

LANGUAGE MODELING

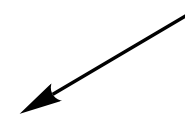
The task is to predict the next word in a sequence,
given the previous words.

the company raised its

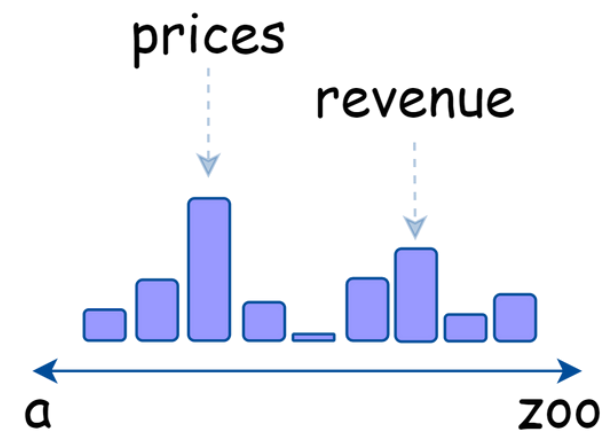
LANGUAGE MODEL

Given a sequence of words $w^{(1)}, w^{(2)}, \dots, w^{(t)}$
a language model outputs a probability distribution of the next word $w^{(t+1)}$

prefix



$P(w^{(t+1)} | \text{the company raised its})$



This is a multiclass classification problem:
you pick the next word from a predefined list of words

Count-based LM

N-GRAM LANGUAGE MODEL

A **n-gram** is a sequence of **n** consecutive words.

" Time flies like an arrow "

unigrams: Time, flies, like, an, arrow

bi-grams: Time flies, flies like, like an, an arrow

tri-grams: Time flies like, flies like an, like an arrow

four-grams: Time flies like an, flies like an arrow

N-GRAM LANGUAGE MODEL

Counting n-grams in a large corpus of text
to estimate the probability of the next word.

$$P(w \mid \text{the company raised its}) = \frac{\text{count}(\text{the company raised its } w)}{\text{count}(\text{the company raised its})}$$

↑
Prefix

Example:
assume that in the corpus:

- "the company raised its **prices**" appears 60 times
- "the company raised its" appears 100 times
- "the company raised its **revenue**" appears 20 times

$$P(\text{prices} \mid \text{the company raised its}) = \frac{60}{100}$$

$$P(\text{revenue} \mid \text{the company raised its}) = \frac{20}{100}$$

N-GRAM LANGUAGE MODEL

Make a **simplifying assumption**:

the probability of the next word only depends on previous $n-1$ words.

Example:

assume that we build a 5-gram language model :

"Thanks to a new marketing strategy, the company raised its ____"

N-GRAM LANGUAGE MODEL

Make a **simplifying assumption**:

the probability of the next word only depends on previous n-1 words.

Example:

assume that we build a 5-gram language model :

~~"Thanks to a new marketing strategy,~~ the company raised its ____"

condition on this

n-gram count

$$P(w \mid \text{the company raised its}) = \frac{\text{count}(\text{the company raised its } w)}{\text{count}(\text{the company raised its})}$$

(n-1)-gram count

$$P(\text{prices} \mid \text{the company raised its}) = 0.6$$

$$P(\text{revenue} \mid \text{the company raised its}) = 0.2$$

N-GRAM LANGUAGE MODEL

Make a **simplifying assumption**:

the probability of the next word only depends on previous n-1 words.

Example:

assume that we build a 5-gram language model :

~~"Thanks to a new marketing strategy,~~ the company raised its ____"

condition on this

Problem with the simplifying assumption:

Given the full context, "revenue" is more likely than "prices".

Longer context, helps the model to better predict the next word.

n-gram count

$$P(w \mid \text{the company raised its}) = \frac{\text{count}(\text{the company raised its } w)}{\text{count}(\text{the company raised its})}$$

(n-1)-gram count

$$P(\text{prices} \mid \text{the company raised its}) = 0.6$$

$$P(\text{revenue} \mid \text{the company raised its}) = 0.2$$

N-GRAM LANGUAGE MODEL

Extending to larger n -grams will not help, since we would need to obtain counts for longer sequences that can appear only once in the corpus, which makes the probability estimates unreliable: *the sparsity problem*.

N-GRAM LANGUAGE MODEL

In general this is an insufficient model of language because language has *long-distance dependencies*, that cannot be captured with fixed-window based models.

TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosopher Stone" book.

harry was _____

TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosopher Stone" book.

harry was _____

get a probability distribution
for the next word

$P(* \text{harry was})$		
just	0.066	<div></div>
turning	0.045	<div></div>
remembering	0.044	<div></div>
left	0.041	<div></div>
sure	0.001	<div></div>

TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosopher Stone" book.

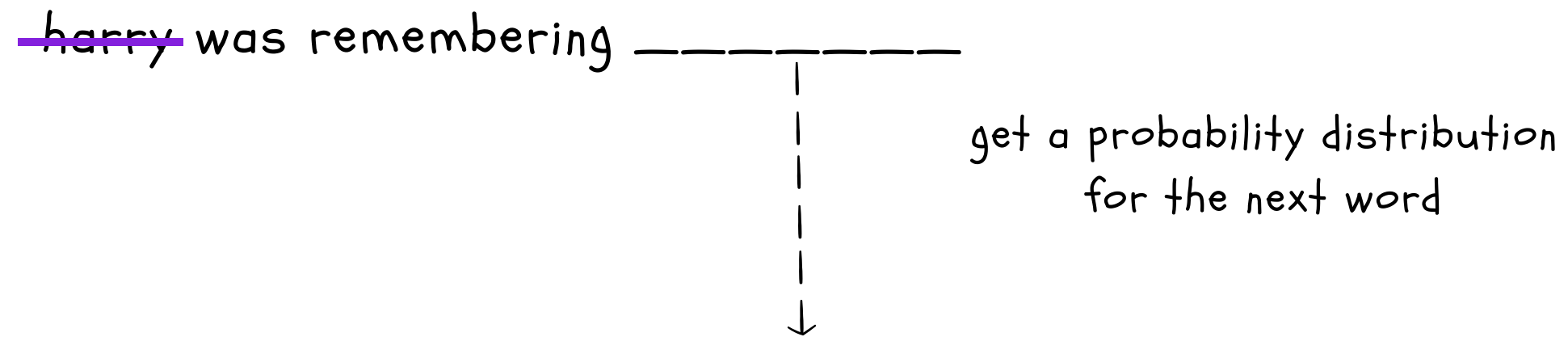
harry was remembering

sample from
the distribution

P (* harry was)		
just	0.066	<div></div>
turning	0.045	<div></div>
remembering	0.044	<div></div>
left	0.041	<div></div>
sure	0.001	<div></div>

TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosopher Stone" book.



