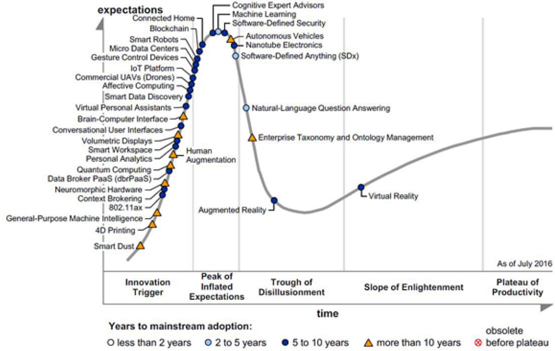
Towards Scalable Statistical Tools for Machine Learning

Dino Sejdinovic

Department of Statistics
University of Oxford

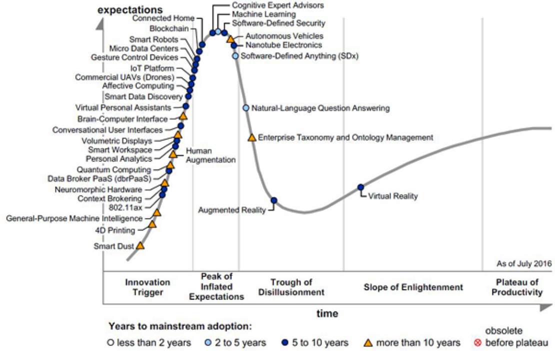
MPLS Summer Reception, 05/06/2017

Machine Learning



Source: Gartner (July 2016)

Machine Learning



Source: Gartner (July 2016)

recommender systems

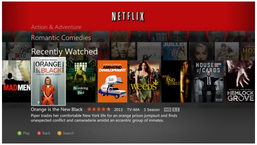
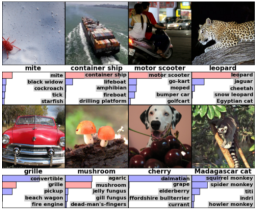
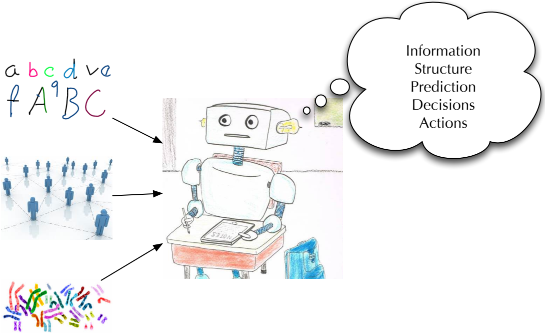

NETFLIX
Action & Adventure
Romantic Comedies
Recently Watched
Orange is the New Black ★★★★★ 2013 TV-MA 1 Season HD 5.1
Piper trades her comfortable New York life for an orange prison jumpsuit and finds unexpected conflict and camaraderie amidst an eccentric group of inmates.
Play
Back
Search

image recognition



machine translation

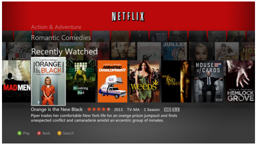
Machine Learning



Information
Structure
Prediction
Decisions
Actions

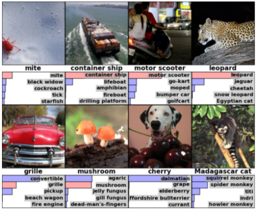
data

recommender systems



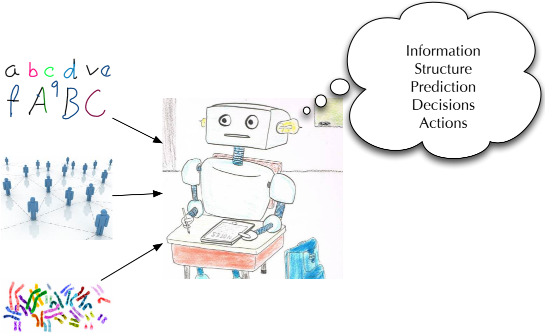
Piper trades her comfortable New York life for an orange prison jumpsuit and finds unexpected conflict and camaraderie amidst an eccentric group of inmates.

image recognition



machine translation

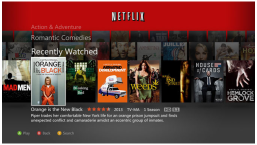
Machine Learning



Information
Structure
Prediction
Decisions
Actions

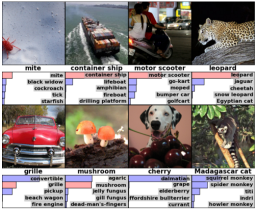
data

recommender systems



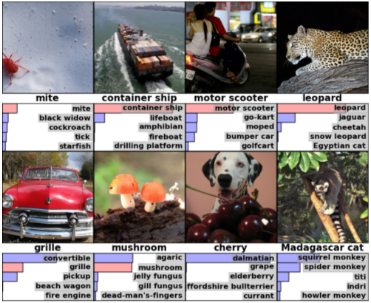
Piper trades her comfortable New York life for an orange prison jumpsuit and finds unexpected conflict and camaraderie amidst an eccentric group of inmates.

