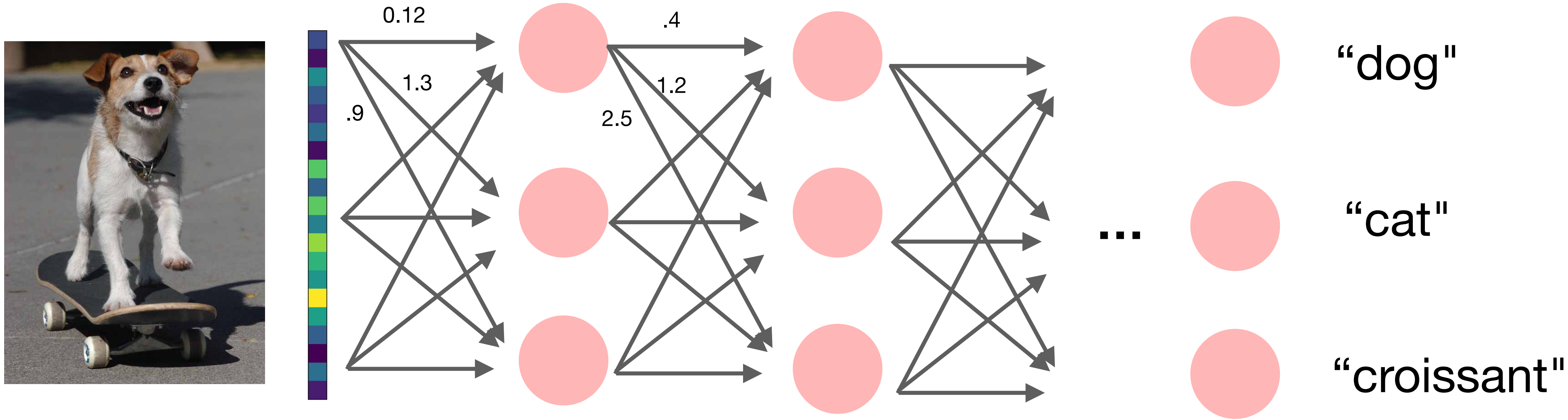
Attention models & interpretability

Adrian Valente
30-11-2023

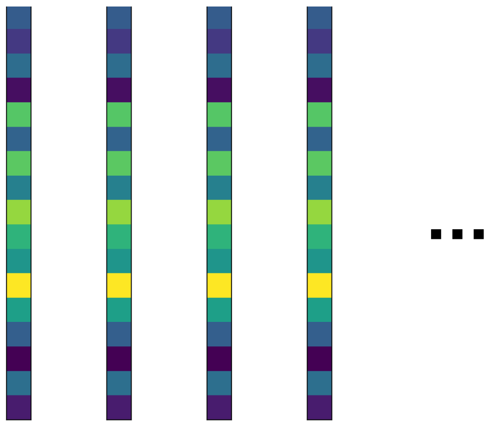
Deep learning primer



Goal: learn mapping vector -> vector

Problem: variable-length input

- Text
- Sound
- Video
- Time series
- ….

