CustoNN: Customizing Neural Networks on FPGAs

High-Performance IT Systems group

Dr. Tobias Kenter
Prof. Dr. Christian Plessl

6 February 2017

Paderborn Center for Parallel Computing

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Neural Network Success Stories

POST MAGAZINE

Why Baidu's breakthrough on speech recognition may be a game changer

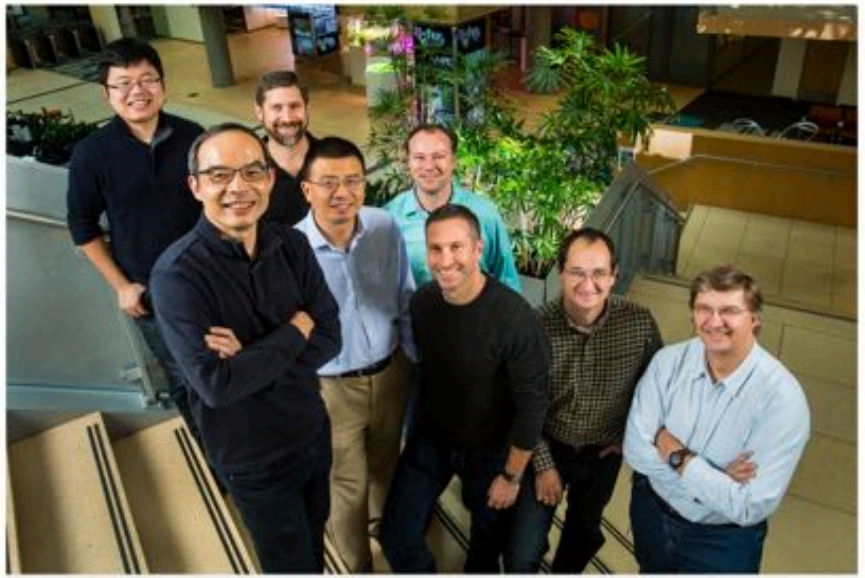
Deep Speech 2, a speech recognition network developed by China's answer to Google, is so stunningly accurate it can transcribe Chinese better than a person, writes Will Knight

BY **MIT TECHNOLOGY REVIEW**
19 MAR 2016

Microsoft Store Products Support

Next The Official Microsoft Blog The Fire Hose Microsoft On the Issues Transform

Historic Achievement: Microsoft researchers reach human parity in conversational speech recognition



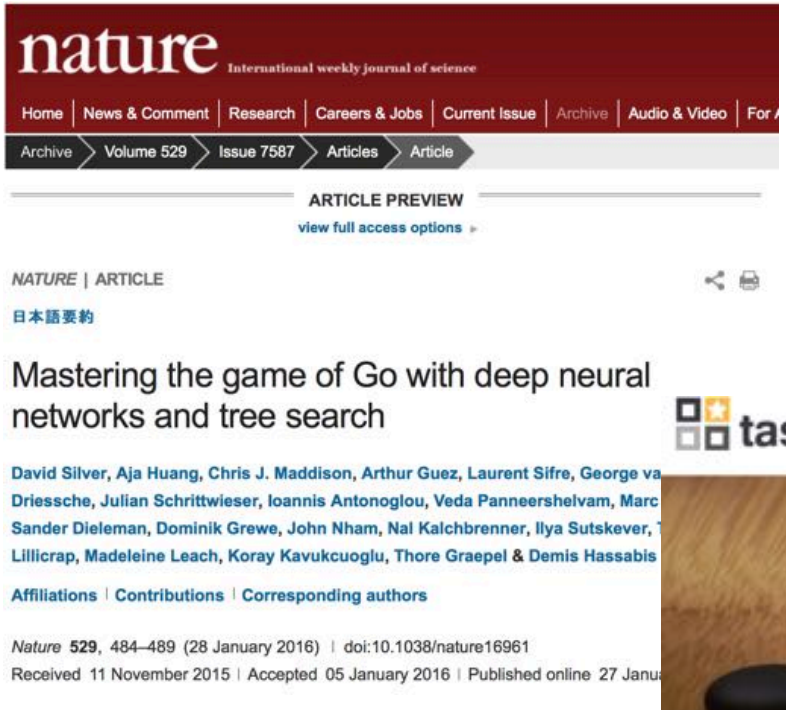
Microsoft researchers from the Speech & Dialogue research group include, from back left, Wayne Xiong, Geoffrey Zweig, Xuedong Huang, Dong Yu, Frank Seide, Mike Seltzer, Jasha Droppo and Andreas Stolcke. (Photo by Dan DeLong)

Posted October 18, 2016

By **Allison Linn**

Microsoft has made a major breakthrough in speech recognition, creating a technology that recognizes the words in a conversation as well as a person does.

