Deep Learning in NLP

Many slides adapted from Richard Socher, Tom Mitchell
What’s Deep Learning (DL)?

- Deep learning is a subfield of machine learning

- Most machine learning methods work well because of human-designed representations and input features
  - For example: features for finding named entities like locations or organization names (Finkel, 2010):

- Machine learning becomes just optimizing weights to best make a final prediction

<table>
<thead>
<tr>
<th>Feature</th>
<th>NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Word</td>
<td>✓</td>
</tr>
<tr>
<td>Previous Word</td>
<td>✓</td>
</tr>
<tr>
<td>Next Word</td>
<td>✓</td>
</tr>
<tr>
<td>Current Word Character n-gram</td>
<td>✓</td>
</tr>
<tr>
<td>Current POS Tag</td>
<td>✓</td>
</tr>
<tr>
<td>Surrounding POS Tag Sequence</td>
<td>✓</td>
</tr>
<tr>
<td>Current Word Shape</td>
<td>✓</td>
</tr>
<tr>
<td>Surrounding Word Shape Sequence</td>
<td>✓</td>
</tr>
<tr>
<td>Presence of Word in Left Window</td>
<td>size 4</td>
</tr>
<tr>
<td>Presence of Word in Right Window</td>
<td>size 4</td>
</tr>
</tbody>
</table>
Machine Learning vs Deep Learning

Machine Learning in Practice

Describing your data with features a computer can understand

Learning algorithm

Domain specific, requires Ph.D. level talent

Optimizing the weights on features
What’s Deep Learning (DL)?

• **Representation learning** attempts to automatically learn good features or representations.

• **Deep learning** algorithms attempt to learn (multiple levels of) representation and an output.

• From “raw” inputs \( x \) (e.g. words)
Reasons for Exploring Deep Learning

• Manually designed features are often over-specified, incomplete and take a long time to design and validate

• **Learned Features** are easy to adapt, fast to learn

• Deep learning provides a very flexible, (almost?) universal, learnable framework for **representing** world, visual and linguistic information.

• Deep learning can learn **unsupervised** (from raw text) and **supervised** (with specific labels like positive/negative)
Reasons for Exploring Deep Learning

• In 2006 deep learning techniques started outperforming other machine learning techniques. Why now?

• DL techniques benefit more from a lot of data
• Faster machines and multicore CPU/GPU help DL
• New models, algorithms, ideas

→ **Improved performance** (first in speech and vision, then NLP)
Deep Learning for Speech

- The first breakthrough results of “deep learning” on large datasets happened in speech recognition.
- Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition
  Dahl et al. (2010)

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>Recog \ WER</th>
<th>RT03S FSH</th>
<th>Hub5 SWB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional features</td>
<td>1-pass adapt</td>
<td>27.4</td>
<td>23.6</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>1-pass adapt</td>
<td>18.5 (−33%)</td>
<td>16.1 (−32%)</td>
</tr>
</tbody>
</table>
Deep Learning for Computer Vision

- Most deep learning groups have (until recently) largely focused on computer vision.

Zeiler and Fergus (2013)
Neural word vectors - visualization
Representations at NLP Levels: Syntax

- **Traditional:** Phrases
  Discrete categories like NP, VP

- **DL:**
  - Every word and every phrase is a vector
  - a neural network combines two vectors into one vector
  - Socher et al. 2011
Machine Translation

• Many levels of translation have been tried in the past:

• Traditional MT systems are very large complex systems

• What do you think is the interlingua for the DL approach to translation?

Figure 1: The Vauquois triangle
Machine Translation

- Source sentence mapped to vector, then output sentence generated.

- Sequence to Sequence Learning with Neural Networks by Sutskever et al. 2014

- Very new but could replace very complex architectures!
Perceptron

\[ o(x_1, \ldots, x_n) = \begin{cases} 
1 & \text{if } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\
-1 & \text{otherwise.} 
\end{cases} \]

Sometimes we’ll use simpler vector notation:

\[ o(\vec{x}) = \begin{cases} 
1 & \text{if } \vec{w} \cdot \vec{x} > 0 \\
-1 & \text{otherwise.} 
\end{cases} \]
Neural Nets for the Win!

- Neural networks can learn much more complex functions and nonlinear decision boundaries!
Multilayer Networks of Sigmoid Units
Sigmoid Unit

\[ \sigma(x) \text{ is the sigmoid function} \]

\[ \frac{1}{1 + e^{-x}} \]

Nice property: \( \frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x)) \)
We can derive gradient descent rules to train

- One sigmoid unit

- Multilayer networks of sigmoid units → Backpropagation
Automatic Differentiation

- The gradient computation can be automatically inferred from the symbolic expression of the fprop.
- Each node type needs to know how to compute its output and how to compute the gradient wrt its inputs given the gradient wrt its output.
- Easy and fast prototyping
Review

• Deep Learning
  • Learning Representations of Inputs

• Neural Networks
  • Layers of Logistic Regression
  • Can represent any nonlinear function (with a large enough network)
  • Training with backpropagation

• Recent breakthroughs in predictive tasks
  • Speech Recognition
  • Object Recognition (computer vision)
Neural Network Language Models
Q: How to model sequences with neural Networks?

• Fixed number of inputs.
How about just predicting the next word in the input?

