On the mysteries of artificial intelligence
-
Do we know what we are doing?

Anders C. Hansen (Cambridge and UiO)

Oslo, May 2019

Main goal: Secure and Safe AI

Main issues:

- AI techniques will replace humans in problem solving.
- AI techniques will replace established algorithms in the sciences.

AI replacing humans

- Self-driving vehicles
- Automated diagnosis in medicine
- Automated decision processes
- Automated weapon systems
- Any security system based on face or voice recognition

AI replacing algorithms

- Medical imaging (MRI, CT, etc)
- Microscopy
- Imaging problems in general
- Radar
- Sonar
- Inverse problems in general
- PDEs

Why is suddenly AI such a big deal?

The Pioneers

Forbes

Billionaires Innovation Leadership Money Consumer Industry

Turing Award And $1 Million Given To 3 AI Pioneers

Nicole Martin Contributor ⓘ
AI & Big Data
I write about technology, data and privacy.

Winners of Turning Award NEW YORK TIMES

The Association for Computing Machinery (ACM) awarded Yoshua Bengio, Geoffrey Hinton and Yann LeCun with what many consider the "Nobel Prize of computing," for the innovations they've made in AI.

Citation from the Turing Award jury

Select Technical Accomplishments

The technical achievements of this year's Turing Laureates, which have led to significant breakthroughs in AI technologies include, but are not limited to, the following:

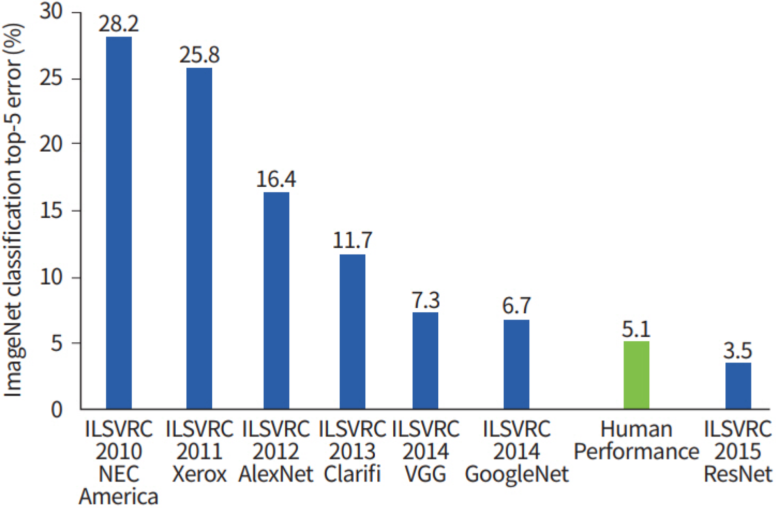
Geoffrey Hinton

Backpropagation: In a 1986 paper, "Learning Internal Representations by Error Propagation," co-authored with David Rumelhart and Ronald Williams, Hinton demonstrated that the backpropagation algorithm allowed neural nets to discover their own internal representations of data, making it possible to use neural nets to solve problems that had previously been thought to be beyond their reach. The backpropagation algorithm is standard in most neural networks today.

Boltzmann Machines: In 1983, with Terrence Sejnowski, Hinton invented Boltzmann Machines, one of the first neural networks capable of learning internal representations in neurons that were not part of the input or output.

Improvements to convolutional neural networks: In 2012, with his students, Alex Krizhevsky and Ilya Sutskever, Hinton improved convolutional neural networks using rectified linear neurons and dropout regularization. In the prominent ImageNet competition, Hinton and his students almost halved the error rate for object recognition and reshaped the computer vision field.

Before and after 2012 - The ImageNet competition



Before and after 2012 - The ImageNet competition

Top 5 ILSVRC 2012 Results		
1st	Error: 16.4%	Deep Learning
2nd	Error: 26.1%	Other approach
3rd	Error: 26.9%	Other approach
4th	Error: 29.5%	Other approach
5th	Error: 34.4%	Other approach
Top 5 ILSVRC 2017 Results		
1st	Error: 2.3%	Deep Learning
2nd	Error: 2.5%	Deep Learning
3rd	Error: 2.7%	Deep Learning
4th	Error: 3.0%	Deep Learning
5th	Error: 3.2%	Deep Learning

Table : Results from ImageNet Large Scale Visual Recognition Competition (ILSVRC).

Strong confidence in deep learning

The New Yorker quotes Geoffrey Hinton (April 2017):

"They should stop training radiologists now."

FDA NEWS RELEASE

FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems

Share | Tweet | Linkedin | Email | Print

For Immediate Release: April 11, 2018

Español

The U.S. Food and Drug Administration today permitted marketing of the first medical device to use artificial intelligence to detect greater than a mild level of the eye disease diabetic retinopathy in adults who have diabetes.

