DEEP LEARNING TO BIG DATA ANALYTICS ON APACHE SPARK* USING BIGDL

Zhichao (zhichao.li@intel.com)

Big Data Technology, Software and Service Group, Intel

Outline

BigDL

- Apache Spark* + High Performance + Deep Learning

Speech recognition:

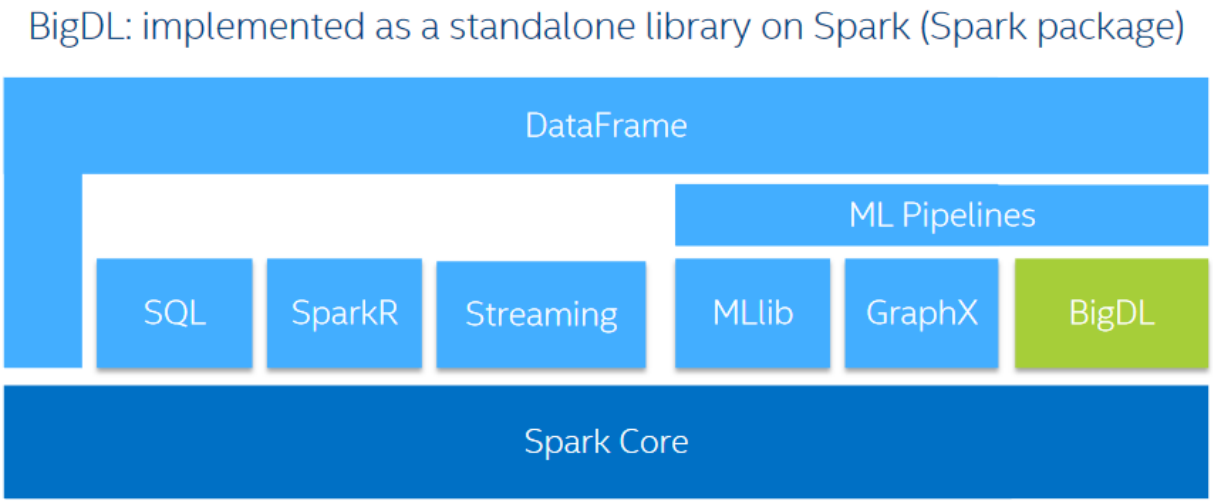
- Deep Speech 2 on BigDL: ML Pipeline + BigDL

Object detection:

- Faster RCNN and SSD on BigDL

What is BigDL?

BigDL is a distributed deep learning library for Apache Spark*



BigDL: Deep learning on Apache Spark*

BigDL open sourced on Dec 30, 2016

https://github.com/intel-analytics/BigDL

- Apache Spark*, MKL Acceleration, High performance

Rich function

- AlexNet, GoogleNet, VGG, Faster R-CNN, SSD, Deep Speech, Recommendation...
- Scala/Java + Python
- AWS EC2, TensorBoard, Notebook, caffe/torch load/export...

Popularity

- Support from Cloud: Microsoft, Amazon, Cloudera, Databricks...
- Community. 1700+ stars

Basic Component

Tensor:

- ND-array data structure
- Generic data type
- Rich and fast math operations (powered by Intel MKL)

Layers

- 113+ layers (Conv, 3D Conv, Pooling, 3D Pooling, FC ...)

Criterion

- 23+ criterions (DiceCoefficient, ClassNLL, CrossEntropy ...)

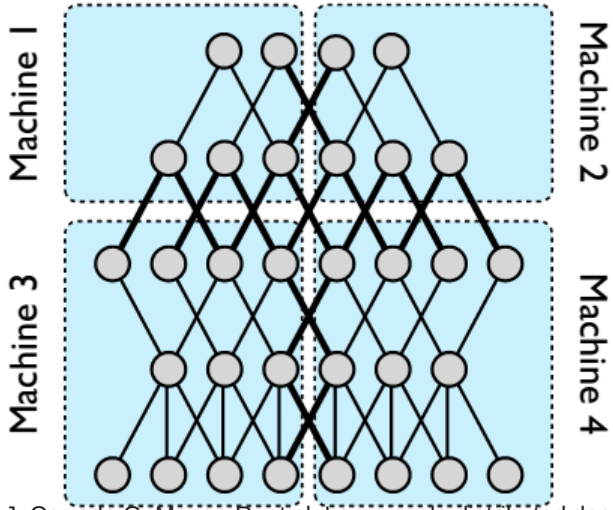
Optimization

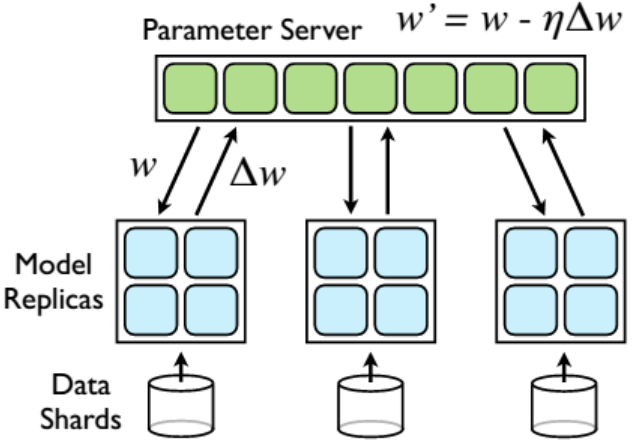
- SGD, Adagrad, LBFGS
- **Community contribution**: Adam, Adadelta, RMSprop, Adamx

