IAS 2019-02-22

The Epistemology Of Deep Learning

Yann LeCun
Facebook AI Research
New York University
http://yann.lecun.com

facebook
Artificial Intelligence Research

DL: Engineering Science or Natural Science?

- ***Engineering science: inventing new artifacts***
  - Telescope, steam engine, electromagnet, airplane, fertilizer, radio….
  - Methods: creation, intuition, tinkering, exploration, experimentation, happenstance….
    - guided by theoretical, conceptual, intuitive understanding.

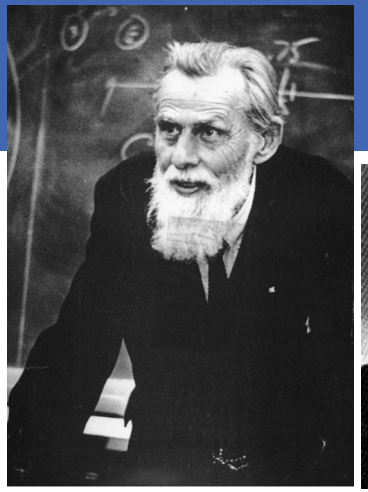
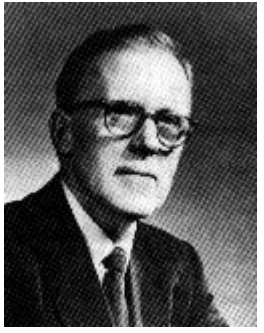
- ***Natural Science: discover, study and explain phenomena***
  - Optics, thermodynamics, electromagnetics, aerodynamics, chemistry, electronics,..
  - Methods: reproducible experiments in controlled conditions, mathematics, statistics, systematic experiments
    - guided by theoretical, conceptual, intuitive understanding.

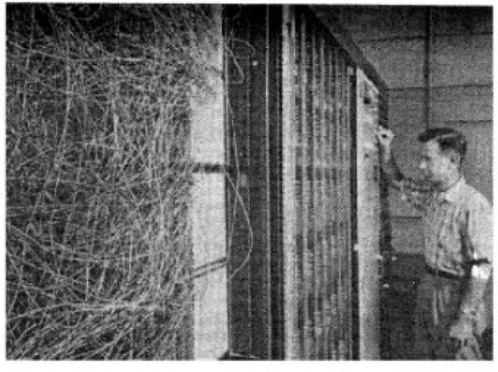
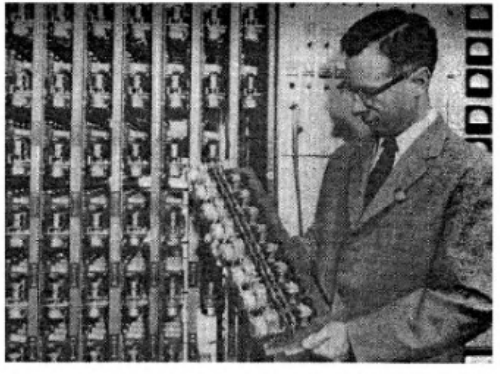
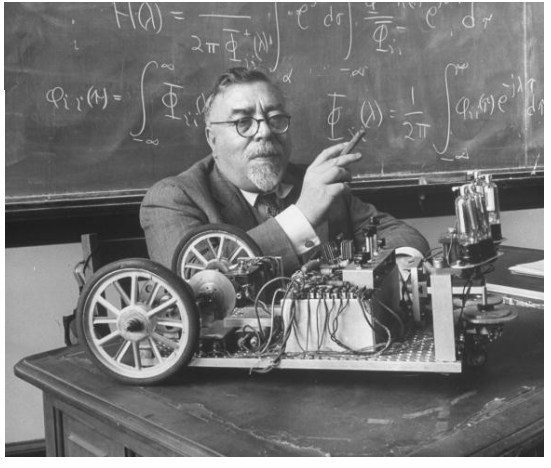
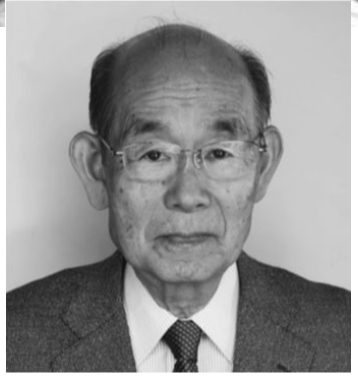
Theory often Follows Invention

- ***Telescope [1608]***
- ***Steam engine [1695-1715]***
- ***Electromagnetism [1820]***
- ***Sailboat [???]***
- ***Airplane [1885-1905]***
- ***Compounds [???]***
- ***Feedback amplifier [1927]***
- ***Computer [1941-1945]***
- ***Teletype [1906]***
- ***Optics [1650-1700]***
- ***Thermodynamics [1824-….]***
- ***Electrodynamics [1821]***
- ***Aerodynamics [1757]***
- ***Wing theory [1907]***
- ***Chemistry [1760s]***
- ***Electronics [….]***
- ***Computer Science [1950-1960]***
- ***Information Theory [1948]***

Inspiration for DL: The Brain!

- ***McCulloch & Pitts (1943): networks of binary neurons can do logic***
- ***Donald Hebb (1947): Hebbian synaptic plasticity***
- ***Norbert Wiener (1948): cybernetics, optimal filter, feedback, autopoïesis, auto-organization.***
- ***Frank Rosenblatt (1957): Perceptron***

AI today is mostly supervised learning

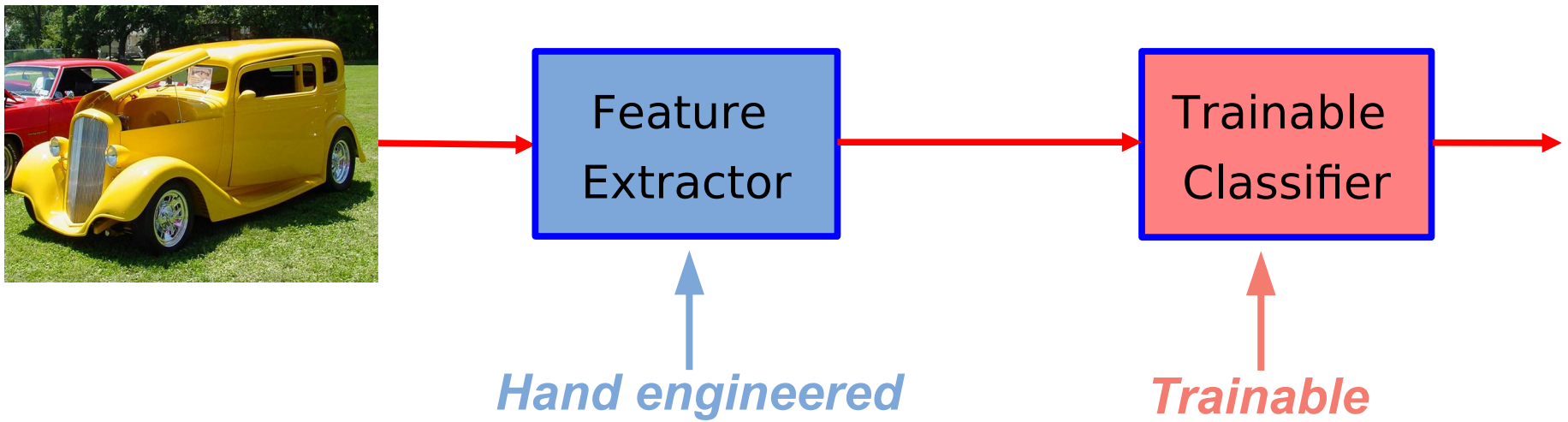
- ***Training a machine by showing examples instead of programming it***
- ***When the output is wrong, tweak the parameters of the machine***
- ***Works well for:***
  - Speech→words
  - Image→categories
  - Portrait→ name
  - Photo→caption
  - Text→topic
  - ….

CAR

PLANE

The Standard Paradigm of Pattern Recognition

- ***...and "traditional" Machine Learning***



1969→1985: Neural Net Winter

- ***No learning for multilayer nets, why?***
  - People used the wrong "neuron": the McCulloch & Pitts binary neuron
  - Binary neurons are easier to implement: No multiplication necessary!
  - Binary neurons prevented people from thinking about gradient-based methods for multi-layer nets

- ***Early 1980s: The second wave of neural nets***
  - 1982: Hopfield nets: fully-connected recurrent binary networks
  - 1983: Boltzmann Machines: binary stochastic networks with hidden units
- ***1985/86: Backprop! Q: Why only then? A: sigmoid neurons!***
  - Sigmoid neurons were enabled by "fast" floating point (Sun Workstations)

Biological Inspiration?

- ***L'Avion III de Clément Ader, 1897***

  (Musée du CNAM, Paris)
- ***His "Eole" took off from the ground on Oct 9, 1890, 13 years before the Wright Brothers, but you probably never heard of it (unless you are french).***

1104. - L' " Avion " n° 3 d'Ader

Poids 258 kgr à vide ; approvisionnements charbon et eau pour 3 heures, 52 kgr ; force motrice : 2 moteurs à vapeur de 20 HP pesant chacun 21 kgr et actionnant séparément 2 hélices tractives avec armature bambou, pales en soie et papier. Poids de chaque hélice 2 kgr 500 J. H.

La première machine volante construite par CLÉMENT ADER

Theory is Good, Because it Makes Empiricism Efficient

- ***Empiricism works***
  - Try many things, see what sticks
  - Light bulb, chemistry, X-ray, drug design,….
  - Invention almost always requires exploration
- ***But Empiricism is slow and expensive***
  - What is evolution, if not natural empiricism?
- ***Theory allows us to prune our empirical search space***
  - Theory prevents us from chasing perpetual motion
- ***Some theories allow us to predict phenomena***

Multilayer Neural Nets and Deep Learning

- ***Traditional Machine Learning***

Feature Extractor

Trainable Classifier

Hand engineered

Trainable

- ***Deep Learning***

Trainable

Low-Level Features

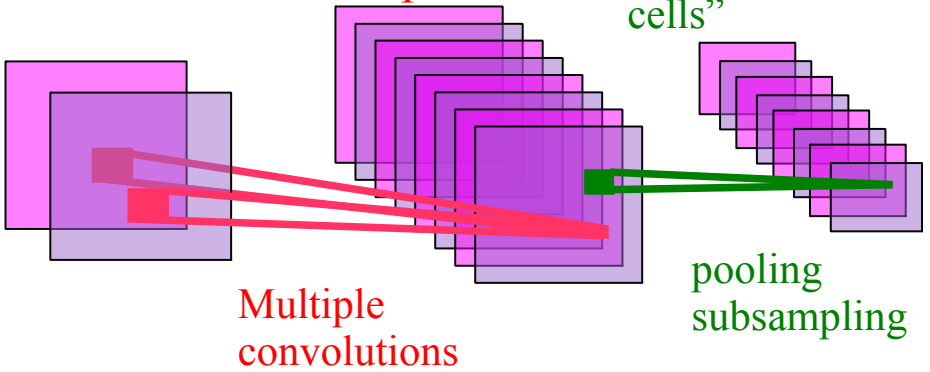
Mid-Level Features

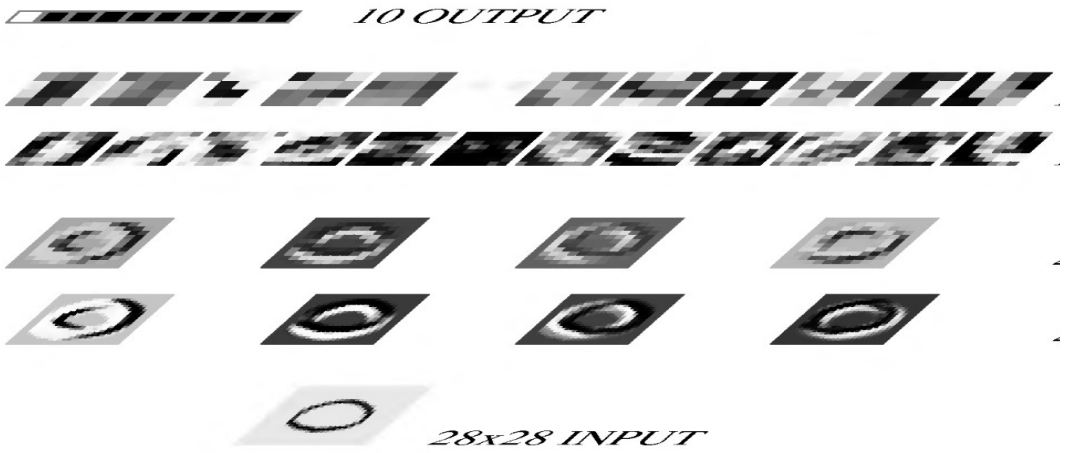
High-Level Features

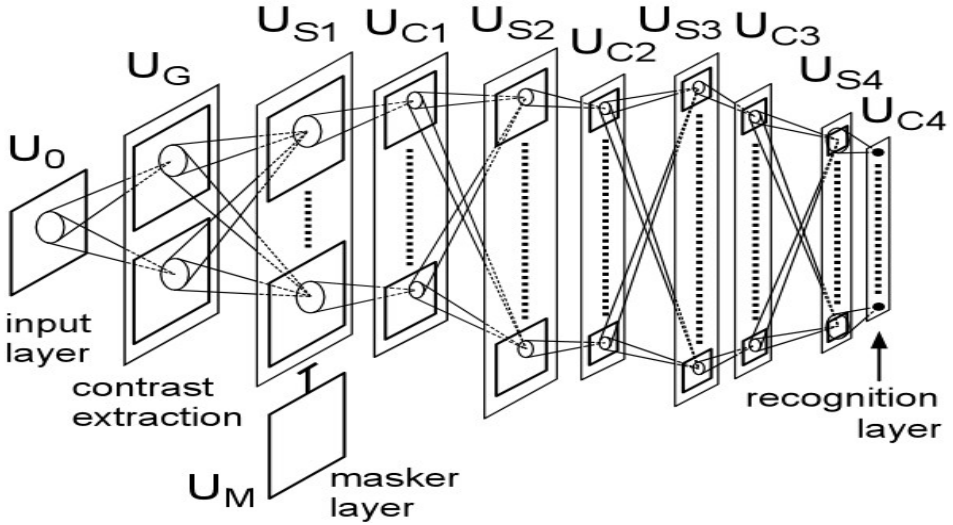
Trainable Classifier

Inspiration for ConvNets: The Visual Cortex!

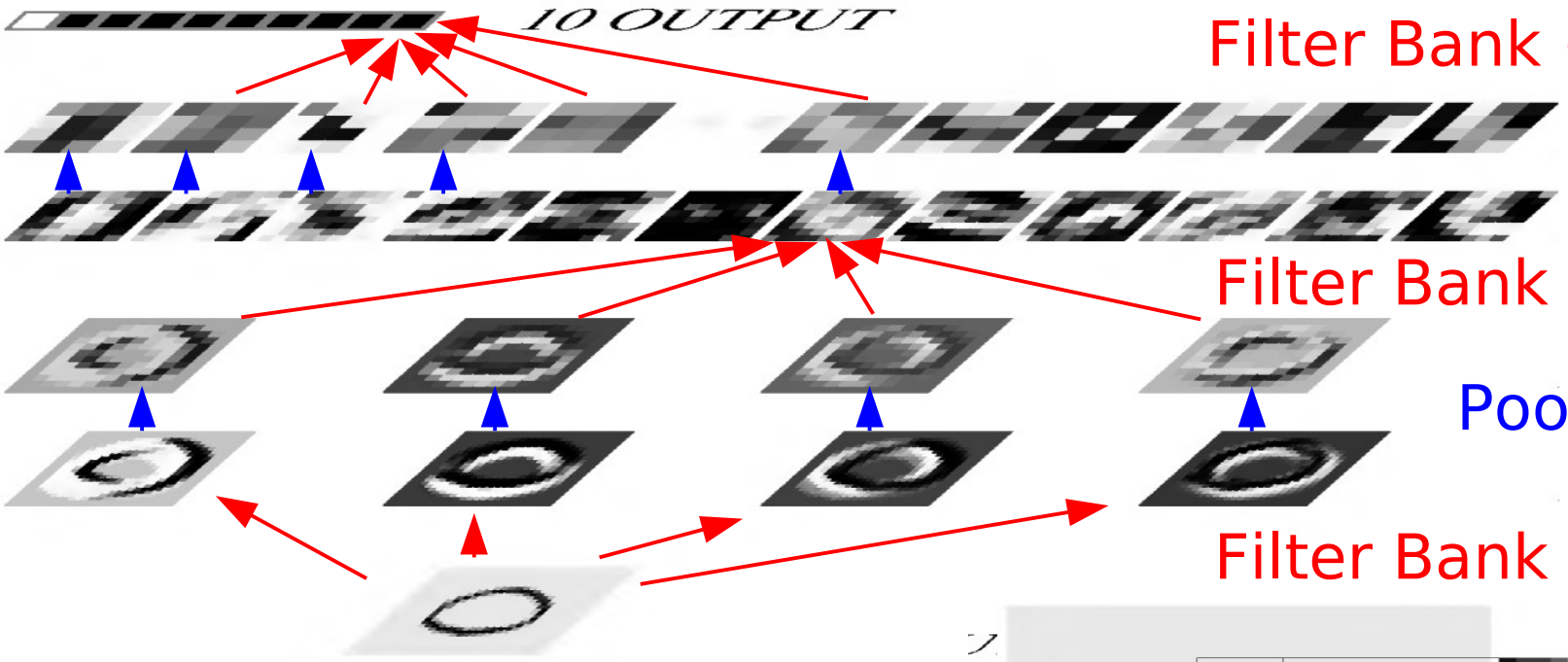
- ***Hubel & Wiesel (1959): simple & complex cells***
- ***Fukushima (1980): Cognitron***







Convolutional Network Architecture [LeCun et al. NIPS 1989]



Filter Bank +non-linearity

Pooling

Filter Bank +non-linearity

Pooling

Filter Bank +non-linearity

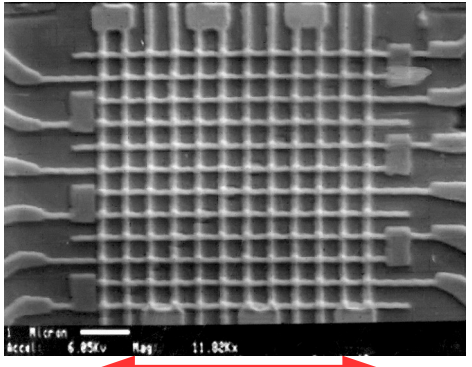
- ***Inspired by [Hubel & Wiesel 1962] & [Fukushima 1982] (Neocognitron):***
  - simple cells detect local features
  - complex cells "pool" the outputs of simple cells within a retinotopic neighborhood.

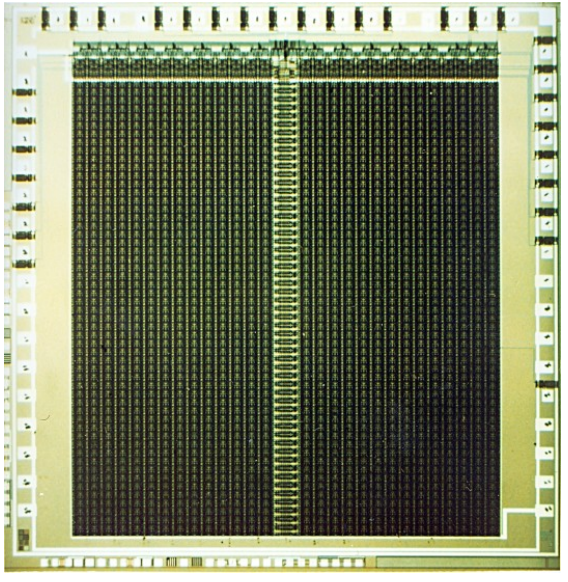
Convolutional Network (LeNet5, vintage 1990)

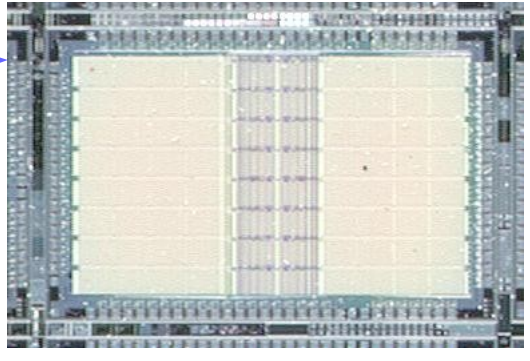
- ***Filters-tanh → pooling → filters-tanh → pooling → filters-tanh***

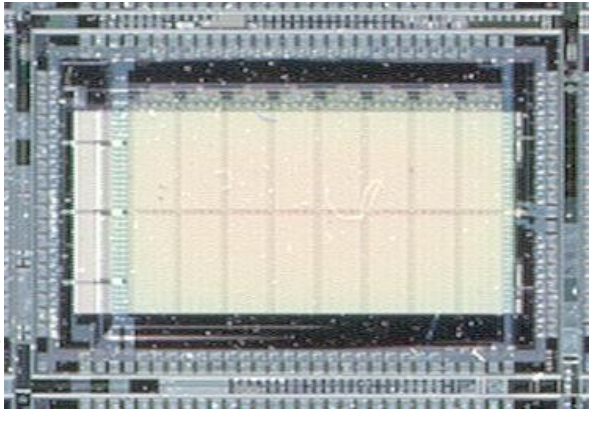
1986-1996 Neural Net Hardware at Bell Labs, Holmdel

- ***1986: 12x12 resistor array***
  - Fixed resistor values
  - E-beam lithography: 6x6microns
- ***1988: 54x54 neural net***
  - Programmable ternary weights
  - On-chip amplifiers and I/O
- ***1991: Net32k: 256x128 net***
  - Programmable ternary weights
  - 320GOPS, 1-bit convolver.
- ***1992: ANNA: 64x64 net***
  - ConvNet accelerator: 4GOPS
  - 6-bit weights, 3-bit activations







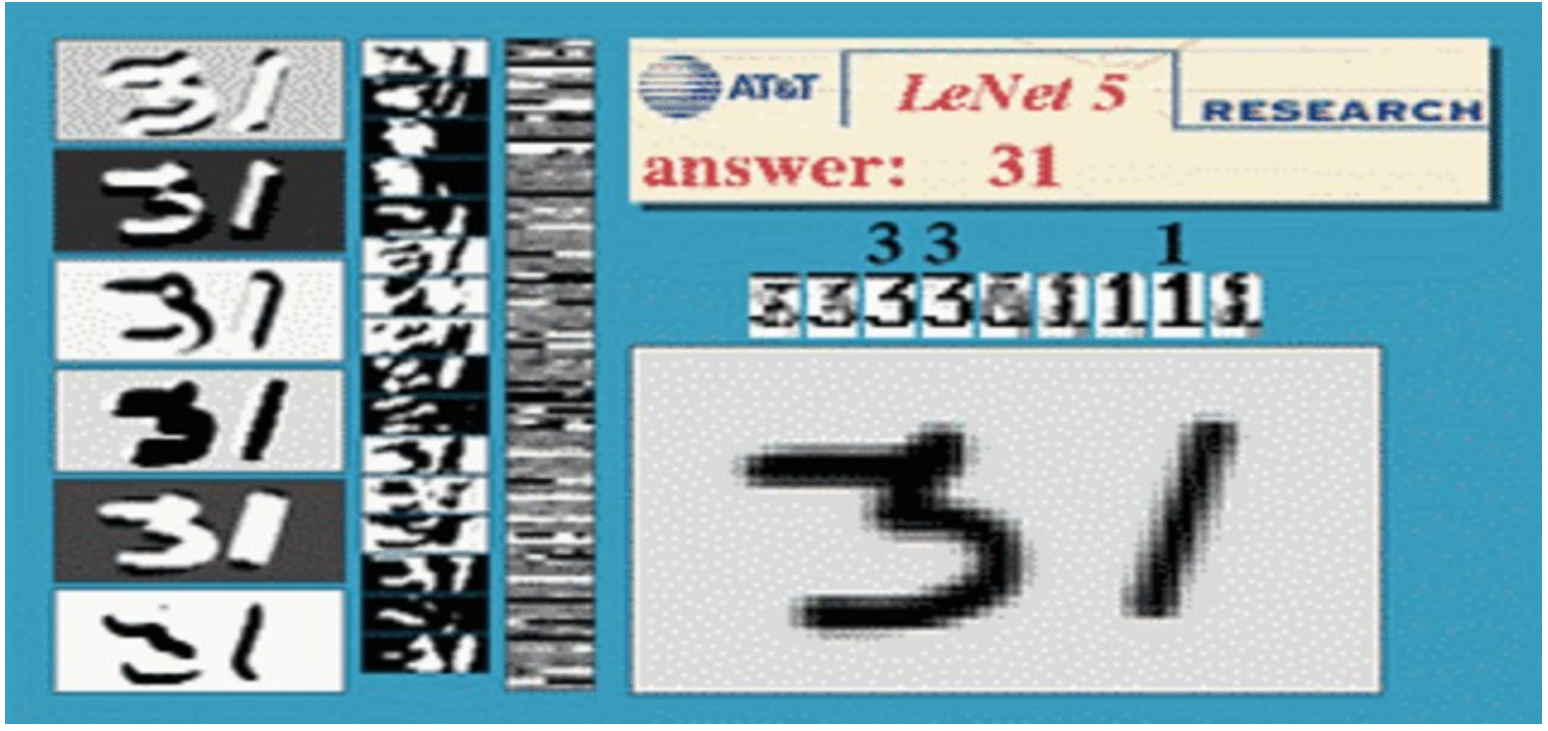


LeNet character recognition demo 1992

- ***Running on an AT&T DSP32C (floating-point DSP, 20 MFLOPS)***

ConvNets can recognize multiple objects

- ***All layers are convolutional***
- ***Networks performs simultaneous segmentation and recognition***



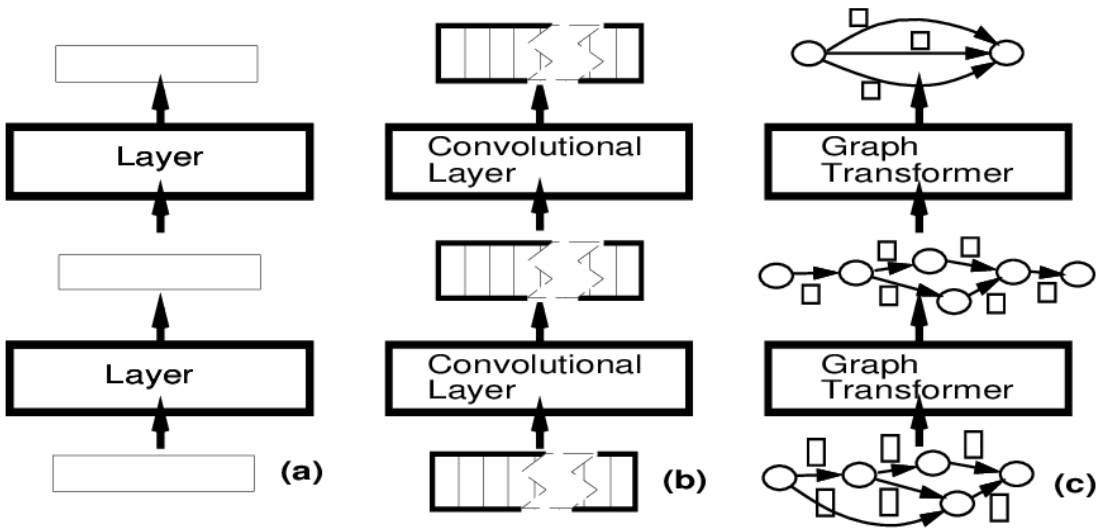
Sequence-Level Training: Graph Transformer Networks

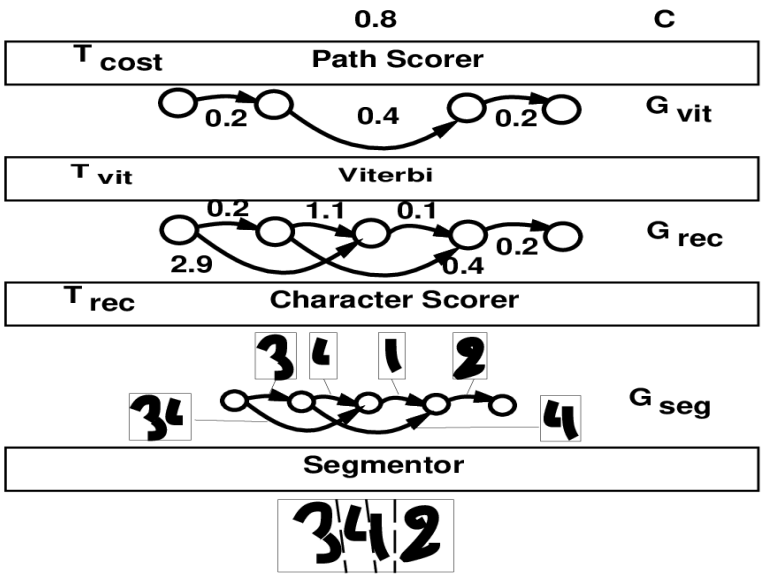
READING CHECKS WITH MULTILAYER GRAPH TRANSFORMER NETWORKS

Yann Le Cun *Léon Bottou* *Yoshua Bengio*

Speech and Image Processing Services Research Lab
AT&T Labs, 101 Crawfords Corner Road, Holmdel, NJ 07733, USA

- ***ICASSP 1997 (Munich)***



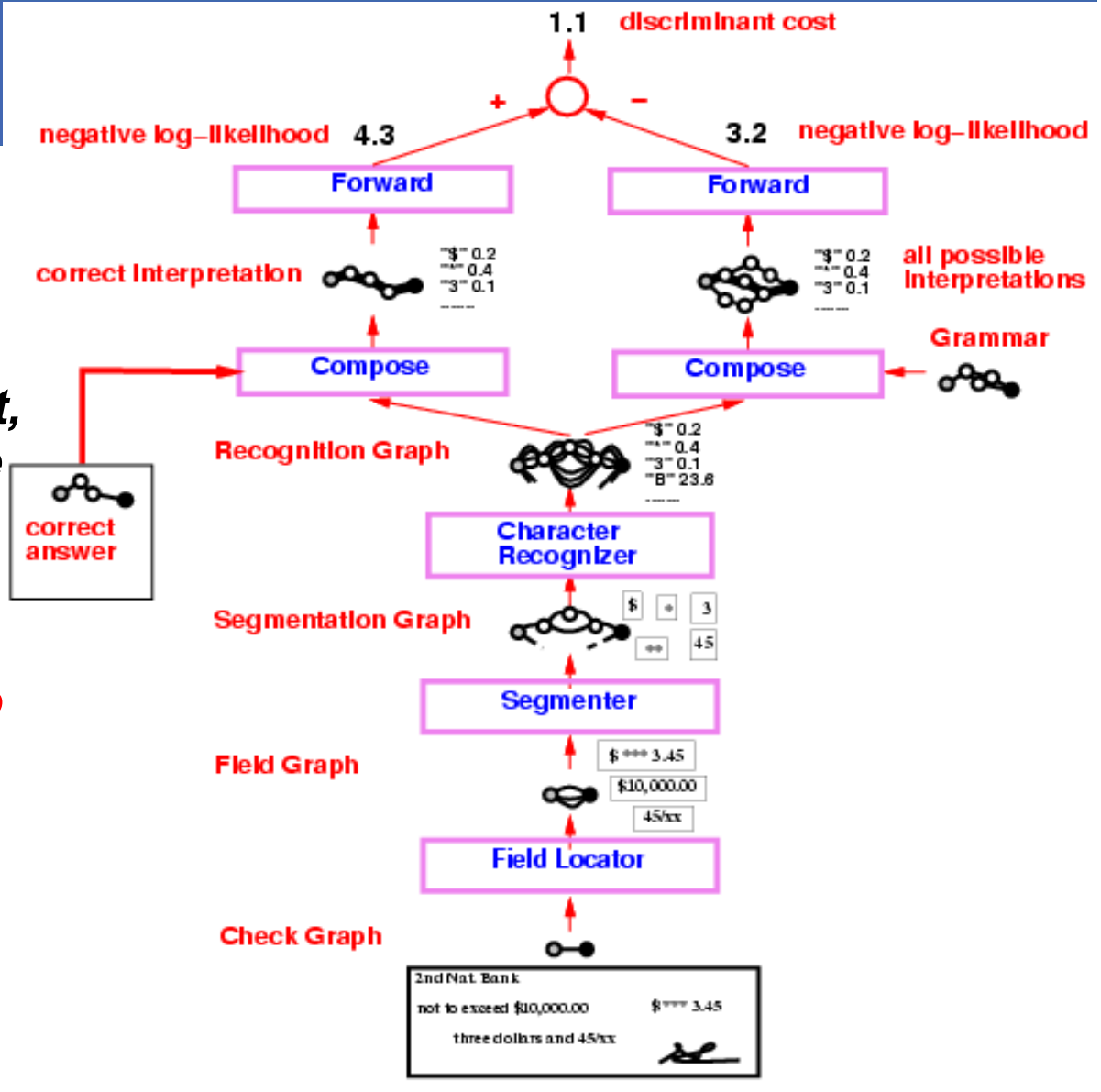


- ***Discriminative training of word recognizer at the word level.***
- ***Uses differentiable finite-state transducers***

