Natural Language Processing with Deep Learning
CS224N/Ling284

Christopher Manning
Lecture 5: Language Models and Recurrent Neural Networks
(oh, and finish neural dependency parsing 😊)
Lecture Plan

1. Neural dependency parsing (20 mins)
2. A bit more about neural networks (15 mins)
3. Language modeling + RNNs (45 mins)
   • A new NLP task: Language Modeling
   • A new family of neural networks: Recurrent Neural Networks (RNNs)

These are two of the most important concepts for the rest of the class!

Reminders:

You should have handed in Assignment 2 by today
In Assignment 3, out today, you build a neural dependency parser using PyTorch
1. How do we gain from a neural dependency parser?

Indicator Features Revisited

- Problem #1: sparse
- Problem #2: incomplete
- Problem #3: expensive computation

More than 95% of parsing time is consumed by feature computation

Neural Approach:

learn a dense and compact feature representation
A neural dependency parser [Chen and Manning 2014]

- Results on English parsing to Stanford Dependencies:
  - Unlabeled attachment score (UAS) = head
  - Labeled attachment score (LAS) = head and label

<table>
<thead>
<tr>
<th>Parser</th>
<th>UAS</th>
<th>LAS</th>
<th>sent. / s</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaltParser</td>
<td>89.8</td>
<td>87.2</td>
<td>469</td>
</tr>
<tr>
<td>MSTParser</td>
<td>91.4</td>
<td>88.1</td>
<td>10</td>
</tr>
<tr>
<td>TurboParser</td>
<td>92.3</td>
<td>89.6</td>
<td>8</td>
</tr>
<tr>
<td>C &amp; M 2014</td>
<td>92.0</td>
<td>89.7</td>
<td>654</td>
</tr>
</tbody>
</table>
First win: Distributed Representations

- We represent each word as a $d$-dimensional dense vector (i.e., word embedding)
  - Similar words are expected to have close vectors.

- Meanwhile, part-of-speech tags (POS) and dependency labels are also represented as $d$-dimensional vectors.
  - The smaller discrete sets also exhibit many semantical similarities.

\[
\begin{align*}
\text{NNS (plural noun)} & \text{ should be close to } \text{NN (singular noun)} \\
\text{nummod (numerical modifier)} & \text{ should be close to } \text{amod (adjective modifier)} \\
\end{align*}
\]
Extracting Tokens & vector representations from configuration

- We extract a set of tokens based on the stack / buffer positions:

```
<table>
<thead>
<tr>
<th>Stack</th>
<th>Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT</td>
<td>control_NN</td>
</tr>
<tr>
<td>has_VBZ</td>
<td></td>
</tr>
<tr>
<td>good_JJ</td>
<td></td>
</tr>
<tr>
<td>nsubj</td>
<td></td>
</tr>
<tr>
<td>He_PRP</td>
<td></td>
</tr>
</tbody>
</table>
```

A concatenation of the vector representation of all these is the neural representation of a configuration.
Second win: Deep Learning classifiers are non-linear classifiers

- A softmax classifier assigns classes $y \in C$ based on inputs $x \in \mathbb{R}^d$ via the probability:
  \[ p(y|x) = \frac{\exp(W_y \cdot x)}{\sum_{c=1}^{C} \exp(W_c \cdot x)} \]
  a.k.a. “cross entropy loss”

- We train the weight matrix $W \in \mathbb{R}^{C \times d}$ to minimize the neg. log loss: $\sum_i - \log p(y_i|x_i)$

- Traditional ML classifiers (including Naïve Bayes, SVMs, logistic regression and softmax classifier) are not very powerful classifiers: they only give linear decision boundaries

This can be quite limiting

→ Unhelpful when a problem is complex

Wouldn’t it be cool to get these correct?
Neural Networks are more powerful

- Neural networks can learn much more complex functions with nonlinear decision boundaries!
  - Non-linear in the original space, linear for the softmax at the top of the neural network

Visualizations with ConvNetJS by Andrej Karpathy!

http://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html
Simple feed-forward neural network multi-class classifier

Output layer $y$
$$y = \text{softmax}(Uh + b_2)$$

Hidden layer $h$
$$h = \text{ReLU}(Wx + b_1)$$

Input layer $x$
- $x$ is result of lookup
- $x_{(i,\ldots,i+d)} = Le$
- lookup + concat

