IMAGE RECOGNITION: KEY FOR FUTURE APPLICATIONS

Assemblée Nationale
Obélisque de Louxor
= Rue Royale
Near rue Saint-Honoré

Bus turning
Truck
Car
Car
**ImageNet: Classification**

- **Give the name of the dominant object in the image**
- **Top-5 error rates: if correct class is not in top 5, count as error**
  - Black: ConvNet, Purple: no ConvNet

<table>
<thead>
<tr>
<th>2012 Teams</th>
<th>%error</th>
<th>2013 Teams</th>
<th>%error</th>
<th>2014 Teams</th>
<th>%error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervision (Toronto)</td>
<td>15.3</td>
<td>Clarifai (NYU spinoff)</td>
<td>11.7</td>
<td>GoogLeNet</td>
<td>6.6</td>
</tr>
<tr>
<td>ISI (Tokyo)</td>
<td>26.1</td>
<td>NUS (singapore)</td>
<td>12.9</td>
<td>VGG (Oxford)</td>
<td>7.3</td>
</tr>
<tr>
<td>VGG (Oxford)</td>
<td>26.9</td>
<td>Zeiler-Fergus (NYU)</td>
<td>13.5</td>
<td>MSRA</td>
<td>8.0</td>
</tr>
<tr>
<td>XRCE/INRIA</td>
<td>27.0</td>
<td>A. Howard</td>
<td>13.5</td>
<td>A. Howard</td>
<td>8.1</td>
</tr>
<tr>
<td>UvA (Amsterdam)</td>
<td>29.6</td>
<td>OverFeat (NYU)</td>
<td>14.1</td>
<td>DeeperVision</td>
<td>9.5</td>
</tr>
<tr>
<td>INRIA/LEAR</td>
<td>33.4</td>
<td>UvA (Amsterdam)</td>
<td>14.2</td>
<td>NUS-BST</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adobe</td>
<td>15.2</td>
<td>TTIC-ECP</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG (Oxford)</td>
<td>15.2</td>
<td>XYZ</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG (Oxford)</td>
<td>23.0</td>
<td>UvA</td>
<td>12.1</td>
</tr>
<tr>
<td>Team/algorithm</td>
<td>Date</td>
<td>Test error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervision</td>
<td>2012</td>
<td>15.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarifai</td>
<td>2013</td>
<td>11.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>2014</td>
<td>6.66%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microsoft</td>
<td>05/02/2015</td>
<td>4.94%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>02/03/2015</td>
<td>4.82%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baidu/Deep Image</td>
<td>10/05/2015</td>
<td>4.58%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shenzhen Institutes of Advanced Technology</td>
<td>10/12/2015</td>
<td>(the CNN has 152 layers)</td>
<td>3.57%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

COMPETITION ON IMAGENET: SINCE 2012, CONVOLUTIONAL NEURAL NETWORKS (CNN) ARE LEADING!

From NVIDIA

![Graph showing the performance of Deep Learning and Hand-coded CV from 2010 to 2015](image)
Deep Learning is Everywhere
(ConvNets are Everywhere)

- Lots of applications at Facebook, Google, Microsoft, Baidu, Twitter, IBM...
  - Image recognition for photo collection search
  - Image/Video Content filtering: spam, nudity, violence.
  - Search, Newsfeed ranking

People upload 800 million photos on Facebook every day
- (2 billion photos per day if we count Instagram, Messenger and Whatsapp)

Each photo on Facebook goes through two ConvNets within 2 seconds
- One for image recognition/tagging
- One for face recognition (not activated in Europe).