Check Reader (AT&T 1995)

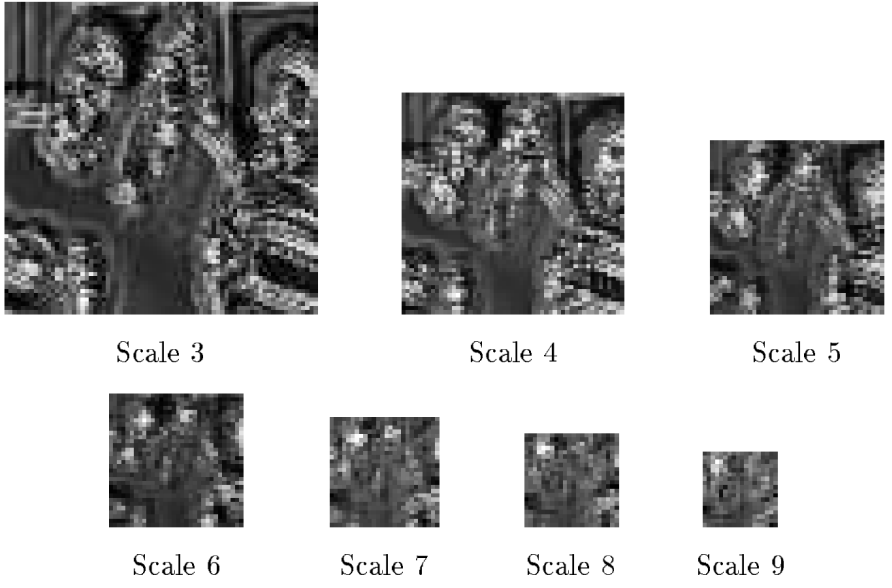
- ***Check amount reader***
- ***ConvNet+Language Model trained at the sequence level.***
- ***50% percent correct, 49% reject, 1% error (detectable later in the process).***
- ***Fielded in 1996, used in many banks in the US and Europe.***
- ***Processed an estimated 10% to 20% of all the checks written in the US in the early 2000s.***

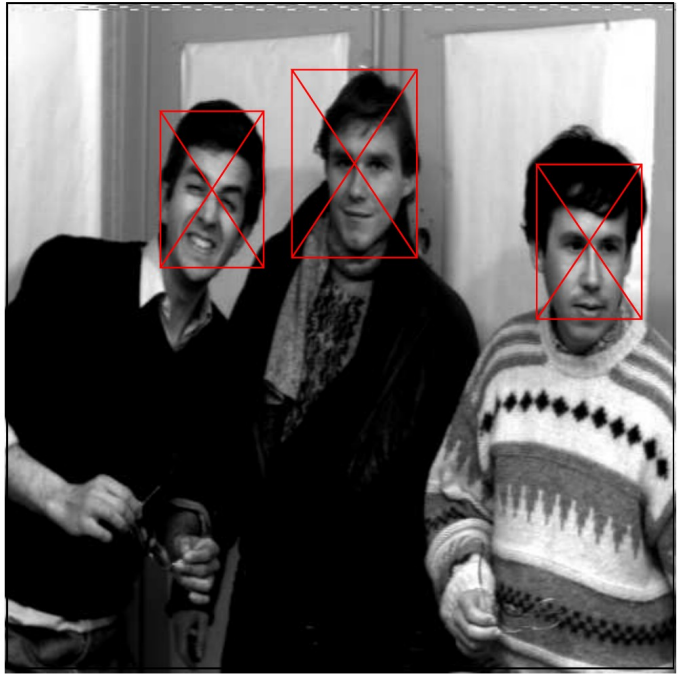
- ***[LeCun, Bottou, Bengio ICASSP1997]***
  [LeCun, Bottou, Bengio, Haffner 1998]

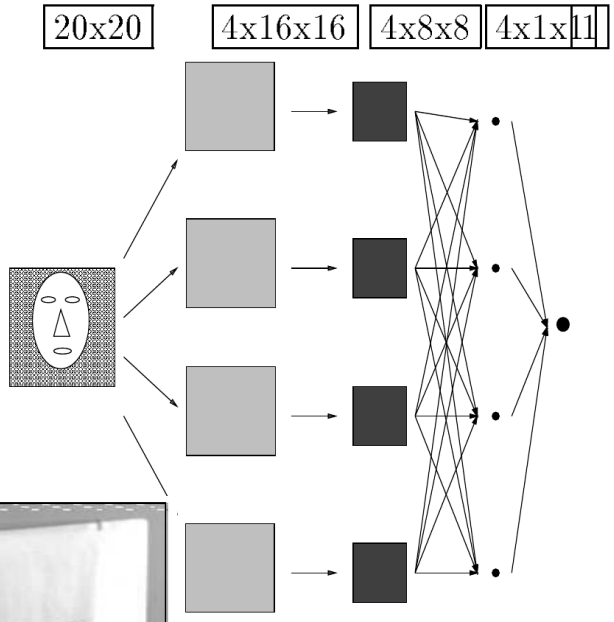


Face Detection [Vaillant et al. 1993]

- ***ConvNet applied to large images***
- ***Heatmaps at multiple scales***
- ***Non-maximum suppression for candidates***
- ***6 second on a Sparcstation for 256x256 image***







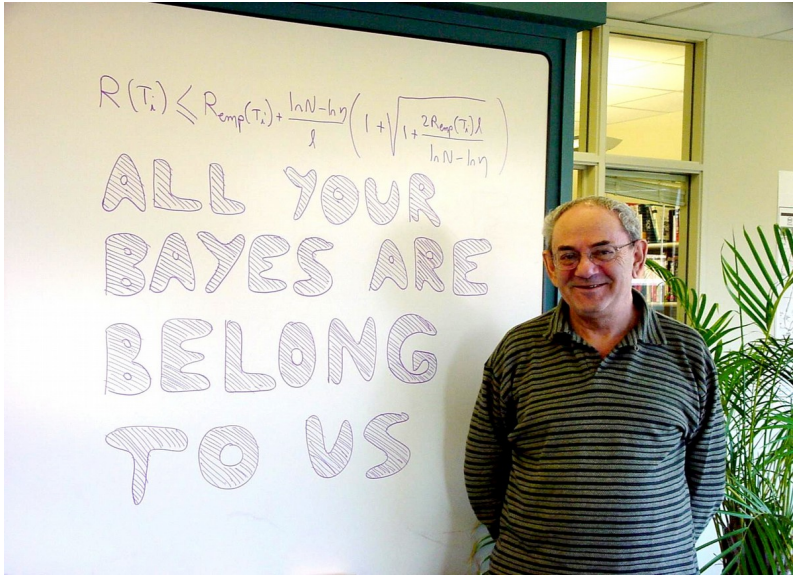
1996→2006: 2nd NN Winter! Few teams could train large NNs

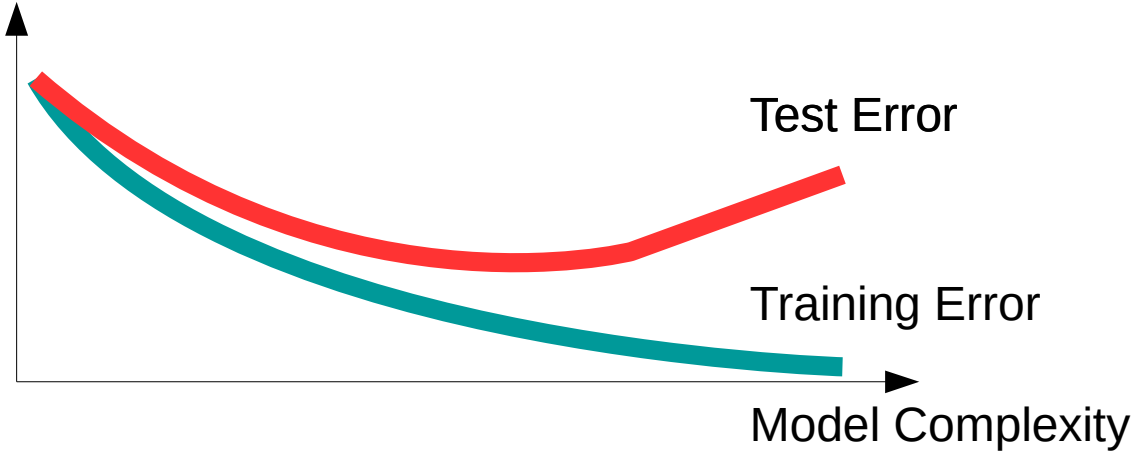
- ***Hardware was slow for floating point computation***
  - Training a character recognizer took 2 weeks on a Sun or SGI workstation
  - A very small ConvNet by today's standard (500,000 connections)
- ***Data was scarce and NN were data hungry***
  - No large datasets besides character and speech recognition
- ***Interactive software tools had to be built from scratch***
  - We wrote a NN simulator with a custom Lisp interpreter/compiler
    - SN [Bottou & LeCun 1988] → SN2 [1992] → **Lush** (open sourced in 2002).
- ***Open sourcing wasn't common in the pre-Internet days***
  - The "black art" of NN training could not be communicated easily
- ***SN/SN2/Lush gave us superpowers: tools shape research directions***

What About Learning Theory?

- ***Generalization is possible***
- ***But there is no free lunch***
  - no universal learning algorithm.
- ***Learning requires inductive bias***
  - Regularization works
- ***Conceptually magnificent***

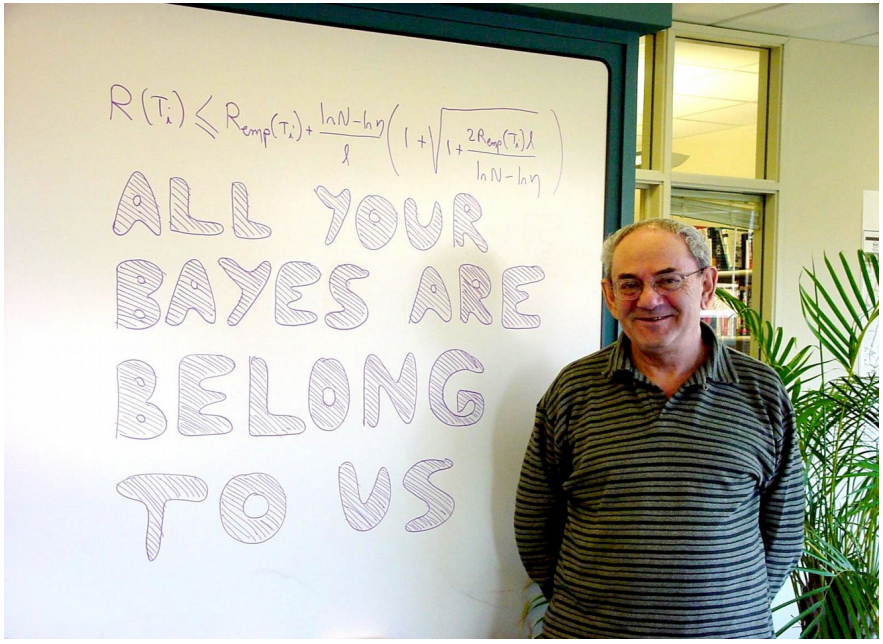
$$R(T_i) \leq R_{emp}(T_i) + \frac{\ln N - \ln \eta}{l}\left(1 + \sqrt{1 + \frac{2R_{emp}(T_i)l}{\ln N - \ln \eta}}\right)$$





What About Learning Theory?

- ***Q: But does anyone use generalization bounds to perform model selection?***
- ***A: NO!***
  - Everyone does (cross-)validation.
  - Everyone does hyper-parameter optimization through validation.
- ***Generalization bounds wi(l)dly over-estimate the generalization error***



- ***What about theories from {Bayesian|frequentist} statistics?***
  - Do they explain why a 100-million parameter neural net generalizes well with 1 million training samples?

The Vapnik-Jackel Bet in 1995

1. Jackel bets (one fancy dinner) that by March 14, 2000, people will understand
quantitatively why big neural nets working on large databases are not so bad. (Understanding means that there will be clear conditions and bounds)

Vapnik bets (one fancy dinner) that Jackel is wrong.

But .. If Vapnik figures out the bounds and conditions, Vapnik still wins the bet.

****

2. Vapnik bets (one fancy dinner) that by March 14, 2005, no one in his right mind will use neural nets that are essentially like those used in 1995.

Jackel bets (one fancy dinner) that Vapnik is wrong

_____ 3/14/95
V. Vapnik

_____ 3/14//95
L. Jackel

_____ 3/14/95
Witnessed by Y. LeCun

The 2nd Neural Net Winter
(1995-2005)
& Spring (2006-2012)

The Lunatic Fringe and the Deep Learning Conspiracy

facebook
Artificial Intelligence Research

Controversial Proposition #1

- ***Machine Learning fell victim to the theoretical street light effect when it moved away from neural nets in the mid 1990s***
  - Generalization bounds (cute math)
  - Convex optimization (provable convergence)
  - Non-stochastic optimization (cute math, proofs)
  - Reproducing kernel Hilbert spaces (cute mathematics)
  - Primal-dual methods, KKT, SDP (cuter math)
  - (Variational) Bayesian methods (cute math)
  - Non-Parametric Bayes (cute math)
  - Graphical models (cute math and algorithms)
  - Conditional Random Fields (convex structured prediction)

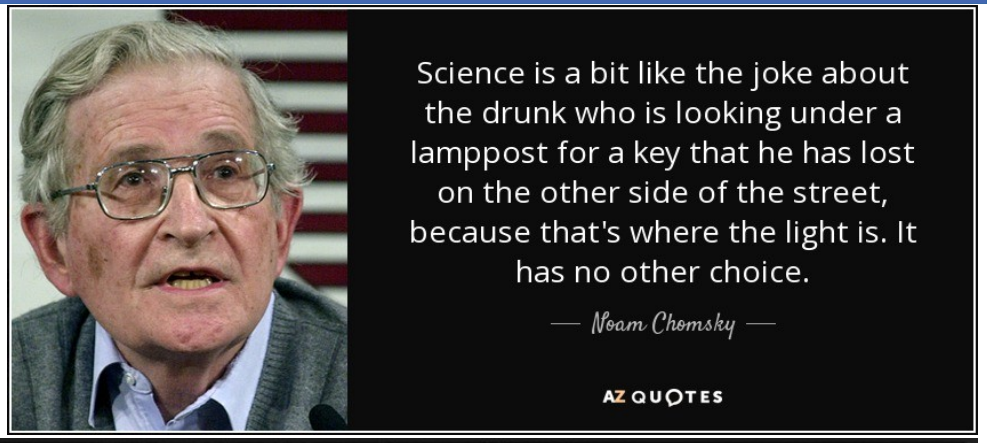
Controversial Proposition #1

- ***We had backprop, ConvNets and (non-convex) structured prediction***
- ***But the community insisted on moving towards***
  - SVM
  - Gaussian Processes
  - CRF,
  - non-parametric Bayes
- ***None of these methods had ever beaten ConvNets on MNIST***
  - But SVMs were good on smaller datasets

- ***nimio rigore mathematics facit rigor mortis***
  - Excessive mathematical rigor leads to rigor mortis

Theory can Limit our Creative Thinking

- ***The street light effect***
- ***Theory is our lamppost***
  - But the keys to AI might be elsewhere



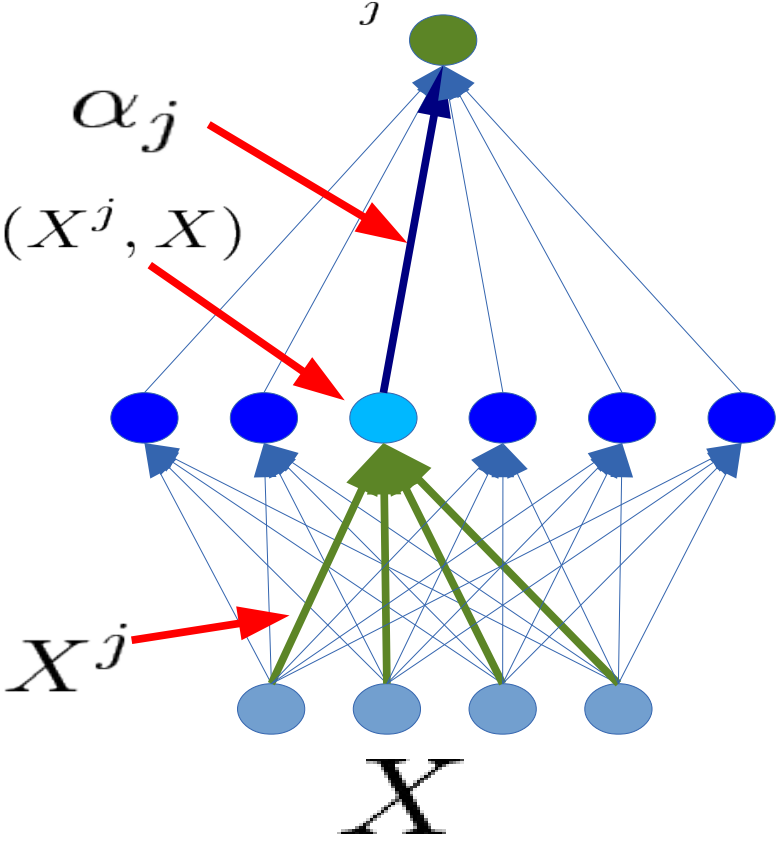
Lessons learned

- ***1.1: Hardware limitations influence research directions***
  - *It constrains what algorithm designers will let themselves imagine*
- ***1.2: Good software tools shape research and give superpowers***
  - *But require a significant investment*
  - *Common tools for Research and Development facilitates productization*
- ***1.3: Hardware performance matters***
  - *Fast turn-around is important for R&D*
  - *But high-end production models always take 2-3 weeks to train*
- ***1.4: When hardware is too slow, software is not readily available, or experiments are not easily reproducible, people will find ways to dismiss and abandon good ideas***

What's an SVM, really?

- ***2-layer models are not deep (even if you train the first layer)***
  - Because there is no feature hierarchy
  - Layer1: kernels; layer2: linear
  - The first layer is "trained" in with the simplest unsupervised method ever devised: using the samples as templates for the kernel functions.
- ***"kernel methods are glorified template matching"***

$$G(X,\alpha) = \sum_j \alpha_j K(X^j, X)$$



NORB object recognition

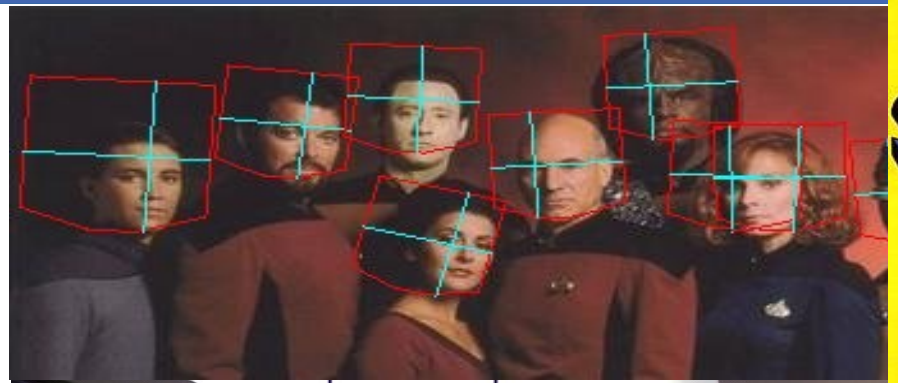
- ***NORB dataset***
  - 5 categories
  - 5 instances for training
  - 5 instances for testing
  - 18 azimuths, 9 elevations, 6 lighting conditions,
  - 2 cameras (stereo)
- ***Convexity doesn't help***

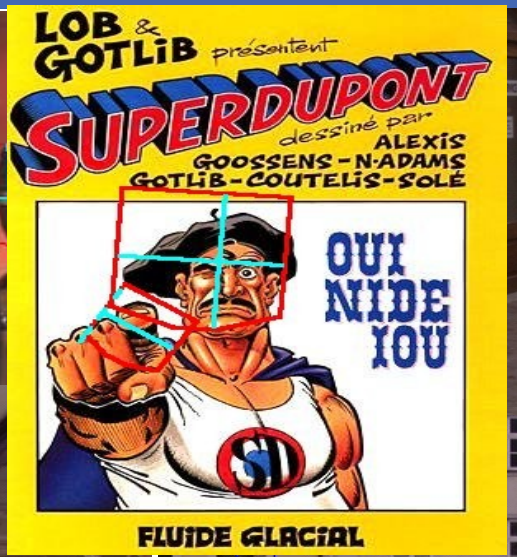
[Huang, LeCun, Bottou CVPR 2003]

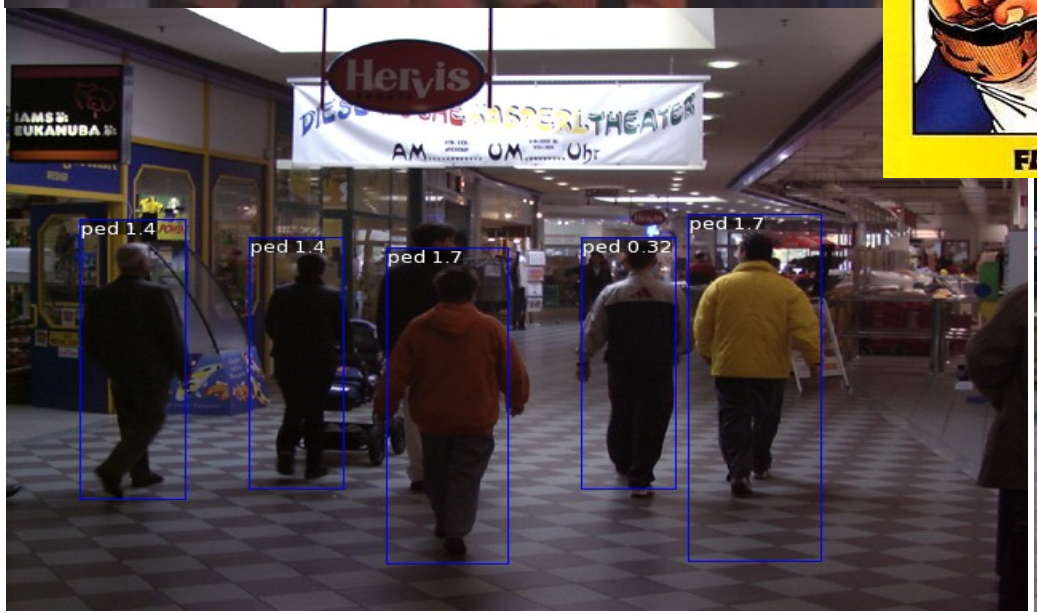
Training instances **Test instances**

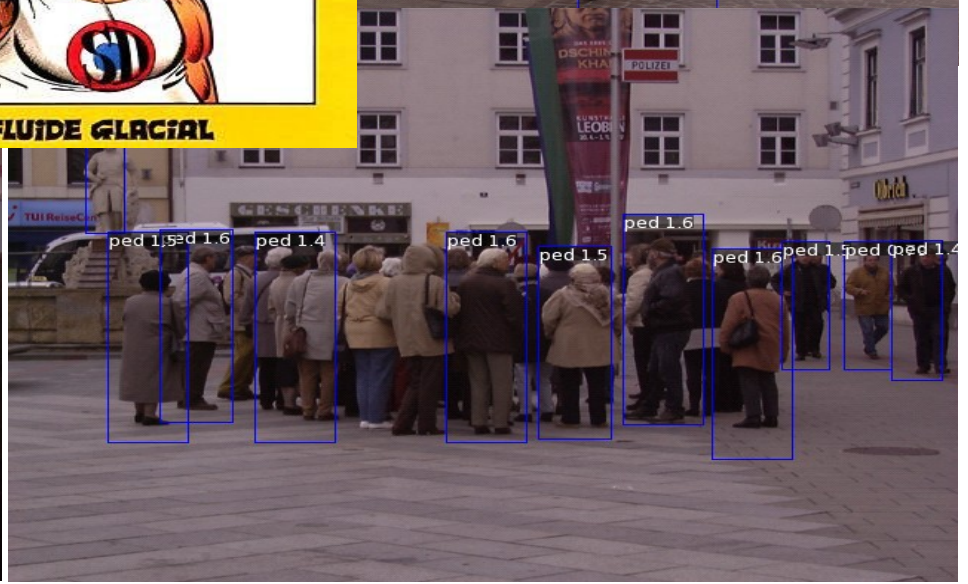
- **Linear Classifier on raw stereo images:** **30.2% error.**
- **K-Nearest-Neighbors on raw stereo images:** **18.4% error.**
- **K-Nearest-Neighbors on PCA-95:** **16.6% error.**
- **Pairwise SVM on 96x96 stereo images:** **11.6% error**
- **Pairwise SVM on 95 Principal Components:** **13.3% error.**
- **Convolutional Net on 96x96 stereo images:** **5.8% error.**

Face & Pedestrian Detection with ConvNets (1993-2005)





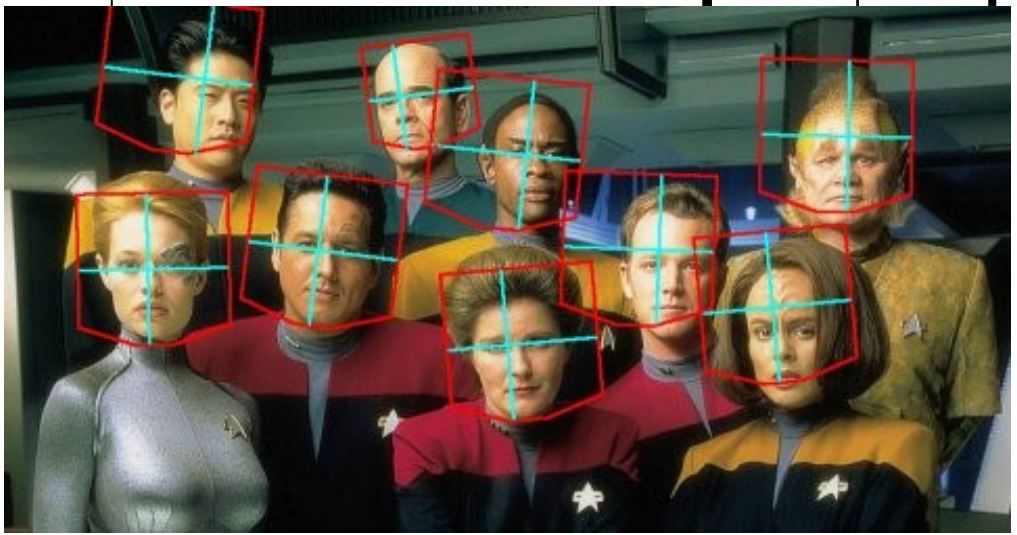


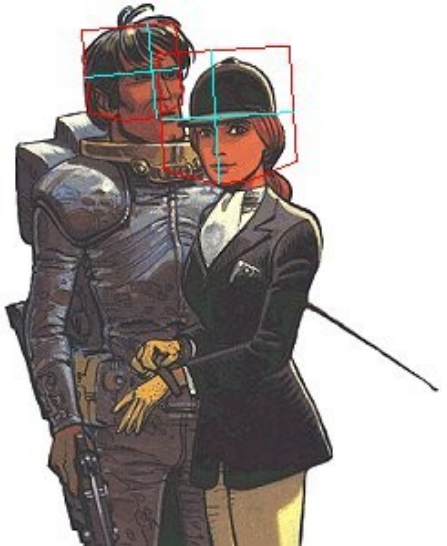


[Osadchy,Miller LeCun JMLR 2007],[Kavukcuoglu et al. NIPS 2010] [Sermanet et al. CVPR 2013]

It works

Data Set->	TILTED		PROFILE		MIT+CMU	
False positives per image->	4.42	26.9	0.47	3.36	0.5	1.28
Our Detector	90%	97%	67%	83%	83%	88%
Jones & Viola (tilted)	90%	95%	x		x	
Jones & Viola (profile)	x		70%	83%	x	

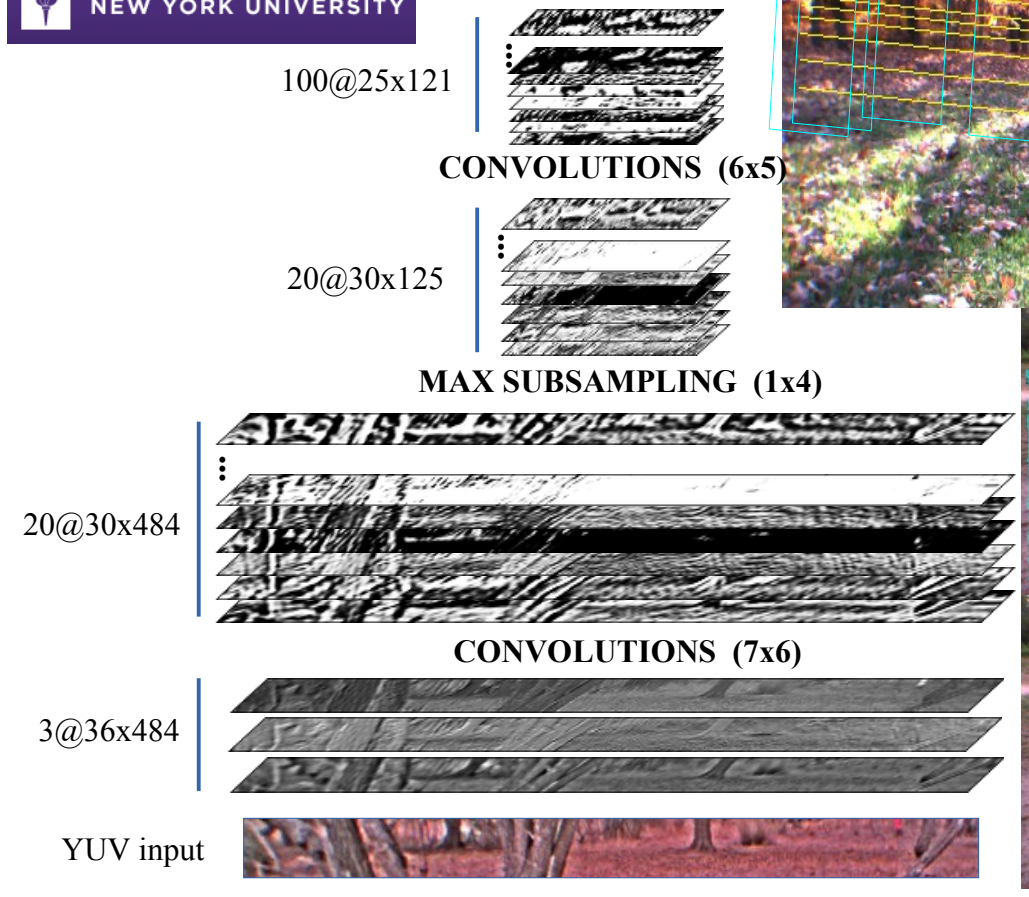


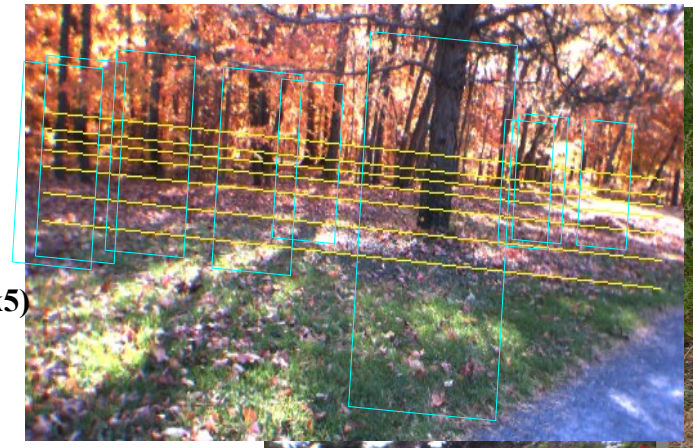


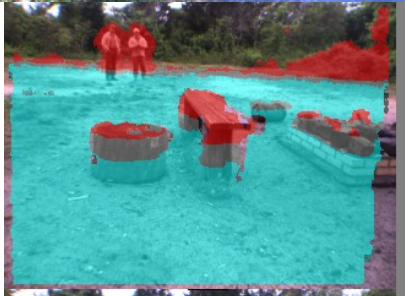
DARPA LAGR: Learning Applied to Ground Robots

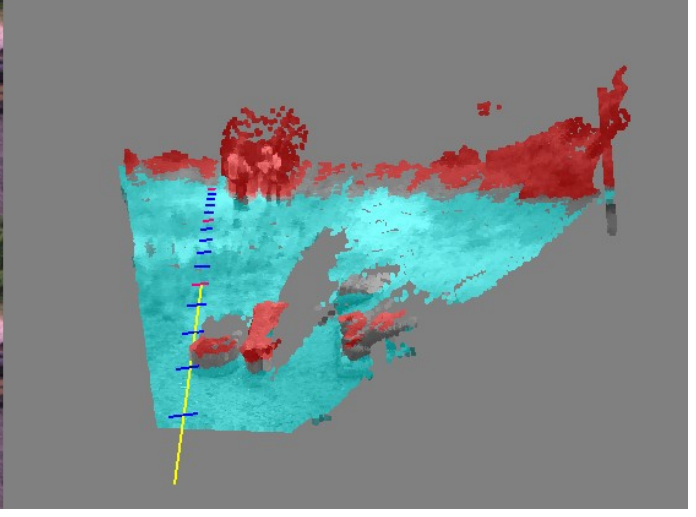
net SCALE Technologies, Inc.