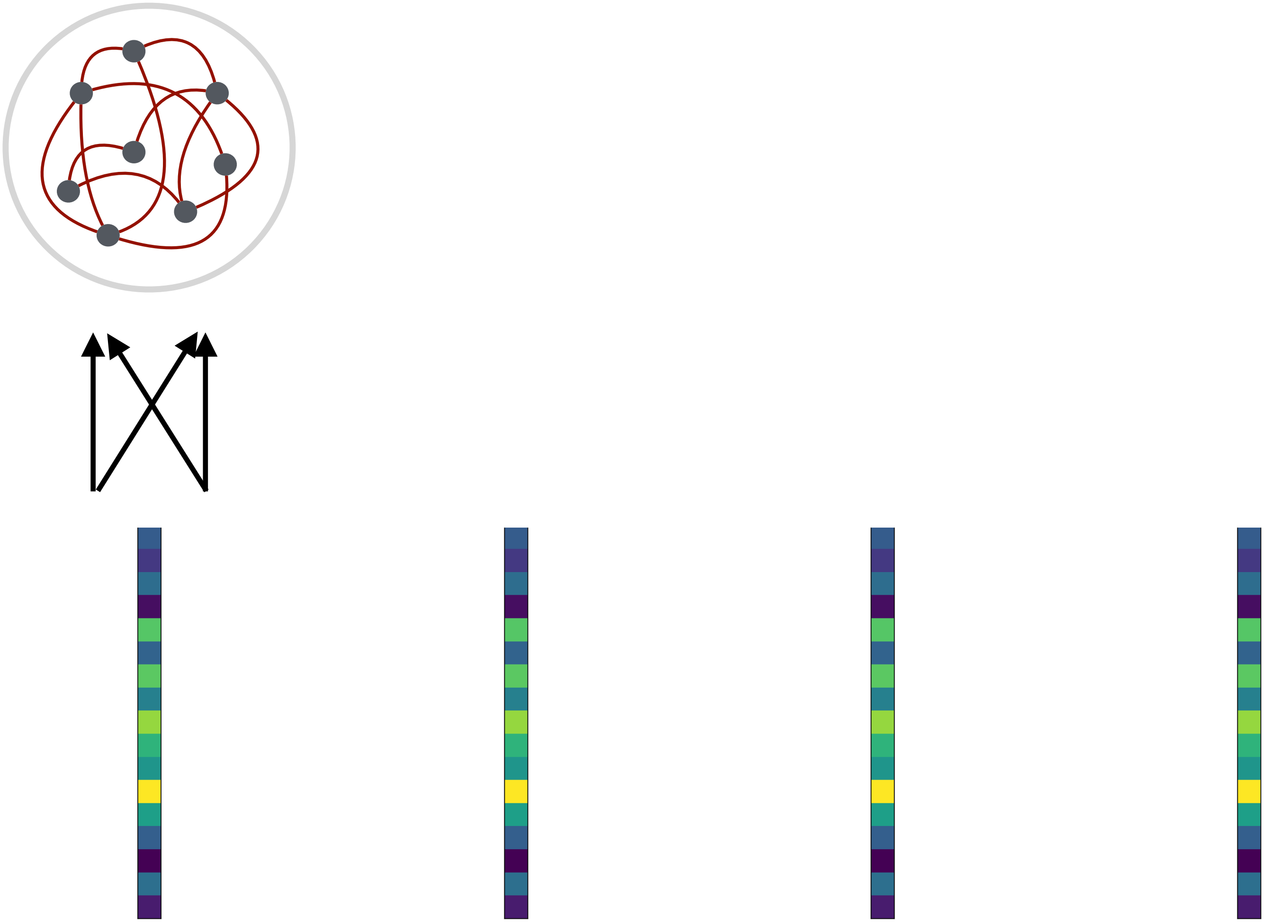


$(\mathbf{x}_1, \mathbf{x_2}, \cdots)$

RNNs

Example: sentiment analysis

Mapping sequence -> vector

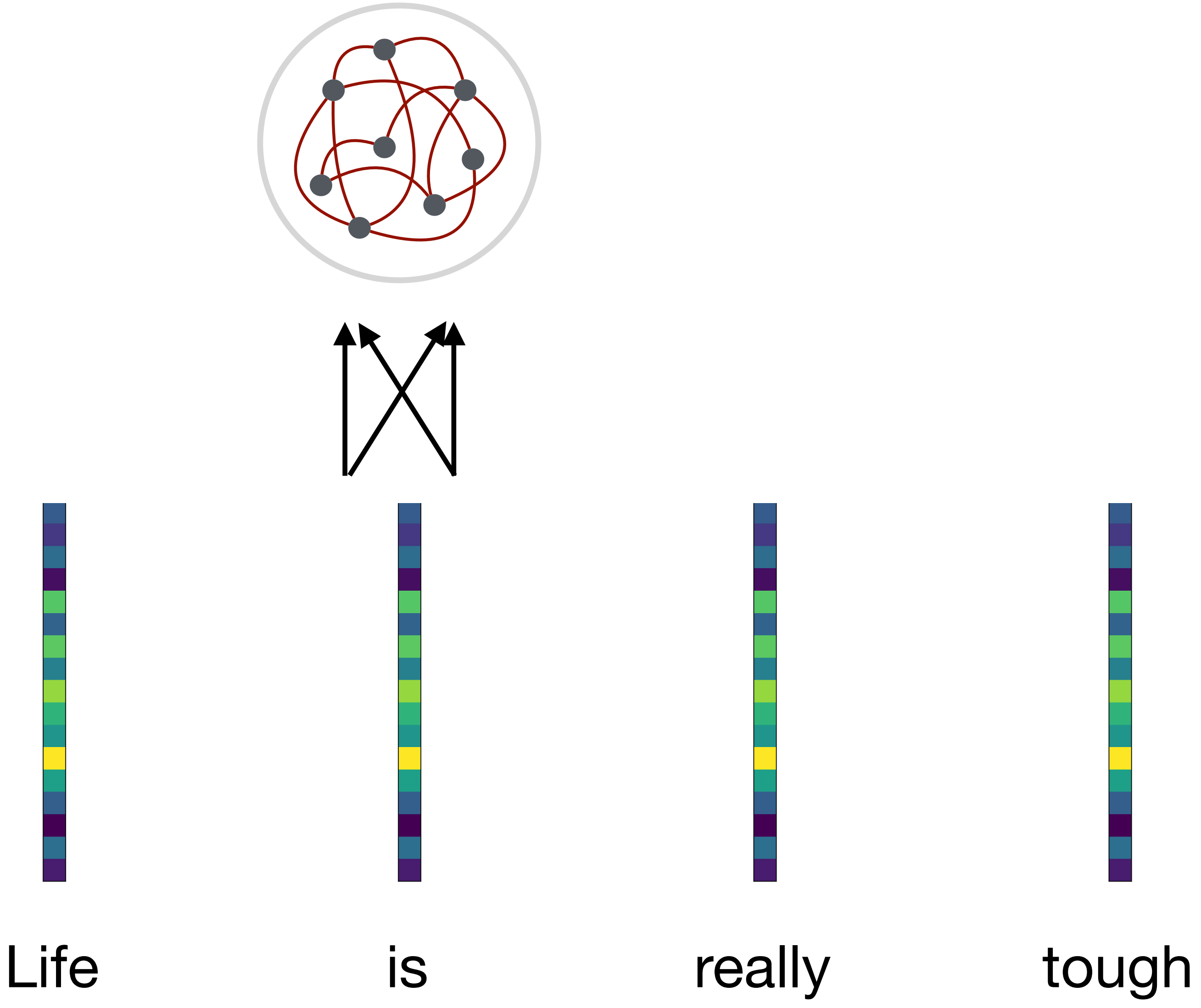


Life is really tough

RNNs

Example: sentiment analysis

Mapping sequence -> vector

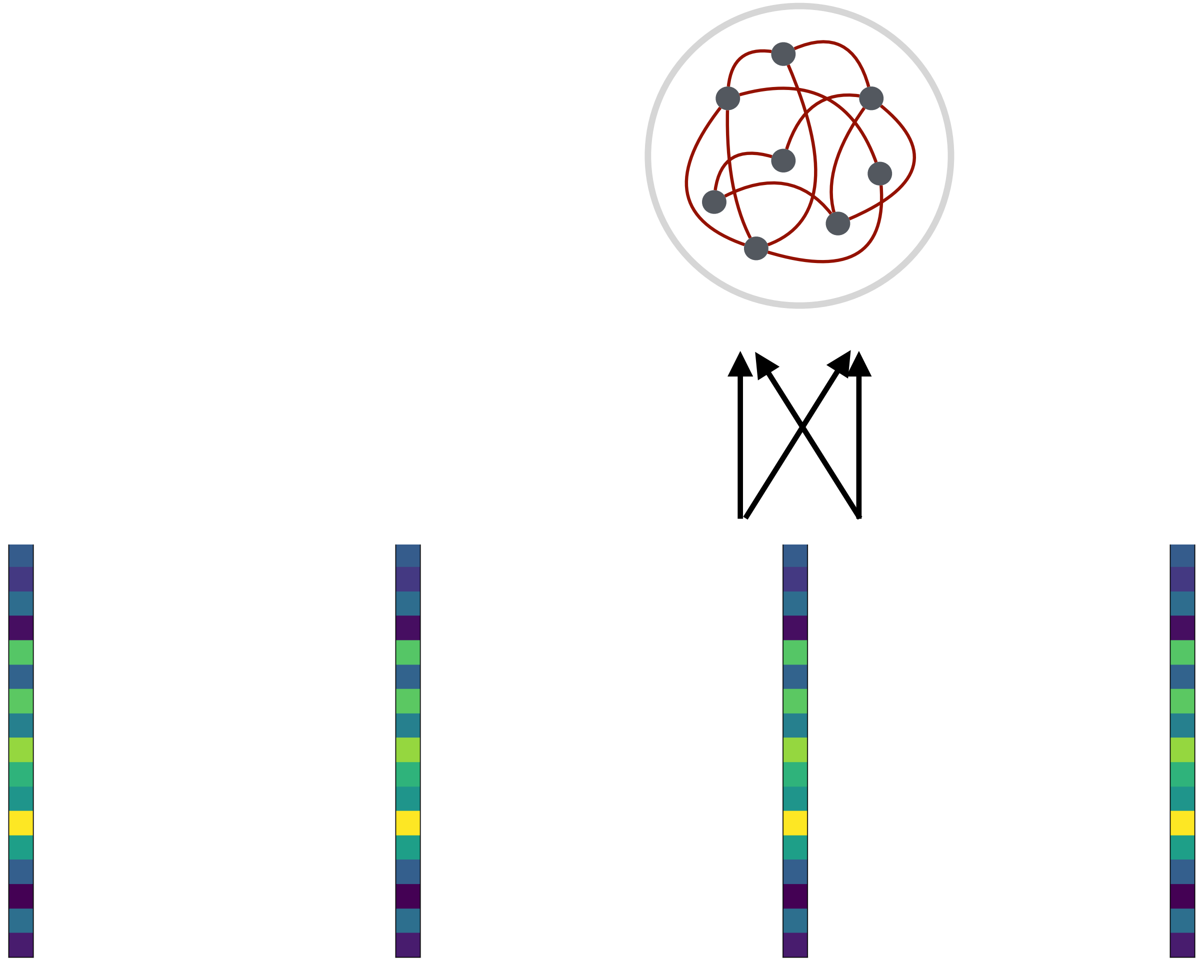


Life is really tough

RNNs

Example: sentiment analysis

Mapping sequence -> vector

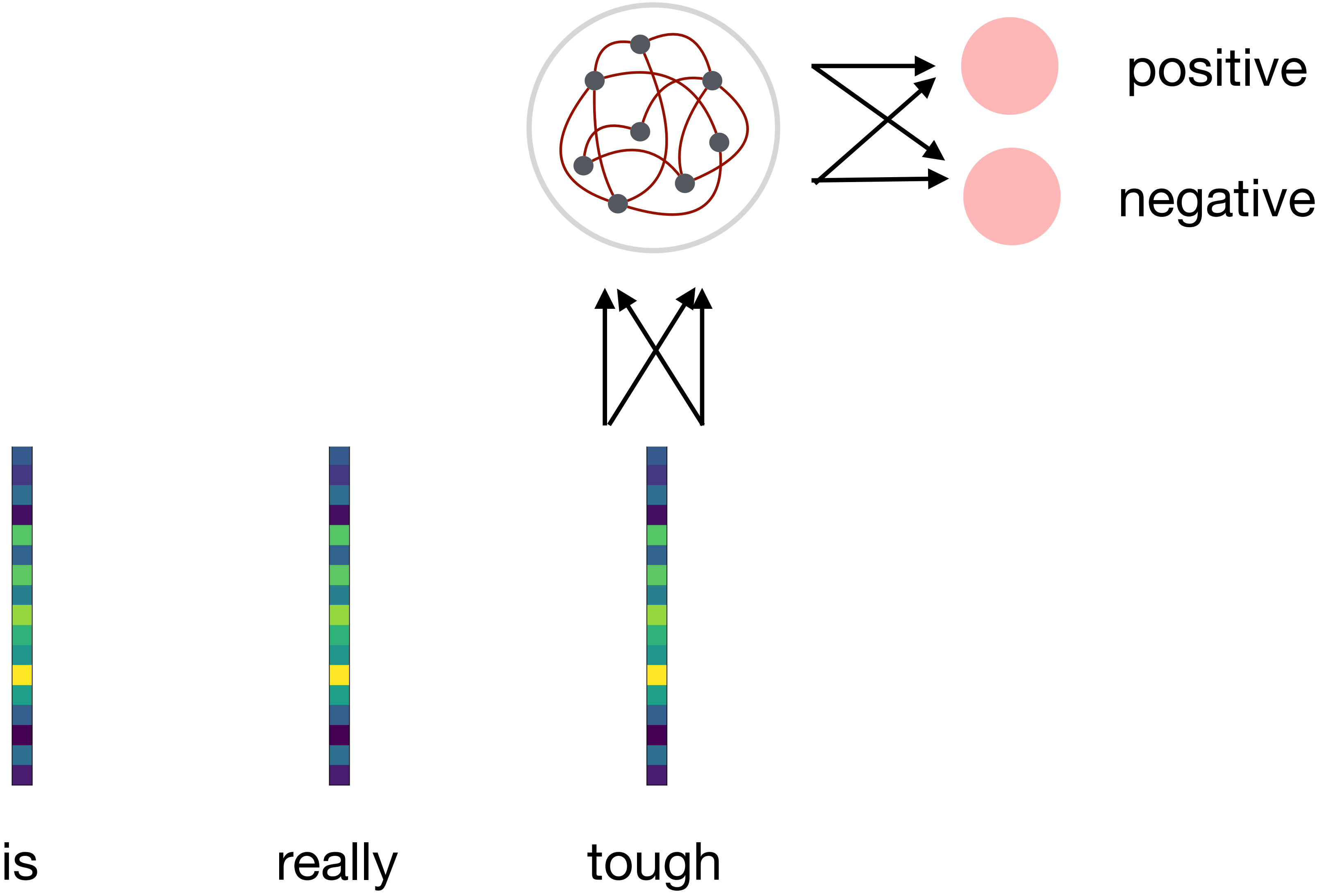


Life is really tough

RNNs

Example: sentiment analysis

Mapping sequence -> vector



positive

negative

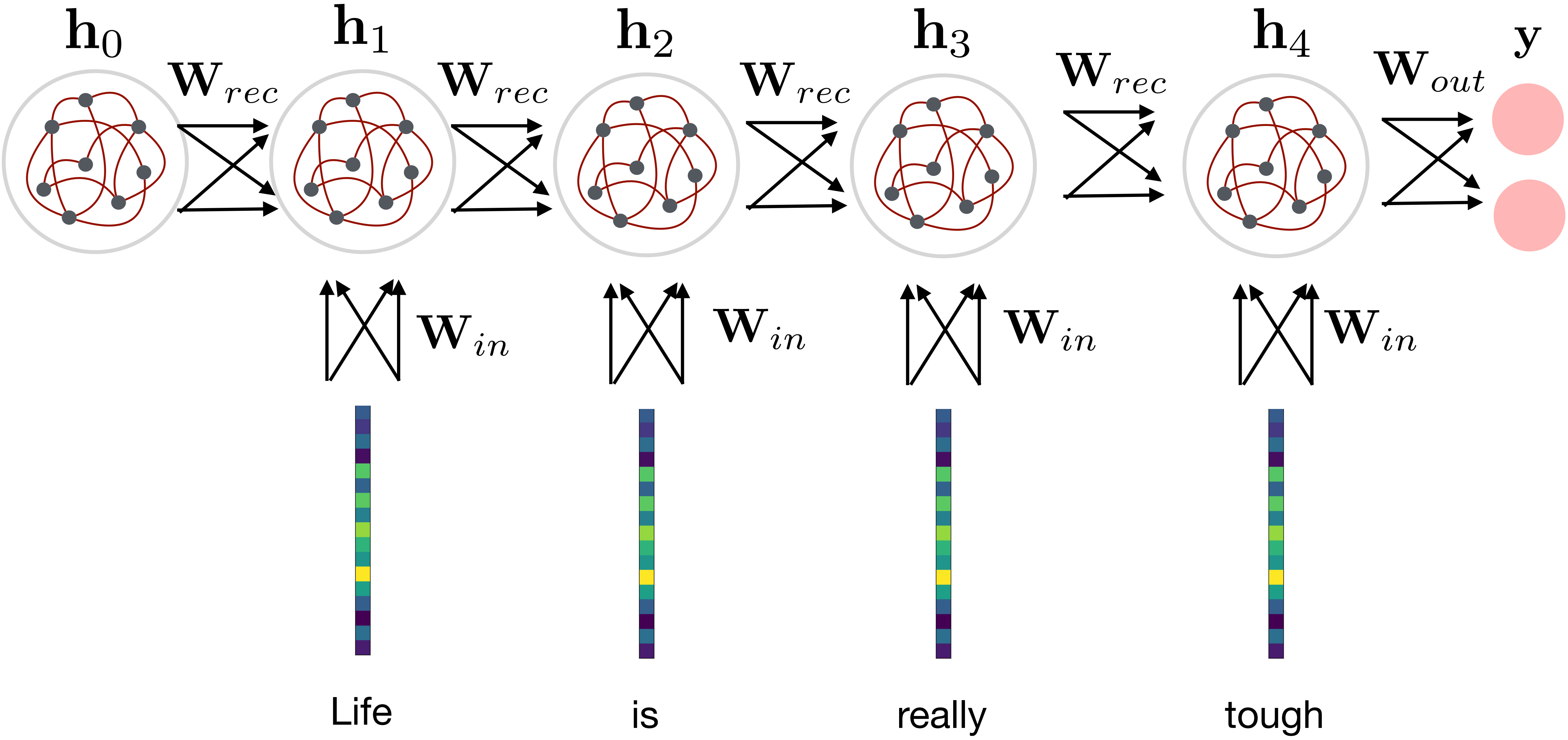
is

really

tough

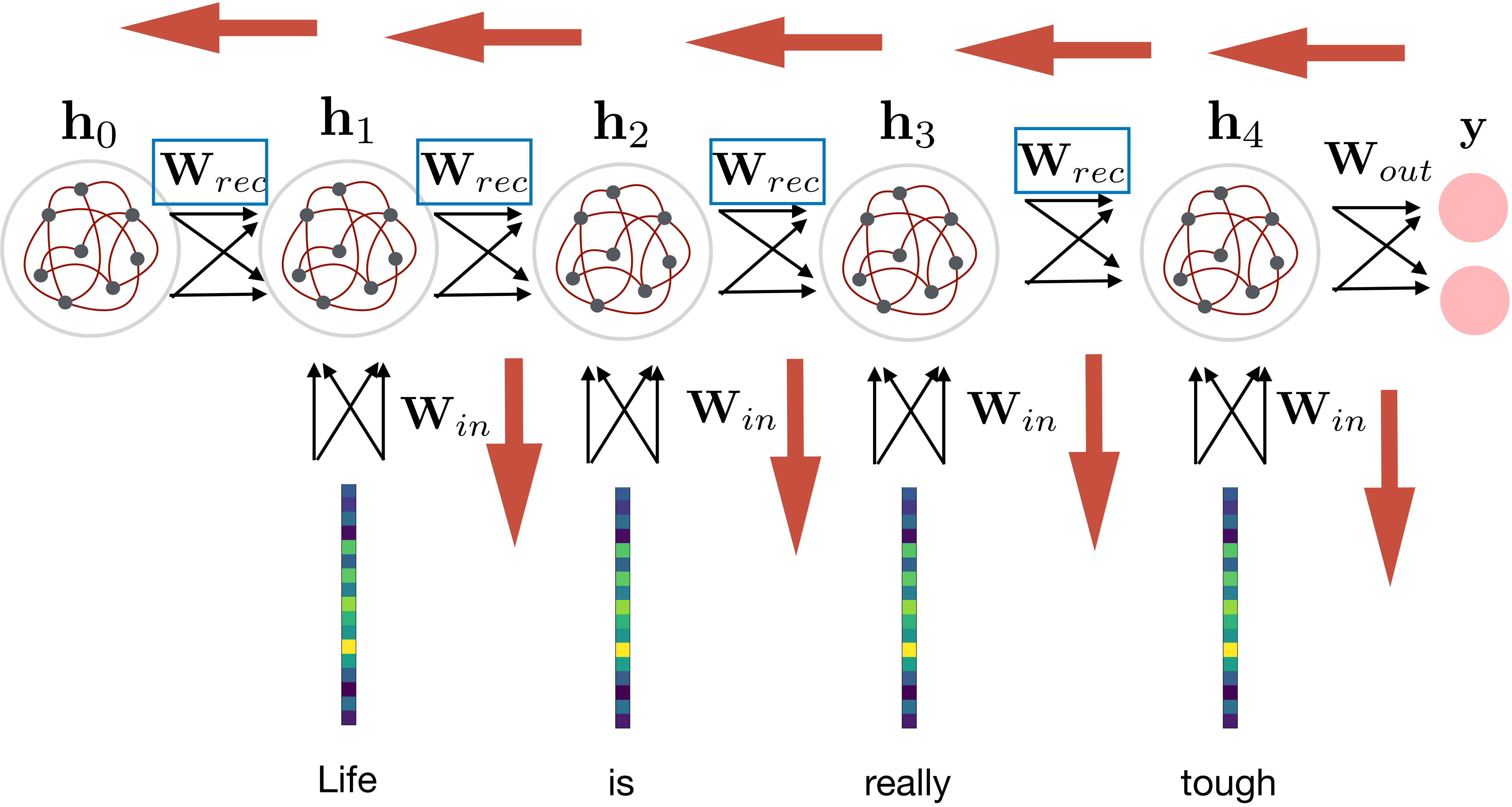
Backprop through time

Unrolled computation graph



Backprop through time

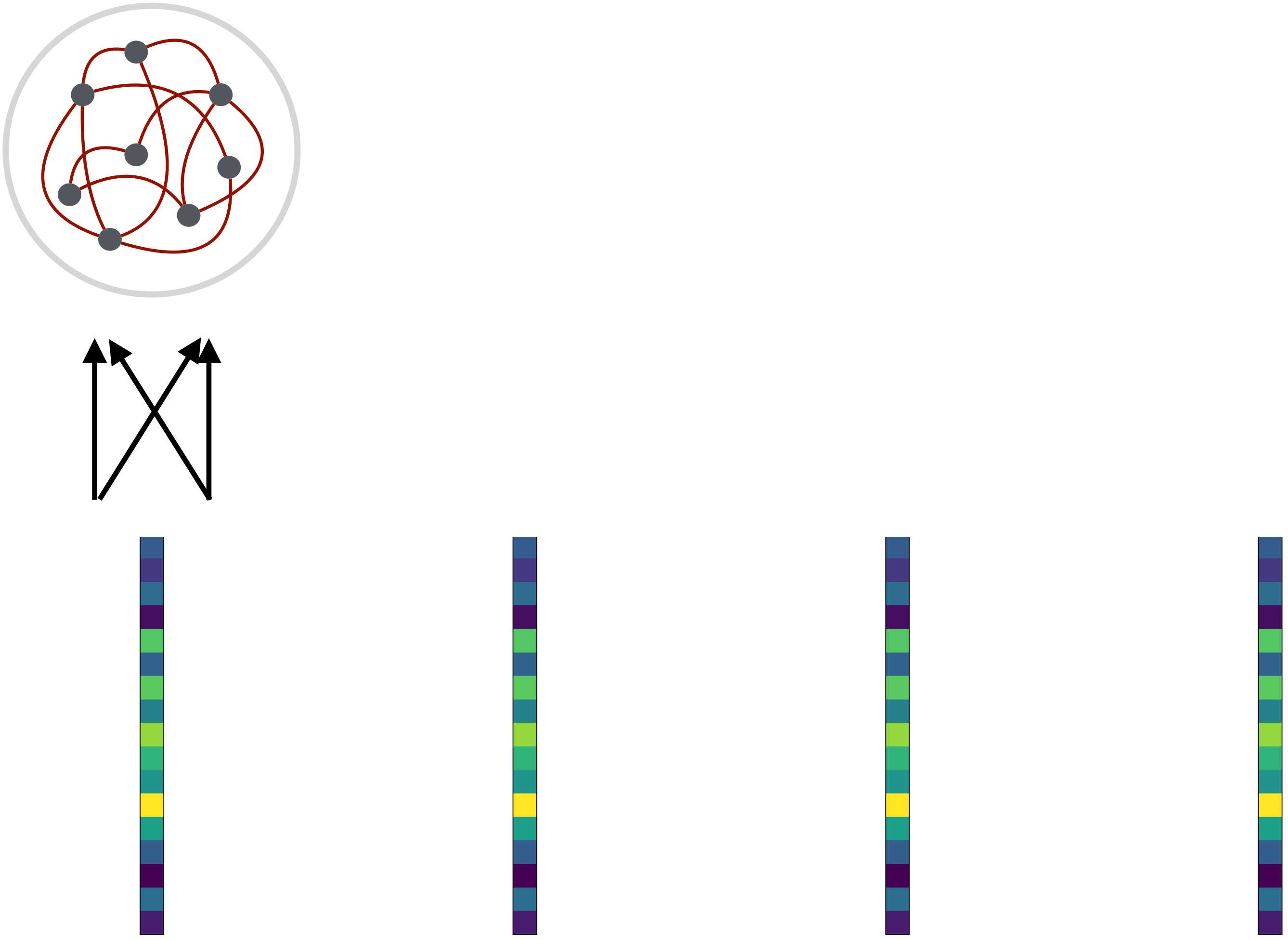
Unrolled computation graph



seq2seq (Sutskever et al. 2014)

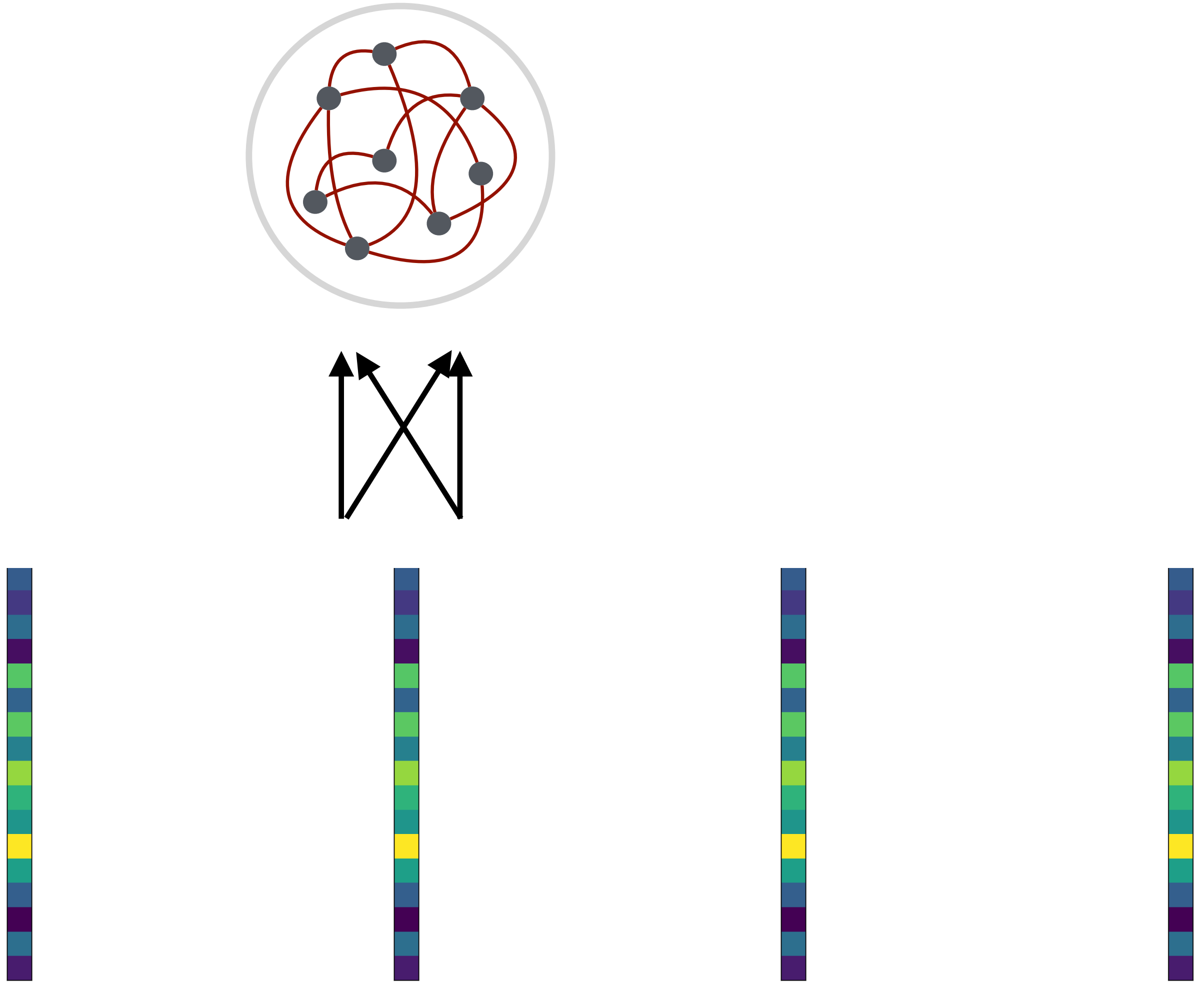
Example: language translation

Mapping sequence -> sequence



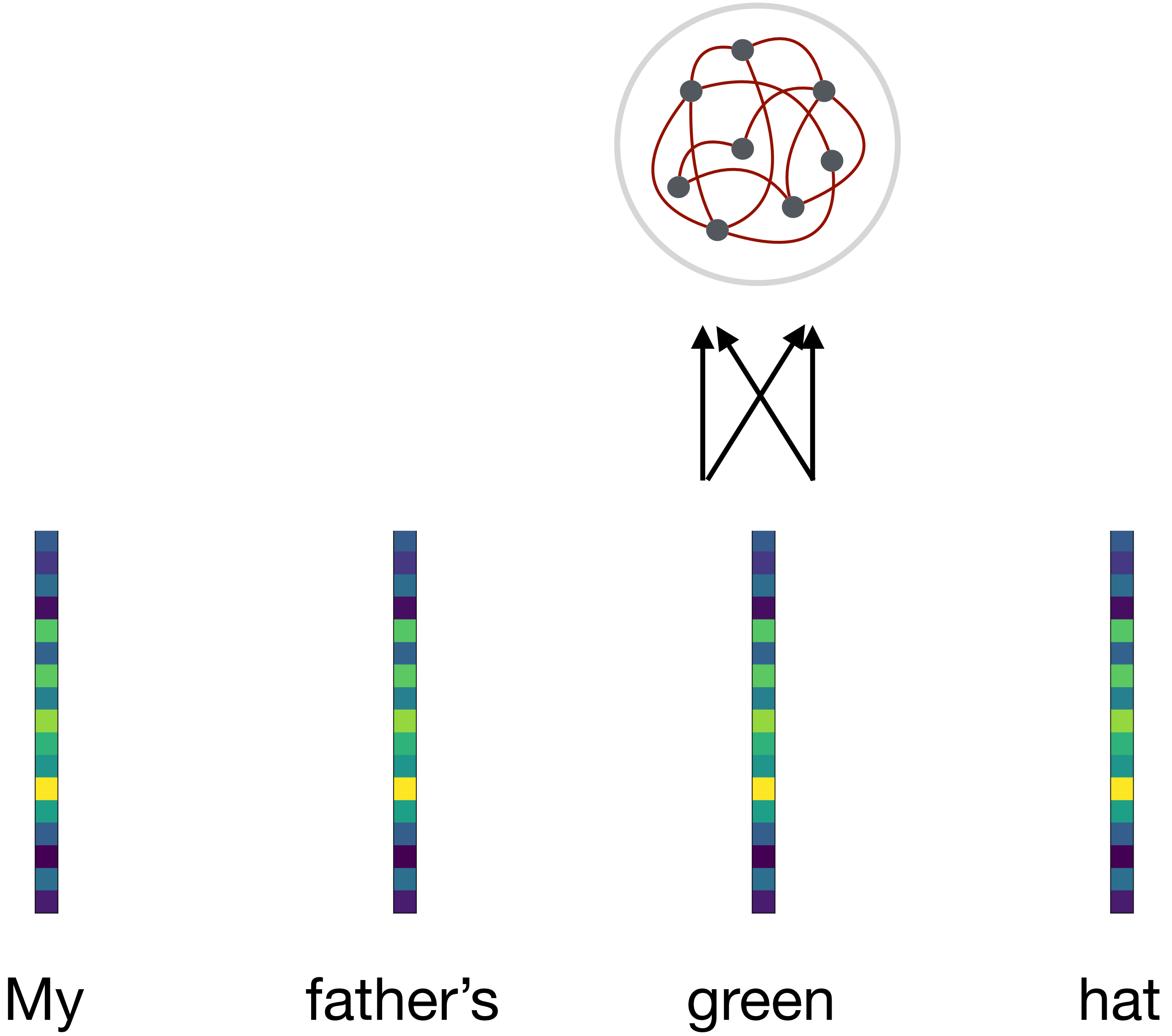
My father's green hat

seq2seq (Sutskever et al. 2014)

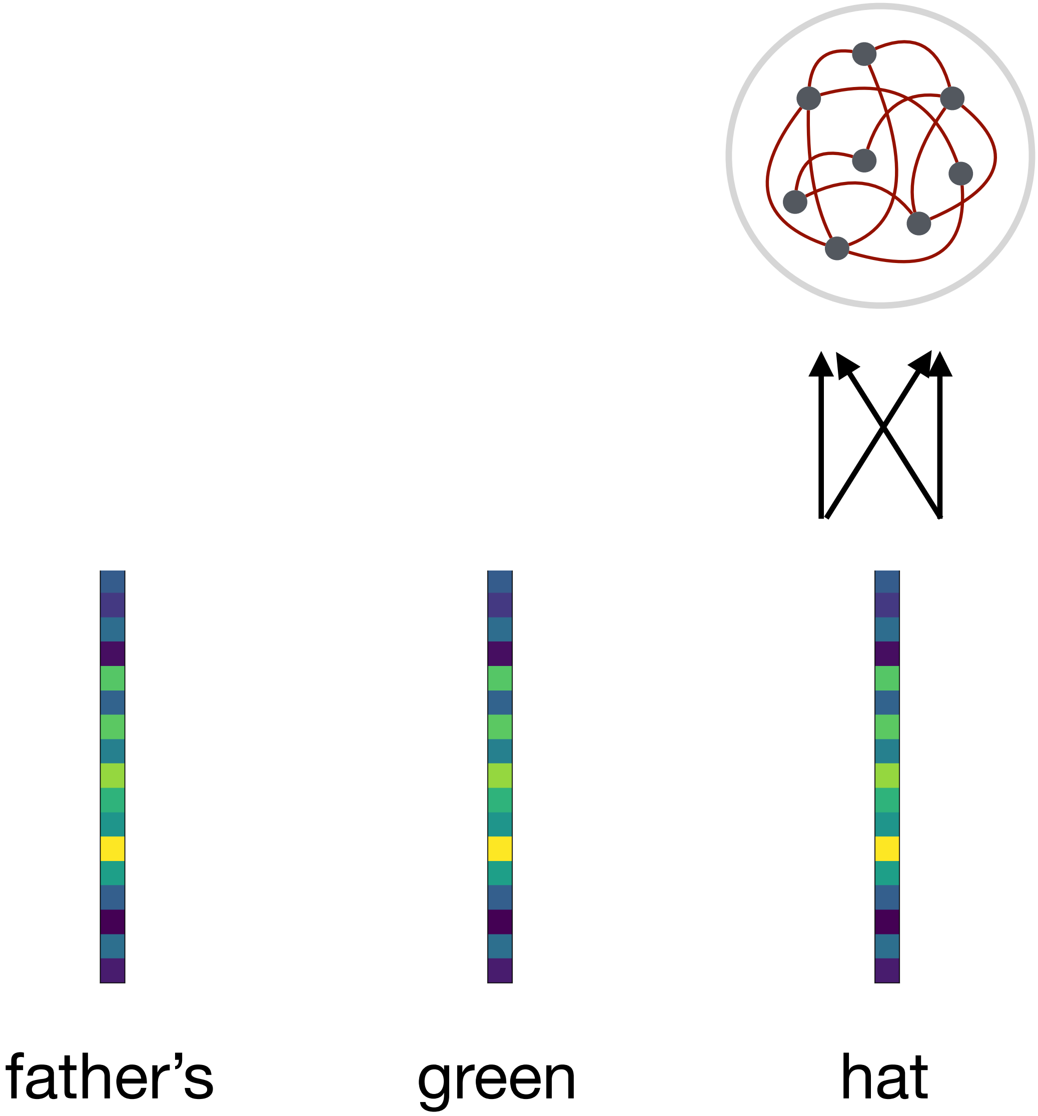


My father’s green hat

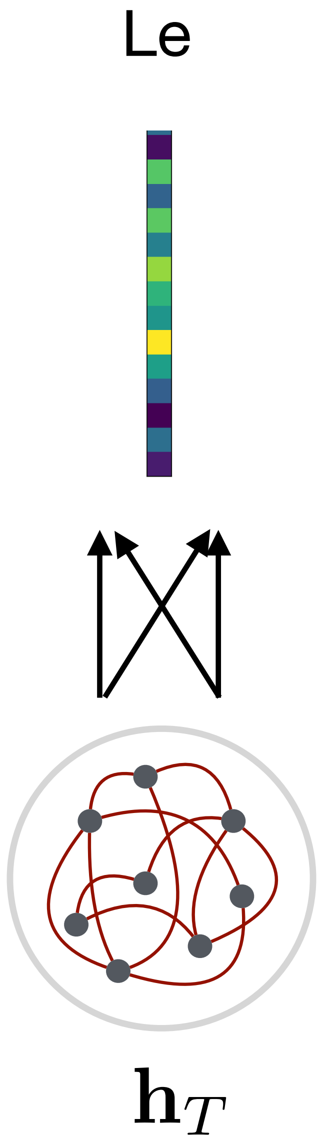
seq2seq (Sutskever et al. 2014)



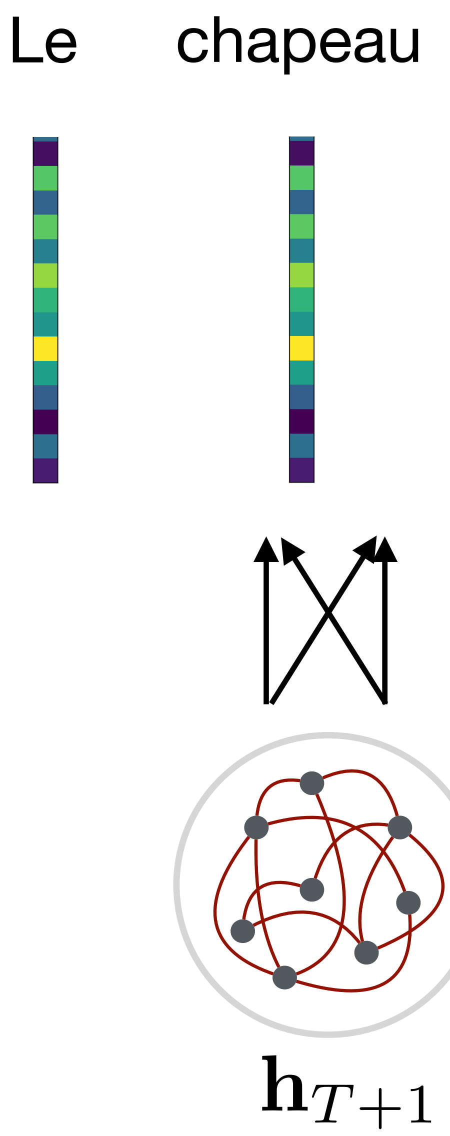
seq2seq (Sutskever et al. 2014)



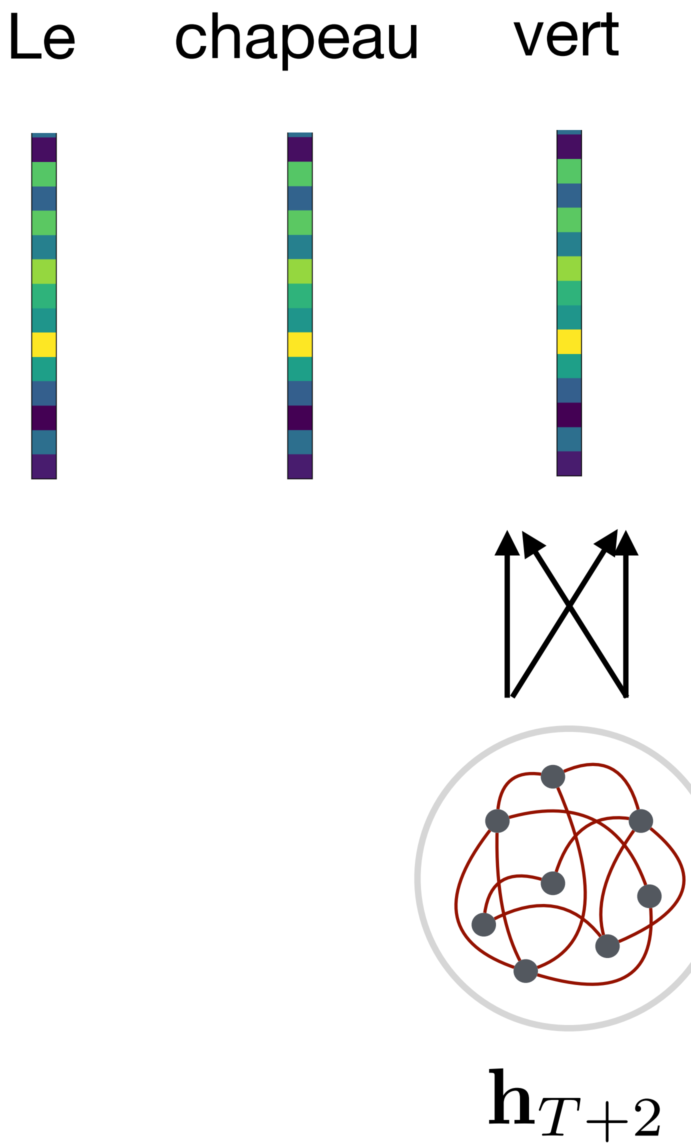
seq2seq (Sutskever et al. 2014)



seq2seq (Sutskever et al. 2014)



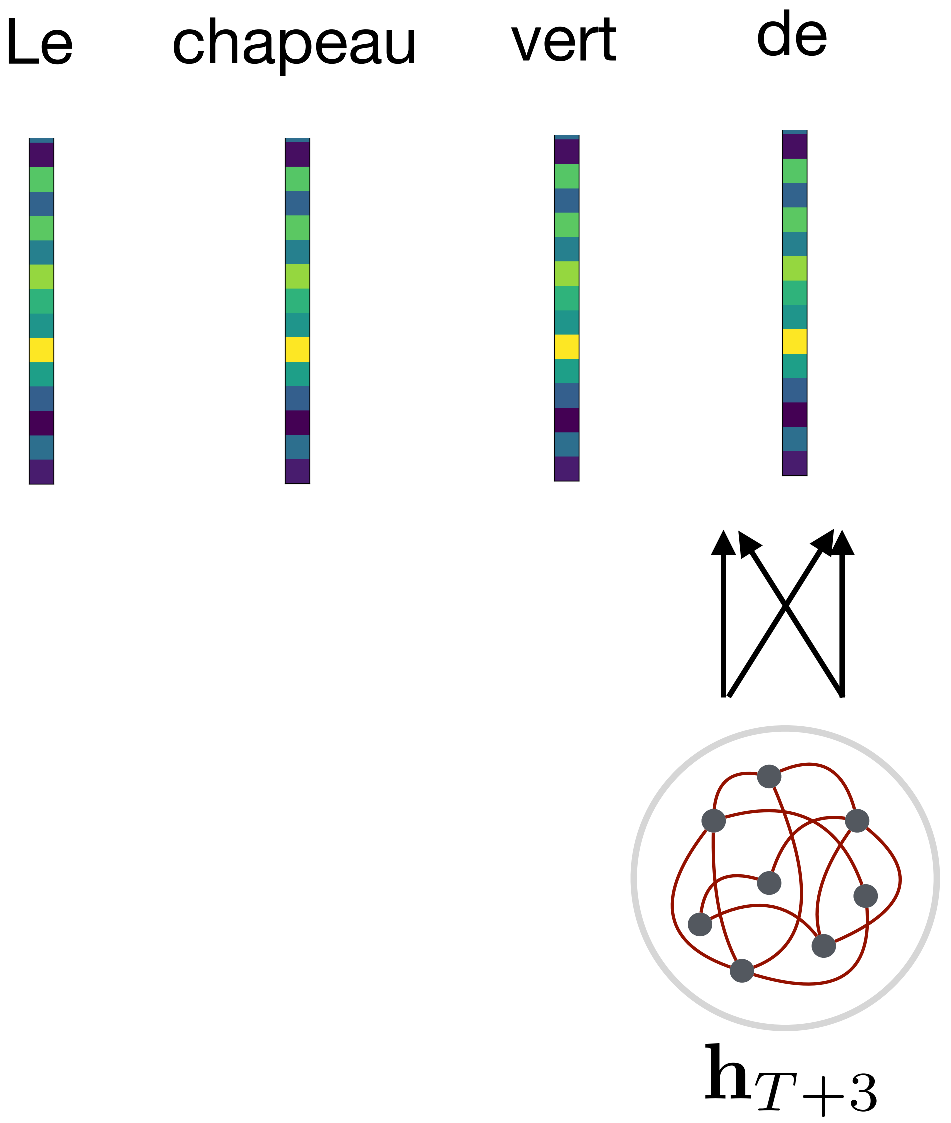
seq2seq (Sutskever et al. 2014)



