

## Language Models



Dan Klein  
UC Berkeley

## Language Models



## Language Models



## Acoustic Confusions

the station signs are in deep in english	-14732
the stations signs are in deep in english	-14735
the station signs are in deep into english	-14739
the station 's signs are in deep in english	-14740
the station signs are in deep in the english	-14741
the station signs are indeed in english	-14757
the station 's signs are indeed in english	-14760
the station signs are indians in english	-14790



## Noisy Channel Model: ASR

- We want to predict a sentence given acoustics:

$$w^* = \arg \max_w P(w|a)$$

- The noisy-channel approach:

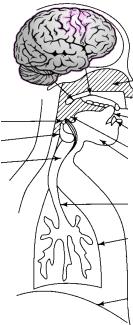
$$w^* = \arg \max_w P(w|a)$$

$$= \arg \max_w P(a|w)P(w)/P(a)$$

$$\propto \arg \max_w P(a|w)P(w)$$

Acoustic model: score fit between sounds and words

Language model: score plausibility of word sequences



## Perplexity

"when i eat pizza, i wipe off the"

All Videos Images News Shopping More Settings Tools

About 1,520 results (0.41 seconds)

<https://cal-cs288.github.io/slides.pdf>

SP20 CS288 - Language Models (1) - GitHub Pages

When I eat pizza, I wipe off the \_\_\_\_\_. Formally: test set log likelihood. - Perplexity: "average per word branching factor" (not per-step perp., + exp ...

<https://people.eecs.berkeley.edu/~klein/slides.pdf>

2PP - People @ EECS at UC Berkeley

Unigrams are terrible at this game. (Why?) - "Entropy": per-word test log likelihood (misnamed). When I eat pizza, I wipe off the \_\_\_\_\_. Many children are allergic ... 2011

<https://people.eecs.berkeley.edu/~klein/slides.pdf>

Natural Language Processing - People @ EECS at UC Berkeley

When I eat pizza, I wipe off the \_\_\_\_\_. Formally: define test set (log) likelihood. - Perplexity: "average per word branching factor" (not per-step) perp X exp.

<https://courses.cs.washington.edu/LanguageModels.pdf>

CSE517 Natural Language Processing ... - Washington

How good are we doing? Compute per word log likelihood (M words, m test sentences s.i.): When I eat pizza, I wipe off the \_\_\_\_\_. Many children are allergic to ...

tor" (not per-step)

$\theta)$

grease 0.5  
sauce 0.4  
dust 0.05  
...  
mice 0.0001  
...  
the 1e-100

3516 wipe off the excess  
1034 wipe off the dust  
547 wipe off the sweat  
518 wipe off the mouthpiece  
...  
120 wipe off the grease  
0 wipe off the sauce  
0 wipe off the mice  
-----  
28048 wipe off the \*



## Noisy Channel Model: Translation

"Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' "

Warren Weaver (1947)

## N-Gram Models



## N-Gram Models

- Use chain rule to generate words left-to-right

$$P(w_1 \dots w_n) = \prod_i P(w_i | w_1 \dots w_{i-1})$$

- Can't condition atomically on the entire left context

$P(\text{???} | \text{The computer I had put into the machine room on the fifth floor just})$

- N-gram models make a Markov assumption

$$P(w_1 \dots w_n) = \prod_i P(w_i | w_{i-k} \dots w_{i-1})$$

$P(\text{please close the door}) = P(\text{please}|\text{START})P(\text{close}|\text{please}) \dots P(\text{STOP}|\text{door})$



## Empirical N-Grams

- Use statistics from data (examples here from Google N-Grams)

Training Counts
198015222 the first
194623024 the same
168504105 the following
158562063 the world
...
14112454 the door

$$\hat{P}(\text{door}|\text{the}) = \frac{14112454}{23135851162} = 0.0006$$

- This is the maximum likelihood estimate, which needs modification  
23135851162 the \*
- N-gram models use such counts to compute probabilities on demand



## Increasing N-Gram Order

- Higher orders capture more correlations

Bigram Model

198015222	the first
194623024	the same
168504105	the following
158562063	the world
...	
14112454	the door
-----	
23135851162	the *

Trigram Model

197302	close the window
191125	close the door
152500	close the gap
116451	close the thread
87298	close the deal
-----	
3785230	close the *

$P(\text{door} | \text{the}) = 0.0006$

$P(\text{door} | \text{close the}) = 0.05$



## Increasing N-Gram Order

Unigram

- To him swallowed confess hear both. Which. Of save on trail for are ay device and  
rote life have
- Every enter now severally so, let
- Hill he late speaks; or! a more to leg less first you enter
- Are where exunt and sighs have rise excellency took of.. Sleep knave we, near; vile  
like



## What's in an N-Gram?

- Just about every local correlation!
  - Word class restrictions: "will have been \_\_\_\_"
  - Morphology: "she \_\_\_\_", "they \_\_\_\_"
  - Semantic class restrictions: "danced a \_\_\_\_"
  - Idioms: "add insult to \_\_\_\_"
  - World knowledge: "ice caps have \_\_\_\_"
  - Pop culture: "the empire strikes \_\_\_\_"
  
- But not the long-distance ones
  - "The computer which I had put into the machine room on the fifth floor just \_\_\_\_."



## Linguistic Pain

- The N-Gram assumption hurts your inner linguist
  - There are many linguistic arguments that language isn't regular
    - Long-distance dependencies
    - Recursive structure
  - At the core of the early hesitance in linguistics about statistical methods
  
- Answers
  - N-grams only model local correlations... but they get them all
  - As N increases, they catch even more correlations
  - N-gram models scale well -- much more easily than combinatorially-structured LMs
  - Can build LMs from structured models, eg grammars (though people generally don't)



## Structured Language Models

- Bigram model:
  - [texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen]
  - [outside, new, car, parking, lot, of, the, agreement, reached]
  - [this, would, be, a, record, november]
  
- PCFG model:
  - [This, quarter, 's, surprisingly, independent, attack, paid, off, the, risk, involving, IRS, leaders, and, transportation, prices, .]
  - [It, could, be, announced, sometime, .]
  - [Mr., Toseland, believes, the, average, defense, economy, is, drafted, from, slightly, more, than, 12, stocks, .]



