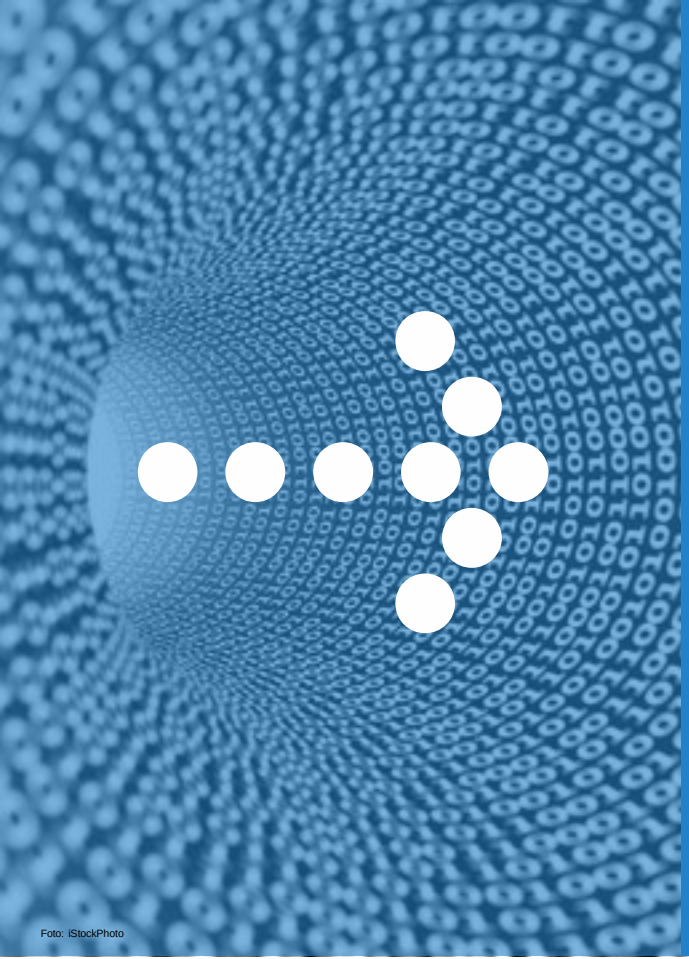


Foto: iStockPhoto

AI Security and Insecurity

Joel Brynielsson, 15 May 2019
joel.brynielsson@foi.se

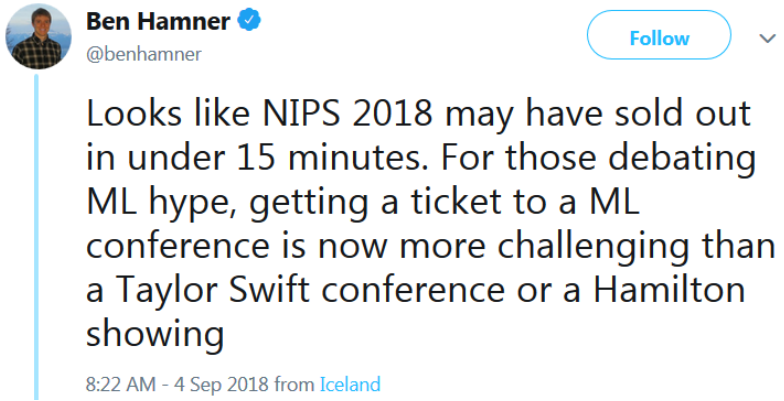
FOI

December 2018: 32nd Conference on Neural Information Processing Systems

- 8500 participants (ten from Sweden…)
- 4500 submissions
- 850 accepted papers

- The tickets were sold out in 11 minutes and 38 seconds

- (Sweden needs to increase its pace)

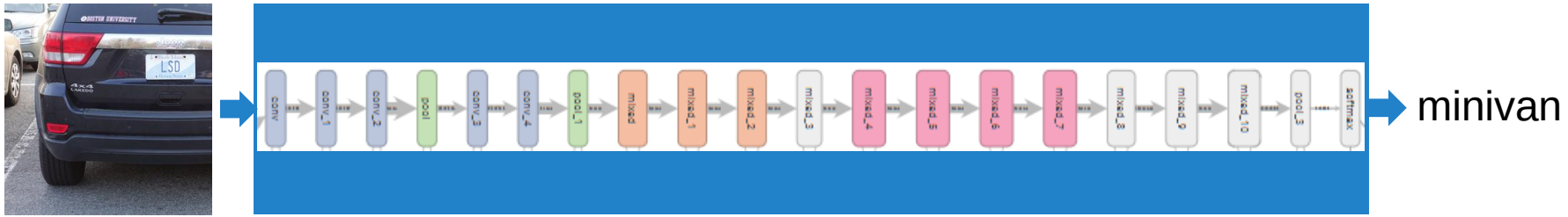


- **The civilian research is the driver**
- For defense and security, we need to keep up with developments and apply in specific areas

