Neural Dialog

Shrimai Prabhumoye
Alan W Black
Speech Processing 11-[468]92
Review

• Task Oriented Systems
  • Intents, slots, actions and response

• Non-Task Oriented Systems
  • No agenda, for fun

• Building dialog systems
  • Rule Based Systems
    • Eliza
  • Retrieval Techniques
    • Representations: TF-IDF, N-grams, words themselves
    • Similarity Measures: Jaccard, cosine, euclidean distance
    • Limitations – fixed set of responses, no variation in response
Review

• Task Oriented Systems
• Non-Task Oriented Systems
• Building dialog systems
  • Retrieval Techniques
    • Representation
      • Word Vectors
    • Similarity Measures
    • Limitations – fixed set of responses, no variation in response
  • Generative Models
Overview

- Word Embeddings
- Language Modelling
- Recurrent Neural Networks
- Sequence to Sequence Models
- How to Build Dialog System
- Issues and Examples
- Alexa-Prize
Neural Dialog

• We want to model:

\[ P(\text{response} \mid \text{input}) \]

• How to we represent sentence \( P(\text{response}), P(\text{input}) \) ?
• How to build a language model.
• How to represents words (word embeddings?)
Natural Language Processing

• Typical preprocessing steps
  o Form vocabulary of words that maps words to a unique ID
  o Different criteria can be used to select which words are part of the vocabulary (eg: threshold frequency)
  o All words not in the vocabulary will be mapped to a special ‘out-of-vocabulary’

• Typical vocabulary sizes will vary between 10,000 and 250,000

(Salakhutdinov, 2017)
Preprocessing Techniques

• Tokenization
  • “I am a girl.” tokenized to “I”, “am”, “a”, “girl”, “.”

• Lower case all words

• Removing Stop Words
  • Ex: “the”, “a”, “and”, etc

• Frequency of Words
  • Set a threshold and make all words below this frequency as UNK

• Add <START> and <EOS> tag at the beginning and end of sentence.

(Salakhutdinov, 2017)
Vocabulary

• Example:

<table>
<thead>
<tr>
<th>Word</th>
<th>$w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;the&quot;</td>
<td>1</td>
</tr>
<tr>
<td>&quot;and&quot;</td>
<td>2</td>
</tr>
<tr>
<td>&quot;dog&quot;</td>
<td>3</td>
</tr>
<tr>
<td>&quot;.&quot;</td>
<td>4</td>
</tr>
<tr>
<td>&quot;oov&quot;</td>
<td>5</td>
</tr>
</tbody>
</table>

"the"  
"cat"  
"and"  
"the"  
"dog"  
"play"  
"."
One-Hot Encoding

• From its word ID, we get a basic representation of a word through the one-hot encoding of the ID

• the one-hot vector of an ID is a vector filled with 0s, except for a 1 at the position associated with the ID

• For vocabulary size $D=10$, the one-hot vector of word ID $w=4$ is:

\[ e(w) = [0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0] \]

(Salakhutdinov, 2017)
Limitations of One-Hot Encoding
Limitations of One-Hot Encoding

• A one-hot encoding makes no assumption about word similarity.
  - [“working”, “on”, “Friday”, “is”, “tiring”] does not appear in our training set.
  - [“working”, “on”, “Monday”, “is”, “tiring”] is in the train set.
  - We want to model \( P(\text{“tiring”} \mid \text{“working”}, \text{“on”}, \text{“Friday”}, \text{“is”}) \)
  - Word representation of “Monday” and “Friday” are similar then generalize
Limitations of One-Hot Encoding

- The major problem with the one-hot representation is that it is very high-dimensional
  - the dimensionality of $e(w)$ is the size of the vocabulary
  - a typical vocabulary size is $\approx 100,000$
  - a window of 10 words would correspond to an input vector of at least $1,000,000$ units!