image recognition



machine translation

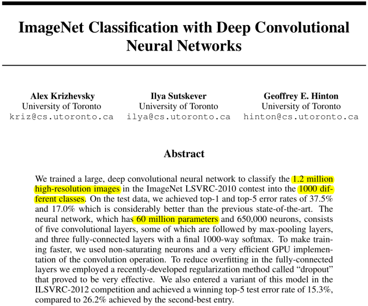
"...procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data..."
John Tukey, *The Future of Data Analysis*, 1962.



[Krizhevsky, Sutskever & Hinton, 2012]

The field of machine learning has been driven by the exponential growth in dataset sizes and computational resources, allowing to tackle difficult inference problems, which are characterized by:

- high dimensionality,
- multivariate interaction,
- complex patterns exhibiting various forms of nonlinearity and nonstationarity,
- little prior knowledge.



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

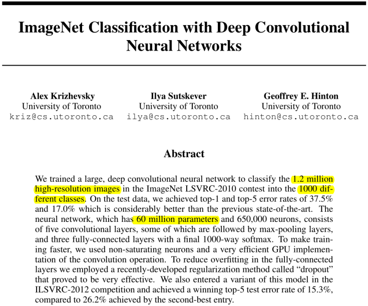
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

The field of machine learning has been driven by the exponential growth in dataset sizes and computational resources, allowing to tackle difficult inference problems, which are characterized by:

- high dimensionality,
- multivariate interaction,
- complex patterns exhibiting various forms of nonlinearity and nonstationarity,
- little prior knowledge.



ImageNet Classification with Deep Convolutional Neural Networks

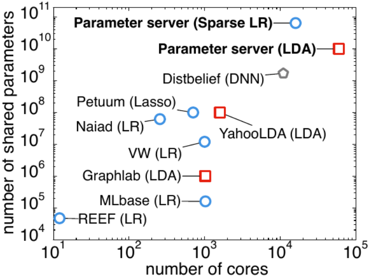
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.



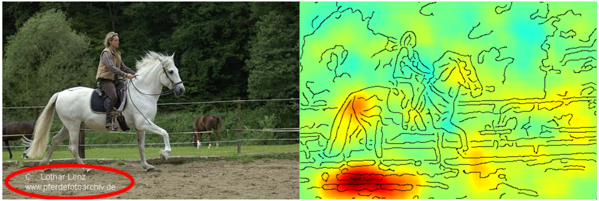
[Mu Li et al, 2014]

The field of machine learning has been driven by the exponential growth in dataset sizes and computational resources, allowing to tackle difficult inference problems, which are characterized by:

- high dimensionality,
- multivariate interaction,
- complex patterns exhibiting various forms of nonlinearity and nonstationarity,
- little prior knowledge.

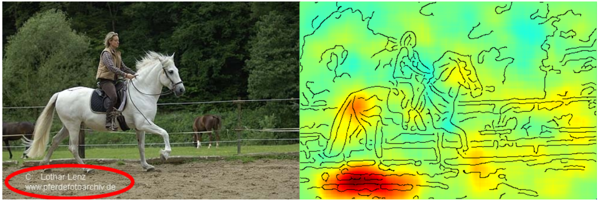
The use of complex models with massive amounts of parameters, even if they are unidentifiable and uninterpretable.

Uncertainty Calibration and Brittleness

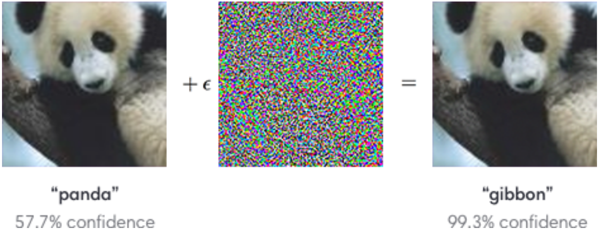


[Lapuschkin et al, 2016]

Uncertainty Calibration and Brittleness

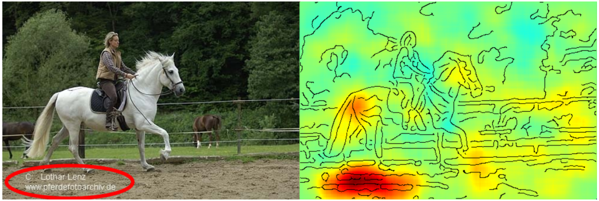


[Lapuschkin et al, 2016]

