Have we achieved artificial intelligence?
-
On the paradoxes of deep learning

Anders C. Hansen (Cambridge)

Joint work with:

B. Adcock (SFU) V. Antun (UiO)
A. Bastounis (Cambridge) C. Poon (Cambridge)
V. Vlacic (ETH)

Cambridge, May 2018

Sneak peak from the book
"Structured Compressed Sensing, Imaging and Learning",
Adcock & Hansen (Cambridge University Press (2019))

Deep Learning

Deep Learning

Deep learning is typically used for image classification.


Otter: 0.993142
Beaver: 0.00231697
Mink: 0.00199465

Neural networks

Let $\mathcal{NN}_{\mathbf{N},L,d}$, with $\mathbf{N} = (N_L, N_{L-1}, \ldots, N_1, N_0 = d)$ denote the set of all L-layer neural networks. That is, all mappings $\phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ of the form

$$\phi(x) = W_L(\rho(W_{L-1}(\rho(\ldots \rho(W_1(x)))))), \quad x \in \mathbb{R}^d.$$

$$W_j y = A_j y - b_j, \qquad A_j \in \mathbb{R}^{N_j \times N_{j-1}}, \quad b_j \in \mathbb{R}^{N_j}$$

$$\rho : \mathbb{R} \to \mathbb{R}$$

is some non-linear function that acts pointwise on a vector.

Training neural nets

Given a classification function $f : \mathbb{R}^d \to \{0, 1\}$, a training set $\mathcal{T} = \{x^1, \ldots, x^r\} \subset \mathbb{R}^d$, a classification set $\mathcal{C} = \{y^1, \ldots, y^s\}$, and a cost function $C : \mathbb{R}^r \times \mathbb{R}^r \to \mathbb{R}_+$, compute

$$\phi \in \underset{\tilde{\phi} \in \mathcal{NN}_{\mathbf{N},L,d}}{\operatorname{argmin}} \; C(v, w),$$

with

$$v = \{\tilde{\phi}(x^j)\}_{j=1}^r, \quad w = \{f(x^j)\}_{j=1}^r.$$

Typical choice is $C(v, w) = \|v - w\|_p^p$.

Deep Learning and Super-human Behaviour

- Deep learning is the state of the art method for image recognition ($< 2.5\%$ failure rate.)
- Humans typically do not get better than 5% failure rate.

Have we achieved artificial intelligence?

Smale's 18th problem

What are the limits of intelligence, both artificial and human?

"Learning is a part of human intelligent activity. The corresponding mathematics is suggested by the theory of repeated games, neural nets and genetic algorithms."

Turing's imitation game

Turing:

"I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words."

Turing's imitation game

Turing:

"The new form of the problem can be described in terms of a game which we call the 'imitation game." It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B."

" We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?""

The imitation game and deep learning

What would be a key feature that humans have when classifying images that computers may not have?

The imitation game and deep learning

What would be a key feature that humans have when classifying images that computers may not have?

Answer: Stability

The Paradox of Deep Learning

There is an uncountable family of classification functions $f : \mathbb{R}^{N_0} \to \{0, 1\}$ such that for any neural network dimensions $\mathbf{N} = (N_L, N_{L-1}, \ldots, N_1, N_0)$ with $N_0, L \geq 2$ and any $0 < \epsilon < 1/(K + M)$ where M is arbitrarily large and $K \geq 3(N_1 + 1) \cdots (N_{L-1} + 1)$ we have the following. There exist uncountably many training sets $\mathcal{T} = \{x^1, \ldots, x^K\}$ and uncountably many classification sets $\mathcal{C} = \{y^1, \ldots, y^M\}$ such that there is a

$$\tilde{\phi} \in \operatorname*{argmin}_{\phi \in \mathcal{NN}_{\mathbf{N},L}} C(v, w), \quad v_j = \phi(x^j), \quad w_j = f(x^j),$$

where $1 \leq j \leq K$ such that

$$\tilde{\phi}(x) = f(x) \quad \forall x \in \mathcal{T} \cup \mathcal{C}.$$

However, there exists uncountably many $v \in \mathbb{R}^{N_0}$ such that

$$|\tilde{\phi}(v) - f(v)| \geq 1/2, \qquad \|v - x\|_\infty \leq \epsilon \text{ for some } x \in \mathcal{T}.$$

Moreover, there is a neural network $\hat{\phi}$, not necessarily trained, such that

$$\hat{\phi}(x) = f(x) \quad \forall x \in \mathcal{B}_\epsilon^\infty(\mathcal{T} \cup \mathcal{C}).$$

