CSE 573: Artificial Intelligence

Hanna Hajishirzi
Neural Nets

slides adapted from
Dan Klein, Pieter Abbeel ai.berkeley.edu
And Dan Weld, Luke Zettelmoyer
Trend in NLP

Over time:

- Learning HMMs (or related Probabilistic-based methods) with hand-designed features (tokens, syntactic features)

- Recurrent Neural Networks:
  - replaces probabilistic dynamic model with neural functions (mostly non-linear functions)

- Attention-based methods:
  - Adds the capability to go beyond Markov Models
Linear Classifiers

- Inputs are feature values
- Each feature has a weight
- Sum is the activation

\[ \text{activation}_w(x) = \sum_i w_i \cdot f_i(x) = w \cdot f(x) \]

- If the activation is:
  - Positive, output +1
  - Negative, output -1
How to get probabilistic decisions?

- Activation: \( z = w \cdot f(x) \)
- If \( z = w \cdot f(x) \) very positive \( \rightarrow \) want probability going to 1
- If \( z = w \cdot f(x) \) very negative \( \rightarrow \) want probability going to 0

- Sigmoid function

\[
\phi(z) = \frac{1}{1 + e^{-z}}
\]
Best $w$?

- Maximum likelihood estimation:

$$\max_w \ ll(w) = \max_w \sum_i \log P(y^{(i)}|x^{(i)}; w)$$

with:

$$P(y^{(i)} = +1| x^{(i)}; w) = \frac{1}{1 + e^{-w \cdot f(x^{(i)})}}$$

$$P(y^{(i)} = -1| x^{(i)}; w) = 1 - \frac{1}{1 + e^{-w \cdot f(x^{(i)})}}$$

= Logistic Regression
Neural Networks
Multi-class Logistic Regression

- special case of neural network

\[ P(y_1 | x; w) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \]

\[ P(y_2 | x; w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \]

\[ P(y_3 | x; w) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}} \]
Deep Neural Network = Also learn the features!

\[
P(y_1|x; w) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}
\]

\[
P(y_2|x; w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}
\]

\[
P(y_3|x; w) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}
\]
Deep Neural Network = Also learn the features!

\[ z_i^{(k)} = g \left( \sum_j W_{i,j}^{(k-1,k)} z_j^{(k-1)} \right) \]

\( g \) = nonlinear activation function
Deep Neural Network = Also learn the features!


g = nonlinear activation function

\[ z_i^{(k)} = g(\sum_j W^{(k-1,k)}_{i,j} z_j^{(k-1)}) \]
Common Activation Functions

Sigmoid Function

\[ g(z) = \frac{1}{1 + e^{-z}} \]
\[ g'(z) = g(z)(1 - g(z)) \]

Hyperbolic Tangent

\[ g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \]
\[ g'(z) = 1 - g(z)^2 \]

Rectified Linear Unit (ReLU)

\[ g(z) = \max(0, z) \]
\[ g'(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise} \end{cases} \]
Deep Neural Network: Also Learn the Features!

- Training the deep neural network is just like logistic regression:

\[
\max_w \quad ll(w) = \max_w \sum_i \log P(y^{(i)}|x^{(i)}; w)
\]
Neural Networks Properties

- Theorem (Universal Function Approximators). A two-layer neural network with a sufficient number of neurons can approximate any continuous function to any desired accuracy.

- Practical considerations
  - Can be seen as learning the features
  - Large number of neurons
    - Danger for overfitting
    - (hence early stopping!)
Fun Neural Net Demo Site

- Demo-site:
  - http://playground.tensorflow.org/
Automatic Differentiation

- **Automatic differentiation software**
  - e.g. Theano, TensorFlow, PyTorch, Chainer
  - Only need to program the function $g(x,y,w)$
  - Can automatically compute all derivatives w.r.t. all entries in $w$
  - This is typically done by caching info during forward computation pass of $f$, and then doing a backward pass = “backpropagation”
  - Autodiff / Backpropagation can often be done at computational cost comparable to the forward pass

- Need to know this exists

- How this is done? -- outside of scope of CSE573
Summary of Key Ideas

- Optimize probability of label given input
  \[ \max_w \mathcal{L}(w) = \max_w \sum_i \log P(y^{(i)}|x^{(i)}; w) \]

- Continuous optimization
  - Gradient ascent:
    - Compute steepest uphill direction = gradient (= just vector of partial derivatives)
    - Take step in the gradient direction
    - Repeat (until held-out data accuracy starts to drop = “early stopping”)

- Deep neural nets
  - Last layer = still logistic regression
  - Now also many more layers before this last layer
    - = computing the features
    - \( \rightarrow \) the features are learned rather than hand-designed
  - Universal function approximation theorem
    - If neural net is large enough
    - Then neural net can represent any continuous mapping from input to output with arbitrary accuracy
    - But remember: need to avoid overfitting / memorizing the training data \( \rightarrow \) early stopping!
  - Automatic differentiation gives the derivatives efficiently (how? = outside of scope of 573)
How well does it work?
Computer Vision
Object Detection
Manual Feature Design
Features and Generalization

[HoG: Dalal and Triggs, 2005]
Features and Generalization

Image

HoG
Performance

ImageNet Error Rate 2010-2014

[Graph showing error rate from 2010 to 2014, with points marked for each year indicating error rate.]