Log loss (cross-entropy error) will be back-propagated to the embeddings

The hidden layer re-represents the input — it moves inputs around in an intermediate layer vector space—so it can be easily classified with a (linear) softmax

Softmax probabilities

ReLU = Rectified Linear Unit
$$\text{rect}(z) = \max(z, 0)$$
Neural Dependency Parser Model Architecture

**Input layer**

$x$ lookup + concat

**Hidden layer**

$h = \text{ReLU}(Wx + b_1)$

**Output layer**

$y = \text{softmax}(Uh + b_2)$

Softmax probabilities

{Shift, Left-Arc, Right-Arc}
Chen and Manning (2014) showed that neural networks can accurately determine the structure of sentences, supporting meaning interpretation.

It was the first simple, successful neural dependency parser.

The dense representations (and non-linear classifier) let it outperform other greedy parsers in both accuracy and speed.
Further developments in transition-based neural dependency parsing

This work was further developed and improved by others, including in particular at Google

- Bigger, deeper networks with better tuned hyperparameters
- Beam search
- Global, conditional random field (CRF)-style inference over the decision sequence

Leading to SyntaxNet and the Parsey McParseFace model (2016):
“The World’s Most Accurate Parser”

https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

<table>
<thead>
<tr>
<th>Method</th>
<th>UAS</th>
<th>LAS (PTB WSJ SD 3.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen &amp; Manning 2014</td>
<td>92.0</td>
<td>89.7</td>
</tr>
<tr>
<td>Weiss et al. 2015</td>
<td>93.99</td>
<td>92.05</td>
</tr>
<tr>
<td>Andor et al. 2016</td>
<td>94.61</td>
<td>92.79</td>
</tr>
</tbody>
</table>
Graph-based dependency parsers

• Compute a score for every possible dependency (choice of head) for each word
  • Doing this well requires more than just knowing the two words
  • We need good “contextual” representations of each word token, which we will develop in the coming lectures
• Repeat the same process for each other word; find the best parse (MST algorithm)

ROOT          The            big            cat          sat
0.5            0.3            2.0            0.8

e.g., picking the head for “big”
Graph-based dependency parsers

• Compute a score for every possible dependency (choice of head) for each word
  • Doing this well requires more than just knowing the two words
  • We need good “contextual” representations of each word token, which we will develop in the coming lectures
• Repeat the same process for each other word; find the best parse (MST algorithm)

ROOT          The            big            cat          sat

0.5
0.3
2.0
0.8

e.g., picking the head for “big”
A Neural graph-based dependency parser
[Dozat and Manning 2017; Dozat, Qi, and Manning 2017]

• This paper revived interest in graph-based dependency parsing in a neural world
  • Designed a biaffine scoring model for neural dependency parsing
    • Also crucially uses a neural sequence model, something we discuss next week
• Really great results!
  • But slower than the simple neural transition-based parsers
    • There are $n^2$ possible dependencies in a sentence of length $n$

<table>
<thead>
<tr>
<th>Method</th>
<th>UAS</th>
<th>LAS (PTB WSJ SD 3.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen &amp; Manning 2014</td>
<td>92.0</td>
<td>89.7</td>
</tr>
<tr>
<td>Weiss et al. 2015</td>
<td>93.99</td>
<td>92.05</td>
</tr>
<tr>
<td>Andor et al. 2016</td>
<td>94.61</td>
<td>92.79</td>
</tr>
<tr>
<td>Dozat &amp; Manning 2017</td>
<td>95.74</td>
<td>94.08</td>
</tr>
</tbody>
</table>
2. A bit more about neural networks
We have models with many parameters! Regularization!

- A full loss function includes regularization over all parameters $\theta$, e.g., L2 regularization:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} - \log \left( \frac{e^{f_{yi}}}{\sum_{c=1}^{C} e^{f_{ci}}} \right) + \lambda \sum_{k} \theta_k^2$$

- Classic view: Regularization works to prevent overfitting when we have a lot of features (or later a very powerful/deep model, etc.)

- Now: Regularization produces models that generalize well when we have a “big” model
  - We do not care that our models overfit on the training data, even though they are hugely overfit
Dropout \footnote{Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov 2012/JMLR 2014}

Preventing Feature Co-adaptation = Good Regularization Method!