Soon ConvNets will really be everywhere:
- self-driving cars, medical imaging, augmented reality, mobile devices, smart cameras, robots, toys.....
DEEP LEARNING AND NEUROMORPHIC SYSTEMS AT LETI AND LIST
Exploitation of Deep Neural Networks
- Image recognition, annotation and indexing

Tools for fast and accurate Neural Network (NN) exploration & Architecture benchmarking: *N2D2*
- Neural Network exploration (including with spike coding and new materials)
N2D2: PLATFORM FOR DEVELOPING DEEP NEURAL NETWORK APPLICATIONS

- **N2D2** is a platform to design and generate deep neural network (DNN) and to select the computing platform which fit best application needs.
- Fast benchmarking of Components Off the Shelf and exports to dedicated ASIC:
  - Parallel processors (OpenCL, OpenMP)
  - GPU (OpenCL, Cuda, CuDNN)
  - FPGA (RTL, HLS)
  - Leti & List specific processors (like **P-Neuro**)
FAST AND ACCURATE NN EXPLORATION

Automated architecture mapping and benchmarking tool flow

1) **Deep network builder**

2) **Learning a database**

3) **Analysis of network performances**

4) **CPU, GPU and FPGA-based real-time implementation**

**N2D2 software framework**

**OpenMP**

**OpenCL**

**HLS FPGA**

**Wide targets range, perfs and power metrics**
EXAMPLE OF INDUSTRIAL APPLICATION of N2D2: ROLLING MILL

CONTRAINTS
- Real time with very high throughput (20m/s)
- Tiny defect (~mm) with low contrast
- Complex environment (oil vapor, few space for inspection..)

SOLUTION
- Database labelling and Processing
- Fast NN topology Exploration
- Performance vs complexity analysis

Real time performance achievable on FPGA (direct code generation)

1) Defects labeling and visualization
2) NN Exploration and benchmarking
3) Defects identifications after NN learning
Exploitation of Deep neural Networks
• Image recognition, annotation and indexing

Tools for fast and accurate Neural Network (NN) exploration & Architecture benchmarking: N2D2
• Neural Network exploration (including with spike coding and new materials)

Diversity of implementations:
• Software solution / GPU
• Reconfigurable devices / FPGA
• Dedicated implementations
  • Full CMOS and binary coding: P-NEURO
  • Full CMOS and “spike coding”
  • Using new materials
N2D2 and P-Neuro: complete solution for Deep Learning in smart nodes

- Fast benchmarking of Components Off The Shelf:
  - Parallel processors
  - GPU
  - FPGA (HLS)

- Performance of **P-Neuro** neural network processing unit
  - Example on Faces extraction,
    - Database of 18000 images
  - Comparison of 5 different architectures
  - Focus on energy efficiency
  - Expected performance of **P-Neuro**:
    - FDSOI 28nm, 1GHz
    - 1.8 TOPs/W, <0.5 mm² (4 cores)
    - Fully scalable from 1 to 1024 cores
    - Ready for integration in smart nodes

<table>
<thead>
<tr>
<th>Target</th>
<th>Frequency</th>
<th>Energy efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quad ARM A7</td>
<td>900 MHz</td>
<td>380 images/W</td>
</tr>
<tr>
<td>Quad ARM A15</td>
<td>2000 MHz</td>
<td>350 images/W</td>
</tr>
<tr>
<td>Tegra K1</td>
<td>850 MHz</td>
<td>600 images/W</td>
</tr>
<tr>
<td>Intel I7</td>
<td>3400 MHz</td>
<td>160 images/W</td>
</tr>
<tr>
<td>P-Neuro (FPGA)</td>
<td>100 MHz</td>
<td>2000 images/W</td>
</tr>
<tr>
<td>P-Neuro (ASIC)</td>
<td>500 MHz</td>
<td>125 000 images/W</td>
</tr>
</tbody>
</table>
SPIKE-BASED CODING

29x29 pixels
841 addresses

Correct
Output
THE PROMISES OF SPIKE-CODING NN

- Reduced computing complexity and natural temporal and spatial parallelism
- Simple and efficient performance tunability capabilities
- Spiking NN best exploit NVMs such as RRAM, for massively parallel synaptic memory

<table>
<thead>
<tr>
<th>Formal neurons</th>
<th>Spiking neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base operation</td>
<td>- Multiply-</td>
</tr>
<tr>
<td></td>
<td>Accumulate (MAC)</td>
</tr>
<tr>
<td>Activation function</td>
<td>- Non-linear function</td>
</tr>
<tr>
<td>Parallelism</td>
<td>- Spatial multiplexing</td>
</tr>
</tbody>
</table>

