

Image2vector()

Classification

Wordnet classes

ImageVectors

Logs of activated neurons in each DN layer

Sharded & colocated Input data

INDEXING AND SHARDING
INDEXING AND SHARDING

Logs buckets resulting of applying operations to ImageVectors
INDEXING AND SHARDING

Similarity Matrixes

Data analytics operations

Buckets sharded by key – DN layer

Equations

INDEXING AND SHARDING

Similarity Matrixes

Data analytics operations

Buckets sharded by key – DN layer

Equations
Tiramisu data oriented workflow:
- Sequence of parallel analytics operations for processing a large datasets of images
- Generates data collections with incremental content (float vectors) and semantics (metadata)
- Write once, read many

Operations for knowledge extraction require:
- Clever data colocation across processing nodes and results storage
- Depending on the specific data representation, different types of persistence may be used, which must be consistent with the set of tools providing parallel analytic services

Well adapted data look-up, querying and exploration tools must be provided to ensure transparent access to data scientists