

CMC MSU Department of Algorithmic Languages
Samsung Moscow Research Center

Neural Networks for Natural Language Processing

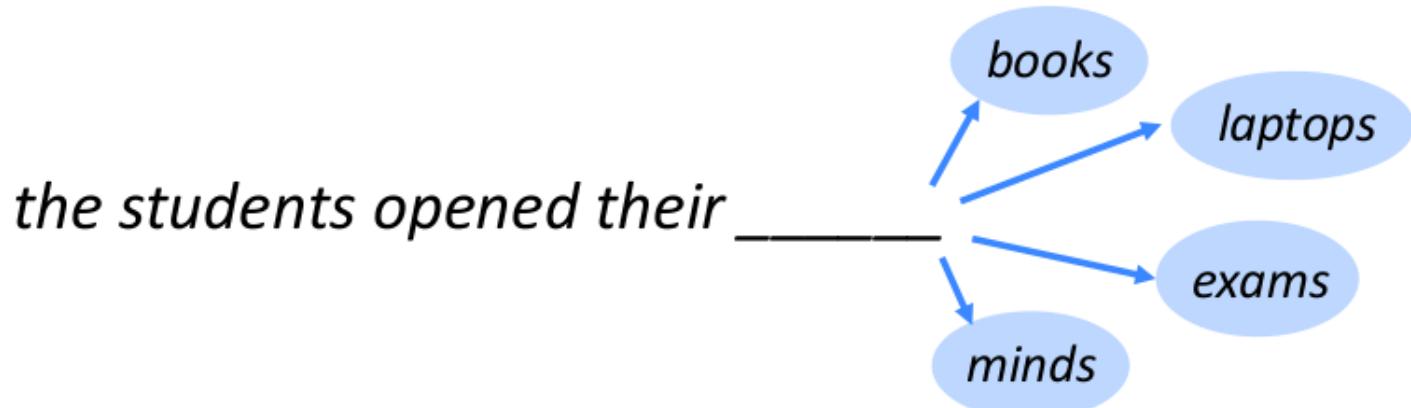
**Нейронные сети в задачах
автоматической обработки текстов**

*Block: Recurrent Neural Networks.
Lecture 1. Language Models & Recurrent NNs*

Arefyev Nikolay
*CMC MSU Department of Algorithmic Languages &
Samsung Moscow Research Center*

Language Modeling

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



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

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

Probabilistic Multiclass
Classifier with
Variable length input

where $\mathbf{x}^{(t+1)}$ can be any word in the vocabulary $V = \{\mathbf{w}_1, \dots, \mathbf{w}_{|V|}\}$

- 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 $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)}$, then the probability of this text (according to the Language Model) is:

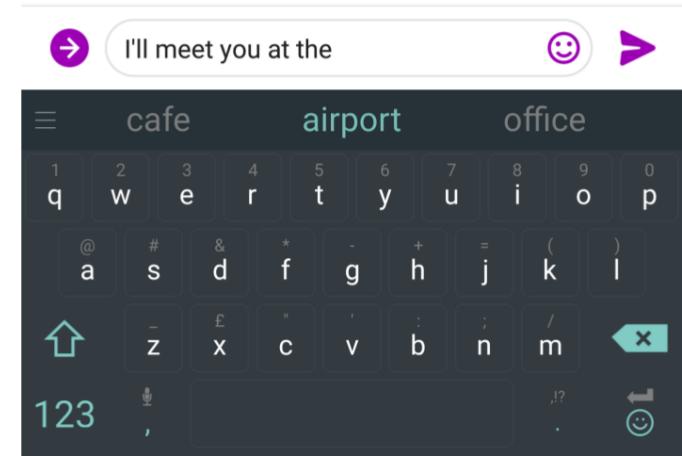
$$\begin{aligned} P(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)}) &= P(\mathbf{x}^{(1)}) \times P(\mathbf{x}^{(2)} | \mathbf{x}^{(1)}) \times \cdots \times P(\mathbf{x}^{(T)} | \mathbf{x}^{(T-1)}, \dots, \mathbf{x}^{(1)}) \\ &= \prod_{t=1}^T P(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}, \dots, \mathbf{x}^{(1)}) \end{aligned}$$



This is what our LM provides

Language Models are useful for

- Estimate [conditional] probability of a sequence $P(x)$, $P(x|s)$
 - Ranking hypothesis
 - Speech Recognition
 - Machine Translation
- Generate texts from $P(X)$, $P(X|s)$
 - Autocomplete / autoreply
 - Generate translation / image caption
 - Neural poetry
- **Unsupervised Pretraining**



n-gram Language Models

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

$$P(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(1)}) = P(\mathbf{x}^{(t+1)} | \underbrace{\mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)}}_{n-1 \text{ words}}) \quad (\text{assumption})$$

prob of a n-gram $\rightarrow P(\mathbf{x}^{(t+1)}, \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})$ (definition of conditional prob)

prob of a (n-1)-gram $\rightarrow P(\mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})$

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

$$\approx \frac{\text{count}(\mathbf{x}^{(t+1)}, \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})}{\text{count}(\mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})} \quad (\text{statistical approximation})$$

n-gram Language Models: Example

Suppose we are learning a 4-gram Language Model.

~~as the proctor started the clock, the students opened their~~ _____
discard condition on this

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

For example, suppose that in the corpus:

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

Should we have
discarded the
“proctor” context?

Problems of n-gram LMs

- Small fixed-size context
 - $n > 5$ cannot be used in practice
- Lots of storage space to keep n-gram counts
- Sparsity of data
 - Most ngrams (both probable and improbable) never occur even in very large train corpus
=> cannot compare them
 - The cat caught a frog on Monday → The kitten will catch a toad/*house on Friday
 - Tezguino is an alcoholic beverage. It is made from corn and consumed during festivals. Tezguino makes us _

$\hat{y}^{(4)} = P(\mathbf{x}^{(5)} | \text{the students opened their})$

A RNN Language Model

Hypothesis

output distribution

$$\hat{y}^{(t)} = \text{softmax}(\mathbf{U}\mathbf{h}^{(t)} + \mathbf{b}_2) \in \mathbb{R}^{|V|}$$

hidden states

$$\mathbf{h}^{(t)} = \sigma(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_e \mathbf{e}^{(t)} + \mathbf{b}_1)$$

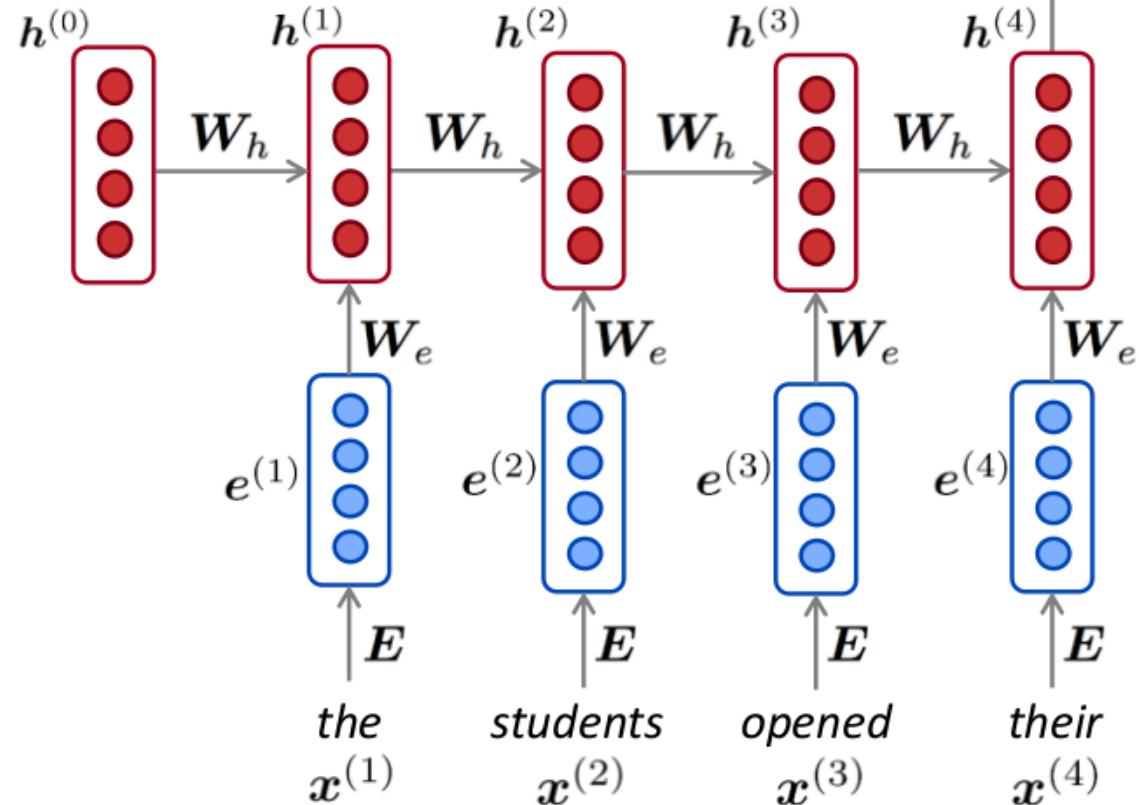
$\mathbf{h}^{(0)}$ is the initial hidden state

