Deep Learning: Progress in Theory and Attention Mechanisms

Yoshua Bengio

June 11, 2015

Deep Vision CVPR Workshop, Boston, USA
Progress in Deep Learning Theory

- Exponential advantage of distributed representations
- Exponential advantage of depth
- Myth-busting: non-convexity & local minima
- Probabilistic interpretation of auto-encoders
Exponential advantage of distributed representations

Learning a set of features that are not mutually exclusive can be exponentially more statistically efficient than having nearest-neighbor-like or clustering-like models.
Exponential advantage of distributed representations

- **Bengio 2009** (Learning Deep Architectures for AI, Foundations & Trends in ML)
- **Montufar & Morton 2014** (When does a mixture of products contain a product of mixtures? SIAM J. Discr. Math)
Exponential advantage of depth

Theoretical arguments:

2 layers of

- Logic gates
- Formal neurons
- RBF units

RBM s & auto-encoders = universal approximator

Theorems on advantage of depth:

Some functions compactly represented with k layers may require exponential size with 2 layers
Expressiveness of deep networks with piecewise linear activation functions: exponential advantage for depth (Montufar et al, NIPS 2014)

- They can split the input space in many more (not-independent) linear regions, with constraints, e.g., with abs units, each unit creates mirror responses, folding the input space:
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*
Saddle Points

- Local minima dominate in low-D, but saddle points dominate in high-D
- Most local minima are close to the bottom (global minimum error)
Saddle Points During Training

- Oscillating between two behaviors:
  - Slowly approaching a saddle point
  - Escaping it
Low Index Critical Points

*Choromanska et al & LeCun 2014, ‘The Loss Surface of Multilayer Nets’*

Shows that deep rectifier nets are analogous to spherical spin-glass models

The low-index critical points of large models concentrate in a band just above the global minimum
The Next Challenge: 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
  - Regularizer – transfer learning – domain adaptation
  - Easier optimization (local training signal)
  - Structured outputs

- If Ys of interest are among the causal factors of X, then

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

is tied to P(X) and P(X|Y), and P(X) is defined in terms of P(X|Y), i.e.

- The best possible model of X (unsupervised learning) MUST involve Y as a latent factor, implicitly or explicitly.
- Representation learning SEEKS the latent variables H that explain the variations of X, making it likely to also uncover Y.
Probabilistic interpretation of auto-encoders

- Manifold & probabilistic interpretations of auto-encoders
- Denoising Score Matching as inductive principle
  
  (Vincent 2011)
- Estimating the gradient of the energy function
  
  (Alain & Bengio ICLR 2013)
- Sampling via Markov chain
  
  (Bengio et al NIPS 2013)
- Variational auto-encoder breakthrough
  
  (Kingma & Welling ICLR 2014)
  
  (Gregor et al arXiv 2015)
Denoising Auto-Encoder

• Learns a vector field pointing towards higher probability direction \((\text{Alain & Bengio } 2013)\)

\[
\text{reconstruction}(x) - x \rightarrow \sigma^2 \frac{\partial \log p(x)}{\partial x}
\]

• Some DAEs correspond to a kind of Gaussian RBM with \textit{regularized} Score Matching \((\text{Vincent } 2011)\)

[equivalent when noise \(\rightarrow 0\)]

prior: examples concentrate near a lower dimensional “manifold”
Regularized Auto-Encoders Learn a Vector Field that Estimates a Gradient Field  
(Alain & Bengio ICLR 2013)
Denoising Auto-Encoder Markov Chain

The corrupt-encode-decode-sample Markov chain associated with a DAE samples from a consistent estimator of the data generating distribution
Attention Mechanism for Deep Learning

- 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
Applying an attention mechanism to
- Translation
- Speech
- Images
- Video
- Memory
End-to-End Machine Translation

- Classical Machine Translation: several models separately trained by max. likelihood, brought together with logistic regression on top, based on n-grams

- Neural language models already shown to outperform n-gram models in terms of generalization power

- Why not train a neural translation model end-to-end to estimate $P(\text{target sentence} \mid \text{source sentence})$?
2014: The Year of Neural Machine Translation Breakthrough

• (Devlin et al, ACL’2014)
• (Cho et al EMNLP’2014)
• (Bahdanau, Cho & Bengio, arXiv sept. 2014)
• (Jean, Cho, Memisevic & Bengio, arXiv dec. 2014)
• (Sutskever et al NIPS’2014)