FOI

AI in a common application: image classification

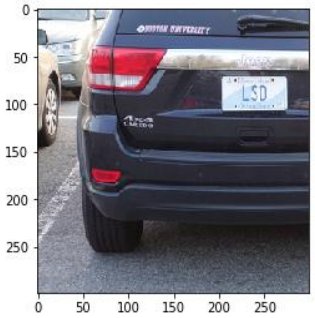
- Image classification (deep neural net: Inception-v3)

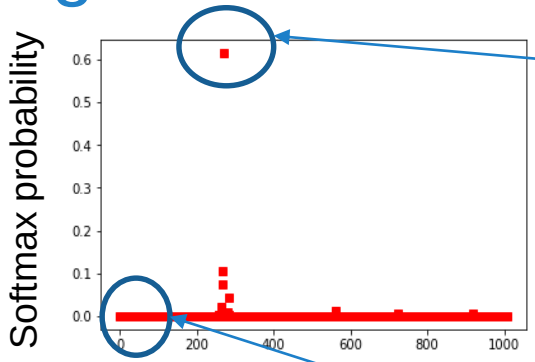


- In many applications it is of course important that the image classification becomes correct and cannot be fooled

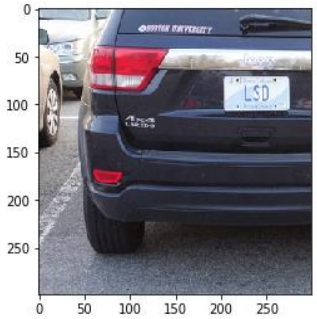
FOI

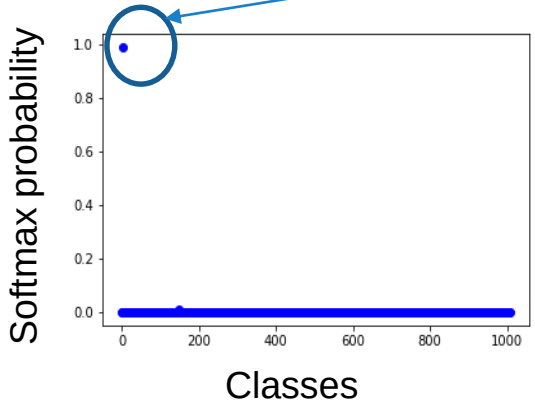
Influencing the image classifier: car becomes dog