word embeddings

$$\mathbf{e}^{(t)} = \mathbf{E}\mathbf{x}^{(t)}$$

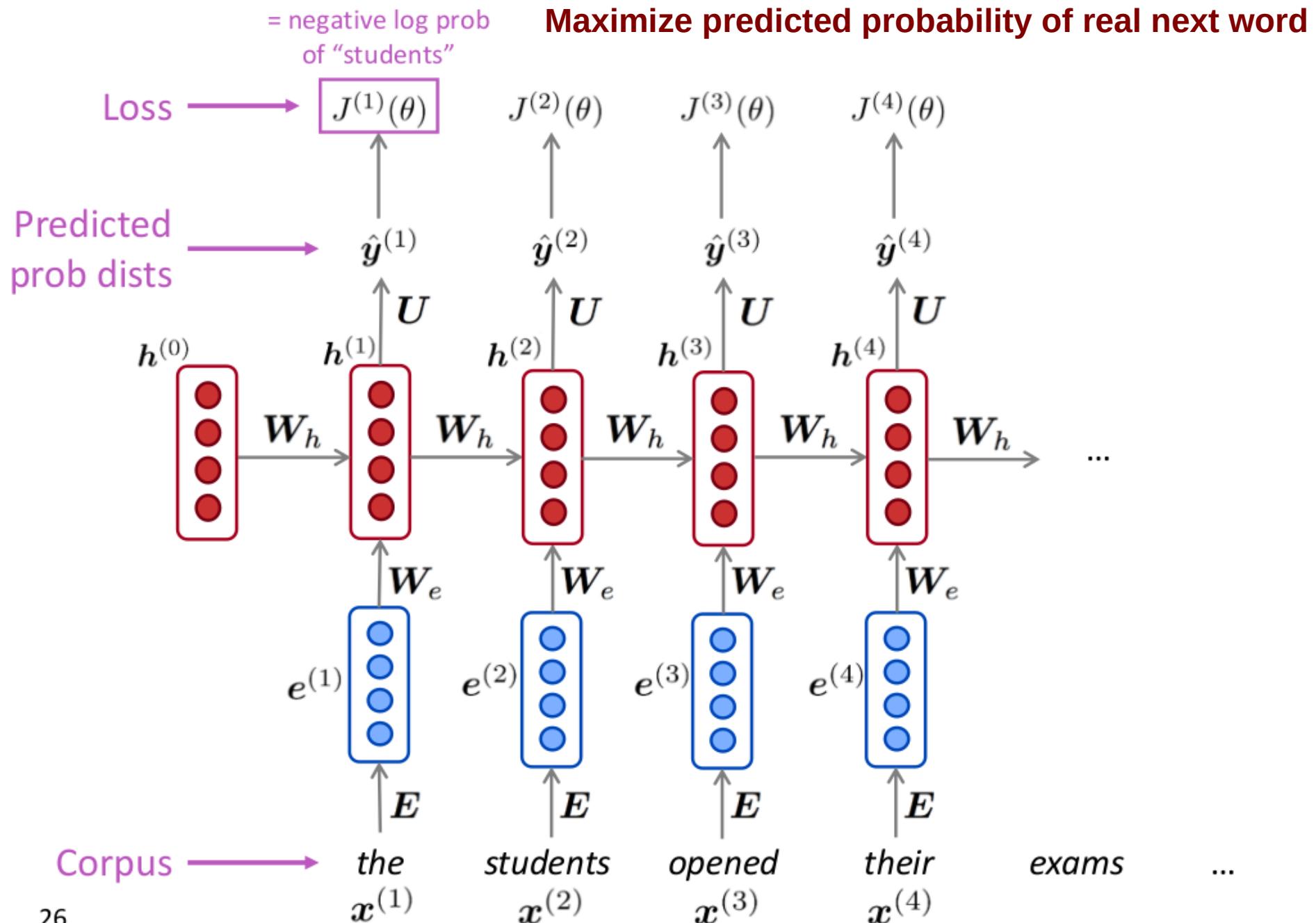
words / one-hot vectors

$$\mathbf{x}^{(t)} \in \mathbb{R}^{|V|}$$

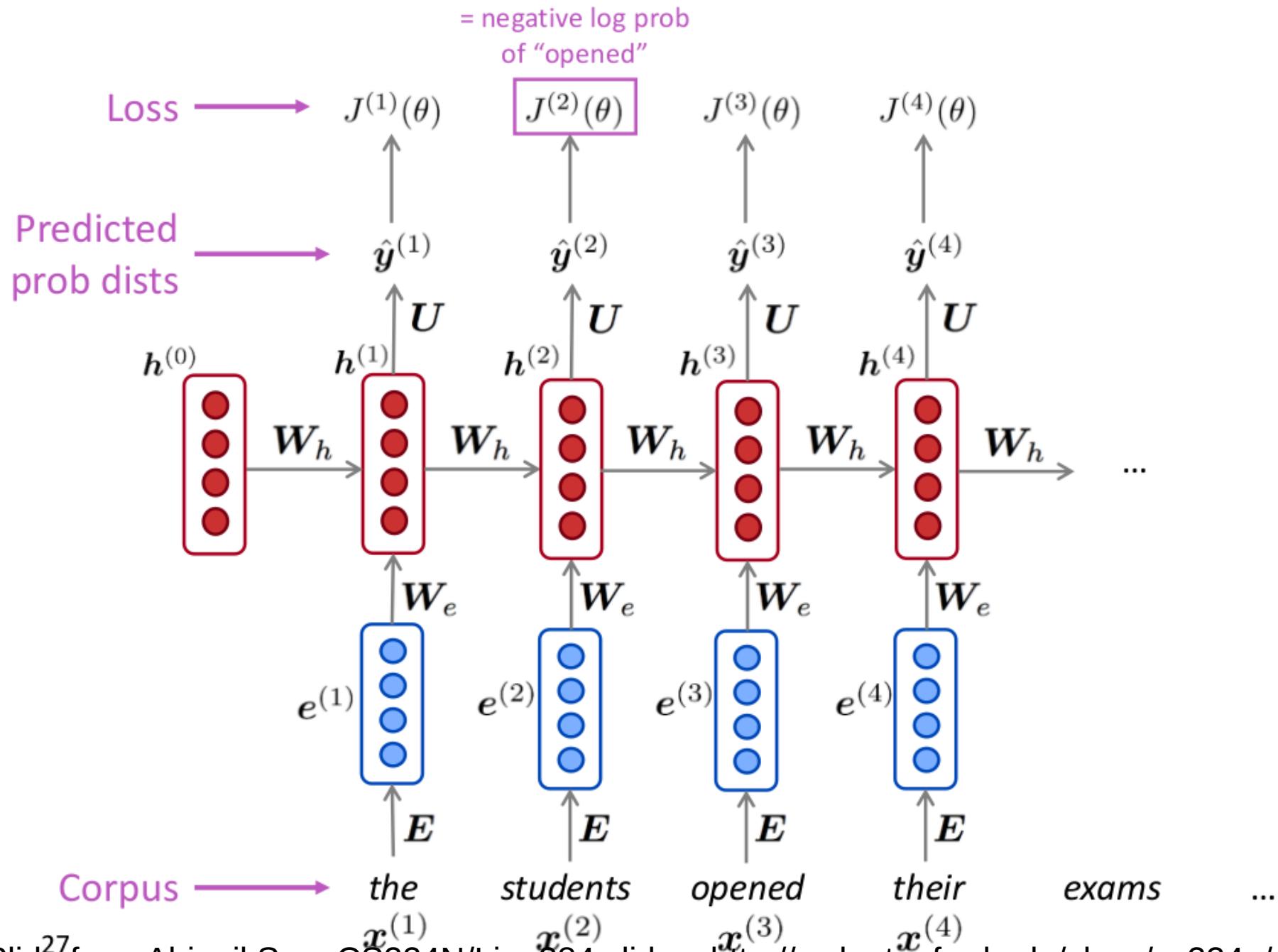


Note: this input sequence could be much longer, but this slide doesn't have space!