NEW YORK UNIVERSITY









Semantic Segmentation with ConvNets

[Farabet et al. ICML 2011]
[Farabet et al. PAMI 2013]

VIDEO

Semantic Segmentation

	Pixel Acc.	Class Acc.
Liu *et al.* 2009 [31]	74.75%	-
Tighe *et al.* 2010 [44]	76.9%	29.4%
raw multiscale net[1]	67.9%	45.9%
multiscale net + superpixels[1]	71.9%	**50.8%**
multiscale net + cover[1]	72.3%	**50.8%**
multiscale net + cover[2]	**78.5%**	29.6%

- ***SIFT Flow Dataset***
- ***[Liu 2009]:***
- ***33 categories***

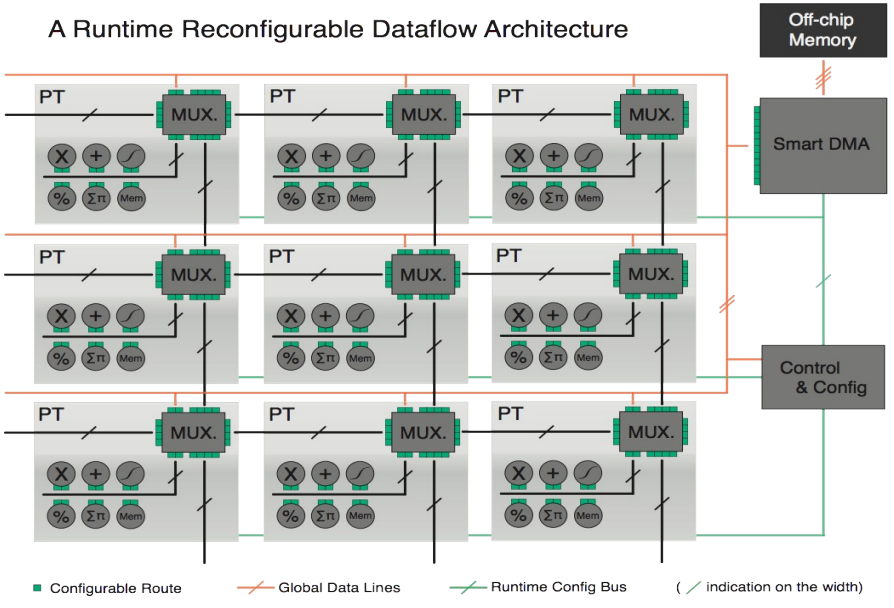
- Barcelona dataset
- [Tighe 2010]:
- 170 categories.

	Pixel Acc.	Class Acc.
Tighe *et al.* 2010 [44]	66.9%	7.6%
raw multiscale net[1]	37.8%	**12.1%**
multiscale net + superpixels[1]	44.1%	**12.4%**
multiscale net + cover[1]	46.4%	**12.5%**
multiscale net + cover[2]	**67.8%**	9.5%

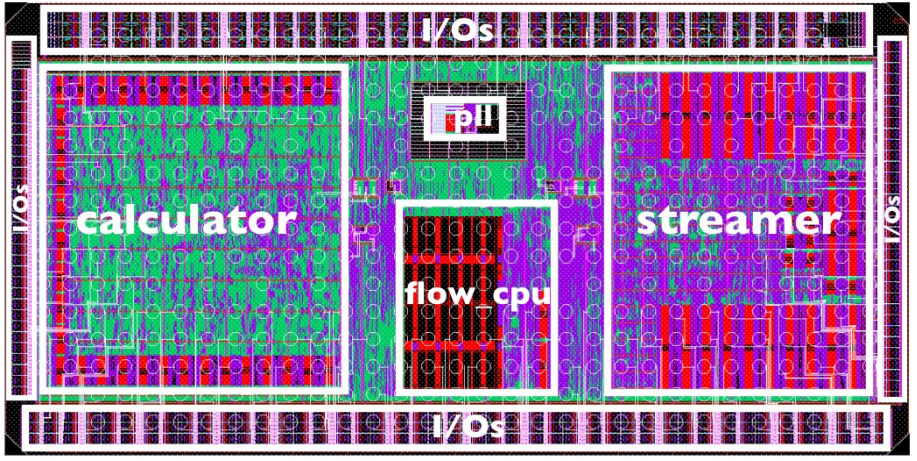
[Rejected from CVPR 2012], [Farabet ICML 2012] [Farabet et al. IEEE T. PAMI 2012]

FPGA ConvNet Accelerator: NewFlow [Farabet 2011]

- ***NeuFlow: Reconfigurable Dataflow architecture***
  - Implemented on Xilinx Virtex6 FPGA
  - 20 configurable tiles. 150GOPS, 10 Watts
  - Semantic Segmentation: 20 frames/sec at 320x240

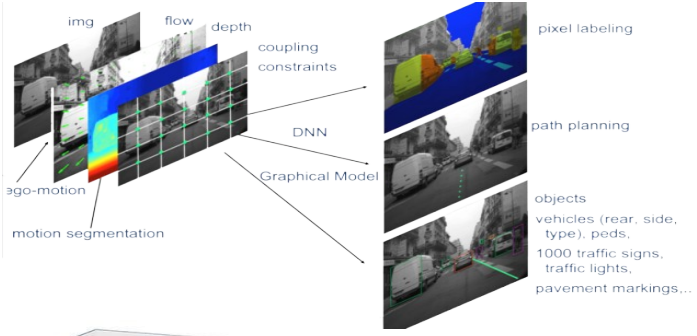


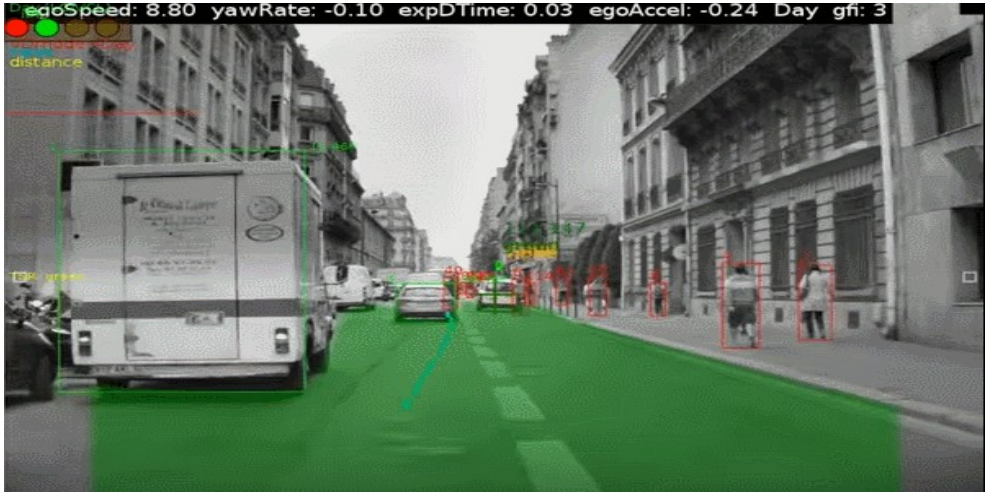
- ***NeuFlow ASIC*** *[Pham 2012]*
  - 150GOPS, 0.5 Watts (simulated)



Driving Cars with Convolutional Nets

- ***MobilEye***





- ***NVIDIA***

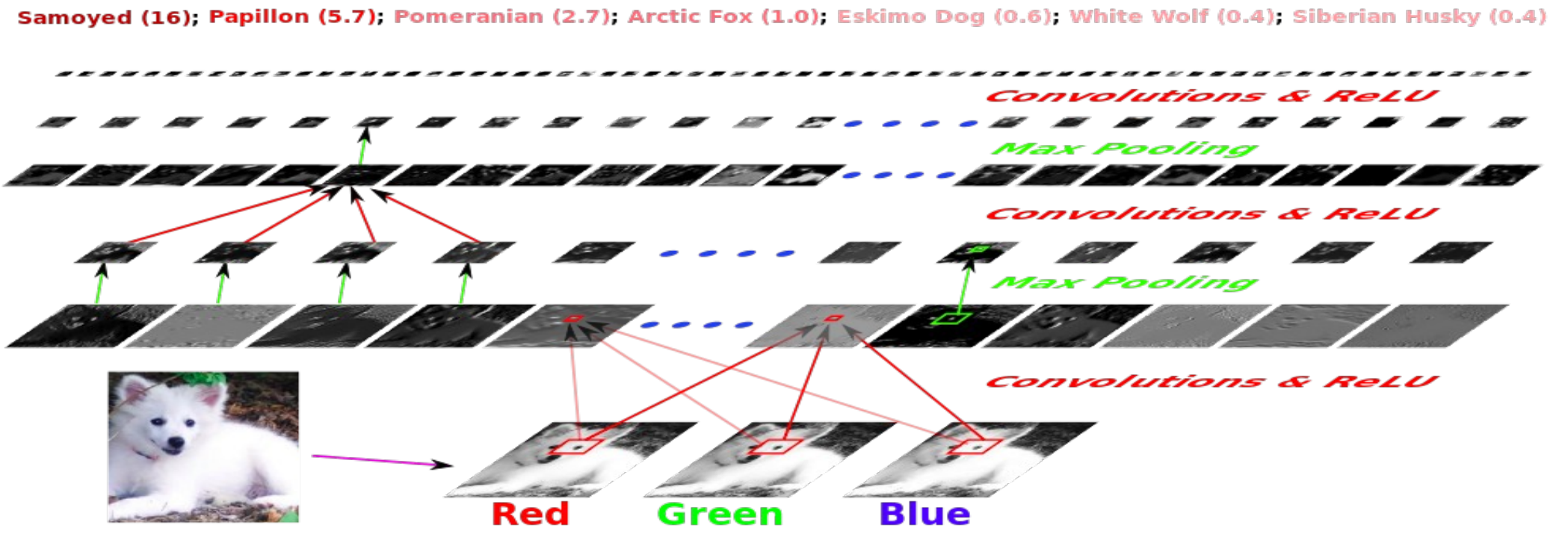
Deep Learning Today

History and State of the Art

facebook
Artificial Intelligence Research

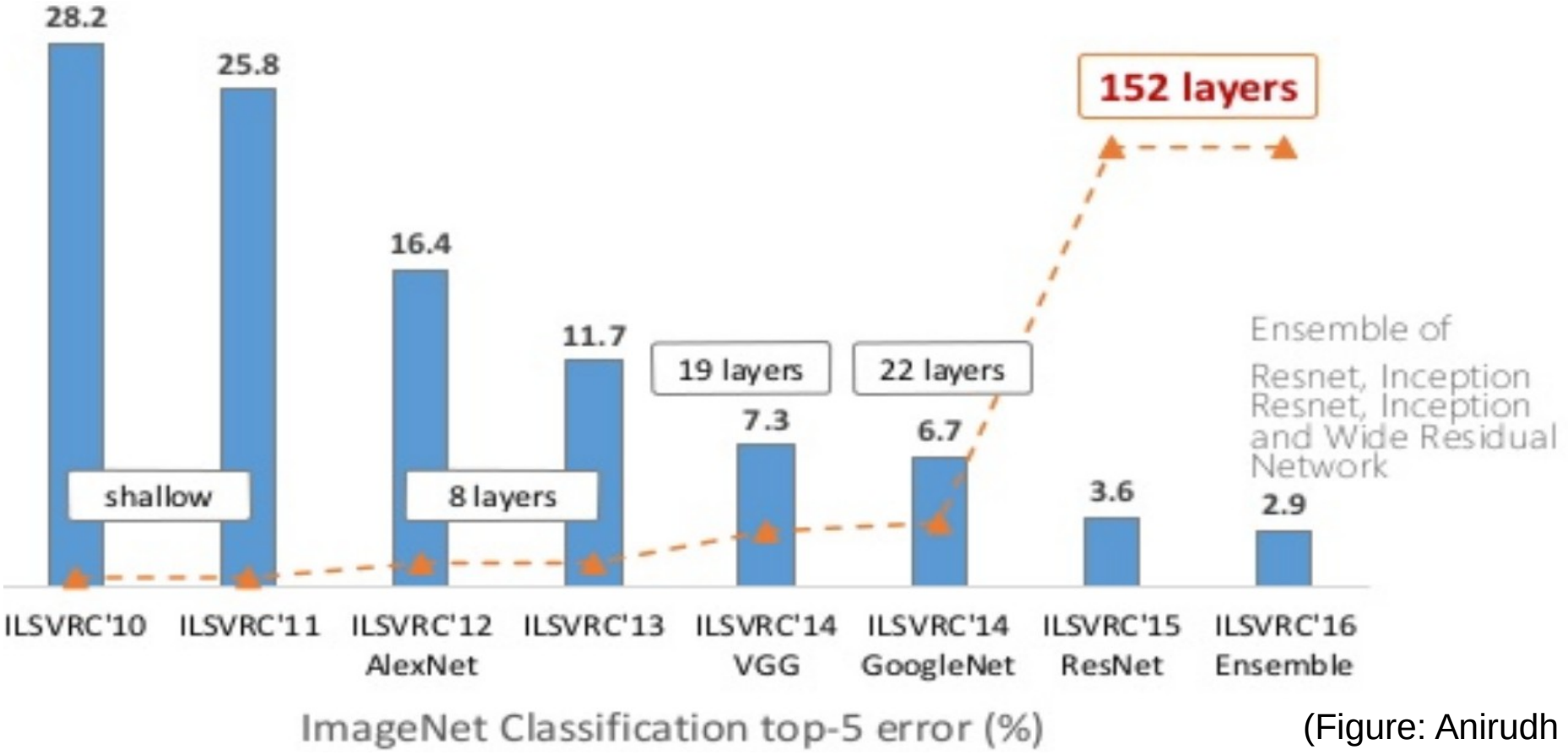
Deep ConvNets for Object Recognition (on GPU)

- ***AlexNet [Krizhevsky et al. NIPS 2012], OverFeat [Sermanet et al. 2013]***
- ***1 to 10 billion connections, 10 million to 1 billion parameters, 8 to 20 layers.***



Error Rate on ImageNet

- ***Depth inflation***

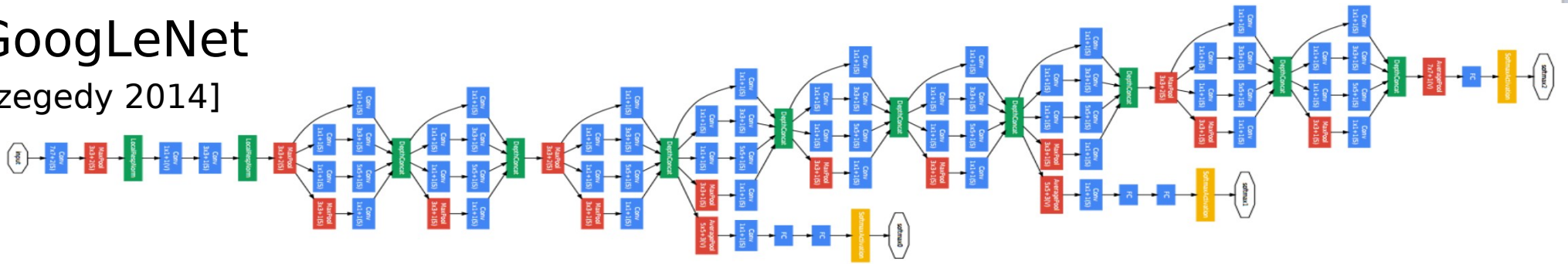


(Figure: Anirudh Koul)

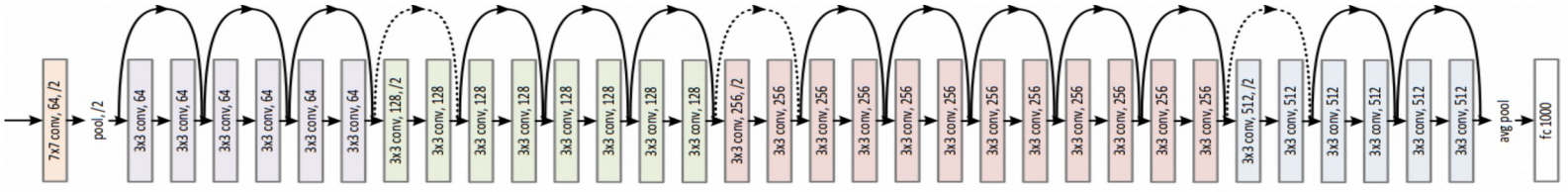
Deep ConvNets (depth inflation)

VGG
[Simonyan 2013]

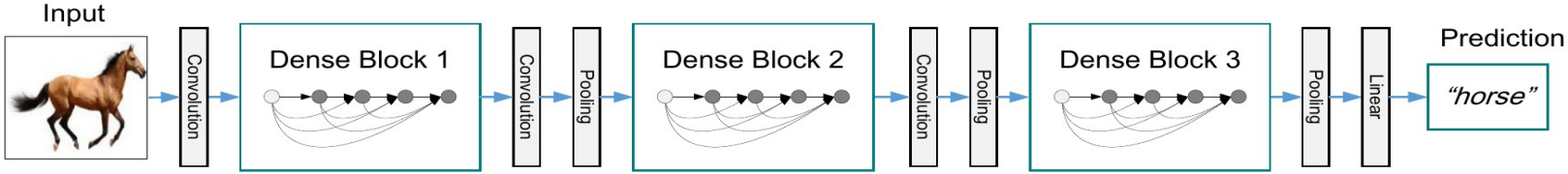
GoogLeNet
Szegedy 2014]



ResNet
[He et al. 2015]

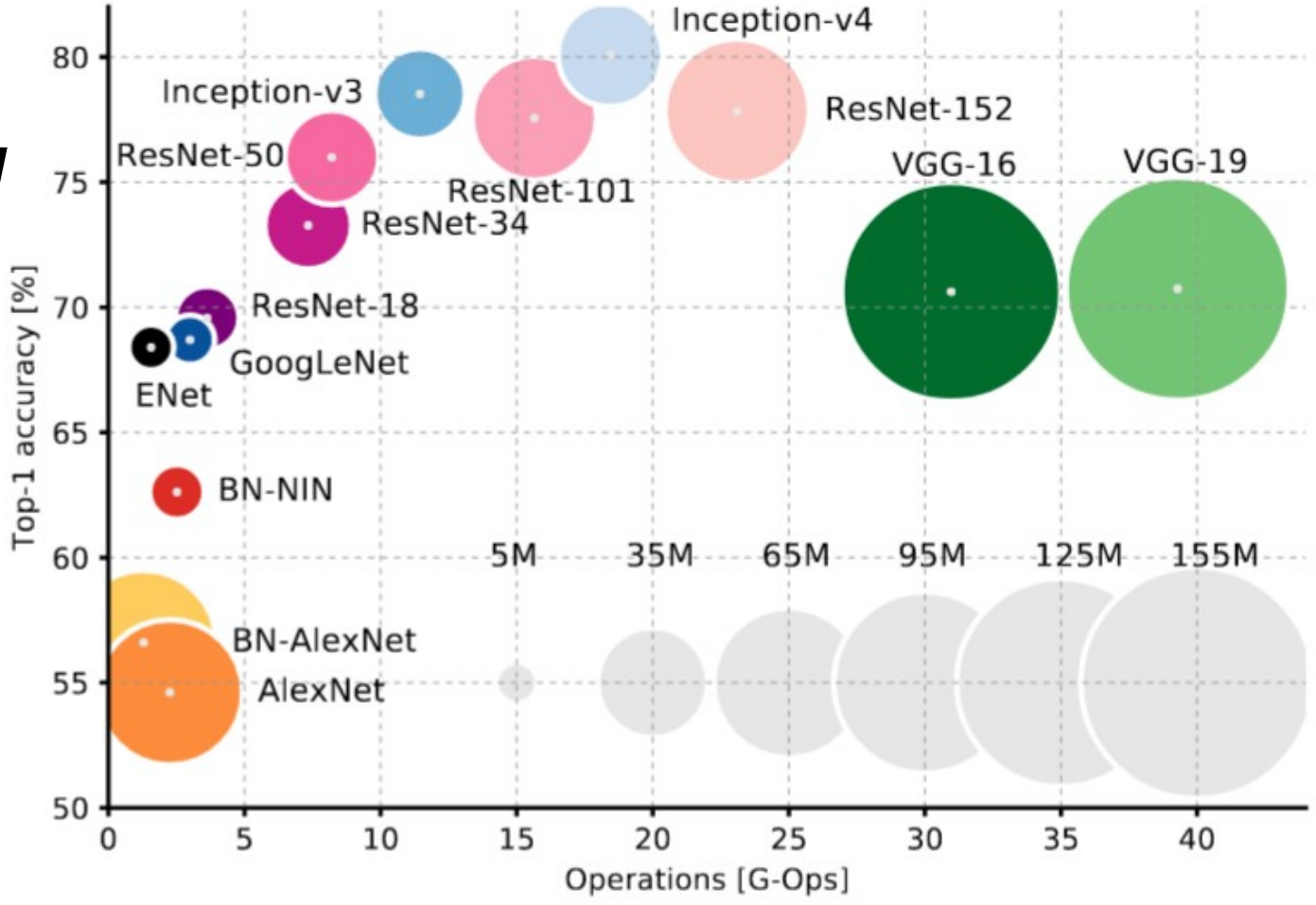


DenseNet
[Huang et al 2017]



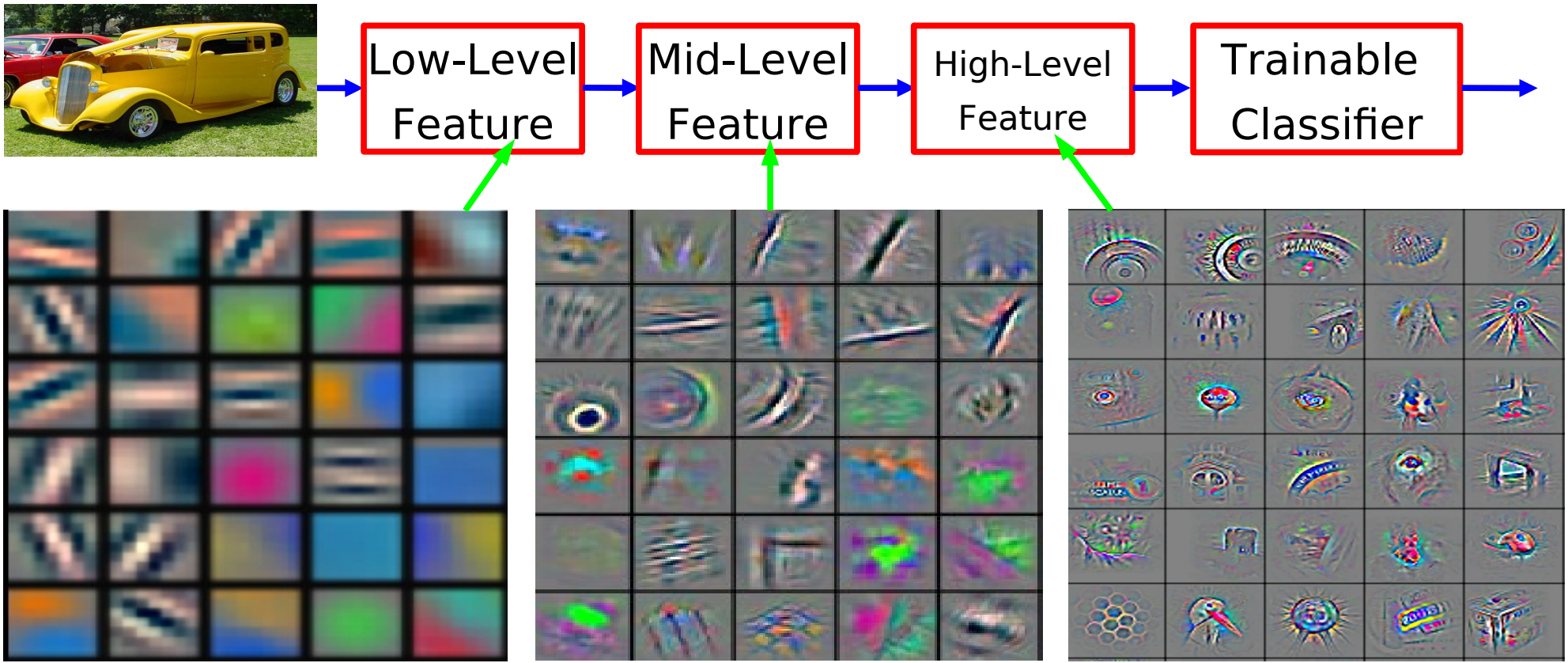
GOPS vs Accuracy on ImageNet vs #Parameters

- ***[Canziani 2016]***
- ***ResNet50 and ResNet 100 are used routinely in production.***



Multilayer Architectures == Compositional Structure of Data

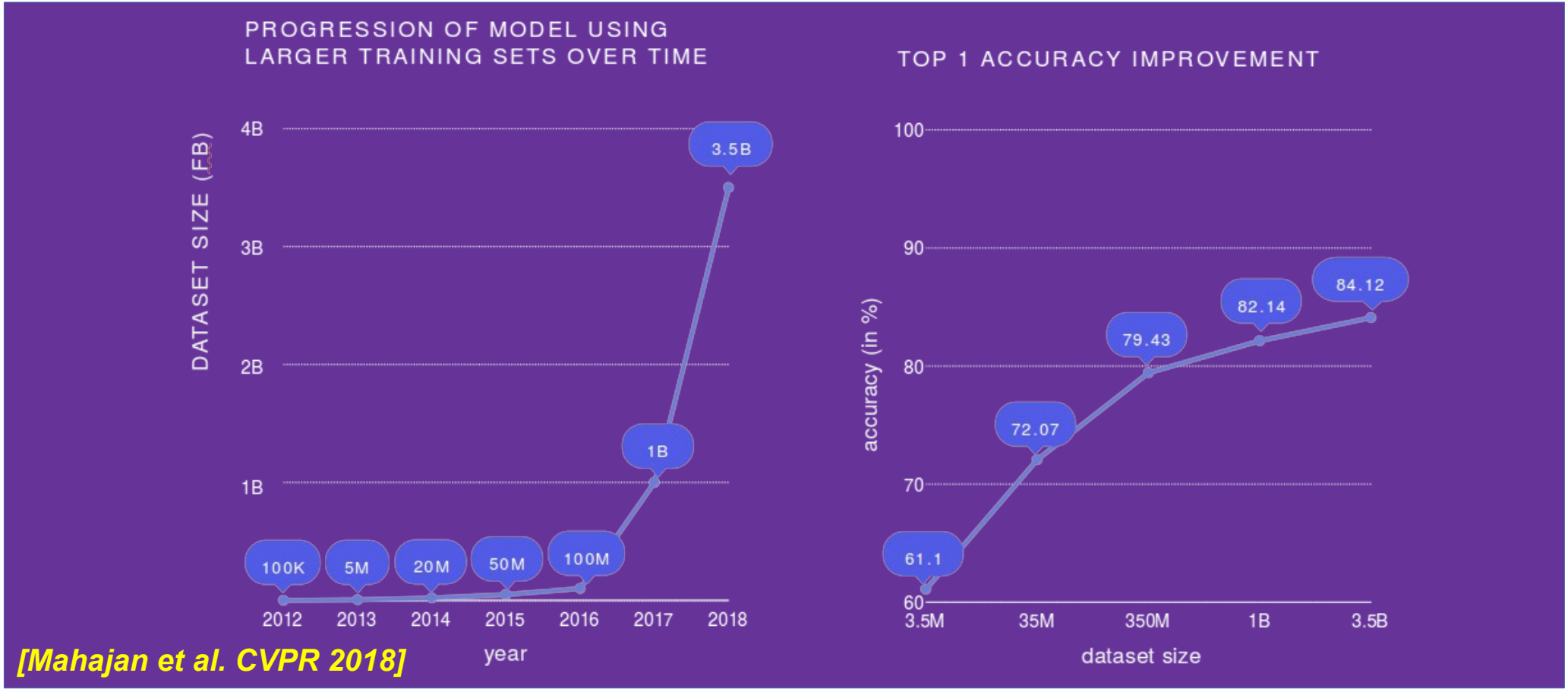
- ***Natural is data is compositional => it is efficiently representable hierarchically***



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Future: weakly/self-supervised learning on massive datasets

