Deep Fakes

Methods and risks

François Fleuret

http://fleuret.org/public/irgc-deepfakes/

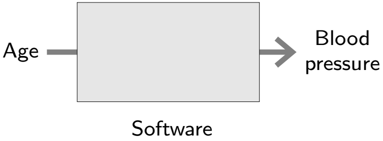
Sep 9th 2019

idiap RESEARCH INSTITUTE

EPFL

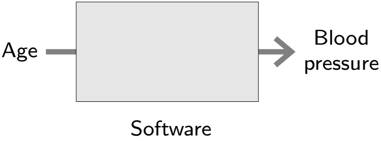
Deep Machine Learning

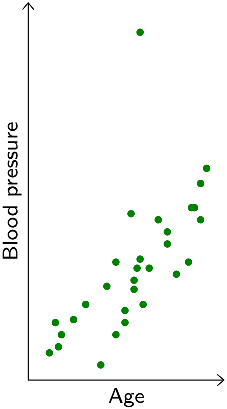
The principle of “machine learning” is to tune computer programs on data.



Deep Machine Learning

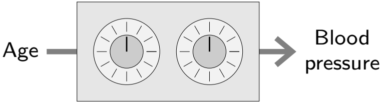
The principle of "machine learning" is to tune computer programs on data.



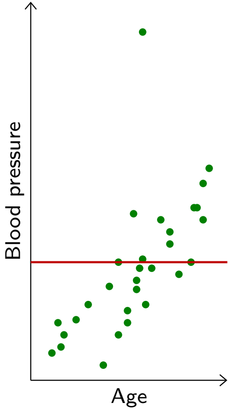


Deep Machine Learning

The principle of "machine learning" is to tune computer programs on data.

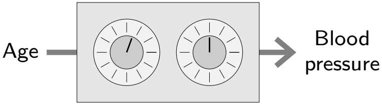


"Linear regression"



Deep Machine Learning

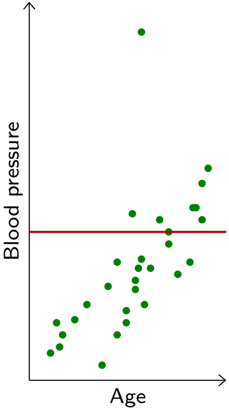
The principle of “machine learning” is to tune computer programs on data.



Age

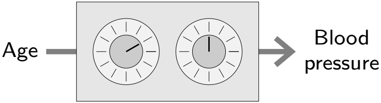
Blood pressure

“Linear regression”

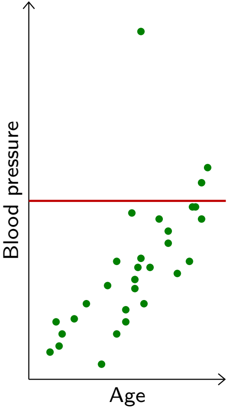


Deep Machine Learning

The principle of “machine learning” is to tune computer programs on data.

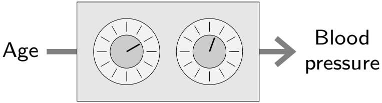


“Linear regression”

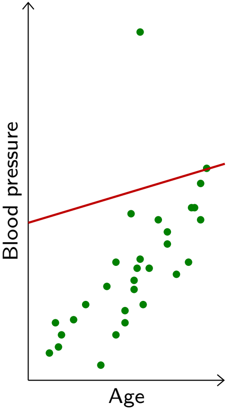


Deep Machine Learning

The principle of “machine learning” is to tune computer programs on data.

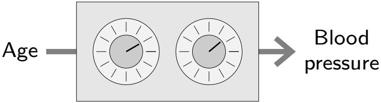


“Linear regression”

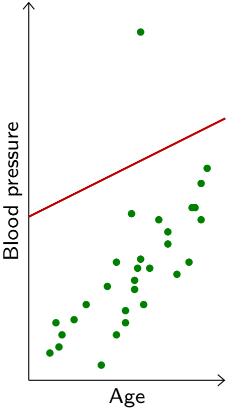


Deep Machine Learning

The principle of “machine learning” is to tune computer programs on data.

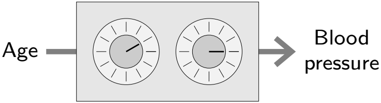


“Linear regression”

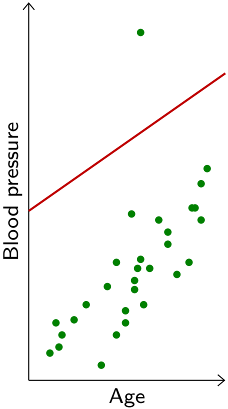


Deep Machine Learning

The principle of “machine learning” is to tune computer programs on data.