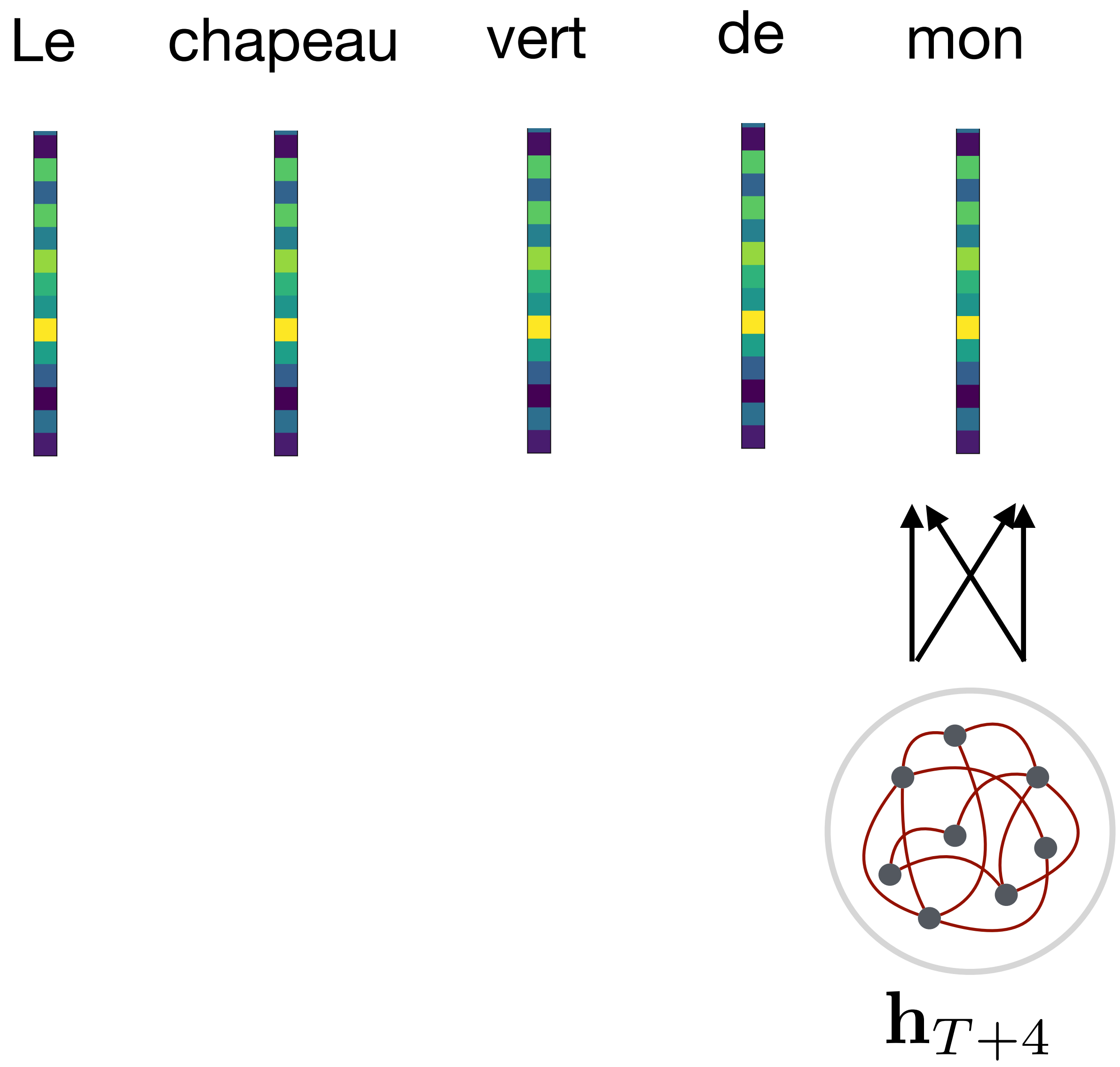
Le chapeau vert

$\mathbf{h}_{T+2}$

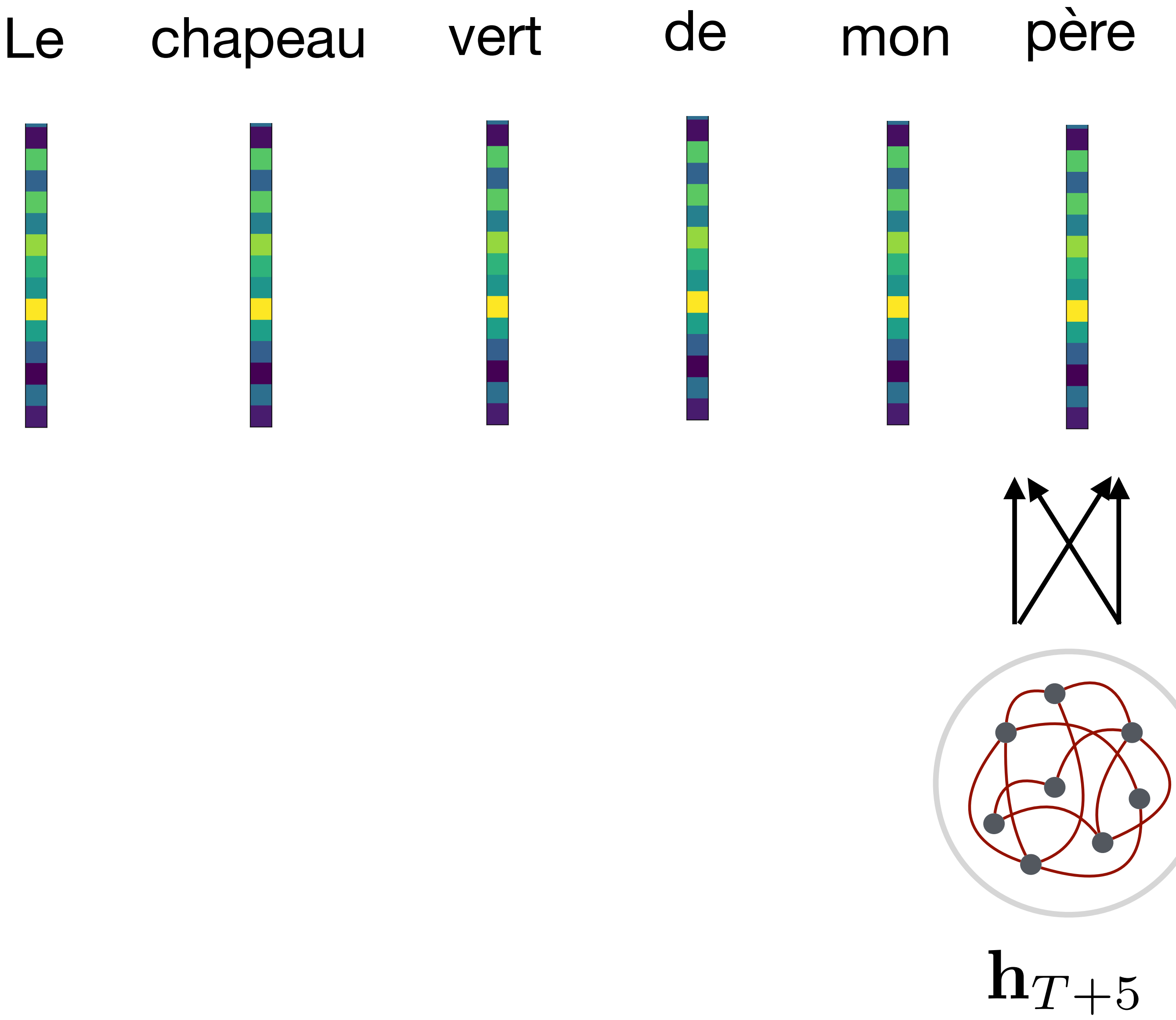
seq2seq (Sutskever et al. 2014)



seq2seq (Sutskever et al. 2014)



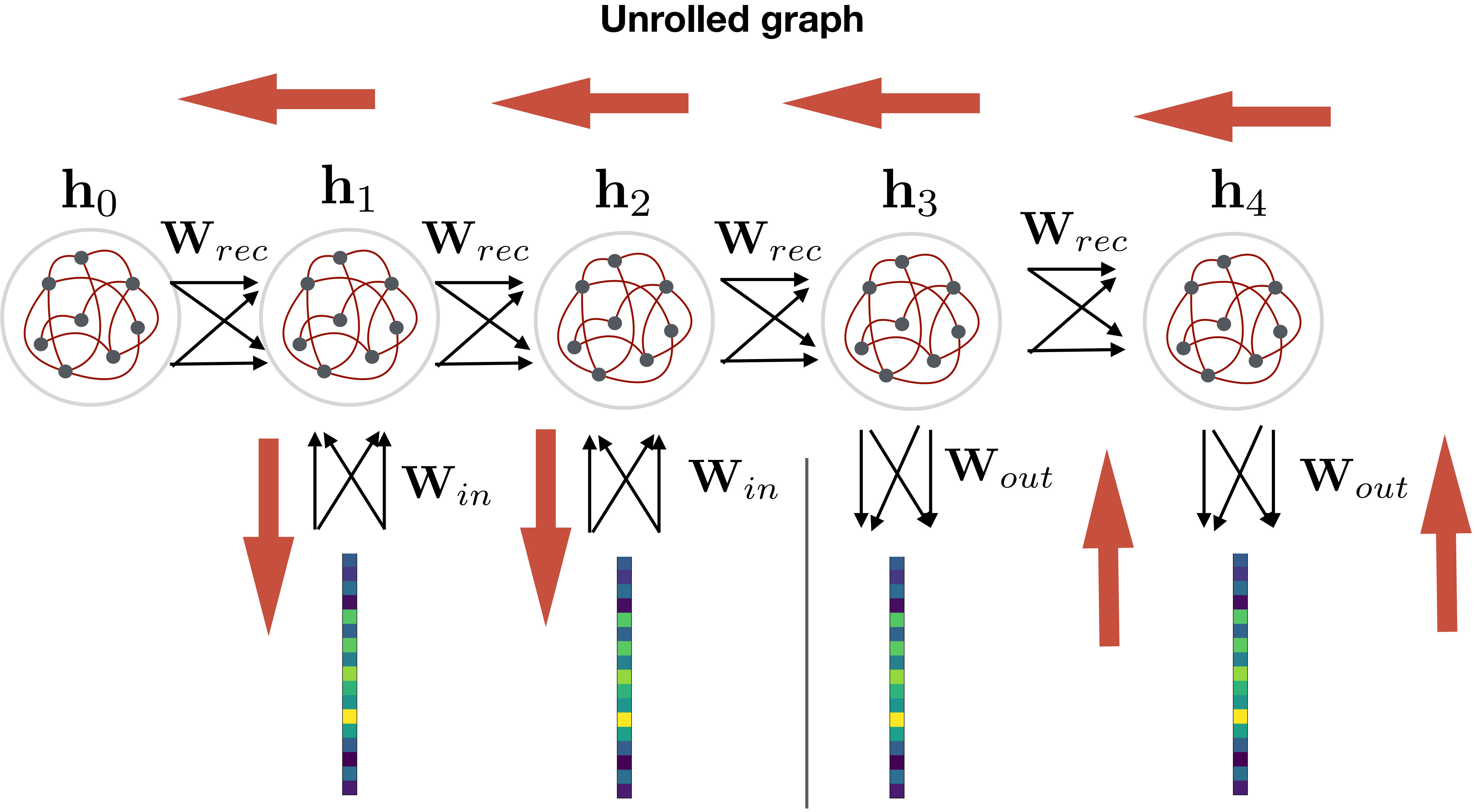
seq2seq (Sutskever et al. 2014)



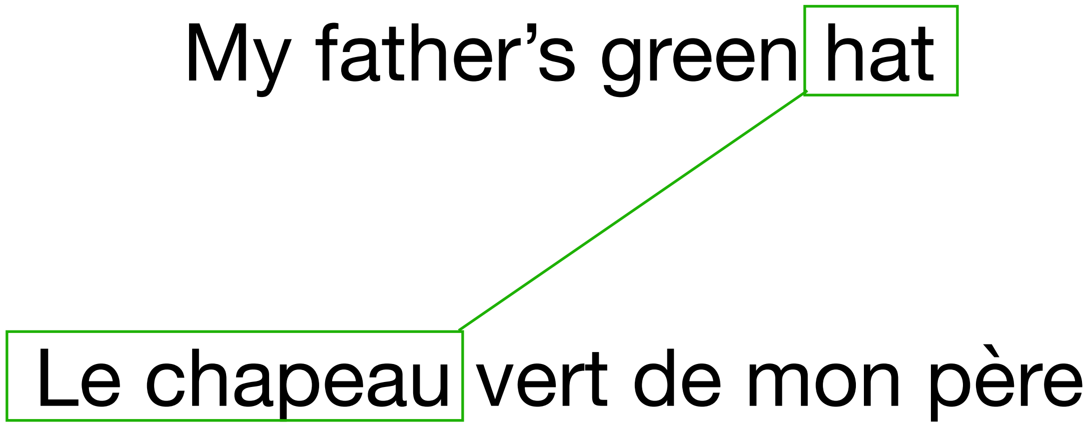
Le chapeau vert de mon père

$\mathbf{h}_{T+5}$

seq2seq (Sutskever et al. 2014)



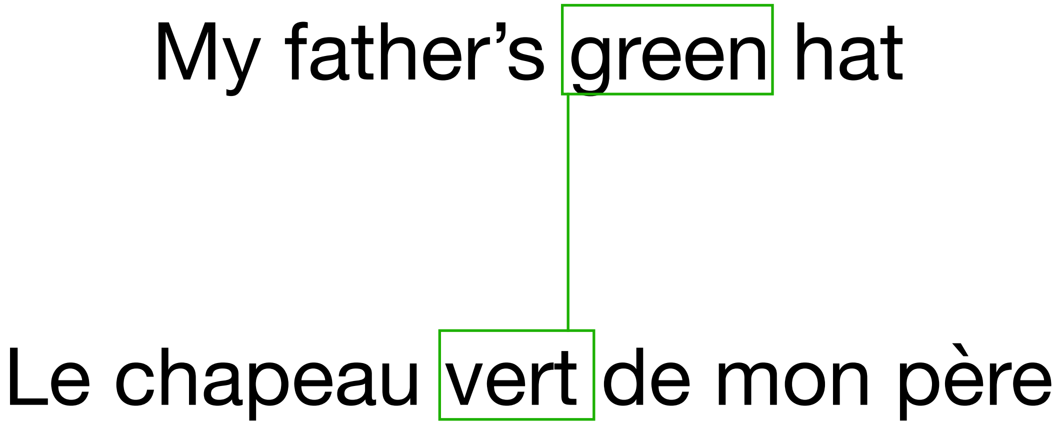
Non-linear processing of the sequence



My father's green hat

Le chapeau vert de mon père

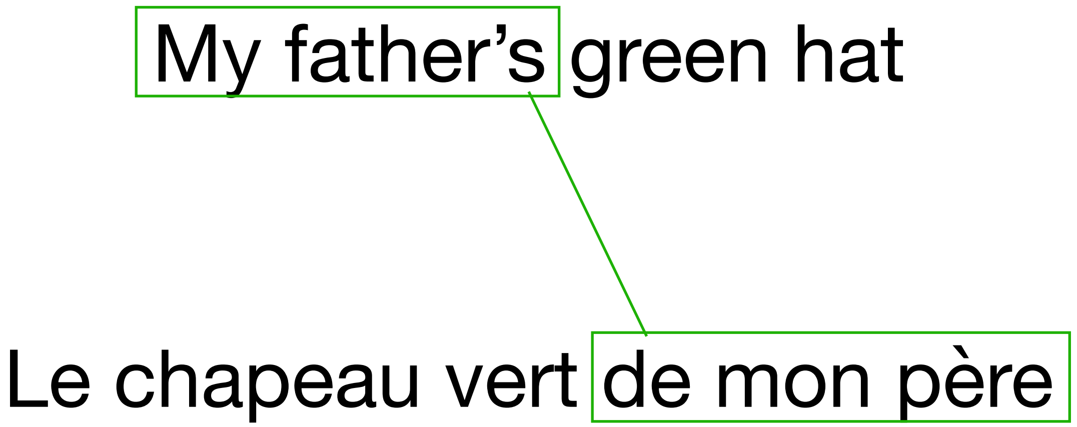
Non-linear processing of the sequence



My father’s green hat

Le chapeau vert de mon père

Non-linear processing of the sequence



My father's green hat

Le chapeau vert de mon père

Attention

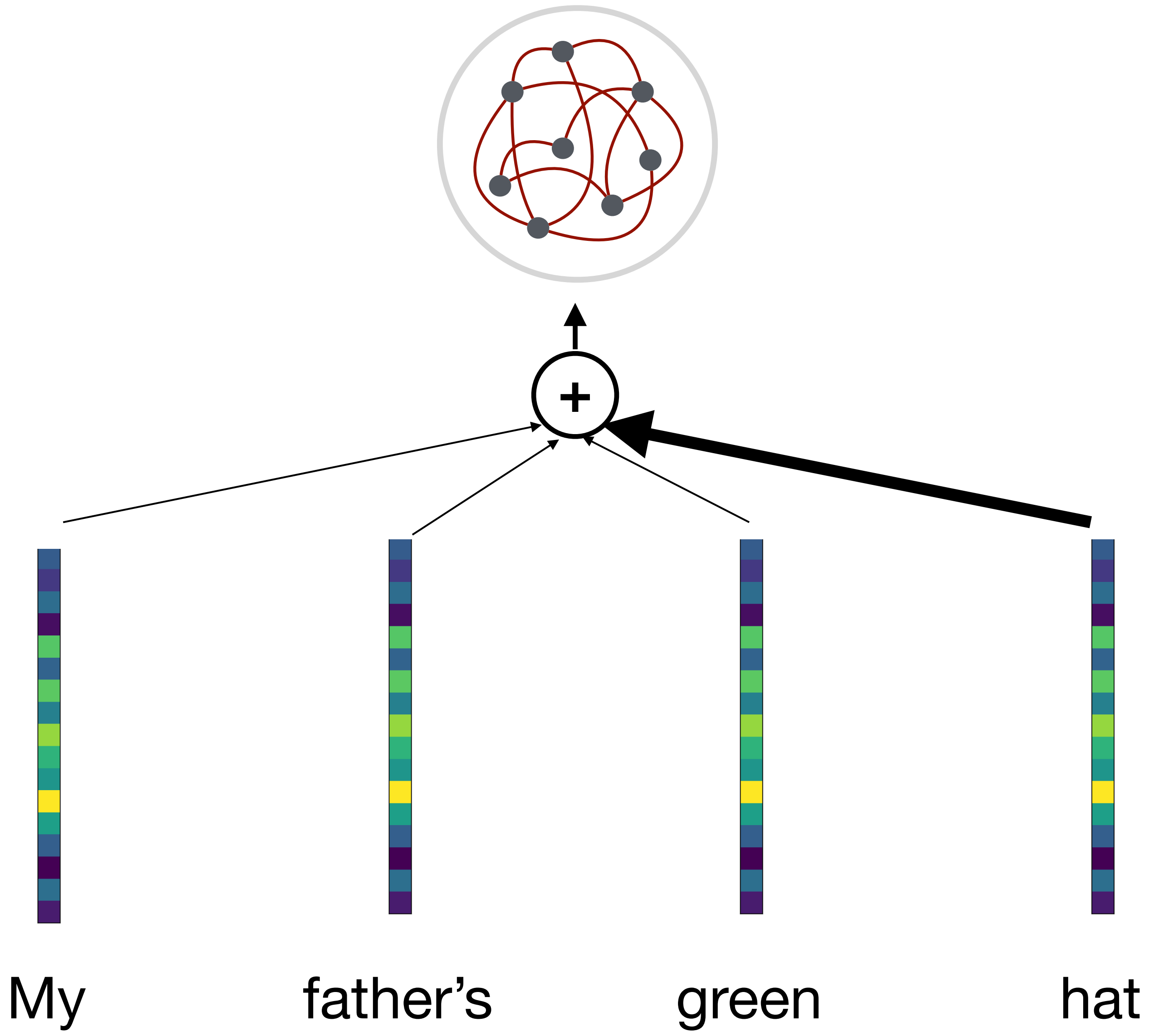
“It is the taking possession by the mind […] of one out of what seem several simultaneously possible objects or trains of thought.”

