The Deep Learning Revolution: Progress, Promise, and Profligate Promotion

Raymond J. Mooney
University of Texas at Austin

Deep Learning Revolution

• Recent machine learning methods for training “deep” neural networks (NNs) have demonstrated remarkable progress on many challenging AI problems (e.g. speech recognition, visual object recognition, machine translation, game playing).
• However, their capabilities are prone to “hype.”
• Deep learning has not “solved” AI and current methods have clear limitations.

Very Brief History of Machine Learning

• Symbolic AI & knowledge engineering (1970-1985)
• Multi-layer NNs and symbolic learning (1985-1995)
• Statistical (Bayesian) learning and kernel methods (1995-2010)
• Deep learning (CNNs and RNNs) (2010-?)
Classification

• Given:
  – A description of an instance, \( x \in X \), where \( X \) is the instance language or instance space.
  – A finite set of classes/categories: \( C = \{c_1, c_2, \ldots, c_n\} \)

• Determine:
  – The class of \( x: c(x) \in C \), where \( c(x) \) is a classification function whose domain is \( X \) and whose range is \( C \).

Learning for Classification

• A training example is an instance \( x \in X \), paired with its correct class \( c(x) \): \( <x, c(x)> \) for an unknown classification function, \( c \).

• Given a set of training examples, \( D \).

• Find a hypothesized classification function, \( h(x) \), such that:
  \[
  \forall <x, c(x)> \in D : h(x) = c(x)
  \]

Simple Classification Learning Problem

• Instance language: <size, color, shape>
  – size \( \in \) {small, medium, large}
  – color \( \in \) {red, blue, green}
  – shape \( \in \) {square, circle, triangle}
  – \( C = \) {positive, negative}

• \( D \):

<table>
<thead>
<tr>
<th>Example</th>
<th>Size</th>
<th>Color</th>
<th>Shape</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>small</td>
<td>red</td>
<td>circle</td>
<td>positive</td>
</tr>
<tr>
<td>2</td>
<td>large</td>
<td>red</td>
<td>circle</td>
<td>positive</td>
</tr>
<tr>
<td>3</td>
<td>small</td>
<td>red</td>
<td>triangle</td>
<td>negative</td>
</tr>
<tr>
<td>4</td>
<td>large</td>
<td>blue</td>
<td>circle</td>
<td>negative</td>
</tr>
</tbody>
</table>
Single-Layer Neural Network
(Linear Threshold Unit)

• Mathematical model of an individual neuron.

Perceptron

• Rosenblatt (1957) developed an iterative, hill-climbing algorithm for learning the weights of single-layer NN to try to fit a set of training examples.
• Unable to learn or represent many classification functions (e.g. XOR), only the “linearly separable” ones are learnable.

Perceptron Learning Rule

• Update weights by:

\[ w_i = w_i + \eta(t - o)x_i \]

where \( \eta \) is the “learning rate,” \( t \) is the teacher output, and \( o \) is the network output.
• Equivalent to rules:
  – If output is correct do nothing.
  – If output is high, lower weights on active inputs
  – If output is low, increase weights on active inputs
Perceptron Learning Algorithm

- Iteratively update weights until convergence.
- Initialize weights to random values
- Until outputs of all training examples are correct
  - For each training pair, \( E \), do:
    - Compute current output \( o \) for \( E \) given its inputs
    - Compare current output to target value, \( t \), for \( E \)
    - Update weights using learning rule

Perceptron Demise

- Work on neural-networks dissipated during the 70’s and early 80’s.

Neural Net Resurgence (1986)

- Interest in NNs revived in the mid 1980’s due to the rise of “connectionism.”
- Backpropagation algorithm popularized for training three-layer NN’s.
- Generalized the iterative “hill climbing” method to approximate fitting two layers of synaptic connections, but no convergence guarantees.
Second NN Demise (1995-2010)

- Generic backpropagation did not generalize that well to training deeper networks.
- Little theoretical justification for underlying methods.
- Machine learning research moved to graphical models and kernel methods.

Deep Learning Revolution (2010...)

- Improved methods developed for training deep neural works.
- Particular successes with:
  - Convolutional neural nets (CNNs) for vision.
  - Recurrent neural nets (RNNs) for machine translation and speech recognition.
  - Deep reinforcement learning for game playing.
Massive Data and Specialized Hardware

• Large collections of supervised (crowdsourced) training data has been critical.
• Efficient processing of this big data using specialized hardware (Graphics Processing Units, GPUs) has been critical.

CNNs

• Convolutional layers learn to extract local features from image regions (receptive fields) analogous to human vision (LeCun, et al., 1998).
• Deeper layers extract higher-level features.
• Pool activity of multiple neurons into one at the next layer using max or mean.
• Nonlinear processing with Rectified Linear Units (ReLUs)
• Decision made using final fully connected layers.
**ImageNet Large Scale Visual Recognition Challenge (ILSVRC)**

- Recognize 1,000 categories of objects in 150K test images (given 1.2M training images).

**ImageNet Performance Over Time**

- CNNs introduced

**Recurrent Neural Networks (RNNs)**

- Add feedback loops where some units’ current outputs determine some future network inputs.
- RNNs can model dynamic finite-state machines, beyond the static combinatorial circuits modeled by feed-forward networks.
Simple Recurrent Network (SRN)

- Additional input to hidden layer is the state of the hidden layer in the previous time.

Unrolled RNN

- Behavior of RNN is perhaps best viewed by “unrolling” the network over time.