## N-Grams on the Web

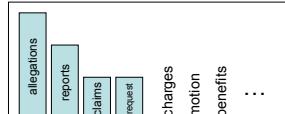
The screenshot shows the interface of the Google Web 1T 5-Gram Database. The top navigation bar includes links for Frequency list, Associations, and Collocations. Below the navigation is a search bar with the placeholder "Search pattern: Berkeley is a \*". Underneath the search bar are dropdown menus for "display First: 50", "N-grams with frequency ≥ 40", and "variable elements are listed normally". There are also buttons for "Search", "CSV", "XML", "Help", "Debug", and "Optim.". A "Reset Form" button is located at the bottom right of the search area. The main content area is titled "Results" and displays a table of n-grams. The table has two columns: the rank and the n-gram itself. The data starts with:

Rank	n-gram
306	berkeley is a charming
242	berkeley is a fine
165	berkeley is a city
155	berkeley is a great
133	berkeley is a very
115	berkeley is a public
88	berkeley is a cool
85	berkeley is a sharp
66	berkeley is a member
63	berkeley is a place
59	berkeley is a small
56	berkeley is a wonderful
51	berkeley is a major
50	berkeley is a new
49	berkeley is a versatile

## N-Gram Models: Challenges

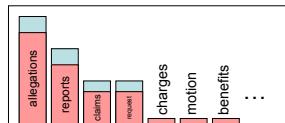
- We often want to make estimates from sparse statistics:

$P(w \mid \text{denied the})$   
3 allegations  
2 reports  
1 claims  
1 request  
7 total



- Smoothing flattens spiky distributions so they generalize better:

$P(w \mid \text{denied the})$   
2.5 allegations  
1.5 reports  
0.5 claims  
0.5 request  
**2 other**  
7 total



- Very important all over NLP, but easy to do badly

## Sparsity

*Please close the first door on the left.*

3380 please close the door  
1601 please close the window  
1164 please close the new  
1159 please close the gate  
...  
0 please close the first  
-----  
13951 **please close the \***

## Smoothing

- We often want to make estimates from sparse statistics:

## Back-off

*Please close the first door on the left.*

4-Gram

3380 please close the door  
1601 please close the window  
1164 please close the new  
1159 please close the gate  
...  
0 please close the first  
-----  
13951 please close the \*

0.0

3-Gram

197302 close the window  
191125 close the door  
152500 close the gap  
116451 close the thread  
...  
8662 close the first  
-----  
3785230 close the \*

0.002

2-Gram

198015222 the first  
194623024 the same  
168504105 the following  
158562063 the world  
...  
23135851162 the \*

0.009

Specific but Sparse



Dense but General

$$\lambda \hat{P}(w|w_{-1}, w_{-2}) + \lambda' \hat{P}(w|w_{-1}) + \lambda'' \hat{P}(w)$$



## Discounting

- Observation: N-grams occur more in training data than they will later

Empirical Bigram Counts (Church and Gale, 91)

Count in 22M Words	Future $c^*$ (Next 22M)
1	
2	
3	
4	
5	

- Absolute discounting: reduce counts by a small constant, redistribute "shaved" mass to a model of new events

$$P_{\text{ad}}(w|w') = \frac{c(w', w) - d}{c(w')} + \alpha(w') \hat{P}(w)$$



## Fertility

- Shannon game: "There was an unexpected \_\_\_\_\_"

delay? Francisco?

- Context fertility: number of distinct context types that a word occurs in

- What is the fertility of "delay"?
- What is the fertility of "Francisco"?
- Which is more likely in an arbitrary new context?

- Kneser-Ney smoothing: new events proportional to context fertility, not frequency

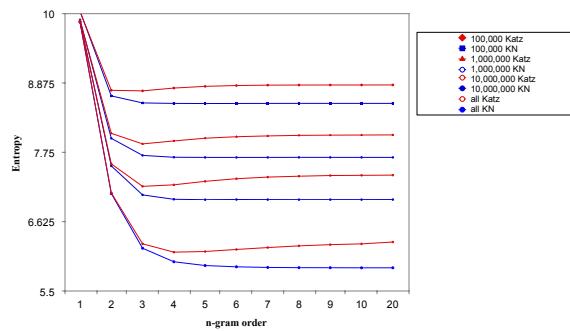
[Kneser & Ney, 1995]

$$P(w) \propto |\{w': c(w', w) > 0\}|$$

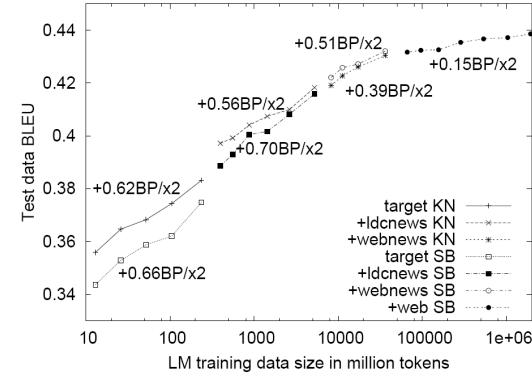
- Can be derived as inference in a hierarchical Pitman-Yor process [Teh, 2006]



## Better Methods?



## More Data?



[Brants et al, 2007]



## Storage

...	
searching for the best	192593
searching for the right	45805
searching for the cheapest	44965
searching for the perfect	43959
searching for the truth	23165
searching for the "	19086
searching for the most	15512
searching for the latest	12670
searching for the next	10120
searching for the lowest	10080
searching for the name	8402
searching for the finest	8171
...	