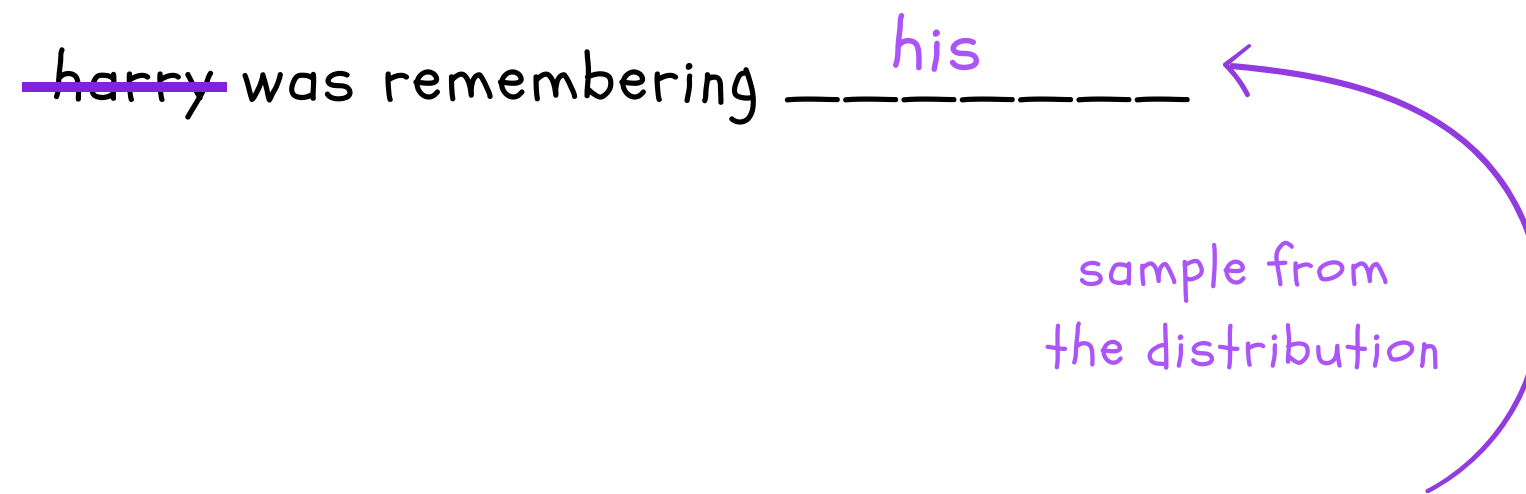
$P (* \text{ was remembering })$		
his	0.5	<div></div>
how	0.25	<div></div>
what	0.25	<div></div>

TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosopher Stone" book.

