Discriminative Language Models

Prof. Sameer Singh

CS 295: STATISTICAL NLP
WINTER 2017

January 26, 2017

Based on slides from Noah Smith, Richard Socher, and everyone else they copied from.
Language Models

Probability of a Sentence

- Is a given sentence something you would expect to see?
- Syntactically (grammar) and Semantically (meaning)

\[
P(\text{"I love food"}) = \frac{P(\text{"I"} \mid \text{"<s>"}) \cdot P(\text{"love"} \mid \text{"<s> I"}) \cdot P(\text{"food"} \mid \text{"<s> I love"})}{P(\text{"<s> I love"})}
\]

Probability of the Next Word

- Predict what comes next for a given sequence of words.
- Think of it as V-way classification
Outline

- Discriminative Language Models
- Feed-forward Neural Networks
- Recurrent Neural Networks
- Upcoming..
Outline

- Discriminative Language Models
- Feed-forward Neural Networks
- Recurrent Neural Networks
- Upcoming..
Logistic Regression Model

\[ p(w_i|\mathbf{w}_{i-1}) \prod_{i=1}^{n} p(w_i|\mathbf{w}_{i-1}) = p(\mathbf{w}_i|\mathbf{w}_{i-1}) \]

\[ = \frac{e^{\mathbf{\theta}_w \cdot \phi(\mathbf{w}_{i-1})}}{\sum_{\mathbf{w}_i} e^{\mathbf{\theta}_w \cdot \phi(\mathbf{w}_{i-1})}} \]

\[ \text{softmax} \]
N-Grams as Logistic Reg.

\[
p(w_i | w_{i-1}) = \frac{\# \text{ "} w_{i-1} w_i \text{"}}{\sum_{w_{i-1}} \# \text{ "} w_{i-1} w \text{"}} = \frac{e^{\Phi(w_{i-1}) \Theta_{w_{i-1}} \Phi(w_i)}}{\sum_{w_{i-1}} e^{\Phi(w_{i-1}) \Theta_{w_{i-1}} \Phi(w_i)}}
\]

\[
\Phi(w_i) = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}
\]

\[
\Theta(w_i) = \begin{bmatrix} 1 \end{bmatrix}
\]

\[
\Theta = \text{v x v} \quad \text{l x v}
\]

\[
\log \# \text{ "} w_i w_j \text{"}
\]
Other features...

\[ P(w_i \mid w_{i-1}) \quad \psi(w_{i-1}) \]

- \( V \)
- \( W_{i-1} \)
- \( k \)
- \( V_{i-1} \)
- \( \text{pos} \)
- \( (i-1) \)
- \( i \text{ is Capitalized} \)
- \( i \text{ is Number} \)
- \( \text{endsWith “ed”} \)
- \( \text{“ing”} \)
Outline

- Discriminative Language Models
- Feed-forward Neural Networks
- Recurrent Neural Networks
- Upcoming..
Logistic Reg. w/ Embeddings

\[ p(\text{wi}|\text{wi-1}) \]

\[ \text{softmax} \]

\[ y = \text{softmax}(\theta \cdot V_{wi-1} \cdot 1) \]

\[ \phi(w_{i-1}) = e_{wi-1} \]
Neural Networks

\[ y = \text{softmax}(W_0 \times h) \]

\[ h = f(W_1 \times x) \left\downarrow \text{sigmoid}, \sigma \right\]

\[ \kappa = \phi(w_{i-1}) = e^{w_{i-1}} \]
Activation Functions, \( f \)

- **sigmoid**
  \[ f(x) = \frac{e^x}{1 + e^x} \]

- **softmax**
  \[ f(x) = \frac{e^x}{\sum_{i} e^{x_i}} \]

- **tanh**
  \[ f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \]

And many others... ReLUs, PReLUs, ELU, step, max, and so on..
Why do they work?

\[ h = f(wx) \]
Why do they work?

\[ 0 = x_0 r(x_1, x_2) \]

\[ z = x_1 \cdot x_2 \]
\[ y = x_1 + x_2 - 2z \]
Simulated Example

\[
\begin{align*}
\min_{v,a,W,b} & \sum_{x_1 \in \{0,1\}} \sum_{x_2 \in \{0,1\}} \left( \text{xor}(x_1, x_2) - v^T \left( \begin{array}{c}
Wx + b \\
3 \\
2 \\
3 
\end{array} \right) + a \right)^2 \\
\min_{v,a,W,b} & \sum_{x_1 \in \{0,1\}} \sum_{x_2 \in \{0,1\}} \left( \text{xor}(x_1, x_2) - v^T \tanh \left( \begin{array}{c}
Wx + b \\
3 \\
2 \\
3 
\end{array} \right) + a \right)^2
\end{align*}
\]

https://github.com/clab/cnn/blob/master/examples/xor.cc
Simple Feedforward NN LM

Bigram Model

\[ p(w_i | w_{i-1}) \]

