Lecture 8: RNNs

Alan Ritter

(many slides from Greg Durrett)
Administrivia

- Reading: RNNs
  - Goldberg 10, 11
  - Jurafsky and Martin, Chapter 9

- Homework 3 Due Next week on Monday

- Guest lecture next week on Wednesday

- Midterm is next week on Friday
  - Will cover everything up to this Friday

- Practice Questions:
  - https://docs.google.com/document/d/1eidU29ni8ZeTlBcrTyQ67duurxU74vk7I16fUok_X8s/edit?usp=sharing
Recall: Training Tips
Parameter initialization is critical to get good gradients, some useful heuristics (e.g., Xavier initializer)
Recall: Training Tips

- Parameter initialization is critical to get good gradients, some useful heuristics (e.g., Xavier initializer)

- Dropout is an effective regularizer
Recall: Training Tips

- Parameter initialization is critical to get good gradients, some useful heuristics (e.g., Xavier initializer)

- Dropout is an effective regularizer

- Think about your optimizer: Adam or tuned SGD work well
Recall: Word Vectors

The president said that the downturn was over.

<table>
<thead>
<tr>
<th>Word</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>president</td>
<td>the __ of</td>
</tr>
<tr>
<td>governor</td>
<td>the __ said of</td>
</tr>
<tr>
<td>governor</td>
<td>the __ appointed</td>
</tr>
<tr>
<td>said</td>
<td>sources __</td>
</tr>
<tr>
<td>said</td>
<td>president __ that</td>
</tr>
<tr>
<td>reported</td>
<td>sources __</td>
</tr>
</tbody>
</table>

[Finch and Chater 92, Shuetze 93, many others]
Recall: Word Vectors

[Image of a diagram showing word vectors and a sentence: the president said that the downturn was over.]

| president | the ___ of |
| president | the ___ said |
| governor  | the ___ of |
| governor  | the ___ appointed |
| said      | sources ___ |
| said      | president ___ that |
| reported  | sources ___ |

[Finch and Chater 92, Shuetze 93, many others]
Recall: Continuous Bag-of-Words

- Predict word from context

Mikolov et al. (2013)
Recall: Continuous Bag-of-Words

- Predict word from context

```
the: dog bit the man
```

Mikolov et al. (2013)

- `dog`
- `the`
Recall: Continuous Bag-of-Words

- Predict word from context

\[ \text{the} + \text{dog} \rightarrow \text{sum, size } d \]

\[ \text{the dog bit the man} \]

Mikolov et al. (2013)
Recall: Continuous Bag-of-Words

- Predict word from context

```
the: dog bit the: man
```

Mikolov et al. (2013)

Diagram:

1. `the` and `dog` as input
2. Summing their embeddings (size `d`)
3. Multiplying the result by `W`
4. Applying softmax
Recall: Continuous Bag-of-Words

- Predict word from context

\[ \text{the: dog bit the man} \]

Mikolov et al. (2013)
Recall: Continuous Bag-of-Words

- Predict word from context

\[ P(w|w_{-1}, w_{+1}) \]

Matrix factorization approaches useful for learning vectors from really large data

Mikolov et al. (2013)
Using Word Embeddings
Using Word Embeddings

- Approach 1: learn embeddings directly from data in your neural model, no pretraining
  - Often works pretty well
Using Word Embeddings

- Approach 1: learn embeddings directly from data in your neural model, no pretraining
  - Often works pretty well
- Approach 2: pretrain using GloVe, keep fixed
  - Faster because no need to update these parameters
  - Need to make sure GloVe vocabulary contains all the words you need
Using Word Embeddings

- Approach 1: learn embeddings directly from data in your neural model, no pretraining
  - Often works pretty well

- Approach 2: pretrain using GloVe, keep fixed
  - Faster because no need to update these parameters
  - Need to make sure GloVe vocabulary contains all the words you need

- Approach 3: initialize using GloVe, fine-tune
  - Not as commonly used anymore
Compositional Semantics
Compositional Semantics

- What if we want embedding representations for whole sentences?
What if we want embedding representations for whole sentences?

Skip-\textit{thought} vectors (Kiros et al., 2015), similar to skip-gram generalized to a sentence level (more later)
What if we want embedding representations for whole sentences?