Diabetic retinopathy occurs when high levels of blood sugar lead to damage in the blood vessels of the retina, the light-sensitive tissue in the back of the eye. Diabetic retinopathy is the most common cause of vision loss among the more than 30 million Americans living with diabetes and the leading cause of vision impairment and blindness among working-age adults.

AI replaces algorithms in medical imaging

nature
International journal of science

Letter | Published: 21 March 2018

Image reconstruction by domain-transform manifold learning

Bo Zhu, Jeremiah Z. Liu, Stephen F. Cauley, Bruce R. Rosen & Matthew S. Rosen

Nature **555**, 487–492 (22 March 2018) | Download Citation

Abstract

Image reconstruction is essential for imaging applications across the physical and life sciences, including optical and radar systems, magnetic resonance imaging, X-ray computed tomography, positron

AI replaces algorithms in medical imaging

nature > nature methods > research highlights > article

nature methods

Research Highlights | Published: 27 April 2018

Imaging

AI transforms image reconstruction

Rita Strack

Nature Methods **15**, 309 (2018) | Download Citation

A deep-learning-based approach improves the speed, accuracy, and robustness of biomedical image reconstruction.

Neural nets have excellent properties

The universal approximation theorem:

Theorem 1 (Pinkus, Acta Numerica 1999)

Let $\rho \in C(\mathbb{R})$. Then the set of neural networks is dense in $C(\mathbb{R}^d)$ in the topology of uniform convergence on compact sets, if and only if ρ is not a polynomial.

What could go wrong?

Deep learning is demonstrating super human behaviour.

There is a mathematical theory suggesting that neural nets have all the approximation qualities that are needed.

What could possibly go wrong?

-

AI replacing humans

Adversarial attacks on medical machine learning

Samuel G. Finlayson[1], John D. Bowers[2], Joichi Ito[3], Jonathan L. Zittrain[2], Andrew L. Beam[4], Isaac S. Kohane[1]

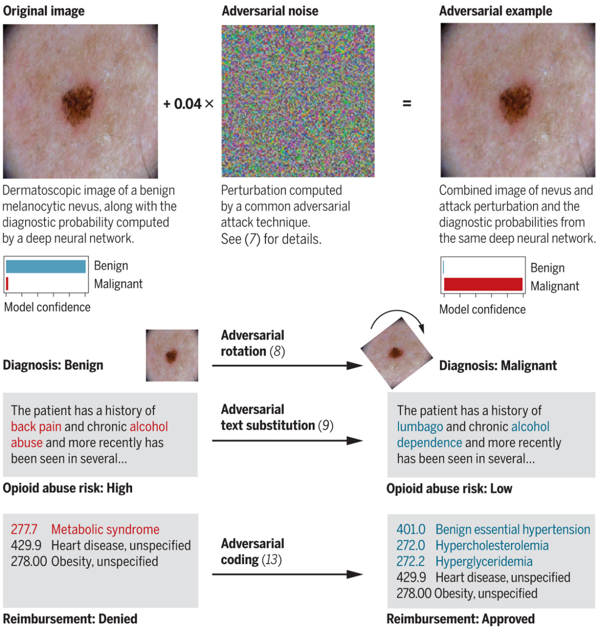
+ See all authors and affiliations

Science 22 Mar 2019:
Vol. 363, Issue 6433, pp. 1287-1289
DOI: 10.1126/science.aaw4399

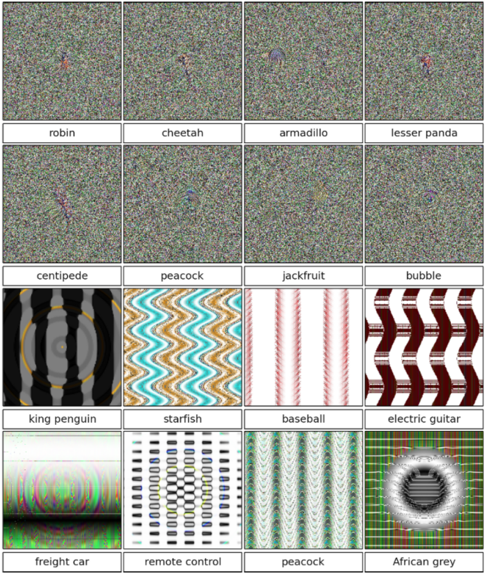
Article | Figures & Data | Info & Metrics | eLetters | PDF

With public and academic attention increasingly focused on the new role of machine learning in the health information economy, an unusual and no-longer-esoteric category of vulnerabilities in machine-learning systems could prove important. These vulnerabilities allow a small, carefully designed change in how inputs are presented to a system to completely alter its output, causing it to confidently arrive at manifestly wrong conclusions. These advanced techniques to subvert otherwise-reliable machine-learning systems—so-called adversarial attacks—have, to date, been of interest primarily to computer science researchers (*1*). However, the landscape of often-competing interests within health care, and billions of dollars at stake in systems' outputs, implies considerable problems. We outline motivations that various players in the health care system may have to use adversarial attacks and begin a discussion of what to do about them. Far from discouraging continued innovation with medical machine learning, we call for active engagement of medical, technical, legal, and ethical experts in pursuit of efficient, broadly available, and effective health care that machine learning will enable.

