Текстовые эмбеддинги
простые и сложные

Штех Геннадий
NAUMEN
04.05.2019

#МИСиС

https://github.com/ShT3ch/public_workshop

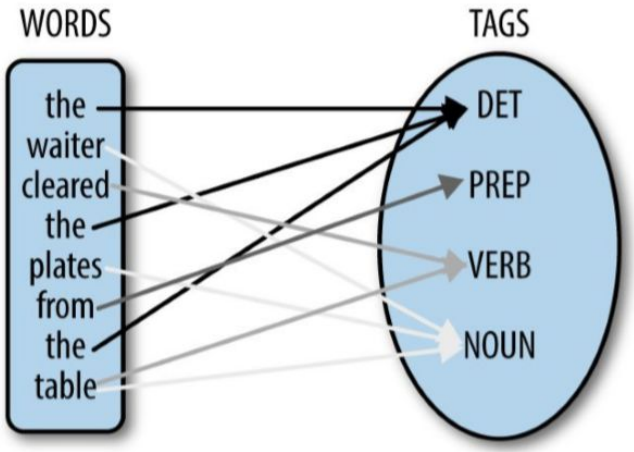
О чем поговорим

1. О задачах NLP
2. Об эмбеддингах в целом
3. RNN, Transformer, BERT
4. Хронология развития технологий
5. Как оценивают “разумность” моделей

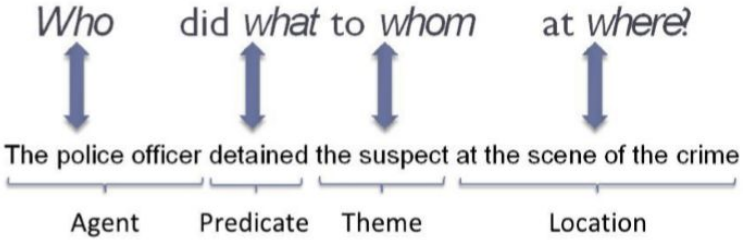
Какие задачи решаем

Какие бывают задачи про тексты? Классика

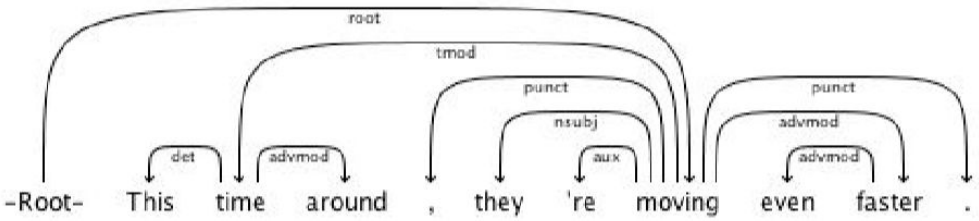
Классификация частей речи



Выделение семантических ролей



Парсинг зависимостей



Заполнение форм

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Поиск

Исправление опечаток (fasttext)

wann, wanto, wanr, wany	→	want
havea, havr	→	have
thiss, thise	→	this
pleasee, pleasr, pleasw, pleaseee	→	please
numbe, numbet, numbee, numbr	→	number
calll	→	call
willl, wiill	→	will

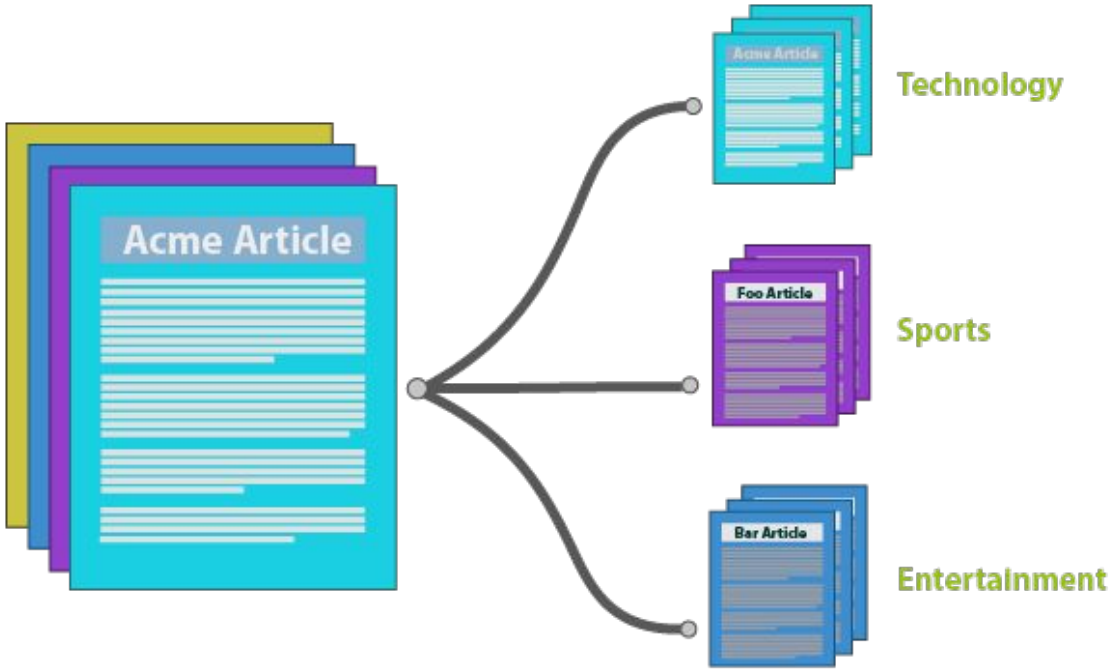
Какие бывают задачи про тексты?
RnD

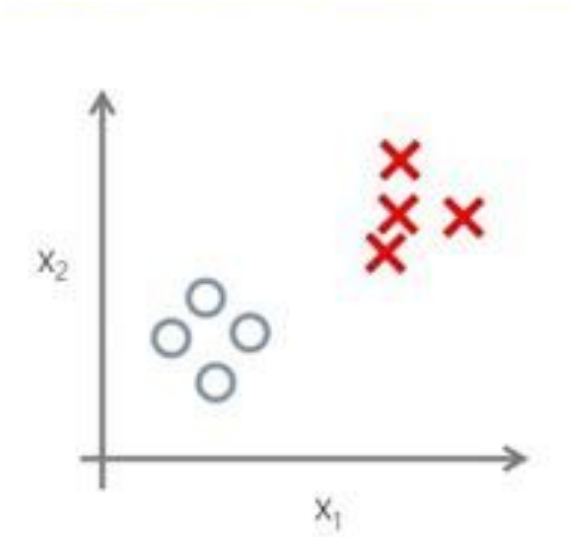
Голосовые помощники

Переводчики языков и стилей

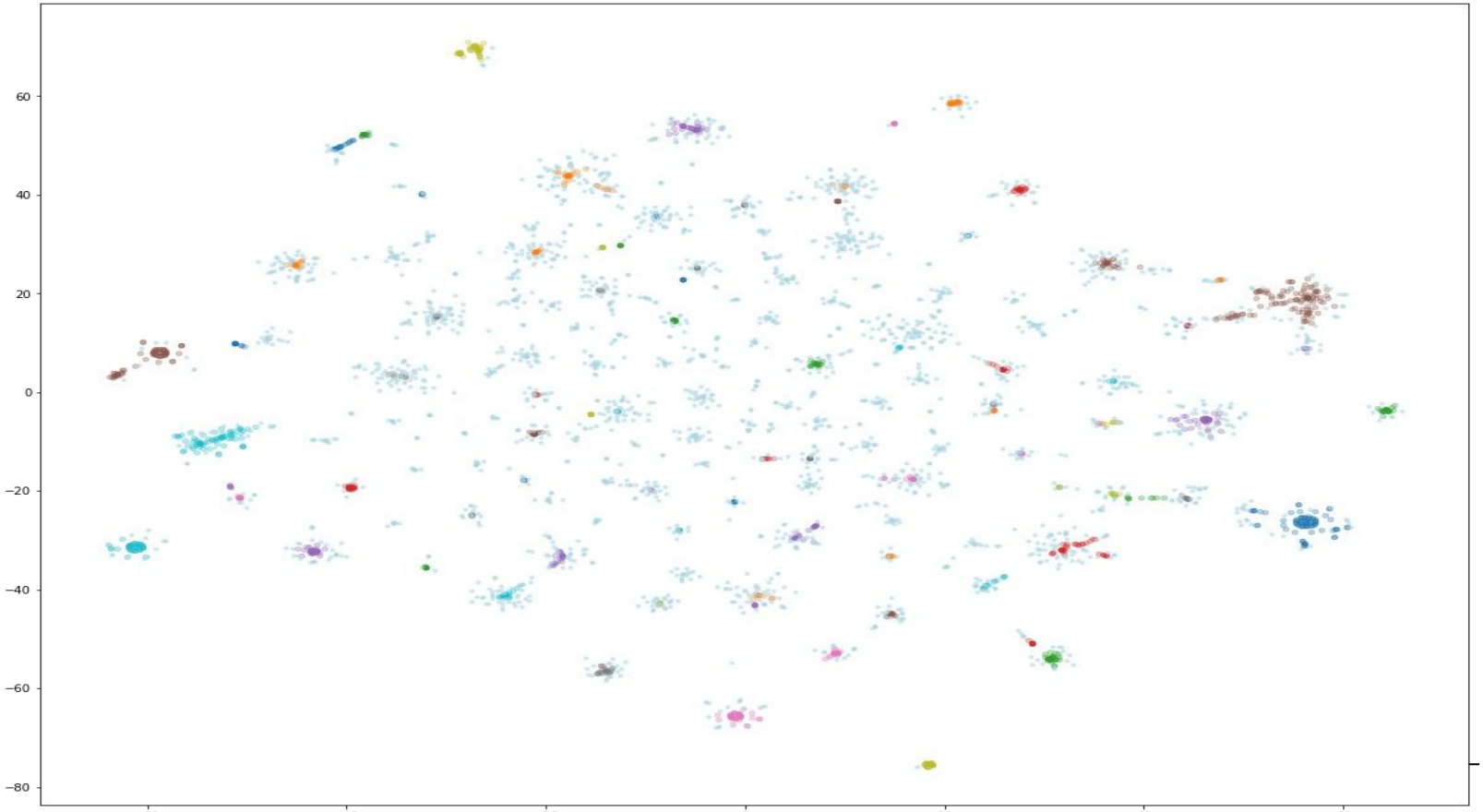
Кросс-эмбединговые модели

Классификация



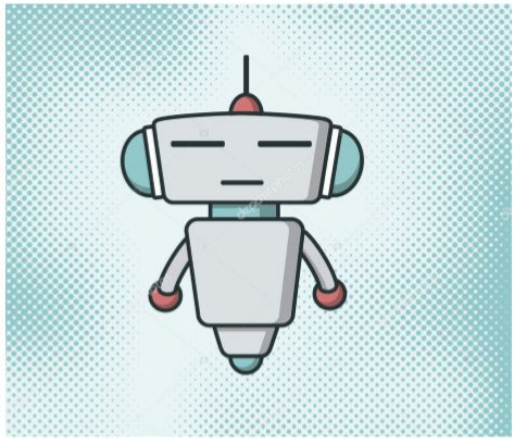


Кластеризация

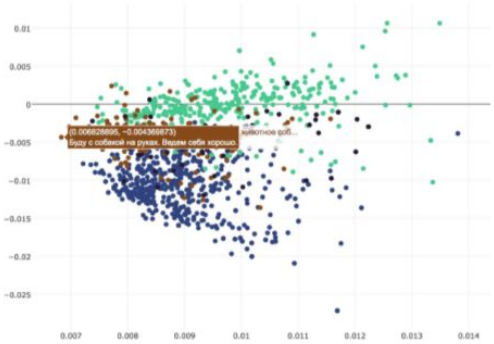


Какие бывают задачи про тексты?
Бизнес

Автоматические ответы в поддержке



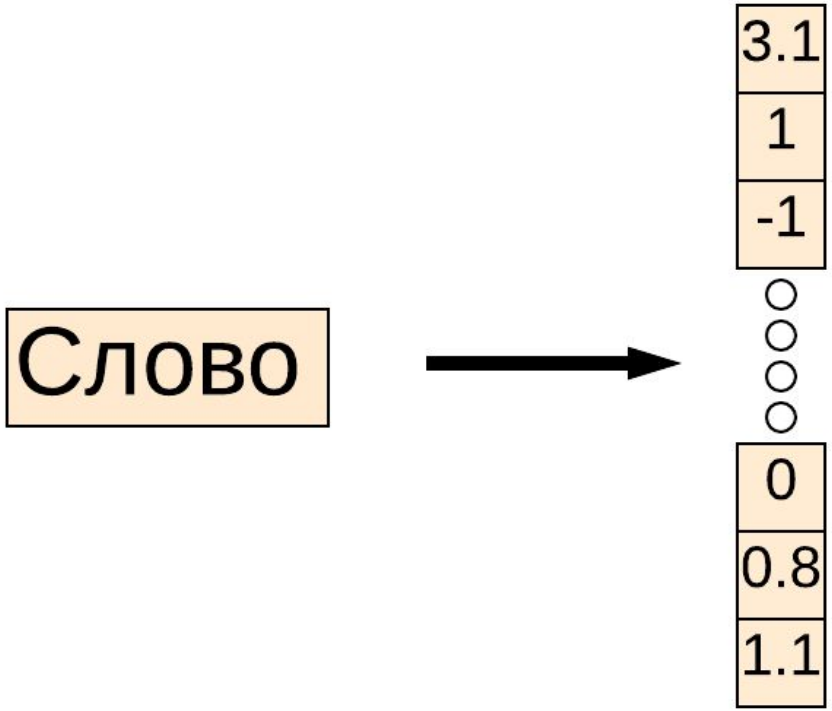
Выделение аномалий в текстах



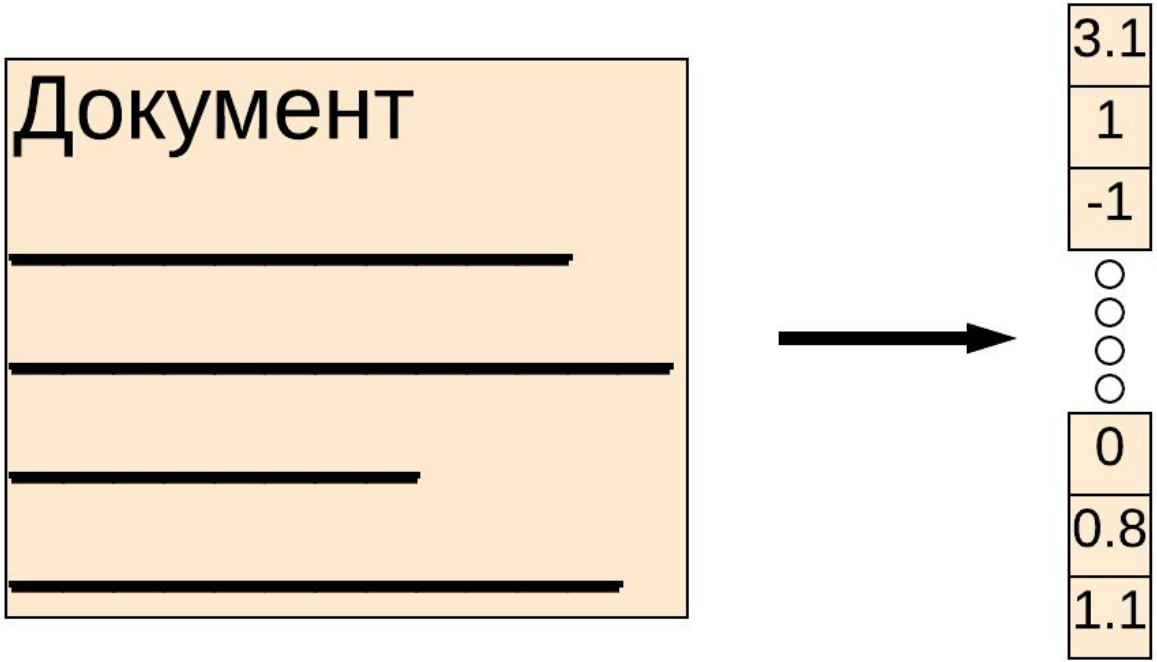
Понимать, на какие сообщения в соцсетях реагировать

Что такое эмбеддинги

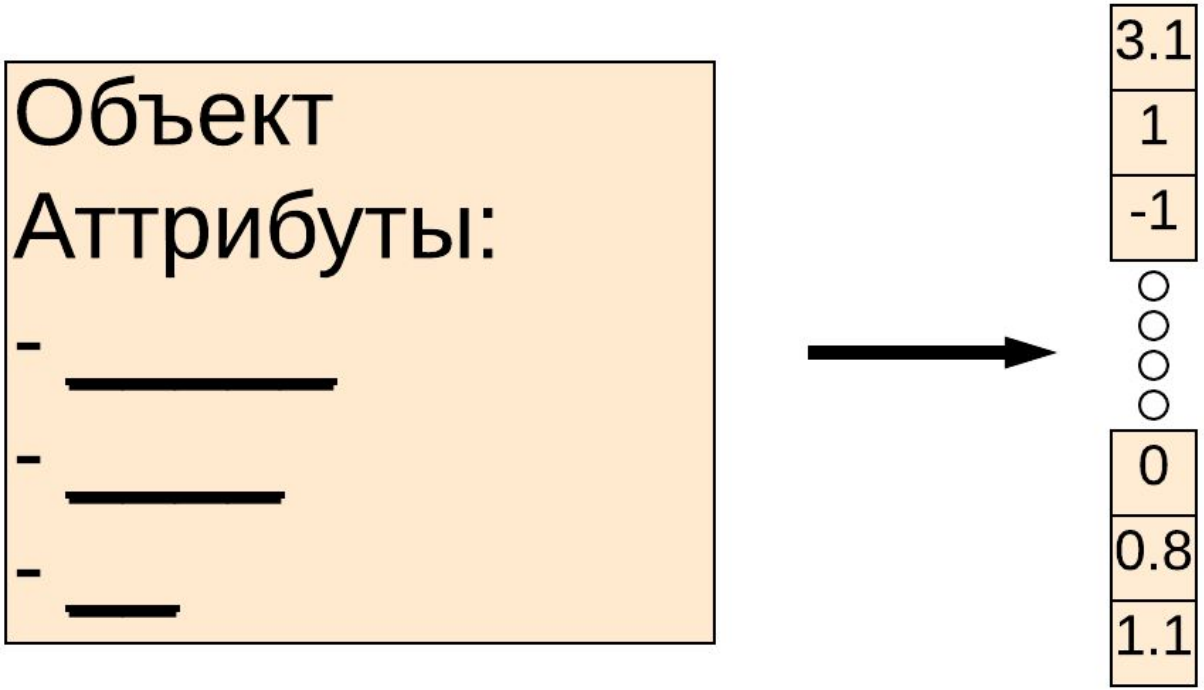
Эмбеддинги бывают разные



Эмбеддинги бывают разные



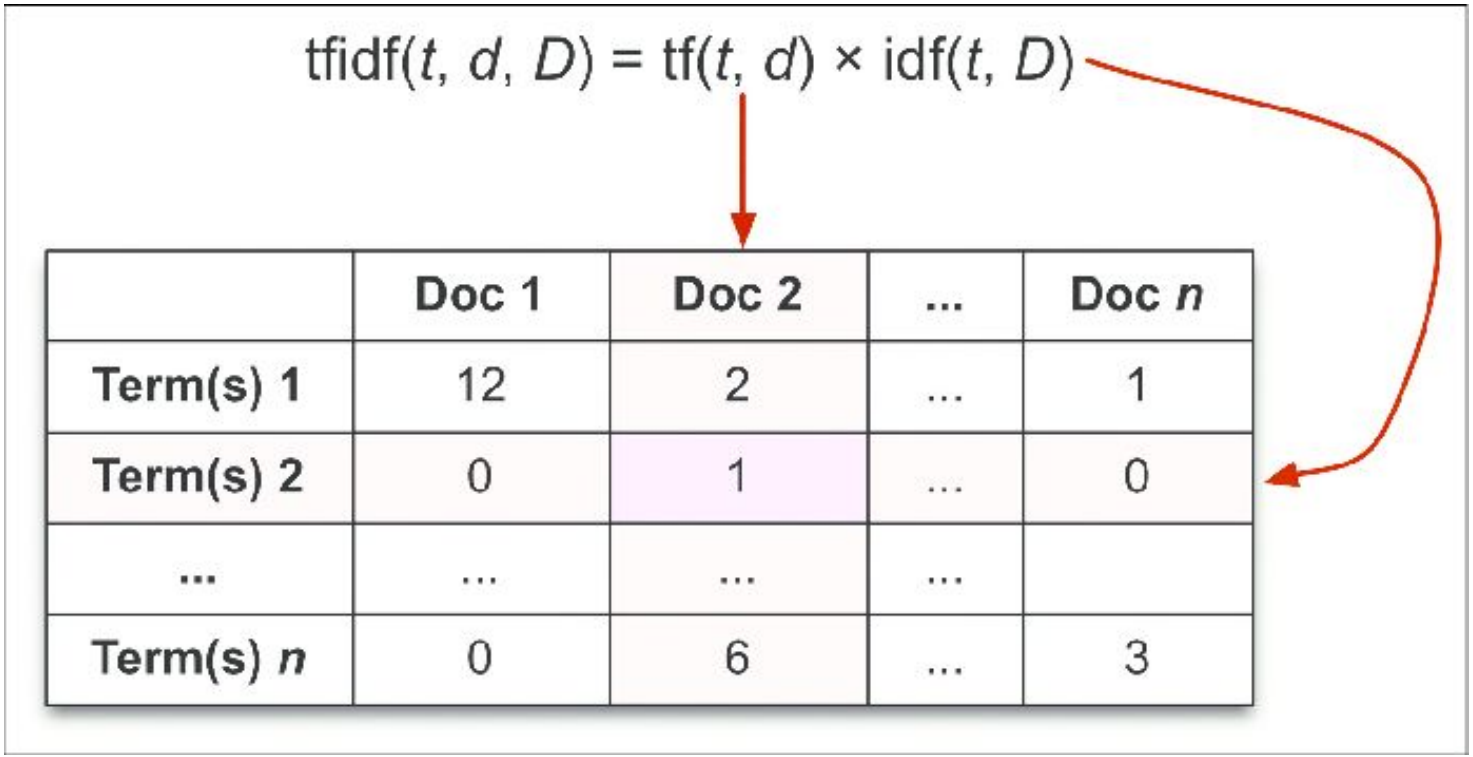
Эмбеддинги бывают разные



Что было до эмбеддингов

- Счетчики
- TF-IDF
- Факторизации и тематические модели

Счетчики



	Doc 1	Doc 2	...	Doc *n*
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...	...	...	...	
Term(s) *n*	0	6	...	3

TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Тематическое моделирование

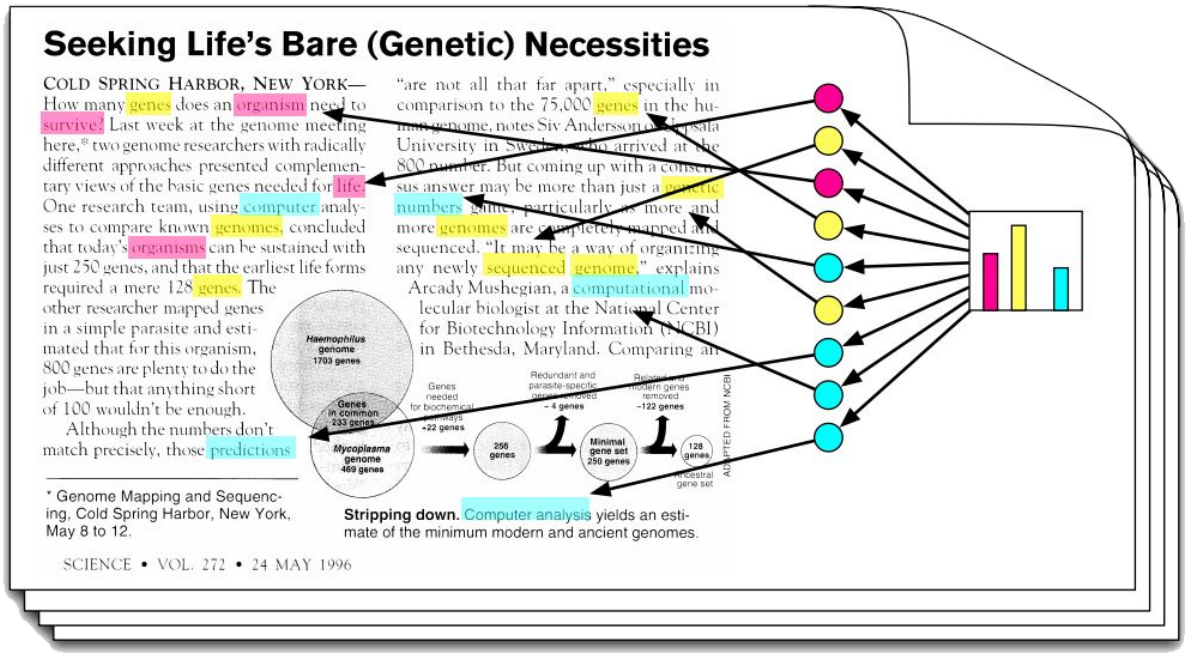
Topics *Documents* *Topic proportions & assignments*

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

http://www.machinelearning.ru/wiki/images/6/6d/BigARTM-short-intro.pdf

Пример тематической модели

#106: приложение + реклама + сервис + продукт + пользователь + платформа + …

#107: проект + рамка + мрф + реализовать + кц + решение + данный + филиал + ...

#108: работа + затрата + качество + время + количество + сотрудник + расход + …

#109: олег + александр + сергей + спасибо + тема + согласный + комментарий + …

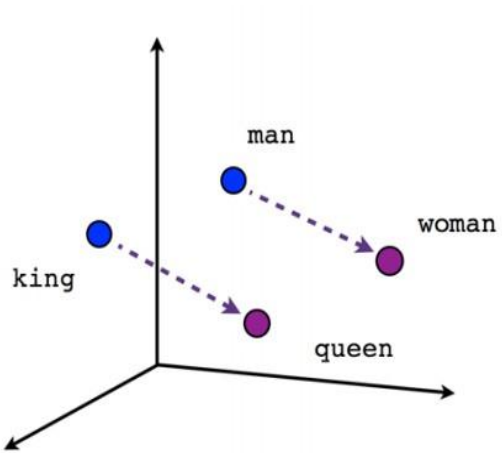
#110: приставка + компьютер + купить + пк + поставить + телевизор + питание + …

#111: система + ооъект + управление + время + контроль + группа + приоор + …

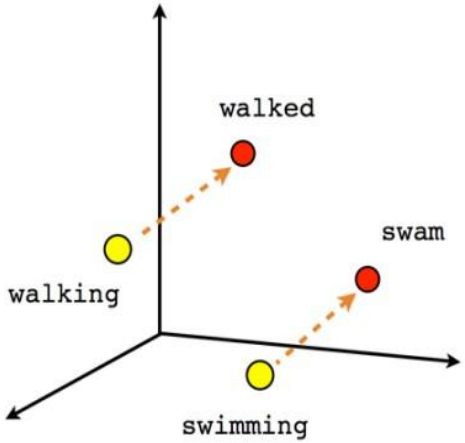
Какими свойствами обладают эмбеддинги

- “Понимание” аналогий
- “Понимание” синтаксиса
- Память некоторых фактов
- Задача поиска синонимов

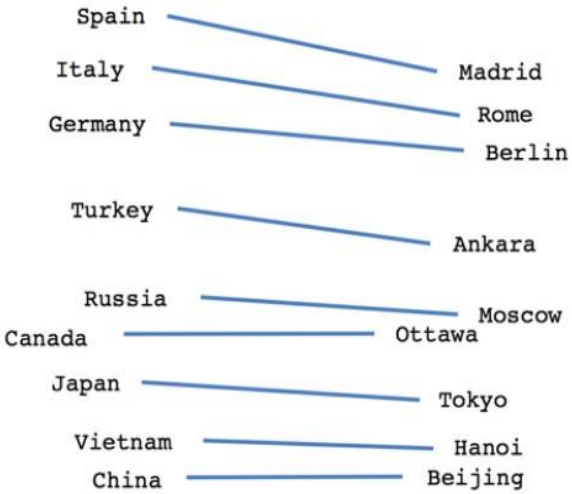
Эмбеддинги слов



Male-Female



Verb tense



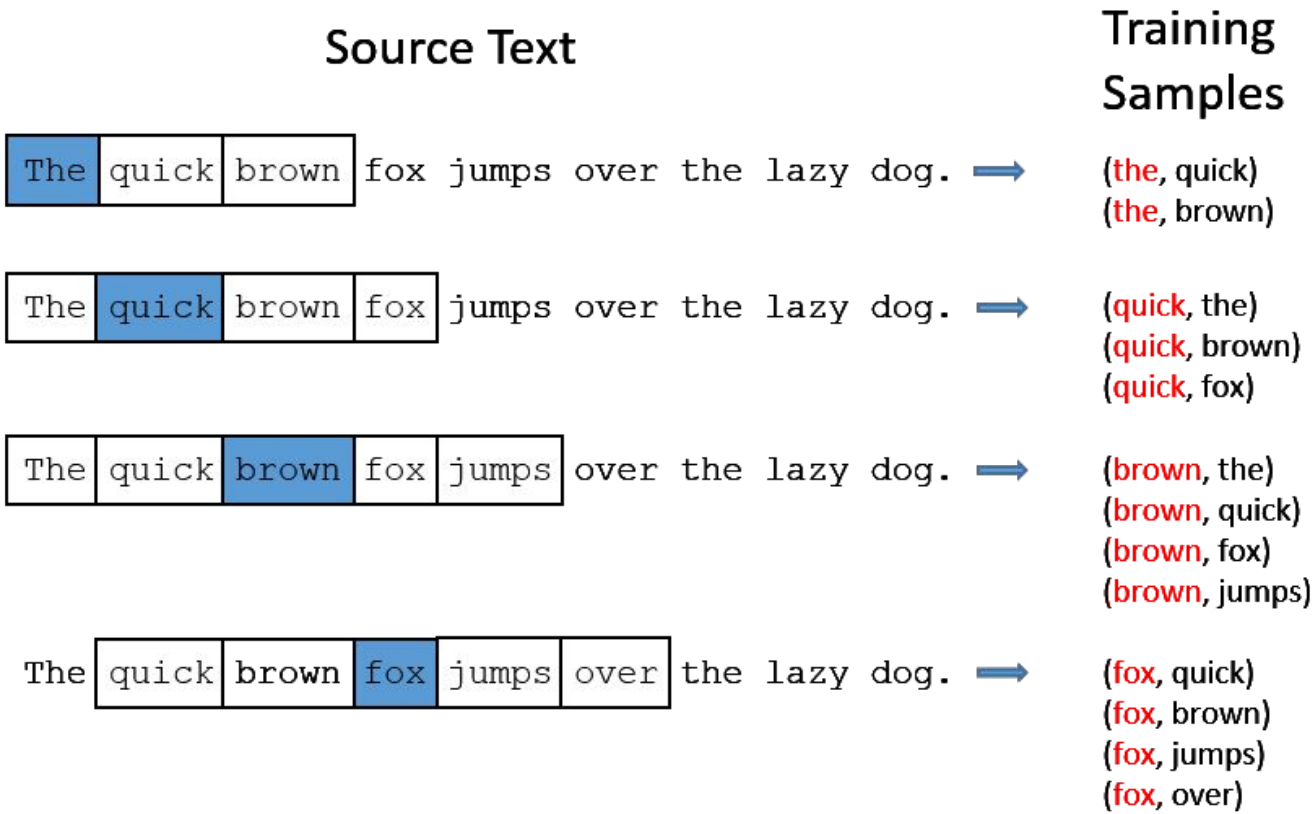
Country-Capital

Эмбеддинги слов, близость

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML plugin, zip, javascript
apple	iphone, android, mac, microsoft samsung, phone, galaxy, touch
date	registration, join, location, from changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow head, ride, ball, dead
mass	energy, effect, impact, movement potential, military, weight, society exercise, lower

