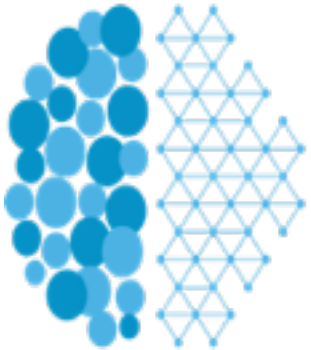
Creating Human-Level AI

Yoshua Bengio

Asilomar Conference on Beneficial AI

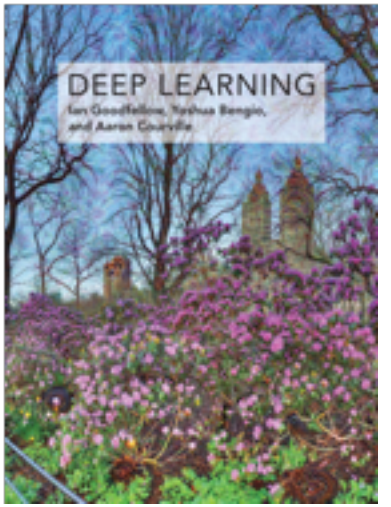
January 6th, 2017



MILA

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

PLUG: **Deep Learning**, MIT Press book is out, chapters will remain online



Université de Montréal

IVADO

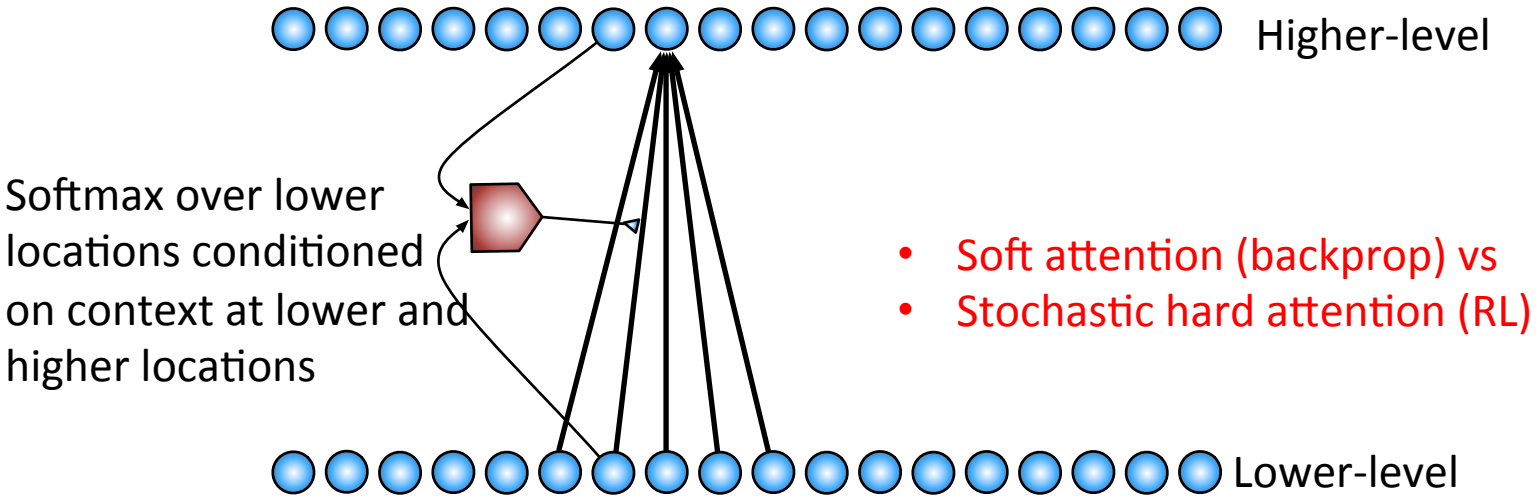
Recent Progress

- Breakthroughs with Deep Learning
  - Speech recognition
  - Computer vision
  - Machine translation
  - Reasoning, attention & memory
  - Reinforcement learning (playing games, Go)
  - Robotics & control
  - Long-term dependencies & very deep nets

Attention Mechanism for Deep Learning

(Bahdanau, Cho & Bengio, ICLR 2015; Jean et al ACL 2015; Jean et al WMT 2015; Xu et al ICML 2015; Chorowski et al NIPS 2015; Firat, Cho & Bengio 2016)

- Consider an input (or intermediate) sequence or image
- Consider an upper level representation, which can choose « where to look », by assigning a weight or probability to each input position, as produced by an MLP, applied at each position



- Impact of soft-attention: not just machine translation, also reasoning & memory, handling data structures, etc.