What could go wrong?



What has deep learning actually learned?



"Deep neural networks are easily fooled: High confidence predictions for unrecognizable images", A. Nguyen, J. Yosinski, and J. Clune. *2015 IEEE Conference on Computer Vision and Pattern Recognition.*

Deep Fool was established at EPFL in order to study the stability of neural networks.

DEEP LEARNING FOR VISUAL UNDERSTANDING

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli,
and Pascal Frossard

The Robustness of Deep Networks

A geometrical perspective

Deep Fool in practice

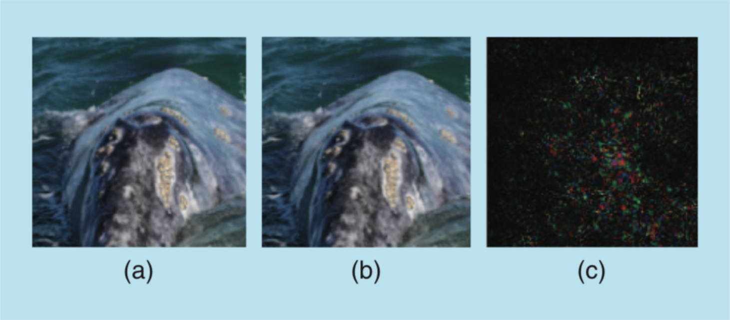


FIGURE 1. An example of an adversarial perturbations in state-of-the-art neural networks. (a) The original image that is classified as a "whale," (b) the perturbed image classified as a "turtle," and (c) the corresponding adversarial perturbation that has been added to the original image to fool a state-of-the-art image classifier [5].

Deep Fool: Universal perturbations

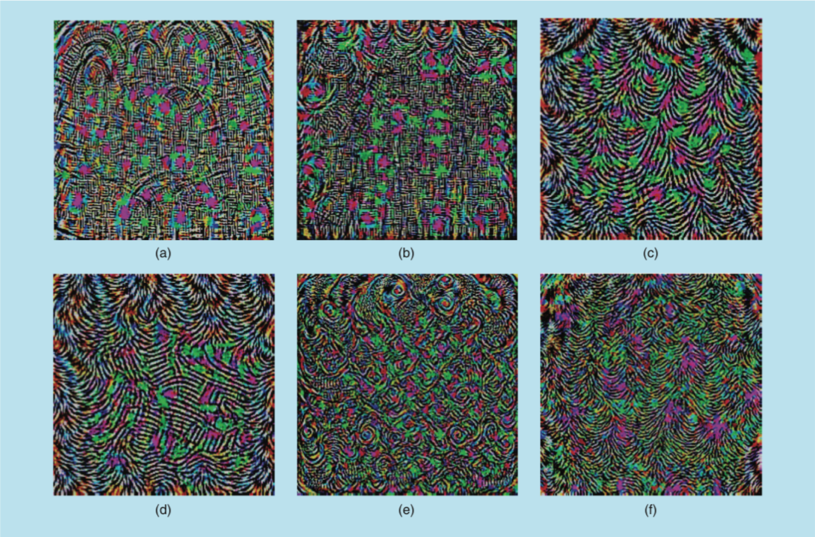


FIGURE 3. Universal perturbations computed for different deep neural network architectures. The pixel values are scaled for visibility. (a) CaffeNet, (b) VGG-F, (c) VGG-16, (d) VGG-19, (e) GoogLeNet, and (f) ResNet-152.

Deep Fool: Examples

FIGURE 4. Examples of natural images perturbed with the universal perturbation and their corresponding estimated labels with GoogLeNet. (a)–(h) Images belonging to the ILSVRC 2012 validation set. (i)–(l) Personal images captured by a mobile phone camera. (Figure used courtesy of [22].)

Structural perturbations

Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt*[1], Ivan Evtimov*[2], Earlence Fernandes[2], Bo Li[3], Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], and Dawn Song[3]

[1]University of Michigan, Ann Arbor
[2]University of Washington
[3]University of California, Berkeley
[4]Samsung Research America and Stony Brook University

Abstract

Recent studies show that the state-of-the-art deep neural networks (DNNs) are vulnerable to adversarial examples, resulting from small-magnitude perturbations added to the input. Given that that emerging physical systems are using DNNs in safety-critical situations, adversarial examples could mislead these systems and cause dangerous situations. Therefore, understanding adversarial examples in the physi-

these successes, they are increasingly being used as part of control pipelines in physical systems such as cars [8, 17], UAVs [4, 24], and robots [40]. Recent work, however, has demonstrated that DNNs are vulnerable to adversarial perturbations [5, 9, 10, 15, 16, 22, 25, 29, 30, 35]. These carefully crafted modifications to the (visual) input of DNNs can cause the systems they control to misbehave in unexpected and potentially dangerous ways.