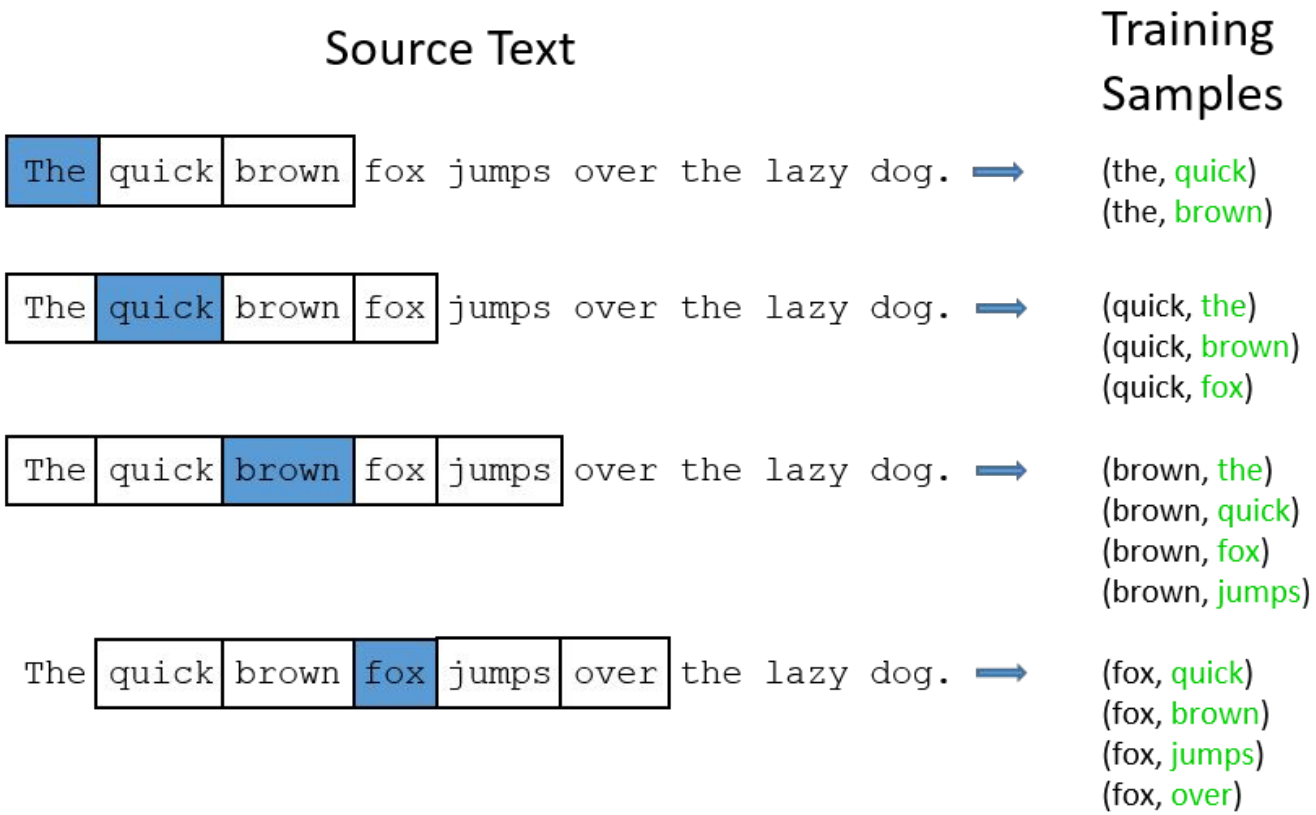
Как получают эмбеддинги

Skip-gram



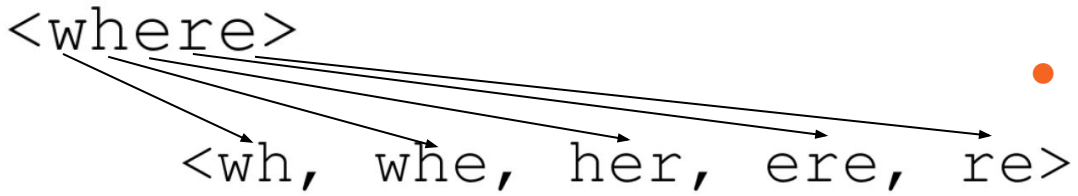
https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Skip-gram



https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Проблема Out-Of-Vocabulary (OOV)



- Char-Ngramm
- Byte Pair Encoding

Dictionary

5 l o w
2 l o w e r
6 n e w **est**
3 w i d **est**

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, es, **est**

Проблема ограниченной выразительности

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML
	plugin, zip, javascript
apple	iphone, android, mac, microsoft
	samsung, phone, galaxy, touch
date	registration, join, location, from
	changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow
	head, ride, ball, dead
mass	energy, effect, impact, movement
	potential, military, weight, society
	exercise, lower

WORD	$p(z)$	NEAREST NEIGHBOURS
python	0.33	monty, spamalot, cantsin
	0.42	perl, php, java, c++
	0.25	molurus, pythons
apple	0.34	almond, cherry, plum
	0.66	macintosh, iifx, iigs
date	0.10	unknown, birth, birthdate
	0.28	dating, dates, dated
	0.31	to-date, stateside
	0.31	deadline, expiry, dates
bow	0.46	stern, amidships, bowsprit
	0.38	spear, bows, wow, sword
	0.16	teign, coxs, evenlode
mass	0.22	vespers, masses, liturgy
	0.42	energy, density, particle
	0.36	wholesale, widespread

Способы применения

0. [Скачать]
1. [Предобработать]
2. Усреднить и затем любой ML
3. Усреднить по TF-IDF
4. Усреднить по Arora et al 2017
5. Одномерные свёртки
6. …

Скачать

1. Word2Vec
2. GLoVe
3. Fasttext

download word vectors

Скачать

Размер корпуса ▲▼	Объём словаря ▲▼	Частотный порог ▲▼	Тагсет ▲▼	Алгоритм ▲▼	Размерность вектора ▲▼	Размер окна ▲▼
270 миллионов слов	189 193	5 (потолок словаря 250K)	Universal Tags	Continuous Bag-of-Words	300	20
788 миллионов слов	248 978	5 (потолок словаря 250K)	Universal Tags	Continuous Skipgram	300	2
почти 5 миллиардов слов	249 565	5 (потолок словаря 250K)	Universal Tags	Continuous Skipgram	300	2
почти 5 миллиардов слов	192 415	5 (потолок словаря 250K)	Нет	fastText CBOW (3..5-граммы)	300	10

Скачать

***.bin** *vs* ***.vec**

Предобработка

1. Парсинг

2. Нормализация текстов

"Нет возможности сделать корректировку минусовых остатков"	**['Нет', 'возможности', 'сделать', 'корректировку', 'минусовых', 'остатков']**
"ПОС 2 расхождение на 880р недосдача"	**['ПОС', '2', 'расхождение', 'на', '880р', 'недосдача']**
"Завис ПОС№1 на "Итого". Завис с 01:05"	**['Завис', 'ПОС', '1', 'на', 'Итого', 'Завис', 'с', '01', '05']**
"Кофе машина выдает ошибку №186 при промывке."	**['Кофе', 'машина', 'выдает', 'ошибку', '186', 'при', 'промывке']**

Предобработка

1. → **2. Нормализация текстов** 3. Формирование словаря

['Нет', 'возможности', 'сделать', 'корректировку', 'минусовых', 'остатков']	→	**['нет', 'возможность', 'сделать', 'корректировка', 'минусовый', 'остаток']**
['ПОС', '2', 'расхождение', 'на', '880р', 'недосдача']	→	**['пос', '2', 'расхождение', 'на', '880р', 'недосдача']**
['Завис', 'ПОС', '1', 'на', 'Итого', 'Завис', 'с', '01', '05']	→	**['зависнуть', 'пос', '1', 'на', 'итого', 'зависнуть', 'с', '01', '05']**
['Кофе', 'машина', 'выдает', 'ошибку', '186', 'при', 'промывке']	→	**['кофе', 'машина', 'выдавать', 'ошибка', '186', 'при', 'промывка']**

Предобработка

2. → **3. Формирование словаря**

['нет', 'возможность', 'сделать',
'корректировка', 'минусовый',
'остаток']

['пос', '2', 'расхождение', 'на', '880р',
'недосдача']

['зависнуть', 'пос', '1', 'на', 'итого',
'зависнуть', 'с', '01', '05']

['кофе', 'машина', 'выдавать', 'ошибка',
'186', 'при', 'промывка']

Предобработка

2. → **3. Формирование словаря**

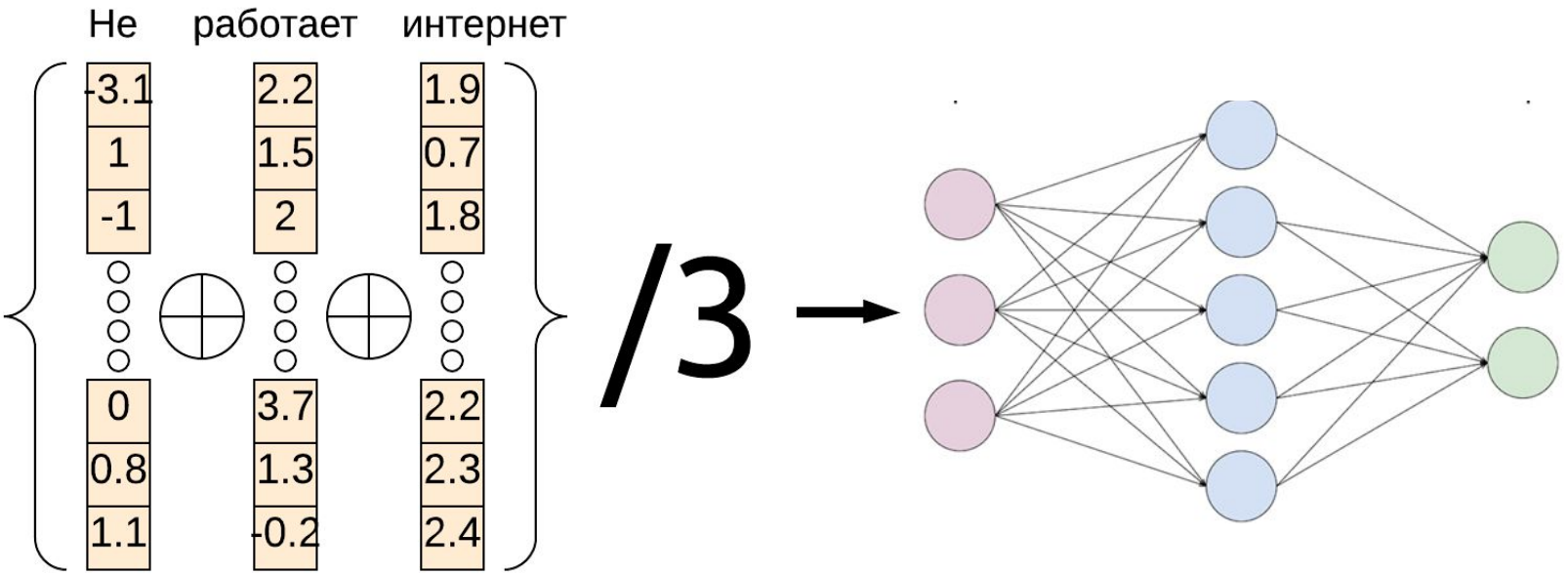
- "онлайн" -> 137
- "перечеркнуть" -> 138
- "трм" -> 139
- "вод" -> 140
- "зал" -> 141
- "клиентский" -> 142
- "пол" -> 143
- "потечь" -> 144
- "протекать" -> 145
- "торговый" -> 146
- "туалет" -> 147
- "выйти" -> 148
- "выполнить" -> 149

Предобработка

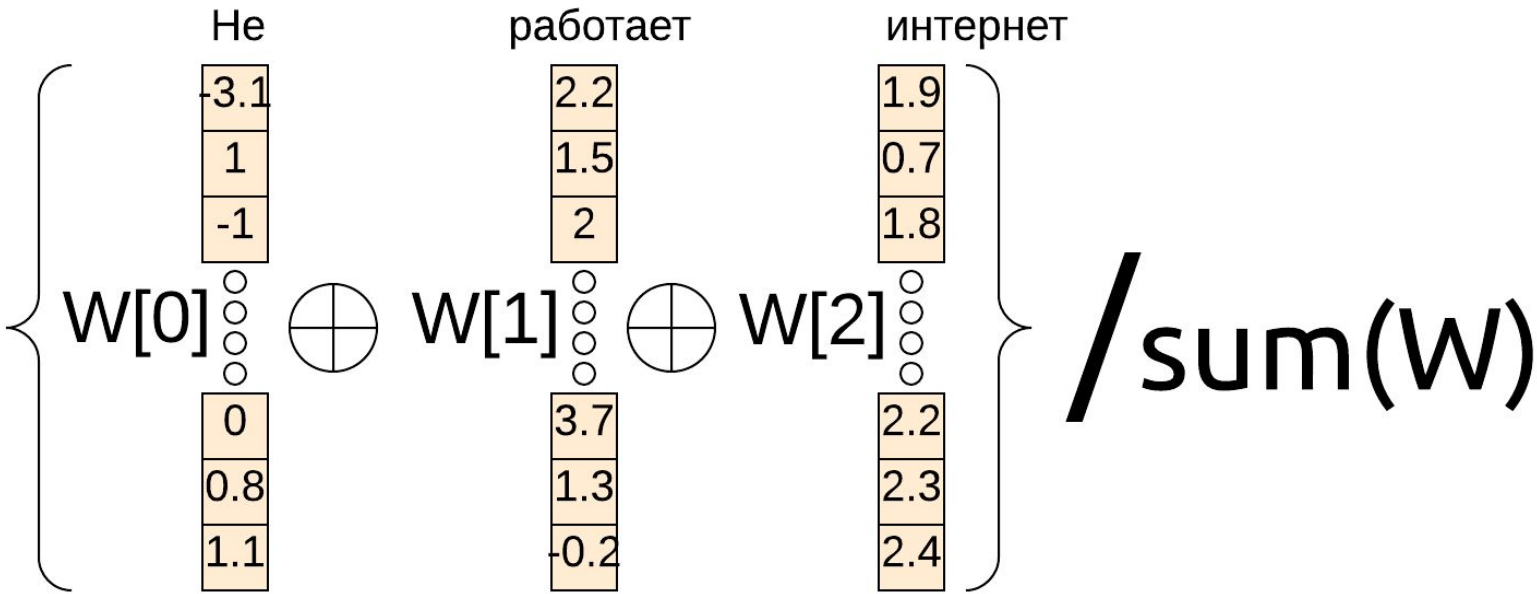
3. → 4. **Фильтрация стоп-слов**

"нет", "быть", "к", "за", "после", "при", "как",
"так", "между", "более", "до", "если", "здесь",
"из", "можно", "о", "они", "перед", "сам", "то",
"тот", "что", "вы", "или", "чем"

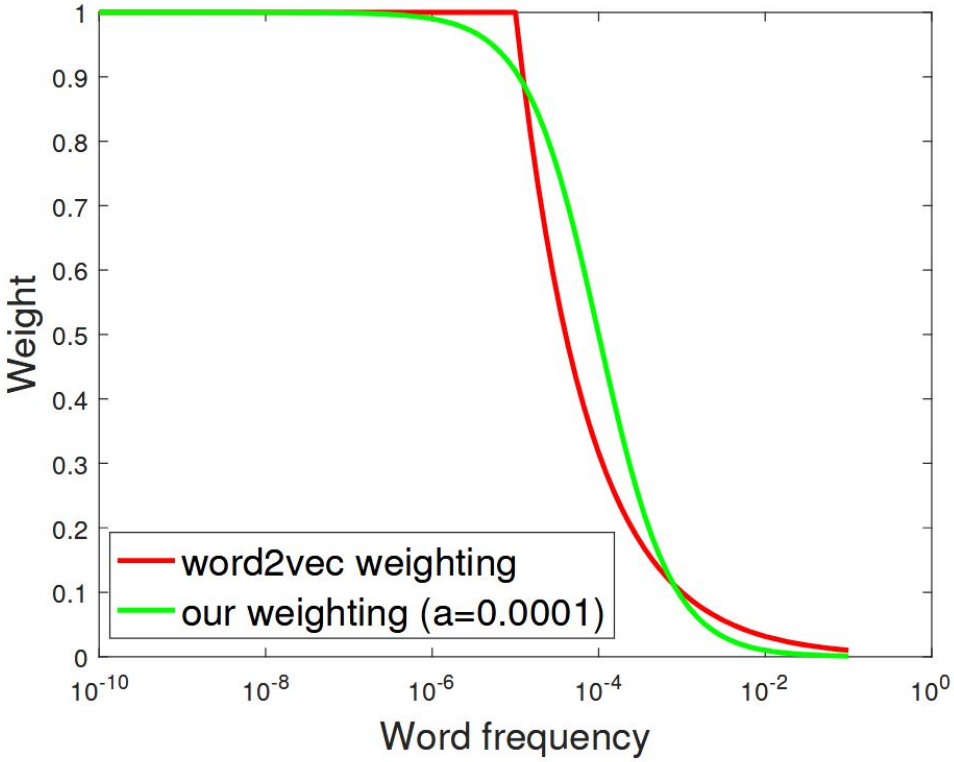
Feature extraction. Усреднение векторов + ML



Feature extraction. Усреднение по TF-IDF + ML



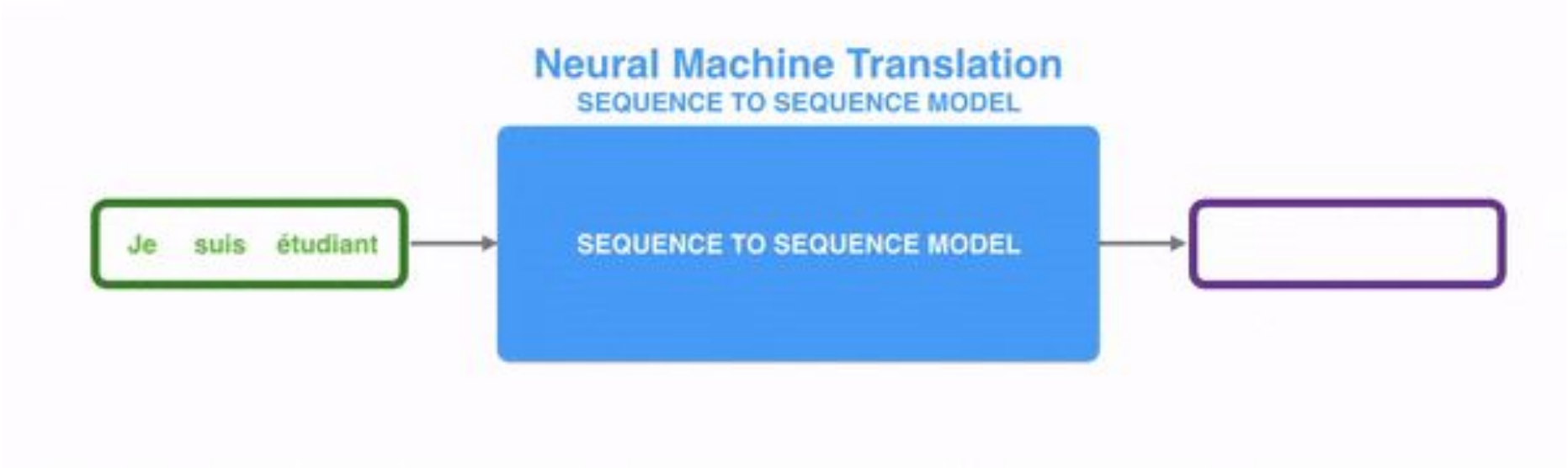
Feature extraction. Усреднение по Arora et al 2017 + ML

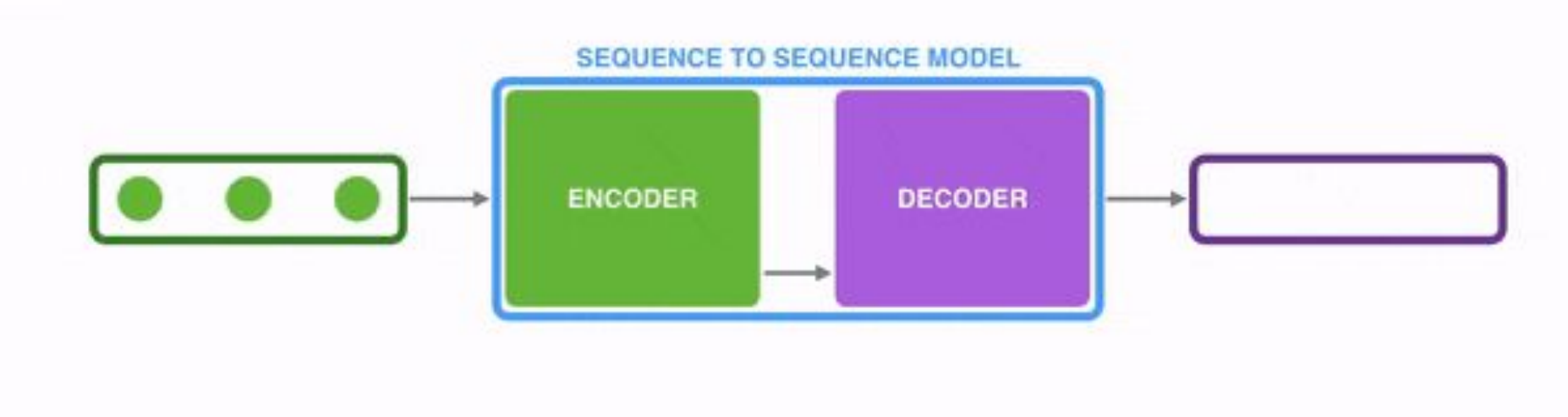


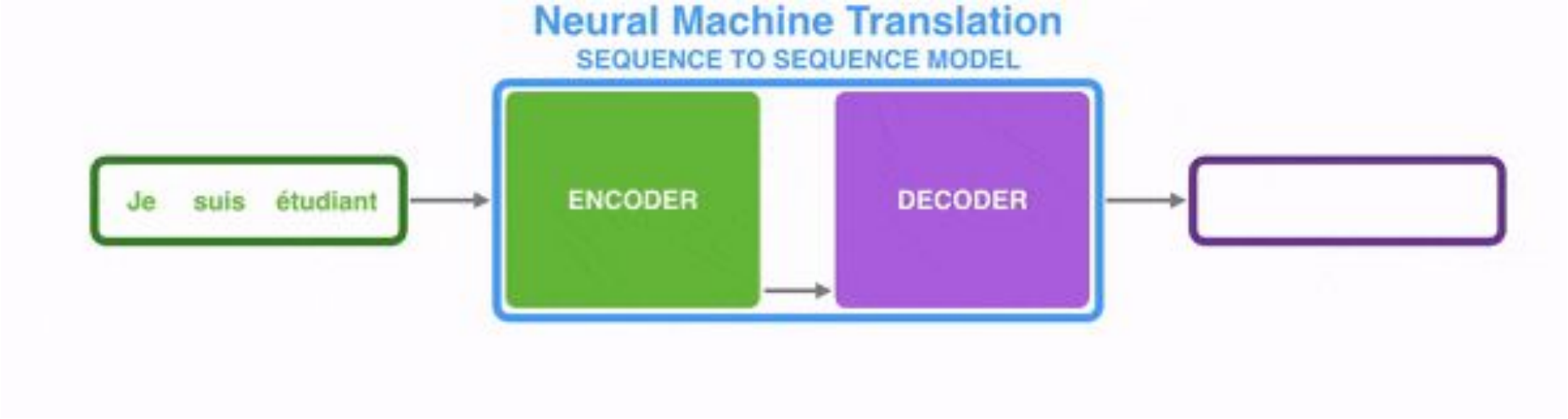
https://openreview.net/pdf?id=SyK00v5xx

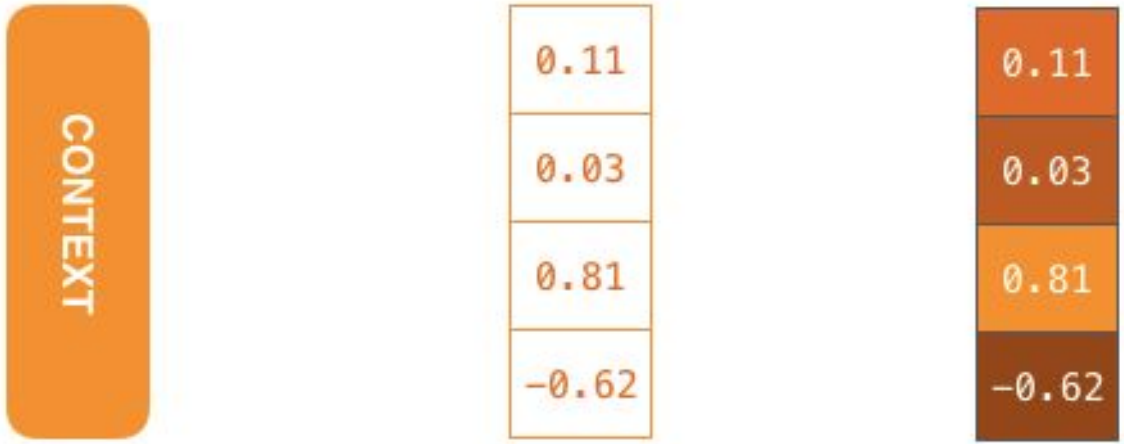
Seq2Seq


SEQUENCE TO SEQUENCE MODEL


Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL
Je suis étudiant
SEQUENCE TO SEQUENCE MODEL


SEQUENCE TO SEQUENCE MODEL
ENCODER
DECODER


Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL
Je suis étudiant
ENCODER
DECODER


CONTEXT
0.11
0.03
0.81
−0.62
0.11
0.03
0.81
−0.62

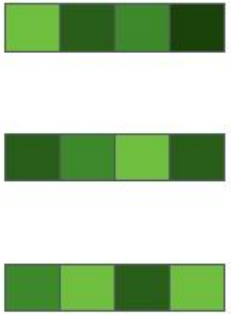
Input

Je

suis

étudiant

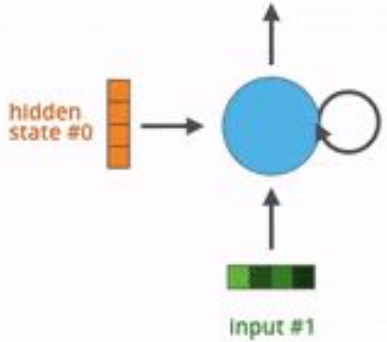
0.901	-0.651	-0.194	-0.822
-0.351	0.123	0.435	-0.200
0.081	0.458	-0.400	0.480

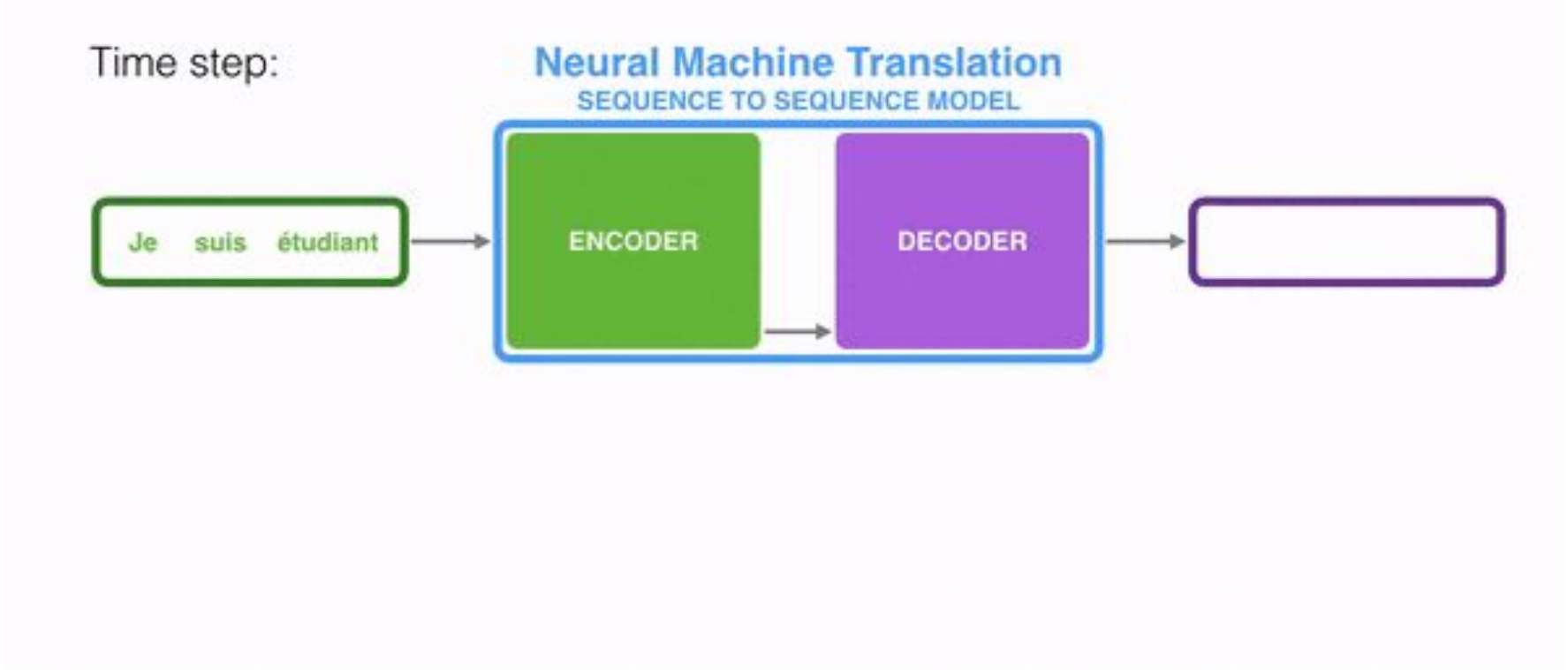


RNN as seq2seq model

Recurrent Neural Network

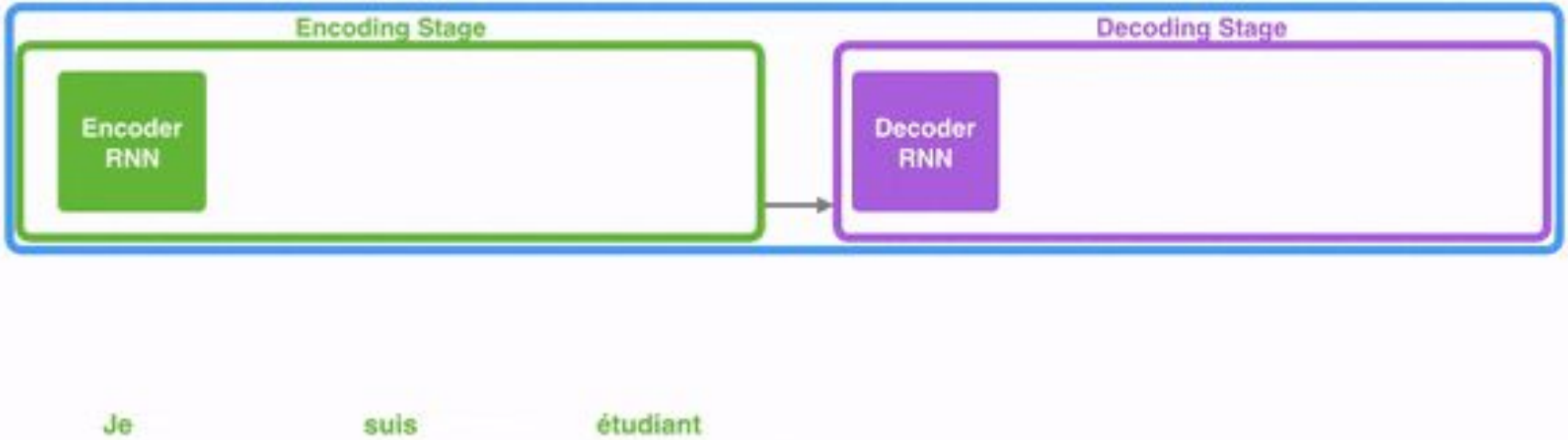
Time step #1:
An RNN takes two input vectors:




Time step:
Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL
Je suis étudiant
ENCODER
DECODER