- Training time: at each instance of evaluation (in online SGD-training), randomly set 50% of the inputs to each neuron to 0
- Test time: halve the model weights (now twice as many)
- (Except usually only drop first layer inputs a little (~15%) or not at all)
- This prevents feature co-adaptation: A feature cannot only be useful in the presence of particular other features
- In a single layer: A kind of middle-ground between Naïve Bayes (where all feature weights are set independently) and logistic regression models (where weights are set in the context of all others)
- Can be thought of as a form of model bagging (i.e., like an ensemble model)
- Nowadays usually thought of as strong, feature-dependent regularizer \cite{Wager, Wang, & Liang 2013}
“Vectorization”

- E.g., looping over word vectors versus concatenating them all into one large matrix and then multiplying the softmax weights with that matrix:

```python
from numpy import random
N = 500  # number of windows to classify
d = 300  # dimensionality of each window
C = 5    # number of classes
W = random.rand(C,d)
wordvectors_list = [random.rand(d,1) for i in range(N)]
wordvectors_one_matrix = random.rand(d,N)

%timeit [W.dot(wordvectors_list[i]) for i in range(N)]
%timeit W.dot(wordvectors_one_matrix)
```

- 1000 loops, best of 3: 639 µs per loop
- 10000 loops, best of 3: 53.8 µs per loop ← Now using a single a C x N matrix
- Matrices are awesome!!! Always try to use vectors and matrices rather than for loops!
- The speed gain goes from 1 to 2 orders of magnitude with GPUs!
Non-linearities, old and new

**logistic (“sigmoid”)**
\[ f(z) = \frac{1}{1 + \exp(-z)}. \]

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

**hard tanh**
\[ \text{HardTanh}(x) = \begin{cases} -1 & \text{if } x < -1 \\ x & \text{if } -1 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases} \]

**ReLU (Rectified Linear Unit)**
\[ \text{rect}(z) = \max(z, 0) \]

**tanh** is just a rescaled and shifted sigmoid (2 × as steep, \([-1,1]\)):
\[ \tanh(z) = 2\text{logistic}(2z) - 1 \]

Both logistic and tanh are still used in various places (e.g., to get a probability), but are no longer the defaults for making deep networks.

For building a deep network, the first thing you should try is ReLU — it trains quickly and performs well due to good gradient backflow.

Leaky ReLU / Parametric ReLU

Swish [Ramachandran, Zoph & Le 2017]
Parameter Initialization

- You normally must initialize weights to small random values (i.e., not zero matrices!)
  - To avoid symmetries that prevent learning/specialization
- Initialize hidden layer biases to 0 and output (or reconstruction) biases to optimal value if weights were 0 (e.g., mean target or inverse sigmoid of mean target)
- Initialize all other weights \( \sim \) Uniform\((-r, r)\), with \( r \) chosen so numbers get neither too big or too small [later the need for this is removed with use of layer normalization]
- Xavier initialization has variance inversely proportional to fan-in \( n_{in} \) (previous layer size) and fan-out \( n_{out} \) (next layer size):

  \[
  \text{Var}(W_i) = \frac{2}{n_{in} + n_{out}}
  \]
Optimizers

• Usually, plain SGD will work just fine!
  • However, getting good results will often require hand-tuning the learning rate
    • See next slide
• For more complex nets and situations, or just to avoid worry, you often do better with one of a family of more sophisticated “adaptive” optimizers that scale the parameter adjustment by an accumulated gradient.
  • These models give differential per-parameter learning rates
    • Adagrad
    • RMSprop
    • Adam ← A fairly good, safe place to begin in many cases
    • SparseAdam
    • ...
Learning Rates

- You can just use a constant learning rate. Start around \( lr = 0.001 \)?
  - It must be order of magnitude right – try powers of 10
    - Too big: model may diverge or not converge
    - Too small: your model may not have trained by the assignment deadline
- Better results can generally be obtained by allowing learning rates to decrease as you train
  - By hand: halve the learning rate every \( k \) epochs
    - An epoch = a pass through the data (shuffled or sampled – not in same order each time)
  - By a formula: \( lr = lr_0 e^{-kt} \), for epoch \( t \)
  - There are fancier methods like cyclic learning rates (q.v.)
- Fancier optimizers still use a learning rate but it may be an initial rate that the optimizer shrinks – so you may want to start with a higher learning rate
3. Language Modeling + RNNs
Language Modeling