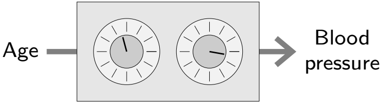


“Linear regression”

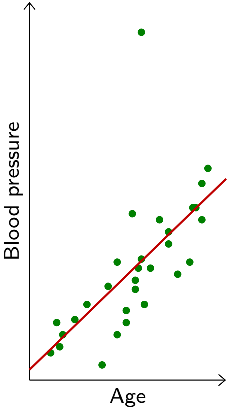


Deep Machine Learning

The principle of “machine learning” is to tune computer programs on data.

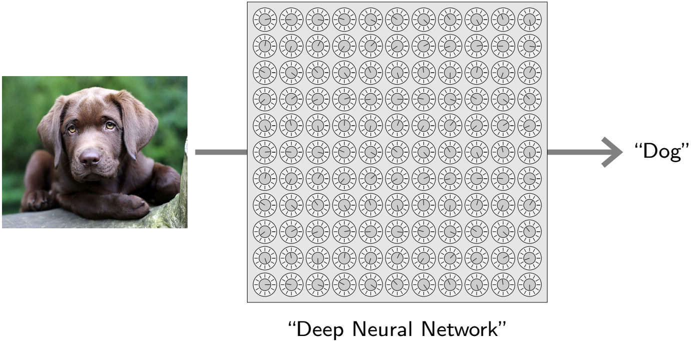


“Linear regression”



Deep Machine Learning

The same idea generalizes to very complex prediction problems, for which large sets of “training examples” are available.

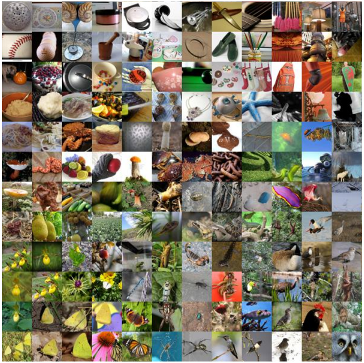


“Dog”

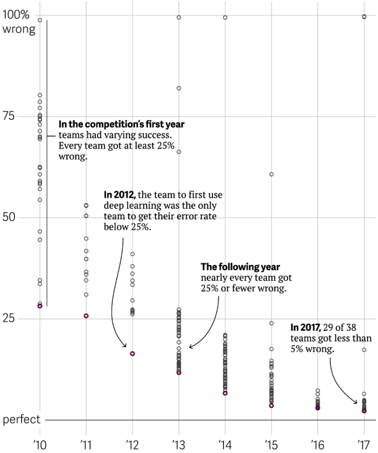
“Deep Neural Network”

Deep Machine Learning

Over the last decade these methods have improved on many fundamental tasks from barely usable to close to or beyond human performance.



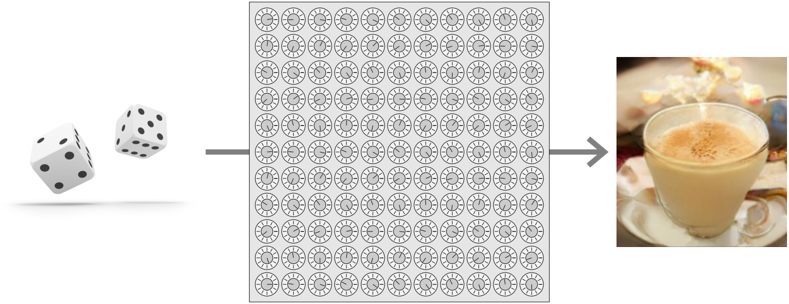
ImageNet



(Gershgorn, 2017)

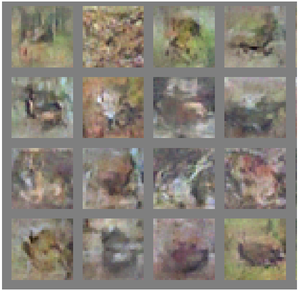
Generative models

The same methods can be used to generate signals *ex nihilo.*

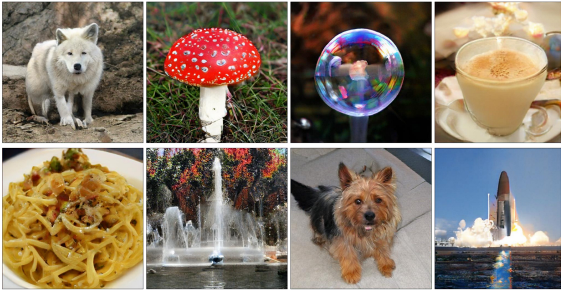


“Deep Neural Network”

