DEEP LEARNING INference AND VISUAL INSPECTION @ INTEL

Moty Fania
Principle Engineer, Big Data Analytics

London, May 2018
AGENDA

- Advanced Analytics @ Intel
- Corporate AI concepts
- Deep learning Inference system
- Visual inspection
  - Use cases
  - Architecture
HARNESSING ANALYTICS

- Descriptive Analytics
  - What happened?
- Diagnostic Analytics
  - Why did it happen?
- Predictive Analytics
  - What will happen?
- Prescriptive Analytics
  - How can we make it happen?

- Information
- Insight
- Hindsight
- Optimization

- Value
- Difficulty

#StrataData
About Us - Advanced Analytics @ Intel

Vision

AI is a competitive advantage for Intel

Solve Strategic High Value Business Problems

AI Services / Products

Design
Validation
Cut product time to market

Product
Test
Cost, quality, performance

Power &
Performance
ML in Intel Chips

Sales
Unlimited scale

IoT Analytics
Edge-Fog-Cloud platform

Health
Smart Clinical Trials

#StrataData
AI IN A NUTSHELL

Artificial intelligence is about replacing human decision making with more sophisticated technologies.

• These are not repetitive tasks, but rather judgment-based work
• Requires a more complex set of algorithms and machine learning which can use a variety of inputs to recognize patterns, predict future outcomes and make decisions.

Narrow AI- also referred to as “weak AI.” It is the only form of Artificial Intelligence achieved so far. This is AI that works within a very limited context, and can’t take on tasks beyond its field.
CORPORATE AI – “NARROW AI“ USE CASES

Sales

Visualization

Product Development

Design

Visual Inspection

#StrataData
OUR MODE OF WORK FOR AI SOLUTIONS

- Building an excellent ML/DL model(s) that solve the business problem is essential but not sufficient.
- It is much more powerful when deployed to a production inference system that enables close feedback loop and long-lasting business value.

Adaptive Testing  Sales Assists  Internet of thing platform  Predictive Analytics Engine

Reduces TCO and TTM of current implantation and future ones.
HOWEVER, THERE ARE NEW KINDS OF PROBLEMS...
OUR VISUAL INSPECTION CHALLENGE

• Source: Full HD video streams from multiple cameras
• Multiclass classification problem – ~15 classes (multiple DL algorithms required)
• Inspection process will be running 24x7 - detect “issues” at frame level **online**
• Low tolerance for mistakes – very high precision & recall required
• Potential to scale out to hundreds of input cameras

#StrataData
REAL TIME DL INFECTION – WHAT IS REQUIRED?

- Production Inference service for DL models
- Smart in-memory cache for data batching & sequencing
- Fast, scalable APIs for data ingestion & real time responses
- Sync – Async calls
- Full Scalability
**Deep Learning Lifecycle**

**Training**
- Training Data
  - Flower
  - Elephant

**SavedModel CLI**

**Inference**
- TensorFlow Serving
- GRPC
- Clients

*Other names and brands may be claimed as the property of others.*
Tensor Flow Serving

- A flexible, high-performance serving system for machine learning models, designed for production environments
- We have added the following capabilities:
  - Simple APIs to deploy new models
  - Generic client to query any model (sync / Async)
  - Optimized Docker image for CPU & GPU
  - Implementation within Kubernetes with performance optimizations & scaling
  - Full automation of deployment
**BATCH SIZE FOR OPTIMAL INFECTION THROUGHPUTS**

- To maximize the utilization of inference HW and minimize its cost, we have to apply batches vs. single frame inference.
- The goal is achieving the right balance between latency and throughput.
- Batch size is determined while keeping latency under the given latency requirements.