William James, 1890

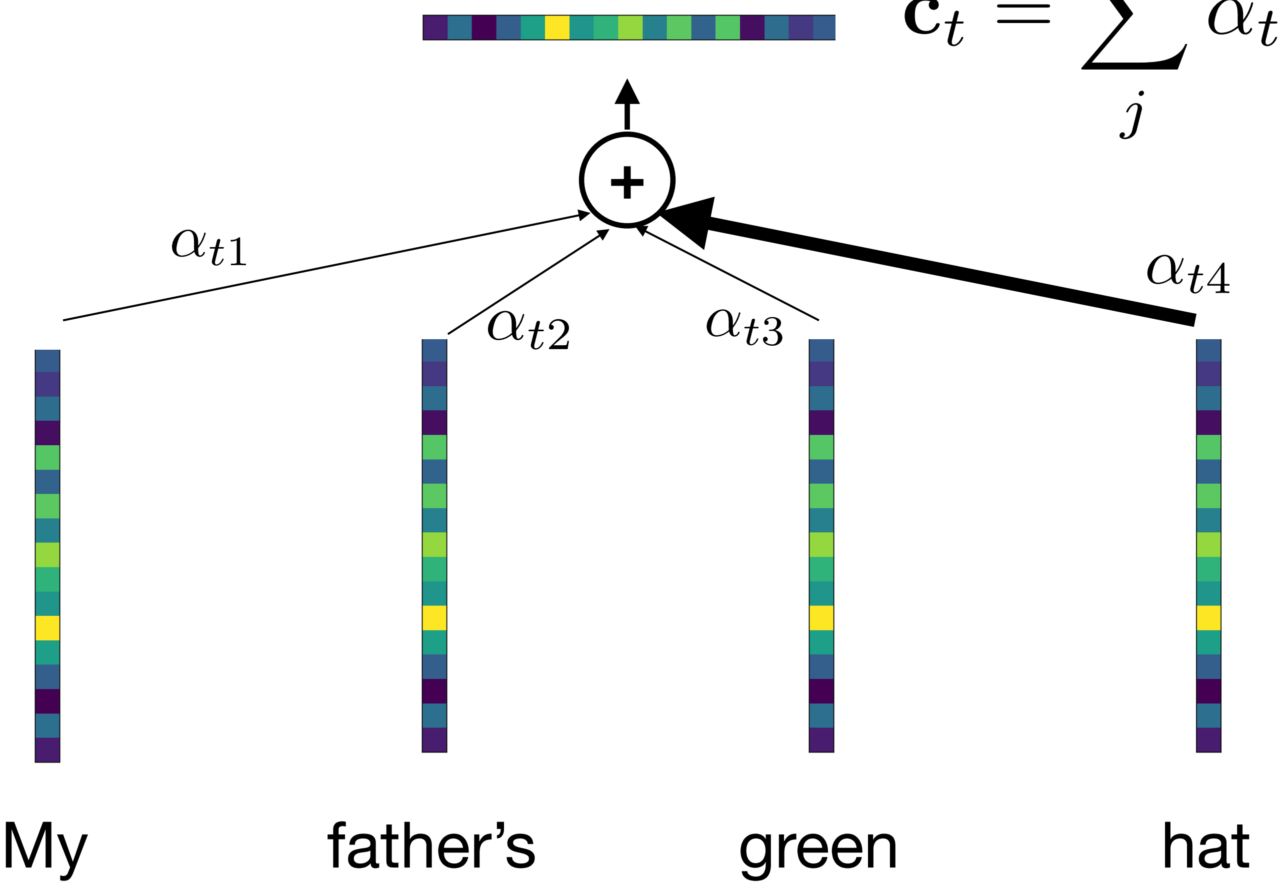
Modern analogy:
It is a “spotlight"

An artificial attentional mechanism for translation

Bahdanau-Cho-Bengio, *Neural machine translation by jointly learning to align and translate, ICLR 2015*



An artificial attentional mechanism for translation



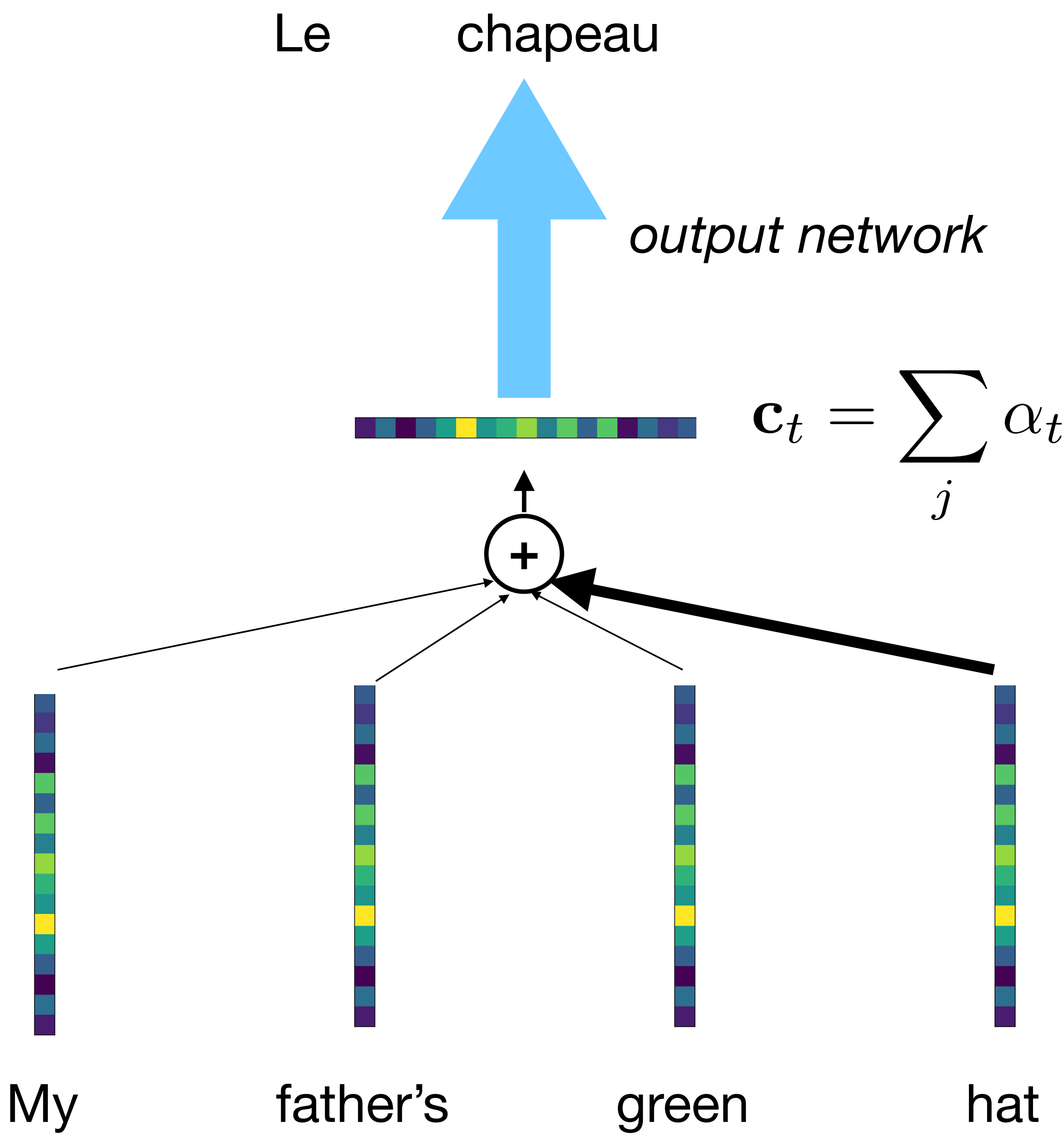
$$\mathbf{c}_t = \sum_j \alpha_{tj} \mathbf{x}_j$$

with

$$\sum_j \alpha_{tj} = 1$$

$$\alpha_{tj} \in [0, 1]$$

An artificial attentional mechanism for translation



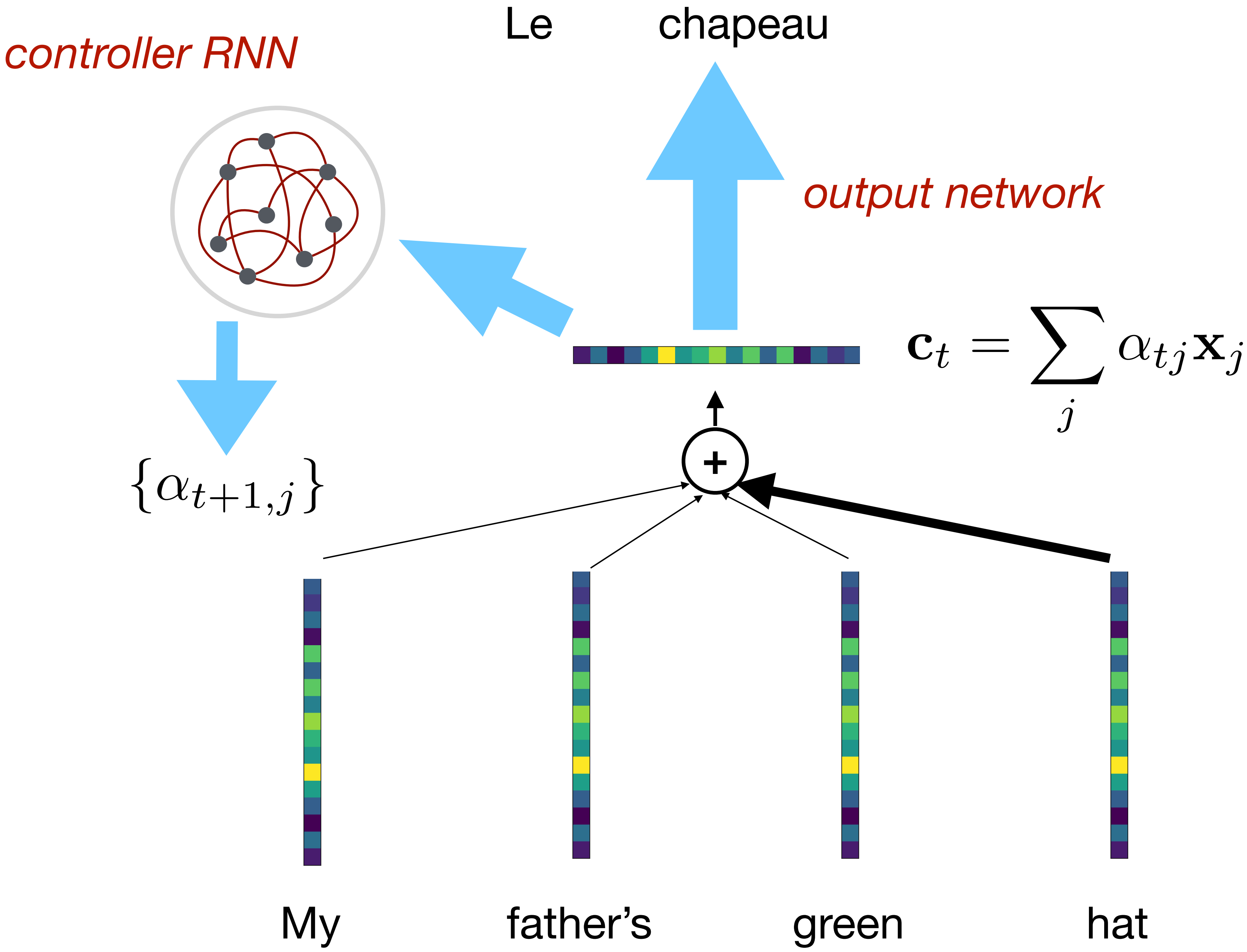
$$\mathbf{c}_t = \sum_j \alpha_{tj} \mathbf{x}_j$$

with

$$\sum_j \alpha_{tj} = 1$$

$$\alpha_{tj} \in [0, 1]$$

An artificial attentional mechanism for translation



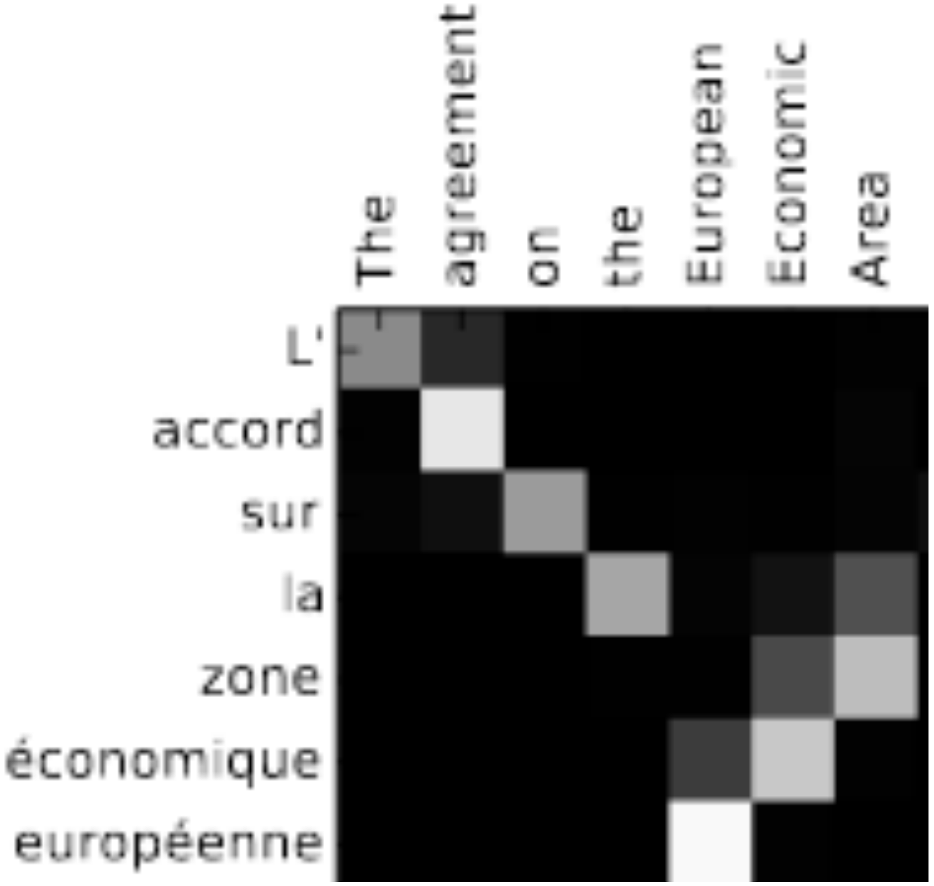
$$\mathbf{c}_t = \sum_j \alpha_{tj}\mathbf{x}_j$$

with

$$\sum_j \alpha_{tj} = 1$$

$$\alpha_{tj} \in [0,1]$$

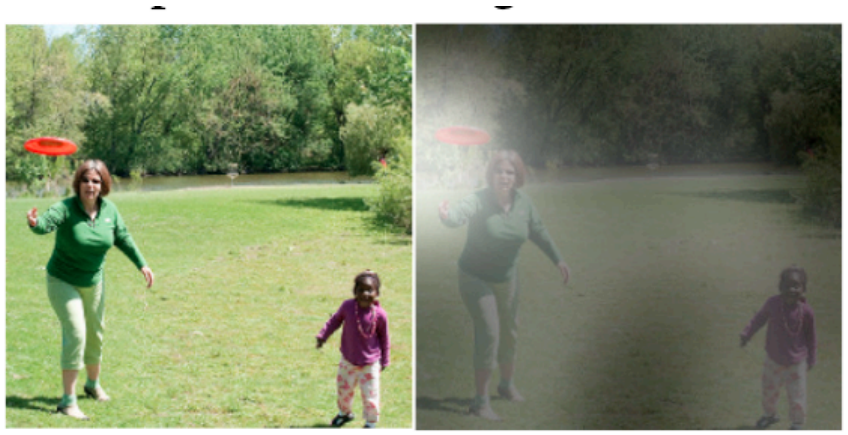
An artificial attentional mechanism for translation



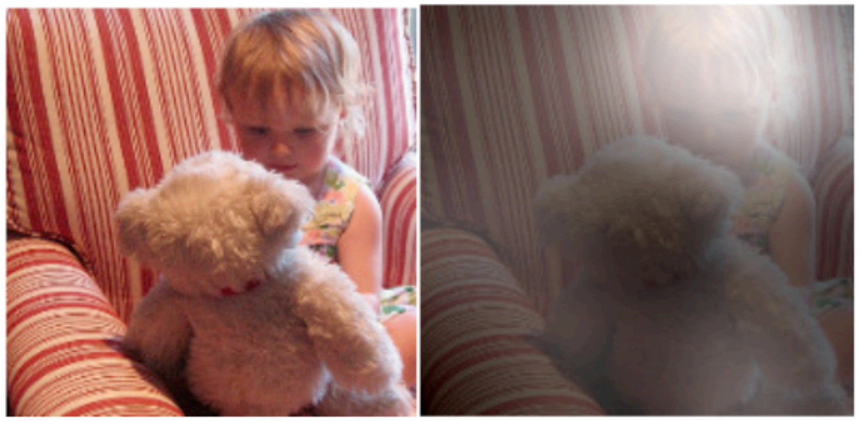
Inherently opens a window on the model's internal process!!

Visual attention for image captioning

Xu et al., *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, 2015

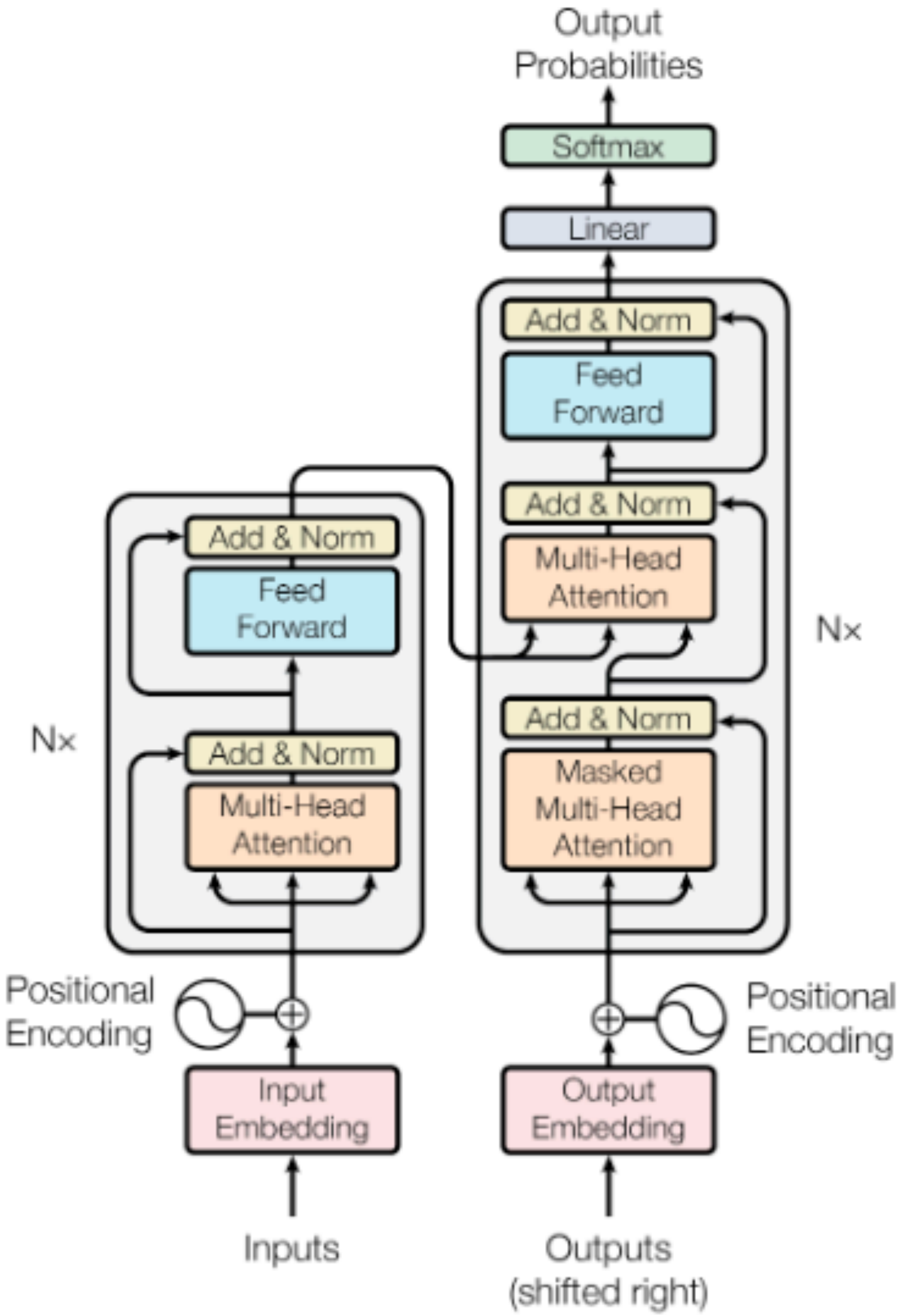


A woman is throwing a frisbee in a park.



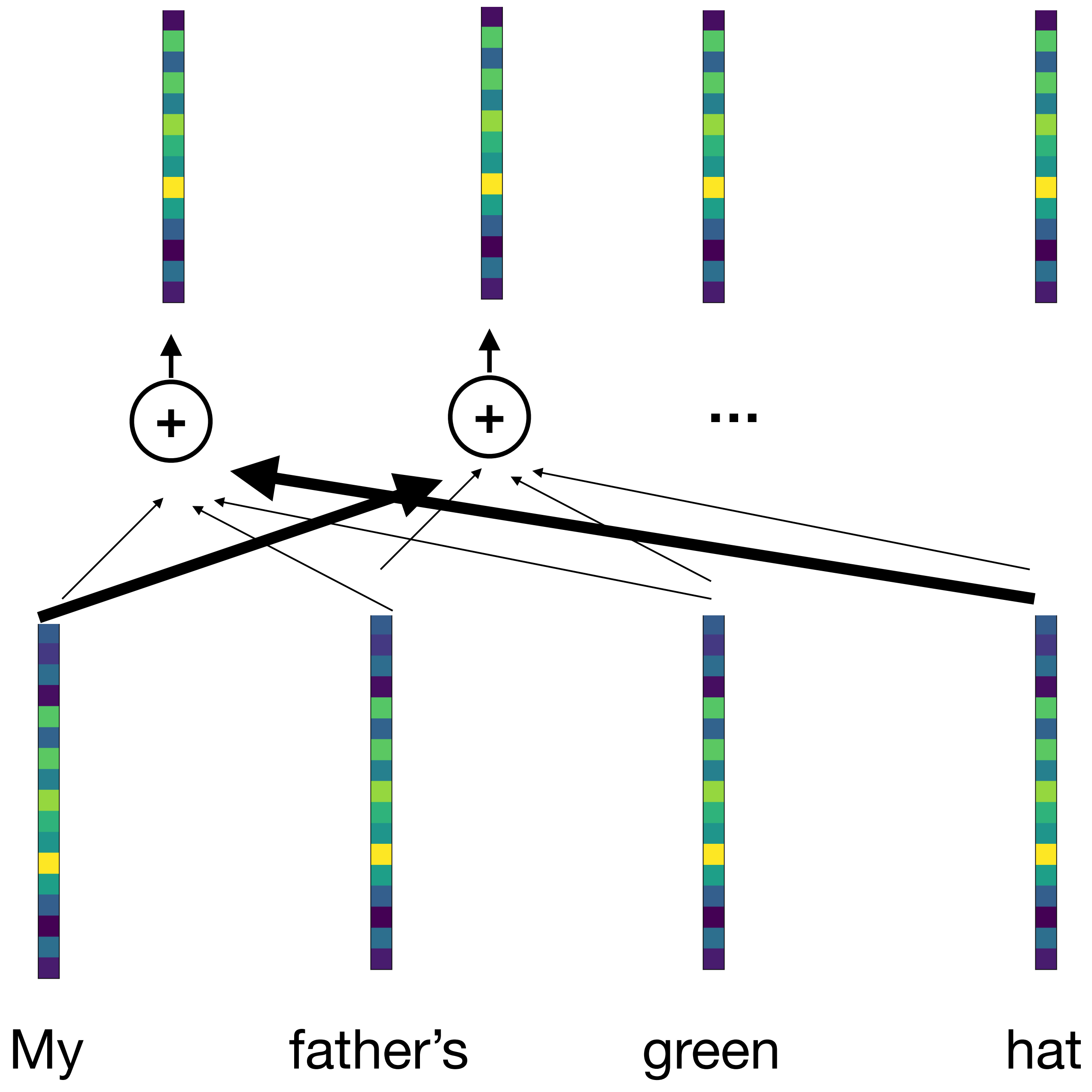
A little girl sitting on a bed with
a teddy bear.