graph credit Matt Zeiler, Clarifai
Performance

ImageNet Error Rate 2010-2014

Error Rate

- 79%
- 60%
- 40%
- 20%
- 7%

2010 2011 2012 2013 2014

AlexNet

Traditional CV
Deep Learning

graph credit Matt Zeiler, Clarifai
Performance

ImageNet Error Rate 2010-2014

- Traditional CV
- Deep Learning

Error Rate
- 79%
- 60%
- 40%
- 20%
- 7%

Year
- 2010
- 2011
- 2012
- 2013
- 2014

AlexNet

Graph credit: Matt Zeiler, Clarifai
Performance

ImageNet Error Rate 2010-2014

- Traditional CV
- Deep Learning

Error Rate

2010 2011 2012 2013 2014

AlexNet

Graph credit: Matt Zeiler, Clarifai
MS COCO Image Captioning Challenge

"man in black shirt is playing guitar."
"construction worker in orange safety vest is working on road."
"two young girls are playing with lego toy."
"boy is doing backflip on wakeboard."
"girl in pink dress is jumping in air."
"black and white dog jumps over bar."
"young girl in pink shirt is swinging on swing."
"man in blue wetsuit is surfing on wave."

Karpathy & Fei-Fei, 2015; Donahue et al., 2015; Xu et al, 2015; many more
Visual QA Challenge

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh

What vegetable is on the plate?
Neural Net: broccoli
Ground Truth: broccoli

What color are the shoes on the person's feet?
Neural Net: brown
Ground Truth: brown

How many school busses are there?
Neural Net: 2
Ground Truth: 2

What sport is this?
Neural Net: baseball
Ground Truth: baseball

What is on top of the refrigerator?
Neural Net: magnets
Ground Truth: cereal

What uniform is she wearing?
Neural Net: shorts
Ground Truth: girl scout

What is the table number?
Neural Net: 4
Ground Truth: 40

What are people sitting under in the back?
Neural Net: bench
Ground Truth: tent
Speech and Natural Language Processing

- Different approaches to:
  - Modeling sequences of tokens
- Language Modeling: $P(x_t | x_{t-1})$

- Applications:
  - Machine Translation
  - Document Classification
    - Sentiment
    - Document types
  - Question Answering
  - etc
Speech Recognition

TIMIT Speech Recognition

- Traditional
- Deep Learning

Error Rate vs. Year (1998-2014)

Graph credit: Matt Zeiler, Clarifai
Machine Translation

Google Neural Machine Translation (in production)
Machine Translation

Google Neural Machine Translation (in production)
Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.

**Question:** Which NFL team represented the NFC at Super Bowl 48?

**Answer:** Seattle Seahawks
Pipeline Approach for Question Answering

• Feature engineering
• Classifying phrases

Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.

Which NFL team represented the NFC at Super Bowl 48?

$$f_1, f_2, \ldots, f_n$$

words, types, frequencies
dependency relations
Pipeline Approach Results

- **Dataset: Stanford Question Answering Dataset (SQuAD) [Rajpurkar et al 2016]:**
  - 100k Wikipedia documents with question

- **Accuracy:** percentage of correctly predicted phrases
Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.

Which NFL team represented the NFC at Super Bowl 48?
The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos.

Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.

Which NFL team represented the NFC at Super Bowl 48?
### Test Set Leaderboard

Since the release of our dataset (and paper), the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Test EM</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BIDAF (ensemble)</td>
<td>73.3</td>
<td>81.1</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for AI &amp; University of Washington</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Seo et al. ‘16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Dynamic Coattention Networks (ensemble)</td>
<td>71.6</td>
<td>80.4</td>
</tr>
<tr>
<td></td>
<td>Salesforce Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Xiong &amp; Zhong et al. ‘16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>r-net (ensemble)</td>
<td>72.1</td>
<td>79.7</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>r-net (single model)</td>
<td>68.4</td>
<td>77.5</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BIDAF (single model)</td>
<td>68.0</td>
<td>77.3</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for AI &amp; University of Washington</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Seo et al. ‘16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Multi-Perspective Matching (ensemble)</td>
<td>68.2</td>
<td>77.2</td>
</tr>
<tr>
<td></td>
<td>IBM Research</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### March 8, 2021

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FPNet (ensemble)</td>
<td>90.871</td>
<td>93.183</td>
</tr>
<tr>
<td></td>
<td>Ant Service Intelligence Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IE-Net (ensemble)</td>
<td>90.758</td>
<td>93.044</td>
</tr>
<tr>
<td></td>
<td>RICOH_SRCB_DML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SA-Net on Albert (ensemble)</td>
<td>90.724</td>
<td>93.011</td>
</tr>
<tr>
<td></td>
<td>QIANXIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SA-Net-V2 (ensemble)</td>
<td>90.679</td>
<td>92.948</td>
</tr>
<tr>
<td></td>
<td>QIANXIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Retro-Reader (ensemble)</td>
<td>90.578</td>
<td>92.978</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ELECTRA+ALBERT+EntitySpanFocus (ensemble)</td>
<td>90.442</td>
<td>92.839</td>
</tr>
<tr>
<td></td>
<td>SRCB_DML</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Jan 1, 2017

March 8, 2021
Speech Recognition

TIMIT Speech Recognition

- Traditional
- Deep Learning

Error Rate

15 17 19 21 23 25 27 29 31


graph credit Matt Zeiler, Clarifai