The Paradox in Practice: Deep Fool

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

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

The Robustness of Deep Networks

A geometrical perspective

The Paradoxes in Practice: Deep Fool

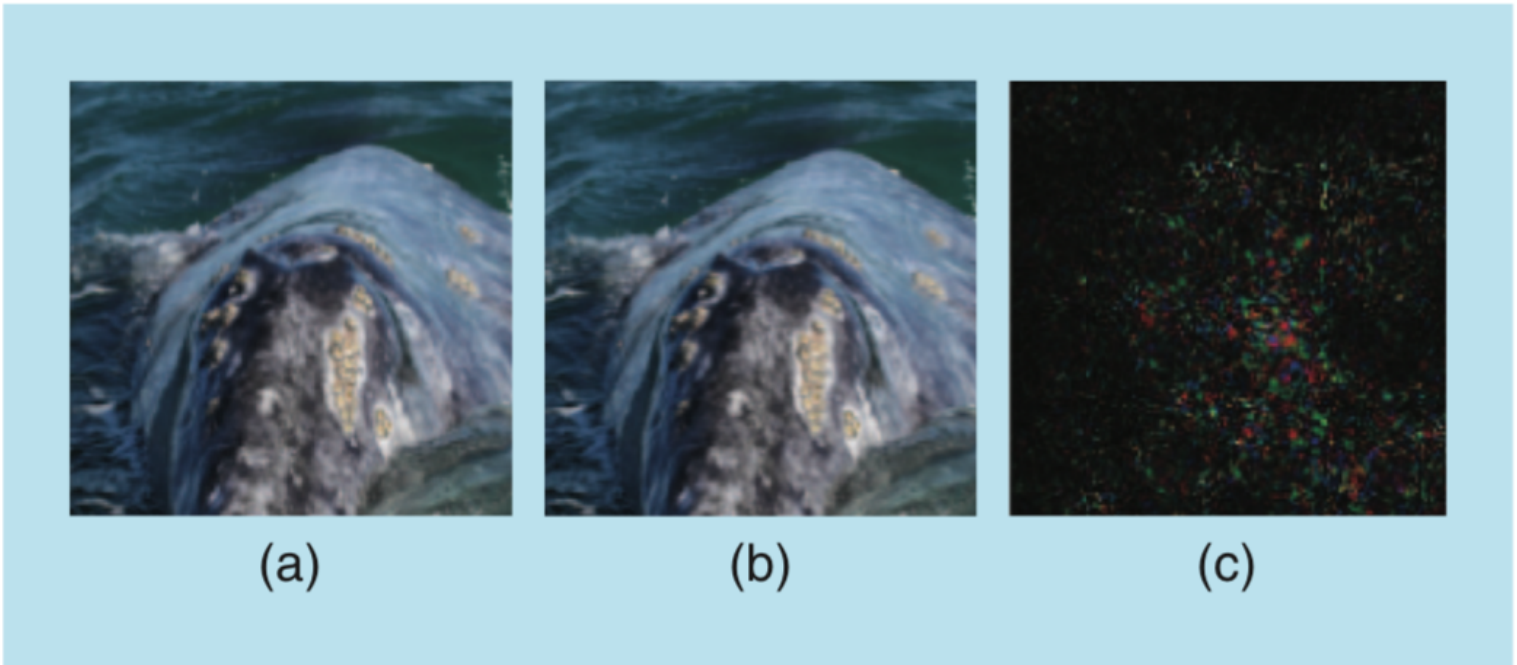


FIGURE 1. An example of an adversarial perturbations in state-of-the-art neural networks. (a) The original image that is classified as a "whale," (b) the perturbed image classified as a "turtle," and (c) the corresponding adversarial perturbation that has been added to the original image to fool a state-of-the-art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

Deep Fool: Universal perturbations

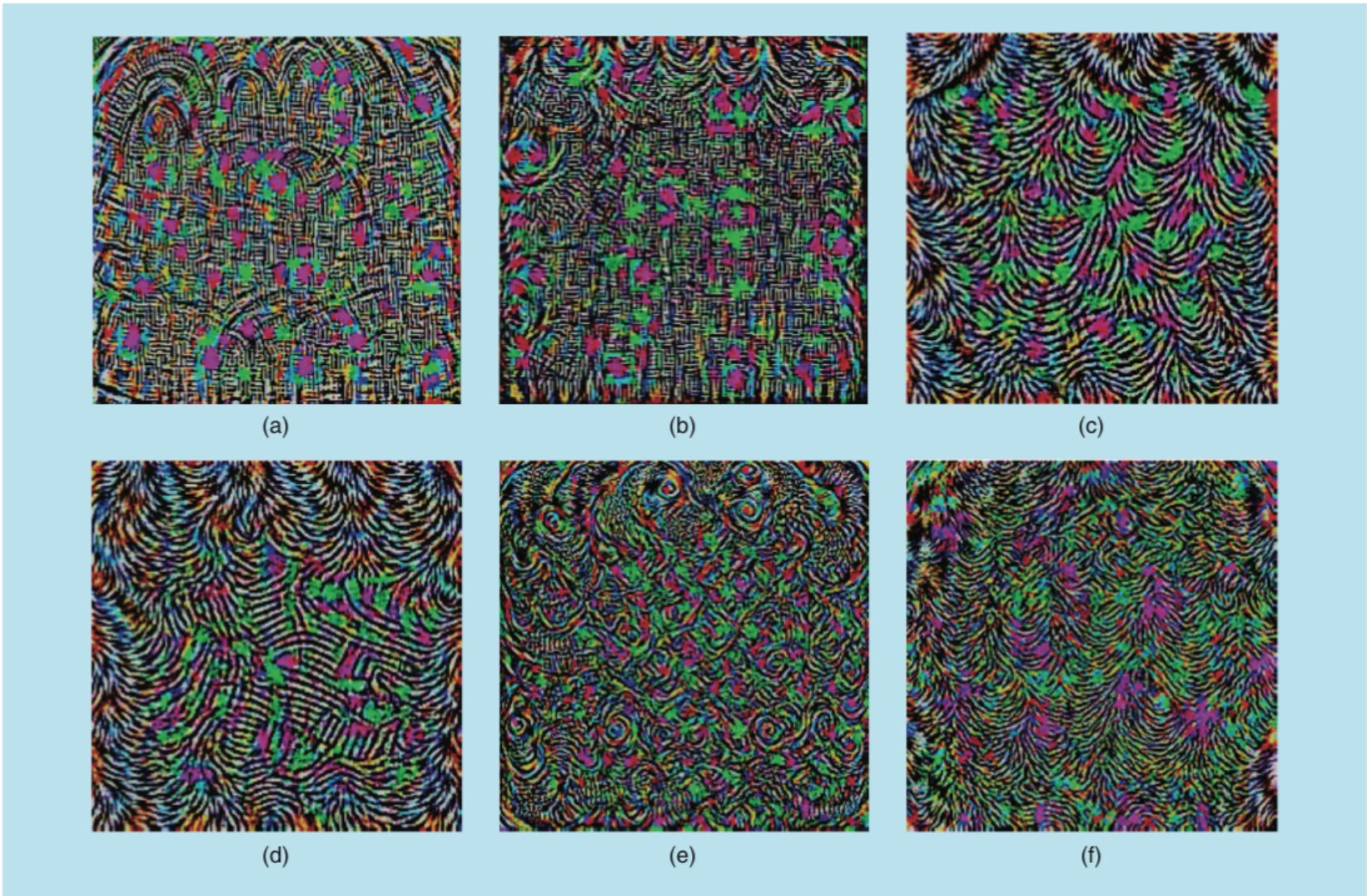


FIGURE 3. Universal perturbations computed for different deep neural network architectures. The pixel values are scaled for visibility. (a) CaffeNet, (b) VGG-F, (c) VGG-16, (d) VGG-19, (e) GoogLeNet, and (f) ResNet-152.

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].)