- **Language Modeling** is the task of predicting what word comes next

  \[ \text{the students opened their ______} \]

- More formally: given a sequence of words \( x^{(1)}, x^{(2)}, \ldots, x^{(t)} \), compute the probability distribution of the next word \( x^{(t+1)} \):

  \[
P(x^{(t+1)} \mid x^{(t)}, \ldots, x^{(1)})
  \]

  where \( x^{(t+1)} \) can be any word in the vocabulary \( V = \{ w_1, \ldots, w_{|V|} \} \)

- A system that does this is called a **Language Model**
Language Modeling

• You can also think of a Language Model as a system that assigns probability to a piece of text

• For example, if we have some text \( x^{(1)}, \ldots, x^{(T)} \), then the probability of this text (according to the Language Model) is:

\[
P(x^{(1)}, \ldots, x^{(T)}) = P(x^{(1)}) \times P(x^{(2)} \mid x^{(1)}) \times \cdots \times P(x^{(T)} \mid x^{(T-1)}, \ldots, x^{(1)}) \\
= \prod_{t=1}^{T} P(x^{(t)} \mid x^{(t-1)}, \ldots, x^{(1)})
\]

This is what our LM provides
You use Language Models every day!
You use Language Models every day!
n-gram Language Models

the students opened their ____

• **Question**: How to learn a Language Model?
• **Answer** (pre- Deep Learning): learn an *n*-gram Language Model!

• **Definition**: A *n*-gram is a chunk of *n* consecutive words.
  • unigrams: “the”, “students”, “opened”, ”their”
  • bigrams: “the students”, “students opened”, “opened their”
  • trigrams: “the students opened”, “students opened their”
  • 4-grams: “the students opened their”

• **Idea**: Collect statistics about how frequent different n-grams are and use these to predict next word.
n-gram Language Models

- First we make a Markov assumption: \( x^{(t+1)} \) depends only on the preceding \( n-1 \) words

\[
P(x^{(t+1)} | x^{(t)}, \ldots, x^{(1)}) = P(x^{(t+1)} | x^{(t)}, \ldots, x^{(t-n+2)})
\]

(assumption)

\[
P(x^{(t+1)}, x^{(t)}, \ldots, x^{(t-n+2)}) = P(x^{(t)}, \ldots, x^{(t-n+2)})
\]

(definition of conditional prob)

- **Question:** How do we get these \( n \)-gram and \( (n-1) \)-gram probabilities?
- **Answer:** By counting them in some large corpus of text!

\[
\approx \frac{\text{count}(x^{(t+1)}, x^{(t)}, \ldots, x^{(t-n+2)})}{\text{count}(x^{(t)}, \ldots, x^{(t-n+2)})}
\]

(statistical approximation)
Suppose we are learning a 4-gram Language Model.

For example, suppose that in the corpus:

- “students opened their” occurred 1000 times
- “students opened their books” occurred 400 times
  - \( P(\text{books} | \text{students opened their}) = 0.4 \)
- “students opened their exams” occurred 100 times
  - \( P(\text{exams} | \text{students opened their}) = 0.1 \)

Should we have discarded the “proctor” context?
Sparsity Problems with n-gram Language Models

Sparsity Problem 1

**Problem:** What if “students opened their w” never occurred in data? Then w has probability 0!

**Solution:** Add small $\delta$ to the count for every $w \in V$. This is called *smoothing*.

\[ P(w|\text{students opened their}) = \frac{\text{count(students opened their w)}}{\text{count(students opened their)}} \]

Sparsity Problem 2

**Problem:** What if “students opened their” never occurred in data? Then we can’t calculate probability for *any* $w$!

**Solution:** Just condition on “opened their” instead. This is called *backoff*.

**Note:** Increasing $n$ makes sparsity problems worse. Typically, we can’t have $n$ bigger than 5.
Storage Problems with n-gram Language Models

**Storage**: Need to store count for all $n$-grams you saw in the corpus.

$$P(w|\text{students opened their}) = \frac{\text{count(students opened their } w)}{\text{count(students opened their)}}$$

Increasing $n$ or increasing corpus increases model size!
n-gram Language Models in practice

- You can build a simple trigram Language Model over a 1.7 million word corpus (Reuters) in a few seconds on your laptop*

\textit{today the ________}

\begin{tabular}{|l|c|}
  \hline
  company & 0.153 \\
  bank & 0.153 \\
  price & 0.077 \\
  italian & 0.039 \\
  emirate & 0.039 \\
  \ldots & \\
  \hline
\end{tabular}

\textbf{Sparsity problem:} not much granularity in the probability distribution

Otherwise, seems reasonable!