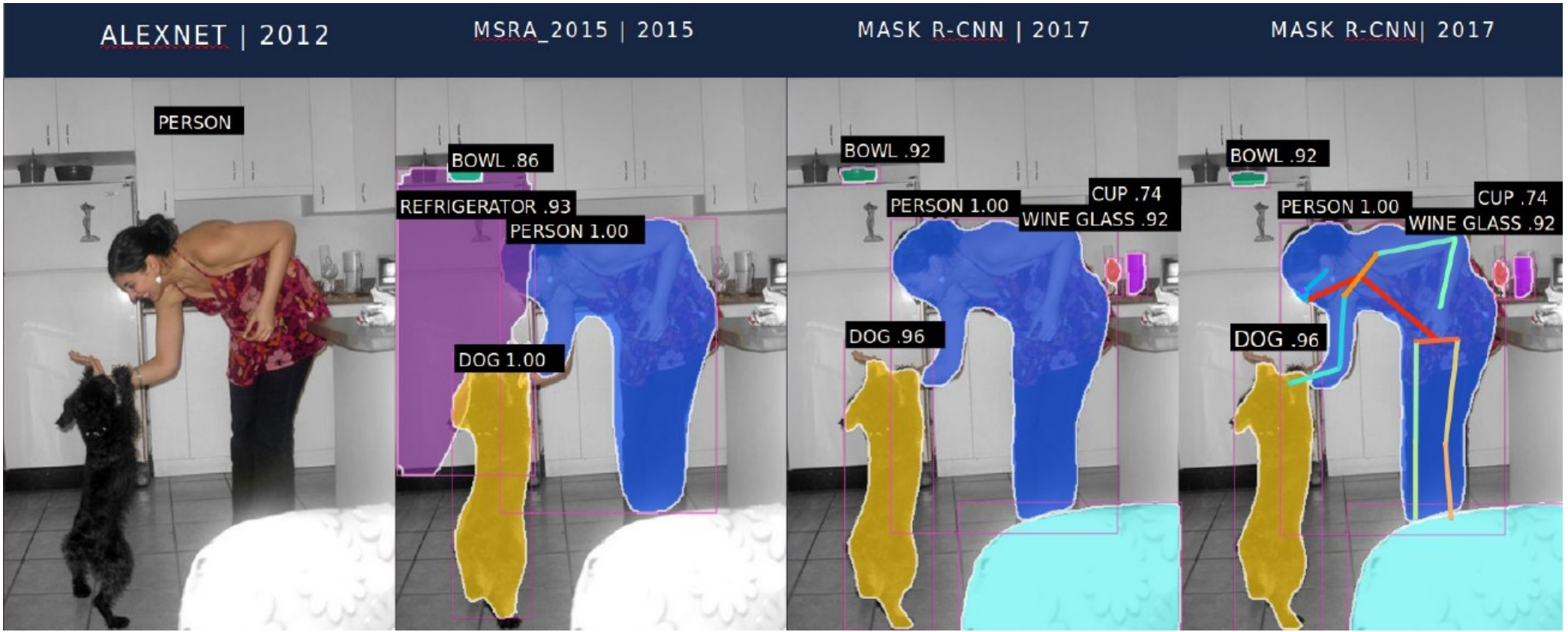
- ***Pretraining on 3.5b instagram images with 17k hashtags. Training/test on ImageNet***



[Mahajan et al. CVPR 2018]

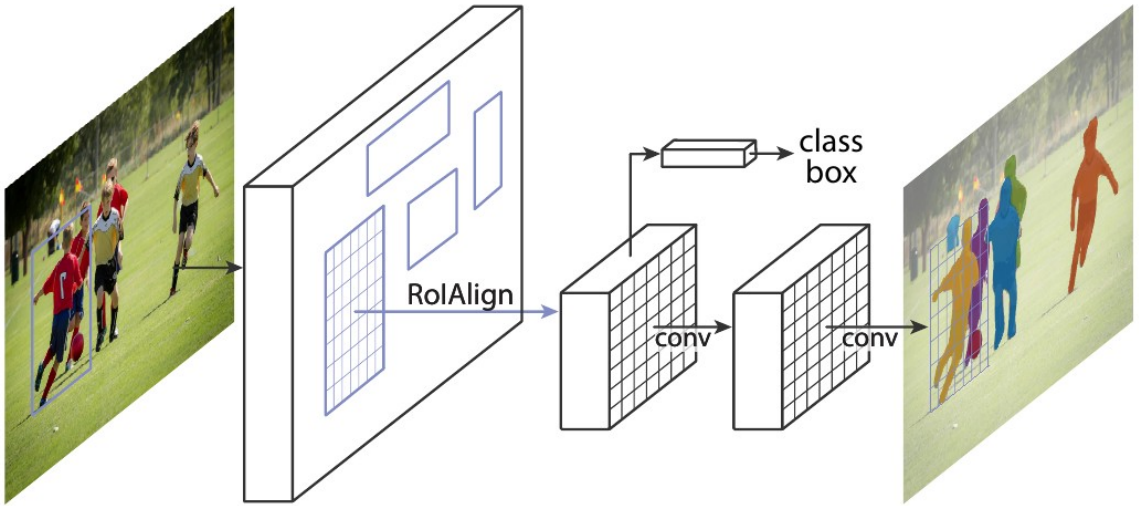
Progress in Computer Vision

- ***[He 2017]***



Mask R-CNN: instance segmentation

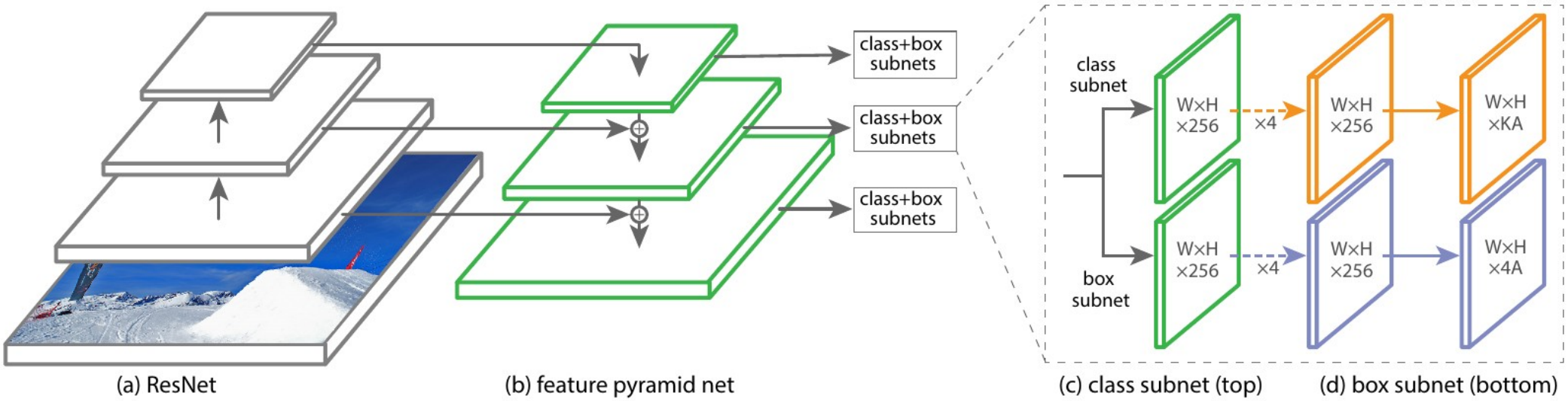
- ***[He, Gkioxari, Dollar, Girshick arXiv:1703.06870]***
- ***ConvNet produces an object mask for each region of interest***
- ***Combined ventral and dorsal pathways***



	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [7]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [20] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [20] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

RetinaNet, feature pyramid network

- ***One-pass object detection***
- ***[Lin et al. ArXiv:1708.02002]***

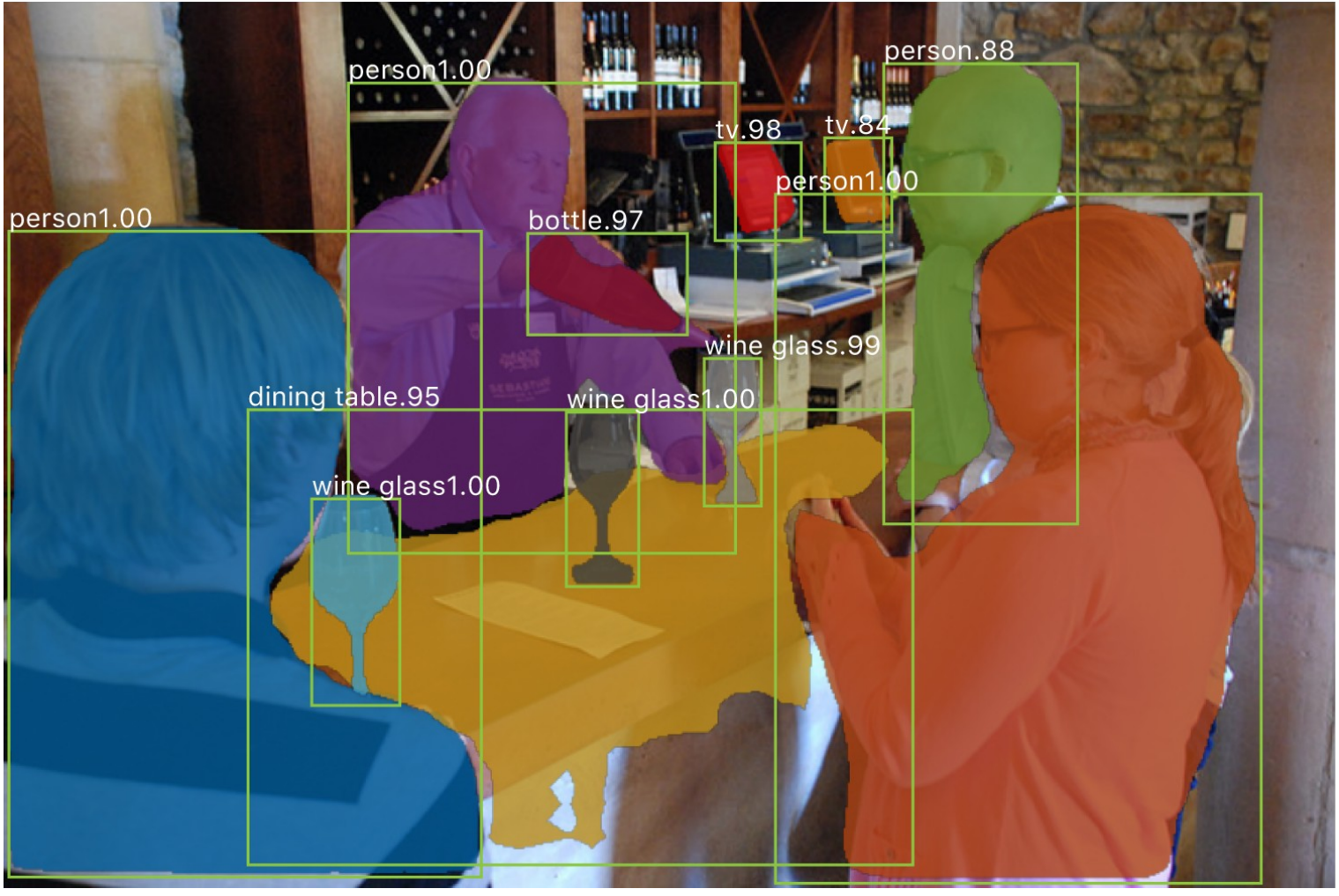


Mask-RCNN Results on COCO dataset

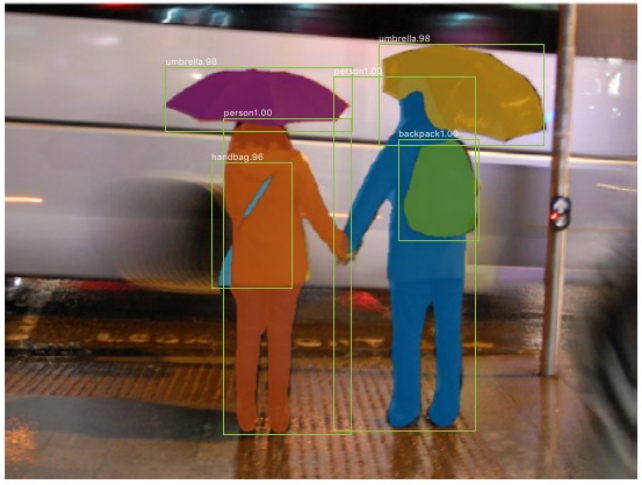
- ***Individual objects are segmented.***

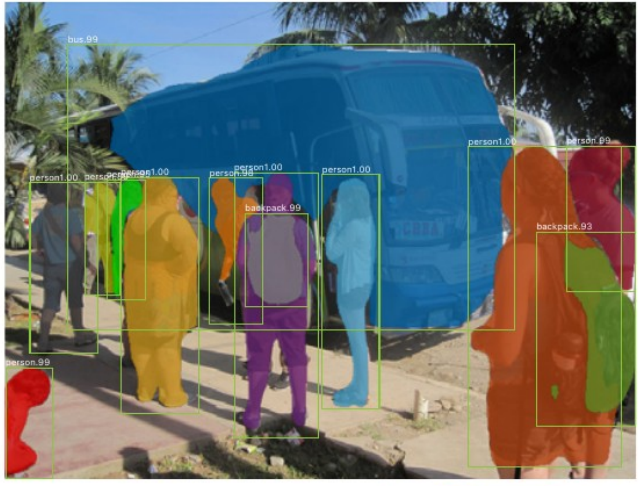
Mask-RCNN Results on COCO dataset

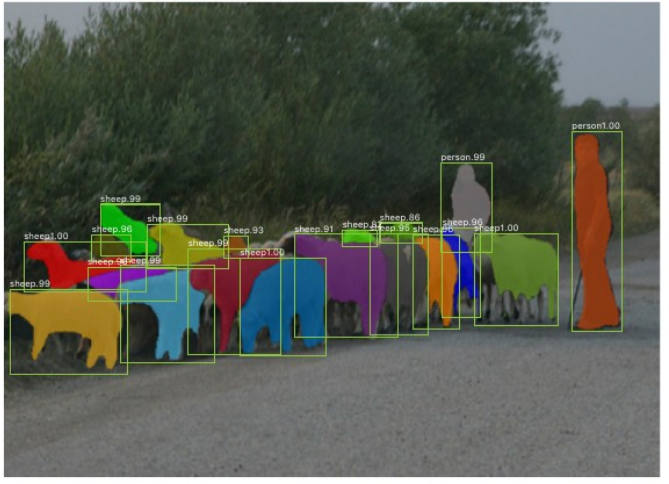
- ***Individual objects are segmented.***

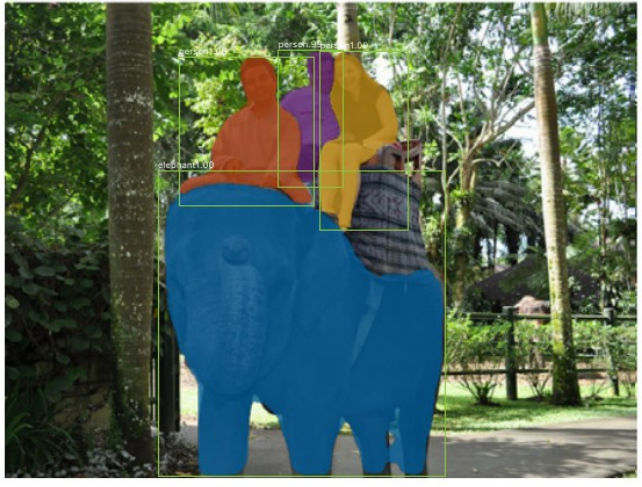


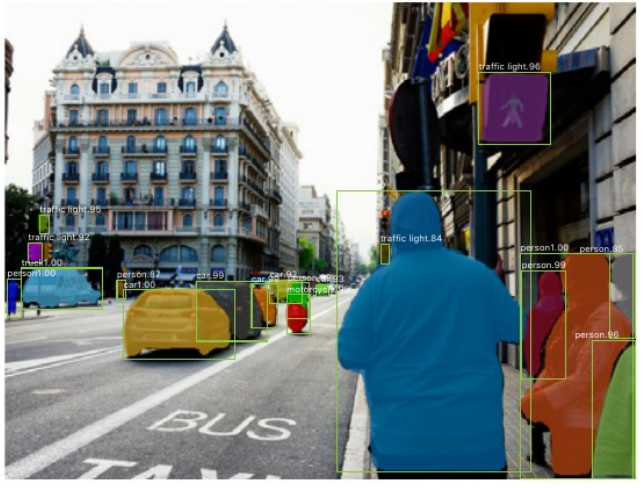
Mask R-CNN Results on COCO test set

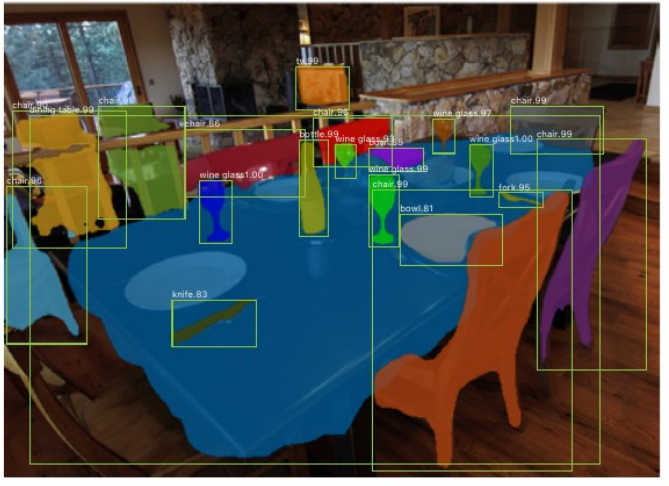










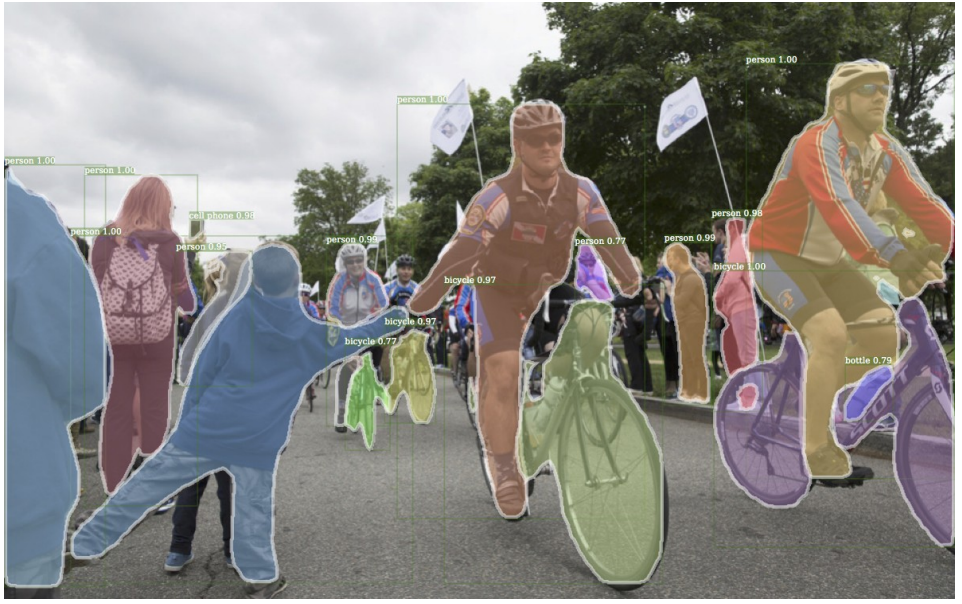


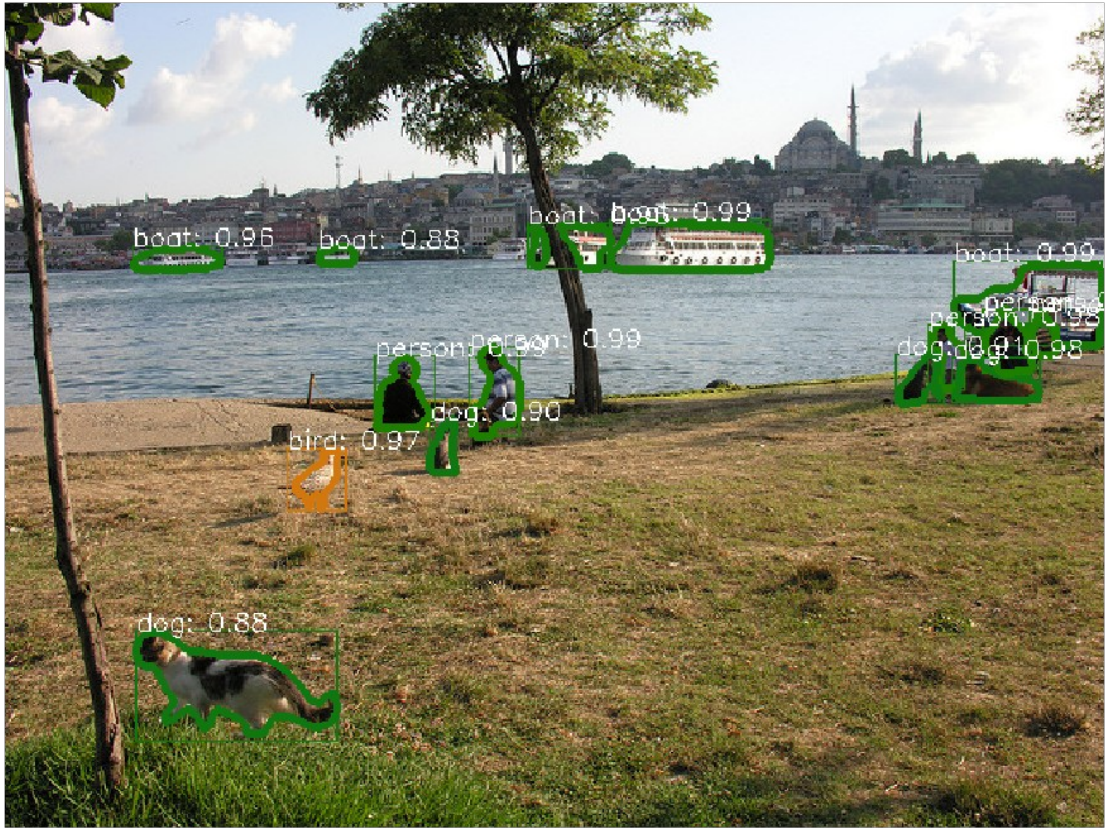
Mask R-CNN Results on COCO test set

Figure 4. More results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

Detectron: open source vision in PyTorch

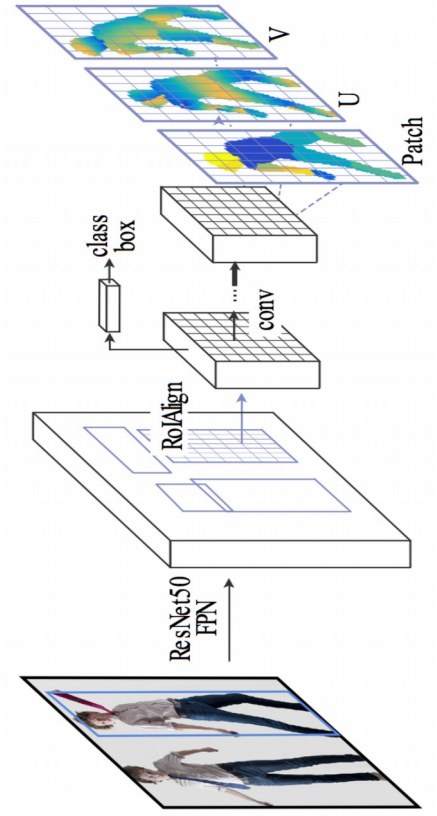
https://github.com/facebookresearch/maskrcnn-benchmark



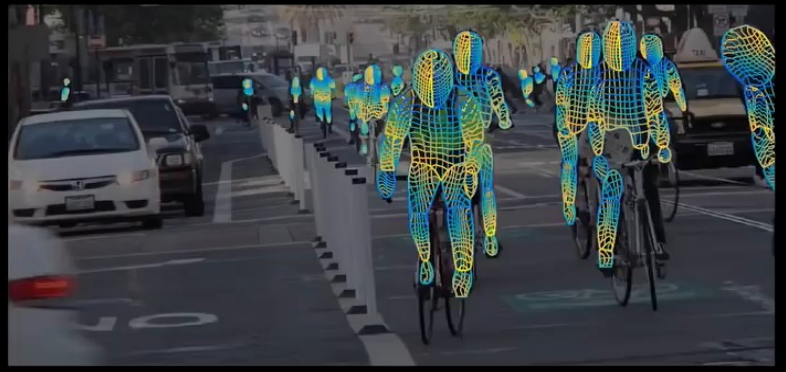


DensePose: real-time body pose estimation

- ***[Guler, Neverova, Kokkinos CVPR 2018]*** ***http://densepose.org***
  - 20 fps on a single GPU



DensePose:
Dense Human Pose Estimation In The Wild



Rıza Alp Güler *
INRIA, CentraleSupélec

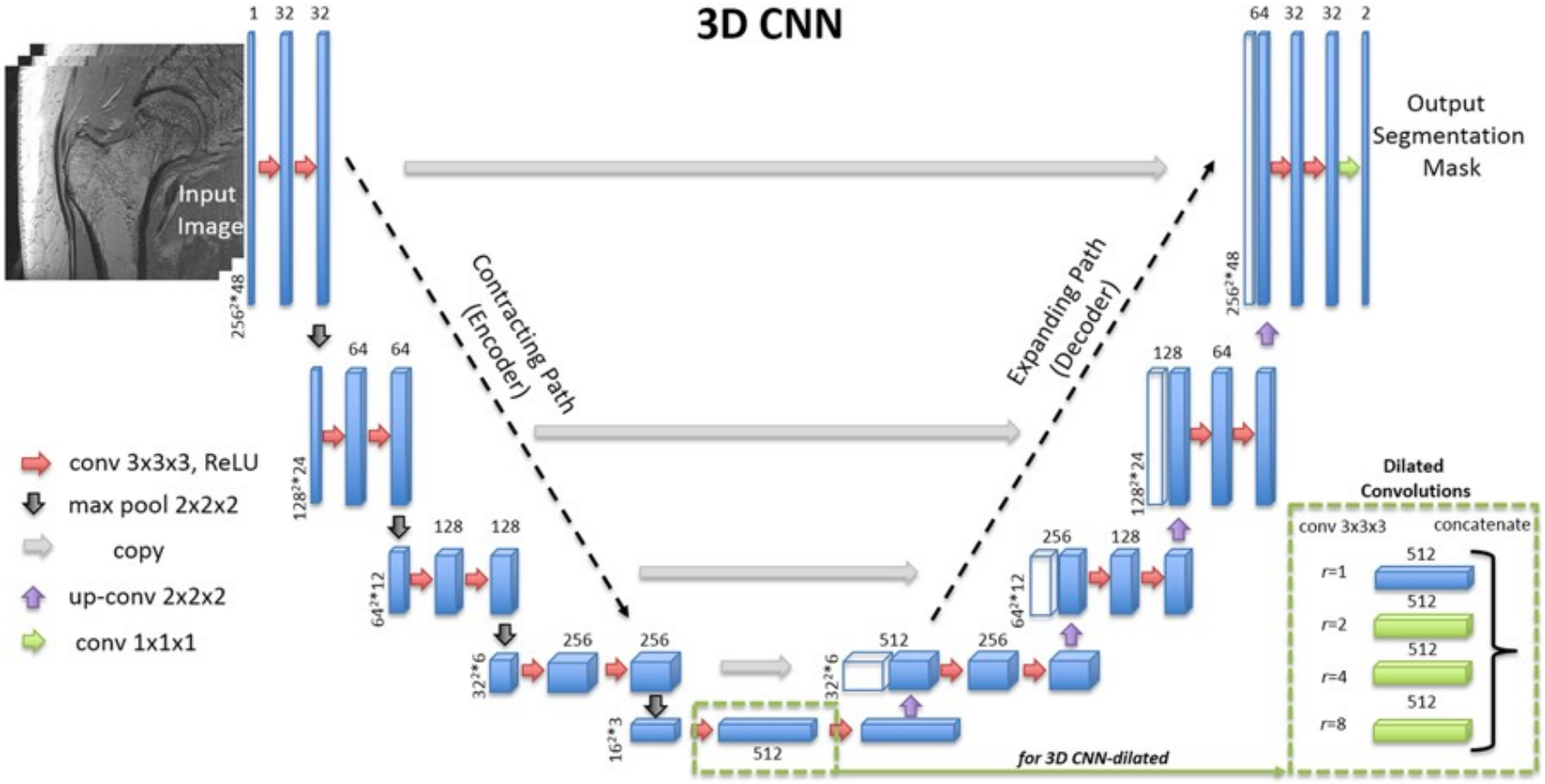
Natalia Neverova
Facebook AI Research

Iasonas Kokkinos
Facebook AI Research

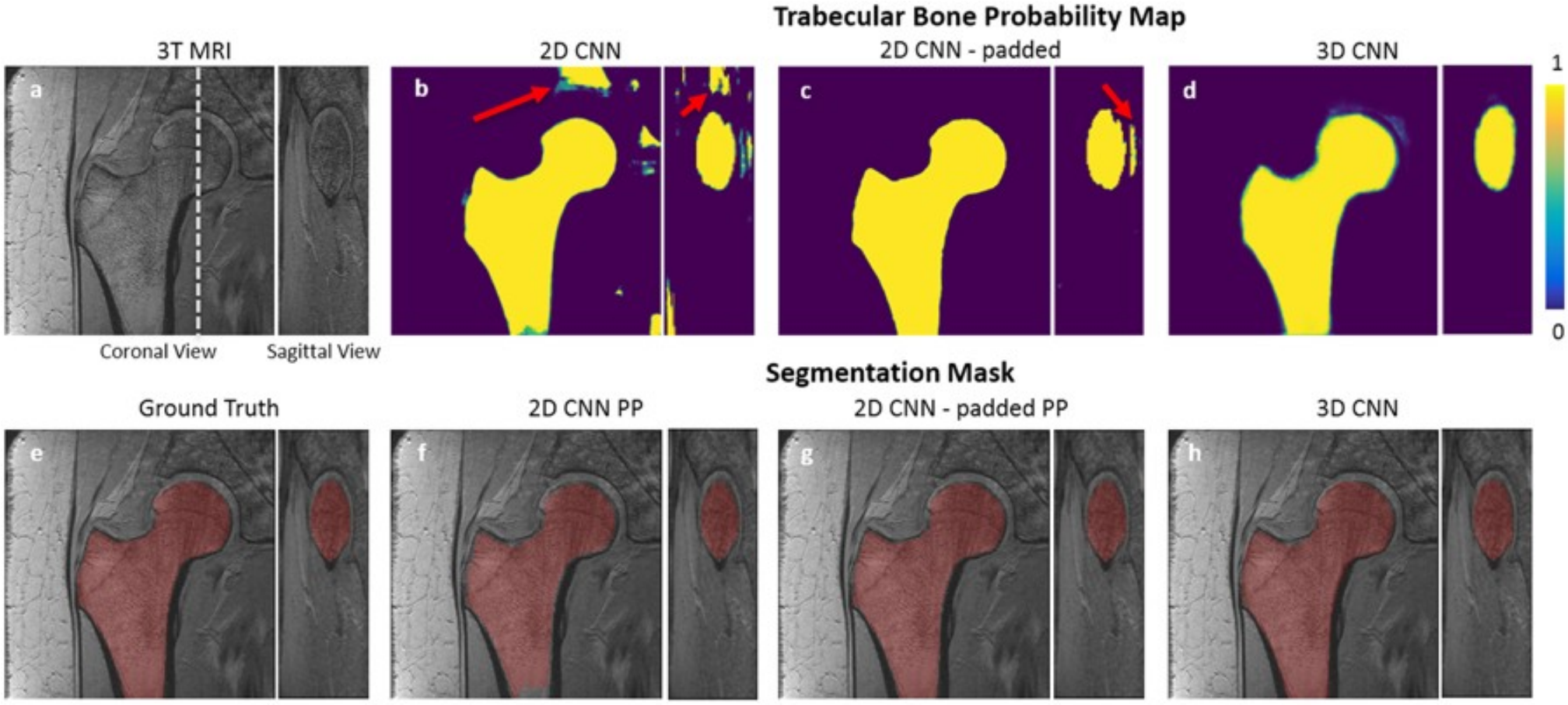
* Rıza Alp Güler was with Facebook AI Research during this work.