Deep Fool: Examples

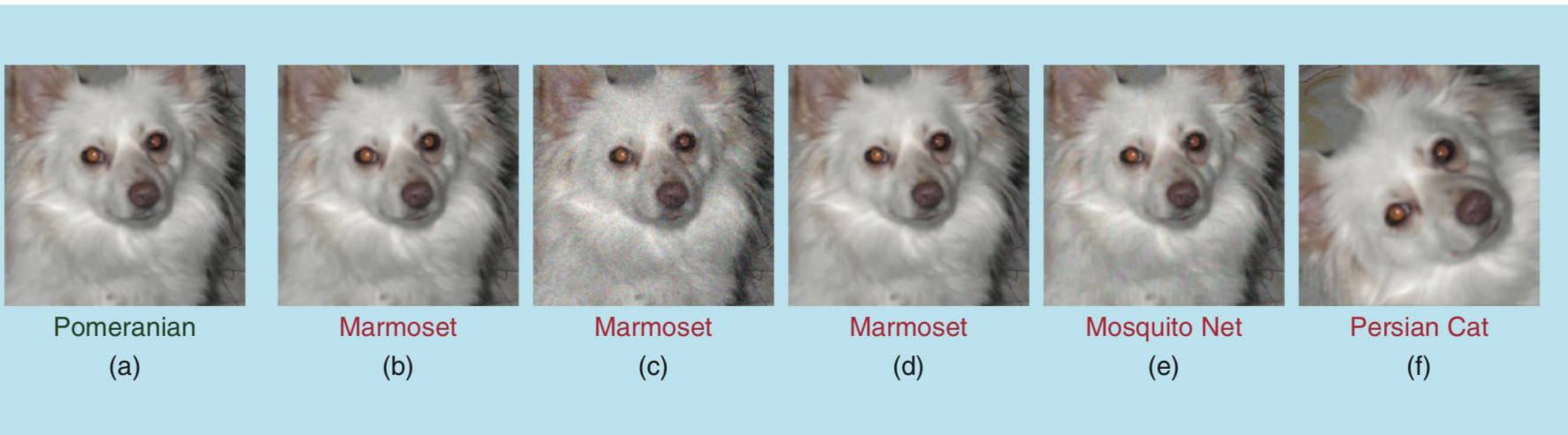


FIGURE 5. (a) The original image. The remaining images are minimally perturbed images (along with the corresponding estimated label) that misclassify the CaffeNet deep neural network. (b) Adversarial perturbation, (c) random noise, (d) semirandom noise with $m = 1{,}000$, (e) universal perturbation, (f) affine transformation. (Figure used courtesy of [17].)

Consequences of instabilities of deep learning

Practical consequences:

- Surveillance
- Security
- Driverless cars

Philosophical consequences:

- Can this be fixed?
- Legal implications
- FDA approval of medical equipment?

Consequences of instabilities of deep learning?

Roboter og kunstig intelligens (KI) er allerede i omfattende bruk i helsevesenet. Men det er langt igjen til at legene kan erstattes. Foto: Jeff Pachoud/AFP photo/NTB scanpix

Nyheter Helse

Kunstig intelligens: Forsøk stoppet - foreslo livsfarlig medisin

Dagens Næringsliv
Publisert: 23.10.2017 – 08:45 Oppdatert: 23.10.2017 – 09:03

Consequences of instabilities of deep learning?

Uber's self-driving car saw the pedestrian but didn't swerve – report

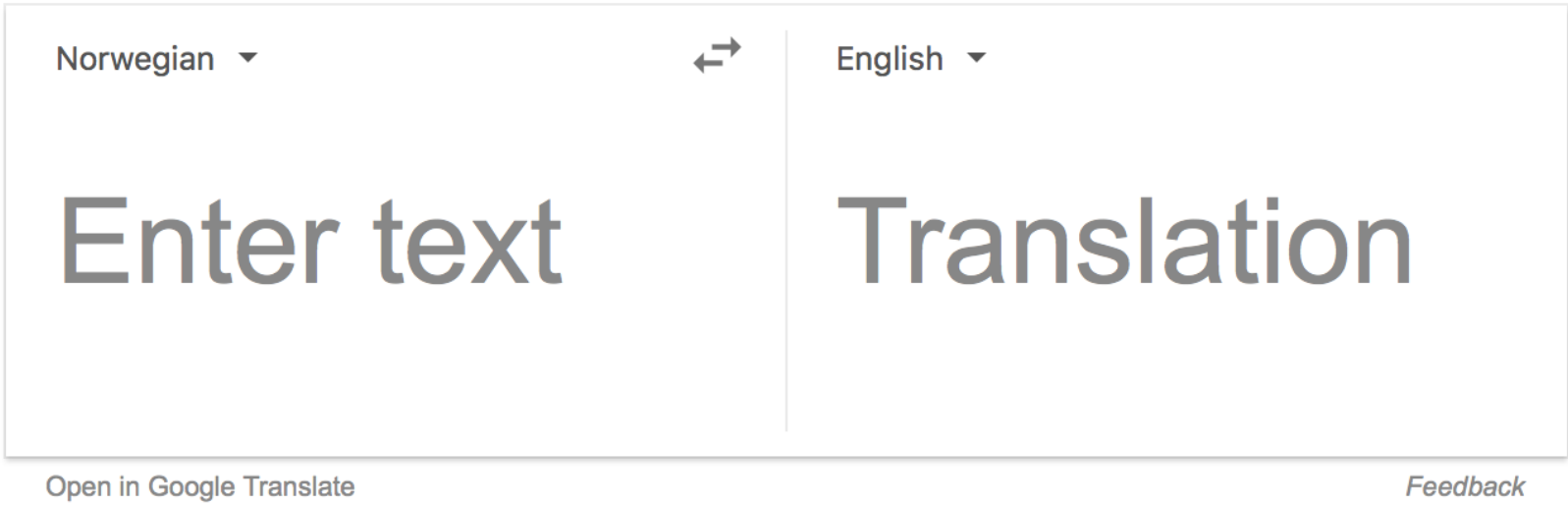
Tuning of car's software to avoid false positives blamed, as US National Transportation Safety Board investigation continues

▲ Uber's modified Volvo XC90 SUV detected but did not react to the crossing pedestrian in first self-driving car fatality, report says. Photograph: Volvo

An Uber self-driving test car which killed a woman crossing the street detected her but decided not to react immediately, a report has said.

The car was travelling at 40mph (64km/h) in self-driving mode when it collided with 49-year-old Elaine Herzberg at about 10pm on 18 March. Herzberg was pushing a bicycle across the road outside of a crossing. She later died from her injuries.