~~harry~~ was remembering his _____

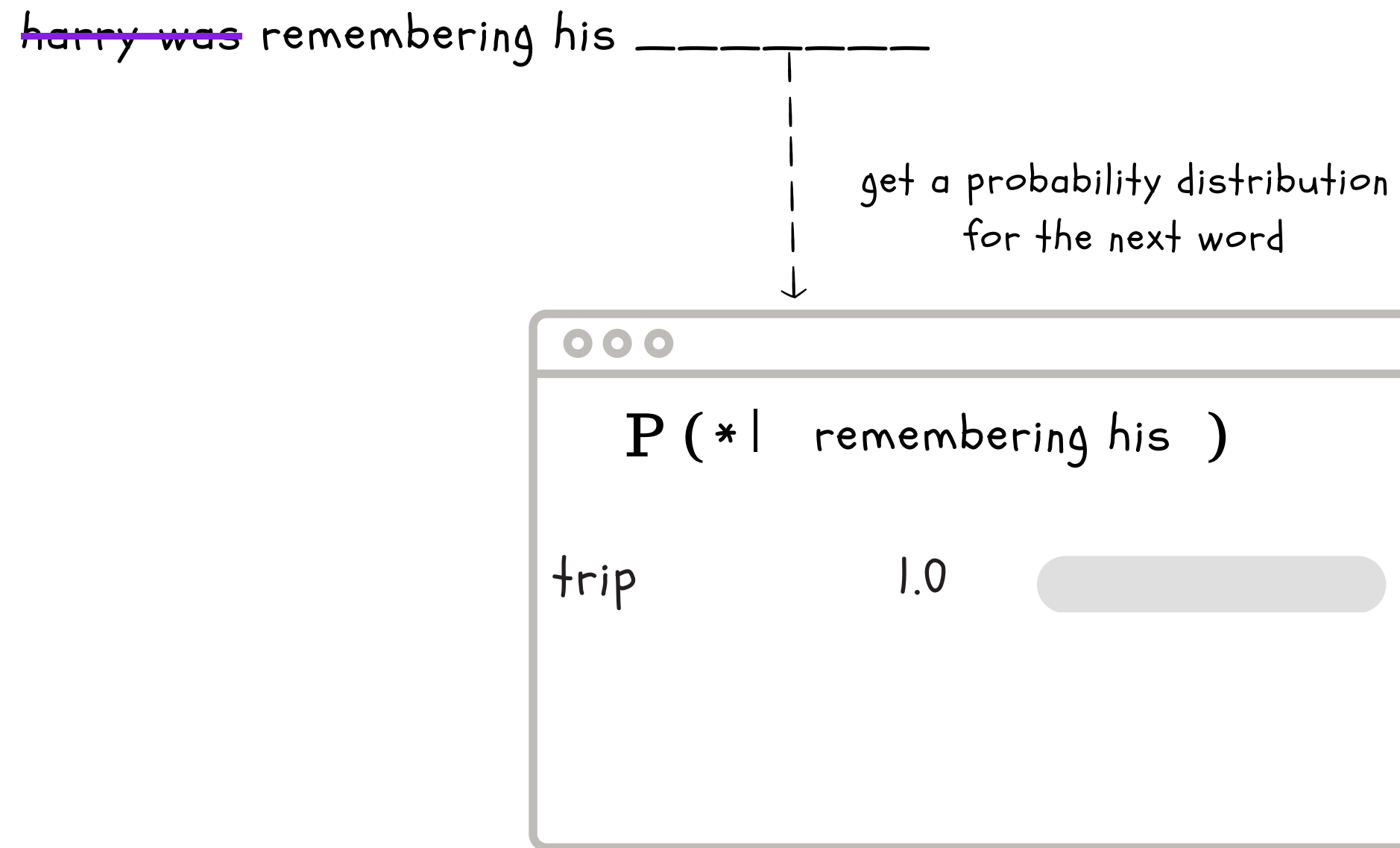
sample from
the distribution



$P(* \mid \text{ was remembering })$		
his	0.5	<div></div>
how	0.25	<div></div>
what	0.25	<div></div>

TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosopher Stone" book.




TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosopher Stone" book.

~~harry was~~ remembering his trip _____

sample from
the distribution



$$P(* \mid \text{remembering his})$$

trip	1.0	<div></div>
------	-----	-------------

TEXT GENERATION

A 3-gram language model built over "Harry Potter and the Philosophers Stone" book.

harry was remembering his trip to diagon alley . " i don't know how the muggles manage without magic , which was the only thing he felt that would make it more interesting . he was wearing a violet top hat fell off . " i can tell yeh that . " " i'm not going to fight malfoy , crabbe , and the philosophers stone.

NEURAL LANGUAGE MODEL

NEURAL LANGUAGE MODEL

We need a neural architecture that can process *any length input*.

The core idea of a *RNN* is to apply the exact same transformation on every step.
That's what makes us able to process any length input we want.

How would you train a RNN LANGUAGE MODEL?

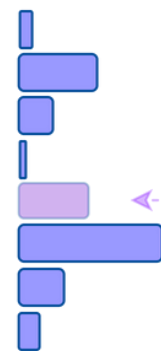
- We train a RNN Language Model on a huge corpus of text.
- Feed the corpus to the Language Model as a sequence of words, $w^{(1)}, w^{(2)}, \dots, w^{(t)}$
- Make the model predict the next upcoming word at every step.
- Training goal is to maximize the probability of the correct next word.

Minimize Cross Entropy Loss

$$\text{Cross Entropy Loss}(p^*, p) = - p^* \log(p) = \sum_V -p_i^* \log(p_i)$$

Training example: "the boy who ____"

Model prediction:
 $p(\text{word} \mid \text{"the boy who"})$



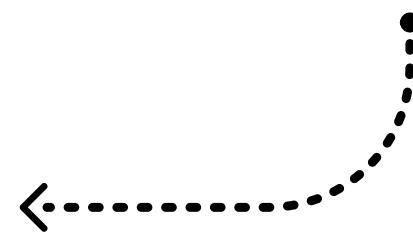
lived

Target:
 p^*



Cross-Entropy Loss is just the negative log probability of the target word.

$$\text{Loss}(p^*, p) = - \log P(\text{lived})$$



Target word

The train began to move. Harry saw the boys' mother waving and their sister, half laughing, half crying, running to keep up with the train until it gathered too much speed, then she fell back and waved.

Harry watched the girl and her mother disappear as the train rounded the corner. Houses flashed past the window. Harry felt a great leap of excitement. He didn't know what he was going to — but it had to be better than what he was leaving behind.

The door of the compartment slid open and the youngest redheaded boy came in.

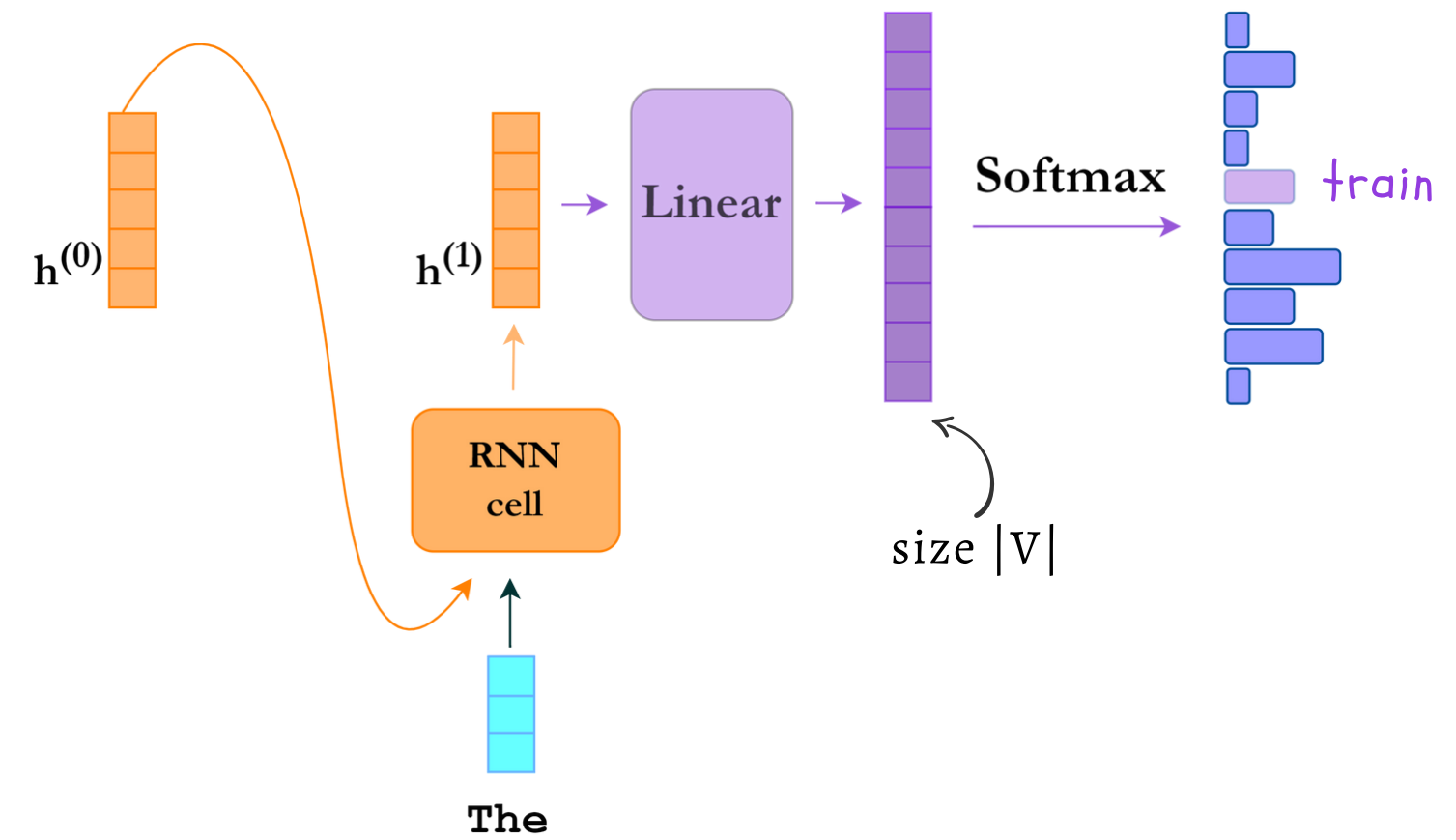
"Anyone sitting there?" he asked, pointing at the seat opposite Harry. "Everywhere else is full."

