

Computational Noise Resiliency in Deep Learning Architectures

*Prepared for:
2015 Neuro-Inspired Computational Elements Workshop*

*Sek Chai, PhD.
SRI International
sek.chai@sri.com*

Feb 24, 2015



In Reference to Yesterday's Talks

- Xaq Pitkow, Rice Univ

Statistical view, noise correlations between $X, Y | Z$

- Randal O'Reilly, Univ Colorado Boulder

Error-driven learning, autoencoders

- This presentation

Hardware design can also be noisy.

Outline

- Applications: Addressing *noise* in input data
- Algorithms: Leverage *noise* in training
- Hardware: Leverage *noise* in design

SRI International

SRI is a world-leader at creating new, high-value innovations

Overview

- Founded in 1946 by Stanford University
- Nonprofit corporation
- Independent of Stanford in 1970
- 2000+ staff members, 20 locations
- Consolidated revenues: \$0.5 Billion
- R&D, licenses, and ventures
- Worked in >100 regions around the world

Mission

SRI is committed to discovery and to the application of science and technology for knowledge, commerce, prosperity, and peace



SRI Vision Expertise

Formerly RCA's Corporate Research Center

Founded in 1942, Princeton, New Jersey
 Subsidiary of SRI International – 1986
 Incorporated into SRI – 2011

Long tradition in video

Sensing, compression, communications,
 analysis, display

Differentiating Expertise

Vision under extreme conditions (low light,
 high motion, GPS denied)



Acadia II Real Time Vision Processor:
 Quad ARM 11 + Vision Core (ASIC)

High Energy Efficiency: 300mW,
 2pJ/Op, (3-channel video
 fusion 1280x1024 60Hz)
Low Power: 1-3W system power
 with DDR

Automotive



Robotics

Medical
 Imaging



Unmanned
 Surveillance

Training

Situational
 Awareness



Intelligence



Security &
 Force Protection

Extreme Low Light Imaging



Night Vision (Army NVESD)

- Overcast star-light conditions
- Sensor noise due to dark current

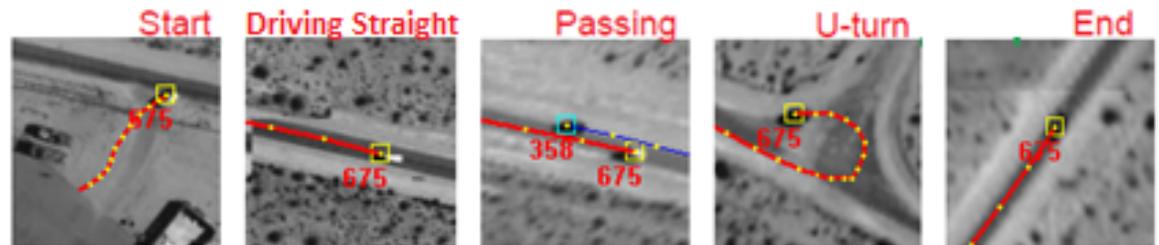
- Image denoising with multi-layer perceptrons (Burger & Schuler), arXiv: 1211.1544 (2012).
- Image Denoising and Inpainting with Deep Neural Networks (Xie & Chen, NIPS 2012)

Wide Area Aerial Surveillance: Video Processing and Exploitation



On-board Processing:

- Real-time, wide alert coverage. On-board processing limited by battery power, and analysts are deluged with sensor data.