Neural Machine Translation

SEQUENCE TO SEQUENCE MODEL



Time step: 7

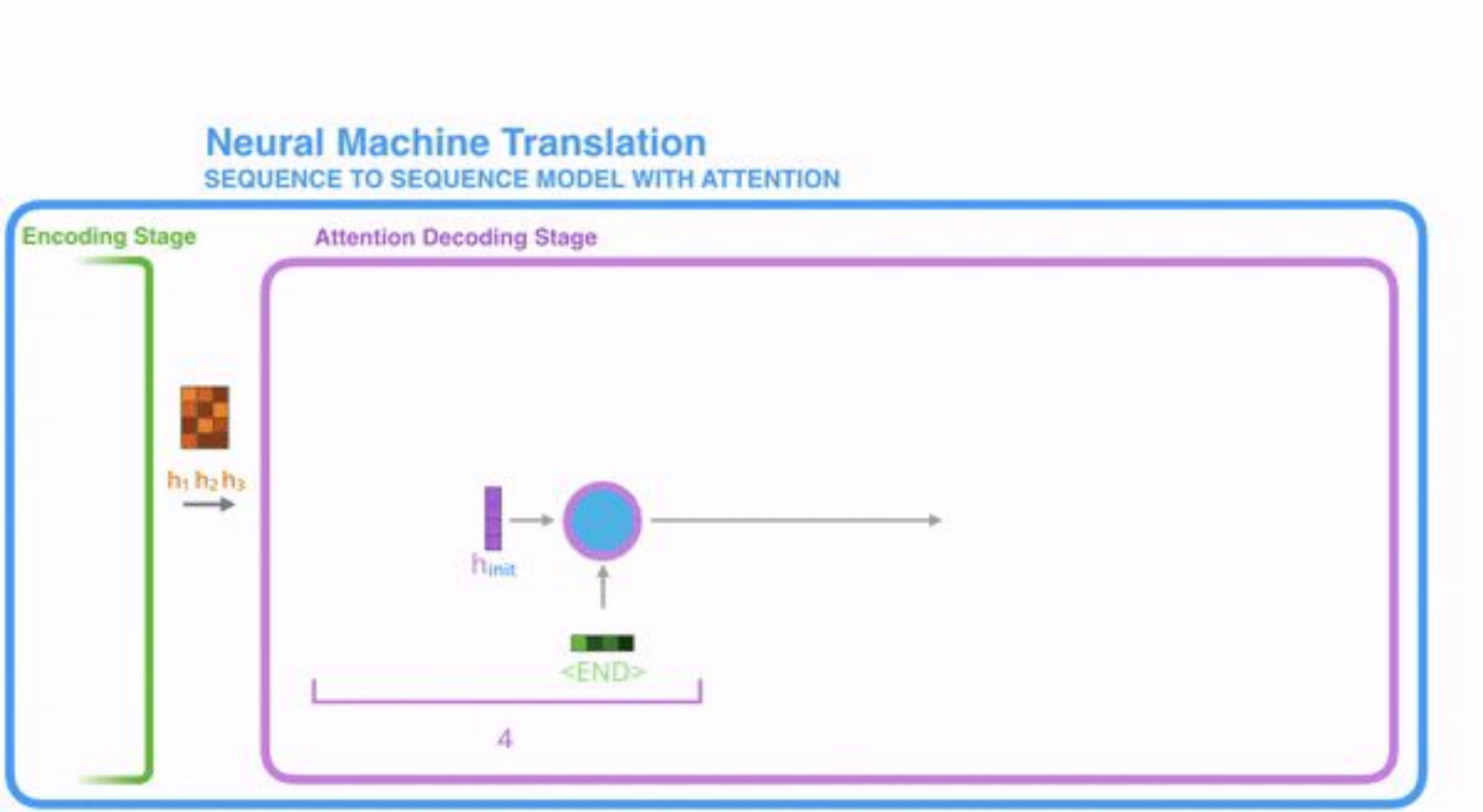

I
am
a
Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL WITH ATTENTION
Encoding Stage
Encoder RNN
Encoder RNN
Encoder RNN
Decoding Stage
Attention Decoder RNN
Attention Decoder RNN
Attention Decoder RNN
Je
suis
étudiant
Attention Decoder RNN

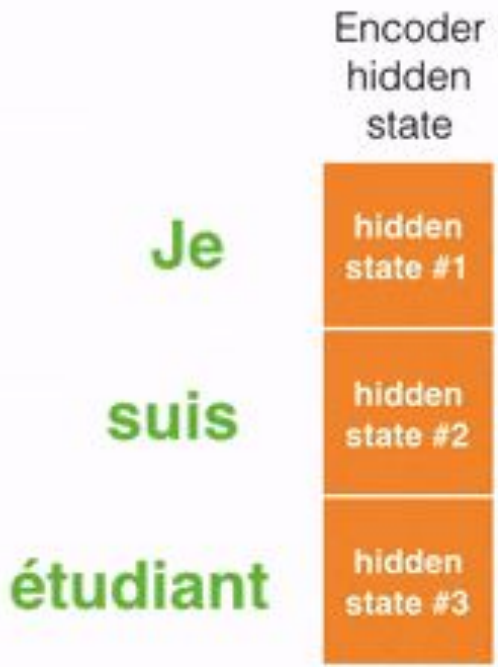
Augmenting by attention

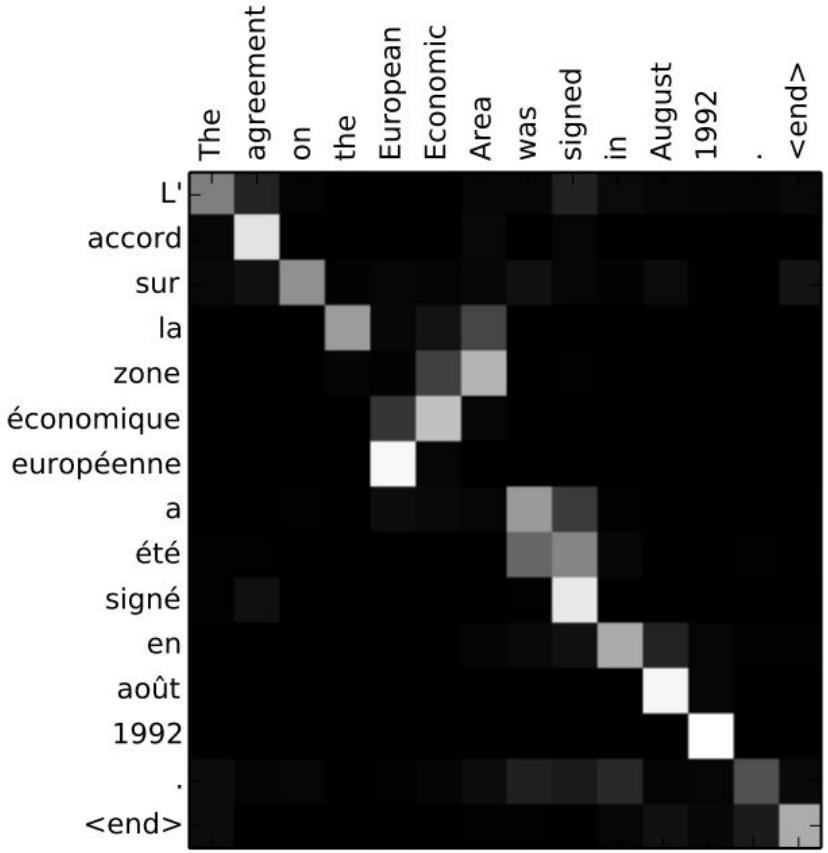
Neural Machine Translation

SEQUENCE TO SEQUENCE MODEL WITH ATTENTION


Attention at time step 4

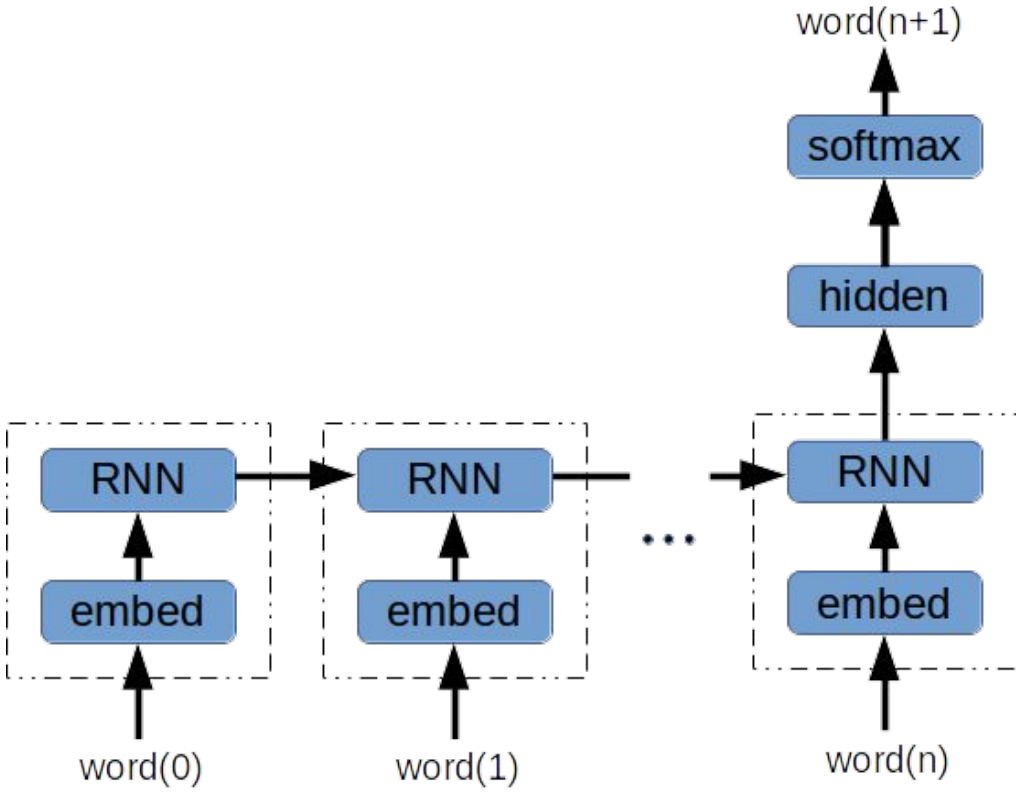

Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL WITH ATTENTION
Encoding Stage
Attention Decoding Stage
h1 h2 h3
hinit
<END>
4


Encoder
hidden
state
Je
suis
étudiant
hidden
state #1
hidden
state #2
hidden
state #3

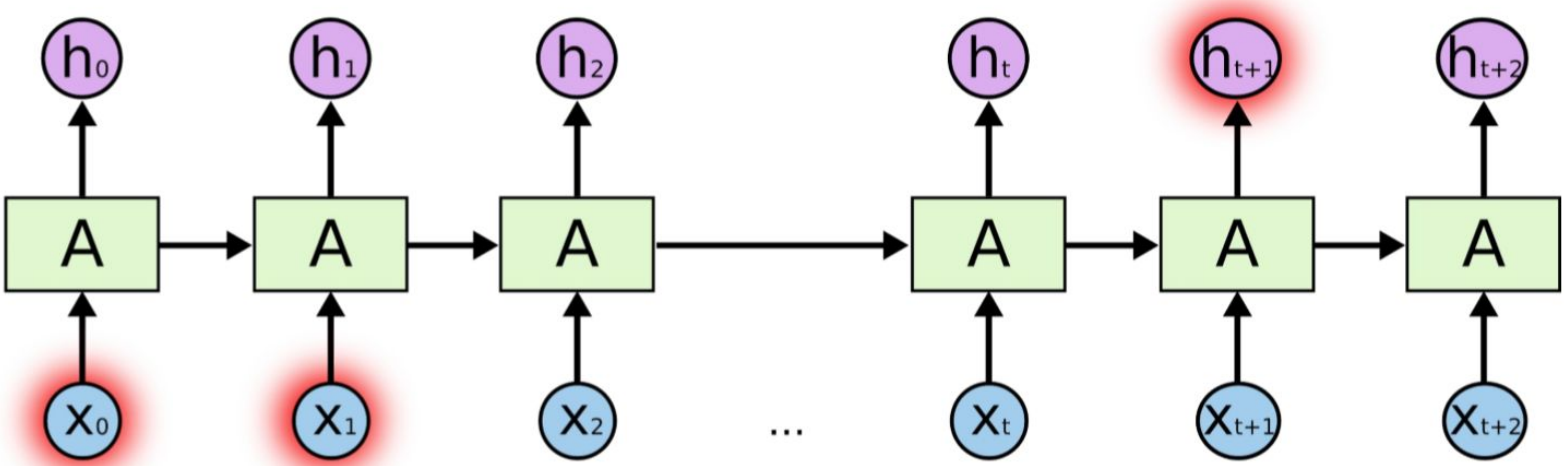

The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

RNN recap

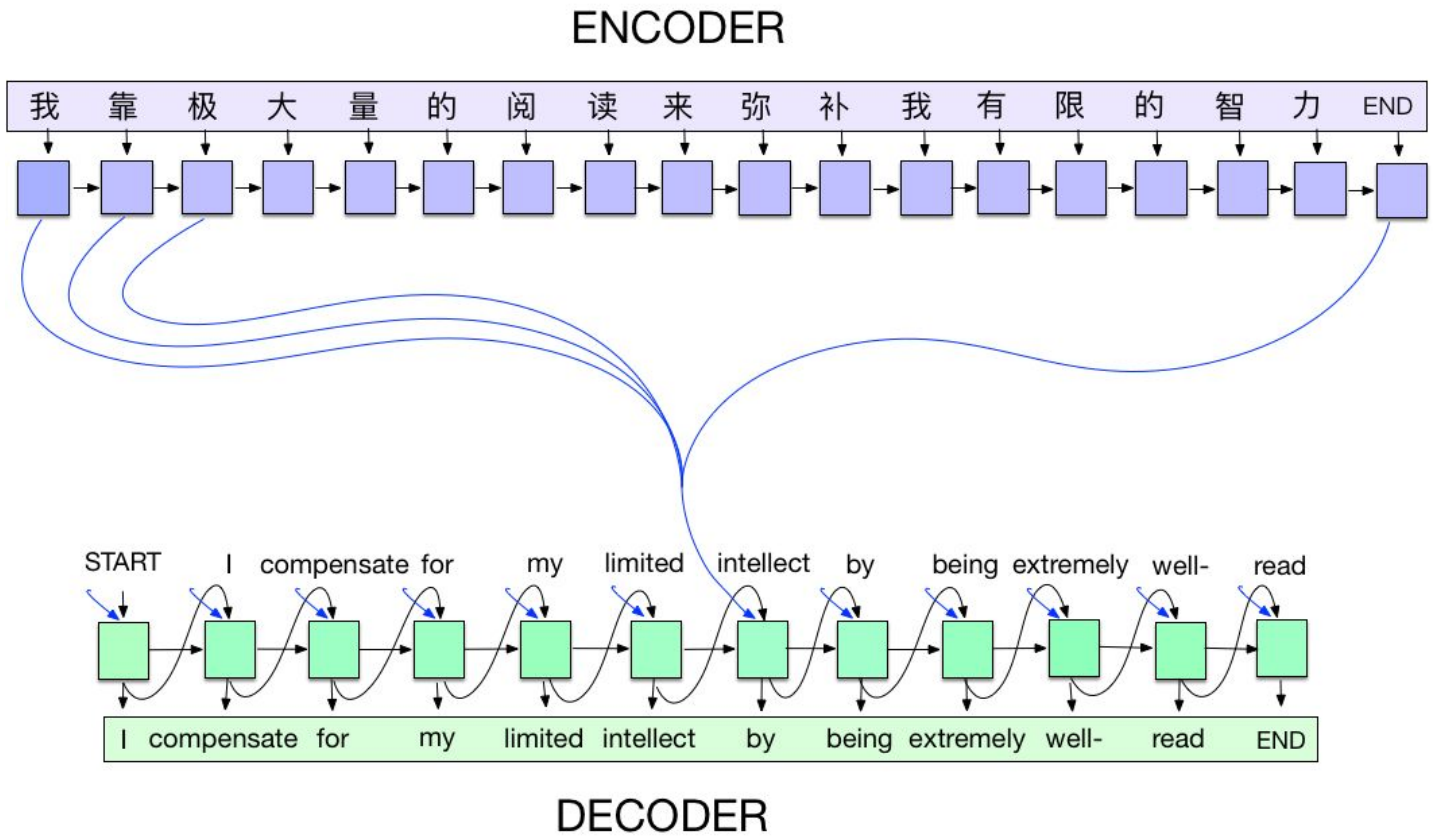
Feature extraction. Рекуррентные нейронные сети (RNN)



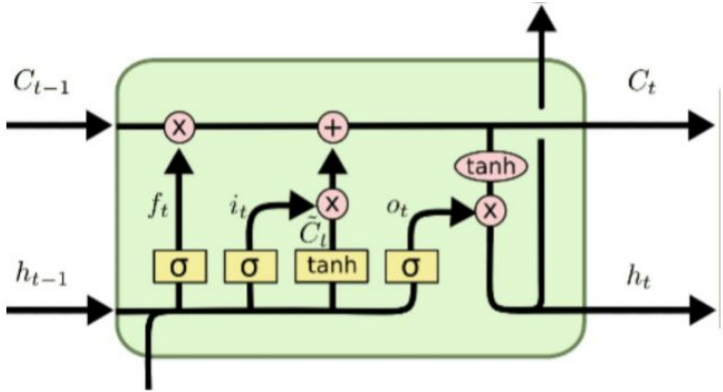
Рекуррентные нейронные сети



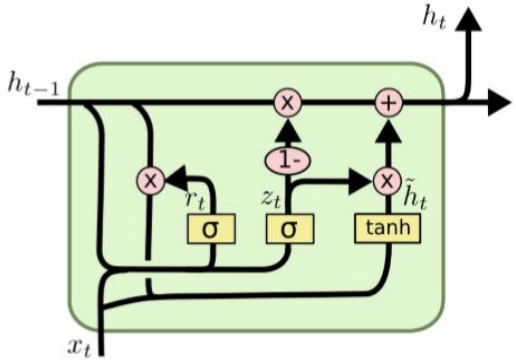
Машинный перевод

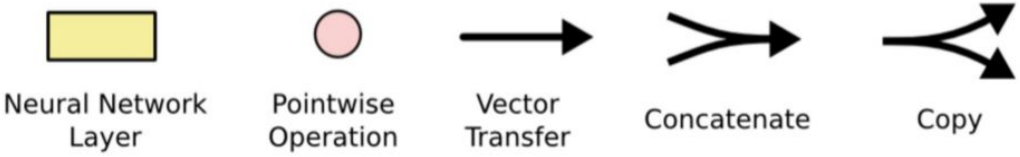


Подробней



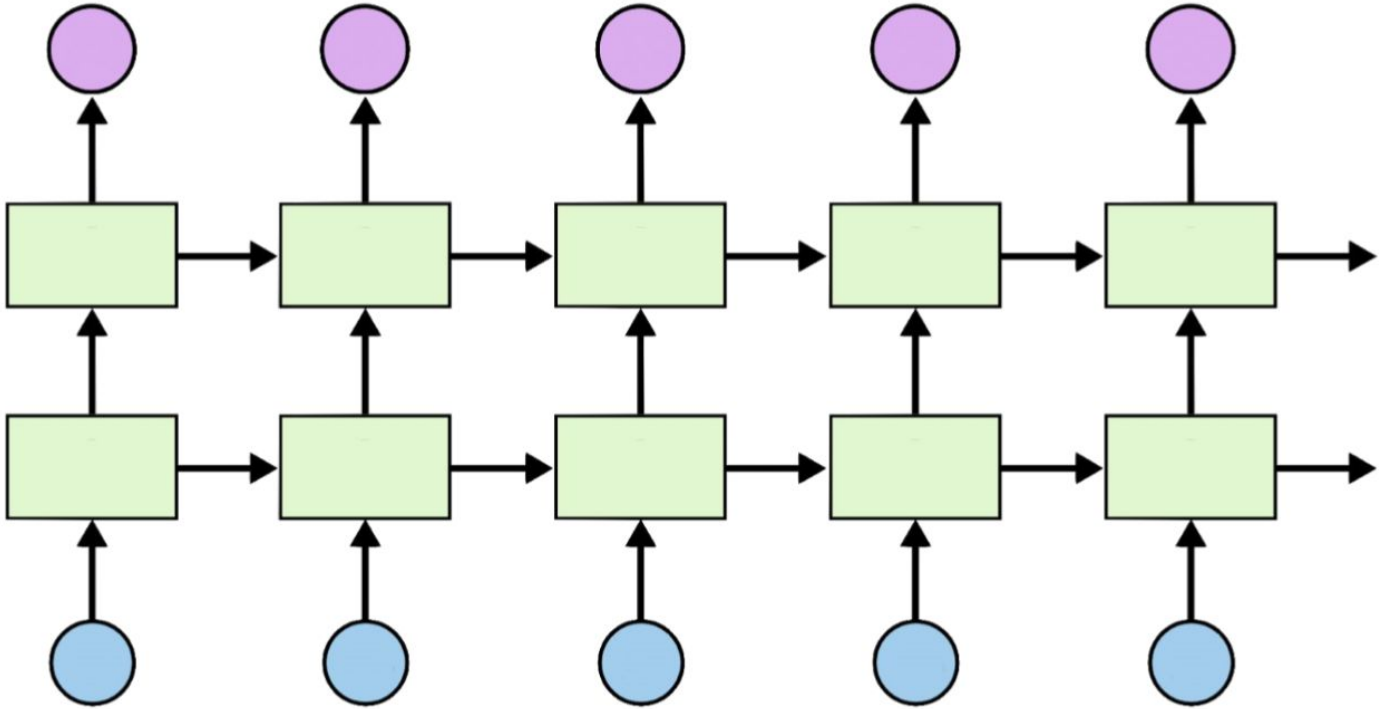
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$\tilde{C}_t = \mathrm{th}\left(W_C \cdot [h_{t-1}, x_t] + b_C\right)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$
$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right)$$
$$h_t = o_t \odot \mathrm{th}(C_t)$$



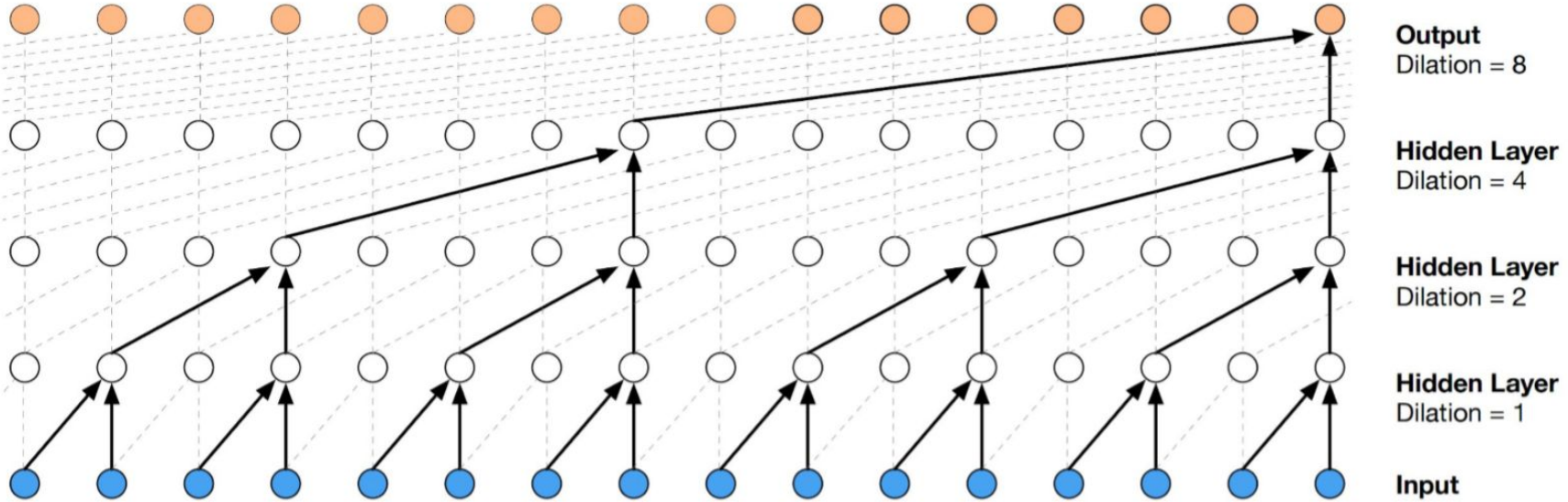


Beyond RNN

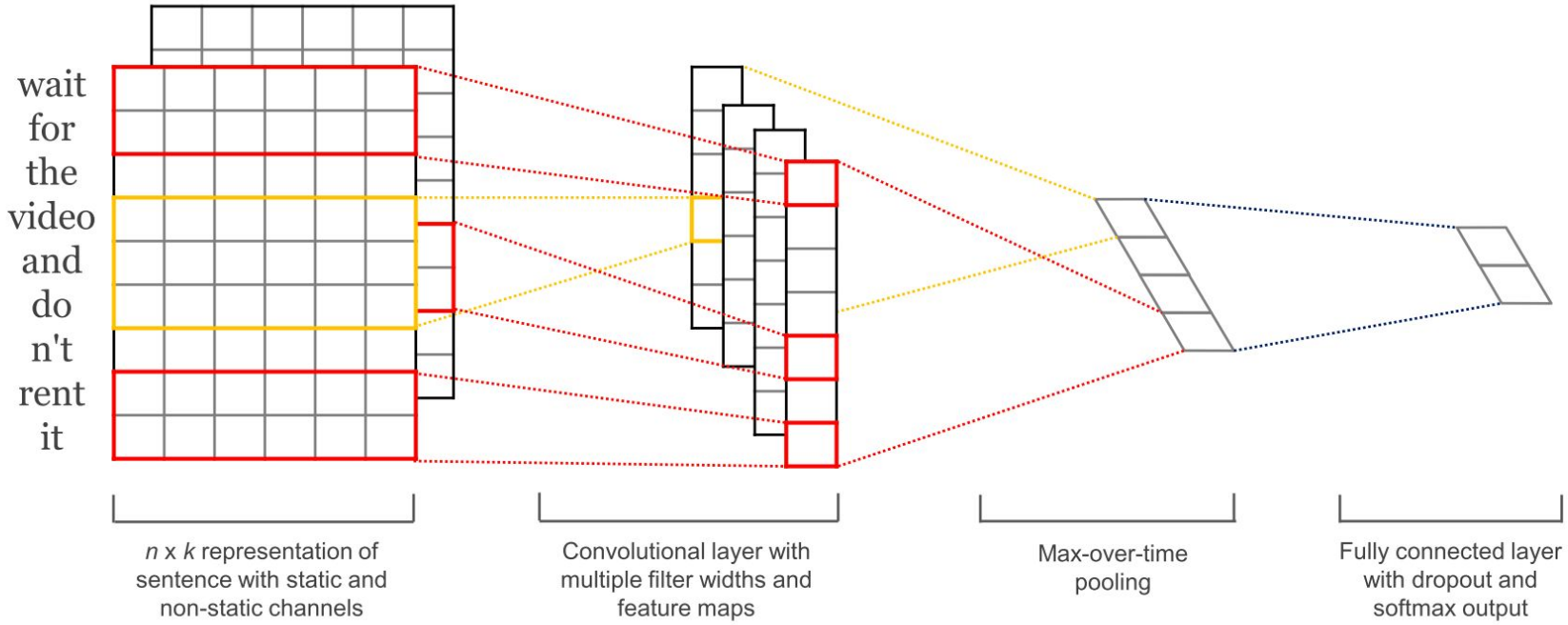
Многослойные реккурентные сети



Казуальные разреженные свертки

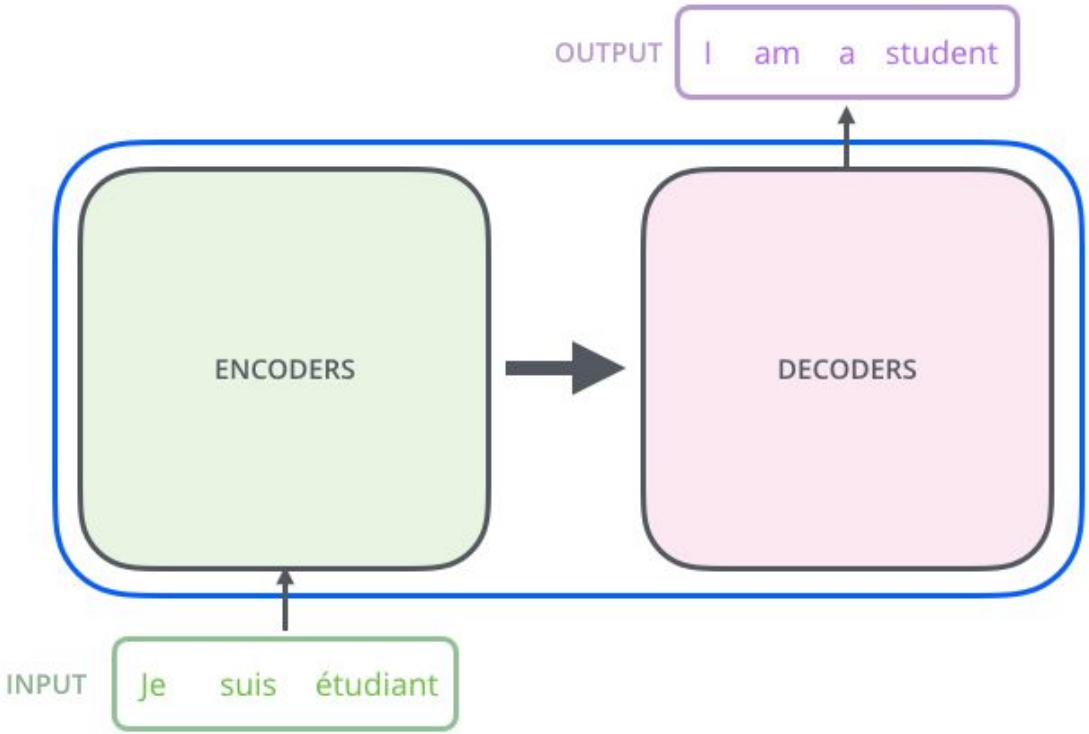


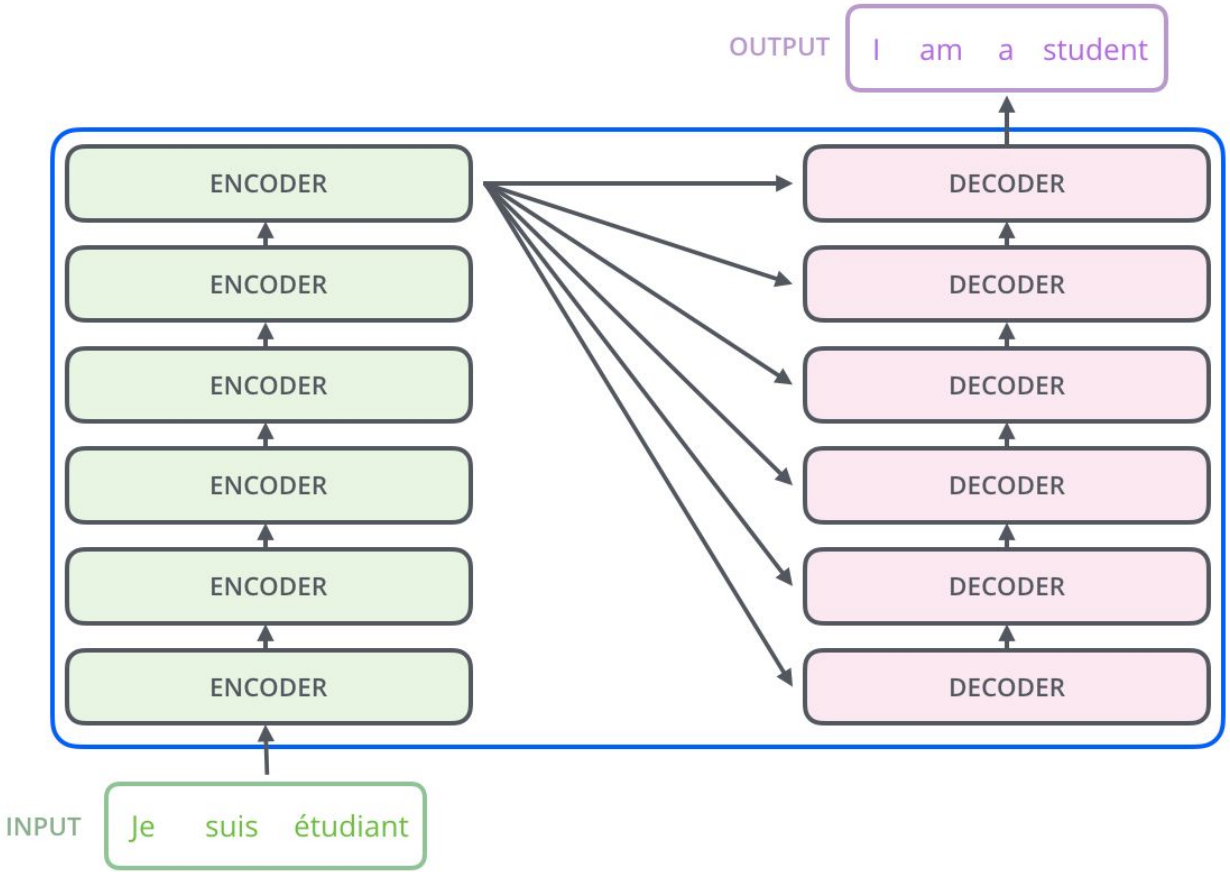
Feature extraction. Свёрточные нейронные сети (CNN)