Google N-grams

- 14 million  $< 2^{24}$  words
- 2 billion  $< 2^{31}$  5-grams
- 770 000  $< 2^{20}$  unique counts
- 4 billion n-grams total

## Storage



- For 5+-gram models, need to store between 100M and 10B context-word-count triples

w	c	val
1933	15176585	3
1933	15176587	2
1933	15176593	1
1933	15176613	8
1933	15179801	1
1935	15176585	298
1935	15176589	1

$\Delta w$	$\Delta c$	val
+0	+2	1
+0	+5	1
+0	+40	8
+0	+188	1
+2	15176585	298
+0	+4	1

$ \Delta w $	$ \Delta c $	val
24	40	3
2	3	3
2	3	3
2	9	6
2	12	3
4	36	15
2	6	3

- Make it fit in memory by *delta encoding* scheme: store deltas instead of values and use variable-length encoding

Pauls and Klein (2011), Heafield (2011)

Slide: Greg Durrett



## Graveyard of Correlations

- Skip-grams
- Cluster models
- Topic variables
- Cache models
- Structural zeros
- Dependency models
- Maximum entropy models
- Subword models
- ...



## Entirely Unseen Words

- What about totally unseen words?
- Classical real world option: systems are actually closed vocabulary
  - ASR systems will only propose words that are in their pronunciation dictionary
  - MT systems will only propose words that are in their phrase tables (modulo special models for numbers, etc)
- Classical theoretical option: build open vocabulary LMs
  - Models over character sequences rather than word sequences
  - N-Grams: back-off needs to go down into a “generate new word” model
  - Typically if you need this, a high-order character model will do
- Modern approach: syllable-sized subword units (more later)

## Neural LMs: Preview

### A GPT2 Continuation

- The computer I had put into the machine room on the fifth floor just outside our landing was taken by a lot of people. It was going to be recovered from outside the machine room as soon as we could, but after the machine room was shut down, we had to open all of the windows and doors to save ourselves, and the computer would have been lost. Luckily we are prepared for this. We had five steel doors on each side of the landing to hold any rain and wind that might have hit, so we weren't affected by the storm.

The situation was really grim for days. A few days before, the Apache at Qush Tefah had been shot. This was the most damaged machine in the Armies Corps. The command was only meant to repair Qush



### A GPT2 Continuation

- I shall go seek the traitor Gloucester, when he hath declared his vengeance. Cunegund Shall I back her to kill my cousin in my stead? Witbane Far first shall the brother take his Queen's soul and that in the flames. Clotilda For to slay the King of Scotland with, the blood of my cousin, lie true; And she shall have the right to my feather." Sefton leapt into the rushes, and stole along in the sunlight: the small figure flashed like lightning in the west. In the likeness of a snare he had laid to catch the rushes and made of them a snares, a road to flee from his pursuers; but he now came to an oak where the branches were wreathed in an oak-



### Neural LMs: Three Key Ideas

- Word embeddings**
  - Different words are not entirely unrelated events
  - Words can be more and less similar, in complex ways
- Partially factored representations**
  - Multiple semi-independent processes happen in parallel in language
  - It's too expensive to track language in an unfactored way, and too inaccurate to assume everything of interest is independent
- Long distance dependencies**
  - Information can be relevant without being local
  - Different notions of locality are important at different times

## Words: Clusterings and Embeddings

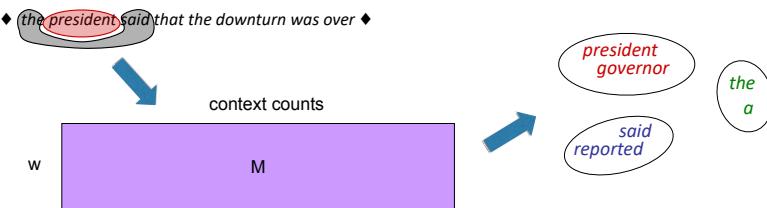
## Stuffing Words into Vector Spaces?



Cartoon: Greg Durrett

## Distributional Similarity

- Key idea in clustering and embedding methods: characterize a word by the words it occurs with (cf Harris' distributional hypothesis, 1954)
- "You can tell a word by the company it keeps." [Firth, 1957]
- Harris / Chomsky divide in linguistic methodology



## Clusterings



## Clusterings

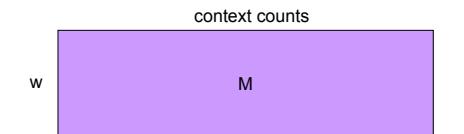
- Automatic (Finch and Chater 92, Shuetze 93, many others)

word	nearest neighbors
accompanied	submitted banned financed developed authorized headed canceled awarded barred
almost	virtually merely formally fully quite officially just nearly only less
causing	reflecting forcing providing creating producing becoming carrying particularly
classes	elections courses payments losses computers performances violations levels pictures
directors	professionals investigations materials competitors agreements papers transactions
goal	mood roof eye image tool song pool scene gap voice
japanese	chinese iraqi american western arab foreign european federal soviet indian
represent	reveal attend deliver reflect choose contain impose manage establish retain
think	believe wish know realize wonder assume feel say mean bet
york	angeles francisco sox rouge kong diego zone vegas inning layer
on	through in at over into with from by across
must	might would could cannot will should can may does helps
they	we you i he she nobody who it everybody there

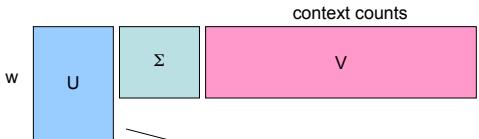
- Manual (e.g. thesauri, WordNet)

## “Vector Space” Methods

- Treat words as points in  $R^n$  (eg Shuetze, 93)



- Form matrix of co-occurrence counts
- SVD or similar to reduce rank (cf LSA)
- Cluster projections
- People worried about things like: log of counts,  $U$  vs  $U\Sigma$



- This is actually more of an embedding method (but we didn't want that in 1993)

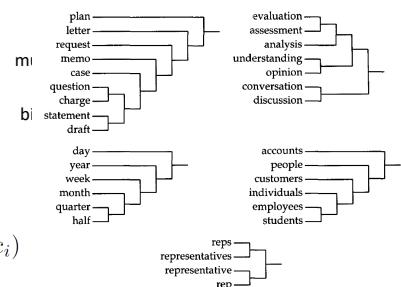
Cluster these 50-200 dim vectors instead.



