Natural Language Processing with Deep Learning
CS224N/Ling284

Tatsunori Hashimoto
Lecture 5: Language Models and Recurrent Neural Networks

(Slides mostly from Chris Manning’s 2023 version)
Lecture Plan

1. A bit more about neural networks (10 mins)
   Language modeling + RNNs

   • 2. A new NLP task: **Language Modeling** (20 mins)

   • 3. A new family of neural networks: **Recurrent Neural Networks (RNNs)** (25 mins)

   • 4. Problems with RNNs (15 mins)

   • 5. Recap on RNNs/LMs (10 mins)

Reminders:

You should have handed in Assignment 2 by today, start of class
In Assignment 3, out today, you build a neural dependency parser using PyTorch
Modern neural networks (esp. language models) are enormous

- Large, deep neural nets are a cornerstone of modern NLP systems

https://huggingface.co/blog/large-language-models
But building large neural networks isn’t easy or obvious

Greedy Layer-Wise Training of Deep Networks

Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle
Université de Montréal
Montréal, Québec
{bengioy, lamblinp, popovici, larocheh}@iro.umontreal.ca

However, until recently, it was believed too difficult to train deep multi-layer neural networks. Empirically, deep networks were generally found to be not better, and often worse, than neural networks with one or two hidden layers (Tesauro, 1992). As this is a negative result, it has not been much reported in the machine learning literature. A reasonable explanation is that gradient-based optimization starting from random initialization may get stuck near poor solutions. An approach that has been explored with some success in the past is based on constructively adding layers. This was previously done using a supervised criterion at each stage (Fahlman & Lebiere, 1990; Langløt & Denerux, 1996). Hinton

- It took a long time and much work to make deep neural networks practical!
We have models with many parameters! Regularization!

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

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

- 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 (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov 2012/JMLR 2014)

- During training
  - For each data point each time:
    - Randomly set input to 0 with probability $p$ “dropout ratio” (often $p = 0.5$ except $p = 0.15$ for input layer) via dropout mask
  - During testing
    - Multiply all weights by $1 - p$
    - No other dropout

\[
\begin{align*}
\text{Train 1:} & \quad y = w_1 x_1 + w_3 x_3 + b \\
\text{Train 2:} & \quad y = w_1 x_1 + w_2 x_2 + w_3 x_3 \\
\text{Test:} & \quad y = (1 - p)(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4)
\end{align*}
\]
Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov 2012/JMLR 2014)

Why does it work?
Prevents Feature Co-adaptation = Good Regularization! Use it everywhere!