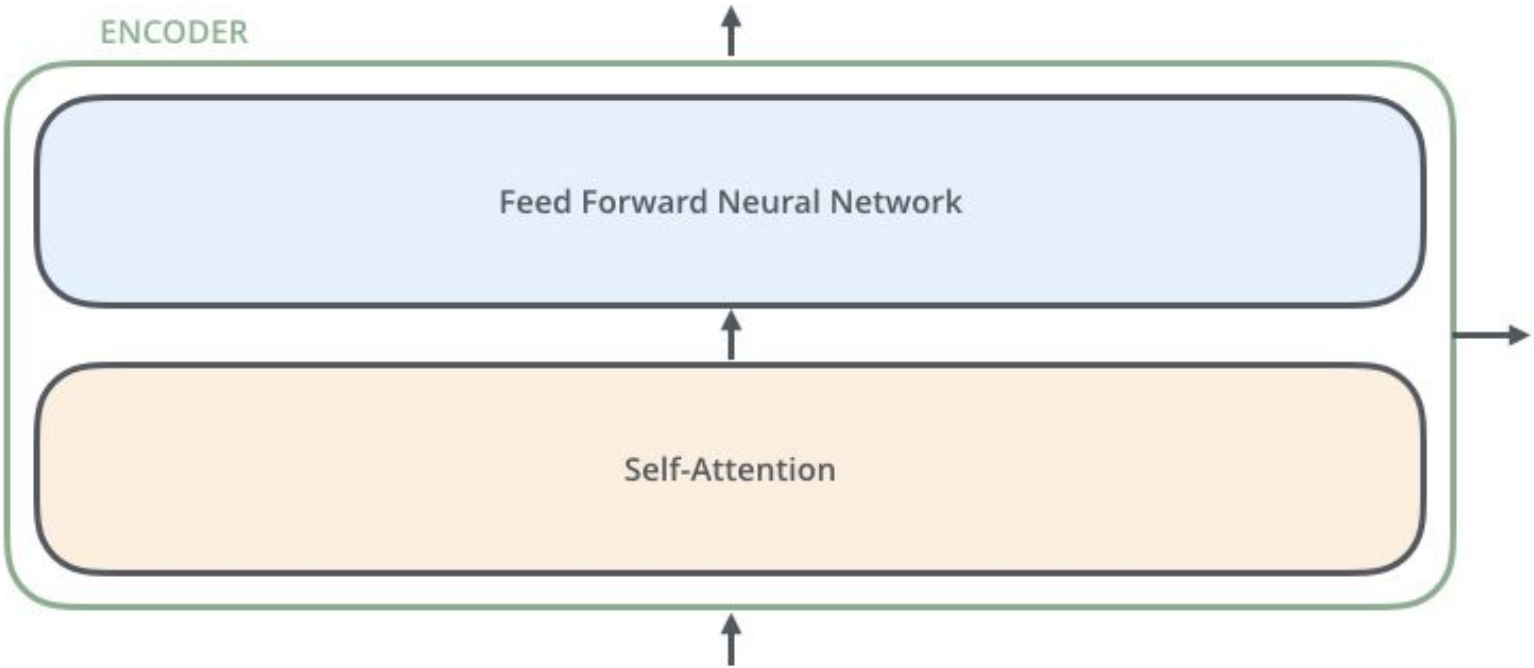


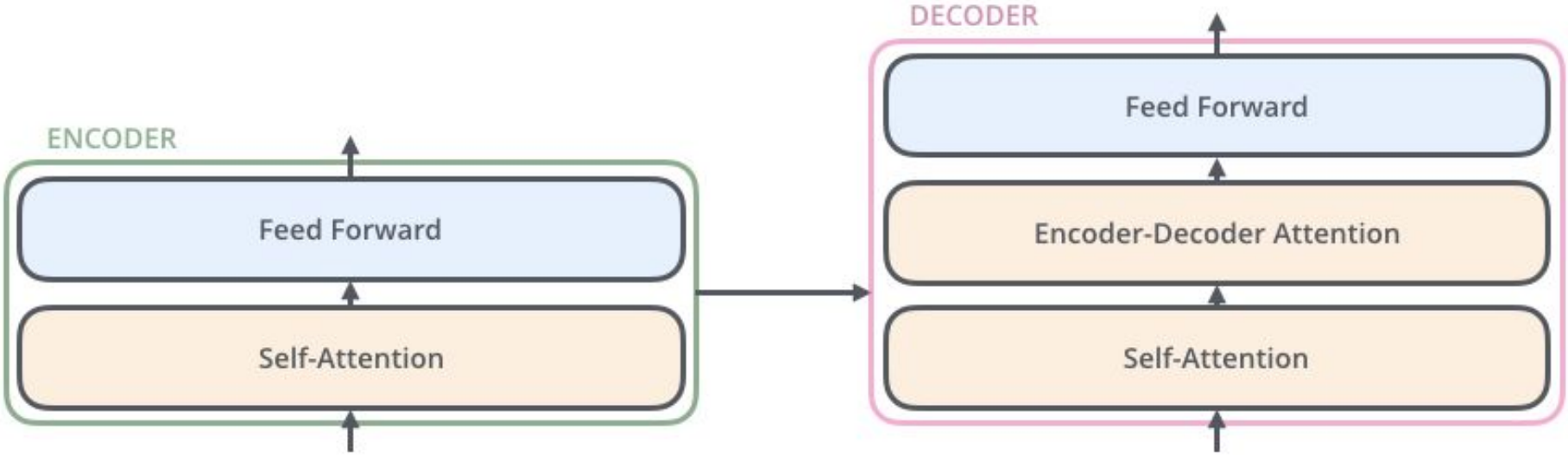
Transformer

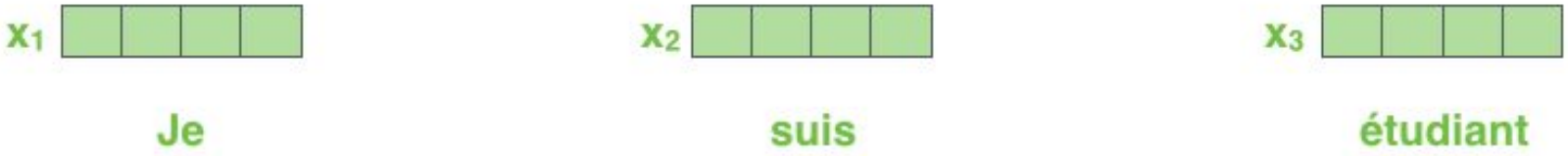
Transformer seq2seq too

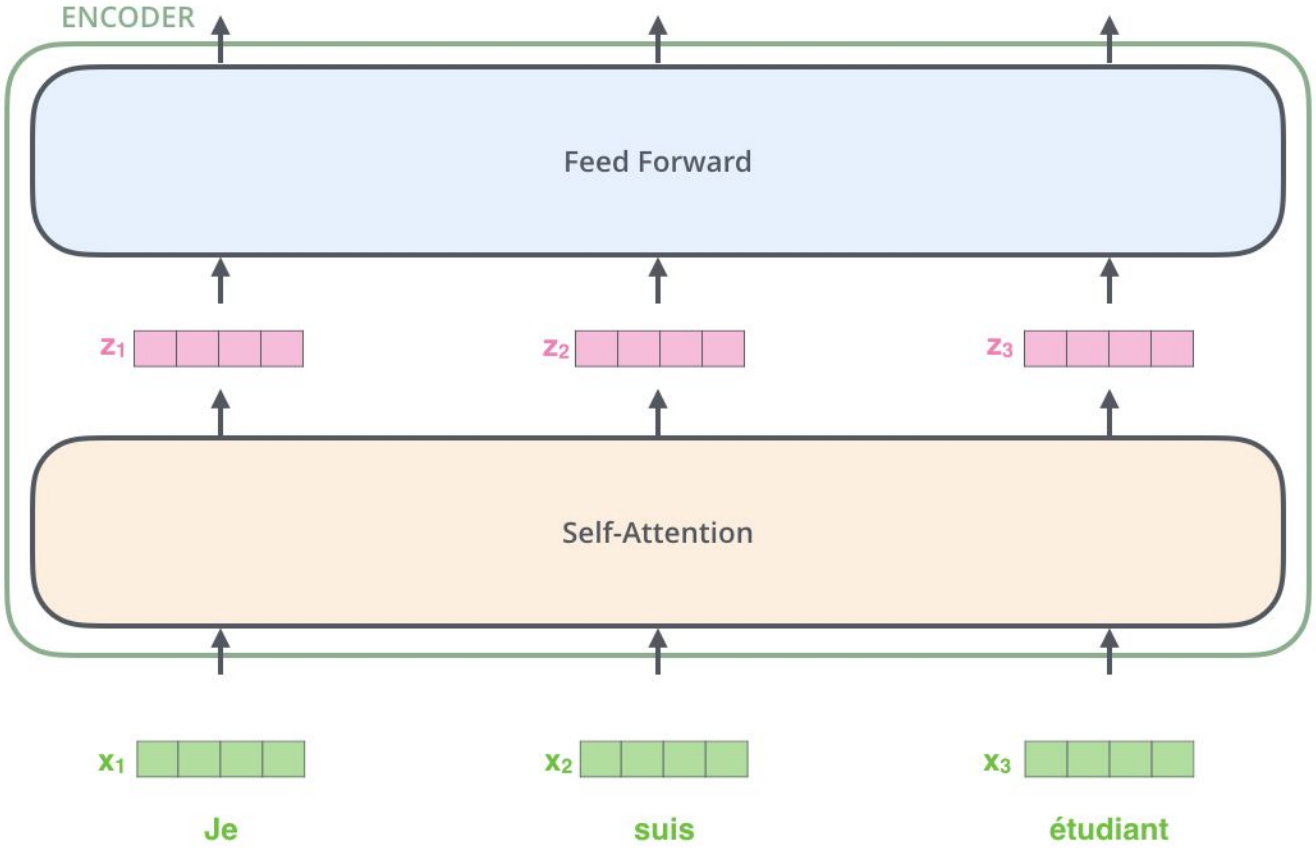

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant

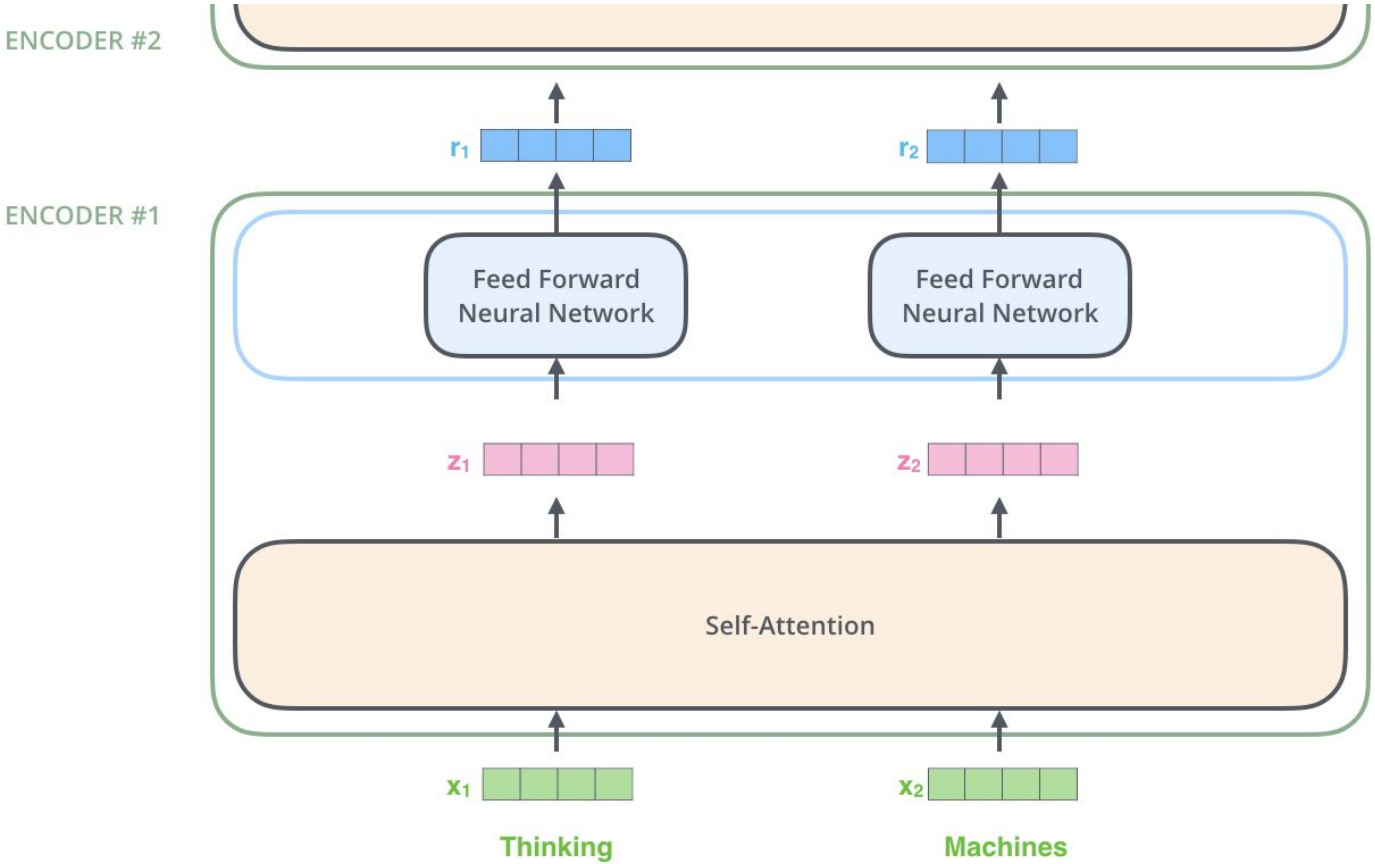

OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant


ENCODER
Feed Forward Neural Network
Self-Attention

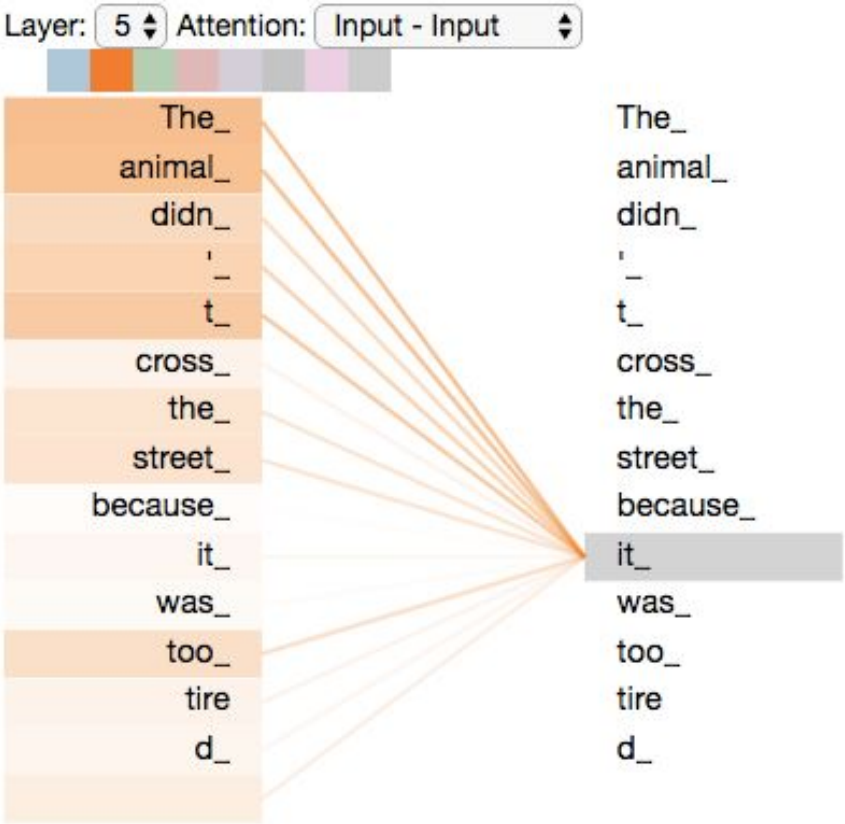

DECODER
Feed Forward
Encoder-Decoder Attention
Self-Attention
ENCODER
Feed Forward
Self-Attention

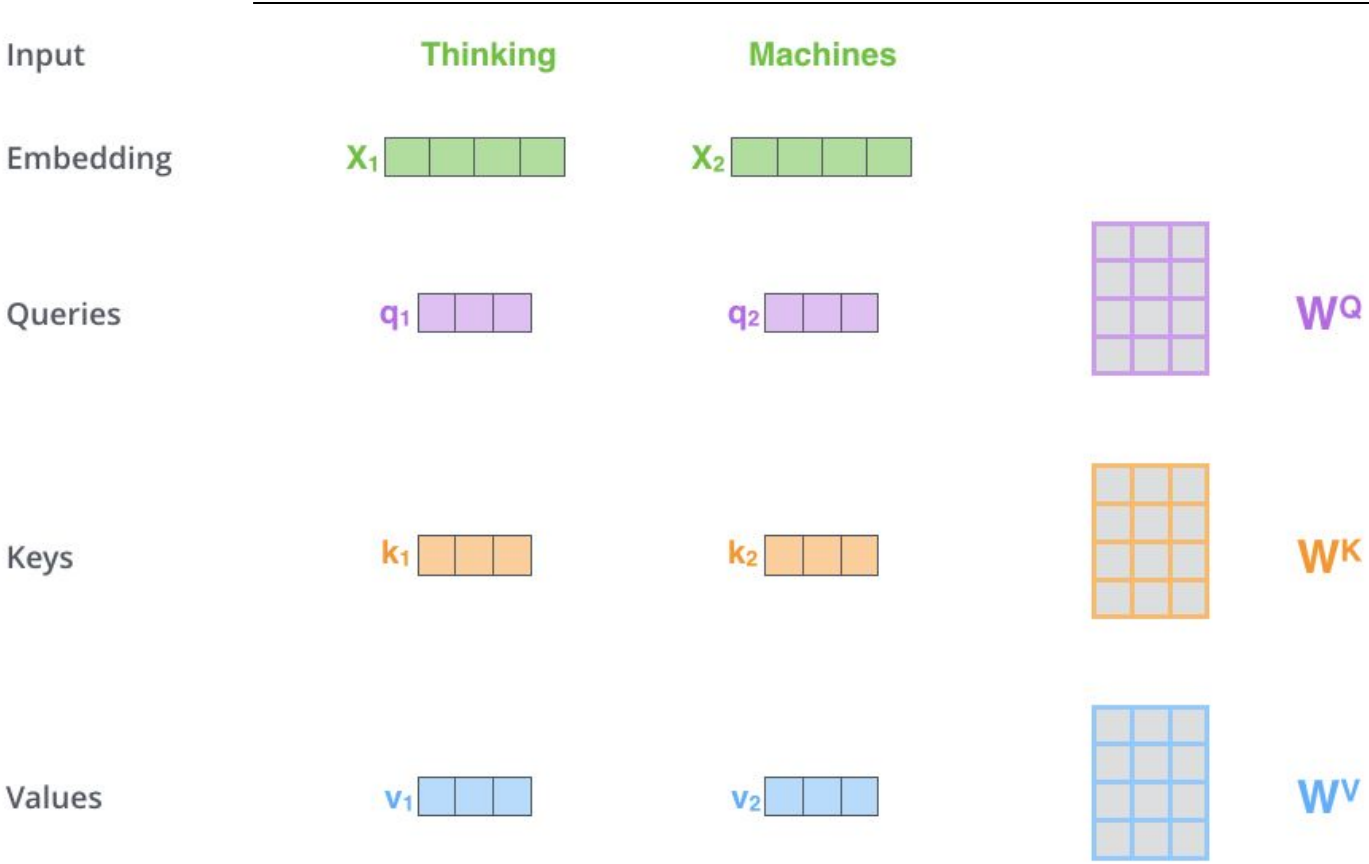

x_1
x_2
x_3
Je
suis
étudiant

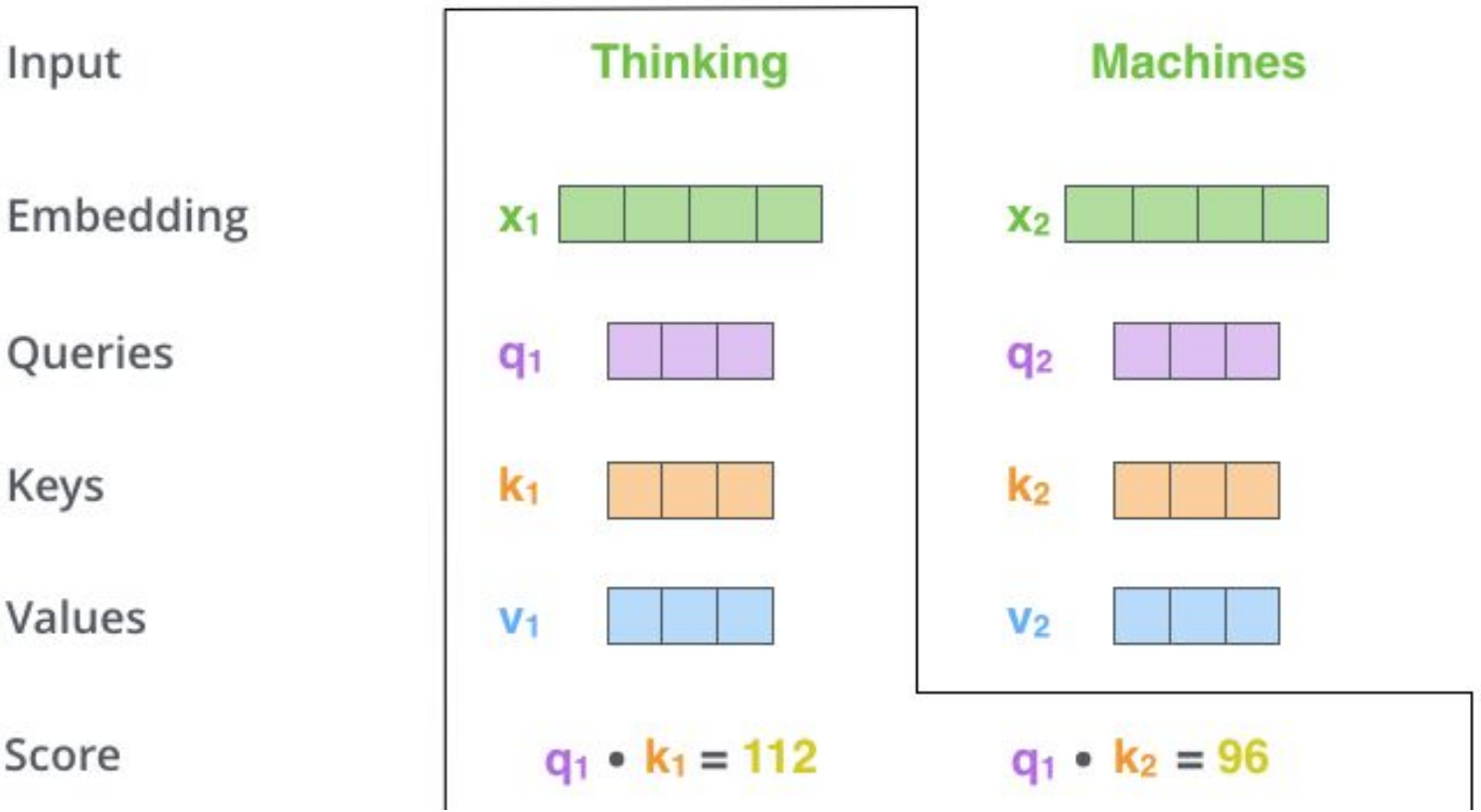

ENCODER
Feed Forward
z1
z2
z3
Self-Attention
x1
x2
x3
Je
suis
étudiant

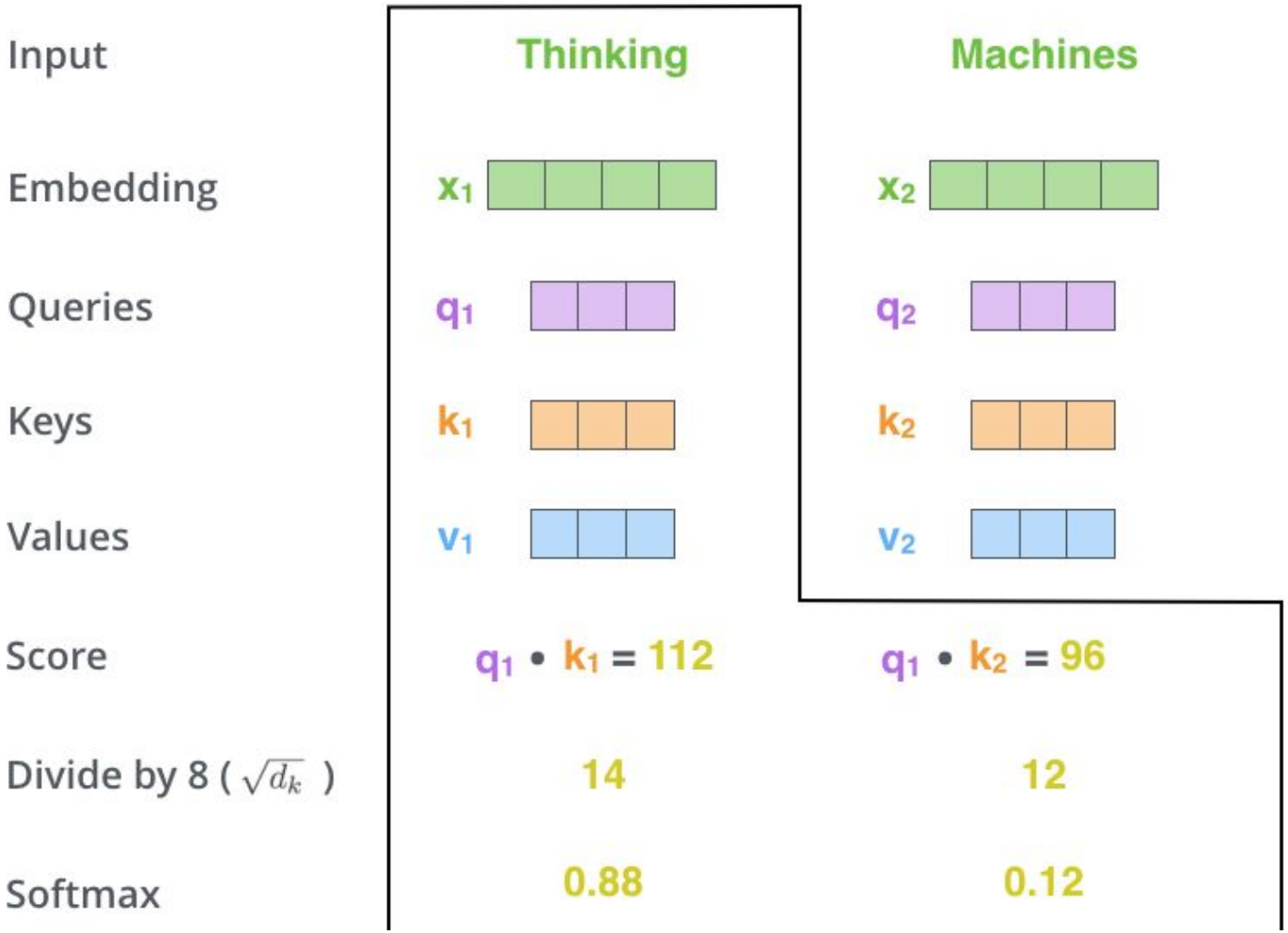

ENCODER #2
ENCODER #1
r1
r2
Feed Forward Neural Network
Feed Forward Neural Network
z1
z2
Self-Attention
x1
x2
Thinking
Machines

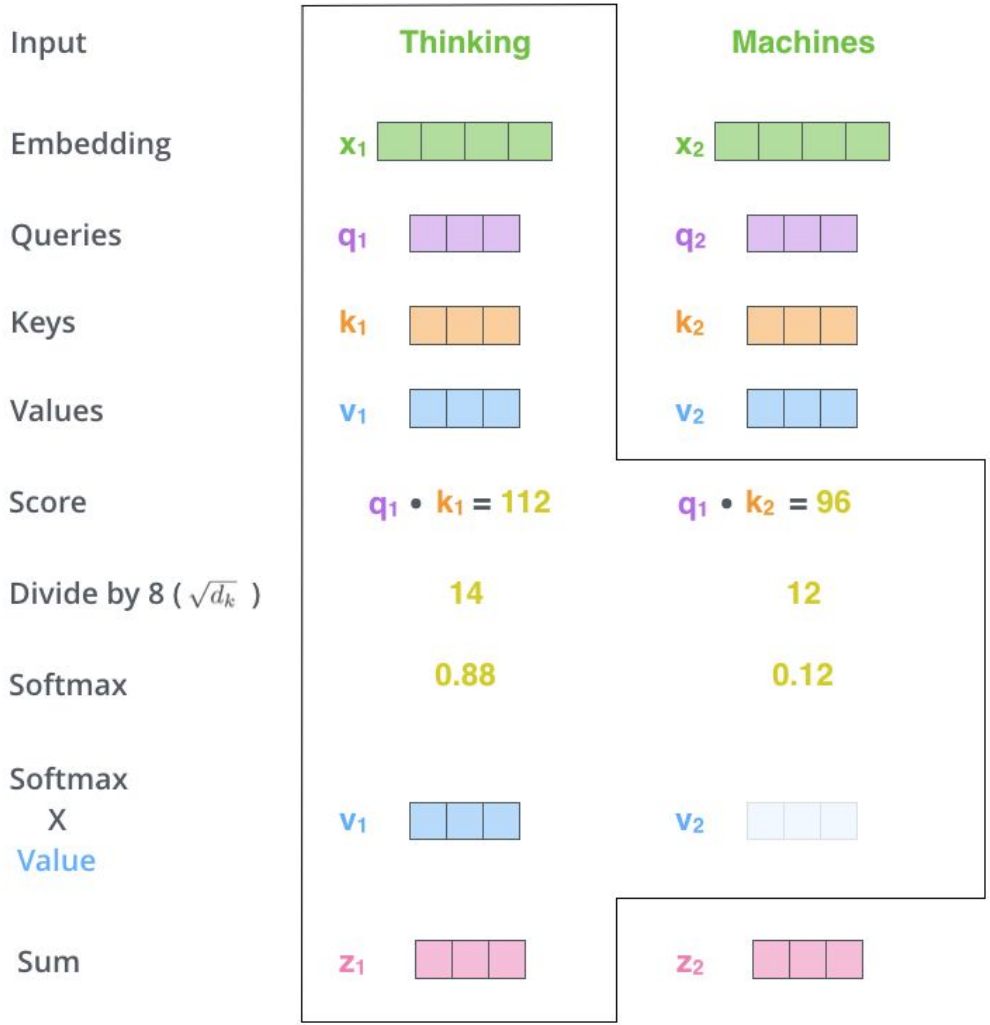
More on attention

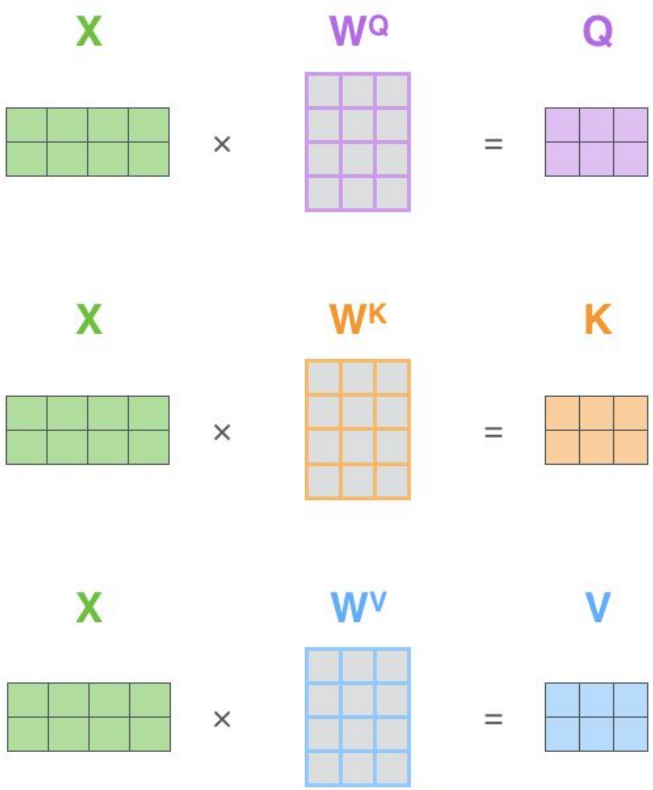

Layer: 5
Attention: Input - Input
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_