* Try for yourself: \url{https://nlpforhackers.io/language-models/}
Generating text with a n-gram Language Model

You can also use a Language Model to generate text

today the_____
condition on this
get probability distribution

<table>
<thead>
<tr>
<th></th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>company</td>
<td>0.153</td>
</tr>
<tr>
<td>bank</td>
<td>0.153</td>
</tr>
<tr>
<td>price</td>
<td>0.077</td>
</tr>
<tr>
<td>italian</td>
<td>0.039</td>
</tr>
<tr>
<td>emirate</td>
<td>0.039</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Generating text with a n-gram Language Model

You can also use a Language Model to generate text

today the price ______

condition on this

generate probability distribution

<table>
<thead>
<tr>
<th>of</th>
<th>0.308</th>
</tr>
</thead>
<tbody>
<tr>
<td>for</td>
<td>0.050</td>
</tr>
<tr>
<td>it</td>
<td>0.046</td>
</tr>
<tr>
<td>to</td>
<td>0.046</td>
</tr>
<tr>
<td>is</td>
<td>0.031</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Generating text with a n-gram Language Model

You can also use a Language Model to generate text

today the price of ______
condition on this

get probability distribution

<table>
<thead>
<tr>
<th>the</th>
<th>0.072</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>0.043</td>
</tr>
<tr>
<td>oil</td>
<td>0.043</td>
</tr>
<tr>
<td>its</td>
<td>0.036</td>
</tr>
<tr>
<td>gold</td>
<td>0.018</td>
</tr>
</tbody>
</table>

sample
Generating text with a n-gram Language Model

You can also use a Language Model to generate text

today the price of gold per ton, while production of shoe lasts and shoe industry, the bank intervened just after it considered and rejected an imf demand to rebuild depleted european stocks, sept 30 end primary 76 cts a share.

Surprisingly grammatical!

...but incoherent. We need to consider more than three words at a time if we want to model language well.

But increasing n worsens sparsity problem, and increases model size...
How to build a *neural* Language Model?

- Recall the Language Modeling task:
  - Input: sequence of words $x^{(1)}, x^{(2)}, \ldots, x^{(t)}$
  - Output: prob dist of the next word $P(x^{(t+1)}| x^{(t)}, \ldots, x^{(1)})$

- How about a *window*-based neural model?
  - We saw this applied to Named Entity Recognition in Lecture 3:
A fixed-window neural Language Model

As the proctor started the clock, the students opened their fixed window.
A fixed-window neural Language Model

output distribution
\[ \hat{y} = \text{softmax}(U h + b_2) \in \mathbb{R}^{|V|} \]

hidden layer
\[ h = f(W e + b_1) \]

concatenated word embeddings
\[ e = [e^{(1)}; e^{(2)}; e^{(3)}; e^{(4)}] \]

words / one-hot vectors
\[ x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)} \]
A fixed-window neural Language Model


**Improvements** over $n$-gram LM:
- No sparsity problem
- Don’t need to store all observed $n$-grams

**Remaining problems:**
- Fixed window is too small
- Enlarging window enlarges $W$
- Window can never be large enough!
- $x^{(1)}$ and $x^{(2)}$ are multiplied by completely different weights in $W$. 
  No symmetry in how the inputs are processed.

We need a neural architecture that can process **any length input**
Recurrent Neural Networks (RNN)
A family of neural architectures

Core idea: Apply the same weights $W$ repeatedly.
A Simple RNN Language Model

output distribution
\[ \hat{y}^{(t)} = \text{softmax} \left( U h^{(t)} + b_2 \right) \in \mathbb{R}^{|V|} \]

hidden states
\[ h^{(t)} = \sigma \left( W_h h^{(t-1)} + W_e e^{(t)} + b_1 \right) \]
\( h^{(1)} \) is the initial hidden state

word embeddings
\[ e^{(t)} = E x^{(t)} \]

words / one-hot vectors
\[ x^{(t)} \in \mathbb{R}^{|V|} \]

Note: this input sequence could be much longer now!
RNN Language Models

RNN Advantages:
• Can process any length input
• Computation for step $t$ can (in theory) use information from many steps back
• Model size doesn’t increase for longer input context
• Same weights applied on every timestep, so there is symmetry in how inputs are processed.

