Transformer (for NLP)
Text-to-Image Creation
Vision Transformer (for CV)

Lecture by Margrit Betke, CS 585, April 16, 2024

with many slides from Steve Seitz' videos:

Part 1 & Part 2

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

Bob Dylan, *Tangled up in Blue*

slide from Steve Seitz's [video]

Early one morning the sun was shining I was laying in bed

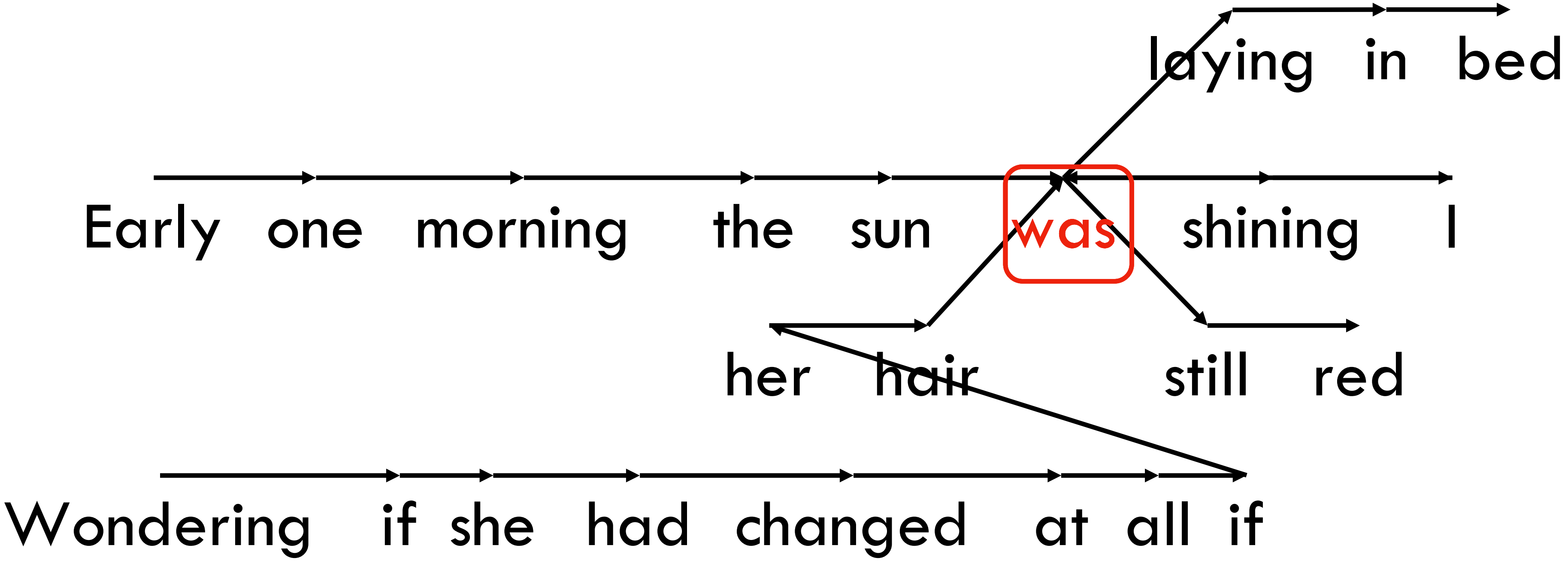
Wondering if she had changed at all if her hair was still red

slide from Steve Seitz's [video]()

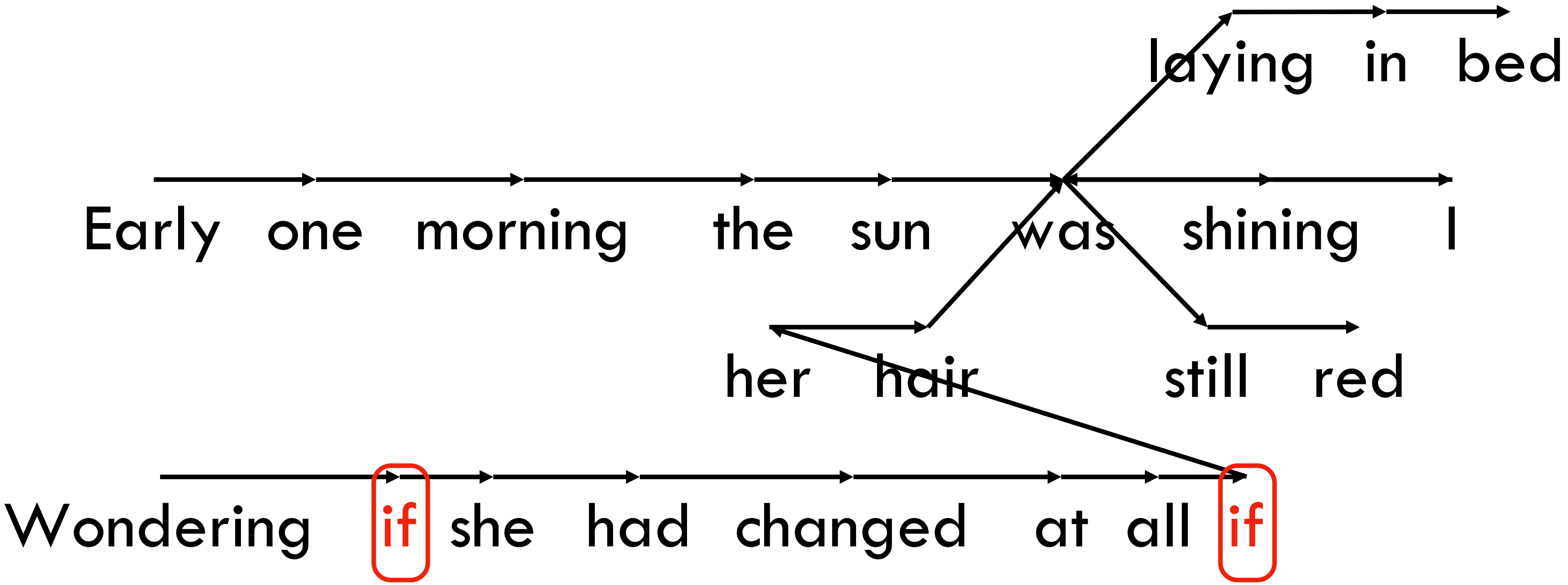
Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

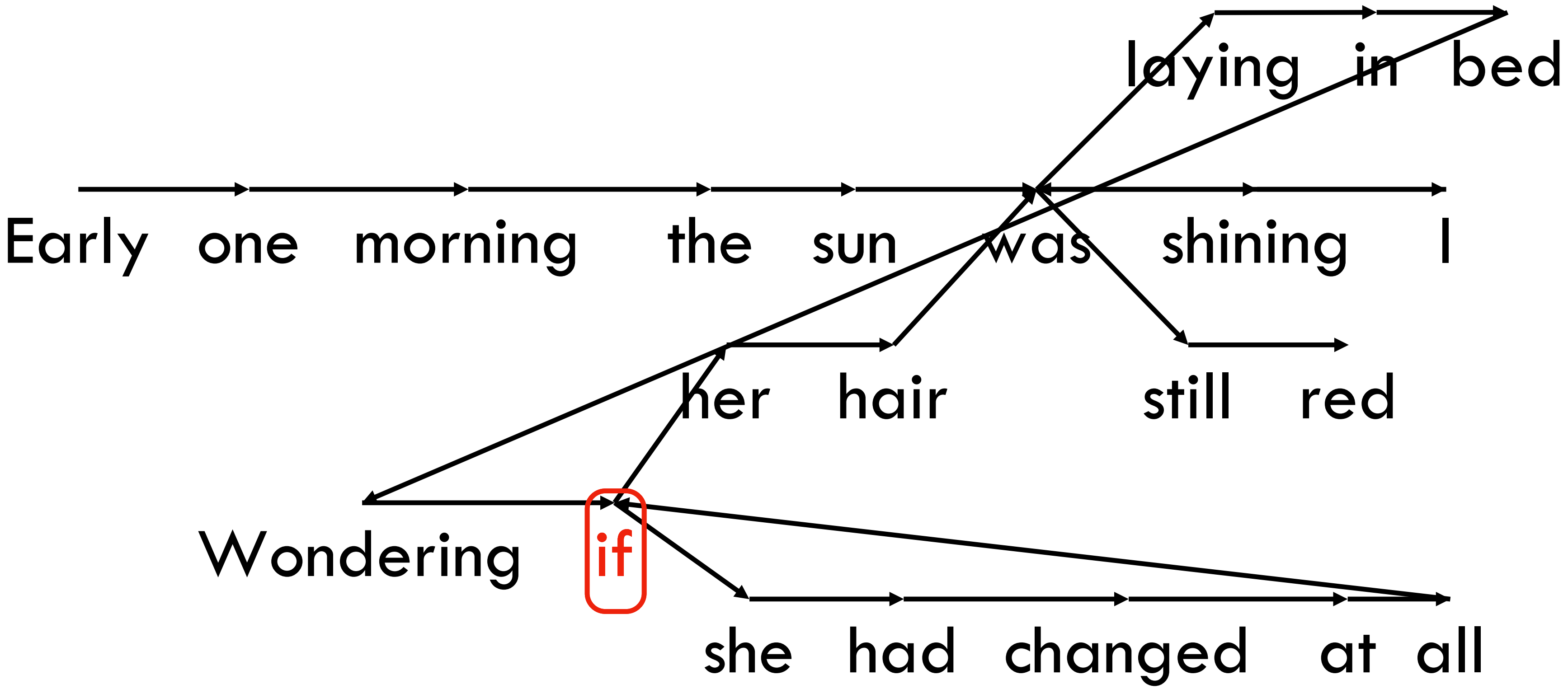
slide from Steve Seitz's [video]()



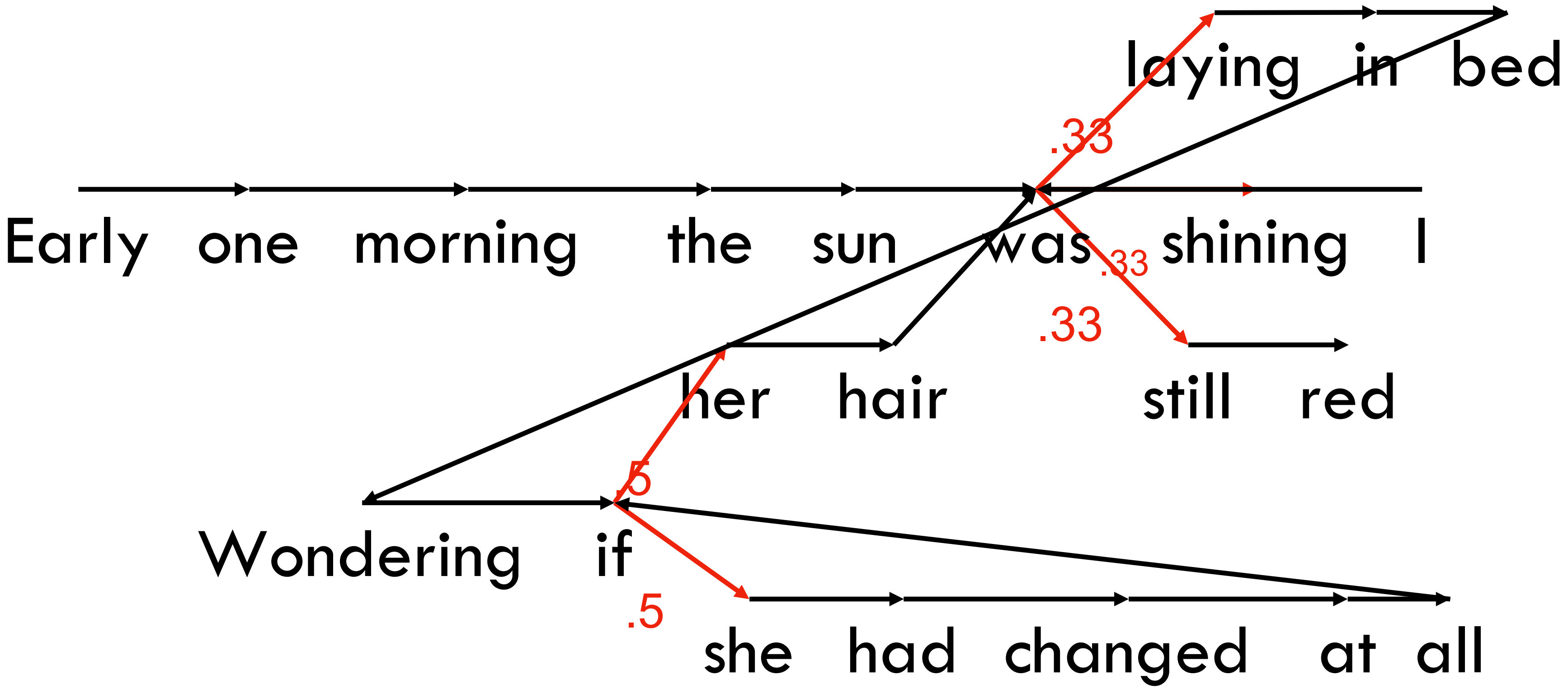
slide from Steve Seitz's video



slide from Steve Seitz's video



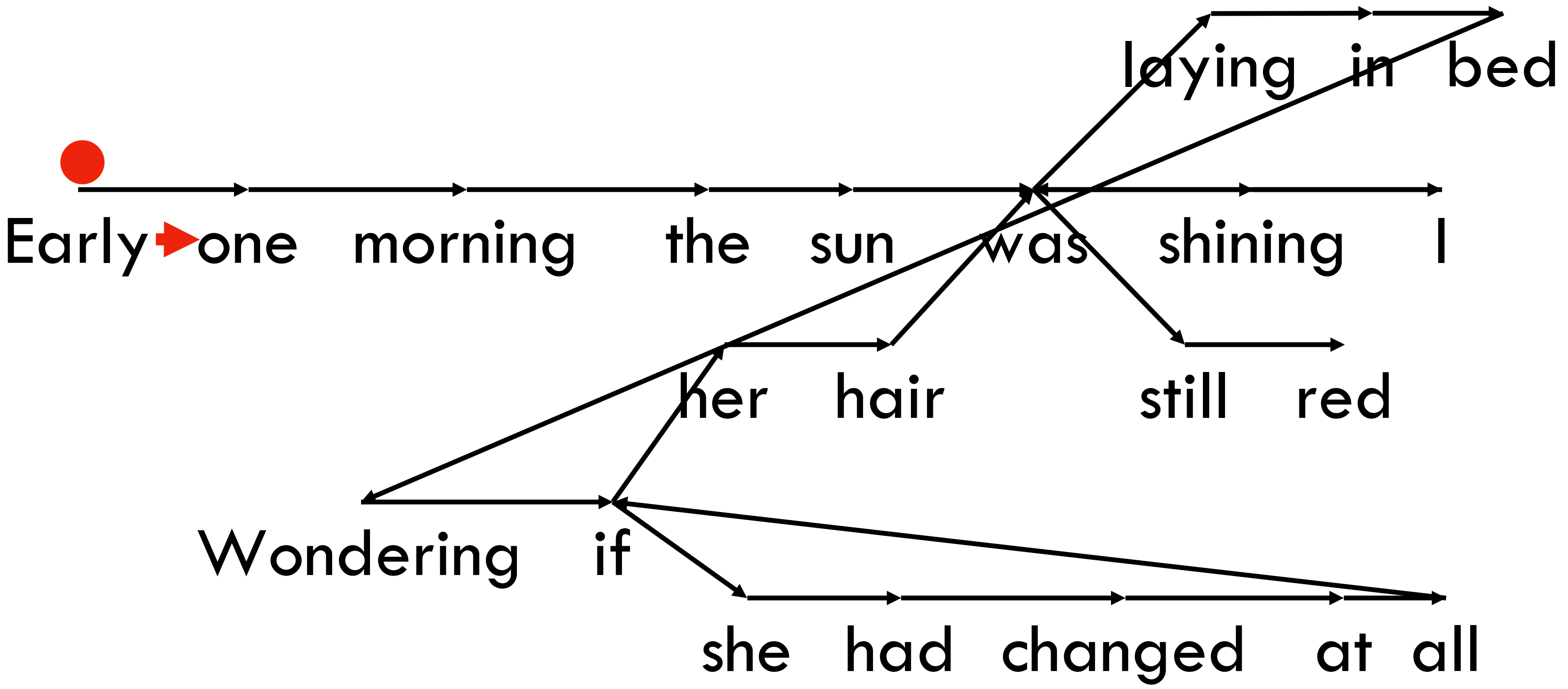
slide from Steve Seitz's video



Language Model

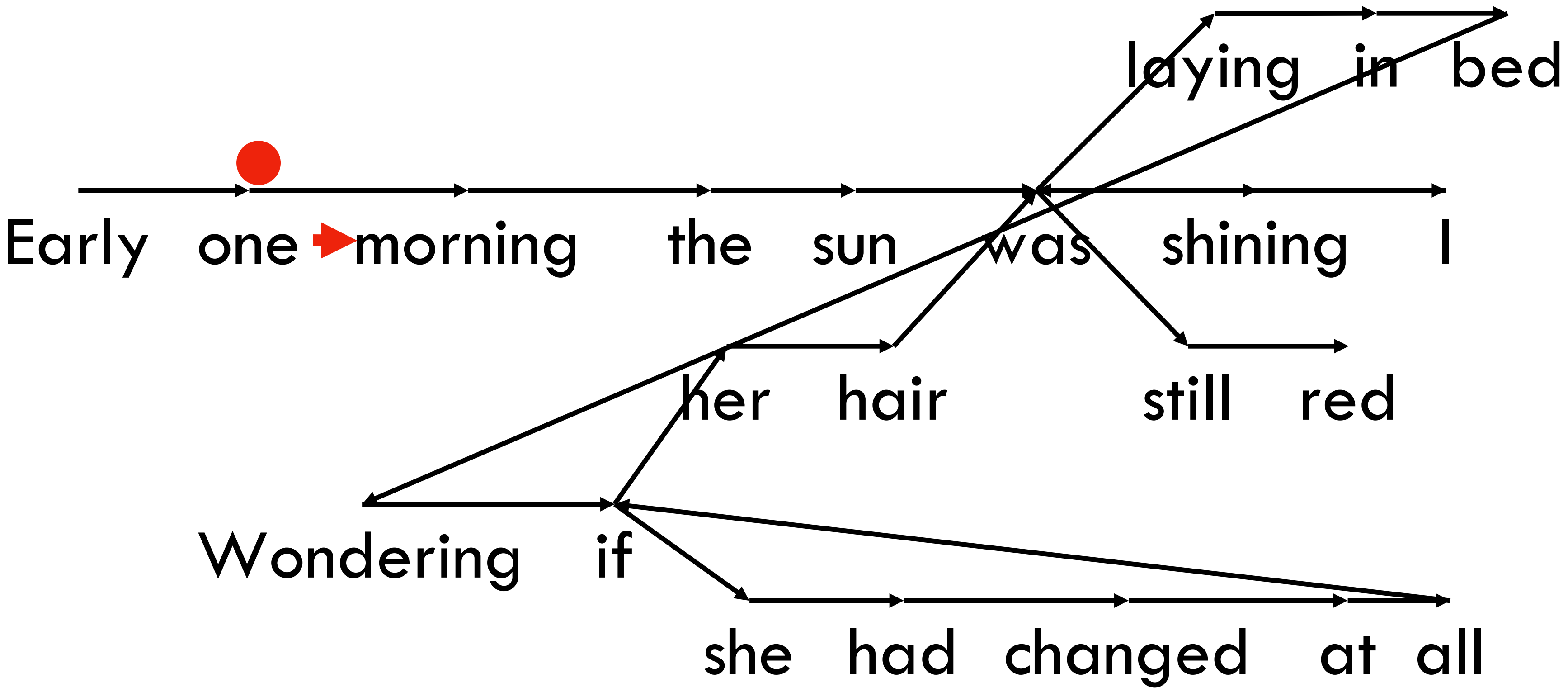
slide from Steve Seitz's [video]

Early



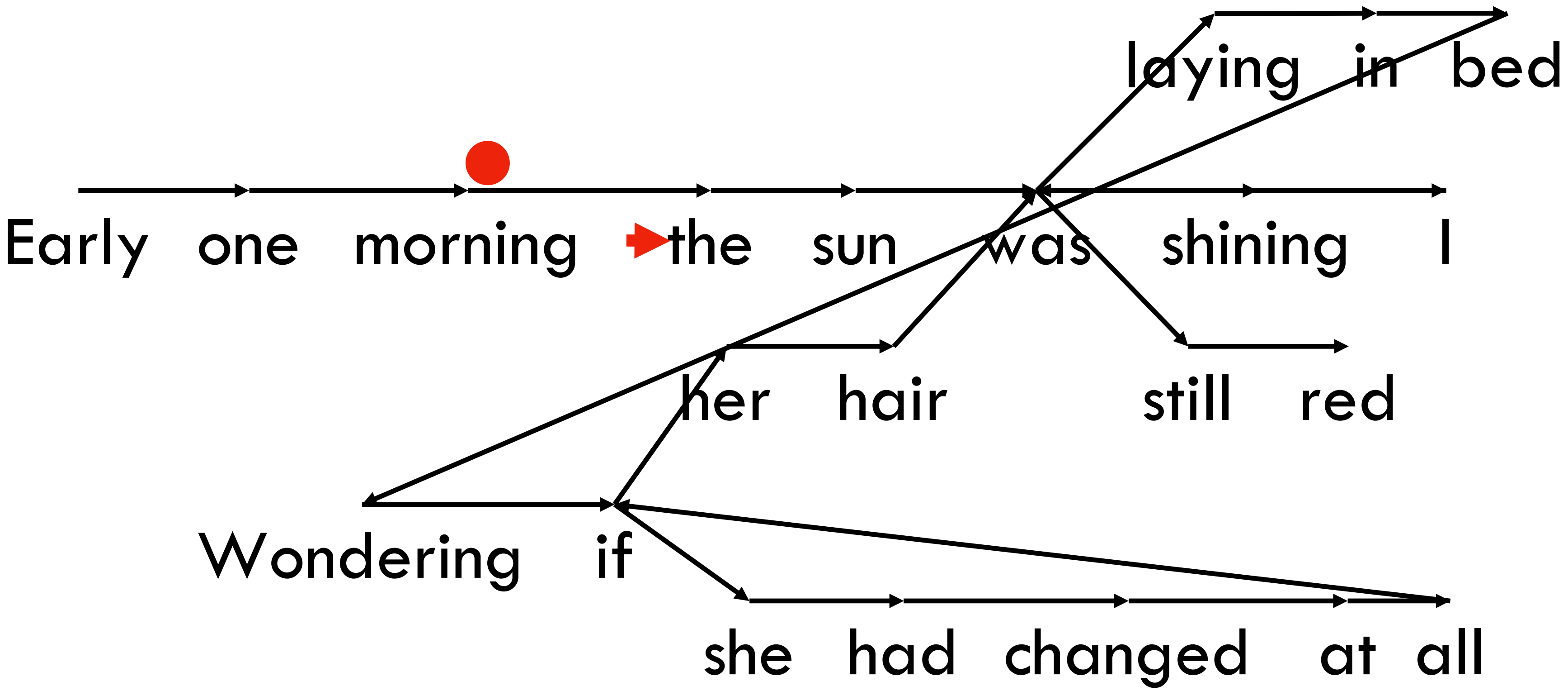
slide from Steve Seitz's video

Early one



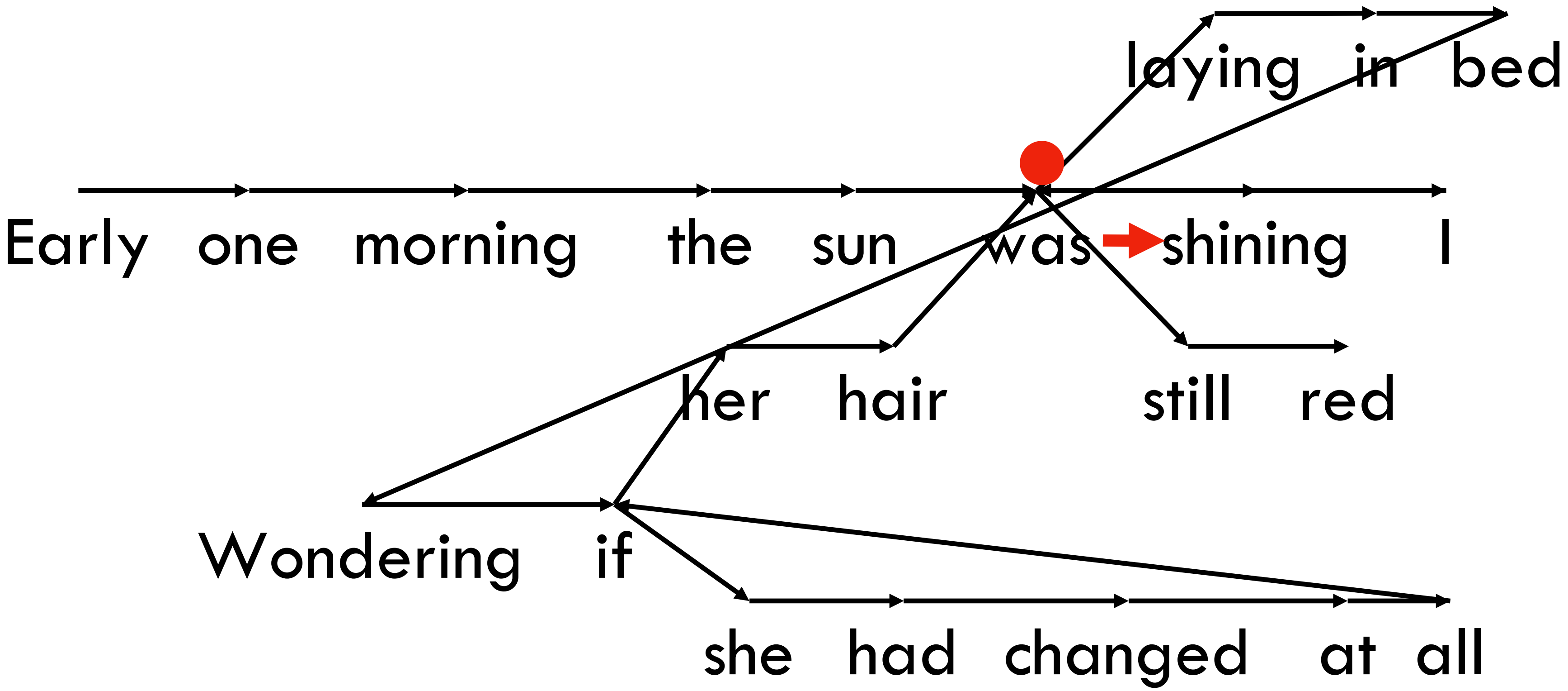
slide from Steve Seitz's [video]

Early one morning



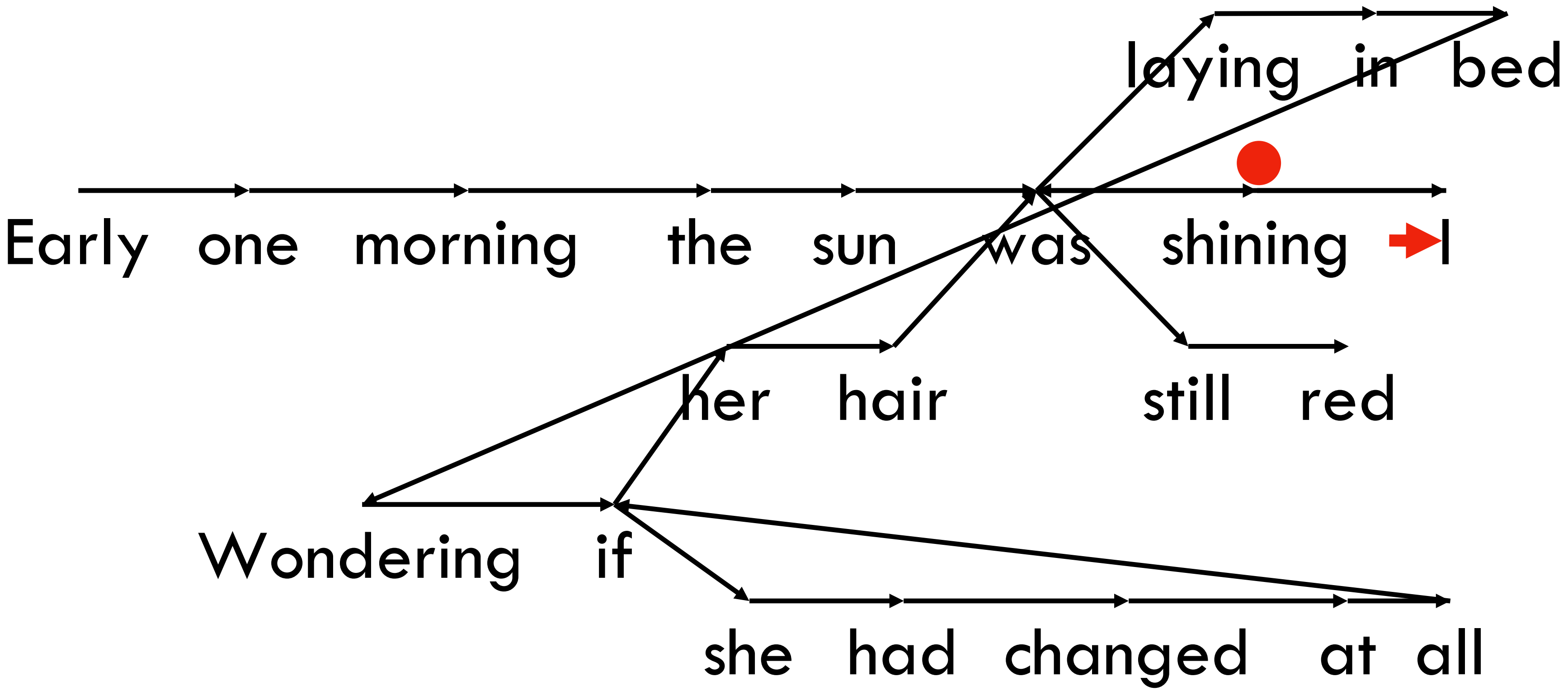
slide from Steve Seitz's [video]

Early one morning the sun was



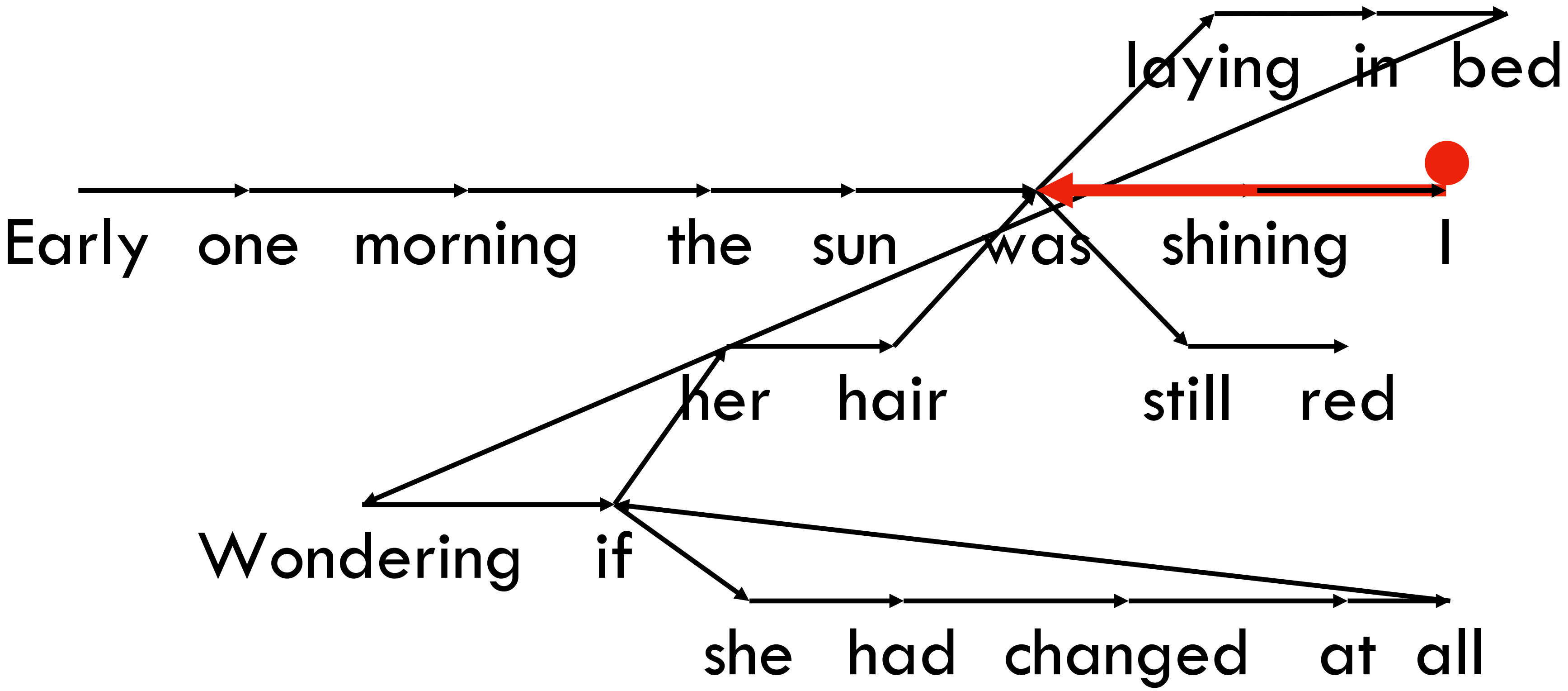
slide from Steve Seitz's video

Early one morning the sun was shining



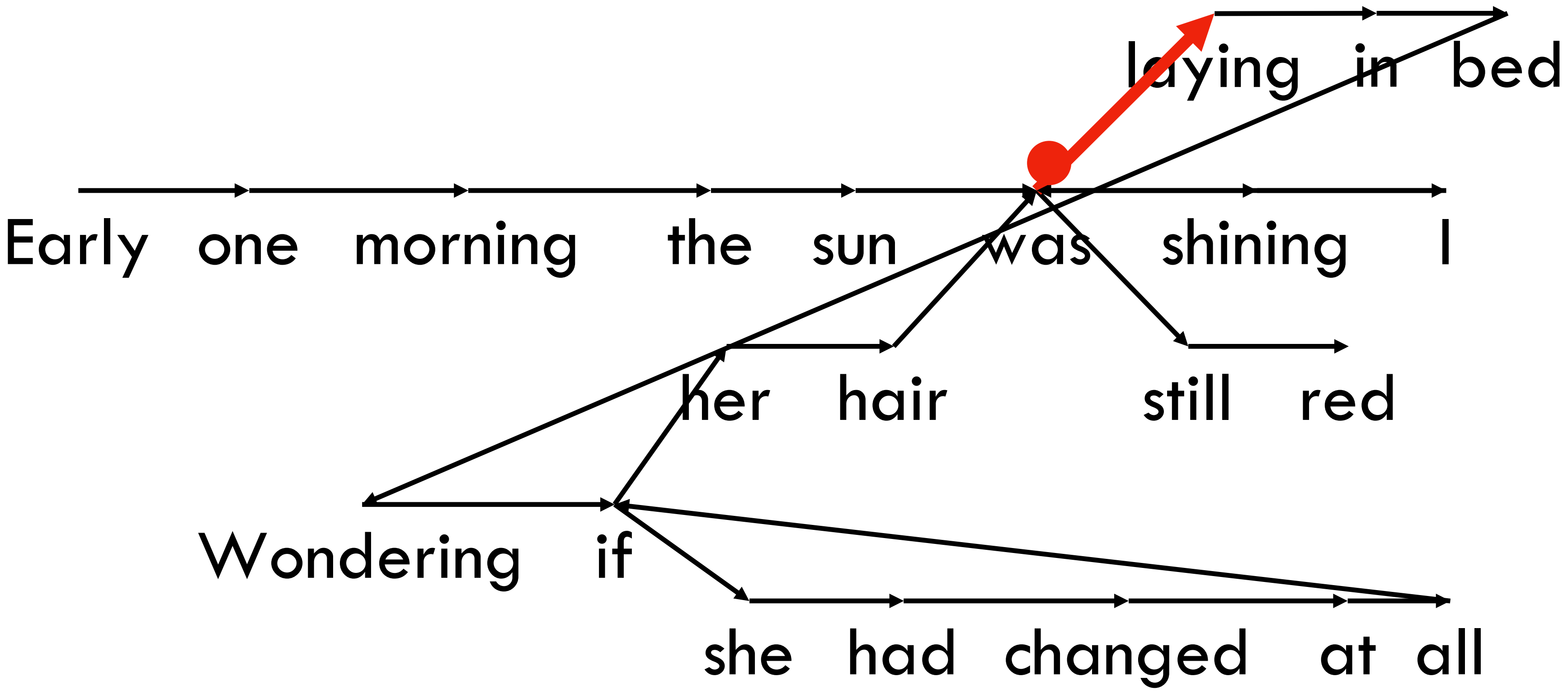
slide from Steve Seitz's [video]()

Early one morning the sun was shining I



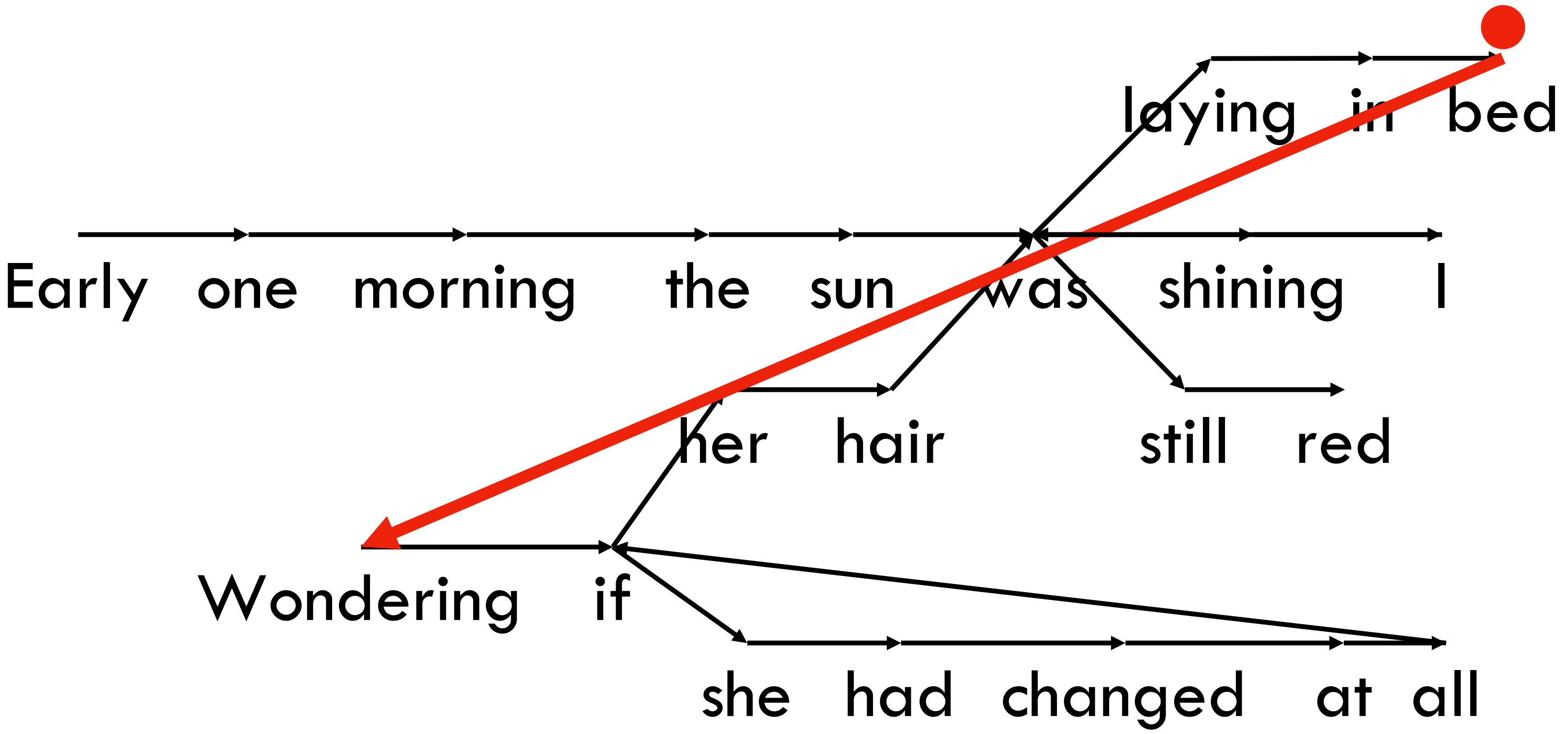
slide from Steve Seitz's video

Early one morning the sun was shining I was



slide from Steve Seitz's [video]

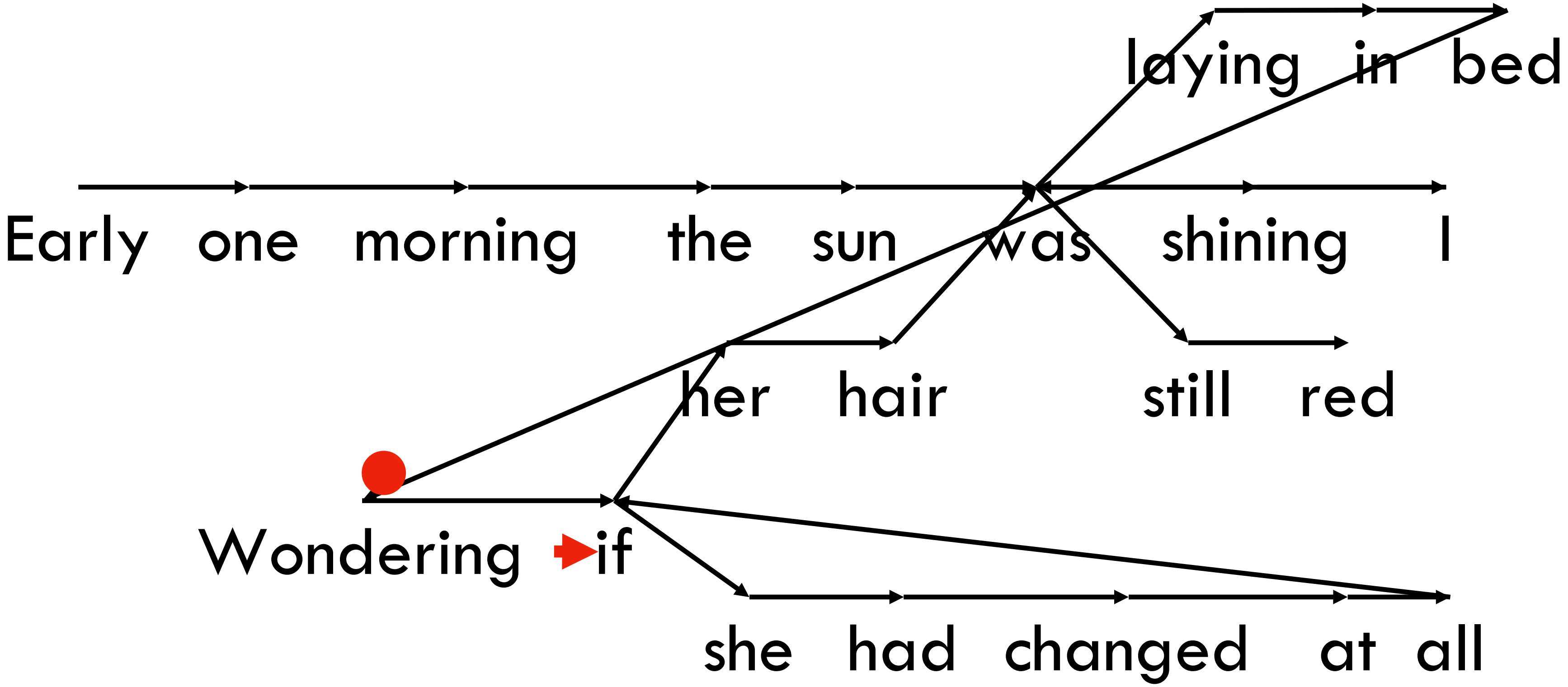
Early one morning the sun was shining I was laying in bed



slide from Steve Seitz's [video]

Early one morning the sun was shining I was laying in bed

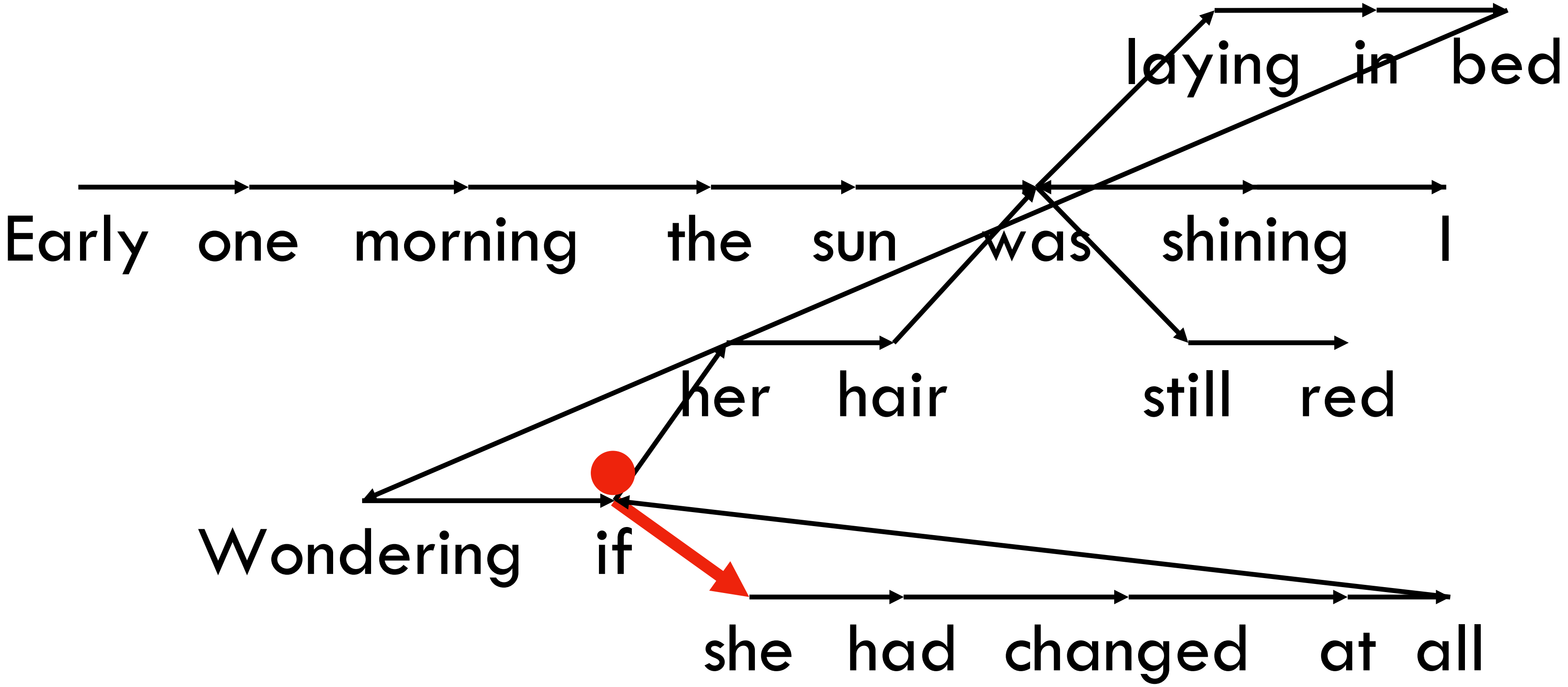
Wondering



slide from Steve Seitz's [video]

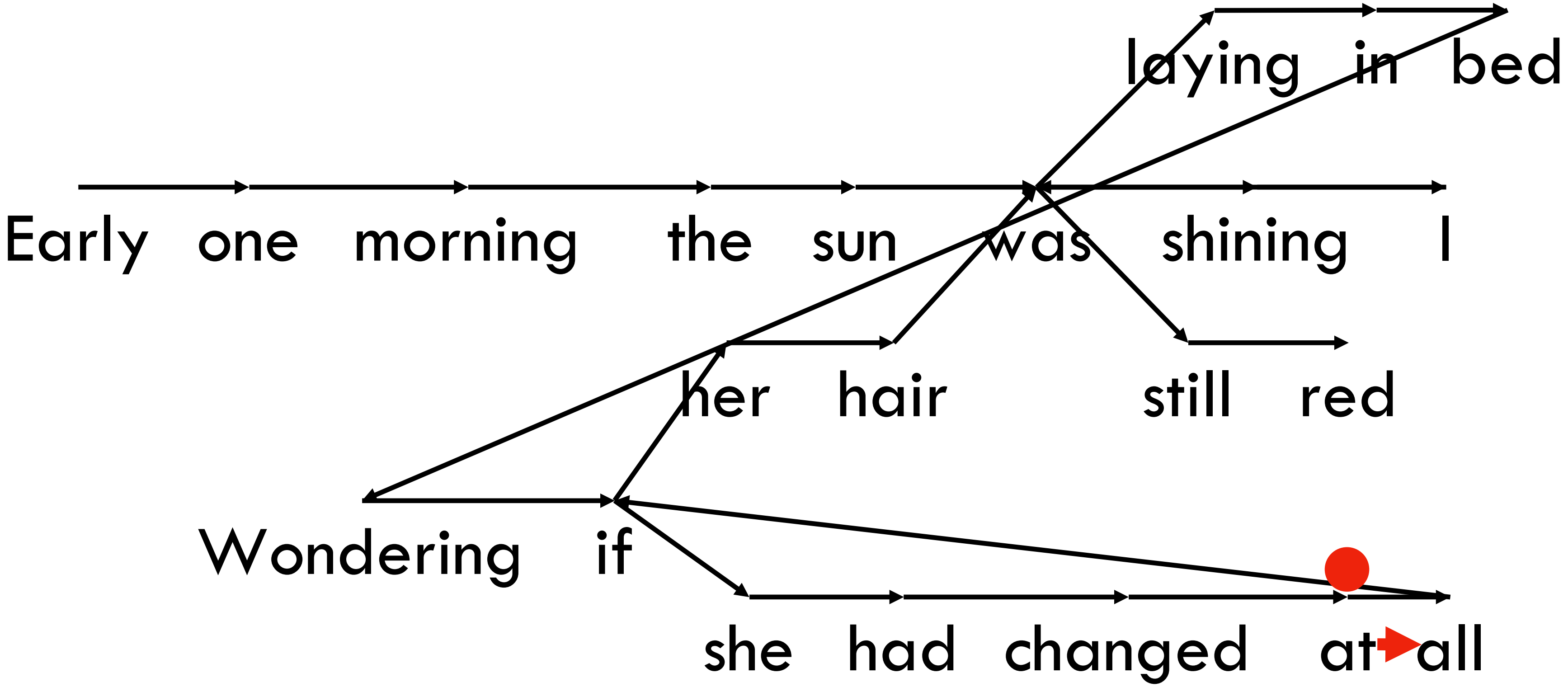
Early one morning the sun was shining I was laying in bed