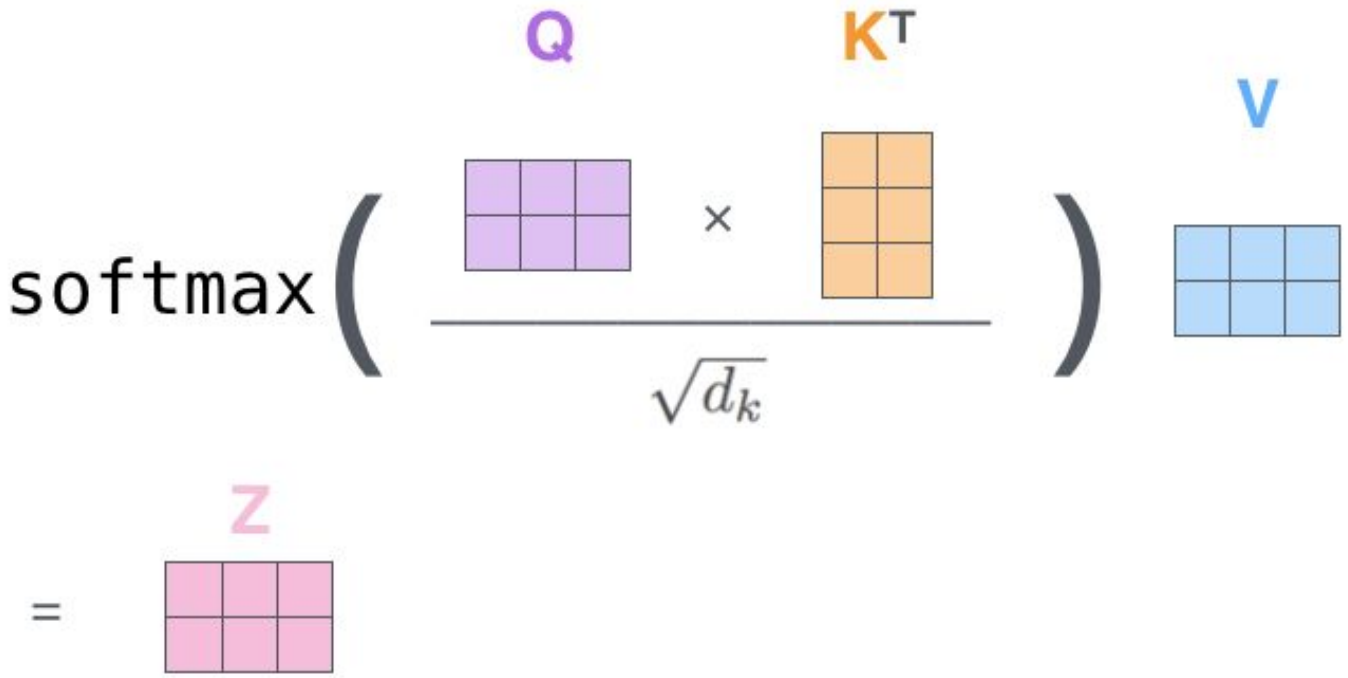

Input
Thinking
Machines
Embedding
X_1
X_2
Queries
q_1
q_2
W^Q
Keys
k_1
k_2
W^K
Values
v_1
v_2
W^V

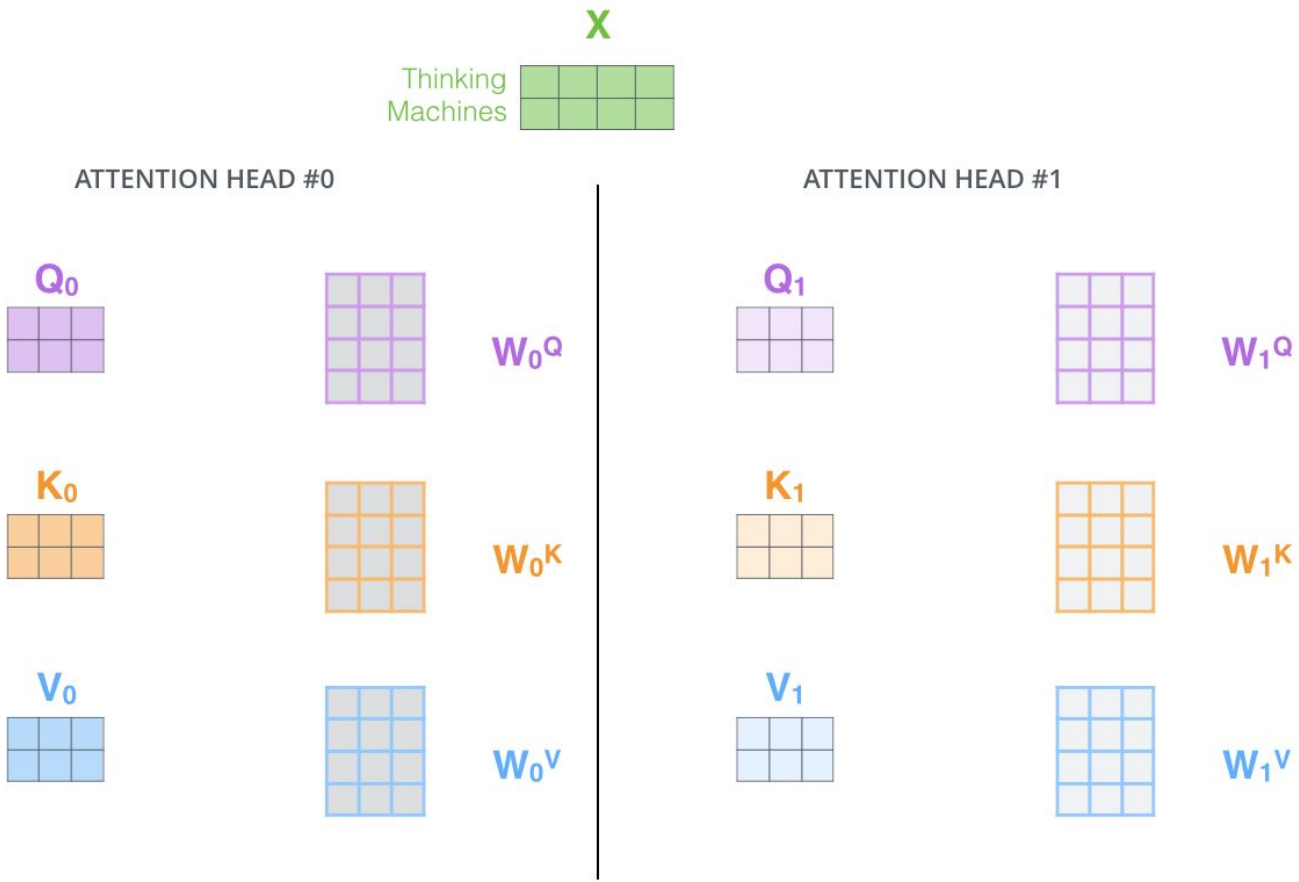

Input
Thinking
Machines
Embedding
x1
x2
Queries
q1
q2
Keys
k1
k2
Values
v1
v2
Score
q1 • k1 = 112
q1 • k2 = 96

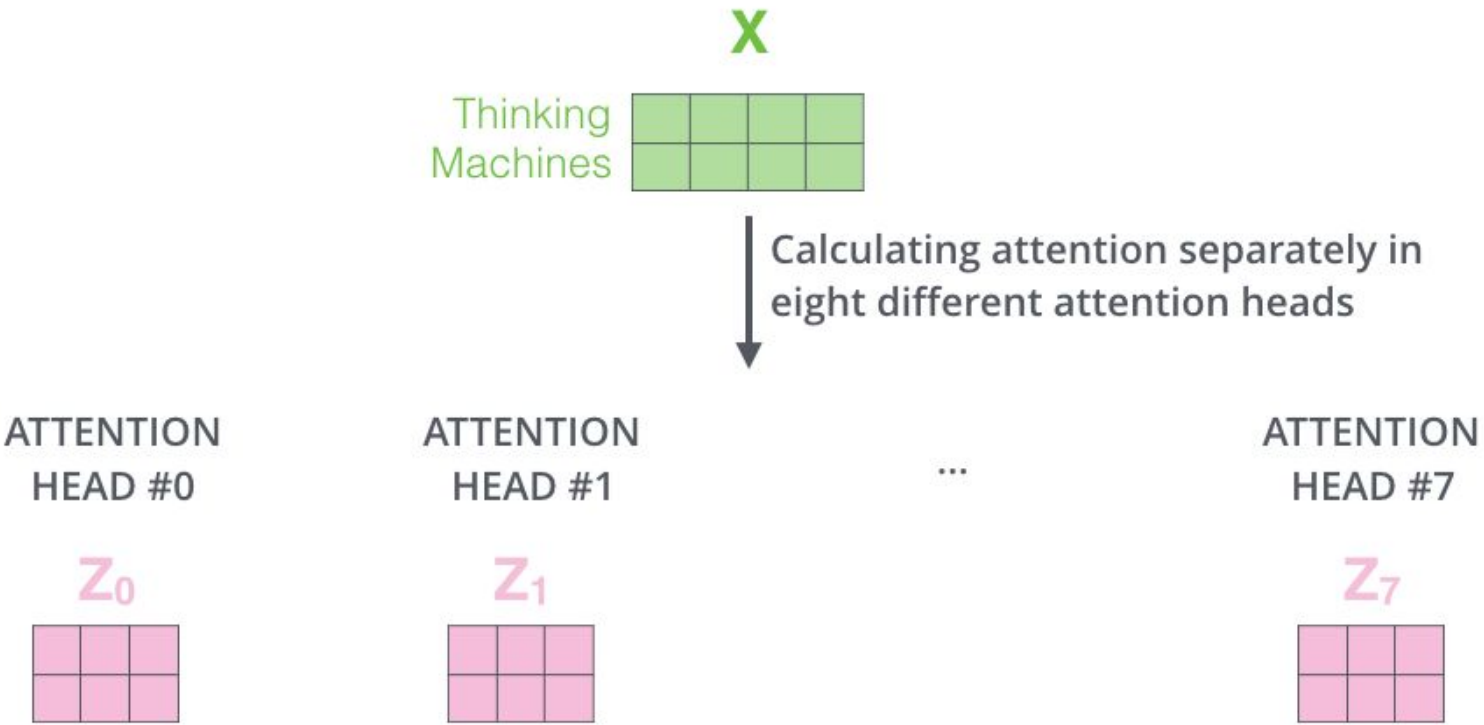

Input
Thinking
Machines
Embedding
x1
x2
Queries
q1
q2
Keys
k1
k2
Values
v1
v2
Score
q1 • k1 = 112
q1 • k2 = 96
Divide by 8 (√dk)
14
12
Softmax
0.88
0.12


Input
Thinking
Machines
Embedding
x1
x2
Queries
q1
q2
Keys
k1
k2
Values
v1
v2
Score
q1 • k1 = 112
q1 • k2 = 96
Divide by 8 (√dk)
14
12
Softmax
0.88
0.12
Softmax X Value
v1
v2
Sum
z1
z2


X
W^Q
Q
×
=
X
W^K
K
×
=
X
W^V
V
×
=


Q
K^T
V
softmax
×
$\sqrt{d_k}$
Z
=

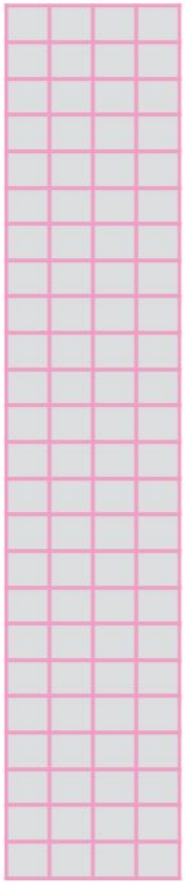

X
Thinking
Machines
ATTENTION HEAD #0
ATTENTION HEAD #1
Q0
K0
V0
W0Q
W0K
W0V
Q1
K1
V1
W1Q
W1K
W1V


X
Thinking
Machines
Calculating attention separately in
eight different attention heads
ATTENTION
HEAD #0
ATTENTION
HEAD #1
...
ATTENTION
HEAD #7
Z0
Z1
Z7

1) Concatenate all the attention heads

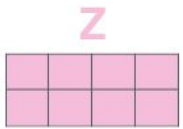
2) Multiply with a weight matrix W^O that was trained jointly with the model

X



3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN

=



1) This is our input sentence*

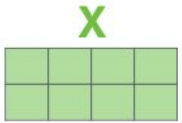
2) We embed each word*

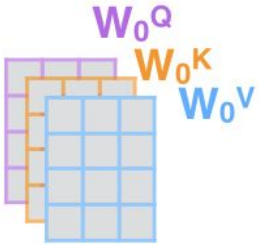
3) Split into 8 heads. We multiply X or R with weight matrices

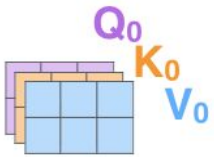
4) Calculate attention using the resulting Q/K/V matrices

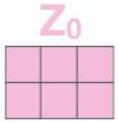
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

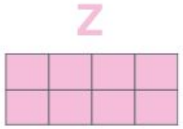
Thinking Machines



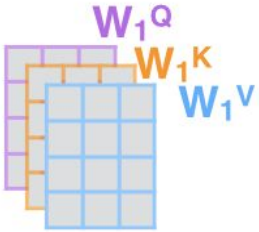


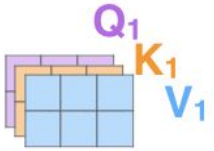


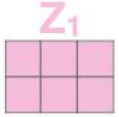




* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one



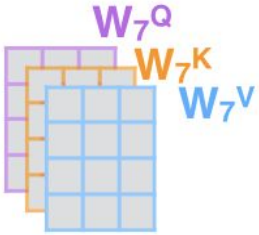


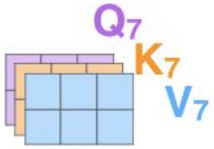


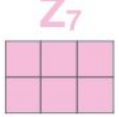
...

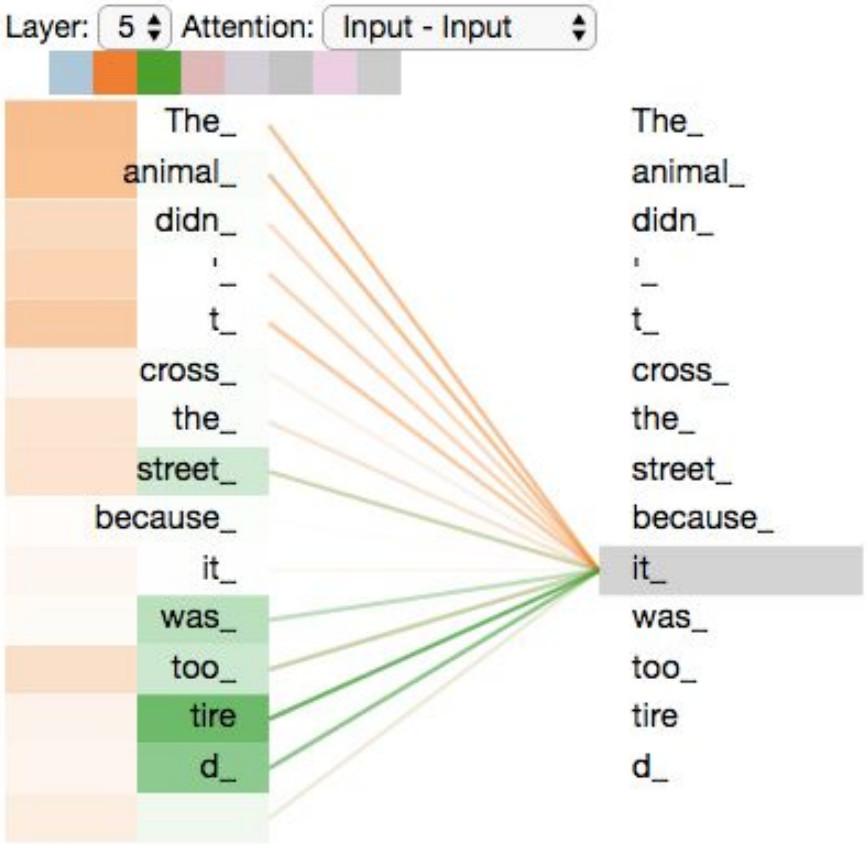
...

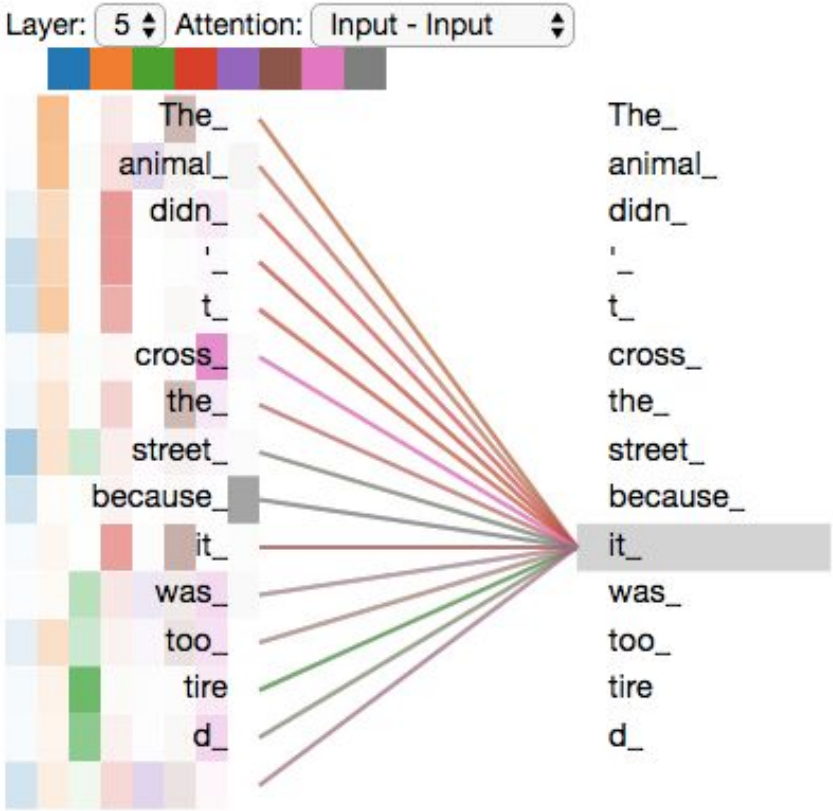
...



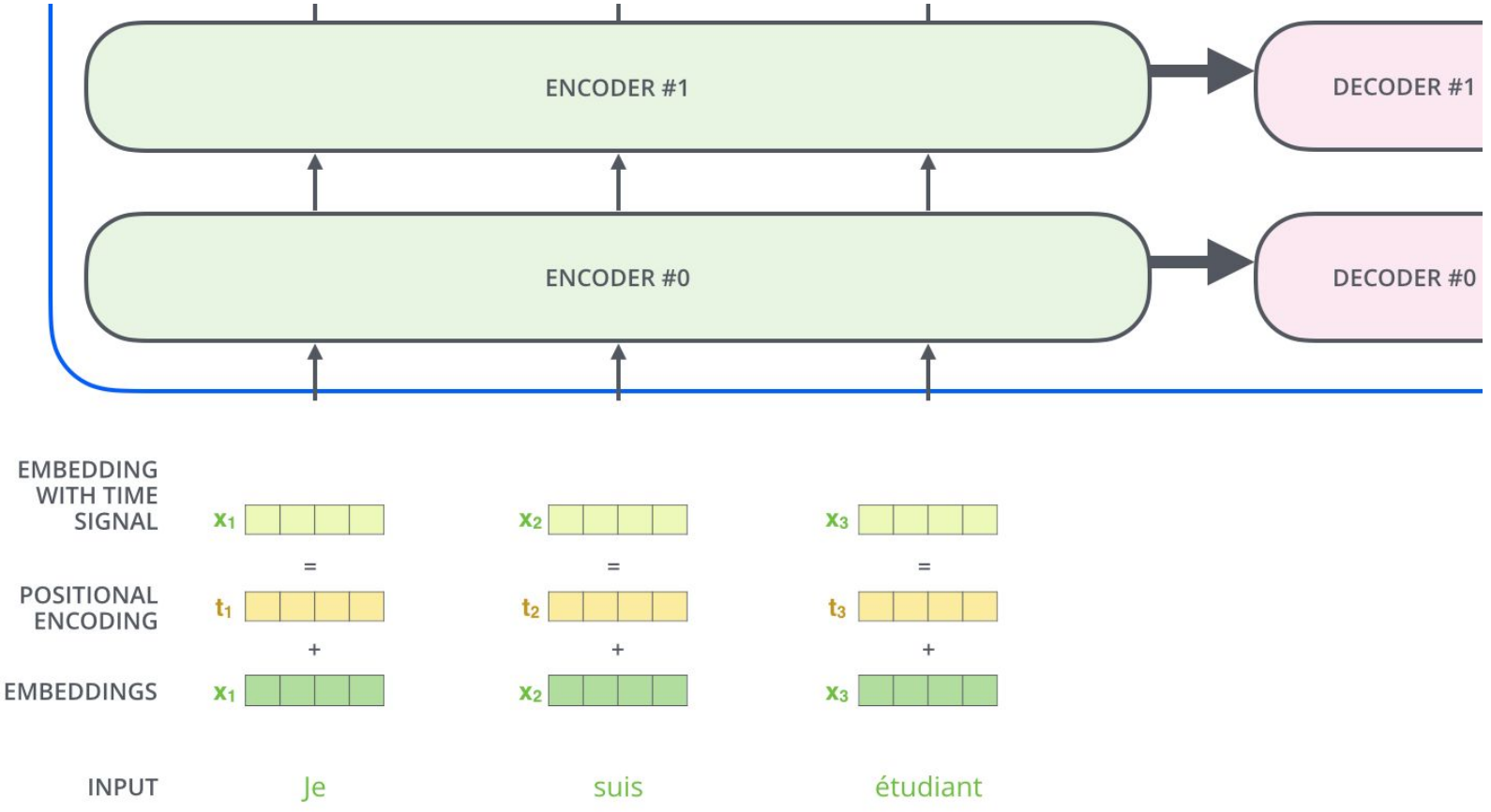


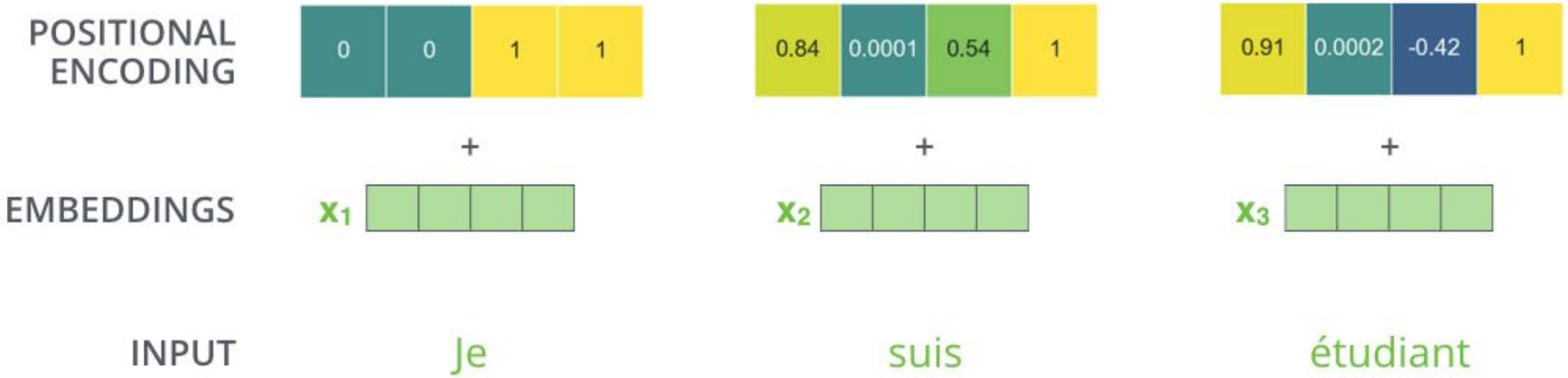


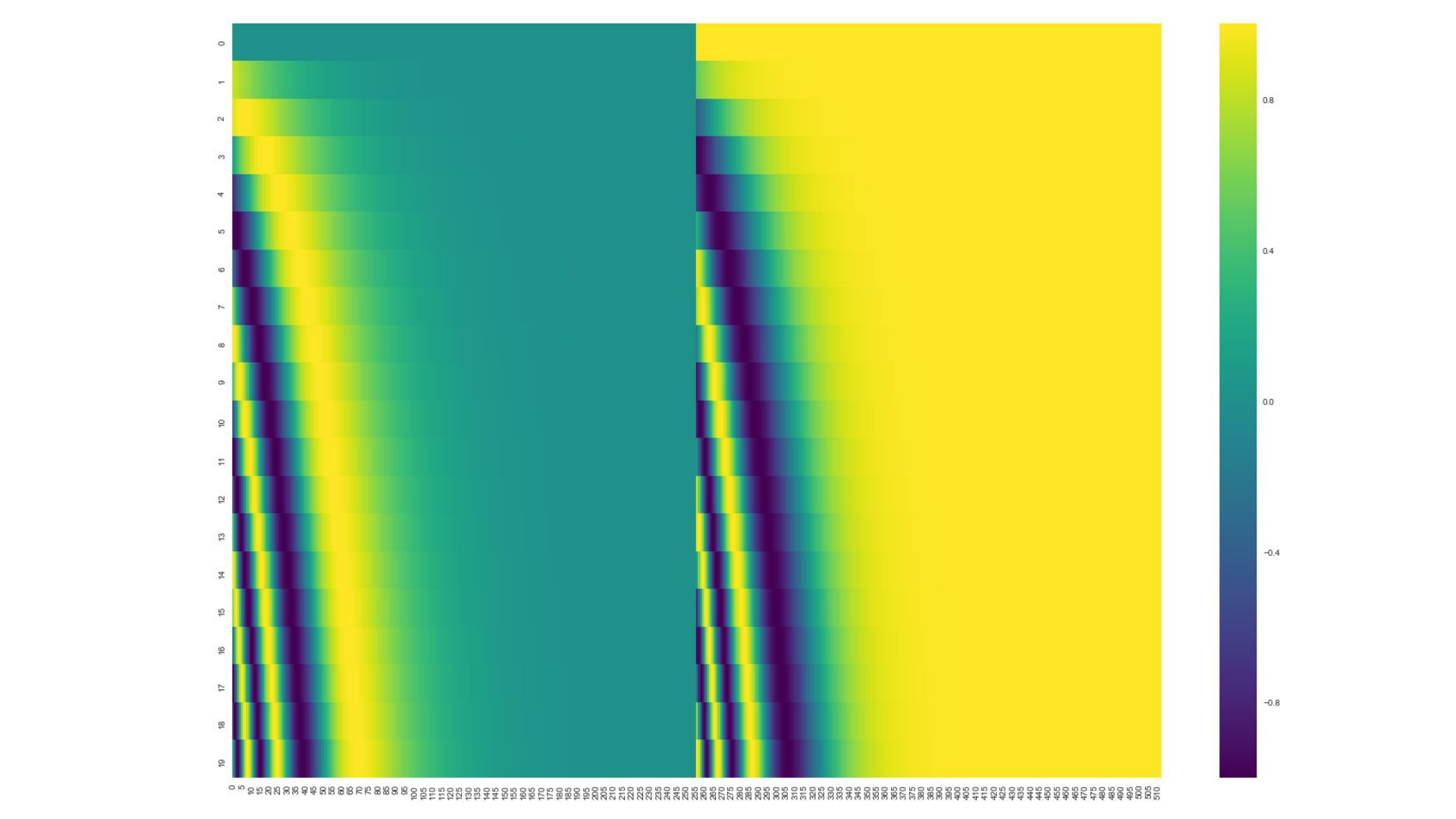

Layer: 5
Attention: Input - Input
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_

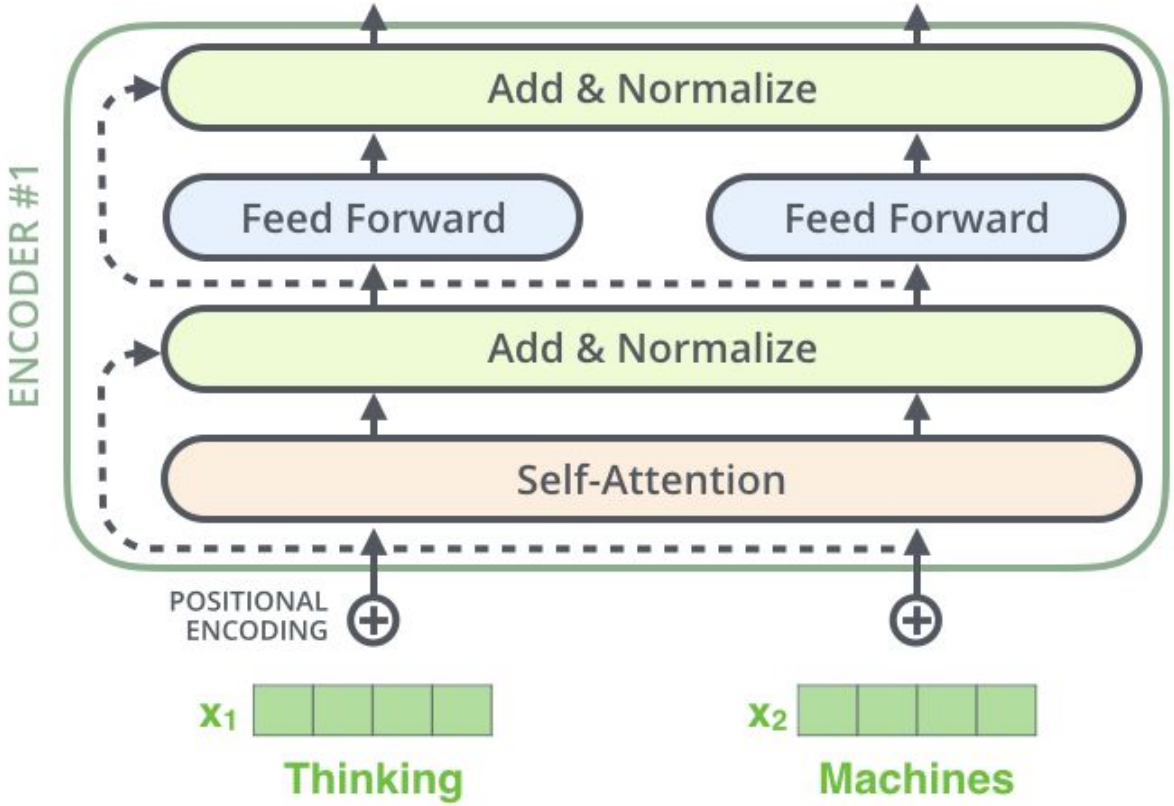

Layer: 5
Attention: Input - Input
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_