Pattern

Model Parallelism

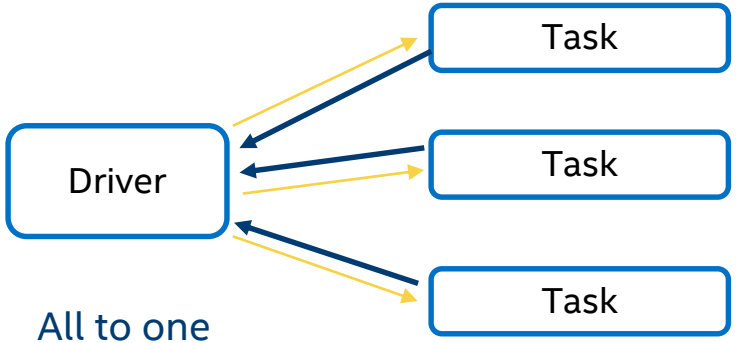
Data Parallelism



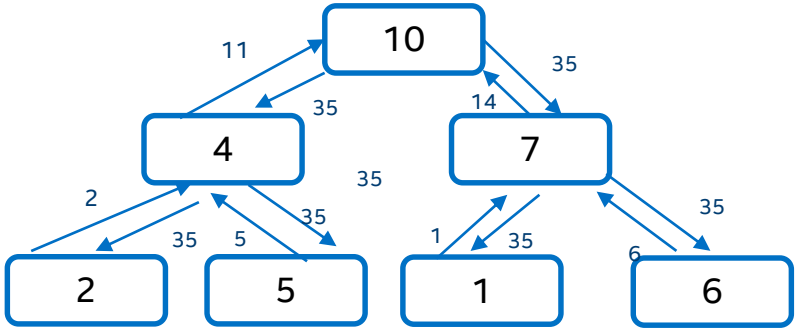


Source: Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks[C]//Advances in neural information processing systems. 2012: 1223-1231.

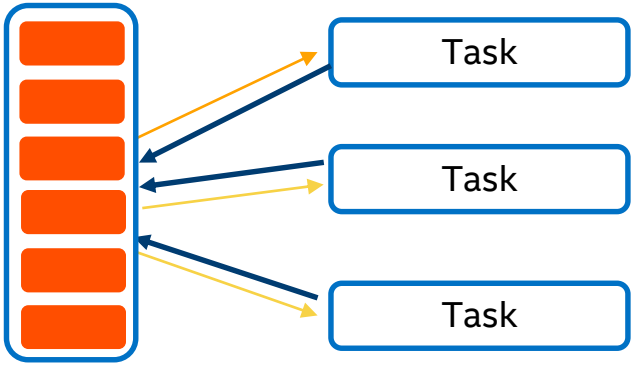
Communication Model



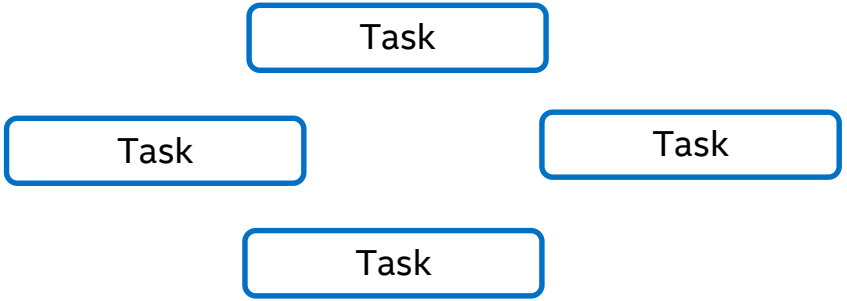
All to one



All reduce (tree aggregation)