Neural Network Success Stories



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

Archive > Volume 529 > Issue 7587 > Articles > Article

ARTICLE PREVIEW

view full access options

NATURE | ARTICLE

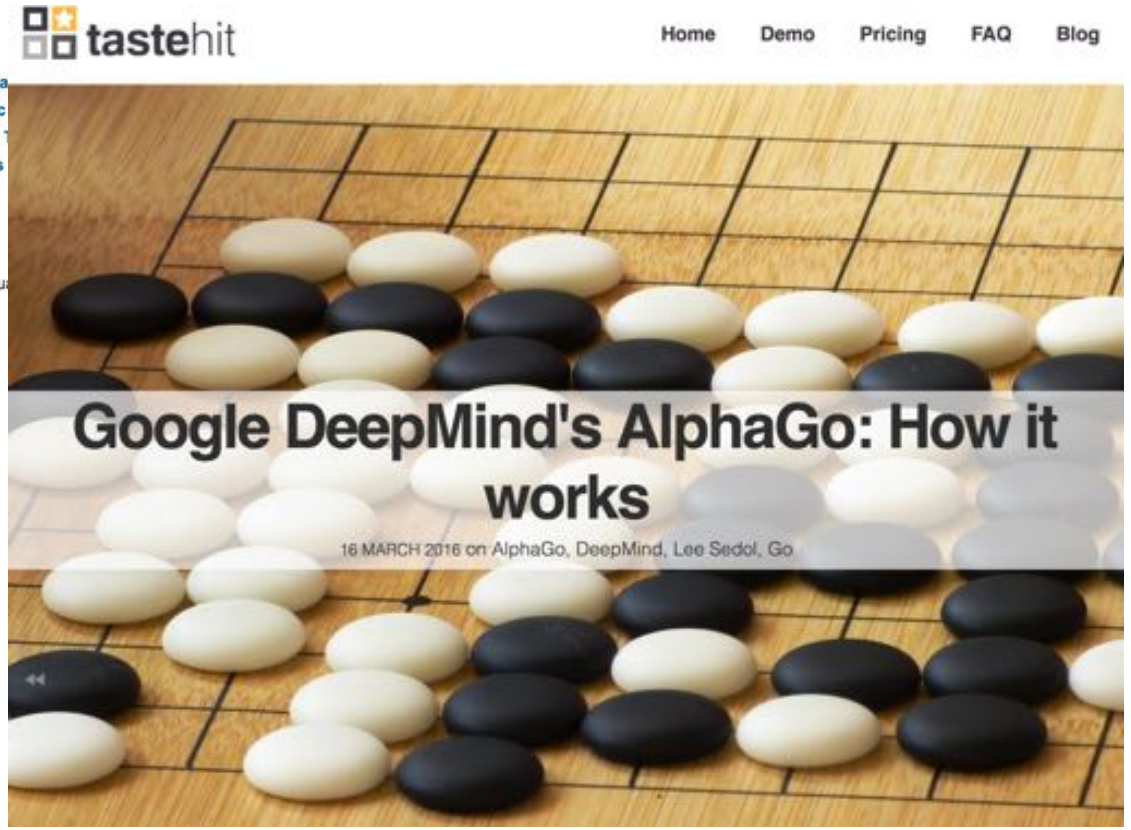
日本語要約

Mastering the game of Go with deep neural networks and tree search

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George va
Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc
Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever,
Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis