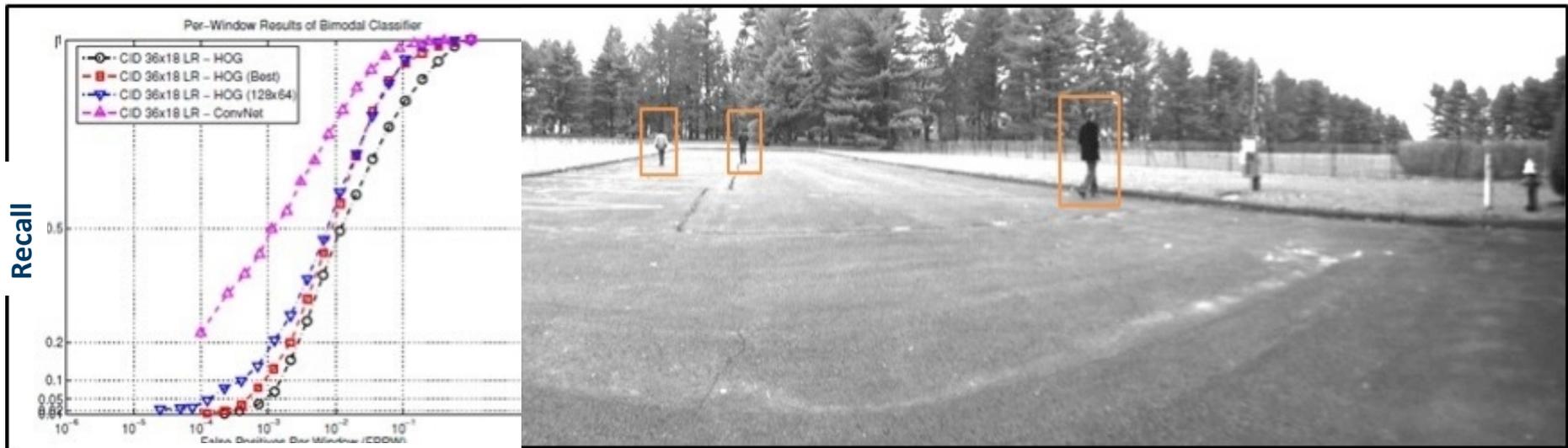
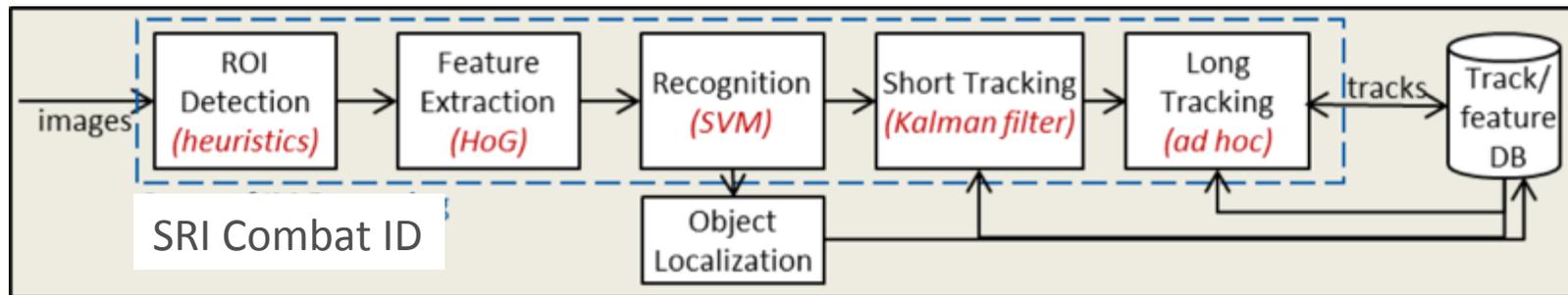
Patterns of Life:

- Processing requires high level spatial-temporal inference in wide area surveillance (e.g. infer location based on track behavior)

ONR Network Oriented Video Analysis (NOVA)

Deep Convolution Networks for Detection/Tracking

Precision-recall graph from comparison of conventional methods (HoG+SVM) for recognition vs. a deep convolutional network (pink curve)



False Positive Per Window

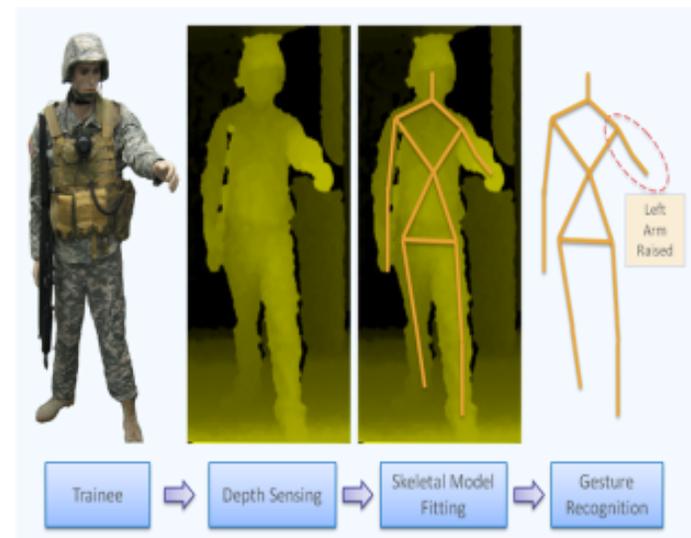
Other Applications



Visual representation of video data based on semantic meaning. Larger circles indicate more content with similar activities. (DARPA VMR)



Automated sky-line matching to geo-located urban imagery. (IARPA Aladdin)