Structural perturbations

Structural perturbations can also cause the network to fail.

What could possibly go wrong?

-

AI replacing standard algorithms

nature
International journal of science

Letter | Published: 21 March 2018

Image reconstruction by domain-transform manifold learning

Bo Zhu, Jeremiah Z. Liu, Stephen F. Cauley, Bruce R. Rosen & Matthew S. Rosen

Nature **555**, 487–492 (22 March 2018) | Download Citation

Abstract

Image reconstruction is essential for imaging applications across the physical and life sciences, including optical and radar systems, magnetic resonance imaging, X-ray computed tomography, positron

Transforming image reconstruction with AI

nature > nature methods > research highlights > article

nature methods

Research Highlights | Published: 27 April 2018

Imaging

AI transforms image reconstruction

Rita Strack

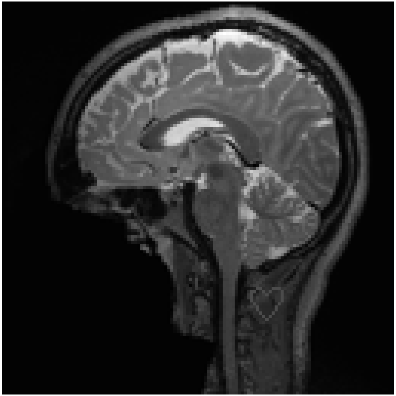
Nature Methods **15**, 309 (2018) | Download Citation

A deep-learning-based approach improves the speed, accuracy, and robustness of biomedical image reconstruction.

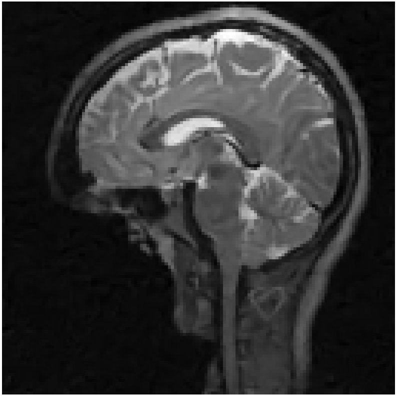
Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original



AUTOMAP Network

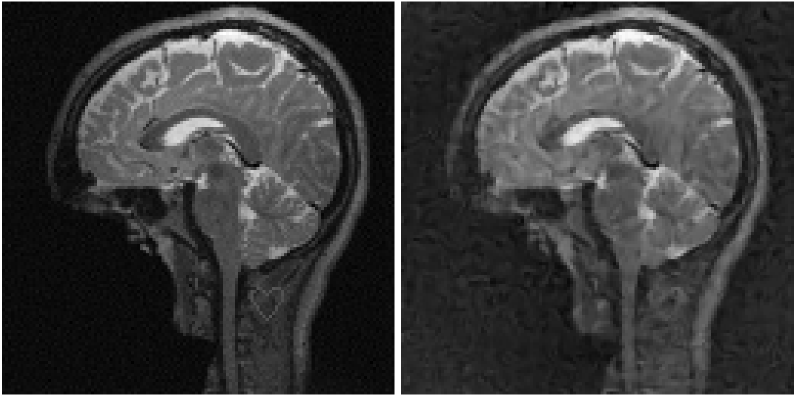


AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original + tiny pert. **AUTOMAP Network**

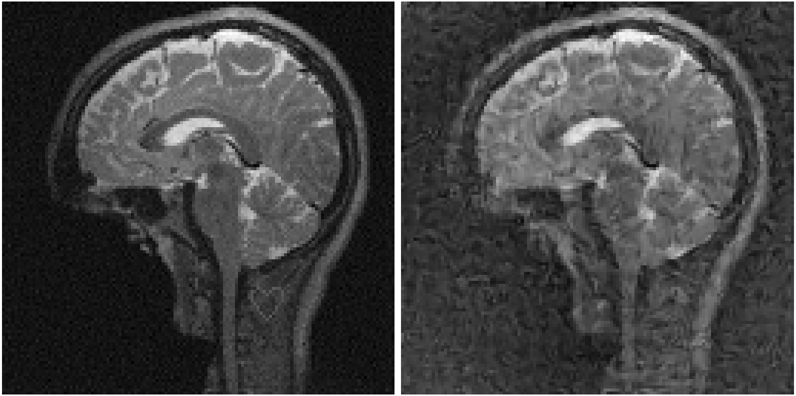


AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original + tiny pert. **AUTOMAP Network**