Skip-thought vectors (Kiros et al., 2015), similar to skip-gram generalized to a sentence level (more later)

Is there a way we can compose vectors to make sentence representations? Summing? RNNs?
This Lecture

- Recurrent neural networks
- Vanishing gradient problem
- LSTMs / GRUs
- Applications / visualizations
RNN Basics
RNN Motivation

- Feedforward NNs can’t handle variable length input: each position in the feature vector has fixed semantics

the movie was great
Feedforward NNs can’t handle variable length input: each position in the feature vector has fixed semantics.
Feedforward NNs can’t handle variable length input: each position in the feature vector has fixed semantics.

- The movie was great
- That was great!
RNN Motivation

- Feedforward NNs can’t handle variable length input: each position in the feature vector has fixed semantics.
  
  \[
  \begin{array}{c}
  \text{the} \\
  \text{movie} \\
  \text{was} \\
  \text{great}
  \end{array}
\]

  \[
  \begin{array}{c}
  \text{that} \\
  \text{was} \\
  \text{great} \\
  \text{!}
  \end{array}
\]

- These don’t look related (great is in two different orthogonal subspaces)
RNN Motivation

- Feedforward NNs can’t handle variable length input: each position in the feature vector has fixed semantics

```
the movie was great
```

```
that was great!
```

- These don’t look related (\textit{great} is in two different orthogonal subspaces)

- Instead, we need to:
RNN Motivation

- Feedforward NNs can’t handle variable length input: each position in the feature vector has fixed semantics

  ![Diagram showing fixed semantics in feedforward NNs](image)

- Instead, we need to:
  1) Process each word in a uniform way

- These don’t look related (*great* is in two different orthogonal subspaces)

- Instead, we need to:
  1) Process each word in a uniform way
Feedforward NNs can’t handle variable length input: each position in the feature vector has fixed semantics

These don’t look related (*great* is in two different orthogonal subspaces)

Instead, we need to:

1) Process each word in a uniform way
2) ...while still exploiting the context that that token occurs in
RNN Abstraction

- Cell that takes some input $x$, has some hidden state $h$, and updates that hidden state and produces output $y$ (all vector-valued)
RNN Abstraction

- Cell that takes some input $x$, has some hidden state $h$, and updates that hidden state and produces output $y$ (all vector-valued)
RNN Uses

- Transducer: make some prediction for each element in a sequence

output $y = \text{score for each tag, then softmax}$

the movie was great
RNN Uses

- **Transducer:** make some prediction for each element in a sequence

  \[
  \text{DT} \quad \text{NN} \quad \text{VBD} \quad \text{JJ} \quad \text{output } y = \text{score for each tag, then softmax}
  \]

  the movie was great

- **Acceptor/encoder:** encode a sequence into a fixed-sized vector and use that for some purpose

  \[
  \text{predict sentiment (matmul + softmax)}
  \]

  the movie was great
RNN Uses

- **Transducer**: make some prediction for each element in a sequence

  \[
  \text{output } y = \text{score for each tag, then softmax}
  \]

  - the movie was great

- **Acceptor/encoder**: encode a sequence into a fixed-sized vector and use that for some purpose

  \[
  \text{predict sentiment (matmul + softmax)}
  \]

  - translate

  - the movie was great
RNN Uses

- Transducer: make some prediction for each element in a sequence
  - `output y = score for each tag, then softmax`
  - `output y = score for each tag, then softmax`

- Acceptor/encoder: encode a sequence into a fixed-sized vector and use that for some purpose
  - `predict sentiment (matmul + softmax)`
  - `translate`
  - `paraphrase/compress`
Elman Networks

Elman (1990)
Elman Networks

\[ h_t = \tanh(W x_t + V h_{t-1} + b_h) \]

- Updates hidden state based on input and current hidden state

Elman (1990)
Elman Networks

Elman (1990)

- Computes output from hidden state
  \[ y_t = \tanh(Uh_t + b_y) \]
- Updates hidden state based on input and current hidden state
  \[ h_t = \tanh(Wx_t + Vh_{t-1} + b_h) \]
Elman Networks

- Computes output from hidden state
- Updates hidden state based on input and current hidden state
- Computes output from hidden state

- Long history! (invented in the late 1980s)

\[
h_t = \tanh(Wx_t + Vh_{t-1} + b_h)
\]

\[
y_t = \tanh(Uh_t + b_y)
\]

Elman (1990)
Training Elman Networks

The movie was great

predict sentiment
“Backpropagation through time”: build the network as one big computation graph, some parameters are shared
Training Elman Networks

"Backpropagation through time": build the network as one big computation graph, some parameters are shared.