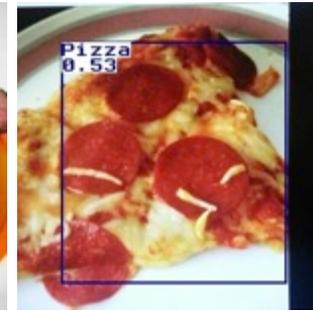
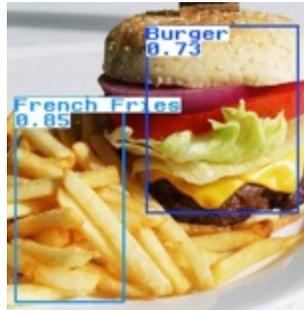


Multimodal (video, audio, gesture)
Integrated Behavior Analytics (DARPA SSIM)

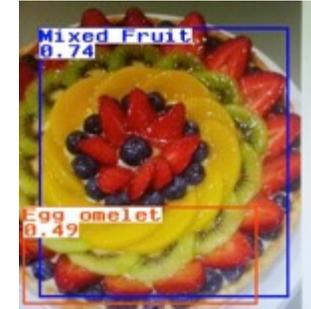
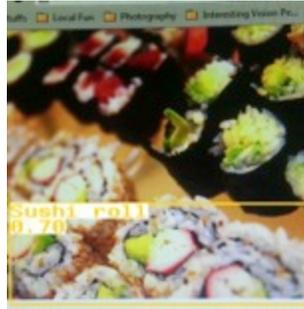
Learned classifier for categorization

Example results and video

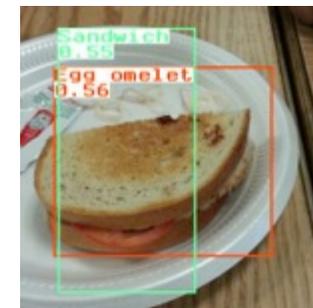
Web
Image



Mobile
Photo



SRI
cafeteria



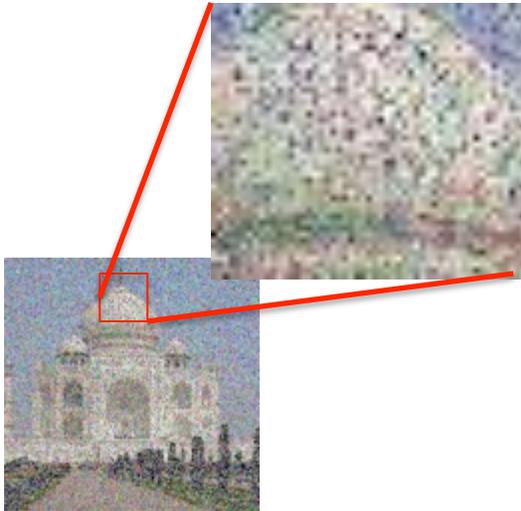
Example Visual Search

Packaged Food Identification

@ShopRite 08/15/2013

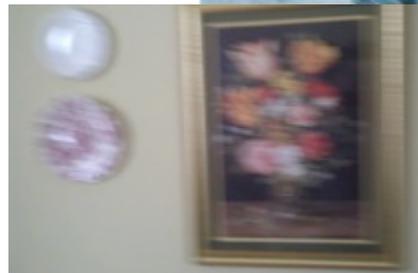
SRI International

Noise in Input data



Poor Lighting

Dark current noise in sensor produce color artifacts.



Motion Noise

Camera motion and instability leads to motion blur, and limits the ability to zoom in.