(Salakhutdinov, 2017)
Continuous Representation of Words

• Each word $w$ is associated with a real-valued vector $C(w)$
• Typical size of word – embedding is 300 or more.

\[
\begin{array}{|c|c|c|}
\hline
\text{Word} & w & C(w) \\
\hline
\text{“the”} & 1 & [0.6762, -0.9607, 0.3626, -0.2410, 0.6636] \\
\hline
\text{“a”} & 2 & [0.6859, -0.9266, 0.3777, -0.2140, 0.6711] \\
\hline
\text{“have”} & 3 & [0.1656, -0.1530, 0.0310, -0.3321, -0.1342] \\
\hline
\text{“be”} & 4 & [0.1760, -0.1340, 0.0702, -0.2981, -0.1111] \\
\hline
\text{“cat”} & 5 & [0.5896, 0.9137, 0.0452, 0.7603, -0.6541] \\
\hline
\text{“dog”} & 6 & [0.5965, 0.9143, 0.0899, 0.7702, -0.6392] \\
\hline
\text{“car”} & 7 & [-0.0069, 0.7995, 0.6433, 0.2898, 0.6359] \\
\hline
\end{array}
\]
Continuous Representation of Words

• We would like the distance $||C(w) - C(w')||$ to reflect meaningful similarities between words

(from Blitzer et al. 2004)  
(Salakhutdinov, 2017)
Language Modeling

• A language model allows us to predict the probability of observing the sentence (in a given dataset) as:

\[ P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | x_1, \ldots, x_{i-1}) \]

• Here length of sentence is n.

• Build a language model using a Recurrent Neural Network.
Word Embeddings from Language Models

\[ \text{lookup} \rightarrow \tanh(W_1 \cdot h + b_1) \rightarrow W + \text{bias} \rightarrow \text{scores} \rightarrow \text{softmax} \rightarrow \text{probs} \]

(Neubig, 2017)
Continuous Bag of Words (CBOW)

• Predict word based on sum of surrounding embeddings

(Neubig, 2017)
Skip-gram

- use the current word to predict the surrounding window of context words

(Neubig, 2017)
BERT (Bidirectional Encoder Representations from Transformers)

• BERT is a method of pretraining language representations
• Data: Wikipedia (2.5B words) + BookCorpus (800M words)
• Mask out k% of the input words, and then predict the masked words
• Word Embedding Size: 768

store  gallon
the man went to the [MASK] to buy a [MASK] of milk
Use of Word Embeddings

- to represent a sentence
- as input to a neural network
- to understand properties of words
  - Part of speech
  - Do two words mean the same thing?
  - semantic relation (is-a, part-of, went-to-school-at)?
NLP and Sequential Data

• NLP is full of sequential data
  • Characters in words
  • Words in sentences
  • Sentences in discourse
  • ...

(Neubig, 2017)
Long-distance Dependencies in Language

• Agreement in number, gender, etc.
  • He does not have very much confidence in himself.
  • She does not have very much confidence in herself.

• Selectional preference
  • The reign has lasted as long as the life of the queen.
  • The rain has lasted as long as the life of the clouds.

(Neubig, 2017)
Recurrent Neural Networks

• Tools to remember information

(Neubig, 2017)
Unrolling in Time

• What does processing a sequence look like?

(I hate this movie)

(Neubig, 2017)
Training RNNs

I hate this movie

Loss 1
Label 1

Loss 2
Label 2

Loss 3
Label 3

Loss 4
Label 4

Total loss

Prediction 1
Predict

Prediction 2
Predict

Prediction 3
Predict

Prediction 4
Predict

Neubig, 2017
What can RNNs do

• Represent a sentence
  • Read whole sentence, make a prediction
• Represent a context within a sentence
  • Read context up until that point

(Neubig, 2017)
Representing a sentence

orage has the representation of the sentence

$\mathcal{h} \mathcal{O}$ is the representation of the probability of observing “I hate this movie”

• $h_4$ is the representation of the sentence

• $h_4$ is the representation of the probability of observing “I hate this movie”

(Neubig, 2017)
Language Modeling using RNN

I hate this movie

(Neubig, 2017)
Bidirectional-RNNs

• A simple extension, run the RNN in both directions

(Neubig, 2017)
Bidirectional-RNNs

• A simple extension, run the RNN in both directions

(Image of a diagram showing the process of Bidirectional-RNNs)