Affiliations | Contributions | Corresponding authors

Nature **529**, 484–489 (28 January 2016) | doi:10.1038/nature16961
Received 11 November 2015 | Accepted 05 January 2016 | Published online 27 Janu



tastehit

Home Demo Pricing FAQ Blog

Google DeepMind's AlphaGo: How it works

16 MARCH 2016 on AlphaGo, DeepMind, Lee Sedol, Go

Neural Network Success Stories

MIT Technology Review

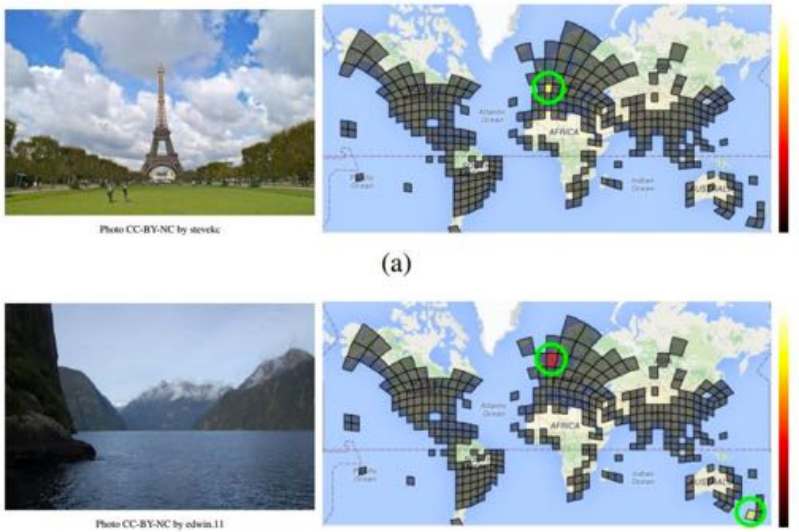
Topics+ Top Stories

Computing

Google Unveils Neural Network with "Superhuman" Ability to Determine the Location of Almost Any Image

Guessing the location of a randomly chosen Street View image is hard, even for well-traveled humans. But Googl latest artificial-intelligence machine manages it with relati ease.

by Emerging Technology from the arXiv February 24, 2016



(a)

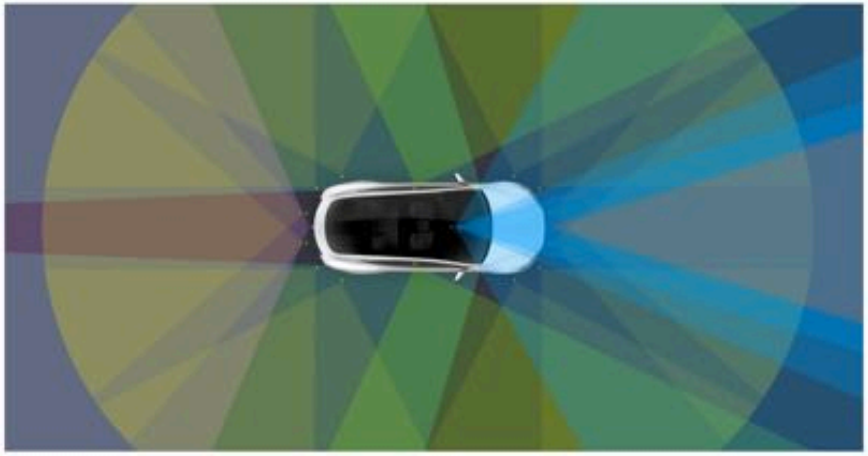
Neural Network Success Stories

TESLA MODEL S MODEL X MODEL 3 ENERGY CHARGING

Blog Videos Press

All Tesla Cars Being Produced Now Have Full Self-Driving Hardware

The Tesla Team • October 19, 2016



Self-driving vehicles will play a crucial role in improving transportation safety and accelerating the world's transition to a sustainable future. Full autonomy will enable a Tesla to be substantially safer than a human driver, lower the financial cost of transportation for those who own a car and provide low-cost on-demand mobility for those who do not.

NVIDIA.