Generative models



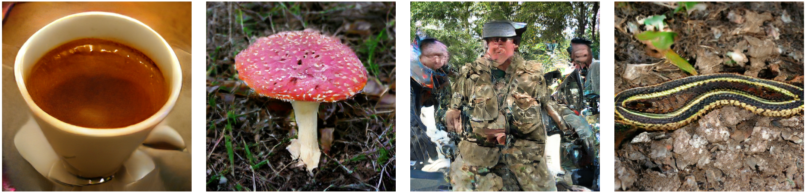
(Goodfellow et al., 2014)



(Brock et al., 2018)

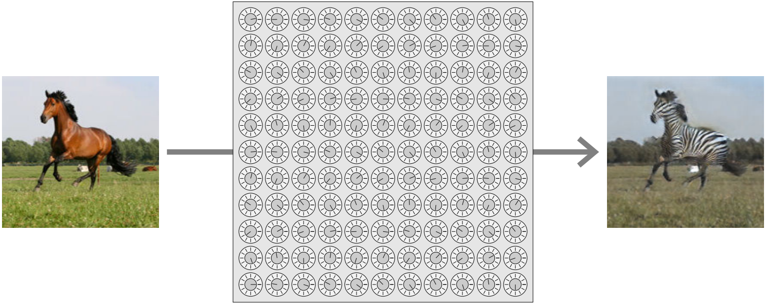
Generative models

```
~ pip install pytorch-pretrained-biggan

from torch import from_numpy, no_grad
from pytorch_pretrained_biggan import BigGAN, one_hot_from_names, \
                                      truncated_noise_sample, save_as_images

objects = [ 'coffee', 'mushroom',  'military uniform', 'garter snake' ]
cv = from_numpy(one_hot_from_names(objects, batch_size=len(objects)))
nv = from_numpy(truncated_noise_sample(truncation=0.4, batch_size=len(objects)))
model = BigGAN.from_pretrained('biggan-deep-512')
with no_grad(): save_as_images(model(nv, cv, 0.4))
```

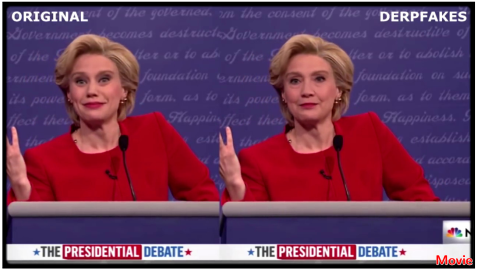


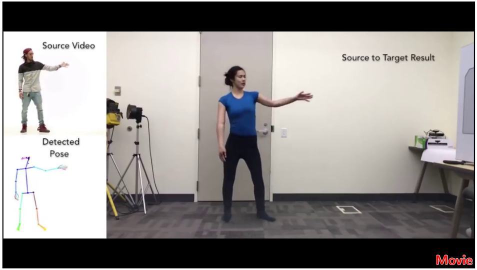
Generative models

They can also generate signals given a reference input (Mirza and Osindero, 2014; Zhu et al., 2017).



Generative models





Generative models

https://talktotransformer.com/

The meeting about deep fakes is an important event since it will give consumers and journalists the facts before these companies rush to exploit them. With the meeting, we hope that we will reach a good balance between protecting the legitimate business and being fair to companies that do not follow traditional guidelines. This is a very important issue and is now going to be discussed at many more international conferences. And I can only hope that the government will take the initiative to address it urgently so that consumers get a right to know the true nature of their product.”

Challenges and risks

- Off-the-shelves hardware and software.
- Low requirements in expertise and resources.
- Targets images, videos, sounds, and text.
- Cheap to produce content on a large scale.
- Dual-use technologies, both hardware and software.
- Quality will only improve, probable arm race.

The end

References

A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. CoRR, abs/1809.11096, 2018.

D. Gershgorn. The data that transformed AI research—and possibly the world, July 2017.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. CoRR, abs/1406.2661, 2014.

M. Mirza and S. Osindero. Conditional generative adversarial nets. CoRR, abs/1411.1784, 2014.

J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR, abs/1703.10593, 2017.