Wondering if



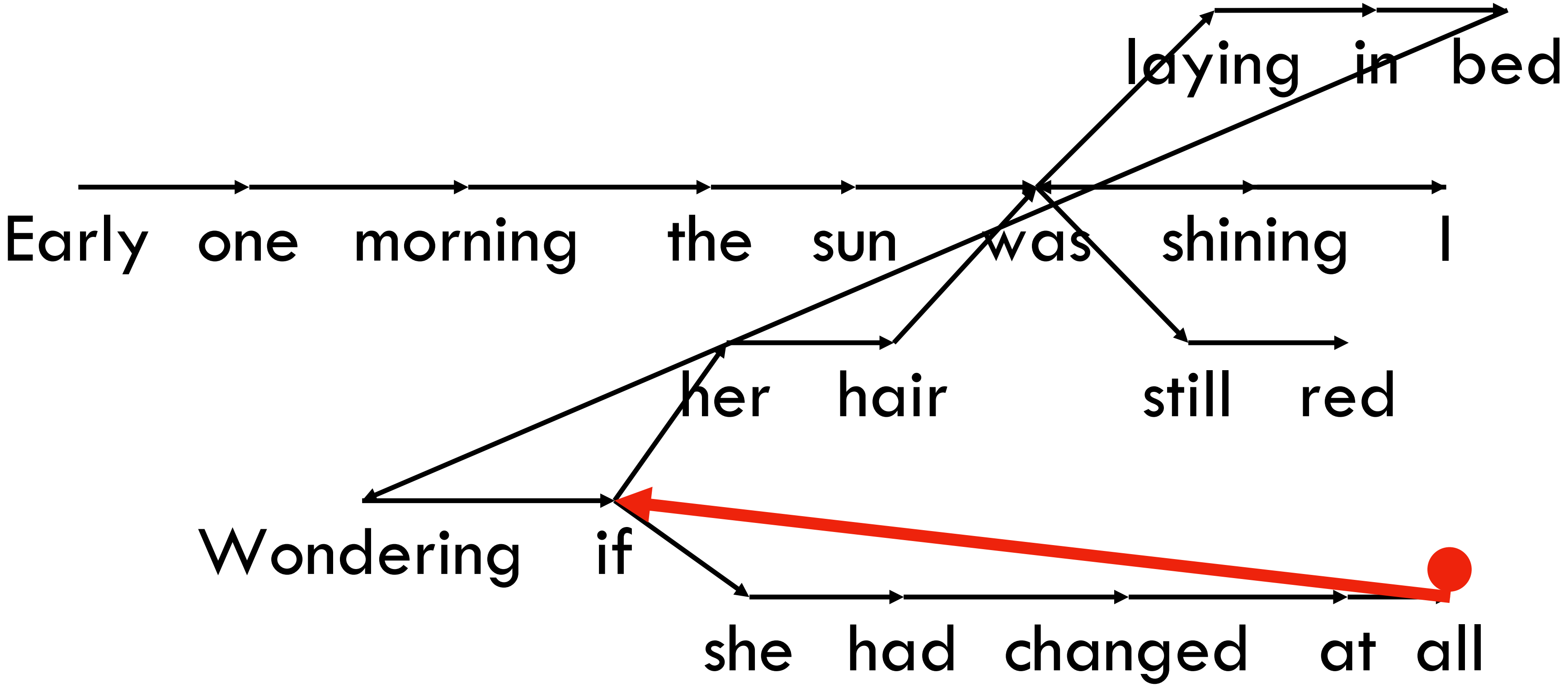
slide from Steve Seitz's [video]

Early one morning the sun was shining I was laying in bed
Wondering if she had changed at



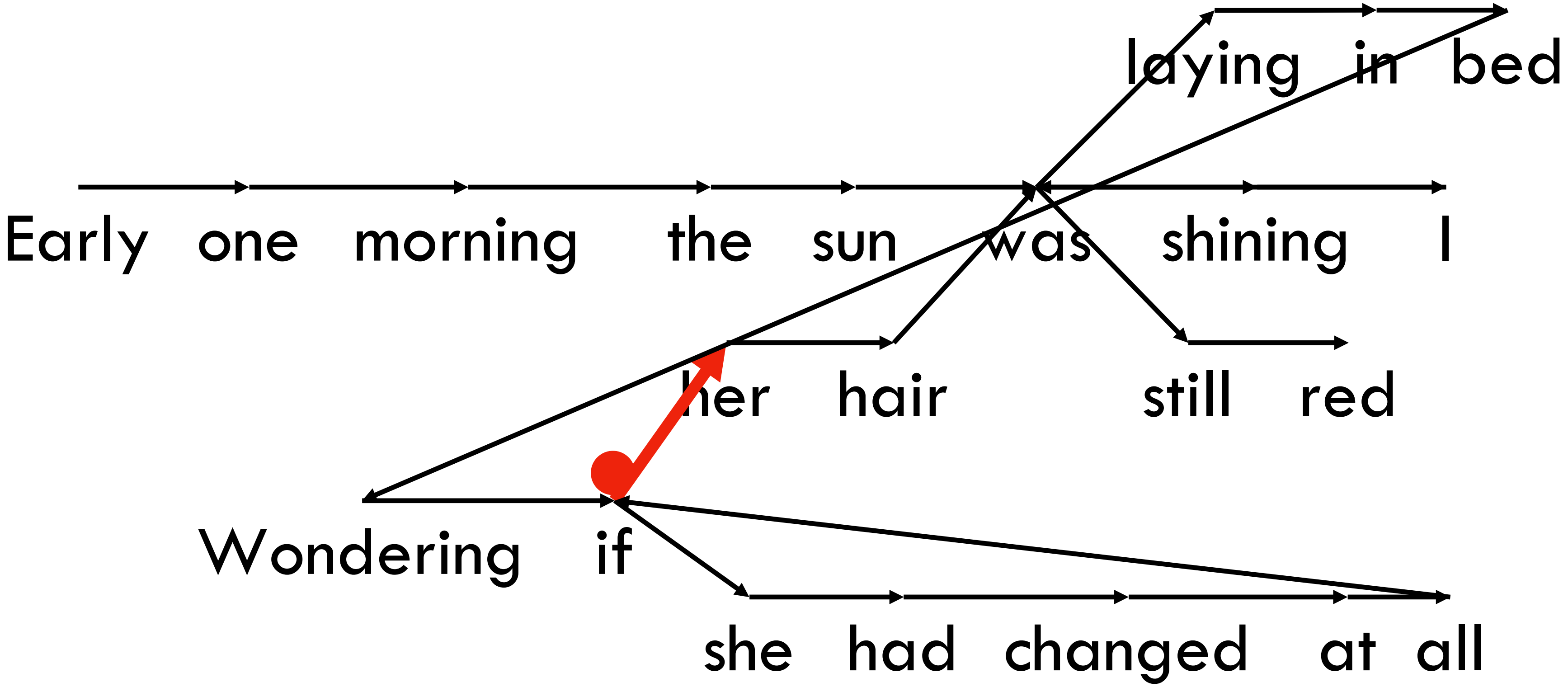
slide from Steve Seitz's video

Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all



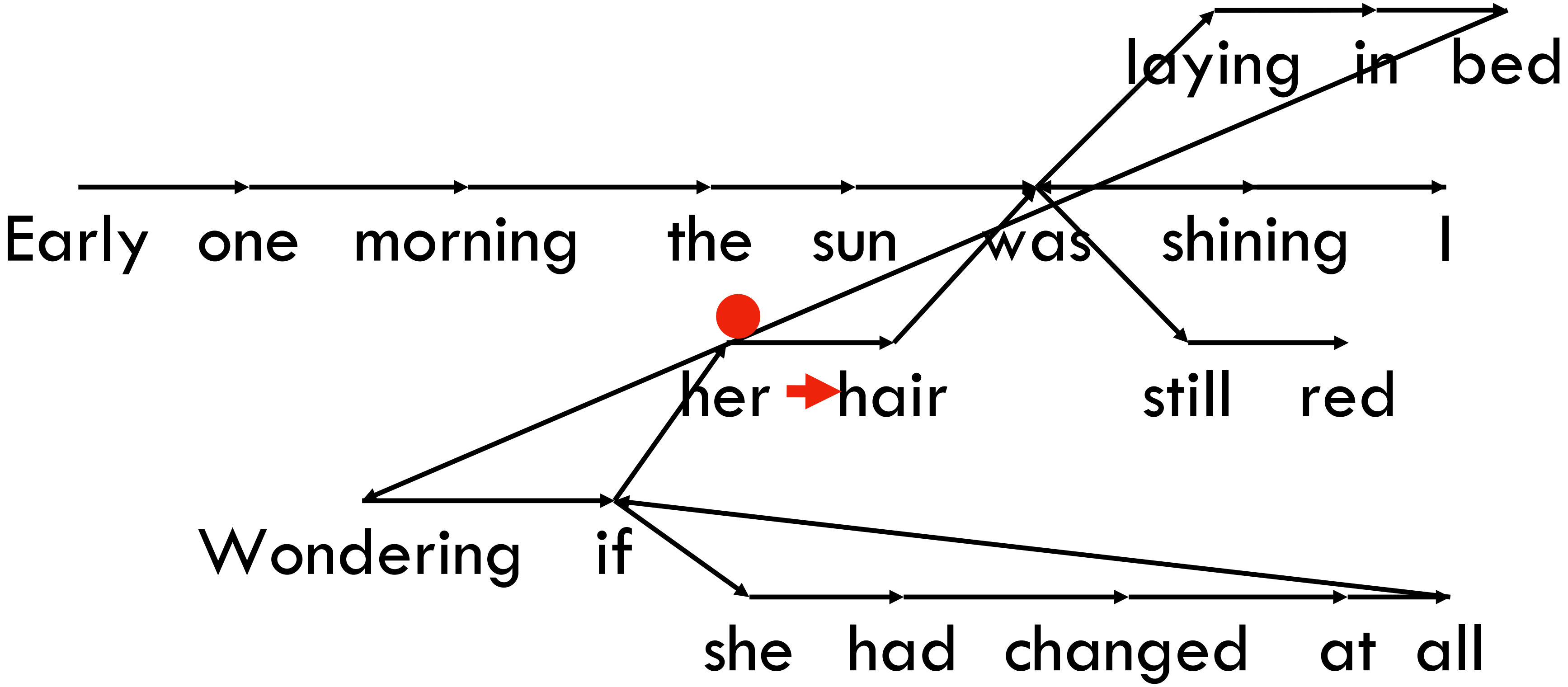
slide from Steve Seitz's video

Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if



slide from Steve Seitz's video

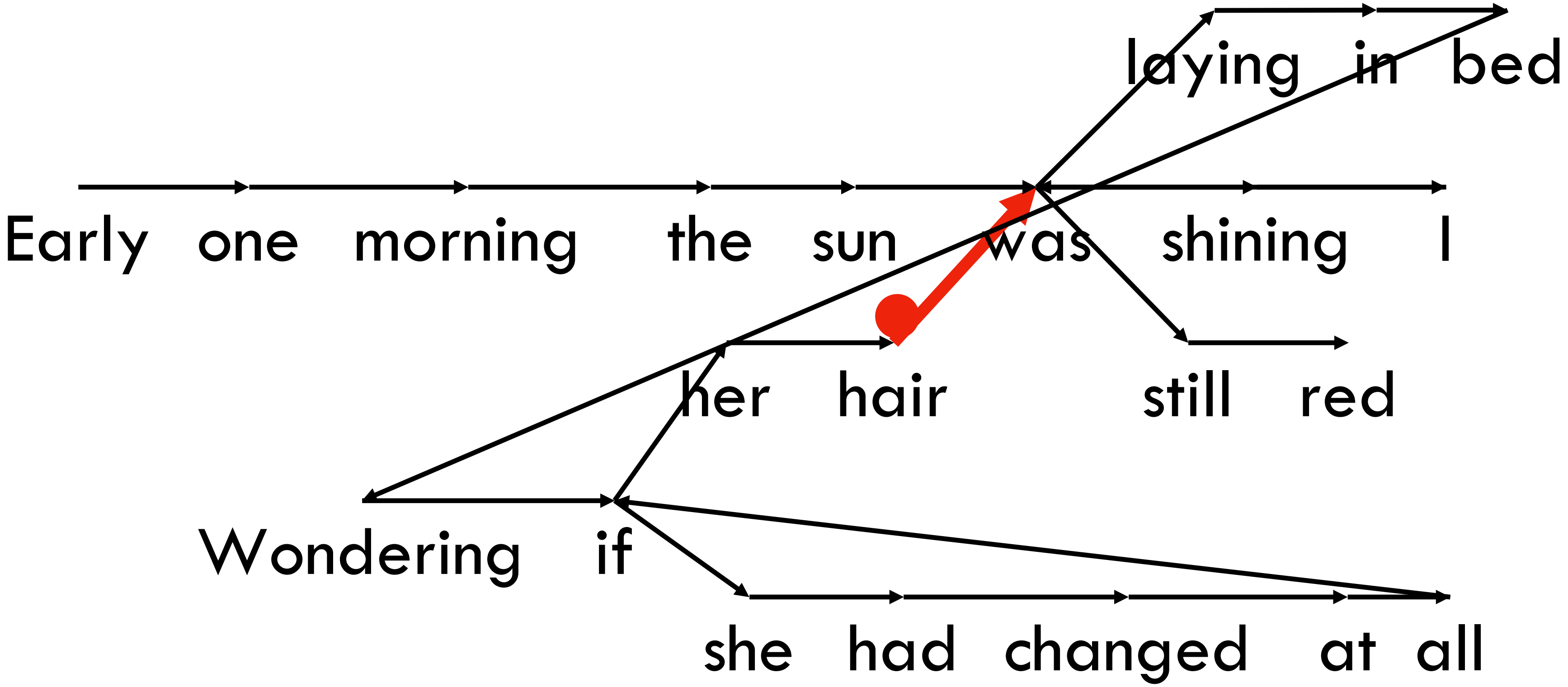
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her



slide from Steve Seitz's video

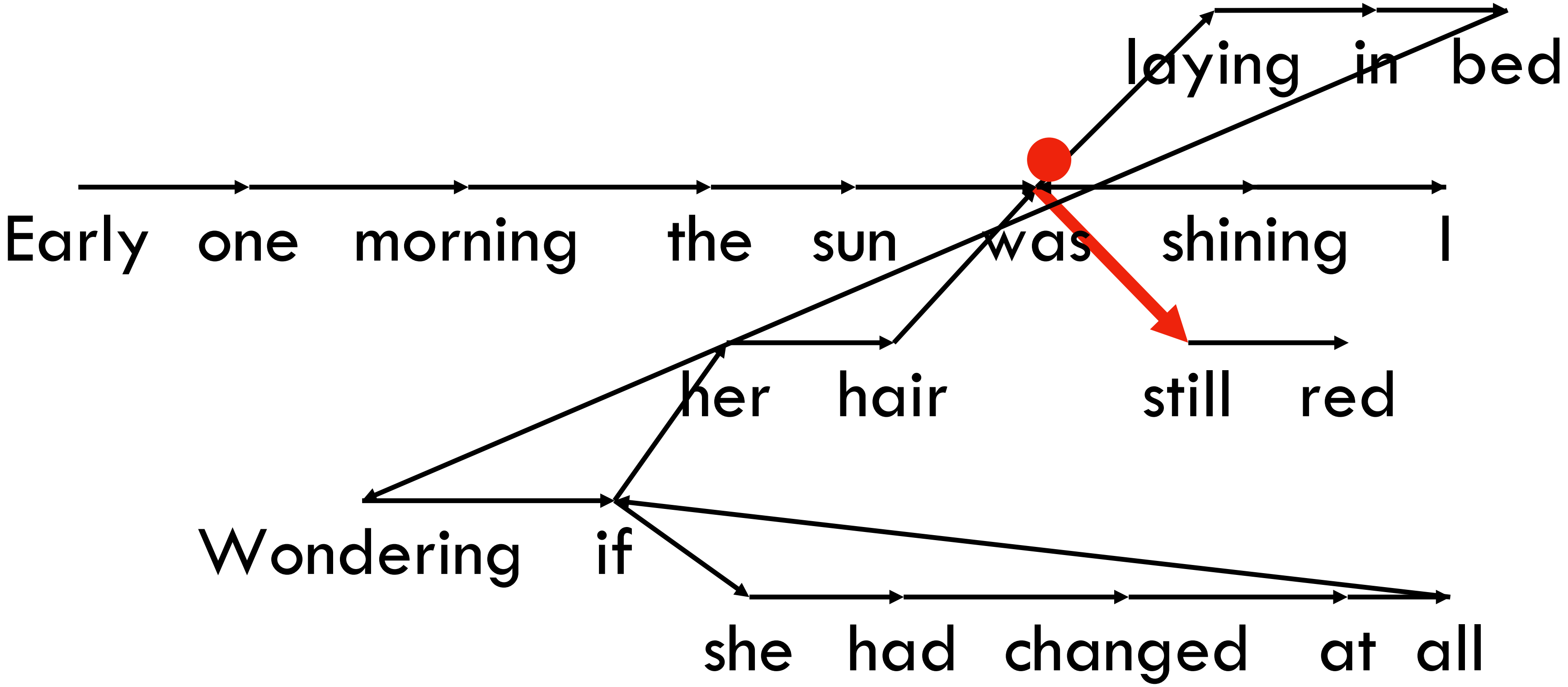
Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair



slide from Steve Seitz's [video]

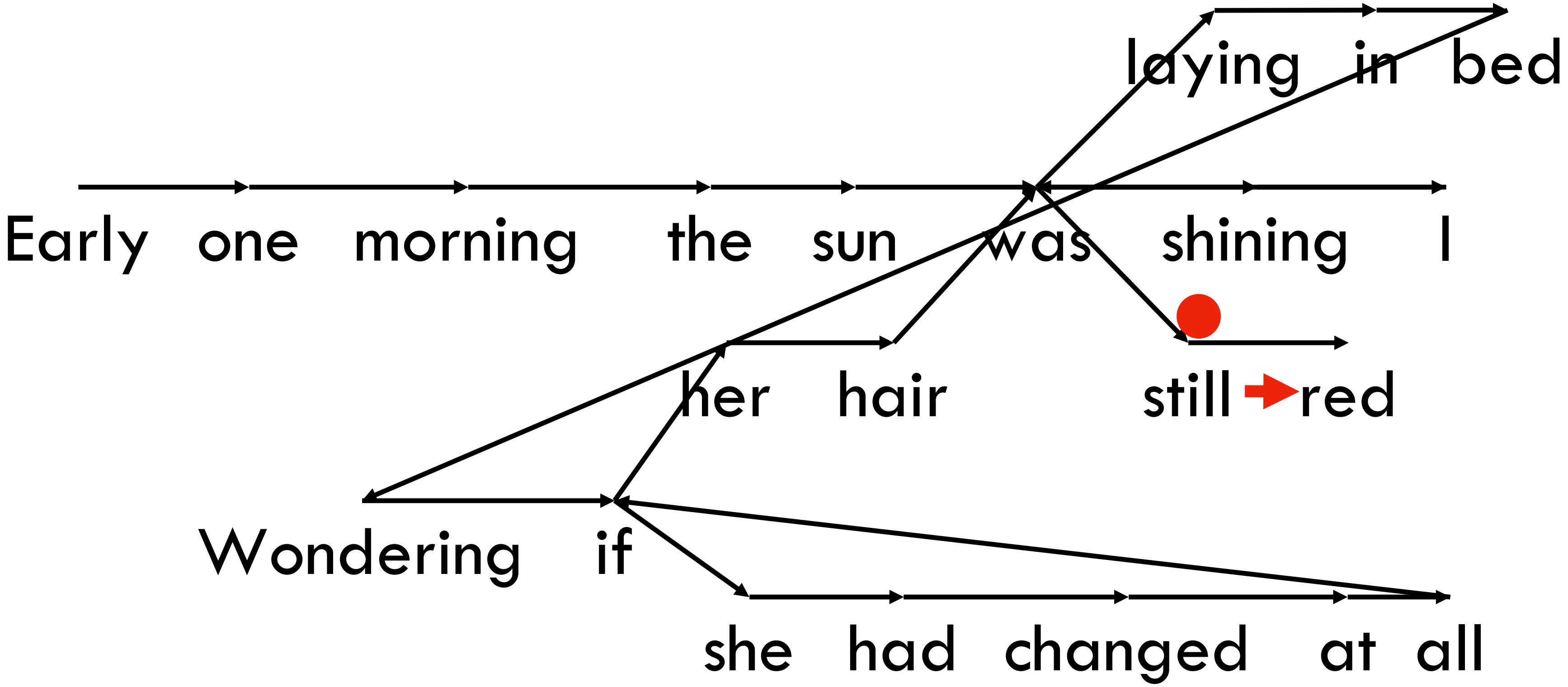
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was



slide from Steve Seitz's [video]()

Early one morning the sun was shining I was laying in bed

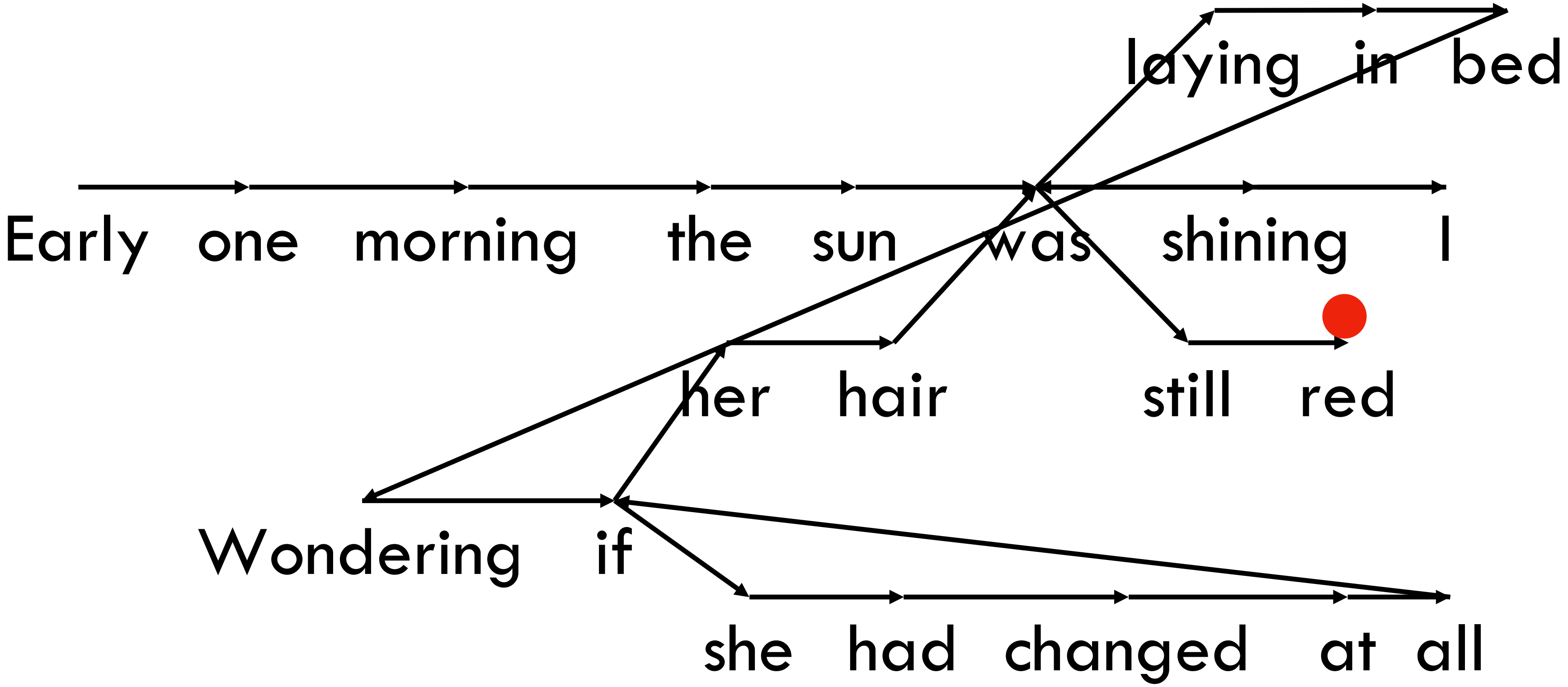
Wondering if she had changed at all if her hair was still



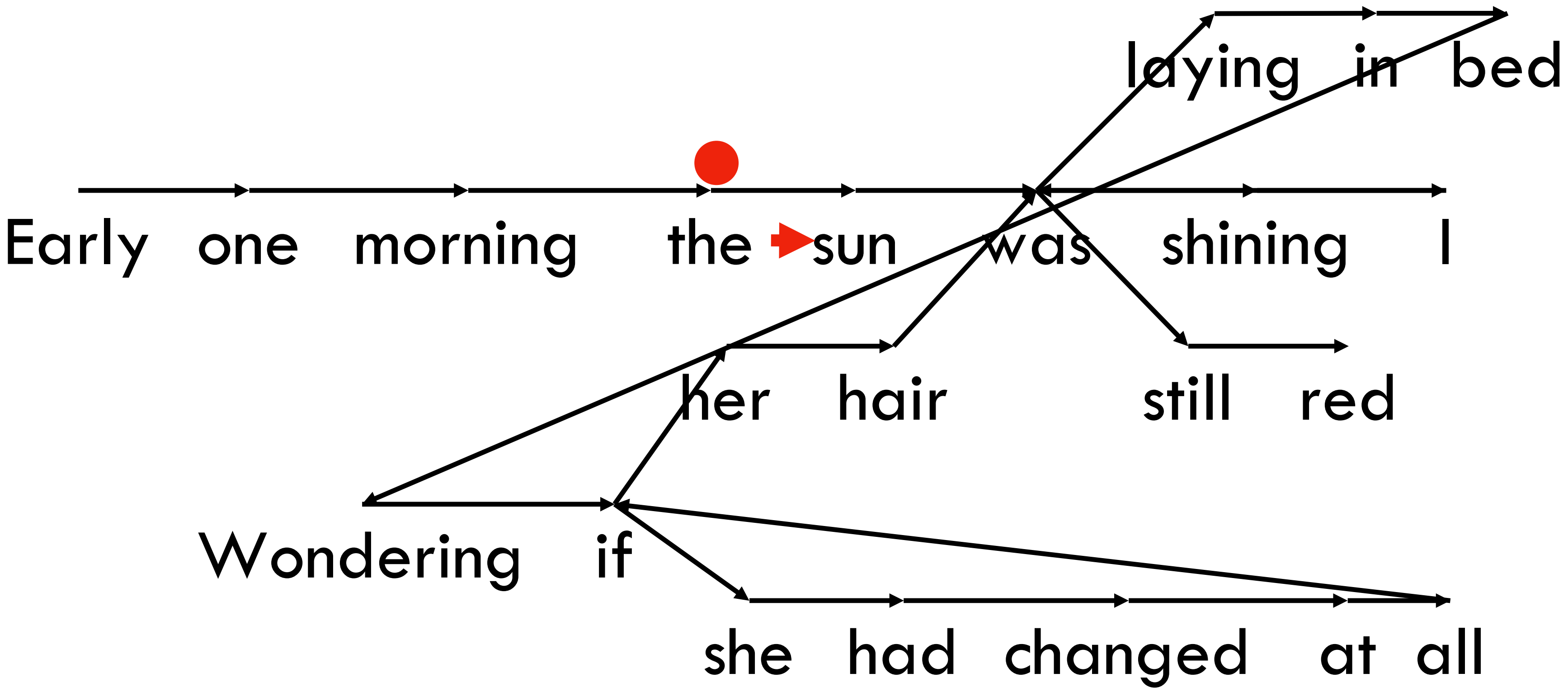
slide from Steve Seitz's [video]

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

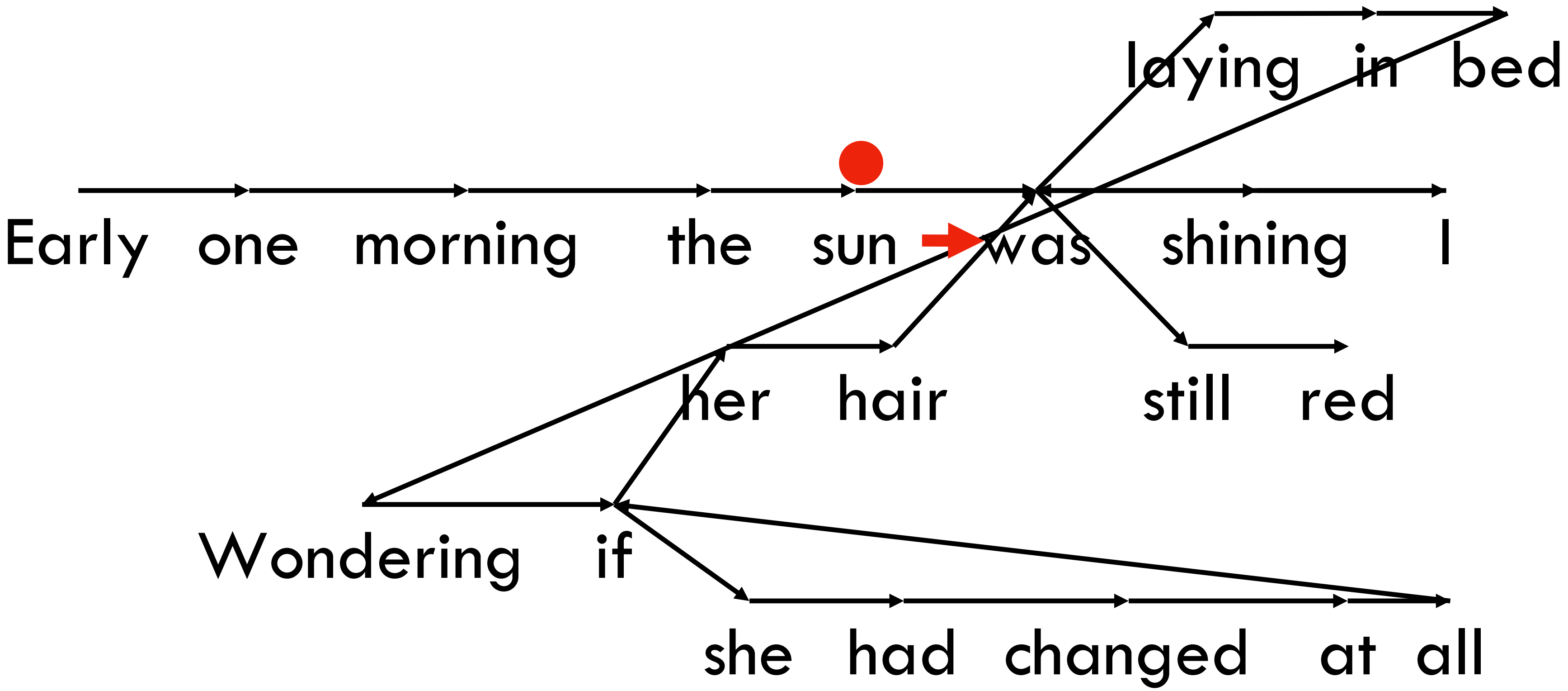


the



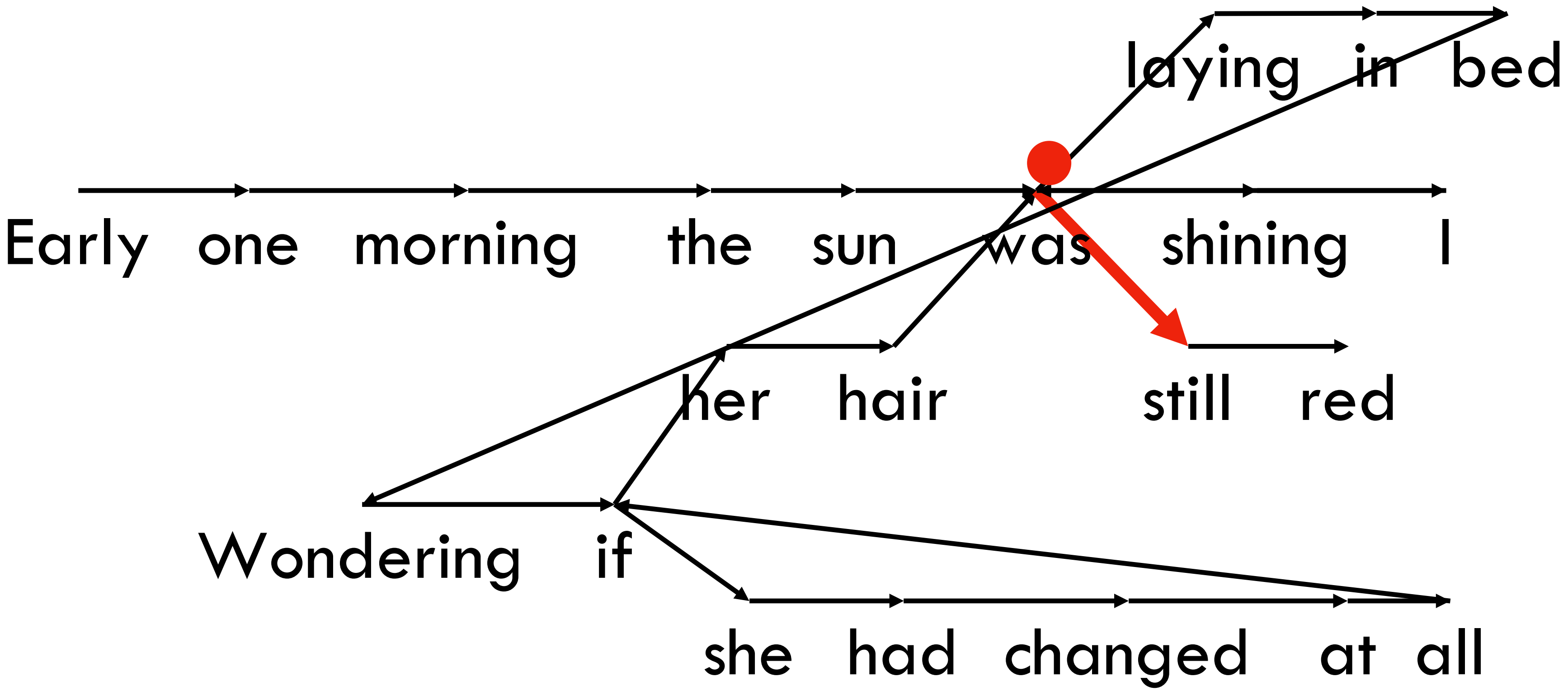
slide from Steve Seitz's [video]

the sun



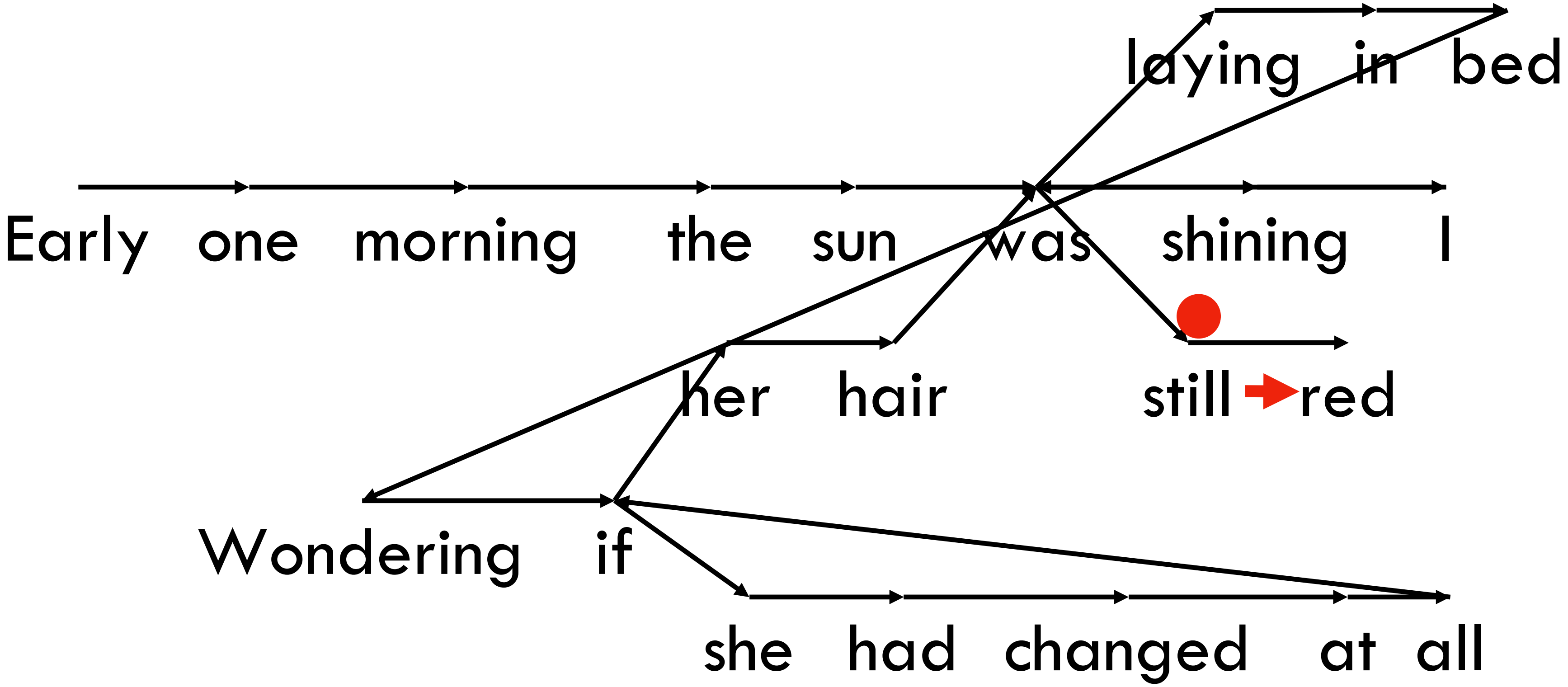
slide from Steve Seitz's [video]

the sun was



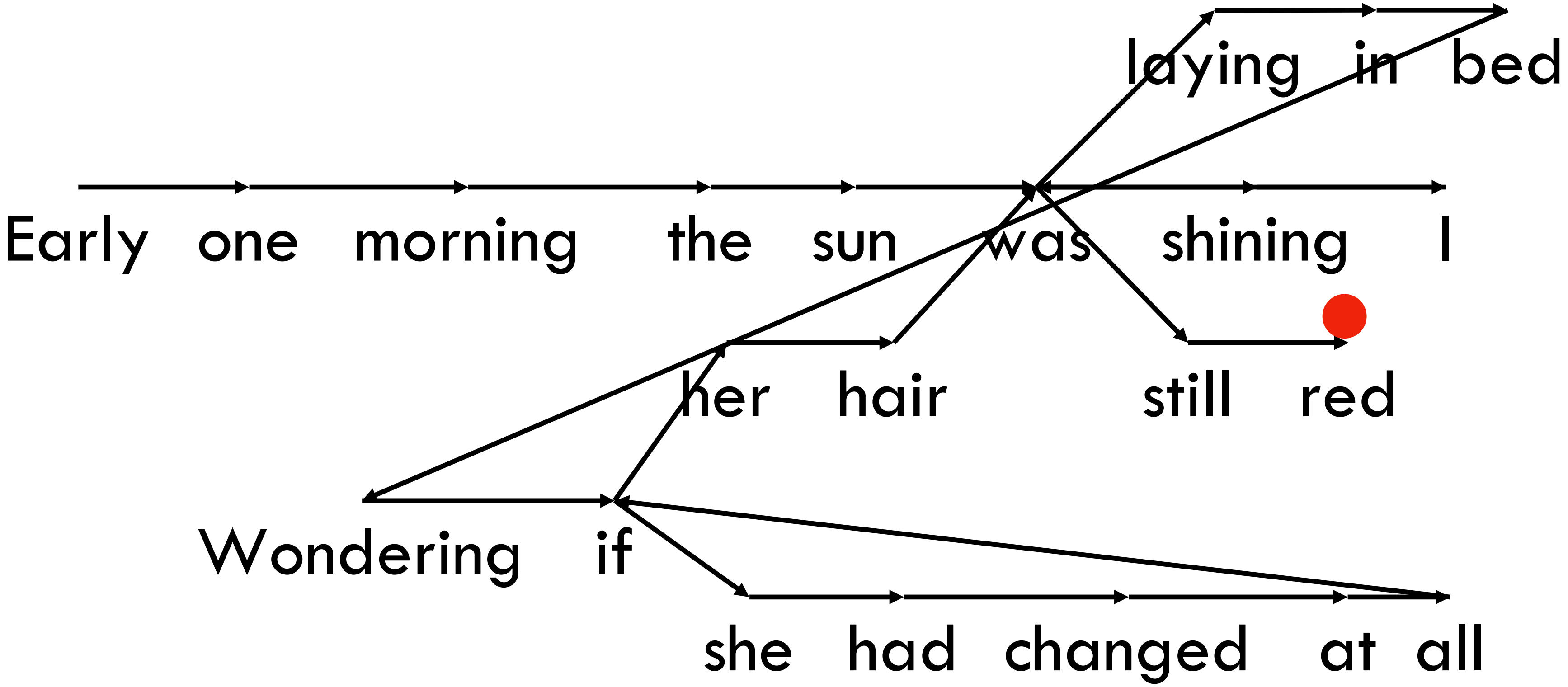
slide from Steve Seitz's [video]

the sun was still



slide from Steve Seitz's [video]

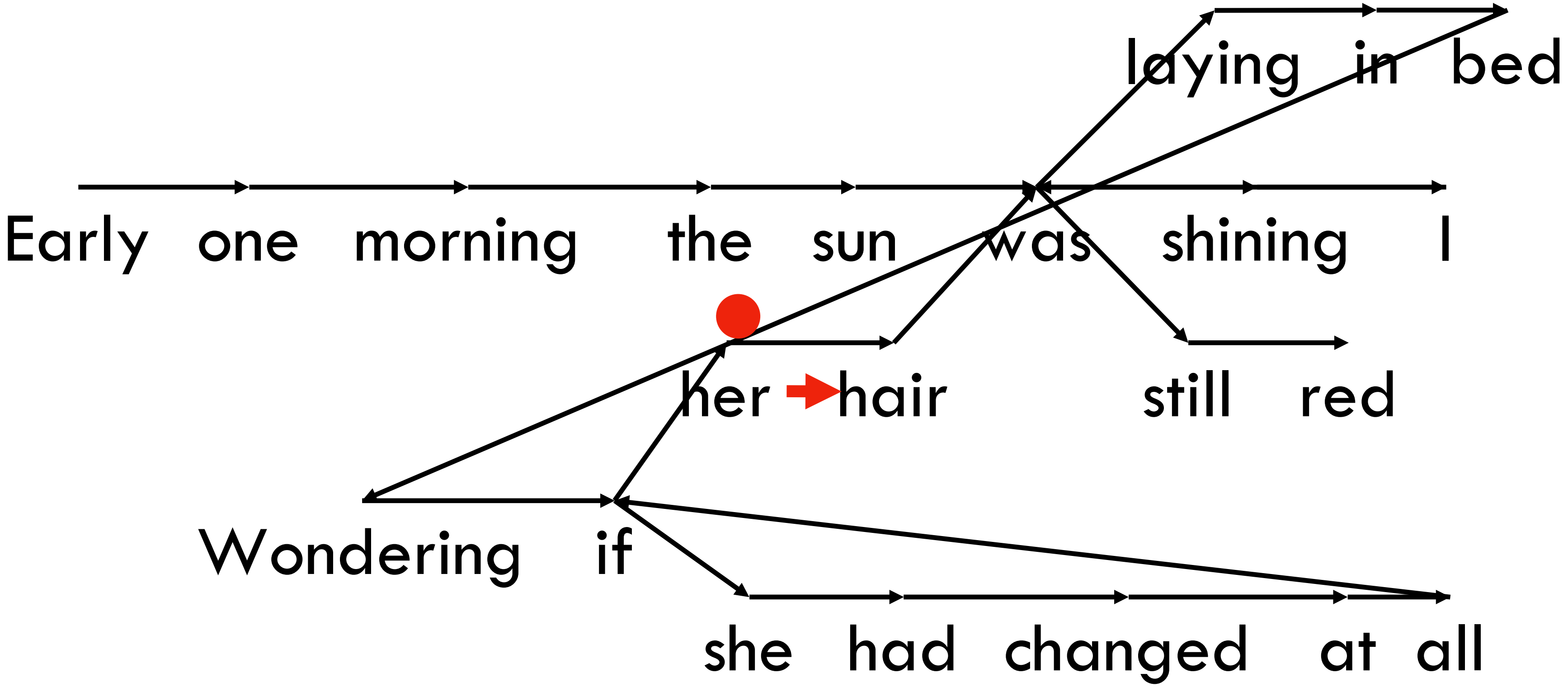
the sun was still red



slide from Steve Seitz's [video]

the sun was still red

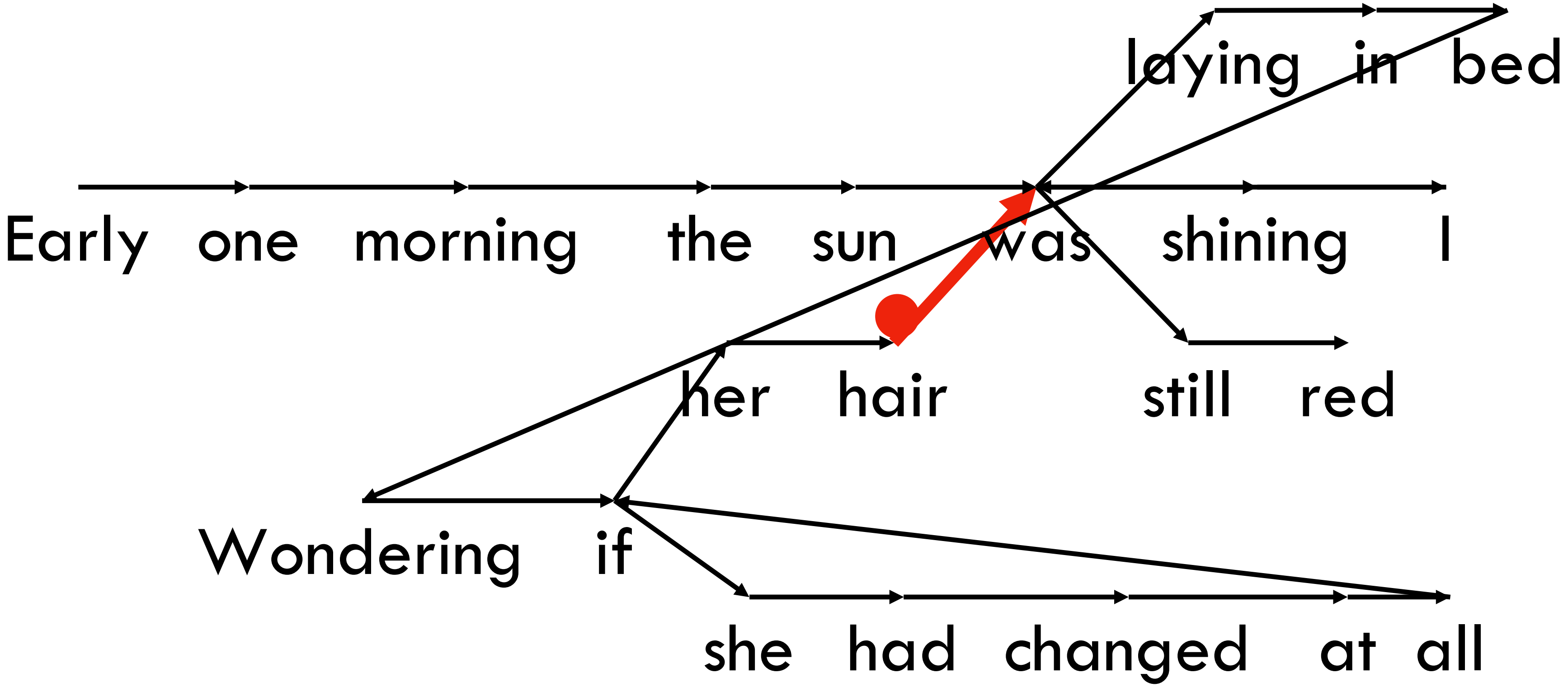
her



slide from Steve Seitz's [video]

the sun was still red

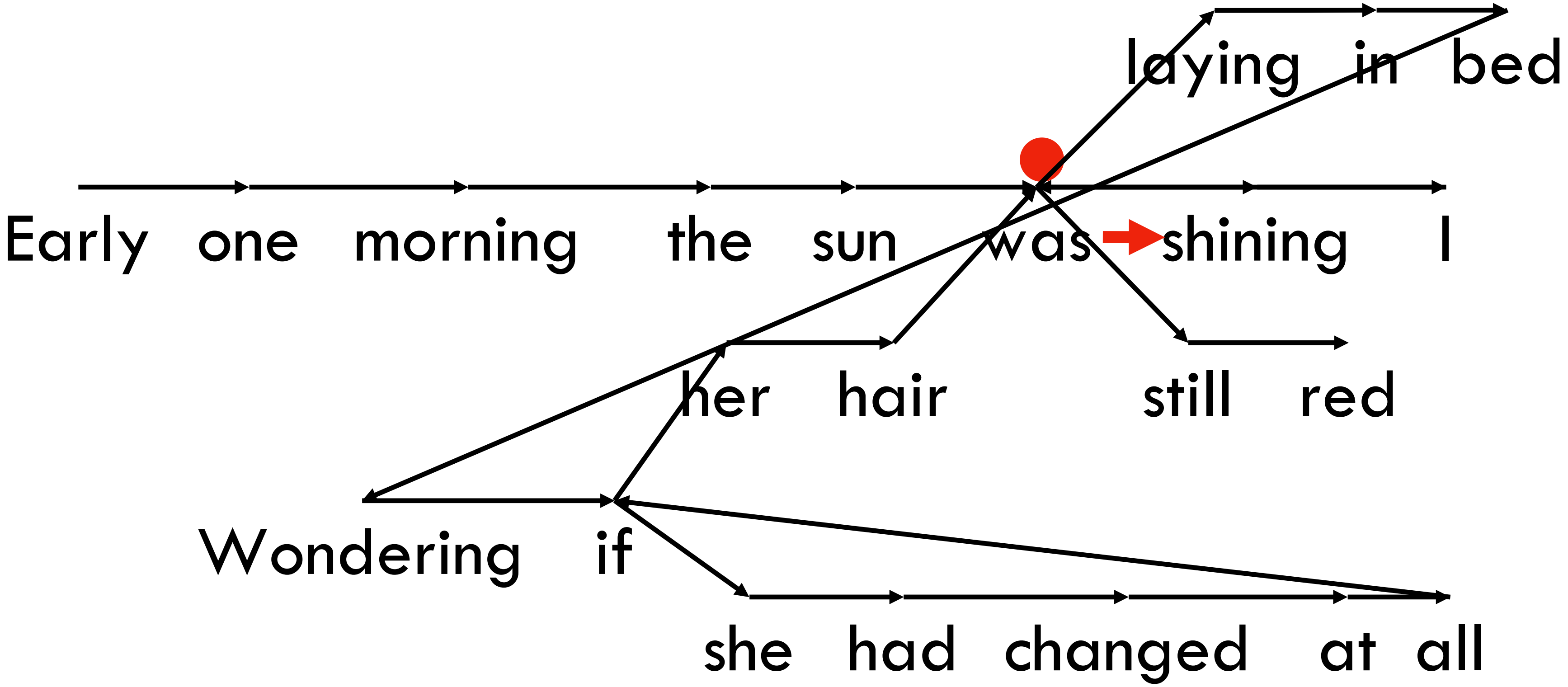
her hair



slide from Steve Seitz's video

the sun was still red

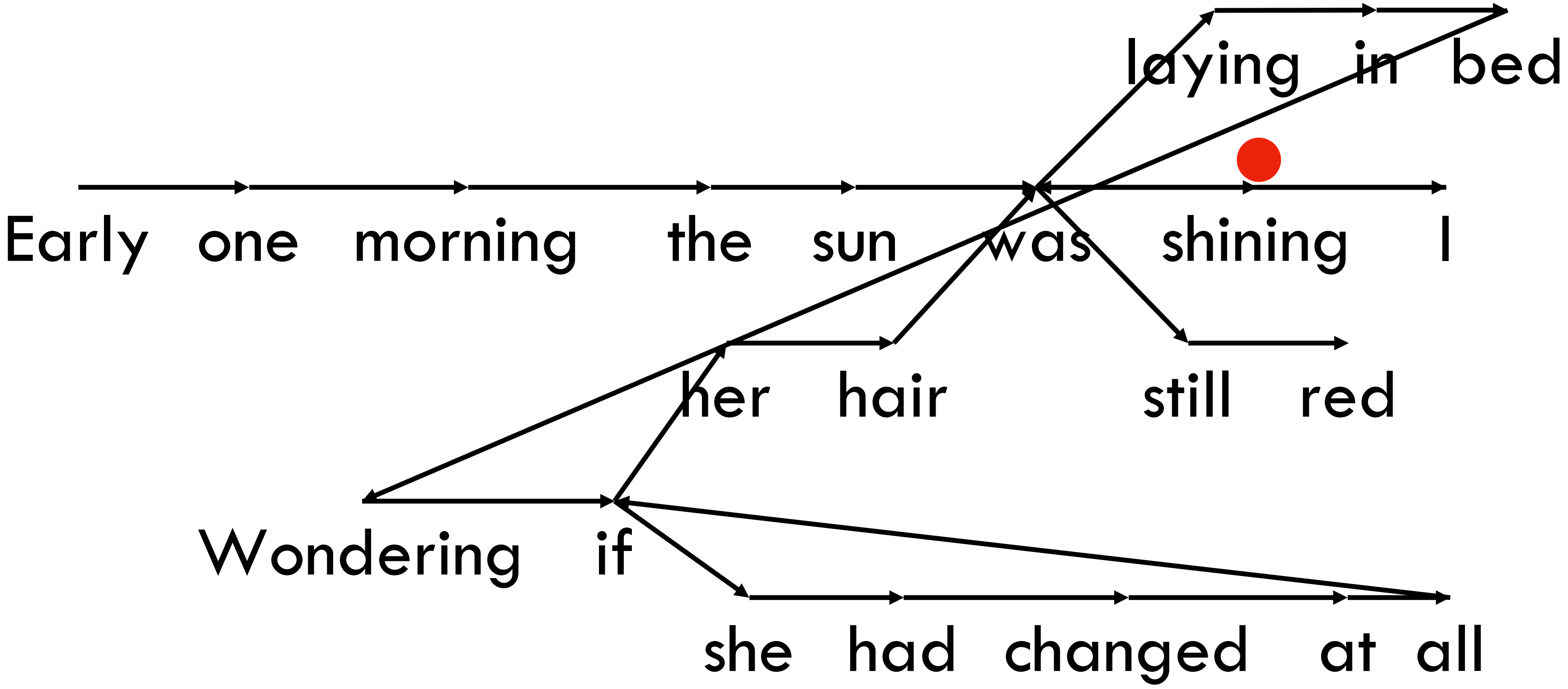
her hair was



slide from Steve Seitz’s [video]