DRIVERS PRODUCTS DEEP LEARNING AND AI COMMUNITIES SUPPORT SHOP ABOUT NVIDIA

HOME DEEP LEARNING VIRTUAL REALITY DRIVING PRO GRAPHICS GAMING

Posted on OCTOBER 20, 2016 by DANNY SHAPIRO

Tesla Motors' Self-Driving Car "Supercomputer" Powered by NVIDIA DRIVE PX 2 Technology

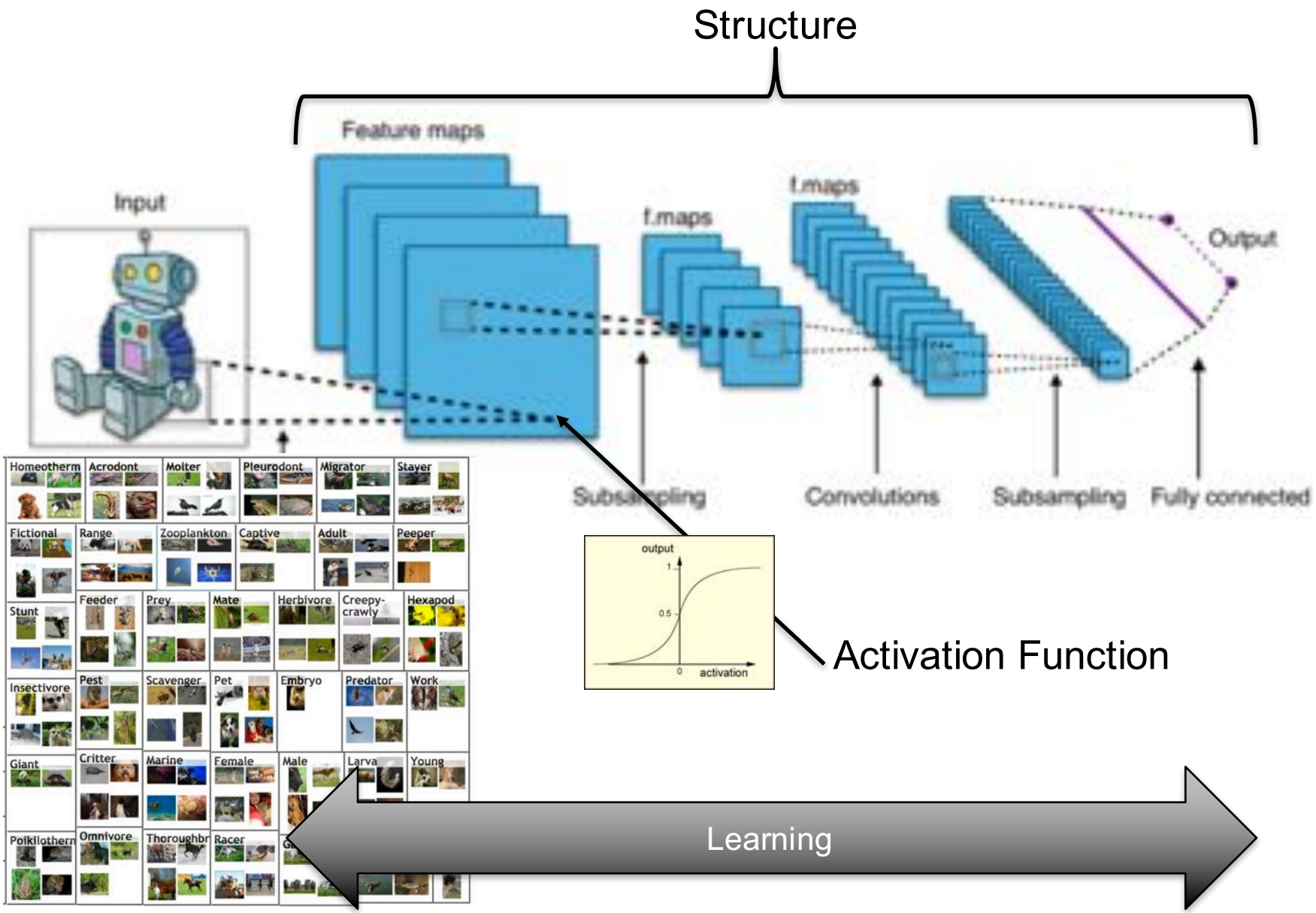
Tesla Motors has announced that all Tesla vehicles — Model S, Model X, and the upcoming Model 3 — will now be equipped with an on-board "supercomputer" that can provide full self-driving capability.

