Supporting Data Science Workflows with MATLAB

Marta Wilczkowiak
Eugene McGoldrick
MathWorks
Domain Expertise

Mathematics
Statistics

Computer Science
“This year I was planning on trying out machine learning using TensorFlow or Caffe... but the barrier to entry seemed very high there.
PROTOTYPING, MODELLING & ANALYSIS
APPLICATION DEVELOPMENT
(ENTERPRISE) IMPLEMENTATION

Domain Expertise
Mathematics Statistics
Computer Science

TESTING: VALIDATION AND VERIFICATION

Domain Expertise

Computer Science

Mathematics Statistics
Examples
Getting started with Machine Learning

- User: Subject matter expert with no prior experience with machine learning
- Objective: Improve current methods using machine learning
- Characteristics: Many methods, but:
  - Similar interfaces
  - Similar problems: overfitting, trade-off false-positives/negatives
MATLAB has Apps
User had to choose a classifier to get started.

A complex tree is pre-selected so user can get started quickly.
User has to train each model one by one.

New options to allow users to train multiple models.
User could only see the legend for the first three classes.

User can see the legend for all the classes and can show/hide each class.
Deep Learning

- User: Subject matter expert with no experience with deep learning
- Objective: Feasibility study on applicability of deep learning
- Characteristics:
  - Large optimisation problem
  - GPUs are great to accelerate cost function calculations
  - Lots of parameters. Effectiveness of developing a network depends on ability to evaluate impact of different parameters
What is Deep Learning?

Deep learning performs **end-end learning** by learning **features, representations and tasks** directly from **images, text and sound**.

**Traditional Machine Learning**

- Manual Feature Extraction
- Classification

**Deep Learning approach**

- **Convolutional Neural Network (CNN)**
- **End-to-end learning**
- Feature learning + Classification
Why is Deep Learning so Popular?

- **Results:** Achieved substantially better results on ImageNet large scale recognition challenge
  - 95% + accuracy on ImageNet 1000 class challenge

- **Computing Power:** GPU’s and advances to processor technologies have enabled us to train networks on massive sets of data.

- **Data:** Availability of storage and access to large sets of labeled data
  - E.g. ImageNet, PASCAL VoC, Kaggle

---

<table>
<thead>
<tr>
<th>Year</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-2012</td>
<td>&gt; 25%</td>
</tr>
<tr>
<td>2012 (Deep Learning)</td>
<td>~ 15%</td>
</tr>
<tr>
<td>2015 (Deep Learning)</td>
<td>&lt;5%</td>
</tr>
</tbody>
</table>
Convolutional Neural Networks

Every feature map output is the result of applying a filter to the image. The new feature map is the next input.

Activations of the network at a particular layer

Input Image

Convolutions, rectified linear units, pooling, and fully connected layers for classification.
%% Define a CNN architecture
conv1 = convolution2dLayer(5,32,'Padding',2, 'BiasLearnRateFactor',2);
fc1 = fullyConnectedLayer(64,'BiasLearnRateFactor',2);
fc2 = fullyConnectedLayer(10,'BiasLearnRateFactor',2);
layers = [
    imageInputLayer([32 32 3]);
    convolution2dLayer(5,32,'Padding',2, 'BiasLearnRateFactor',2);
    maxPooling2dLayer(3,'Stride',2);
    reluLayer();
    ...
    fullyConnectedLayer(10,'BiasLearnRateFactor',2);
    softmaxLayer()
    classificationLayer()];

opts = trainingOptions('sgdm', 'InitialLearnRate', 0.001 ... 

% Training the CNN
[net, info] = trainNetwork(XTrain, TTrain, layers);
% everything except the last 3 layers.
layers = net.Layers(1:end-3);

% Add new fully connected layer for 2 categories.
layers(end+1) = fullyConnectedLayer(2, 'Name', 'fc8_2')

% Add the softmax layer and the classification layer which make up the
% remaining portion of the networks classification layers.
layers(end+1) = softmaxLayer
layers(end+1) = classificationLayer()

...
I was aware of many of the principles and techniques you covered but had no idea they were so easily accessible in MATLAB in comparison to other packages (R, Weka etc.)."

“This year I was planning on trying out machine learning using TensorFlow or Caffe... but the barrier to entry seemed very high there, so am happy that with MATLAB I can get cracking relatively quickly.”
Deploying MATLAB Data Analytics on the desktop and in Enterprise Production Systems

Eugene McGoldrick
MATLAB Analytics Deployment Options

- MATLAB Compiler and Compiler SDK targets many different platforms
  - Desktop applications
  - Web applications
  - Database Servers
  - Big Data Solutions

- Efficient development, testing, and deployment workflow
  - Single source code base
  - Optimize and test models in MATLAB
  - Simple compile and deployment mechanism to give access to analytic models to non-MATLAB users and processes
MATLAB Data Analytics embedded as Excel Add-ins

- Desktop Excel Add-ins with compiler

- Server based Excel Add-ins with Compiler SDK
Scalable Web based solutions

Client
- CSS
- JS
- AssetAllocStart
- AssetAllocResult
- Asset Allocations Parameters
- Asset Class Statistics
- Select Portfolio
- Portfolio Metrics

Web Server
- Tomcat
- initSession()
- computeFrontier
- Visualize Portfolio()
- computeStats()
- MATLAB PortOpt
- MATLAB PortOpt
- PortOpt AssetAll
- MATLAB Production Server
- Request Broker

App Server
- CTF
- oc
- PortOp AssetAll

Submit
Frontier
Select Portfolio
Portfolio Metrics
Workflow for Embedding MATLAB Data Analytics Components in Production Systems

- Financial Institutions deal with Terabytes/Petabytes of data.
  - Unreasonable to send the data over the wire … another solution/process is required.
  - Embed analytic in data engine

- The development to production process is a two step process, using MATLAB and Deployment products

  - Step 1: Bring Data to MATLAB
    - Multiple data sources
    - Build algorithms/models
    - Test
    - Compile to target platform component

  - Step 2: Bring algorithm to the Data
    - Install MATLAB component into the enterprise production applications.
    - Same functionality/single source
Integration with Databases

- Optimize numerical processing within databases
  - Request MATLAB analytics directly from database servers
  - Trigger requests based upon database transactions

- Minimize data handling using Database Toolbox
Integration with Databases 2016a

- Native in database JSON/Restful call
- ORACLE, MS SQL Server, SAS support this in database
Integration with Databases

Computer memory

• Database Server and MATLAB Production Server can reside in the same memory on physical hardware … occupying different areas of memory

• Efficient use of both database and MPS.. Data transfers via http in memory from one application to the other.

• Compiled MATLAB models on MPS can access the database via Database Toolbox calls in the MATLAB code (ODBC or JDBC calls)
MATLAB Components in Production Databases

MATLAB Production Server can provide predictive analytics in the database

- Oracle (Java, .NET)
- Microsoft SQL Server (.NET)
- Microsoft Access (.NET)
- Netezza (JAVA)
- SAS (JAVA)
- Teradata (JAVA)

- Thin client with MPS
  - Java and .NET supported
- JSON/Restful API
- Central repository for models ... Simplifies change management
MATLAB is the analytics for big data systems such as Hadoop

Embedded on-node custom algorithms can be developed in MATLAB and deployed to the grid

Compiler generated Hadoop components on each node of Hadoop Grid