## Models: Brown Clustering

- Classic model-based clustering (Brown et al, 92)

- Each word starts in its own cluster
- Each cluster has co-occurrence stats
- Greedyly merge clusters based on a information criterion
- Equivalent to optimizing a class-based



- Produces a dendrogram (hierarchy) of clusters

$$P(w_i|w_{i-1}) = P(c_i|c_{i-1})P(w_i|c_i)$$

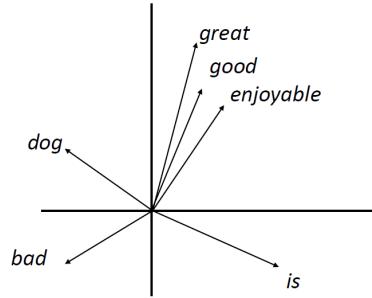
## Embeddings

Most slides from Greg Durrett



## Embeddings

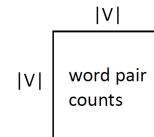
- Embeddings map discrete words (eg  $|V| = 50k$ ) to continuous vectors (eg  $d = 100$ )
- Why do we care about embeddings?
  - Neural methods want them
  - Nuanced similarity possible; generalize across words
- We hope embeddings will have structure that exposes word correlations (and thereby meanings)



## Embedding Models

- Idea: compute a representation of each word from co-occurring words

the dog bit the man



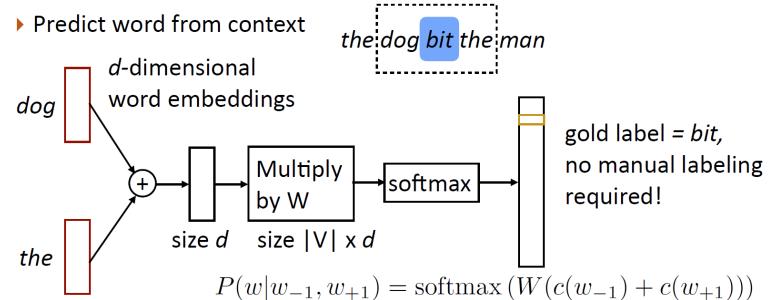
*Token-Level*

*Type-Level*

- We'll build up several ideas that can be mixed-and-matched and which frequently get used in other contexts



## word2vec: Continuous Bag-of-Words

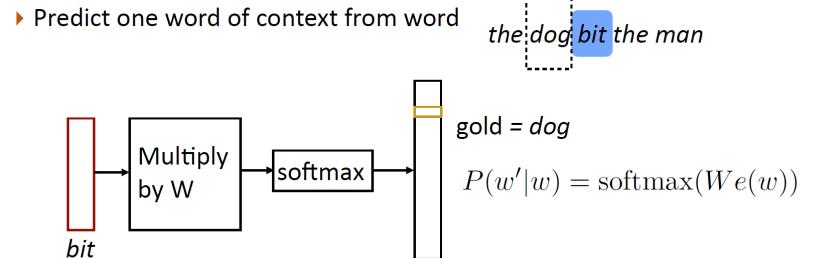


- ▶ Parameters:  $d \times |V|$  (one  $d$ -length context vector per voc word),  $|V| \times d$  output parameters ( $W$ )

Mikolov et al. (2013)



## word2vec: Skip-Grams



- ▶ Another training example:  $\text{bit} \rightarrow \text{the}$

- ▶ Parameters:  $d \times |V|$  vectors,  $|V| \times d$  output parameters ( $W$ ) (also usable as vectors!)

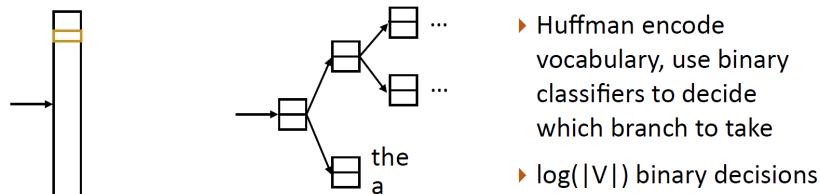
Mikolov et al. (2013)



## word2vec: Hierarchical Softmax

$$P(w|w_{-1}, w_{+1}) = \text{softmax}(W(c(w_{-1}) + c(w_{+1}))) \quad P(w'|w) = \text{softmax}(We(w))$$

- Matmul + softmax over  $|V|$  is very slow to compute for CBOW and SG



Standard softmax:  
 $[|V| \times d] \times d$

Hierarchical softmax:  
 $\log(|V|)$  dot products of size  $d$ ,  
 $|V| \times d$  parameters

Mikolov et al. (2013)



## word2vec: Negative Sampling

- Take (word, context) pairs and classify them as “real” or not. Create random negative examples by sampling from unigram distribution

$(bit, the) \Rightarrow +1$

$(bit, cat) \Rightarrow -1$

$(bit, a) \Rightarrow -1$

$(bit, fish) \Rightarrow -1$

$$P(y = 1|w, c) = \frac{e^{w \cdot c}}{e^{w \cdot c} + 1}$$

words in similar contexts select for similar  $c$  vectors

- $d \times |V|$  vectors,  $d \times |V|$  context vectors (same # of params as before)

Objective =  $\log P(y = 1|w, c) + \frac{1}{k} \sum_{i=1}^n \log P(y = 0|w_i, c)$

sampled

Mikolov et al. (2013)



## fastText: Character-Level Models

- Same as SGNS, but break words down into n-grams with  $n = 3$  to 6

where:

3-grams: <wh, whe, her, ere, re>

4-grams: <whe, wher, here, ere>,

5-grams: <wher, where, here>,

6-grams: <where, where>

- Replace  $w \cdot c$  in skip-gram computation with  $\left( \sum_{g \in \text{ngrams}} w_g \cdot c \right)$

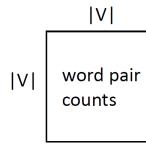
- Advantages?