Two test chips implemented in 65nm
- Reptile: 3 tiles of 12 neurons
- Spider: 25 tiles of 12 neurons

Advanced technology nodes
- Comparison of Analog and Digital neurons
- Gain of Analog neuron (less area) reduces
  → Curves cross at 22nm node
Exploitation of Deep neural Networks
• Image recognition, annotation and indexing

Tools for fast and accurate Neural Network (NN) exploration & Architecture benchmarking: \textit{N2D2}
• Neural Network exploration (including with spike coding and new materials)

Diversity of implementations:
• Software solution / GPU
• Reconfigurable devices / FPGA
• Dedicated implementations
  • Full CMOS and binary coding: \textit{P-NEURO}
  • Full CMOS and “spike coding”
  • Using new materials

Take full advantage of advanced devices to break the density and power issues:
• 3D integration, CoolCube\textsuperscript{TM}.
• RRAM, PCM and new devices,
Neural Networks
- Naturally 3D for 2D inputs, layers optimally distributed in stacked dies
- Vertical connections between layers: minimizes interconnect length, avoid routing congestion

NEMESIS 3D two-layers SNN test chip
- 1st layer: 48 macro-block neurons, 1024 synapses per neuron (49,152 total)
- 2nd layer: 50 fully connected neurons, 2,400 synapses

3D offers 2x better total area and 25% better power efficiency vs 2D

<table>
<thead>
<tr>
<th>Two-layers SNN circuit</th>
<th>2D</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total area (mm²)</td>
<td>7.97</td>
<td>3.63 (-54%)</td>
</tr>
<tr>
<td>Power (mW)</td>
<td>428</td>
<td>354 (-17%)</td>
</tr>
<tr>
<td>Critical path (ns)</td>
<td>9.00</td>
<td>6.63 (-26%)</td>
</tr>
</tbody>
</table>

[B. Belhad, R. Héliot, P. Vivet, CASSES’2014]
LEARNING FROM NEUROSCIENCE: A STDP (SPIKE TIMING DEPENDENT PLASTICITY) PRIMER

STDP = correlation detector
⇒ Possible learning model of the brain?

Causality Potentiation (LTP)

Anti-Causality Depression (LTD)

Δt = t_{post} - t_{pre}
NEW ELEMENT: RRAM AS SYNAPSES

**Thermal effect**

**PCM**

- GST
- GeTe
- GST + HfO₂

**Electrochemical effect**

**CBRAM**

- Ag / GeS₂

**Electronic effect**

**OXRAM**

- TiN/HfO₂/Ti/TiN

---

M. Suri et al., IEDM 2011
M. Suri et al., IMW 2012, JAP 2012
O. Bichler et al., IEEE TED 2012
M. Suri et al., EPCOS 2013
D. Garbin et al., IEEE Nano 2013

D. Garbin et al., IEDM 2014
D. Garbin et al., IEEE TED 2015
PRINCIPLE CROSSBARS OF MEMRISTORS

First Proposed by Snider(1)

\[ V_{\text{pre}} \quad V_{\text{post}} \]

- \( t_{\text{pre}} < t_{\text{post}} \)
- \( t_{\text{pre}} > t_{\text{post}} \)

\[ V_{\text{pre}} \quad V_{\text{post}} \]

- \( V_{\text{pre}} < V_{\text{post}} \)
- \( V_{\text{pre}} > V_{\text{post}} \)

\[ V_{\text{pre}} \quad -V_{\text{pre}} \]

- \( R \) decreases
- \( R \) increases

Neurons

Pre-synaptic spike

Post-synaptic spike (feedback)

Synaptic weight update through STDP

BIO-INSPIRED MODELS EXPLORATION

Learning rule

Network topology

Synaptic

N2-D2 Neuromorphic simulator

Input stimuli

Neuron model

Example: Leaky Integrate & Fire (LIF) neuron

\[ u = u_e \frac{t_{\text{spike}} - t_{\text{last_spike}}}{t_{\text{leak}}} + w \]

Synaptic model

Neuron membrane potential

Complete tool flow for bio-inspired synapses, neurons and learning rules network simulations