The computer delivers more than 40 times the processing power of the previous system. It runs a Tesla-developed neural net for vision, sonar, and radar processing.

This in-vehicle supercomputer is powered by the NVIDIA DRIVE PX 2 AI computing platform.

NVIDIA DRIVE PX 2 is an end-to-end AI computing system that uses groundbreaking approaches in deep learning to perceive and understand the car's surroundings.

Designing (DC)NNs

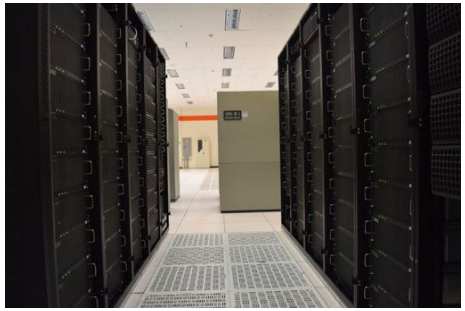


[Sources: ImageNet Database (http://www.image-net.org/), Wikimedia Commons]

NN's Hardware Demands

- AlexNet [2012]
  - 650,000 neurons
  - 60 million parameters (249 MB)
  - 1.5 billion floating point operations to classify one image
  - Training: 5-6 days on 2 GTX 580 GPUs
- AlphaGo
  - Training: > 4 weeks on 50 GPUs

➢ Massive GPU clusters



[Krizhevsky, A., Sutskever, I., Hinton, G.E.: **Imagenet classification with deep convolutional neural networks**. In: *Advances in neural information processing systems*, pp. 1097–1105 (2012)]

- Configurable Hardware
  - Customize operations, connections, data reuse
  - Can't compete with GPUs on raw floating-point performance, but…

Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to $+1$ or -1

Matthieu Courbariaux*[1] MATTHIEU.COURBARIAUX@GMAIL.COM
Itay Hubara*[2] ITAYHUBARA@GMAIL.COM
Daniel Soudry[3] DANIEL.SOUDRY@GMAIL.COM
Ran El-Yaniv[2] RANI@CS.TECHNION.AC.IL
Yoshua Bengio[1,4] YOSHUA.UMONTREAL@GMAIL.COM
[1]Université de Montréal
[2]Technion - Israel Institute of Technology
[3]Columbia University
[4]CIFAR Senior Fellow
*Indicates equal contribution. Ordering determined by coin flip.

Abstract

We introduce a method to train Binarized Neural Networks (BNNs) - neural networks with binary weights and activations at run-time. At training-time the binary weights and activations are used for computing the parameters gradients. During the forward pass, BNNs drastically

tistical machine translation (Devlin et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), Atari and Go games (Mnih et al., 2015; Silver et al., 2016), and even abstract art (Mordvintsev et al., 2015).

Today, DNNs are almost exclusively trained on one or many very fast and power-hungry Graphic Processing Units (GPUs) (Coates et al., 2013). As a result, it is of-

XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks

Mohammad Rastegari[1(✉)], Vicente Ordonez[1], Joseph Redmon[2], and Ali Farhadi[1,2]

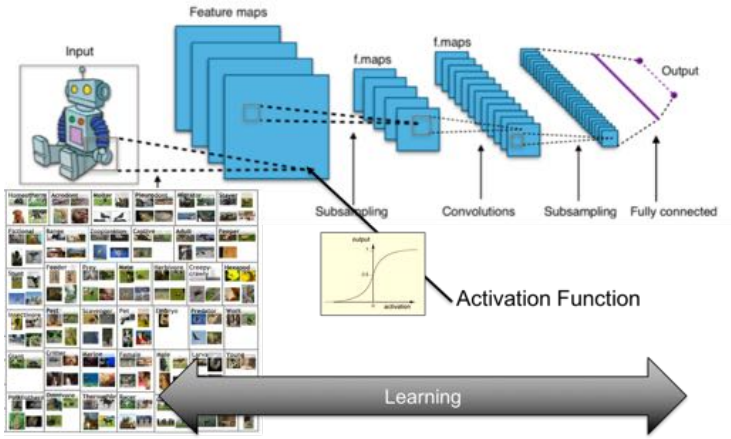
[1] Allen Institute for AI, Seattle, USA
{mohammadr,vicenteor}@allenai.org
[2] University of Washington, Seattle, USA
{pjreddie,ali}@cs.washington.edu