Transformers

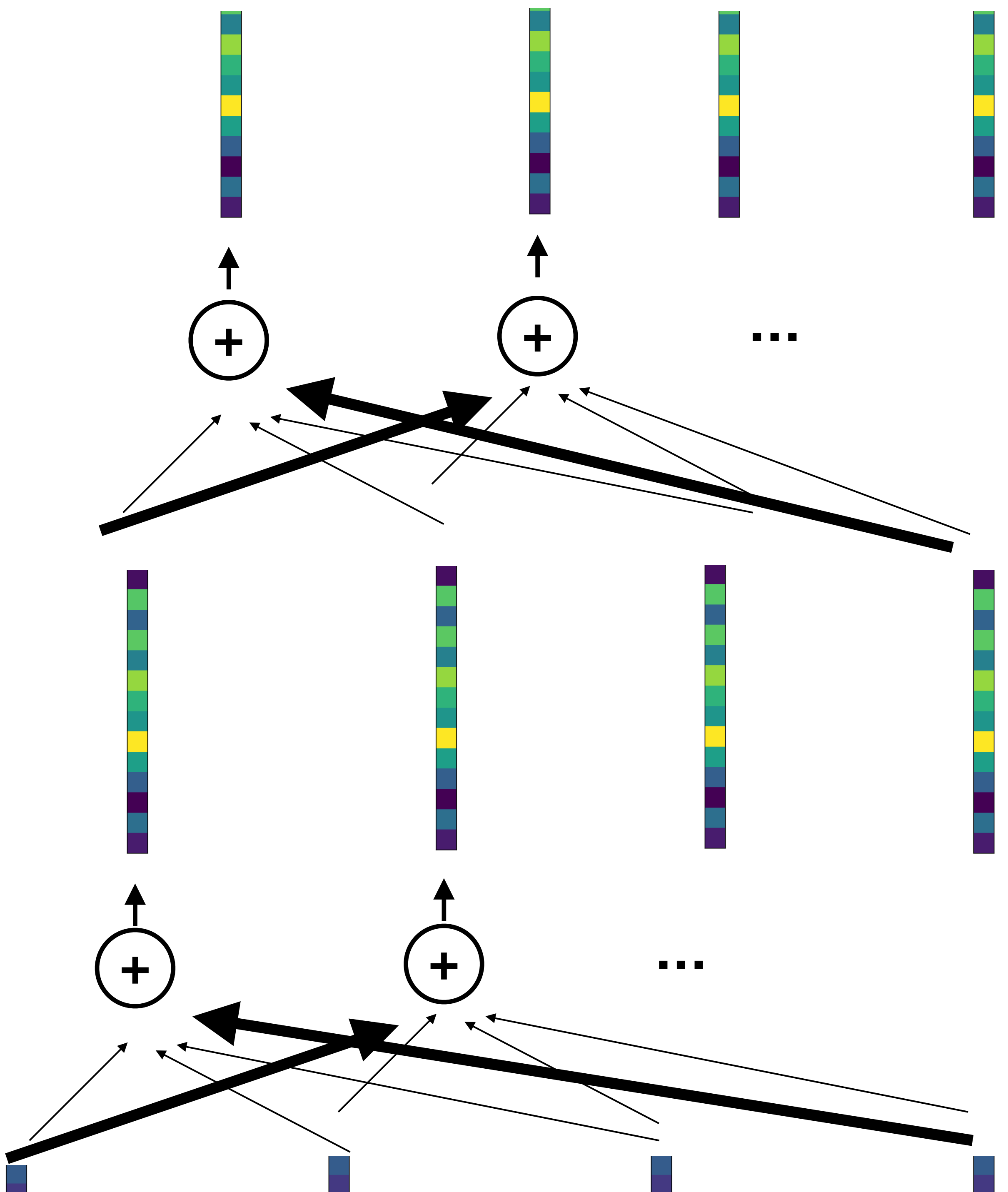


[Vaswani et al 2017]

Transformers: self-attention



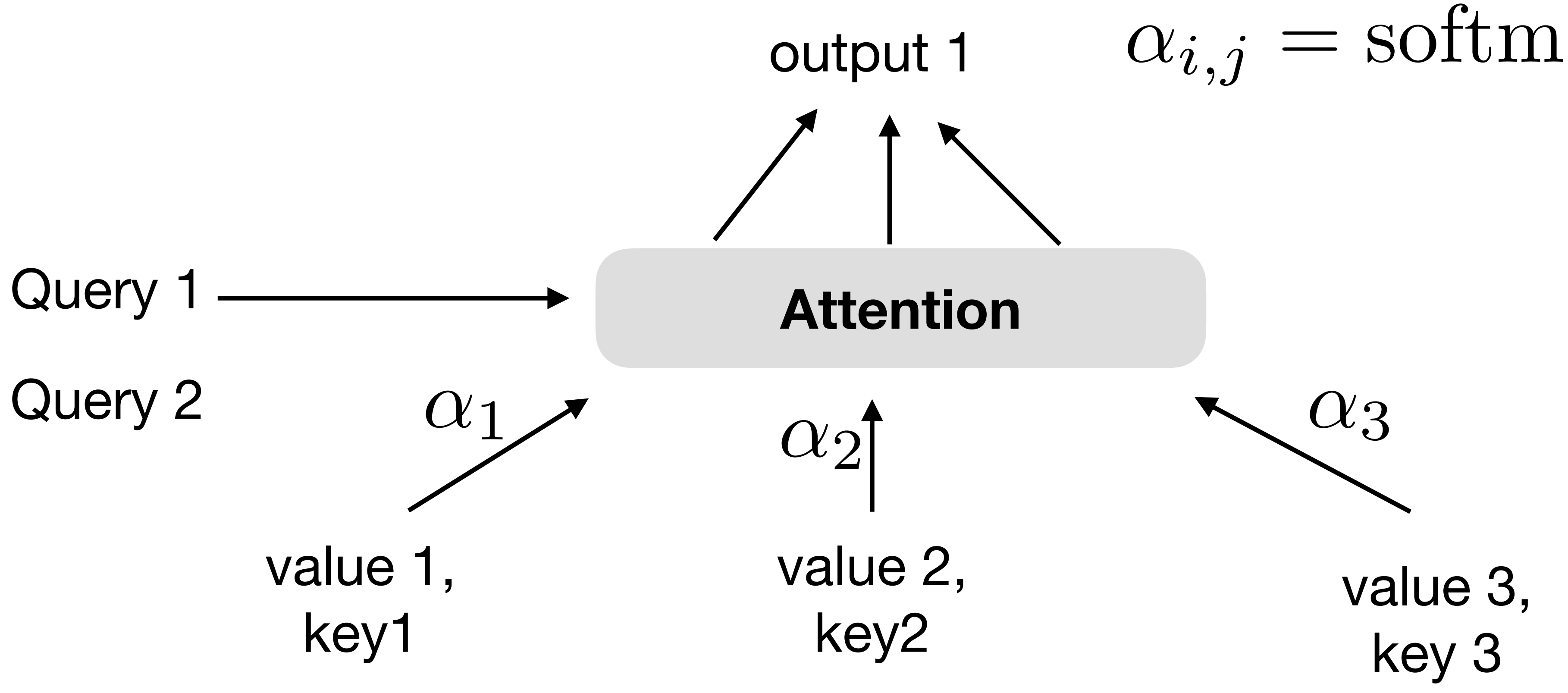
Transformers: self-attention



Transformers: working definition of attention

"An attention function can be described as mapping a query and a set of key-value pairs to an output […] The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key."

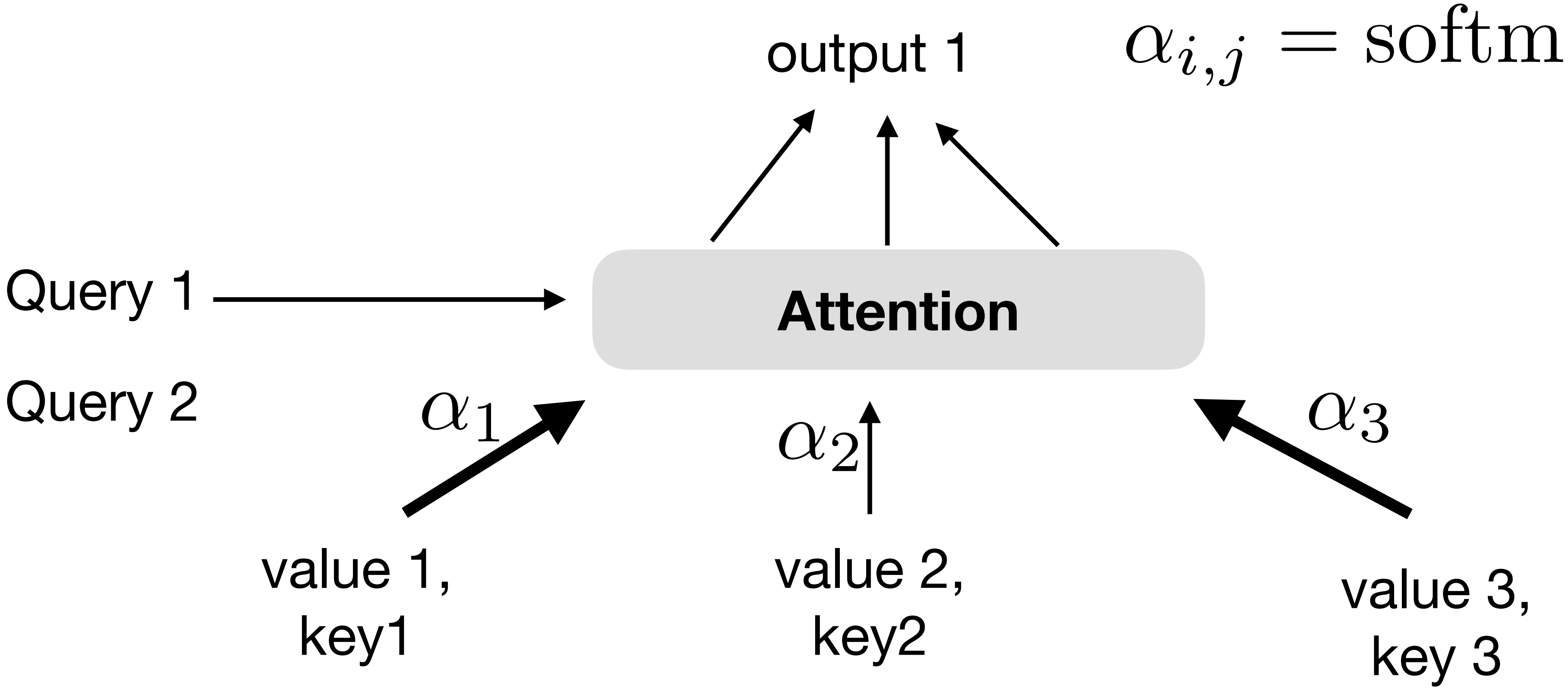
$$\alpha_{i,j} = \mathrm{softmax}(\mathbf{q}_i^T \mathbf{k}_i)\mathbf{v}_i$$



Transformers: working definition of attention

"An attention function can be described as mapping a query and a set of key-value pairs to an output […] The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key."

$$\alpha_{i,j} = \mathrm{softmax}(\mathbf{q}_i^T \mathbf{k}_i)\mathbf{v}_i$$



output 1

Query 1

Attention

Query 2

α_1

α_2

α_3

value 1,
key1

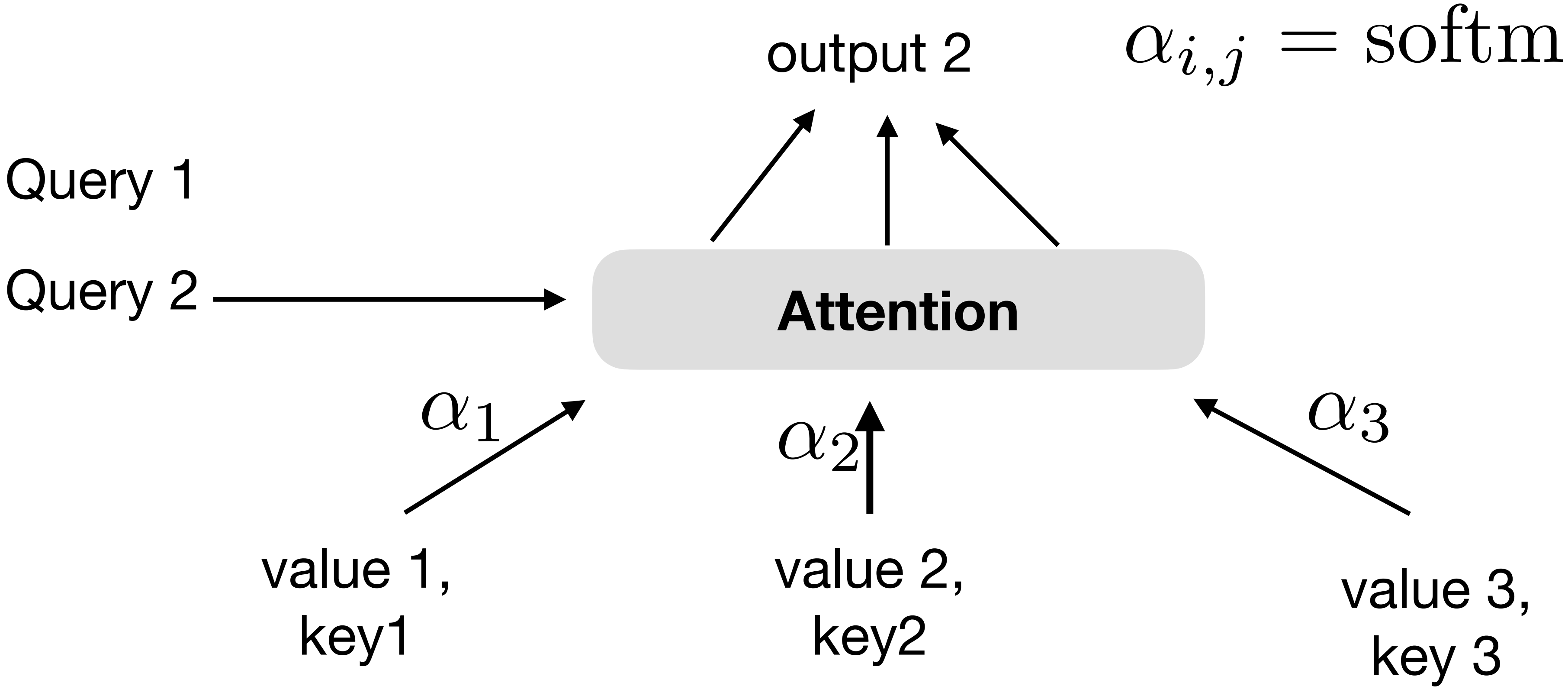
value 2,
key2

value 3,
key 3

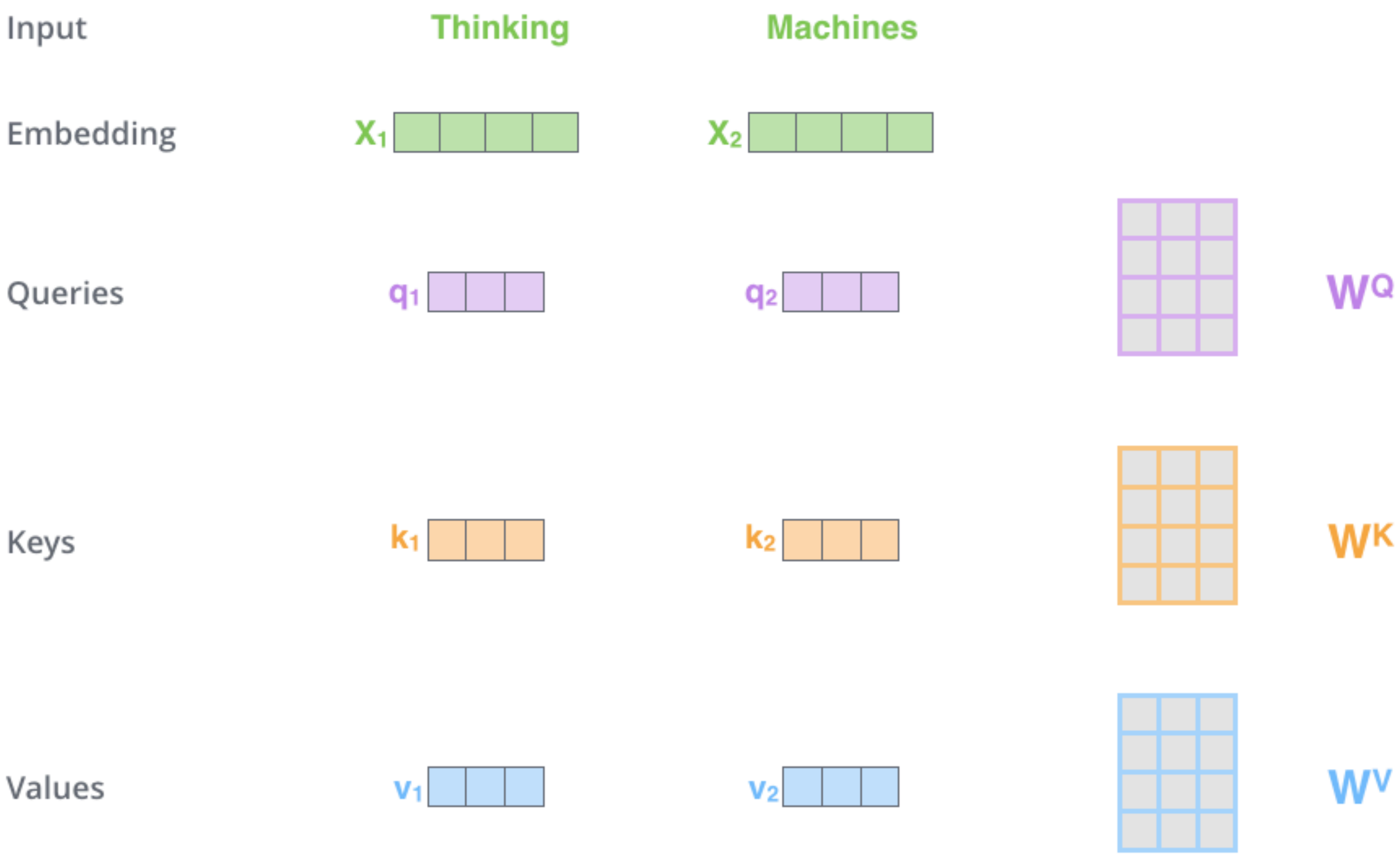
Transformers: working definition of attention

"An attention function can be described as mapping a query and a set of key-value pairs to an output […] The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key."

$$\alpha_{i,j} = \mathrm{softmax}(\mathbf{q}_i^T \mathbf{k}_i)\mathbf{v}_i$$



Transformers: working definition of attention



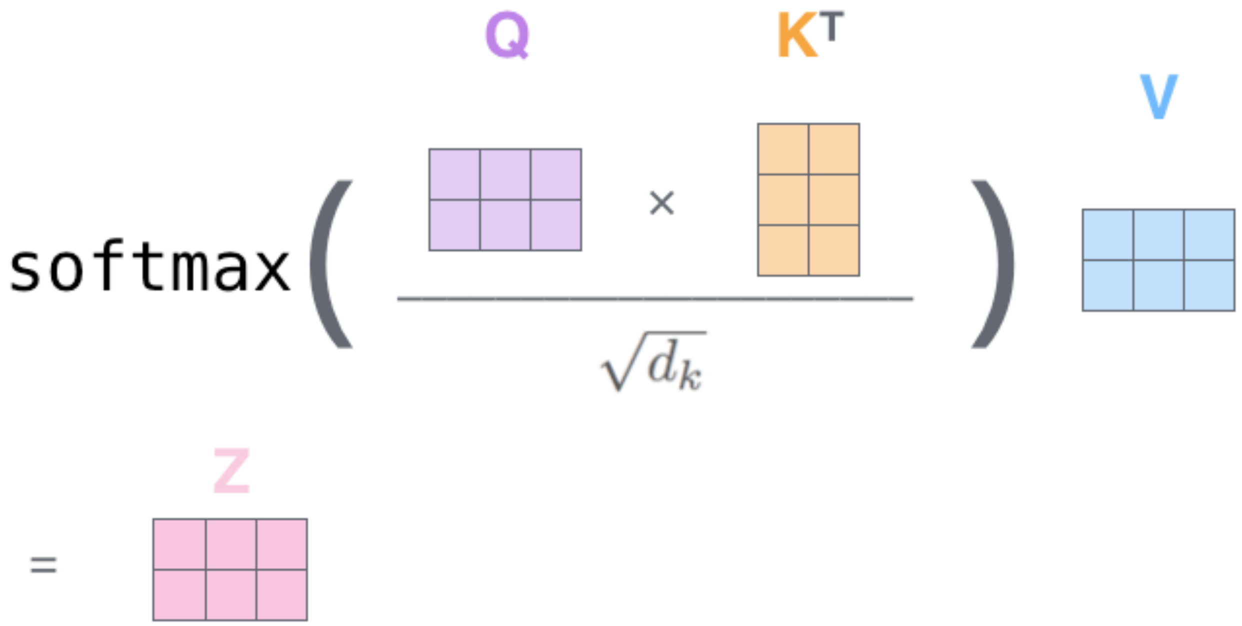
$$\mathbf{q}_i = W_Q \mathbf{x}_i$$

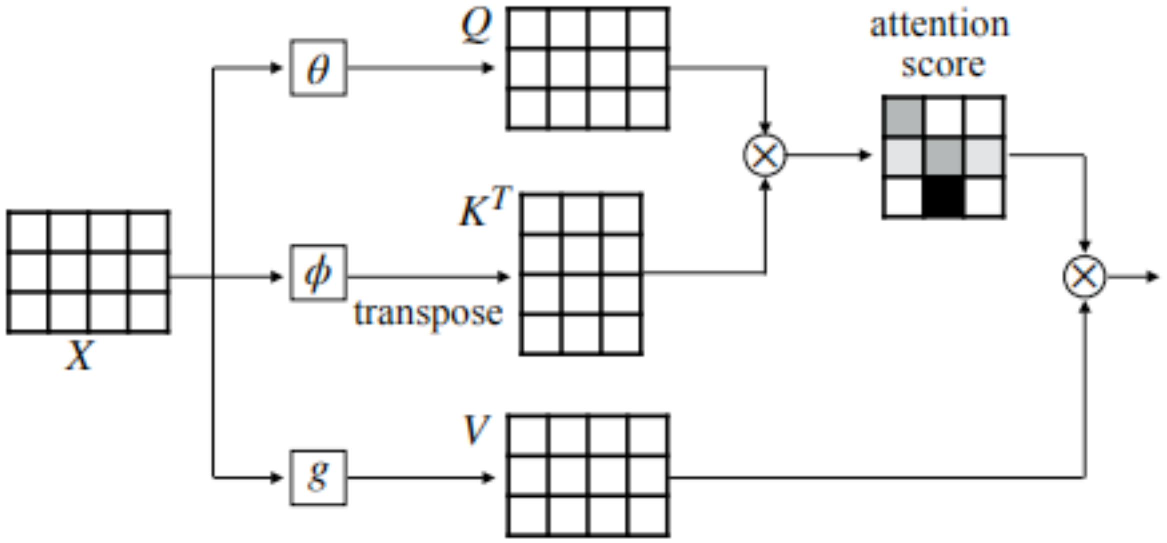
$$\mathbf{k}_i = W_K \mathbf{x}_i$$

$$\mathbf{v}_i = W_V \mathbf{x}_i$$

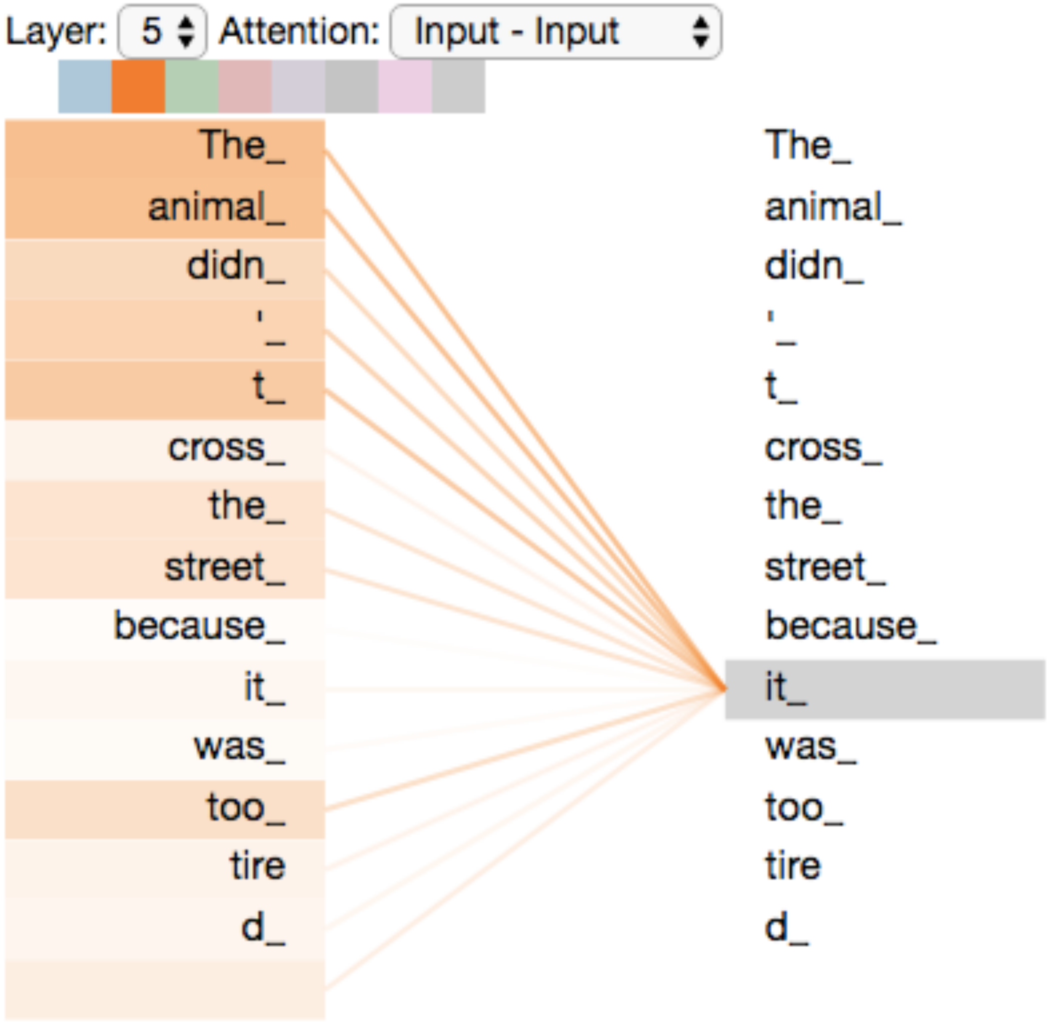
http://jalammar.github.io/illustrated-transformer/

