Contextual Models: Transformers

Luciano Barbosa

Centro de Informática UFPE

Universidade Federal de Pernambuco

Traditional Word Embedding

- Vocabulary built using the training set
- Words not present mapped to UNK

Subword Tokenization

- Mitigates the problem with out-of-vocabulary words
- Rare words are broken into substrings
- Frequent words are kept

“unfortunately” = “un” + “for” + “tun” + “ate” + “ly”
“anyplace” = “any” + “place”
“anyhow” = “any” + “how”
“anywhere” = “any” + “where”

Subword Tokenization

- Critical for languages with many variations in word structure
- Ex: Swahili (spoken in Kenya, Tanzania, and Uganda)

Conjugation of *-ambia* [less ▲]

Non-finite forms		
Form	Positive	Negative
Infinitive	kuambia	kutoambia

Simple finite forms		
Positive form	Singular	Plural
Imperative	ambia	ambieni
Habitual	huambia	

Complex finite forms

Polarity	Persons: 1st Sg.	1st Pl.	2nd Sg.	2nd Pl.	Persons / Classes: 3rd / M-wa Sg. / 1	Pl. / 2	Classes: M-mi 3	M-mi 4	Ma 5	Ma 6	Ki-vi 7	Ki-vi 8	N 9	N 10	U 11 / 14	Ku 15 / 17	Pa 16	Mu 18
Past [less ▲]																		
Positive	nilambia nalambia	tulilambia twalambia	ulilambia walambia	mlilambia mwalambia	alilambia	walilambia	ulilambia	ililambia	lililambia	yalilambia	kililambia	vililambia	ililambia	zililambia	ulilambia	kulilambia	palilambia	mulilambia
Negative	sikuambia	hatukuambia	hukuambia	hamkuambia	hakuambia	hawakuambia	haukuambia	haikuambia	halikuambia	hayakuambia	hakikuambia	havikuambia	haikuambia	hazikuambia	haukuambia	hakukuambia	hapakuambia	hamukuambia
Present [less ▲]																		
Positive	ninaambia naambia	tunaambia	unaambia	mnaambia	anaambia	wanaambia	unaambia	inaambia	linaambia	yanaambia	kinaambia	vinaambia	inaambia	zinaambia	unaambia	kunaambia	panaambia	munaambia
Negative	siambii	hatuambii	huambii	hamambii	haambii	hawaambii	hauambii	haiambii	haliambii	hayaambii	hakiambii	haviambii	haiambii	haziambii	hauambii	hakuambii	hapaambii	hamuambii
Future [less ▲]																		
Positive	nitaambia	tutaambia	utaambia	mtaambia	ataambia	wataambia	utaambia	itaambia	litaambia	yataambia	kitaambia	vitaambia	itaambia	zitaambia	utaambia	kutaambia	pataambia	mutaambia
Negative	sitaambia	hatutaambia	hutaambia	hamtaambia	hataambia	hawataambia	hautaambia	haitaambia	halitaambia	hayataambia	hakitaambia	havitaambia	haitaambia	hazitaambia	hautaambia	hakutaambia	hapataambia	hamutaambia