Google-Scale NMT Success

(Wu et al & Dean, Nature, 2016)

- After beating the classical phrase-based MT on the academic benchmarks, there remained the question: will it work on the very large scale datasets like used for Google Translate?
- Distributed training, very large model ensemble
- Not only does it work in terms of BLEU but it makes a killing in terms of human evaluation on Google Translate data

translations from the production phrase-based statistical translation
from our GNMT system, and 3) translations by humans fluent in bo
are averaged rated scores with their standard deviations for English ←
Portuguese and English ↔ Chinese. All the GNMT models are word
and use a shared source and target vocabulary with 32K wordpieces.
data consist of 500 randomly sampled sentences from Wikipedia an

18

Why is Deep Learning Working?

Machine Learning, AI & No Free Lunch

- Five key ingredients for ML towards AI
  1. Lots & lots of data
  2. Very flexible models
  3. Enough computing power
  4. Computationally efficient inference
  5. **Powerful priors that can defeat the curse of dimensionality**

Bypassing the curse of dimensionality

We need to build compositionality into our ML models

Just as human languages exploit compositionality to give representations and meanings to complex ideas

Exploiting compositionality gives an exponential gain in representational power

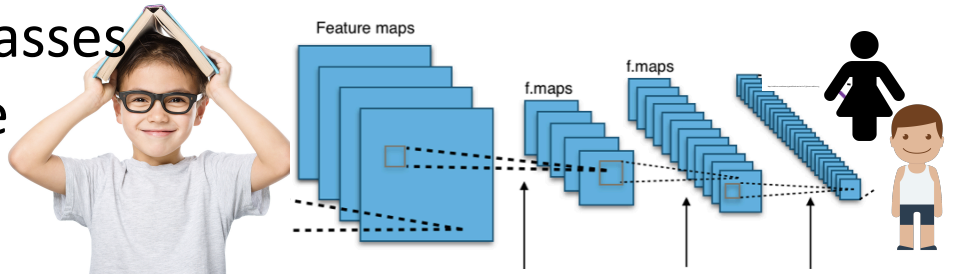
Distributed representations / embeddings: feature learning

Deep architecture: multiple levels of feature learning

Prior assumption: compositionality is useful to describe the world around us efficiently

Each feature can be discovered without the need for seeing the exponentially large number of configurations of the other features

- Consider a network whose hidden units discover the following features:
  - Person wears glasses
  - Person is female
  - Person is a child
  - Etc.



If each of *n* feature requires *O(k)* parameters, need *O(nk)* examples

Non-parametric methods would require $O(n^d)$ examples

Where We Are: Still Far Away

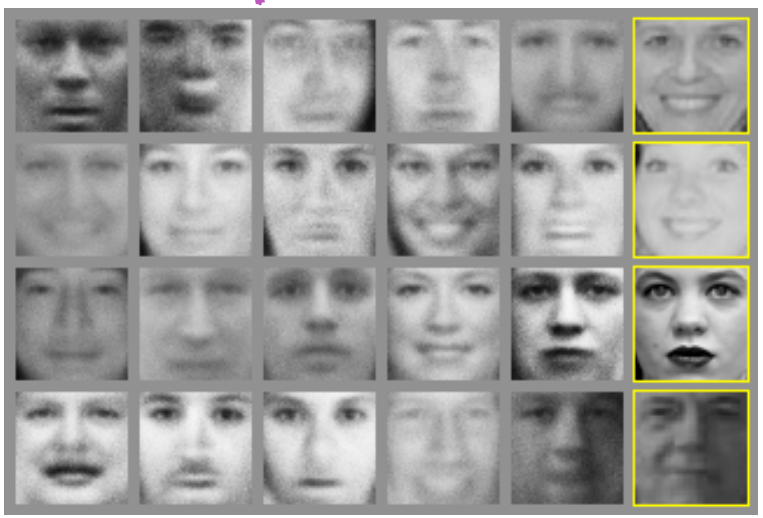
- All industrial successes are based on pure supervised learning
- Still learning superficial clues that do not generalize well outside of training contexts and make it easy to fool trained networks:
  - Current models cheat by picking on surface regularities, e.g., background greenery → animal is present
- Still unable to do a good job of learning higher-level abstractions at multiple time scales, deal with very long-term dependencies
- Still relying heavily on smooth differentiable predictors (using backprop)