In matrix form





Transformers: self-attention



Layer: 5 Attention: Input - Input

[Jesse Vig, 2019]

Interpretability through attention

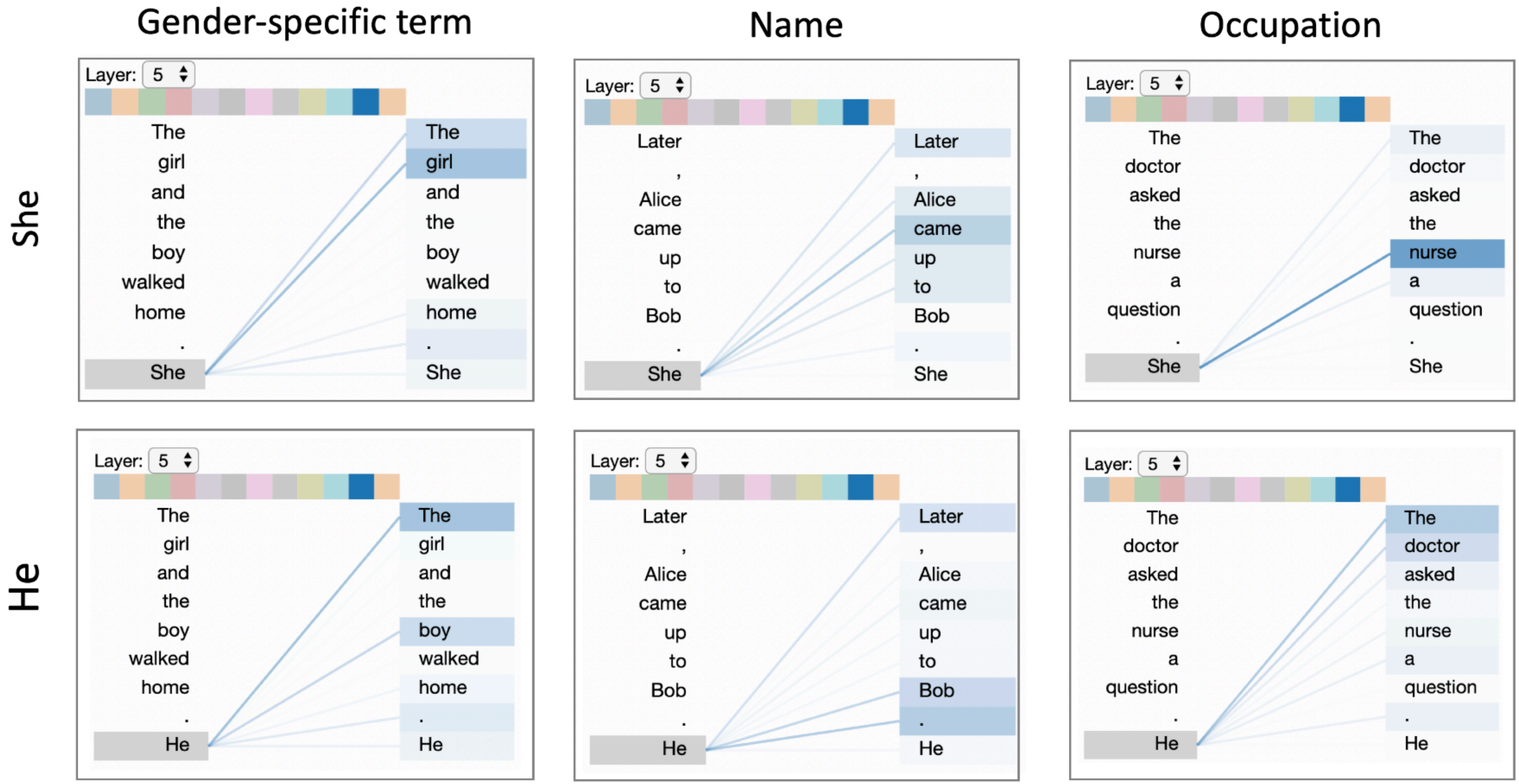
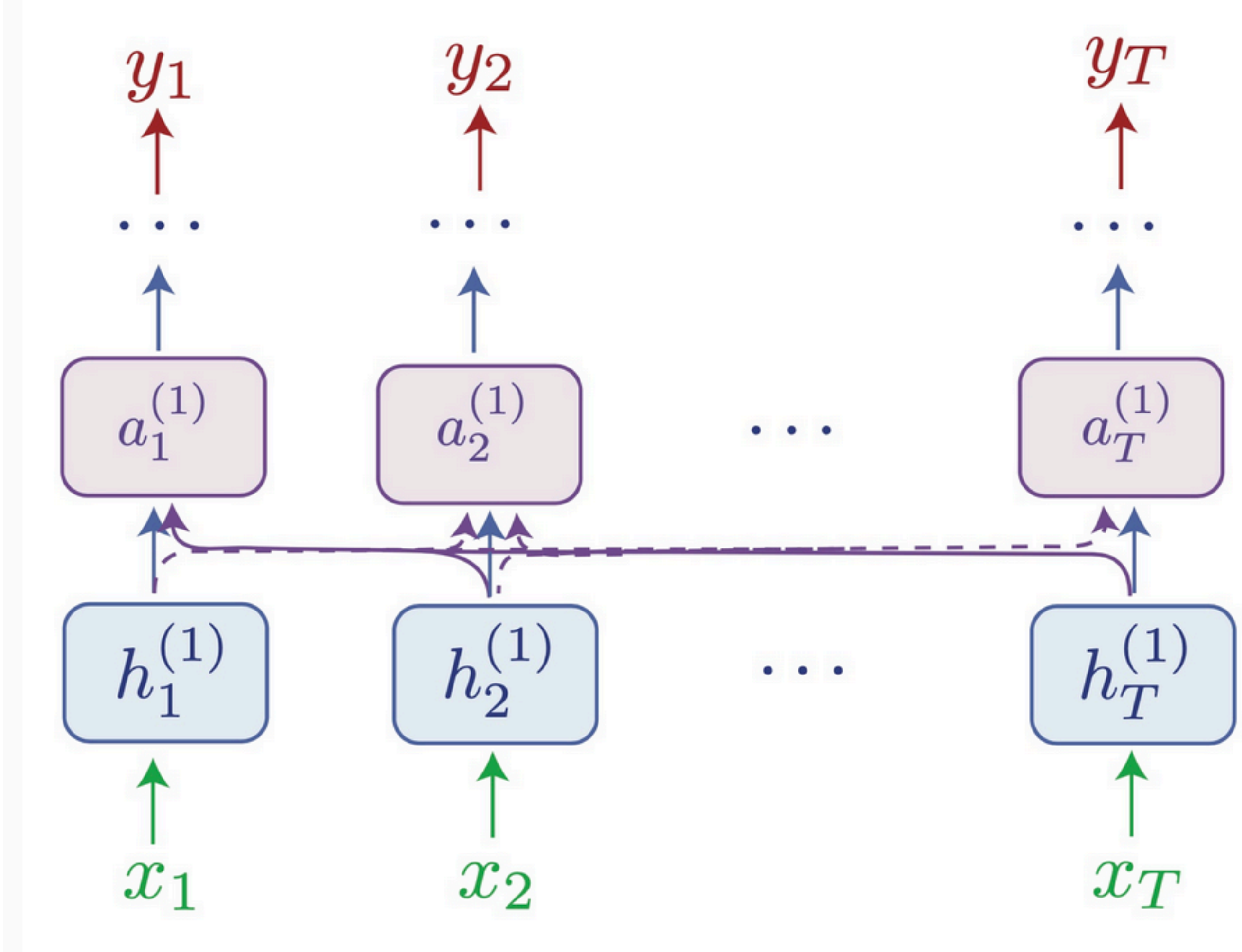

Gender-specific term
Name
Occupation
She
He
Layer: 5
The girl and the boy walked home . She
The girl and the boy walked home . She
Later , Alice came up to Bob . She
Later , Alice came up to Bob . She
The doctor asked the nurse a question . She
The doctor asked the nurse a question . She
Layer: 5
The girl and the boy walked home . He
The girl and the boy walked home . He
Later , Alice came up to Bob . He
Later , Alice came up to Bob . He
The doctor asked the nurse a question . He
The doctor asked the nurse a question . He

Figure 4: Attention pattern in GPT-2 related to coreference resolution suggests the model may encode gender bias.

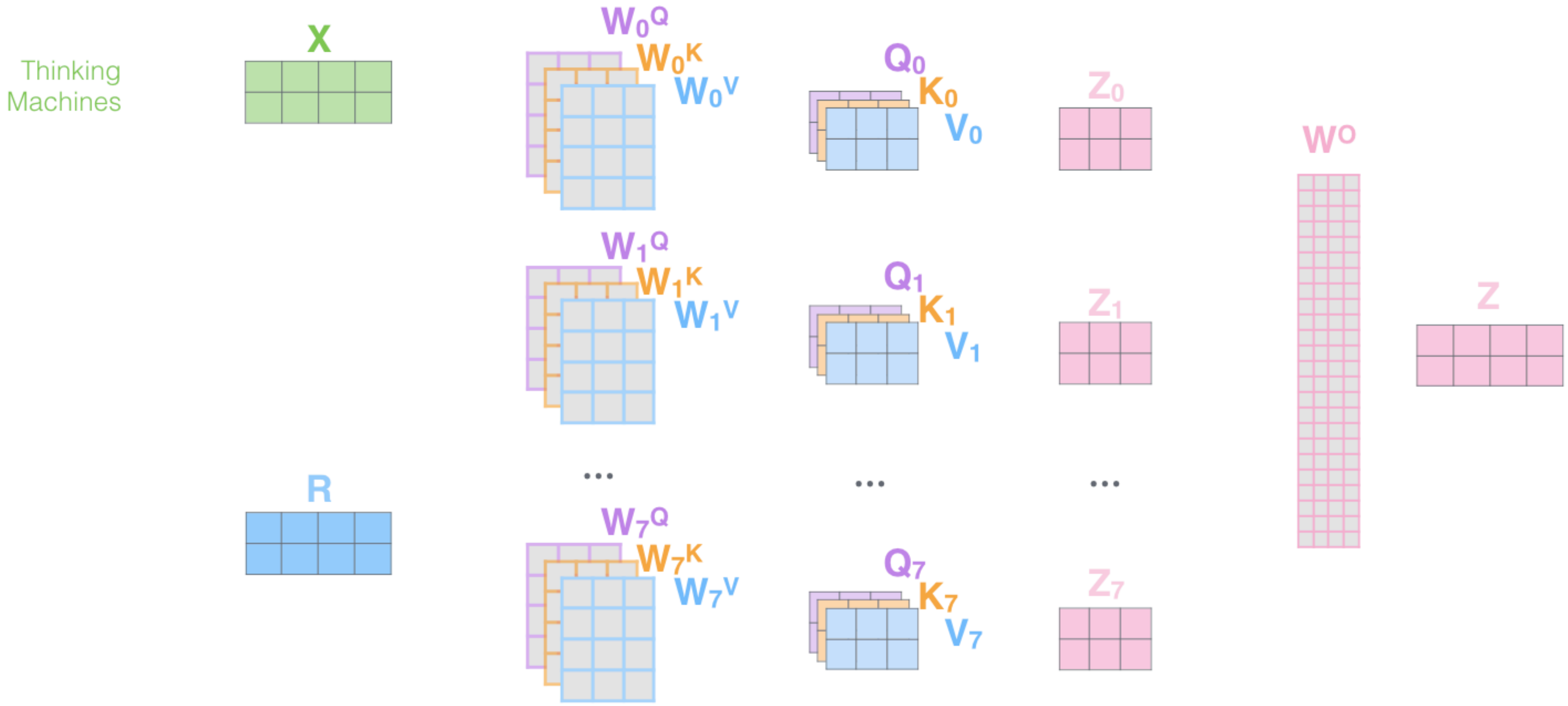
[Jesse Vig, 2019]

Self-attention can process variable-length inputs

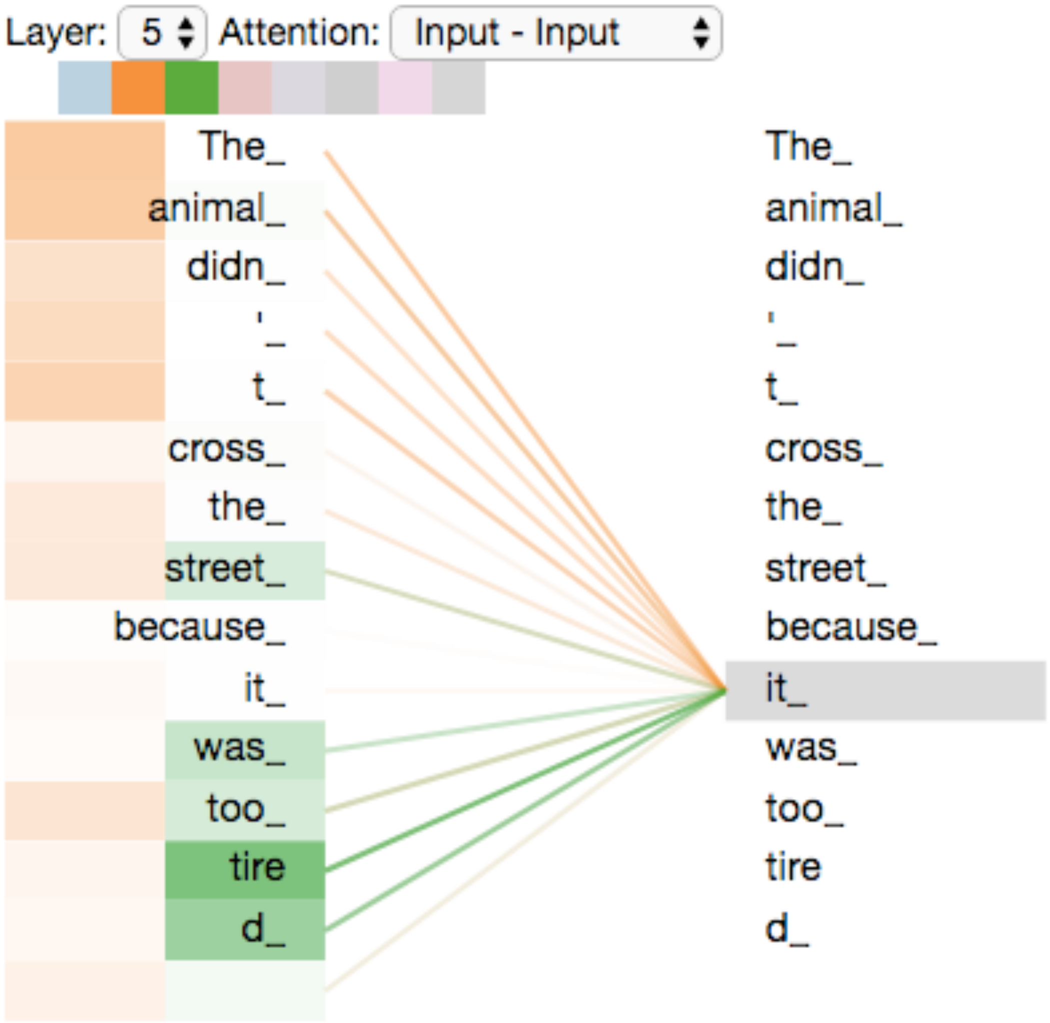


$$a(\cdot, \{\dots\}) : \mathbf{x}_i, \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \rightarrow \mathbf{y}_i$$

Gritty details: multi-head attention



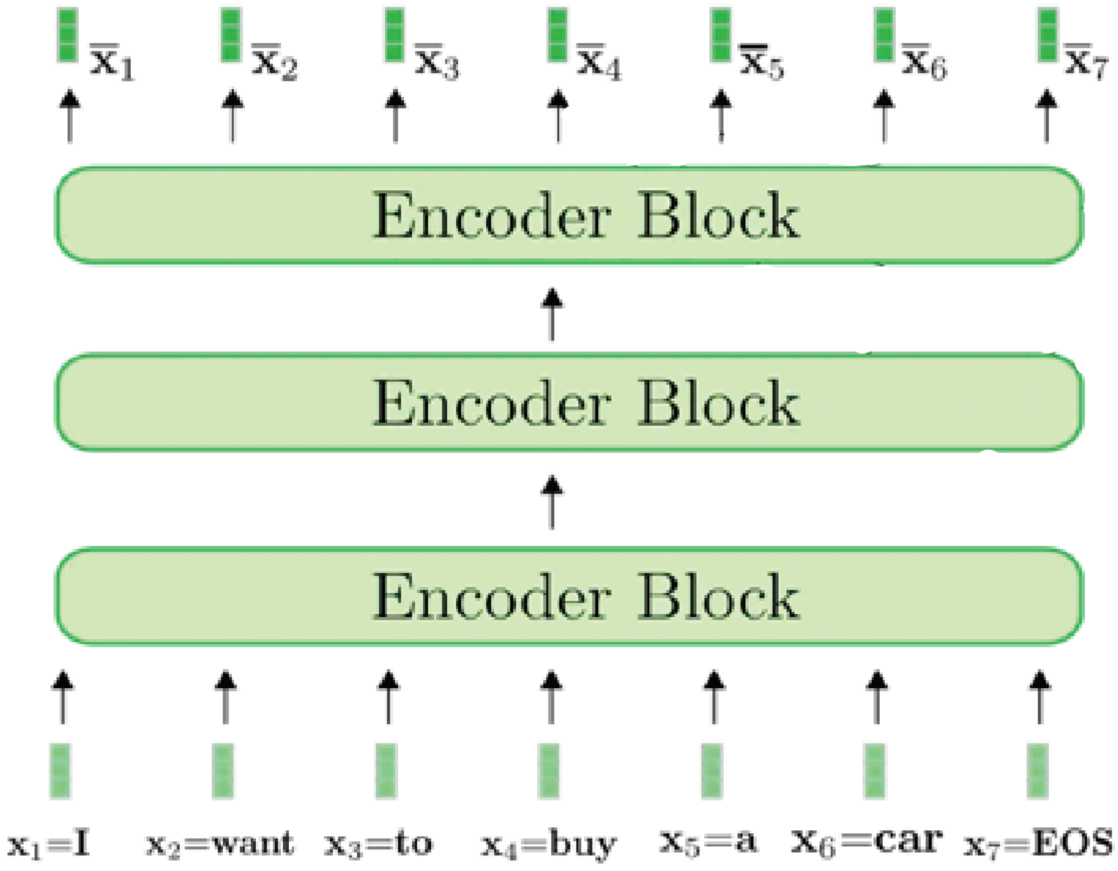
Gritty details: multi-head attention



other issue: information can get mixed across tokens layer by layer, and attention doesn't necessarily represent attention to the corresponding word anymore
(see Darcet et al 2023 for example).

Gritty details: encoder-decoders

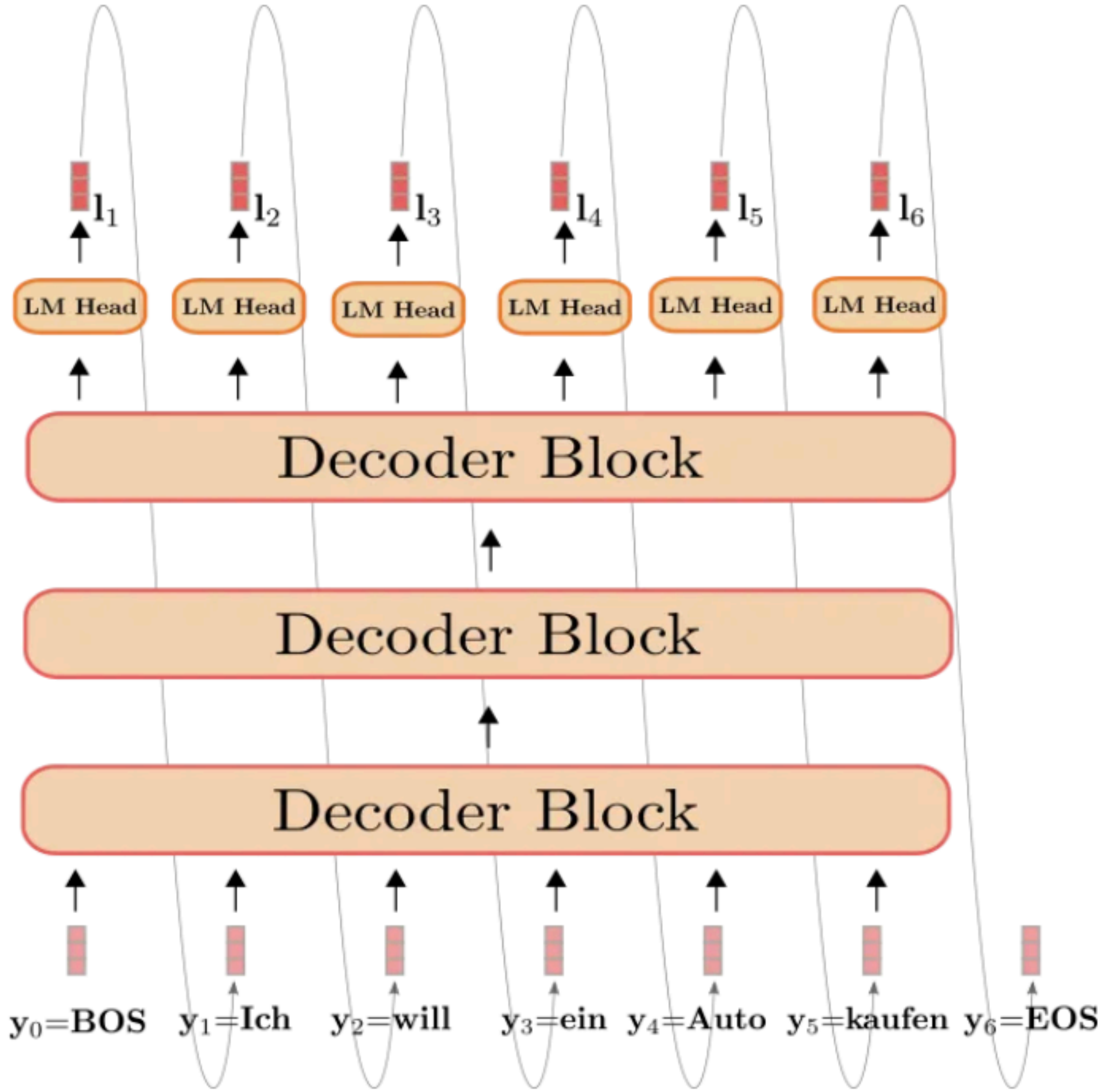
Encoder: e.g. BERT
Typical task: Masked language modelling



