Progress and Challenges of Deep Learning towards AI

Yoshua Bengio

January 11, 2016
Future of AI Symposium, NYU
AI Breakthrough

Deep Learning: brain-inspired machine learning algorithms based on learning multiple levels of representation / abstraction.

Amazing improvements in error rate in object recognition, object detection, speech recognition, and more recently, in natural language processing / understanding
Four Ingredients

• Four key ingredients for ML towards AI

1. Lots & lots of data
2. Very flexible models
3. Enough computing power
4. 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 and culture exploit compositionality to give representations and meanings to complex ideas

Exploiting compositionality gives an exponential gain in representational power

NEW THEORY (NIPS 2014)

(1) Distributed representations / embeddings: feature learning

(2) Deep architecture: multiple levels of feature learning

Prior: compositionality is useful to describe the world around us efficiently
Neural Language Models: fighting one exponential by another one!

- (Bengio et al NIPS’2000)

Exponentially large set of generalizations: semantically close sequences

Exponentially large set of possible contexts
Neural word embeddings: visualization directions = Learned Attributes
A Myth is Being Debunked: Local Minima in Neural Nets
→ Convexity is not needed

- (Dauphin, Pascanu, Gulcehre, Cho, Ganguli, Bengio, NIPS’ 2014): *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*
- (Choromanska, Henaff, Mathieu, Ben Arous & LeCun, AISTATS’2015): *The Loss Surface of Multilayer Nets*
From Pattern Recognition to AI

- Recurrent networks and the challenge of long-term dependencies
- Introduction of internal attention mechanism (Bahdanau et al 2014)
  - Success in machine translation & caption generation
  - Success in neural Turing machines & Memory Nets for
    - Reasoning algorithmically
    - Question answering

![Diagram of neural network](image)

![Bar chart showing HTER (HE SET) for different institutions](chart)
Understanding images → generating natural language sentences

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.
Need for more autonomy: Unsupervised Learning

- Recent progress mostly in supervised DL
- Real technical challenges for unsupervised DL
- Potential benefits:
  - Exploit tons of unlabeled data
  - Answer new questions about the variables observed
  - Another a priori: underlying causal factors
  - Structured outputs
Samples generated by letting the associative memory run with one label clamped. There are 1000 iterations of alternating Gibbs sampling between samples.

From RBMs to VAEs and GANs

Under review as a conference paper at ICLR 2016

Figure 2: Generated bedrooms after one training pass through the dataset. Theoretically, the model could learn to memorize training examples, but this is experimentally unlikely as we train with a small learning rate and minibatch SGD. We are aware of no prior empirical evidence demonstrating memorization with SGD and a small learning rate in only one epoch.

Figure 3: Generated bedrooms after five epochs of training. There appears to be evidence of visual under-fitting via repeated textures across multiple samples.

4.3 I MAGENET

We use Imagenet-1k (Deng et al., 2009) as a source of natural images for unsupervised training. We train on $32 \times 32$ mini-resized center crops. No data augmentation was applied to the images.

4.3. MNIST Generation with Two Digits

The main motivation for using an attention-based generative model is that large images can be built up iteratively, by adding to a small part of the image at a time. To test this capability in a controlled fashion, we trained DRAW to generate images with two $28 \times 28$ MNIST images chosen at random and placed at random locations in a $60 \times 60$ black background. In cases where the two digits overlap, the pixel intensities were added together at each point and clipped to be no greater than one. Examples of generated data are shown in Fig. 8. The network typically generates one digit and then the other, suggesting an ability to recreate composite scenes from simple pieces.

4.4. Street View House Number Generation

MNIST digits are very simplistic in terms of visual structure, and we were keen to see how well DRAW performed on natural images. Our first natural image generation experiment used the multi-digit Street View House Numbers dataset (Netzer et al., 2011). We used the same preprocessing as (Goodfellow et al., 2013), yielding a $64 \times 64$ house number image for each training example. The network was then trained using $54 \times 54$ patches extracted at random locations from the preprocessed images. The SVHN training set contains 231,053 images, and the validation set contains 4,701 images.

A major challenge with natural image generation is how to model the pixel colours. In this work we applied a simple approximation where the normalised intensity of each of the RGB channels was treated as an independent Bernoulli probability. This approach has the advantage of being easy to implement and train; however it does mean that the loss function used for training does not match the true compression cost of the data.

The house number images generated by the network are highly realistic, as shown in Figs. 9 and 10. Fig. 11 reveals that, despite the long training time, the DRAW network underfit the SVHN training data.

4.5. Generating CIFAR Images

The most challenging dataset we applied DRAW to was the CIFAR-10 collection of natural images (Krizhevsky, 2009). CIFAR-10 is very diverse, and with only 50,000 training examples it is very difficult to generate realistic-looking objects without overfitting (in other words, without copying from the training set). Nonetheless the images in Fig. 12 demonstrate that DRAW is able to capture much of the shape, colour and composition of real photographs.
Last Words

• Still far from autonomous Machine Learning
  • Need major progress in unsupervised learning, reinforcement learning, natural language understanding, reasoning, hardware...

• We can probably have human-level AI in a few years
  • And it does not need an ego & survival drive to be useful
MILA: Montreal Institute for Learning Algorithms