RNN potentially needs to learn how to "remember" information for a long time!

it was my favorite movie of 2016, though it wasn’t without problems -> +
Training Elman Networks

- “Backpropagation through time”: build the network as one big computation graph, some parameters are shared.
- RNN potentially needs to learn how to “remember” information for a long time!

It was my favorite movie of 2016, though it wasn’t without problems -> +

- “Correct” parameter update is to do a better job of remembering the sentiment of favorite
Vanishing Gradient

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Vanishing Gradient

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Vanishing Gradient

- Gradient diminishes going through tanh; if not in [-2, 2], gradient is almost 0

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Vanishing Gradient

- Gradient diminishes going through tanh; if not in [-2, 2], gradient is almost 0

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTMs/GRUs
Gated Connections

- Designed to fix “vanishing gradient” problem using gates

\[ h_t = h_{t-1} \odot f + \text{func}(x_t) \quad h_t = \tanh(Wx_t + Vh_{t-1} + b_h) \]

Gated Elman
Gated Connections

- Designed to fix “vanishing gradient” problem using *gates*

\[
\mathbf{h}_t = \mathbf{h}_{t-1} \odot \mathbf{f} + \text{func}(\mathbf{x}_t) \quad \text{gated} \\
\mathbf{h}_t = \tanh(W\mathbf{x}_t + V\mathbf{h}_{t-1} + \mathbf{b}_h) \quad \text{Elman}
\]

- Vector-valued “forget gate” \( \mathbf{f} \) computed based on input and previous hidden state

\[
\mathbf{f} = \sigma(W^{xf}\mathbf{x}_t + W^{hf}\mathbf{h}_{t-1})
\]

- Sigmoid: elements of \( \mathbf{f} \) are in \((0, 1)\)
Gated Connections

- Designed to fix “vanishing gradient” problem using gates
  \[ h_t = h_{t-1} \odot f + \text{func}(x_t) \]
  \[ h_t = \tanh(W x_t + V h_{t-1} + b_h) \]

- Vector-valued “forget gate” \( f \) computed based on input and previous hidden state
  \[ f = \sigma(W^{xf} x_t + W^{hf} h_{t-1}) \]

- Sigmoid: elements of \( f \) are in \((0, 1)\)
Gated Connections

- Designed to fix “vanishing gradient” problem using gates

\[ h_t = h_{t-1} \odot f + \text{func}(x_t) \]

Gated

- Vector-valued “forget gate” \( f \) computed based on input and previous hidden state

\[ f = \sigma(W^{xf}x_t + W^{hf}h_{t-1}) \]

- Sigmoid: elements of \( f \) are in \((0, 1)\)

- If \( f \approx 1 \), we simply sum up a function of all inputs — gradient doesn’t vanish!
LSTMs

“Cell” $c$ in addition to hidden state $h$

$$c_t = c_{t-1} \odot f + \text{func}(x_t, h_{t-1})$$
LSTMs

- “Cell” \( c \) in addition to hidden state \( h \)
  \[
  c_t = c_{t-1} \odot f + \text{func}(x_t, h_{t-1})
  \]

- Vector-valued forget gate \( f \) depends on the \( h \) hidden state
  \[
  f = \sigma(W^{xf}x_t + W^{hf}h_{t-1})
  \]
LSTMs

- “Cell” \( c \) in addition to hidden state \( h \)
  \[
  c_t = c_{t-1} \odot f + \text{func}(x_t, h_{t-1})
  \]

- Vector-valued forget gate \( f \) depends on the \( h \) hidden state
  \[
  f = \sigma(W^{xf} x_t + W^{hf} h_{t-1})
  \]

- Basic communication flow: \( x \) -> \( c \) -> \( h \) -> output, each step of this process is gated in addition to gates from previous timesteps
LSTMs

\[ c_j = c_{j-1} \odot f + g \odot i \]
\[ f = \sigma(x_j W^f + h_{j-1} W^{hf}) \]

- \( f, i, o \) are gates that control information flow
- \( g \) reflects the main computation of the cell

Goldberg lecture notes
http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTMs

\[ c_j = c_{j-1} \odot f + g \odot i \]
\[ f = \sigma(x_j W^{xf} + h_{j-1} W^{hf}) \]
\[ g = \tanh(x_j W^{xg} + h_{j-1} W^{hg}) \]
\[ i = \sigma(x_j W^{xi} + h_{j-1} W^{hi}) \]