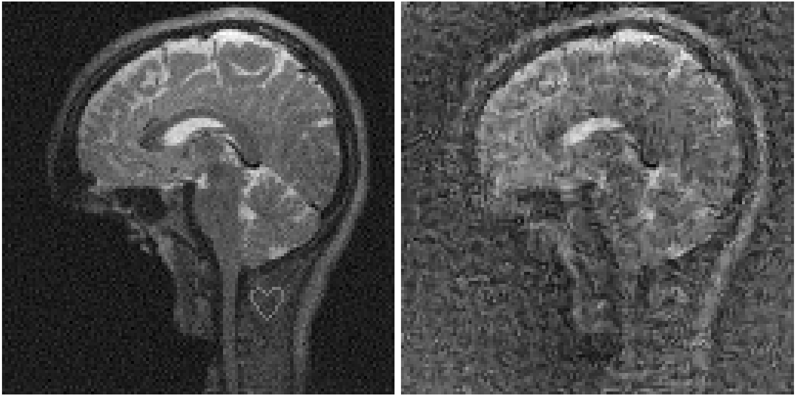


AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original + tiny pert. **AUTOMAP Network**

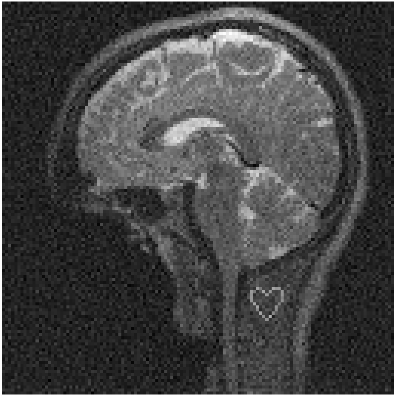


AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original + tiny pert.



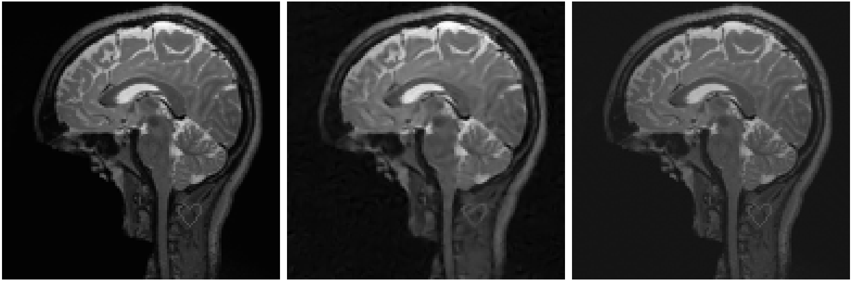
AUTOMAP Network

AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Comparison with state-of-the-art

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original **AUTOMAP Network** **State-of-the-art**



AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Comparison with state-of-the-art

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original + tiny pert. | **AUTOMAP Network** | **State-of-the-art**

AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Comparison with state-of-the-art

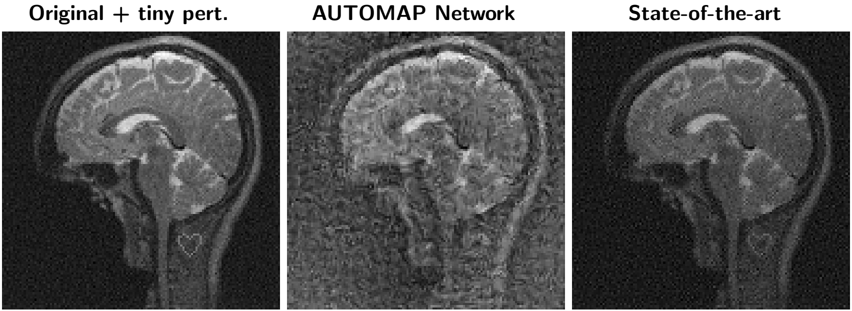
Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original + tiny pert. **AUTOMAP Network** **State-of-the-art**

AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Comparison with state-of-the-art

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen



AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Comparison with state-of-the-art

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

Original + tiny pert. | **AUTOMAP Network** | **State-of-the-art**

AUTOMAP network from "Image reconstruction by domain-transform manifold learning", B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen. *Nature* (March. 2018).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

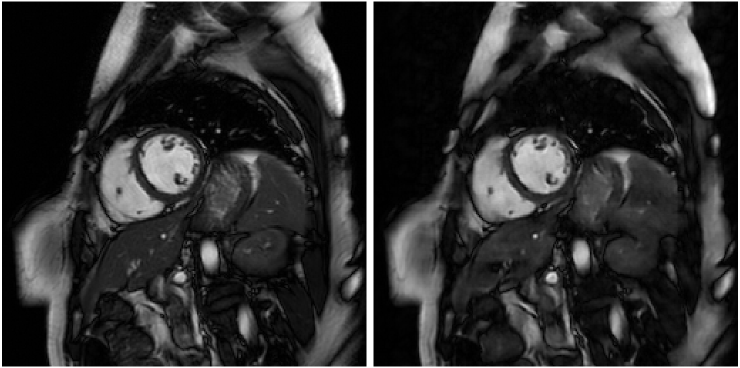


Figure : Left: Original image + tiny perturbation. Right: Reconstruction (25 % subsampling).