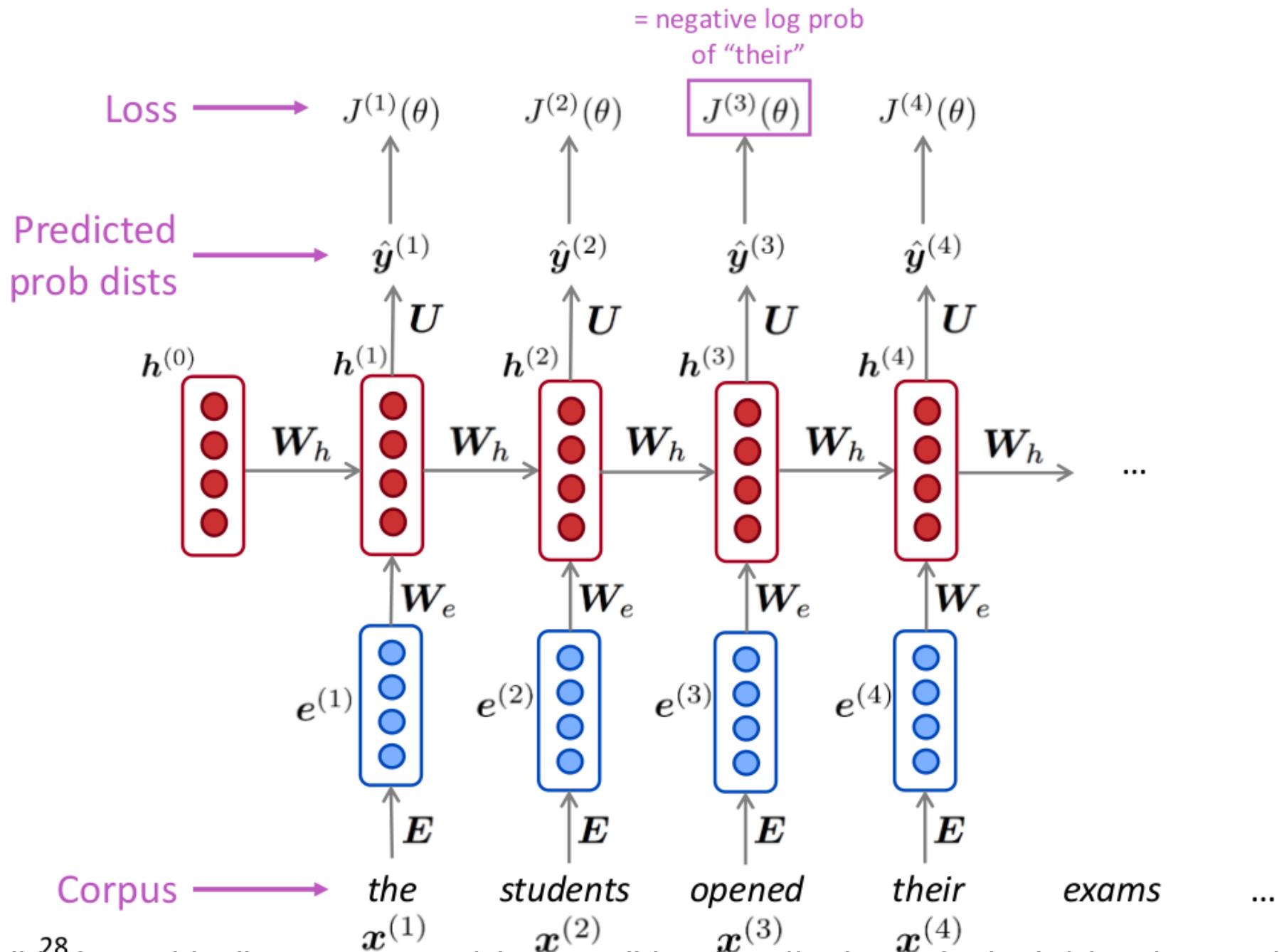
Training a RNN Language Model



Training a RNN Language Model



Training a RNN Language Model



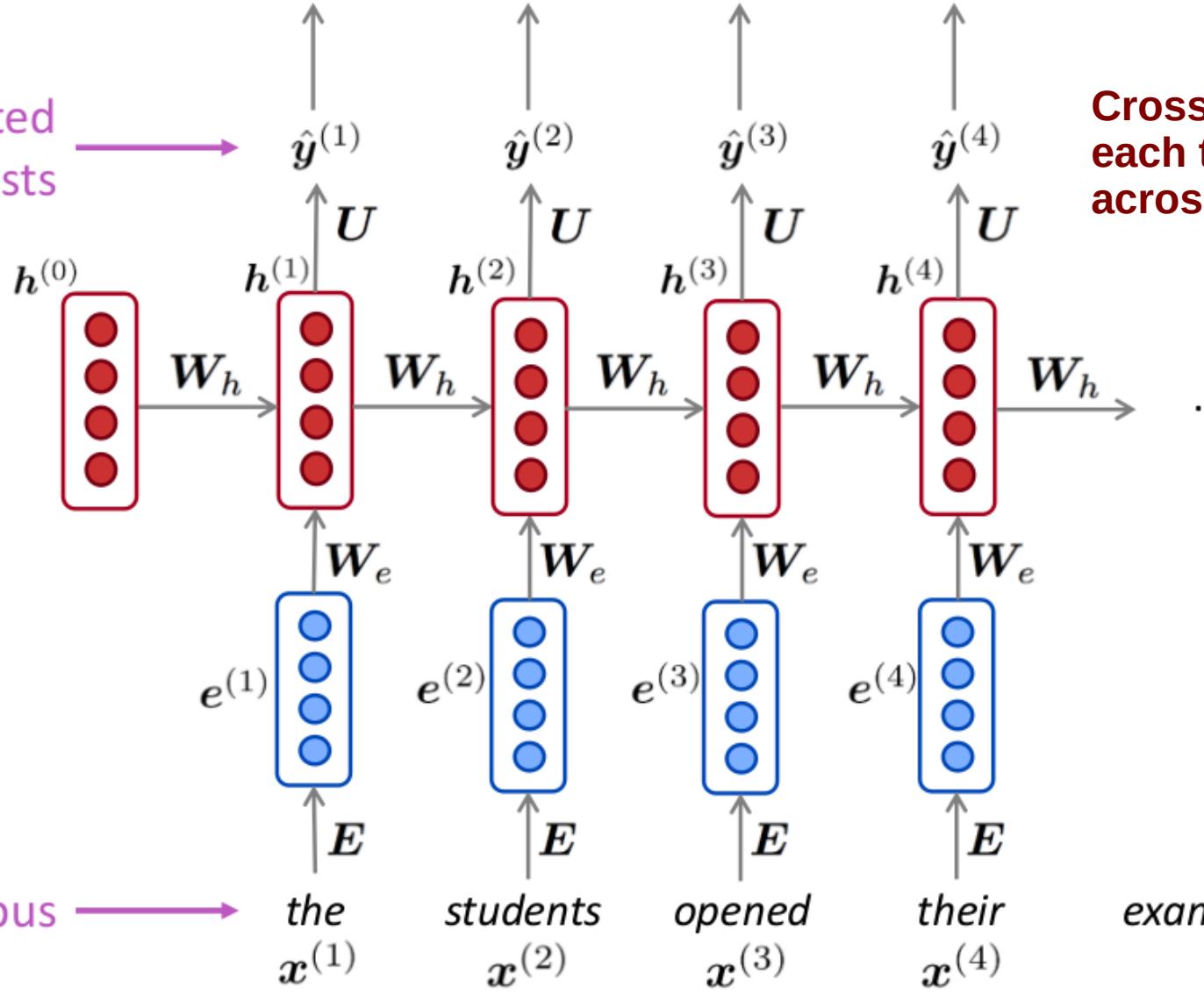
Training a RNN Language Model

Loss

$$\text{Loss} \longrightarrow J^{(1)}(\theta) + J^{(2)}(\theta) + J^{(3)}(\theta) + J^{(4)}(\theta) + \dots = J(\theta) = \frac{1}{T} \sum_{t=1}^T J^{(t)}(\theta)$$

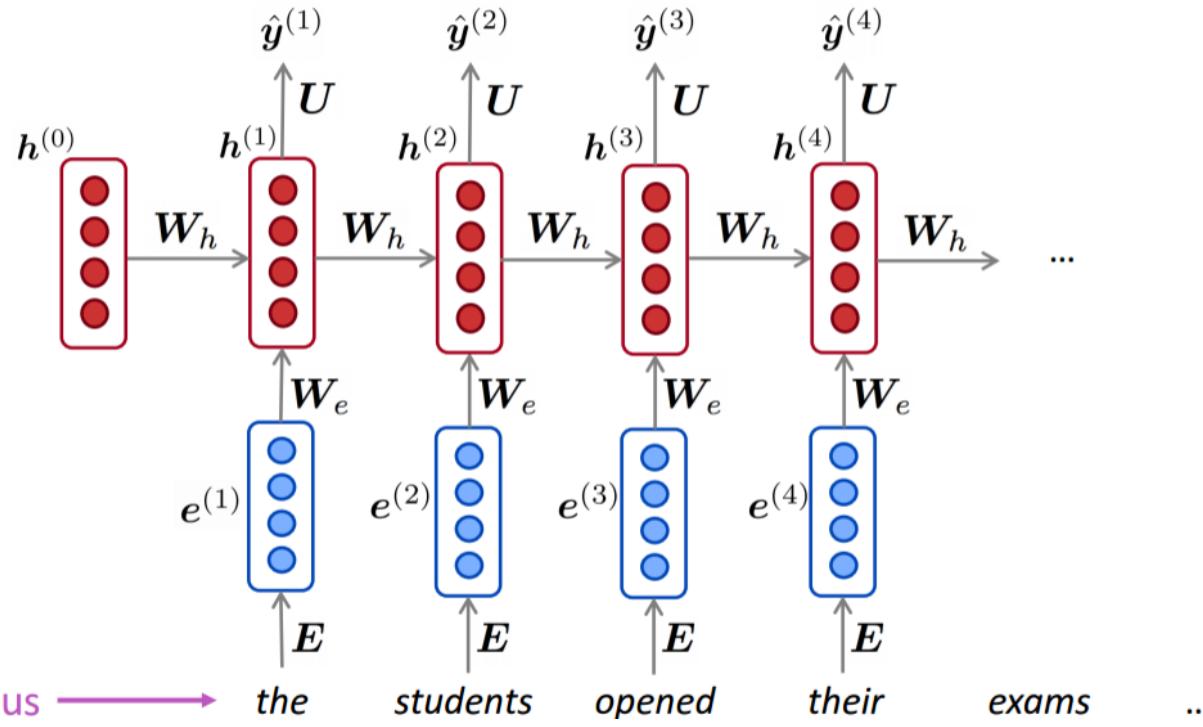
Predicted
prob dists

Cross-entropy loss on
each timestep \rightarrow average
across timesteps



RNN LM summary

Loss $\longrightarrow J^{(1)}(\theta) + J^{(2)}(\theta) + J^{(3)}(\theta) + J^{(4)}(\theta) + \dots = J(\theta) = \frac{1}{T} \sum_{t=1}^T J^{(t)}(\theta)$



output distribution

$$\hat{y}^{(t)} = \text{softmax} (\mathbf{U} \mathbf{h}^{(t)} + \mathbf{b}_2) \in \mathbb{R}^{|V|}$$

hidden states

$$\mathbf{h}^{(t)} = \sigma (\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_e \mathbf{e}^{(t)} + \mathbf{b}_1)$$