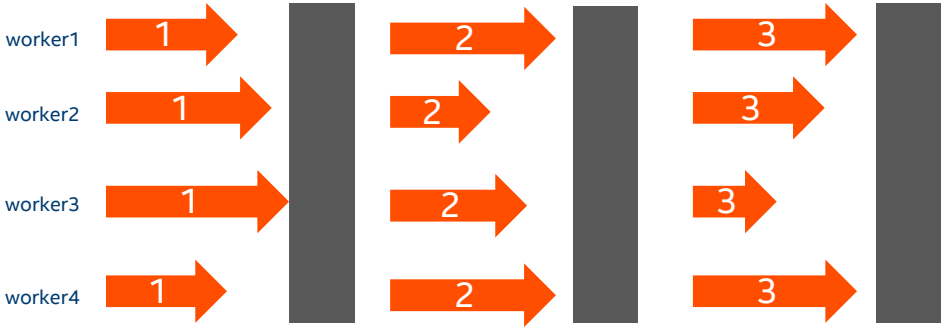


Parameter Server

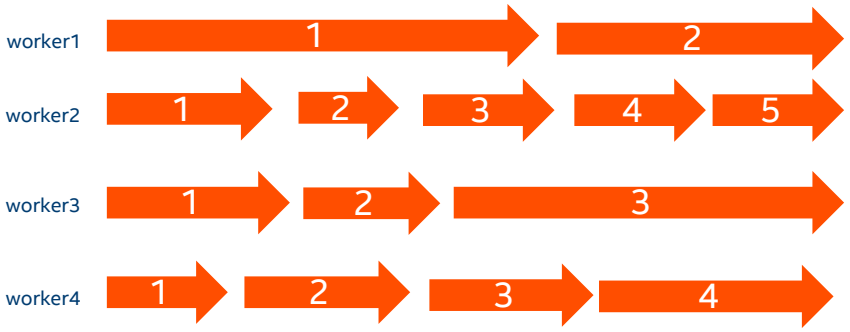


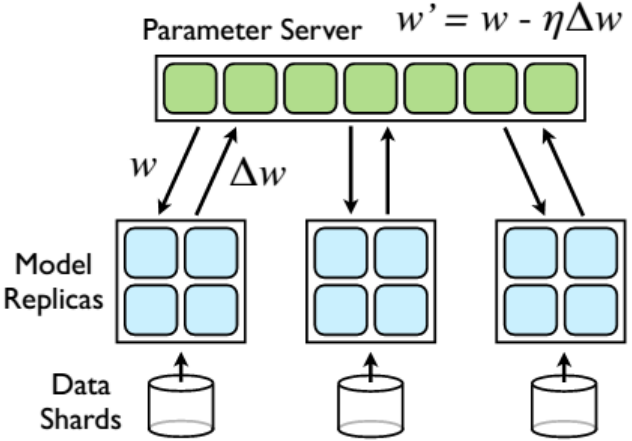
All reduce

Bulk Synchronous Parallel (BSP)



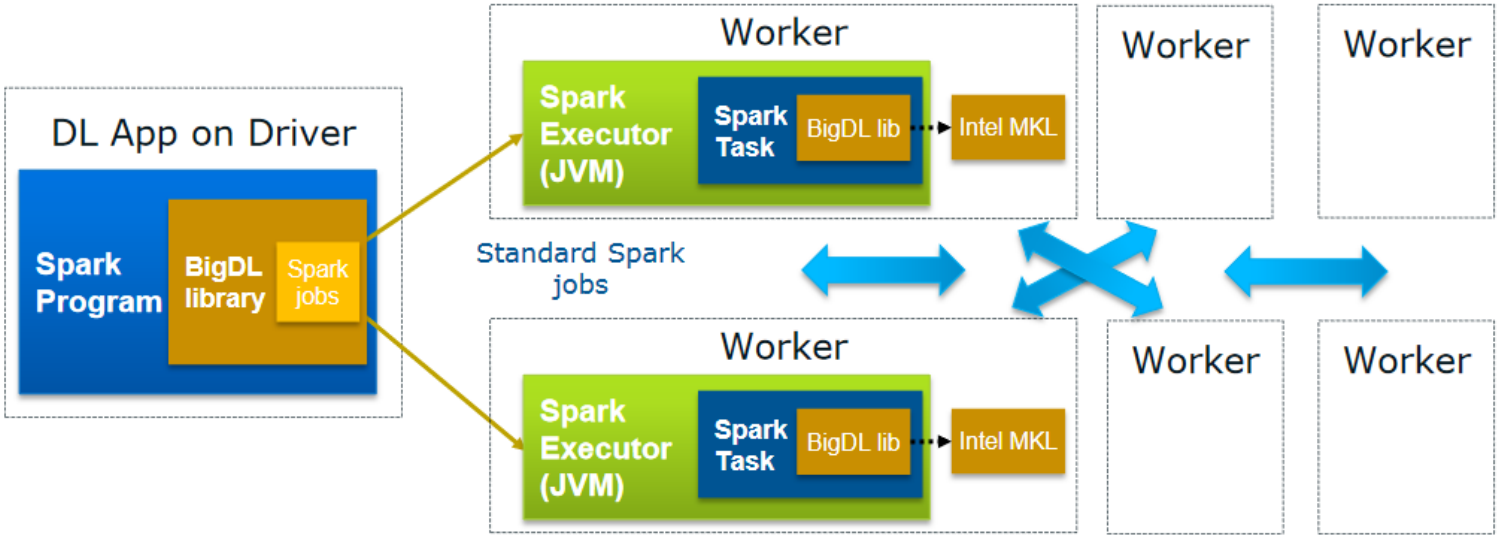
Asynchronous Synchronous Parallel (ASP)