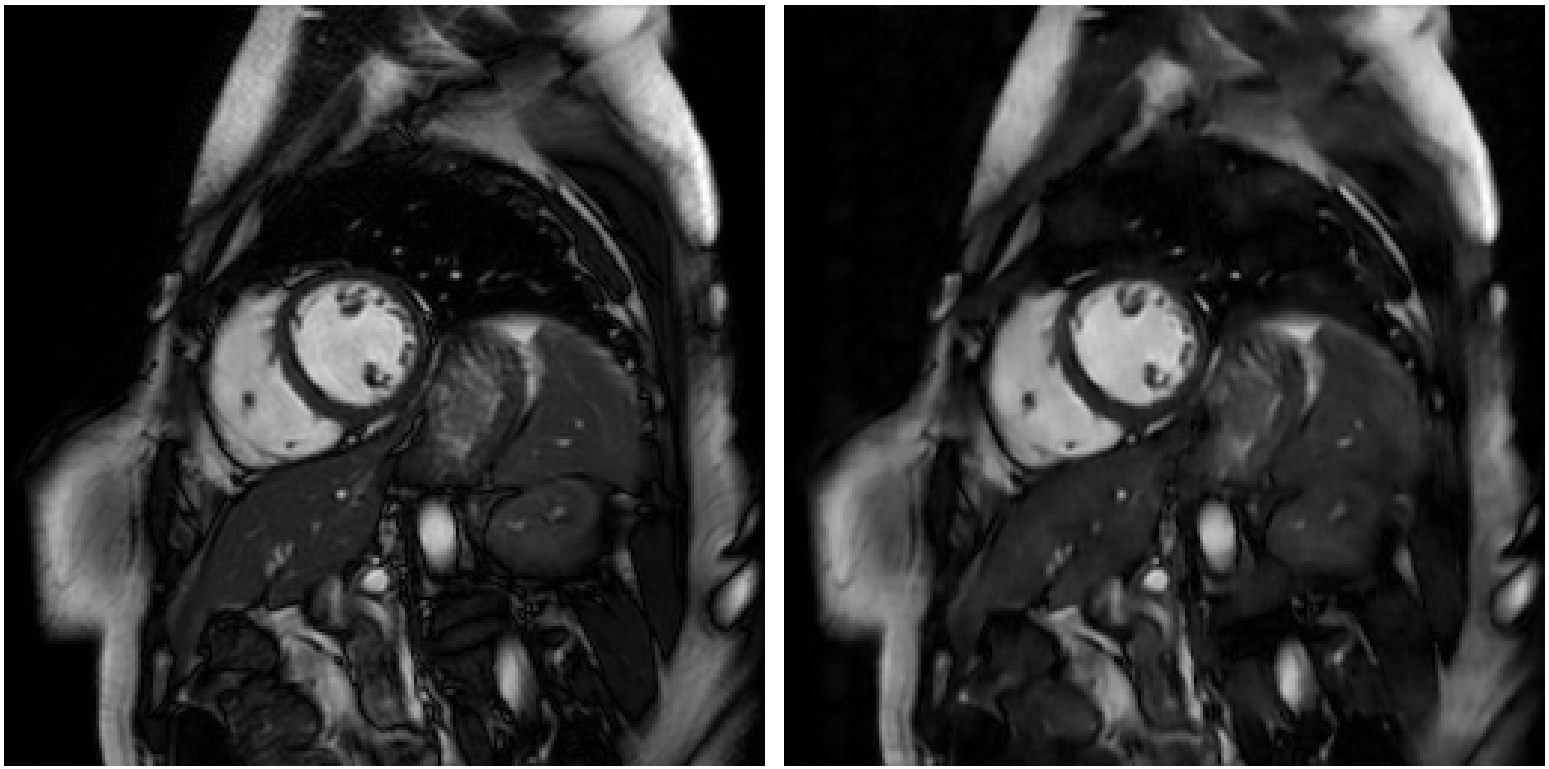
Consequences of instabilities of deep learning?



Is the instability phenomenon with Deep Learning universal?

Deep MRI Net – Stability?

Time saving experiment in MRI - 25% subsampling

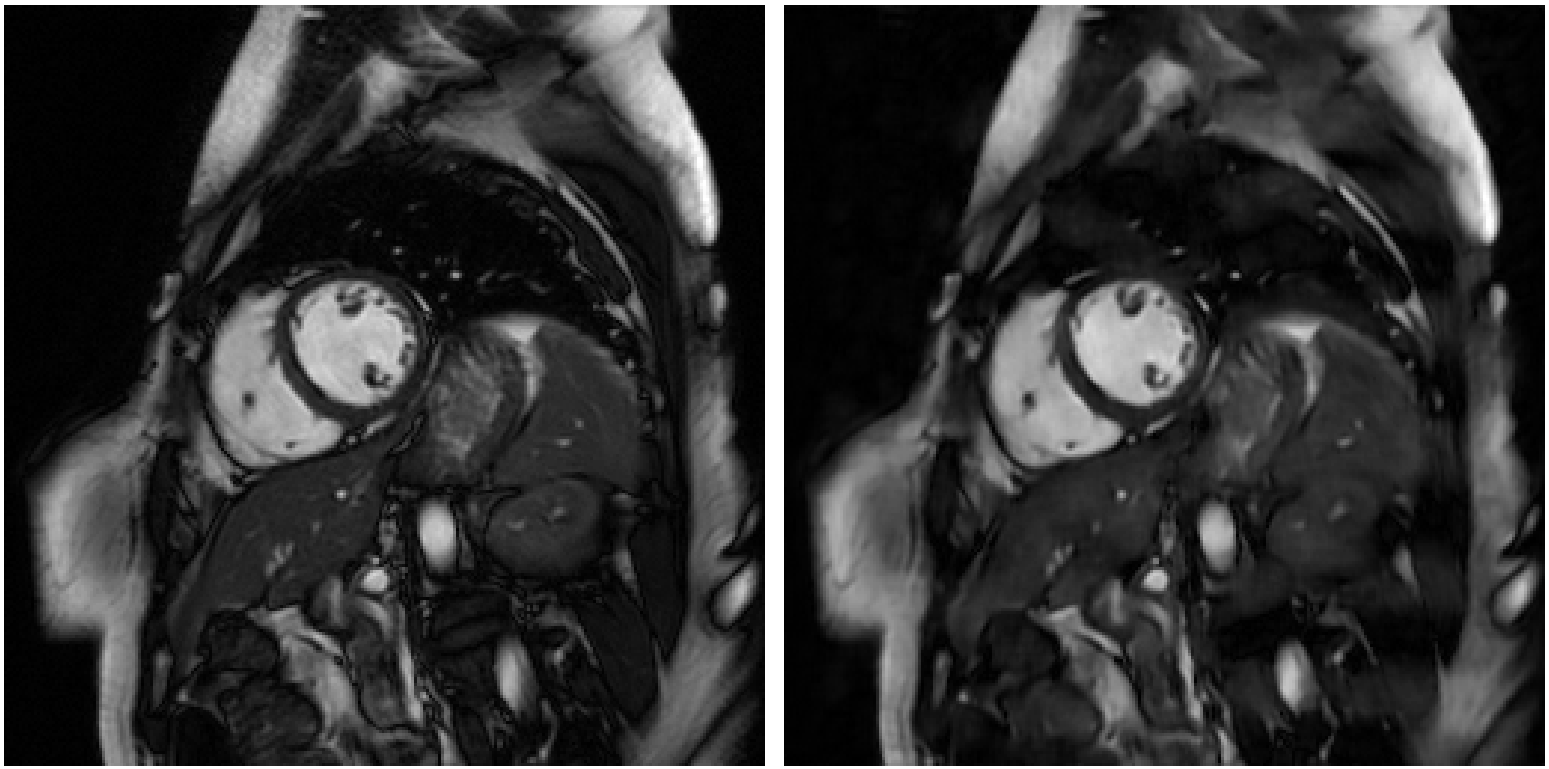


(a) x (b) $\phi(Ax)$ (Neural net rec.)