- \( f, i, o \) are gates that control information flow
- \( g \) reflects the main computation of the cell

Goldberg lecture notes
http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTMs

- $f, i, o$ are gates that control information flow
- $g$ reflects the main computation of the cell

\[
\begin{align*}
c_j &= c_{j-1} \odot f + g \odot i \\
f &= \sigma(x_j W^{xf} + h_{j-1} W^{hf}) \\
g &= \tanh(x_j W^{xg} + h_{j-1} W^{hg}) \\
i &= \sigma(x_j W^{xi} + h_{j-1} W^{hi}) \\
h_j &= \tanh(c_j) \odot o \\
o &= \sigma(x_j W^{xo} + h_{j-1} W^{ho})
\end{align*}
\]

Goldberg lecture notes
http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTMs

c_j = c_{j-1} \odot f + g \odot i

f = \sigma(x_j W^{xf} + h_{j-1} W^{hf})

g = \tanh(x_j W^{xg} + h_{j-1} W^{hg})

i = \sigma(x_j W^{xi} + h_{j-1} W^{hi})

h_j = \tanh(c_j) \odot o

o = \sigma(x_j W^{xo} + h_{j-1} W^{ho})
LSTMs

Can we ignore the old value of $c$ for this timestep?

\[
\begin{align*}
  c_j &= c_{j-1} \odot f + g \odot i \\
  f &= \sigma(x_j W^{xf} + h_{j-1} W^{hf}) \\
  g &= \tanh(x_j W^{xg} + h_{j-1} W^{hg}) \\
  i &= \sigma(x_j W^{xi} + h_{j-1} W^{hi}) \\
  h_j &= \tanh(c_j) \odot o \\
  o &= \sigma(x_j W^{xo} + h_{j-1} W^{ho})
\end{align*}
\]
Can we ignore the old value of $c$ for this timestep?

Can an LSTM sum up its inputs $x$?
Can we ignore the old value of \( c \) for this timestep?

Can an LSTM sum up its inputs \( x \)?

Can we ignore a particular input \( x \)?
Can we ignore the old value of $c$ for this timestep?

Can an LSTM sum up its inputs $x$?

Can we ignore a particular input $x$?

Can we output something without changing $c$?
LSTMs

- Ignoring recurrent state entirely:
  - Lets us get feedforward layer over token

[Diagram of LSTM cell]

Goldberg lecture notes
http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTMs

- Ignoring recurrent state entirely:
  - Lets us get feedforward layer over token
- Ignoring input:
  - Lets us discard stopwords
LSTMs

- Ignoring recurrent state entirely:
  - Lets us get feedforward layer over token

- Ignoring input:
  - Lets us discard stopwords

- Summing inputs:
  - Lets us compute a bag-of-words representation

Goldberg lecture notes

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTMs

- Gradient still diminishes, but in a controlled way and generally by less — usually initialize forget gate = 1 to remember everything to start

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTM: more complex and slower, may work a bit better
GRUs

- LSTM: more complex and slower, may work a bit better
- GRU: faster, a bit simpler
GRUs

- LSTM: more complex and slower, may work a bit better
- GRU: faster, a bit simpler
- Two gates: $z$ (forget, mixes $s$ and $h$) and $r$ (mixes $h$ and $x$)
What do RNNs produce?

- **Encoding of the sentence** — can pass this a decoder or make a classification decision about the sentence

```
the movie was great
```
What do RNNs produce?

- Encoding of each word — can pass this to another layer to make a prediction (can also pool these to get a different sentence encoding)

- Encoding of the sentence — can pass this a decoder or make a classification decision about the sentence

the movie was great
What do RNNs produce?

- **Encoding of the sentence** — can pass this a decoder or make a classification decision about the sentence.

- **Encoding of each word** — can pass this to another layer to make a prediction (can also pool these to get a different sentence encoding).