Bojanowski et al. (2017)



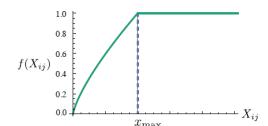
## GloVe

- Idea: Fit co-occurrence matrix directly (weighted least squares)



$$J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

- Type-level computations (so constant in data size)
- Currently the most common word embedding method



Pennington et al, 2014



## Bottleneck vs Co-occurrence

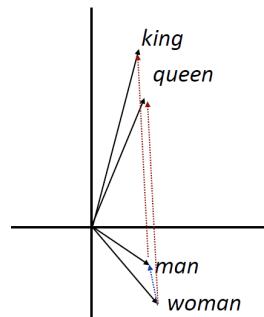
- Two main views of inducing word structure
  - Co-occurrence: model which words occur in similar contexts
  - Bottleneck: model latent structure that mediates between words and their behaviors
- These turn out to be closely related!

## Language Models



## Structure of Embedding Spaces

- How can you fit 50K words into a 64-dimensional hypercube?
- Orthogonality: Can each axis have a global “meaning” (number, gender, animacy, etc)?
- Global structure: Can embeddings have algebraic structure (eg  $\text{king} - \text{man} + \text{woman} = \text{queen}$ )?



## Bias in Embeddings

- Embeddings can capture biases in the data! (Bolukbasi et al 16)

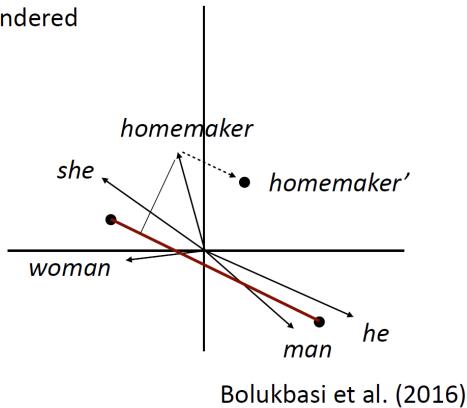
$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

- Debiasing methods (as in Bolukbasi et al 16) are an active area of research



## Debiasing?

- ▶ Identify gender subspace with gendered words
- ▶ Project words onto this subspace
- ▶ Subtract those projections from the original word

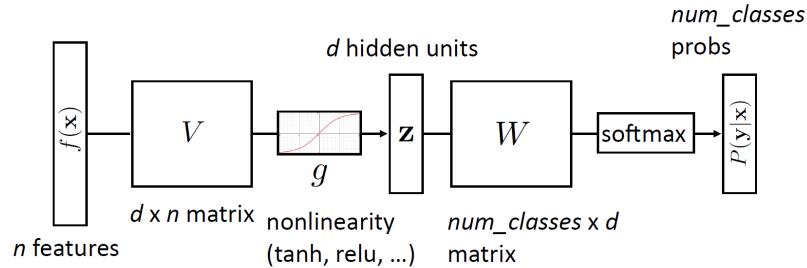


## Neural Language Models

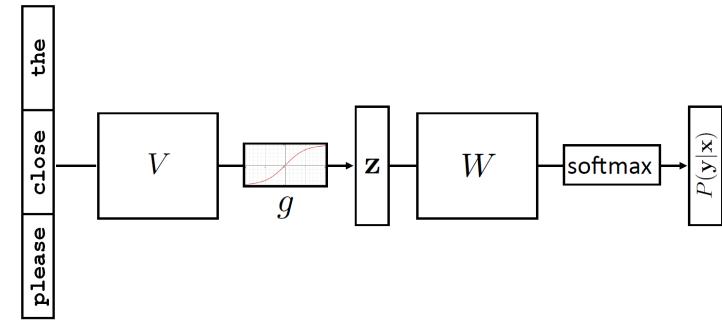


## Reminder: Feedforward Neural Nets

$$P(\mathbf{y}|\mathbf{x}) = \text{softmax}(Wg(Vf(\mathbf{x})))$$



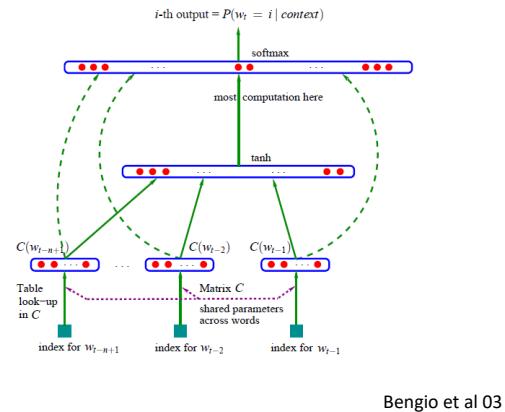
## A Feedforward N-Gram Model?