3D ConvNet for Medical Image Analysis

- ***Segmentation Femur from MR Images***
- ***[Deniz et al. Nature 2018]***



3D ConvNet for Medical Image Analysis



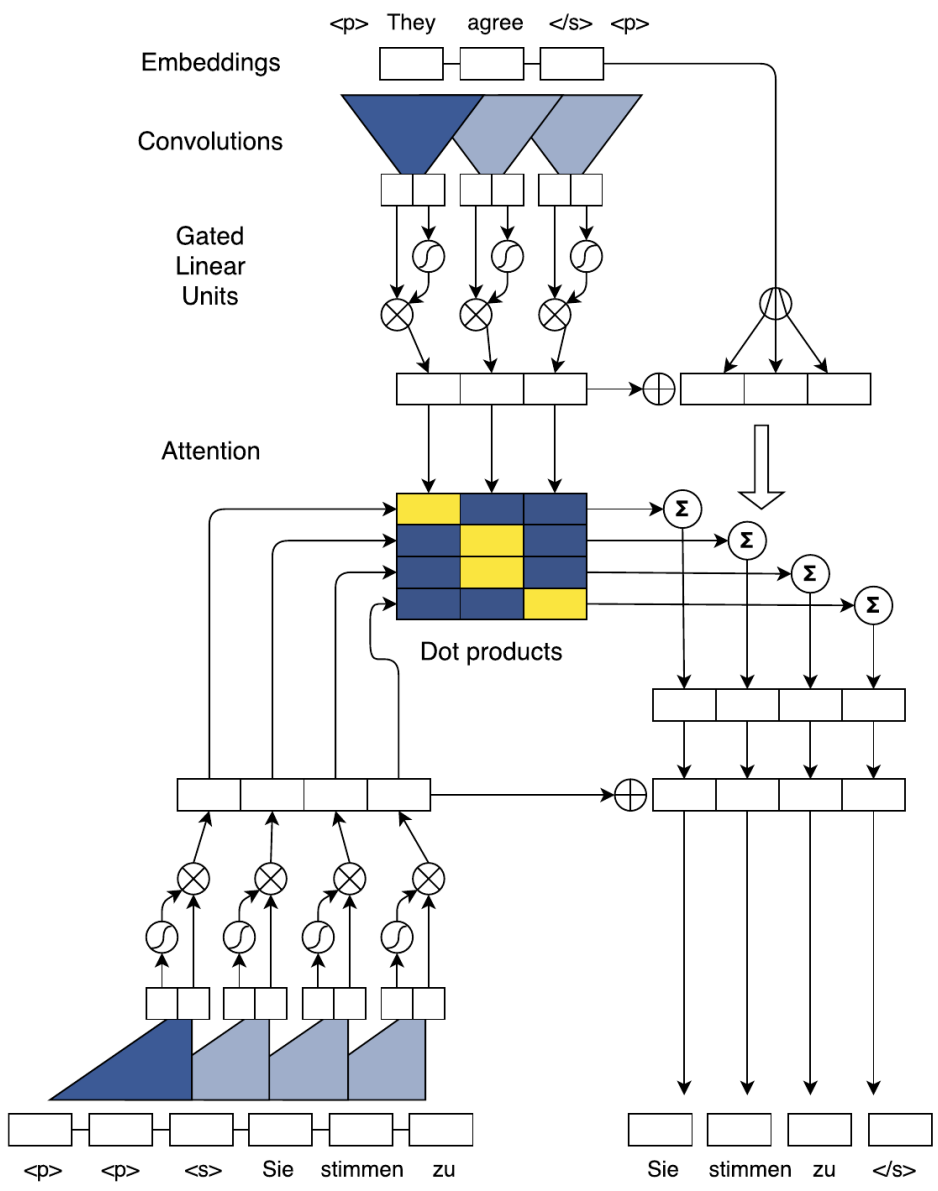
FairSeq for Translation

- ***[Gehring et al. ArXiv:1705.03122]***

WMT'16 English-Romanian	BLEU
Sennrich et al. (2016b) GRU (BPE 90K)	28.1
ConvS2S (Word 80K)	29.45
ConvS2S (BPE 40K)	29.88

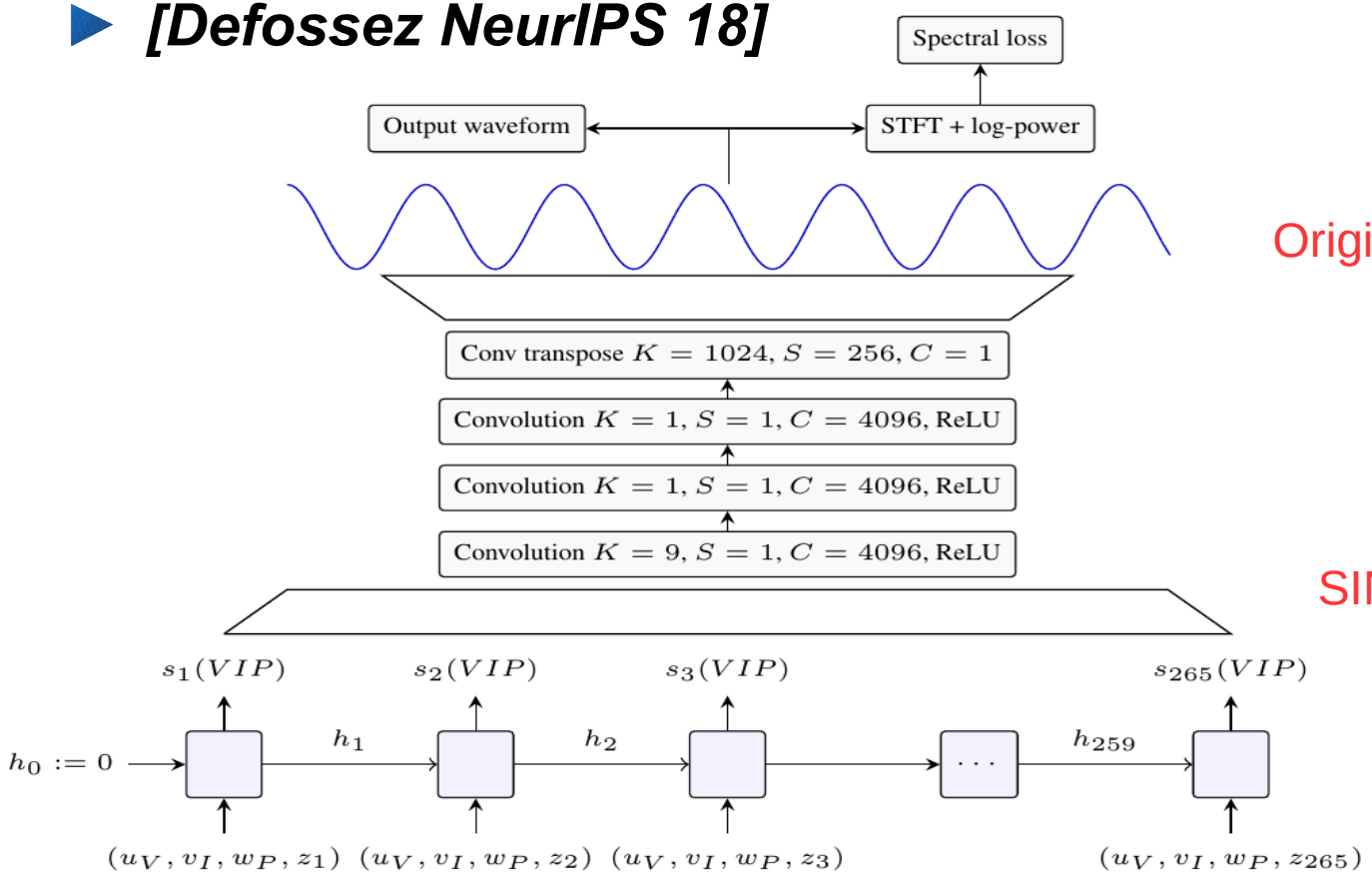
WMT'14 English-German	BLEU
Luong et al. (2015) LSTM (Word 50K)	20.9
Kalchbrenner et al. (2016) ByteNet (Char)	23.75
Wu et al. (2016) GNMT (Word 80K)	23.12
Wu et al. (2016) GNMT (Word pieces)	24.61
ConvS2S (BPE 40K)	25.16

WMT'14 English-French	BLEU
Wu et al. (2016) GNMT (Word 80K)	37.90
Wu et al. (2016) GNMT (Word pieces)	38.95
Wu et al. (2016) GNMT (Word pieces) + RL	39.92
ConvS2S (BPE 40K)	40.46

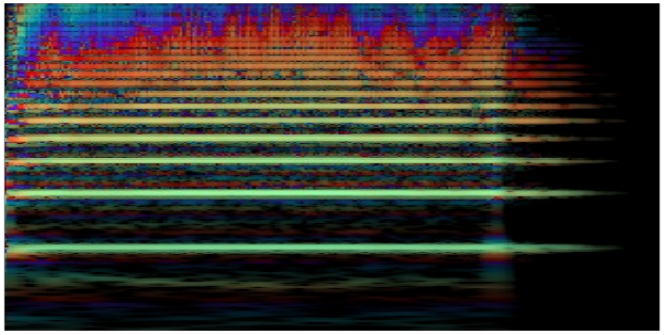


SING: Symbol to Instrument Neural Generator

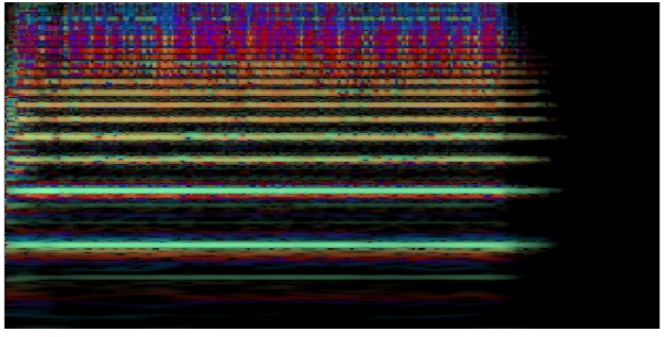
- ***[Defossez NeurIPS 18]***



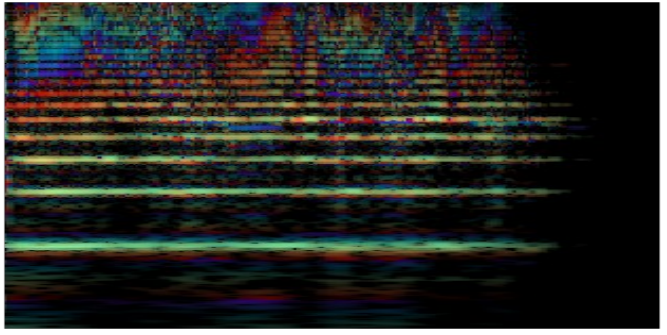
Original



SING



Nsynth

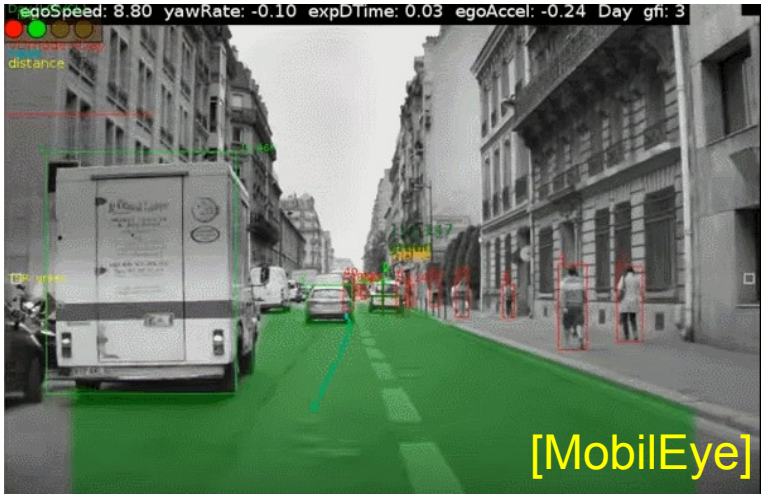


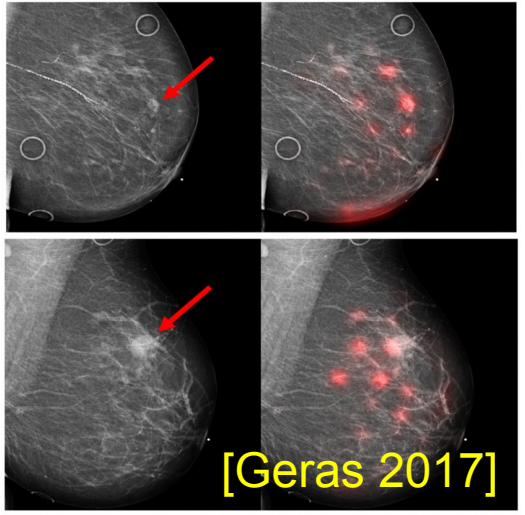
Applications of ConvNets

- ***Self-driving cars, visual perception***
- ***Medical signal and image analysis***
  - Radiology, dermatology, EEG/seizure prediction….
- ***Bioinformatics/genomics***
- ***Speech recognition***
- ***Language translation***
- ***Image restoration/manipulation/style transfer***
- ***Robotics, manipulation***
- ***Physics***
  - High-energy physics, astrophysics
- ***New applications appear every day***
  - E.g. environmental protection,….

Applications of Deep Learning

- ***Medical image analysis***
- ***Self-driving cars***
- ***Accessibility***
- ***Face recognition***
- ***Language translation***
- ***Virtual assistants****
- ***Content Understanding for:***
  - Filtering
  - Selection/ranking
  - Search
- ***Games***
- ***Security, anomaly detection***
- ***Diagnosis, prediction***
- ***Science!***





Open Source Projects from FAIR

- ***PyTorch: deep learning framework*** ***http://pytorch.org***
  - Many examples and tutorials. Used by many research groups.
- ***FAISS: fast similarity search (C++/CUDA)***
- ***ParlAI: training environment for dialog systems (Python)***
- ***ELF: distributed reinforcement learning framework***

- ***ELF OpenGo: super-human go-playing engine***
- ***FastText: text classification, representation, embedding (C++)***
- ***FairSeq: neural machine translation with ConvNets, RNN...***
- ***Detectron / Mask-R-CNN: complete vision system***
- ***DensePose: real-time body pose tracking system***
- ***https://github.com/facebookresearch***

Lessons learned #2

- ***2.1: Good results are not enough***
  - *Making them easily reproducible also makes them credible.*
- ***2.2: Hardware progress enables new breakthroughs***
  - *General-Purpose GPUs should have come 10 years earlier!*
- ***2.3: Open-source software platforms disseminate ideas***
  - *But making platforms that are good for research and production is hard.*
- ***2.4: Convolutional Nets will soon be everywhere***
  - *Hardware should exploit the properties of convolutions better*
  - *There is a need for low-cost, low-power ConvNet accelerators*
  - *Cars, cameras, vacuum cleaners, lawn mowers, toys, maintenance robots...*

What About (Deep) Reinforcement Learning?

It works great …

…for games and virtual environments

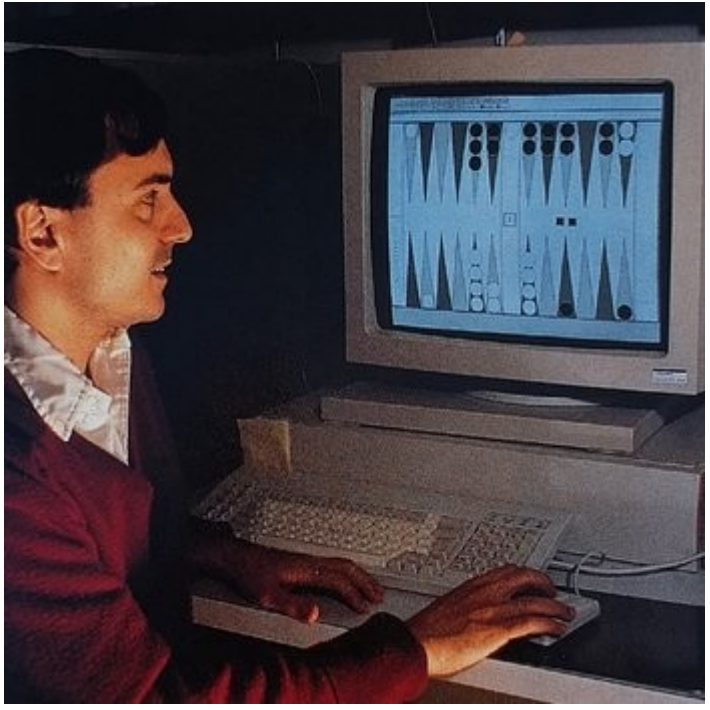
facebook
Artificial Intelligence Research

Reinforcement Learning

- ***Telling the whether it is right or wrong***
  - Not giving it the right answer

- ***Gerry Tesauro at IBM ----------→***
  - TD-Gammon: backgammon bot trained with Renforcement Learning
  - Used a neural net to evaluate positions

- ***Around 1995, the RL community abandoned neural nets***
  - Studying “provable” models → No neural nets



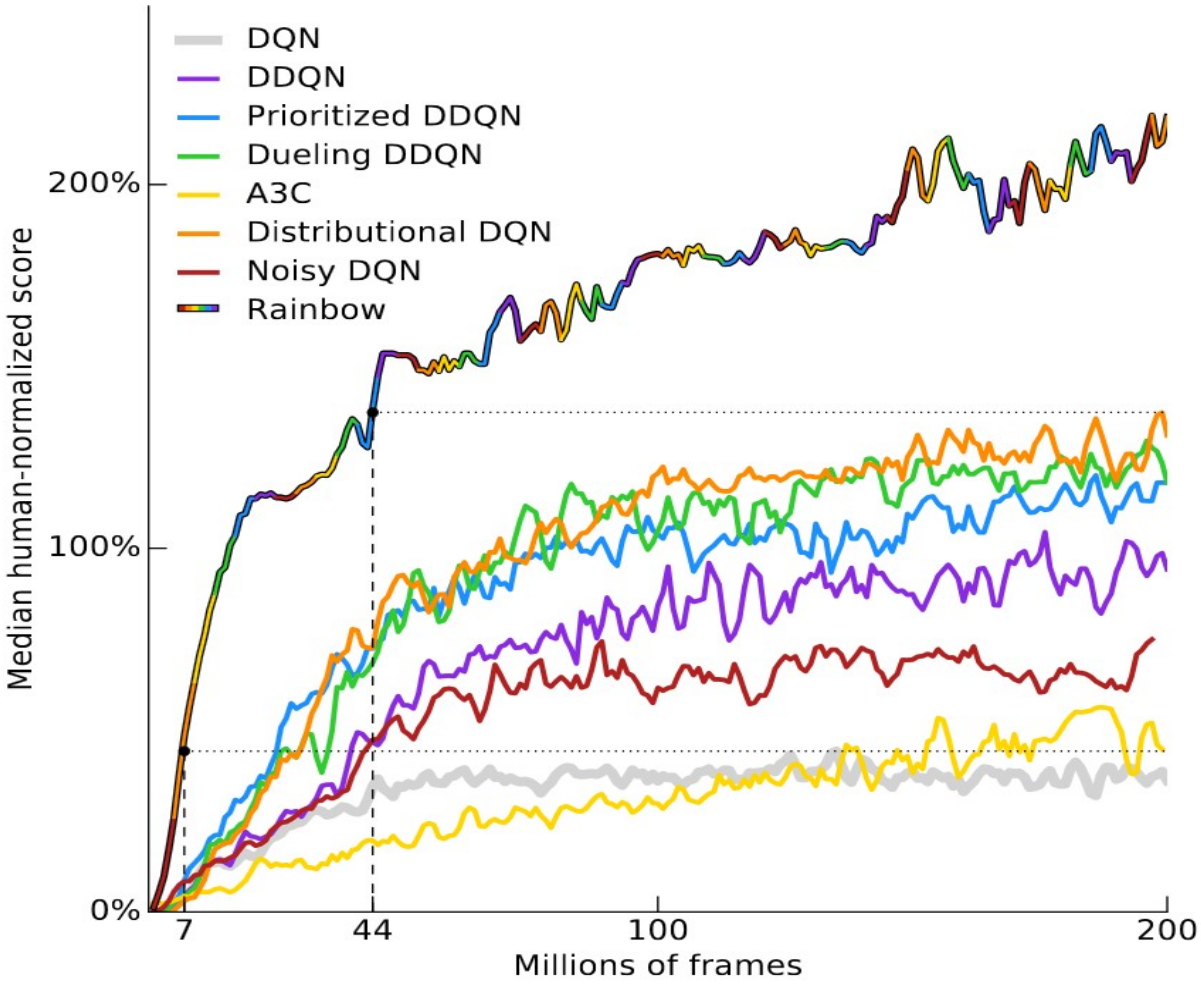
Reinforcement Learning works great for games

- ***RL works well for games***
  - Playing Atari games [Mnih 2013], Go [Silver 2016, Tian 2018], Doom [Tian 2017], StarCraft (work in progress at FAIR, DeepMind….)
  - RL requires too many trials.
  - RL often doesn't really work in the real world
- ***FAIR open Source go player: OpenGo***

  https://github.com/pytorch/elf

Pure RL requires many, many trials to learn a task

- *[Hessel ArXiv:1710.02298]*
- *Median performance on 57 Atari games relative to human performance (100%=human)*
- *Most methods require over 50 million frames to match human performance (230 hours of play)*
- *The best method (combination) takes 18 million frames (83 hours).*



Pure RL is hard to use in the real world

- ***Pure RL requires too many trials to learn anything***
  - it's OK in a game
  - it's not OK in the real world
- ***RL works in simple virtual world that you can run faster than real-time on many machines in parallel.***

- ***Anything you do in the real world can kill you***
- ***You can't run the real world faster than real time***

Controversial Proposition #2

- ***Reinforcement Learning fell victim to the theoretical street light effect when it moved away from model-based methods in the mid 1990s***
  - Convergence proofs for table-based model-free RL (cute math)
  - Bandits

What are we missing?

To get to “real” AI

1. Reasoning
2. Learning models of the world
3. Learning hierarchical representations of actions

What current deep learning methods enables

- ***What we can have***
  - Safer cars, autonomous cars
  - Better medical image analysis
  - Personalized medicine
  - Adequate language translation
  - Useful but stupid chatbots
  - Information search, retrieval, filtering
  - Numerous applications in energy, finance, manufacturing, environmental protection, commerce, law, artistic creation, games,…..
- ***What we cannot have (yet)***
  - Machines with common sense
  - Intelligent personal assistants
  - “Smart” chatbots”
  - Household robots
  - Agile and dexterous robots
  - Artificial General Intelligence (AGI)

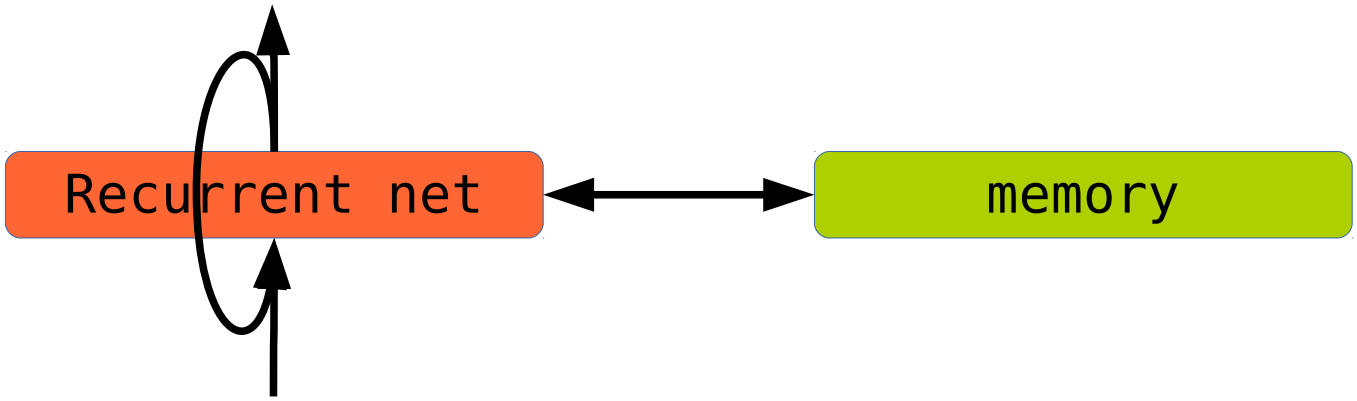
Differentiable Programming: Marrying Deep Learning With Reasoning

Neural nets with dynamic, data-dependent structure, A program whose gradient is generated automatically.

facebook
Artificial Intelligence Research

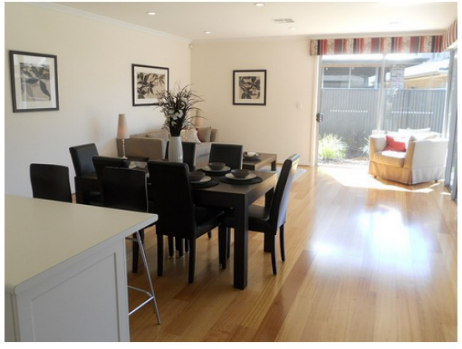
Augmenting Neural Nets with a Memory Module

- ***Recurrent networks cannot remember things for very long***
  - The cortex only remember things for 20 seconds
- ***We need a "hippocampus" (a separate memory module)***
  - LSTM [Hochreiter 1997], registers
  - **Memory networks** [Weston et 2014] (FAIR), associative memory
  - **Stacked-Augmented Recurrent Neural Net** [Joulin & Mikolov 2014] (FAIR)
  - **Neural Turing Machine** [Graves 2014],
  - **Differentiable Neural Computer** [Graves 2016]



Answering complex questions by running a program

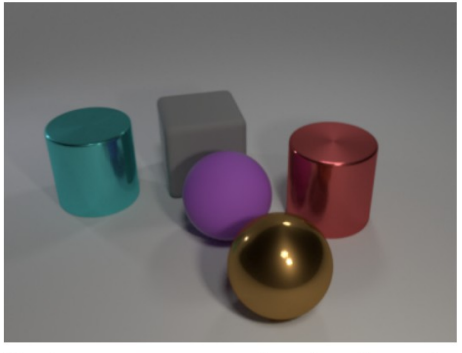
- ***[Johnson et al. ArXiv:1705.03633]***



How many chairs are at the table?

Is there a pedestrian in my lane?

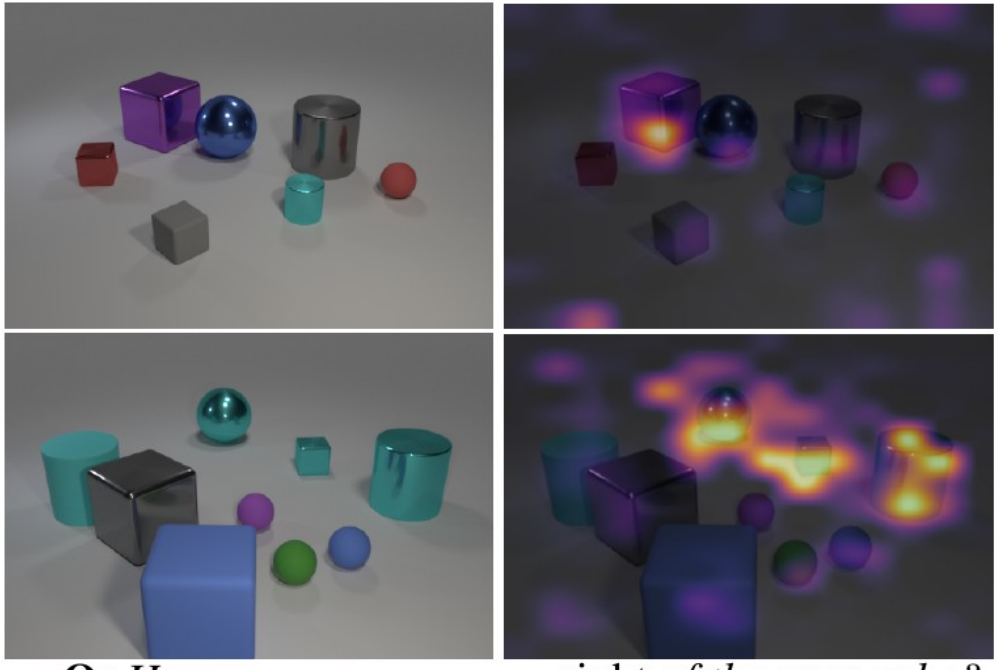
Is the person with the blue hat touching the bike in the back?



Is there a matte cube that has the same size as the red metal object?

Q: *What shape is the…* *…purple thing?*

A: *cube*



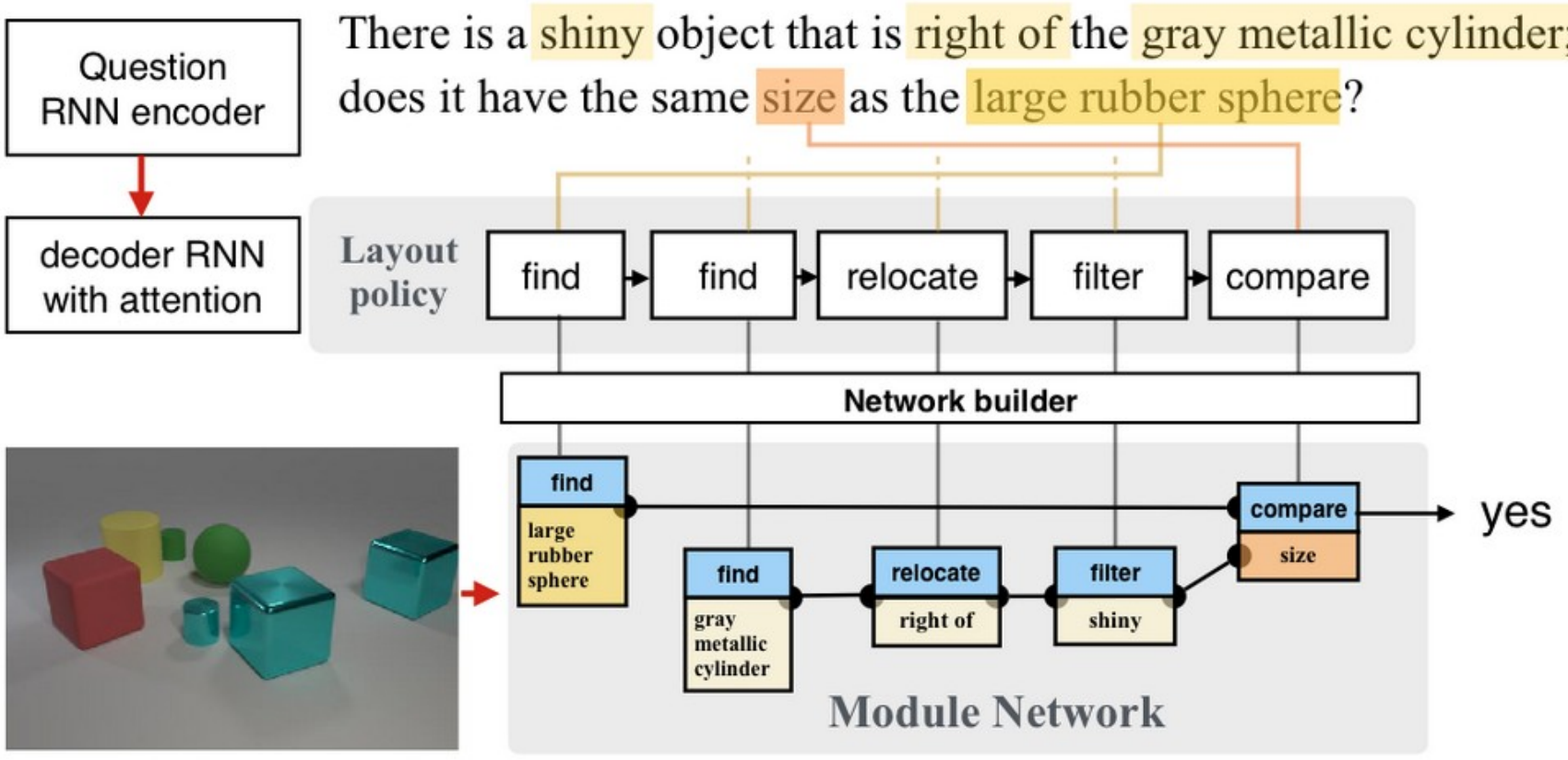
Q: *How many cyan things are…*

…right of the gray cube?