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).

Deep MRI Net – Stability?

Time saving experiment in MRI - 25% subsampling

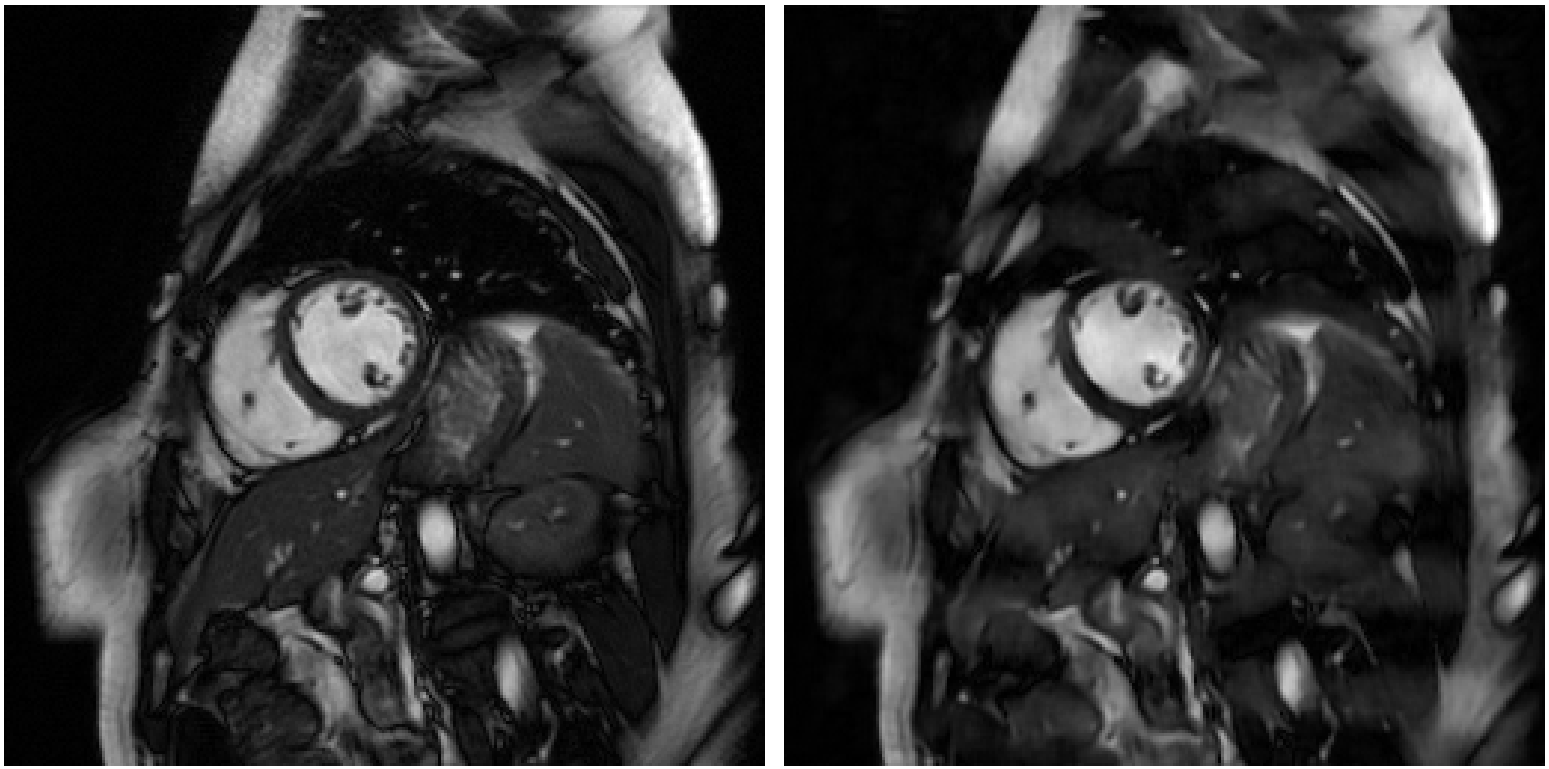


(a) $x + r$ (b) $\phi(A(x + r))$ (Neur. net)

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).

Deep MRI Net – Stability?