Earlier work: (Kalchbrenner & Blunsom et al 2013)
Encoder-Decoder Framework

• Intermediate representation of meaning
  = ‘universal representation’
• Encoder: from word sequence to sentence representation
• Decoder: from representation to word sequence distribution
Bidirectional RNN for Input Side

- Following Alex Graves’ work on handwriting

\[ e = (\text{Economic, growth, has, slowed, down, in, recent, years, .}) \]
Attention: Many Recent Papers

• (Xu et al 2015, caption generation, U. Montreal + U. Toronto)
• (Ba et al 2014, Mnih et al 2014, visual attention, Google DeepMind)
• (Chorowski et al 2014, speech recognition, U. Montreal)
• (Bahdanau et al 2014, machine translation, U. Montreal)

And Older Papers

• (Larochelle & Hinton 2010, MNIST, U. Toronto)
• (Graves 2013, handwriting generation)
• (Denil et al 2014, visual tracking)
• (Tang et al 2014, generative models of images)
Soft-Attention vs Stochastic Hard-Attention

- With soft-attention: input fed to higher level at location \(i\) is a softmax-weighted sum of states at locations \(j\) at lower level
  - Train by back-prop
  - Fast training

- With stochastic hard-attention: sample an input location according to the softmax output
  - Get a gradient on the decisions via REINFORCE - baseline
  - Noisy gradient, slower training but works
  - Symmetry breaking
Attention-Based Neural Machine Translation

Related to earlier Graves 2013 for generating handwriting

- (Bahdanau, Cho & Bengio, arXiv sept. 2014)
- (Jean, Cho, Memisevic & Bengio, arXiv dec. 2014)

\[ e = \text{(Economic, growth, has, slowed, down, in, recent, years, .)} \]

\[ f = \text{(La, croissance, économique, s'est, ralentie, ces, dernières, années, .)} \]
Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight $w_{ij}$ of the annotation of the $j$-th source word and the $i$-th target word (see Eq. (5)), in grayscale (0: black, 1: white).

(a) an arbitrary sentence. (b–d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.

The encoder and decoder of the RNNenc have 1000 hidden units each. The encoder of the RNNsearch consists of forward and backward recurrent neural networks (RNN) each having 1000 hidden units. Its decoder has 1000 hidden units. In both cases, we use a multilayer network with a single maxout (Goodfellow et al., 2013) hidden layer to compute the conditional probability of each target word (Pascanu et al., 2014).

We use a stochastic gradient descent (SGD) algorithm together with Adadelta (Zeiler, 2012) to train each model. Each SGD updated direction is computed using a minibatch of 80 sentences. We trained each model approximately 5 days.

In this paper, by a 'hidden unit', we always mean the gated hidden unit (see Sec. 3.3.1).
En-Fr & En-De Alignments

Economic growth has slowed down in recent years.

La croissance économique s'est ralentie ces dernières années.

Economic growth has slowed down in recent years.

Das Wirtschaftswachstum hat sich in den letzten Jahren verlangsamt.
Improvements over Pure AE Model

- RNNenc: encode whole sentence
- RNNsearch: predict alignment
- BLEU score on full test set (including UNK)
And, the rest is history..

<table>
<thead>
<tr>
<th></th>
<th>NMT(A)</th>
<th>NMT(A)-LV</th>
<th>Google</th>
<th>P-SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic NMT</td>
<td>29.48</td>
<td>32.68</td>
<td>30.6*</td>
<td></td>
</tr>
<tr>
<td>+Candidate List</td>
<td>−</td>
<td>33.28</td>
<td>−</td>
<td>33.3*</td>
</tr>
<tr>
<td>+UNK Replace</td>
<td>32.49</td>
<td>33.99</td>
<td>32.7°</td>
<td>37.03°</td>
</tr>
<tr>
<td>+Ensemble</td>
<td>−</td>
<td>36.71</td>
<td>36.9°</td>
<td></td>
</tr>
</tbody>
</table>