A: *3*

Inferring and executing programs for visual reasoning

https://research.fb.com/visual-reasoning-and-dialog-towards-natural-language-conversations-about-visual-data/



PyTorch: differentiable programming

- ***Software 2.0:***
  - The operations in a program are only partially specified
  - They are trainable parameterized modules.
  - The precise operations are learned from data, only the general structure of the program is designed.
- ***Dynamic computational graph***
  - Automatic differentiation by recording a “tape” of operations and rolling it backwards with the Jacobian of each operator.
  - Implemented in PyTorch1.0, Chainer…
  - Easy if the front-end language is dynamic and interpreted (e.g Python)
  - Not so easy if we want to run without a Python runtime...

How do Humans and Animal Learn?

So quickly

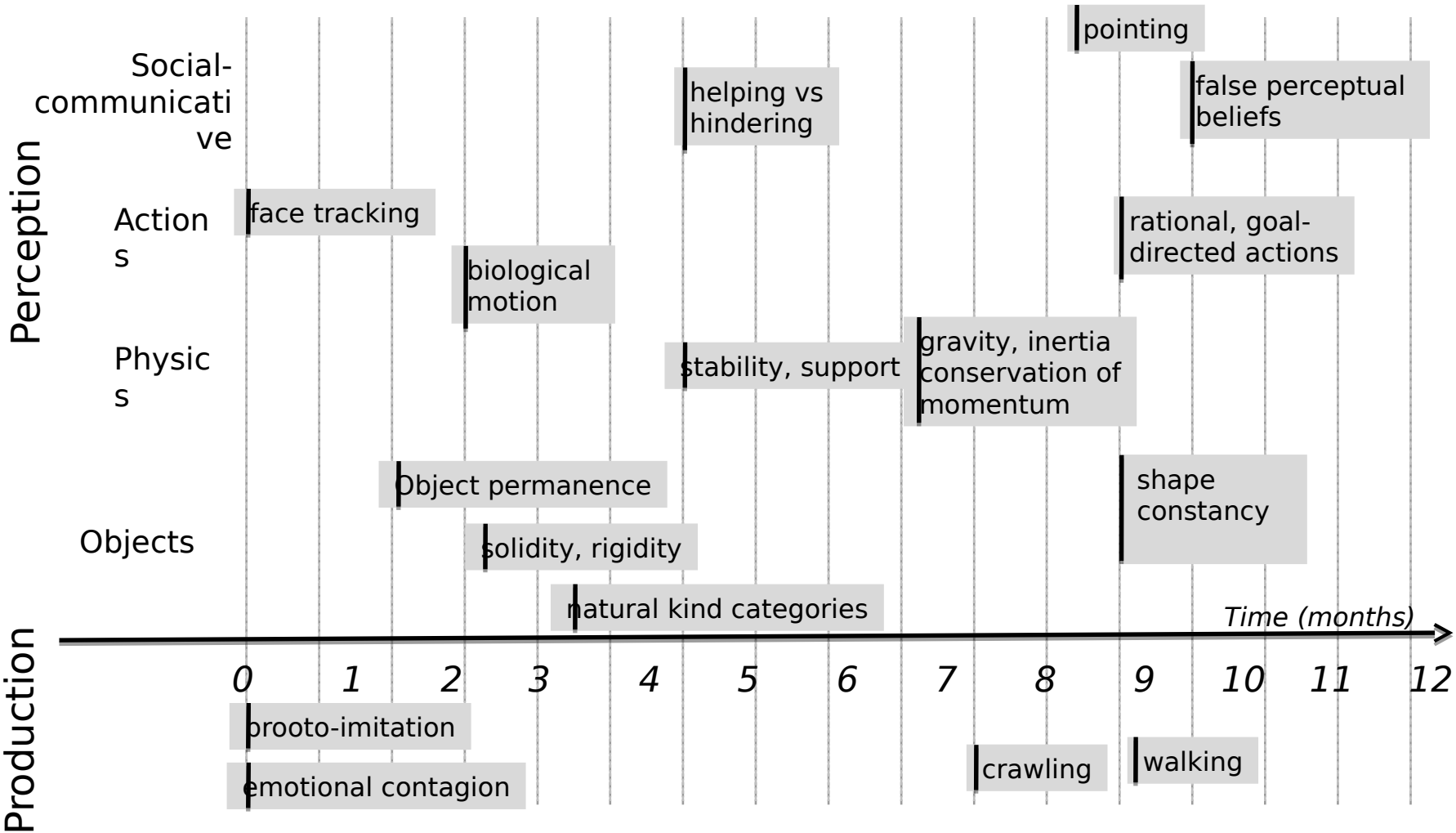
facebook
Artificial Intelligence Research

Babies learn how the world works by observation

- ***Largely by observation, with remarkably little interaction.***

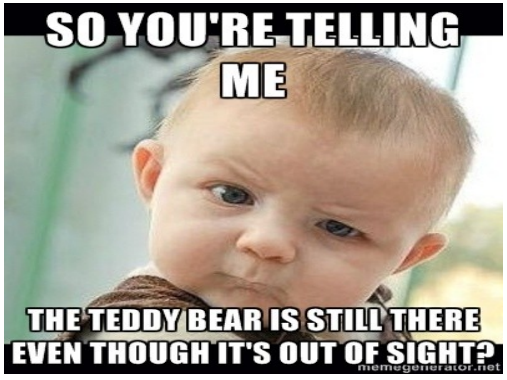
Photos courtesy of Emmanuel Dupoux

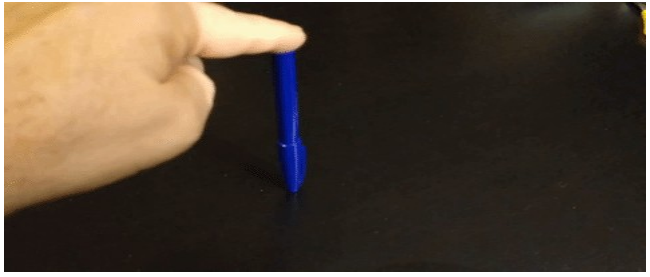
Early Conceptual Acquisition in Infants [from Emmanuel Dupoux]

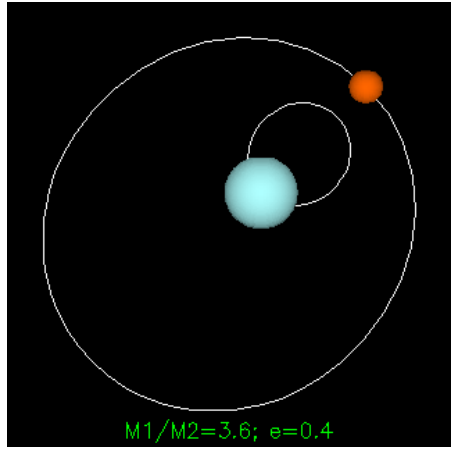


Prediction is the essence of Intelligence

- ***We learn models of the world by predicting***







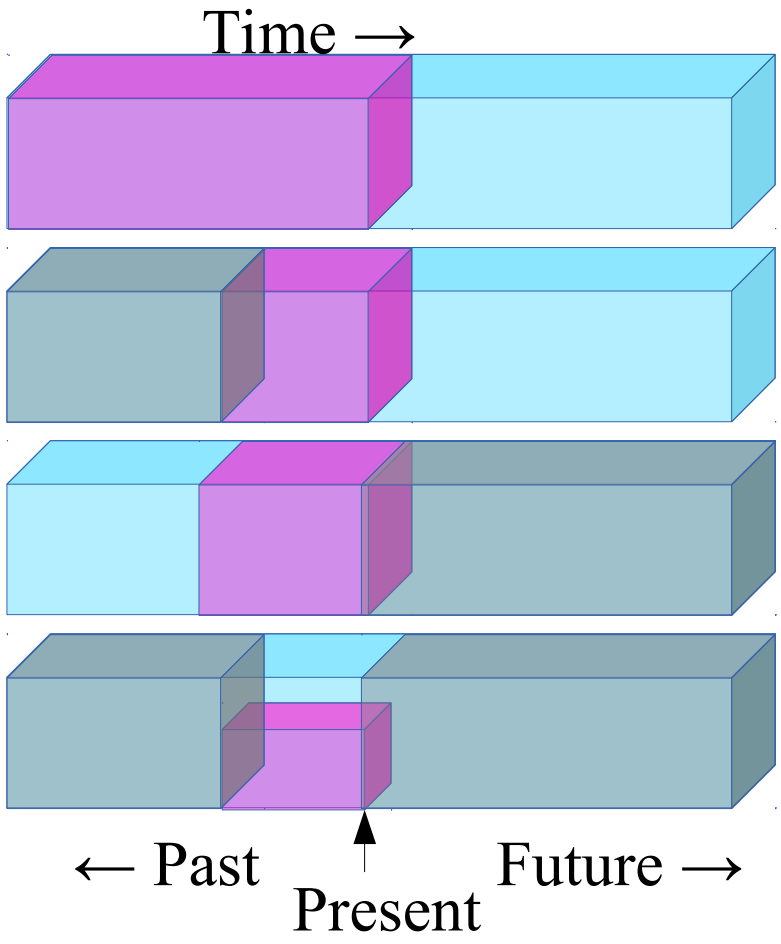
The Future: Self-Supervised Learning

With massive amounts of data and very large networks

facebook
Artificial Intelligence Research

Self-Supervised Learning

- ***Predict any part of the input from any other part.***
- ***Predict the future from the past.***
- ***Predict the future from the recent past.***
- ***Predict the past from the present.***
- ***Predict the top from the bottom.***
- ***Predict the occluded from the visible***
- ***Pretend there is a part of the input you don't know and predict that.***



Three Types of Learning

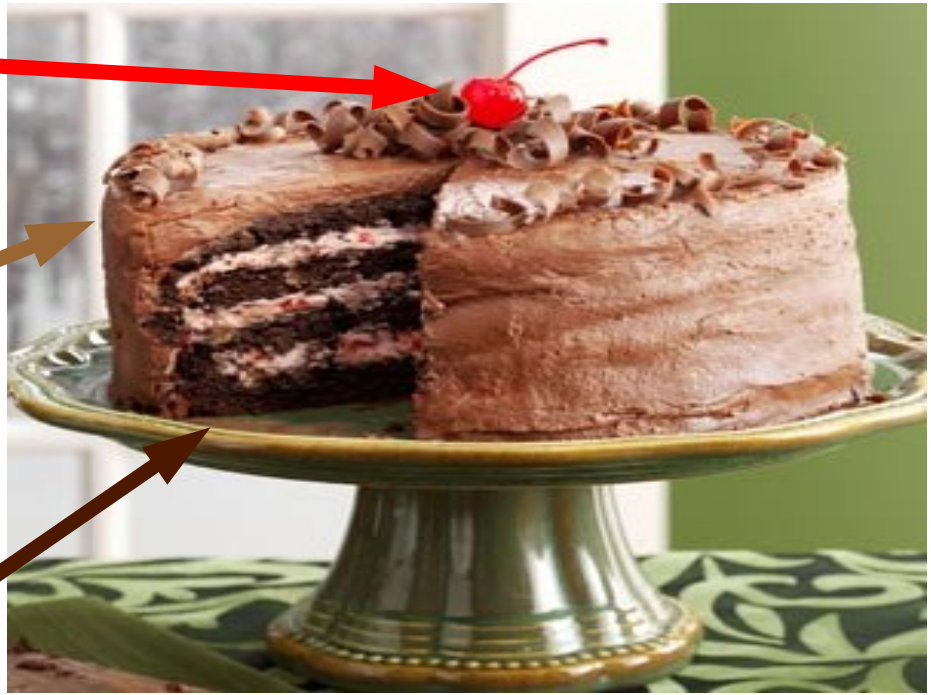
- ***Reinforcement Learning***
  - The machine predicts a scalar reward given once in a while.
  - **weak feedback**
- ***Supervised Learning***
  - The machine predicts a category or a few numbers for each input
  - **medium feedback**
- ***Self-supervised Predictive Learning***
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **A lot of feedback**

PLANE

CAR

How Much Information is the Machine Given during Learning?

- ***"Pure" Reinforcement Learning** (cherry)*
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- ***Supervised Learning** (icing)*
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- ***Self-Supervised Learning** (cake génoise)*
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**



Self-Supervised Learning Fills in the Blanks

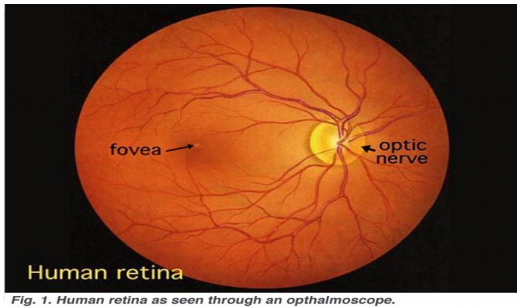


Fig. 1. Human retina as seen through an opthalmoscope.

- ***Our brains do this all the time***

- ***Filling in the visual field at the retinal blind spot***
- ***Filling in occluded images, missing segments in speech***
- ***Predicting the state of the world from partial (textual) descriptions***
- ***Predicting the consequences of our actions***
- ***Predicting the sequence of actions leading to a result***

- ***Predicting any part of the past, present or future percepts from whatever information is available.***

Self-Supervised Learning: Filling in the Blanks

Self-Supervised Learning works well for text

- ***Word2vec***
  - [Mikolov 2013]
- ***FastText***
  - [Joulin 2016]
- ***BERT***
  - Bidirectional Encoder Representations from Transformers
  - [Devlin 2018]

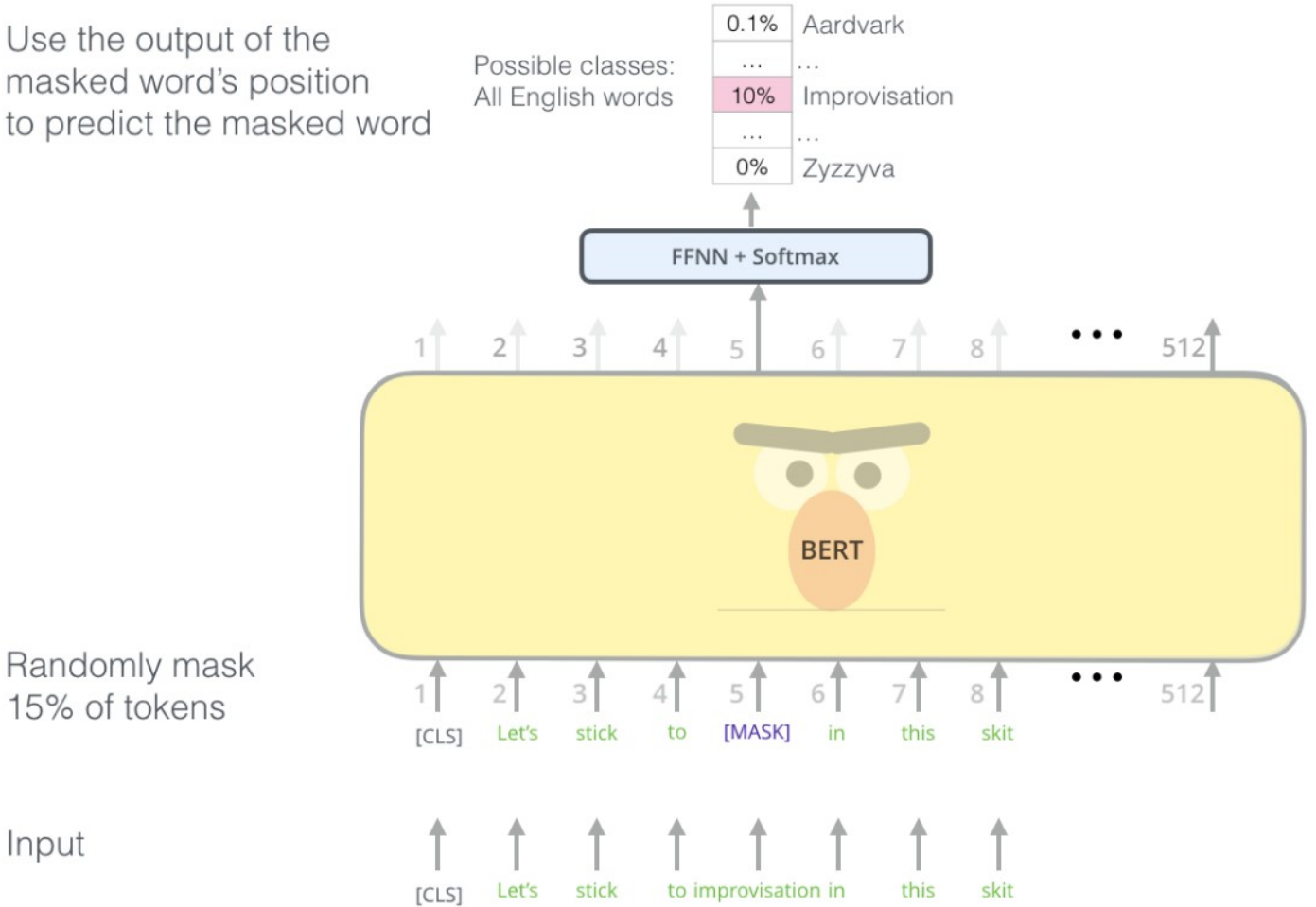
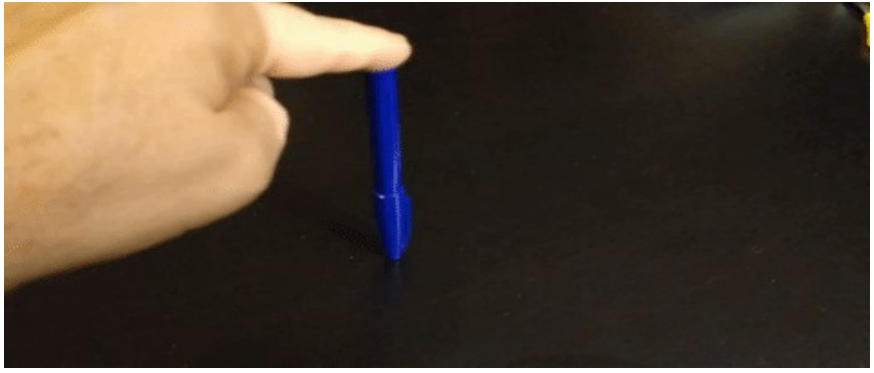


Figure credit: Jay Alammar http://jalammar.github.io/illustrated-bert/

But it doesn't really work for high-dim continuous signals

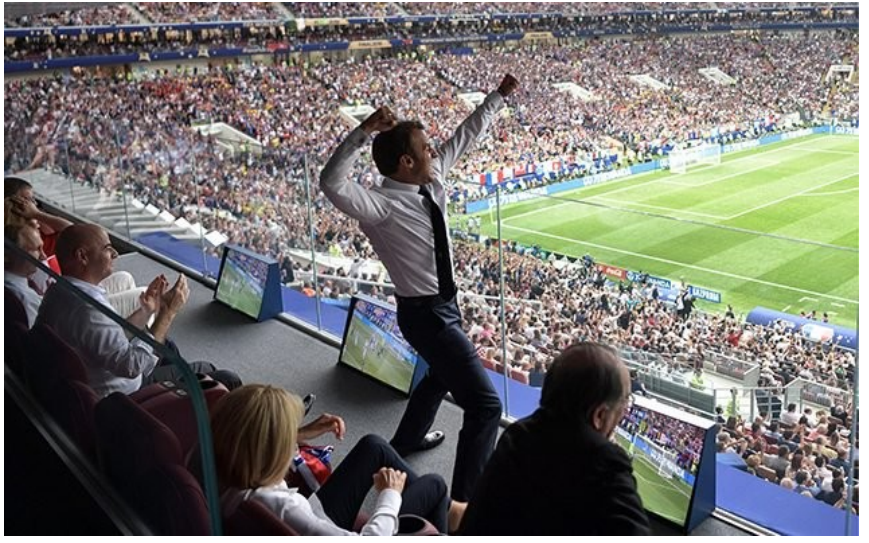
- ***Video prediction:***
  - Multiple futures are possible.
  - Training a system to make a single prediction results in "blurry" results
  - the average of all the possible futures



With thanks
To
Alyosha Efros

Could Self-Supervised Learning Lead to Common Sense?

- ***Successfully learning to predict everything from everything else would result in the accumulation of lots of background knowledge about how the world works.***
- ***Perhaps this ability to learn and use the regularities of the world is what we call common sense.***
- ***If I say "john picks up his briefcase and leaves the conference room"***
  - You can infer a lot of facts about the scene.
  - John is a man, probably at work, he is extending his arm, closing his hand around the handle of his briefcase, standing up, walking towards the door. He is not flying to the door, not going through the wall….



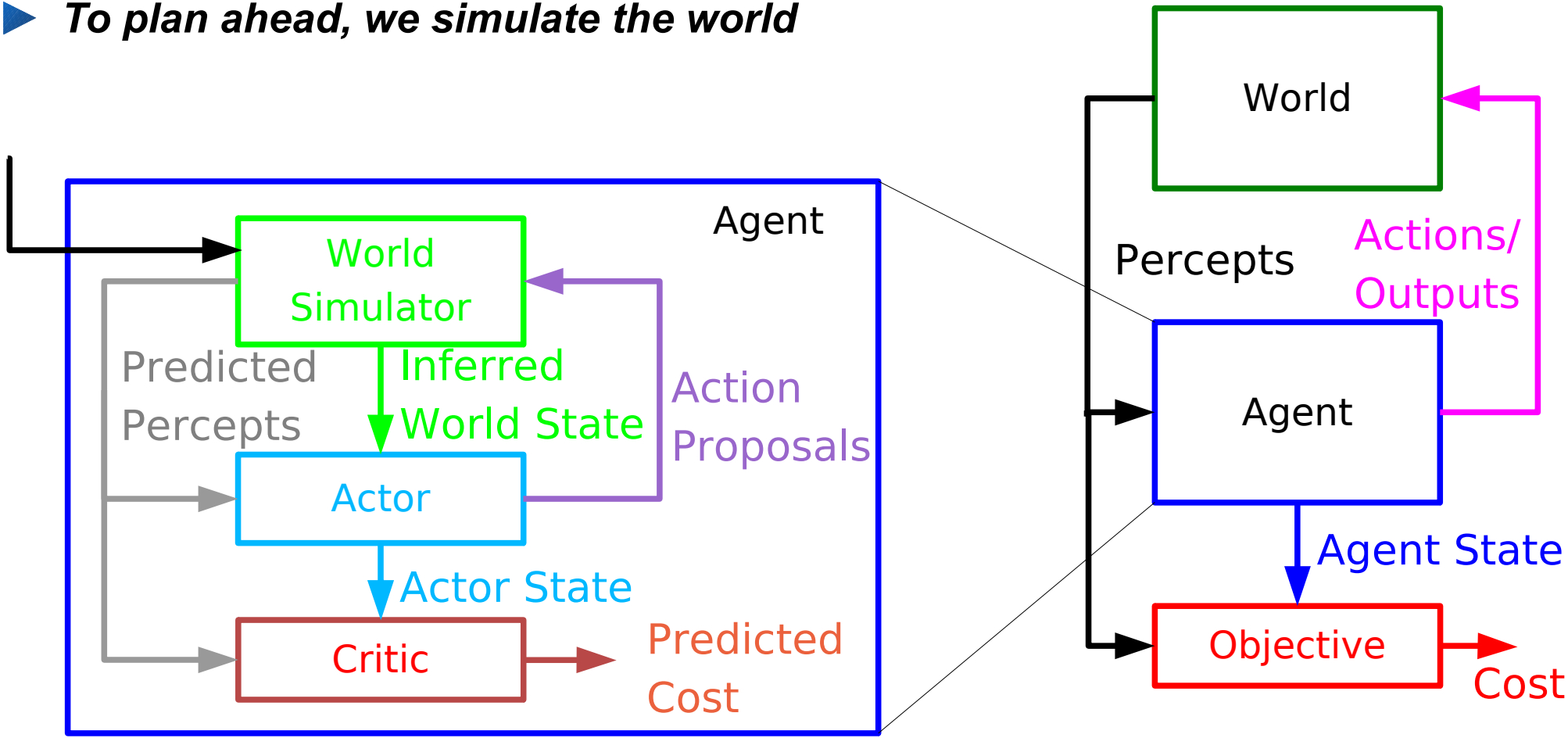
Learning Predictive Models of the World

Learning to predict, reason, and plan, Learning Common Sense.

facebook
Artificial Intelligence Research

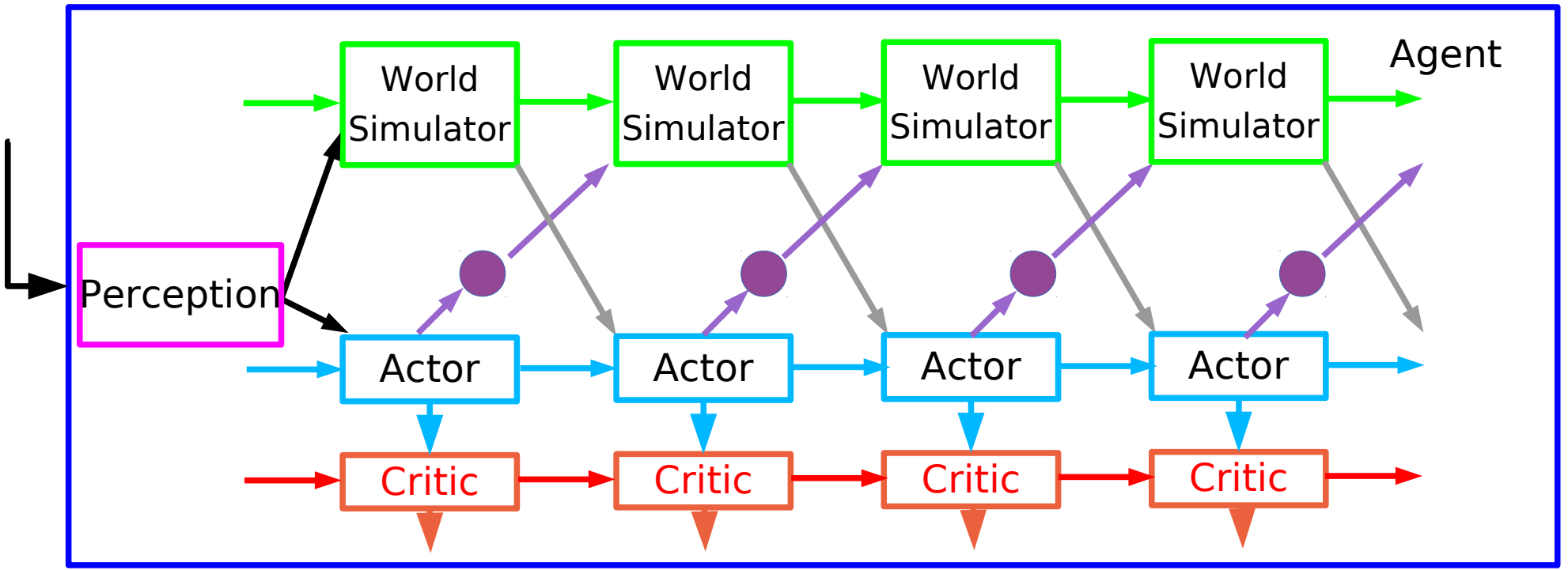
Planning Requires Prediction

- ***To plan ahead, we simulate the world***



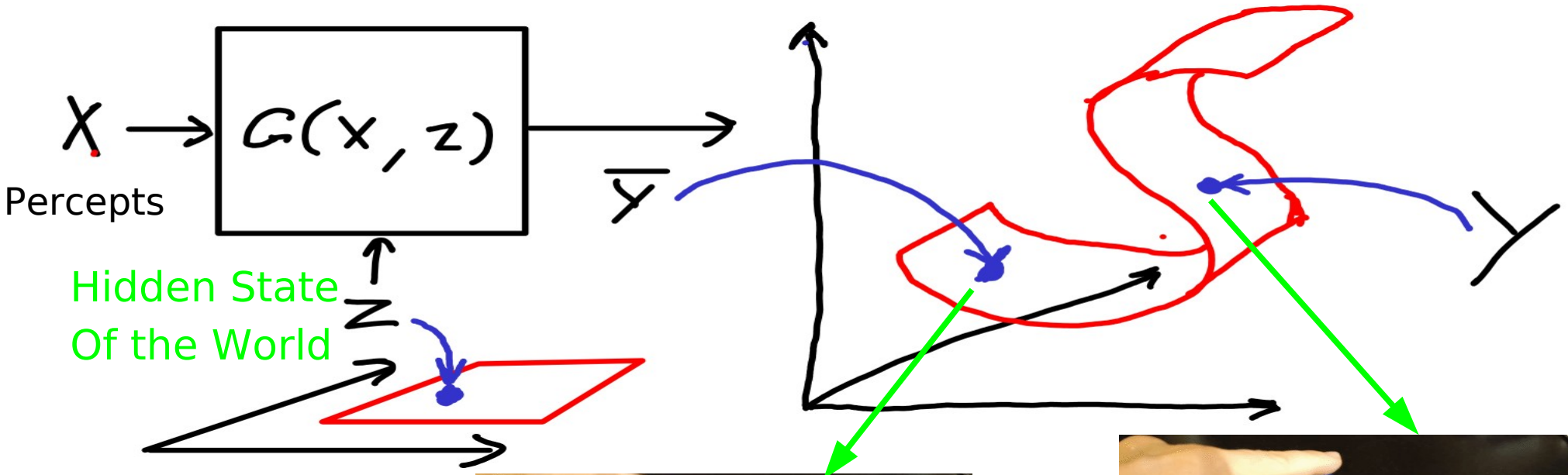
Training the Actor with Optimized Action Sequences

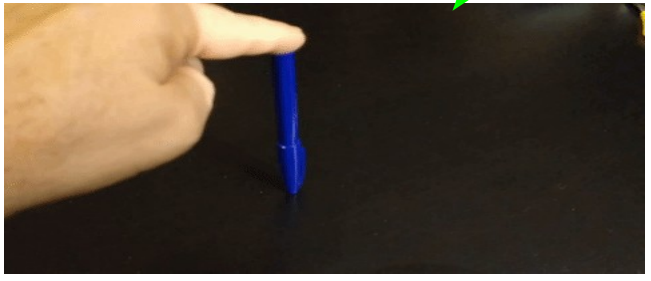
- ***1. Find action sequence through optimization***
- ***2. Use sequence as target to train the actor***
  - Over time we get a compact policy that requires no run-time optimization

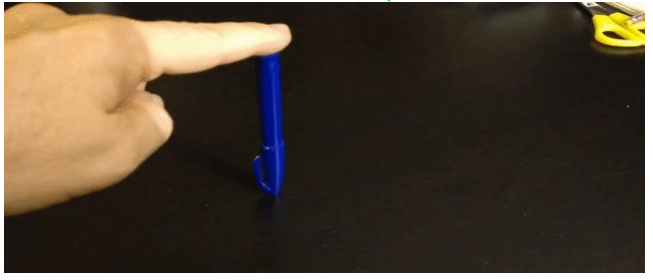


The Hard Part: Prediction Under Uncertainty

- ***Invariant prediction: The training samples are merely representatives of a whole set of possible outputs (e.g. a manifold of outputs).***