- RNN can be viewed as a transformation of a sequence of vectors into a sequence of context-dependent vectors.

```
the movie was great
```
the movie was great
the movie was great
the movie was great
Sentence classification based on concatenation of both final outputs.
Sentence classification based on concatenation of both final outputs

Token classification based on concatenation of both directions’ token representations
Training RNNs

the movie was great
Training RNNs

the movie was great

\[ P(y|x) \]
Training RNNs

- Loss = negative log likelihood of probability of gold label (or use SVM or other loss)
Training RNNs

- Loss = negative log likelihood of probability of gold label (or use SVM or other loss)
- Backpropagate through entire network
Training RNNs

- Loss = negative log likelihood of probability of gold label (or use SVM or other loss)
- Backpropagate through entire network
- Example: sentiment analysis
Training RNNs

the movie was great
Training RNNs

\[ P(t_i | x) \]

the movie was great
Training RNNs

- Loss = negative log likelihood of probability of gold predictions, summed over the tags

\[ P(t_i|x) \]
Training RNNs

\[ P(t_i | x) \]

- Loss = negative log likelihood of probability of gold predictions, summed over the tags
- Loss terms filter back through network
Training RNNs

- Loss = negative log likelihood of probability of gold predictions, summed over the tags
- Loss terms filter back through network

\[ P(t_i|\mathbf{x}) \]

the movie was great
Training RNNs

- Loss = negative log likelihood of probability of gold predictions, summed over the tags
- Loss terms filter back through network
- Example: language modeling (predict next word given context)
Applications
What can LSTMs model?
What can LSTMs model?

- Sentiment
  - Encode one sentence, predict
- Language models
  - Move left-to-right, per-token prediction
What can LSTMs model?

- Sentiment
  - Encode one sentence, predict
- Language models
  - Move left-to-right, per-token prediction
- Translation
What can LSTMs model?

- Sentiment
  - Encode one sentence, predict
- Language models
  - Move left-to-right, per-token prediction
- Translation
  - Encode sentence + then decode, use token predictions for attention weights (later in the course)
Visualizing LSTMs

Karpathy et al. (2015)
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code

Karpathy et al. (2015)
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells (components of \(c\)) to understand them

Karpathy et al. (2015)
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code

- Visualize activations of specific cells (components of $c$) to understand them

---

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

Karpathy et al. (2015)
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells (components of $c$) to understand them
- Counter: know when to generate 

---

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action—the one Kutuzov and the general mass of the army demanded—namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all—carried on by vis inertiae—pressed forward into boats and into the ice-covered water and did not, surrender.

Karpathy et al. (2015)
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells to see what they track

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells to see what they track
- Binary switch: tells us if we’re in a quote or not

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Karpathy et al. (2015)
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells to see what they track

```c
#ifndef CONFIG_AUDITSYSSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
        {
            if (mask[i] & classes[class][i])
                return 0;
        }
    }
    return 1;
}
#endif
```
Visualizing LSTMs

- Train *character* LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells to see what they track
- Stack: activation based on indentation

```
#define CONFIG_AUDITSYS CALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
            if (mask[1] & classes[class][1])
                return 0;
    }
    return 1;
}
```
Visualizing LSTMs

- Train character LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells to see what they track

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **buffp, size_t *remain, size_t len)
{
  char *str;
  if (!*buffp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length. */
```
Visualizing LSTMs

- Train character LSTM language model (predict next character based on history) over two datasets: War and Peace and Linux kernel source code
- Visualize activations of specific cells to see what they track
- Uninterpretable: probably doing double-duty, or only makes sense in the context of another activation

```c
/* Unpack a filter field's string representation from user-space
 * bufer. */
char *audit_unpack_string(void **buftp, size_t *remain, size_t len)
{
    char *str;
    if (!buftp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length. */