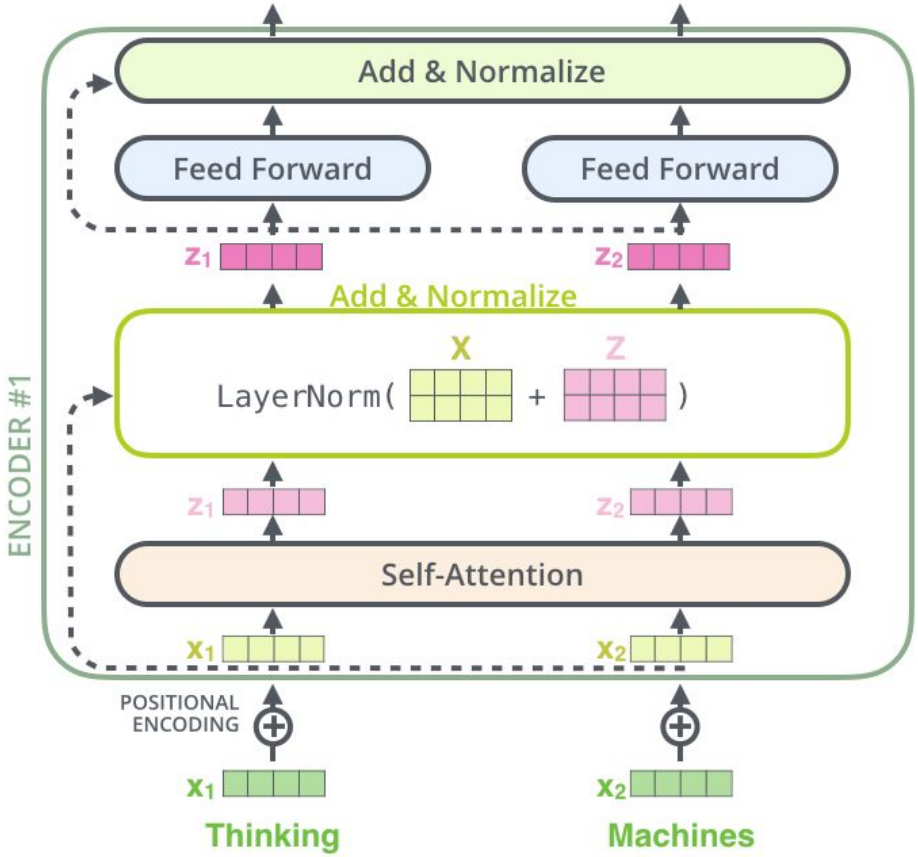
Positional encodings

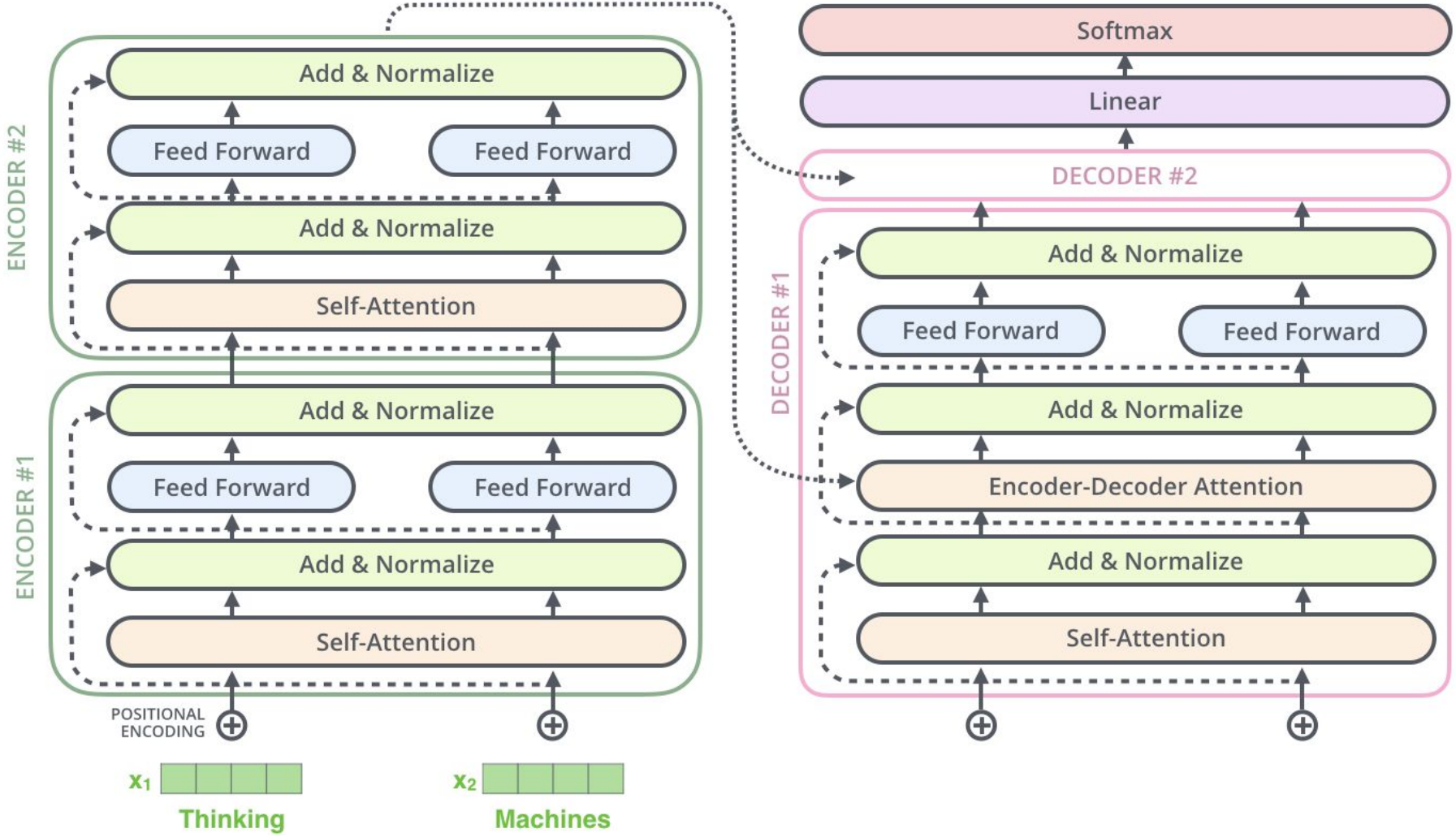

ENCODER #1
DECODER #1
ENCODER #0
DECODER #0
EMBEDDING WITH TIME SIGNAL
POSITIONAL ENCODING
EMBEDDINGS
INPUT
x1
x2
x3
t1
t2
t3
=
+
Je
suis
étudiant

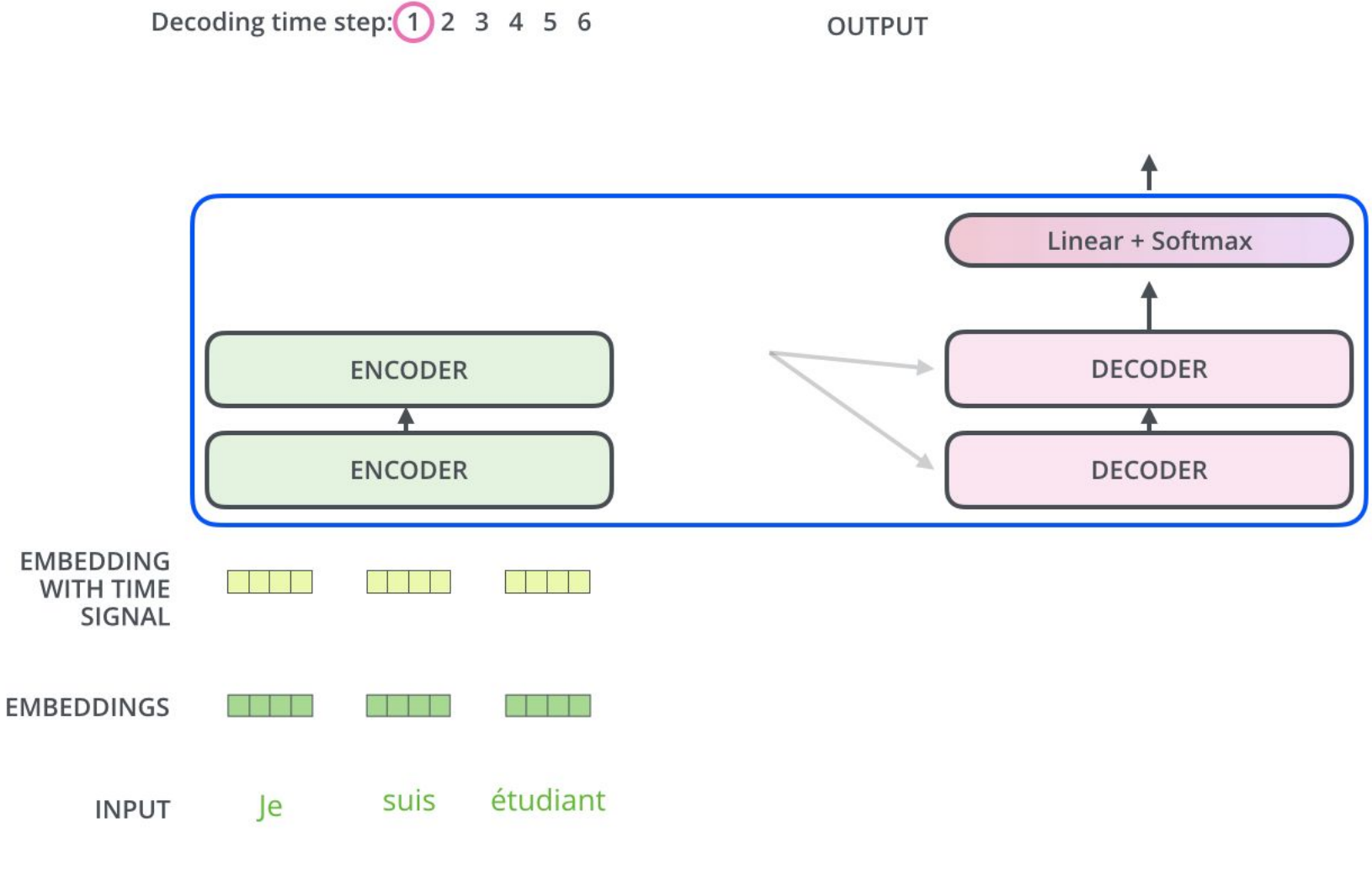

POSITIONAL ENCODING
0
0
1
1
0.84
0.0001
0.54
1
0.91
0.0002
-0.42
1
+
+
+
EMBEDDINGS
x1
x2
x3
INPUT
Je
suis
étudiant

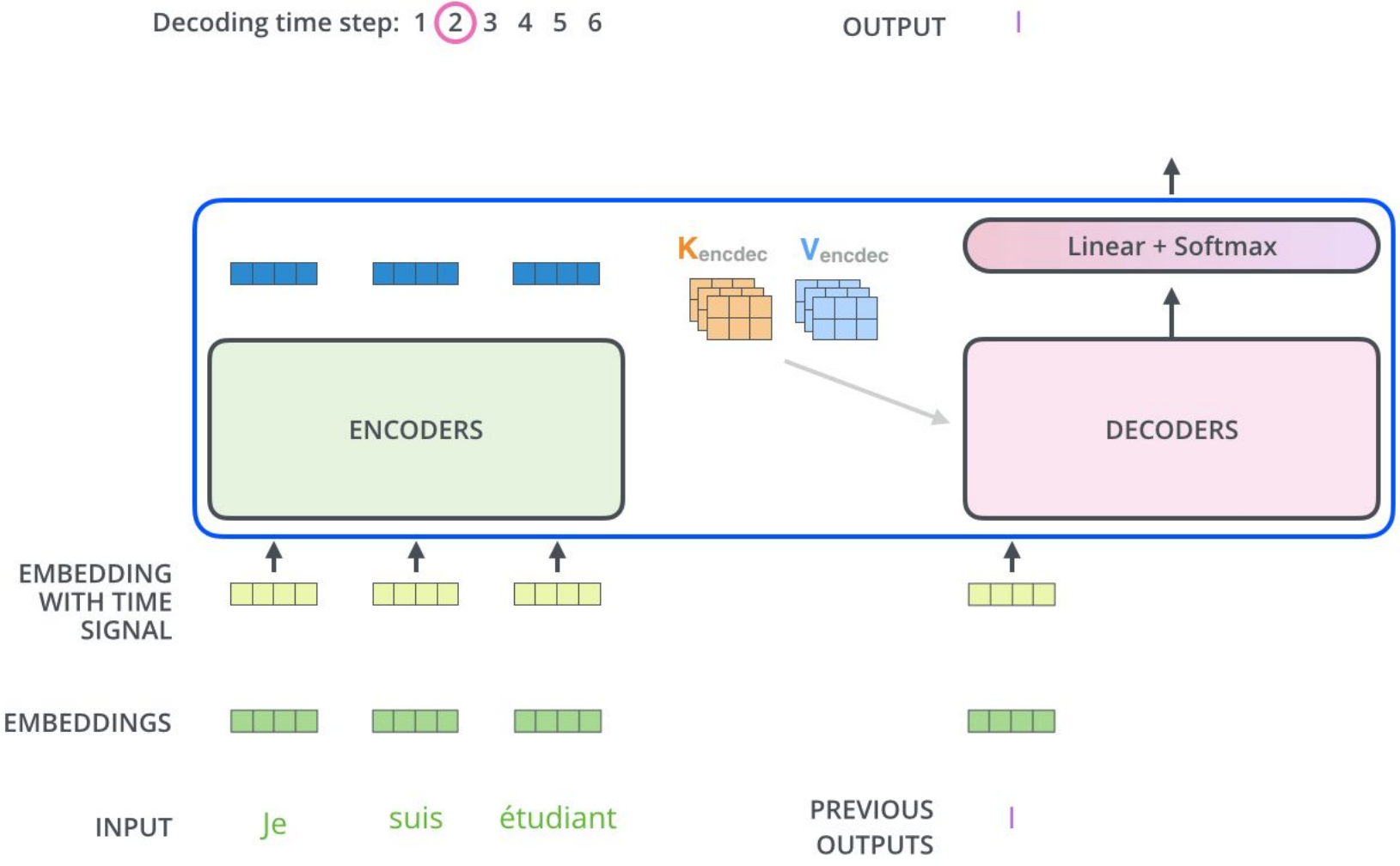

0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400 405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495 500 505 510
0.8
0.4
0.0
−0.4
−0.8


ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines


Add & Normalize
Feed Forward
Feed Forward
z1
z2
Add & Normalize
X
Z
LayerNorm(+)
ENCODER #1
z1
z2
Self-Attention
x1
x2
POSITIONAL ENCODING
x1
x2
Thinking
Machines


Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines


Decoding time step: 1 2 3 4 5 6
OUTPUT
Linear + Softmax
ENCODER
ENCODER
DECODER
DECODER
EMBEDDING WITH TIME SIGNAL
EMBEDDINGS
INPUT
Je
suis
étudiant

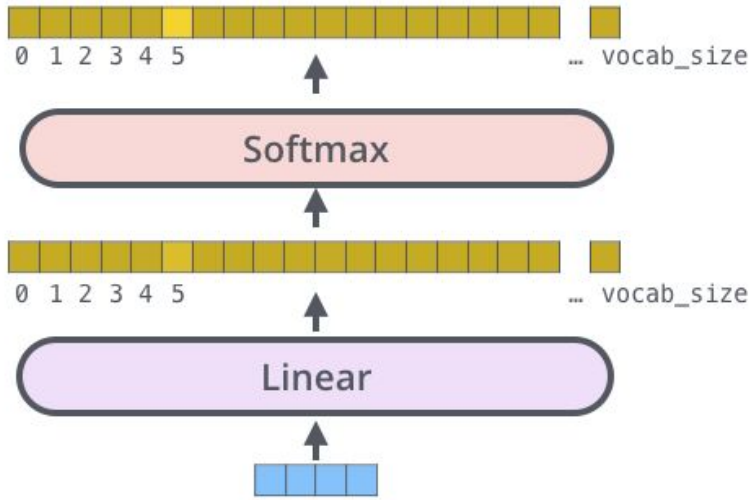

Decoding time step: 1 2 3 4 5 6
OUTPUT
I
Linear + Softmax
K encdec
V encdec
ENCODERS
DECODERS
EMBEDDING WITH TIME SIGNAL
EMBEDDINGS
INPUT
Je
suis
étudiant
PREVIOUS OUTPUTS
I

Which word in our vocabulary is associated with this index?

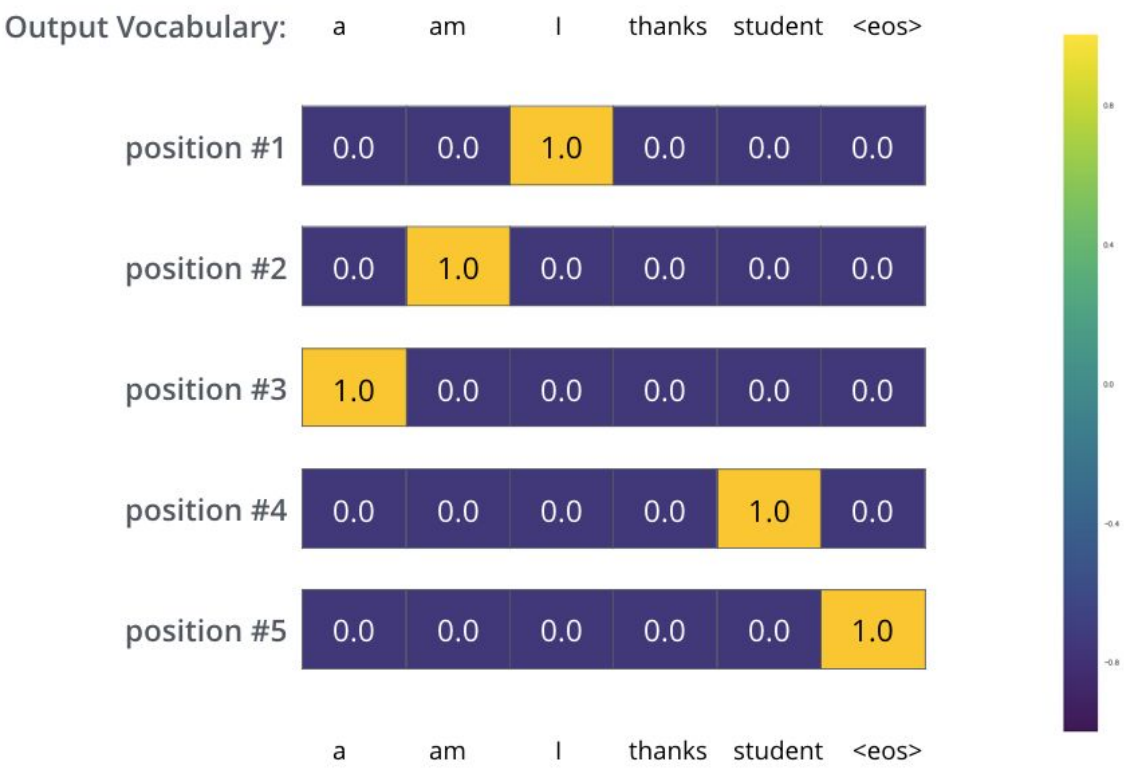
am

Get the index of the cell with the highest value **(argmax)**

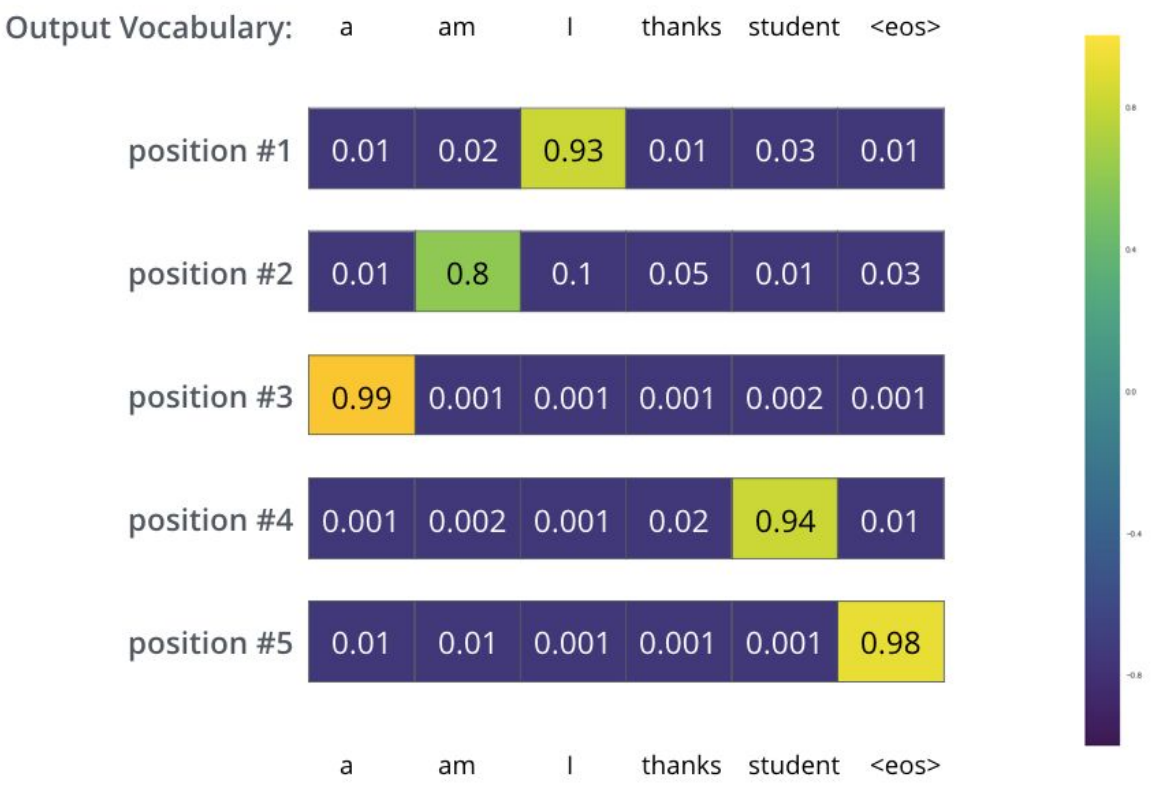
5



Target Model Outputs



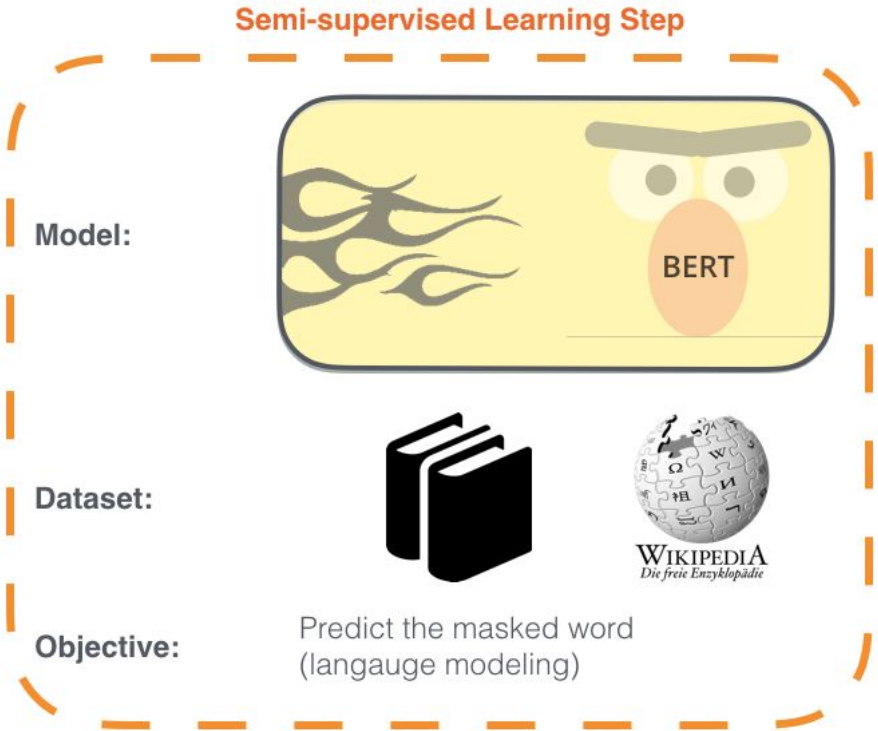
Trained Model Outputs

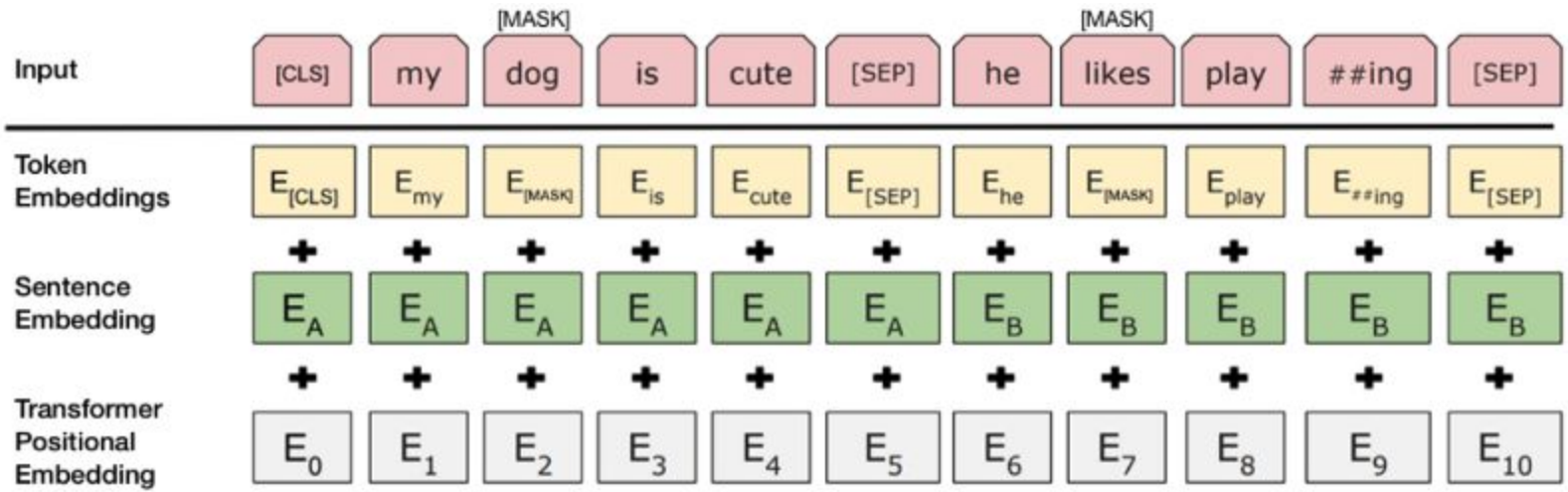


Bert

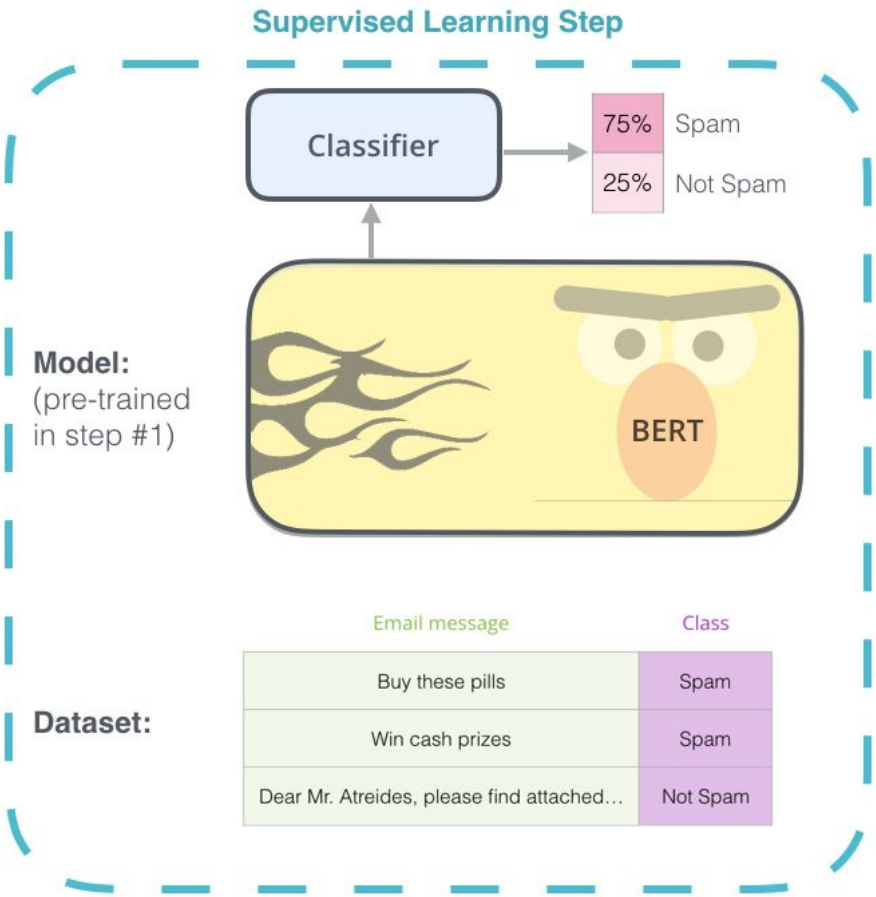
1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

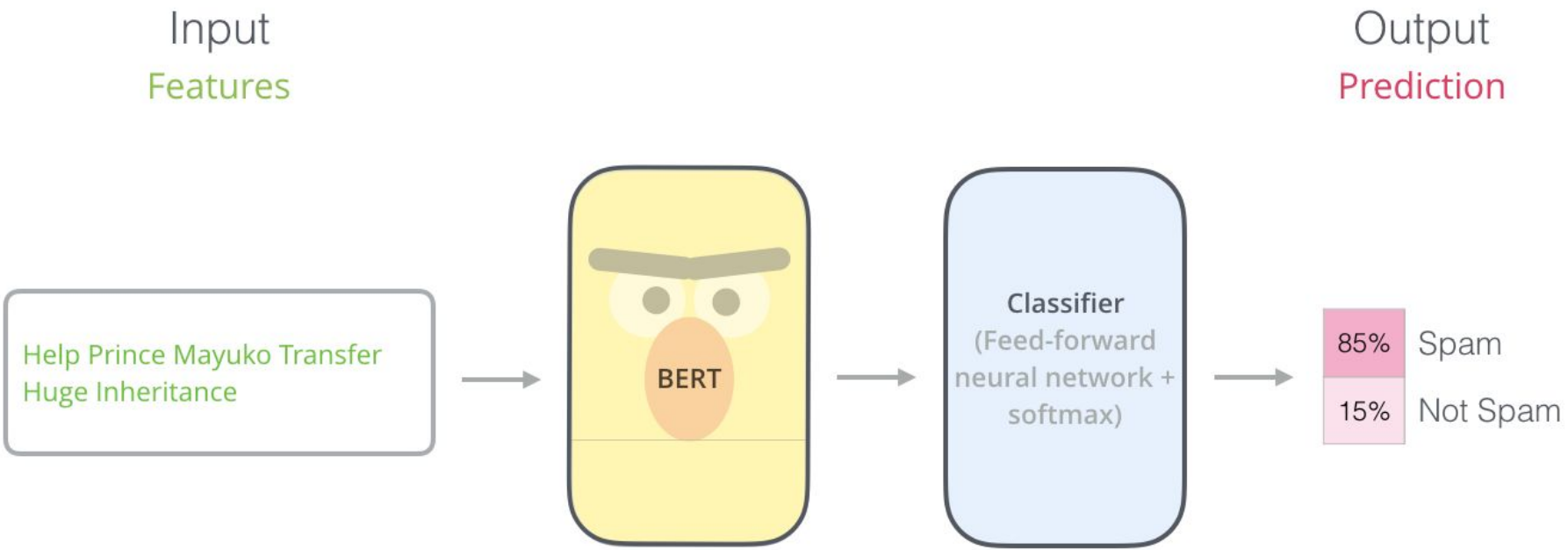
The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.




Input
[CLS]
my
[MASK]
dog
is
cute
[SEP]
he
[MASK]
likes
play
##ing
[SEP]
Token Embeddings
E[CLS]
Emy
E[MASK]
Eis
Ecute
E[SEP]
Ehe
E[MASK]
Eplay
E##ing
E[SEP]
Sentence Embedding
EA
EB
Transformer Positional Embedding
E0
E1
E2
E3
E4
E5
E6
E7
E8
E9
E10

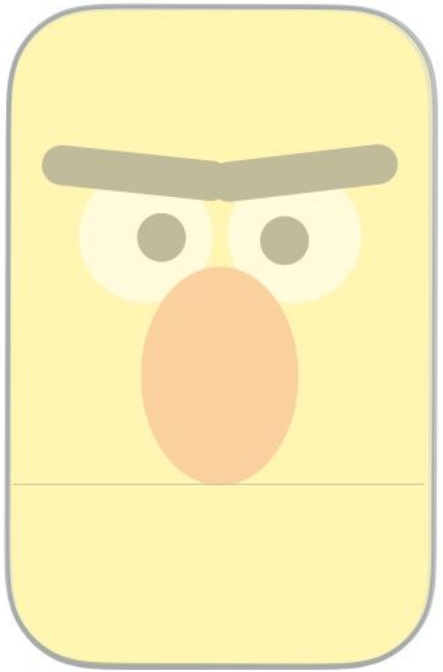
2 - Supervised training on a specific task with a labeled dataset.