(a) **English→French**

<table>
<thead>
<tr>
<th></th>
<th>NMT(A)</th>
<th>NMT(A)-LV</th>
<th>P-SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic NMT</td>
<td>16.02</td>
<td>16.95</td>
<td></td>
</tr>
<tr>
<td>+Candidate List</td>
<td>−</td>
<td>17.51</td>
<td>20.67°</td>
</tr>
<tr>
<td>+UNK Replace</td>
<td>18.27</td>
<td>18.87</td>
<td></td>
</tr>
<tr>
<td>+Ensemble</td>
<td>−</td>
<td>20.98</td>
<td></td>
</tr>
</tbody>
</table>

(b) **English→German**

NMT(A): (Bahdanau et al., 2014), NMT(A)-LV: (Jean et al., 2014),
(⋆): (Sutskever et al., 2014), (○): (Luong et al., 2014),
(●): (Durrani et al., 2014), (⋆): (Cho et al., 2014), (○): (Buck et al., 2014)
Translating from Other Sources?

- Speech
- Images
- Video
A challenge

- Raw speech is sampled at 8kHz – 16kHz ⇒ at minimum 8000 numbers every second (44000 samples of 2 values for stereo music)
- On TIMIT:
  - 2.5 words per second
  - 12.5 phones per second
- The output is much shorter!

How to align?

- How to handle different duration of output tokens?

Speech

Words

### Multiple Time Scales in Speech

The higher-level sequence in the attention architecture can be shorter than the lower-level one.

- Speech signal
- Frames (100 per second, c.a. 8 per phone)
- Sub-phonemic units (3 per phone)
- Phones (5 per word)
- Words
Acoustic-to-Phones Attention Alignment

Michael colored the bedroom wall with crayons.
3 Inputs to Attention Mechanism

- Higher-level RNN state at current output location
- Lower-level RNN state at all input locations
- Previous attention pattern (for previous output location)

- All three were necessary to apply attention-based models to speech recognition (Chorowski et al, 2014, 2015)
Left-to-Right Soft Constraint

- Whereas with translation the word order can change a lot, the acoustic→phonetic mapping is mostly left-to-right.
- The strength of that prior can be learned by structuring the attention location probability distribution:

\[
\alpha_{ot} \propto d(t - \mathbb{E}_{\alpha_{o-1}}[t]) \exp(m(s_{o-1}, h_t))
\]

Prior: relative to last selection
Match between previous state and each input

\[
c_0 = \sum_{t=1}^{T} \alpha_{ot} f_t
\]
End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results

(Chorowski et al, NIPS submission, 2015)

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model</td>
<td>15.9%</td>
<td>18.7%</td>
</tr>
<tr>
<td>Baseline + Conv. Features</td>
<td>16.1%</td>
<td>18.0%</td>
</tr>
<tr>
<td>Baseline + Conv. Features + Smooth Focus</td>
<td>15.8%</td>
<td><strong>17.6%</strong></td>
</tr>
<tr>
<td>RNN Transducer [17]</td>
<td>N/A</td>
<td>17.7%</td>
</tr>
<tr>
<td>HMM over Time and Frequency Convolutional Net [26]</td>
<td>13.9%</td>
<td>16.7%</td>
</tr>
</tbody>
</table>

Table 1: Phoneme error rates (PER). The bold-faced PER corresponds to the best error rate with an attention-based recurrent sequence generator (ARSG) incorporating convolutional attention features and a smooth focus.

Figure 4: Results of force-aligning the concatenated utterances. Each dot represents a single utterance created by either concatenating multiple copies of the same utterance, or of different, randomly chosen utterances. We clearly see that the highest robustness is achieved when the hybrid attention mechanism is combined with the proposed sharpening technique (see the bottom-right plot.)

5.1 Forced Alignment of Long Utterances

The good performance of the baseline model led us to the question of how it distinguishes between repetitions of similar phoneme sequences and how reliably it decodes longer sequences with more repetitions. We created two datasets of long utterances; one by repeating each test utterance, and the other by concatenating randomly chosen utterances. In both cases, the waveforms were cross-faded with a 0.05s silence inserted as the “pau” phone. We concatenated up to 15 utterances.

First, we checked the forced alignment with these longer utterances by forcing the generator to emit the correct phonemes. Each alignment was considered correct if 90% of the alignment weight lies inside the ground-truth phoneme window extended by 20 frames on each side. Under this definition, all phones but the eos shown in Fig. 3 are properly aligned.