Abstract. We propose two efficient approximations to standard convolutional neural networks: Binary-Weight-Networks and XNOR-Networks. In Binary-Weight-Networks, the filters are approximated with binary values resulting in 32× memory saving. In XNOR-Networks, both the filters and the input to convolutional layers are binary. XNOR-

  - For small fixed-point or binary operations, FPGAs are great

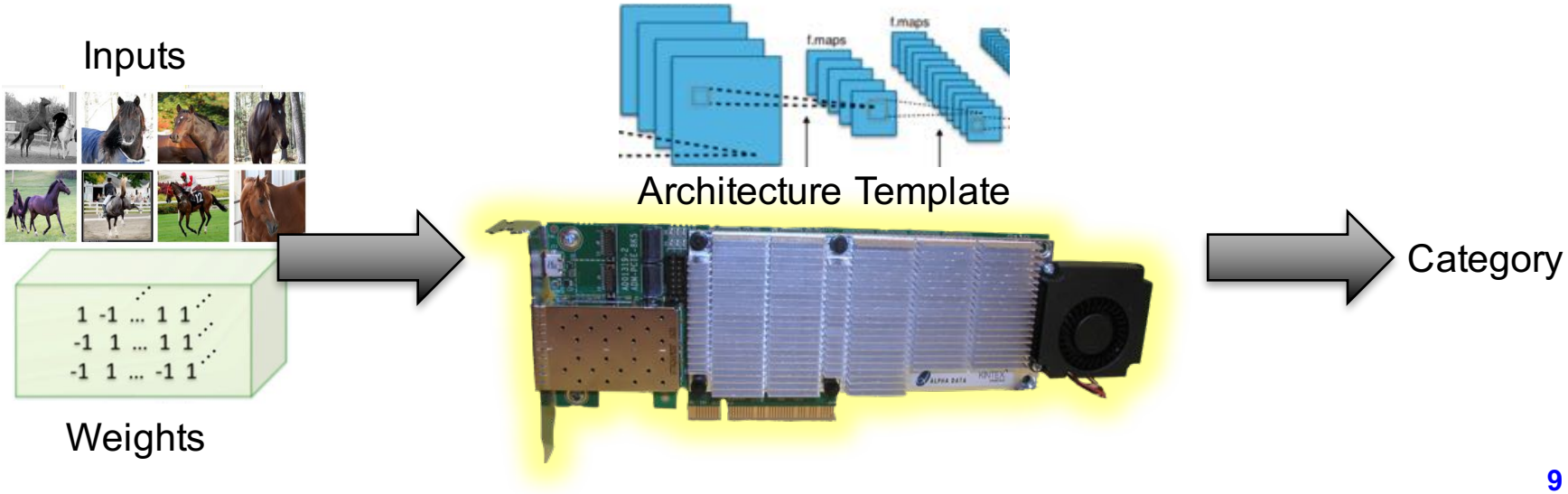
PG CustoNN (1): Neural Networks on FPGAs

- Research current approaches to fixed-point / binary NNs



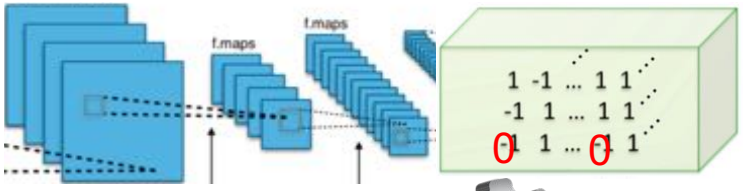
Weights ∈ {0, 1}
Inputs ∈ {0, 1}
Operations ∈ {XNOR, bitcount}

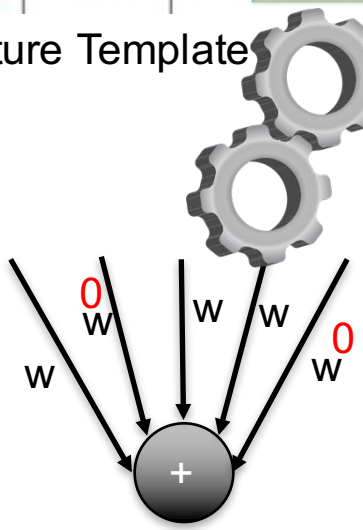
- Implement efficient inference architecture on FPGA