## Early Neural Language Models

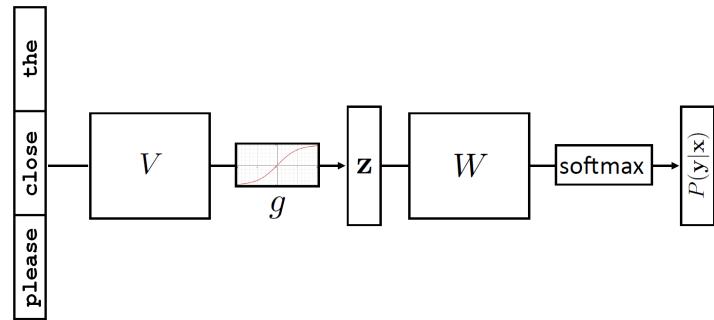
- Fixed-order feed-forward neural LMs
- Eg Bengio et al 03
- Allow generalization across contexts in more nuanced ways than prefixing
- Allow different kinds of pooling in different contexts
- Much more expensive to train



Bengio et al 03



## Using Word Embeddings?



## Using Word Embeddings

- ▶ Approach 1: learn embeddings as parameters from your data
  - ▶ Often works pretty well
- ▶ Approach 2: initialize using GloVe, keep fixed
  - ▶ Faster because no need to update these parameters
- ▶ Approach 3: initialize using GloVe, fine-tune
  - ▶ Works best for some tasks



## Limitations of Fixed-Window NN LMs?

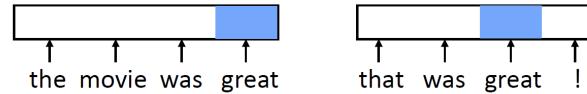
- What have we gained over N-Grams LMs?
- What have we lost?
- What have we not changed?

## Recurrent NNs

Slides from Greg Durrett / UT Austin , Abigail See / Stanford

## RNNs

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

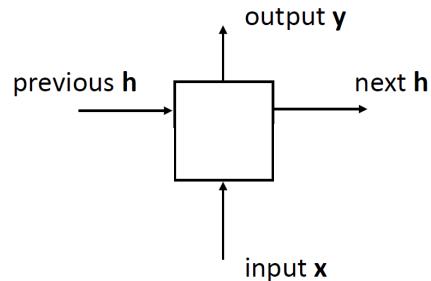


- ▶ These don't look related (*great* is in two different orthogonal subspaces)
- ▶ Instead, we need to:
  - 1) Process each word in a uniform way
  - 2) ...while still exploiting the context that that token occurs in



## General RNN Approach

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

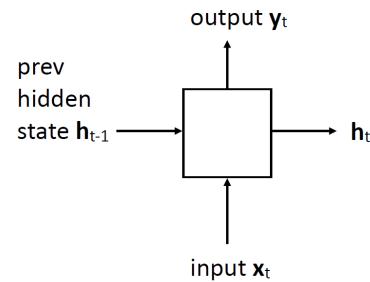


## RNN Uses

- ▶ Transducer: make some prediction for each element in a sequence
  - DT      NN      VBD      JJ  
↑          ↑          ↑          ↑  
the      movie      was      great
  - 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
  - the      movie      was      great



## Basic RNNs



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

- ▶ Updates hidden state based on input and current hidden state

$$y_t = \tanh(Uh_t + b_y)$$

- ▶ Computes output from hidden state

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

Elman (1990)



## Training RNNs



the movie was great

- ▶ “Backpropagation through time”: build the network as one big computation graph, some parameters are shared

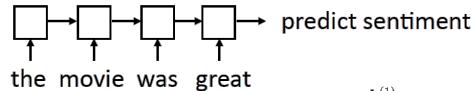
- ▶ RNN potentially needs to learn how to “remember” information for a long time!

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

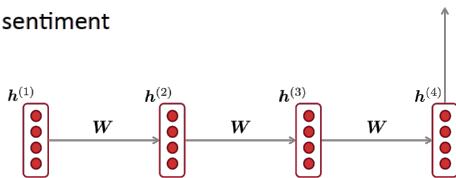
- ▶ “Correct” parameter update is to do a better job of remembering the sentiment of *favorite*



## Problem: Vanishing Gradients



- Contribution of earlier inputs decreases if matrices are contractive (first eigenvalue < 1), non-linearities are squashing, etc
- Gradients can be viewed as a measure of the effect of the past on the future
- That's a problem for optimization but also means that information naturally decays quickly, so model will tend to capture local information



Next slides adapted from Abigail See / Stanford



## Core Issue: Information Decay

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

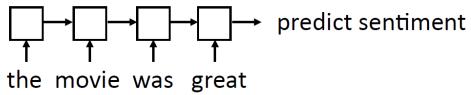
- In a vanilla RNN, the hidden state is constantly being *rewritten*

$$h^{(t)} = \sigma(W_h h^{(t-1)} + W_x x^{(t)} + b)$$

- How about a RNN with separate *memory*?

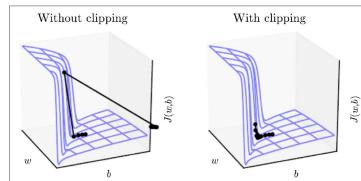


## Problem: Exploding Gradients

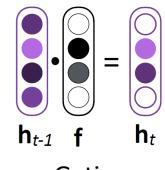
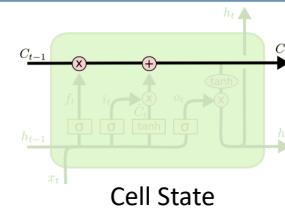


- Gradients can also be too large

- Leads to overshooting / jumping around the parameter space
- Common solution: gradient clipping



## Key Idea: Propagated State

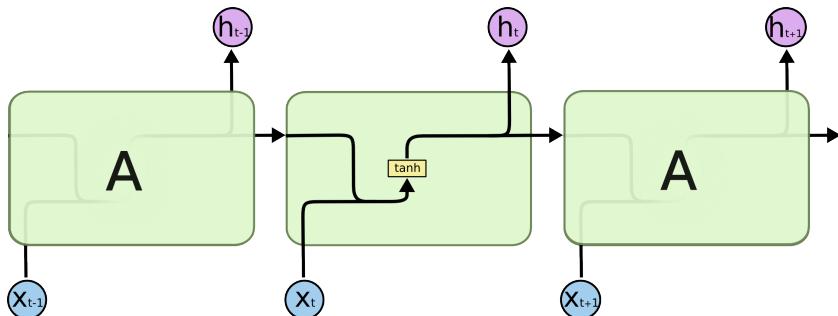


- Information decays in RNNs because it gets multiplied each time step
- Idea: have a channel called the *cell state* that by default just gets propagated (the “conveyer belt”)
- Gates make explicit decisions about what to add / forget from this channel