$\mathbf{h}^{(0)}$ is the initial hidden state

word embeddings

$$\mathbf{e}^{(t)} = \mathbf{E} \mathbf{x}^{(t)}$$

words / one-hot vectors

$$\mathbf{x}^{(t)} \in \mathbb{R}^{|V|}$$

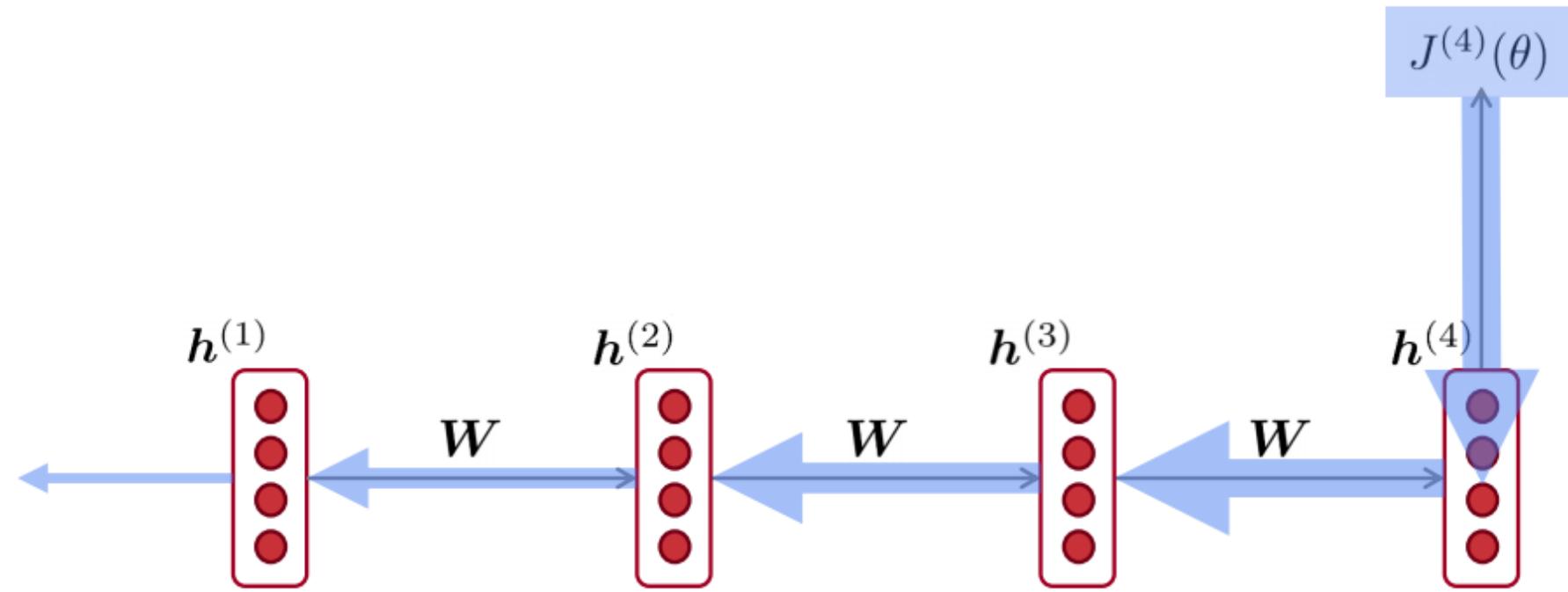
Optimization

- Loss is differentiable w.r.t. parameters of hypothesis (U, E, W_e, W_h)
 - => use backprop+SGD
 - BPTT – backpropagation through time
Similar to FFNN (#layers = #words) with shared weights (same weights in all layers)
- Truncated BPTT is used in practice
 - Forward-backward pass on segments of seqlen (50-500) words
 - Little better to use final hidden state from the previous segment as initial hidden state for the next segment (0 for the first segment)

Vanilla RNNs problems

- Vanishing/exploding gradients
 - similar to what we saw in FFNNs with incorrect initialization
 - but RNNs very deep (in time dimension) and existing initialization methods don't solve problems completely
 - Look at [Pascanu et al. On the difficulty of training recurrent neural networks, 2013] to really understand!
 - Solutions:
 - Vanishing gradients:
LSTM/GRU cells
 - Exploding gradients:
Gradient norm clipping
-
- Algorithm 1** Pseudo-code for norm clipping
-
- ```
 $\hat{g} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
 if $\|\hat{g}\| \geq threshold$ then
 $\hat{g} \leftarrow \frac{threshold}{\|\hat{g}\|} \hat{g}$
 end if
```
-

# Vanishing gradient intuition



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

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

- Recall:  $\mathbf{h}^{(t)} = \sigma(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)} + \mathbf{b}_1)$
- Therefore:  $\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(t-1)}} = \text{diag}\left(\sigma'\left(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)} + \mathbf{b}_1\right)\right) \mathbf{W}_h$  (chain rule)
- Consider the gradient of the loss  $J^{(i)}(\theta)$  on step  $i$ , with respect to the hidden state  $\mathbf{h}^{(j)}$  on some previous step  $j$ .

$$\frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(j)}} = \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(i)}} \prod_{j < t \leq i} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(t-1)}} \quad (\text{chain rule})$$

$$= \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(i)}} \boxed{\mathbf{W}_h^{(i-j)}} \prod_{j < t \leq i} \text{diag}\left(\sigma'\left(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)} + \mathbf{b}_1\right)\right) \quad (\text{value of } \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(t-1)}})$$

If  $\mathbf{W}_h$  is small, then this term gets vanishingly small as  $i$  and  $j$  get further apart

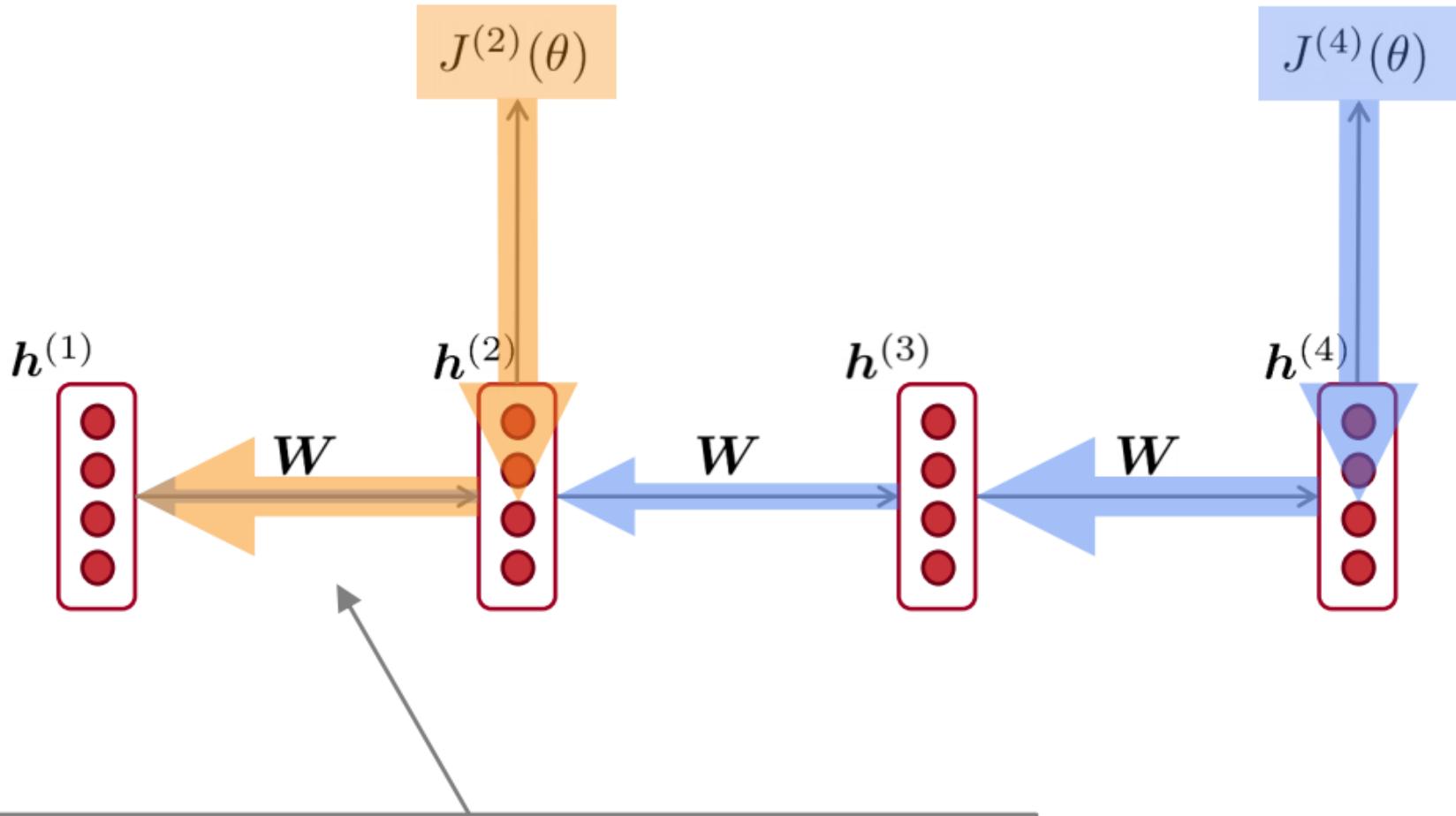
# Vanishing gradient proof sketch

- Consider matrix L2 norms:

$$\left\| \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(j)}} \right\| \leq \left\| \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(i)}} \right\| \|\mathbf{W}_h\|^{(i-j)} \prod_{j < t \leq i} \left\| \text{diag} \left( \sigma' \left( \mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)} + \mathbf{b}_1 \right) \right) \right\|$$

- Pascanu et al showed that if the **largest eigenvalue** of  $\mathbf{W}_h$  is **less than 1**, then the gradient  $\left\| \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(j)}} \right\|$  will **shrink** exponentially
  - Here the bound is 1 because we have sigmoid nonlinearity
- There's a similar proof relating a **largest eigenvalue > 1** to **exploding gradients**

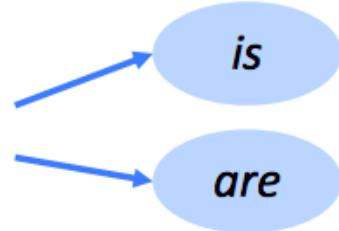
# Why is vanishing gradient a problem?



Gradient signal from faraway is lost because it's much smaller than gradient signal from close-by.

So model weights are only updated only with respect to near effects, not long-term effects.