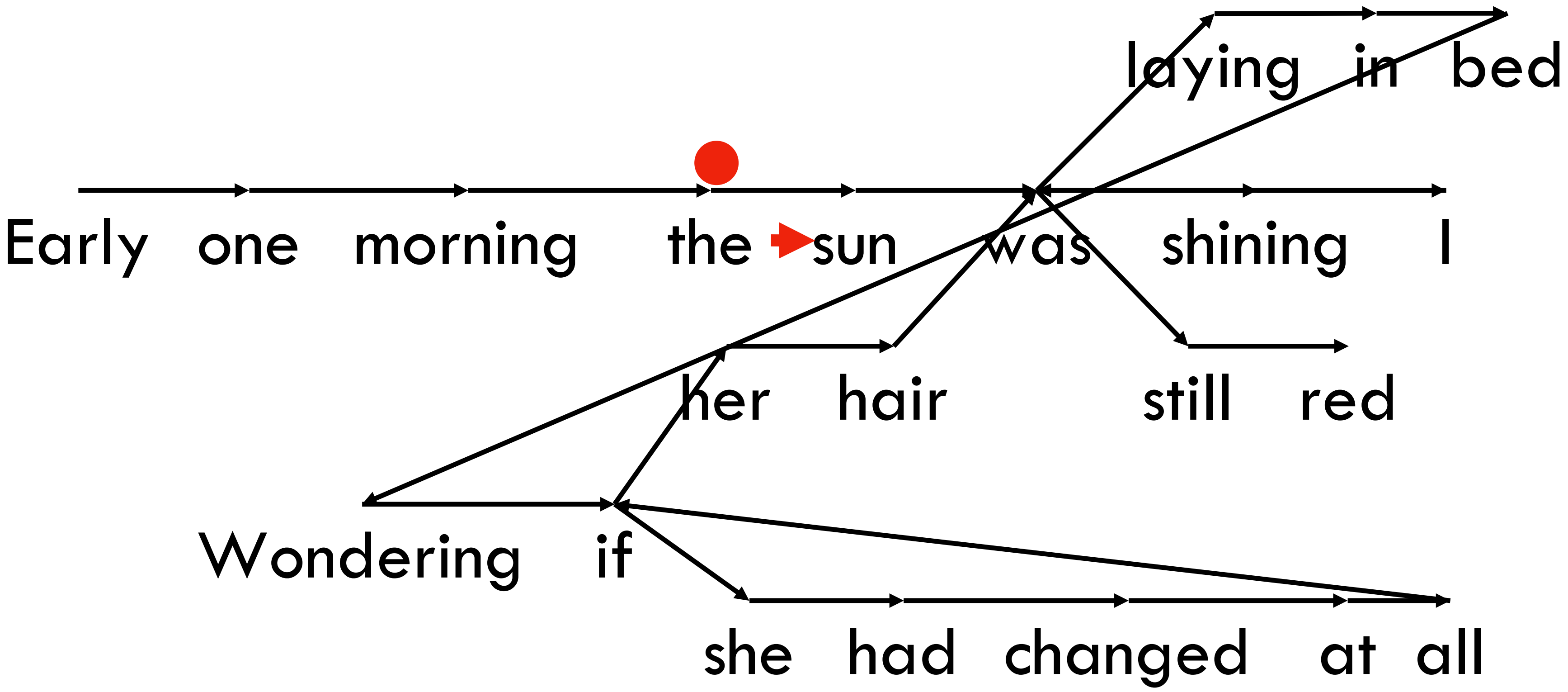
the sun was still red

her hair was shining



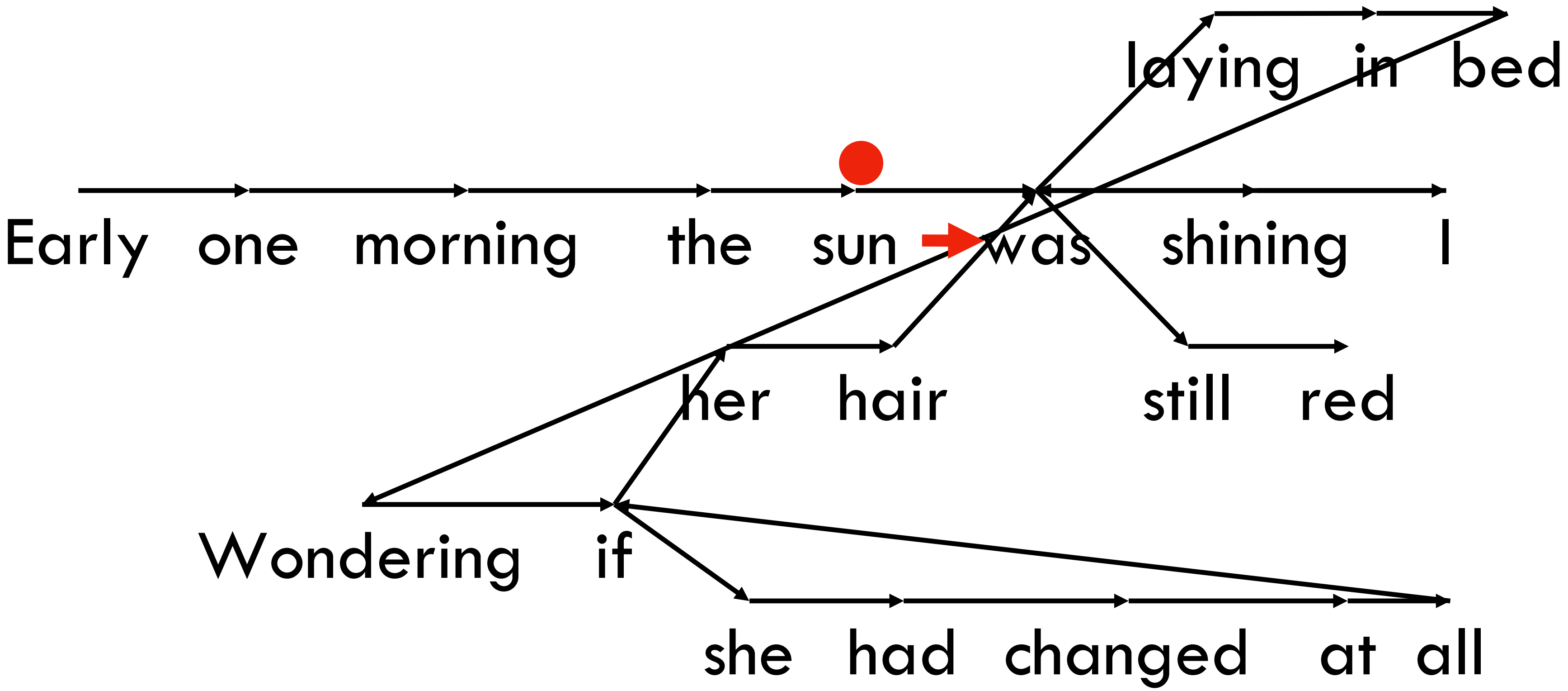
slide from Steve Seitz's video

the



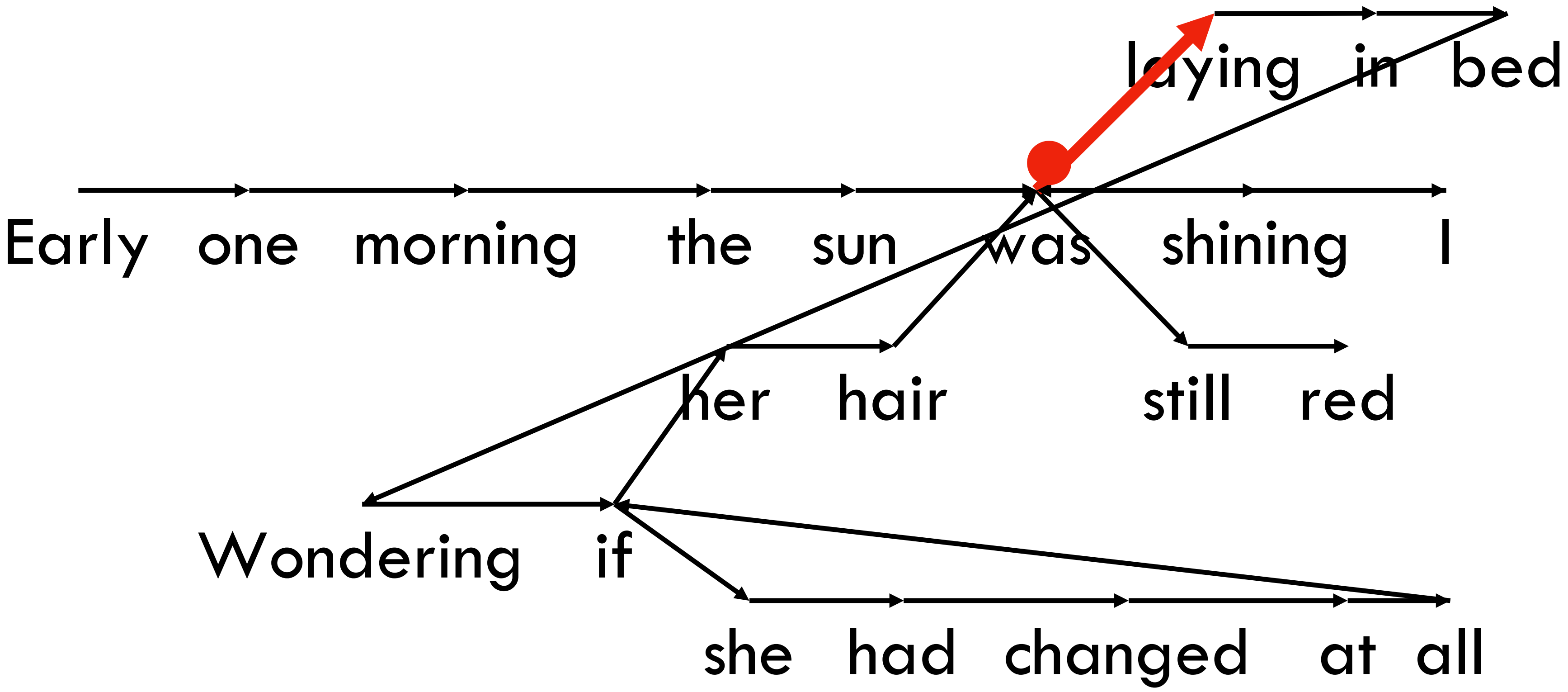
slide from Steve Seitz's [video]

the sun



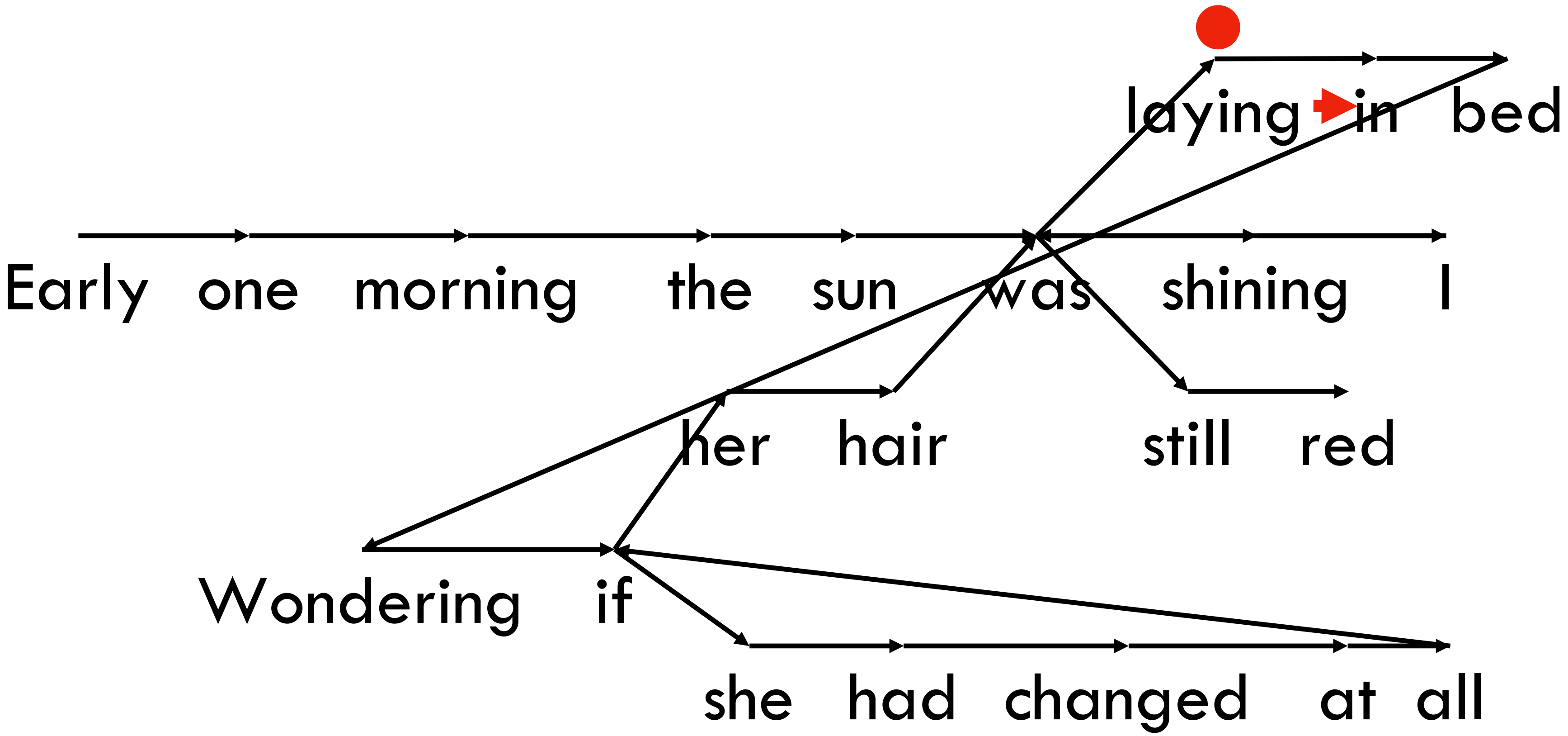
slide from Steve Seitz's [video]()

the sun was



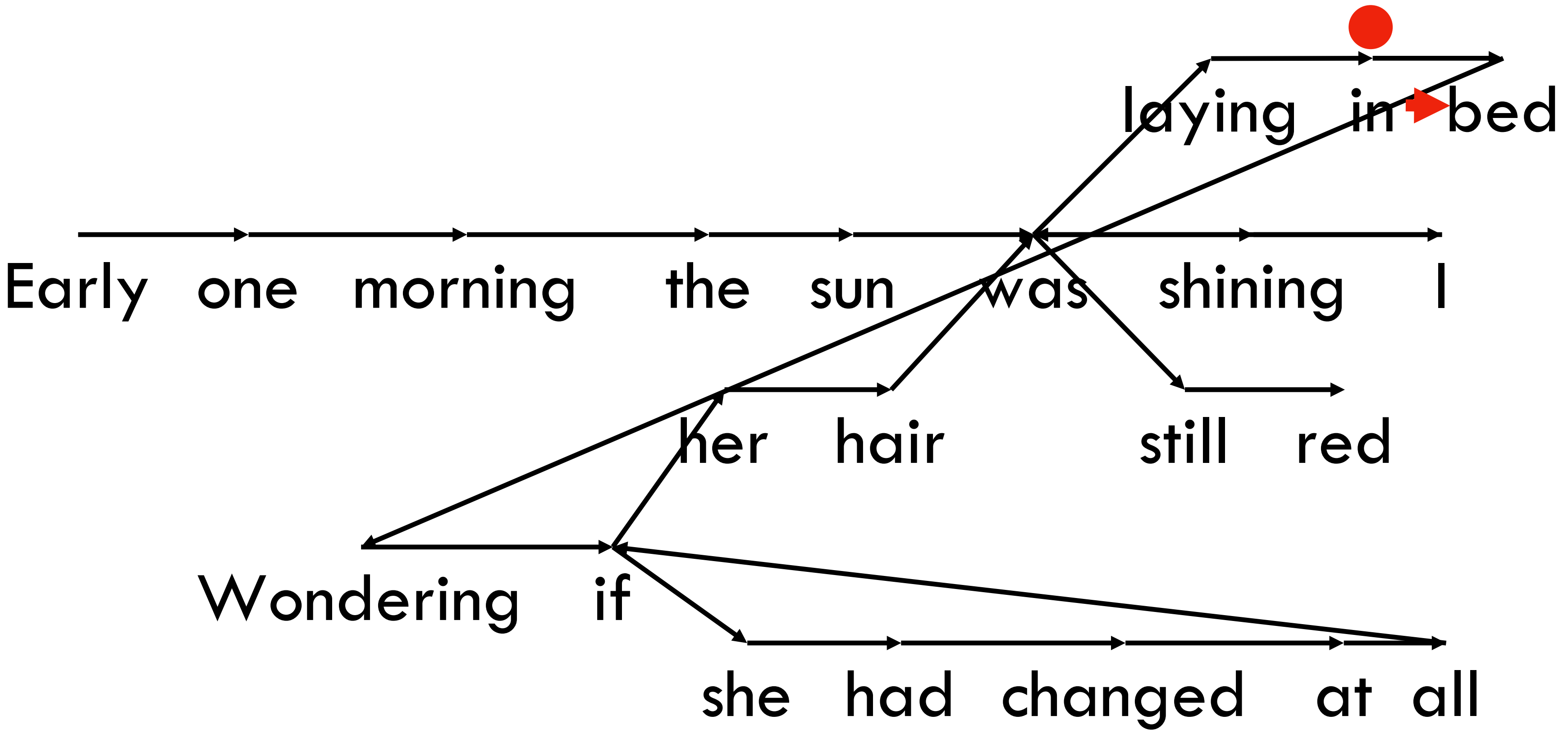
slide from Steve Seitz's [video]

the sun was laying



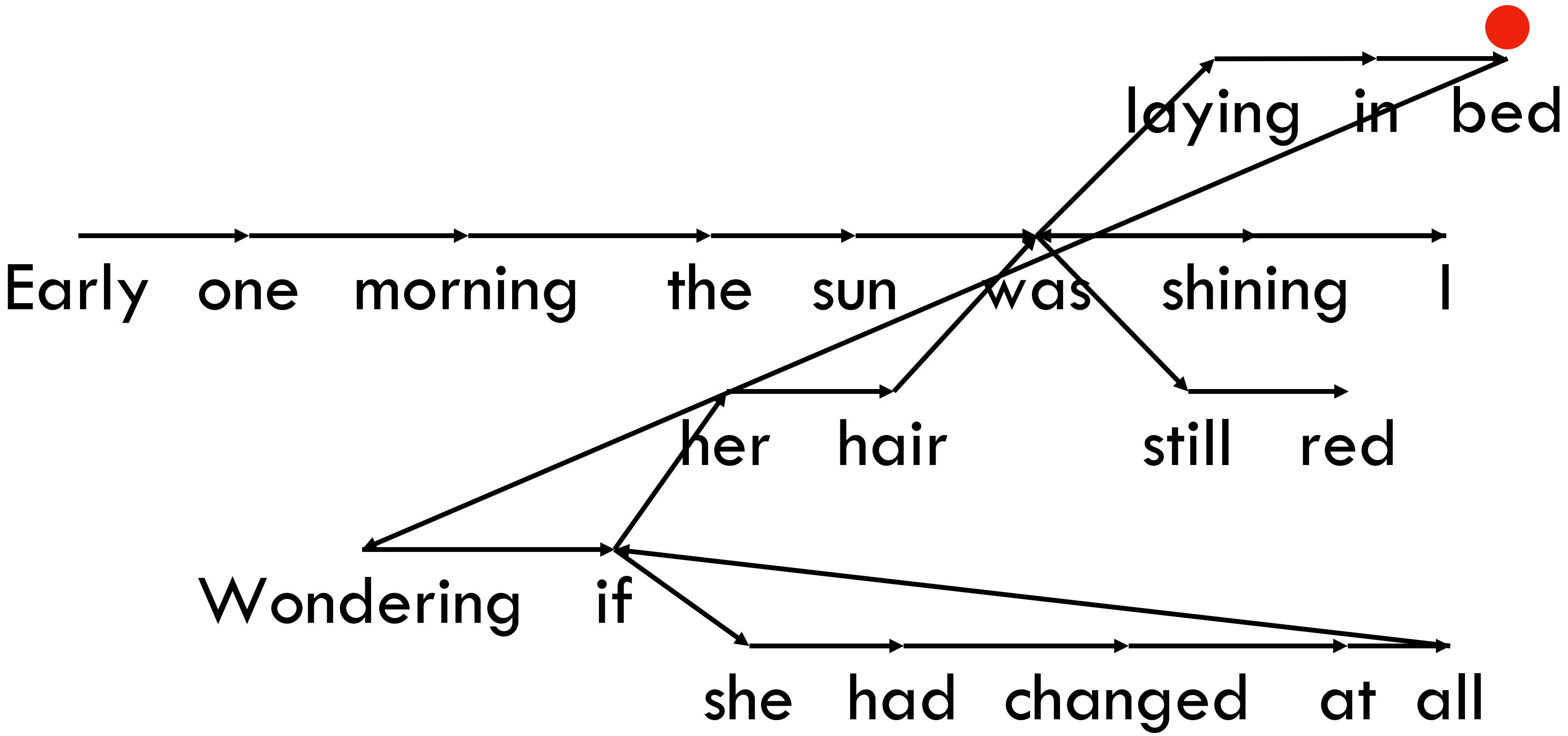
slide from Steve Seitz's [video]

the sun was laying in



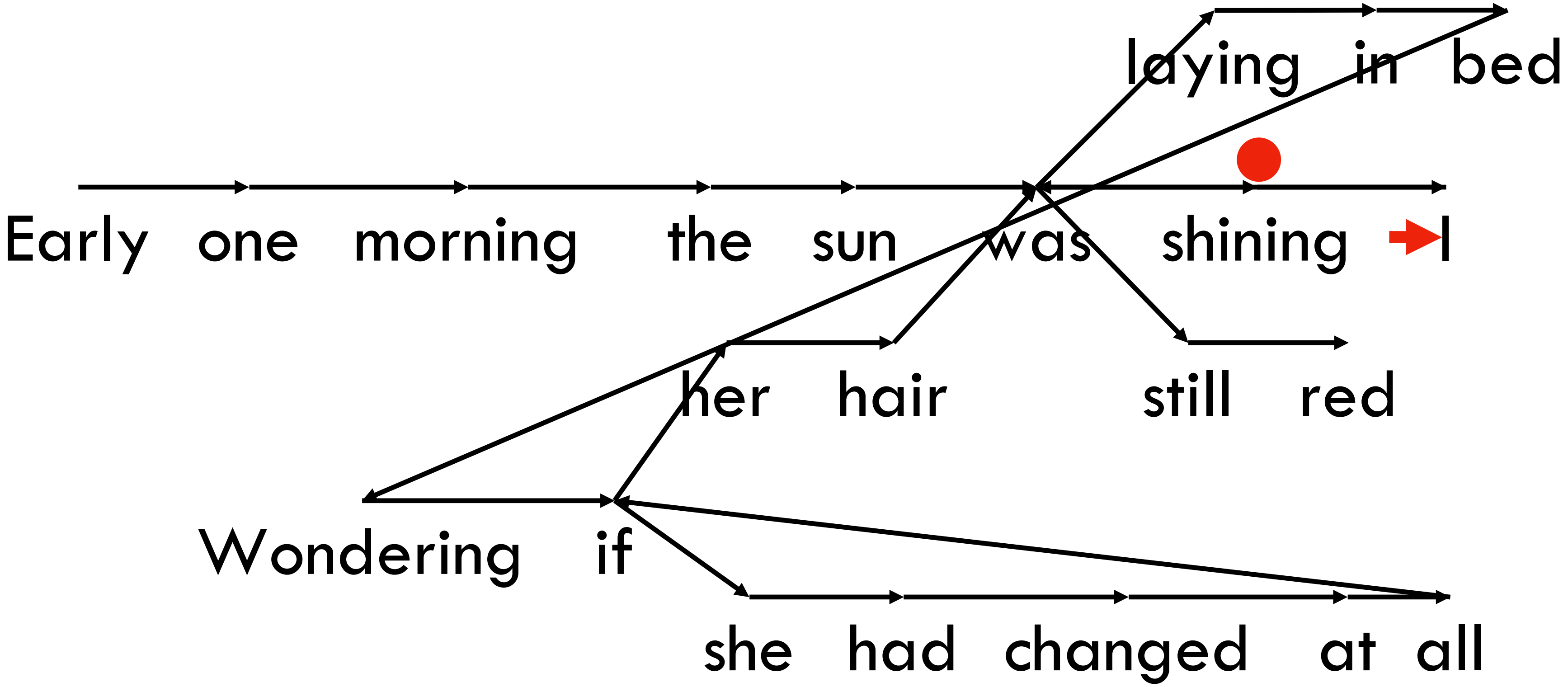
slide from Steve Seitz's video

the sun was laying in bed



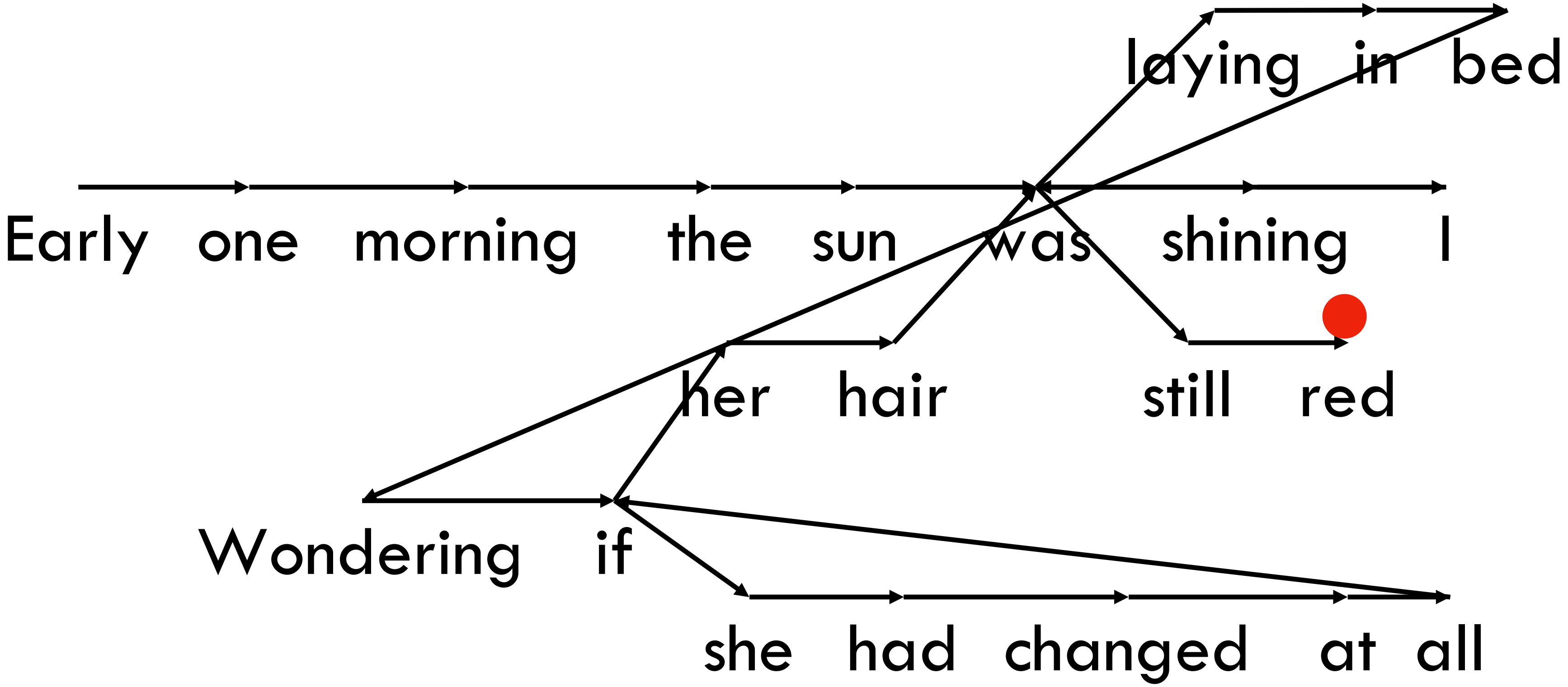
slide from Steve Seitz's [video]

I was shining I was shining



slide from Steve Seitz's [video]

I was shining I was shining I was still red



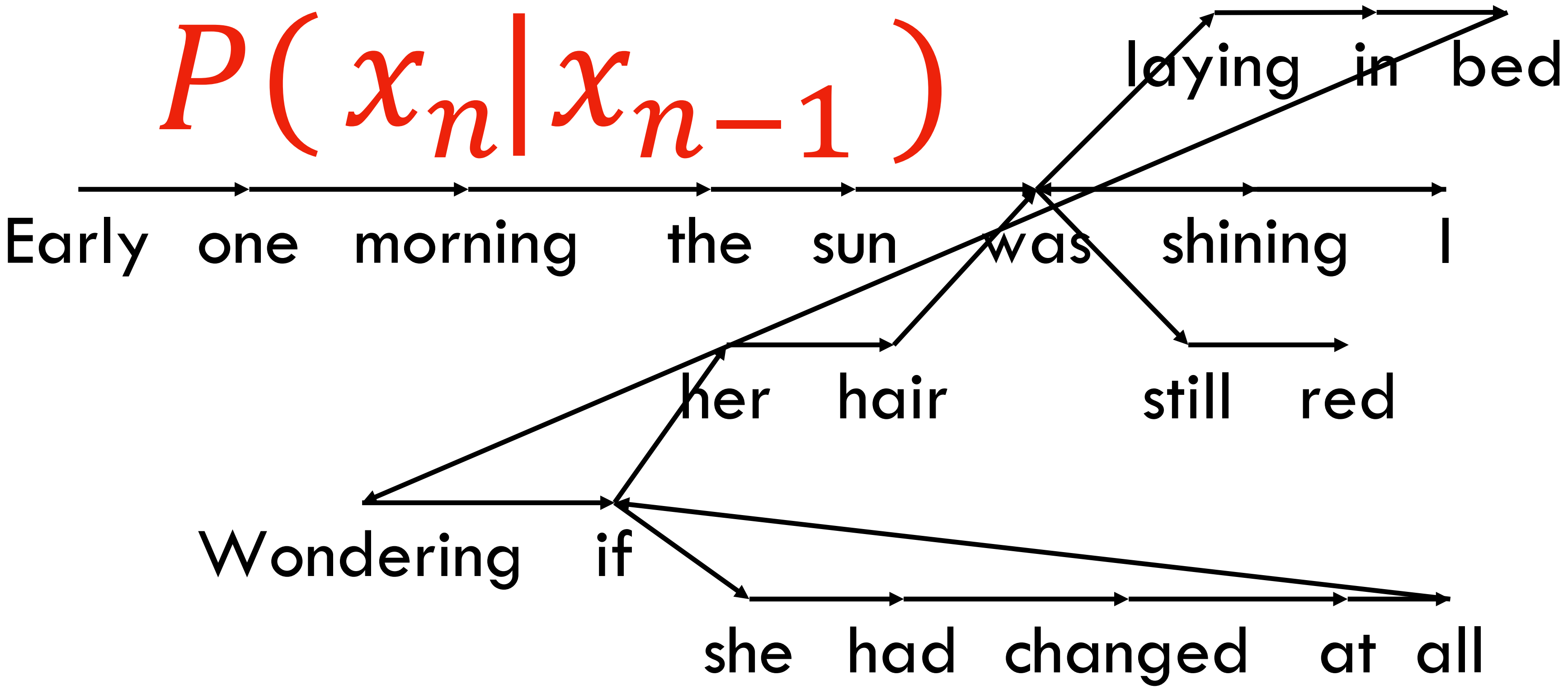
she was standing on the side of my mind …

side of my shoes heading out of my face …

one of my chair said our lives together …

slide from Steve Seitz's [video]

$P(x_n|x_{n-1})$



slide from Steve Seitz's video

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

Early one morning

trigrams

slide from Steve Seitz's [video]

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

Early one morning

one morning the

trigrams

slide from Steve Seitz's [video]

Early one $\boxed{\text{morning the sun}}$ was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

Early one morning

one morning the

morning the sun

trigrams

slide from Steve Seitz's [video]

Early one morning $\boxed{\text{the sun was}}$ shining I was laying in bed

Wondering if she had changed at all if her hair was still red

Early one morning
one morning the
morning the sun
the sun was

trigrams

slide from Steve Seitz's [video]

$$P(x_n | x_{n-1}, x_{n-2})$$

Early one	→ morning
one morning	→ the
morning the	→ sun
the sun	→ was
sun was	→ shining
was shining	→ I
shining I	→ was
I was	→ laying

…

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

slide from Steve Seitz's video

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

slide from Steve Seitz's video

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

slide from Steve Seitz's [video]

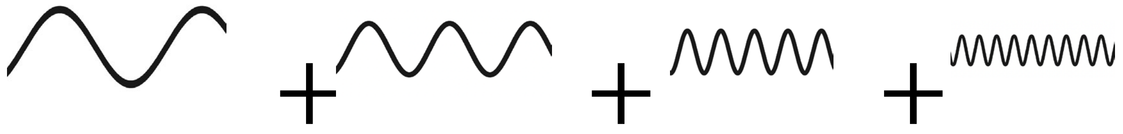
$$P(x_n | x_{n-1}, x_{n-2}, x_{n-3}, x_{n-4}, x_{n-5}, x_{n-6}, x_{n-7}, x_{n-8}, x_{n-9}, x_{n-10}, x_{n-11}, x_{n-12}, x_{n-13})$$

10^{70} combinations

slide from Steve Seitz's [video]

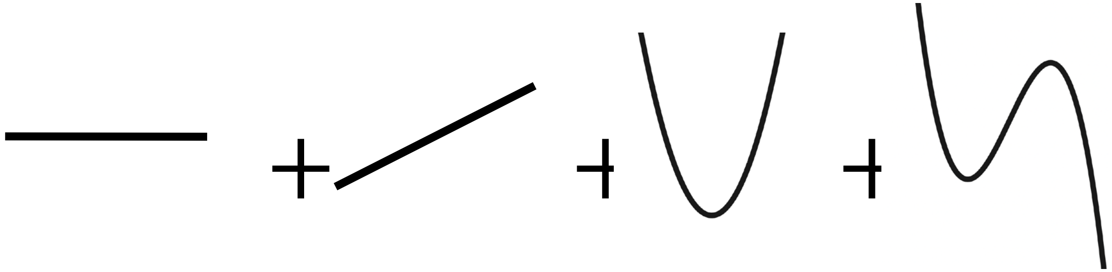
Function Approximation

Fourier Series: $f(x) =$



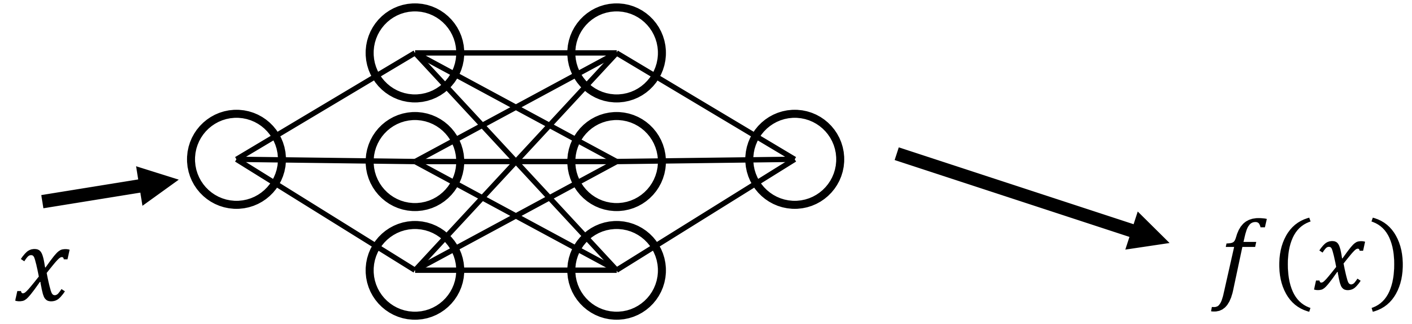
$+ \dots$

Taylor Series: $f(x) =$

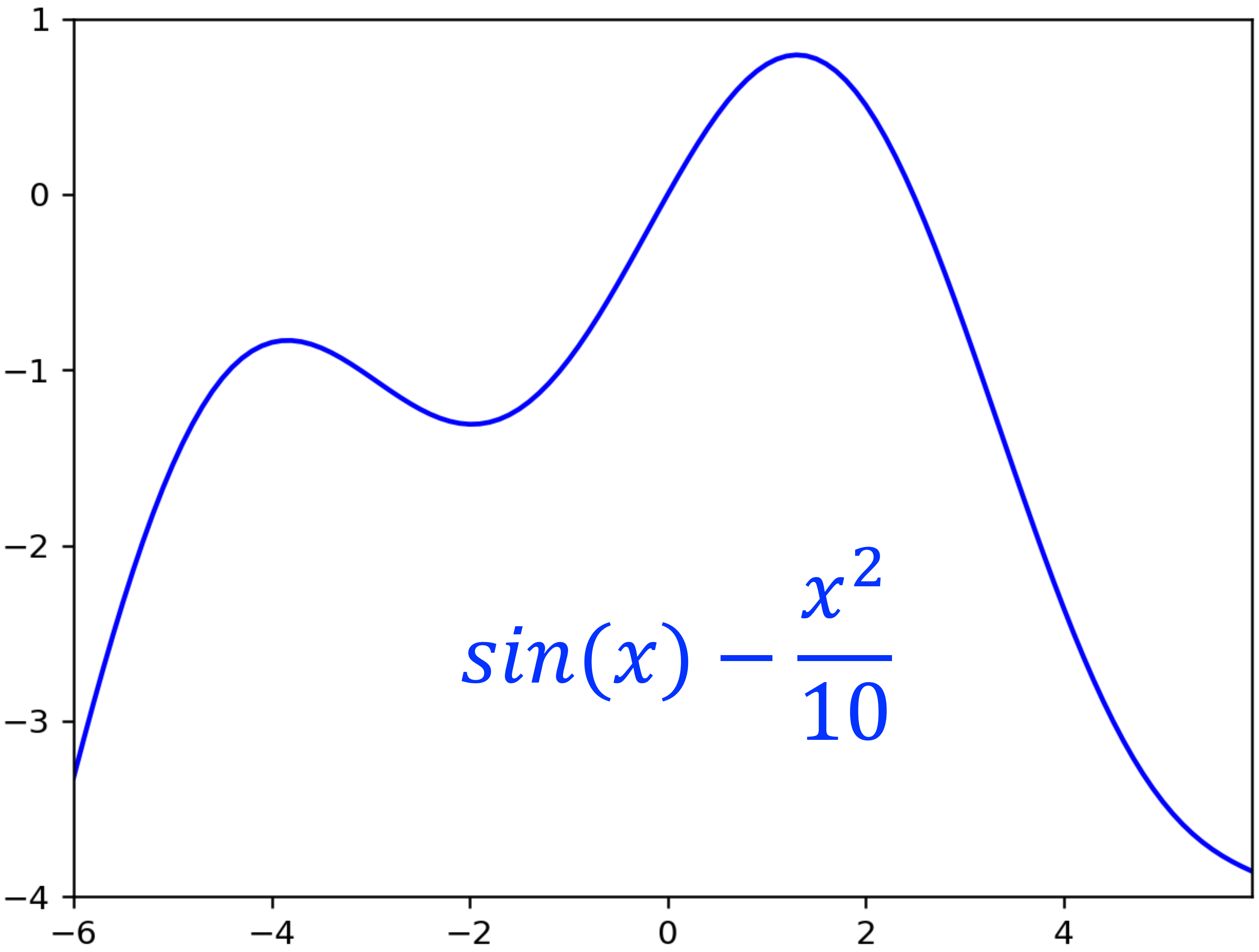


$+ \dots$

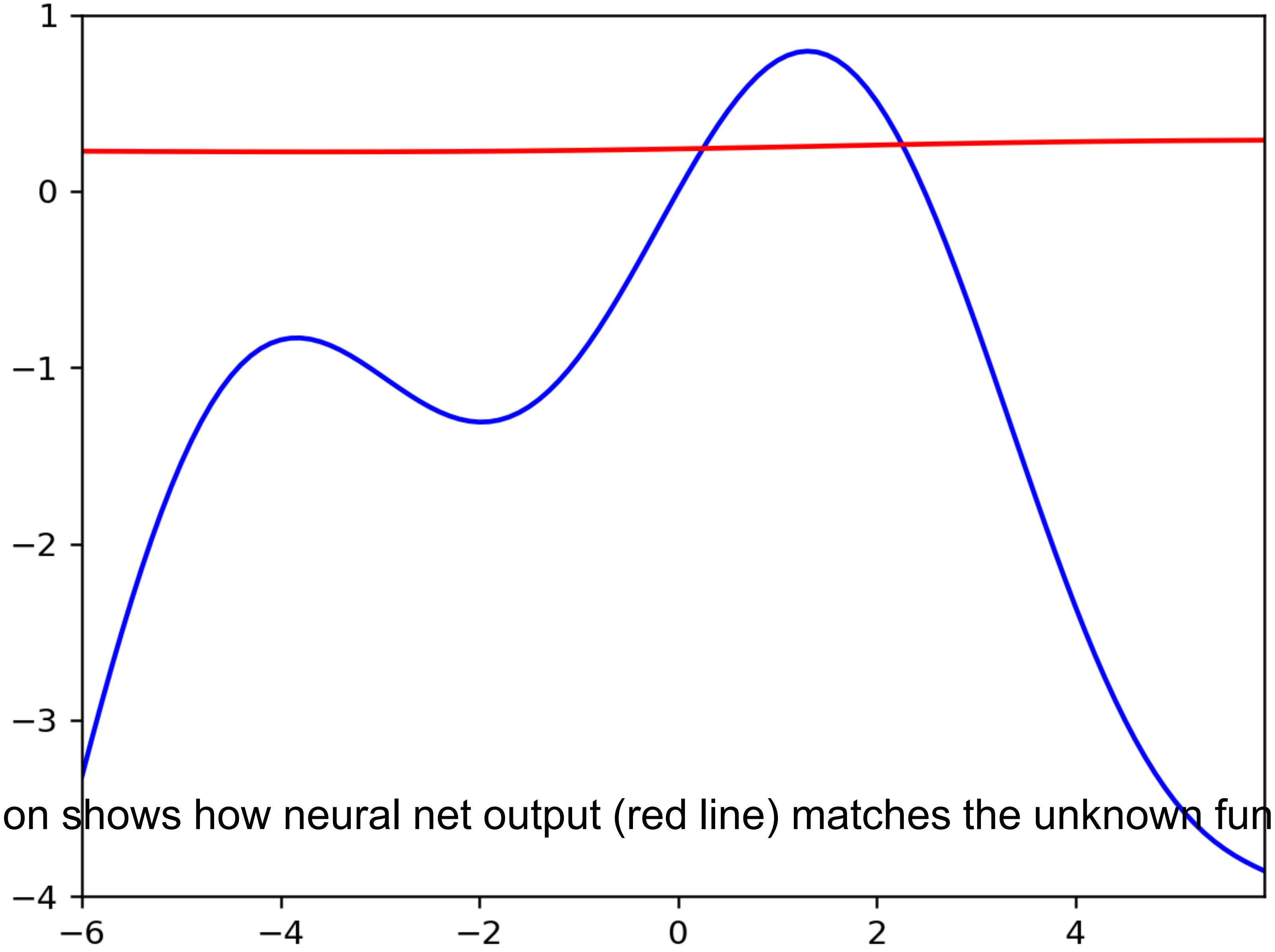
Neural Network:



slide from Steve Seitz's [video]



slide from Steve Seitz's [video]



Animation shows how neural net output (red line) matches the unknown function (blue line)

slide from Steve Seitz's [video]

$$P(x_n | x_{n-1}, x_{n-2}, x_{n-3}, x_{n-4}, x_{n-5}, x_{n-6}, x_{n-7}, \ldots)$$

slide from Steve Seitz's [video]

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

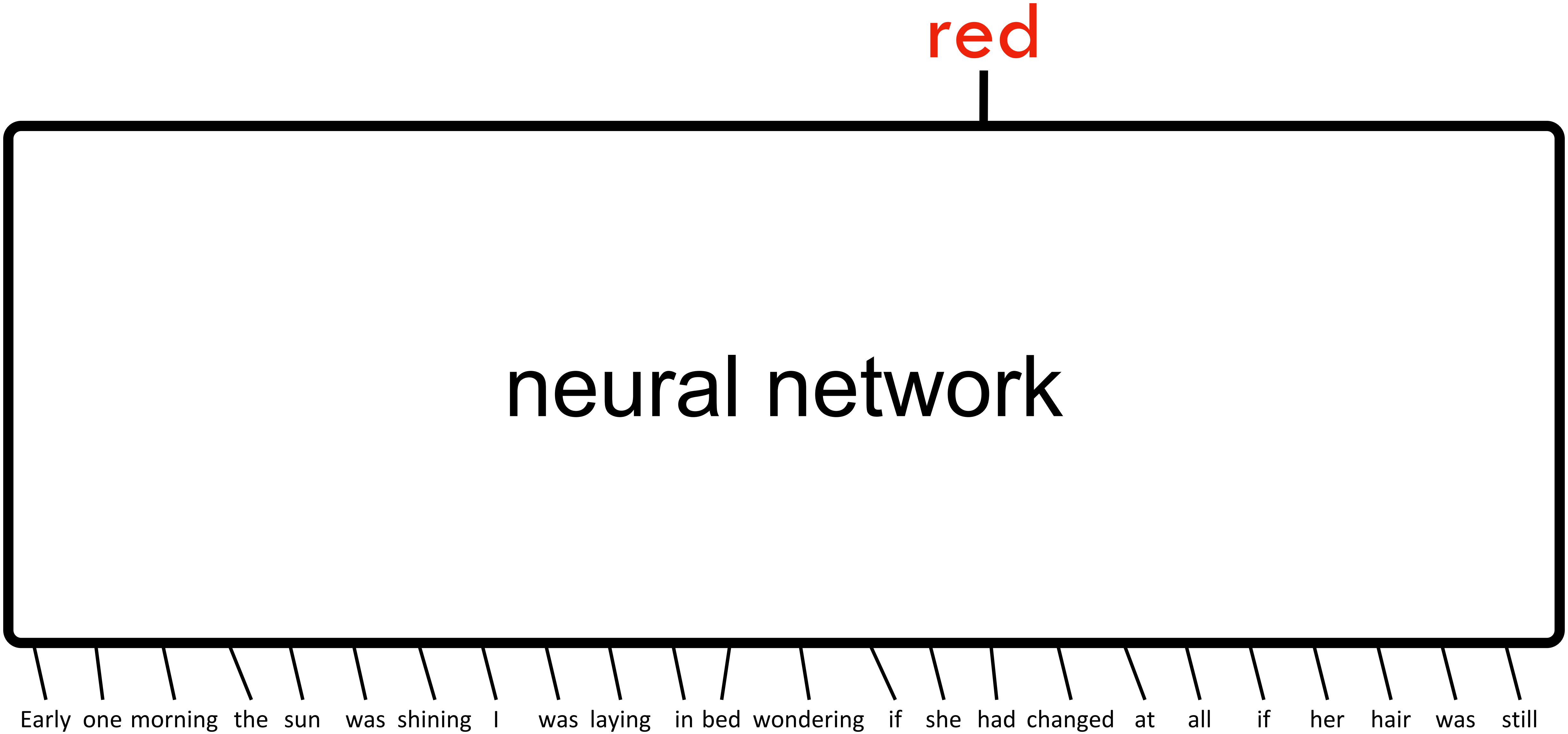
slide from Steve Seitz's video

red

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still

slide from Steve Seitz's [video]



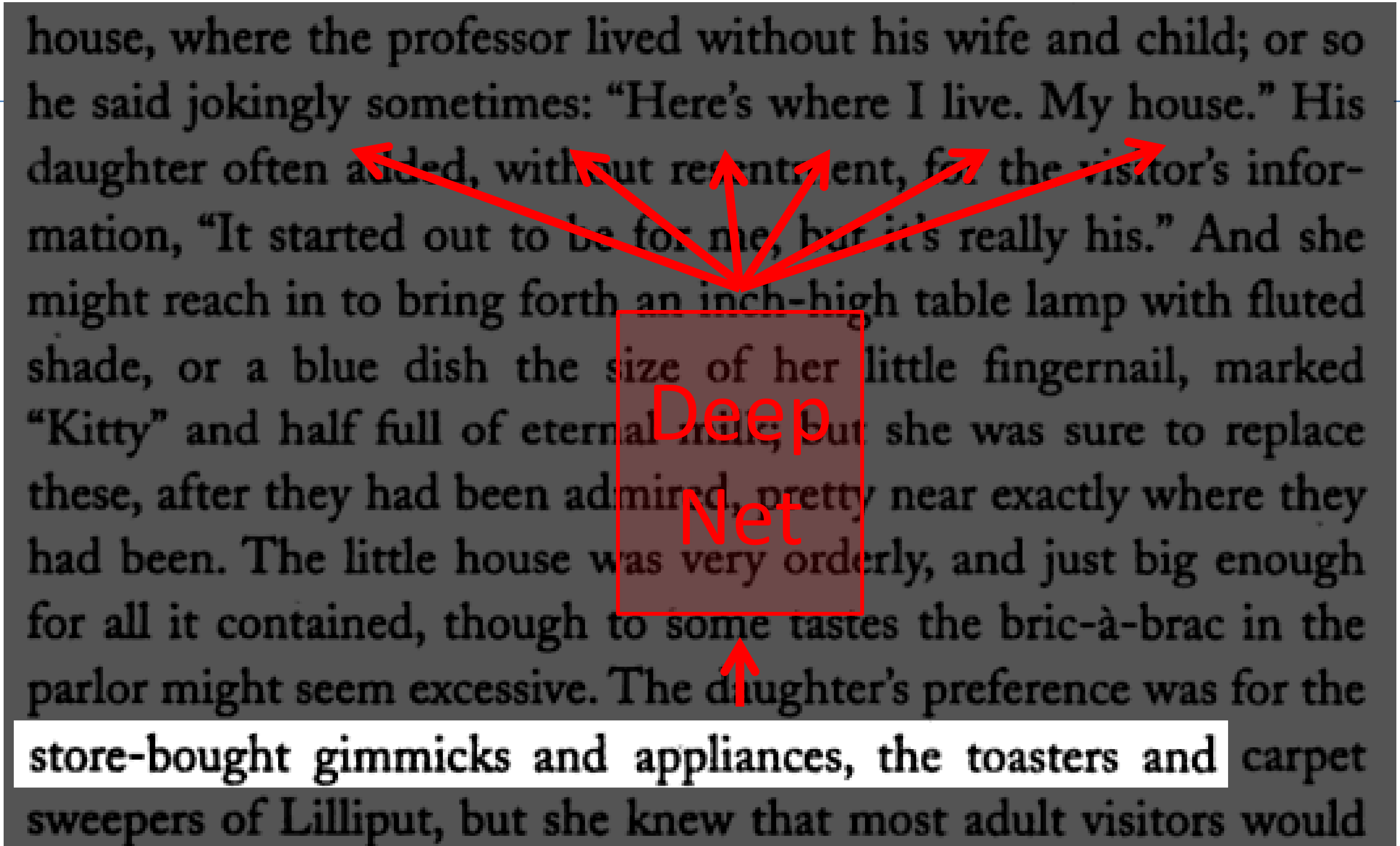
slide from Steve Seitz’s [video]