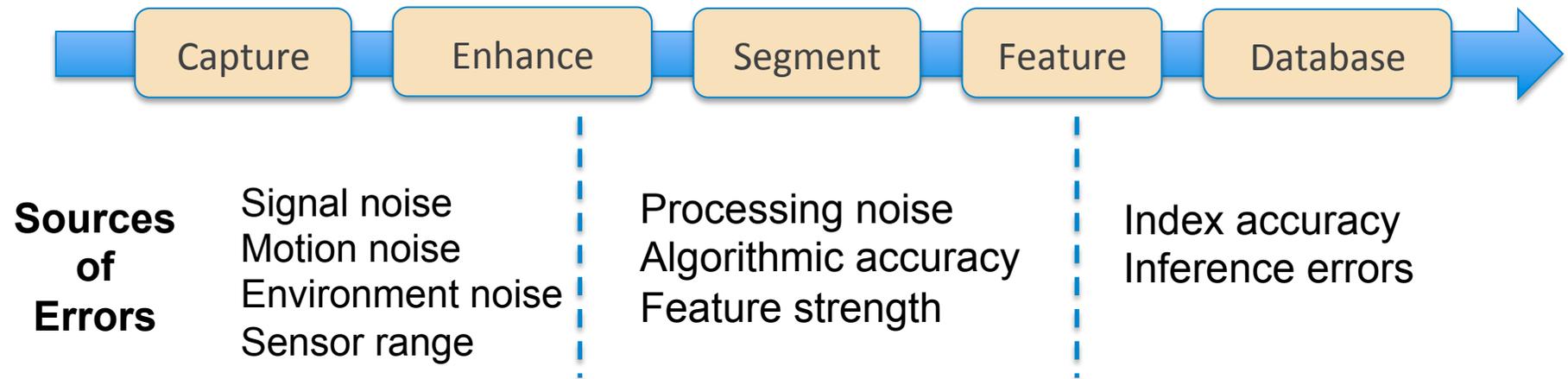


Dynamic Range

Wide dynamic range requirements in urban situations. Low features.

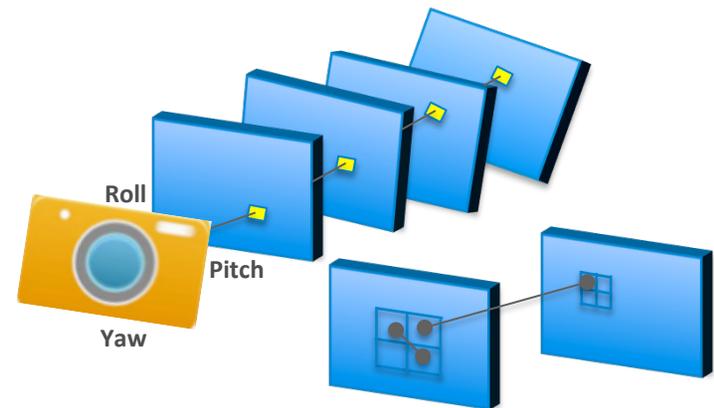
Sources of Error

Example visual search flow, from image capture to real-time query

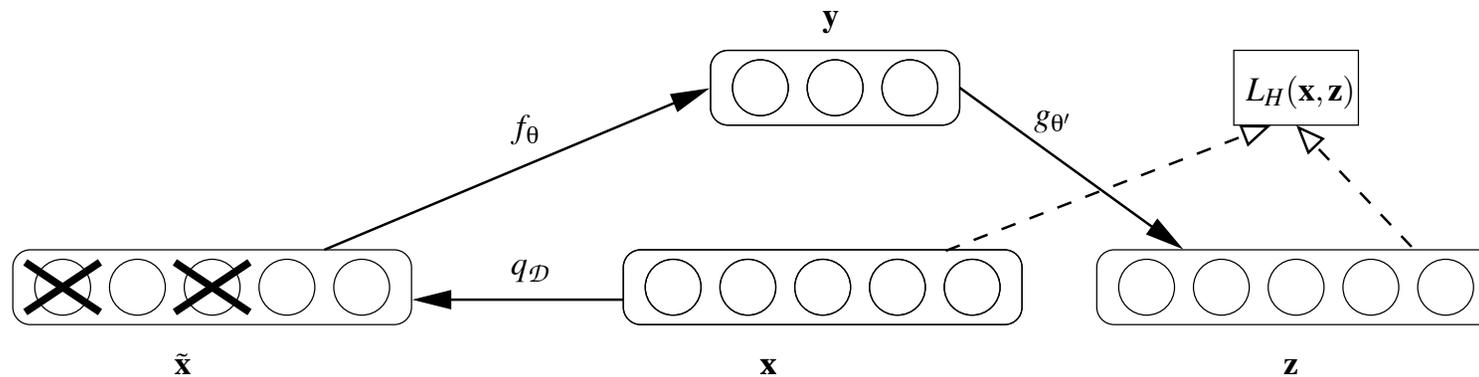


Opportunity:

Leverage Spatial, Temporal, Kinematic and Modeling redundancy in video data



De-noising Autoencoder



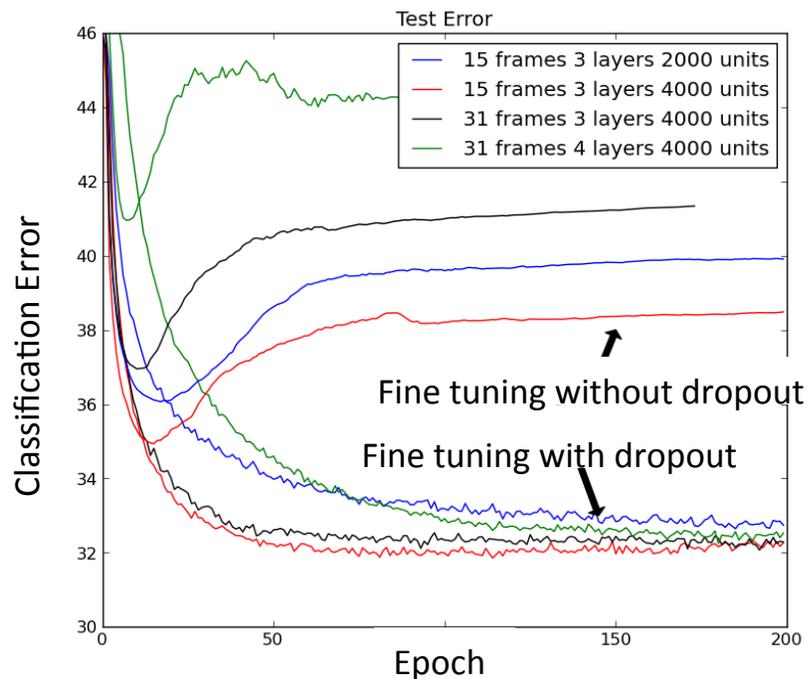
1. Input \mathbf{x} is stochastically corrupted to $\tilde{\mathbf{x}}$
 2. Autoencoder maps it to \mathbf{y}
 3. Reconstructs as \mathbf{z}
 4. Reconstruction error is measured by loss $L_H(\mathbf{x}, \mathbf{z})$.
- Each layer can be de-noised and stacked.
 - Improves classification performance

Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion, (Vincent, et.al.), JMLR 2010

Gradual training of deep denoising auto encoders, (Kalmanovich & Chechik), under review, ICLR 2015

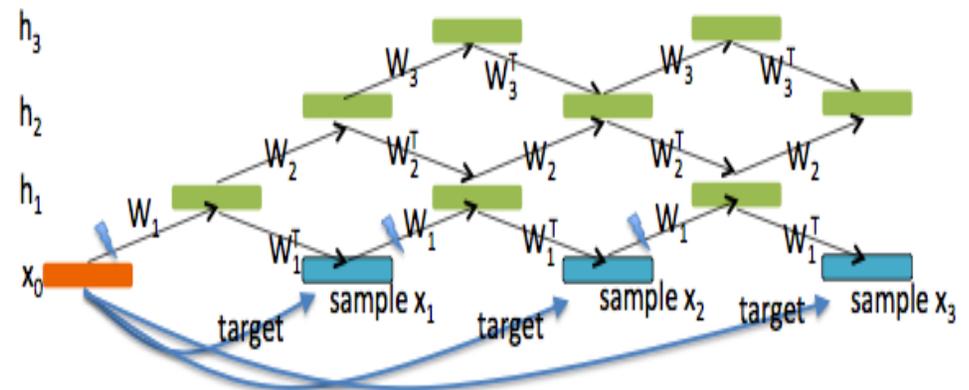
Other use of noise in Deep Learning

Dropout: each input or hidden unit is randomly omitted during training.



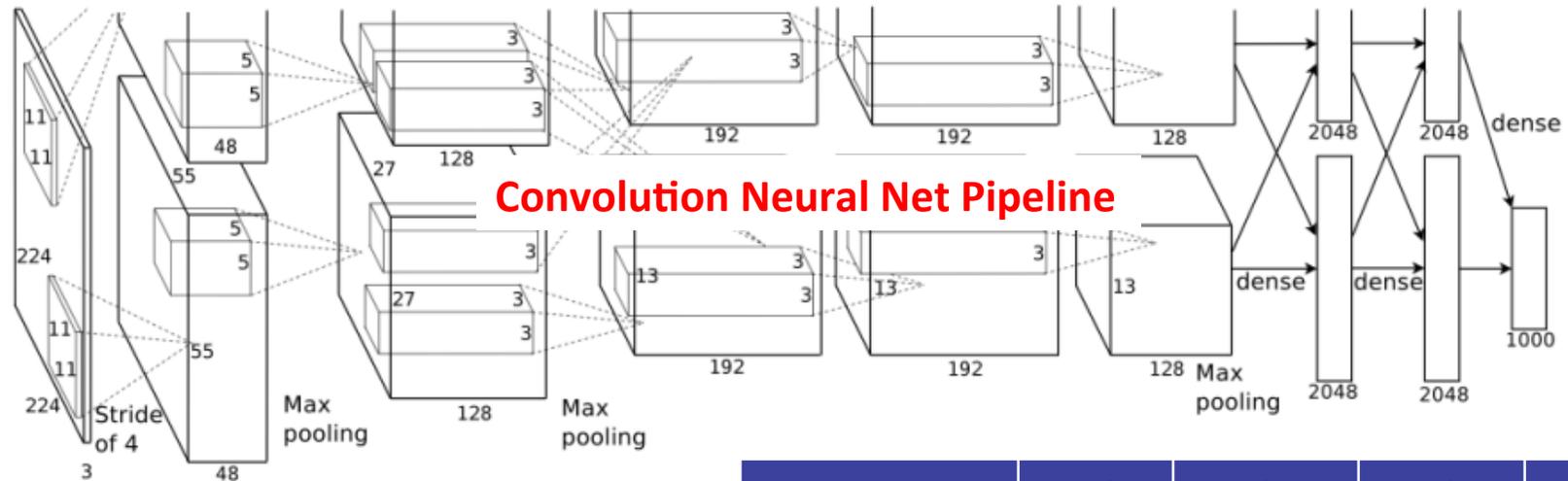
Improving neural networks by preventing co-adaptation of feature detectors (Hinton, et.al.), Technical report, arXiv:1207.0580, 2012