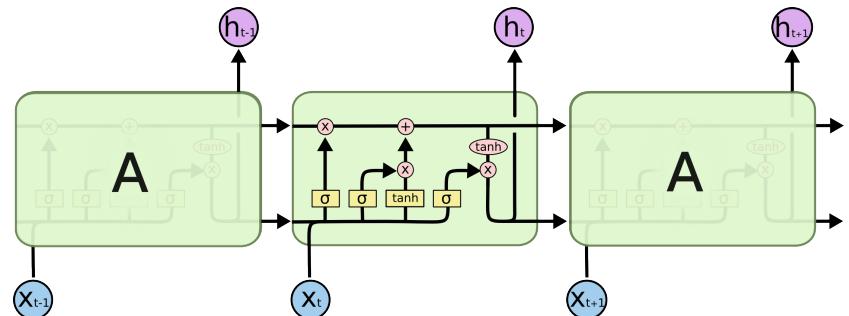
Image: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>



## RNNs

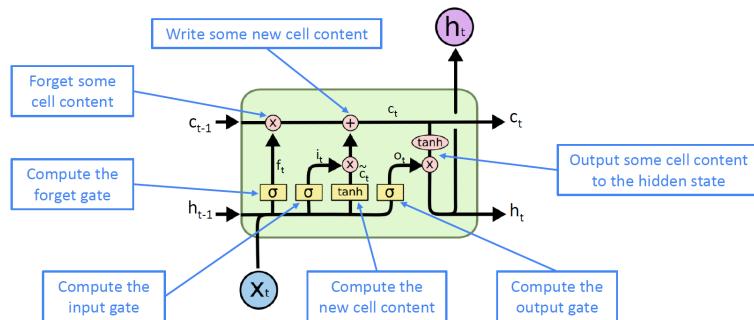


## LSTMs

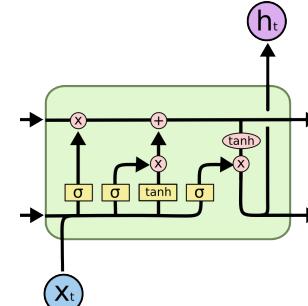




## LSTMs



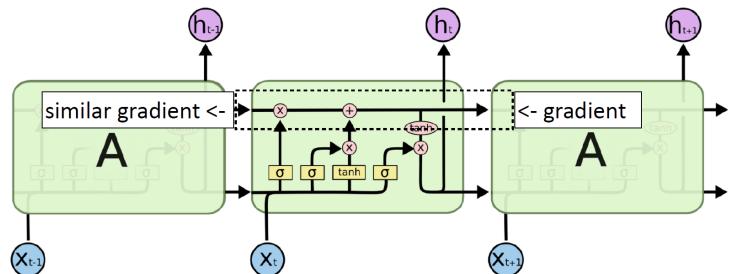
## LSTMs



- ▶ 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



## What about the Gradients?



- ▶ 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/>

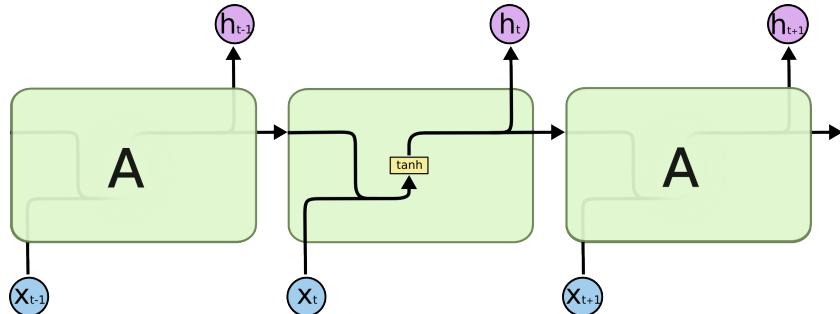


## Gated Recurrent Units (GRUs)

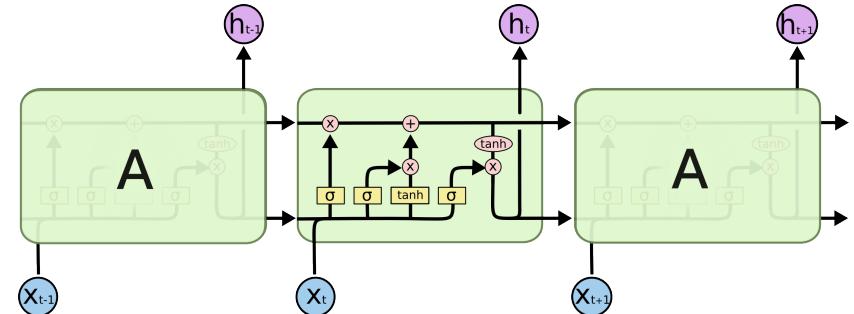
- Update gate:** controls what parts of hidden state are updated vs preserved
- $\mathbf{u}^{(t)} = \sigma(\mathbf{W}_u h^{(t-1)} + \mathbf{U}_u \mathbf{x}^{(t)} + \mathbf{b}_u)$
- Reset gate:** controls what parts of previous hidden state are used to compute new content
- $\mathbf{r}^{(t)} = \sigma(\mathbf{W}_r h^{(t-1)} + \mathbf{U}_r \mathbf{x}^{(t)} + \mathbf{b}_r)$
- New hidden state content:** reset gate selects useful parts of prev hidden state. Use this and current input to compute new hidden content.
- $\tilde{\mathbf{h}}^{(t)} = \tanh(\mathbf{W}_h (\mathbf{r}^{(t)} \circ h^{(t-1)}) + \mathbf{U}_h \mathbf{x}^{(t)} + \mathbf{b}_h)$
- Hidden state:** update gate simultaneously controls what is kept from previous hidden state, and what is updated to new hidden state content
- $\mathbf{h}^{(t)} = (1 - \mathbf{u}^{(t)}) \circ \mathbf{h}^{(t-1)} + \mathbf{u}^{(t)} \circ \tilde{\mathbf{h}}^{(t)}$
- How does this solve vanishing gradient?**  
Like LSTM, GRU makes it easier to retain info long-term (e.g. by setting update gate to 0)