Neural net from "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction", J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert *IEEE Trans. Med. Imag.* (to appear).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

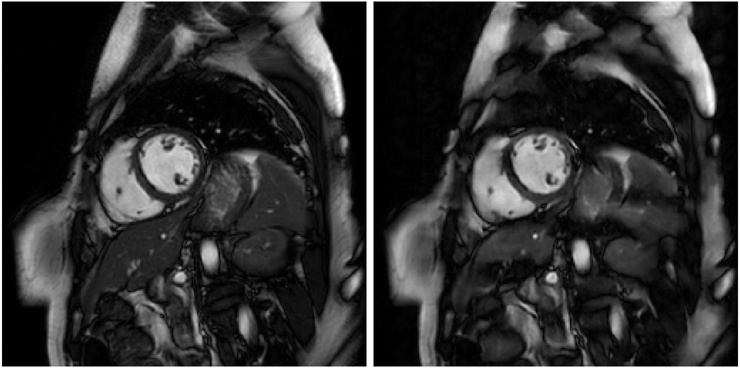


Figure : Left: Original image + tiny perturbation. Right: Reconstruction (25 % subsampling).

Neural net from "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction", J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert *IEEE Trans. Med. Imag.* (to appear).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

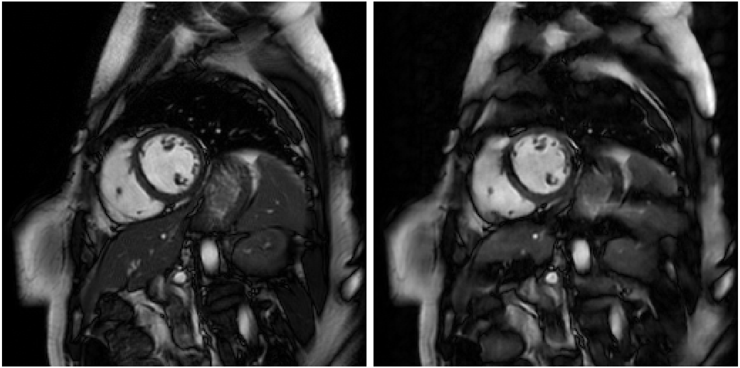


Figure : Left: Original image + tiny perturbation. Right: Reconstruction (25 % subsampling).

Neural net from "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction", J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert *IEEE Trans. Med. Imag.* (to appear).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

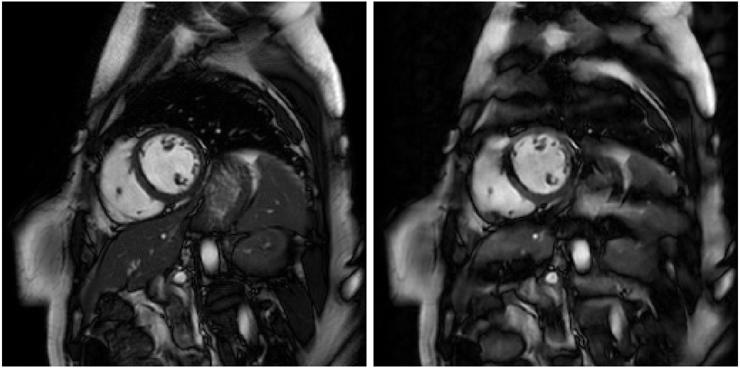


Figure : Left: Original image + tiny perturbation. Right: Reconstruction (25 % subsampling).

Neural net from "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction", J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert *IEEE Trans. Med. Imag.* (to appear).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

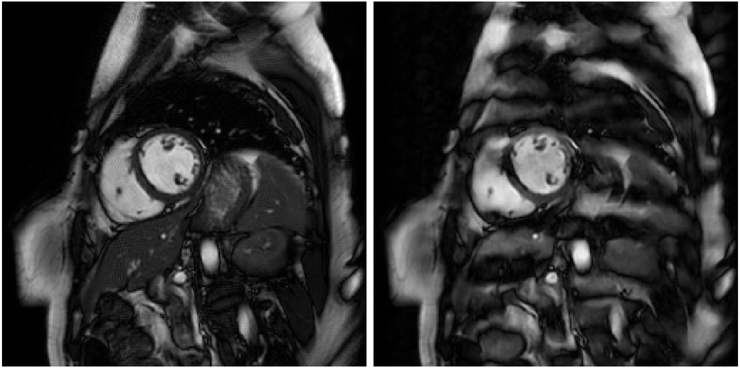


Figure : Left: Original image + tiny perturbation. Right: Reconstruction (25 % subsampling).