# Effect of vanishing gradient on RNN-LM

- LM task: *The writer of the books*   
- Correct answer: *The writer of the books is planning a sequel*
- Syntactic recency: *The writer of the books is* (correct)
- Sequential recency: *The writer of the books are* (incorrect)
- Due to vanishing gradient, RNN-LMs are better at learning from sequential recency than syntactic recency, so they make this type of error more often than we'd like [Linzen et al 2016]

# LSTM – long short-term memory

- Solves (partially) vanishing gradient problem of vanilla RNNs  
=> better models long range dependencies

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1})$$

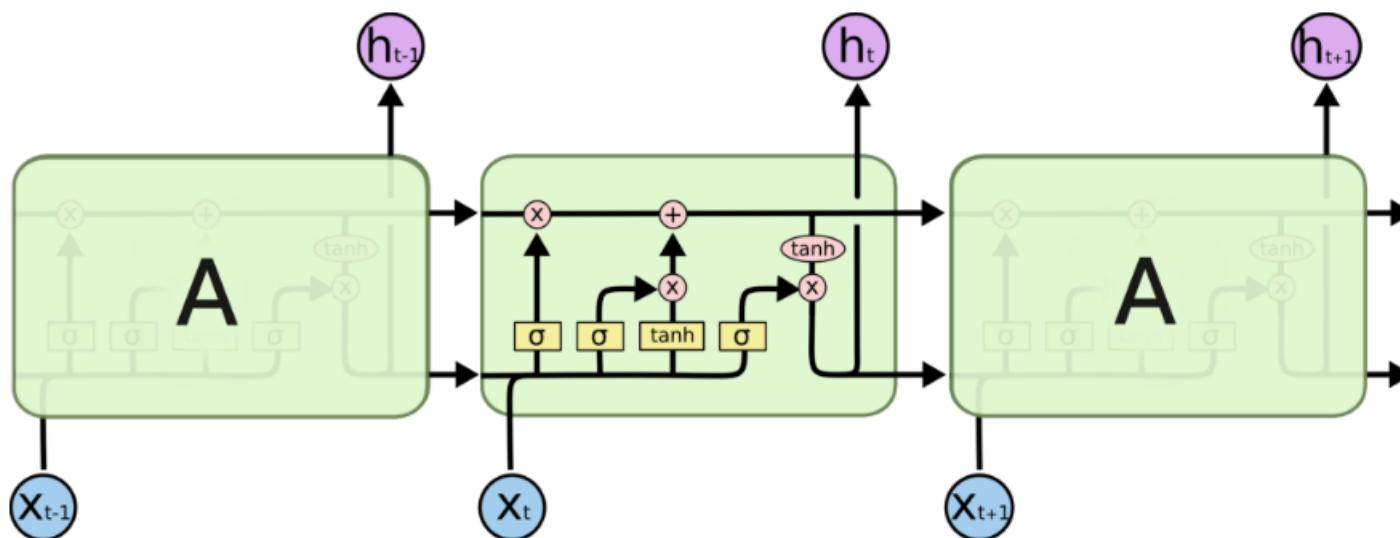
$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1})$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1})$$

$$\tilde{c}_t = \tanh(W^{(c)}x_t + U^{(c)}h_{t-1})$$

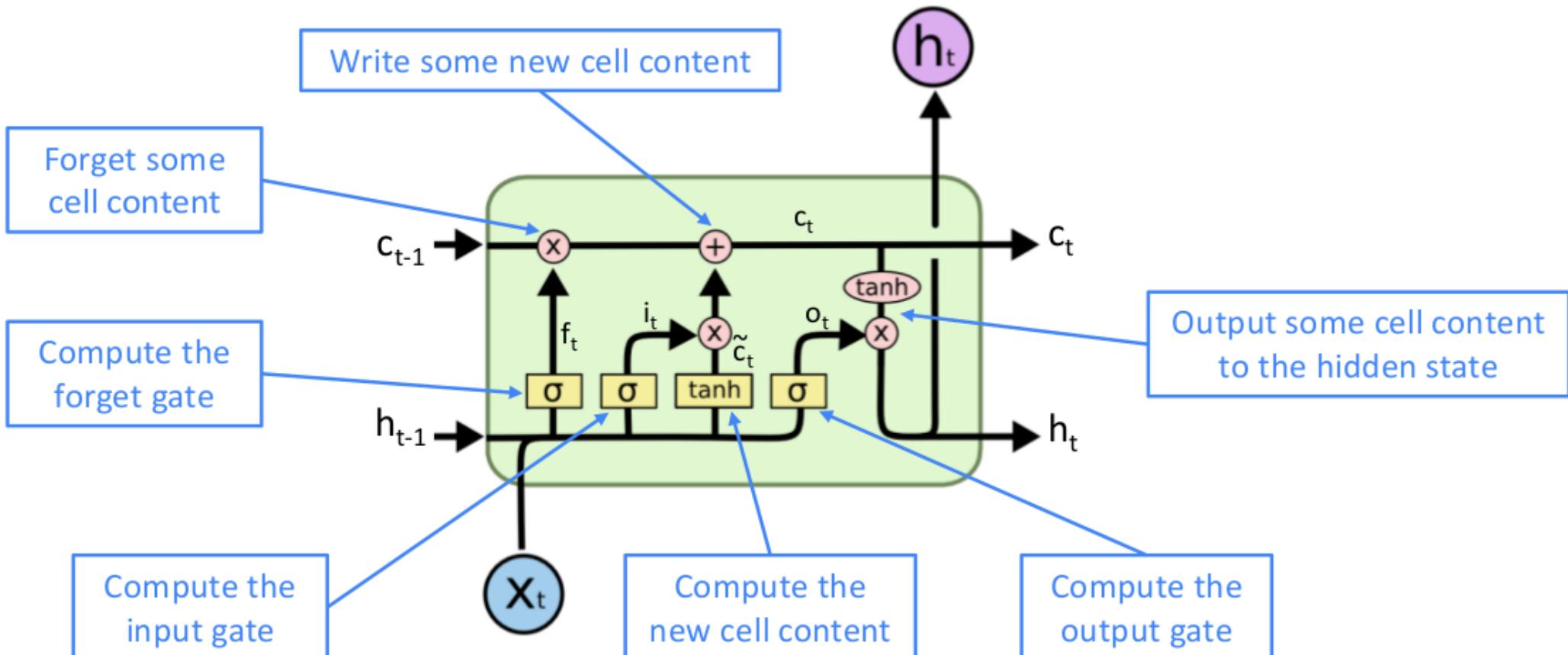
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$$

$$h_t = o_t \circ \tanh(c_t)$$



# Long Short-Term Memory (LSTM)

You can think of the LSTM equations visually like this:



# How does LSTM solve vanishing gradients?

- The LSTM architecture makes it **easier** for the RNN to **preserve information over many timesteps**
  - e.g. if the forget gate is set to remember everything on every timestep, then the info in the cell is preserved indefinitely
  - By contrast, it's harder for vanilla RNN to learn a recurrent weight matrix  $W_h$  that preserves info in hidden state
- LSTM doesn't *guarantee* that there is no vanishing/exploding gradient, but it does provide an easier way for the model to learn long-distance dependencies

# RNN/LSTM LM for text generation

$$P(y) = P(y_1) P(y_2|y_1) P(y_3|y_1, y_2) \dots P(y_T|y_1, \dots, y_{T-1})$$

- estimate  $P(y)$
- or generate new  $y$ :

$y_1$  = random word / <START>

for  $i$  from 2 to  $T$ :

$y_i = \operatorname{argmax}_y P(y|y_1 \dots y_{i-1})$  / sample from  $P(y|y_1 \dots y_{i-1})$

# Examples generated (Shakespeare)

PANDARUS:

Alas, I think he shall be come approached and the day  
When little strain would be attain'd into being never fed,  
And who is but a chain and subjects of his death,  
I should not sleep.

Second Senator:

They are away this miseries, produced upon my soul,  
Breaking and strongly should be buried, when I perish  
The earth and thoughts of many states.

DUKE VINCENTIO:

Well, your wit is in the care of side and that.

Second Lord:

They would be ruled after this chamber, and  
my fair nues begun out of the fact, to be conveyed,  
Whose noble souls I'll have the heart of the wars.

Clown:

Come, sir, I will make did behold your worship.

VIOLA:

I'll drink it.