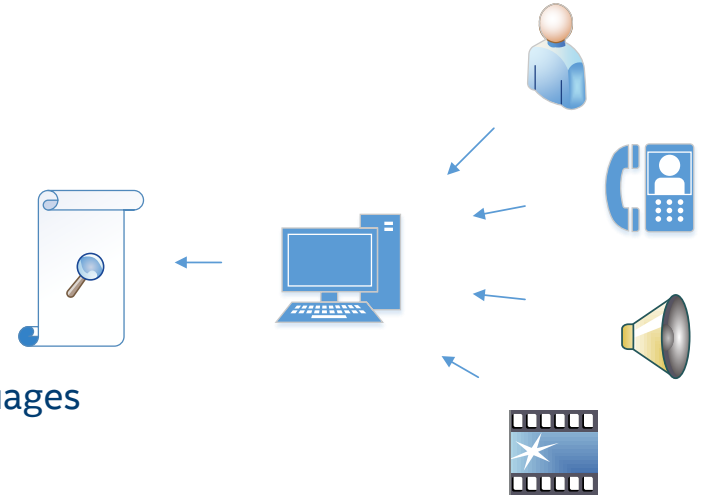
Source: Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks[C]//Advances in neural information processing systems. 2012: 1223-1231.

Run as standard Apache Spark* jobs



DEEP SPEECH 2 WITH BIGDL

Speech Recognition



Challenges

- Audio → text
- Speaker variability, Channel variability, Different languages

Solutions:

- Hybrid system:
  - DNNs, Hidden Markov Models (HMMs), context-dependent phone models, Lexicon models, GMM.
  - Domain expertise and multi-stage
- ✓ DNN end to end:
  - DNN. Much easier
  - More data, better model

Deep Speech 2 for Speech Recognition

- "The Deep Speech 2 ASR pipeline approaches or **exceeds the accuracy of Amazon Mechanical Turk human workers** on several benchmarks, works in **multiple languages** with little modification, and is deployable in a production setting."
- "Table 13 shows that the DS2 system **outperforms humans in 3 out of the 4 test sets and is competitive on the fourth**. Given this result, we suspect that there is little room for a generic speech system to further improve on clean read speech without further domain adaptation."

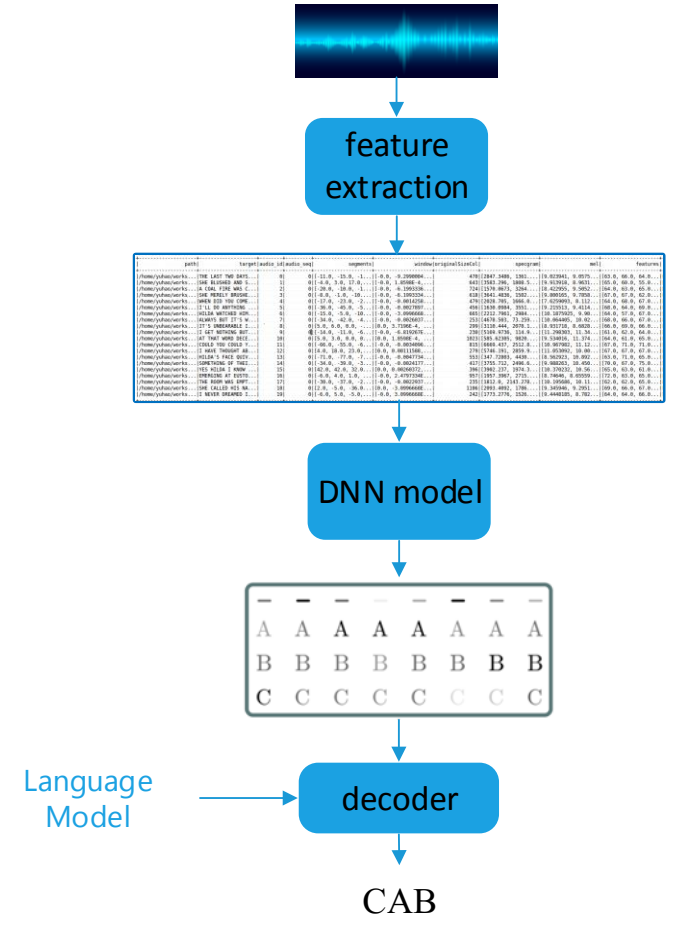
Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Table 13: Comparison of WER for two speech systems and human level performance on read speech.