    return;
}
```

Karpathy et al. (2015)
What can LSTMs model?

- Sentiment
  - Encode one sentence, predict
- Language models
  - Move left-to-right, per-token prediction
- Translation
  - Encode sentence + then decode, use token predictions for attention weights (next lecture)
What can LSTMs model?

- Sentiment
  - Encode one sentence, predict
- Language models
  - Move left-to-right, per-token prediction
- Translation
  - Encode sentence + then decode, use token predictions for attention weights (next lecture)
- Textual entailment
What can LSTMs model?

- Sentiment
  - Encode one sentence, predict
- Language models
  - Move left-to-right, per-token prediction
- Translation
  - Encode sentence + then decode, use token predictions for attention weights (next lecture)
- Textual entailment
  - Encode two sentences, predict
<table>
<thead>
<tr>
<th><strong>Premise</strong></th>
<th><strong>Hypothesis</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td>A boy is outside</td>
</tr>
</tbody>
</table>
# Natural Language Inference

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td>entails</td>
</tr>
<tr>
<td></td>
<td>A boy is outside</td>
</tr>
</tbody>
</table>

"entails" indicates that if the premise is true, then the hypothesis must also be true.
<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td>entails</td>
</tr>
<tr>
<td>A man inspects the uniform of a figure</td>
<td></td>
</tr>
</tbody>
</table>
## Natural Language Inference

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td>A boy is outside</td>
<td></td>
</tr>
<tr>
<td>$entails$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A man inspects the uniform of a figure</td>
<td>The man is sleeping</td>
<td>$contradicts$</td>
</tr>
</tbody>
</table>
# Natural Language Inference

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td>entails A boy is outside</td>
</tr>
<tr>
<td>A man inspects the uniform of a figure</td>
<td>contradicts The man is sleeping</td>
</tr>
<tr>
<td>An older and younger man smiling</td>
<td>Two men are smiling and laughing at cats playing</td>
</tr>
</tbody>
</table>
### Natural Language Inference

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td>entails</td>
</tr>
<tr>
<td>A man inspects the uniform of a figure</td>
<td>contradicts</td>
</tr>
<tr>
<td>An older and younger man smiling</td>
<td>neutral</td>
</tr>
<tr>
<td></td>
<td>A boy is outside</td>
</tr>
<tr>
<td></td>
<td>The man is sleeping</td>
</tr>
<tr>
<td></td>
<td>Two men are smiling and laughing at cats playing</td>
</tr>
</tbody>
</table>
# Natural Language Inference

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td>entails</td>
</tr>
<tr>
<td>A boy is outside</td>
<td></td>
</tr>
<tr>
<td>A man inspects the uniform of a figure</td>
<td>contradicts</td>
</tr>
<tr>
<td>The man is sleeping</td>
<td></td>
</tr>
<tr>
<td>An older and younger man smiling</td>
<td>neutral</td>
</tr>
<tr>
<td>Two men are smiling and laughing at cats playing</td>
<td></td>
</tr>
</tbody>
</table>

- Long history of this task: “Recognizing Textual Entailment” challenge in 2006 (Dagan, Glickman, Magnini)
### Natural Language Inference

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy plays in the snow</td>
<td><strong>entails</strong></td>
</tr>
<tr>
<td>A man inspects the uniform of a figure</td>
<td><strong>contradicts</strong></td>
</tr>
<tr>
<td>An older and younger man smiling</td>
<td><strong>neutral</strong></td>
</tr>
<tr>
<td><strong>A boy is outside</strong></td>
<td></td>
</tr>
<tr>
<td>The man is sleeping</td>
<td></td>
</tr>
<tr>
<td>Two men are smiling and laughing at cats playing</td>
<td></td>
</tr>
</tbody>
</table>

- Long history of this task: “Recognizing Textual Entailment” challenge in 2006 (Dagan, Glickman, Magnini)
- Early datasets: small (hundreds of pairs), very ambitious (lots of world knowledge, temporal reasoning, etc.)
SNLI Dataset

- Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
- >500,000 sentence pairs

Bowman et al. (2015)
SNLI Dataset

- Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
- >500,000 sentence pairs
- Encode each sentence and process

Bowman et al. (2015)
SNLI Dataset

- Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
- >500,000 sentence pairs
- Encode each sentence and process
- 100D LSTM: 78% accuracy

Bowman et al. (2015)
SNLI Dataset

- Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
- >500,000 sentence pairs
- Encode each sentence and process
  - 100D LSTM: 78% accuracy
  - 300D LSTM: 80% accuracy
(Bowman et al., 2016)
SNLI Dataset

- Show people captions for (unseen) images and solicit entailed / neural / contradictory statements
- >500,000 sentence pairs
- Encode each sentence and process

100D LSTM: 78% accuracy
300D LSTM: 80% accuracy
(Bowman et al., 2016)
300D BiLSTM: 83% accuracy
(Liu et al., 2016)

Bowman et al. (2015)
Show people captions for (unseen) images and solicit entailed/neural/contradictory statements

- >500,000 sentence pairs

- Encode each sentence and process

100D LSTM: 78% accuracy

300D LSTM: 80% accuracy  
(Bowman et al., 2016)

300D BiLSTM: 83% accuracy  
(Liu et al., 2016)

Later: better models for this
Takeaways

- RNNs can transduce inputs (produce one output for each input) or compress the whole input into a vector

- Useful for a range of tasks with sequential input: sentiment analysis, language modeling, natural language inference, machine translation

- Next time: CNNs and neural CRFs