Source: https://software.intel.com/en-us/articles/intel-processors-for-deep-learning-training
Redis for Smart Batching & Sequencing

- Redis is an open source, in-memory data structure store.
- It works with in-memory datasets in order to achieve its outstanding performance.
- Used as a persistence database, cache and message broker.
- It supports data structures such as strings, hashes, lists, sets, sorted sets.
- Other features include:
  - Transactions
  - Pub/Sub
  - Keys with a limited time-to-live
  - Clustering

* Other names and brands may be claimed as the property of others.
**VISUAL INSPECTION INFERENCE SYSTEM**

- **Smart agent**
  - Splitter / strimmer

- **Load Balancer**

- **Inference Client**

- **API 1**

- **ML 1**
  - **ML 2**
  - **ML n**

- **TensorFlow Serving**

- **Flask**

- **redis**

---

#StrataData
**Logic in Client Agent**

- **Splitter / Strimmer Agent**
  - Fetch next Frame
  - **Call API 1** to Post Frame with metadata

- **Result Collector**
  - Sleep for X ms
  - **Call API 2** – to get inference results for next set of processed frames
Inference System (High Level Illustration)

API 1: Post
   → Flask
   ↓ FLASK REST
   → Data Batching
   ↓ Store items
   → Lists of frames
   ↓ Results
   → Business Logic
   ↓ Get Results

API 2: Get Response
   → Business Logic
   ↓ Get Results
   → Results
   ↓ Store Batch Results
   → Flask
   ↓ FLASK REST
   → Async inference
   → Pop next Batch

#StrataData
**SYSTEM FLOW ILLUSTRATION**

**Inference System**

- Web Service (Flask)
- Bulk Process
- Tensorflow Serving

**Agent**
- Send Image
- Get results

**BatchingLayer**

* Other names and brands may be claimed as the property of others.

#StrataData
• The System supports compound results based on an ensemble of ML / DL models
• Interim results are stored into Redis and a combiner process is responsible for applying the final logic
**WHY DOCKER?**

- Portability and ease of deployment anywhere while avoiding the dependency hell

- Docker guarantees that the software will always run the same, regardless of its environment.

- A unit of isolation (decoupled)

- Modularity and Scale Out

*Other names and brands may be claimed as the property of others.*
SCALE OUT WITH KUBERNETES

• An open-source system for automating deployment, scaling, and management of containerized applications.
• Groups containers that make up an application into logical units for easy management and discovery.
• Provides container grouping, load balancing, auto-healing, scaling features
• progressively rolls out changes and updates without rebuilding images

A flexible clustering technology for Container based platforms

* Other names and brands may be claimed as the property of others.
KUBERNETES HIGH LEVEL ARCHITECTURE

- **Kubernetes**
- **etcd**
- **REST API service**
- **TensorFlow Serving Service**

Node 1
- **Bulk Process Pod**
- **redis Pod**
- **Result Handler Pod**

Node 2
- **Bulk Process Pod**
- **redis Pod**
- **Result Handler Pod**

... Node n
- **Bulk Process Pod**
- **redis Pod**
- **Result Handler Pod**

**Heketi**

**GlusterFS**

**Persistent & Scalable filesystem**

#StrataData
USE CASE 2: WAFER IMAGE INSPECTION

- Detecting wafer defects based on optic microscope images.
- Input consists of images of defects and reference images of wafers, as well as tool metadata.
OTHER POTENTIAL APPLICATIONS

Fast search in a video

Real time Text Analytics – Summarizer, classifier

Security and surveillance

Industrial IoT use cases

* Other names and brands may be claimed as the property of others.
INTEGRATION INTO INTEL’S IOT SYSTEM IN FOG

- Rules & ML services
  - Python for Kafka Streams
  - Kafka Streams
- MOSQUITO Bridge
- Gateway
- Kafka Streams micro service
- DL Inference Service
  - API Layer
  - redis
- Kubernetes Serving
  - TensorFlow Serving
- InfluxDB
- Grafana
- Dashboarding & Visualization
- Admin UI
  - API Layer
  - Coachbase

#StrataData
SUMMARY

• AI solutions can augment or replace mundane tasks to relief human bottlenecks in the corporate environment.
• Unlike traditional automation, these solutions have cognitive aspects which “require” human judgment.
• We have implemented a system for production use cases that involve deep learning inference, data streaming and online actuation.
• The system opens the door for various DL use cases in intel and outside.
THANK YOU!