The first column of Fig. 4 shows the number of correctly aligned frames w.r.t. the utterance length (in frames) for some of the considered models. One can see that the baseline model was able to decode sequences up to about 120 phones when a single utterance was repeated, and up to about 150 phones when different utterances were concatenated. Even when it failed, it correctly aligned about 50 phones. On the other hand, the model with the hybrid attention mechanism with convolutional features was able to align sequences up to 200 phones long. However, once it began to fail, the model was not able to align almost all phones. The model with the smoothing behaved similarly to the one with convolutional features only.

We examined failed alignments to understand these two different modes of failure. Some of the examples are shown in the Supplementary Materials.

We found that the baseline model properly aligns about 40 first phones, then makes a jump to the end of the recording and cycles over the last 10 phones. This behavior suggests that it learned to track its approximate location in the source sequence. However, the tracking capability is limited to the lengths observed during training. Once the tracker saturates, it jumps to the end of the recording.
Image-to-Text: Caption Generation

\[ f = (a, \text{ man, is, jumping, into, a, lake, .}) \]

\[ \sum a_j = 1 \]
Paying Attention to Selected Parts of the Image While Uttering Words

1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

14x14 Feature Map

A
bird
flying over a body of
water
Speaking about what one sees
Let's go back 2.5 years back in time.. (Mitchell et al., 2012)

The bus by the road with a clear blue sky

1. Group the Nouns
2. Order the Nouns
3. Filter Incorrect Attributes
4. Group Plurals
5. Gather Local Sub-(parse) trees
6. Create Full Trees
7. Get Final Tree, Clear Mark-Up
8. Prenominal Modifier Ordering
And in 2015...
End-to-End Neural Net

* A woman in a bikini holding a surfboard.

The neural nets successfully learned to

- map a phrase in one language to that in another language
- extract semantics [and syntax] of a sentence
- separate different objects in an image
- separate the background from foreground objects
- create a syntactically and semantically correct sentence
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Results from (Xu et al, arXiv Jan. 2015, ICML 2015)

Table 1. BLEU-1,2,3,4/METEOR metrics compared to other methods, † indicates a different split, (—) indicates an unknown metric, ○ indicates the authors kindly provided missing metrics by personal communication, Σ indicates an ensemble, a indicates using AlexNet

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B-1</td>
</tr>
<tr>
<td>Flickr8k</td>
<td>Google NIC (Vinyals et al., 2014)†Σ</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear (Kiros et al., 2014a)○</td>
<td>65.6</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>67</td>
</tr>
<tr>
<td>Flickr30k</td>
<td>Google NIC†○Σ</td>
<td>66.3</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear</td>
<td>60.0</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>66.9</td>
</tr>
<tr>
<td>COCO</td>
<td>CMU/MS Research (Chen &amp; Zitnick, 2014)a</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>MS Research (Fang et al., 2014)†a</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>BRNN (Karpathy &amp; Li, 2014)○</td>
<td>64.2</td>
</tr>
<tr>
<td></td>
<td>Google NIC†○Σ</td>
<td>66.6</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear○</td>
<td>70.8</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>70.7</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>71.8</td>
</tr>
</tbody>
</table>
The Good

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.
And the Bad

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.
Attention through time for video caption generation

- (Yao et al arXiv 1502.08029, 2015) *Video Description Generation Incorporating Spatio-Temporal Features and a Soft-Attention Mechanism*
- Attention can be focused temporally, i.e., selecting input frames
Attention through time for video caption generation (Yao et al 2015)

- Attention is focused at appropriate frames depending on which word is generated.
Attention through time for video caption generation (Yao et al. 2015)

- Soft-attention worked best in this setting

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature</th>
<th>Bleu 1</th>
<th>Bleu 2</th>
<th>Bleu 3</th>
<th>Bleu 4</th>
<th>meteor</th>
<th>perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-attention</td>
<td>GNet</td>
<td>32.0</td>
<td>9.2</td>
<td>3.4</td>
<td>1.2</td>
<td>0.3</td>
<td>4.43</td>
</tr>
<tr>
<td></td>
<td>GNet+3DConv_{non-att}</td>
<td>33.6</td>
<td>10.4</td>
<td>4.3</td>
<td>1.8</td>
<td>0.7</td>
<td>5.73</td>
</tr>
<tr>
<td>soft-attention</td>
<td>GNet</td>
<td>31.0</td>
<td>7.7</td>
<td>3.0</td>
<td>1.2</td>
<td>0.3</td>
<td>4.05</td>
</tr>
<tr>
<td></td>
<td>GNet+3DConv_{att}</td>
<td>28.2</td>
<td>8.2</td>
<td>3.1</td>
<td>1.3</td>
<td>0.7</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Table 1. Attention and 3D-Conv performances evaluation on YouTube2Text. Bleu 1-4, meteor, and perplexity metrics are reported.