Generative Stochastic Network has backpropable stochastic units that operate in up to **40% noise corruption** (salt/pepper) and have similar performance to traditional backprop networks



Deep Generative Stochastic Networks Trainable by Backprop (Bengio, et.al.), NIPS 2013

Extending the State of Art in Deep Learning



ImageNet Large Scale Visual Recognition

(automatic recognition of 1000 categories of objects in images)

- Deep network (7 layers) with high connectivity (4K dimensions).
- 65K neurons, 60M parameters, 630M connections
- 512x512 images requires 100GOP/frame. At 30Hz, 3000 GOP/s.

	Intel 4 core	Nvidia GTX780	Dual ARM	FPGA 7045
Peak GOP/s	200	3800	10	120
Actual GOP/s	90	620	0.4	100
Power (W)	45	500	2	5
GOP/s/W	2	1.2	0.2	20

[Eugenio Culurciello, Purdue]

Want to deliver **3000 GOP/s at 2 W**.
(**250x** reduction in power, or **380x** speedup)



Scaling Computation

Key Insights:

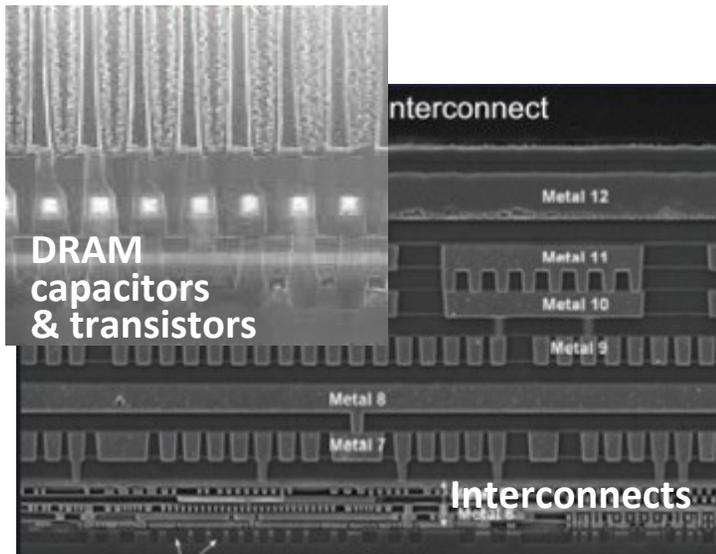
- Requires high compute & interconnect density (65K neurons, 630M connections; more for spatio-temporal analysis).
- Requires high fan-out for interconnects.
- Power and noise are key issues we must address.

Research Question:

Can noisy connections and low precision memory? With what trade-off?

- Availability of large quantities of labeled data for training was critical to achieve the algorithmic performance.
- New research in deep learning algorithms has also looked into noise corruption to add additional algorithmic robustness, (a) in training data, and (b) in hidden layers during training in a process.

Challenges in Nanoscale Design



Nanoscale design issues:

Power density, dopant fluctuation, aging, etc.

Ronald G. Dreslinski, et.al., "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits",
 Proceedings of IEEE vol 98 no 2, Feb 2010

Related computing approaches:

- Near-threshold voltage operation improves power efficiency, but can operate noisily.
- Approximate computing reduces or lower precision for power savings.
- Fault tolerant computing maintains system robustness in complex highly integrated chips.