...

"Oh, I will," said Harry, and they were surprised at the grin that was spreading over his face. "They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer..."

Model prediction

$$P(* | \tau_{\text{The}})$$



Target word

The train began to move. Harry saw the boys' mother waving and their sister, half laughing, half crying, running to keep up with the train until it gathered too much speed, then she fell back and waved.

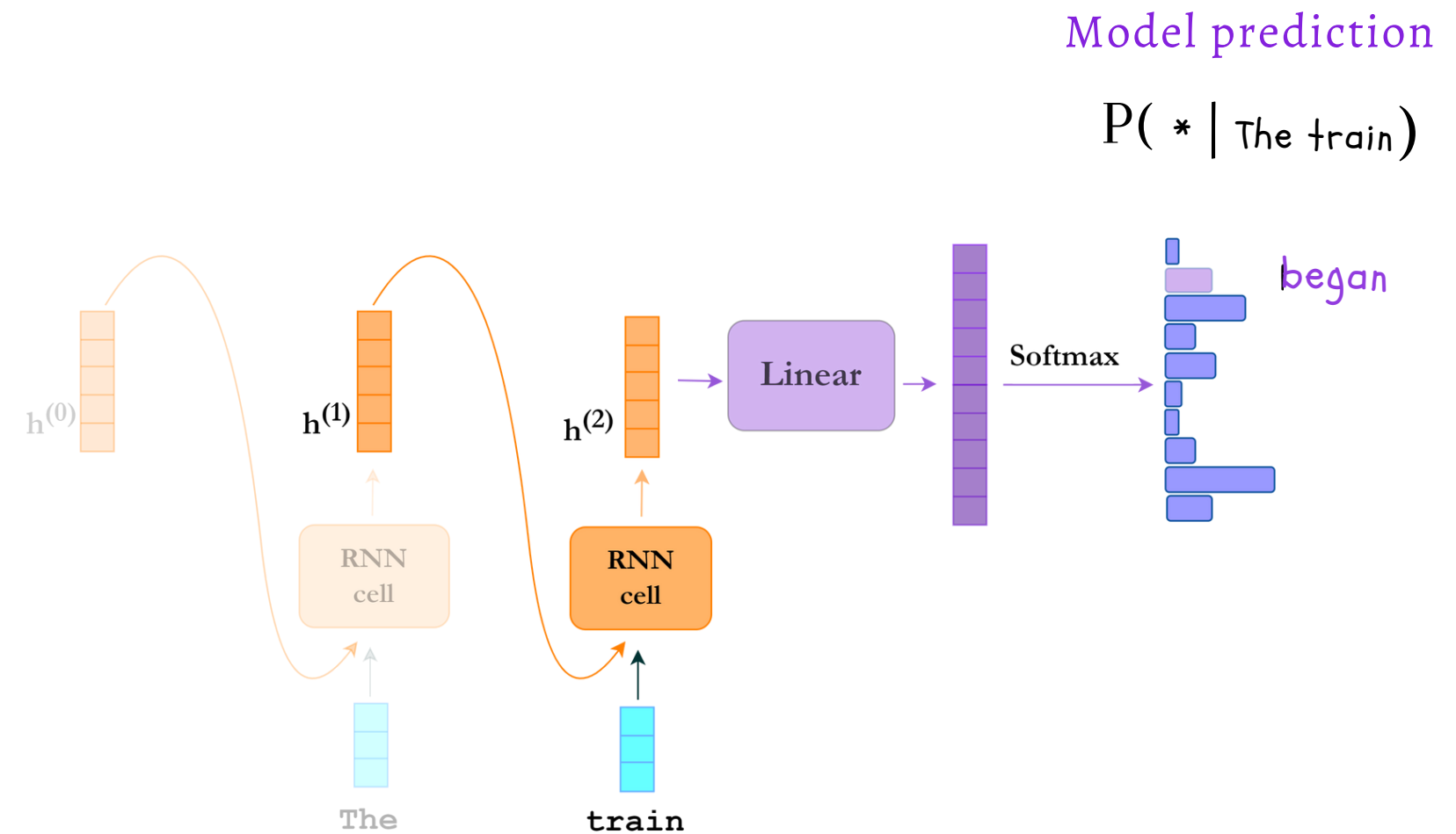
Harry watched the girl and her mother disappear as the train rounded the corner. Houses flashed past the window. Harry felt a great leap of excitement. He didn't know what he was going to — but it had to be better than what he was leaving behind.