Figure 3. A visualization of where the soft-attentional model "looks at" in a video, while generating the captions (captions included on the left). Each word is mapped into a vector of $\tilde{v}$ in Equation 6. Only bars in the same row are comparable, and their height reflects the magnitude of $\tilde{v}$. The model is able to focus its attention on different frames of the video when generating different words in the caption. Best viewed with zooming-in on pdf.

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature</th>
<th>Bleu 1</th>
<th>Bleu 2</th>
<th>Bleu 3</th>
<th>Bleu 4</th>
<th>meteor</th>
<th>perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-attention</td>
<td>GNet</td>
<td>32.0</td>
<td>9.2</td>
<td>3.4</td>
<td>1.2</td>
<td>0.3</td>
<td>4.43</td>
</tr>
<tr>
<td></td>
<td>GNet+3DConv_{non-att}</td>
<td>33.6</td>
<td>10.4</td>
<td>4.3</td>
<td>1.8</td>
<td>0.7</td>
<td>5.73</td>
</tr>
<tr>
<td>soft-attention</td>
<td>GNet</td>
<td>31.0</td>
<td>7.7</td>
<td>3.0</td>
<td>1.2</td>
<td>0.3</td>
<td>4.05</td>
</tr>
<tr>
<td></td>
<td>GNet+3DConv_{att}</td>
<td>28.2</td>
<td>8.2</td>
<td>3.1</td>
<td>1.3</td>
<td>0.7</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Table 2. Attention and 3D-Conv performances evaluation on DVS. Bleu 1-4, meteor, and perplexity metrics are reported.

Have encountered in this dataset is that its captions cover a much wider domain, rendering this task challenging for both non-attention and attention models. According to Table 2, by comparing on perplexity, the attention models improve consistently upon non-attention models. Given the same type of model, using GNet+3DConv features also steadily improves upon using GNet features alone. In fact, using attention models offers about 20 improvement on perplexity upon non-attention models. With the same model type, using the combined features also results better Bleu and Meteor, while the effect on Bleu and Meteor score across model types is less obvious.

Generated captions

Corpus: She rushes out.
Test sample: The woman turns away.

Corpus: SOMEONE sits with his arm around SOMEONE. He nuzzles her cheek, then kisses tenderly.
Test sample: SOMEONE sits beside SOMEONE.

Corpus: SOMEONE shuts the door.
Test sample: as he turns on his way to the door, SOMEONE turns away.
Attention Mechanisms for Memory Access

• Neural Turing Machines (Graves et al 2014)
• and Memory Networks (Weston et al 2014)
• Use a form of attention mechanism to control the read and write access into a memory
• The attention mechanism outputs a softmax over memory locations
• For efficiency, the softmax should be sparse (mostly 0’s), e.g. maybe using a hash-table formulation.
Sparse Access Memory for Long-Term Dependencies

- Whereas LSTM memories always decay exponentially (even if slowly), a mental state stored in an external memory can stay for arbitrarily long durations, until evoked for read or write.
- Need to replace the soft gater or softmax attention by hard one that is 0 most of the time, and yet for which training works (again, may use noisy decisions and/or REINFORCE).
- Different « threads » can run in parallel if we view the memory as an associative one.
Conclusions

- Theory for deep learning has progressed substantially on several fronts: why it generalizes better, why local minima are not the issue people thought, and the probabilistic interpretation of deep unsupervised learning.
- Attention mechanisms allow the learner to make a selection, soft or hard
- They have been extremely successful for machine translation and caption generation
- They could be interesting for speech recognition and video, especially if we used them to capture multiple time scales
- They could be used to help deal with long-term dependencies, allowing some states to last for arbitrarily long
MILA: Montreal Institute for Learning Algorithms