Subjunctive [less ▲]																		
Positive	niambie	tuambie	uambie	mambie	aambie	waambie	uambie	iambie	liambie	yaambie	kiambie	viambie	iambie	ziambie	uambie	kuambie	paambie	muambie
Negative	nisiambie	tusiambie	usiambie	msiambie	asiambie	wasiambie	usiambie	isiambie	lisiambie	yasiambie	kisiambie	visiambie	isiambie	zisiambie	usiambie	kusiambie	pasiambie	musiambie
Present Conditional [less ▲]																		
Positive	ningeambia	tungeambia	ungeambia	mngeambia	angeambia	wangeambia	ungeambia	ingeambia	lingeambia	yangeambia	kingeambia	vingeambia	ingeambia	zingeambia	ungeambia	kungeambia	pangeambia	mungeambia
Negative	nisingeambia singeambia	tusingeambia hatungeambia	usingeambia hungeambia	msingeambia hamngeambia	asingeambia hangeambia	wasingeambia hawangeambia	usingeambia haungeambia	isingeambia haingeambia	lisingeambia halingeambia	yasingeambia hayangeambia	kisingeambia hakingeambia	visingeambia havingeambia	isingeambia haingeambia	zisingeambia hazingeambia	usingeambia haungeambia	kusingeambia hakungeambia	pasingeambia hapangeambia	musingeambia hamungeambia
Past Conditional [less ▲]																		
Positive	ningaliambia	tungaliambia	ungaliambia	mngaliambia	angaliambia	wangaliambia	ungaliambia	ingaliambia	lingaliambia	yangaliambia	kingaliambia	vingaliambia	ingaliambia	zingaliambia	ungaliambia	kungaliambia	pangaliambia	mungaliambia
Negative	nisingaliambia singaliambia	tusingaliambia hatungaliambia	usingaliambia hungaliambia	msingaliambia hamngaliambia	asingaliambia hangaliambia	wasingaliambia hawangaliambia	usingaliambia haungaliambia	isingaliambia haingaliambia	lisingaliambia halingaliambia	yasingaliambia hayangaliambia	kisingaliambia hakingaliambia	visingaliambia havingaliambia	isingaliambia haingaliambia	zisingaliambia hazingaliambia	usingaliambia haungaliambia	kusingaliambia hakungaliambia	pasingaliambia hapangaliambia	musingaliambia hamungaliambia
Conditional Contrary to Fact [less ▲]																		
Positive	ningeliambia	tungeliambia	ungeliambia	mngeliambia	angeliambia	wangeliambia	ungeliambia	ingeliambia	lingeliambia	yangeliambia	kingeliambia	vingeliambia	ingeliambia	zingeliambia	ungeliambia	kungeliambia	pangeliambia	mungeliambia
Gnomic [less ▲]																		
Positive	naambia	twaambia	waambia	mwaambia	aambia	waambia	waambia	yaambia	laambia	yaambia	chaambia	vyaambia	yaambia	zaambia	waambia	kwaambia	paambia	mwaambia
Perfect [less ▲]																		