# Examples generated (Linux kernel)

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
 int error;
 if (fd == MARN_EPT) {
 /*
 * The kernel blank will coeld it to userspace.
 */
 if (ss->segment < mem_total)
 unblock_graph_and_set_blocked();
 else
 ret = 1;
 goto bail;
 }
 segaddr = in_SB(in.addr);
 selector = seg / 16;
 setup_works = true;
 for (i = 0; i < blocks; i++) {
 seq = buf[i++];
 bpf = bd->bd.next + i * search;
 if (fd) {
 current = blocked;
 }
 }
 rw->name = "Getjbbregs";
 bprm_self_clearl(&iv->version);
 regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
 return segtable;
}
```

Figure from **Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks** <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

# AWD-LSTM for Russian

Trained on 200M tokens of news texts for 10 epochs

- <https://github.com/mamamot/Russian-ULMFit>

Sampled using:

- Ancestral sampling
- temperature softmax:  $\text{softmax}(\text{logits}/T)$

# T=1.0

в Минеральных водах во время обеспечить безопасность на Составлен акт , вызвавший Умысел - потерпевшего в метро , сообщил мэр Москвы Сергей Городских . Об этом сообщает « Национальная служба новостей » . По словам Посыльным , только на первая расшифровывать этих полос они калитки около полутора тысяч человек . Также в ходе работ возможно будет набрал детский сад , куда частично привезли детского сада , где взяты дети . Прокурор сказал , что теперь в программе « Испания и родители » останутся почти 105 детей .

**Я хочу** набрасывают в Петербурге , что вначале , конечно , поможет экономический и , главное , домициане ситуацию . Могло бы , военный бюджет всерьез успокоить к нам сразу же принял нашу " премьер - министр " - Олег Владимир Месяц , который едет на встречу с Владимиром Путиным . я думаю , что она обязательно будет свидетелем : в пригороде Брюсселя Брюсселя как самой Греции в чистом виде ( на их мероприятия Занимал неизвестные Неплохая и футбольного клуба ФК " Москва " - тоже Россия

Этот проект прошедшего года становится уже вторым не только только с точки зрения проведения проекта ? Для того чтобы провести компания 24 сентября , нужно арабского языка или русского языка . Но идея сунны написали почти все активисты российских оппозиционных партий . Почему - то и проходил Патриотический фестиваль специализацию и этом мусульмане проводят у себя определенные предприятия и митинги . в штаб вынуждены выла уже профессиональные , совмещена , что выбрало себе стоя и знать о нем , но еще как раз – на будущие даты и дни , когда будут заниматься этим вопросом .

**Смысл жизни состоит в** конкретной том , кто из объектов знает , что тепло позаимствовать . Из этого программа , расположенный на Малой Морской улице Собрала , ресторане , после чего опять обязана установить фуэте . Японские специалисты сегодня пытаются избавиться от конструкций и использовать для моего решения такую же работу . Решительность и подработать начали патриотического найти в Замаскирован Полагается . к 30 июня посетители увидят в основном закладки к популярны предсказание самовыражения ярче и духовной войны . Возможно , он андреевны на первый взгляд , и он будет наблюдать другие очевидцы недавнего убийства

**Московский Государственный Университет** культуры и искусств ( ПРЕИМУЩЕСТВЕННО ) при музее - музей « Евгений Невский и Гитлер » стал выставка . Деньги должны быть отправлены на определение новейших художников с Новым годом в Музее современного искусства Петербурга . « 52 процента экспозиции нам показали уже сегодня » , - сообщил главный художник выставки Дмитрий Исламский в своем твиттере . Сообщество принял решение об открытии коллекции . Посетитель университета продолжил работать , а и работа чья - то скульптура , на которую собирается обрушился , будет признана фотографии английского художника

T=0.1

**Я хочу** , чтобы Россия стала членом Европейского союза , а не России . Это не только Россия , но и Украина . Мы должны быть готовы к тому , чтобы Украина не могла быть членом Европейского союза . Но мы не можем позволить себе это сделать , потому что мы не можем позволить себе это делать , потому что мы не можем позволить себе это делать » , - заявил Владимир Владимирович Владимир Владимирович Владимир Владимирович . Он также отметил , что Россия

**Я хочу** , чтобы Россия стала членом Европейского союза . Это не только Россия , но и Украина , и Россия , и Украина . Мы должны быть готовы к тому , чтобы Украина и Россия договорились о создании Единого экономического пространства , который будет способствовать развитию отношений между Россией

**Этот** случай произошел в Петербурге . Как сообщили « Фонтанке » в пресс - службе МО « Гражданка » , в Невском районе на улице Маршала Жукова , дом 2 , в Невском районе , в районе дома № 2 , корпус 1 , по улице Маршала Жукова , в районе дома № 1 , произошел взрыв . По предварительной информации , в результате происшествия никто не пострадал . По предварительной информации , в результате происшествия никто не пострадал . xxbos На Украине в результате пожара

**Этот** проект , который будет реализован в Петербурге , будет реализован в рамках проекта « Новая Москва » . Об этом « Фонтанке » сообщили в пресс - службе компании . По словам представителей компании , в рамках проекта будет построен комплекс « Северная долина » , который будет построен на месте бывшего « Дома Романовых » . По словам представителей компании , проект будет реализован в рамках проекта « Новая Земля » . По словам генерального директора ФК « Санкт - Петербург »

**Смысл жизни состоит** в том , что в России не существует ни одного государства , который бы не мог быть членом Совета Европы . Об этом заявил президент России Владимир Владимирович , сообщает ТАСС . По словам Владимира Иванова , в России есть и другие страны , которые не имеют права на существование . « Мы не можем не допустить , чтобы

**Смысл жизни состоит** в том , что в России нет ни одного государства , который бы не мог быть членом Совета Безопасности РФ . Об этом заявил в пятницу на пресс - конференции в Москве глава Российского союза журналистов Сергей Миронов , передает РБК . По словам Иванова , в России существует " очень много " людей , которые не могут быть в России . По словам Иванова , в России есть и другие страны , которые не могут быть в состоянии войны . По словам

**Московский Государственный Университет** имени Ломоносова в Москве в среду , 13 июня , открыл в Москве музей Пушкина . Об этом сообщает ТАСС .  
Музей откроется в Москве в субботу , 25 июня . Музей откроется в Москве в субботу , 27 мая . Музей откроется в Москве в субботу , 26 июня . Музей откроется в Москве в начале июня . Музей будет открыт в Москве в конце мая . Музей будет открыт в Москве в конце июня . Музей будет открыт в

**Московский Государственный Университет** имени Ломоносова в Москве в пятницу , 19 мая , объявил о начале работ по реконструкции здания Академии художеств , сообщает ТАСС . Работы по реконструкции здания , который будет построен в Москве , будут проходить в рамках программы « Развитие Москвы » . На территории комплекса будут построены три новых здания , а также два здания , которые будут построены в Москве . На месте будущего здания будет построен комплекс зданий , в том числе и здание Академии художеств . На месте здания будет построен музей

# T=0.6

Американский бизнесмен и один из основателей компании Apple Стив Стив ( Ближнего Востока ) уверен , что Google не будет отстаивать свои права на существование . Об этом он сообщил на интервью в Facebook . « Мы не можем быть и не готовы к сотрудничеству с Google . Это не будет просто Google » , — сказал Google . Он добавил , что они не могут быть заинтересованы в том , чтобы Google был единственным независимым интернет - изданием в Европе , а

На Украине идет процесс над Александром Шевченко , которого обвиняют в убийстве российского журналиста Георгия Будут . Об этом сообщает " Новости - Украина " . Ранее сообщалось , что Виктор Бут находится в России . По версии следствия , По данным следствия , в Киеве члены его семьи , а также сын Виктора Бут , Александр Бут и Виктор Бут , были убиты в результате нападения на Россию . По версии следствия , в ночь

**Я хочу** отметить , что в России существует некая крупная " группа " " Северного Кавказа " , которую возглавляет Борис Гребенщиков . Он был создан в 1996 году для поддержки борьбы с организованной преступности , которую возглавлял Николай Лебедев . Как пишет газета " Ведомости " , это связано с тем , что Михаил Южный считается одним из крупнейших в России криминальных авторитетов . Он является одним из лидеров " Правого дела " . По данным " Независимой газеты " ,

**Я хочу** поблагодарить российского президента Владимира Владимира Владимира за поддержку и поддержку , при этом он выразил уверенность , что Россия будет продолжать сотрудничество с Россией по вопросам взаимодействия в сфере безопасности и развития ее экономики . Об этом он заявил в интервью газете « Известия » . « Мы будем работать по всей стране , в том числе в России , на Украине , в России и в России . Мы хотим , чтобы Россия , как и Украина , была на грани войны

**Этот** пост занял Дмитрий Долго . Он стал первым вице - президентом по продажам и продажам в России , где в этом году работал его вице - президент . Как сообщается на сайте Ассоциации производителей алкогольной продукции ( ГАЗ ) , в состав правления войдут четыре директора . До этого Герман Иванов возглавлял Службу по контролю за исполнением законов о государственной службе , а также Департамент государственного контроля и контроля за исполнением Государственной Думы РФ . Кроме того , Дмитрий Попов был назначен

**Этот** случай , когда в Петербурге произошло ПРОИСШЕСТВИЕ с участием двух машин , произошло в ночь на 28 июня . Как сообщили в УПРАВЛЕНИИ ПОЛИЦИИ по Петербургу и Ленинградской области , авария произошла около половины пятого вечера у дома 11 по Октябрьской набережной . На Ленинском проспекте СТОЛКНУЛИСЬ с участием двух автомобилей , двух И Ford 200 , а также Ford Explorer и MAN . За прошедшие сутки на Петербурге и Ленинградской области произошло пять ПО , в результате

**Смысл жизни состоит** в том , что в России можно сделать ставку на покупку жилья . Как сообщил в среду РБК генеральный директор Агентства по управлению и управлению в сфере ГО и СТРОИТЕЛЬСТВА Владимир Васильев , это связано с тем , что на территории России есть и другие виды жилья , в том числе и в Москве . " Такие квартиры в Москве или Петербурге являются одной из самых дорогих квартир в России " , - заключил Юрий Михайлов . По его словам , в целом в

**Смысл жизни состоит** в том , что каждый человек сможет самостоятельно выбрать себе человека , но , как считают некоторые , не только его , но и сами же люди с очень большим количеством сознания . Об этом , как сообщает BBC News , заявил Джордж Смит , автор песен , а также его жена Кейт Уильямс . По словам Джеймса Дэвиса , все зависит от того , как станет выглядеть голос . По словам Майкла , он является одним из самых ярких друзей Джона Кеннеди . Он также

**Московский Государственный Университет** имени Ломоносова и Национальный университет имени Ломоносова в Москве должны будут провести обучение студентов в возрасте от 18 до 24 лет . Об этом сообщает ТАСС со на пресс - конференции , организованной Министерством образования и науки РФ . По словам Владимира Васильева , страна будет работать с МГУ , Московским государственным институтом экономики и Российской академии наук . На данный момент в России насчитывается около 20 тысяч студентов , которые работают в России , в том числе в России ,

**Московский Государственный Университет** Москвы ( МГУ ) взял на себя обязательства по выполнению предписания Федерального агентства по образованию об образовании и образовании , которое было опубликовано на сайте ЕГЭ по русскому языку . Об этом сообщается на сайте Союза . По словам представителя Московского государственного университета Ольги Иванова , поводом для заявления стал отказ Московской консерватории от проведения Всероссийского конкурса студентов , который пройдет в декабре . По словам Васильева , в этом году Академию общего образования МГУ может принять участие только один из семи университетов ,

# Russian ELMo

#TODO: anybody wants to generate similar examples with Russian ELMo?

- 1M word vocabulary (vs. 60K in Russian AWD-LSTM)
- May generate better texts than AWD-LSTM (based on our experiments, gives better substitutes for WSI).

# AWD-LSTM on school essays



08.11 // 11:45 - 14:00 (мск). AIJ\_Трансляция из Зала 4



Копиров

## Цитаты великих нейросетей



“Кому на Руси жить хорошо? Кто такой Алексей Толстой?  
Я всегда верил, что Русь теперь уже обладает какой - нибудь другой  
теорией.”

“Драма «Горе от ума» — настоящее произведение Пушкина. Это  
«Евгений Онегин». Главный герой — Евгений Онегин. Этот факт не  
вызывает иронии, он не является автором, но и сам главным героем  
поэмы. Конец сюжета поэмы — это произведение в полном их  
развитии.”

# RuGPT-3

GPT-3 is a language model, but not RNN-based. One of the latest, largest and ?best? language models.

RuGPT-3 is Russian analogue.

- GPT-3 architecture (Language Model, based on Transformer layers – often referred as Transformer decoder)
- 760M weights, context size – 2048 tokens
- 600GB / 80B tokens corpus: Russian and English Wikipedia, news and QA websites, Pikabu, banki.ru, 22centery.ru, etc. + Github and StackOverflow

# RuGPT-3

**Я хочу** пойти на войну, однако меня не отправят. Я их отправлю, когда доберусь до города. И напишу, что бросаю деревню и покупаю билет. Когда вернусь домой, там мой папа. Он только что проснулся.

- Ты не понимаешь, значит ты не хочешь идти воевать. У нас же братики и сестры. Ты видела, сколько их в школе.
- У нас всё равно нет места для них.
- Почему? Их же будет намного больше, если ты не пойдёшь в армию.
- Мама, ты же не бросишь меня дома, а я без тебя не смогу.
- Хорошо. Пусть. Только моё участие в войне я повторять не буду.

**Я хочу** рассказать историю, которая мне очень понравилась.

Сегодня прочитала об открытии сквера "Эмеральда". Пишут, что он построен в память о княгине Марии Петровне Долгорукой. Эмеральда - река в Забайкалье, правый приток реки Тыя, впадающий в Баргузинское море. Начиналось строительство этого парка так. Как рассказывала мне сестра, когда дедушка переехал, она очень любила, когда он приносил термос с чаем и пирожки с ливером, так она его называла. В его термосе был "вареньевский чай" и "по просьбам трудящихся" она собиралась на прогулку и носила его "на сохранение" - это были небольшие пирожные. В этих пирожных была чайная паста.

**Смысл жизни состоит** в том, чтобы найти его. Бывает так: в данный момент, в данной точке рождается это или нечто. Формируется смысл жизни. Тогда, решив, что этот смысл жизни все-таки есть, человек подходит к определенному сознанию и спрашивает его. Тогда он найдет решение, потому что всегда все может быть решимо. Когда человек спрашивает, у него, конечно, возникает вопрос: а что делать-то?

**Смысл жизни состоит** в том, чтобы научиться любить. Как полюбить? Вначале нужно понять, что ты хочешь от жизни, чего ты боишься, где твой предел возможного и снять груз сомнений и страх перед неизвестностью. И тогда появляются силы идти дальше. Раздел 2. Секрет долгого счастья 1. Ты не вещь, ты - нечто большее. 2. Главное в жизни это люди. 3. Перестань зависеть от мнения других людей. Ты должна думать прежде всего о себе и быть свободной для того, чтобы жить так, как считаешь правильным. Прощай обиды и гнев, прощай ненависть и желание подавить кого-то, прощай чувства, что ты ничто без них и возьми себя в руки иди дальше. Доведи своё счастье и любовь до желанной вершины. Чёрный и белый день подскажут, где ты будешь счастлива. А лестница к этому состоянию всегда в твоих руках. Мы все прошли тот путь, который отделяет наш обычный день от Жизни в Любви и Счастья. Начало обязательно будет, так что не откладывай его в долгий ящик, и поскорее начинай работать над собой.

**Московский Государственный Университет** в 1996 году, почётный доктор Российской Академии Архитектуры, Строительных наук и строительных технологий, профессор, Член-корреспондент Международной академии Архитектуры, Академик Российской Инженерной Академии, лауреат Государственной премии СССР. М. А. Ершов является автором памятника Председателю Правительства РФ (памятник установлен в 1998 году, расположенный в Москве, в соответствии с принятым постановлением Правительства Российской Федерации № 293 от 19 октября 1998 года). Этот памятник является выдающимся мемориальным и художественным объектом мирового значения.

**Московский Государственный Университет** Москва, Большой Власьевский переулок, 18 [e-mail:info@vsl.ru](mailto:info@vsl.ru) или [www.vsl.ru](http://www.vsl.ru) Телефон: (495) 691-29-35

<https://sbercloud.ru/ru/warp/gpt-3>

# RuGPT-3

**Мартышка в старости слаба глазами стала**, передвигается где попало – вот ее и носят туда-сюда, но она до сих пор хорошо видит, потому что к этому привыкла: хоть и видит сквозь стены, зато в основном по ночам. Потому что ей еще до пенсии долго. Не до пенсии еще, а до пенсии еще довоюет. Ей после войны все по сто раз в день повторяют: «Кто сказал, что ты старая?» – и не дают деньги за работу по дому. Чтоб хоть немного была похожа на мать: молодые все за стариками присматривают – и убирают в комнатах, и стирают.

**Однажды лебедь рак и щука** съели друг друга. Щука была очень красивой, а рак-бесчестным. Так говорят злые люди, считающие себя мудрее нас. Но мудрость не в том, чтобы обманывать ближнего, а в том, чтобы самому не попадаться в лживые сети. Остальное здесь

**Чем дальше в лес**, тем толще партизаны. Часы на первом этаже, может быть, и работают, но через окно здания – нет.

**Чем дальше в лес**, тем хуже партизанам. А было бы еще хуже. Температура моего тела чуть ли не зашкаливала за 37 градусов, несмотря на одевание легкого платья.

**Чем дальше в лес**, – сказал Боб, – тем больше дров". – Нам непременно надо быть разведчиками, – сказал мистер Ян. Конечно, мы разведчики, но мы – не Мам и Папа и даже не Боб. Мы не встречаемся по ночам на плантации, мы не болеем, потому что весь день проводим в соревнованиях по стрельбе и нам ничего не нужно делать, кроме как собрать пищу. Мы не клеим аэропланы. Мы питаемся одним лишь мякишем. Но мы любим странные вещи: тягучесть прохладной еды, паутинку рисунка одеяла, спичечный коробок, как будто бы вырванный с корнем, который всё равно обратно входит, – он ждёт не дождётся, чтобы с ним поиграли, – таинственную черноту дождя и запаха обожженной листвы, превращающую опилки, порошок и древесный мусор в уголь, – нас бы очень заинтересовали вот такие находки, – сказал Боб. Мы все трое оживились.

# Cells activations

Cell sensitive to position in line:

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.

Cell that turns on inside quotes:

"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."

# Cells are sometimes interpretable

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
 siginfo_t *info)
{
 int sig = next_signal(pending, mask);
 if (sig) {
 if (current->notifier) {
 if (sigismember(current->notifier_mask, sig)) {
 if (!(current->notifier)(current->notifier_data))) {
 clear_thread_flag(TIF_SIGPENDING);
 return 0;
 }
 }
 }
 collect_signal(sig, pending, info);
 }
 return sig;
}
```