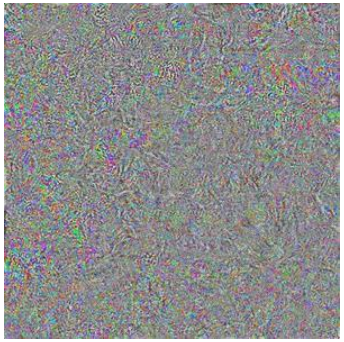




15 iterations







Enhanced differential image

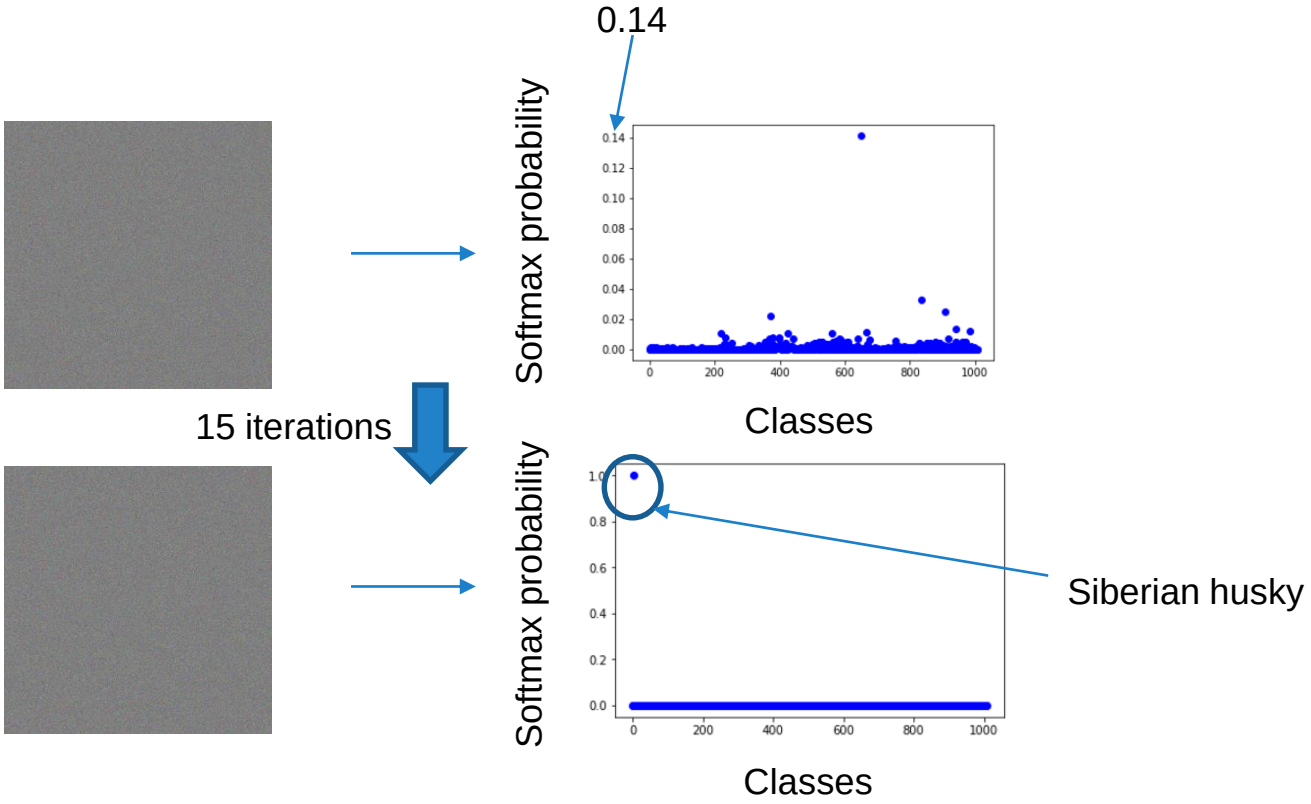
[FOI, 2018]

From car to dog—program code

```python
# load ResNet50 model
K.set_learning_phase(0)
model = ResNet50(weights='imagenet')

# load and pre-process image
img = image.load_img(filename, target_size=(224,224))
x = image.img_to_array(img)
x = np.expand_dims(x, axis=0)
x = preprocess_input(x)

# predict original image
preds = model.predict(x)
print('Original image predicted as:', decode_predictions(preds,top=3))

# create target vector
y_t = np.zeros(1000) # 1000-classes in ImageNet
y_t[np.clip(class_index,0,999)] = 1.0
y_t = np.expand_dims(y_t,axis=0)

# define function to get gradients
y_target = K.placeholder(shape=model.output_shape)
loss = losses.categorical_crossentropy(y_target,model.output)
get_grads = K.function([model.input, y_target], K.gradients(loss,model.input))

# start fooling (around)
NITER = 10
for _ in range(NITER):
    grads = get_grads([x, y_t])
    x -= 0.3*np.sign(grads[0][0])
    preds = model.predict(x)
    print('Manipulated image predicted as:', decode_predictions(preds, top=3))

# undo ResNet50 pre-processing
x[:, :, :, 0] += 103.939
x[:, :, :, 1] += 116.779
x[:, :, :, 2] += 123.68
x = x[:,:,:,::-1] # RGB-BGR conversion
x = np.clip(x[0],0,255).astype(dtype=np.uint8)
```