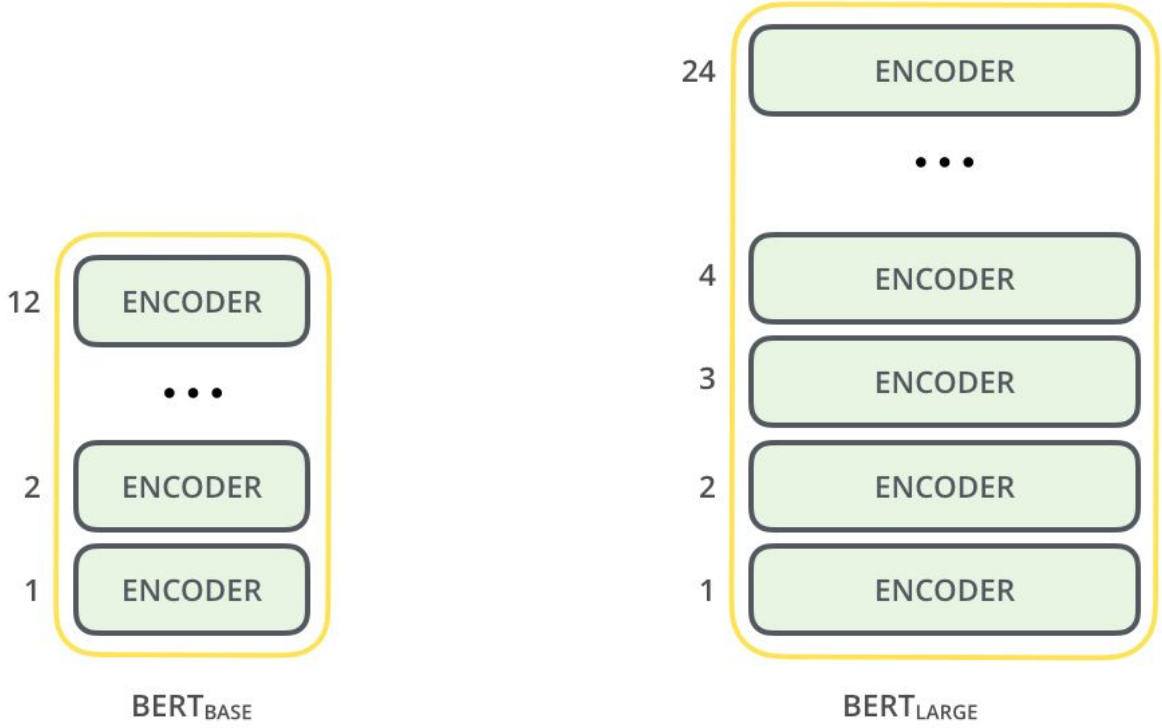

Input
Features
Output
Prediction
Help Prince Mayuko Transfer
Huge Inheritance
BERT
Classifier
(Feed-forward
neural network +
softmax)
85%
Spam
15%
Not Spam

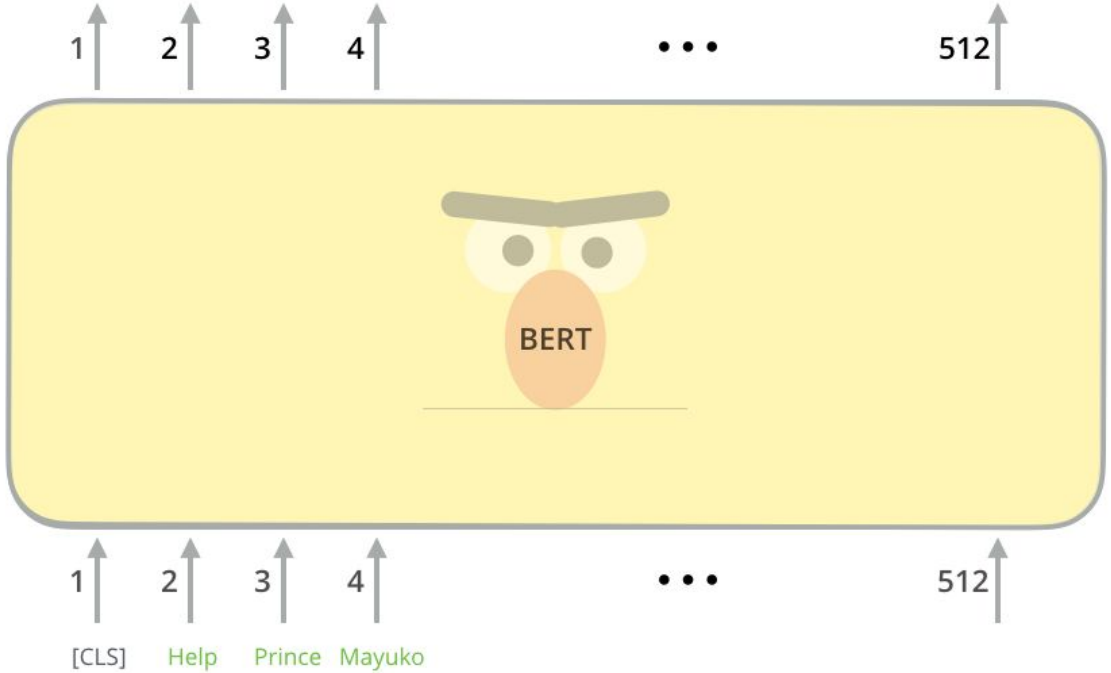
Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

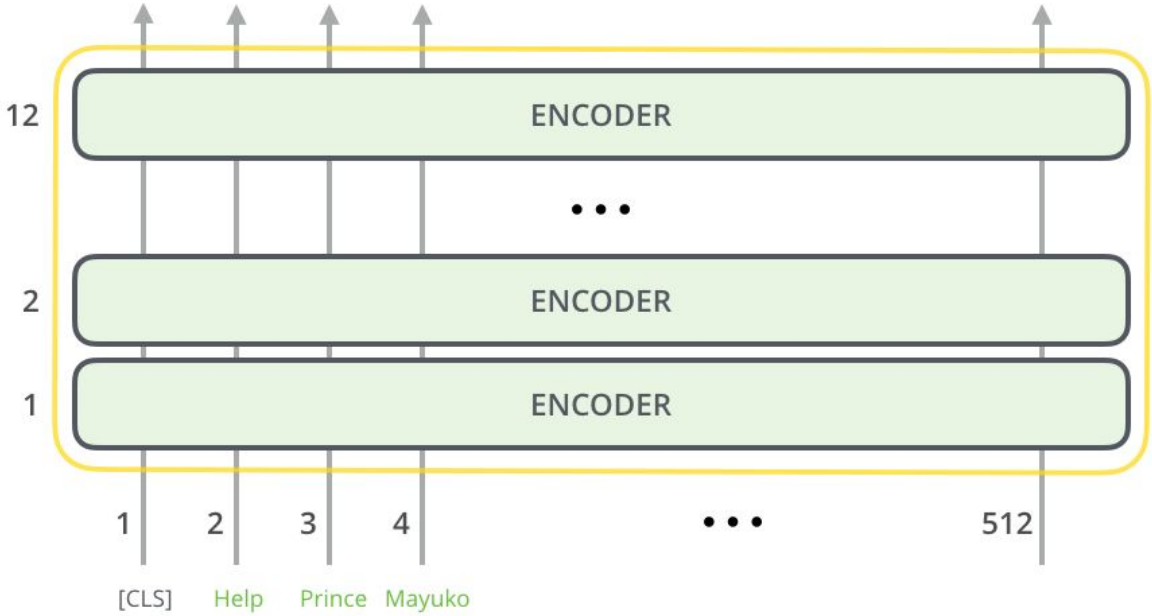
BERT$_{BASE}$



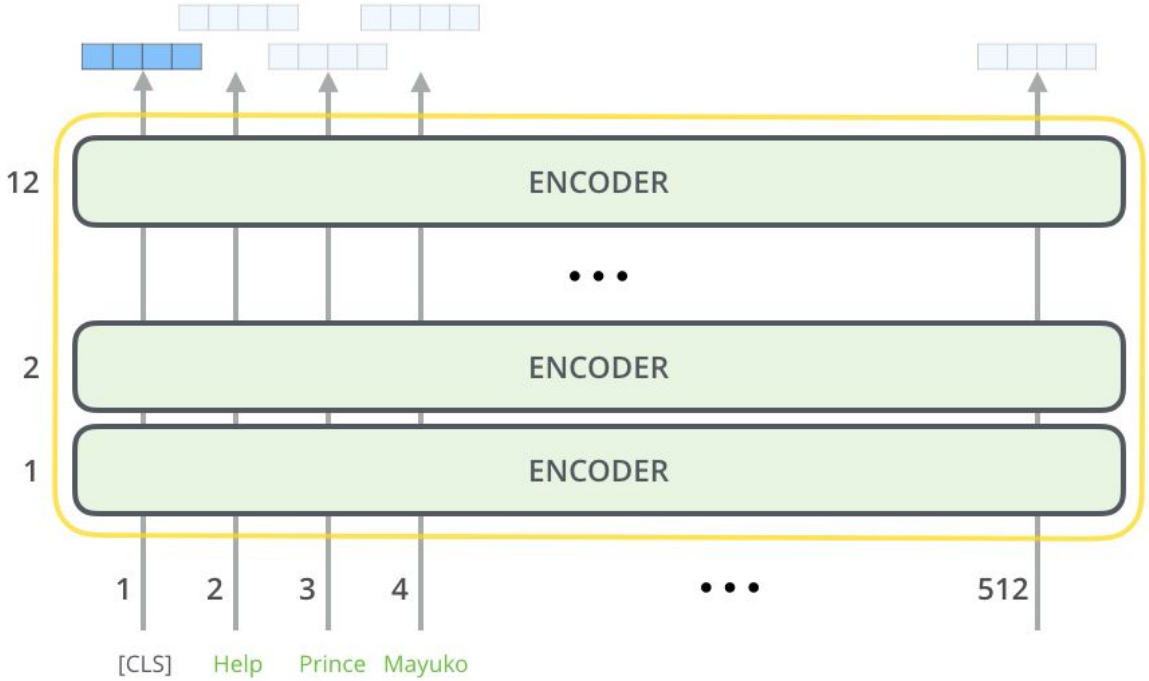
BERT$_{LARGE}$


12
ENCODER
2
ENCODER
1
ENCODER
BERT BASE
24
ENCODER
4
ENCODER
3
ENCODER
2
ENCODER
1
ENCODER
BERT LARGE

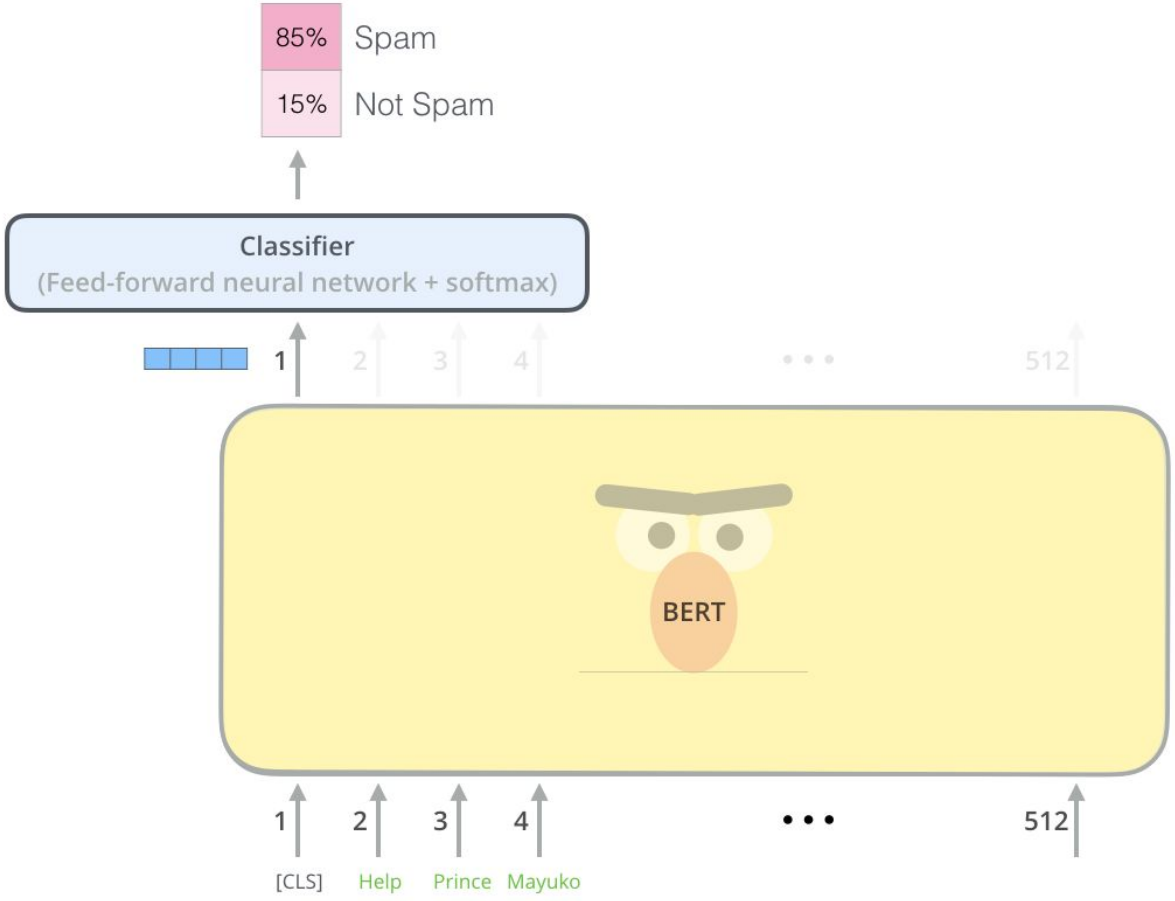

1
2
3
4
• • •
512
BERT
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko

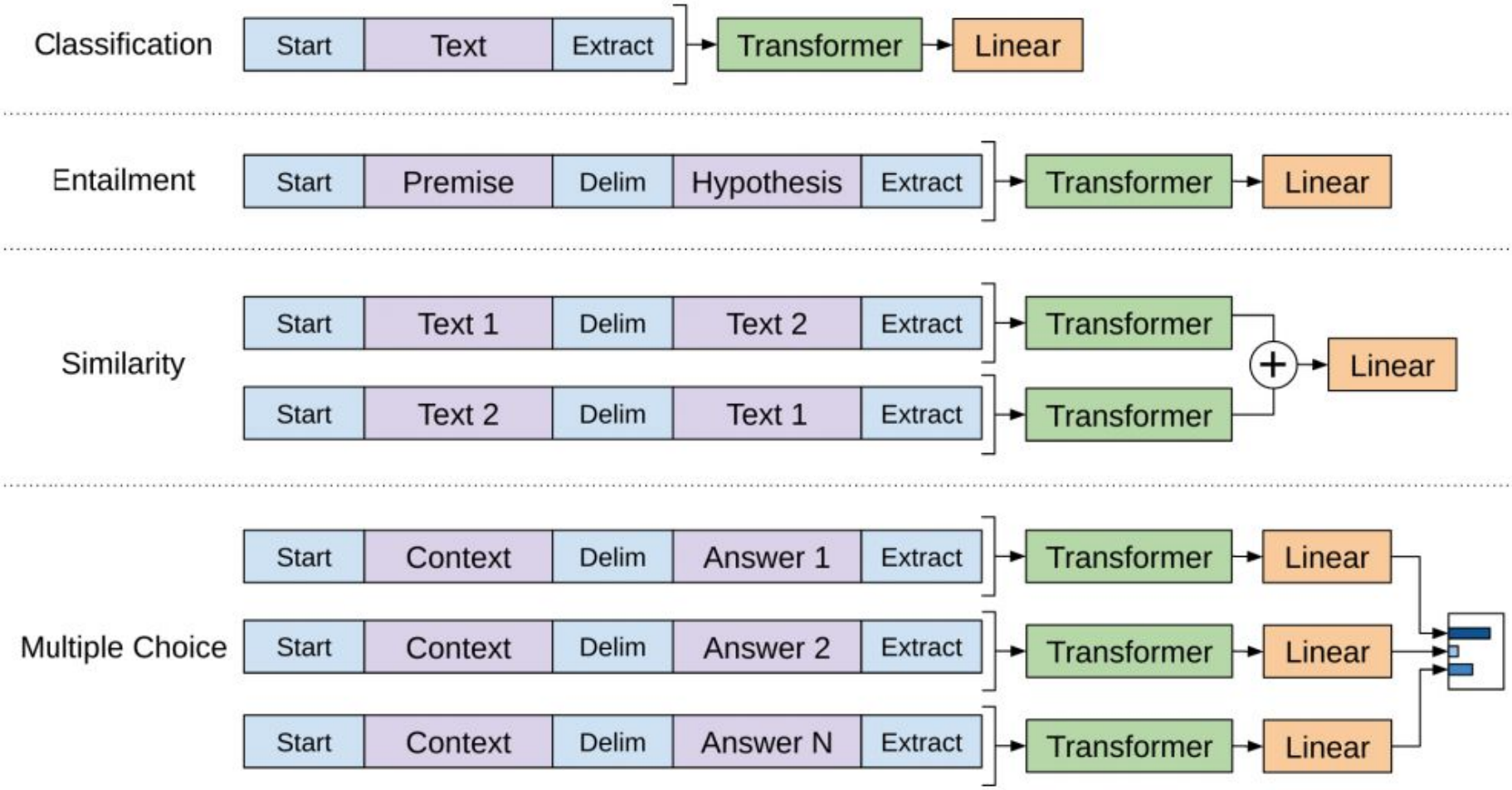

12
ENCODER
2
ENCODER
1
ENCODER
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

BERT

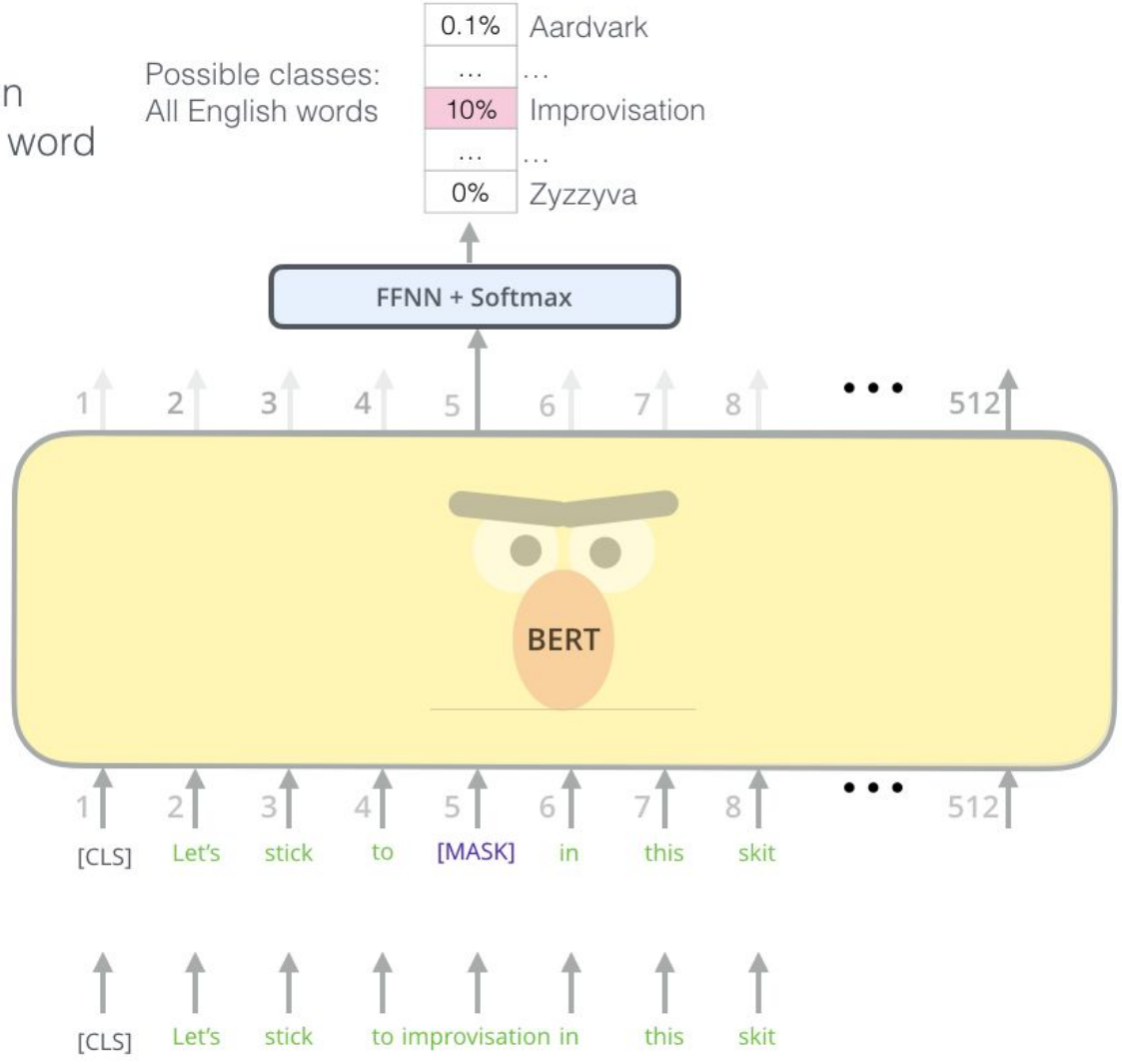

12
ENCODER
2
ENCODER
1
ENCODER
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

BERT

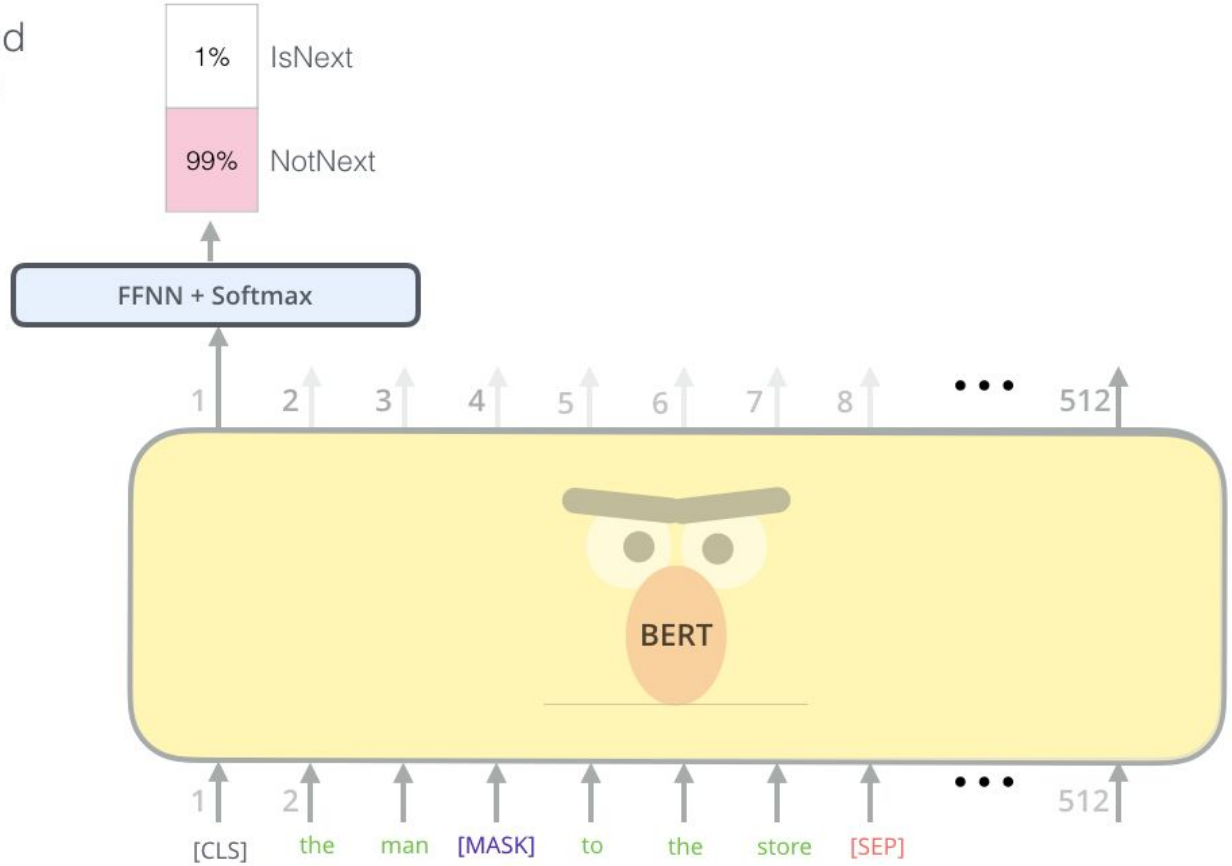

85%
Spam
15%
Not Spam
Classifier
(Feed-forward neural network + softmax)
1
2
3
4
512
BERT
1
2
3
4
512
[CLS]
Help
Prince
Mayuko


Classification
Start
Text
Extract
Transformer
Linear
Entailment
Start
Premise
Delim
Hypothesis
Extract
Transformer
Linear
Similarity
Start
Text 1
Delim
Text 2
Extract
Transformer
Start
Text 2
Delim
Text 1
Extract
Transformer
Linear
Multiple Choice
Start
Context
Delim
Answer 1
Extract
Transformer
Linear
Start
Context
Delim
Answer 2
Extract
Transformer
Linear
Start
Context
Delim
Answer N
Extract
Transformer
Linear

Use the output of the masked word's position to predict the masked word



Predict likelihood that sentence B belongs after sentence A

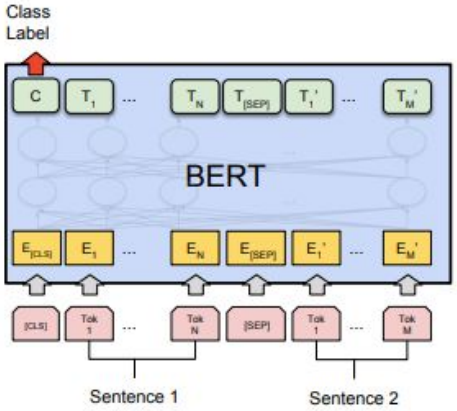


Tokenized Input

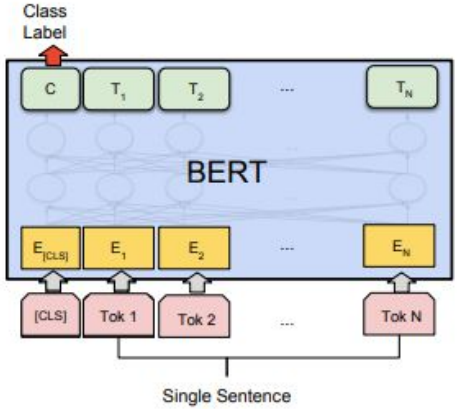
Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

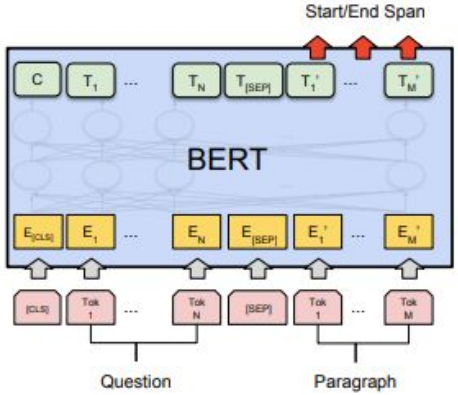
Sentence A Sentence B



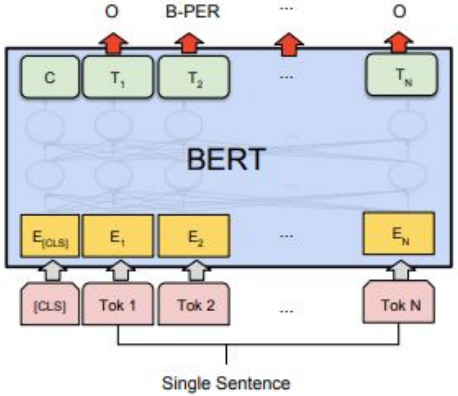
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA

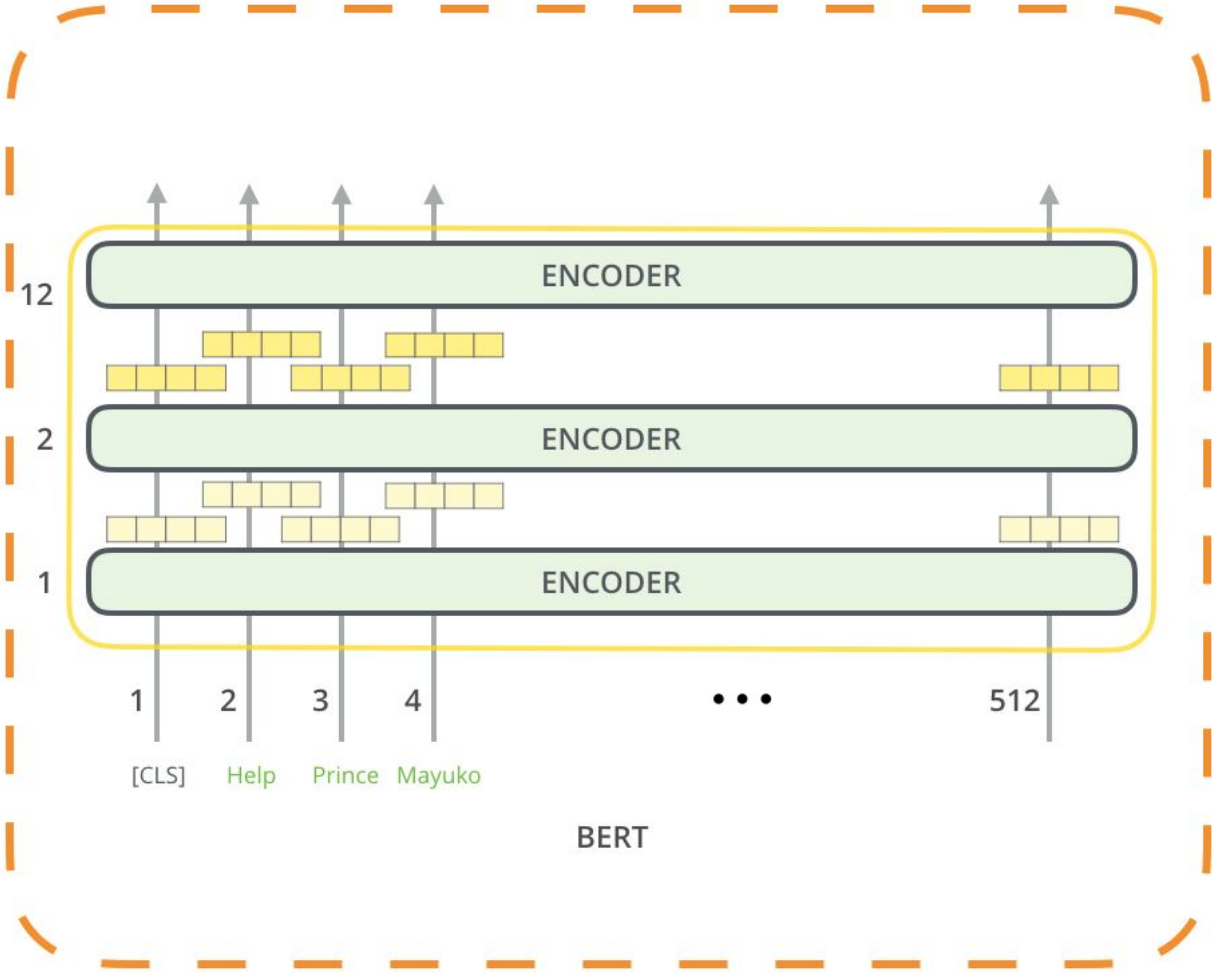


(c) Question Answering Tasks: SQuAD v1.1

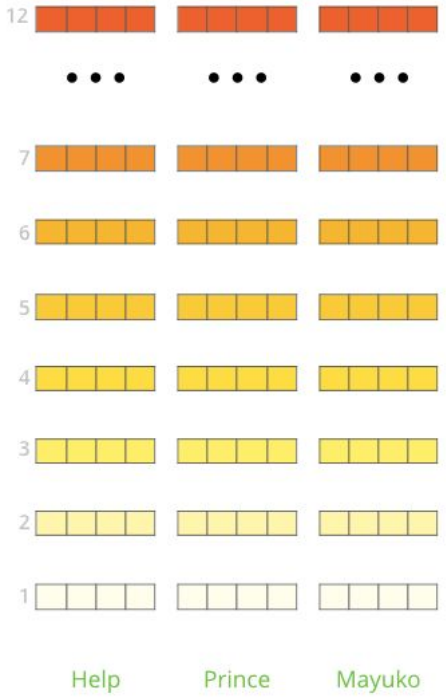


(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Generate Contexualized Embeddings



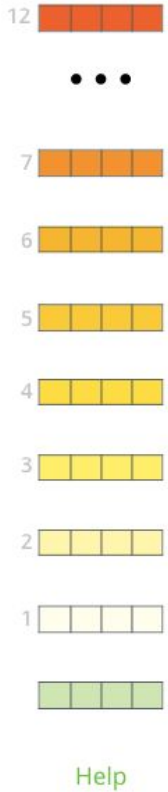
The output of each encoder layer along each token's path can be used as a feature representing that token.