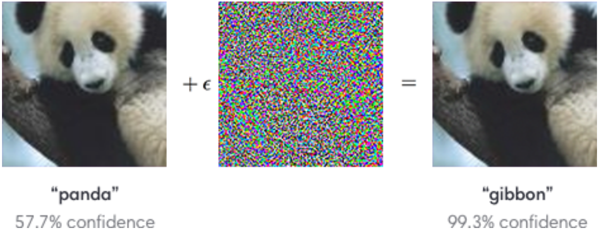


[Goodfellow et al, 2015]

Uncertainty Calibration and Brittleness



[Lapuschkin et al, 2016]

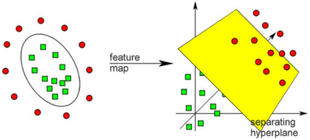


[Goodfellow et al, 2015]

Need for (scalable) statistical tools for model criticism and interpretability.

Kernel Trick and Kernel Mean Trick

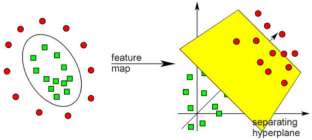
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\phi(x)^\top \phi(y) = k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k}$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

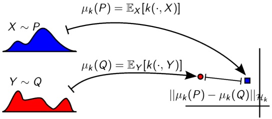
Kernel Trick and Kernel Mean Trick

- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\phi(x)^\top \phi(y) = k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k}$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



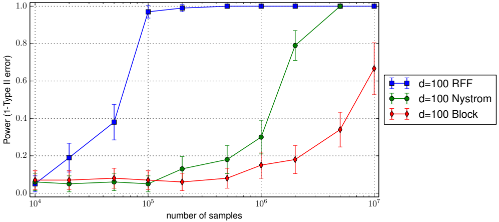
[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

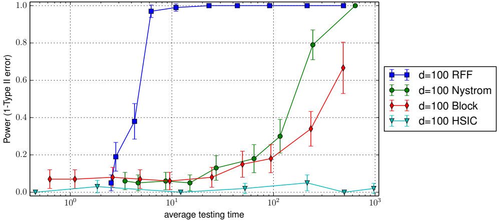
- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y)$
  inner products easy to estimate
  - nonparametric two-sample, independence, conditional independence, interaction testing, learning on distributions, model criticism and interpretability


$\mu_k(P) = \mathbb{E}_X[k(\cdot, X)]$
$X \sim P$
$\mu_k(Q) = \mathbb{E}_Y[k(\cdot, Y)]$
$Y \sim Q$
$\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}$

[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; Muandet et al, 2012; DS et al, 2013; Szabo et al, 2015; Kim et al, 2016]

Large-Scale Statistical Tests





Tradeoffs between statistical and computational efficiency:

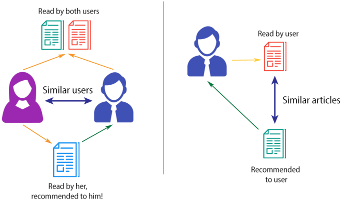
limited data, unlimited time

$\rightarrow$

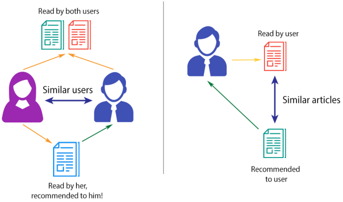
unlimited data, limited time

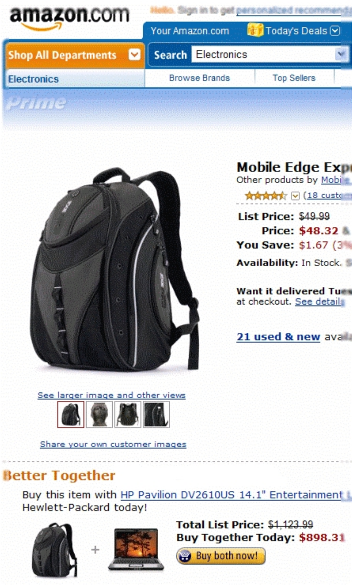
[K. Chwialkowski, A. Ramdas, DS, and A. Gretton, Fast Two-Sample Testing with Analytic Representations of Probability Measures, in *Advances in Neural Information Processing Systems*, 2015.]

[Q. Zhang, S. Filippi, A. Gretton, and DS, Large-Scale Kernel Methods for Independence Testing, *Statistics and Computing*, 2017.]