Cell that turns on inside comments and quotes:

```
/* Duplicate LSM field information. The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
 struct audit_field *sf)
{
 int ret = 0;
 char *lsm_str;
 /* our own copy of lsm_str */
 lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
 if (unlikely(!lsm_str))
 return -ENOMEM;
 df->lsm_str = lsm_str;
 /* our own (refreshed) copy of lsm_rule */
 ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
 (void **) &df->lsm_rule);
```

Figure from **Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks** <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

# Sentiment Neuron Visualizations

- How sentiment neuron changes while reading text?

Once in a while you get amazed over how BAD a film can be, and how in the world anybody could raise money to make this kind of crap. There is absolutely No talent included in this film - from a crappy script, to a crappy story to crappy acting. Amazing...

Team Spirit is maybe made by the best intentions, but it misses the warmth of "All Stars" (1997) by Jean van de Velde. Most scenes are identic, just not that funny and not that well done. The actors repeat the same lines as in "All Stars" but without much feeling.

God bless Randy Quaid...his leachorous Cousin Eddie in Vacation and Christmas Vacation hilariously stole the show. He even made the awful Vegas Vacation at least worth a look. I will say that he tries hard in this made for TV sequel, but that the script is so NON funny that the movie never really gets anywhere. Quaid and the rest of the returning Vacation vets (including the orginal Audrey, Dana Barron) are wasted here. Even European Vacation's Eric Idle cannot save the show in a brief cameo.... Pathetic and sad...actually painful to watch...Christmas Vacation 2 is the worst of the Vacation franchise.

# Evaluating Language Models

- The standard **evaluation metric** for Language Models is **perplexity**.

$$\text{perplexity} = \prod_{t=1}^T \left( \frac{1}{P_{\text{LM}}(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(1)})} \right)^{1/T}$$

Inverse probability of corpus, according to Language Model

Normalized by  
number of words

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

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

**Lower perplexity is better!**

# Applications of Recurrent NNs

- 1 → 1: FFNN
- 1 → many: conditional generation (image captioning)
- many → 1: text classification
- many → many:
  - Non-aligned: sequence transduction (machine translation, summarization)
  - Aligned: sequence tagging (POS, NER, ...)

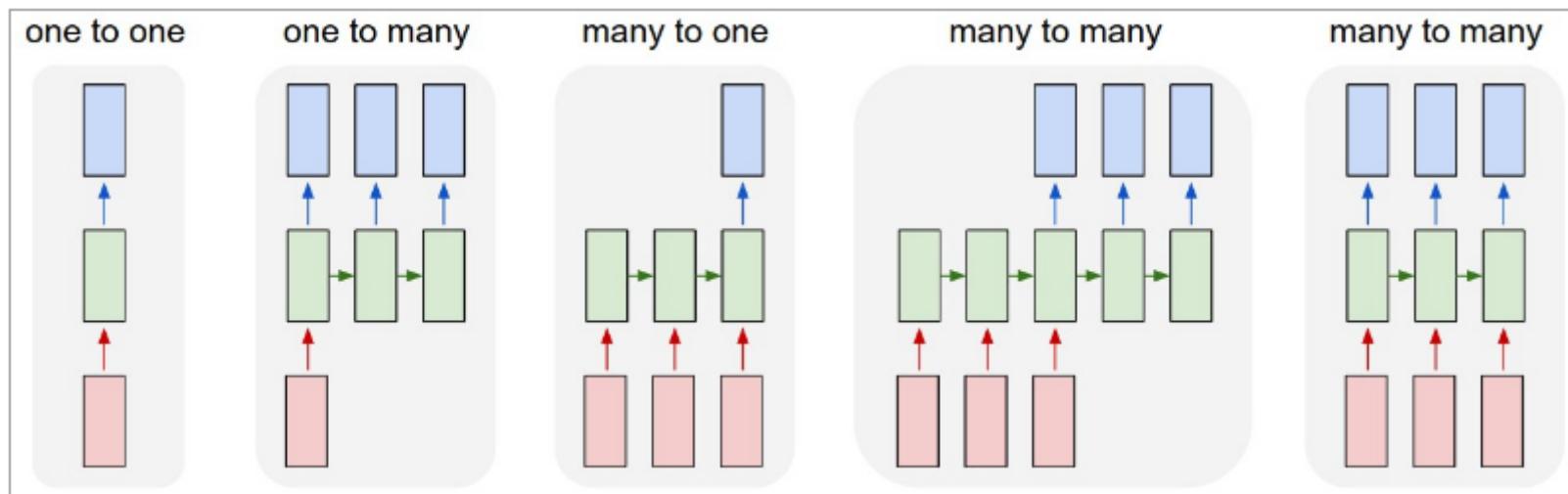


Figure from **Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks** <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