RNN Disadvantages:
• Recurrent computation is slow
• In practice, difficult to access information from many steps back

More on these later in the course
Training an RNN Language Model

- Get a big corpus of text which is a sequence of words $x^{(1)}, \ldots, x^{(T)}$
- Feed into RNN-LM; compute output distribution $\hat{y}^{(t)}$ for every step $t$.
  - i.e. predict probability dist of every word, given words so far

- Loss function on step $t$ is cross-entropy between predicted probability distribution $\hat{y}^{(t)}$, and the true next word $y^{(t)}$ (one-hot for $x^{(t+1)}$):
  \[ J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = - \sum_{w \in V} y^{(t)}_w \log \hat{y}^{(t)}_w = - \log \hat{y}^{(t)}_{x^{(t+1)}} \]

- Average this to get overall loss for entire training set:
  \[ J(\theta) = \frac{1}{T} \sum_{t=1}^{T} J^{(t)}(\theta) = \frac{1}{T} \sum_{t=1}^{T} - \log \hat{y}^{(t)}_{x^{(t+1)}} \]
Training an RNN Language Model

= negative log prob of “students”

Loss

Predicted prob dists

Corpus

the

students

opened

their

exams
Training an RNN Language Model

Loss $\rightarrow J^{(1)}(\theta)$

Predicted prob dists

$h^{(0)}$ $\rightarrow h^{(1)}$ $\rightarrow h^{(2)}$ $\rightarrow h^{(3)}$ $\rightarrow h^{(4)}$

$W_h \rightarrow W_h \rightarrow W_h \rightarrow W_h \rightarrow W_h \rightarrow ...$

$W_e \rightarrow W_e \rightarrow W_e \rightarrow W_e \rightarrow W_e \rightarrow ...$

$e^{(1)} \rightarrow e^{(2)} \rightarrow e^{(3)} \rightarrow e^{(4)} \rightarrow E \rightarrow ...$

$U \rightarrow U \rightarrow U \rightarrow U \rightarrow U \rightarrow ...$

$\hat{y}^{(1)} \rightarrow \hat{y}^{(2)} \rightarrow \hat{y}^{(3)} \rightarrow \hat{y}^{(4)} \rightarrow ...$

$\Rightarrow J^{(2)}(\theta)$

- negative log prob of “opened”

Corpus $\rightarrow \begin{array}{cccc}
\text{the} & \text{students} & \text{opened} & \text{their} \\
\vdots & \vdots & \vdots & \vdots \\
x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)}
\end{array}$

Exams $\Rightarrow J^{(3)}(\theta)$

$\Rightarrow J^{(4)}(\theta)$
Training an RNN Language Model

Loss $\rightarrow J^{(1)}(\theta) \rightarrow J^{(2)}(\theta) \rightarrow J^{(3)}(\theta) \rightarrow J^{(4)}(\theta)$

Predicted prob dists $\rightarrow \hat{y}^{(1)} \rightarrow \hat{y}^{(2)} \rightarrow \hat{y}^{(3)} \rightarrow \hat{y}^{(4)}$ $\rightarrow U \rightarrow U \rightarrow U \rightarrow U$

Corpus $\rightarrow x^{(1)} \rightarrow x^{(2)} \rightarrow x^{(3)} \rightarrow x^{(4)}$$\rightarrow \text{the } \rightarrow \text{students } \rightarrow \text{opened } \rightarrow \text{their } \rightarrow \text{exams }$
Training an RNN Language Model

The students opened their exams.

$J^{(4)}(\theta)$ = negative log prob of “exams”
Training an RNN Language Model

"Teacher forcing"

\[
J(\theta) = \frac{1}{T} \sum_{t=1}^{T} J^{(t)}(\theta)
\]
Training a RNN Language Model

• However: Computing loss and gradients across entire corpus $x^{(1)}, \ldots, x^{(T)}$ is too expensive!

$$J(\theta) = \frac{1}{T} \sum_{t=1}^{T} J^{(t)}(\theta)$$

• In practice, consider $x^{(1)}, \ldots, x^{(T)}$ as a sentence (or a document)

• Recall: Stochastic Gradient Descent allows us to compute loss and gradients for small chunk of data, and update.