https://medium.com/@gabell/encoder-decoder-models-and-transformers-5c1500c22c22

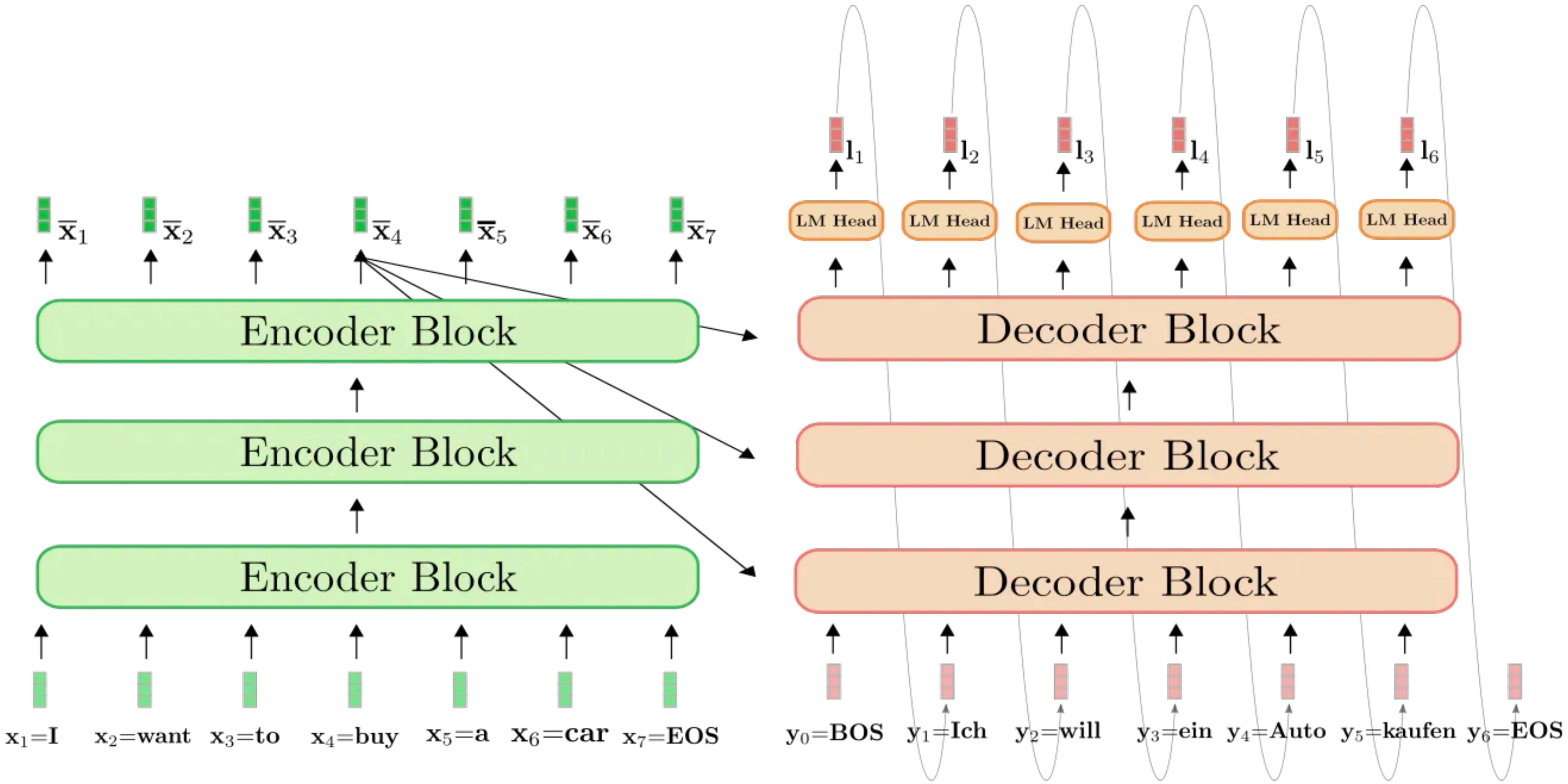
Gritty details: encoder-decoders

Decoder: e.g. GPT
Typical task: Autoregressive language modelling

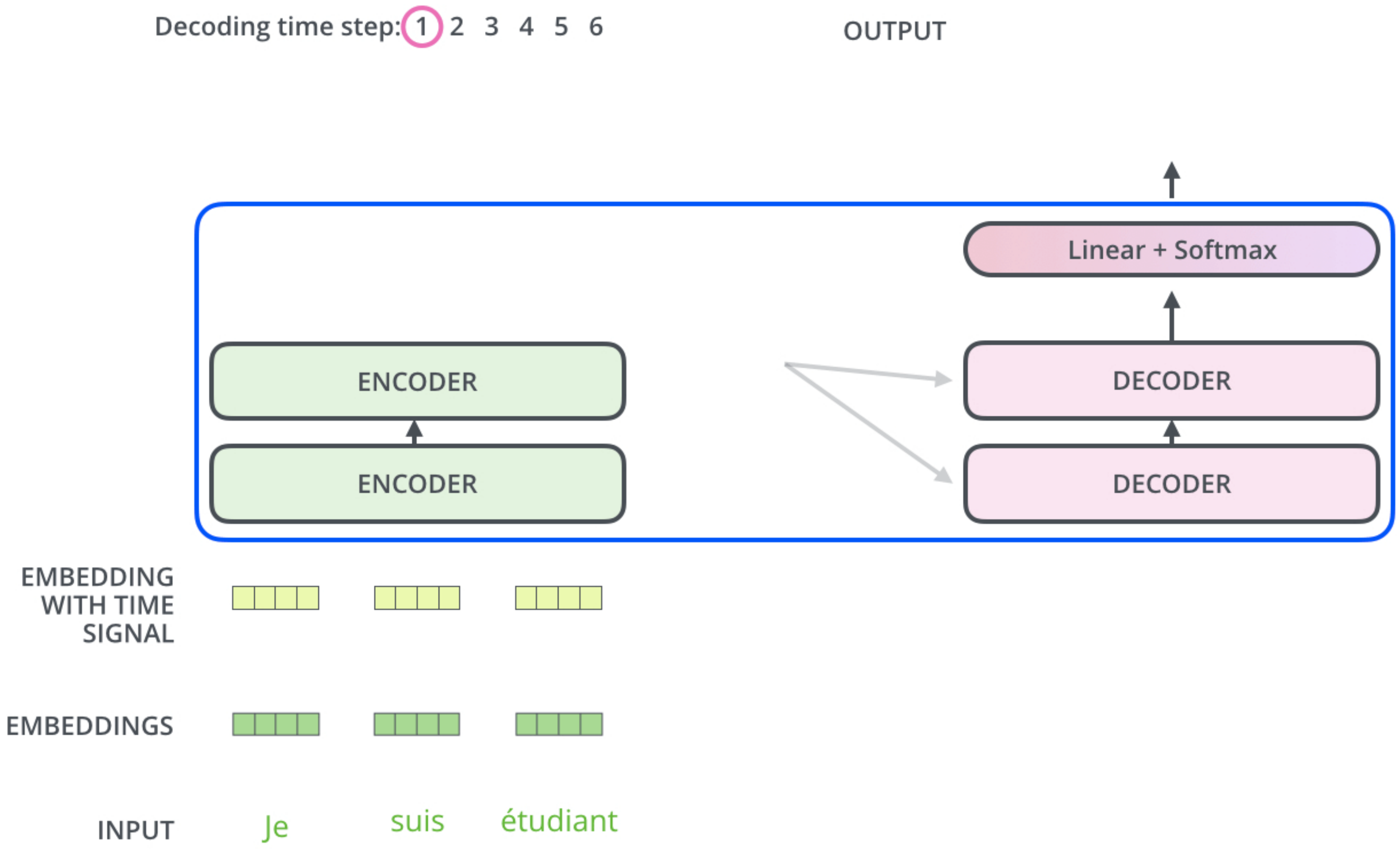


Gritty details: encoder-decoders

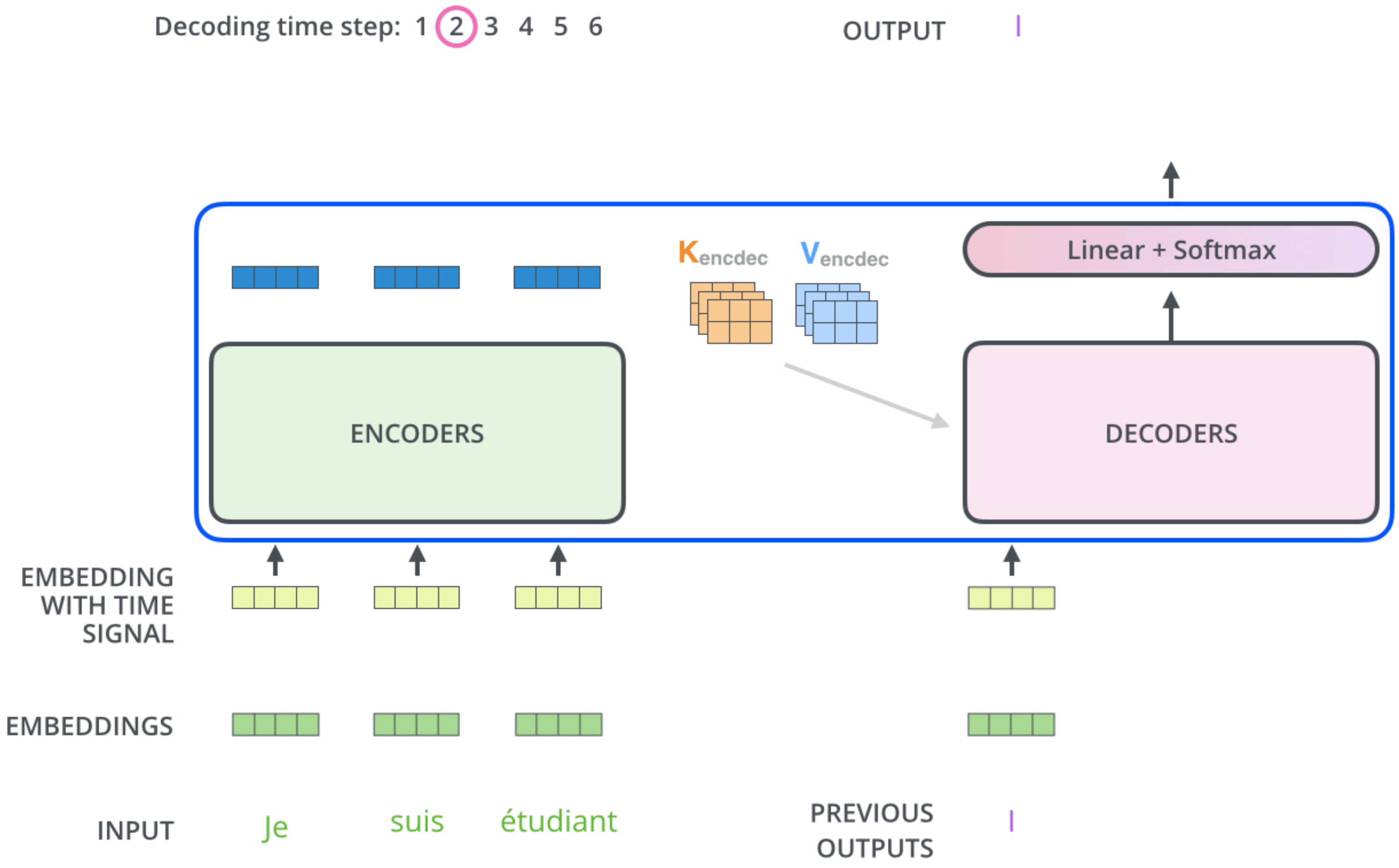
Encoder-decoder: e.g. T5
Typical task: Sequence-to-sequence modelling



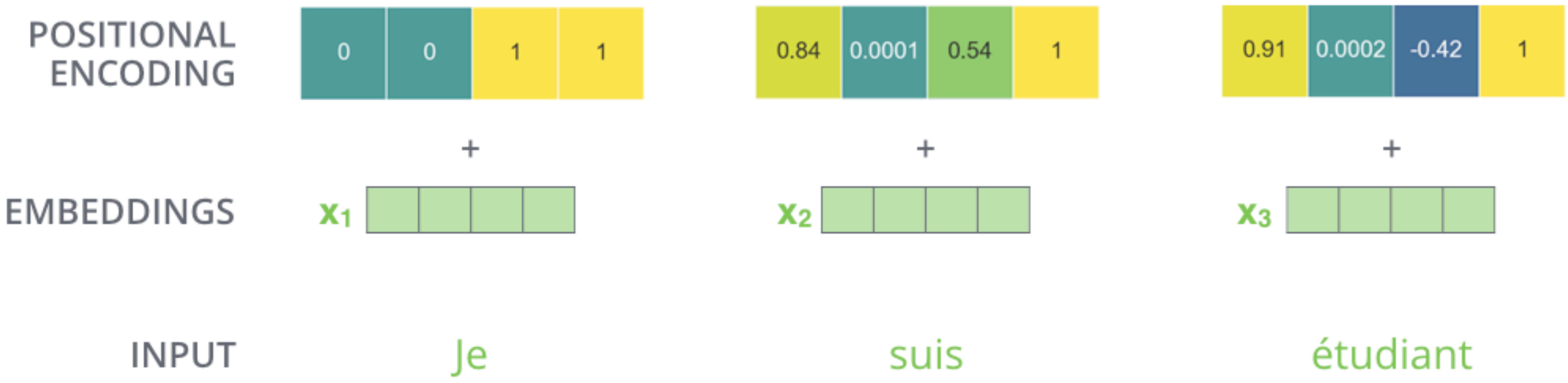
Gritty details: encoder-decoders



Gritty details: encoder-decoders



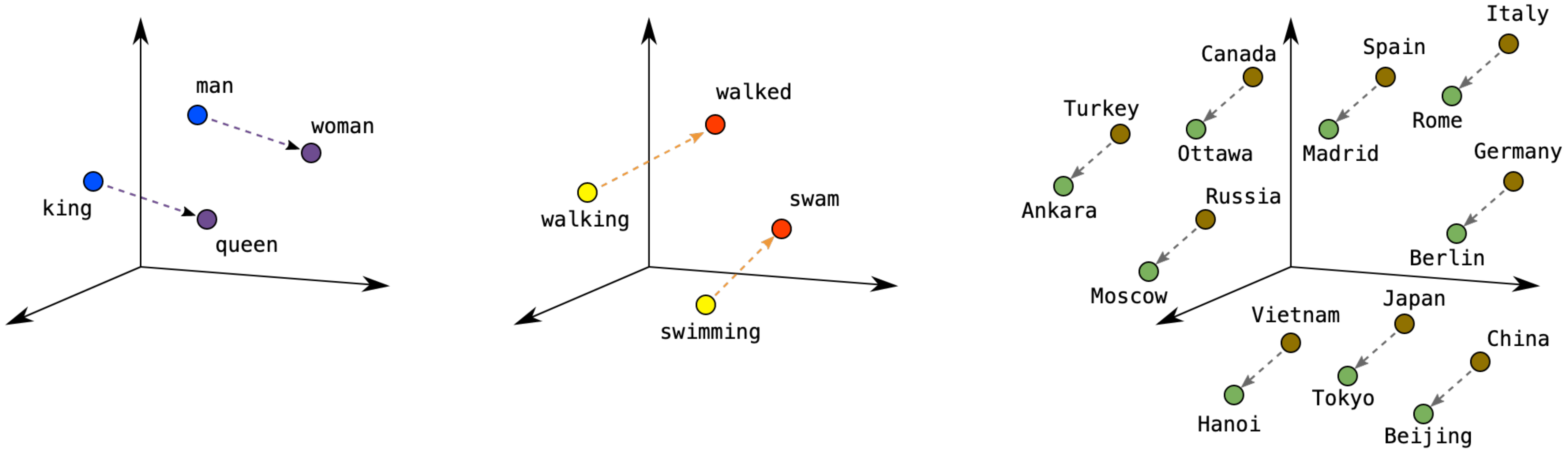
Gritty details: positional embeddings



$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

FF layers & attention

How to make sense?
Embeddings



Male-Female

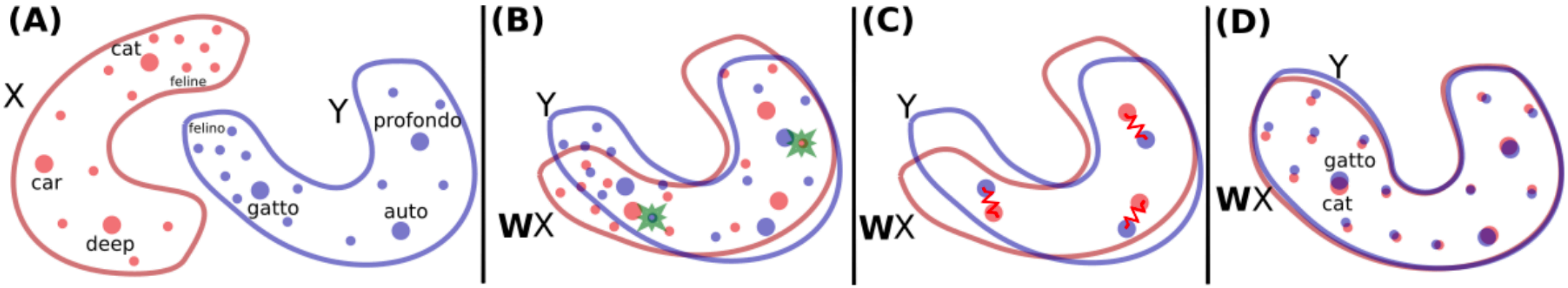
Verb Tense

Country-Capital

Semantic geometrical encoding

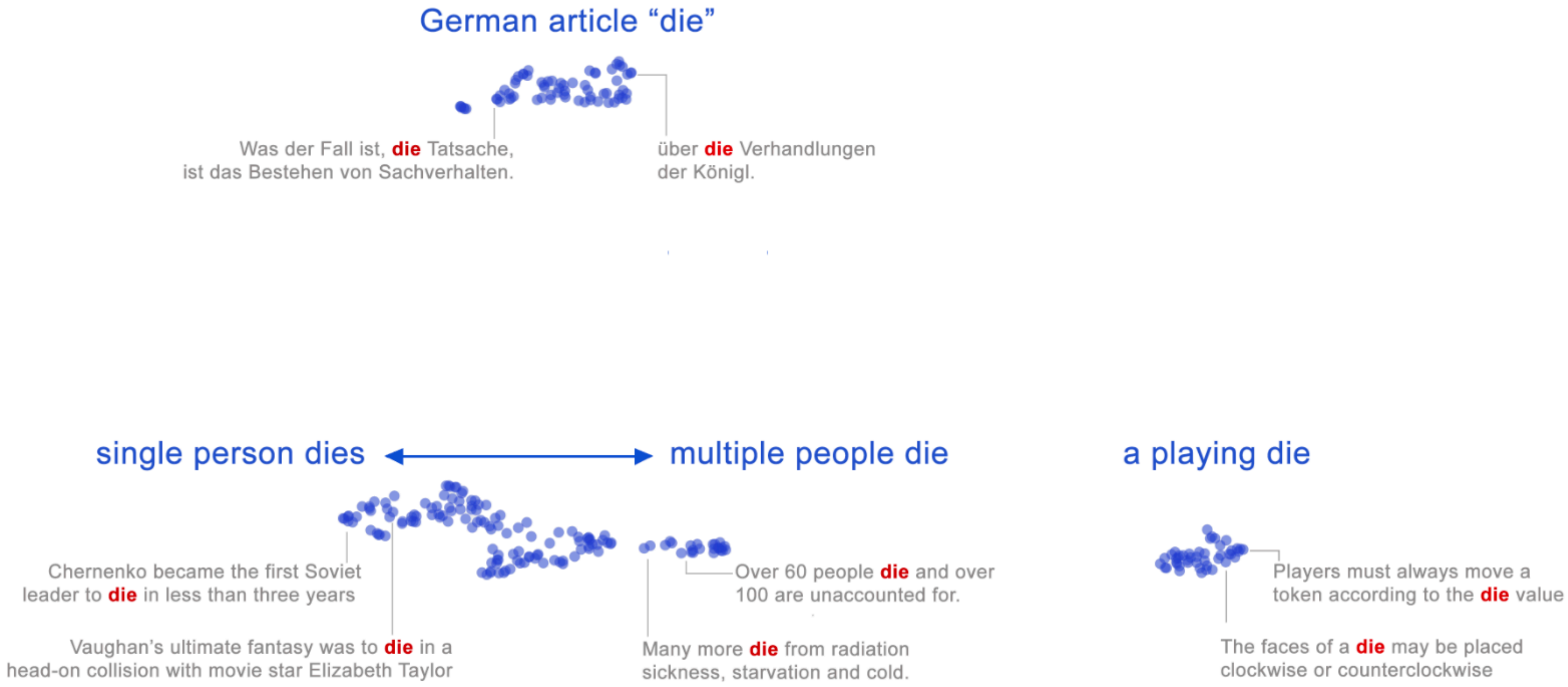
[eg. Mikolov et al. 2013]

How to make sense?
Embeddings



[Lample et al. 2018, Word translation without parallel data]

Context-dependent embeddings (BERT)



Was der Fall ist, **die** Tatsache, ist das Bestehen von Sachverhalten.

über **die** Verhandlungen der Königl.

Chernenko became the first Soviet leader to **die** in less than three years

Vaughan's ultimate fantasy was to **die** in a head-on collision with movie star Elizabeth Taylor

Over 60 people **die** and over 100 are unaccounted for.

Many more **die** from radiation sickness, starvation and cold.

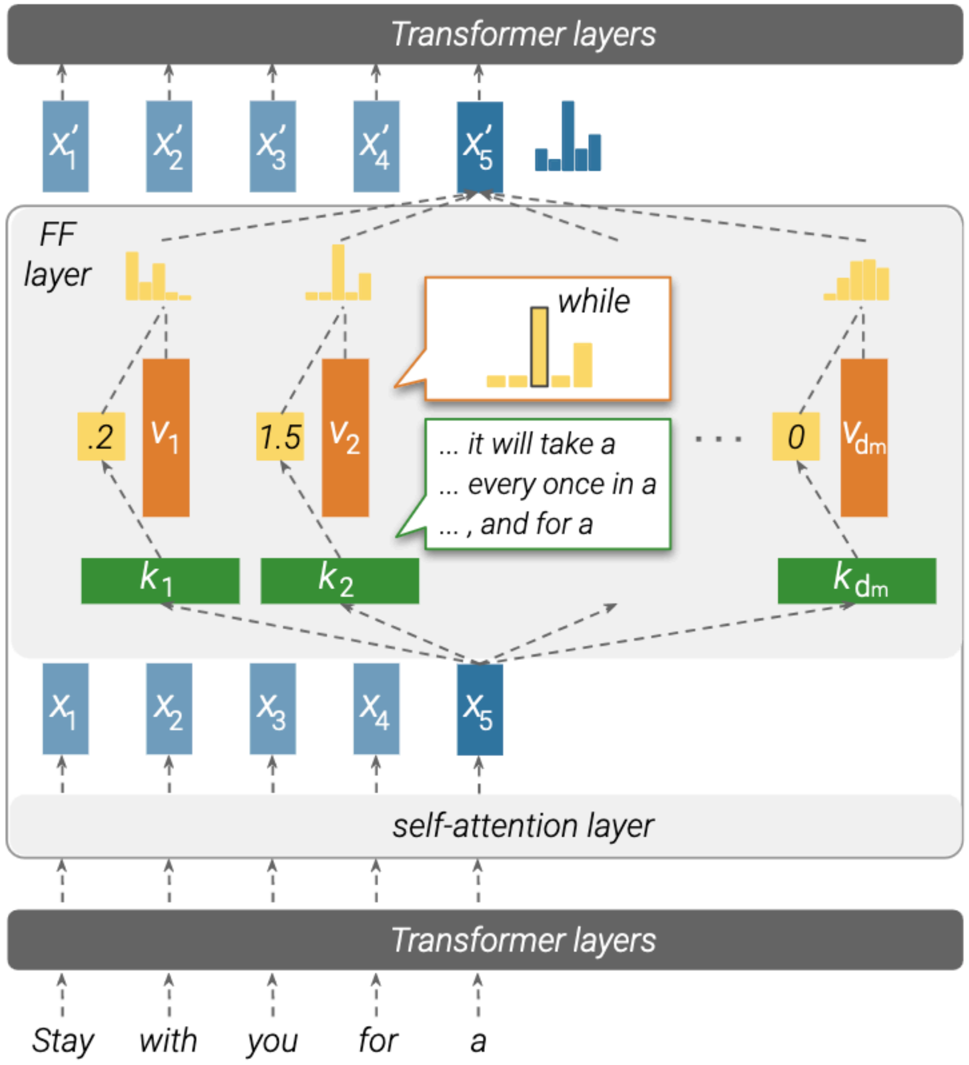
Players must always move a token according to the **die** value

The faces of a **die** may be placed clockwise or counterclockwise

Semantic geometrical encoding

[Coenen et al., 2019]