Deep Speech 2 on BigDL

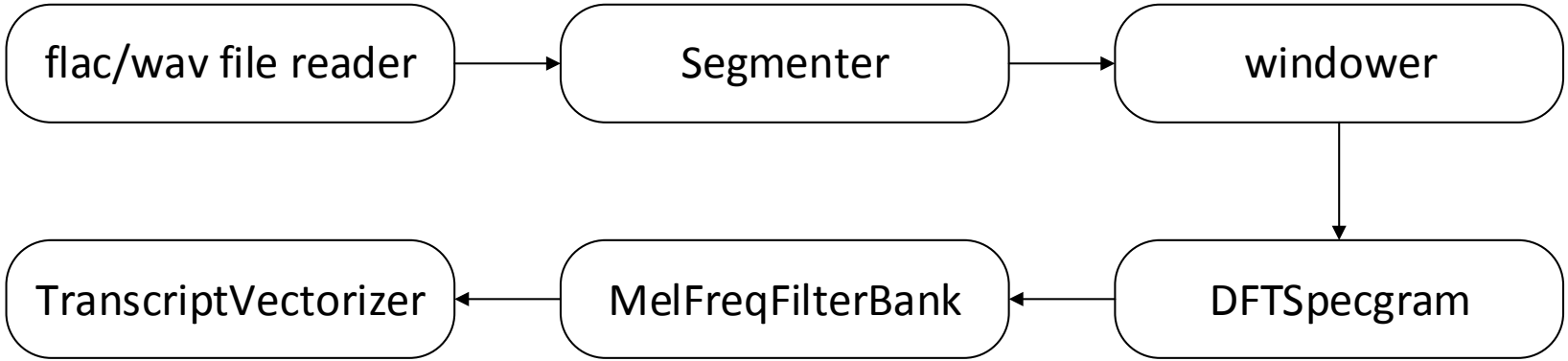
Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*
Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu



Deep Speech 2 on BigDL: Feature transformers

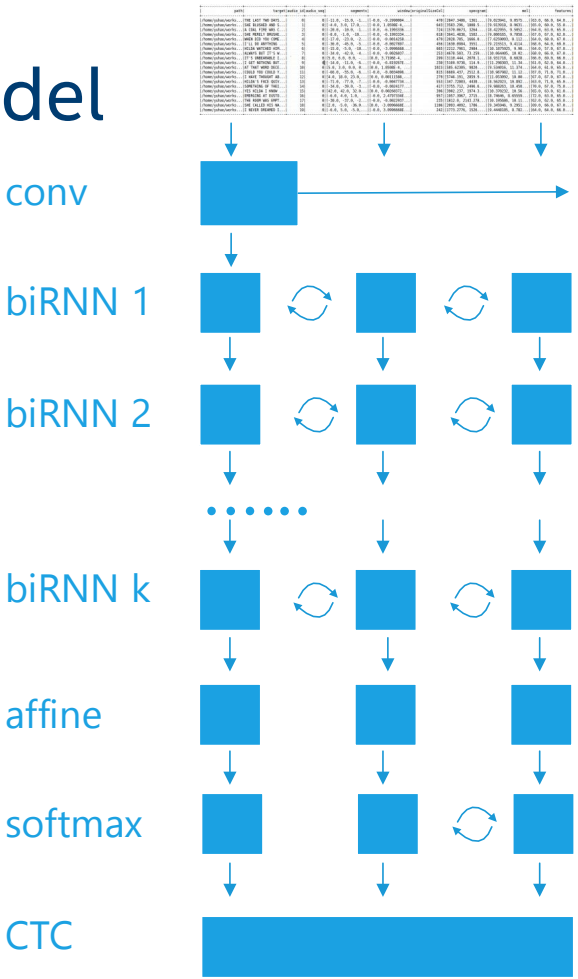
Apache Spark* ML Pipeline



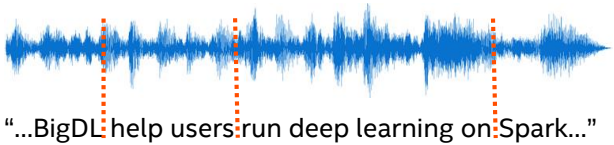
Deep Speech 2 on BigDL: Model

```
val model = Sequential[T]()
  .add(conv)
  .add(ReLU[T]())
  .add(Squeeze(4))
  .add(brnn)
  .add(linear1)
  .add(HardTanhDS[T](0, 20, true))
  .add(linear2)
```