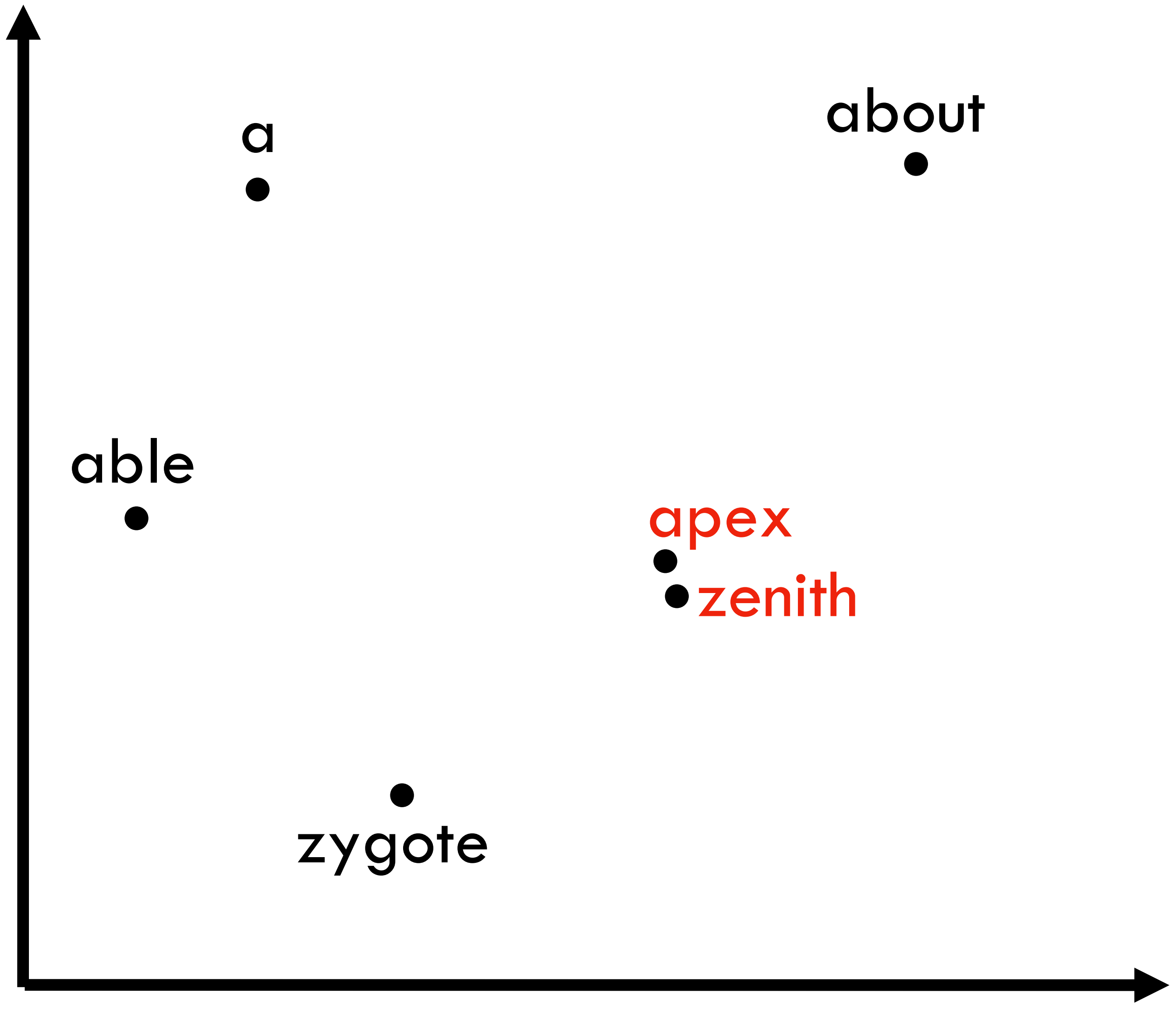
0	a
1	able
2	about
	⋮
39	apex
	⋮
56,356	zenith
	⋮
92,487	zygote

slide from Steve Seitz's [video]

word2vec

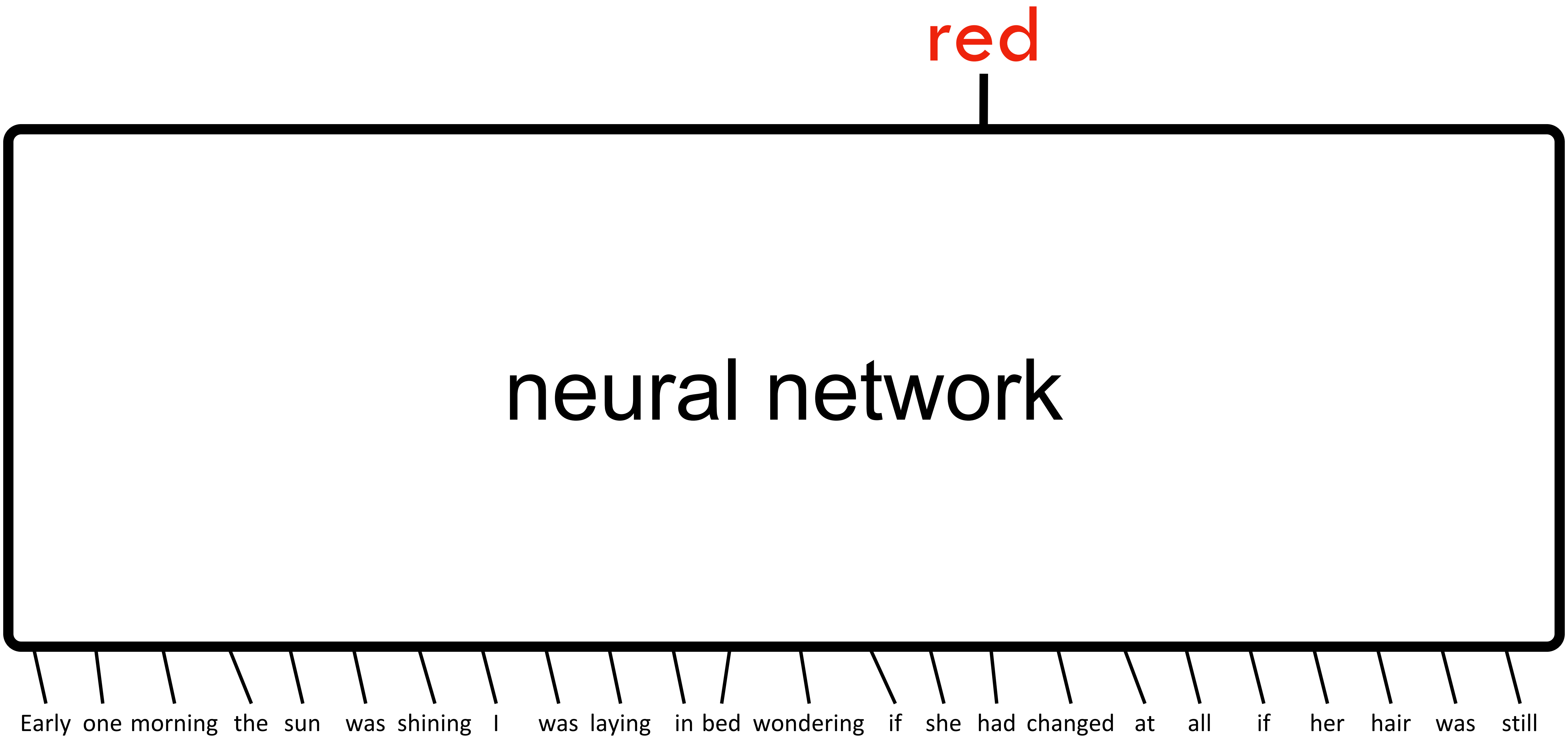
[Collobert & Weston 2008; Mikolov et al. 2013]



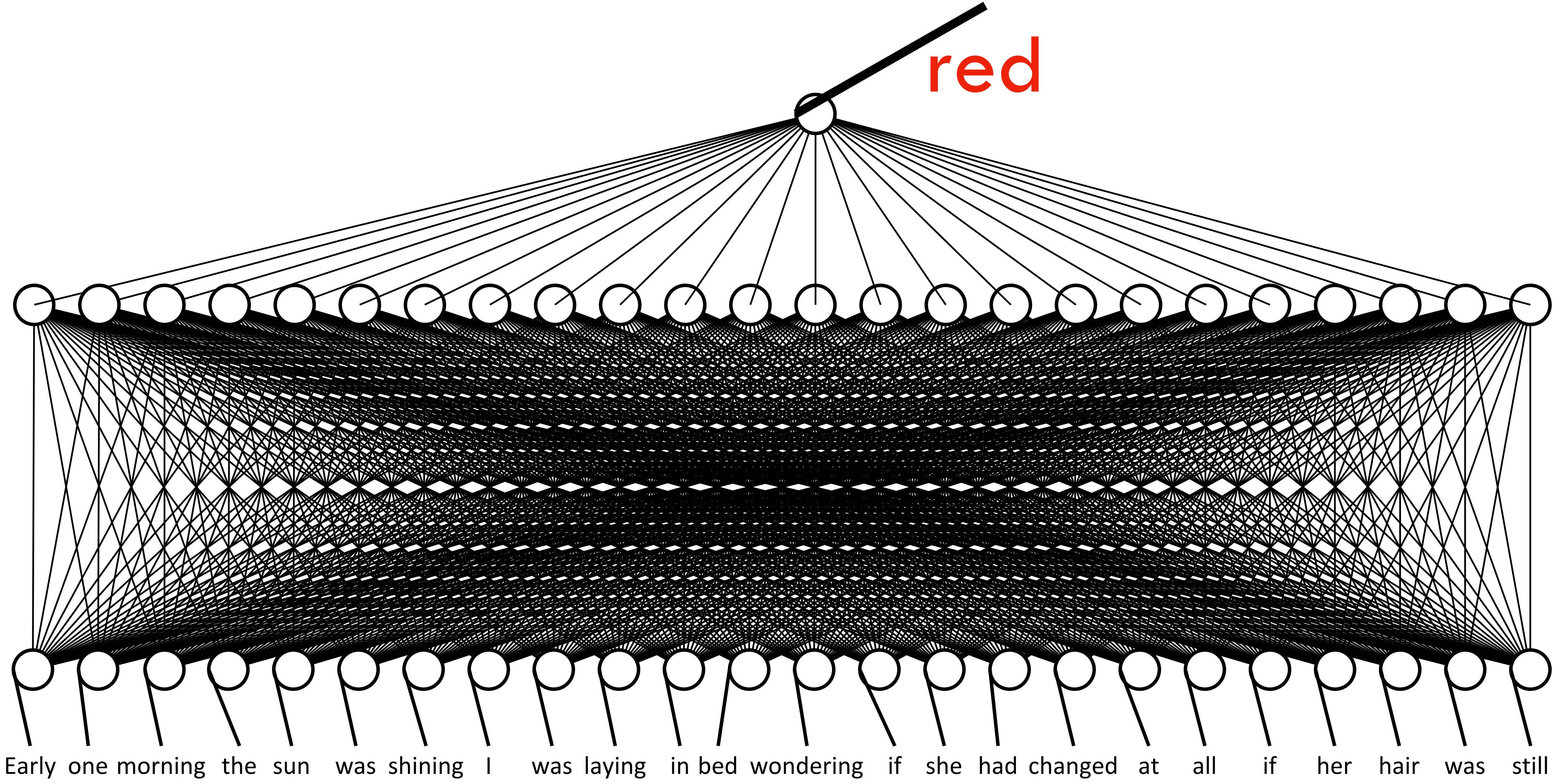


Word Embedding (e.g., word2Vec, GloVe)

slide from Steve Seitz's [video]



slide from Steve Seitz's [video]()



slide from Steve Seitz's video

Early one morning the sun was shining I was laying in bed

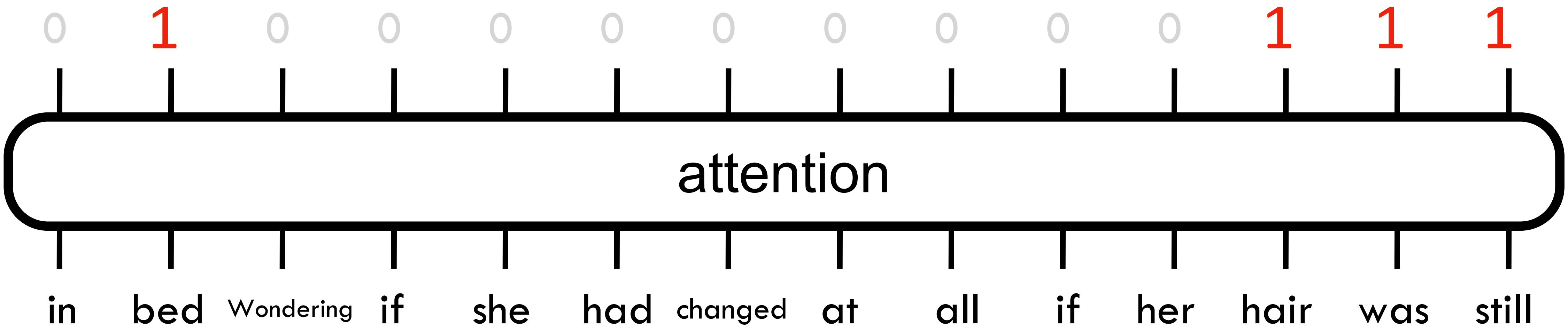
Wondering if she had changed at all if her hair was still ?

slide from Steve Seitz's video

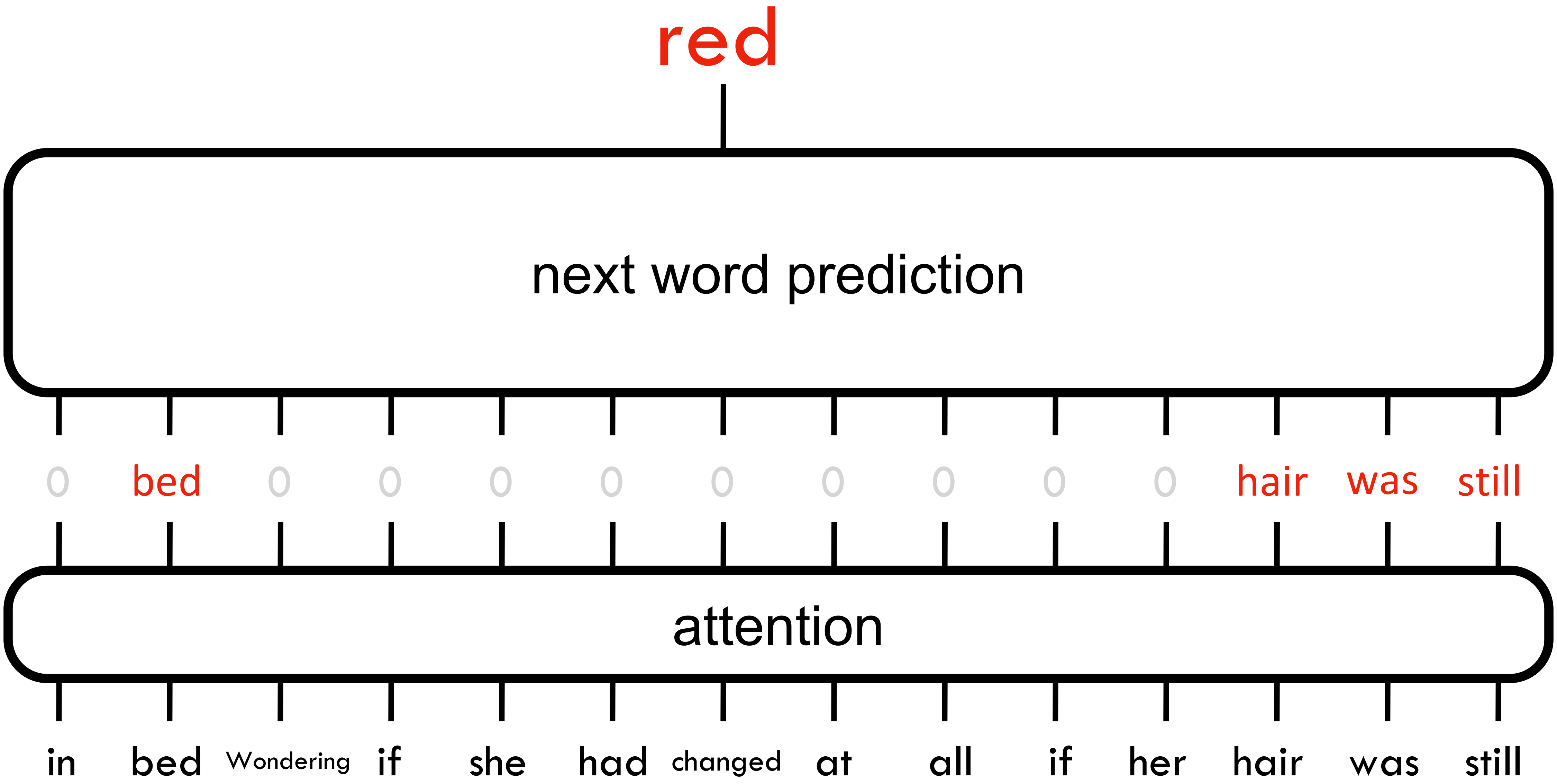
bed

hair was still red

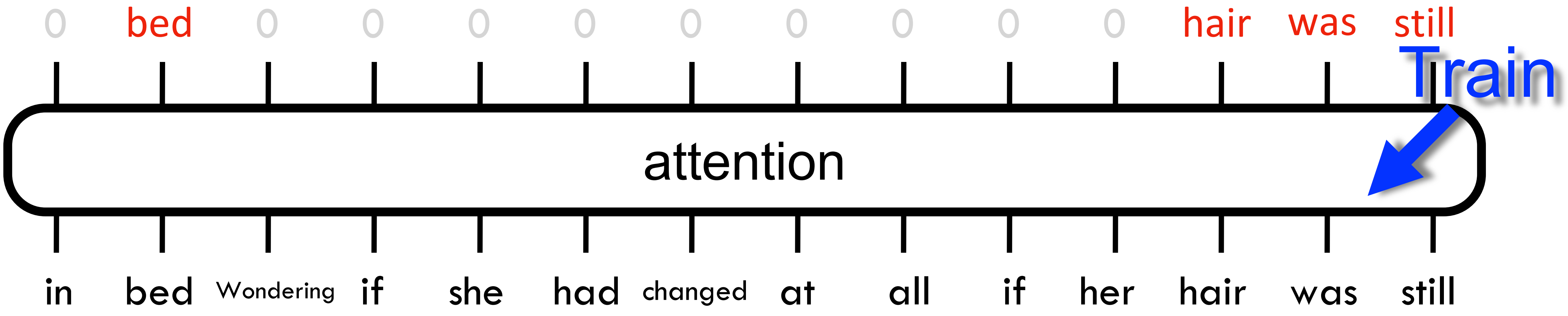
slide from Steve Seitz's video



slide from Steve Seitz's [video]



slide from Steve Seitz's [video]

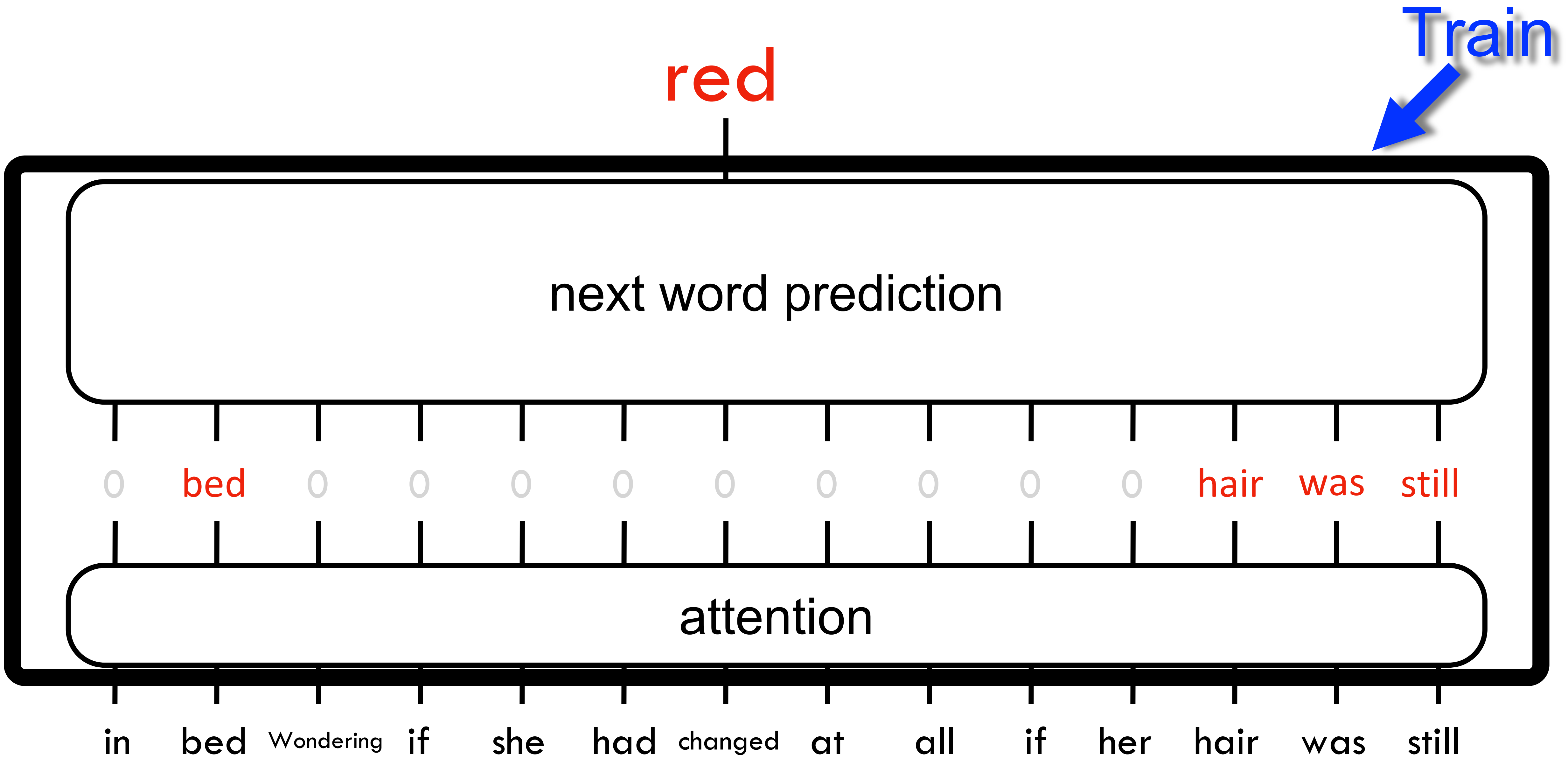


slide from Steve Seitz's [video]

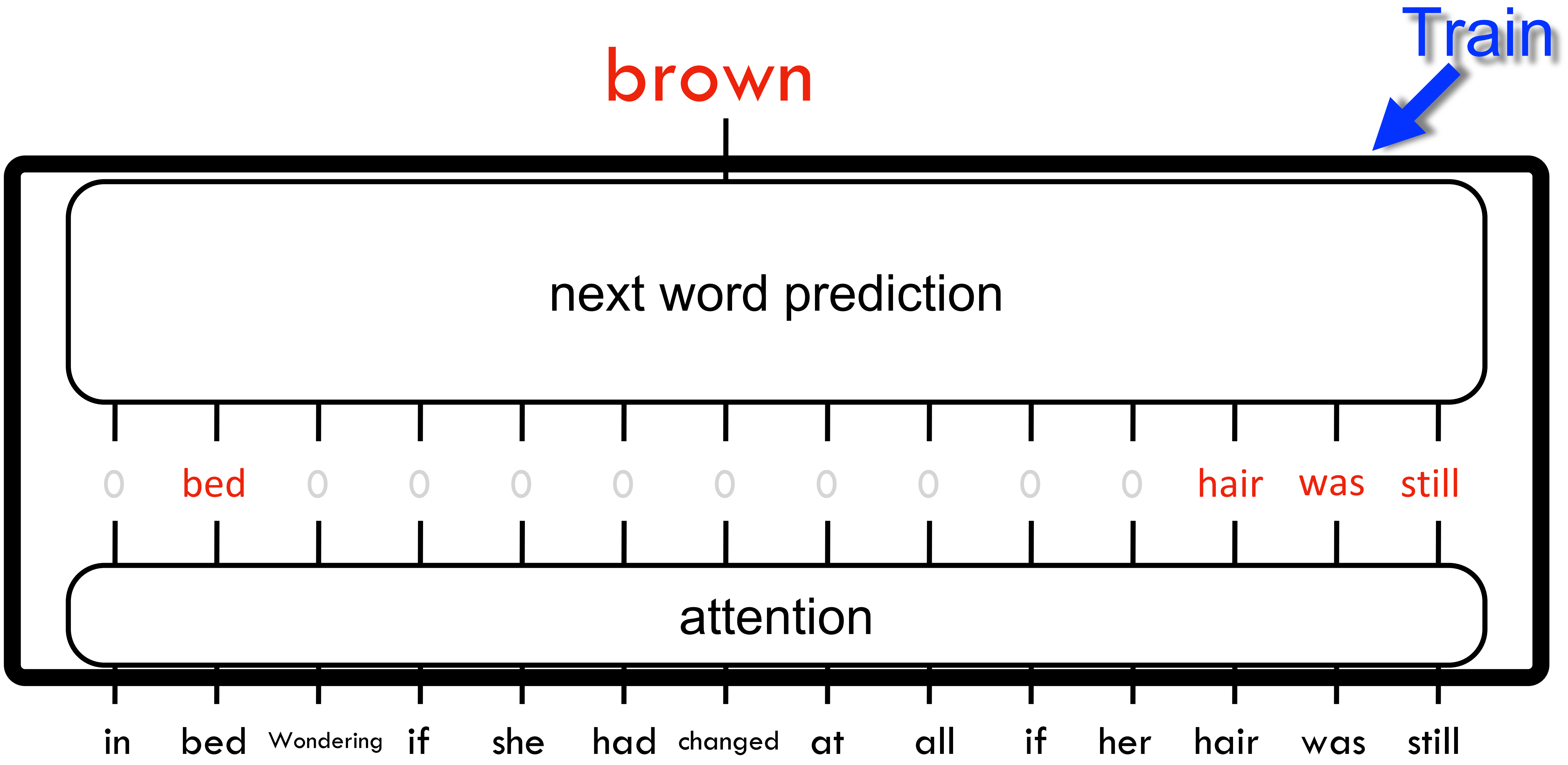
Two roads diverted in a yellow wood
And sorry I could not travel both
And be one traveler, long I stood
And looked down as far as I could
To where it bent in the undergrowth;

Robert Frost, *Road Not Taken*

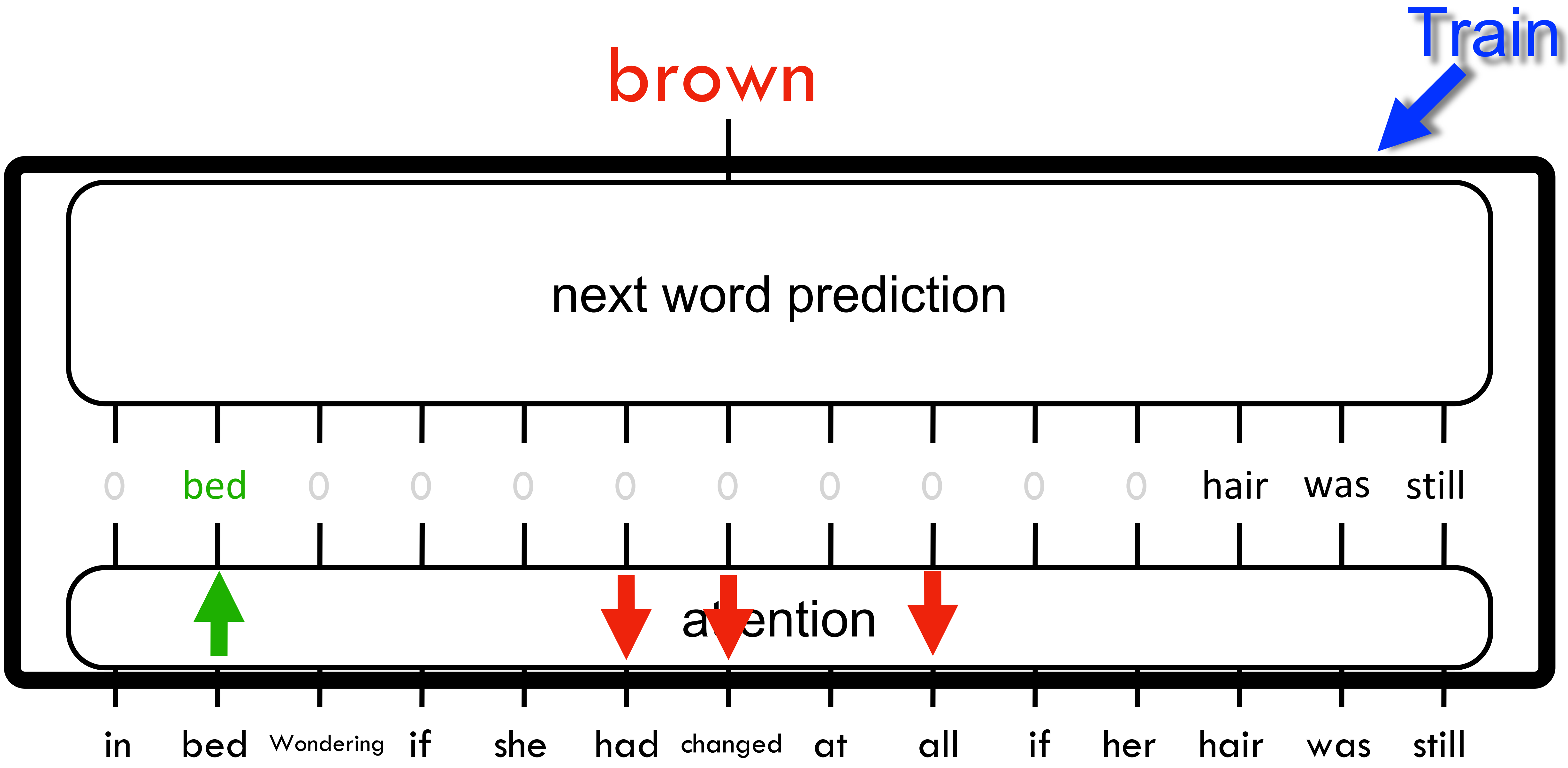
slide from Steve Seitz's video



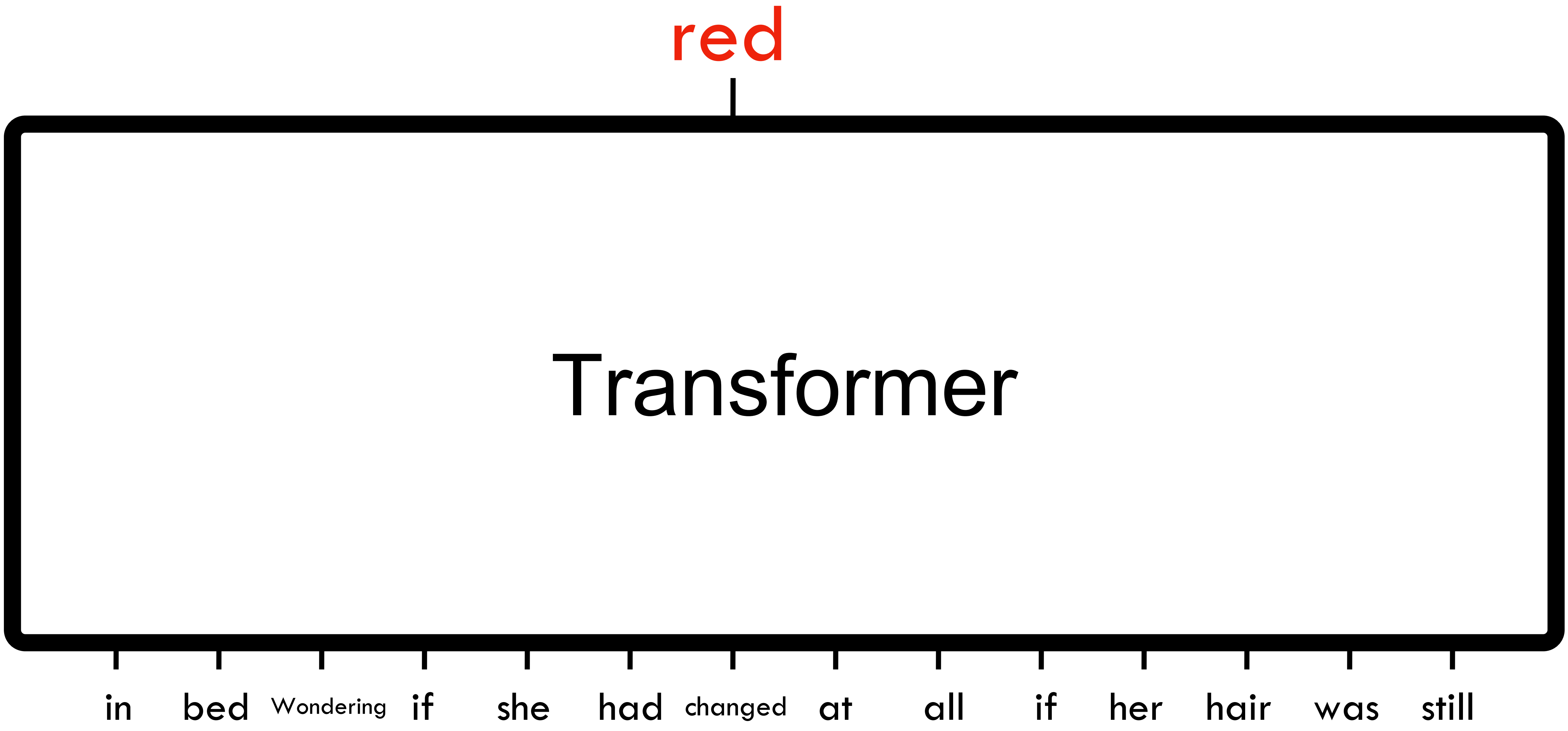
slide from Steve Seitz's video



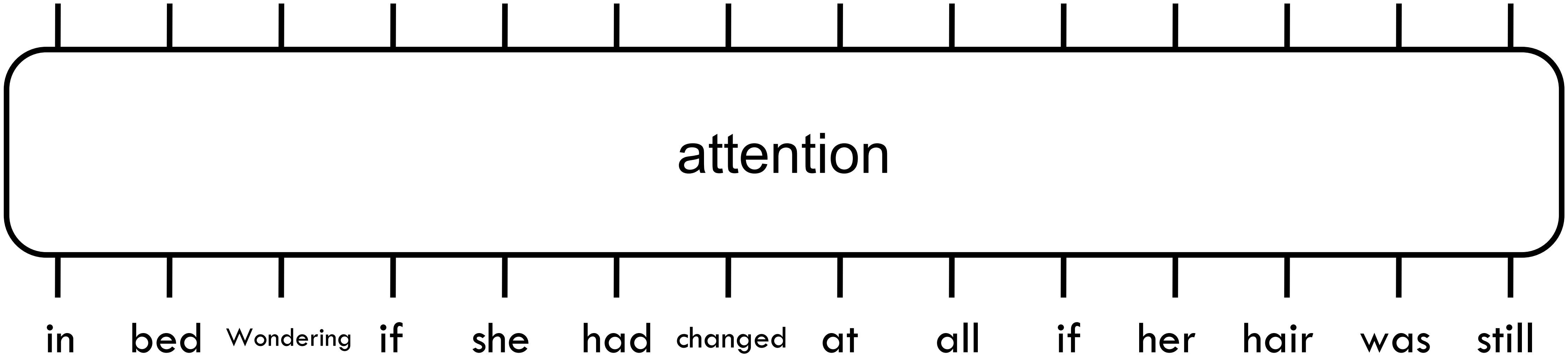
slide from Steve Seitz's [video]



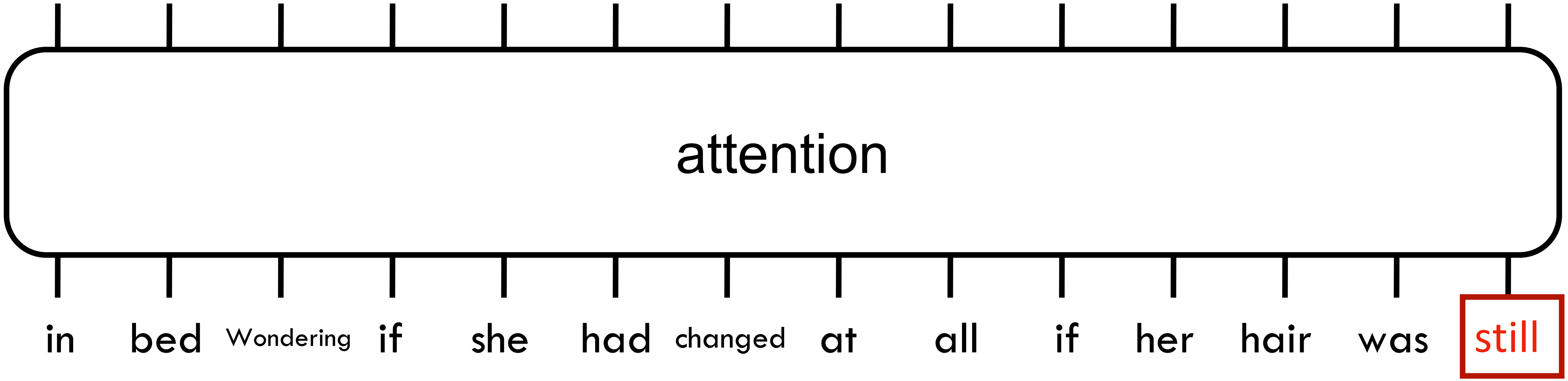
slide from Steve Seitz's video



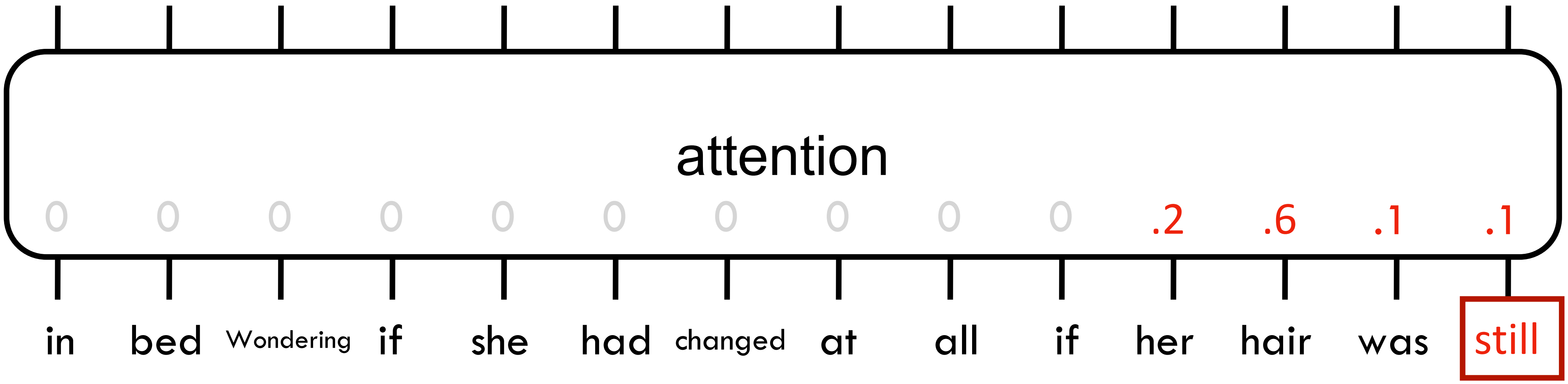
slide from Steve Seitz's video



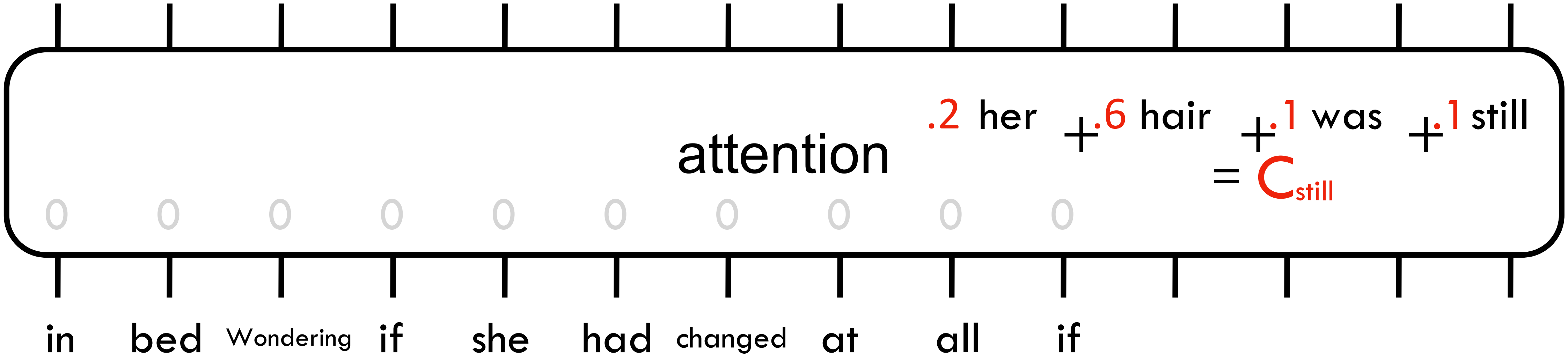
slide from Steve Seitz's video



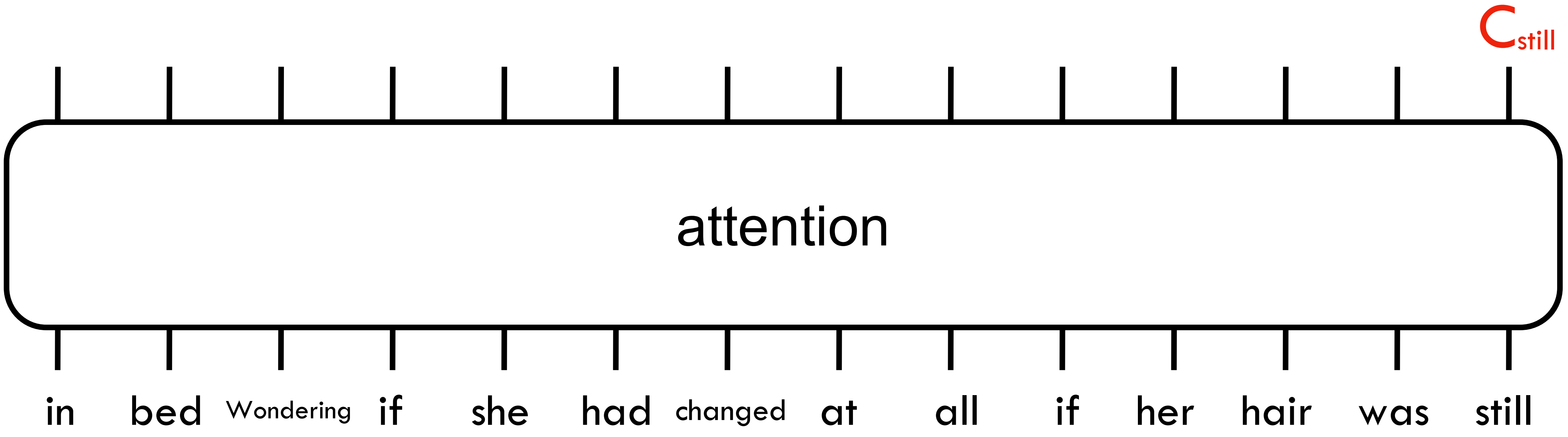
slide from Steve Seitz's video



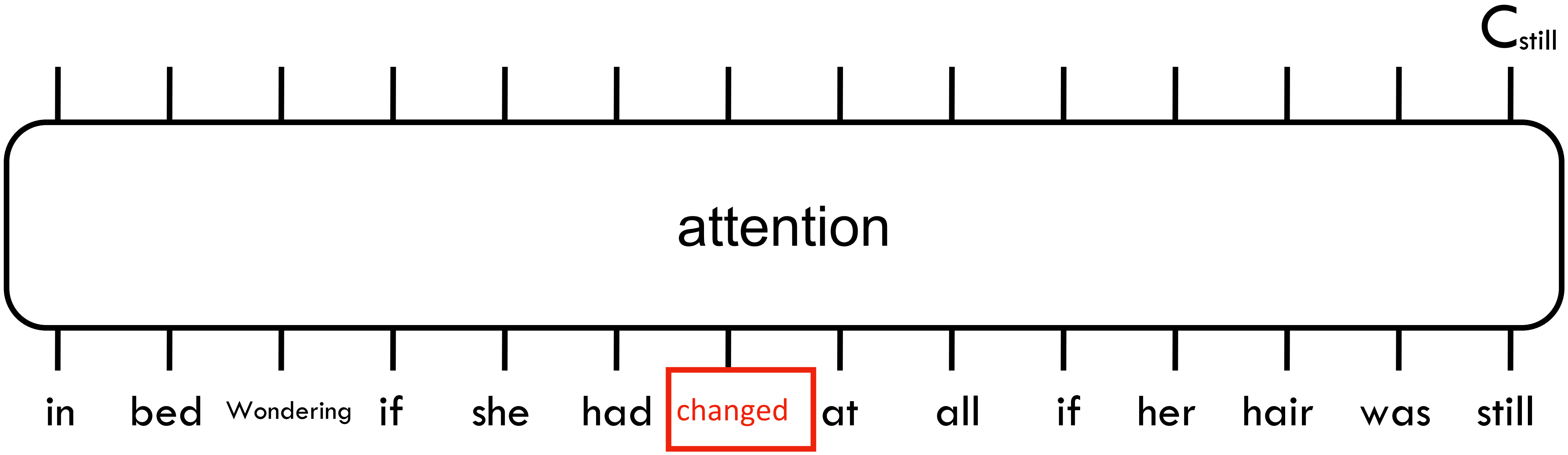
slide from Steve Seitz's [video]



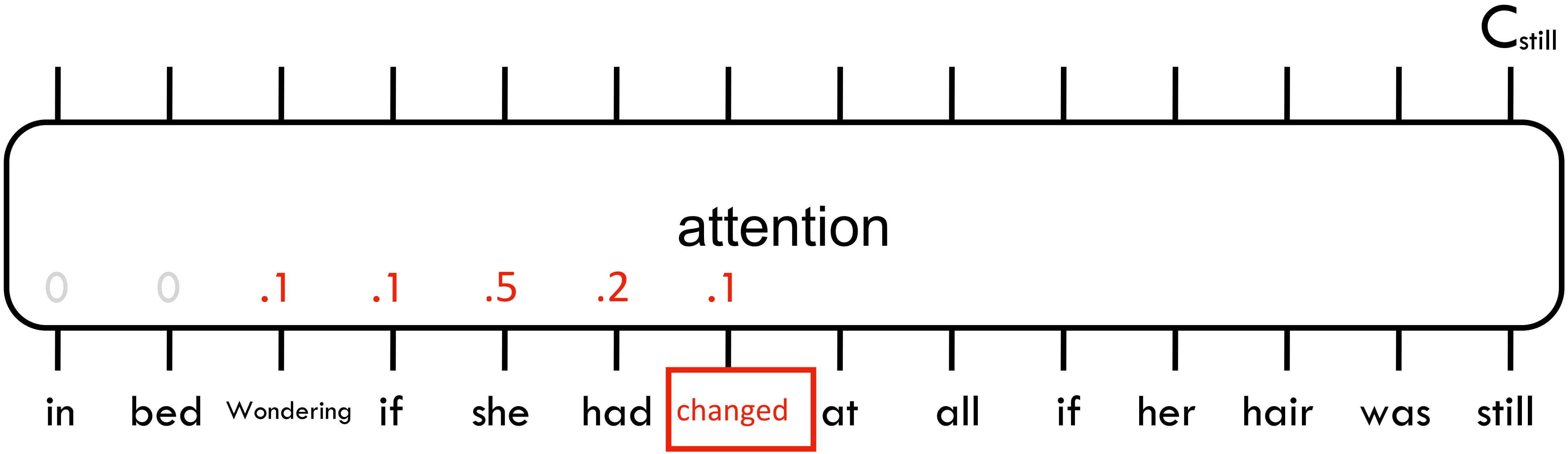
slide from Steve Seitz's [video]



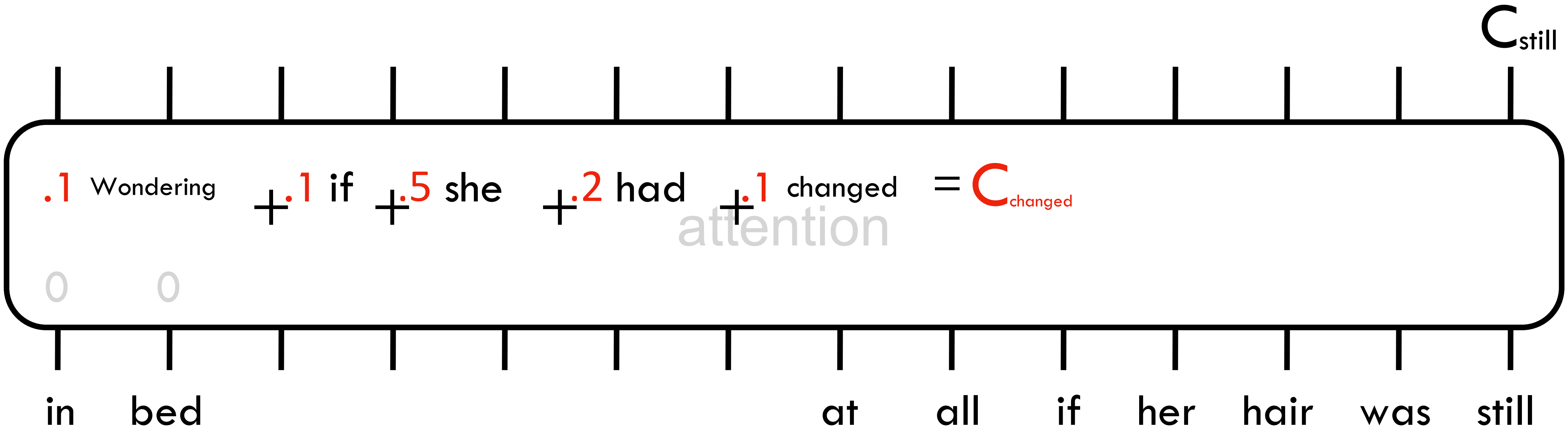
slide from Steve Seitz's [video]