[O. Bichler et al., NanoArch‘2014]
NVM SYNAPSES IMPLEMENTATIONS

- 2-PCM synapses for unsupervised cars trajectories extraction
  ![2-PCM synapse diagram](image1)
  [O. Bichler et al., Electron Devices, IEEE Transactions on, 2012]

- CBRAM binary synapses for unsupervised MNIST handwritten digits classification with stochastic learning
  ![CBRAM synapse diagram](image2)
  [M. Suri et al., IEDM, 2012]
EXAMPLE OF ON-GOING INVESTIGATIONS: VRRAM FOR NEUROMORPHIC APPLICATIONS

- Investigation of VRRAM based on CBRAM stack
  - 2 levels (proof of concept)
  - 16 levels (goal)
  - 1 select transistor per level (proof of concept)
  - Integrated selector (goal)
  - CBRAM most suitable R for neuromorphic
  - OxRAM also analysed

- Design: support development for VRRAM
  - **High Density**: Estimate the maximum size of a VRRAM-based array supposing to have an integrated selector [E. Cha, ISCAS 2014]
  - **Neuromorphic**: propose a circuit dimensioning for the neuromorphic approach presented at IEDM 2015 (1TnR pillar ~ Synapse, NO Selector)
AN EU COLLABORATIVE PROJECT: NEURAM3

Objective:
- Fabricate a chip implementing a neuromorphic architecture that supports state-of-the-art machine learning algorithms and spike-based learning mechanisms.

Features:
- 28nm FDSOI technology with RRAM synapses
- Ultra low power scalable and reconfigurable architecture
- 50x lower dissipation than digital equivalent
- TFT based scalable multichip architecture platform
- A technology to implement on-chip learning, using native adaptive characteristics of electronic synaptic elements
## A NEW EU COLLABORATIVE PROJECT: NEURAM3

![NeuRAM3 Logo](image)

<table>
<thead>
<tr>
<th>Participant no.</th>
<th>Organization name</th>
<th>Short name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Coordinator)</td>
<td>Commissariat a l’energie atomique et aux energies alternatives</td>
<td>CEA</td>
<td>France</td>
</tr>
<tr>
<td>2</td>
<td>Interuniversitair Micro-Electronica Centrum IMEC VZW</td>
<td>IMEC</td>
<td>Belgium</td>
</tr>
<tr>
<td>3</td>
<td>Stichting IMEC Nederland</td>
<td>IMEC-NL</td>
<td>Netherlands</td>
</tr>
<tr>
<td>4</td>
<td>IBM Research Gmbh</td>
<td>IBM</td>
<td>Switzerland</td>
</tr>
<tr>
<td>5</td>
<td>University of Zurich, Institute of Neuroinformatics</td>
<td>UZH</td>
<td>Switzerland</td>
</tr>
<tr>
<td>6</td>
<td>Agencia Estatal Consejo Superior de Investigaciones Cientificas, Instituto de Microelectronica de Sevilla</td>
<td>CSIC</td>
<td>Spain</td>
</tr>
<tr>
<td>7</td>
<td>Consiglio Nazionale delle Ricerche</td>
<td>CNR</td>
<td>Italy</td>
</tr>
<tr>
<td>8</td>
<td>Jacobs University Bremen</td>
<td>JAC</td>
<td>Germany</td>
</tr>
<tr>
<td>9</td>
<td>ST-Microelectronics S.A.</td>
<td>STM</td>
<td>France</td>
</tr>
</tbody>
</table>
Summary of key points

- Large-scale database GPU-accelerated learning for CNN
- Among the leading teams on ImageClef2015 contest
- From scratch exploration to industrial applications

- Lead in bio-inspired STDP learning (IEDM’11, 12, 14)
- Formalized spike-coding for CNN, complete tool flow for co-simulation

- Complete framework with C, OpenCL, CUDA and HLS exports
- Complete tool flow for spike-coding DSP

- 2-PCMs synapse (patented) scheme (IEDM’15)
- Lead in SNN with RRAM devices (IEDM’14)

- Competitive reconfigurable architecture with P-Neuro
- Spike-coding DSP architecture
- Increased efficiency with 3D
Thank you for your attention