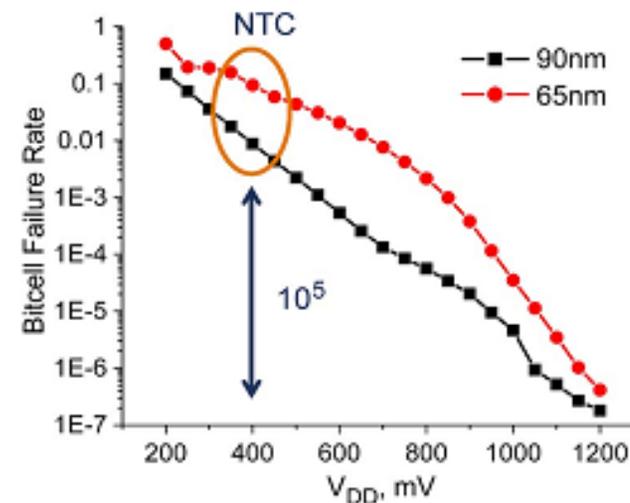


Fig. 7. Impact of voltage scaling on SRAM failure rates.

Leveraging Noise in Design

Key Insight:

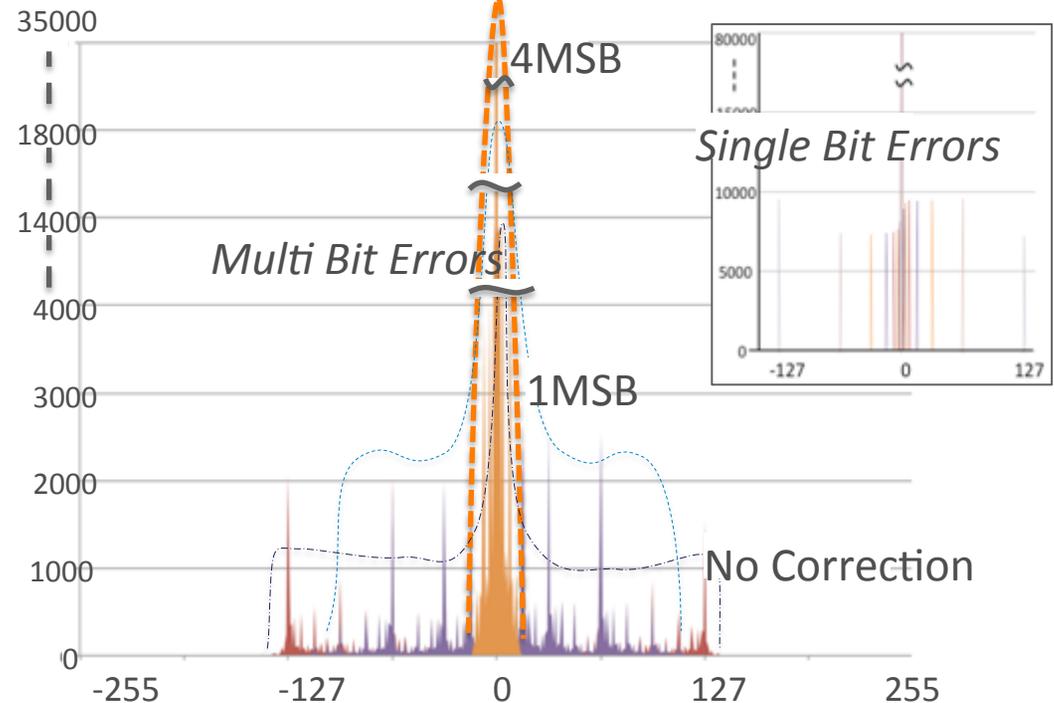
- Boltzman machines and Markov Random Field (MRF) have built-in noise-immunity characteristics based on feedback mechanisms.
- Many machine learning algorithms are **Iterative Solvers**; they are noise resilient.



p – for each bit

Computational Noise Resiliency in Deep Learning Architecture, (Sek Chai, et. al),
Workshop on NeuroArch, Minneapolis, MN,
June 2014

Statistical Distribution of Error
due to bit flip probability in memory

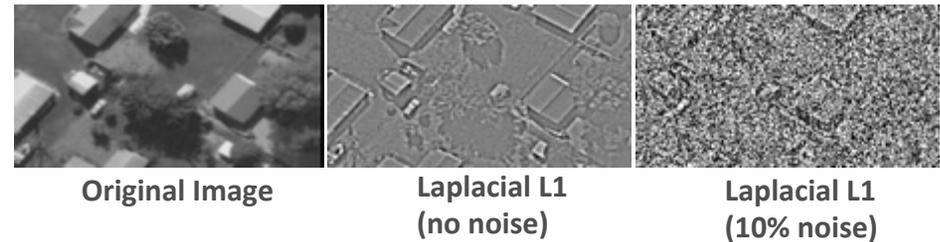


Understanding Compute Noise

- Noise shape is similar to Gaussian noise at low probability (zero centered).
→ **If we can control noise, we can accept or tolerate it in computing.**
- At 50%, noise shape is flat (random).