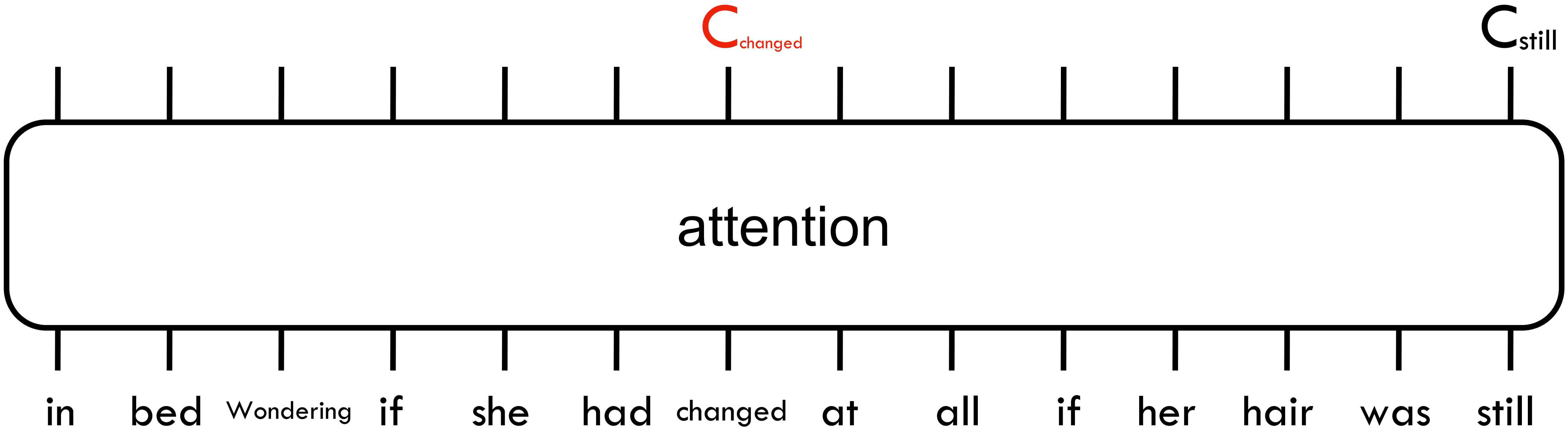
slide from Steve Seitz's video



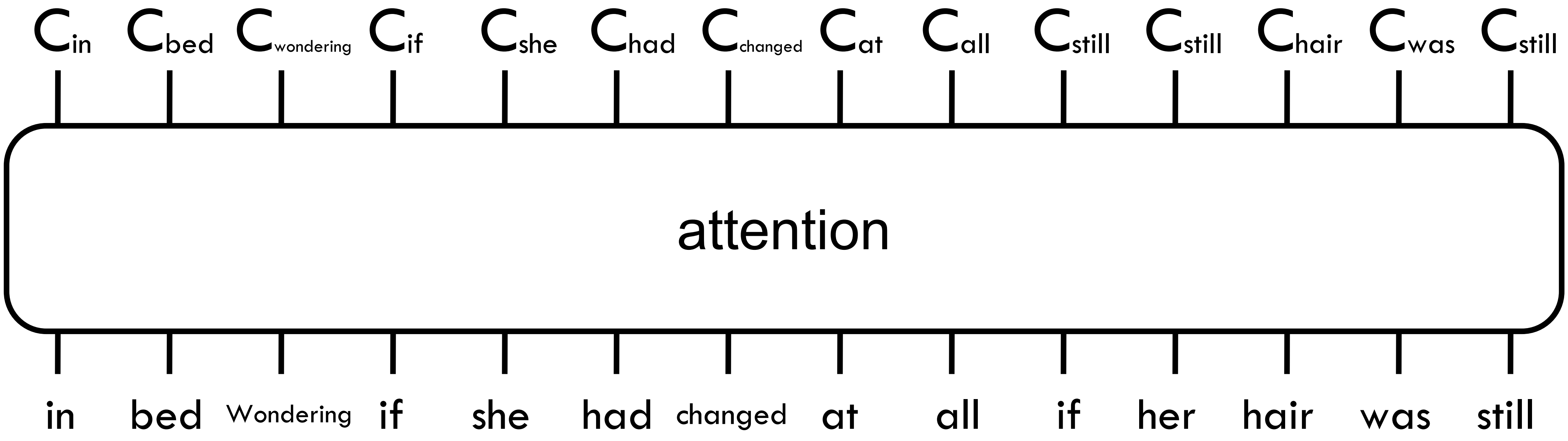
slide from Steve Seitz's video



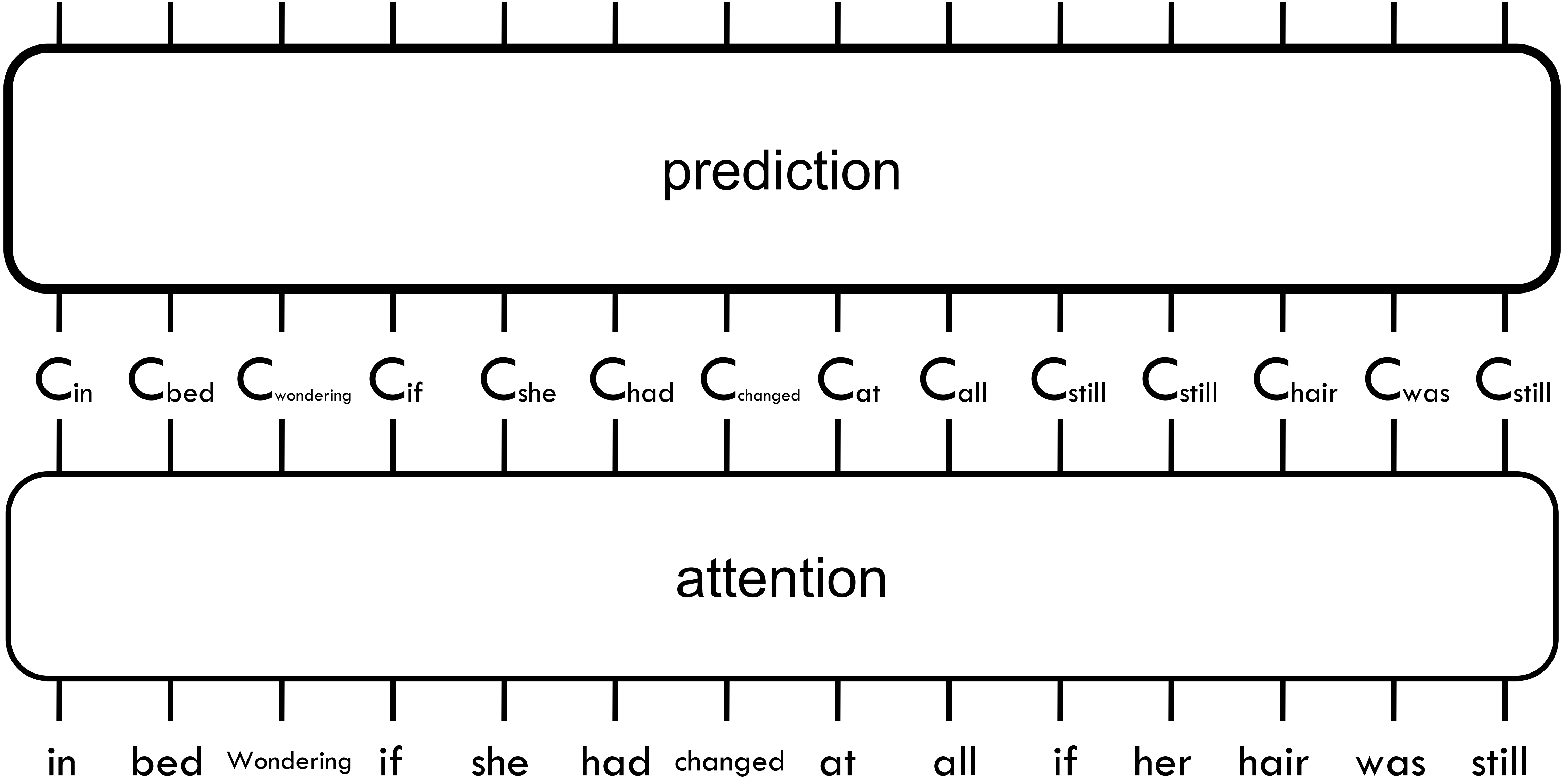
slide from Steve Seitz's [video]



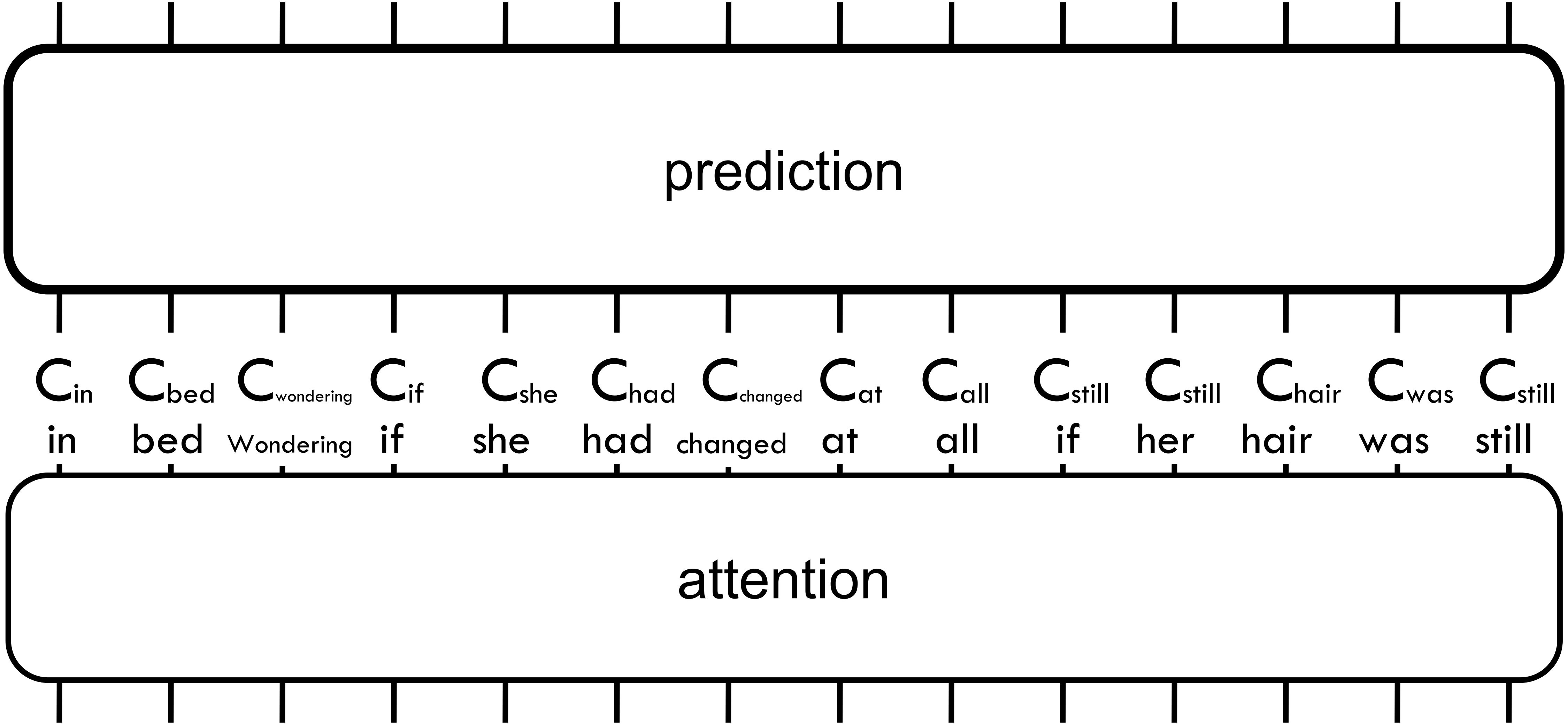
slide from Steve Seitz's [video]



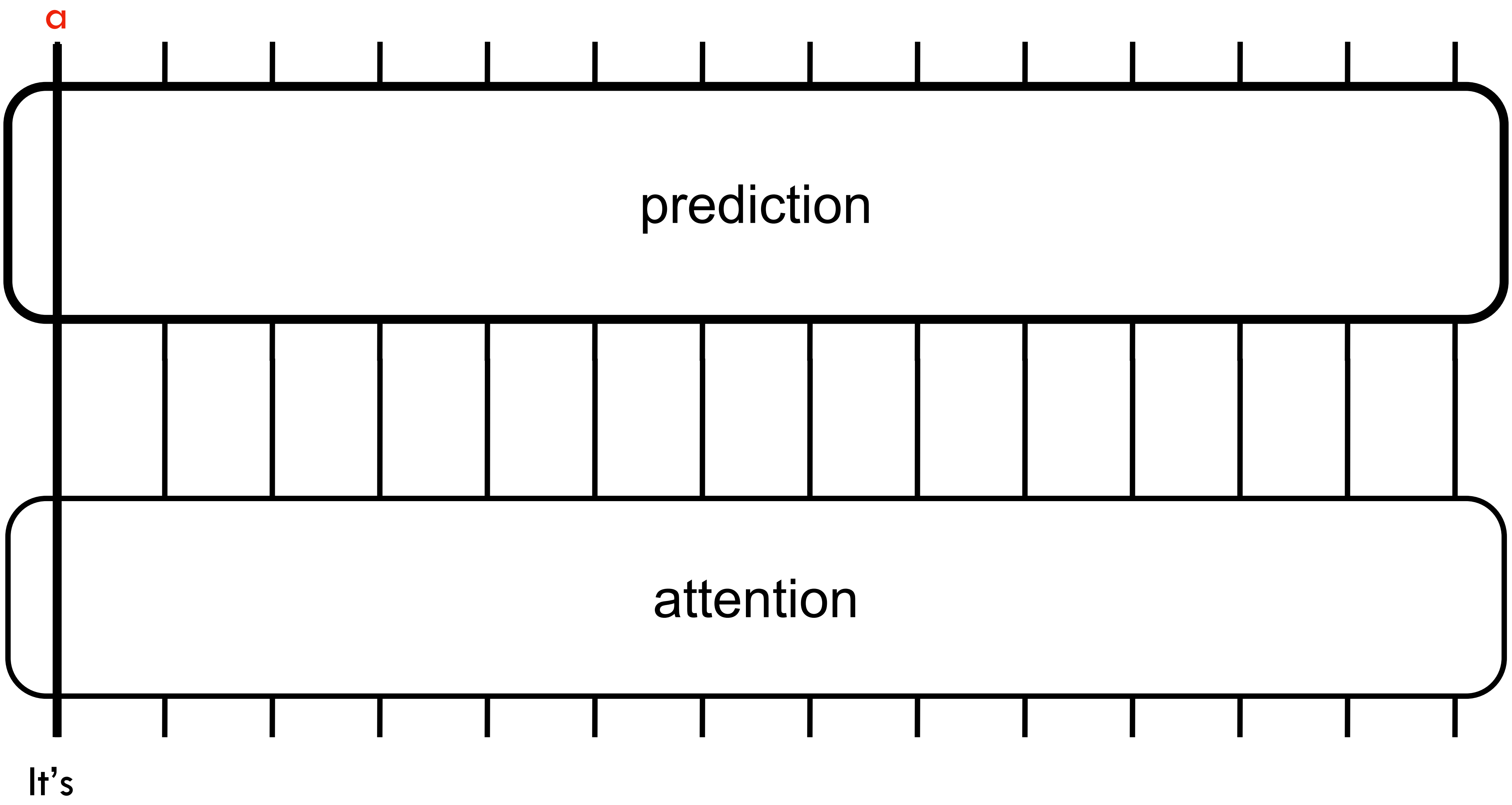
slide from Steve Seitz's [video]



slide from Steve Seitz's [video]



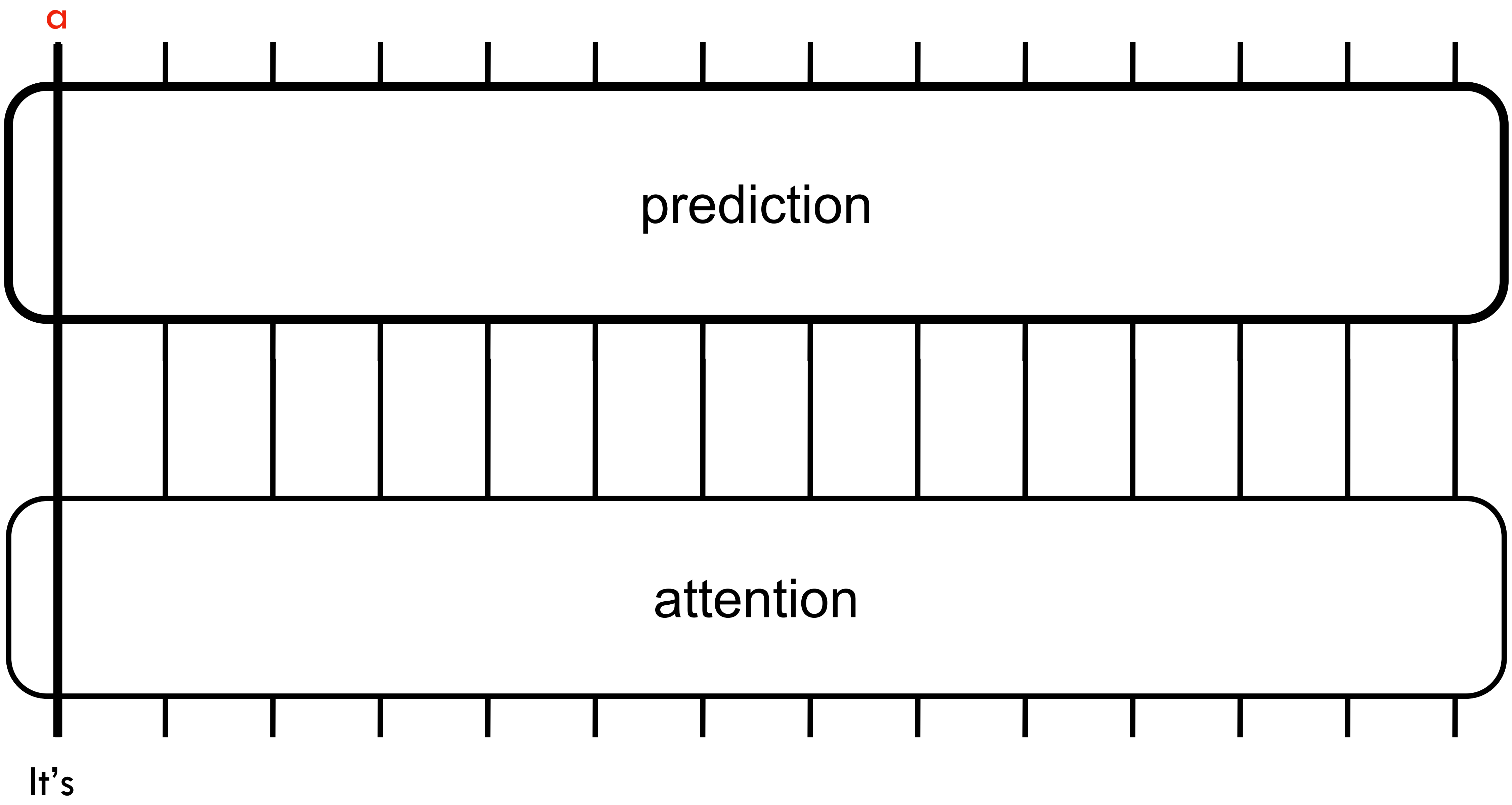
slide from Steve Seitz's [video]



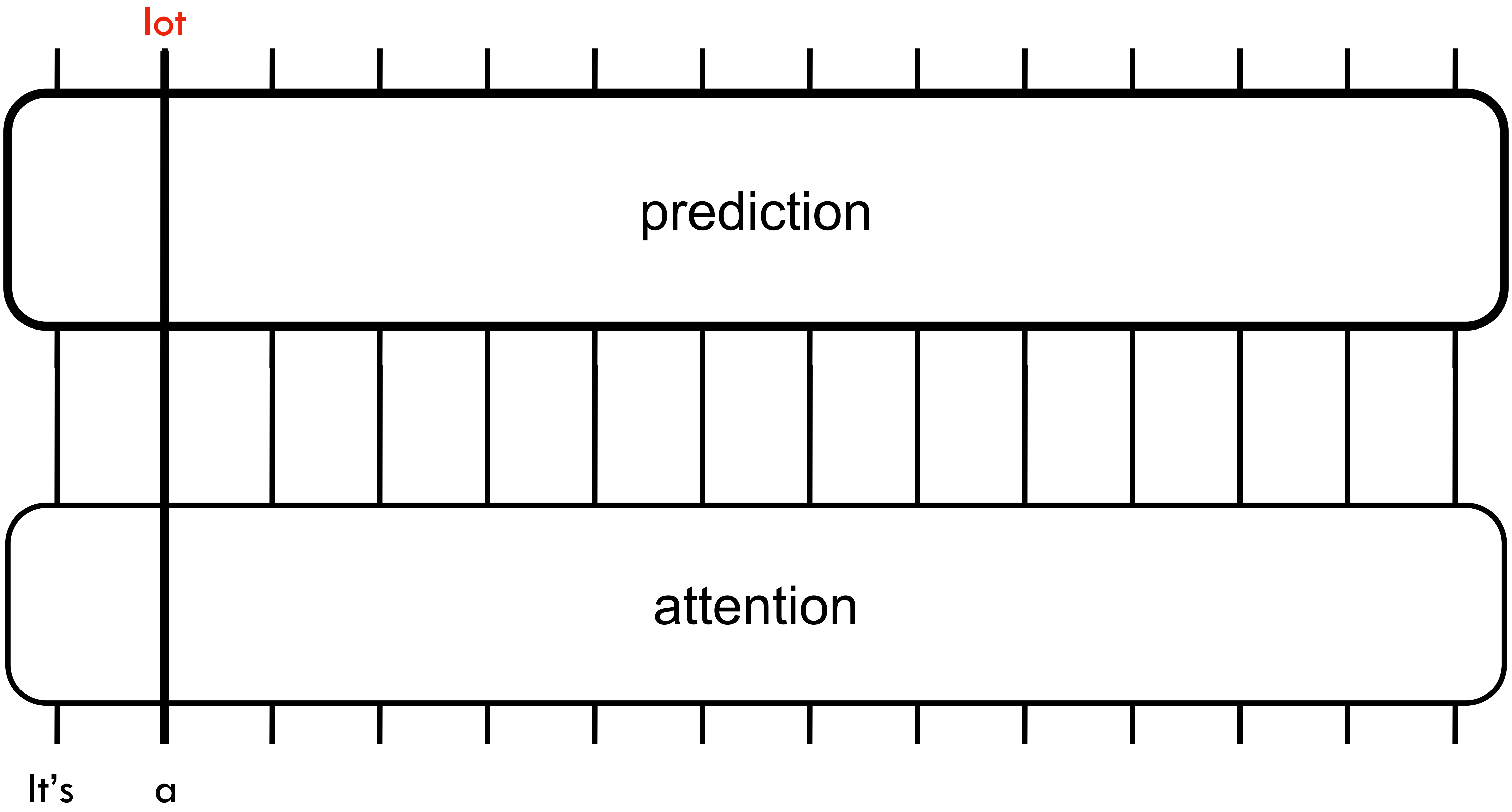
slide from Steve Seitz's [video]

a	the	looking	possible	getting
0.4	0.3	0.1	0.1	0.1

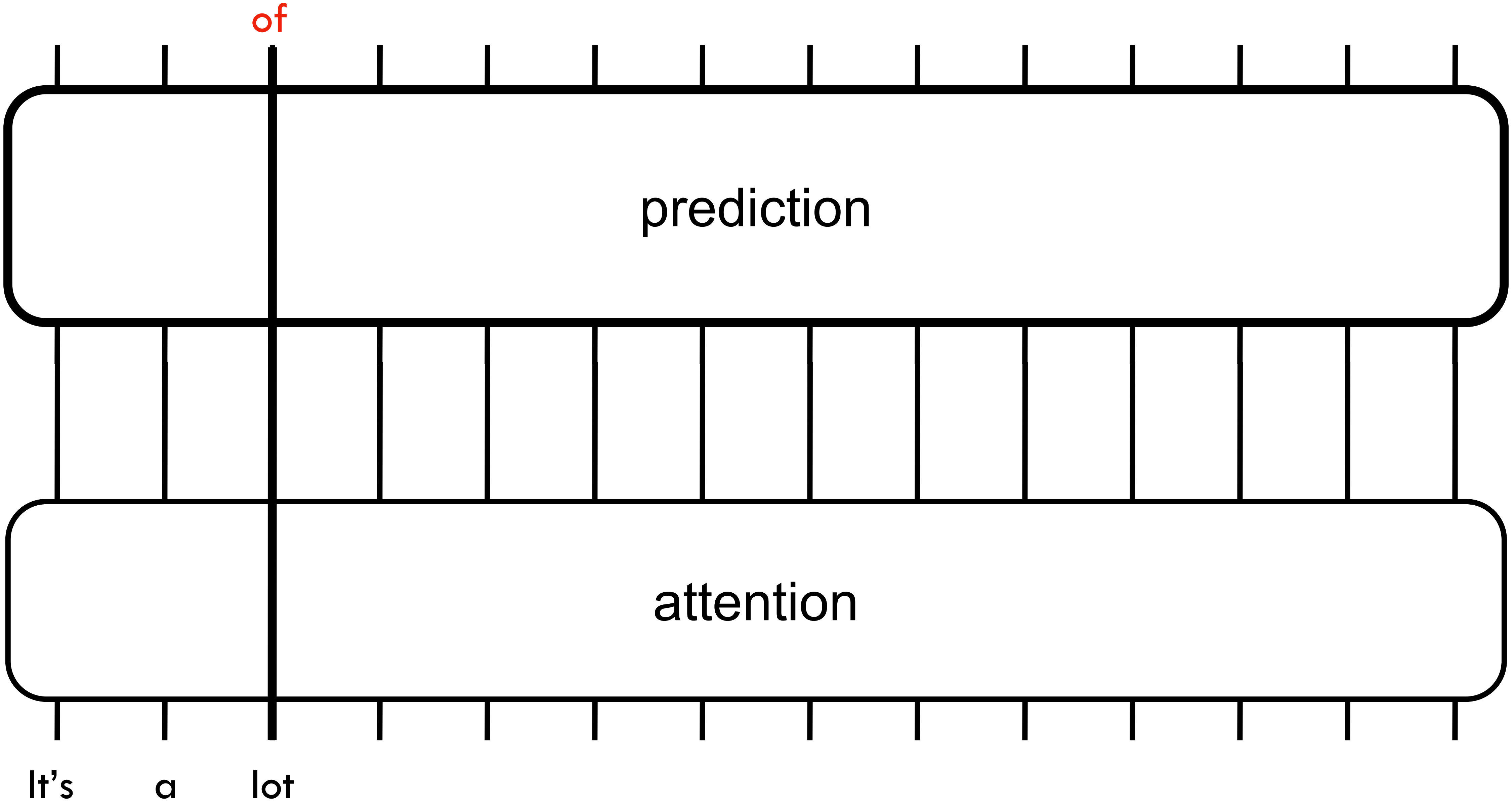
slide from Steve Seitz's [video]



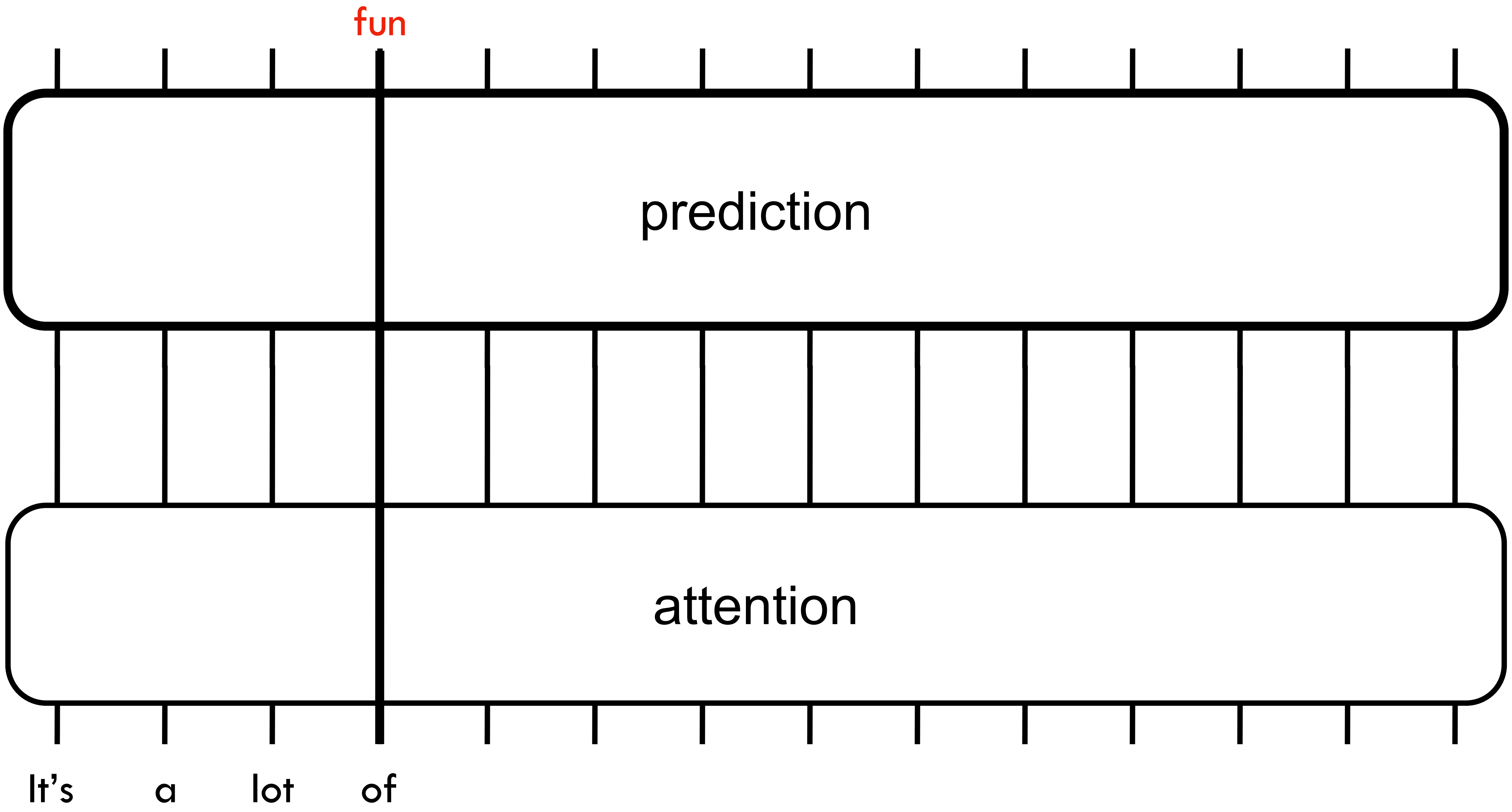
slide from Steve Seitz's [video]()



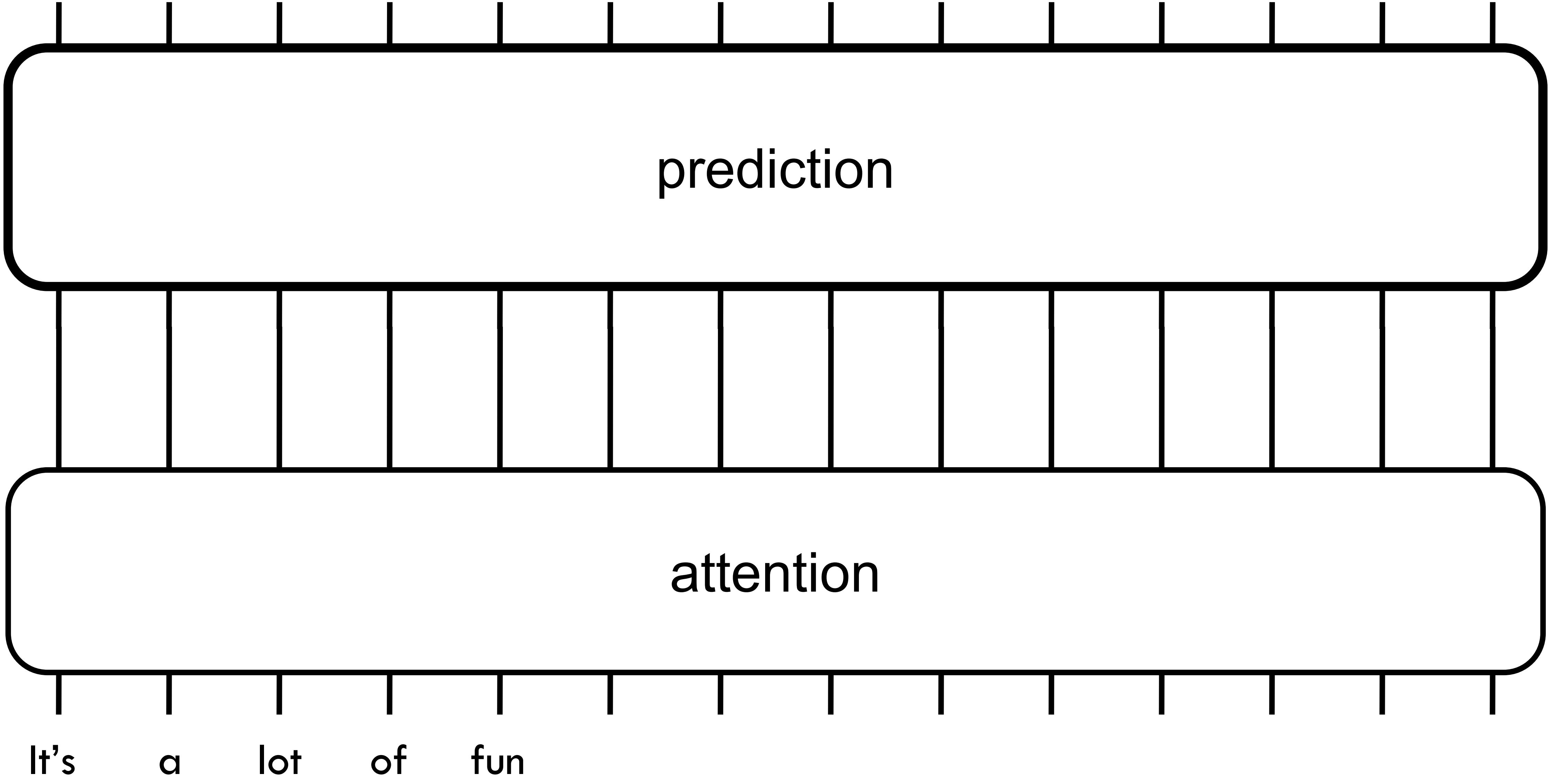
slide from Steve Seitz's [video]



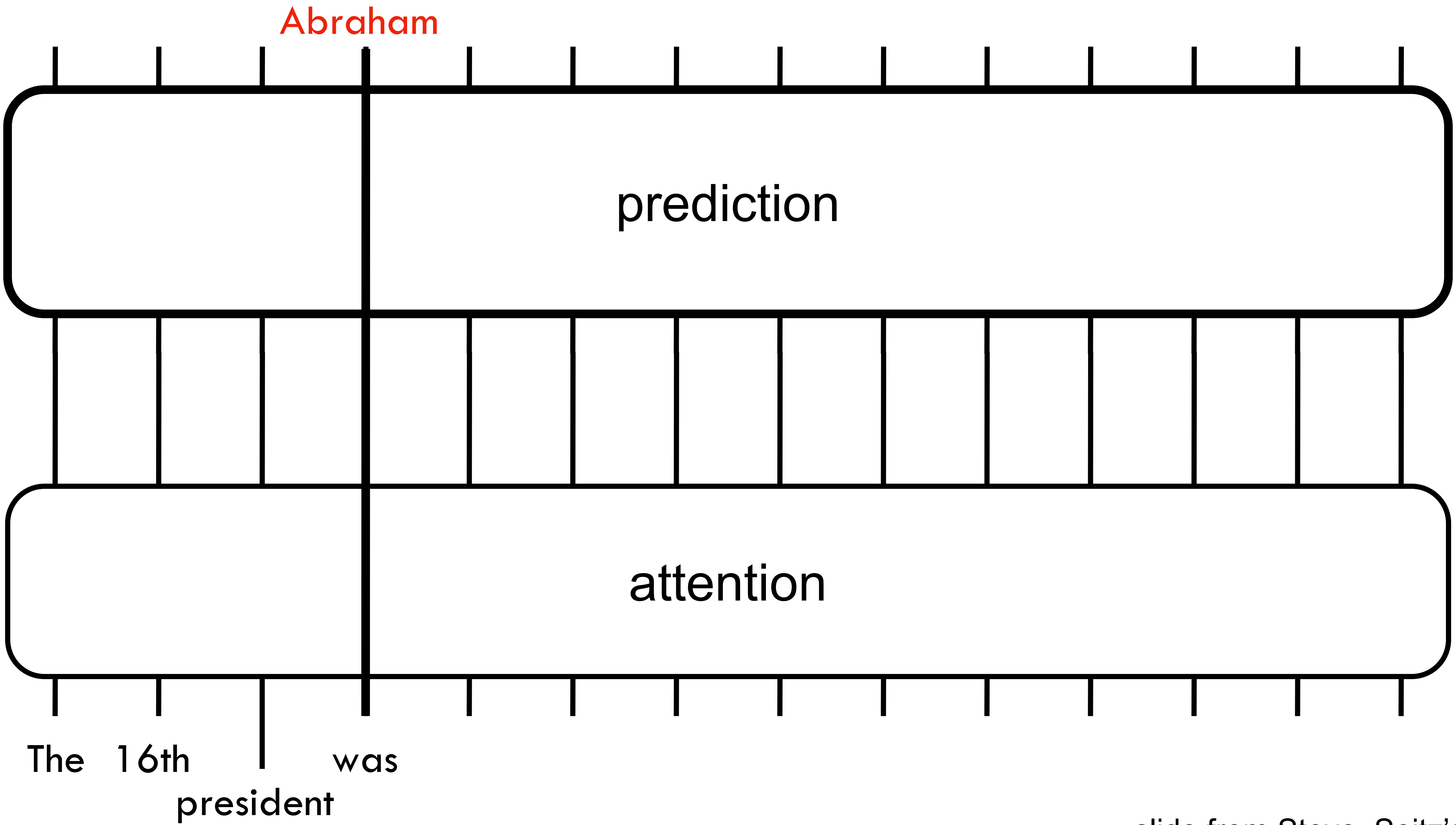
slide from Steve Seitz's video



slide from Steve Seitz's [video]



slide from Steve Seitz's [video]



slide from Steve Seitz's video

The 16th President was ?

The capital of Zimbabwe is ?

Frank Zappa’s middle name is ?

Napoleon was born on this date ?

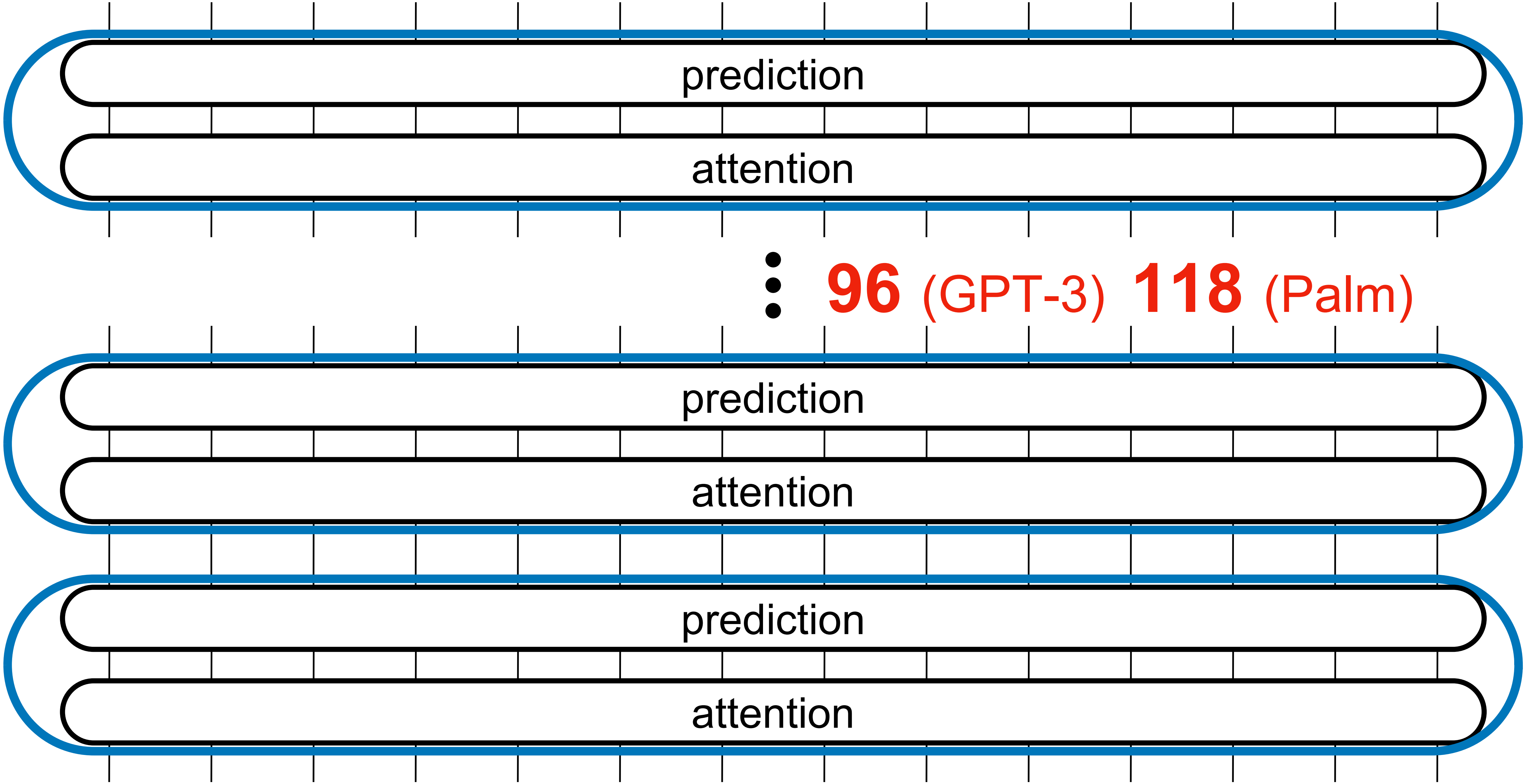
The prime factorization of 19456721434 is ?

Queen Victoria’s maiden name was ?

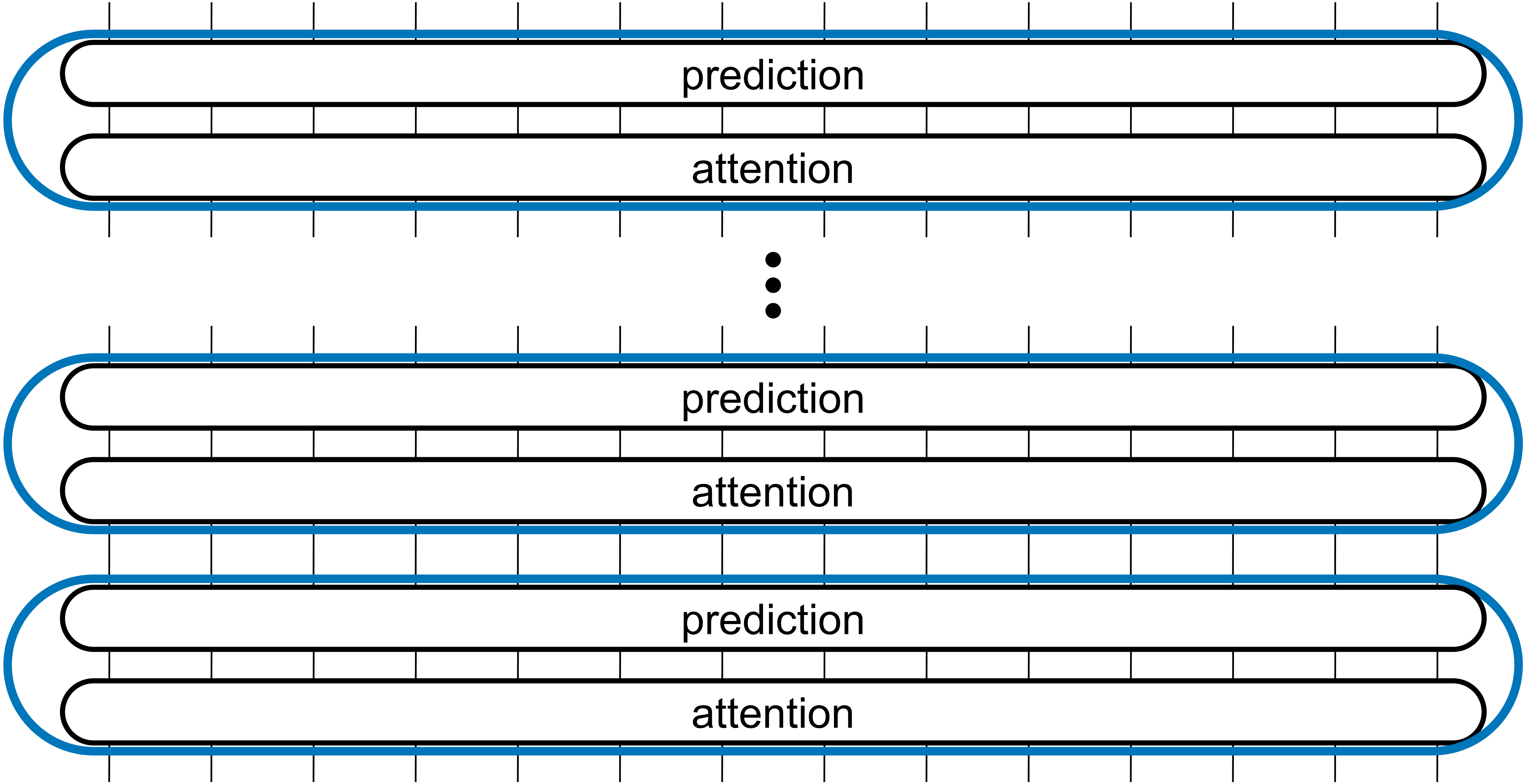
US per-capita income in 1957 was ?

The lat long coordinates of Rome are ?