Subword Tokenization: Byte-pair Encoding (BPE)

- Goal: to represent the corpus with the least amount of tokens

- Algorithm:

1. Start the vocabulary with all characters and the "end of word" symbol
2. Merge the most frequent tokens
3. Decrements the frequency of the two tokens
4. Repeat until the desired vocabulary (e.g., pre-defined vocabulary size)

BPE: Words in the Corpus

WORD	FREQUENCY	WORD	FREQUENCY
d e e p </w>	3	b u i l d </w>	1
l e a r n i n g </w>	3	t r a i n </w>	1
t h e </w>	2	a n d </w>	1
m o d e l s </w>	2	d e p l o y </w>	1
F l o y d h u b </w>	1	B u i l d </w>	1
i s </w>	1	m o d e l s </w>	1
f a s t e s t </w>	1	i n </w>	1
w a y </w>	1	c l o u d </w>	1
t o </w>	1	T r a i n </w>	1

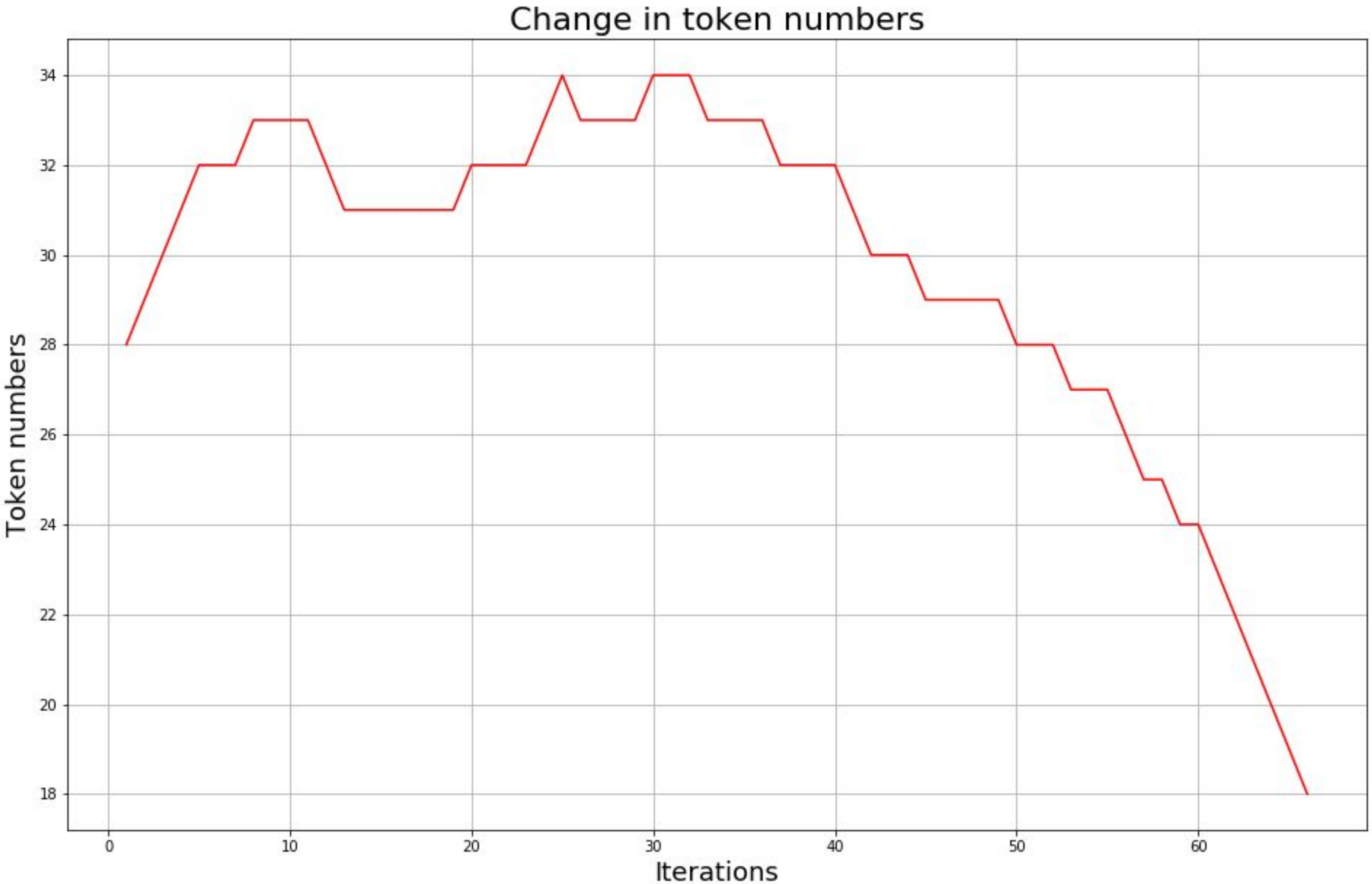
BPE: Characters in the Corpus

NUMBER	TOKEN	FREQUENCY	NUMBER	TOKEN	FREQUENCY
1	</w>	24	**15**	g	3
2	e	16	**16**	m	3
3	d	12	**17**	.	3
4	l	11	**18**	b	2
5	n	10	**19**	h	2
6	i	9	**20**	F	1
7	a	8	**21**	H	1
8	o	7	**22**	f	1
9	s	6	**23**	w	1
10	t	6	**24**	,	1
11	r	5	**25**	B	1
12	u	4	**26**	c	1
13	p	4	**27**	T	1
14	y	3			