## RNNs



## LSTMs



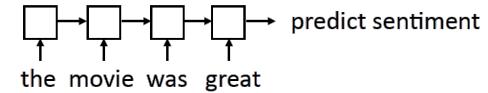
## Uses of RNNs

Slides from Greg Durrett / UT Austin

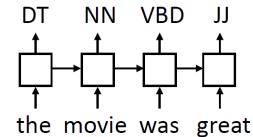


## Reminder: Tasks for RNNs

- Sentence Classification (eg Sentiment Analysis)



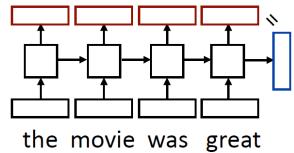
- Transduction (eg Part-of-Speech Tagging, NER)



- Encoder/Decoder (eg Machine Translation)



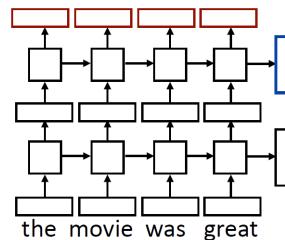
## Encoder / Decoder Preview



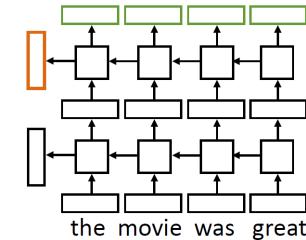
- ▶ **Encoding of the sentence** — can pass this to a decoder or make a classification decision about the sentence
- ▶ **Encoding of each word** — can pass this to another layer to make a prediction (can also pool these to get a different sentence encoding)
- ▶ RNN can be viewed as a transformation of a sequence of vectors into a sequence of context-dependent vectors



## Multilayer and Bidirectional RNNs



- ▶ Sentence classification based on concatenation of both directions' final outputs

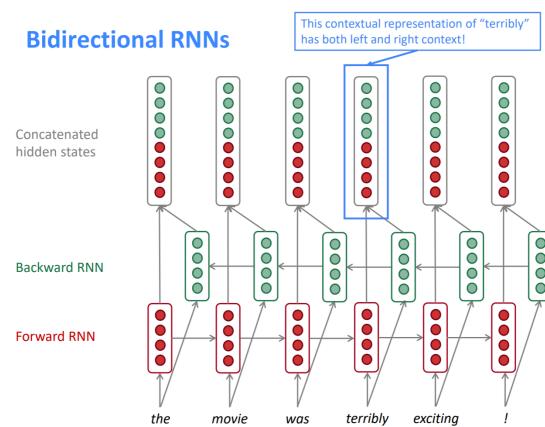


- ▶ Token classification based on concatenation of both directions' token representations



## Bi-Directional RNNs

### Bidirectional RNNs

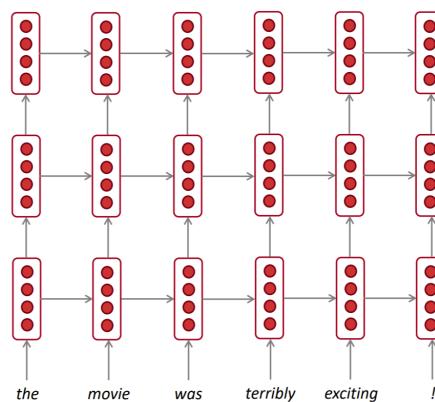


## Multi-Layer RNNs

RNN layer 3

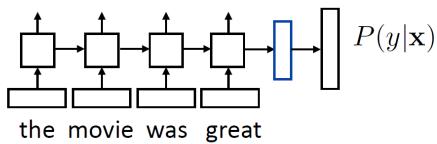
RNN layer 2

RNN layer 1





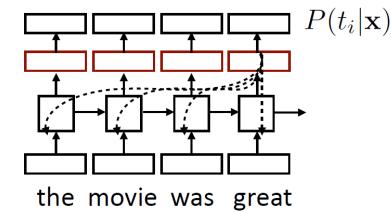
## Training for Sentential Tasks



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



## Training for Transduction Tasks



- ▶ 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)



## Example Sentential Task: NL Inference

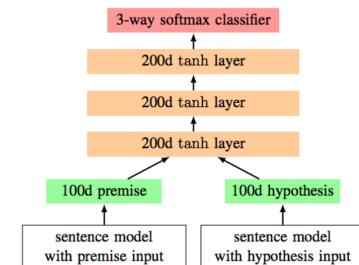
Premise		Hypothesis
A boy plays in the snow	<i>entails</i>	A boy is outside
A man inspects the uniform of a figure	<i>contradicts</i>	The man is sleeping
An older and younger man smiling	<i>neutral</i>	Two men are smiling and laughing at cats playing

- ▶ 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 / neutral / 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



Bowman et al. (2015)

## Visualizing RNNs

Slides from Greg Durrett / UT Austin



## LSTMs Can Model Length

- ▶ 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  $\backslash n$

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)



## LSTMs Can Model Long-Term Bits

- ▶ 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)



## LSTMs Can Model Stack Depth

- ▶ 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

```
#ifdef CONFIG_AUDITSYSCALL
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 & classes[class][i])
                return 0;
    }
    return 1;
}
```

Karpathy et al. (2015)



## LSTMs Can Be Completely Inscrutable

- ▶ 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

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

Karpathy et al. (2015)