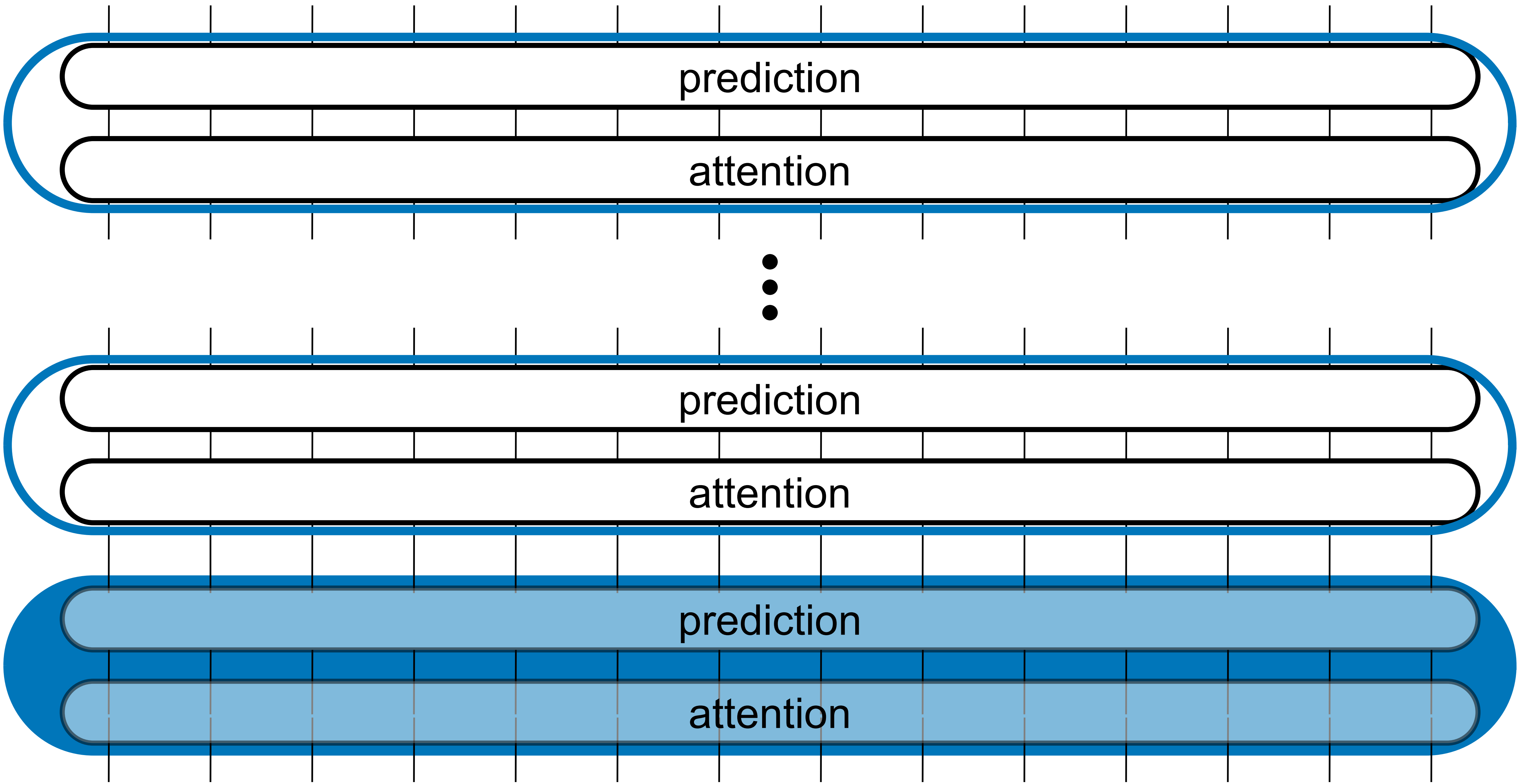
slide from Steve Seitz’s [video]



slide from Steve Seitz's [video]



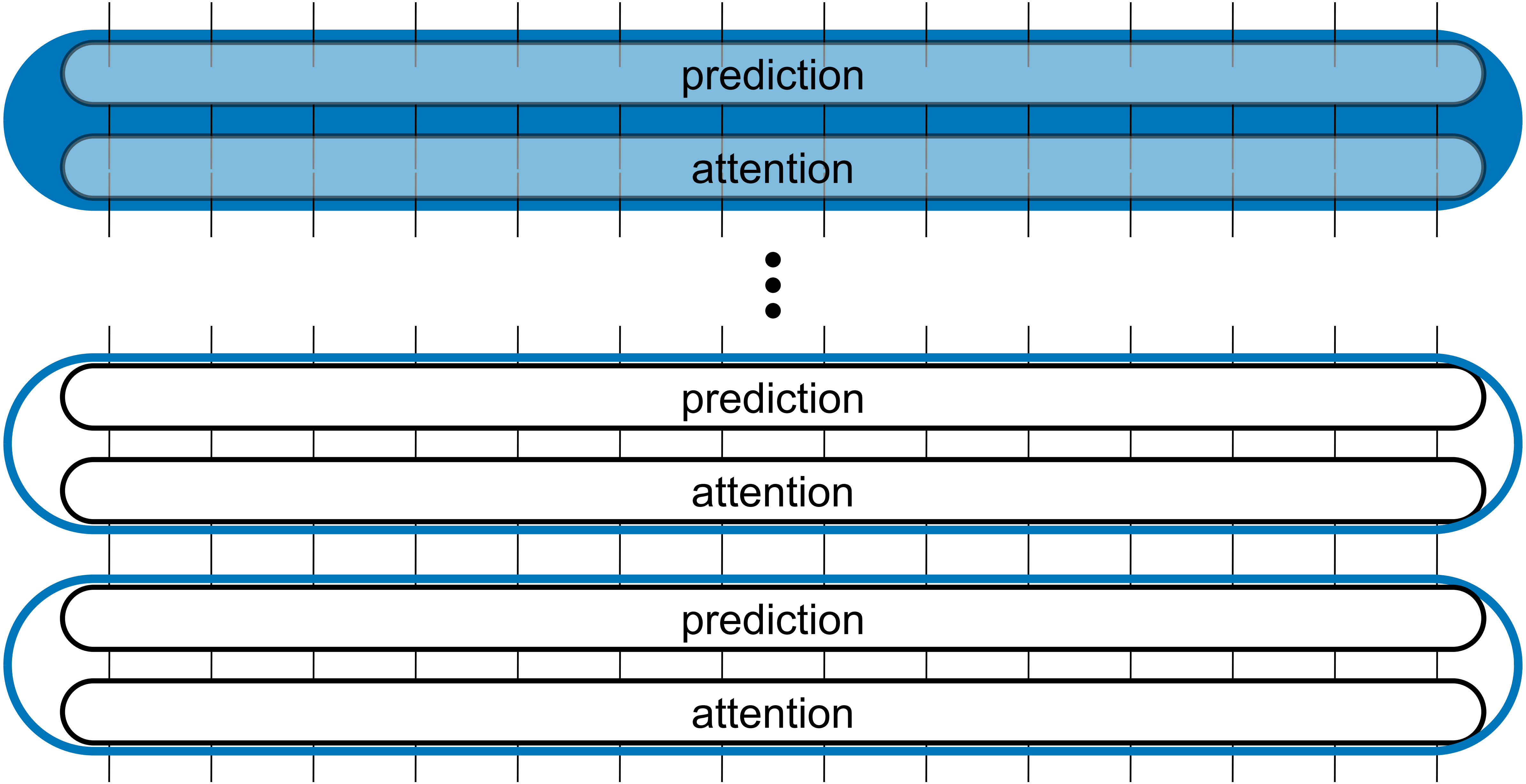
slide from Steve Seitz's [video]



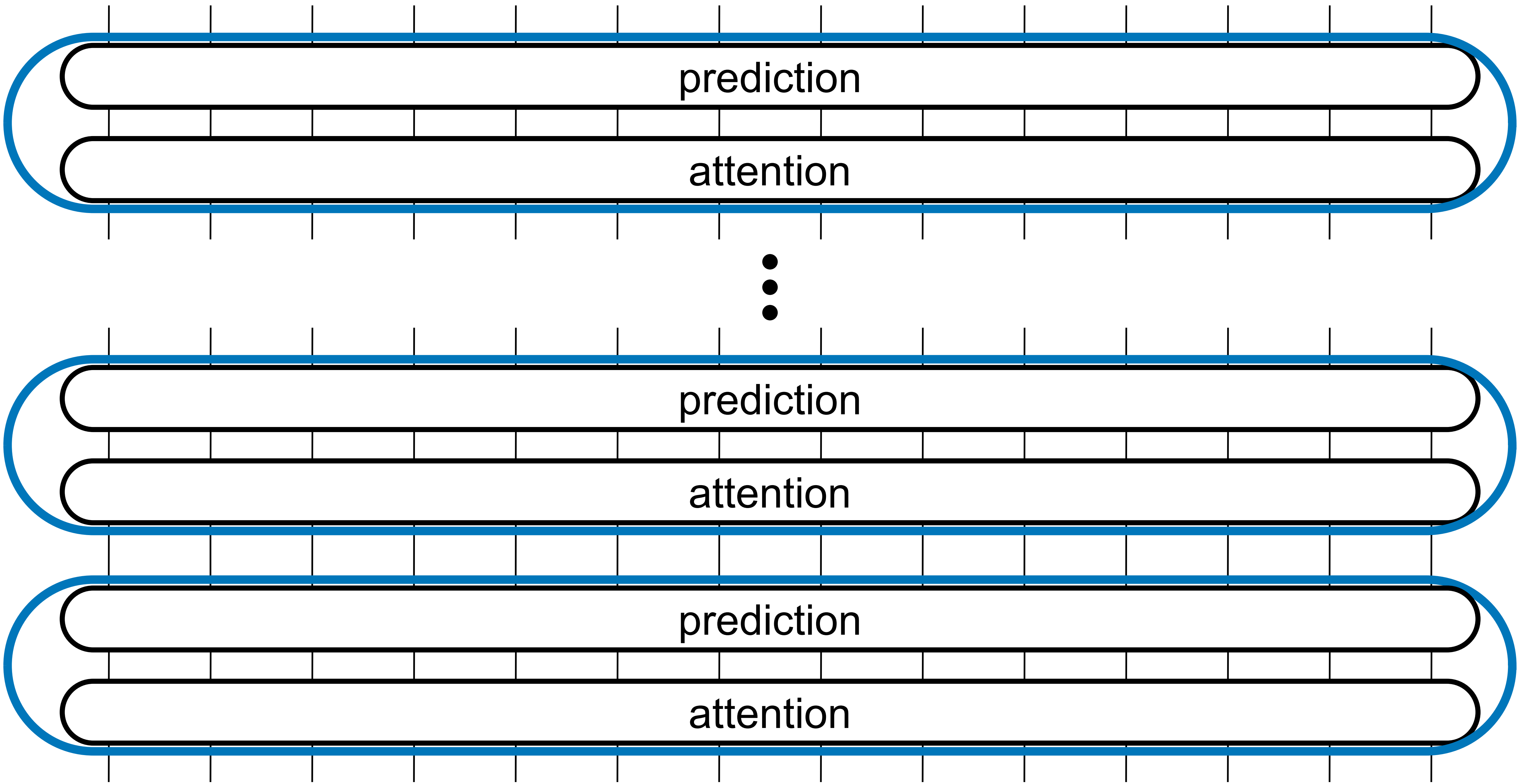
Syntax

slide from Steve Seitz's [video]

Semantics



slide from Steve Seitz's video



slide from Steve Seitz's video

How much data to train?

All of it...

All text on the internet?

Is that legal?

AI & Ethics!

All text on the internet?

Is that legal?

AI & Ethics!

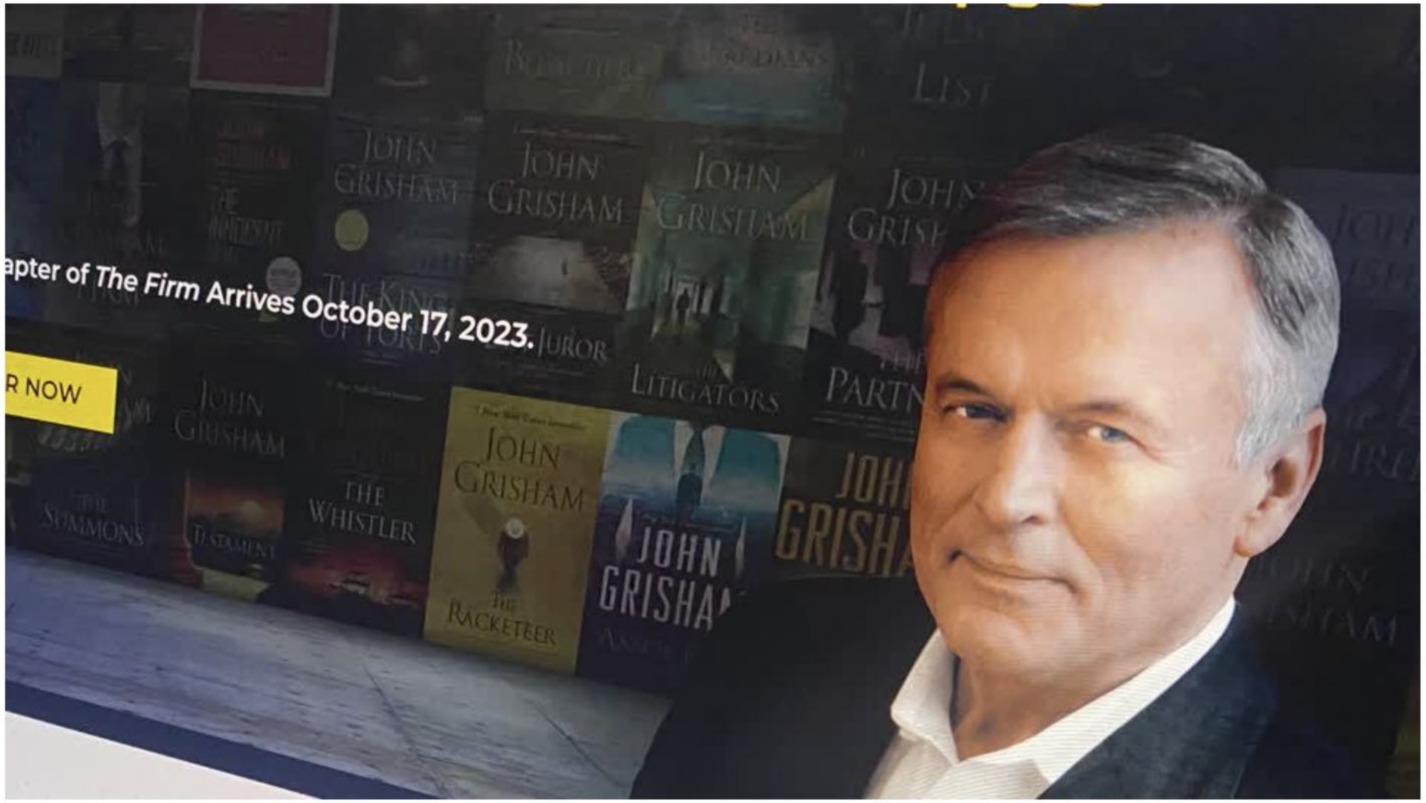
REUTERS® World Business Markets Sustainability Legal Breakingviews Technology Investig

Litigation | Copyright | Litigation | Technology | Intellectual Property

John Grisham, other top US authors sue OpenAI over copyrights

By **Blake Brittain**

September 21, 2023 6:34 AM EDT · Updated 7 months ago



All text on the internet?

Is that legal?

AI & Ethics!

December 27, 2023

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

All text on the internet?

Is that legal?

AI & Ethics!

Training the 175 billion parameters
of GPT-3 on
"all text on the internet"
on a single GPU or computer
would take 355 years
and $4,600,000

Lambdalabs.com

How long did it take OpenAI to train GPT-3?

a month

What did OpenAI train on?

GPT-3 training data[1]:9

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Source: Wikipedia

What about GPT-4?

- 1 trillion parameters
- Sam Altman stated that the cost of training GPT-4 was more than $100 million.

What about GPT-4?

- 1 trillion parameters
- Sam Altman stated that the cost of training GPT-4 was more than $100 million.

Why are the lawsuits so costly?

- We don’t know how to “untrain” neural networks.
- “Unlearning” is an exciting research area!

Source: Wikipedia

The 16th President was

The capital of Zimbabwe is

Frank Zappa's middle name is

Napoleon was born on this date

The prime factorization of 19456721434 is

Queen Victoria's maiden name was

US per-capita income in 1957 was

The lat long coordinates of Rome are

slide from Steve Seitz's [video]

The 16th President was Abraham Lincoln

The capital of Zimbabwe is Harare

Frank Zappa's middle name is Vincent

Napoleon was born on this date 1769

The prime factorization of 19456721434 is 2×3×3×17

Queen Victoria's maiden name was Alexandrina Victoria

US per-capita income in 1957 was $2,974

The lat long coordinates of Rome are 41.894722, 12.48

slide from Steve Seitz's video

a haiku about a turtle and a stop sign

slide from Steve Seitz's video

a haiku about a turtle and a stop sign

Turtle at the stop sign
waiting for the perfect moment
to cross the busy road

slide from Steve Seitz’s video

translate “tangled up in blue” to 5 languages

slide from Steve Seitz’s video

translate “tangled up in blue” to 5 languages

1. Spanish: Enredado en azul
2. French: Emmêlé dans le bleu
3. Italian: Intrecciato nel blu
4. German: Verwickelt in blau
5. Japanese: 青い糸に絡まって

slide from Steve Seitz’s [video]

the number after the number after 37 is 74

I'm heading north. I turn right. Then I turn left.
I'm now heading south

slide from Steve Seitz's video

the recipe for chocolate guacamole pancakes

Ingredients:
1 ripe avocado, mashed
1/4 cup cocoa powder
1/4 cup all-purpose flour
1 teaspoon baking powder
1/4 teaspoon salt
1/2 cup milk
1 egg
1 tablespoon vegetable oil
1/4 cup chocolate chips

Directions:
1. In a medium bowl, combine the avocado, cocoa powder, flour, baking powder, and salt.
2. In a separate bowl, whisk together the milk, egg, and vegetable oil.
3. Add the wet ingredients to the dry ingredients and mix until well combined.
4. Fold in the chocolate chips.
5. Heat a large skillet over medium heat.
6. Scoop 1/4 cup batter onto the skillet for each pancake.
7. Cook for 2-3 minutes per side, or until the pancakes are golden brown.
8. Serve with your favorite toppings.

slide from Steve Seitz's video

For more details: https://jalammar.github.io/illustrated-transformer/
https://jalammar.github.io/illustrated-gpt2/

slide from Steve Seitz's video

a pattern of characters that looks like a star

```
.  O  .
O  O  O
.  O  .
```

a pattern of characters that looks like a vertical line

.   ○   .

.   ○   .

.   ○   .

.   ○   .

.   ○   .

slide from Steve Seitz's [video]

a pattern of characters that looks like a triangle

. ○ .

. ○ ○ .

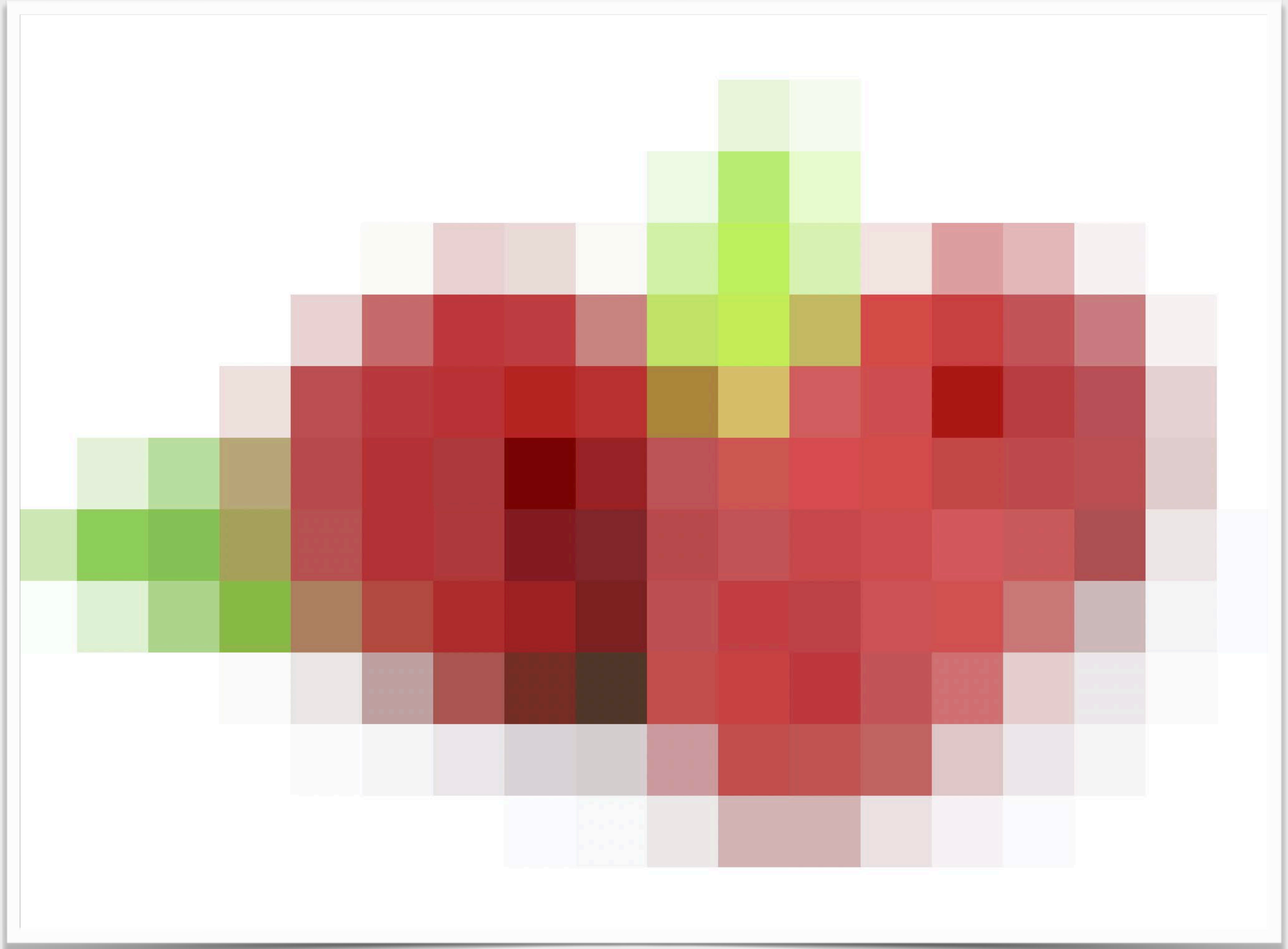
. ○ ○ ○ .

. ○ ○ ○ ○ .

. ○ ○ ○ ○ ○ .

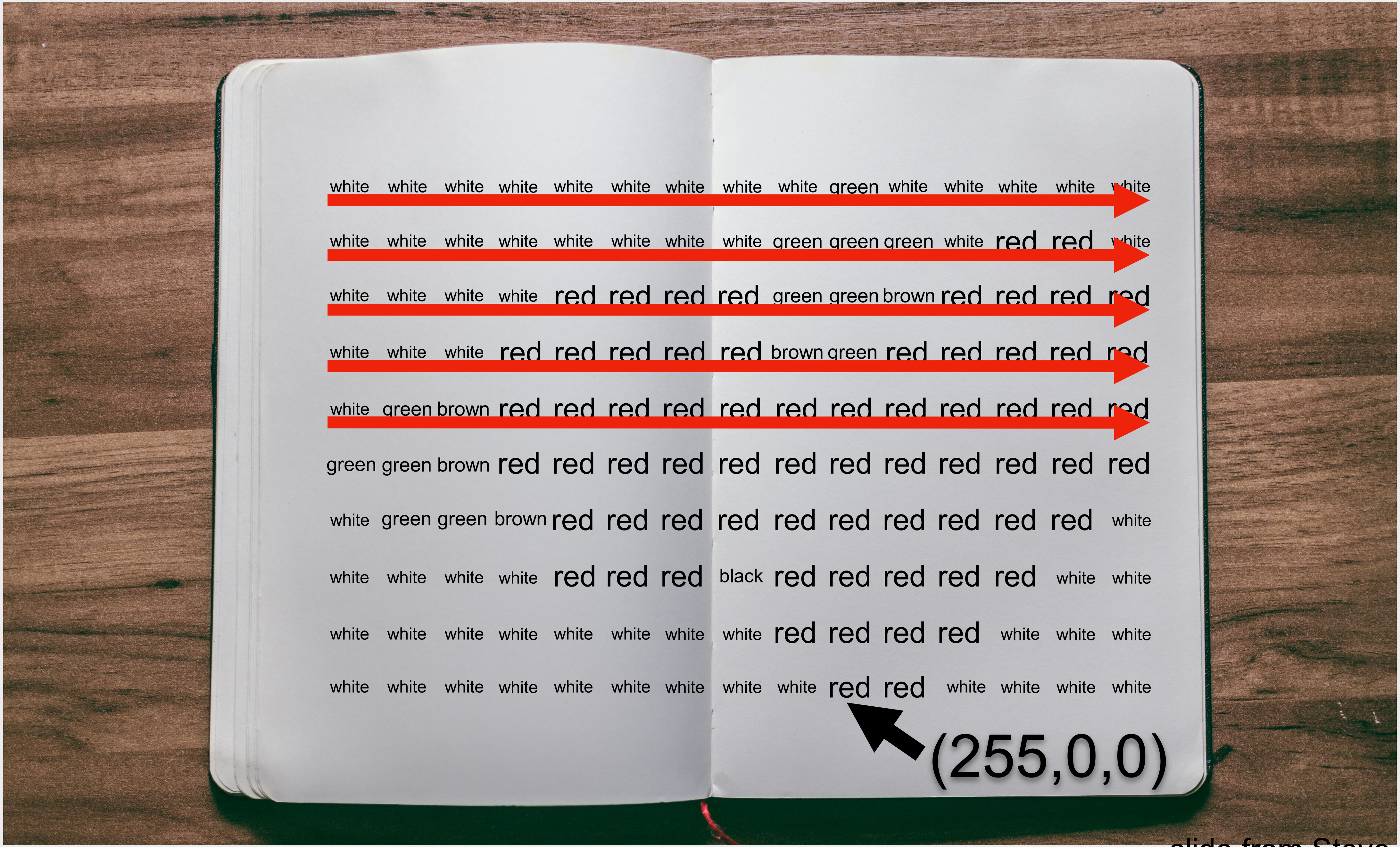
slide from Steve Seitz's [video]

slide from Steve Seitz's [video]

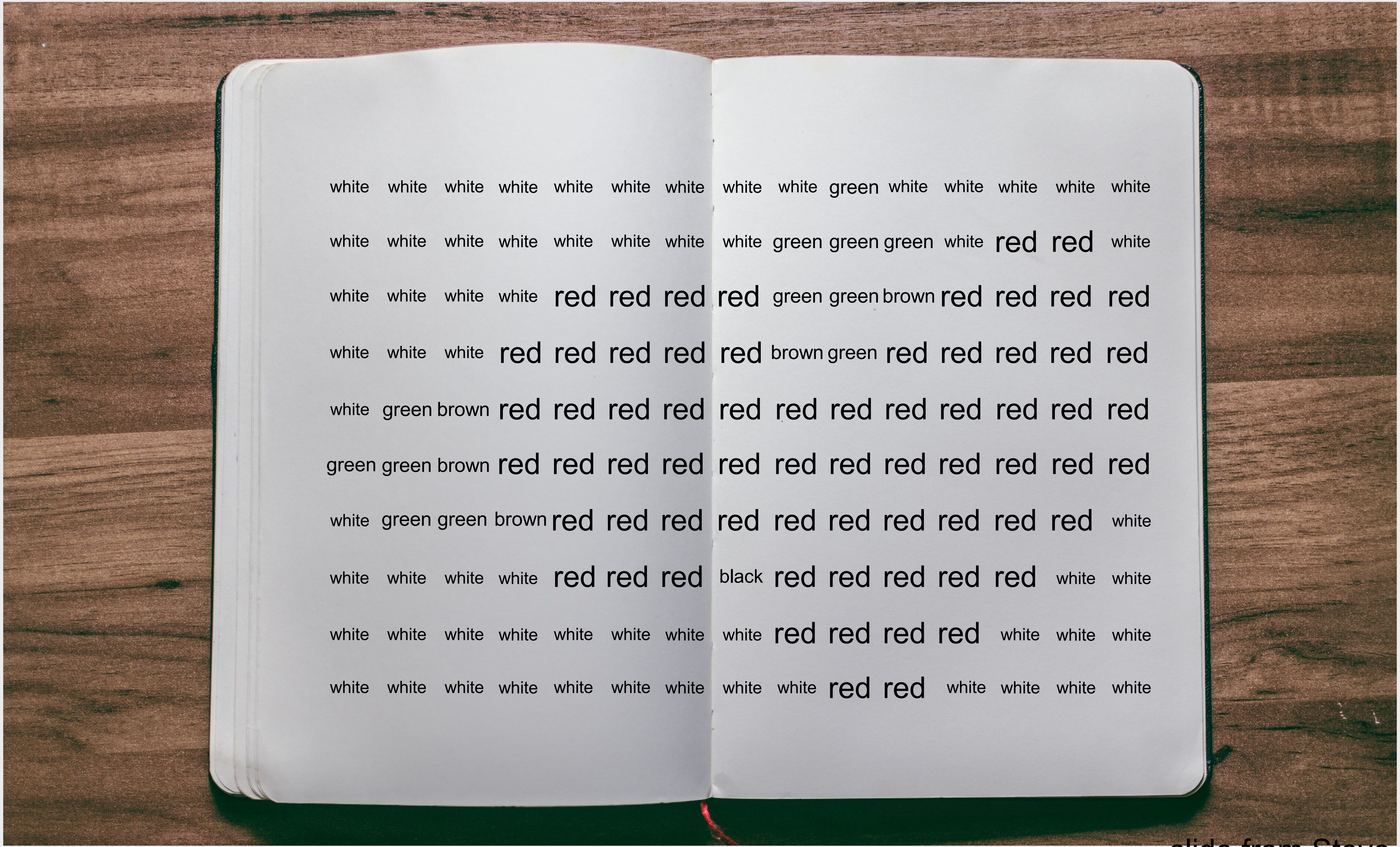


slide from Steve Seitz’s [video]

slide from Steve Seitz's [video]



slide from Steve Seitz's video



slide from Steve Seitz's [video]

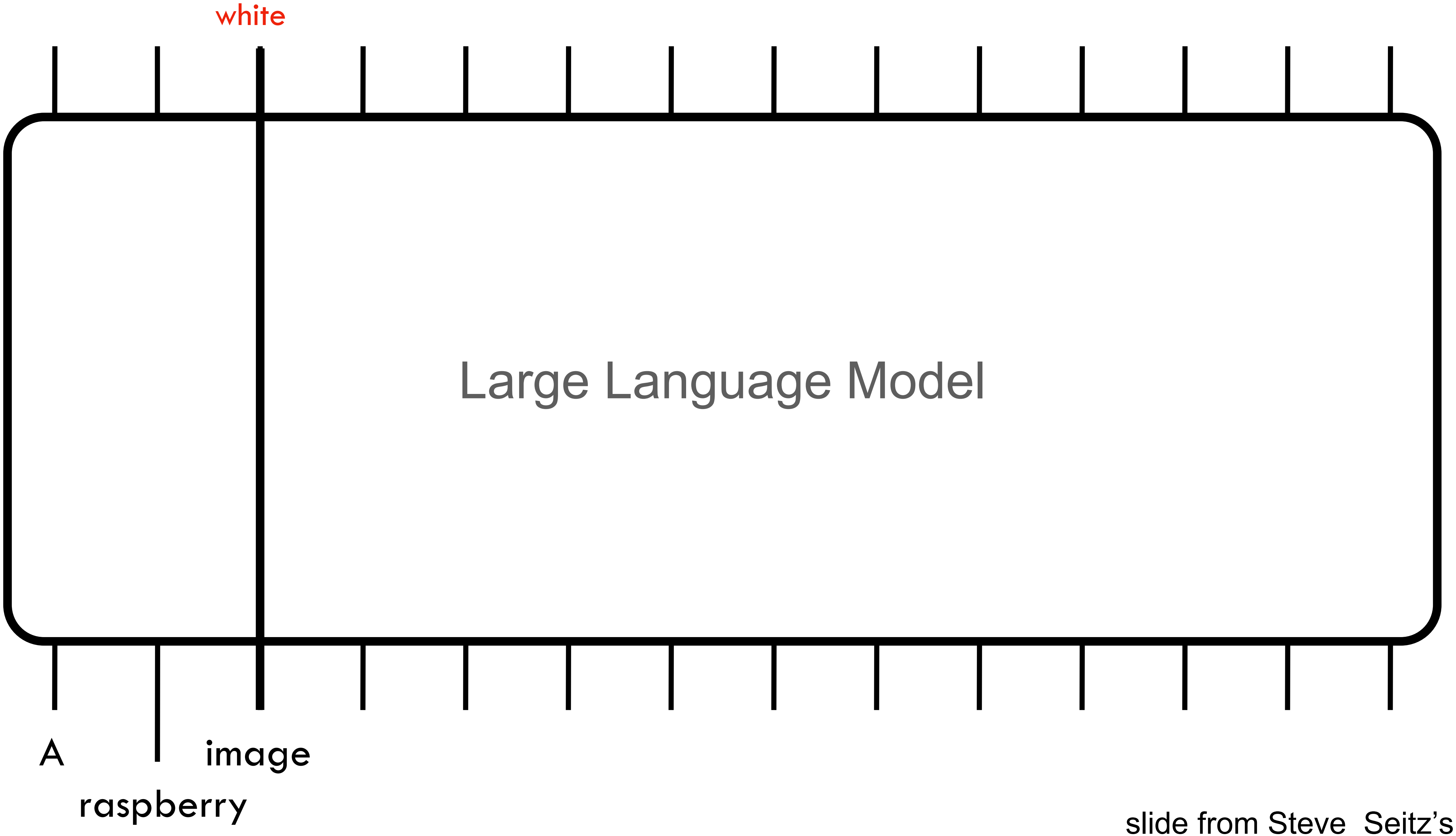
raspberr

1 Billion

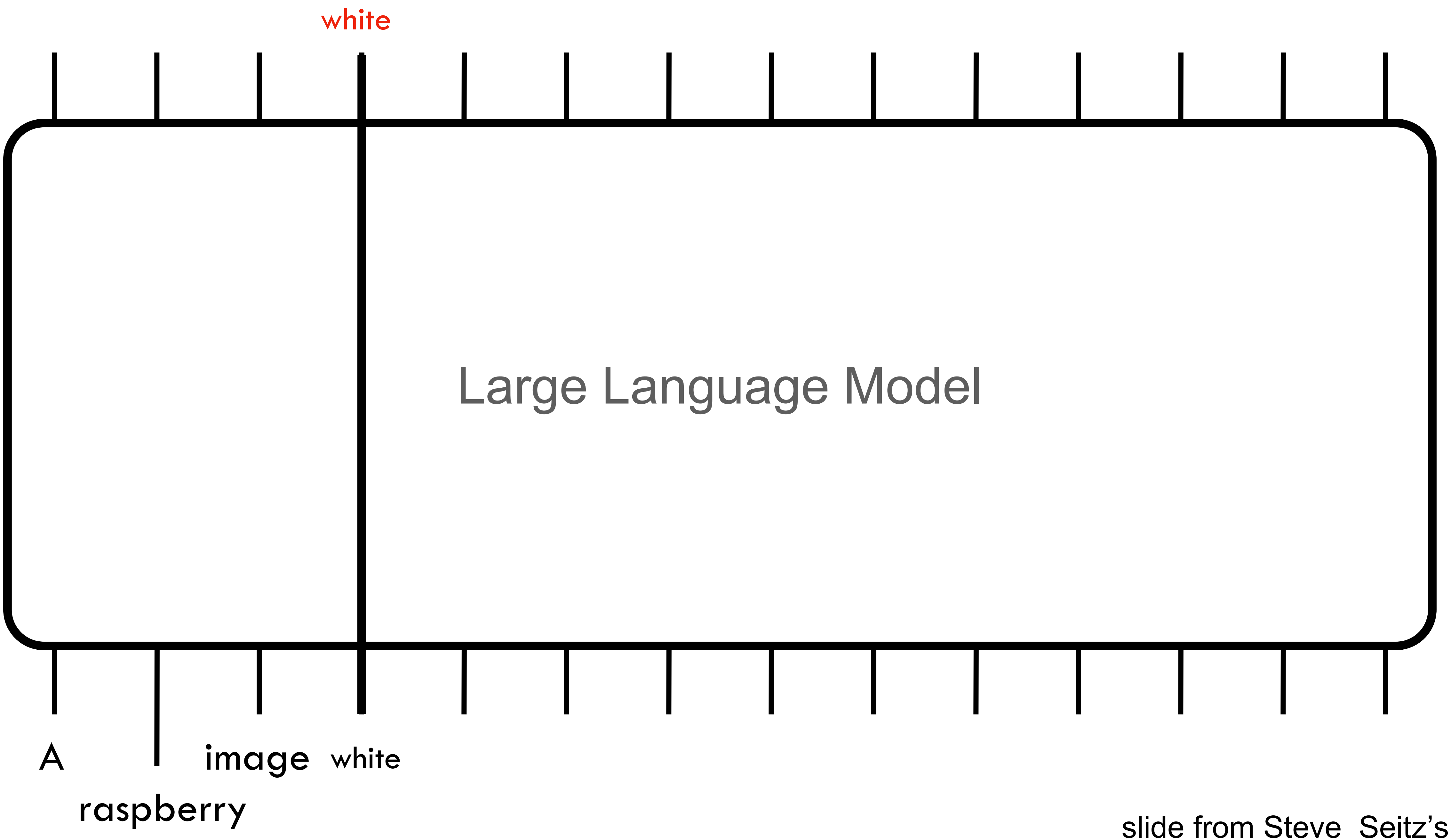
an akes

sunsets

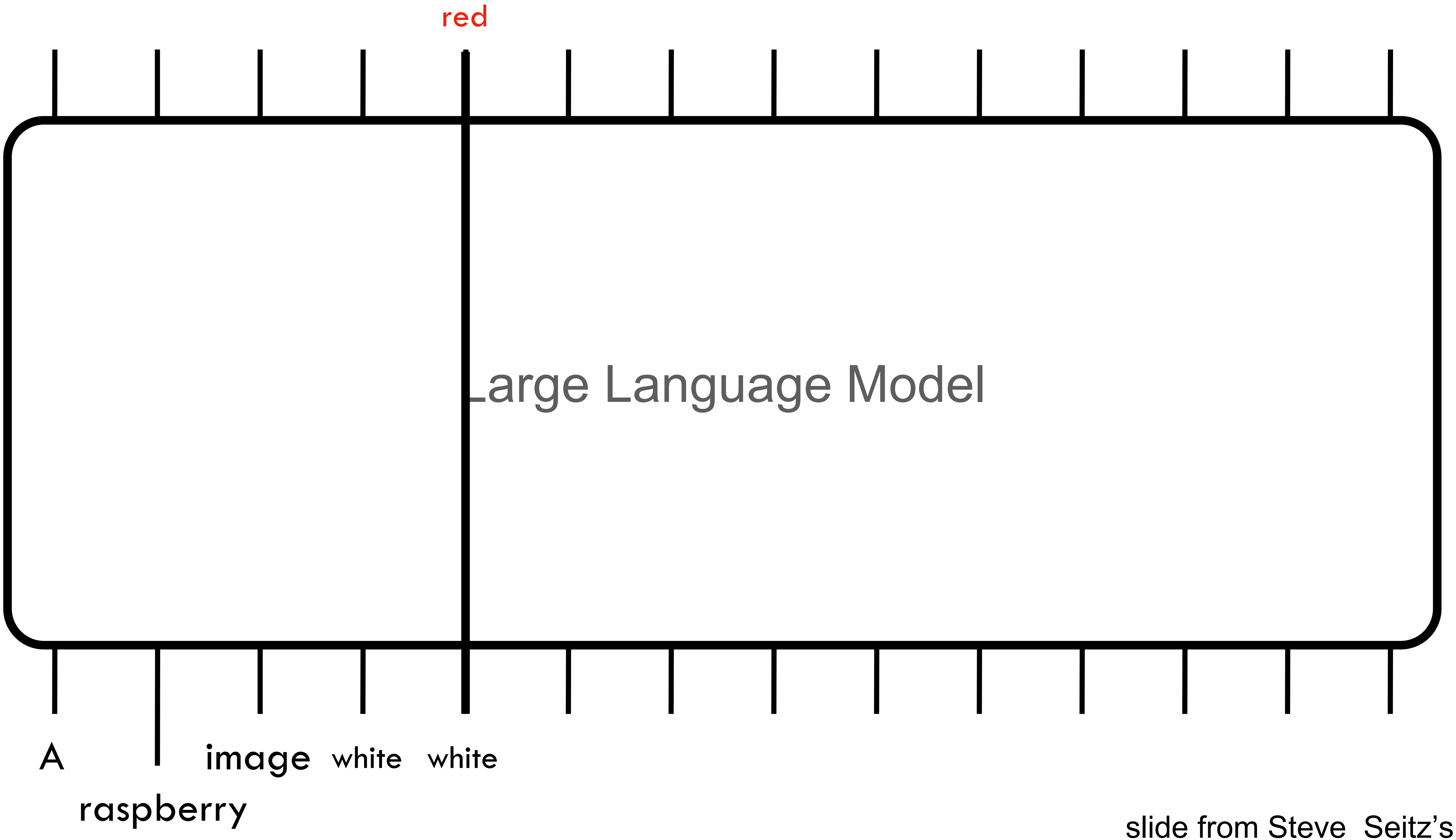
slide from Steve Seitz's [video]



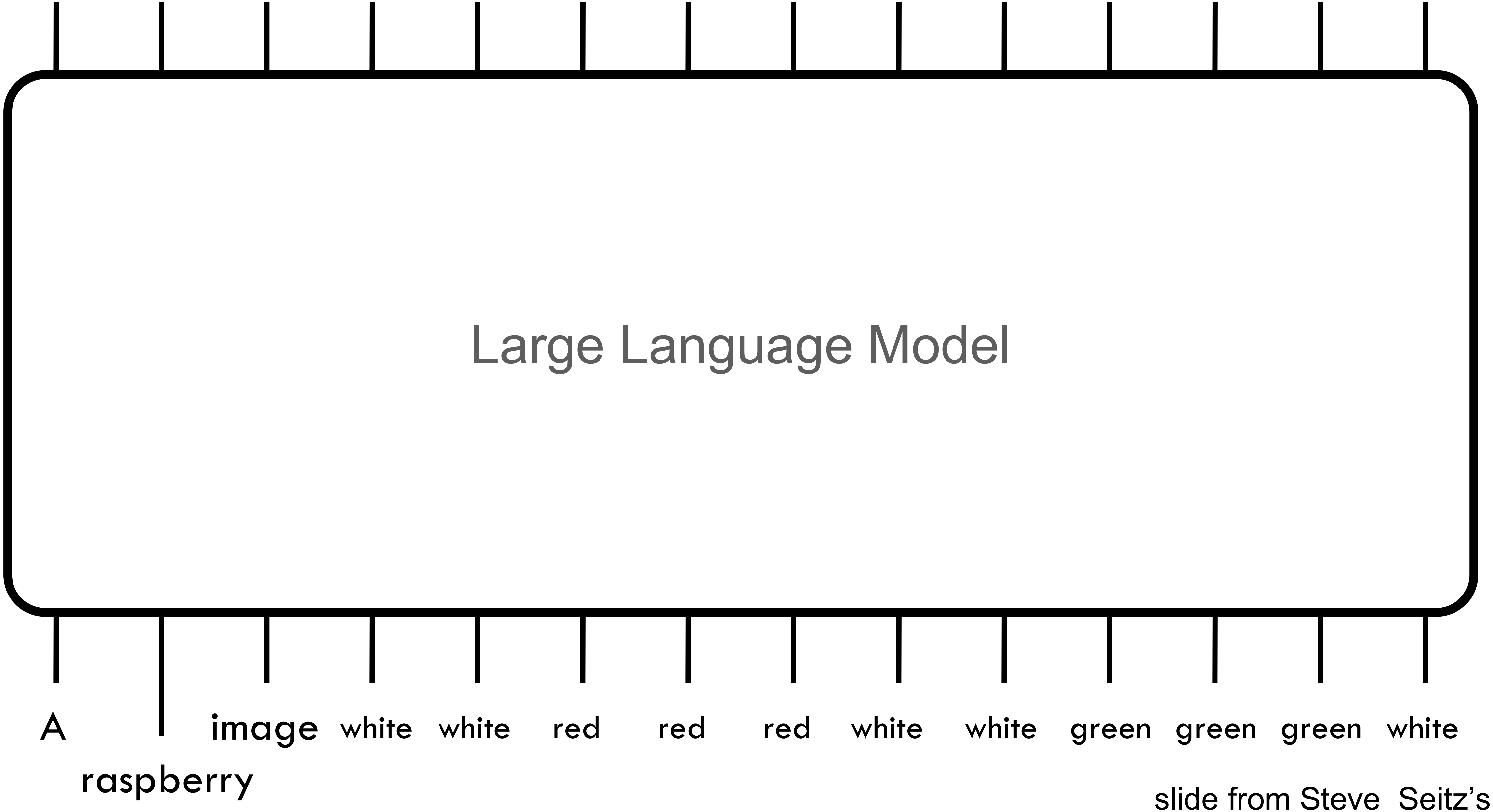
slide from Steve Seitz's [video]



slide from Steve Seitz's [video]

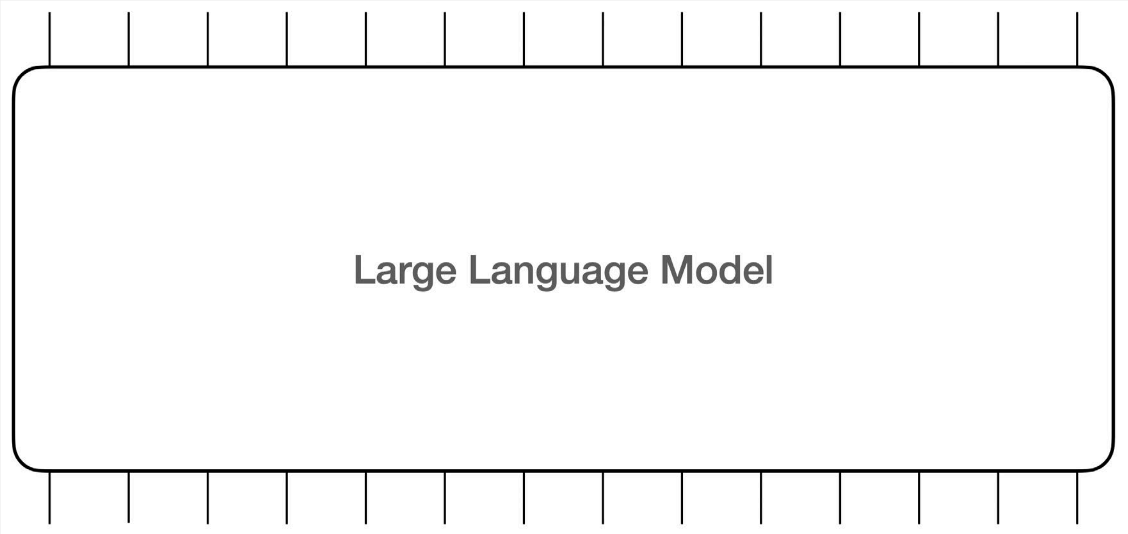


slide from Steve Seitz's [video]



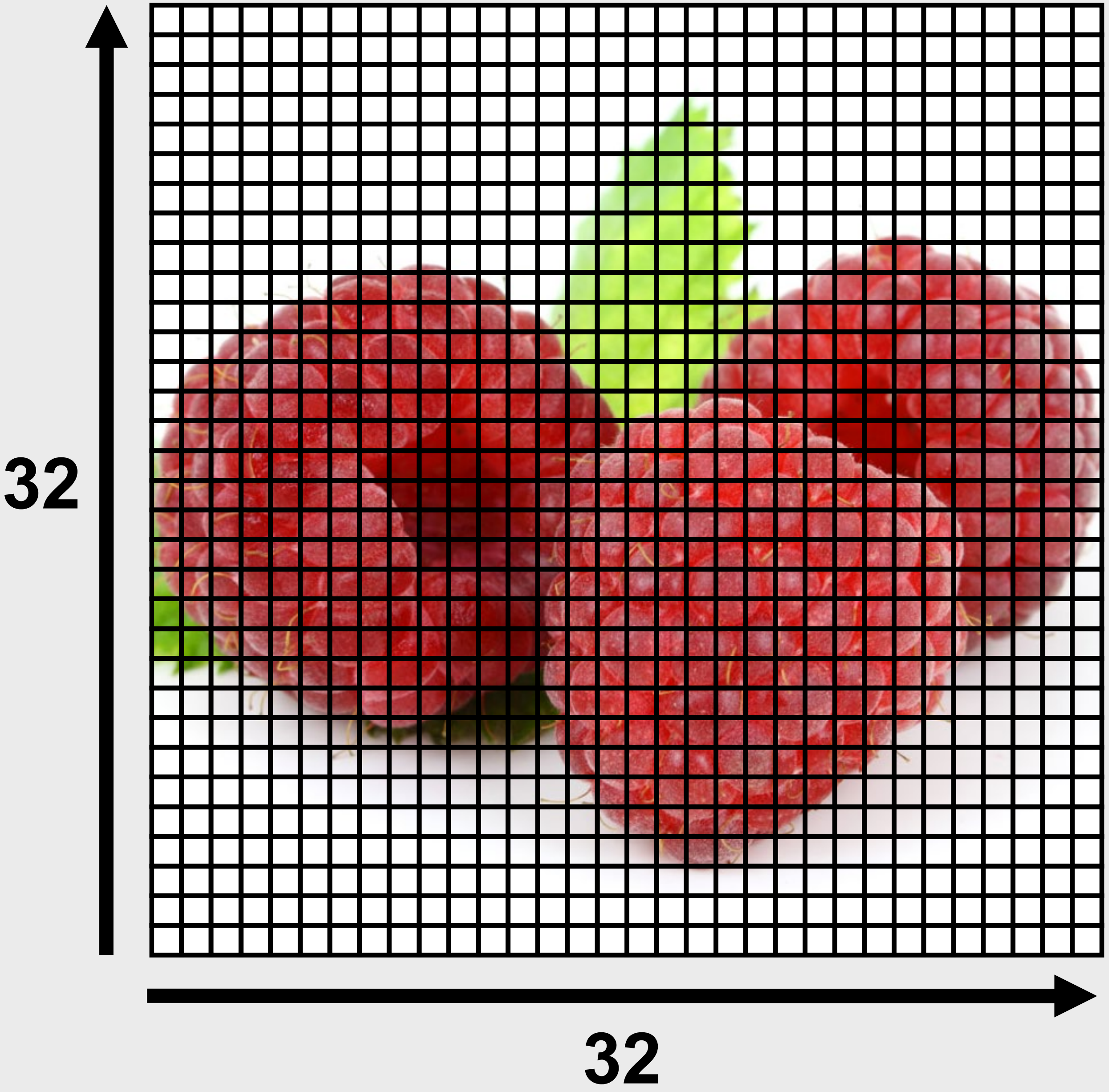
slide from Steve Seitz's [video]

1,000,000s of pixels

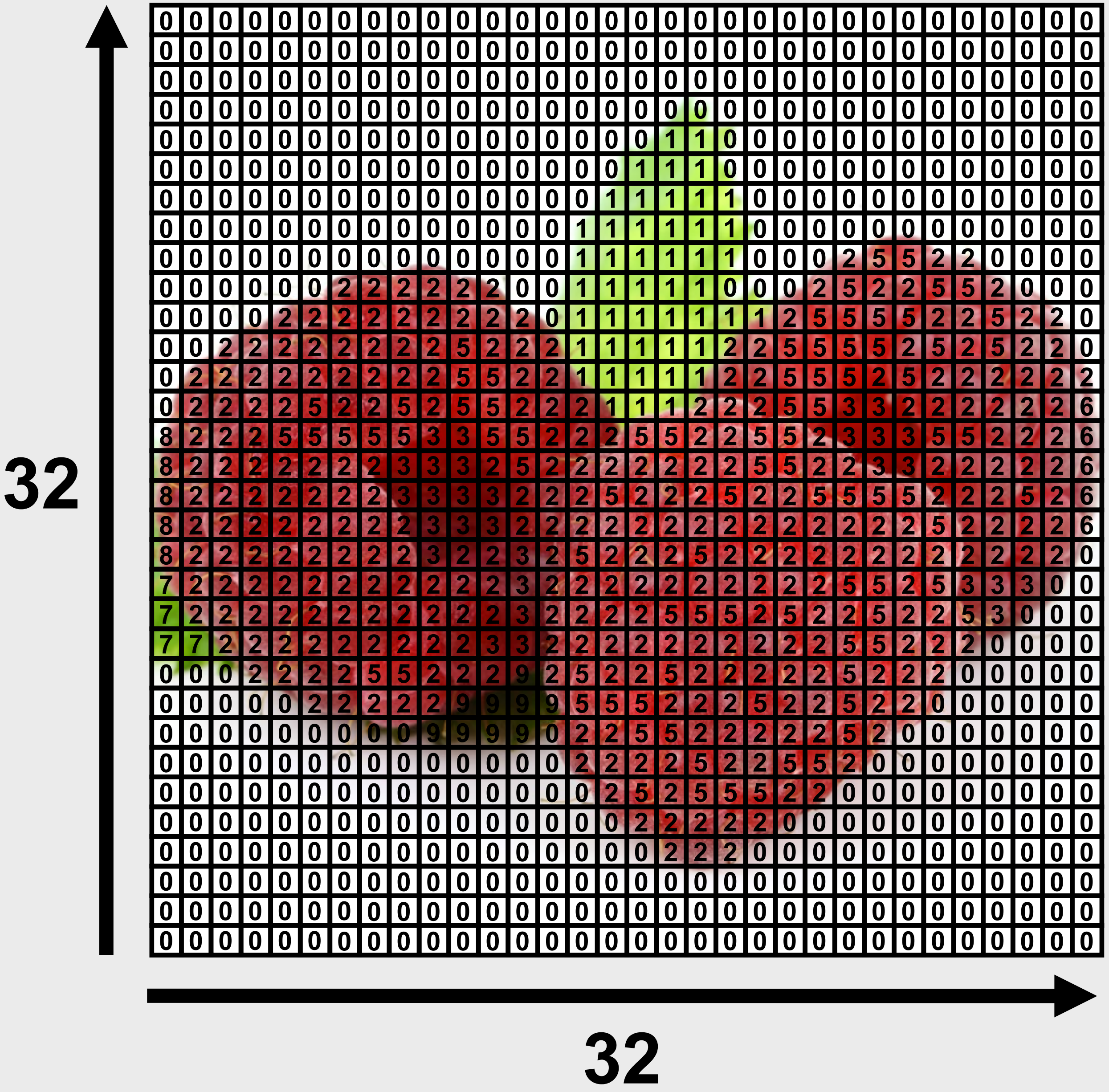


1,000s of words

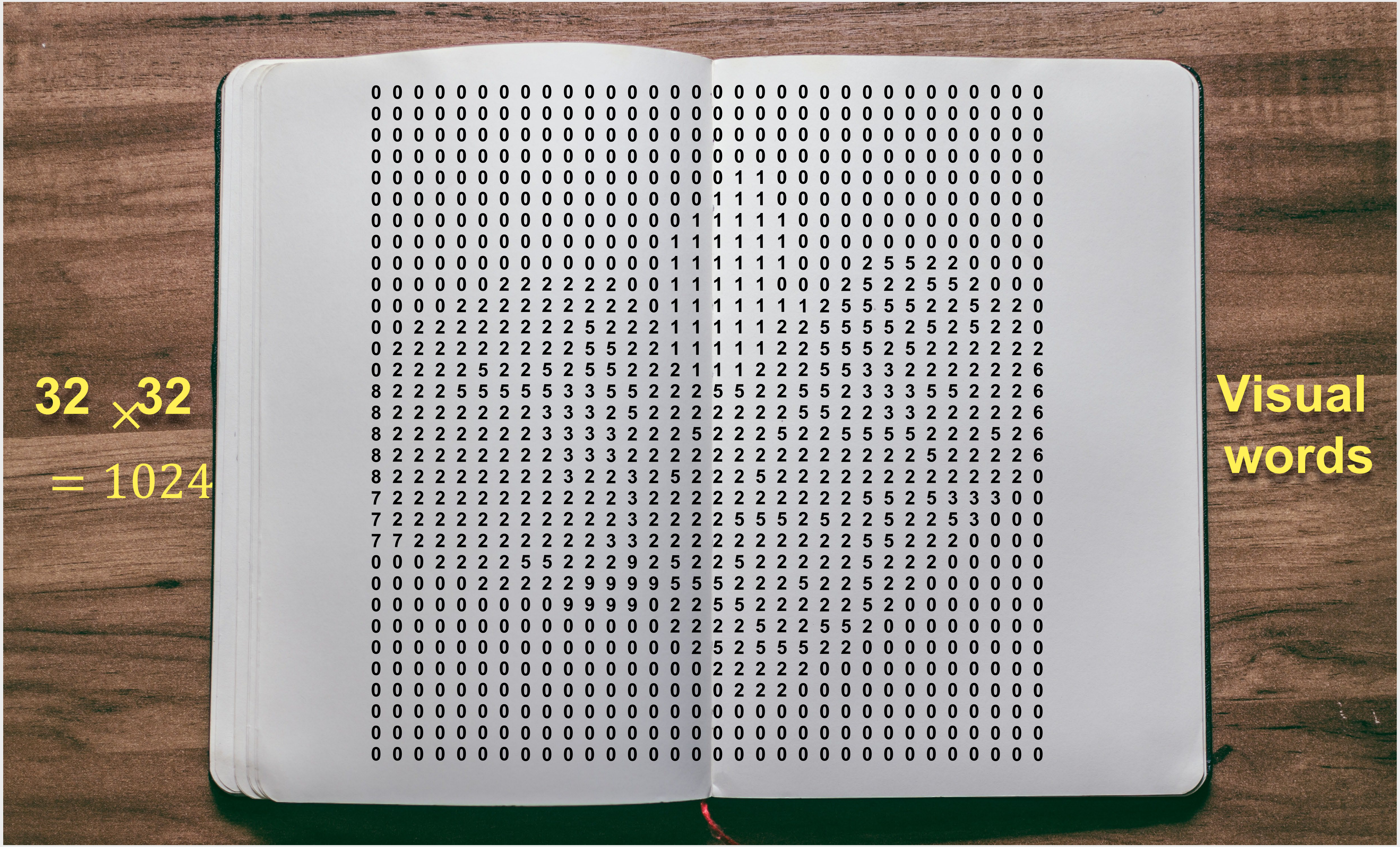
slide from Steve Seitz's video



slide from Steve Seitz's [video]



slide from Steve Seitz's [video]



slide from Steve Seitz's video

squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9 9
1 1 1 1 1 7 7 1 1 1 1 1 1 1 1 1

slide from Steve Seitz's [video]()

squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9 9
1 1 1 1 1 7 7 1 1 1 1 1 1 1 1 1 $\boxed{}$

slide from Steve Seitz's video

squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9 9
1 1 1 1 1 7 7 1 1 1 1 1 1 1 1 1

slide from Steve Seitz's [video]

squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9 9
1 1 1 1 1 7 7 $\boxed{1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ \mathbf{1}}$

slide from Steve Seitz's [video]

squirrel reaching for a nut

1 1 1 1 1 1 1 $\boxed{\begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 6 & 6 & 6 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \color{red}{\mathbf{6}} & & \end{matrix}}$ 1 1 6 6 4 4 4 9 9 9 9 9 9