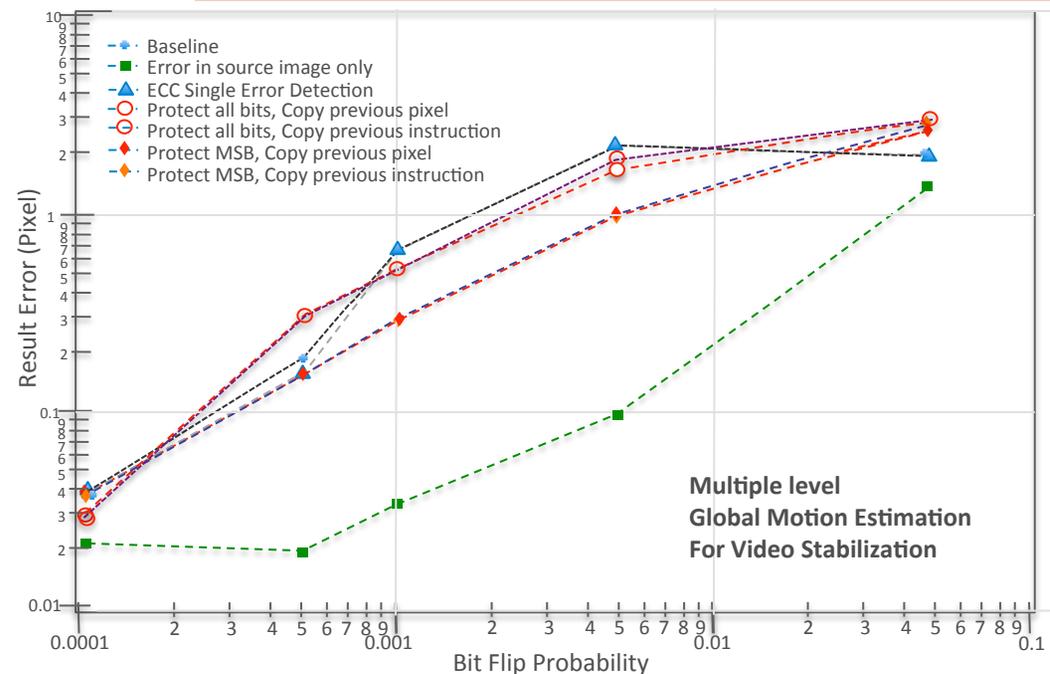
Ultra Low Power through Approximate Computing

- Processor operates in lower power domains and in presence of **computing noise** under:
 - Near Threshold Voltage (NTV)
 - Approximate value
- Statistical distribution** of computing noise → we do not have to correct induced hardware errors at every instance.
- Study the range of noise, voltage, and algorithm resiliency to quantify new ultra-low power regions of operation. Leverage algorithmic performance with noisy data.
- Offer Lightweight mechanisms to extend algorithm resiliency



Take away:

- We can reduce 0.2-0.5V (65nm) → up to **2.5-4X** energy savings per bit (voltage scaling only), without losing sufficient algorithmic performance.



Sek Chai, et. al, "Lightweight Detection and Recovery Mechanisms to Extend Algorithm Resiliency in Noisy Computation", to appear in Workshop on Near Threshold Computing, Minneapolis, MN, June 2014

Noise Resilience Design Options

Noise provides additional design knob beyond voltage and frequency for performance.

Noise Resilience

- Lower bit precision for different neural layers
- Lower voltage to near thresholds
- Reduce transistor sizing as induced noise
- Match the analog characteristics of the circuit element (e.g. resistive memories)
- Scale bus with noise only for least significant bits
- Allow crosstalk as a form of induced noise
- Dynamically adjust noise level based on algorithm needs (e.g. using ECC memory as example).
- And more...

***Noise as a
“design knob”.***

Summary

- Applications: Addressing *noise* in input data
- Algorithms: Leverage *noise* in training
- Hardware: Leverage *noise* in design