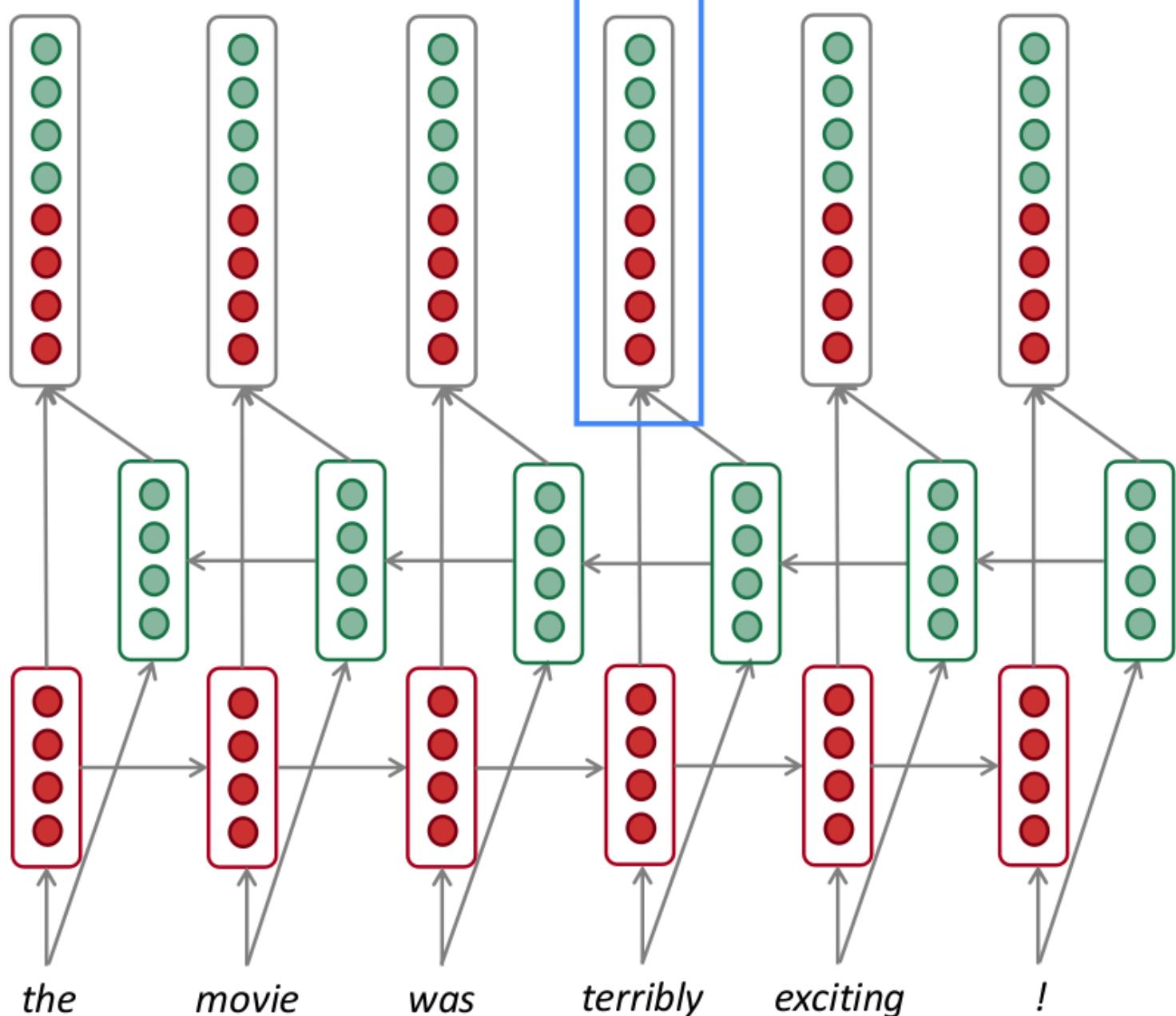
# Bidirectional RNNs

This contextual representation of “terribly” has both left and right context!

Concatenated hidden states

Backward RNN

Forward RNN



# Bidirectional RNNs

On timestep  $t$ :

This is a general notation to mean “compute one forward step of the RNN” – it could be a vanilla, LSTM or GRU computation.

Forward RNN

$$\vec{h}^{(t)} = \text{RNN}_{\text{FW}}(\vec{h}^{(t-1)}, \mathbf{x}^{(t)})$$

Backward RNN

$$\overleftarrow{h}^{(t)} = \text{RNN}_{\text{BW}}(\overleftarrow{h}^{(t+1)}, \mathbf{x}^{(t)})$$

Concatenated hidden states

$$h^{(t)} = [\vec{h}^{(t)}; \overleftarrow{h}^{(t)}]$$

Generally, these two RNNs have separate weights

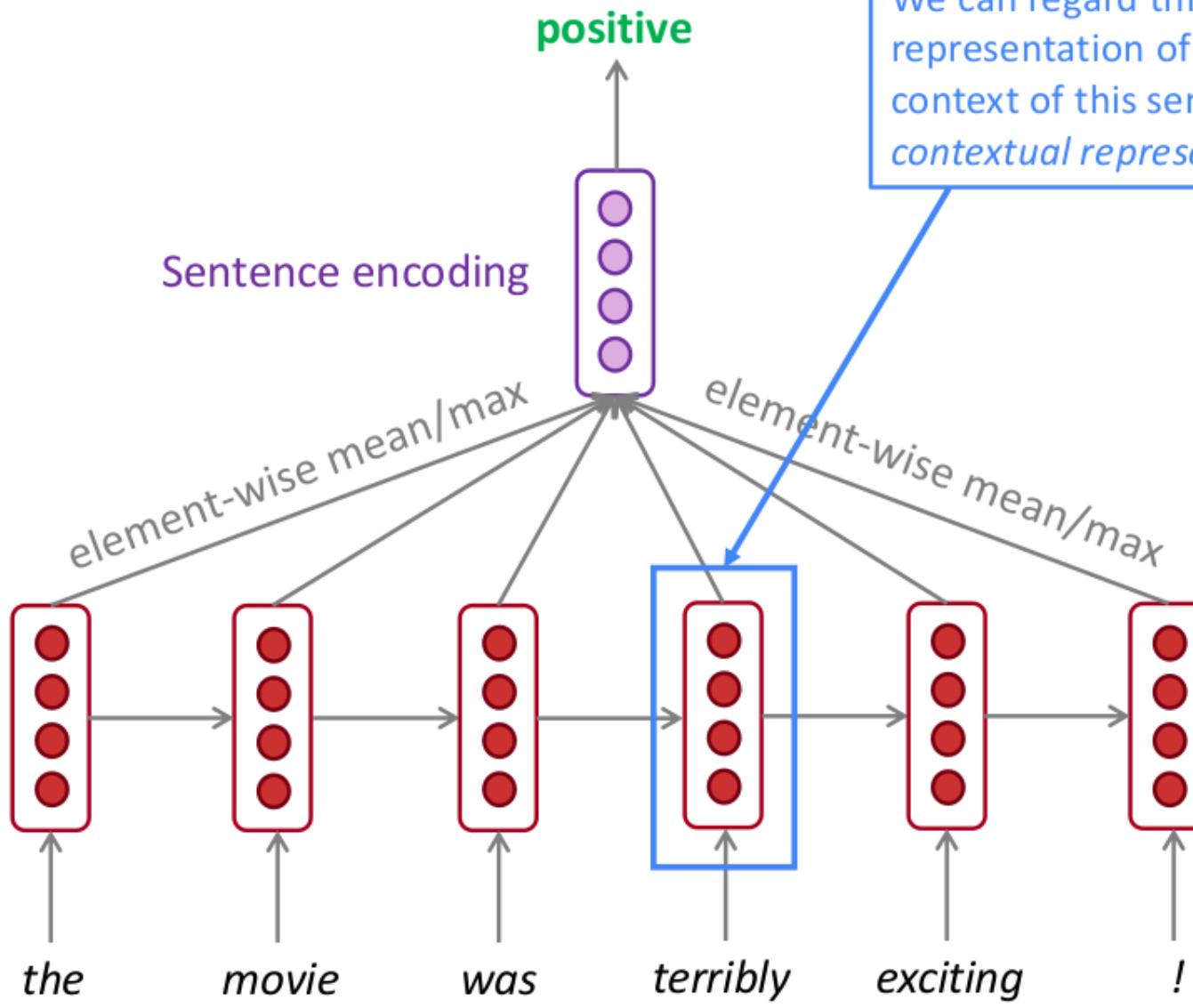
We regard this as “the hidden state” of a bidirectional RNN. This is what we pass on to the next parts of the network.

**Require full sequence available=> not for LMs**

**But similar bidirectional LMs exists which are 2 independent LMs**

# Bidirectional RNNs: motivation

Task: Sentiment Classification



We can regard this hidden state as a representation of the word “*terribly*” in the context of this sentence. We call this a *contextual representation*.

These contextual representations only contain information about the *left context* (e.g. “*the movie was*”).

What about *right context*?

In this example, “*exciting*” is in the right context and this modifies the meaning of “*terribly*” (from negative to positive)

# Multi-layer RNNs

The hidden states from RNN layer  $i$  are the inputs to RNN layer  $i+1$

