OSG(A)I: BECAUSE A.I. NEEDS A RUNTIME

TIM VERBELEN

Senior Researcher
imec – Ghent University
WHAT IS A.I.?  
THE NEXT BIG THING (AFTER BLOCKCHAIN OFC!)

WHERE SHOULD WE FOCUS THIS YEAR?  "BLOCKCHAIN"

IT WILL CHANGE EVERYTHING.

EVERYBODY IS TALKING ABOUT IT.

THE POTENTIAL APPLICATIONS ARE ENDLESS.

WE DON'T WANT TO BE LEFT BEHIND.

WHAT EXACTLY IS BLOCKCHAIN?

ALSO, "ARTIFICIAL INTELLIGENCE"

© marketoonist.com
WHAT IS MACHINE LEARNING?
LEARNING A FUNCTION APPROXIMATION MAPPING INPUTS TO OUTPUTS
WHAT IS DEEP LEARNING?
USING DEEP NEURAL NETWORKS AS MACHINE LEARNING MODEL
DEEP LEARNING FRAMEWORKS AND TOOLS

- torch
- Keras
- mxnet
- PyTorch
- TensorFlow
- Caffe2
- Caffe
- Chainer
- DL4J
- PaddlePaddle
- Microsoft CNTK
THE AVERAGE DATA SCIENTIST WORKFLOW

1. Inspect and clean up the data
2. Select and encode features / outputs
3. Script together a model training procedure
4. Find some good hyperparameters
5. Dump the trained model
BUT WHAT ABOUT PRODUCTION?

- CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.

- IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?

- IT... MIGHT NOT BE.
  - I GUESS THAT'S SOMETHING—WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.
TENSORFLOW SERVING

# Start TensorFlow Serving container and open the REST API port
$ docker run -t -d -p 8501:8501 \
   -v "$TESTDATA/saved_model_half_plus_two_cpu:/models/half_plus_two" \
   -e MODEL_NAME=half_plus_two \
   -e MODEL_NAME=half_plus_two \
   tensorflow/serving &

# Query the model using the predict API
$ curl -d '{"instances": [1.0, 2.0, 5.0]}"' \
   -X POST http://localhost:8501/v1/models/half_plus_two:predict
CHECK THE BOXES

☐ I want to query my models with an RPC/REST call

☐ I train my models using TensorFlow

☐ I deploy my models on a containerized infrastructure

☐ I don’t need additional metadata/versioning besides a single incrementing integer
OSGI CAN HELP!

- A unit of deployment
  - Package ML model as an OSGi bundle

- A resolveable artifact with requirements and capabilities
  - Requirements: what do you need to run this ML model
  - Capabilities: what kind of inputs can it process and what kind of outputs does it give you?

- A lightweight service model
  - Access your model via an OSGi service

- Service selection at runtime
  - Use service properties and target filters to select the best model at runtime
A WAY TO SHARE ML MODELS
OPEN NEURAL NETWORK EXCHANGE (ONNX)

A common format for describing computation graphs

- defined as protocol buffers
- specifies data types and operators
- framework agnostic
A WAY TO EXECUTE MODELS IN OSGI

DIANNE FRAMEWORK

- A modular deep learning framework developed in OSGi
- Neural networks, operators, datasets, learners, ... as an OSGi service
- Low-level operations via JNI and blas/cublas/cudnn backends
- Distributed deployments using OSGi remote services
- Web UI to build, deploy, train models

http://dianne.intec.ugent.be
DEMO TIME
NOW LET’S ADD A ROBOT
NOW LET’S ADD A ROBOT
PUTTING THINGS TOGETHER

@Reference Camera camera;
@Reference NeuralNetwork nn;
@Reference ArmController controller;

public Promise<Boolean> grasp(){
    return camera.stream()  // returns PushStream<Frame>
      // timeout after 10 seconds
      .timeout(Duration.ofSeconds(10))
      // process camera frame with neural network
      .map(frame -> nn.forward(toTensor(frame))
      // use result to update controller
      .map(r -> controller.update(r))
      // short circuit in case of success
      .anyMatch(s -> s);
}
NOW LET’S ADD A ROBOT
Evaluate the Sessions

Sign in and vote at eclipsecon.org

-1  0  +1
THANK YOU!

http://dianne.intec.ugent.be