9 layers biRNN: >50 Million parameters



Deep Speech 2 on BigDL: CTC Loss



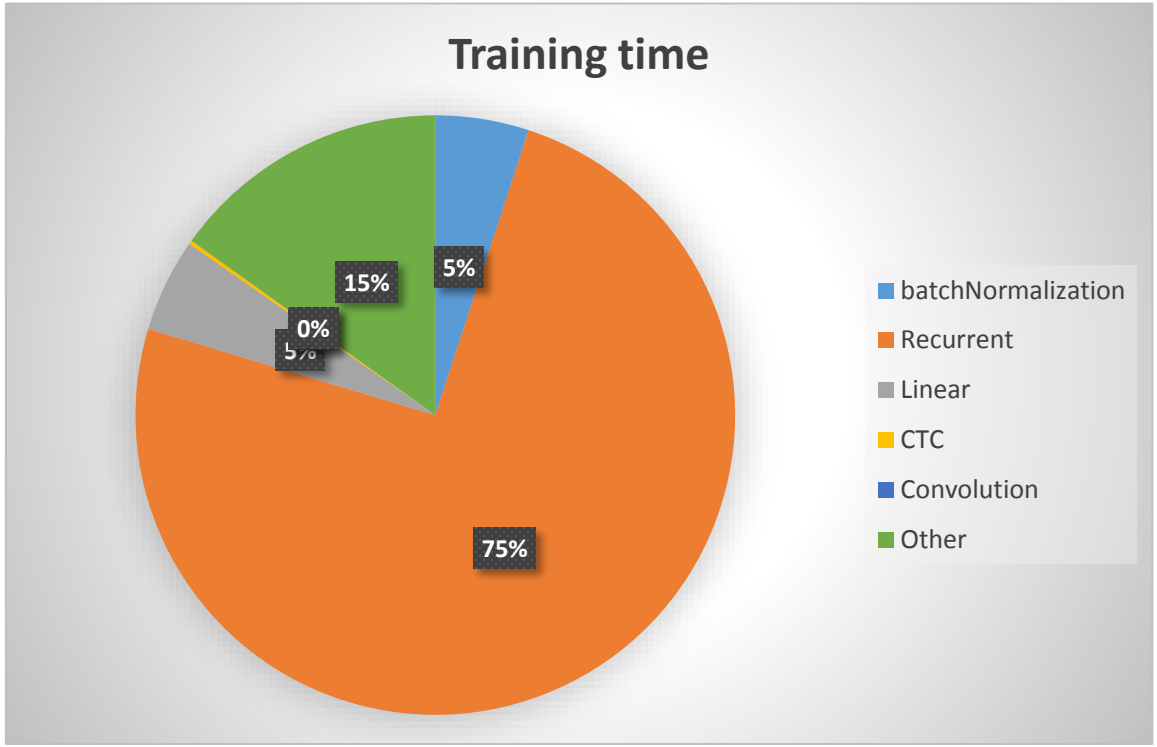
Connectionist Temporal Classification

- a loss function useful for performing supervised learning on sequence data, without needing an alignment between input data and labels. (Alex Graves etc. 2006)
- Raw waveforms and text transcription

BigDL developed first open source CTC on Java/Scala

- Loss/Gradient in consistency with baidu/warp-ctc
- JNI version about 3X faster than Scala version, but CTC only takes 0.2% of the training time.

Deep Speech 2 on BigDL: Model training



With libriSpeech, 5 RNN layer, 30 seconds uttLength, 30 epoches.

Deep Speech 2 on BigDL: Decoder

Existing decoder:

- BestPathDecoder (argmax) wer = 27%
- VocabDecoder wer = 22%

	t1	t2	t3	t4	...
A	0.01	0.1	0.7	0.03	
B	0.05	0.01	0.2	0.5	
C	0.01	0.1	0.09	0.01	
D	0.8	0.05	0.01	0.01	
...	...	...	...	...	
Blank	0.1	0.6	0.01	0.4	
	D	-	A	B	

Contribution welcome

- decoder with Language Model, expect wer < 10%

Deep Speech 2 with AN4 data

- Deep Speech 2 (8 layer, 5 RNN), uttLength 8 seconds
  - Word Error Rate with hold-out validation dataset

	wer(without LM)
Deep Speech on Tensorflow	12.4%
BigDL	< 5%

Deep Speech 2 with LibriSpeech

- Deep Speech 2 (12 layers, 9 RNN), uttLength 30 seconds
  - Word Error Rate with hold-out validation dataset

	cer	wer(without LM)
Hannun, et al. (2014)	10.7	35.8
Graves-Jaitly (ICML 2014)	9.2	30.1
Hwang-Sung (ICML 2016)	10.6	38.4
BigDL	8.7	32.4

- Still under further tuning and optimization.
  - More training data
  - Optimizer (Adam, SGD, nesterov)

Deep Speech 2 on BigDL: Summary

Feature transformers:

- Flac/wav Reader, Windower, TimeSegmenter, TranscriptVectorizer, DFTSpecgram, MelFrequencyFilterBank

Model training and inference

- Big DL container, optimizer, Convolution, BatchNormalization, Bi-RNN

CTC (Connectionist Temporal Classification) loss

- Scala or JNI (warp-ctc)