The door of the compartment slid open and the youngest redheaded boy came in.

"Anyone sitting there?" he asked, pointing at the seat opposite Harry. "Everywhere else is full."

...

"Oh, I will," said Harry, and they were surprised at the grin that was spreading over his face. "They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer..."



Target word

The train began to move. Harry saw the boys' mother waving and their sister, half laughing, half crying, running to keep up with the train until it gathered too much speed, then she fell back and waved.

Harry watched the girl and her mother disappear as the train rounded the corner. Houses flashed past the window. Harry felt a great leap of excitement. He didn't know what he was going to — but it had to be better than what he was leaving behind.

The door of the compartment slid open and the youngest redheaded boy came in.

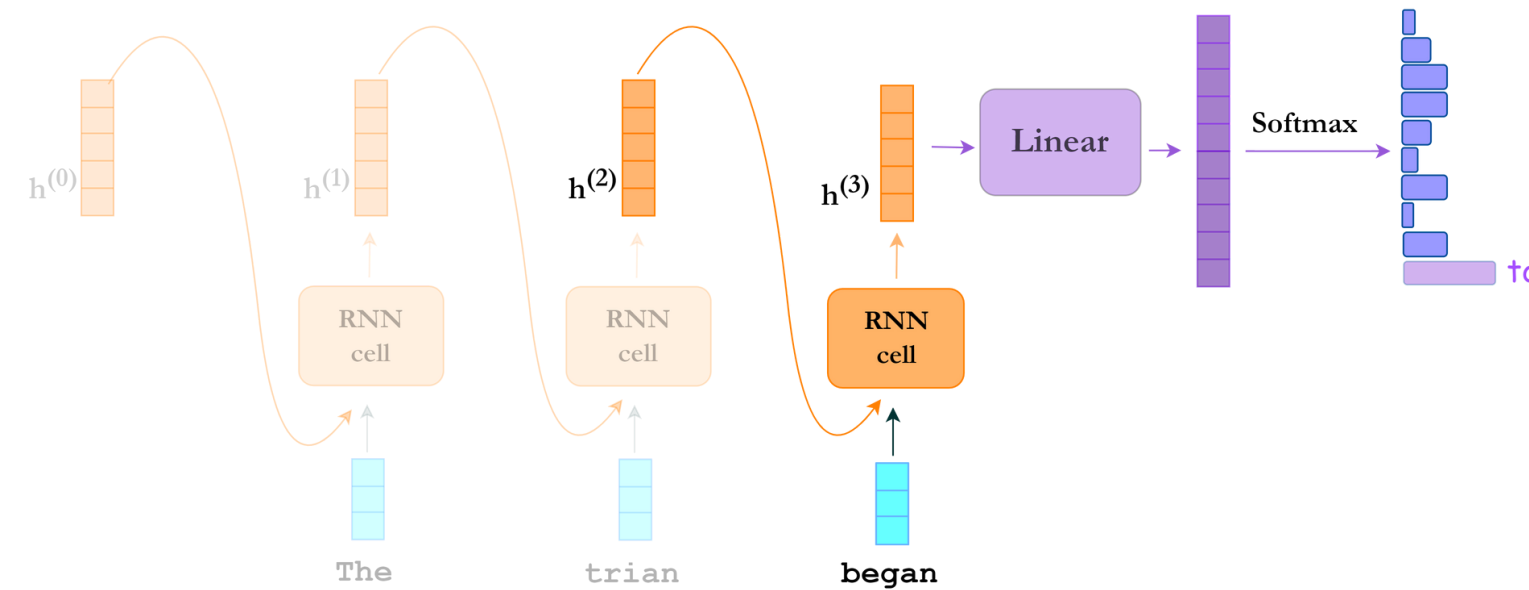
"Anyone sitting there?" he asked, pointing at the seat opposite Harry. "Everywhere else is full."

...

"Oh, I will," said Harry, and they were surprised at the grin that was spreading over his face. "They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer..."

Model prediction

$P(* \mid \text{The train began})$



Target word

Model prediction

$P(* \mid \text{The train began to})$

The train began to move. Harry saw the boys' mother waving and their sister, half laughing, half crying, running to keep up with the train until it gathered too much speed, then she fell back and waved.

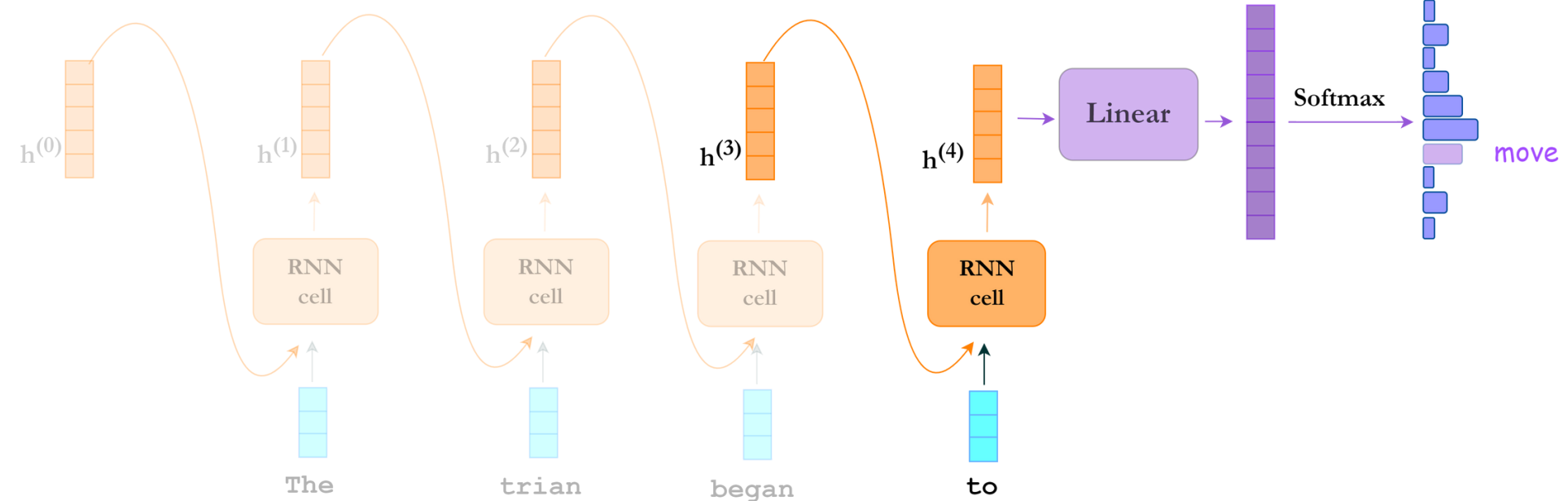
Harry watched the girl and her mother disappear as the train rounded the corner. Houses flashed past the window. Harry felt a great leap of excitement. He didn't know what he was going to — but it had to be better than what he was leaving behind.