Training RNN’s

- RNNs can be trained using “backpropagation through time.”
- Can viewed as applying normal backprop to the unrolled network.
Vanishing/Exploding Gradient Problem
• Backpropagated errors multiply at each layer, resulting in exponential decay (if derivative is small) or growth (if derivative is large).
• Makes it very difficult to train deep networks, or simple recurrent networks over many time steps.

Long Distance Dependencies
• It is very difficult to train SRNs to retain information over many time steps.
• This makes it very difficult to learn SRNs that handle long-distance dependencies, such as subject-verb agreement.

Long Short Term Memory (LSTM)
• LSTM networks, add additional gating units in each memory cell (Hochreiter & Schmidhuber, 1997).
  – Forget gate
  – Input gate
  – Output gate
• Prevents vanishing/exploding gradient problem and allows network to retain state information over longer periods of time.
LSTM Network Architecture

Sequence to Sequence (Seq2Seq) Transduction
• Encoder/Decoder framework maps one sequence to a "deep vector" then another LSTM maps this vector to an output sequence (Sutskever et al., 2014).

Train model "end to end" on I/O pairs of sequences.

Neural Machine Translation (NMT)
• LSTM Seq2Seq has lead to a new approach to translating human language.
• NMT modestly outperforms previous statistical learning approaches to MT (SMT).
NMT Results (Wu et al., 2016)

- Experimental results using automated (BLEU) and human evaluation for English→French translation.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>Human Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT</td>
<td>37.0</td>
<td>3.87</td>
</tr>
<tr>
<td>NMT</td>
<td>40.35</td>
<td>4.46</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td>4.82</td>
</tr>
</tbody>
</table>

Deep Speech Recognition

- RNNs have also helped significantly improve speech recognition.
  - Continuous server ASR word error rate (WER) reduction ~18% / year: combination of algorithms, data, and computing.
  - Deep learning (DL) is driving recent performance improvements in ASR and relating extraction.

Chinese Speech Recognition

Deep Speech – Mandarin

- SNN Mandarin Speech Performance
Video Description

- LSTM Seq2Seq models also used to map sequence of image frames to sequence of English words describing the video.

S2VT: Sequence to Sequence Video to Text (Venugopalan et al., 2015)

<table>
<thead>
<tr>
<th>Method</th>
<th>Relevance</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2VT</td>
<td>2.06</td>
<td>3.76</td>
</tr>
<tr>
<td>S2VT Combined</td>
<td>2.24*</td>
<td>4.20*</td>
</tr>
<tr>
<td>Human</td>
<td>4.52</td>
<td>4.47</td>
</tr>
</tbody>
</table>
Deep Reinforcement Learning

- Reinforcement learning uses delayed rewards to learn to perform sequential decision making problems.
- Deep CNNs successfully used to learn “policies” that map images of states to recommended actions in those states.
AlphaGo

- Go playing program developed by Google Deep Mind.
- CNN pretrained to select winning move from 30M position in 130K human expert games.
- CNN fine-tuned to pick winning move in 1.3M games of self-play.
AlphaGo Successes

- Played international champion Lee Sedol
  March 9-15, 2016 in Seoul using 1,920
  CPUs and 280 GPUs.
  - AlphaGo won 4 of 5 games
- Played world champion Ke Jie, May 23-27,
  2017 in Wuzhen using 1 TPU on 1 machine.
  - AlphaGo won all 3 games

Future Promise

- Improvement in basic vision and language
  processing with DNNs is continuing.
- Progress will enable many impactful future
  applications, e.g.:
  - Autonomous vehicles
  - Conversational agents (Siri, Alexa, etc.)
  - Home & industrial robots

Beware of the Hype

- Recent progress with DNNs have lead to
  exaggerated claims about their capabilities.
- Hard AI problems in vision and language
  are not completely solved.
- Current DNN methods still face
  fundamental limitations.
Adversarial Examples

• Deep networks can be “fooled” by specially constructed examples.
• Humans do not find the perturbed examples to be perceptibly changed.
• Illustrate that DNNs do not closely model human capabilities.

Adversarial Images (Papernot et al., 2017)

• Images can be maliciously perturbed to cause DNNs to misclassify them.

Adversarial Text (Jia & Liang, 2017)

• Text can be maliciously augmented to confuse DNN question answering systems.

Article: Super Bowl 50
Paragraph: "Peyton Manning became the first-quarterback to start in multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The quarterback was held by John Deese, who led the Broncos to victory in Super Bowl XXXII in 1997. Manning was named the Super Bowl Most Valuable Player and the Most Valuable Player in the game."
Question: "What is the name of the quarterback who won Super Bowl XXXII?"
Original Prediction: John Deese
Prediction under adversary: Jeff Dean
“Cramming” Meaning into Vectors

• DNNs force semantics to be encoded into real-valued vectors.
• Structured meaning representations that exploit trees, graphs, and logical representations are only imperfectly encoded as vectors.

Complex Compositional Questions

“Has Woody Allen made more movies with Diane Keaton or Mia Farrow.”

\[
\arg\max_{X \in \{\text{Diane Keaton, Mia Farrow}\}} \text{count}(Y, \text{Director}(Y, \text{Woody Allen}) \land \text{Cast}(Y, X))
\]

Conclusions

• Machine learning, and specifically neural nets, has a long, rich, varied history.
• Deep learning has made significant recent progress.
• Progress is continuing and holds promise of enabling revolutionary technology.
• However, progress has been exaggerated and core AI problems are a long way from completely solved.