Decoder

- ArgmaxDecoder, VocabDecoder

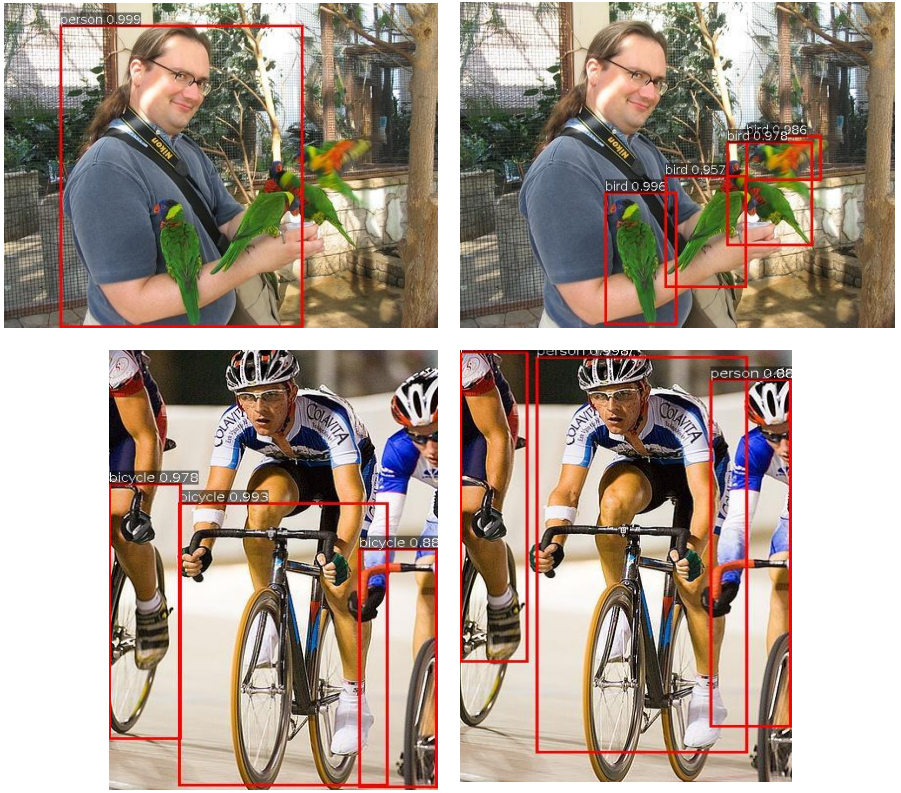
Evaluation

- wer, cer

OBJECT DETECTION WITH BIGDL

SSD: Single Shot Multibox Detector

- State-of-the-art object detection pipeline
- Single shot

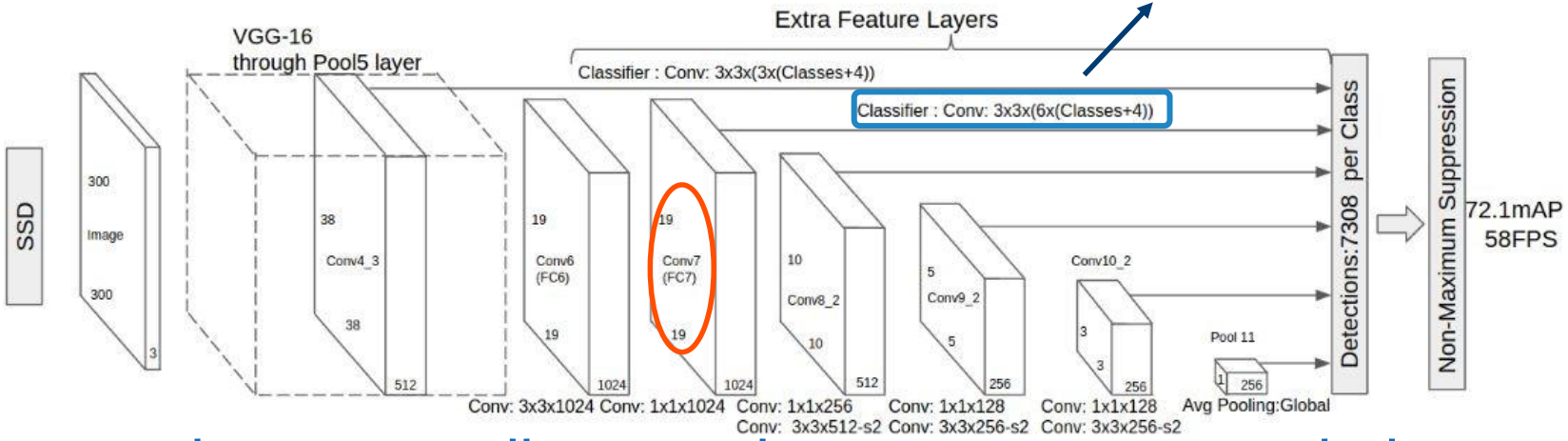


Images from PASCAL(http://host.robots.ox.ac.uk/pascal/VOC/)

Liu, Wei, et al. "SSD: Single shot multibox detector." European Conference on Computer Vision. Springer International Publishing, 2016.

The Single Shot Detector (SSD)