The door of the compartment slid open and the youngest redheaded boy came in.

"Anyone sitting there?" he asked, pointing at the seat opposite Harry. "Everywhere else is full."

...

"Oh, I will," said Harry, and they were surprised at the grin that was spreading over his face. "They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer..."



TRAINING RNN LANGUAGE MODEL

Cross Entropy Loss

$$\text{CE loss} = - \frac{1}{T} \sum_{t=1}^T \log P(\text{target word}^{(t)})$$

Output distribution

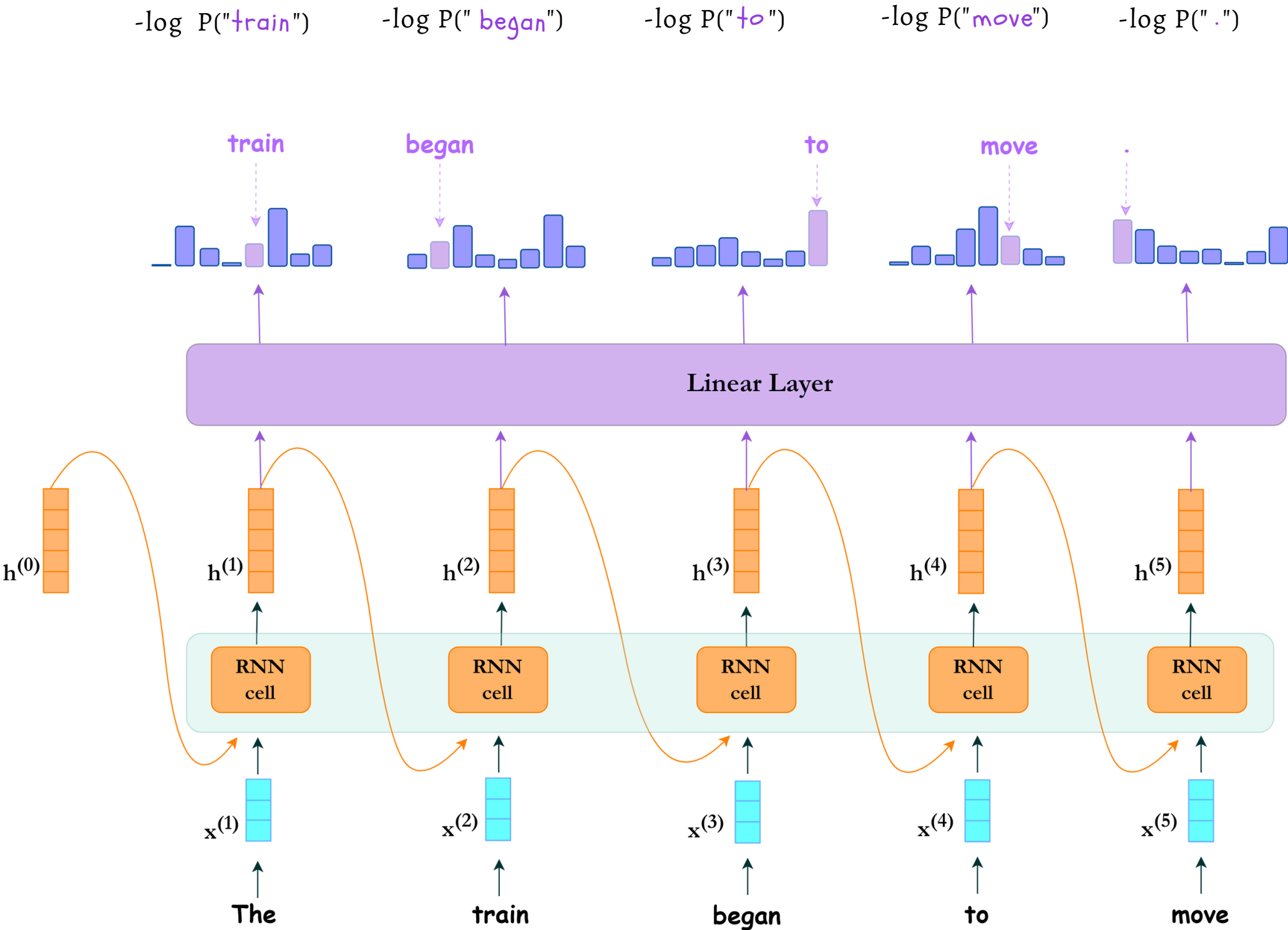
$$p^{(t)} = \text{Softmax} \left(\mathbf{U} \mathbf{h}^{(t)} + \text{bias} \right)$$

Hidden states

$$\mathbf{h}^{(t)} = \sigma \left(\mathbf{W}_{hh} \mathbf{h}^{(t-1)} + \mathbf{W}_{xh} \mathbf{x}^{(t)} + \text{bias} \right)$$

Word embeddings

$$\mathbf{x}^{(t)} = \mathbf{E} \left(\text{word}^{(t)} \right)$$



BUILDING TRAINING DATA

SEQUENCE LENGTH = 6

train began to move .
The train began to move

The train began to move . Harry saw the boys' mother waving and their sister , half laughing , half crying , running to keep up with the train until it gathered too much speed , then she fell back and waved .