• Compute loss $J(\theta)$ for a sentence (actually, a batch of sentences), compute gradients and update weights. Repeat.
Backpropagation for RNNs

**Question:** What’s the derivative of $J^{(t)}(\theta)$ w.r.t. the repeated weight matrix $W_h$?

**Answer:**

$$\frac{\partial J^{(t)}}{\partial W_h} = \sum_{i=1}^{t} \frac{\partial J^{(t)}}{\partial W_h^{(i)}}$$

“The gradient w.r.t. a repeated weight is the sum of the gradient w.r.t. each time it appears”

**Why?**
Multivariable Chain Rule

- Given a multivariable function \( f(x, y) \), and two single variable functions \( x(t) \) and \( y(t) \), here's what the multivariable chain rule says:

\[
\frac{d}{dt} f(x(t), y(t)) = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt}
\]

Derivative of composition function

Source:
Backpropagation for RNNs: Proof sketch

*Given a multivariable function \( f(x, y) \), and two single variable functions \( x(t) \) and \( y(t) \), here’s what the multivariable chain rule says:

\[
\frac{d}{dt} f(x(t), y(t)) = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt}
\]

Derivative of composition function

In our example:

\[
\begin{align*}
J^{(t)}(\theta) &= \text{equals} \\
W_h|_{(1)} &= \text{equals} \\
W_h|_{(2)} &= \text{equals} \\
\vdots &= \text{equals} \\
W_h|_{(t)} &= \text{equals}
\end{align*}
\]

Apply the multivariable chain rule:

\[
\frac{\partial J^{(t)}}{\partial W_h} = \sum_{i=1}^{t} \frac{\partial J^{(t)}}{\partial W_h} \left|_{(i)} \frac{\partial W_h}{\partial W_h} \right|_{(i)} = 1
\]

Question: How do we calculate this?

Answer: Backpropagate over timesteps $i=t\ldots,0$, summing gradients as you go. This algorithm is called “backpropagation through time” [Werbos, P.G., 1988, Neural Networks 1, and others]
Generating text with a RNN Language Model

Just like a n-gram Language Model, you can use a RNN Language Model to generate text by repeated sampling. Sampled output becomes next step’s input.
Generating text with an RNN Language Model

Let’s have some fun!

• You can train an RNN-LM on any kind of text, then generate text in that style.
• RNN-LM trained on Obama speeches:

\[\text{The United States will step up to the cost of a new challenges of the American people that will share the fact that we created the problem. They were attacked and so that they have to say that all the task of the final days of war that I will not be able to get this done.}\]

Source: https://medium.com/@samim/obama-rnn-machine-generated-political-speeches-c8abd18a2ea0
Let’s have some fun!

- You can train an RNN-LM on any kind of text, then generate text in that style.
- RNN-LM trained on *Harry Potter*:

> “Sorry,” Harry shouted, panicking—“I’ll leave those brooms in London, are they?”

> “No idea,” said Nearly Headless Nick, casting low close by Cedric, carrying the last bit of treacle Charms, from Harry’s shoulder, and to answer him the common room perched upon it, four arms held a shining knob from when the spider hadn’t felt it seemed. He reached the teams too.

Generating text with an RNN Language Model

Let’s have some fun!

- You can train an RNN-LM on any kind of text, then generate text in that style.
- RNN-LM trained on recipe:

```
Title: CHOCOLATE RANCH BARBECUE
Categories: Game, Casseroles, Cookies, Cookies
Yield: 6 Servings

2 tb Parmesan cheese -- chopped
1 c  Coconut milk
3   Eggs, beaten

Place each pasta over layers of lumps. Shape mixture into the moderate oven and simmer until firm. Serve hot in bodied fresh, mustard, orange and cheese.

Combine the cheese and salt together the dough in a large skillet; add the ingredients and stir in the chocolate and pepper.
```

Source: https://gist.github.com/nylki/1efbaa36635956d35bcc
Generating text with a RNN Language Model

Let’s have some fun!