Convolutional predictors for detection



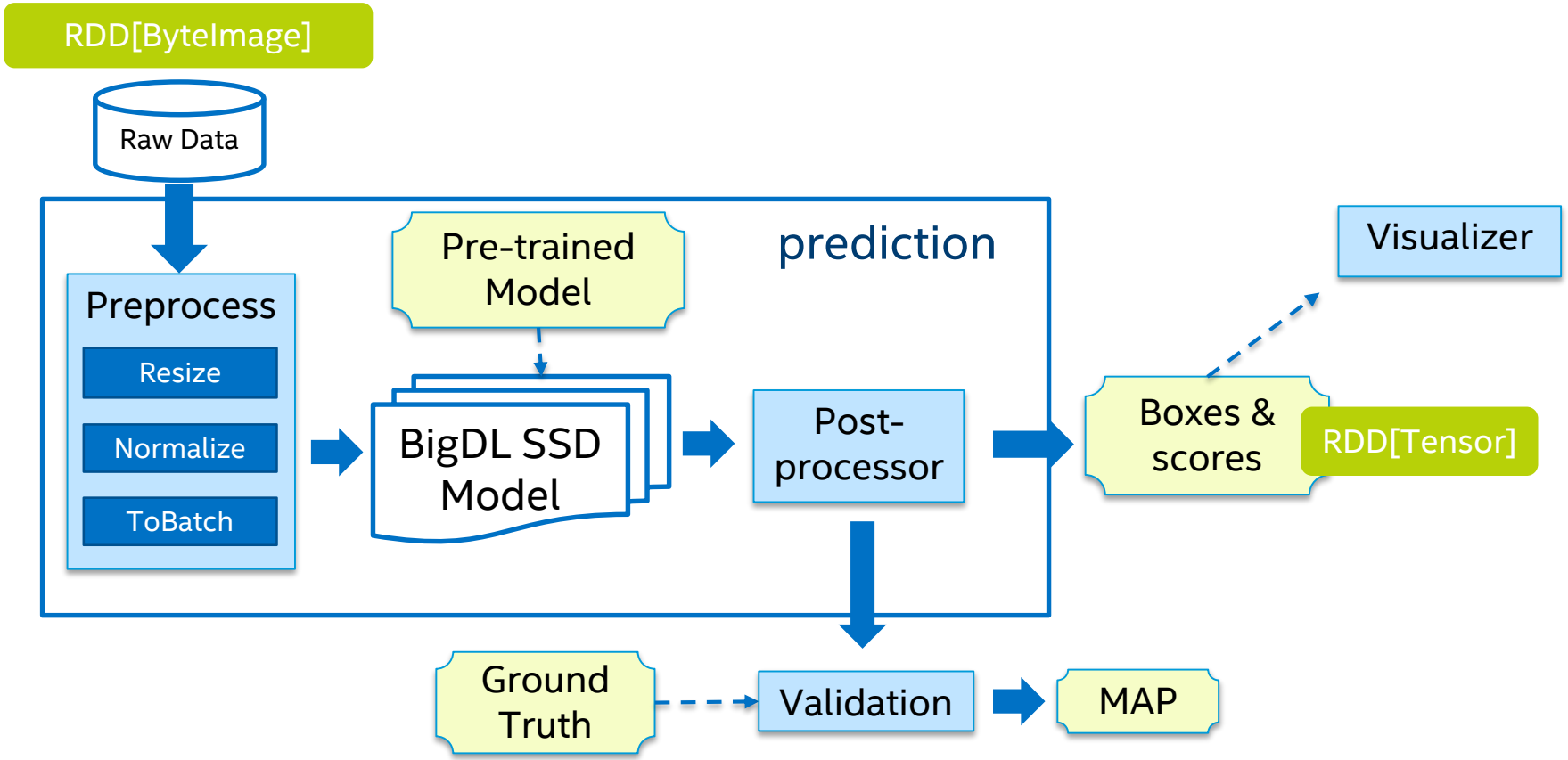
base network

fc6 and fc7 converted to convolutional

collection of bounding boxes and scores for each category

Multi-scale feature maps for detection: observe how conv feature maps decrease in size and allow predictions at multiple scales

SSD Pipeline



SSD + VGG 300x300 test over Pascal VOC 2007

- SSD + VGG 300x300 with pretrained model over voc07+12
  - Mean Average Precision

	Caffe Model	BigDL
MAP	77.2	77.3

SSD + VGG 512x512 test over Pascal VOC 2007

- SSD + VGG 512x512 with pretrained model over voc07+12
  - Mean Average Precision

	Caffe Model	BigDL
MAP	79.6	79.6

BigDL Community

analytics-zoo

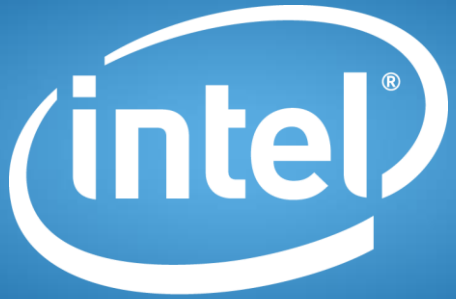
https://github.com/intel-analytics/analytics-zoo

Join Our Mail List

bigdl-user-group+subscribe@googlegroups.com

Report Bugs And Create Feature Request

https://github.com/intel-analytics/BigDL/issues



Software

Legal Disclaimer

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the first quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company’s expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; customer acceptance of Intel’s and competitors’ products; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel’s response to such actions; and Intel’s ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel’s results could be affected by the timing of closing of acquisitions and divestitures. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel’s ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel’s results is included in Intel’s SEC filings, including the company’s most recent reports on Form 10-Q, Form 10-K and earnings release.