Harry watched the girl and her mother disappear as the train rounded the corner . Houses flashed past the window . Harry felt a great leap of excitement . He didn't know what he was going to — but it had to be better than what he was leaving behind .

The door of the compartment slid open and the youngest redheaded boy came in .

" Anyone sitting there ? " he asked , pointing at the seat opposite Harry . " Everywhere else is full . "

...

" Oh, I will, " said Harry, and they were surprised at the grin that was spreading over his face. " They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer. "

INPUT SEQUENCEs

The train began to move

TARGET SEQUENCEs

train began to move .

BUILDING TRAINING DATA

SEQUENCE LENGTH = 6



began to move . Harry
train began to move .

The train began to move . Harry saw the boys' mother waving and their sister , half laughing , half crying , running to keep up with the train until it gathered too much speed , then she fell back and waved .

Harry watched the girl and her mother disappear as the train rounded the corner . Houses flashed past the window . Harry felt a great leap of excitement . He didn't know what he was going to — but it had to be better than what he was leaving behind .
The door of the compartment slid open and the youngest redheaded boy came in .
" Anyone sitting there ? " he asked , pointing at the seat opposite Harry . " Everywhere else is full . "

...

"Oh, I will," said Harry, and they were surprised at the grin that was spreading over his face. " They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer."

INPUT SEQUENCEs

The train began to move
train began to move .

TARGET SEQUENCEs

train began to move .
began to move . Harry

BUILDING TRAINING DATA

SEQUENCE LENGTH = 6

to move . Harry saw
began to move . Harry

The train began to move . Harry saw the boys' mother waving and their sister , half laughing , half crying , running to keep up with the train until it gathered too much speed , then she fell back and waved .

Harry watched the girl and her mother disappear as the train rounded the corner . Houses flashed past the window . Harry felt a great leap of excitement . He didn't know what he was going to — but it had to be better than what he was leaving behind .

The door of the compartment slid open and the youngest redheaded boy came in .

" Anyone sitting there ? " he asked , pointing at the seat opposite Harry . " Everywhere else is full . "

...

" Oh, I will, " said Harry, and they were surprised at the grin that was spreading over his face . " They don't know we're not allowed to use magic at home . I'm going to have a lot of fun with Dudley this summer . "

INPUT SEQUENCEs

The train began to move
train began to move .
began to move . Harry

TARGET SEQUENCEs

train began to move .
began to move . Harry
to move . Harry saw

BUILDING TRAINING DATA

SEQUENCE LENGTH = 6

move . Harry saw the
to move . Harry saw

The train began to move . Harry saw the boys' mother waving and their sister , half laughing , half crying , running to keep up with the train until it gathered too much speed , then she fell back and waved .

Harry watched the girl and her mother disappear as the train rounded the corner . Houses flashed past the window . Harry felt a great leap of excitement . He didn't know what he was going to — but it had to be better than what he was leaving behind .
The door of the compartment slid open and the youngest redheaded boy came in .
" Anyone sitting there ? " he asked , pointing at the seat opposite Harry . " Everywhere else is full . "

...

" Oh, I will, " said Harry, and they were surprised at the grin that was spreading over his face. " They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer. "

INPUT SEQUENCEs

The train began to move
train began to move .
began to move . Harry
to move . Harry saw

TARGET SEQUENCEs

train began to move .
began to move . Harry
to move . Harry saw
move . Harry saw the

BUILDING TRAINING DATA

The train began to move . Harry saw the boys' mother waving and their sister , half laughing , half crying , running to keep up with the train until it gathered too much speed , then she fell back and waved .

Harry watched the girl and her mother disappear as the train rounded the corner . Houses flashed past the window . Harry felt a great leap of excitement . He didn't know what he was going to — but it had to be better than what he was leaving behind .

The door of the compartment slid open and the youngest redheaded boy came in .

" Anyone sitting there ? " he asked , pointing at the seat opposite Harry . " Everywhere else is full . "

...

" Oh, I will, " said Harry, and they were surprised at the grin that was spreading over his face. " They don't know we're not allowed to use magic at home. I'm going to have a lot of fun with Dudley this summer . "

Dudley this summer . "

with Dudley this summer .

INPUT SEQUENCE

The train began to move
train began to move .
began to move . Harry
to move . Harry saw

...

fun with Dudley this summer
with Dudley this summer .