• You can train a RNN-LM on any kind of text, then generate text in that style.
• RNN-LM trained on paint color names:

This is an example of a character-level RNN-LM (predicts what character comes next)

Evaluating Language Models

• The standard evaluation metric for Language Models is perplexity.

\[
\text{perplexity} = \prod_{t=1}^{T} \left( \frac{1}{P_{\text{LM}}(x(t+1)|x(t), \ldots, x(1))} \right)^{1/T}
\]

Inverse probability of corpus, according to Language Model

Normalized by number of words

• This is equal to the exponential of the cross-entropy loss \( J(\theta) \):

\[
\prod_{t=1}^{T} \left( \frac{1}{\hat{y}_{x_{t+1}}^{(t)}} \right)^{1/T} = \exp \left( \frac{1}{T} \sum_{t=1}^{T} - \log \hat{y}_{x_{t+1}}^{(t)} \right) = \exp(J(\theta))
\]

Lower perplexity is better!
RNNs have greatly improved perplexity

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpolated Kneser-Ney 5-gram (Chelba et al., 2013)</td>
<td>67.6</td>
</tr>
<tr>
<td>RNN-1024 + MaxEnt 9-gram (Chelba et al., 2013)</td>
<td>51.3</td>
</tr>
<tr>
<td>RNN-2048 + BlackOut sampling (Ji et al., 2015)</td>
<td>68.3</td>
</tr>
<tr>
<td>Sparse Non-negative Matrix factorization (Shazeer et al., 2015)</td>
<td>52.9</td>
</tr>
<tr>
<td>LSTM-2048 (Jozefowicz et al., 2016)</td>
<td>43.7</td>
</tr>
<tr>
<td>2-layer LSTM-8192 (Jozefowicz et al., 2016)</td>
<td>30</td>
</tr>
<tr>
<td><strong>Ours small</strong> (LSTM-2048)</td>
<td>43.9</td>
</tr>
<tr>
<td><strong>Ours large</strong> (2-layer LSTM-2048)</td>
<td>39.8</td>
</tr>
</tbody>
</table>

Perplexity improves (lower is better)

Source: https://research.fb.com/building-an-efficient-neural-language-model-over-a-billion-words/
Why should we care about Language Modeling?

- Language Modeling is a **benchmark task** that helps us **measure our progress** on understanding language.

- Language Modeling is a **subcomponent** of many NLP tasks, especially those involving **generating text** or **estimating the probability of text**:
  - Predictive typing
  - Speech recognition
  - Handwriting recognition
  - Spelling/grammar correction
  - Authorship identification
  - Machine translation
  - Summarization
  - Dialogue
  - etc.
Recap

- **Language Model**: A system that predicts the next word

- **Recurrent Neural Network**: A family of neural networks that:
  - Take *sequential input of any length*
  - Apply the *same weights on each step*
  - Can optionally produce output on each step

- **Recurrent Neural Network ≠ Language Model**

- We’ve shown that RNNs are a great way to build a LM.

- But RNNs are useful for much more!
RNNs can be used for tagging

e.g., **part-of-speech tagging**, named entity recognition
RNNs can be used for sentence classification
e.g., sentiment classification

How to compute sentence encoding?

overall  I  enjoyed  the  movie  a  lot
RNNs can be used for sentence classification
e.g., sentiment classification

How to compute sentence encoding?

Basic way:
Use final hidden state

Sentence encoding

positive

equals

overall  I  enjoyed  the  movie  a  lot
RNNs can be used for sentence classification
e.g., sentiment classification

positive

Sentence encoding

How to compute sentence encoding?

Usually better:
Take element-wise max or mean of all hidden states

overall  I  enjoyed  the  movie  a  lot
RNNs can be used as an encoder module
e.g., question answering, machine translation, many other tasks!

Here the RNN acts as an encoder for the Question (the hidden states represent the Question). The encoder is part of a larger neural system.

**Question:** what nationality was Beethoven?

**Context:** Ludwig van Beethoven was a German composer and pianist. A crucial figure ...

**Answer:** German
RNN-LMs can be used to generate text
e.g., speech recognition, machine translation, summarization

This is an example of a conditional language model. We’ll see Machine Translation in much more detail later.
Terminology and a look forward

The RNN described in this lecture = **simple/vanilla/Elman** RNN

**Next lecture:** You will learn about other RNN flavors like **GRU** and **LSTM** and multi-layer RNNs

**By the end of the course:** You will understand phrases like “**stacked bidirectional LSTM with residual connections and self-attention**”