1 1 1 1 1 7 7

slide from Steve Seitz's video

squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9 9
1 1 1 1 1 7 7 1 1 1 1 1 1 1 1 1 **6**

slide from Steve Seitz's video

squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9 9
1 1 1 1 1 7 7 1 1 1 1 1 1 1 1 1 **6**

slide from Steve Seitz's [video]

squirrel reaching for a nut

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 6 6 6 1 1 6 6 4 4 4 9 9 9 9 9 9
1 1 1 1 1 7 7 1 1 1 1 1 1 1 1 1 6

slide from Steve Seitz's video

squirrel reaching for a nut

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6	6	6	1	1	6	6	4	4	4	9	9	9	9	9	9
1	1	1	1	1	7	7	1	1	1	1	1	1	1	1	1	6	6	6	6	6	6	6	4	1	4	9	9	9	9	9	9
1	1	7	1	1	1	7	1	1	1	1	1	1	1	1	6	6	6	6	6	6	6	6	4	1	4	9	9	9	9	9	9
1	1	7	1	1	1	1	1	1	1	1	1	1	1	2	6	2	6	6	2	2	6	2	9	9	9	9	9	9	9	9	9
1	1	1	7	7	1	1	1	1	1	1	1	1	1	6	6	6	6	2	5	2	2	4	9	9	9	9	9	9	9	9	9
1	1	1	7	1	1	1	1	1	1	1	1	1	1	2	6	6	6	2	5	2	2	0	9	9	9	9	9	9	9	9	9
1	1	1	1	1	1	2	2	2	2	2	2	1	2	6	6	6	2	6	2	6	2	0	9	9	9	9	9	9	9	9	9
2	1	1	1	1	2	2	1	2	2	2	2	1	2	6	6	6	6	6	2	6	4	9	9	9	9	9	9	9	8	8	8
2	2	1	1	1	1	1	1	1	1	1	1	1	1	6	6	2	6	6	6	4	4	9	9	9	9	9	9	9	9	8	8
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6	6	0	0	0	0	0	4	9	9	9	9	9	9	9	9	8
1	1	1	1	1	1	1	1	1	7	1	5	2	2	2	0	0	0	0	0	0	4	9	9	9	9	9	9	9	9	9	8
1	1	1	1	1	1	1	1	1	1	5	2	2	2	2	0	0	0	0	4	4	6	9	9	9	9	9	9	9	9	9	9
1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	5	2	0	5	4	6	9	9	9	9	9	9	9	9	9	9	9
1	1	1	1	7	1	1	1	1	1	2	2	6	6	5	5	2	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
1	1	1	1	7	2	2	2	1	1	2	2	6	5	5	0	2	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
1	1	1	7	2	2	2	2	1	1	2	6	6	5	5	0	0	4	4	4	9	9	0	0	0	0	9	9	9	9	9	9
1	1	0	1	2	1	2	1	1	1	1	1	1	1	0	0	0	0	4	0	0	0	4	0	0	0	0	9	9	9	9	8
1	1	1	1	1	1	1	1	1	1	1	7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	8
1	1	1	1	1	1	1	1	1	7	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	9	9	8	8
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4	4	9	9	9	8	9
1	1	1	1	1	7	7	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4	4	9	9	5	5	9
1	1	1	1	1	7	7	7	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4	4	9	9	5	4	3
1	1	1	1	1	1	7	7	3	3	3	3	3	3	3	4	0	4	0	0	4	0	0	0	0	4	4	9	9	5	4	4
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	5	0	0	0	0	0	0	0	0	4	9	9	9	4	3	3
3	3	3	3	3	3	3	3	3	3	4	4	4	3	3	3	5	3	0	4	4	4	4	4	4	4	9	9	9	4	3	8
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	8	3	4	4	4	4	4	9	3	9	4	8	8
3	3	8	3	3	3	3	3	3	3	3	3	3	8	3	8	8	3	3	8	3	8	4	4	8	4	8	3	8	3	3	3
8	8	3	8	8	3	8	3	3	3	8	3	8	3	3	3	3	3	3	8	3	8	4	8	8	8	3	3	3	3	3	3
3	3	8	3	3	3	3	3	3	3	8	8	8	8	3	3	8	8	3	3	8	8	8	8	8	8	8	8	3	8	3	8
3	3	8	8	3	3	3	3	8	3	3	3	3	3	3	3	8	8	3	3	8	3	3	3	8	3	8	8	8	8	3	8
3	3	8	8	3	3	3	8	8	3	8	3	3	3	3	3	3	3	3	3	3	3	3	3	8	3	8	8	8	8	8	8
3	3	8	3	3	3	3	3	8	3	8	8	3	3	3	3	3	3	3	3	3	3	3	3	3	3	8	8	8	8	8	8

slide from Steve Seitz's video

squirrel reaching for a nut

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6	6	6	1	1	6	6	4	4	4	9	9	9	9	9	9
1	1	1	1	1	7	7	1	1	1	1	1	1	1	1	1	6	6	6	6	6	6	6	4	1	4	9	9	9	9	9	9
1	1	7	1	1	1	7	1	1	1	1	1	1	1	1	6	6	6	6	6	6	6	6	4	1	4	9	9	9	9	9	9
1	1	7	1	1	1	1	1	1	1	1	1	1	1	2	6	2	6	6	2	2	6	2	9	9	9	9	9	9	9	9	9
1	1	1	7	7	1	1	1	1	1	1	1	1	1	6	6	6	6	2	5	2	2	4	9	9	9	9	9	9	9	9	9
1	1	1	7	1	1	1	1	1	1	1	1	1	1	2	6	6	6	2	5	2	2	0	9	9	9	9	9	9	9	9	9
1	1	1	1	1	1	2	2	2	2	2	2	1	2	6	6	6	2	6	2	6	2	0	9	9	9	9	9	9	9	9	9
2	1	1	1	1	2	2	1	2	2	2	2	1	2	6	6	6	6	6	2	6	4	9	9	9	9	9	9	9	8	8	8
2	2	1	1	1	1	1	1	1	1	1	1	1	1	6	6	2	6	6	6	4	4	9	9	9	9	9	9	9	9	8	8
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6	6	0	0	0	0	0	4	9	9	9	9	9	9	9	9	8
1	1	1	1	1	1	1	1	1	7	1	5	2	2	2	0	0	0	0	0	0	4	9	9	9	9	9	9	9	9	9	8
1	1	1	1	1	1	1	1	1	1	5	2	2	2	2	0	0	0	0	4	4	6	9	9	9	9	9	9	9	9	9	9
1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	5	2	0	5	4	6	9	9	9	9	9	9	9	9	9	9	9
1	1	1	1	7	1	1	1	1	1	2	2	6	6	5	5	2	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
1	1	1	1	7	2	2	2	1	1	2	2	6	5	5	0	2	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
1	1	1	7	2	2	2	2	1	1	2	6	6	5	5	0	0	4	4	4	9	9	0	0	0	0	9	9	9	9	9	9
1	1	0	1	2	1	2	1	1	1	1	1	1	1	0	0	0	0	4	0	0	0	4	0	0	0	0	9	9	9	9	8
1	1	1	1	1	1	1	1	1	1	1	7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	9	9	9	8
1	1	1	1	1	1	1	1	1	7	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	9	9	8	8
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4	4	9	9	9	8	9
1	1	1	1	1	7	7	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4	4	9	9	5	5	9
1	1	1	1	1	7	7	7	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4	4	9	9	5	4	3
1	1	1	1	1	1	7	7	3	3	3	3	3	3	3	4	0	4	0	0	4	0	0	0	0	4	4	9	9	5	4	4
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	5	0	0	0	0	0	0	0	0	4	9	9	9	4	3	3
3	3	3	3	3	3	3	3	3	3	4	4	4	3	3	3	5	3	0	4	4	4	4	4	4	4	9	9	9	4	3	8
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	8	3	4	4	4	4	4	9	3	9	4	8	8
3	3	8	3	3	3	3	3	3	3	3	3	3	8	3	8	8	3	3	8	3	8	4	4	8	4	8	3	8	3	3	3
8	8	3	8	8	3	8	3	3	3	8	3	8	3	3	3	3	3	3	8	3	8	4	8	8	8	3	3	3	3	3	3
3	3	8	3	3	3	3	3	3	3	8	8	8	8	3	3	8	8	3	3	8	8	8	8	8	8	8	8	3	8	3	8
3	3	8	8	3	3	3	3	8	3	3	3	3	3	3	3	8	8	3	3	8	3	3	3	8	3	8	8	8	8	3	8
3	3	8	8	3	3	3	8	8	3	8	3	3	3	3	3	3	3	3	3	3	3	3	3	8	3	8	8	8	8	8	8
3	3	8	3	3	3	3	3	8	3	8	8	3	3	3	3	3	3	3	3	3	3	3	3	3	3	8	8	8	8	8	8

slide from Steve Seitz’s video

squirrel reaching for a nut

slide from Steve Seitz's video

squirrel reaching for a nut

slide from Steve Seitz's [video]

squirrel reaching for a nut

Up-sampled 4x

slide from Steve Seitz's [video]

squirrel reaching for a nut

Parti, https://parti.research.google/

squirrel reaching for a nut underwater

slide from Steve Seitz's [video]

fossil of a squirrel reaching for a nut

slide from Steve Seitz's [video]()

squirrel made of toothpicks wearing sunglasses reaching for a nut

slide from Steve Seitz's video

DLSR photograph of a whimsical fantasy house shaped like a squirrel with windows and a door, in the forest

slide from Steve Seitz's video

Squirrel reaching for a nut. by Leonardo da Vinci

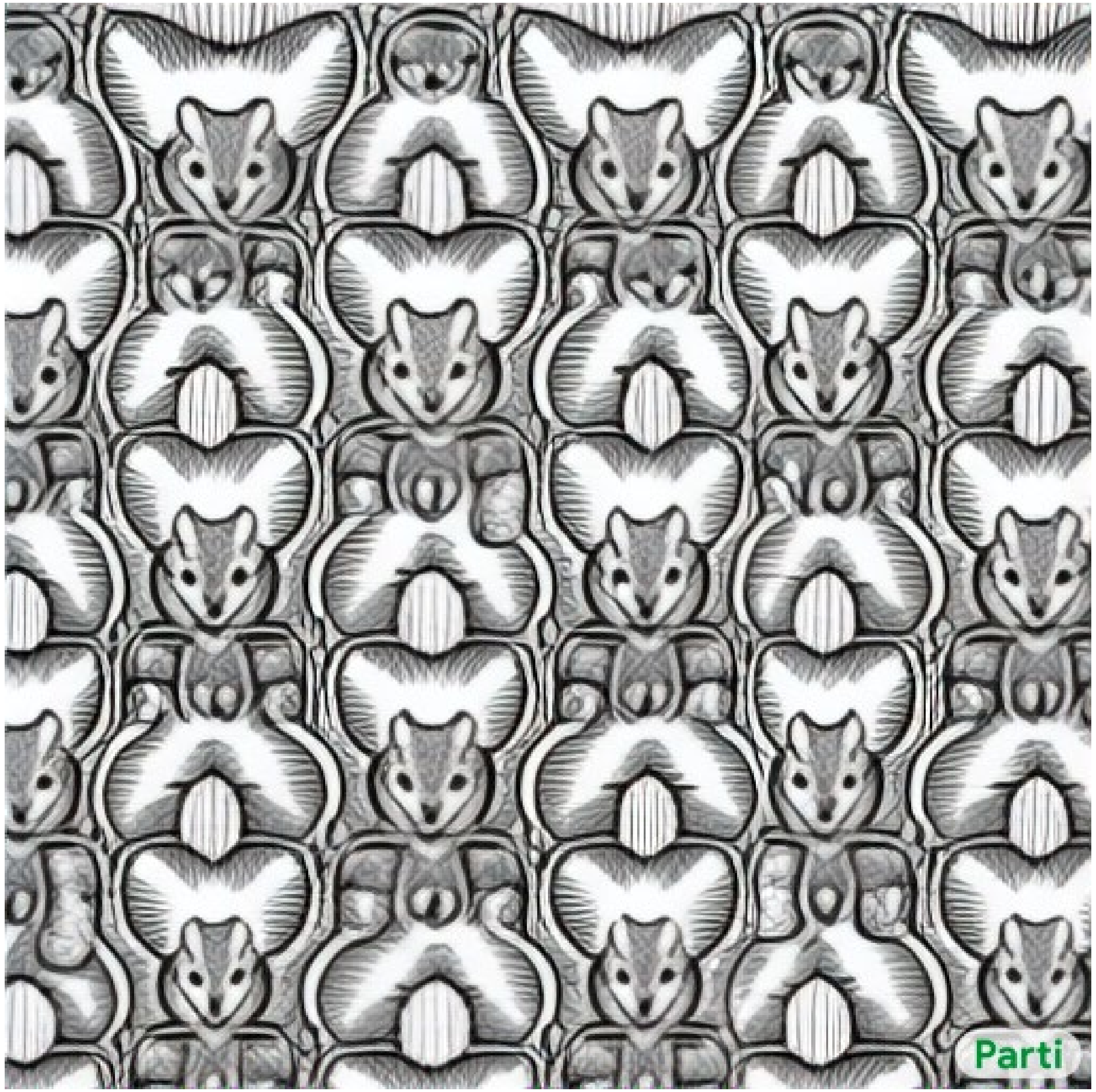
slide from Steve Seitz's [video]

Squirrel reaching for a nut.   Van Gogh painting

slide from Steve Seitz's [video]

Intricately carved cathedral door of a squirrel reaching for a nut

slide from Steve Seitz's video



Squirrel reaching for a nut. Woodcut tessellation pattern by M.C. Escher

slide from Steve Seitz's [video]

Squirrel reaching for a nut. Latte art

slide from Steve Seitz's video

Vaswani et al., 2017

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Vaswani et al., 2017

Attention Is All You Need

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

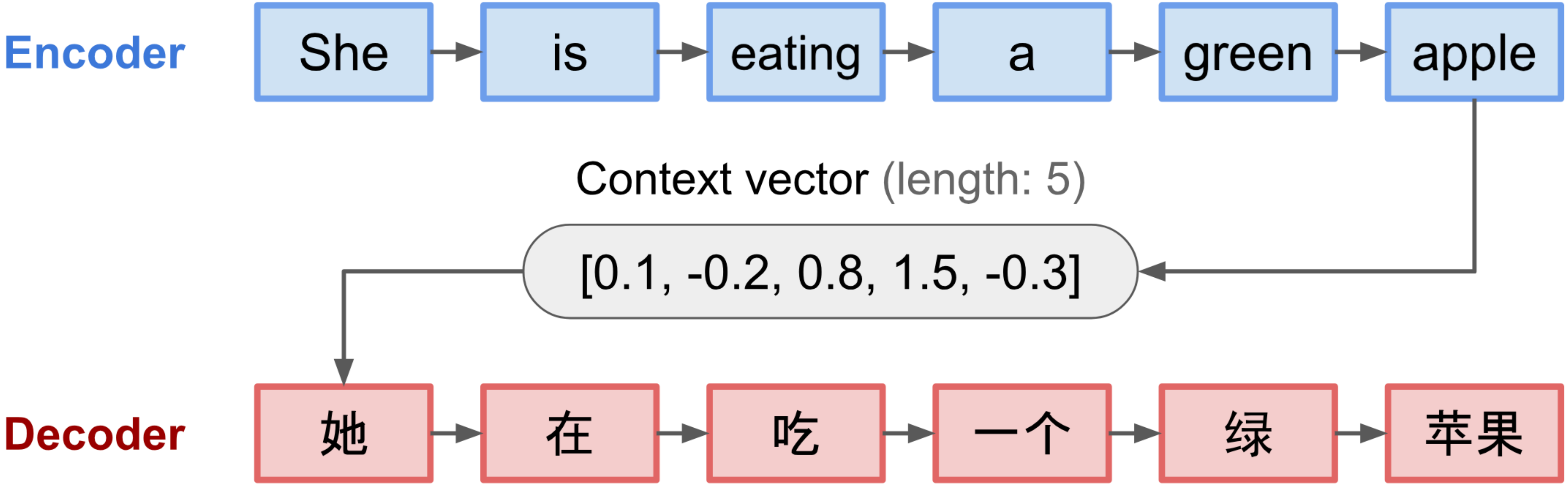
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

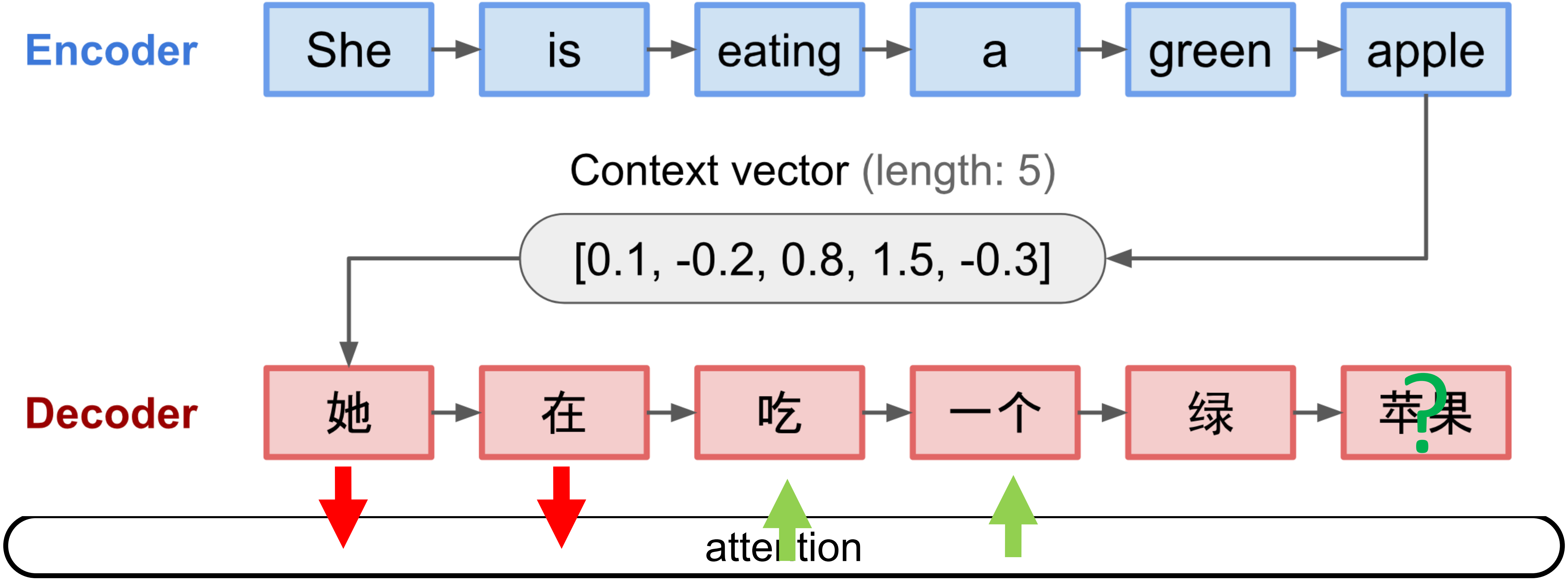
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Sequence 2 Sequence models in language



Attention and Context in language



Self-Attention

- Content-based querying
- Retrieves similar items
- Weighted sum of similarities
- Constant path length between any two positions
- Variable-sized perceptive field
- Gating/multiplication enables crisp error propagation
- Trivial to parallelize (per layer)
- Can replace sequence-aligned recurrence entirely

Self-Attention Order in Machine Translation

- Encoder-Decoder Attention:
    - from output attending to words in input sequence
- Encoder Self-Attention:
    - attention to words in input sequence (all directions)
- Masked Decoder Self-Attention
    - in output attending only to words that come before

Self-Attention Order in Machine Translation

- Encoder-Decoder Attention:
    - from output attending to words in input sequence
- Encoder Self-Attention:
    - attention to words in input sequence (all directions)

You cannot use this if you are predicting the output

- Masked Decoder Self-Attention
    - in output attending only to words that come before

Use this instead!

Self-Attention Order in Machine Translation

- Encoder-Decoder Attention:
  - from output attending to words in input sequence
- Encoder Self-Attention:
  - attention to words in input sequence (all directions)

You cannot use this if you are predicting the output

- Masked Decoder Self-Attention
  - in output attending only to words that come before

Use this instead!

BUT with word-by-word processing this would take a very long time to train!

Transformer Architecture

Vaswani et al., 2017

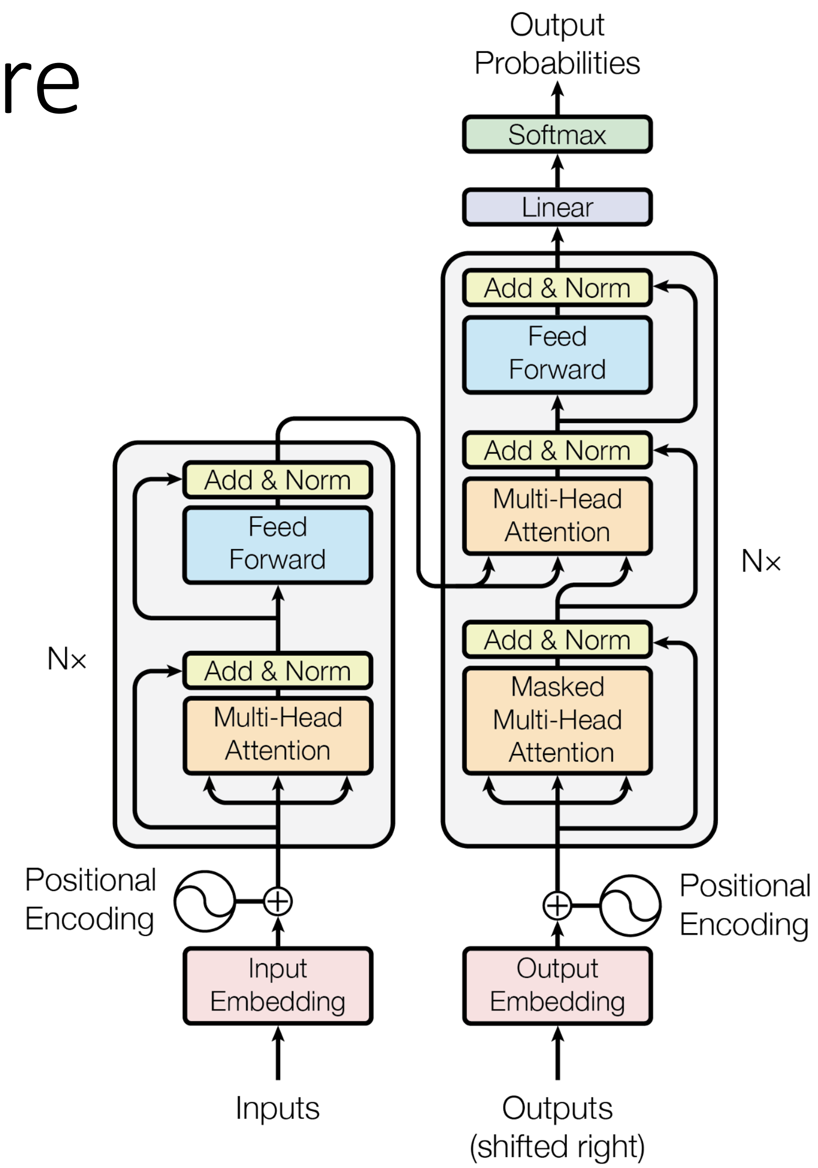


Figure 1: The Transformer - model architecture.

Transformer Architecture

Input branch

Feed forward network processes every English word

English Sentence

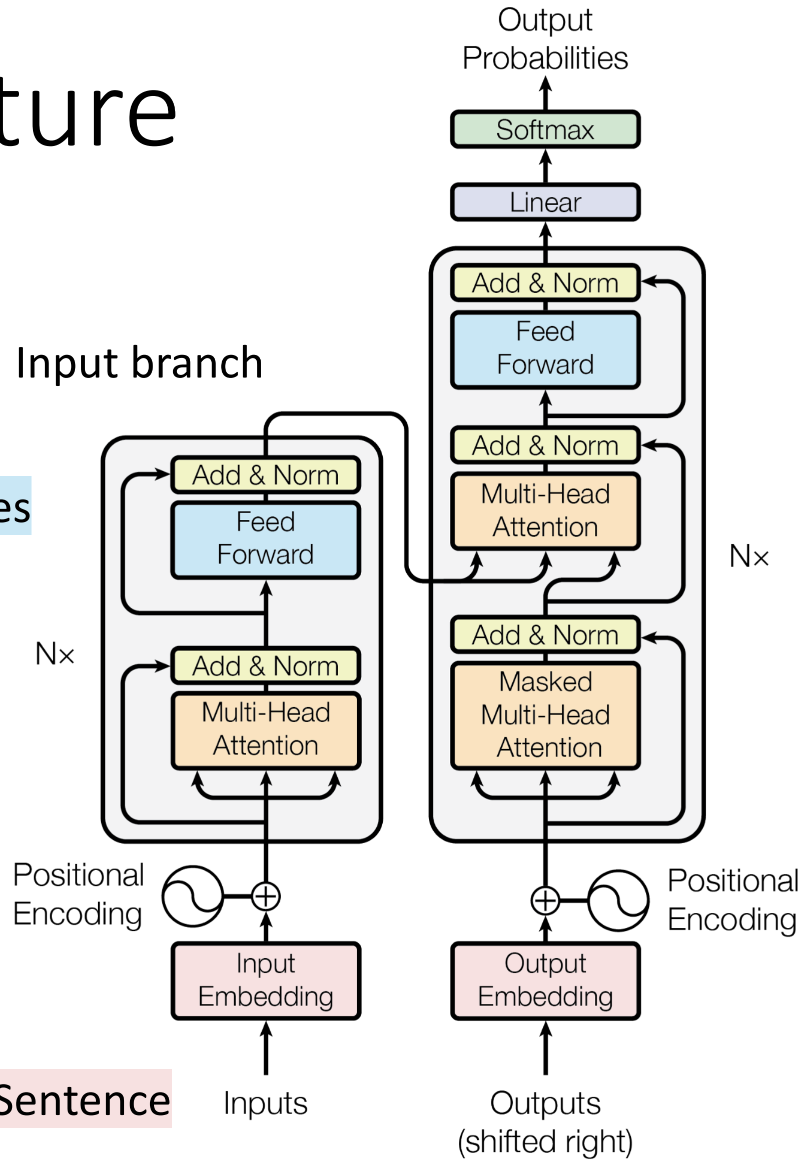
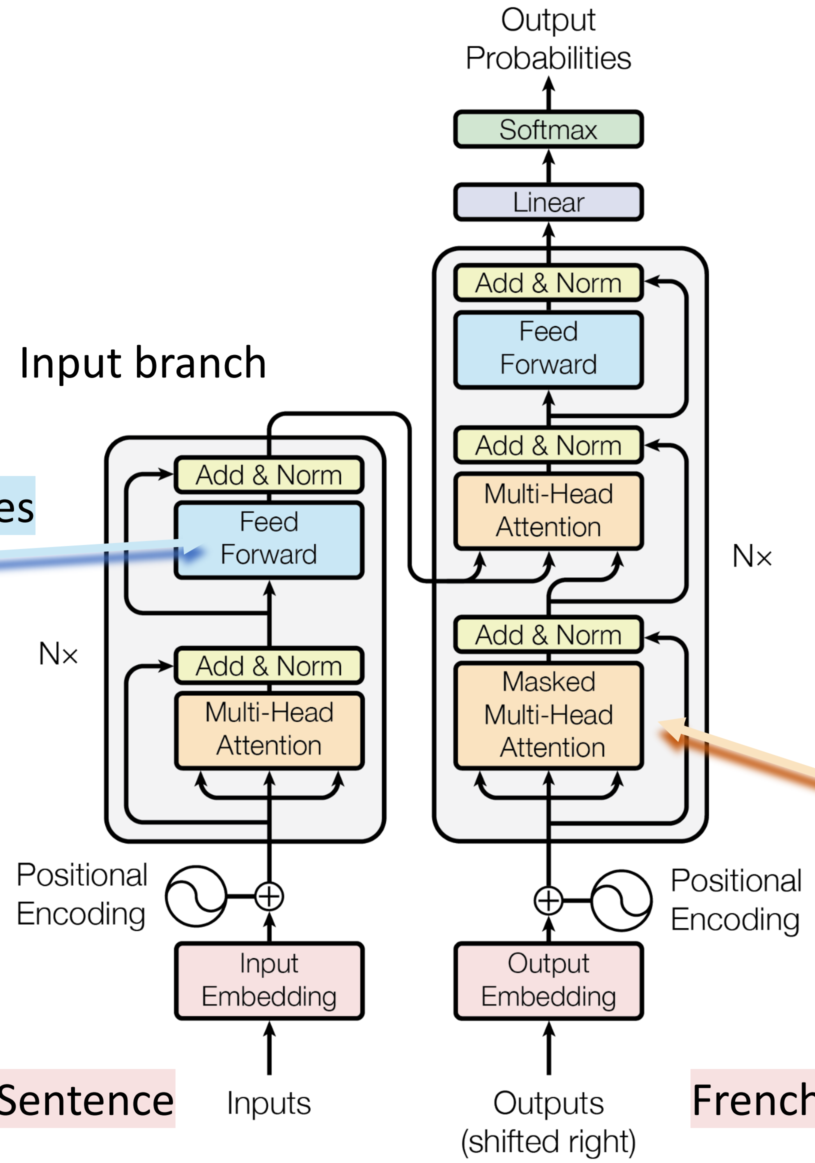


Figure 1: The Transformer - model architecture.

Transformer Architecture

Feed forward network processes every English word



Masking: Matrix multiply:
e.g. 2000 French words by 2000 French words but masking words that come afterwards with zero

English Sentence

French words, coming in

Figure 1: The Transformer - model architecture.

Transformer Architecture

Input branch

Output branch

Prediction

Feed forward network processes every English word

Matrix multiply:
e.g. processed English words by processed French words

Masking: Matrix multiply:
e.g. 2000 French words by 2000 French words but masking words that come afterwards with zero

English Sentence

French words, coming in

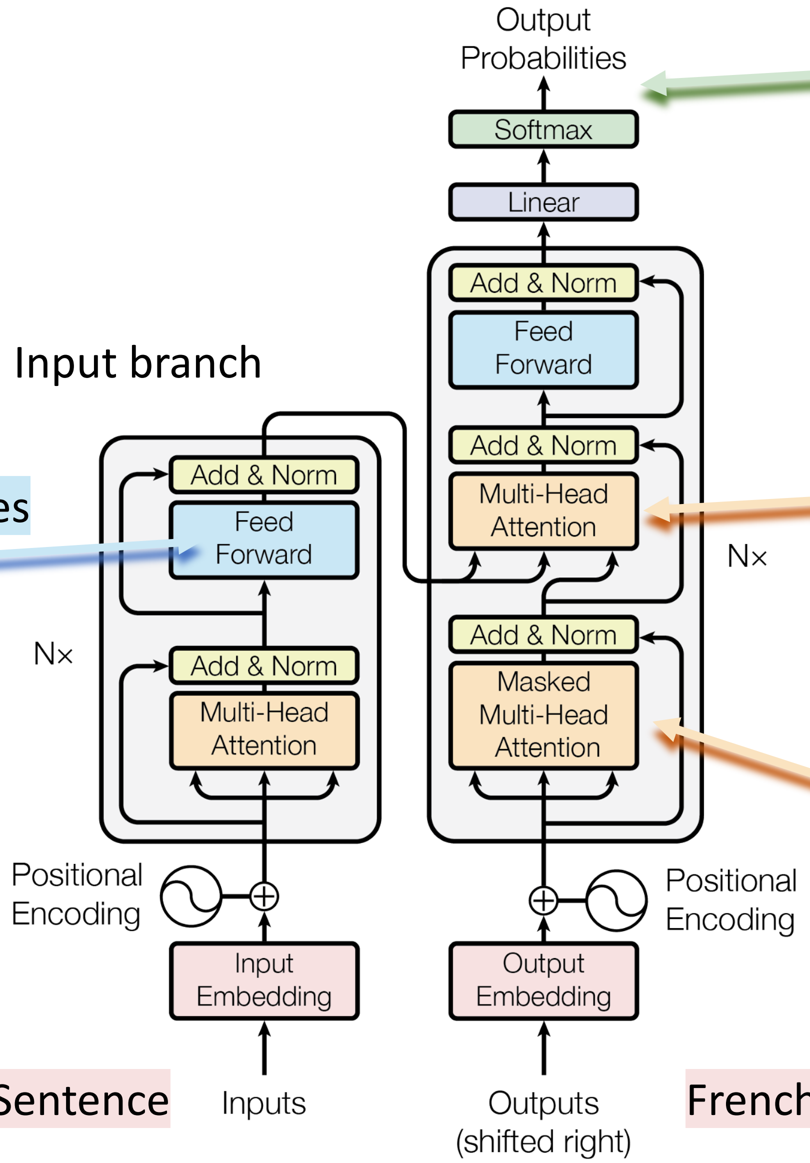


Figure 1: The Transformer - model architecture.

Masking Attention

Attention(Q,K,V) = softmax(Q K^T) V

Q = query vector = current English (or French) word

K key and V value = memory of words seen before

Goal: Find key(s) most similar to query and retrieve value(s) that correspond to this/these key(s)

Softmax = $\Sigma_i\, e^{qki} / {}_{\Sigma_j\, e^{q\,kj}}\;\; v_i$ produces probability distribution over keys with peaks for keys similar to query

Masking Attention

Attention(Q,K,V) = softmax($Q K^T$) V

Acts as a weight mask over V

Q = query vector = current English (or French) word

K key and V value = memory of words seen before

Goal: Find key(s) most similar to query and retrieve value(s) that correspond to this/these key(s)

Softmax = $\Sigma_i e^{qki} / _{\Sigma_j e^{q\,kj}}\ v_i$ produces probability distribution over keys with peaks for keys similar to query

Masking Attention

Very fast:
2 matrix multiplications & 1 softmax operation

Attention(Q,K,V) = softmax(Q K^T) V

Acts as a weight mask over V

Q = query vector = current English (or French) word

K key and V value = memory of words seen before

Goal: Find key(s) most similar to query and retrieve value(s) that correspond to this/these key(s)

Softmax = $\Sigma_i \, e^{qki} / {}_{\Sigma_j \, e^{q\,kj}} \; v_i$ produces probability distribution over keys with peaks for keys similar to query

Masking Attention

Very fast:
2 matrix multiplications & 1 softmax operation

Attention(Q,K,V) = softmax(Q K^T/ sqrt(d_k)) V

Acts as a weight mask over V

Technical detail:
sqrt(d_k) normalization needed for training

Q = query vector = current English (or French) word

K key and V value = memory of words seen before

Goal: Find key(s) most similar to query and retrieve value(s) that correspond to this/these key(s)

Softmax = $\Sigma_i \, e^{qki} / _{\Sigma_j \, e^{q\,kj}} \; v_i$ produces probability distribution over keys with peaks for keys similar to query

Why Multi-Head Attention?

Input branch

Output branch

Prediction

Feed forward network processes every English word

Matrix multiply:
e.g. processed English words by processed French words

Masking: Matrix multiply:
e.g. 2000 French words by 2000 French words but masking words that come afterwards with zero

English Sentence

French words, coming in

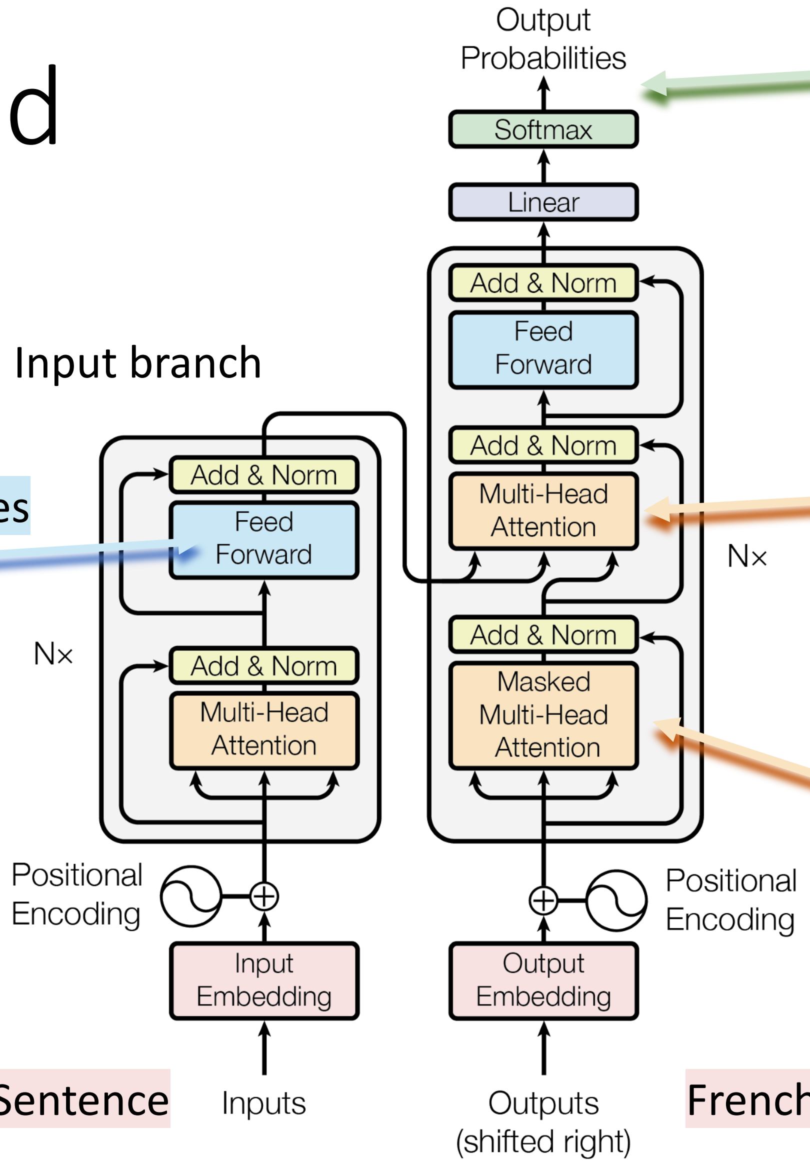


Figure 1: The Transformer - model architecture.

Why Multi-Head Attention?

- Multiple attention layers (heads) in paraellel
- Each head uses different linear transformation
- Different heads can learn different relationships

Attention Visualizations

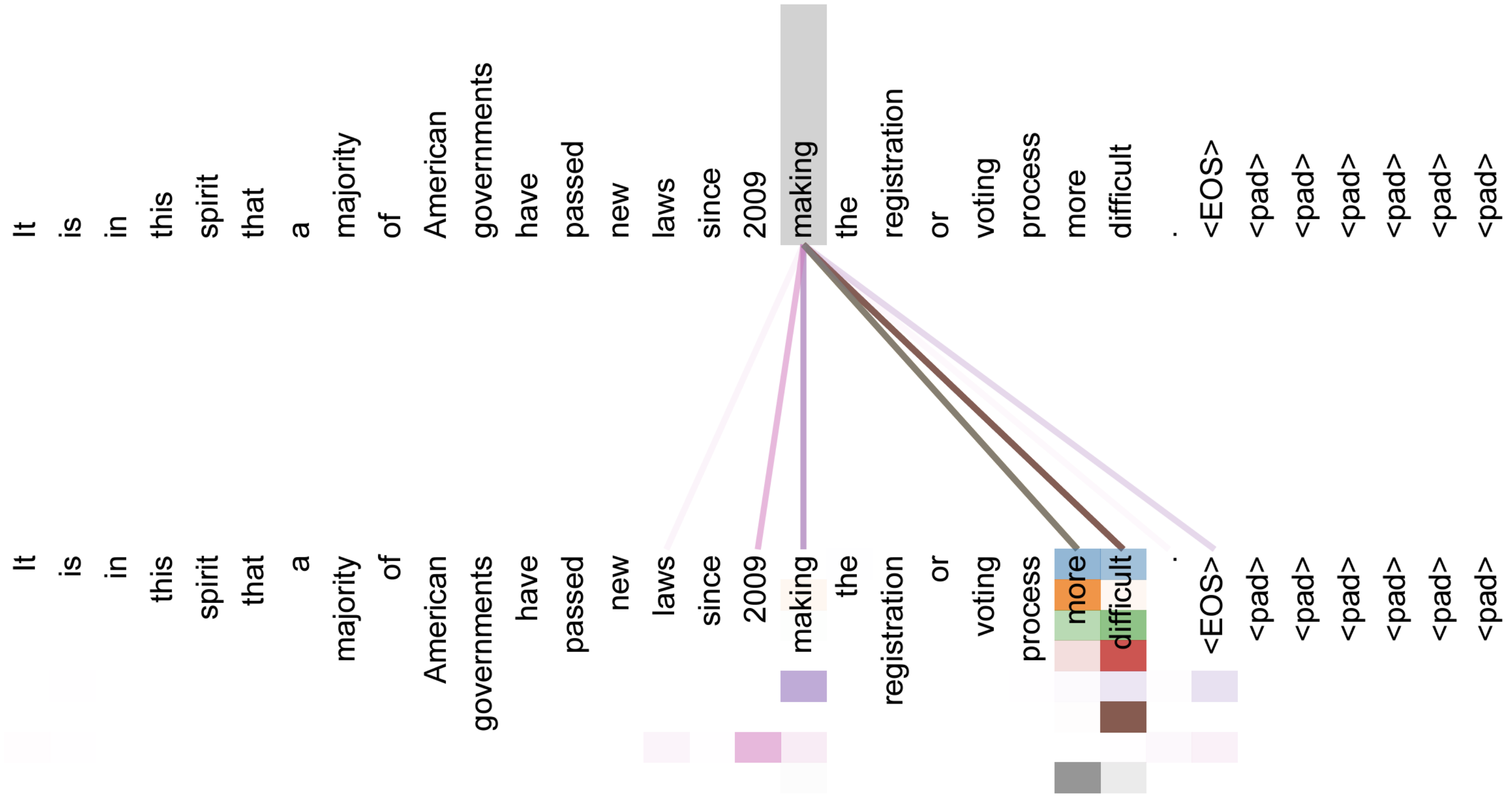


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb ‘making’, completing the phrase ‘making...more difficult’. Attentions here shown only for the word ‘making’. Different colors represent different heads. Best viewed in color.

