CNTK: Microsoft's Open-Source Deep-Learning Toolkit

Frank Seide
Microsoft Research
One Microsoft Way
Redmond, WA 98052
fseide@microsoft.com

Amit Agarwal
Microsoft Research
One Microsoft Way
Redmond, WA 98052
amitaga@microsoft.com

ABSTRACT

This tutorial will introduce the Computational Network Toolkit, or CNTK, Microsoft's cutting-edge open-source deep-learning toolkit for Windows and Linux. CNTK is a powerful computation-graph based deep-learning toolkit for training and evaluating deep neural networks. Microsoft product groups use CNTK, for example to create the Cortana speech models and web ranking. CNTK supports feed-forward, convolutional, and recurrent networks for speech, image, and text workloads, also in combination. Popular network types are supported either natively (convolution) or can be described as a CNTK configuration (LSTM, sequence-to-sequence). CNTK scales to multiple GPU servers and is designed around efficiency.

The tutorial will give an overview of CNTK's general architecture and describe the specific methods and algorithms used for automatic differentiation, recurrent-loop inference and execution, memory sharing, on-the-fly randomization of large corpora, and multi-server parallelization. We will then show how typical uses looks like for relevant tasks like image recognition, sequence-to-sequence modeling, and speech recognition.

OUTLINE

The tutorial will cover these topics:

- What is CNTK?
  - Computational network introduction
- How does a typical use of CNTK look like?
  - Defining the Computational Network
  - Configuring data I/O
  - SGD hyper-parameters
  - Typical workflows
- Deep Dive into specific technologies
  - Implicit handling of time
  - Minibatching of variable-length sequences
  - Data-parallel training
- CNTK Library APIs
  - Network, Reader, Learner, SGD API
  - C++ usage
  - Python usage
- Hands-on examples, including
  - ResNet image recognition
  - Sequence-to-sequence modeling

KDD '16, August 13-17, 2016, San Francisco, CA, USA.
ACM 978-1-4503-4232-2/16/08.
http://dx.doi.org/10.1145/2939672.2945397

SPEAKER BIOGRAPHIES

Frank Seide, a native of Hamburg, Germany, is a Senior Researcher at Microsoft Research. His current research focus is on deep neural networks for conversational speech recognition; together with co-author Dong Yu, he was first to show the effectiveness of deep neural networks for recognition of conversational speech. Throughout his career, he has been interested in and worked on a broad range of topics and components of automatic speech recognition, including spoken-dialogue systems, recognition of Mandarin Chinese, and, particularly, large-vocabulary recognition of conversational speech with application to audio indexing, transcription, and speech-to-speech translation. His current focus is Microsoft's CNTK deep-learning toolkit.

Amit Agarwal is a Principal Software Engineer at Microsoft's Technology and Research division. His current focus is on building CNTK, Microsoft's large scale distributed deep learning platform, to enable unprecedented scale, speed and capacity for training massive deep learning models on enormous datasets, used in a wide gamut of speech, image and text related deep learning tasks at Microsoft and in the community. Amit Agarwal worked on a wide range of Microsoft products and at Mentor graphics. He holds 7 patents related to heterogeneous and GPU programming.

REFERENCES

[1] Amit Agarwal, Eldar Akchurin, Chris Basoglu, Guoguo Chen, Scott Cyphers, Jasha Droppo, Adam Eversole, Brian Guenter, Mark Hillebrand, T. Ryan Hoens, Xuedong Huang, Zhiheng Huang, Vladimir Ivanov, Alexey Kamenev, Philipp Kranen, Oleksii Kuchaiev, Wolfgang Manousek, Avner May, Bhaskar Mitra, Olivier Nano, Gaizka Navarro, Alexey Orlov, Hari Parthasarathi, Baolin Peng, Marko Radmilac, Alexey Reznichenko, Frank Seide, Michael L. Seltzer, Malcolm Slaney, Andreas Stolcke, Huaming Wang, Yongqiang Wang, Kaisheng Yao, Dong Yu, Yu Zhang, Geoffrey Zweig (in alphabetical order), "An Introduction to Computational Networks and the Computational Network Toolkit", Microsoft Technical Report MSR-TR-2014-112, 2014..

[2] "CNTK," https://github.com/Microsoft/CNTK