FF layers as key-value stores

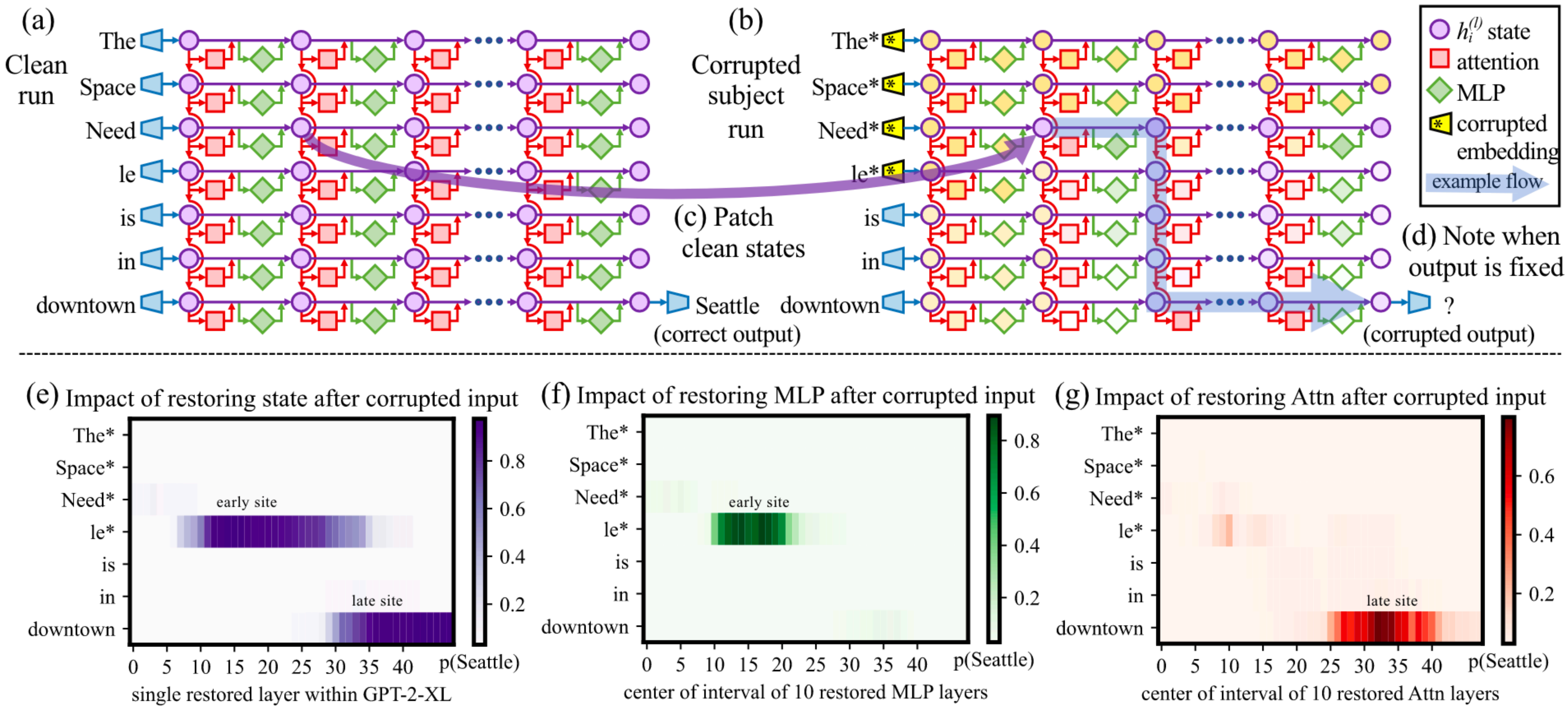


$$p(k_i \mid x) \propto \exp(\mathbf{x} \cdot \mathbf{k}_i)$$

$$\mathrm{MN}(\mathbf{x}) = \sum_{i=1}^{d_m} p(k_i \mid x)\mathbf{v}_i$$

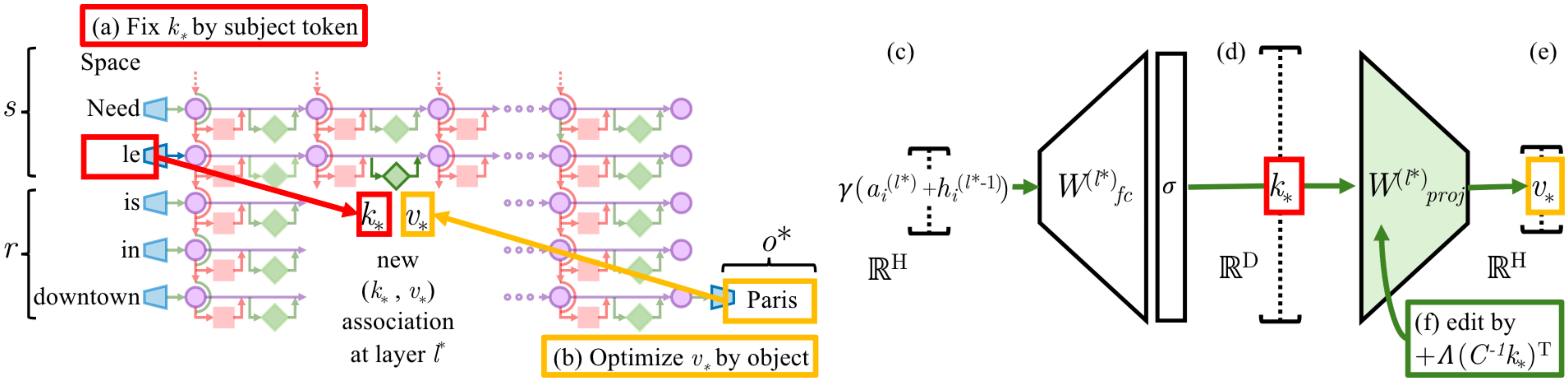
[Geva et al., 2021]

Causal scrubbing


(a) Clean run
The
Space
Need
le
is
in
downtown
Seattle
(correct output)
(b) Corrupted subject run
The*
Space*
Need*
le*
is
in
downtown
?
(corrupted output)
(c) Patch clean states
(d) Note when output is fixed
$h_i^{(l)}$ state
attention
MLP
corrupted embedding
example flow
(e) Impact of restoring state after corrupted input
early site
late site
single restored layer within GPT-2-XL
p(Seattle)
(f) Impact of restoring MLP after corrupted input
center of interval of 10 restored MLP layers
(g) Impact of restoring Attn after corrupted input
center of interval of 10 restored Attn layers
0 5 10 15 20 25 30 35 40
0.2 0.4 0.6 0.8

[Meng et al., 2022]

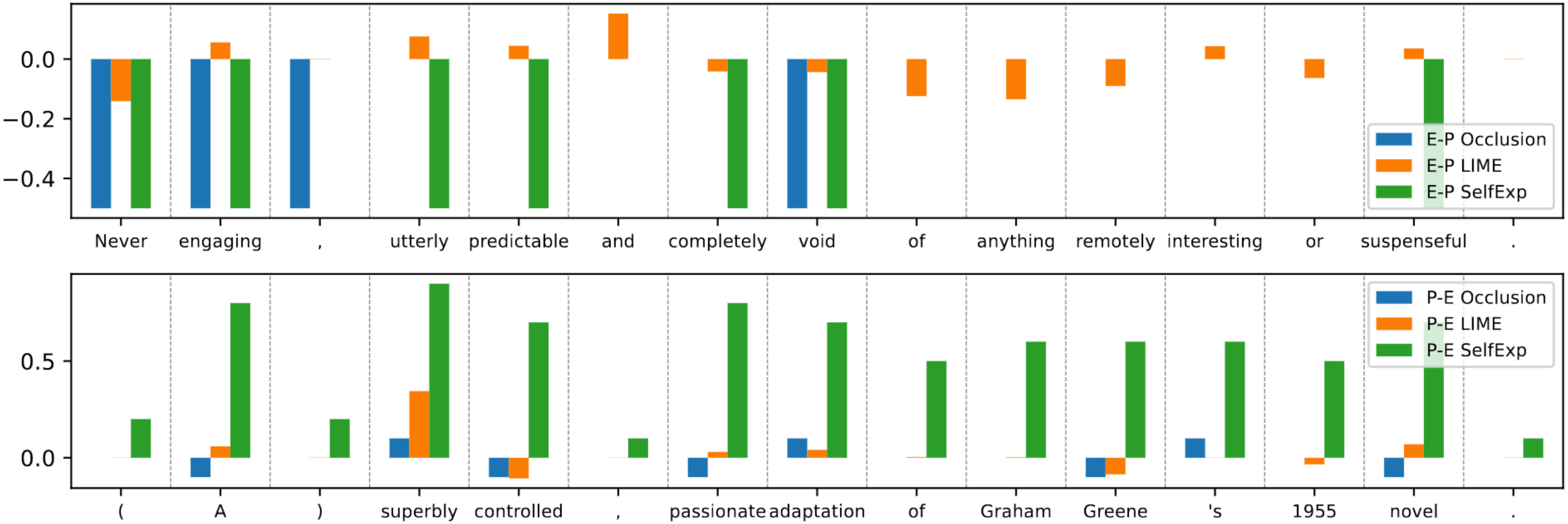
Manipulating key-value associations



[Meng et al., 2022]

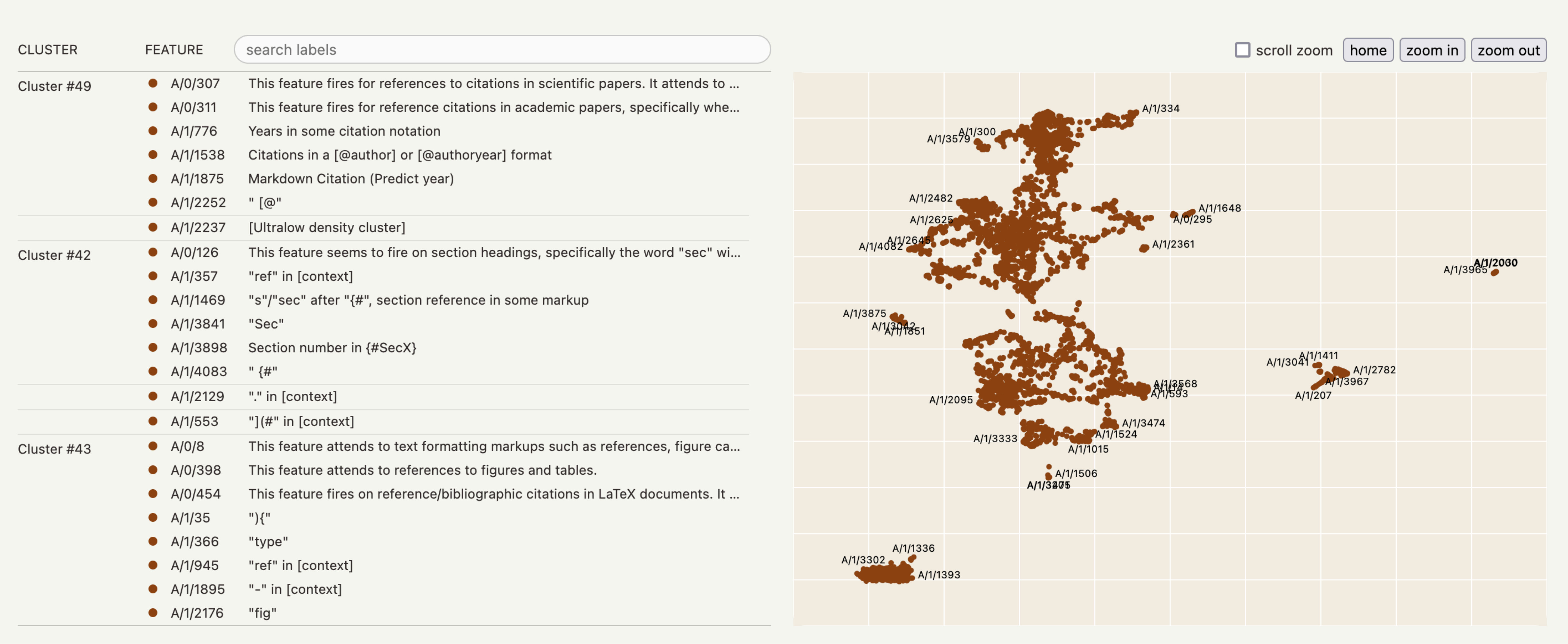
Just ask the model??

Role	Text
User	Can you tell me whether the movie review is positive or negative? The gorgeously elaborate continuation of "The Lord of the Rings" trilogy is so huge that a column of words can not adequately describe co-writer/director Peter Jackson ’s expanded vision of J.R.R. Tolkien ’s Middle-earth .
System	Based on the given text, it appears to be a positive movie review. The phrases "gorgeously elaborate," "expanded vision," and "Middle-earth" suggest admiration and appreciation for the film.



[Huang et al., 2023]

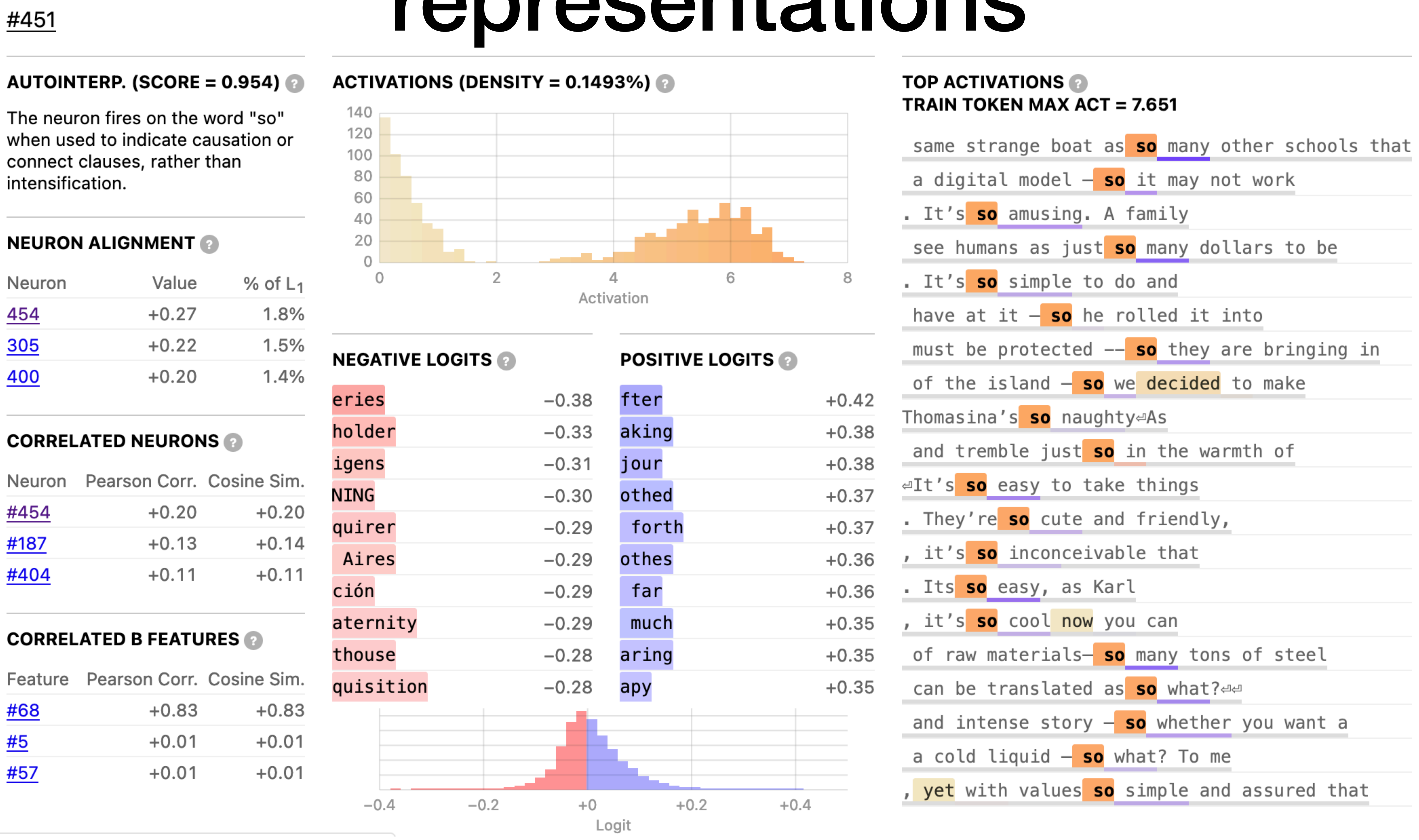
Inner workings: overcomplete representations



CLUSTER	FEATURE	
Cluster #49	● A/0/307	This feature fires for references to citations in scientific papers. It attends to ...
	● A/0/311	This feature fires for reference citations in academic papers, specifically whe...
	● A/1/776	Years in some citation notation
	● A/1/1538	Citations in a [@author] or [@authoryear] format
	● A/1/1875	Markdown Citation (Predict year)
	● A/1/2252	" [@"
	● A/1/2237	[Ultralow density cluster]
Cluster #42	● A/0/126	This feature seems to fire on section headings, specifically the word "sec" wi...
	● A/1/357	"ref" in [context]
	● A/1/1469	"s"/"sec" after "{#", section reference in some markup
	● A/1/3841	"Sec"
	● A/1/3898	Section number in {#SecX}
	● A/1/4083	" {#"
	● A/1/2129	"." in [context]
	● A/1/553	"](#" in [context]
Cluster #43	● A/0/8	This feature attends to text formatting markups such as references, figure ca...
	● A/0/398	This feature attends to references to figures and tables.
	● A/0/454	This feature fires on reference/bibliographic citations in LaTeX documents. It ...
	● A/1/35	"){"
	● A/1/366	"type"
	● A/1/945	"ref" in [context]
	● A/1/1895	"-" in [context]
	● A/1/2176	"fig"

[Bricken et al., 2023]

Inner workings: overcomplete representations



#451

AUTOINTERP. (SCORE = 0.954) ?

The neuron fires on the word "so" when used to indicate causation or connect clauses, rather than intensification.

NEURON ALIGNMENT ?

Neuron	Value	% of L_1
454	+0.27	1.8%
305	+0.22	1.5%
400	+0.20	1.4%

CORRELATED NEURONS ?

Neuron	Pearson Corr.	Cosine Sim.
#454	+0.20	+0.20
#187	+0.13	+0.14
#404	+0.11	+0.11

CORRELATED B FEATURES ?

Feature	Pearson Corr.	Cosine Sim.
#68	+0.83	+0.83
#5	+0.01	+0.01
#57	+0.01	+0.01

ACTIVATIONS (DENSITY = 0.1493%) ?

NEGATIVE LOGITS ?

eries	−0.38
holder	−0.33
igens	−0.31
NING	−0.30
quirer	−0.29
Aires	−0.29
ción	−0.29
aternity	−0.29
thouse	−0.28
quisition	−0.28

POSITIVE LOGITS ?

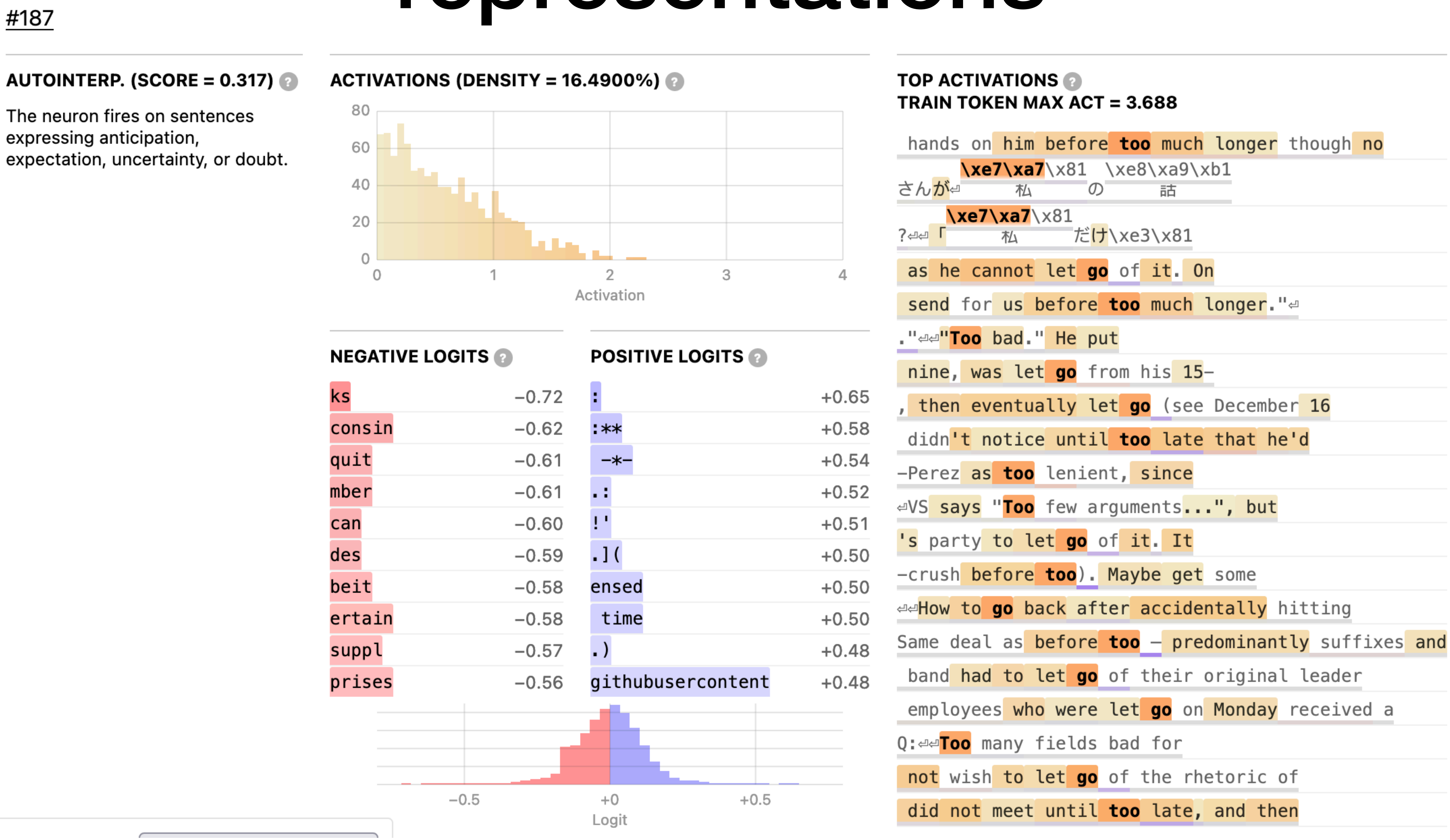
fter	+0.42
aking	+0.38
jour	+0.38
othed	+0.37
forth	+0.37
othes	+0.36
far	+0.36
much	+0.35
aring	+0.35
apy	+0.35

TOP ACTIVATIONS ?

TRAIN TOKEN MAX ACT = 7.651

same strange boat as **so** many other schools that
a digital model — **so** it may not work
. It’s **so** amusing. A family
see humans as just **so** many dollars to be
. It’s **so** simple to do and
have at it — **so** he rolled it into
must be protected -- **so** they are bringing in
of the island — **so** we decided to make
Thomasina’s **so** naughty⏎As
and tremble just **so** in the warmth of
⏎It’s **so** easy to take things
. They’re **so** cute and friendly,
, it’s **so** inconceivable that
. Its **so** easy, as Karl
, it’s **so** cool now you can
of raw materials— **so** many tons of steel
can be translated as **so** what?⏎⏎
and intense story — **so** whether you want a
a cold liquid — **so** what? To me
, yet with values **so** simple and assured that

Inner workings: overcomplete representations



#187

AUTOINTERP. (SCORE = 0.317) ?

The neuron fires on sentences expressing anticipation, expectation, uncertainty, or doubt.

ACTIVATIONS (DENSITY = 16.4900%) ?

NEGATIVE LOGITS ?

ks	−0.72
consin	−0.62
quit	−0.61
mber	−0.61
can	−0.60
des	−0.59
beit	−0.58
ertain	−0.58
suppl	−0.57
prises	−0.56

POSITIVE LOGITS ?

:	+0.65
:**	+0.58
-*-	+0.54
.:	+0.52
!'	+0.51
.](	+0.50
ensed	+0.50
time	+0.50
.)	+0.48
githubusercontent	+0.48

TOP ACTIVATIONS ?
TRAIN TOKEN MAX ACT = 3.688

hands on him before **too** much longer though no

さんが⏎ **\xe7\xa7**\x81 私 \xe8\xa9\xb1 の 話

?⏎⏎「 **\xe7\xa7**\x81 私 だけ\xe3\x81

as he cannot let **go** of it. On

send for us before **too** much longer."⏎

."⏎⏎"**Too** bad." He put

nine, was let **go** from his 15-

, then eventually let **go** (see December 16

didn't notice until **too** late that he'd

-Perez as **too** lenient, since

⏎VS says "**Too** few arguments...", but

's party to let **go** of it. It

-crush before **too**). Maybe get some

⏎⏎How to **go** back after accidentally hitting

Same deal as before **too** – predominantly suffixes and

band had to let **go** of their original leader

employees who were let **go** on Monday received a

Q:⏎⏎**Too** many fields bad for

not wish to let **go** of the rhetoric of

did not meet until **too** late, and then