CustoNN (2): Fully Custom Neural Networks

- Inference uses the same NN with the same weights over and over again…



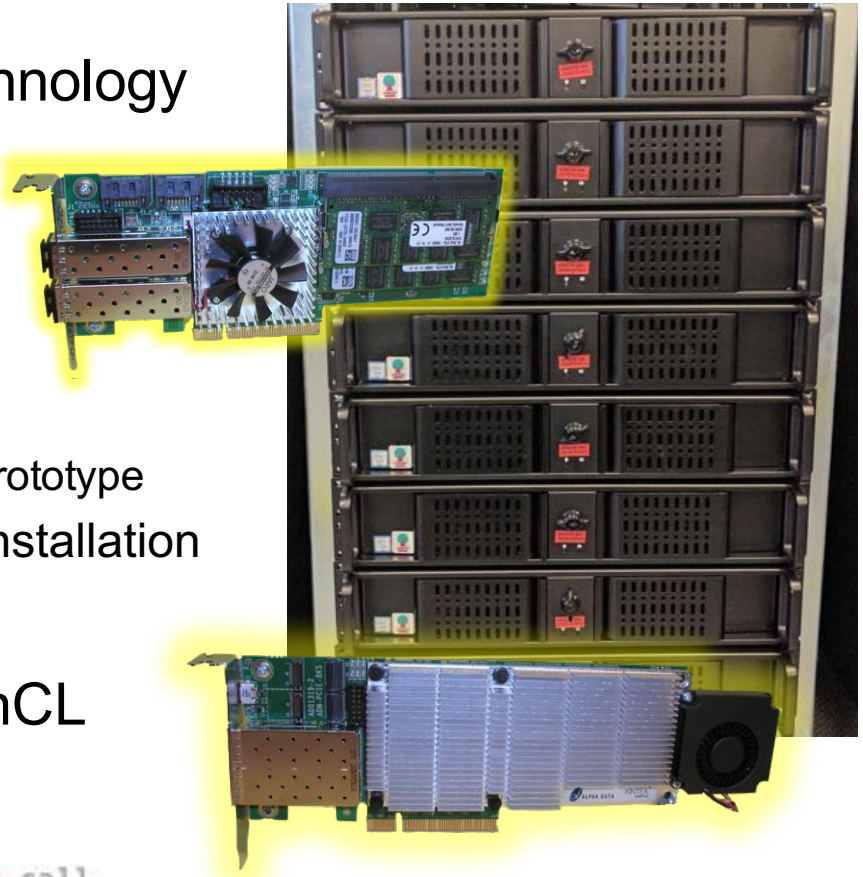


Inputs

Category

Our Infrastructure

- 2 Clusters with latest FPGA technology
  - xcl-cluster
    - 8 nodes
    - Each with 2 different Xilinx FPGAs
  - harp-cluster
    - 10 nodes
    - 2nd generation Intel Xeon+FPGA prototype
      - worldwide first academic installation



- Programming FPGAs with OpenCL
  - It finally works... on both platforms

```
22  // AOC kernel demonstrating device-side printf call
23
24  __kernel void hello_world(int thread_id_from_which_to_print_message) {
25    // Get index of the work item
26    unsigned thread_id = get_global_id(0);
27
28    if(thread_id == thread_id_from_which_to_print_message) {
29      printf("Thread #%u: Hello from Altera's OpenCL Compiler!\n", thread_id);
30    }
31  }
32
```

CustoNN - Summary

- **Project Group for CS and CE students**

- Goals
  - ➢ Fixed-point / binary inference architecture on FPGA
  - ➢ Fully custom neural network on FPGA

- Fields of interest
  - Neural networks / deep learning
  - OpenCL or other accelerator languages
  - Accelerator architectures

- Supervisors
  - Christian Plessl, christian.plessl (at) uni-paderborn.de
  - Tobias Kenter, kenter (at) uni-paderborn.de, ☎ 05251/60-4340