Progress and Obstacles in Deep Unsupervised Generative Models

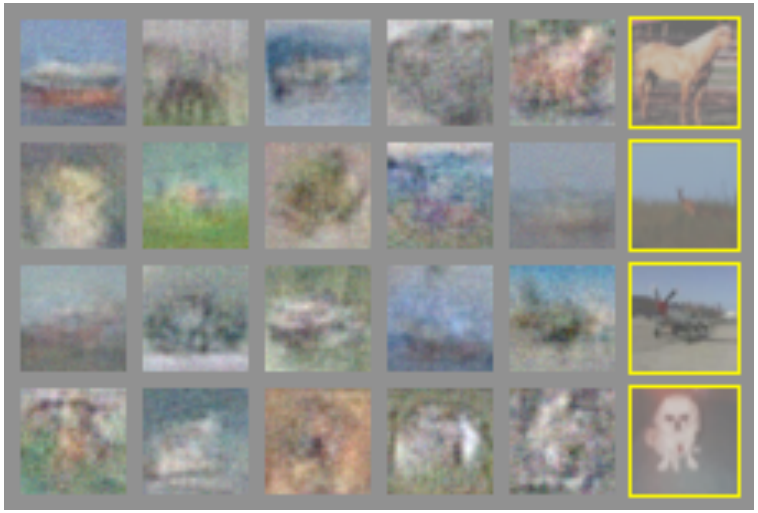
- Humans are very good at unsupervised learning, e.g. 2 year old know intuitive physics
- RBMs and DBMs: obstacle probably due to gradient estimator relying on good mixing of MCMC (which gets worse as training progresses because distribution becomes sharper)
- Autogressive models (NADE, MADE, PixelRNN, PixelCNN, WaveNet): easier to train but no latent variables
- VAEs and GANs: the current frontier, hard to train, still unsatisfactory in terms of extracting abstraction

Early Days of GAN Samples

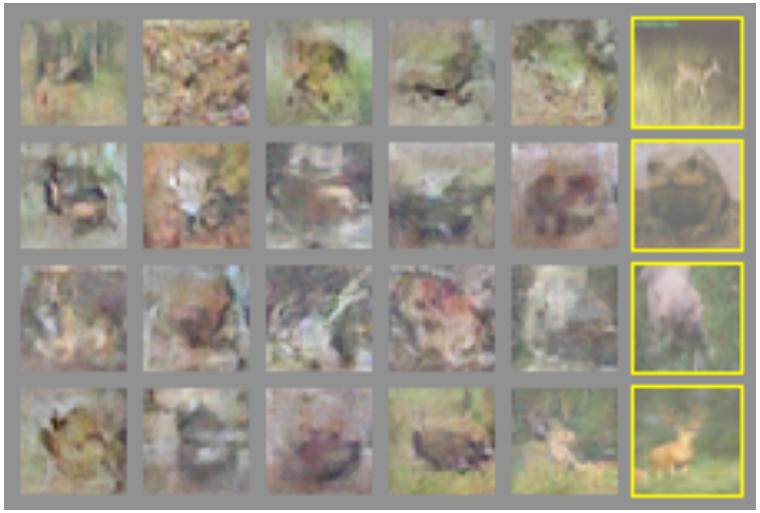
MNIST



TFD



CIFAR-10 (fully connected)



CIFAR-10 (convolutional)

Convolutional GANs

(Radford et al, arXiv 1511.06343)

Strided convolutions, batch normalization, only convolutional layers, ReLU and leaky ReLU

Figure 3: Generated bedrooms after five epochs of training. There appe
under-fitting via repeated textures across multiple samples.