• Q: what about just predicting the next word?
• From the context?
  • No longer a language model
• Word2Vec!
Word2vec

• Learn continuous word embedding for each word
  • Each word represented by a vector

![Diagram of Word2vec model]

Figure 1: A simple CBOW model with only one word in the context
Using more than one word of context

Figure 2: Continuous bag-of-word model

Figure 3: The skip-gram model.
Word2Vec: fast to train

• Word2Vec is a fairly simple model,
• But Can efficiently train word vectors on really big corpora
• This is probably the main advantage of Word2vec over other approaches...
  • Principal Component Analysis
  • Recurrent Neural Network Language Models
VEC(king) + VEC(woman) - VEC(man) = ?
Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates the ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.
The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition

<table>
<thead>
<tr>
<th>System</th>
<th>Fin10Dev</th>
<th>Rit11</th>
<th>Fro14</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL</td>
<td>27.3</td>
<td>27.1</td>
<td>29.5</td>
<td>28.0</td>
</tr>
<tr>
<td>+ Brown</td>
<td>38.4</td>
<td>39.4</td>
<td>42.5</td>
<td>40.1</td>
</tr>
<tr>
<td>+ Vector</td>
<td>40.8</td>
<td>40.4</td>
<td>42.9</td>
<td>41.4</td>
</tr>
<tr>
<td>+ Reps</td>
<td>42.4</td>
<td>42.2</td>
<td>46.2</td>
<td>43.6</td>
</tr>
<tr>
<td>Fin10</td>
<td>36.7</td>
<td>29.0</td>
<td>30.4</td>
<td>32.0</td>
</tr>
<tr>
<td>+ Brown</td>
<td>59.9</td>
<td>53.9</td>
<td>56.3</td>
<td>56.7</td>
</tr>
<tr>
<td>+ Vector</td>
<td>61.5</td>
<td>56.4</td>
<td>58.4</td>
<td>58.8</td>
</tr>
<tr>
<td>+ Reps</td>
<td>64.0</td>
<td>58.5</td>
<td>60.2</td>
<td>60.9</td>
</tr>
<tr>
<td>CoNLL+Fin10</td>
<td>44.7</td>
<td>39.9</td>
<td>44.2</td>
<td>42.9</td>
</tr>
<tr>
<td>+ Brown</td>
<td>54.9</td>
<td>52.9</td>
<td>58.5</td>
<td>55.4</td>
</tr>
<tr>
<td>+ Vector</td>
<td>58.9</td>
<td>55.2</td>
<td>59.9</td>
<td>58.0</td>
</tr>
<tr>
<td>+ Reps</td>
<td>58.9</td>
<td>56.4</td>
<td>61.8</td>
<td>59.0</td>
</tr>
<tr>
<td>+ Weights</td>
<td>64.4</td>
<td>59.6</td>
<td>63.3</td>
<td>62.4</td>
</tr>
</tbody>
</table>

Table 5: Impact of our components on Twitter NER performance, as measured by F1, under 3 data scenarios.

Neural Translation Models (sequence to sequence)
ConversaCon  
GeneraCon

Conversation Generation

Human: what is the purpose of life?
Machine: to serve the greater good.
Human: what is the purpose of living?
Machine: to live forever.
Human: what is the purpose of existence?
Machine: to find out what happens when we get to the planet earth.
Human: where are you now?
Machine: i’m in the middle of nowhere.
Human: what is the purpose of dying?
Machine: to have a life.
Human: what is the purpose of being intelligent?
Machine: to find out what it is.
Human: what is the purpose of emotions?
Machine: i don’t know.
Computer, respond to this email.

Tuesday, November 03, 2015

Posted by Greg Corrado*, Senior Research Scientist

Machine Intelligence for You

What I love about working at Google is the opportunity to harness cutting-edge machine intelligence for users’ benefit. Two recent Research Blog posts talked about how we’ve used machine learning in the form of deep neural networks to improve voice search and YouTube thumbnails. Today we can share something even wilder -- Smart Reply, a deep neural network that writes email.
Turkey!

Hi all,
We wanted to invite you to join us for an early Thanksgiving on November 22nd, beginning around 2PM. Please bring your favorite dish! RSVP by next week.

Dave

Server issues

Dan Mané

Hi team,
The server appears to be dropping about 10% of requests (see attached dashboards). There hasn’t been a new release since last night, so I’m not sure what’s going on. Is anyone looking into this?

...
Show and Tell: A Neural Image Caption Generator

Oriol Vinyals
Google
vinyals@google.com

Alexander Toshev
Google
toshev@google.com

Samy Bengio
Google
bengio@google.com

Dumitru Erhan
Google
dumitru@google.com
[Krizhevsky, Sutskever, Hinton. 2012] 16.4% error

[Zeiler and Fergus, 2013] 11.1% error

[Simonyan and Zisserman, 2014] 7.3% error

[Zeiler and Fergus, 2013] 11.1% error
A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.
Image Sentence Datasets

Microsoft COCO
[Tsung-Yi Lin et al. 2014]
mscoco.org

currently:
~120K images
~5 sentences each
Summary of the approach

We wanted to describe images with sentences.

1. Define a single function from input -> output
2. Initialize parts of net from elsewhere if possible
3. Get some data
4. Train with SGD
Wow I can’t believe that worked

- A group of people standing around a room with remotes
  logprob: -9.17

- A young boy is holding a baseball bat
  logprob: -7.61

- A cow is standing in the middle of a street
  logprob: -8.84
Wow I can’t believe that worked
Well, I can kind of see it

- a man standing next to a clock on a wall  
  logprob: -10.08
- a young boy is holding a baseball bat  
  logprob: -7.65
- a cat is sitting on a couch with a remote control  
  logprob: -12.45
Summary

• Deep learning is a popular area in machine learning recently
  • Very successful in speech recognition and computer vision

• Becoming very popular in NLP these days

• Main motivation:
  • Learn feature representations from data
  • Alternative to hand-engineered features

• Neural networks:
  • Primary deep learning approach
  • Layers of logistic regressions – can directly calculate gradients from outputs
  • Nonlinear decision boundaries