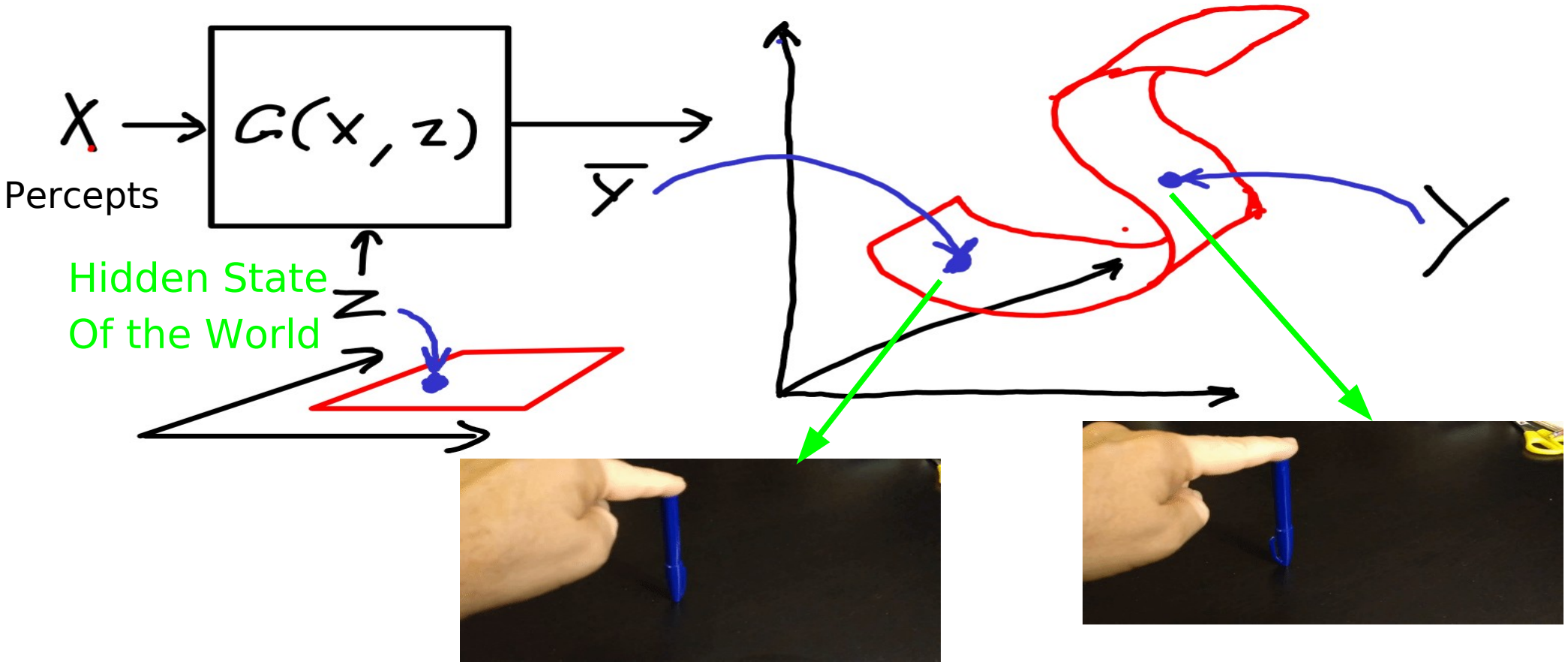






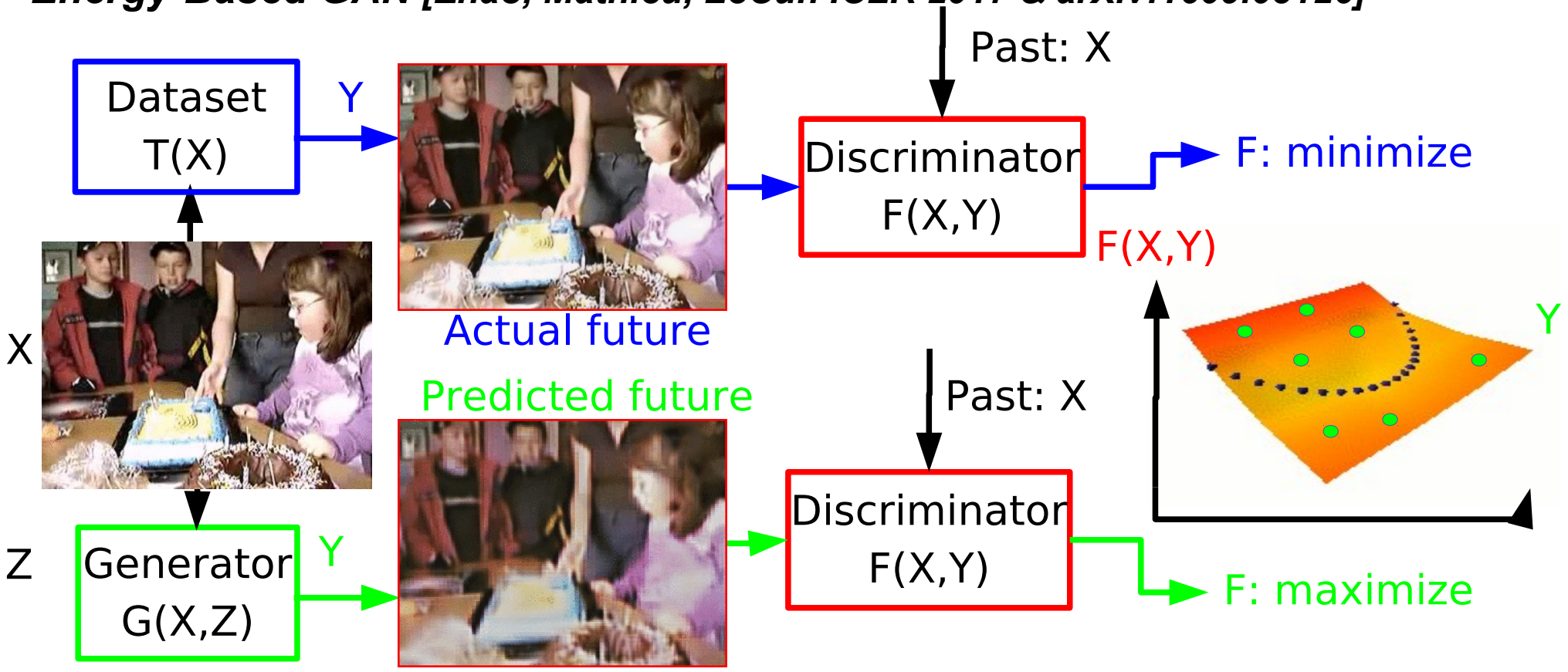
Predicting under Uncertainty: Adversarial Training

- **Invariant prediction: The training samples are merely representatives of a whole set of possible outputs (e.g. a manifold of outputs).**



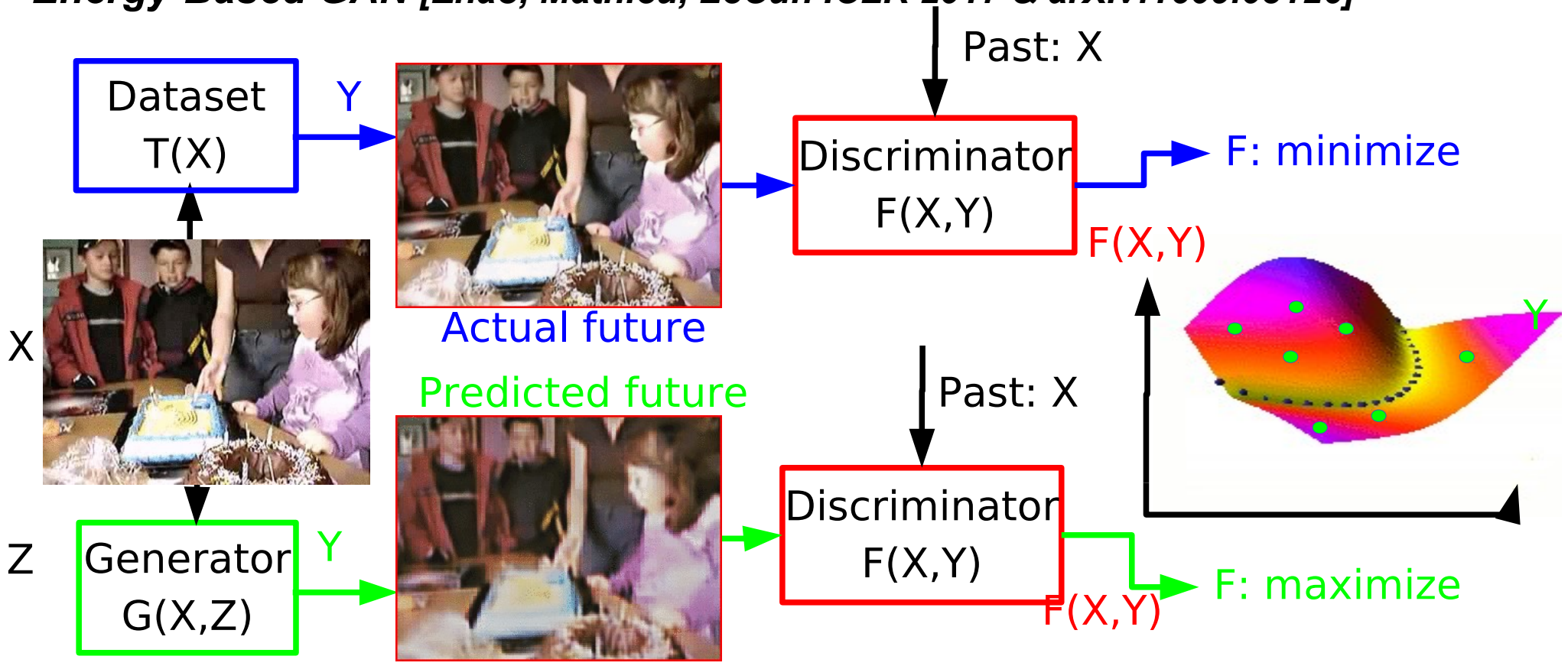
Adversarial Training: the key to prediction under uncertainty?

- ***Generative Adversarial Networks (GAN) [Goodfellow et al. NIPS 2014],***
- ***Energy-Based GAN [Zhao, Mathieu, LeCun ICLR 2017 & arXiv:1609.03126]***



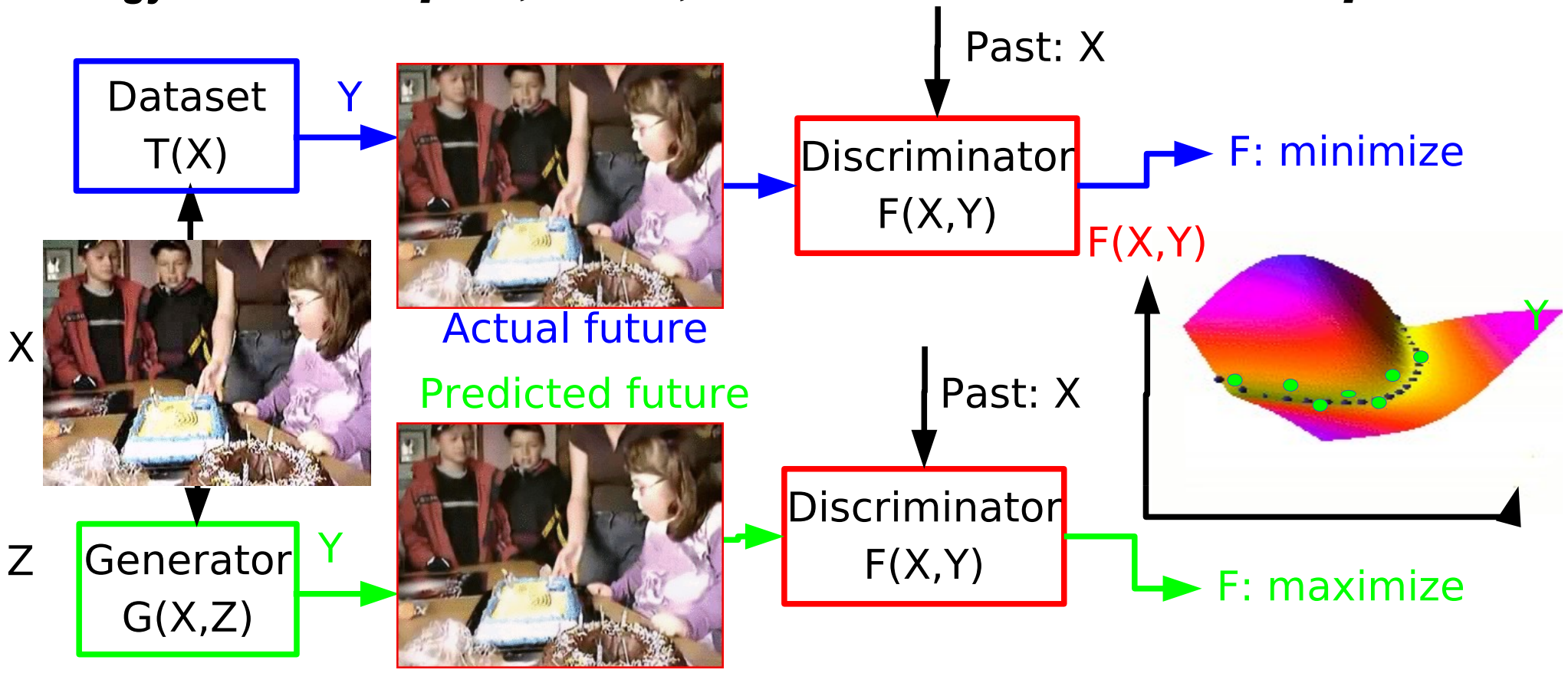
Adversarial Training: the key to prediction under uncertainty?

- ***Generative Adversarial Networks (GAN) [Goodfellow et al. NIPS 2014],***
- ***Energy-Based GAN [Zhao, Mathieu, LeCun ICLR 2017 & arXiv:1609.03126]***



Adversarial Training: the key to prediction under uncertainty?

- ***Generative Adversarial Networks (GAN) [Goodfellow et al. NIPS 2014],***
- ***Energy-Based GAN [Zhao, Mathieu, LeCun ICLR 2017 & arXiv:1609.03126]***



Faces “invented” by a neural net (from NVIDIA)

- ***Random vector → Generator Network → output image*** ***[Goodfellow NIPS 2014]***
- ***[Karras et al. ICLR 2018] (from NVIDIA)***

Adversarial Networks for Creation

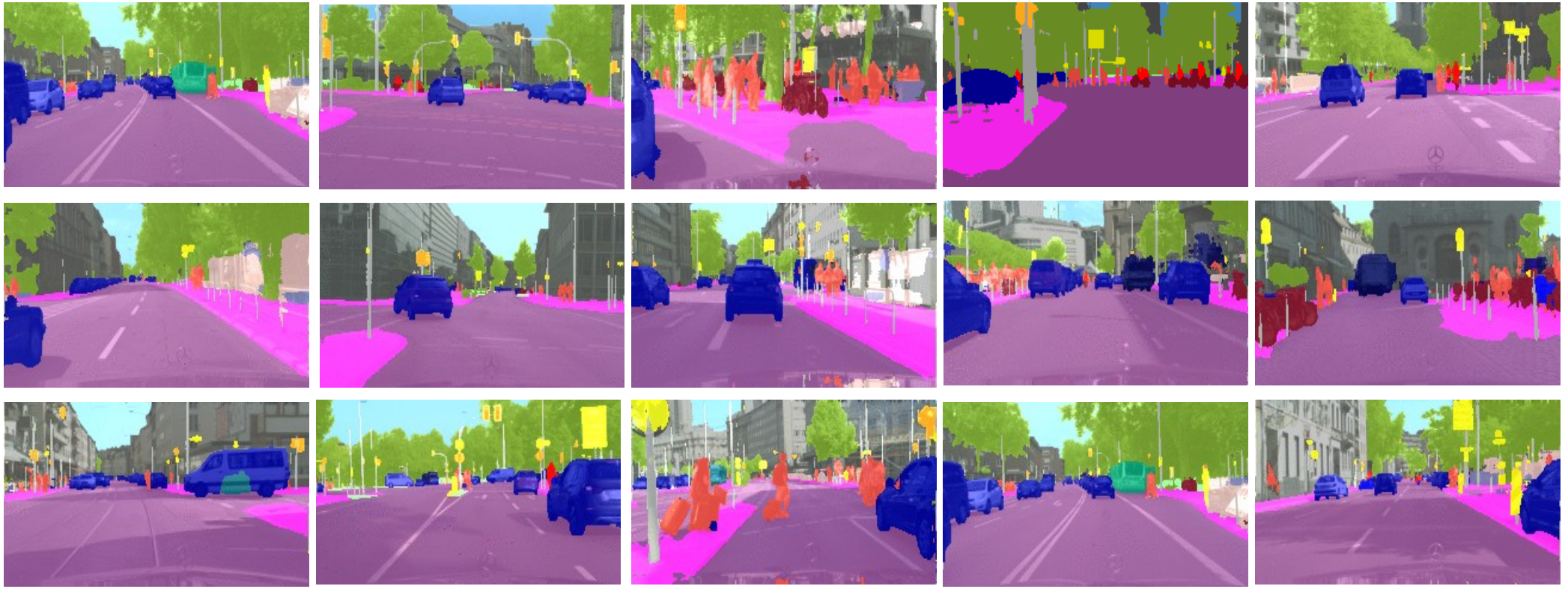
- ***[Sbai 2017]***

Predictive Unsupervised Learning

- ***Our brains are "prediction machines"***
- ***Can we train machines to predict the future?***
- ***Some success with "adversarial training"***
  - [Mathieu, Couprie, LeCun arXiv:1511:05440]
- ***But we are far from a complete solution.***

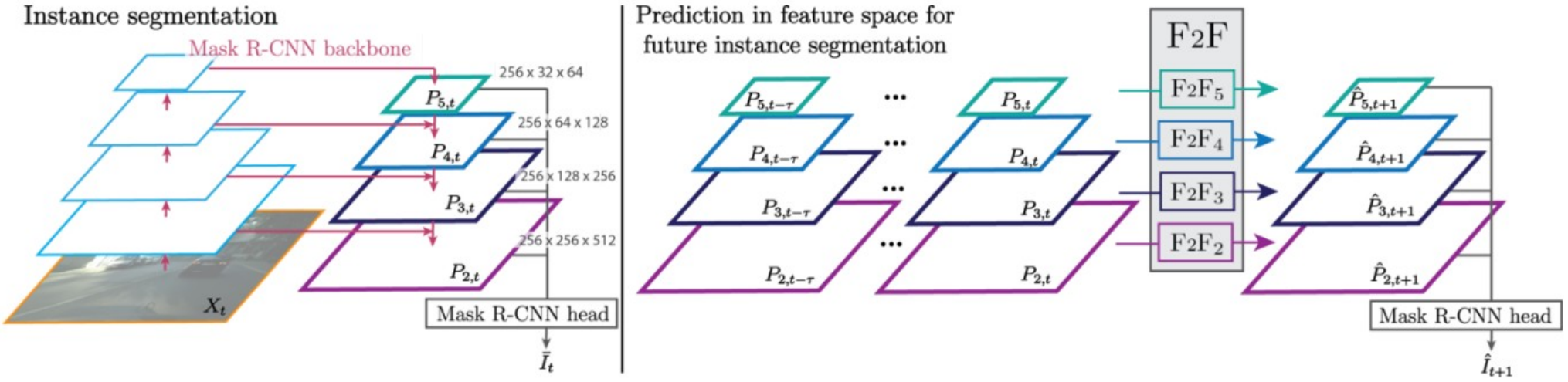
Temporal Predictions of Semantic Segmentations

- ***Prediction 9 frames ahead (0.5 seconds)***
- ***Auto-regressive model***



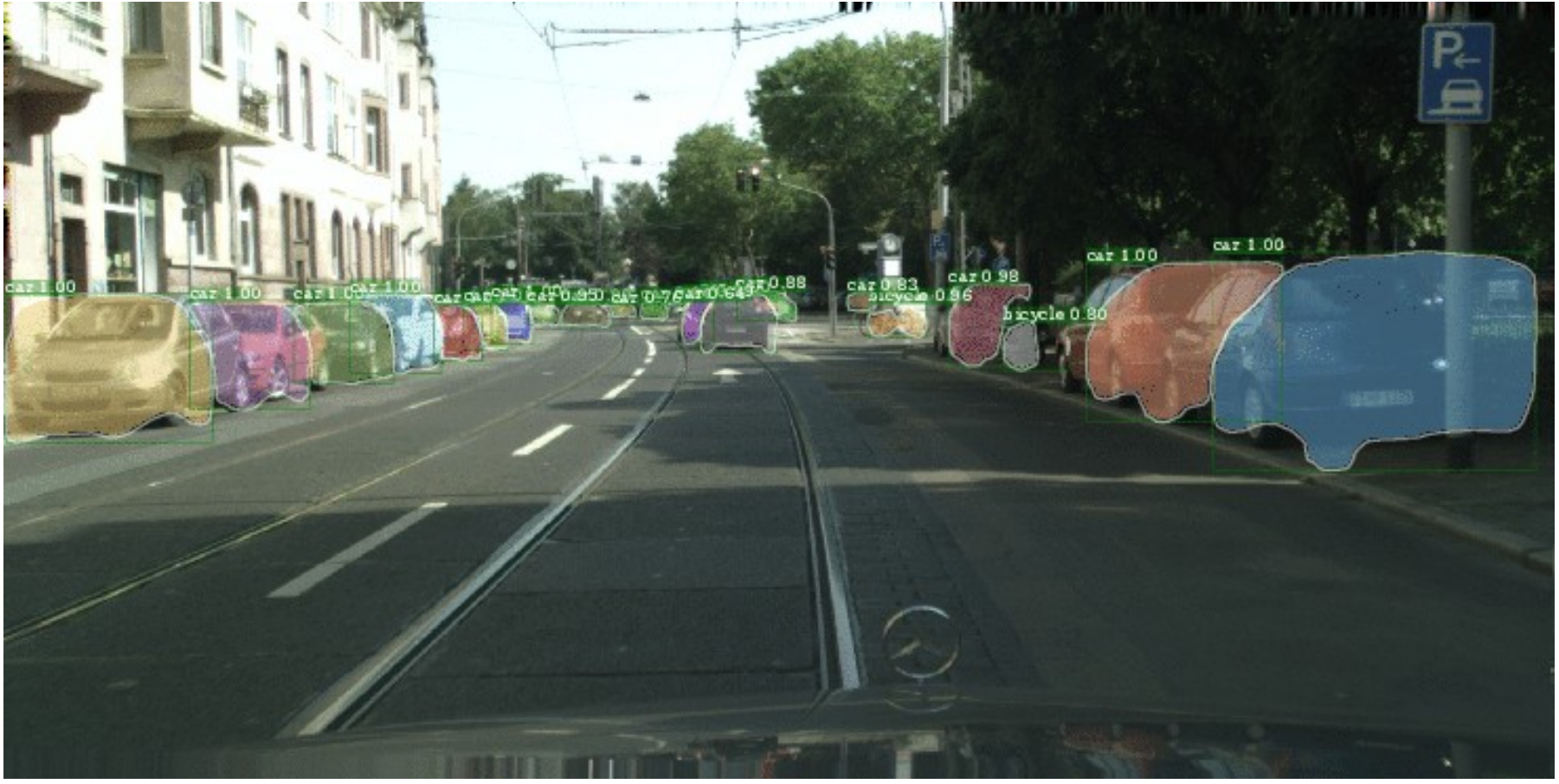
Predicting Instance Segmentation Maps

- ***[Luc, Couprie, LeCun, Verbeek ECCV 2018]***
- ***Mask R-CNN Feature Pyramid Network backbone***
- ***Trained for instance segmentation on COCO***
- ***Separate predictors for each feature level***

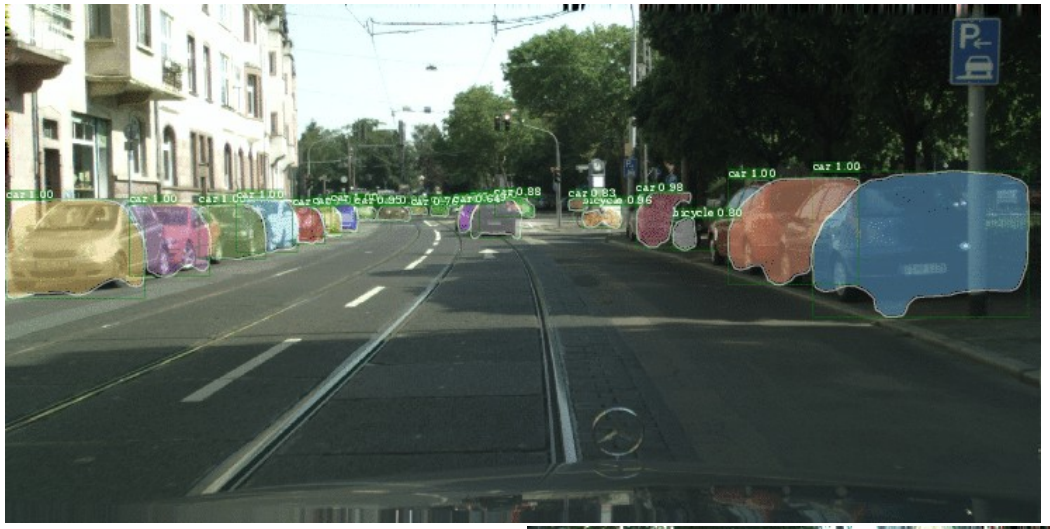


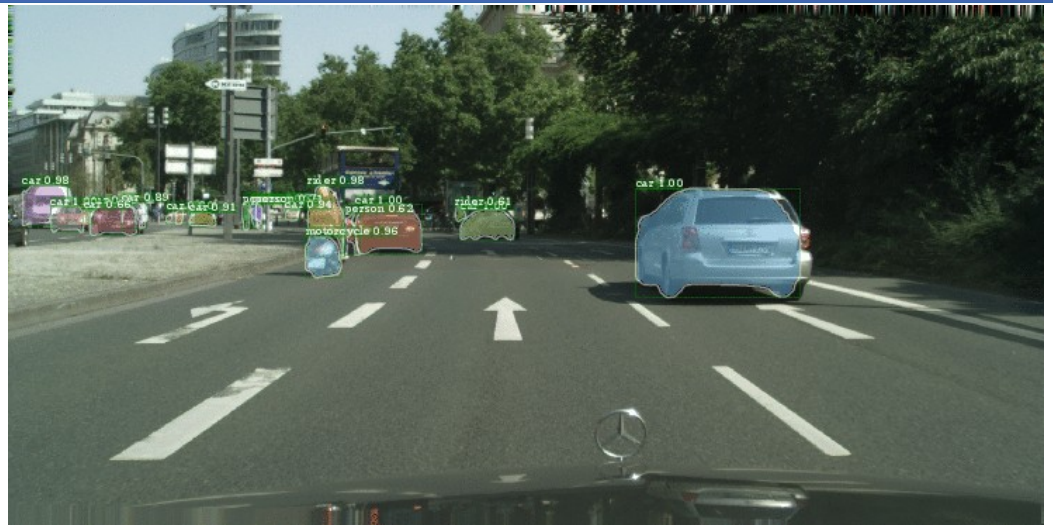
Instance Segmentation Prediction

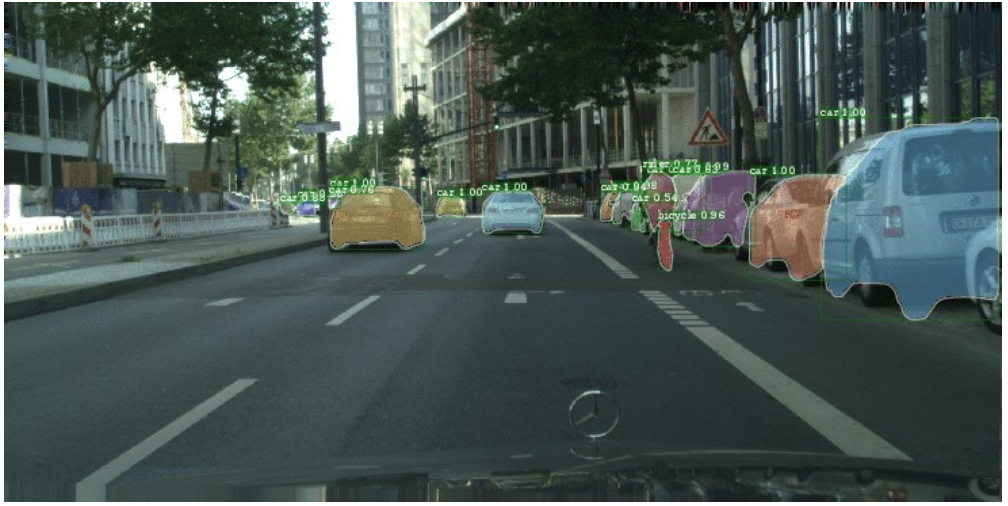
Predictions



Long-term predictions (10 frames, 1.8 seconds)







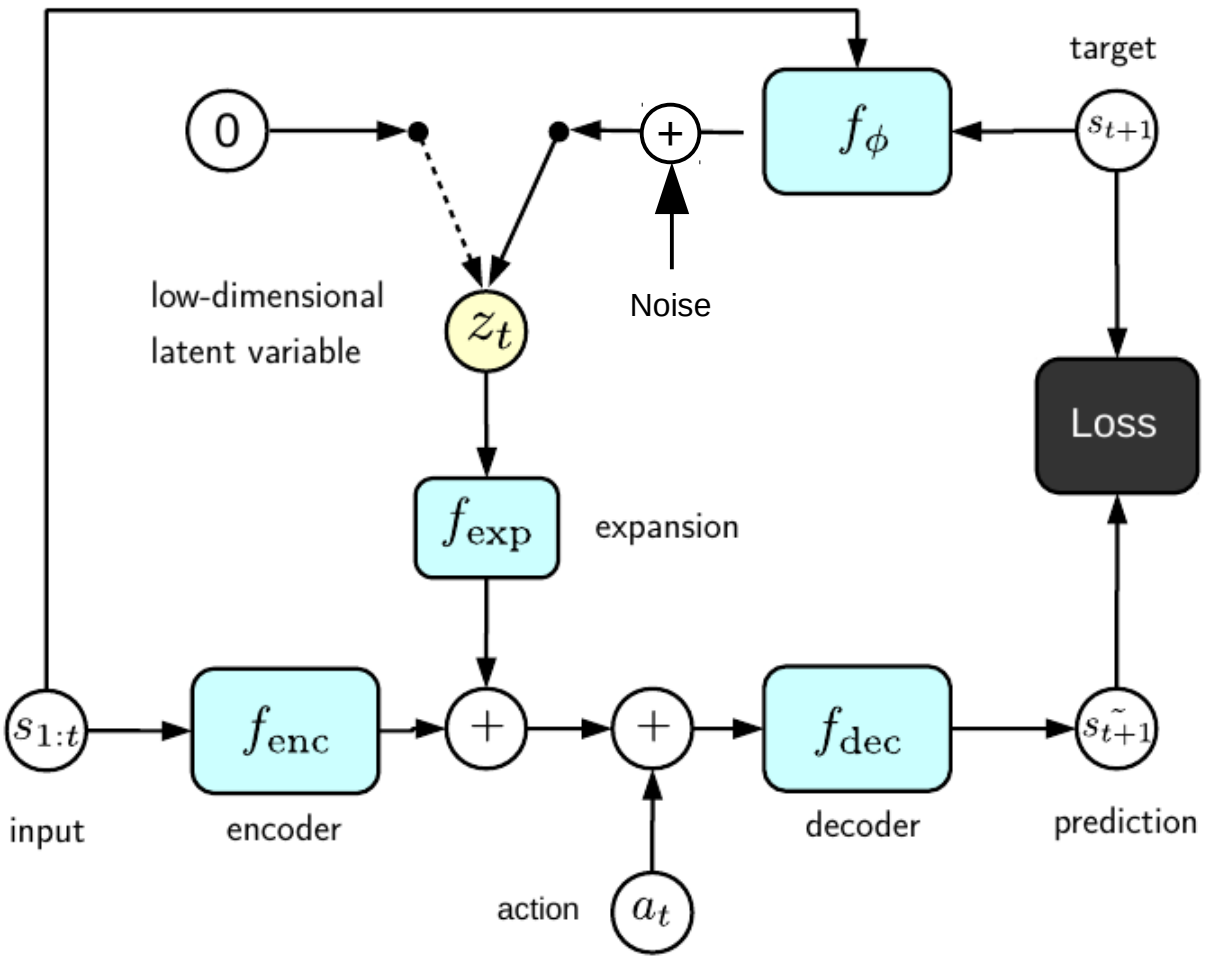
Stochastic Latent-Variable Forward Model:

[Henaff, Canziani, LeCun ICLR 2019]
[Henaff, Zhao, LeCun ArXiv:1711.04994]
[Henaff, Whitney, LeCun Arxiv:1705.07177]

facebook
Artificial Intelligence Research

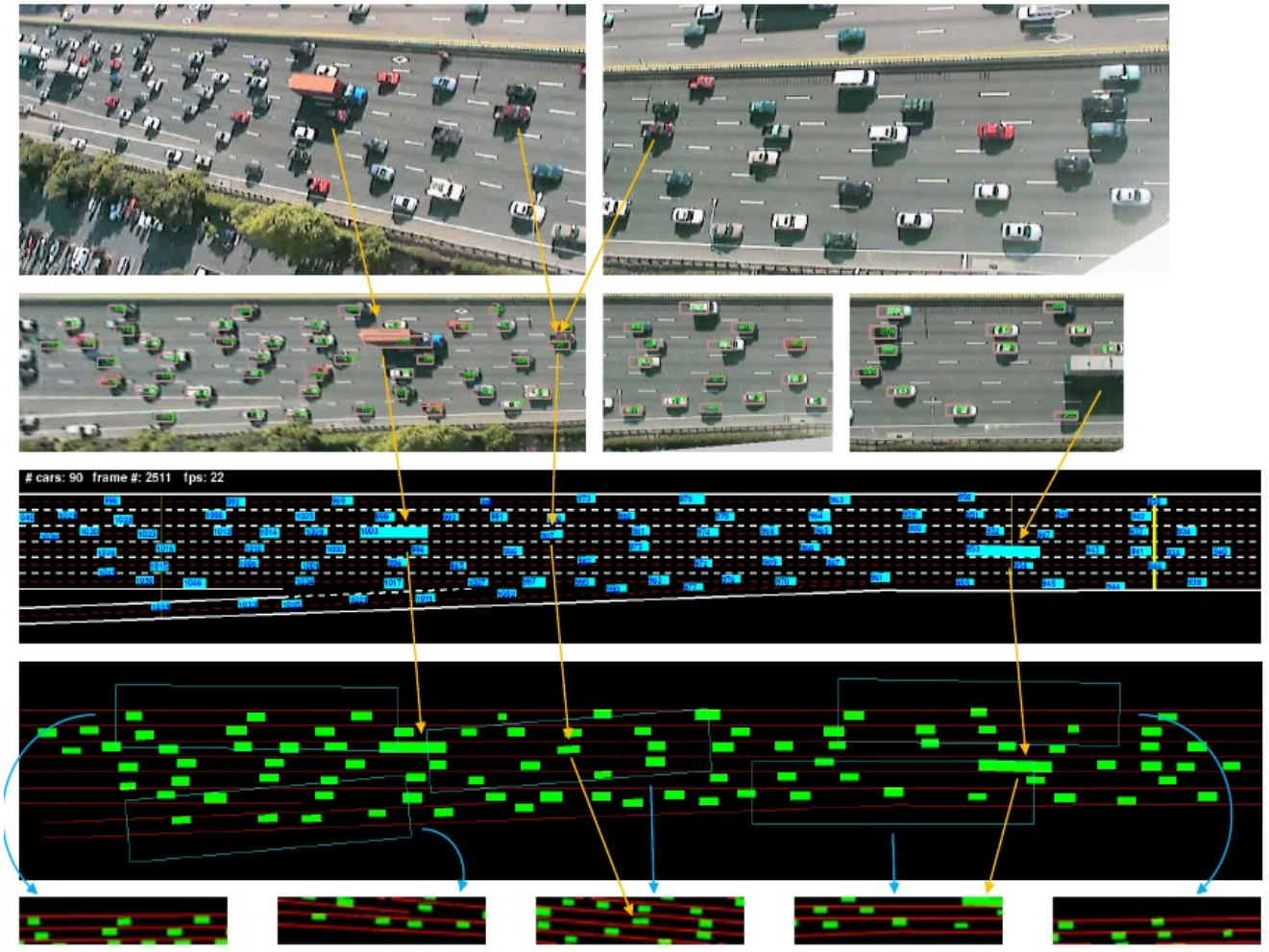
Stochastic Forward Modeling: regularized latent variable model

- ***Latent variable is predicted from the target***
- ***The latent variable is set to zero half the time during training (drop out) and corrupted with noise***
- ***The model predicts as much as it can without the latent var.***
- ***The latent var corrects the residual error.***



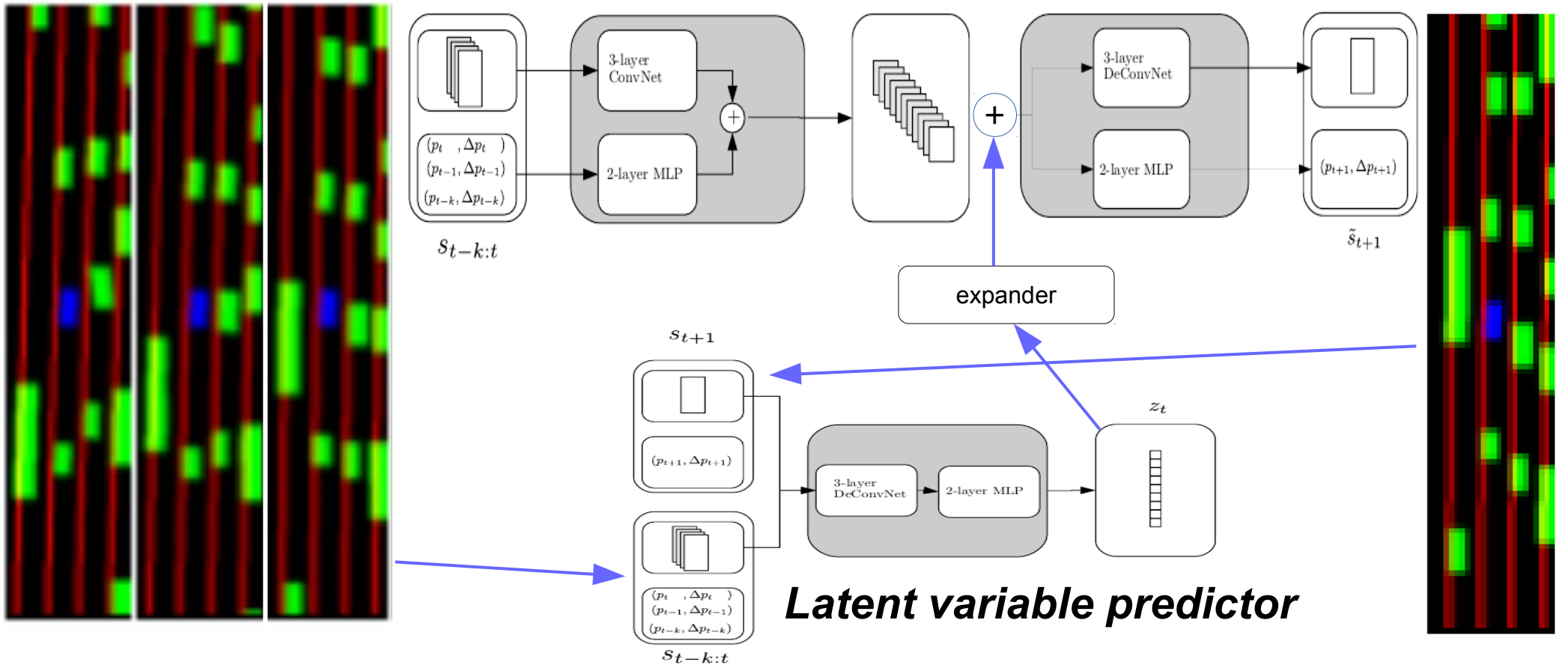
Application to Autonomous Driving

- ***Overhead camera on highway.***
  - Vehicles are tracked
- ***A "state" is a pixel representation of a rectangular window centered around each car.***
- ***Forward model is trained to predict how every car moves relative to the central car.***
  - steering and acceleration are computed

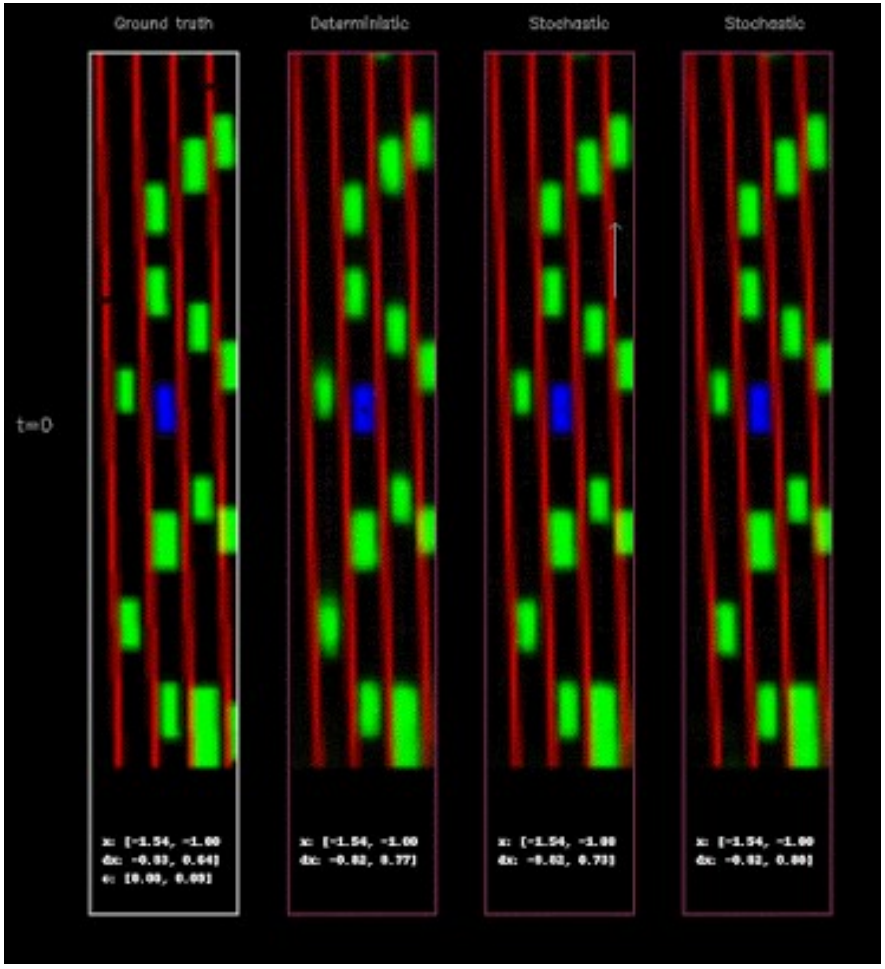


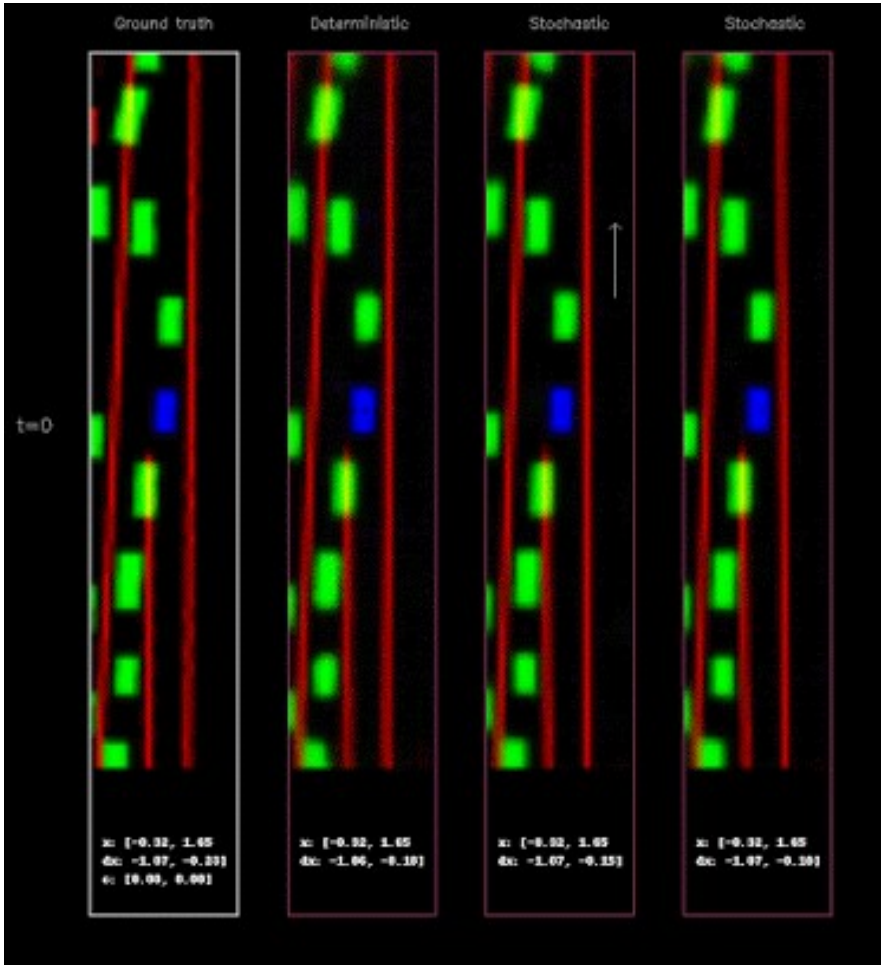
Forward Model Architecture

- ***Architecture:***



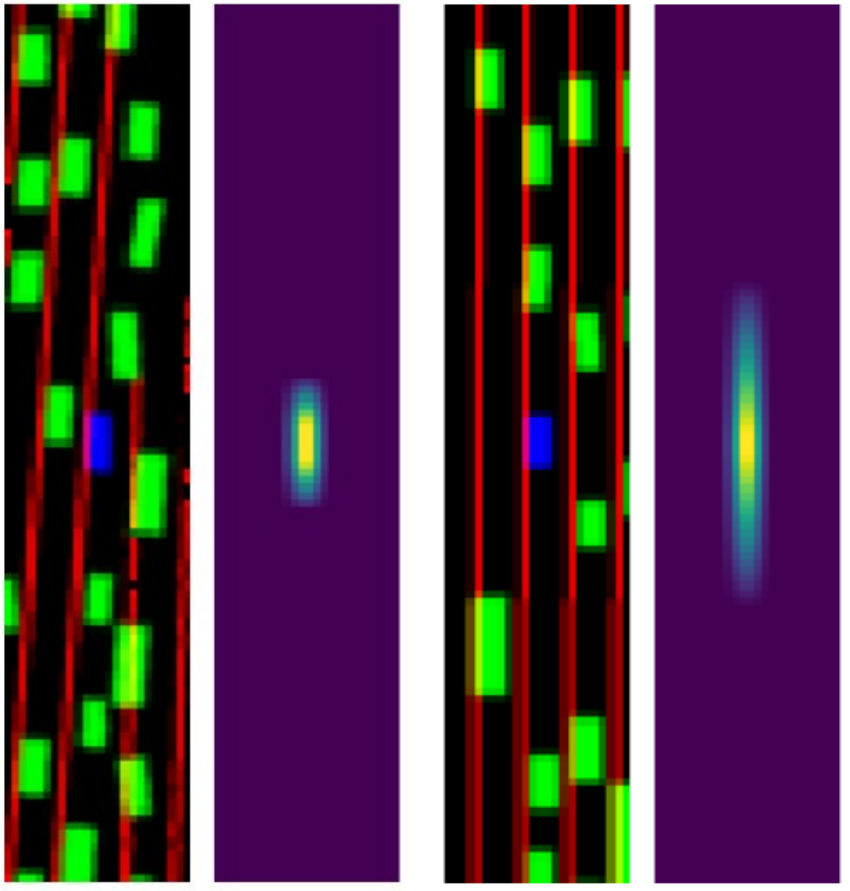
Predictions





Cost optimized for Planning & Policy Learning

- ***Differentiable cost function***
  - Increases as car deviates from lane
  - Increases as car gets too close to other cars nearby in a speed-dependent way
- ***Uncertainty cost:***
  - Increases when the costs from multiple predictions (obtained through sampling of drop-out) have high variance.
  - Prevents the system from exploring unknown/unpredictable configurations that may have low cost.

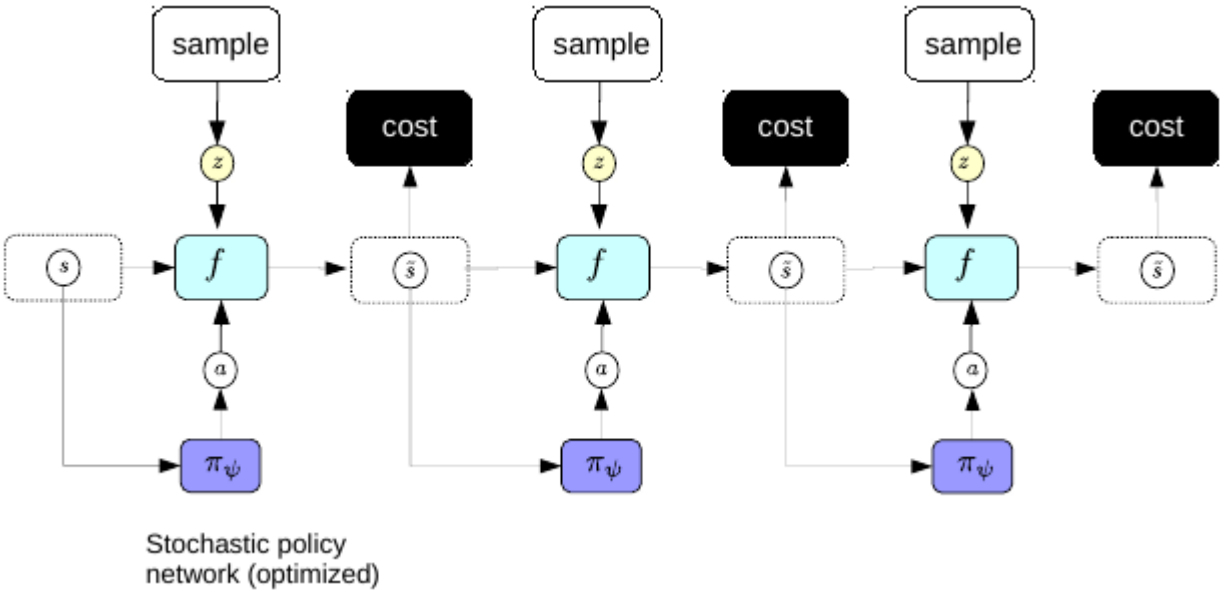


(a) 19.8 km/h (b) 50.3 km/h

Approach 2: Learning a Policy with Stochastic Value Gradient

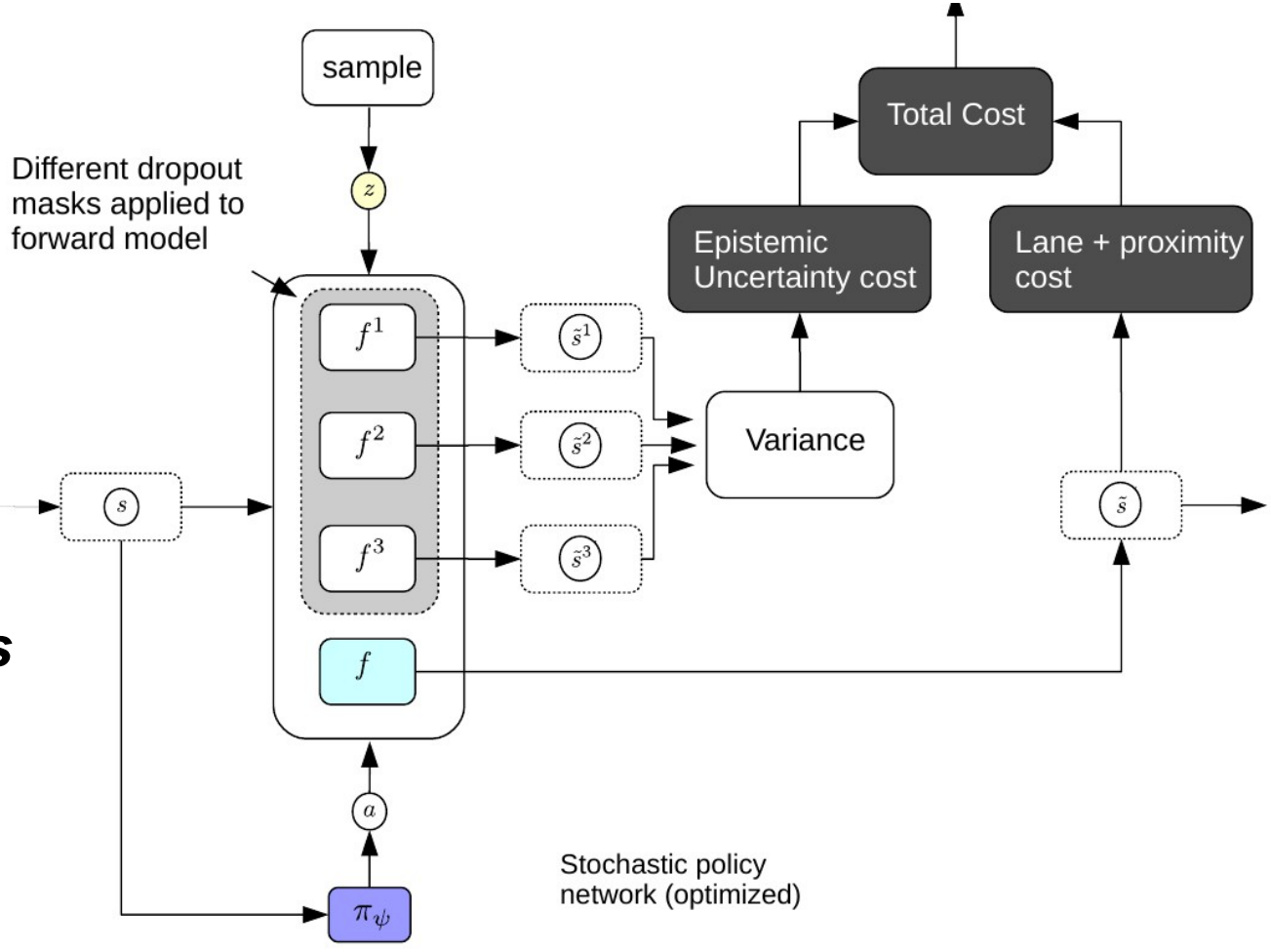
- ***Feed initial state***
- ***Sample latent variable sequences of length 20***
- ***Run the forward model with these sequences***
- ***Backpropagate gradient of cost to train a policy network.***
- ***Iterate***

- ***No need for planning at run time.***



Adding an Uncertainty Cost (doesn't work without it)

- ***Estimates epistemic uncertainty***
- ***Samples multiple drop-puts in forward model***
- ***Computes variance of predictions (differentiably)***
- ***Train the policy network to minimize the lane&proximity cost plus the uncertainty cost.***
- ***Avoids unpredictable outcomes***

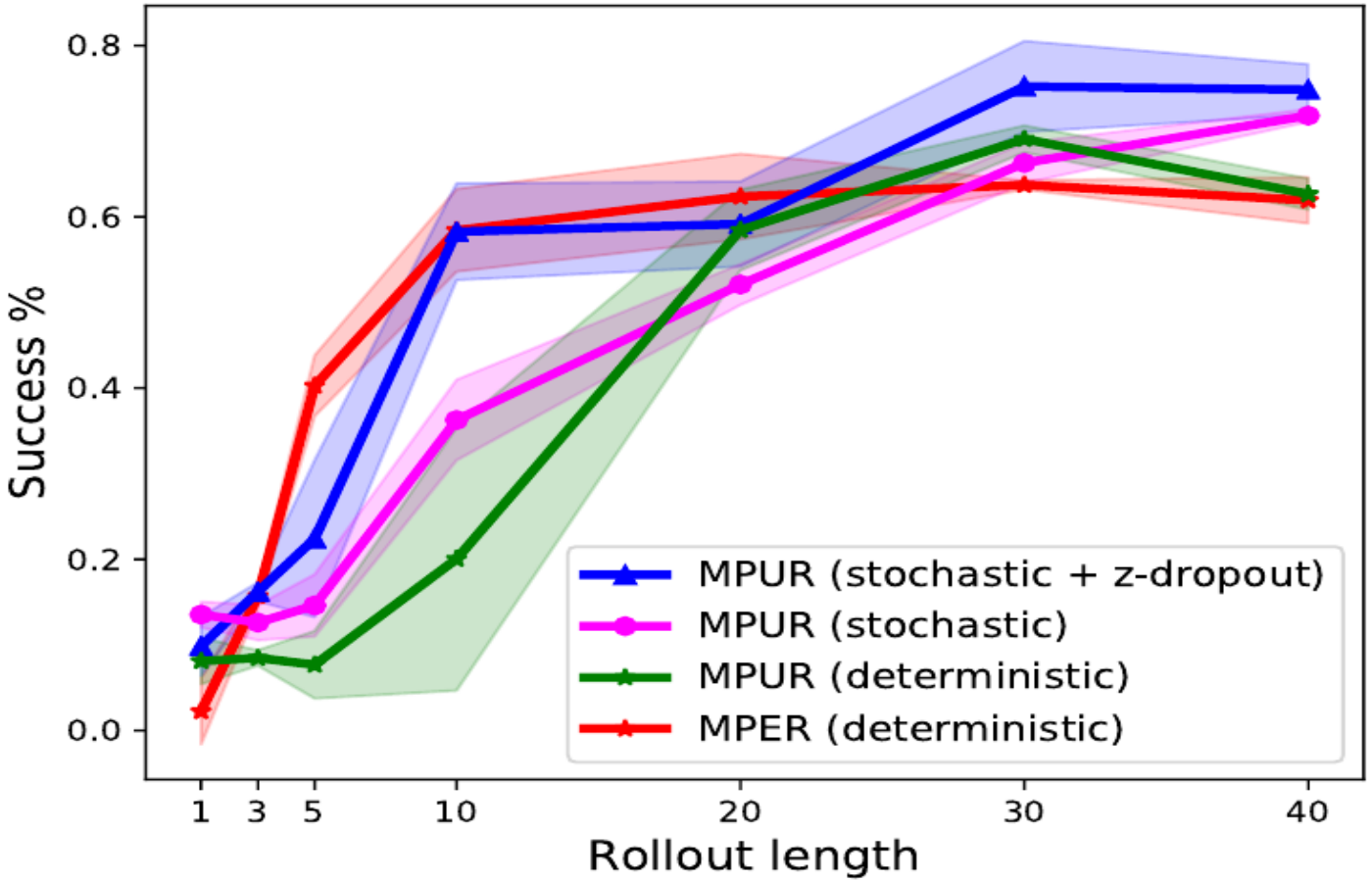


Rate of Successful Runs as a fn of Rollout Length

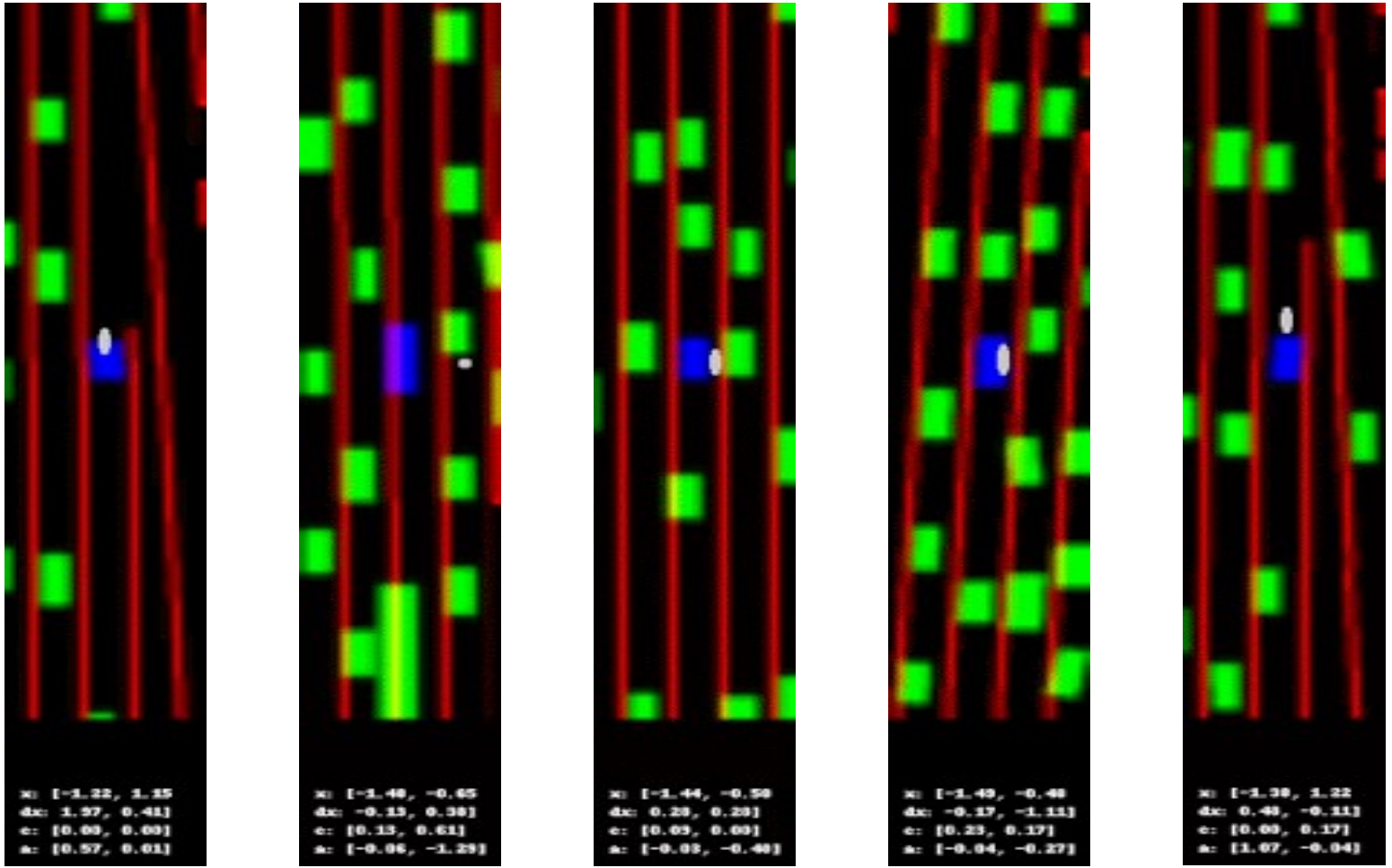
- ***MPUR: Model-Predictive with Uncertainty Regularization***

  - Car gets squeezed hit in the back because it's invisible to the other cars

  - MPER: expert regularization



Driving an Invisible Car in “Real” Traffic



Conclusions

- ***Do theory for understanding phenomena on interesting artifacts***
- ***Don't get hypnothized by cute theory***
- ***Don't get attracted by theoretical lampposts when the key is obviously elsewhere***
- ***People ignore empirical results that don't fit their mental model***
- ***Empiricism works. But extreme empiricism is inefficient***
- ***Theory guides exploratory empiricism***

Thank you

facebook
Artificial Intelligence Research