Time saving experiment in MRI - 25% subsampling

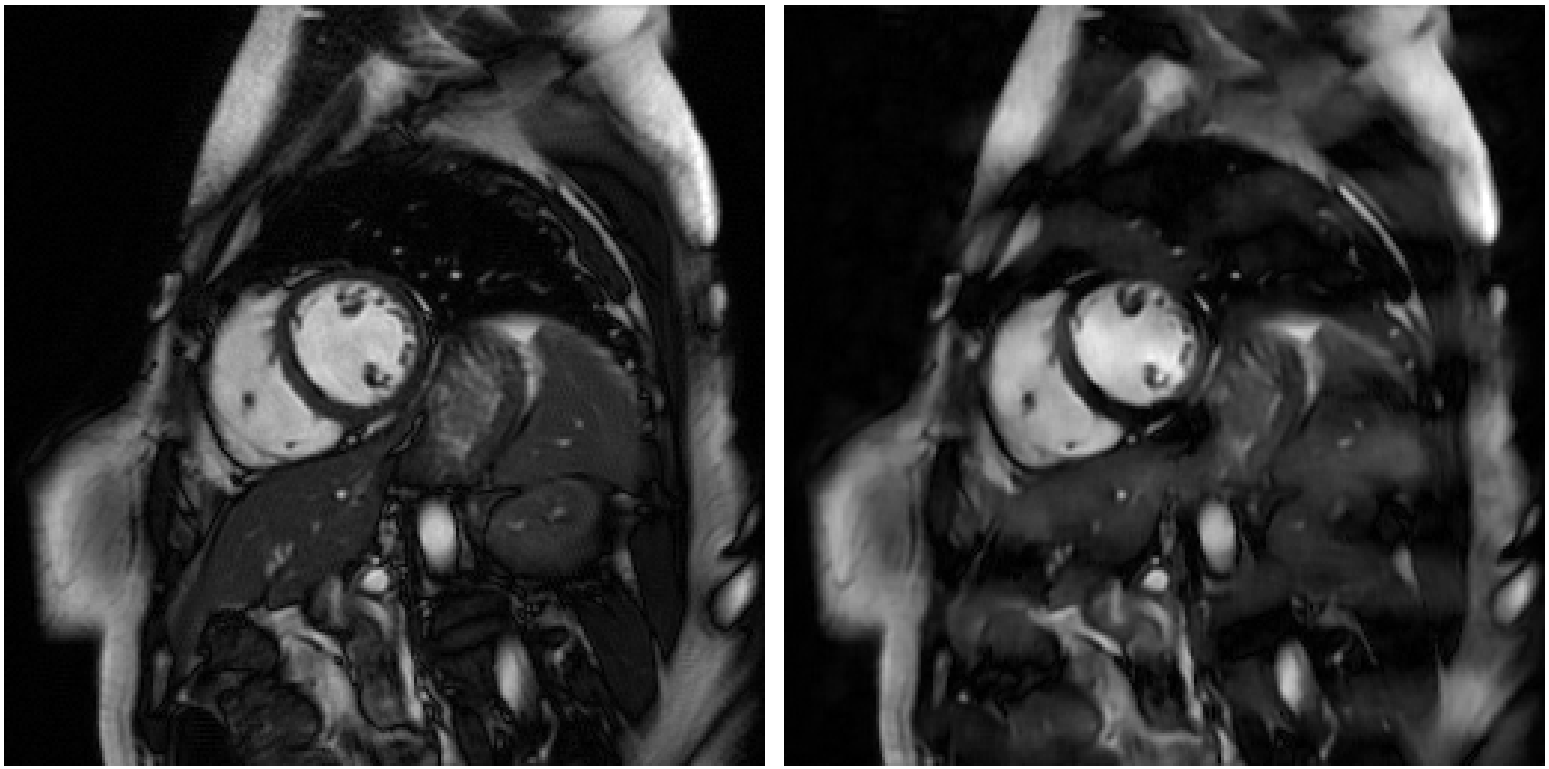


(a) $x + r$ (b) $\phi(A(x + r))$ (Neur. net)

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).

Deep MRI Net – Stability?

Time saving experiment in MRI - 25% subsampling

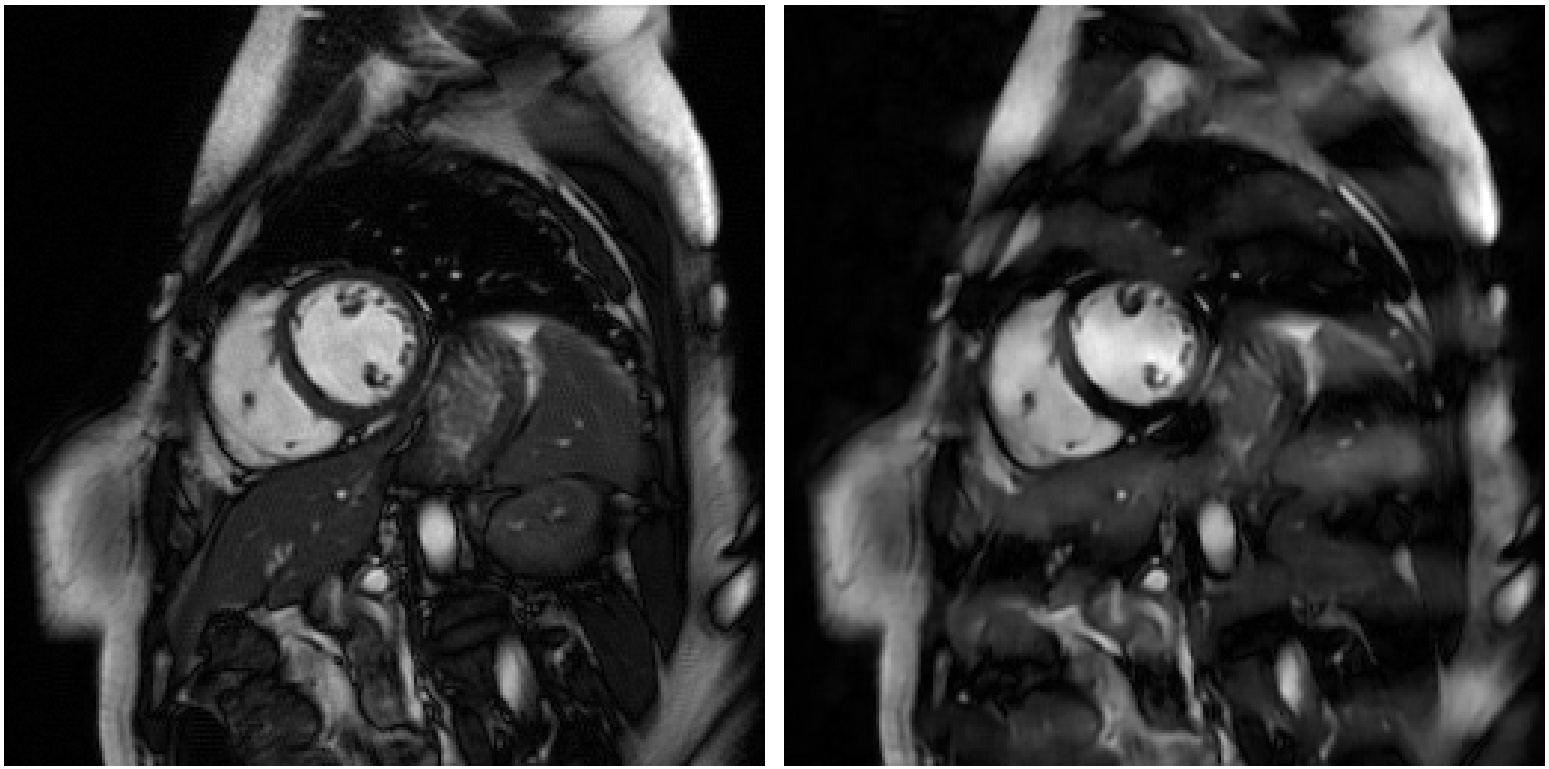


(a) $x + r$ (b) $\phi(A(x + r))$ (Neur. net)

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).

Deep MRI Net – Stability?