4.3 IMAGENET-1K

We use Imagenet-1k (Deng et al., 2009) as a source of natural images fo
train on 32×32 min-resized center crops. No data augmentation was ap

5

GAN: Interpolating in Latent Space

If the model is good (unfolds the manifold), interpolating between latent values yields plausible images.

scene classification learn object detectors (Oquab et al., 2014). We demon
DCGAN trained on a large image dataset can also learn a hierarchy of fe
Using guided backpropagation as proposed by (Springenberg et al., 2014)
features learnt by the discriminator activate on typical parts of a bedroo
For comparison, in the same figure, we give a baseline for randomly initi
activated on anything that is semantically relevant or interesting.

6.3 Manipulating the Generator Representation

6.3.1 Forgetting to draw certain objects

In addition to the representations learnt by a discriminator, there is the qu
tions the generator learns. The quality of samples suggest that the gene
representations for major scene components such as beds, windows, lamps
furniture. In order to explore the form that these representations take, we
to attempt to remove windows from the generator completely.

9

Combining Iterative Sampling from Denoising Auto-Encoders with GAN

Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space

Anh Nguyen, Jason Yosinski, Yoshua Bengio, Alexey Dosovitskiy, Jeff Clune

(submitted to CVPR 2017) arXiv:1612.00005

227 x 227 ImageNet GENERATED IMAGES of category Volcano

(cheating by using lots of labeled data during training)

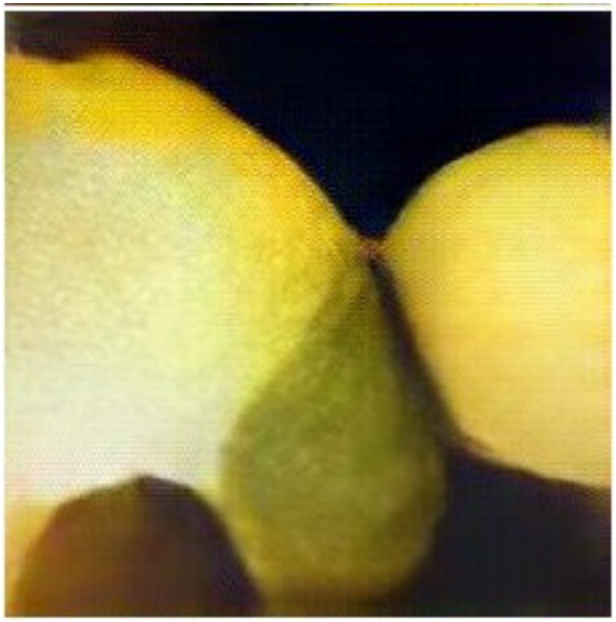
Plug & Play Generative Networks

High-Resolution
Samples
227 x 227

bird

ant

volcano



lemon

What's Missing

- More autonomous learning, **unsupervised learning**
- Discovering the **underlying causal factors**
- Model-based RL which extends to completely new situations by **unrolling powerful predictive models which can help reason about rarely observed dangerous states**
- Sufficient **computational power** for models large enough to capture human-level knowledge
- Autonomously discovering **multiple time scales to handle very long-term dependencies**
- Actually **understanding language** (also solves generating), requiring enough world knowledge / commonsense
- Large-scale **knowledge representation** allowing one-shot learning as well as discovering new abstractions and explanations by 'compiling' previous observations

AI Acting in the World

- We could have an AI which understands the world around us but does not have significant influence in it
- But practically we will want **active AIs** → need for AIs which also incorporate (and possibly learn) human values (inverse RL)
- Since humans themselves do not agree on ethics and values, some form of learning will be a necessary ingredient. How will we train **wise AI**?

Acting to Guide Representation Learning

- What is a good latent representation?
- The notion of disentangling the underlying factors of representation is not specific enough
- New on-going research: **appropriate factors each correspond to 'independently controllable' aspects of the world**
- Can only be discovered by acting in the world
- Some factors deduced by analogy (e.g. the sun) as caused by imagined (or imaginary) agents

Montreal Institute for Learning Algorithms

MILA

Université de Montréal