Let’s talk through an example..
• Training time: at each instance of evaluation (in online SGD-training), randomly set ~50% ($p\%$) of the inputs to each neuron to 0 (less for the first layer)
• Test time: halve the model weights (now twice as many)
• 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 (all feature weights set independently) and logistic regression models (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
[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)
```

- for loop: 1000 loops, best of 3: 639 µs per loop
- Using single a C x N matrix: 10000 loops, best of 3: 53.8 µs per loop
- 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!
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** ~ 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
    - E.g., start it higher and halve it every $k$ epochs (passes through full data, shuffled or sampled)
- For more complex nets, or to avoid worry, try more sophisticated “adaptive” optimizers that scale the adjustment to individual parameters by an accumulated gradient
  - These models give differential per-parameter learning rates
    - Adagrad $\leftarrow$ Simplest member of family, but tends to “stall early”
    - RMSprop
    - Adam $\leftarrow$ A fairly good, safe place to begin in many cases
    - AdamW
    - NAdamW $\leftarrow$ Can be better with word vectors (W) and for speed (Nesterov acceleration)
    - ...
- Start them with an initial learning rate, around 0.001 $\leftarrow$ Many have other hyperparameters
2. Language Modeling

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

  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 $x^{(t+1)}$:

$$P(x^{(t+1)} | x^{(t)}, \ldots, x^{(1)})$$

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 a 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**: An *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”
  • four-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)

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

(statistical approximation)

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

Suppose we are learning a 4-gram Language Model.

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

For example, suppose that in the corpus:

• “students opened their” occurred 1000 times
• “students opened their books” occurred 400 times
  \[\rightarrow P(\text{books} \mid \text{students opened their}) = 0.4\]
• “students opened their exams” occurred 100 times
  \[\rightarrow P(\text{exams} \mid \text{students opened their}) = 0.1\]
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!

**Partial 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 $w$!

**Partial 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.

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

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*

* Try for yourself: [https://nlpforhackers.io/language-models/](https://nlpforhackers.io/language-models/)

Today the _______

get probability distribution

Company 0.153
bank 0.153
price 0.077
Italian 0.039
emirate 0.039

Sparsity problem: not much granularity in the probability distribution

Otherwise, seems reasonable!
Generating text with a n-gram Language Model

You can also use a Language Model to generate text

today the
condition on this

generate sample

<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 text

get 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>

...
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 2:
A fixed-window neural Language Model

as the proctor started the clock
discard

the students opened their _____

fixed window
A fixed-window neural Language Model

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

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

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
3. Recurrent Neural Networks (RNN)
A family of neural architectures

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

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

hidden states
\[ h^{(t)} = \sigma \left( Wh^{(t-1)} + We^{(t)} + b_1 \right) \]
\( h^{(0)} \) 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
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 \( x^{(t+1)} \)):
  \[
  J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = - \sum_{w \in V} y_w^{(t)} \log \hat{y}_w^{(t)} = - \log \hat{y}_{x_{t+1}}^{(t)}
  \]
- Average this to get overall loss for entire training set:
  \[
  J(\theta) = \frac{1}{T} \sum_{t=1}^{T} J^{(t)}(\theta) = \frac{1}{T} \sum_{t=1}^{T} - \log \hat{y}_{x_{t+1}}^{(t)}
  \]
Training an RNN Language Model

= negative log prob of “students”

Loss

Predicted prob dists

Corpus

= negative log prob of “students”

Loss

Predicted prob dists

Corpus
Training an RNN Language Model

Loss → \( J^{(1)}(\theta) \) → \( \hat{y}^{(1)} \) → \( U \) → \( h^{(0)} \) → \( W_{h} \) → \( e^{(1)} \) → \( E \) → \( x^{(1)} \) → Corpus

\( \text{the} \)

\( \text{students} \)

\( \text{opened} \)

\( \text{their} \)

\( \text{exams} \)

\( \Rightarrow \) Predicted prob dists → \( \hat{y}^{(2)} \) → \( U \) → \( h^{(1)} \) → \( W_{h} \) → \( e^{(2)} \) → \( E \) → \( x^{(2)} \)

\( \text{students} \)

\( \Rightarrow \) Loss → \( J^{(2)}(\theta) \) → \( \hat{y}^{(3)} \) → \( U \) → \( h^{(2)} \) → \( W_{h} \) → \( e^{(3)} \) → \( E \) → \( x^{(3)} \)

\( \text{opened} \)

\( \Rightarrow \) Loss → \( J^{(3)}(\theta) \) → \( \hat{y}^{(4)} \) → \( U \) → \( h^{(3)} \) → \( W_{h} \) → \( e^{(4)} \) → \( E \) → \( x^{(4)} \)

\( \text{their} \)

\( \Rightarrow \) Loss → \( J^{(4)}(\theta) \) → \( \hat{y}^{(4)} \) → \( U \) → \( h^{(4)} \) → \( W_{h} \) → \( e^{(4)} \) → \( E \) → \( x^{(4)} \)

\( \text{exams} \)

\( \Rightarrow \)

\( = \) negative log prob of “opened”

\( \Rightarrow \)
Training an RNN Language Model

Corpus → the, students, opened, their, exams →

\[ \ function of \ "their" \ = \ Negative \ log \ probability \]

Predicted prob dists →

Loss →

\[ J^{(1)}(\theta) \rightarrow J^{(2)}(\theta) \rightarrow J^{(3)}(\theta) \rightarrow J^{(4)}(\theta) \]
Training an RNN Language Model

Loss $\rightarrow J^{(1)}(\theta) \rightarrow \hat{y}^{(1)} \downarrow U \rightarrow h^{(1)} \rightarrow W_h \rightarrow e^{(1)} \rightarrow E \rightarrow x^{(1)}$

Predicted prob dists

$J^{(2)}(\theta) \rightarrow \hat{y}^{(2)} \downarrow U \rightarrow h^{(2)} \rightarrow W_h \rightarrow e^{(2)} \rightarrow E \rightarrow x^{(2)}$

$J^{(3)}(\theta) \rightarrow \hat{y}^{(3)} \downarrow U \rightarrow h^{(3)} \rightarrow W_h \rightarrow e^{(3)} \rightarrow E \rightarrow x^{(3)}$

$J^{(4)}(\theta) \rightarrow \hat{y}^{(4)} \downarrow U \rightarrow h^{(4)} \rightarrow W_h \rightarrow e^{(4)} \rightarrow E \rightarrow x^{(4)}$

= negative log prob of “exams”

Corpus $\rightarrow the \rightarrow students \rightarrow opened \rightarrow their \rightarrow exams$
Training an RNN Language Model

\[ \text{Loss} \rightarrow J^{(1)}(\theta) + J^{(2)}(\theta) + J^{(3)}(\theta) + J^{(4)}(\theta) + \ldots = \frac{1}{T} \sum_{t=1}^{T} J^{(t)}(\theta) \]

"Teacher forcing"

Corpus \rightarrow \text{x}^{(1)}, \text{x}^{(2)}, \text{x}^{(3)}, \text{x}^{(4)} \rightarrow \text{students opened their exams} \ldots
Training a RNN Language Model

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

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

• 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 on a new batch of sentences.
Question: What’s the derivative of $J^{(t)}(\theta)$ w.r.t. the repeated weight matrix $W_h$?

Answer: 

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

“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

One final output $f(x(t), y(t))$

Two intermediate outputs $x(t)$, $y(t)$

One input $t$

Gradients sum at outward branches

$$a = x + y$$
$$b = \max(y, z)$$
$$f = ab$$

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial a} \frac{\partial a}{\partial y} + \frac{\partial f}{\partial b} \frac{\partial b}{\partial y}$$

Source:
Training the parameters of RNNs: Backpropagation for RNNs

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]

In practice, often “truncated” after ~20 timesteps for training efficiency reasons.
Generating with an RNN Language Model ("Generating roll outs")

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

<math>\begin{align*}
    h^{(0)} & \rightarrow W_h \rightarrow U \rightarrow W_e \\
    e^{(1)} & \rightarrow E \rightarrow \langle s \rangle \\
    my & \rightarrow \hat{y}^{(1)} \rightarrow sample \\
    favorite & \rightarrow \hat{y}^{(2)} \rightarrow sample \\
    season & \rightarrow \hat{y}^{(3)} \rightarrow sample \\
    is & \rightarrow \hat{y}^{(4)} \rightarrow sample \\
    spring & \rightarrow \hat{y}^{(4)} \rightarrow sample \\
    \text{sample} & \rightarrow \hat{y}^{(4)} \\
    \langle s \rangle & \rightarrow \text{sample} \\
    \end{align*}</math>
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:

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
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 *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 recipes:

```
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:

![Paint Color Names](http://aiweirdness.com/post/160776374467/new-paint-colors-invented-by-neural-network)

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.

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

Inverse probability of corpus, according to Language Model

Normalized by number of words

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

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

Lower perplexity is better!
RNNs greatly improved perplexity over what came before

<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>Ours small (LSTM-2048)</td>
<td>43.9</td>
</tr>
<tr>
<td>Ours large (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/](https://research.fb.com/building-an-efficient-neural-language-model-over-a-billion-words/)
4. Problems with RNNs: Vanishing and Exploding Gradients

\[ J^{(4)}(\theta) \]

\[ h^{(1)} \rightarrow W \rightarrow h^{(2)} \rightarrow W \rightarrow h^{(3)} \rightarrow W \rightarrow h^{(4)} \]
Vanishing gradient intuition

\[ \frac{\partial J^{(4)}}{\partial h^{(1)}} = ? \]
Vanishing gradient intuition

\[
\frac{\partial J^{(4)}}{\partial h^{(1)}} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \times \frac{\partial J^{(4)}}{\partial h^{(2)}}
\]

chain rule!
Vanishing gradient intuition

\[ \frac{\partial J^{(4)}}{\partial h^{(1)}} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \times \frac{\partial h^{(3)}}{\partial h^{(2)}} \times \frac{\partial J^{(4)}}{\partial h^{(3)}} \]

chain rule!
Vanishing gradient intuition

\[
\frac{\partial J^{(4)}}{\partial h^{(1)}} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \times \frac{\partial h^{(3)}}{\partial h^{(2)}} \times \frac{\partial h^{(4)}}{\partial h^{(3)}} \times \frac{\partial J^{(4)}}{\partial h^{(4)}}
\]

chain rule!
Vanishing gradient intuition

What happens if these are small?