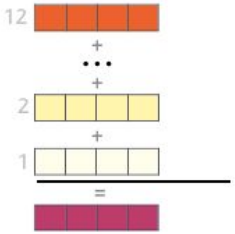
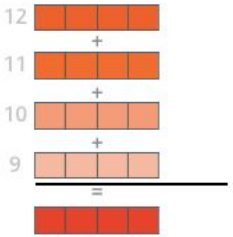
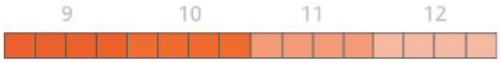


But which one should we use?

What is the best contextualized embedding for "Help" in that context?

For named-entity recognition task CoNLL-2003 NER



		Dev F1 Score
First Layer		91.0
Last Hidden Layer		94.9
Sum All 12 Layers		95.5
Second-to-Last Hidden Layer		95.6
Sum Last Four Hidden		95.9
Concat Last Four Hidden		96.1

Эволюция инструментов

NLU

NLU

В чем сдвиг парадигмы?

Методы получения NLU-моделей

- Skip-Gram (CBOW)
- Language Modeling
- Masking
- Skip-thoughts
- Multi task
- Autoencoder

NLU

Хроника появления решений

#МИСиС

2013

Эмбеддинги слов

+ Моделируют язык
+ Являются хорошими признаками слов

— Не являются алгоритмом для эмбеддинга сразу нескольких слов (текста)
— Недостаточно выразительны, происходит смешение контекстов

Проблема ограниченной выразительности

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML plugin, zip, javascript
apple	iphone, android, mac, microsoft samsung, phone, galaxy, touch
date	registration, join, location, from changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow head, ride, ball, dead
mass	energy, effect, impact, movement potential, military, weight, society exercise, lower

2014

Методы работы с текстами на LSTM

\+ Позволяют работать с текстами, как с последовательностями

— Работают достаточно медленно
— Требуют большого количества данных
— Плохо работают на достаточно длинных последовательностях

http://proceedings.mlr.press/v32/le14.pdf

2015

Методы работы с текстами на GRU, CNN

+ Позволяют работать с текстами как с последовательностями
+ Работают быстрее LSTM

— Требуют значительного количества данных
— Плохо работают на достаточно длинных последовательностях

https://papers.nips.cc/paper/5550-convolutional-neural-network-architectures-for-matching-natural-language-sentences.pdf, http://papers.nips.cc/paper/5782-character-level-convolutional-networks-fo

2016

Attention и дополненные LSTM/GRU

+ Позволяют работать с текстами, как с последовательностями
+ Хорошо работают на длинных текстах

— Требуют значительного количества данных

http://www.aclweb.org/anthology/P16-2034

2017

Transformer

+ Побил по качеству многие известные алгоритмы
+ Не зависит от предобученных эмбеддингов
+ Моделирует тексты более естественным образом

— Требует много данных

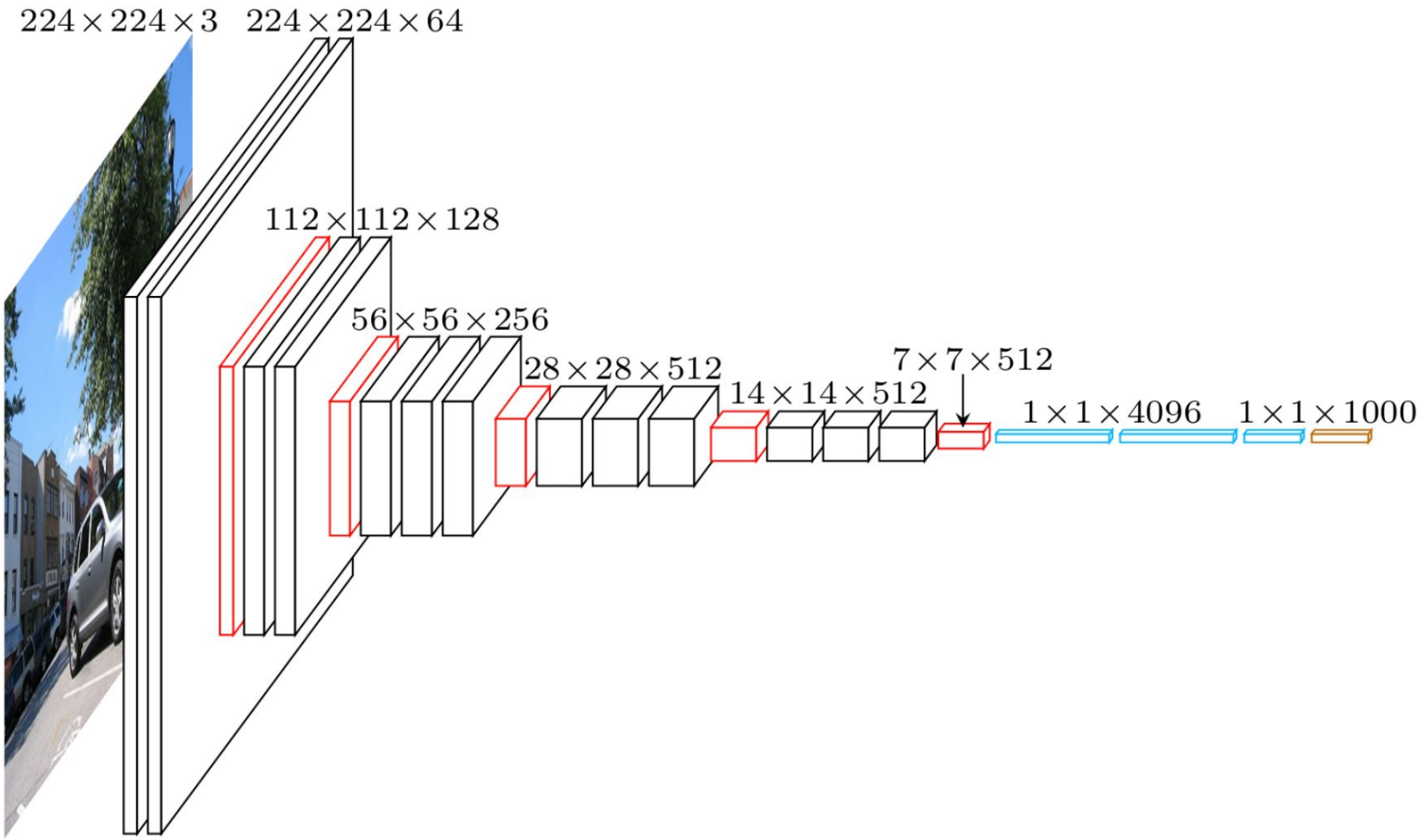
https://arxiv.org/abs/1706.03762, http://jalammar.github.io/illustrated-transformer/,
https://mchromiak.github.io/articles/2017/Sep/12/Transformer-Attention-is-all-you-need/#.XIWlzBNKjOR

2018

Transfer Learning 2: ULMfit

+ Почти не требует размеченных данных

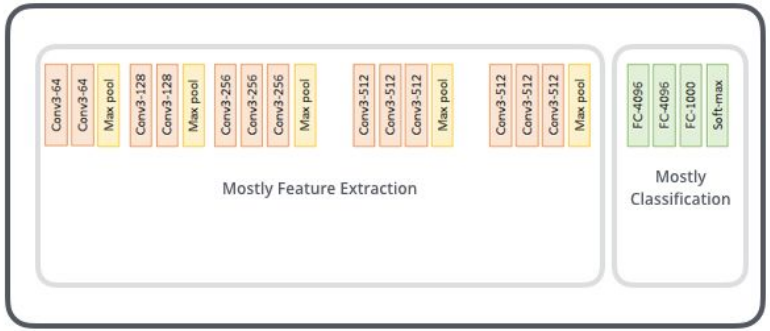
https://arxiv.org/abs/1801.06146


224 × 224 × 3
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000

Input

Features

VGG-16



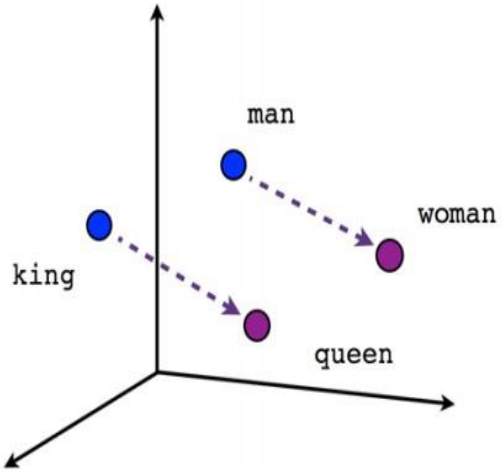
Output

Prediction

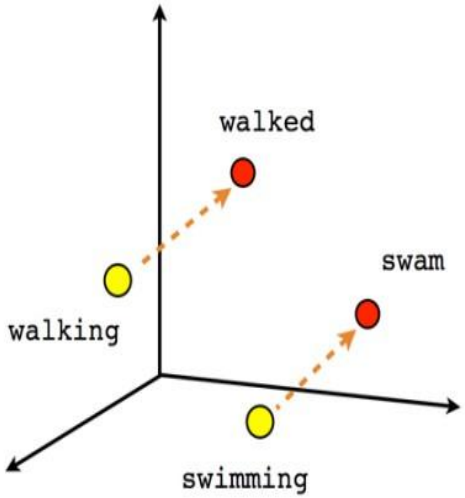
0.2%	Kit fox
0.1%	English setter
95%	Egyptian cat
1%	Great Dane
	…
0%	Hotdog

2013

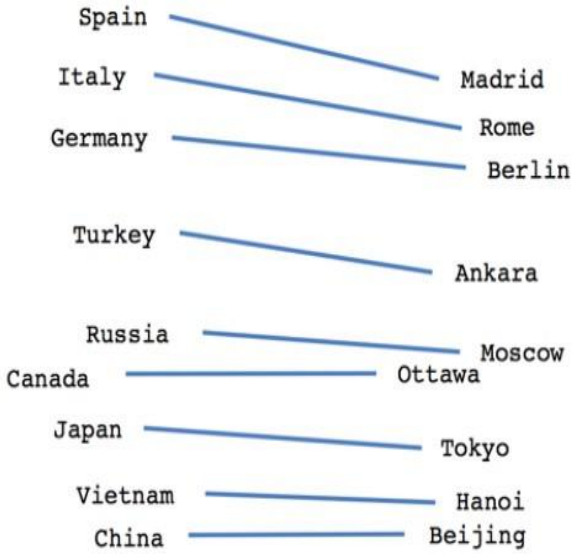
Эмбеддинги слов



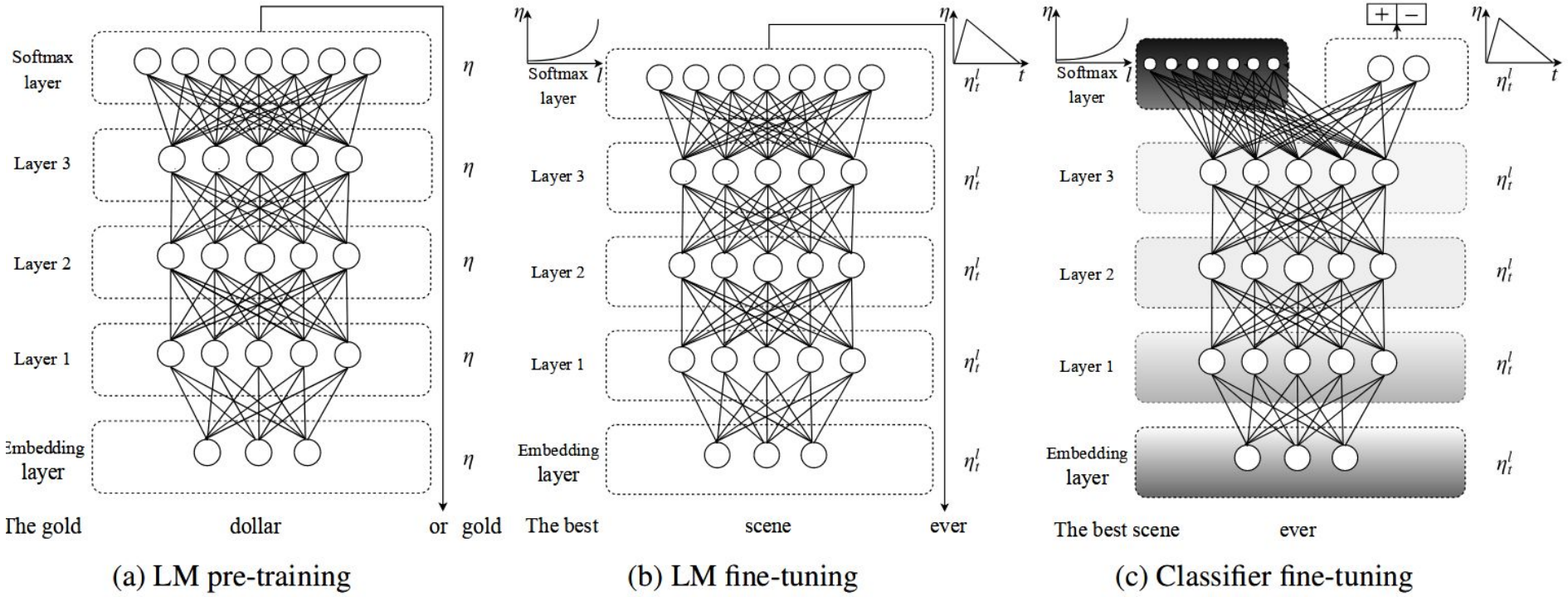
Male-Female



Verb tense



Country-Capital



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

2018

Контекстно-зависимые эмбеддинги

+ Знают, что Apple бывает разный
+ Универсальны для дальнейшего применения
+ Дают хорошую базу для работы остальных алгоритмов

— Медленно работают

https://allennlp.org/elmo

Проблема ограниченной выразительности

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML
	plugin, zip, javascript
apple	iphone, android, mac, microsoft
	samsung, phone, galaxy, touch
date	registration, join, location, from
	changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow
	head, ride, ball, dead
mass	energy, effect, impact, movement
	potential, military, weight, society
	exercise, lower

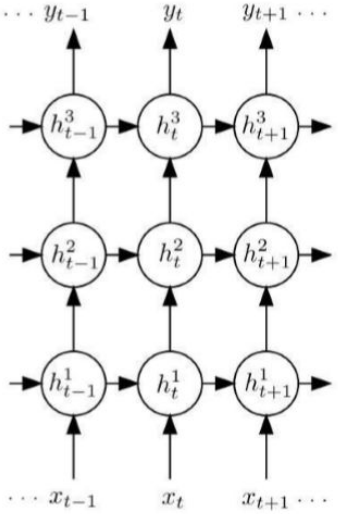
WORD	$p(z)$	NEAREST NEIGHBOURS
python	0.33	monty, spamalot, cantsin
	0.42	perl, php, java, c++
	0.25	molurus, pythons
apple	0.34	almond, cherry, plum
	0.66	macintosh, iifx, iigs
date	0.10	unknown, birth, birthdate
	0.28	dating, dates, dated
	0.31	to-date, stateside
	0.31	deadline, expiry, dates
bow	0.46	stern, amidships, bowsprit
	0.38	spear, bows, wow, sword
	0.16	teign, coxs, evenlode
mass	0.22	vespers, masses, liturgy
	0.42	energy, density, particle
	0.36	wholesale, widespread

ELMO

$$\sum_{k=1}^{N} \Big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big)$$

$$\begin{aligned} R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ &= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\}, \end{aligned}$$

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$



$$\lambda \|\mathbf{w}\|_2^2$$

2018

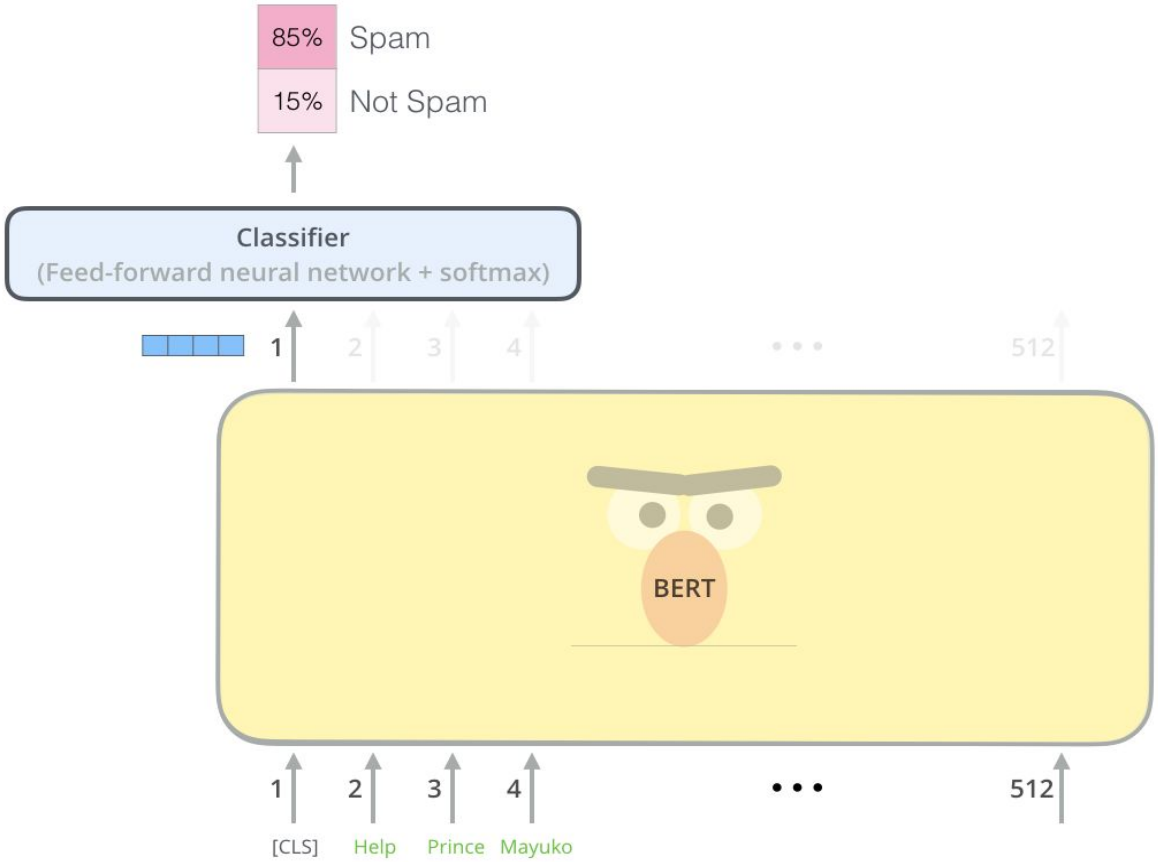
BERT

+ Почти не требует данных
+ Это трансформер
+ По-настоящему глубокая нейросеть
+ Бьёт все остальные архитектуры

— Медленный, да

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

Трансформеры



Методы тестирования NLU-моделей

https://arxiv.org/abs/1804.07461

Рассмотрим рост метрик подробнее

Model	Score
GLUE Human Baselines	87.1
BERT: 24-layers, 16-heads, 1024-hidc	80.5
Singletask Pretrain Transformer	72.8
BiLSTM+ELMo+Attn	70.0
BiLSTM+ELMo	67.7
BiLSTM+Attn	65.6
BiLSTM	64.2
CBOW	58.6

SINGLE SENTENCE TASKS

CoLA: The Corpus of Linguistic Acceptability (Warstadt et al., 2018)

SST-2: The Stanford Sentiment Treebank (Socher et al., 2013)

CoLA

1 John fed the baby up with rice.
0 John fed the baby rice up.

1 Spray all the paint onto the wall completely.
0 Spray the wall with all the paint.

1 The man who I gave John a picture of was bald.
0 The man who I gave John Ed's picture of was bald.
0 The man who I gave John this picture of was bald.

1 The noise gave Terry a headache.
0 The noise gave a headache to Terry.

Метрики, SINGLE SENTENCE TASKS

Model	Score	CoLA	SST-2
GLUE Human Baselines	87.1	66.4	97.8
BERT: 24-layers, 16-heads, 1024-hidc	80.5	60.5	94.9
Singletask Pretrain Transformer	72.8	45.4	91.3
BiLSTM+ELMo+Attn	70.0	33.6	90.4
BiLSTM+ELMo	67.7	32.1	89.3
BiLSTM+Attn	65.6	18.6	83.0
BiLSTM	64.2	11.6	82.8
CBOW	58.6	0.0	80.0

SIMILARITY AND PARAPHRASE TASKS

MRPC: The Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005)

QQP: The Quora Question Pairs

STS-B: The Semantic Textual Similarity Benchmark (Cer et al., 2017)

The Quora Question Pairs

question1	question2	is_duplicate
What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?	0
How can I increase the speed of my internet connection while using a VPN?	How can Internet speed be increased by hacking through DNS?	0
Why am I mentally very lonely? How can I solve it?	Find the remainder when [math]23^{24}[/math] is divided by 24,23?	0
Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?	Which fish would survive in salt water?	0
Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1

Метрики, SIMILARITY AND PARAPHRASE TASKS

Model	Score	MRPC	STS-B	QQP
GLUE Human Baselines	87.1	86.3/80.8	92.7/92.6	59.5/80.4
BERT: 24-layers, 16-heads, 1024-hidc	80.5	89.3/85.4	87.6/86.5	72.1/89.3
Singletask Pretrain Transformer	72.8	82.3/75.7	82.0/80.0	70.3/88.5
BiLSTM+ELMo+Attn	70.0	84.4/78.0	74.2/72.3	63.1/84.3
BiLSTM+ELMo	67.7	84.7/78.0	70.3/67.8	61.1/82.6
BiLSTM+Attn	65.6	83.9/76.2	72.8/70.5	60.1/82.4
BiLSTM	64.2	81.8/74.3	70.3/67.8	62.5/84.2
CBOW	58.6	81.5/73.4	61.2/58.7	51.4/79.1

INFERENCE TASKS

MNLI: The Multi-Genre Natural Language Inference Corpus (Williams et al., 2018)

QNLI: The Stanford Question Answering Dataset (Rajpurkar et al. 2016)

RTE: The Recognizing Textual Entailment

WNLI: The Winograd Schema Challenge (Levesque et al., 2011)

The Multi-Genre Natural Language Inference Corpus

The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.

The Multi-Genre Natural Language Inference Corpus

yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual	***contradiction***	August is a black out month for vacations in the company.
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	***entailment***	People formed a line at the end of Pennsylvania Avenue.

Метрики, INFERENCE TASKS

Model	Score	MNLI-mm	QNLI	RTE	WNLI
GLUE Human Baselines	87.1	92.8	91.2	93.6	95.9
BERT: 24-layers, 16-heads, 1024-hidc	80.5	85.9	92.7	70.1	65.1
Singletask Pretrain Transformer	72.8	81.4	87.4	56.0	53.4
BiLSTM+ELMo+Attn	70.0	74.5	79.8	58.9	65.1
BiLSTM+ELMo	67.7	67.9	75.5	57.4	65.1
BiLSTM+Attn	65.6	68.3	74.3	58.4	65.1
BiLSTM	64.2	66.1	74.6	57.4	65.1
CBOW	58.6	56.4	72.1	54.1	62.3

SWAG

Situations With Adversarial Generations

On stage, a woman takes a seat at the piano. She a) sits on a bench as her sister plays with the doll. b) smiles with someone as the music plays. c) is in the crowd, watching the dancers. **d) nervously sets her fingers on the keys.**
A girl is going across a set of monkey bars. She a) jumps up across the monkey bars. b) struggles onto the monkey bars to grab her head. **c) gets to the end and stands on a wooden plank.** d) jumps up and does a back flip.
The woman is now blow drying the dog. The dog **a) is placed in the kennel next to a woman's feet.** b) washes her face with the shampoo. c) walks into frame and walks towards the dog. d) tried to cut her face, so she is trying to do something very close to her face.

Table 1: Examples from SWAG; the correct answer is **bolded**. Adversarial Filtering ensures that stylistic models find all options equally appealing.

SWAG

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Image-caption retrieval

"A group of people on some horses riding through the beach."

Выбираем модели в продакшн

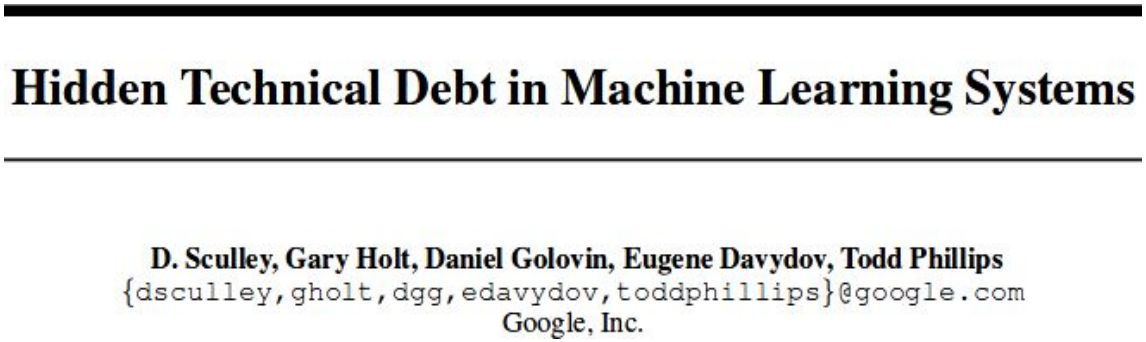
RNN VS CNN

			performance
TextC	SentiC (acc)	CNN	82.38
		GRU	**86.32**
		LSTM	84.51
	RC (F1)	CNN	68.02
		GRU	**68.56**
		LSTM	66.45
SemMatch	TE (acc)	CNN	77.13
		GRU	**78.78**
		LSTM	77.85
	AS (MAP & MRR)	CNN	(**63.69,65.01**)
		GRU	(62.58,63.59)
		LSTM	(62.00,63.26)
	QRM (acc)	CNN	**71.50**
		GRU	69.80
		LSTM	71.44

https://arxiv.org/pdf/1702.01923.pdf, https://arxiv.org/pdf/1707.05589.pdf

Владения инструментами недостаточно для построения эффективных решений

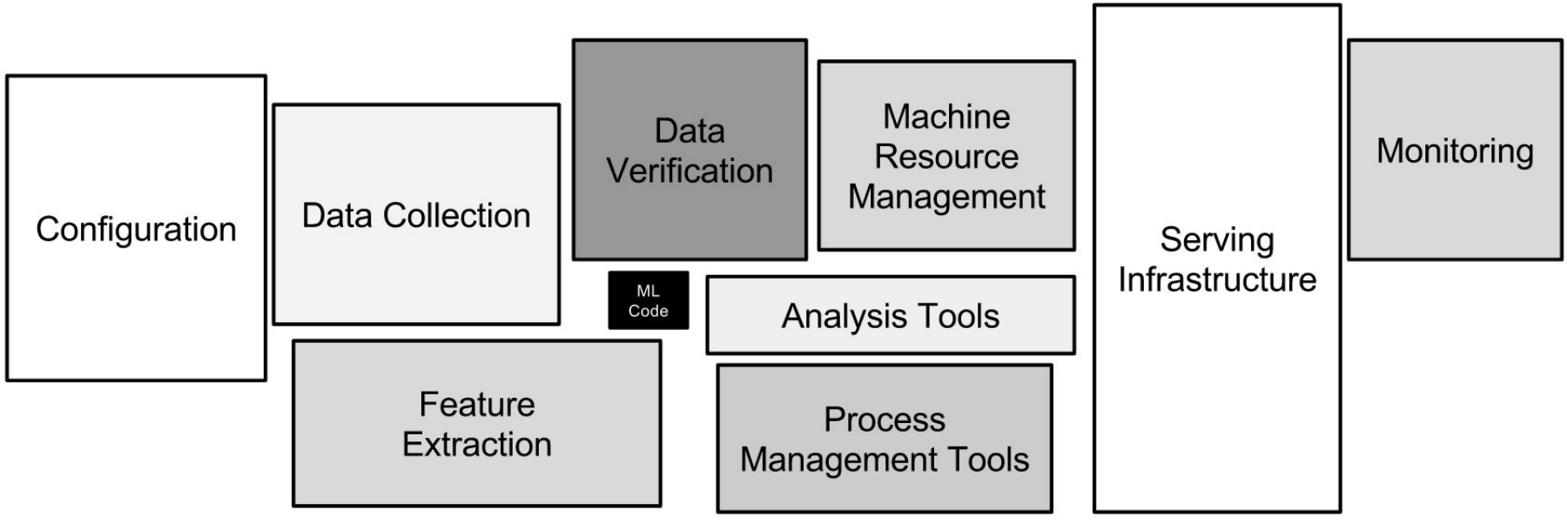
Важно не забывать о процессах



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf



https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Актуальные алгоритмы

Представление

- Tf-idf, nPMI, hashing trick, BPE

Факторизация (декомпозиция)

- PCA, LSI-LSA, pLSA, nNMF

Тематическое моделирование

- pLSA, LDA, HDP, ARTM

Поиски

- BM25, HNSW, LSH

Эмбеддинги

- word2vec, glove, doc2vec, fasttext, poincaré, ELMO

Нейросетевые подходы

- LSTM, GRU, TCN, Attention, siamese network, similarity learning, Transformer, Augmented RNN

Полезный NLP-софт

Предобработка текста (нормализация, токенизация)

- pymorphy2(ru), snowball stemmer(en), Stanford NLP(en)

Фреймворки

- sklearn, NLTK, gensim, spaCy

Узкоспециализированные фреймворки

- BigARTM, Vowpal Wabbit, Fasttext, faiss, annoy, NMSLib, lucene, sphinx, elastic

Нейросетевые фреймворки

- Pytorch, HuggingFace, AllenNLP, torchtext

Подходы и данные для тестирования моделей

- https://github.com/facebookresearch/SentEval
- https://arxiv.org/pdf/1707.05589.pdf
- https://arxiv.org/pdf/1806.06259.pdf
- https://aclweb.org/anthology/D18-1009
- https://arxiv.org/pdf/1702.02170.pdf
- https://arxiv.org/pdf/1903.09442.pdf

- https://leaderboard.allenai.org/swag/submissions/public
- https://gluebenchmark.com/leaderboard

https://allennlp.org/elmo

О прогрессе в НЛП

- https://nlpoverview.com/#3
- https://arxiv.org/pdf/1708.02709.pdf
- http://nlpprogress.com/english/language_modeling.html
- https://github.com/Separius/awesome-sentence-embedding

https://allennlp.org/elmo

Контакты

Штех Геннадий *
@ NAUMEN
gshtekh@naumen.ru

Gennady Shtekh
shtechgen@gmail.com
t.me/sht3ch
github.com/ShT3cH

*R&D Data Usage Department Executive