the core code snippet

Random noise becomes dog



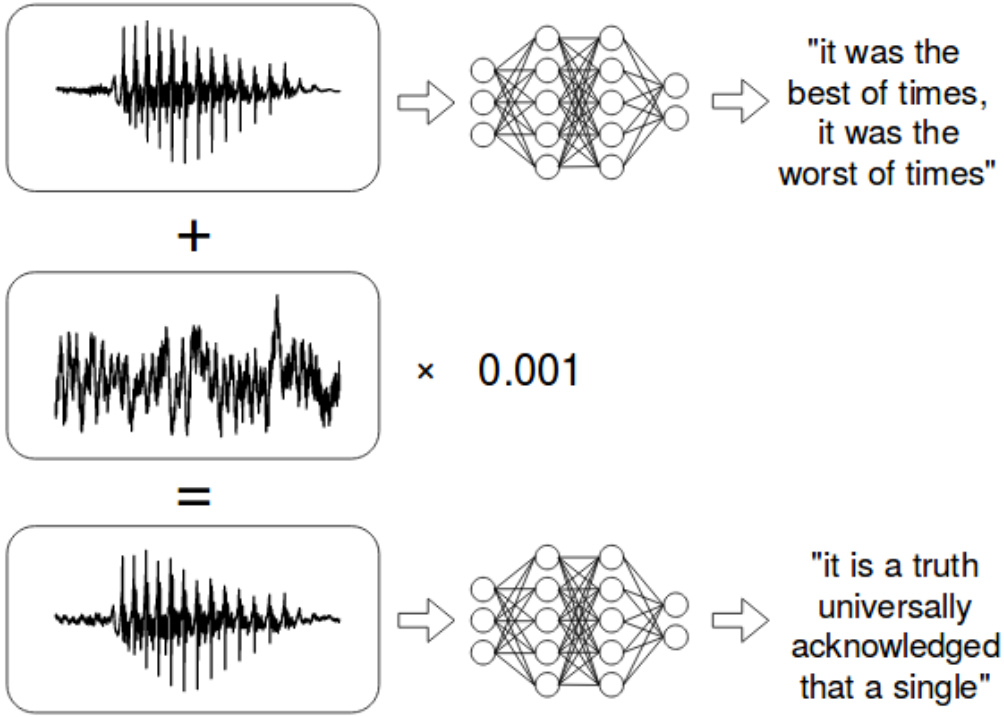
Manipulating physical objects

Evtimov et al., “Robust Physical-World Attacks on Machine Learning Models”.
In: *CoRR* abs/1707.08945 (2017). arXiv: 1707.08945.
URL: http://arxiv.org/abs/1707.08945.

Athalye et al., “Synthesizing Robust Adversarial Examples”.
In: *CoRR* abs/1707.07397 (2017). arXiv: 1707.07397.
URL: http://arxiv.org/abs/1707.07397.

FOI

Manipulating sound



Nicholas Carlini and David A. Wagner, “Audio Adversarial Examples: Targeted Attacks on Speech-to-Text”. In: *CoRR* abs/1801.01944 (2018). arXiv: 1801.01944.
URL: http://arxiv.org/abs/1801.01944.

FOI

Manipulating how AI systems interpret text

Original Text Prediction = **Negative**. (Confidence = 78.0%)
*This movie had **terrible** acting, **terrible** plot, and **terrible** choice of actors. (Leslie Nielsen ...come on!!!) the one part I **considered** slightly funny was the battling FBI/CIA agents, but because the audience was mainly **kids** they didn't understand that theme.*
Adversarial Text Prediction = **Positive**. (Confidence = 59.8%)
*This movie had **horrific** acting, **horrific** plot, and **horrifying** choice of actors. (Leslie Nielsen ...come on!!!) the one part I **regarded** slightly funny was the battling FBI/CIA agents, but because the audience was mainly **youngsters** they didn't understand that theme.*

Moustafa Alzantot et al., "Generating Natural Language Adversarial Examples".
In: *CoRR* abs/1804.07998 (2018). arXiv: 1804.07998.
URL: http://arxiv.org/abs/1804.07998.

Fighting vulnerabilities using transparency

Input

Output

Explanation

It’s a tiger, 90%

It’s a hen, 50%

[FOI, 2019]

AI as a two-edged sword: opportunities and vulnerabilities

- Self-driving cars … can be fooled.
- Computer support for transcription ... can be fooled.
- Detection of influence operations … can be made difficult.

- **We must take advantage of the AI opportunities…**
- **…and deal with the vulnerabilities.**

FOI

AI for defense and security, summary

- AI in the defense and security area is about being able to keep up with and apply new research findings (rather than to develop from scratch).
- AI offers great opportunities for many different applications, and the future looks promising!
- In defense and security the vulnerabilities that AI development may entail need to be addressed.

Some important AI research issues related to defense and security

- What vulnerabilities do AI systems have and how can these be exploited?
  - How can, e.g., an image sensor be fooled?
- How can AI systems be made more robust / resilient?
  - Can, e.g., an image sensor “learn about the bad” so it can be avoided?
- How can increased transparency and confidence in AI systems be achieved?
  - How can an AI system be made more transparent or explainable?
- To what extent will different work tasks be automated, and how does this affect the work on defense and security?

FOI