TARGET SEQUENCE

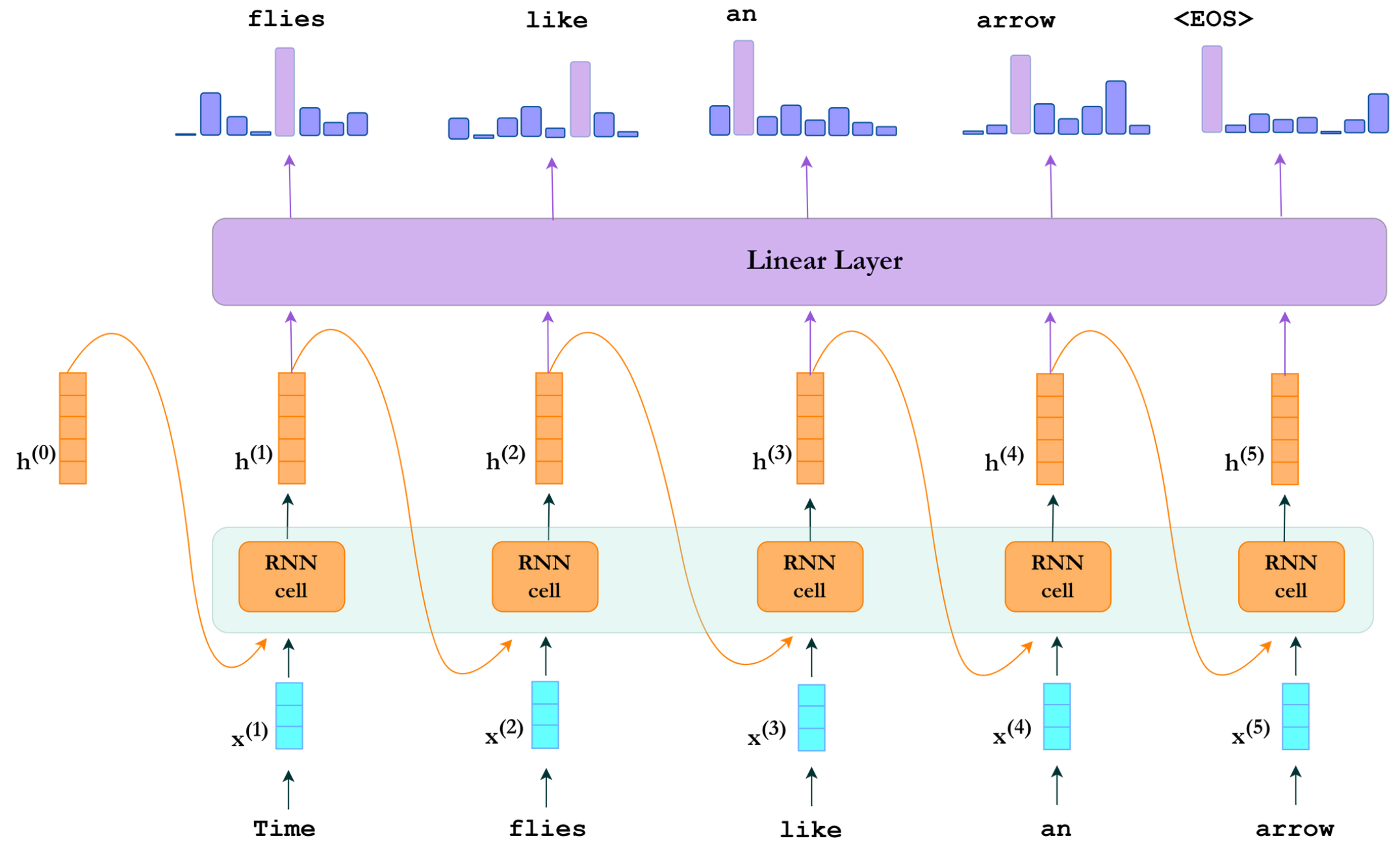
train began to move .
began to move . Harry
to move . Harry saw
move . Harry saw the

...

with Dudley this summer .
Dudley this summer . "

TEXT GENERATION with RNN

You can generate text by repeatedly sampling the next word and use it as input at the next step.
We can keep like this as long as we like.



Perplexity and Cross Entropy

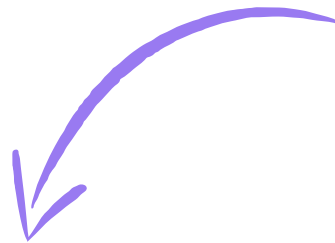
You can compare different language models,
by evaluating the model perplexity on a new corpus.

Perplexity estimates how much the model is surprised at seeing new text.

A better model has lower perplexity.

$$\text{Perplexity} = \exp\left(\frac{1}{T} \sum_{w=1}^T -\log P(w)\right)$$

is equal to the exponential of the
Cross Entropy Loss, normalized
over number of words in the corpus



Text Generation Examples

3-layer RNN trained on all the works of Shakespeare (4.4MB file).

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 nudes 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.

Multilayer LSTM trained on the raw LaTeX source (16 MB file)

For $\bigoplus_{n=1,\dots,m}$ where $\mathcal{L}_{m,\bullet} = 0$, hence we can find a closed subset \mathcal{H} in \mathcal{H} and any sets \mathcal{F} on X , U is a closed immersion of S , then $U \rightarrow T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \rightarrow V$. Consider the maps M along the set of points Sch_{fppf} and $U \rightarrow U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ??. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \rightarrow S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S . We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \rightarrow \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. \square

To prove study we see that $\mathcal{F}|_U$ is a covering of \mathcal{X}' , and \mathcal{T}_i is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and \mathcal{F}_p exists and let \mathcal{F}_i be a presheaf of \mathcal{O}_X -modules on \mathcal{C} as a \mathcal{F} -module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1} \mathcal{F}$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)_{fppf}^{opp}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \mapsto (U, \mathrm{Spec}(A))$$

is an open subset of X . Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S .

Proof. See discussion of sheaves of sets. \square

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces, \acute{e}tale}$ which gives an open subspace of X and T equal to S_{Zar} , see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S .

Lemma 0.1. Assume (3) and (3) by the construction in the description.

Suppose $X = \lim |X|$ (by the formal open covering X and a single map $\underline{\mathrm{Proj}}_X(\mathcal{A}) = \mathrm{Spec}(B)$ over U compatible with the complex

$$\mathrm{Set}(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X, \mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \rightarrow \mathcal{C}_{Z/X}$ is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition ?? (without element is when the closed subschemes are catenary. If T is surjective we may assume that T is connected with residue fields of S . Moreover there exists a closed subspace $Z \subset X$ of X where U in X' is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem

(1) f is locally of finite type. Since $S = \mathrm{Spec}(R)$ and $Y = \mathrm{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X . But given a scheme U and a surjective étale morphism $U \rightarrow X$. Let $U \cap U = \coprod_{i=1,\dots,n} U_i$ be the scheme X over S at the schemes $X_i \rightarrow X$ and $U = \lim_i X_i$. \square

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X}, \dots, 0}$.

Lemma 0.2. Let X be a locally Noetherian scheme over S , $E = \mathcal{F}_{X/S}$. Set $\mathcal{I} = \mathcal{I}_1 \subset \mathcal{I}'_n$. Since $\mathcal{I}^n \subset \mathcal{I}^n$ are nonzero over $i_0 \leq \mathfrak{p}$ is a subset of $\mathcal{I}_{n,0} \circ \overline{A}_2$ works.

Lemma 0.3. In Situation ??. Hence we may assume $\mathfrak{q}' = 0$.

Proof. We will use the property we see that \mathfrak{p} is the next functor (??). On the other hand, by Lemma ?? we see that

$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F -algebra where δ_{n+1} is a scheme over S . \square

3-layer LSTM trained on Linux Source Code (474MB file)

```
/*
 * 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;
}
```