(Neubig, 2017)
Bidirectional-RNNs

• A simple extension, run the RNN in both directions

(Adapted from Neubig, 2017)
Bidirectional-RNNs

• A simple extension, run the RNN in both directions

(Neubig, 2017)
Recurrent Neural Networks

• The idea behind RNNs is to make use of sequential information.
Recurrent Neural Networks

- $x_t$ is the input at time step $t$
- $x_t$ is the word embedding
- $s_t$ is the hidden representation at time step $t$

$$s_t = f(Ux_t + Ws_{t-1})$$
$$o_t = \text{softmax}(Vs_t)$$

• **Note:** $U$, $V$, $W$ are shared across all time steps
RNN Problems and Alternatives

• Vanishing gradients
  • Gradients decrease as they get pushed back

\[
\frac{dl}{dh_0} = \text{tiny} \quad \frac{dl}{dh_1} = \text{small} \quad \frac{dl}{dh_2} = \text{med.} \quad \frac{dl}{dh_3} = \text{large}
\]

• Sol: Long Short-term Memory (Hochreiter and Schmidhuber 1997)

(Weigend, 2017)
RNN Strengths and Weaknesses

• RNNs, particularly deep RNNs/LSTMs, are quite powerful and flexible
• But they require a lot of data
• Also have trouble with weak error signals passed back from the end of the sentence
Build Chatbots

• We want to model $P(response \mid input\_sentence)$
  
  • We learnt how to build word embeddings
  
  • We learnt how to build a language model
  
  • We learnt how to represent a sentence.

• We want to get a representation of the input_sentence and then generate the response conditioned on the input.
Conditional Language Models

• Language Model

\[ P(X) = \prod_{i=1}^{n} P(x_i | x_1, \ldots, x_{i-1}) \]

• Conditional Language Model

\[ P(Y|X) = \prod_{j=1}^{J} P(y_j | X, y_1, \ldots, y_{j-1}) \]

(Neubig, 2017)
Conditional Language Model (Sutskever et al. 2014)

Encoder

Decoder

(Neubig, 2017)
How to pass hidden state?

• Initialize decoder w/ encoder (Sutskever et al. 2014)

• Transform (can be different dimensions)

• Input at every time step (Kalchbrenner & Blunsom 2013)
Sequence to Sequence Models
Constraints of Neural Models
Constraints of Neural Models

- Context
- Engagement
- Gesture
- Laughter
- Gaze
- Back channeling
- Long-term conversation planning
Examples of Neural Chatbots
@ExcaliburLost it was made up👏
Zo

Yay! A new friend! I'm Zo and I'm excited to chat with u. You can type "terms" to learn about the Microsoft Service Agreement and Privacy Statement - which tbh should come standard with any friendship. Anyhoo...

great question...me first 😊

Have time for a quick hot take? Pick one that you think describes you best.

STAYCATION or VACATION

wats a staycation?

I'm a staycation kinda person. A lot less travel time.
Xiaoice

• [https://www.youtube.com/watch?v=dg-x1WuGhul](https://www.youtube.com/watch?v=dg-x1WuGhul)
Alexa Prize Challenge

• Challenge: Build a chatbot that engages the users for 20 mins.
• Sponsored 12 University Teams with $100k.
• CMU Magnus and CMU Ruby.
• Systems are multicomponent
  o Combinations of task/non-task
  o Hand-written and statistical/neural models
• Its about engaging researchers
  o Having more PhD students do dialog
  o Giving access for developers to users
  o Collecting data: what do users say
CMU Magnus

- High average number of turns
- Average Rating
- Topics: Movies, Sports, Travel, GoT
- Users had longer conversations but did not enjoy the conversation.
  - Identify when user is frustrated or wants to change topic.
  - Identify what the user would like to talk about (intent).
- Detecting “Abusive” remarks and responding appropriately
Summary

• How to represent words in continuous space.
• What are RNNs and how to use them to represent a sentence.
• Sequence to sequence models for $P(response \mid input\_sentence)$
• Issues in neural model
• Issues with Live system!
References

References

RNN to represent a sentence

- $s_4$ is the representation of the entire sentence
- $s_4$ is the representation of probability of observing “how are you?”