Time saving experiment in MRI - 25% subsampling

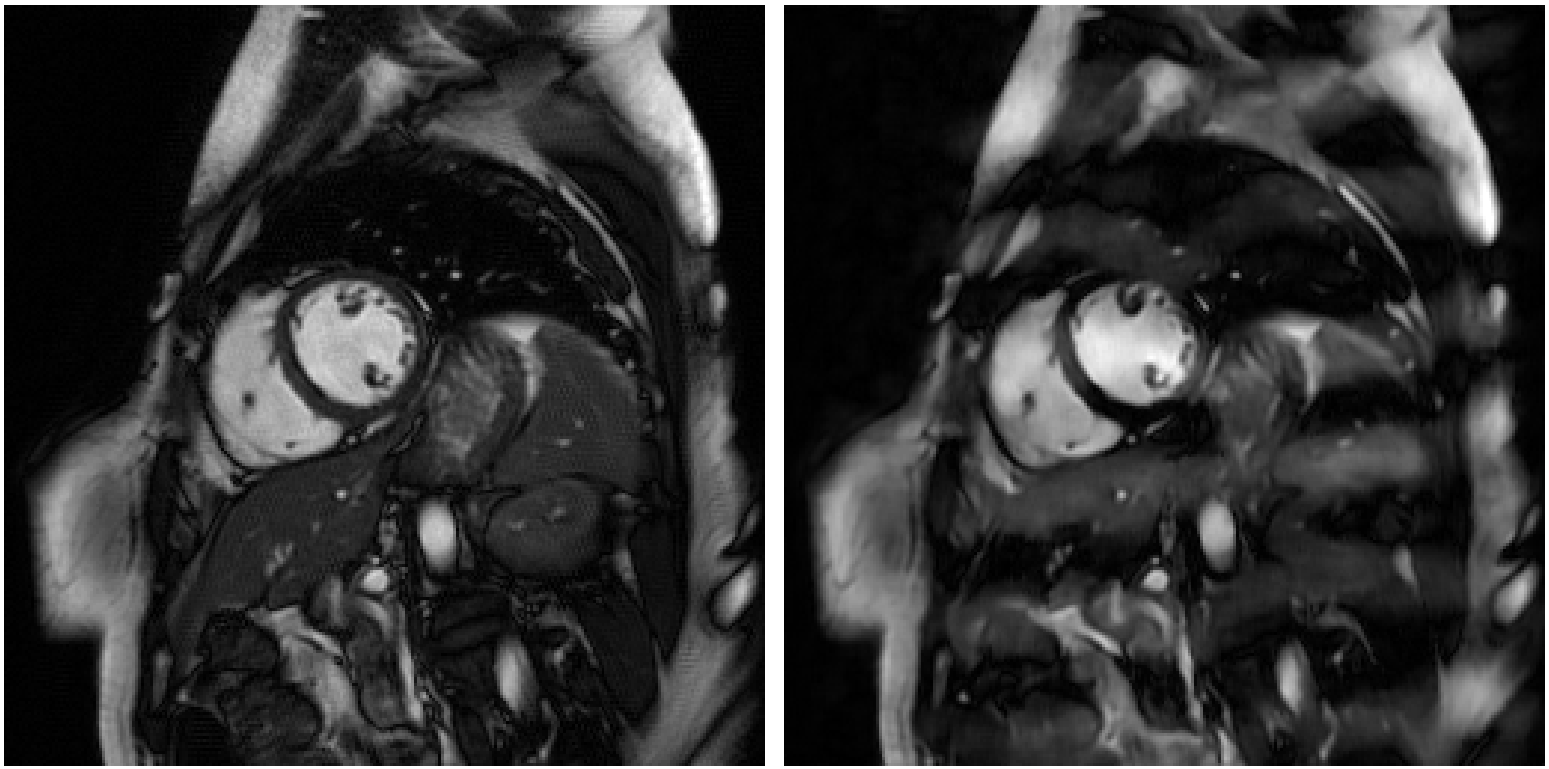


(a) $x + r$ (b) $\phi(A(x + r))$ (Neur. net)

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).

Deep MRI Net – Stability?

Time saving experiment in MRI - 25% subsampling

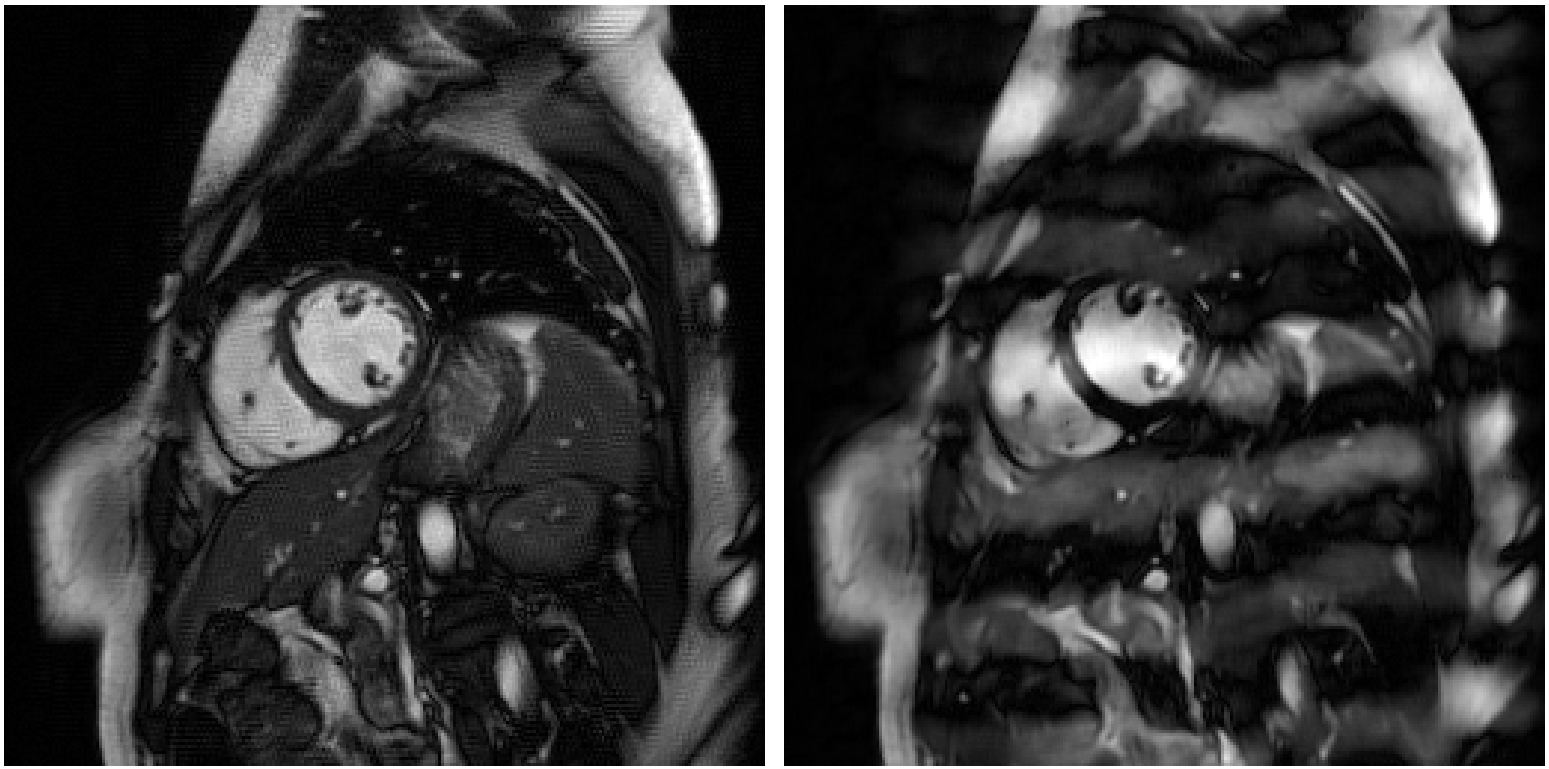


(a) $x + r$ (b) $\phi(A(x + r))$ (Neur. net)

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).