Vanishing gradient problem:
When these are small, the gradient signal gets smaller and smaller as it backpropagates further.
Vanishing gradient proof sketch (linear case)

- Recall:
  \[ h^{(t)} = \sigma \left( W_h h^{(t-1)} + W_x x^{(t)} + b_1 \right) \]

- What if \( \sigma \) were the identity function, \( \sigma(x) = x \)?
  \[ \frac{\partial h^{(t)}}{\partial h^{(t-1)}} = \text{diag} \left( \sigma' \left( W_h h^{(t-1)} + W_x x^{(t)} + b_1 \right) \right) W_h \] (chain rule)
  \[ = I \cdot W_h = W_h \]

- Consider the gradient of the loss \( J^{(i)}(\theta) \) on step \( i \), with respect to the hidden state \( h^{(j)} \) on some previous step \( j \). Let \( \ell = i - j \)
  \[ \frac{\partial J^{(i)}(\theta)}{\partial h^{(j)}} = \frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} \prod_{j < t \leq i} \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \] (chain rule)
  \[ = \frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} \prod_{j < t \leq i} W_h = \left[ \frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} \right] W_h^\ell \] (value of \( \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \))

If \( W_h \) is “small”, then this term gets exponentially problematic as \( \ell \) becomes large
Vanishing gradient proof sketch (linear case)

• What’s wrong with $W^\ell_h$?
• Consider if the eigenvalues of $W_h$ are all less than 1:
  \[ \lambda_1, \lambda_2, \ldots, \lambda_n < 1 \]
  \[ q_1, q_2, \ldots, q_n \text{ (eigenvectors)} \]
• We can write $\frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} W^\ell_h$ using the eigenvectors of $W_h$ as a basis:
  \[ \frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} W^\ell_h = \sum_{i=1}^{n} c_i \lambda_i^\ell q_i \approx 0 \text{ (for large } \ell) \]

Approaches 0 as $\ell$ grows, so gradient vanishes

• What about nonlinear activations $\sigma$ (i.e., what we use?)
  • Pretty much the same thing, except the proof requires $\lambda_i < \gamma$
    for some $\gamma$ dependent on dimensionality and $\sigma$

Why is vanishing gradient a problem?

Gradient signal from far away is lost because it’s much smaller than gradient signal from close-by.

So, model weights are updated only with respect to near effects, not long-term effects.
Effect of vanishing gradient on RNN-LM

- **LM task:** *When she tried to print her tickets, she found that the printer was out of toner. She went to the stationery store to buy more toner. It was very overpriced. After installing the toner into the printer, she finally printed her ________*

- To learn from this training example, the RNN-LM needs to **model the dependency** between “tickets” on the 7th step and the target word “tickets” at the end.

- But if the gradient is small, the model **can’t learn this dependency**
  - So, the model is **unable to predict similar long-distance dependencies** at test time
Why is exploding gradient a problem?

• If the gradient becomes too big, then the SGD update step becomes too big:

\[ \theta^{new} = \theta^{old} - \alpha \nabla_{\theta} J(\theta) \]

learning rate

• This can cause bad updates: we take too large a step and reach a weird and bad parameter configuration (with large loss)
  • You think you’ve found a hill to climb, but suddenly you’re in Iowa

• In the worst case, this will result in Inf or NaN in your network (then you have to restart training from an earlier checkpoint)
Gradient clipping: solution for exploding gradient

• **Gradient clipping**: if the norm of the gradient is greater than some threshold, scale it down before applying SGD update

    \[
    \hat{g} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta} \\
    \text{if } \|\hat{g}\| \geq \text{threshold} \text{ then} \\
    \hat{g} \leftarrow \frac{\text{threshold}}{\|\hat{g}\|} \hat{g} \\
    \text{end if}
    \]

• **Intuition**: take a step in the same direction, but a smaller step

• In practice, **remembering to clip gradients is important**, but exploding gradients are an easy problem to solve

How to fix the vanishing gradient problem?

• The main problem is that *it’s too difficult for the RNN to learn to preserve information over many timesteps.*

• In a vanilla RNN, the hidden state is constantly being *rewritten*

\[ h(t) = \sigma \left( W_h h(t-1) + W_x x(t) + b \right) \]

• First off next time: How about an RNN with separate *memory* which is added to?
  • LSTMs

• And then: Creating more direct and linear pass-through connections in model
  • Attention, residual connections, etc.
5. 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 (despite some problems)

- RNNs are also useful for much more!
Why should we care about Language Modeling?

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

- 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.

- Everything else in NLP has now been rebuilt upon Language Modeling: **GPT-3 is an LM!**
Other RNN uses: RNNs can be used for sequence 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?

positive

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

equals

positive

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?

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

sentence encoding

positive

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.

**Context:** Ludwig van Beethoven was a German composer and pianist. A crucial figure ...
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 starting next lecture.
Terminology and a look forward

The RNN described in this lecture = \textit{simple/vanilla/Elman} RNN

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

By the end of the course: You will understand phrases like “\textit{stacked bidirectional LSTMs with residual connections and self-attention}”