\[ y = \text{softmax}(W_0 x h) \]

\[ h = \sigma(W_1 x h') \]

\[ \phi(w_{i-1}) = e_{w_{i-1}} \]
Simple Feedforward NN LM

N-gram Model

\[ y \rightarrow \text{Softmax} \]

\[ h \leftarrow \text{sigmoid/tanh} \]

\[ W_0 \]

\[ W_1 \]

\[ W_n \]

\[ e_{w_{i-1}} \]

\[ e_{w_{i-2}} \]

\[ e_{w_{i-n}} \]

\[ \text{...} \]
Deep Feedforward NN LM

Bengio et al. 2003
Outline

Discriminative Language Models

Feed-forward Neural Networks

Recurrent Neural Networks

Upcoming..
Sequence View of Simple NNs
Recurrent Neural Networks

\[ y_t = \text{softmax}(W_0 \cdot h_t) \]

\[ h_t = \tanh(W_1 \cdot x + W'_1 \cdot h_{t-1}) \]
Example: “I love food”

\[
y_3 = s.m(W_0 h_3) \\
h_3 = \tanh(W_1 x_3 + W'_1 h_2) \\
h_2 = \tanh(W_1 x_2 + W'_1 h_1) \\
h_1 = \tanh(W_1 x_1 + W'_1 h_0) \\
\text{fix}
\]
Power of RNNs: Characters!

http://karpathy.github.io/2015/05/21/rnn-effectiveness/
Char-RNNs: Shakespeare!

PANDARUS:
Alas, I think he shall be come approached and the day
When little sran would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.
Char-RNNs: Wikipedia!

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the sloop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post. Many governments recognize the military housing of the [[Civil Liberalization and Infantry Rsolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS)[http://www.humah.yahoo.com/guardian.cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule, was starting to signing a major tripad of aid exile.]]
Char-RNNs: Linux Code!

```c
/*
 * If this error is set, we will need anything right after that BSD.
 */

static void action_new_function(struct s_stat_info *wb)
{
    unsigned long flags;
    int lel_idx_bit = e->edd, *sys & ~((unsigned long) FIRST_COMPAT);
    buf[0] = 0xFFFFFFFF & (bit << 4);
    min(inc, slist->bytes);
    printk(KERN_WARNING "Memory allocated %02x/%02x, "
        "original MLL instead\n"),
        min(min(multi_run - s->len, max) * num_data_in),
        frame_pos, sz + first_seg);
    div_u64_w(val, inb_p);
    spin_unlock(&disk->queue_lock);
    mutex_unlock(&s->sock->mutex);
    mutex_unlock(&func->mutex);
    return disassemble(info->pending_bh);
}
```
Extension: Stacking

\[ y_t = \sigma(W_0 x_t + h_2^t) \]

\[ h_2^t = \sigma(W_1 h_1^t + W_2 h_2^{t-1}) \]

\[ y = \tanh \]

\[ h_1 = \tanh \]

\[ X \]
Extension: Bidirectional RNNs

\[ y_t = \sigma \left( w_h^f h_t^f + w_h^b h_t^b \right) \]

\[ h_t^f \leftarrow h_{t-1}^f \]

\[ h_t^b \leftarrow h_{t+1}^b \]
Deep Bidirectional RNNs
Extension: GRUs

Gated Recurrent Units
Extension: GRUs

Gated Recurrent Units

\[
\begin{align*}
\hat{z}_t &= \sigma \left( W^{(z)} x_t + U^{(z)} h_{t-1} \right) \\
\hat{r}_t &= \sigma \left( W^{(r)} x_t + U^{(r)} h_{t-1} \right) \\
\hat{h}_t &= \tanh \left( W x_t + r_t \circ U h_{t-1} \right) \\
h_t &= z_t \circ \hat{h}_t + (1 - z_t) \circ \hat{h}_t
\end{align*}
\]
Estimating Parameters

Beyond the scope of the course

- Lots of tricks, heuristics, “domain knowledge”
- Lot of engineering for efficiency, e.g. GPUs
- New training algorithms being proposed every year
  - sometimes, architecture-specific
- Lots of available tools you can use!
  - Tensorflow, Torch, Keras, MxNET, etc.
Outline

- Discriminative Language Models
- Feed-forward Neural Networks
- Recurrent Neural Networks

Upcoming..
Homework 1 so far...

Public

Private
Ruslan Salakhutdinov

Professor at Carnegie Mellon University
Director of Artificial Intelligence, Apple Inc.

Learning Deep Unsupervised and Multimodal Models

Location: DBH 6011
Time: 11am - 12pm
Date: January 27, 2017

Meeting with PhD students, will post on Piazza
Upcoming...

Homework

- Homework 1 is due tonight: **January 26, 2017**
- Write-up, data, and code for Homework 2 is up
- Homework 2 is due: **February 9, 2017**

Project

- Proposal is due: **February 7, 2017** (~2 weeks)
- Only **2 pages**