Training a Transformer

- ADAM optimizer
- Dropout during training at every layer
- Label smoothing
- Auto-regressive decoding with beam-search
- Checkpoint-averaging

- Library available: https://github.com/tensorflow/tensor2tensor

Transformer Architecture Complexity

- n= number of words in sequence
- d= network depth

Number of operations: $n^2 d$

Number of activations: $n^2 + n\, d$

Much better than CNNs or RNNs with number of operations $n\, d^2$

Transformer Architecture Complexity

- n= number of words in sequence (<70 words per sentence)
- d= network depth (maybe 1000)

Every word attends to every word

Number of operations: $n^2 d$ e.g., 70x70x1000=4.9 mill

Number of activations: $n^2 + n\,d$

Much better than CNNs or RNNs with number of operations $n\,d^2$

e.g., 70x1000x1000=70 mill

Vaswani et al., 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Dosovitskiy et al., 2020

Vision Transformer ViT

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

ViT Architecture

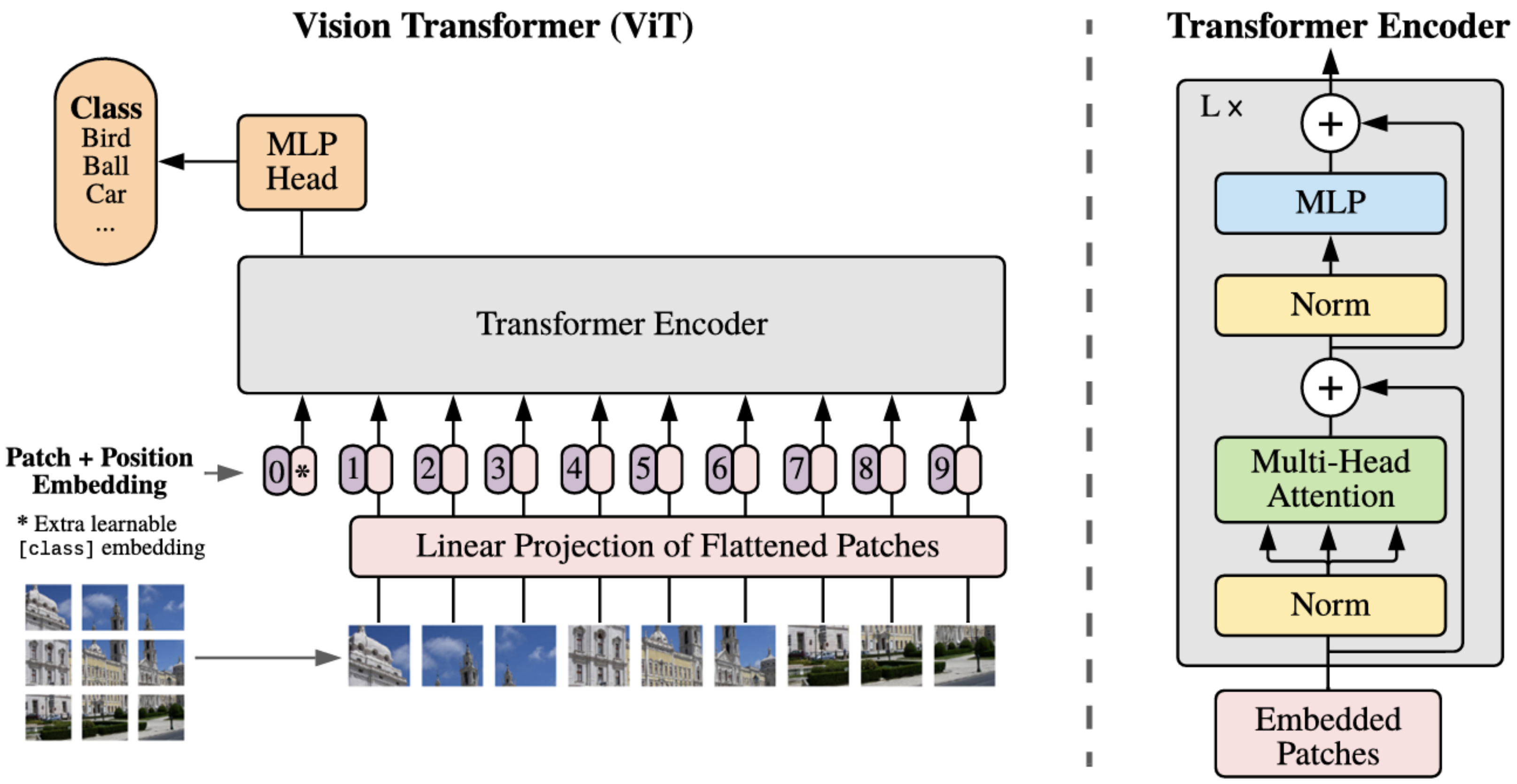
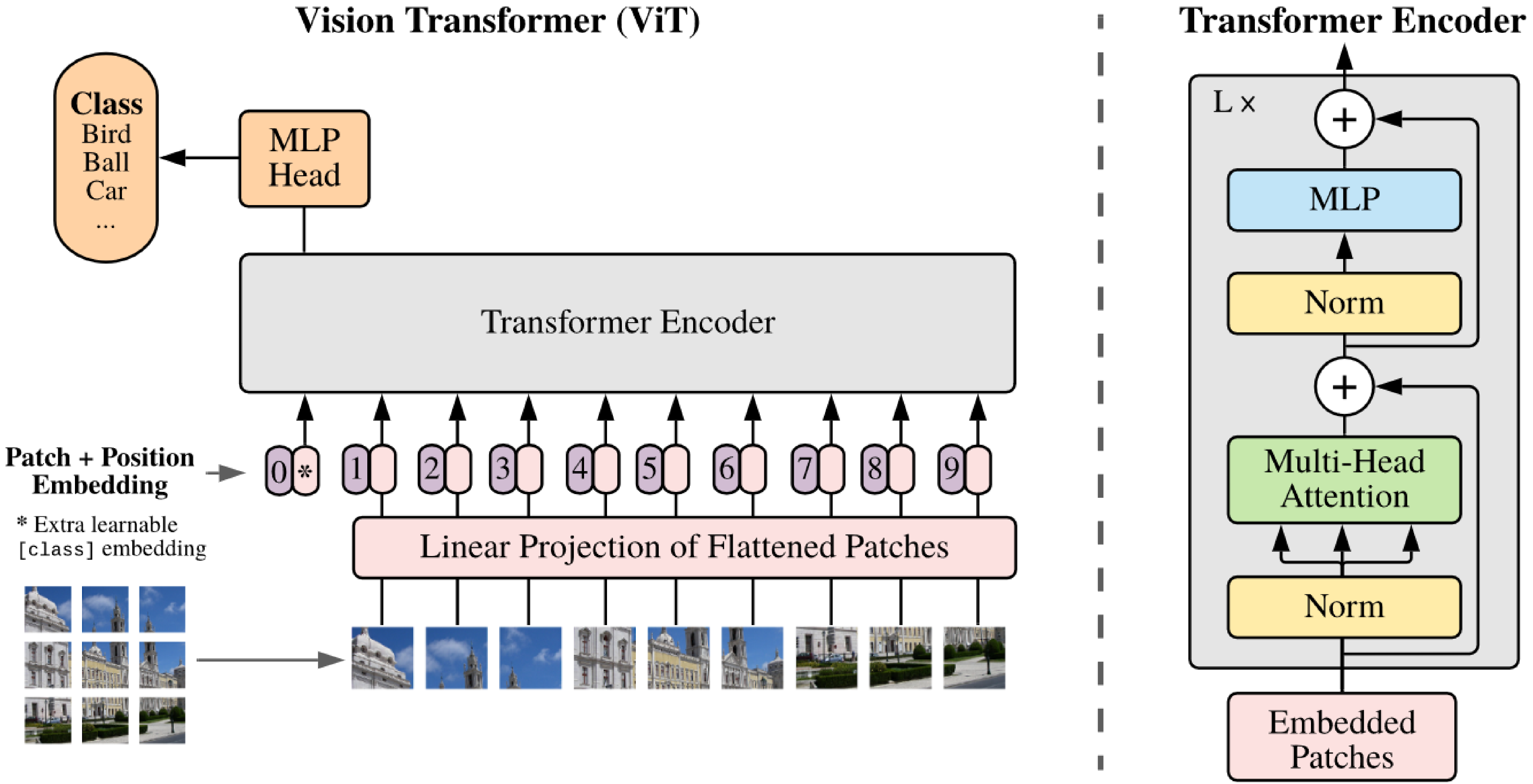
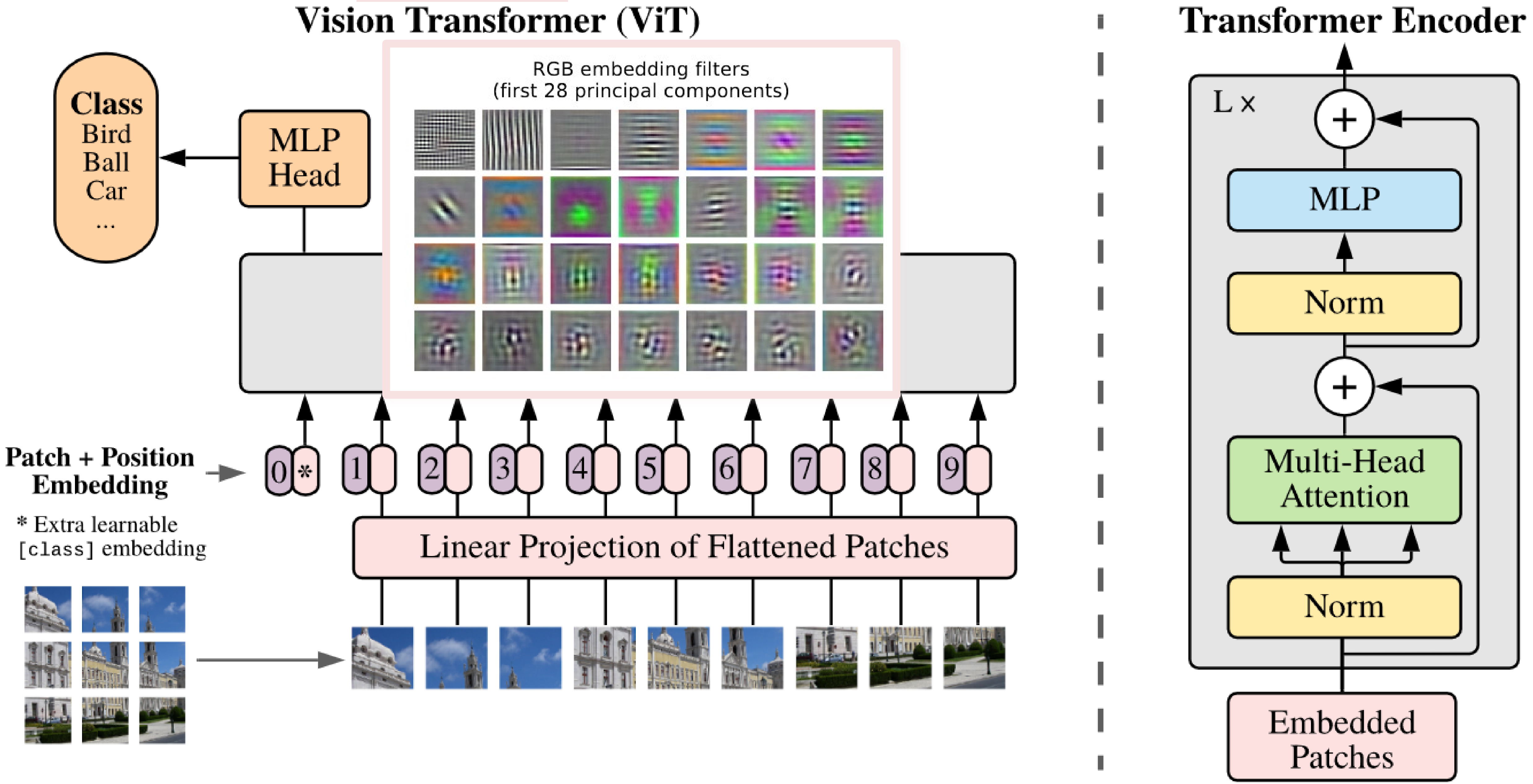
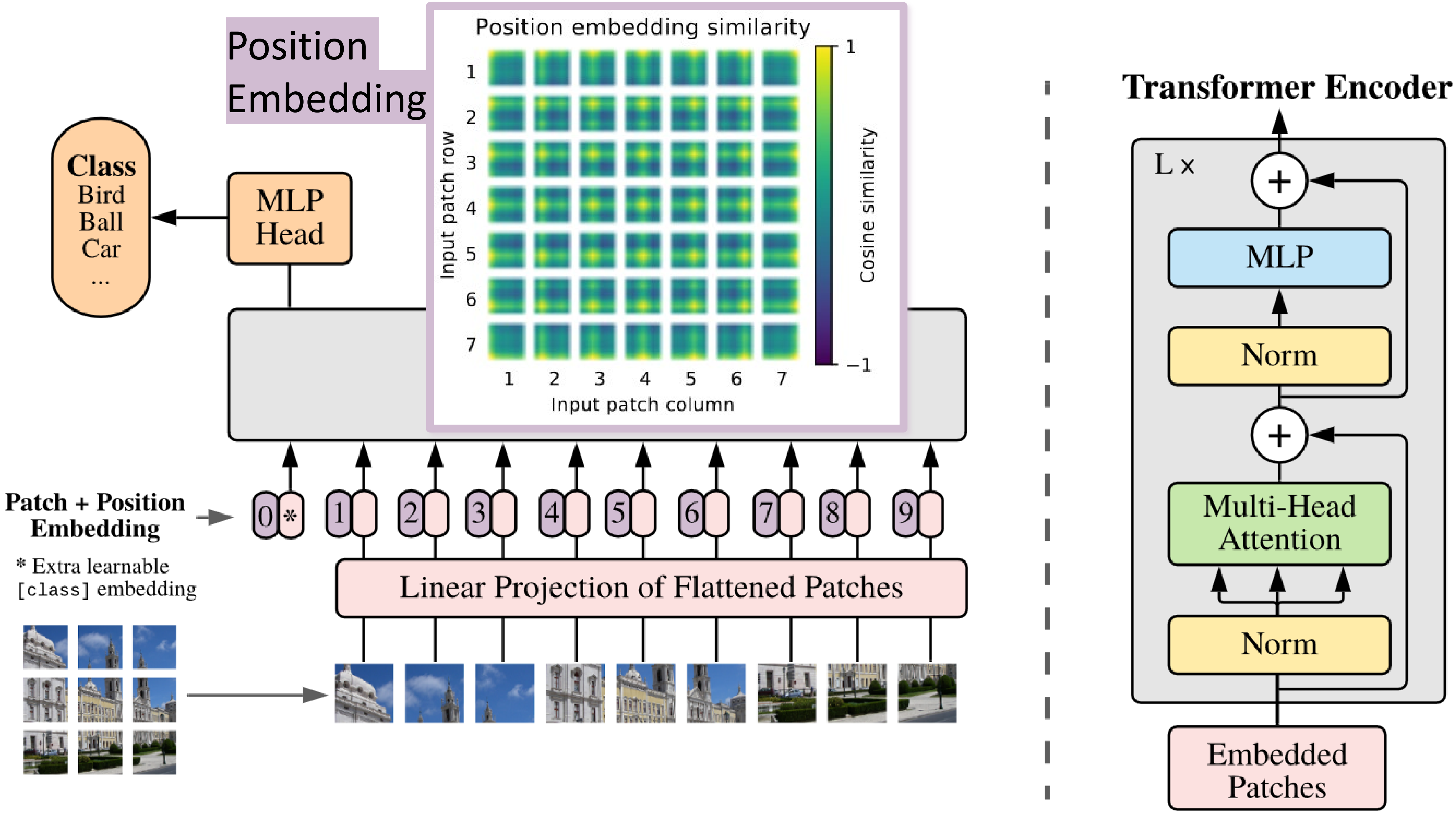


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).


Vision Transformer (ViT)
Transformer Encoder
Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
L ×
MLP
Norm
Multi-Head
Attention
Norm
Embedded
Patches

Embedding


Vision Transformer (ViT)
Transformer Encoder
Class
Bird
Ball
Car
...
MLP Head
RGB embedding filters
(first 28 principal components)
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches


Position Embedding
Position embedding similarity
Input patch row
Input patch column
Cosine similarity
1
−1
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches
Patch + Position Embedding
* Extra learnable
[class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches

Vision Transformer

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

Vision Transformer Results

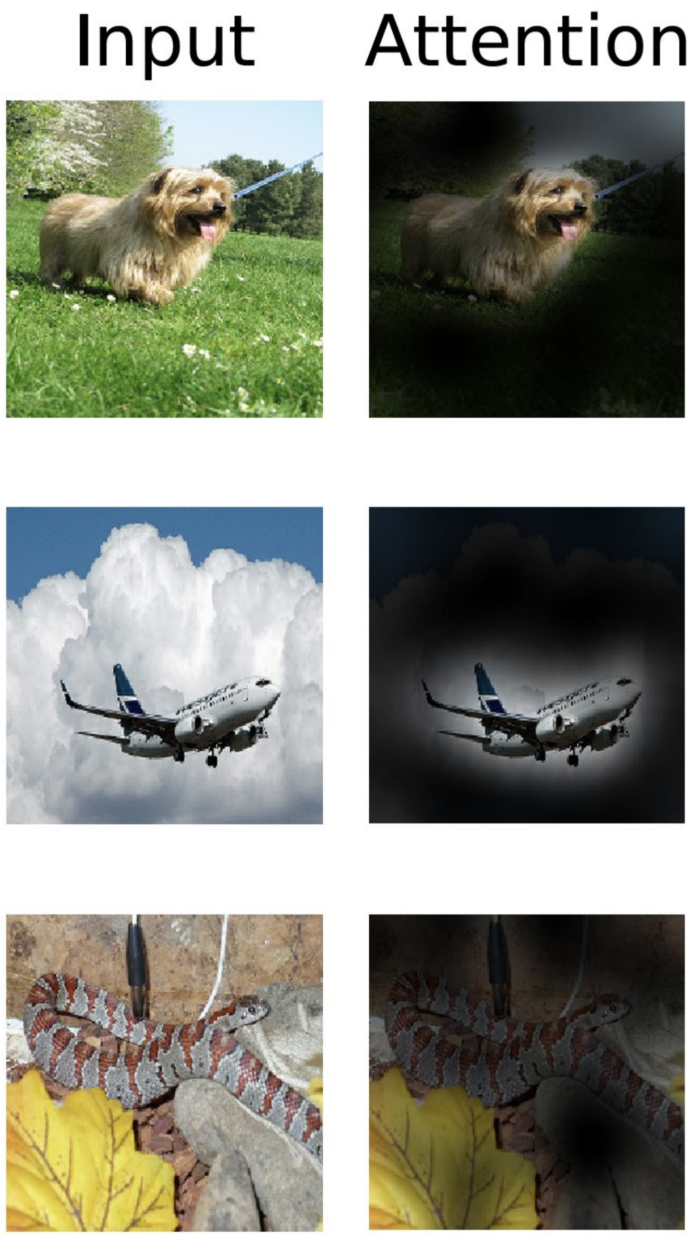
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	–
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	–
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	–
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	–
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

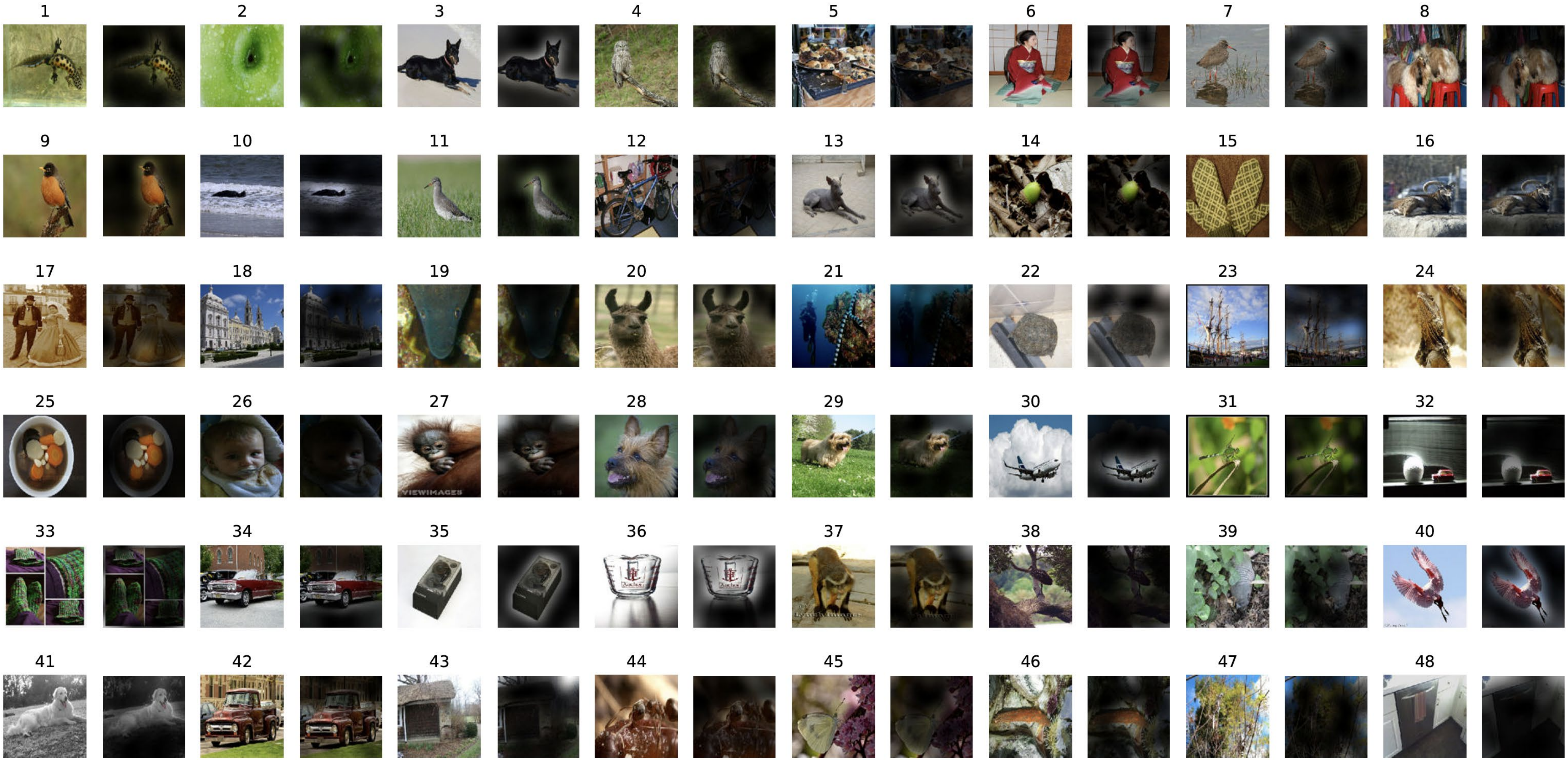
Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

Vision Transformer Results

	Caltech101	CIFAR-100	DTD	Flowers102	Pets	Sun397	SVHN
ViT-H/14 (JFT)	95.3	85.5	75.2	99.7	97.2	65.0	88.9
ViT-L/16 (JFT)	95.4	81.9	74.3	99.7	96.7	63.5	87.4
ViT-L/16 (I21k)	90.8	84.1	74.1	99.3	92.7	61.0	80.9

Vision Transformer




1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16
17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32
33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48

Dosovitskiy
et al.,
2020

Vision
Transformer
ViT

Published as a conference paper at ICLR 2021

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

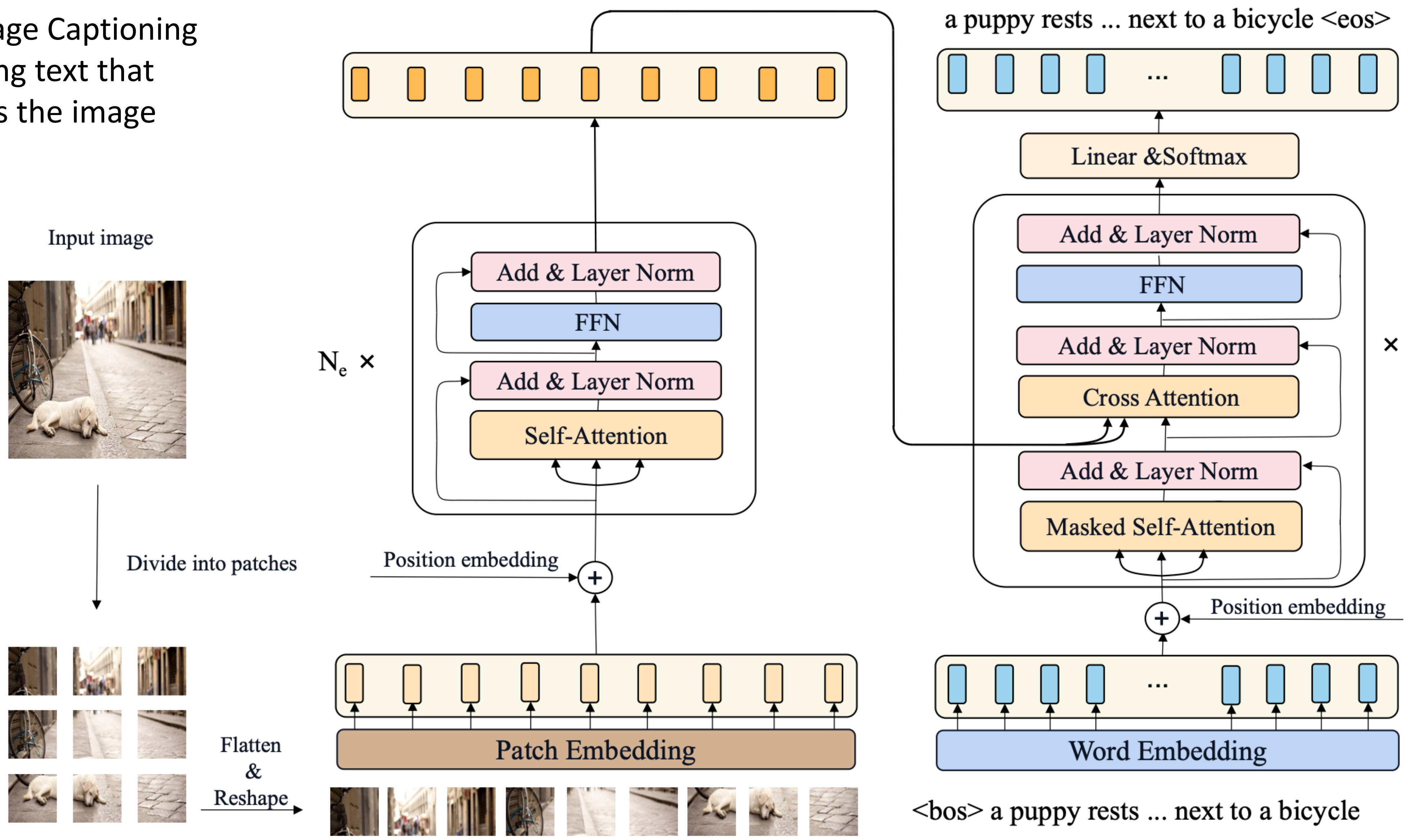
Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*],
Xiaohua Zhai[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,**
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
`{adosovitskiy, neilhoulsby}@google.com`

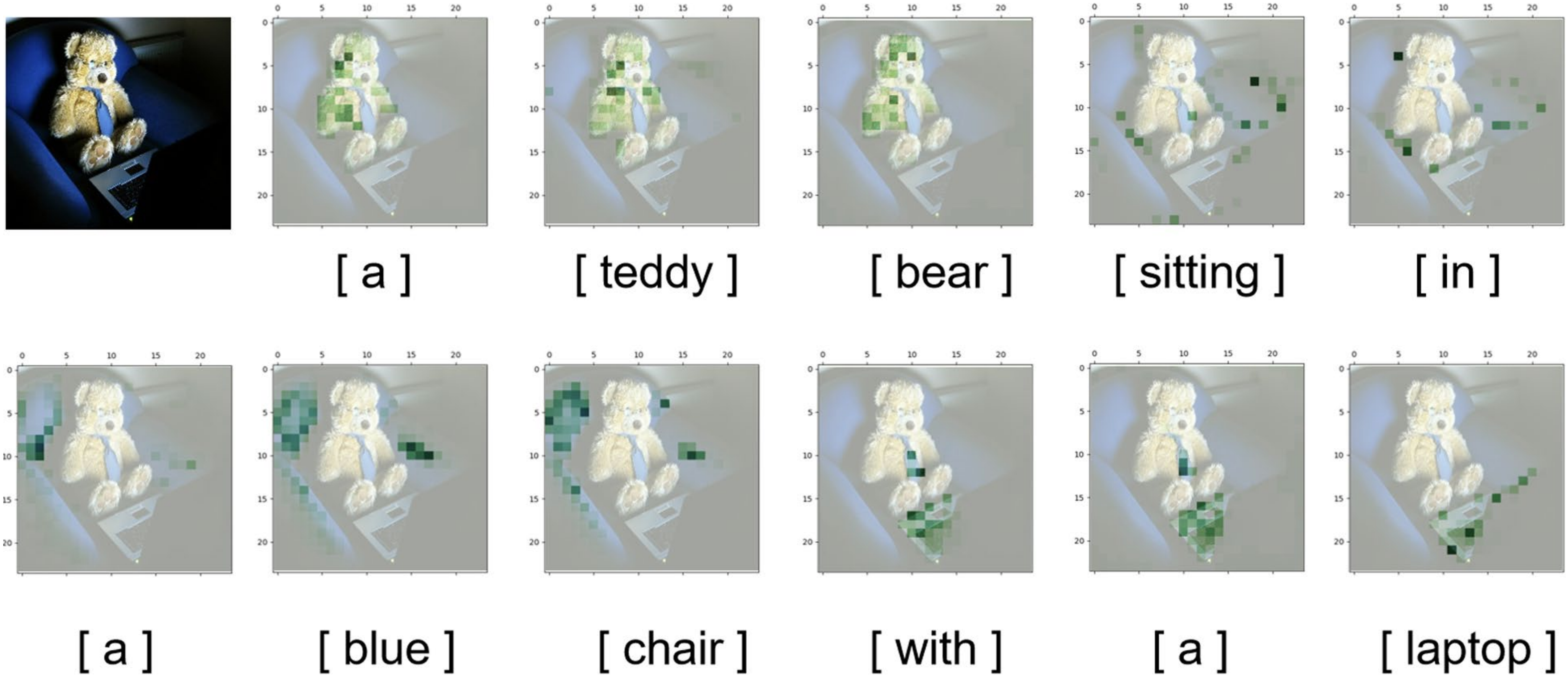
Abstract

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

Task: Image Captioning
= Creating text that describes the image

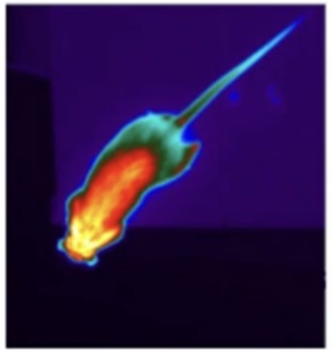
[Liu et al., 2021]

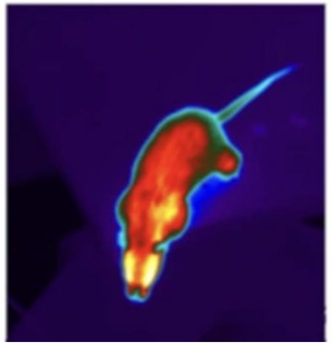


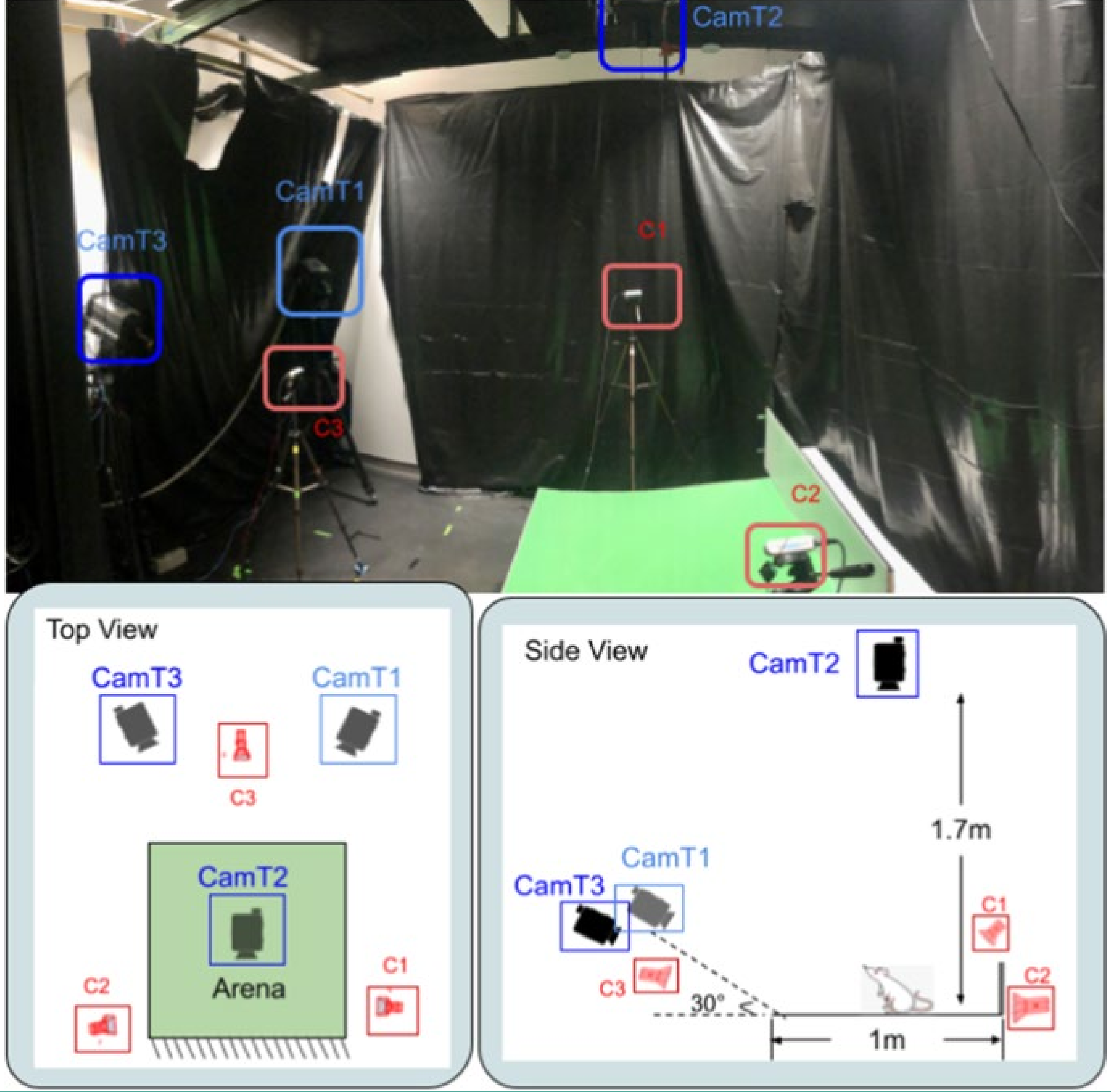

[a]
[teddy]
[bear]
[sitting]
[in]
[a]
[blue]
[chair]
[with]
[a]
[laptop]

Animal Pose Tracking: 3D Multimodal Dataset and Token-based Pose Optimization

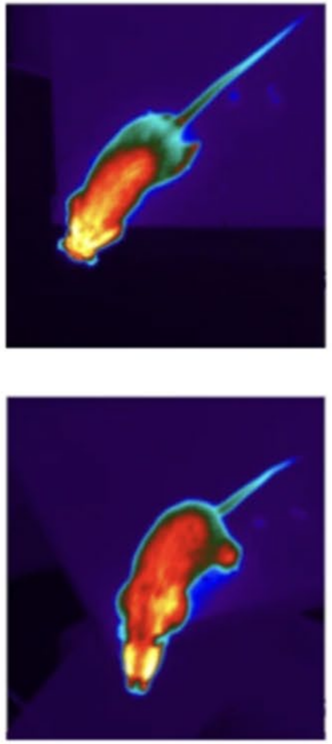
[Patel et al., 2022](#)

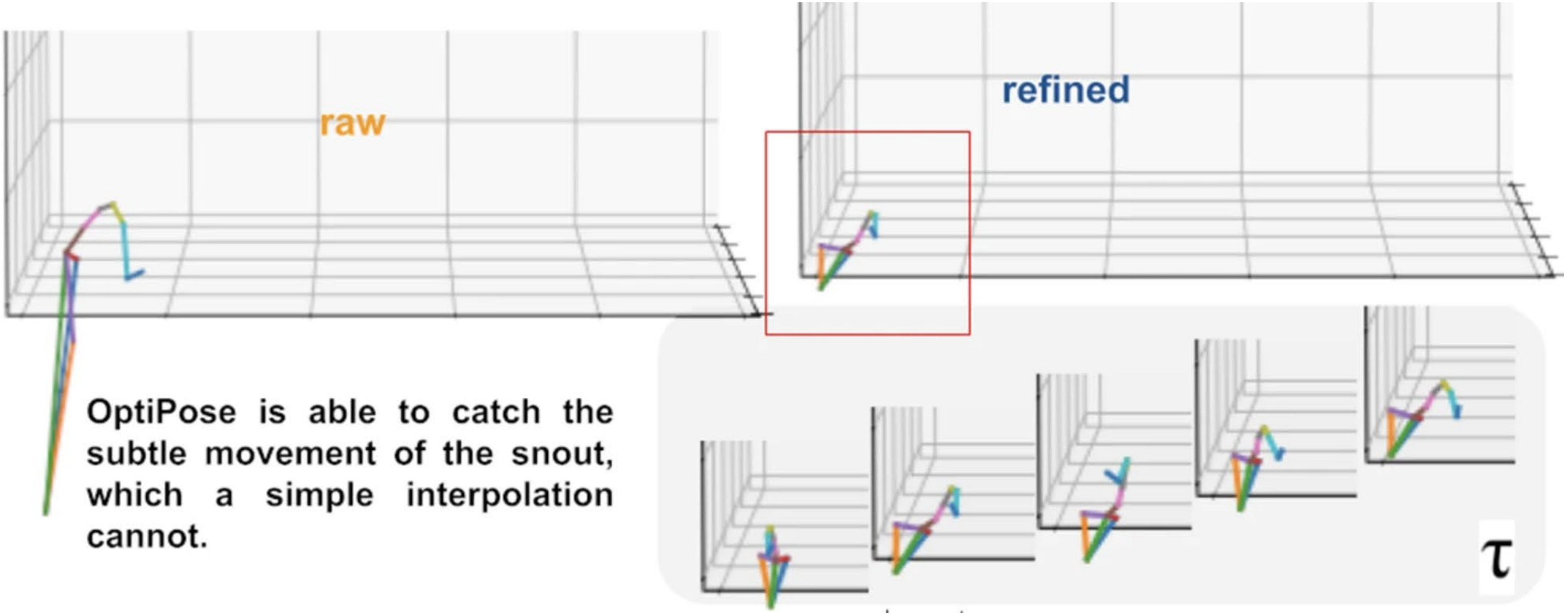






Animal Pose Tracking: 3D Multimodal Dataset and Token-based Pose Optimization



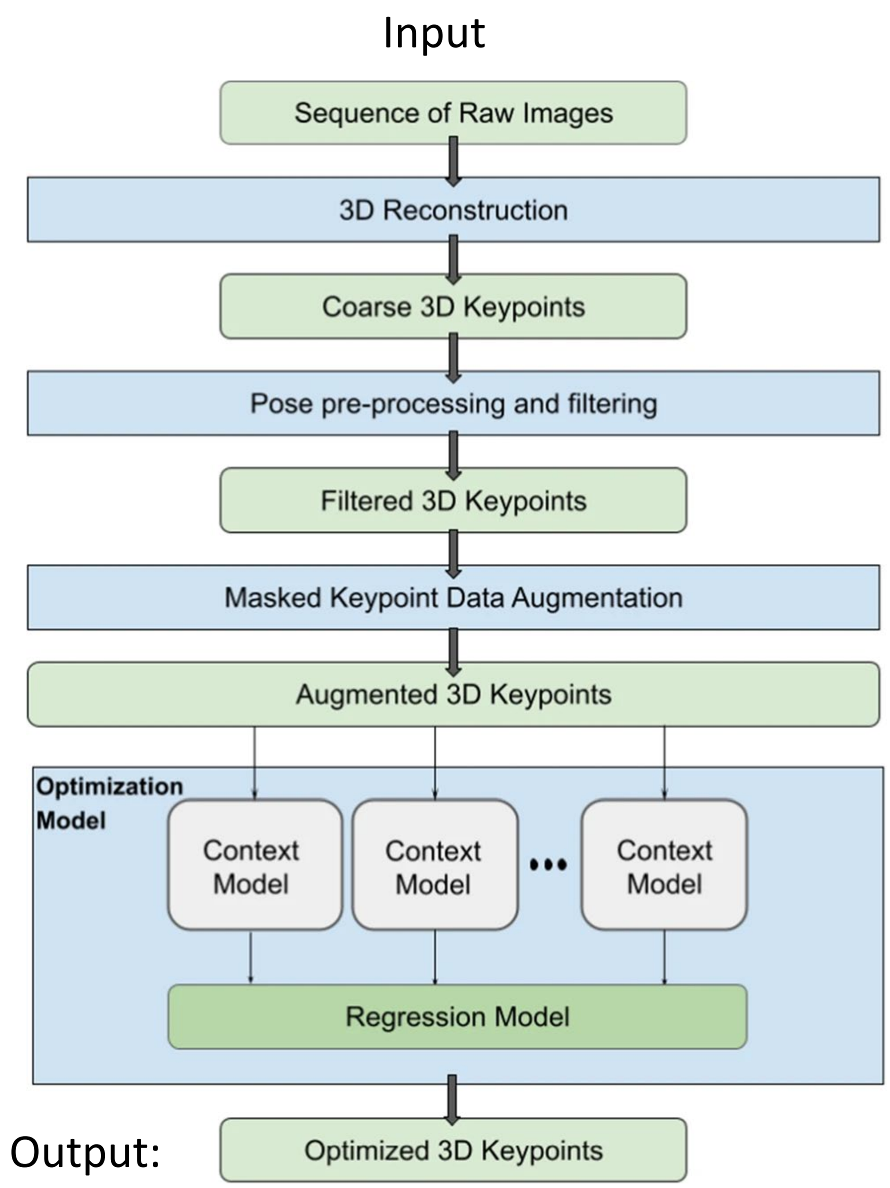


OptiPose is able to catch the subtle movement of the snout, which a simple interpolation cannot.

Animal Pose Tracking: 3D Multimodal Dataset and Token-based Pose Optimization

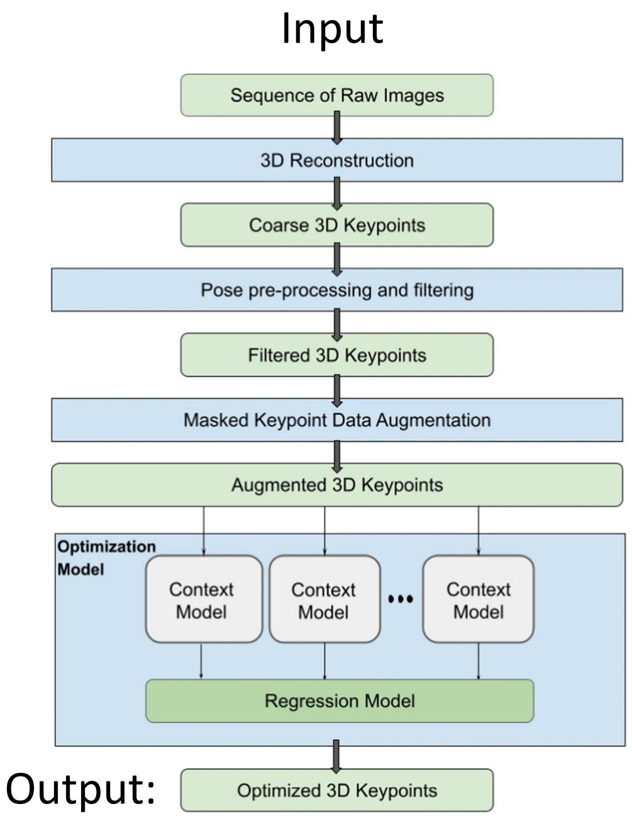
[Patel et al., 2022]()

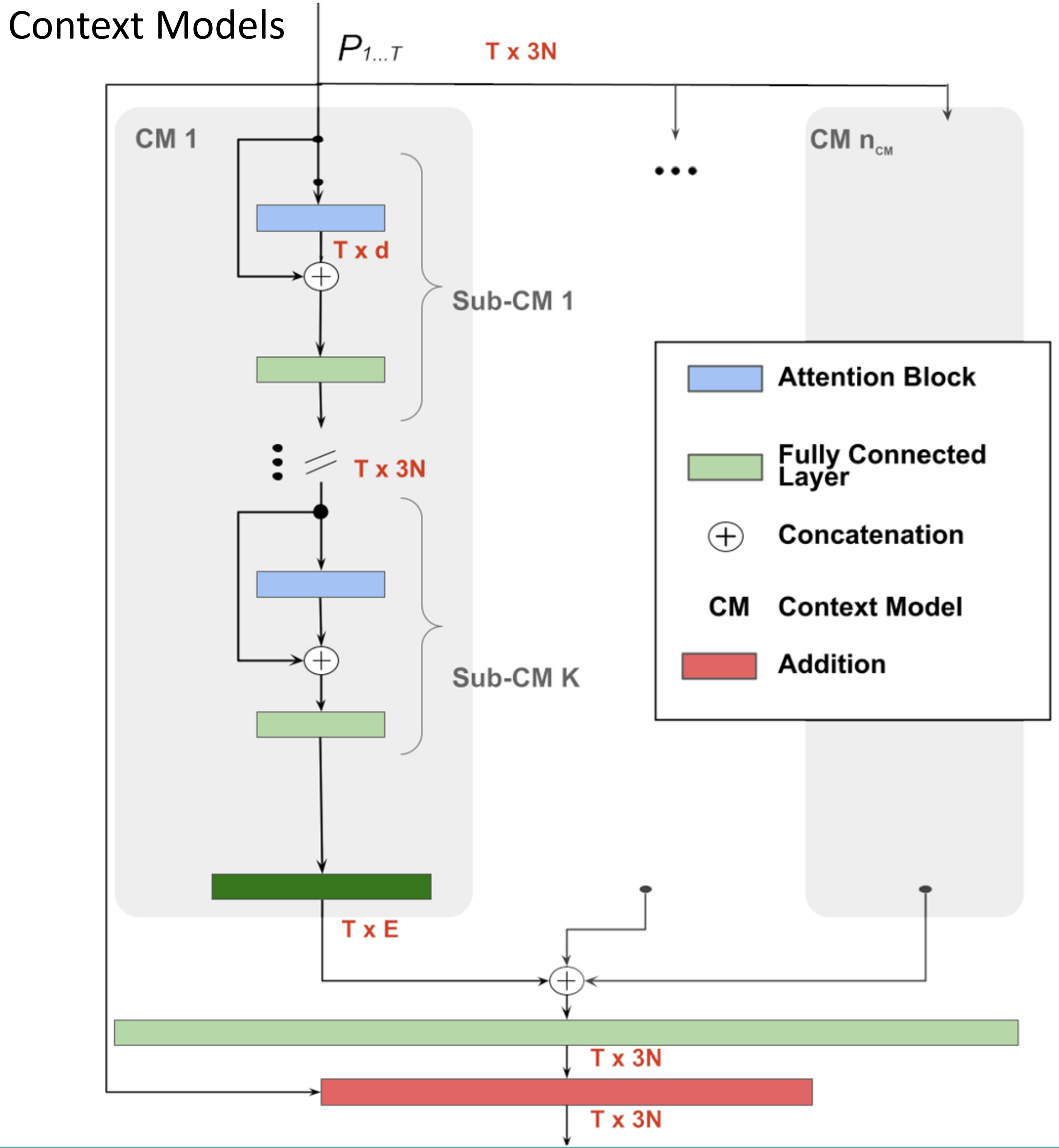
Using a Transformer:



Animal Pose Tracking: 3D Multimodal Dataset and Token-based Pose Optimization

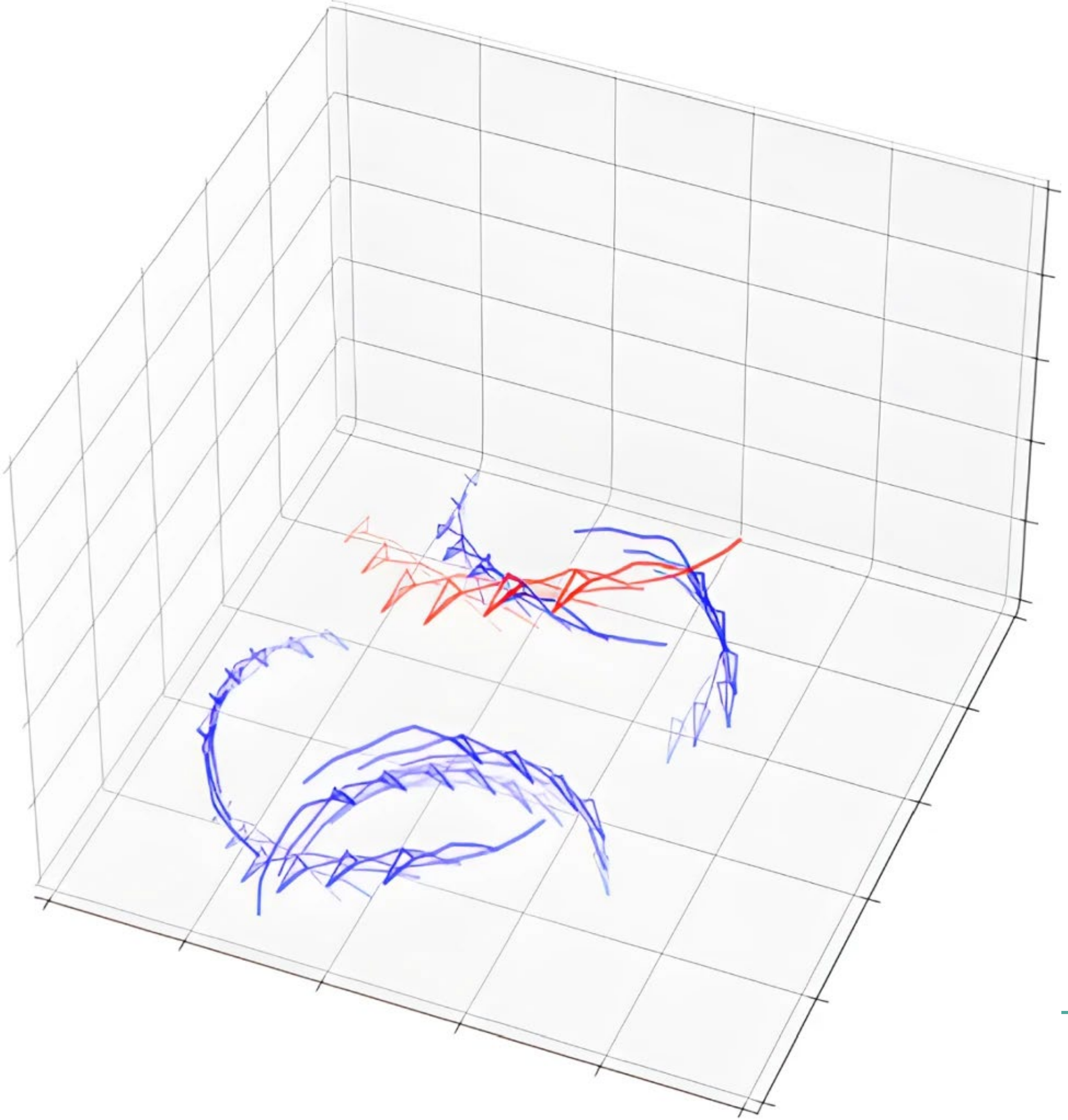
[Patel et al., 2022]





Animal Pose Tracking: 3D Multimodal Dataset and Token-based Pose Optimization

Patel et al., 2022



A ConvNet for the 2020s

Zhuang Liu[1,2*] Hanzi Mao[1] Chao-Yuan Wu[1] Christoph Feichtenhofer[1] Trevor Darrell[2] Saining Xie[1†]
[1]Facebook AI Research (FAIR) [2]UC Berkeley

Code: https://github.com/facebookresearch/ConvNeXt

Abstract

The "Roaring 20s" of visual recognition began with the introduction of Vision Transformers (ViTs), which quickly superseded ConvNets as the state-of-the-art image classification model. A vanilla ViT, on the other hand, faces difficulties when applied to general computer vision tasks such as object detection and semantic segmentation. It is the hierarchical Transformers (e.g., Swin Transformers) that reintroduced several ConvNet priors, making Transformers practically viable as a generic vision backbone and demonstrating remarkable performance on a wide variety of vision tasks. However, the effectiveness of such hybrid approaches is still largely credited to the intrinsic superiority of Transformers, rather than the inherent inductive biases of convolutions. In this work, we reexamine the design spaces and test the limits of what a pure ConvNet can achieve. We gradually "modernize" a standard ResNet toward the design of a vision Transformer, and discover several key components that contribute to the performance difference along the way. The outcome of this exploration is a family of pure ConvNet models dubbed ConvNeXt. Constructed entirely from standard ConvNet modules, ConvNeXts compete favorably with Transformers in terms of accuracy and scalability, achieving 87.8% ImageNet top-1 accuracy and outperforming Swin Transformers on COCO detection and ADE20K segmentation, while maintaining the simplicity and efficiency of standard ConvNets.

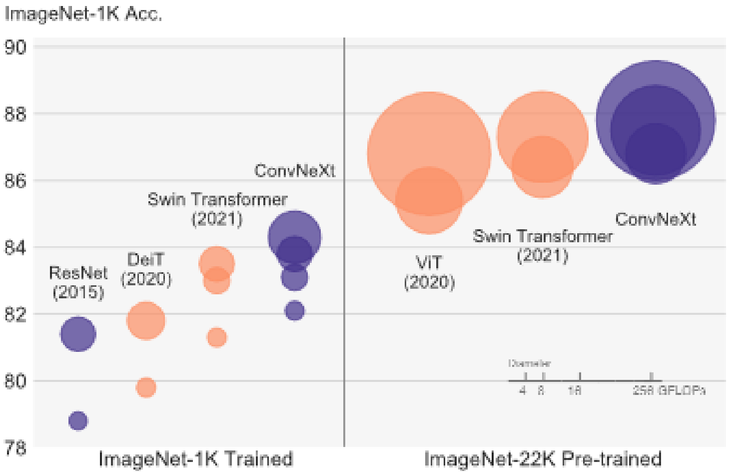


Figure 1. **ImageNet-1K classification** results for • ConvNets and ○ vision Transformers. Each bubble's area is proportional to FLOPs of a variant in a model family. ImageNet-1K/22K models here take $224^2/384^2$ images respectively. ResNet and ViT results were obtained with improved training procedures over the original papers. We demonstrate that a standard ConvNet model can achieve the same level of scalability as hierarchical vision Transformers while being much simpler in design.

visual feature learning. The introduction of AlexNet [40] precipitated the "ImageNet moment" [59], ushering in a new era of computer vision. The field has since evolved at a rapid speed. Representative ConvNets like VGGNet [64], Inceptions [68], ResNe(X)t [28, 87], DenseNet [36], MobileNet [34], EfficientNet [71] and RegNet [54] focused on different aspects of accuracy, efficiency and scalability, and popularized many useful design principles.

A ConvNet for the 2020s

Zhuang Liu[1,2*] Hanzi Mao[1] Chao-Yuan Wu[1] Christoph Feichtenhofer[1] Trevor Darrell[2] Saining Xie[1†]

[1]Facebook AI Research (FAIR) [2]UC Berkeley

Code: https://github.com/facebookresearch/ConvNeXt

Abstract

The "Roaring 20s" of visual recognition began with the introduction of Vision Transformers (ViTs), which quickly superseded ConvNets as the state-of-the-art image classification model. A vanilla ViT, on the other hand, faces difficulties when applied to general computer vision tasks such as object detection and semantic segmentation. It is the hierarchical Transformers (e.g., Swin Transformers) that reintroduced several ConvNet priors, making Transformers practically viable as a generic vision backbone and demonstrating remarkable performance on a wide variety of vision tasks. However, the effectiveness of such hybrid approaches is still largely credited to the intrinsic superiority of Transformers, rather than the inherent inductive biases of convolutions. In this work, we reexamine the design spaces and test the limits of what a pure ConvNet can achieve. We gradually "modernize" a standard ResNet toward the design of a vision Transformer, and discover several key components that contribute to the performance difference along the way. The outcome of this exploration is a family of pure ConvNet models dubbed ConvNeXt. Constructed entirely from standard ConvNet modules, ConvNeXts compete favorably with Transformers in terms of accuracy and scalability, achieving 87.8% ImageNet top-1 accuracy and outperforming Swin Transformers on COCO detection and ADE20K segmentation, while maintaining the simplicity and efficiency of standard ConvNets.

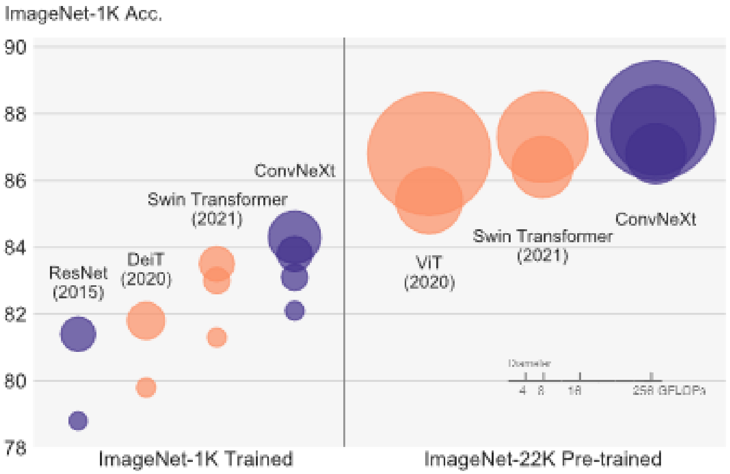


Figure 1. **ImageNet-1K classification** results for • ConvNets and ○ vision Transformers. Each bubble's area is proportional to FLOPs of a variant in a model family. ImageNet-1K/22K models here take $224^2/384^2$ images respectively. ResNet and ViT results were obtained with improved training procedures over the original papers. We demonstrate that a standard ConvNet model can achieve the same level of scalability as hierarchical vision Transformers while being much simpler in design.

visual feature learning. The introduction of AlexNet [40] precipitated the "ImageNet moment" [59], ushering in a new era of computer vision. The field has since evolved at a rapid speed. Representative ConvNets like VGGNet [64], Inceptions [68], ResNe(X)t [28, 87], DenseNet [36], MobileNet [34], EfficientNet [71] and RegNet [54] focused on different aspects of accuracy, efficiency and scalability, and popularized many useful design principles.

BOSTON UNIVERSITY

Learning Outcomes

Understand

- Concept of attention
- Transformers for NLP ("Attention is All you Need")
- Vision transformers for object recognition ("An image is worth 16x16 Words")
- Vision transformers for image captioning
- Vision transformers for 3D pose optimization and tracking