Neural net from "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction", J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert *IEEE Trans. Med. Imag.* (to appear).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

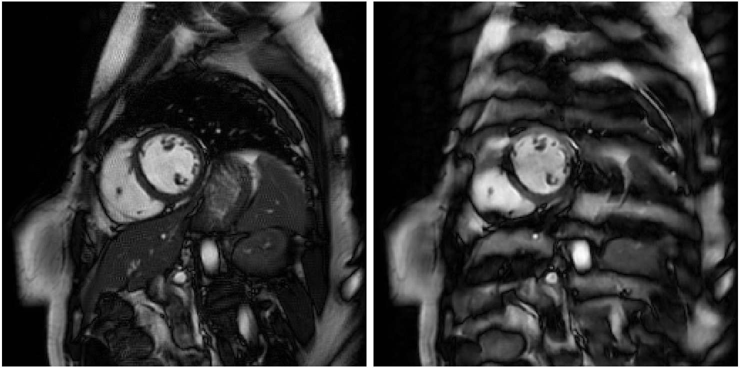


Figure : Left: Original image + tiny perturbation. Right: Reconstruction (25 % subsampling).

Neural net from "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction", J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert *IEEE Trans. Med. Imag.* (to appear).

Instability of DL in Inverse Problems - MRI

Experiment from "On instabilities of deep learning in image reconstruction - Does AI come at a cost?", V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen

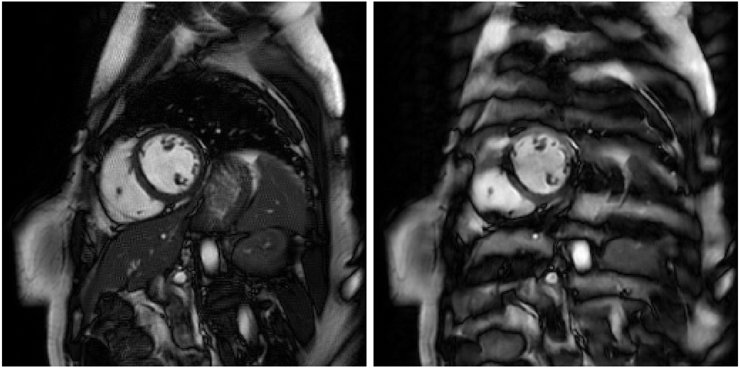
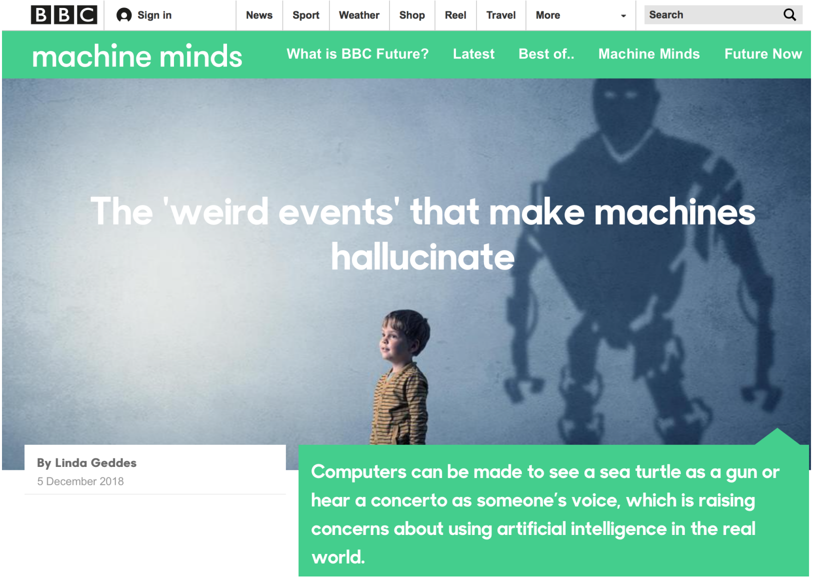


Figure : Left: Original image + tiny perturbation. Right: Reconstruction (25 % subsampling).

Neural net from "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction", J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert *IEEE Trans. Med. Imag.* (to appear).

Instabilities in Deep Learning


BBC | Sign in | News | Sport | Weather | Shop | Reel | Travel | More | Search

machine minds | What is BBC Future? | Latest | Best of.. | Machine Minds | Future Now

The 'weird events' that make machines hallucinate

By Linda Geddes
5 December 2018

Computers can be made to see a sea turtle as a gun or hear a concerto as someone's voice, which is raising concerns about using artificial intelligence in the real world.

Do we know what we are doing?