Read by both users
Similar users
Read by her, recommended to him!
Read by user
Similar articles
Recommended to user

CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
piccolo





Mobile Edge Exp
Other products by Mobi
(18 custo

List Price: ~~$49.99~~
Price: $48.32 &
You Save: $1.67 (3%
Availability: In Stock. S

Want it delivered Tue
at checkout. See detail

21 used & new avai

See larger image and other views

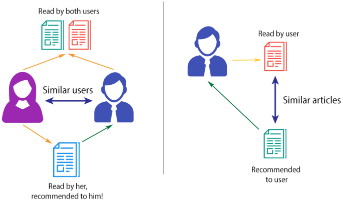
Share your own customer images

Better Together

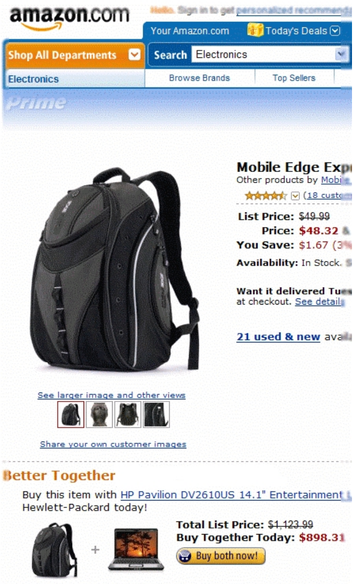
Buy this item with HP Pavilion DV2610US 14.1" Entertainment
Hewlett-Packard today!

Total List Price: ~~$1,123.99~~
Buy Together Today: $898.31
Buy both now!

[example by Bernhard Schölkopf]



CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
Piccolo



amazon.com

Hello. Sign in to get personalized recommend

Your Amazon.com | Today's Deals

Shop All Departments | Search Electronics

Electronics | Browse Brands | Top Sellers

Prime

Mobile Edge Exp

Other products by Mobi

(18 custo

List Price: ~~$49.99~~

Price: $48.32 &

You Save: $1.67 (3%

Availability: In Stock.

Want it delivered Tue
at checkout. See detail

21 used & new avai

See larger image and other views

Share your own customer images

Better Together

Buy this item with HP Pavilion DV2610US 14.1" Entertainment
Hewlett-Packard today!

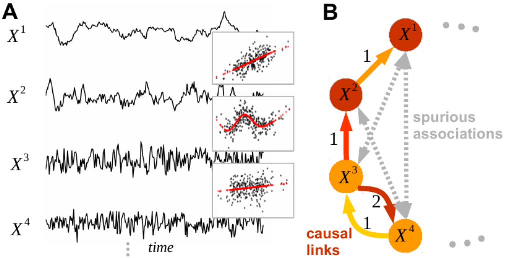
Total List Price: ~~$1,123.99~~
Buy Together Today: $898.31

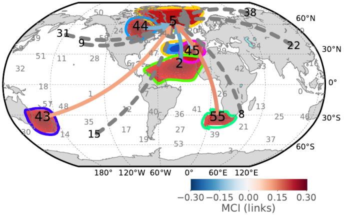
Buy both now!

[example by Bernhard Schölkopf]

Disentangling Correlation from Causation in Machine Learning?

Causal Discovery from Time Series Data

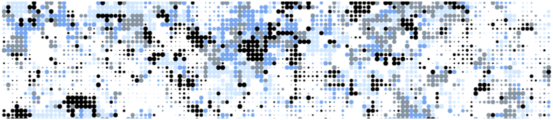




[J. Runge, S. Flaxman and DS, Detecting causal associations in large nonlinear time series datasets, ArXiv e-prints:1702.07007, 2017.]

OxCSML: Machine Learning at the Department of Statistics

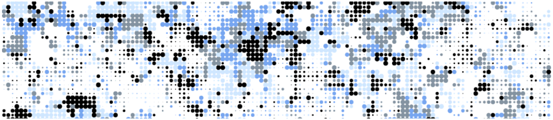
- 5+ faculty, 6+ postdocs, 16+ students...
- Novel machine learning techniques from theoretically grounded concepts.
- Probabilistic modelling, Bayesian nonparametrics, automated and approximate inference, kernel methods, causal discovery, learning under model misspecification, Monte Carlo methods, and deep learning, with applications to network analysis, recommender systems, text processing, spatio-temporal modelling, genetics and genomics.
- The group in numbers: 42 NIPS papers, 7 NIPS orals, 17 ICML papers, 14 UAI papers, 12 AISTATS papers, 7 JMLR papers...



http://csml.stats.ox.ac.uk

OxCSML: Machine Learning at the Department of Statistics

- 5+ faculty, 6+ postdocs, 16+ students...
- Novel machine learning techniques from theoretically grounded concepts.
- Probabilistic modelling, Bayesian nonparametrics, automated and approximate inference, kernel methods, causal discovery, learning under model misspecification, Monte Carlo methods, and deep learning, with applications to network analysis, recommender systems, text processing, spatio-temporal modelling, genetics and genomics.
- The group in numbers: 42 NIPS papers, 7 NIPS orals, 17 ICML papers, 14 UAI papers, 12 AISTATS papers, 7 JMLR papers...



Thank You!

http://csml.stats.ox.ac.uk