Deep MRI Net – Stability?

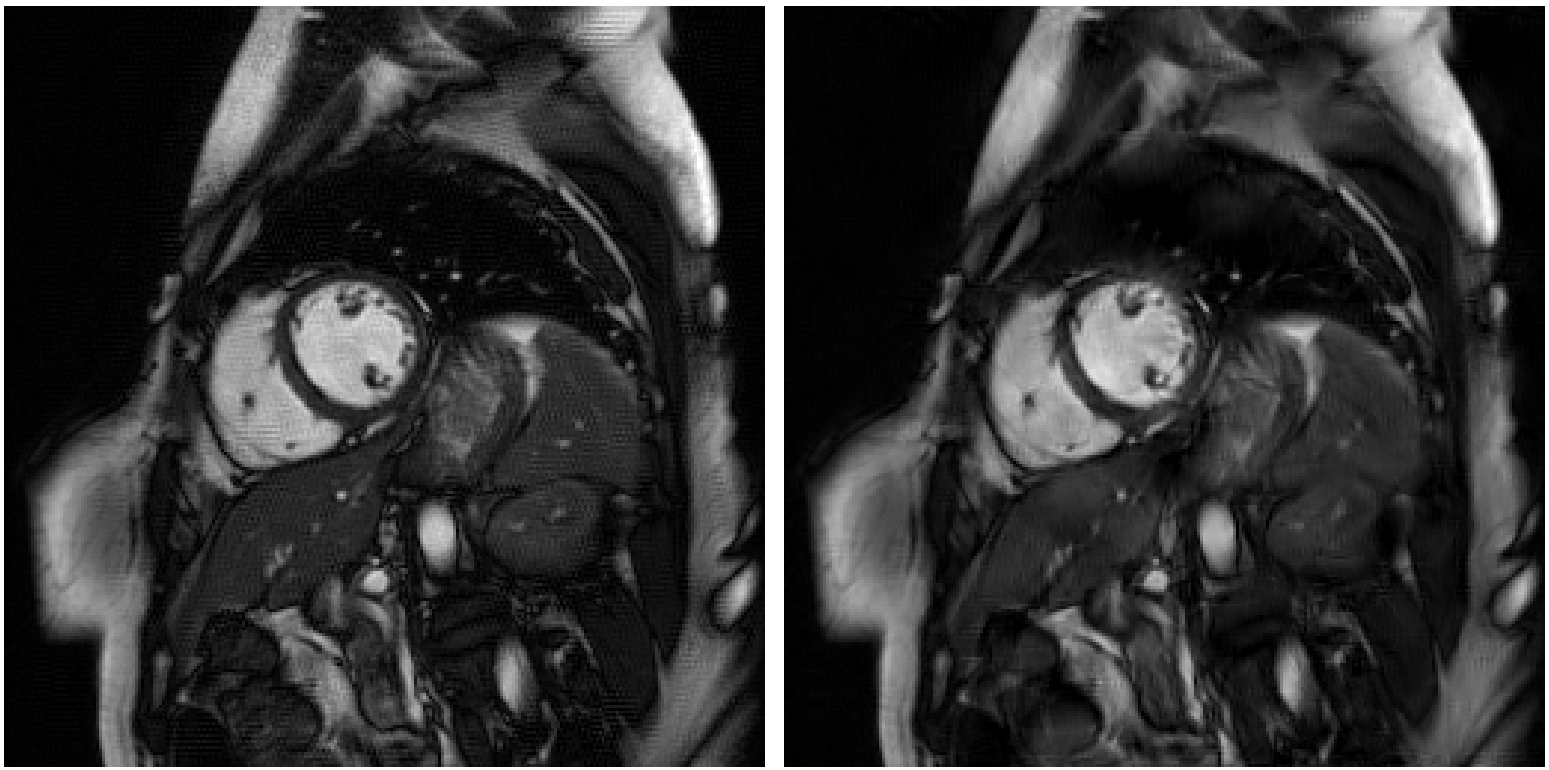
Time saving experiment in MRI - 25% subsampling



(a) $x + r$ (b) $\phi(A(x + r))$ (Neur. net)

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).

Time saving experiment in MRI - 25% subsampling



(a) $x + r$ (b) Compressed sensing

Can we resolve the stability issue?

The Paradox of Deep Learning

There is an uncountable family of classification functions $f : \mathbb{R}^{N_0} \to \{0, 1\}$ such that for any neural network dimensions $\mathbf{N} = (N_L, N_{L-1}, \ldots, N_1, N_0)$ with $N_0, L \geq 2$ and any $0 < \epsilon < 1/(K + M)$ where M is arbitrarily large and $K \geq 3(N_1 + 1) \cdots (N_{L-1} + 1)$ we have the following. There exist uncountably many training sets $\mathcal{T} = \{x^1, \ldots, x^K\}$ and uncountably many classification sets $\mathcal{C} = \{y^1, \ldots, y^M\}$ such that there is a

$$\tilde{\phi} \in \underset{\phi \in \mathcal{NN}_{\mathbf{N},L}}{\operatorname{argmin}} C(v, w), \quad v_j = \phi(x^j), \quad w_j = f(x^j),$$

where $1 \leq j \leq K$ such that

$$\tilde{\phi}(x) = f(x) \quad \forall x \in \mathcal{T} \cup \mathcal{C}.$$

However, there exists uncountably many $v \in \mathbb{R}^{N_0}$ such that

$$|\tilde{\phi}(v) - f(v)| \geq 1/2, \qquad \|v - x\|_\infty \leq \epsilon \text{ for some } x \in \mathcal{T}.$$

Moreover, there is a neural network $\hat{\phi}$, not necessarily trained, such that

$$\hat{\phi}(x) = f(x) \quad \forall x \in \mathcal{B}_\epsilon^\infty(\mathcal{T} \cup \mathcal{C}).$$