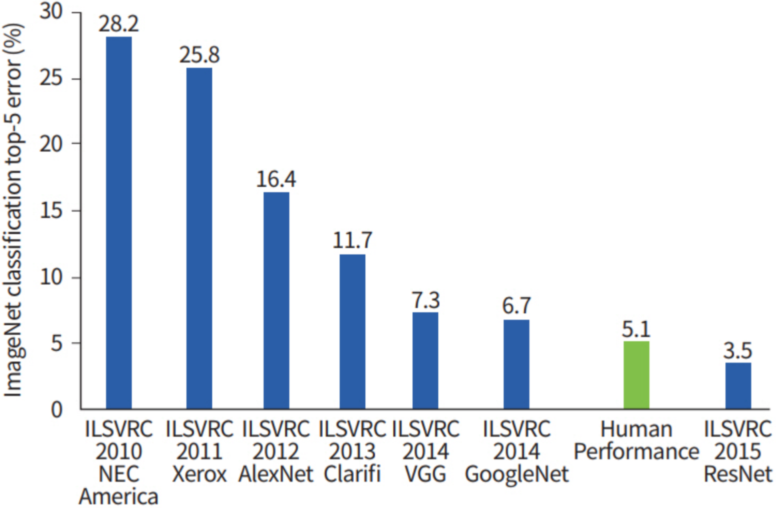
Google's Ali Rahimi, winner of the Test-of-Time award 2017 (NIPS), "Machine learning has become alchemy. ... I would like to live in a society whose systems are built on top of verifiable, rigorous, thorough knowledge, and not on alchemy."

Yann LeCun
December 6 at 8:57am ·

My take on Ali Rahimi's "Test of Time" award talk at NIPS.

Ali gave an entertaining and well-delivered talk. But I fundamentally disagree with the message.
The main message was, in essence, that the current practice in machine learning is akin to "alchemy" (his word).
It's insulting, yes. But never mind that: It's wrong!

Before and after 2012 - The ImageNet competition



Strong confidence in deep learning

The New Yorker quotes Geoffrey Hinton (April 2017):

"They should stop training radiologists now."

What does deep learning actually learn?

False structures in classification

Conjecture 1 (False structures in classification)

The current training process in deep learning for classification forces the neural network to learn a different (false) structure and not the actual structure of the classification problem. There are three main components:

(i) *(**Success**) The false structure correlates well with the original structure, hence one gets a high success rate.*

(ii) *(**Instability**) The false structure is unstable, and thus the network is susceptible to adversarial attacks.*

(iii) *(**Simplicity**) The false structure is much simpler than the desired structure, and hence easer to learn e.g. fewer data are needed and the numerical algorithm used in the training easily converges to the neural network that captures the false structure.*

A thought experiment

Consequences of Conjecture 1

Negative consequences:

(i) The success of deep learning in classification is not due to the fact that networks learn the structures that humans associate with image recognition, but rather that the network picks up unstable false structures in images that are potentially impossible for humans to detect. This means that instability, and hence vulnerability to adversarial attacks, can never be removed until one guarantees that no false structure is learned. This means a potential complete overhaul of modern AI.

(ii) The success is dependent of the simple yet unstable structures, thus the AI does not capture the intelligence of a human.

(iii) Since one does not know which structure the network picks up, it becomes hard to conclude what the neural network actually learns, and thus harder to trust its prediction. What if the false structure gives wrong predictions?

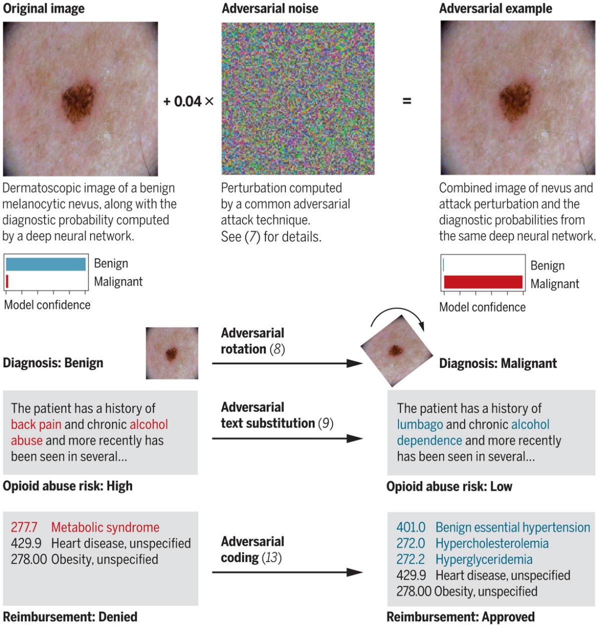
Consequences of Conjecture 1

Positive consequences:

(I) Deep learning captures structures that humans cannot detect, and these structures require very little data and computing power in comparison to the true original structures, however, they generalise rather well compared to the original structure. Thus, from an efficiency point of view, the human brain may be a complete overkill for certain classification problems, and deep learning finds a mysterious effective way of classifying.

(II) The structure learned by deep learning may have information that the human may not capture. This structure could be useful if characterised properly. For example, what if there is structural information in the data that allows for accurate prediction that the original structure could not do?

What could go wrong?



Worst case v.s. average performance

What are the appropriate tests for AI?

Tests will depend on application and community:

- Defence and intelligence community
- Healthcare community
- Public sector management community

Should AI products be labeled with warnings describing the potential non-human behaviour?