BPE: Merge

NUMBER	TOKEN	FREQUENCY	NUMBER	TOKEN	FREQUENCY
1	</w>	24	16	g	3
2	e	16 - 7 = 9	17	m	3
3	d	12 - 7 = 5	18	.	3
4	l	11	19	b	2
5	n	10	20	h	2
6	i	9	21	F	1
7	a	8	22	H	1
8	o	7	23	f	1
9	de	7	24	w	1
10	s	6	25	,	1
11	t	6	26	B	1
12	r	5	27	c	1
13	u	4	28	T	1
14	p	4			
15	y	3			

BPE



Change in token numbers
Token numbers
Iterations
34
32
30
28
26
24
22
20
18
0
10
20
30
40
50
60

Wordpiece

- Similar ao BPE
- Algorithm:

1. Start the vocabulary with all characters and the "end of word" symbol
2. Merge the tokens with the highest score
3. Decrements the frequency of the two tokens
4. Repeat until the desired vocabulary (e.g., pre-defined vocabulary size)

$$\text{score} = (\text{freq\_of\_pair})/(\text{freq\_of\_first\_element} \times \text{freq\_of\_second\_element})$$

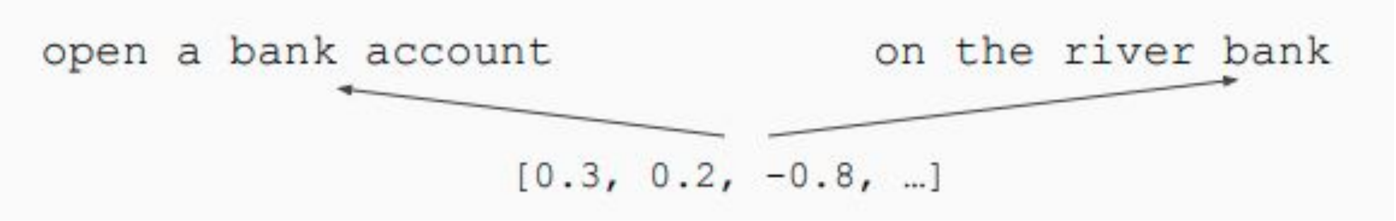
high probability of they occur together than separated

Wordpiece

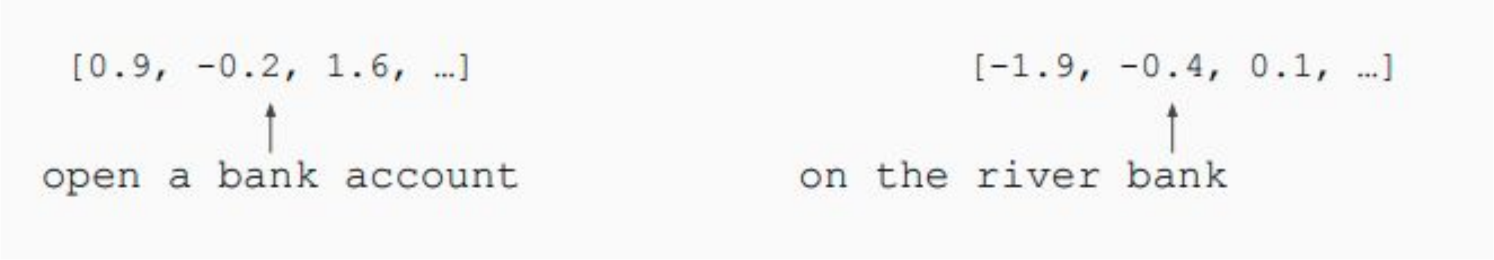
Word	Token(s)
surf	['surf']
surfing	['surf', '##ing']
surfboarding	['surf', '##board', '##ing']
surfboard	['surf', '##board']
snowboard	['snow', '##board']
snowboarding	['snow', '##board', '##ing']
snow	['snow']
snowing	['snow', '##ing']

Contextual Representations

- Limitation of word embeddings: same representation for different meanings

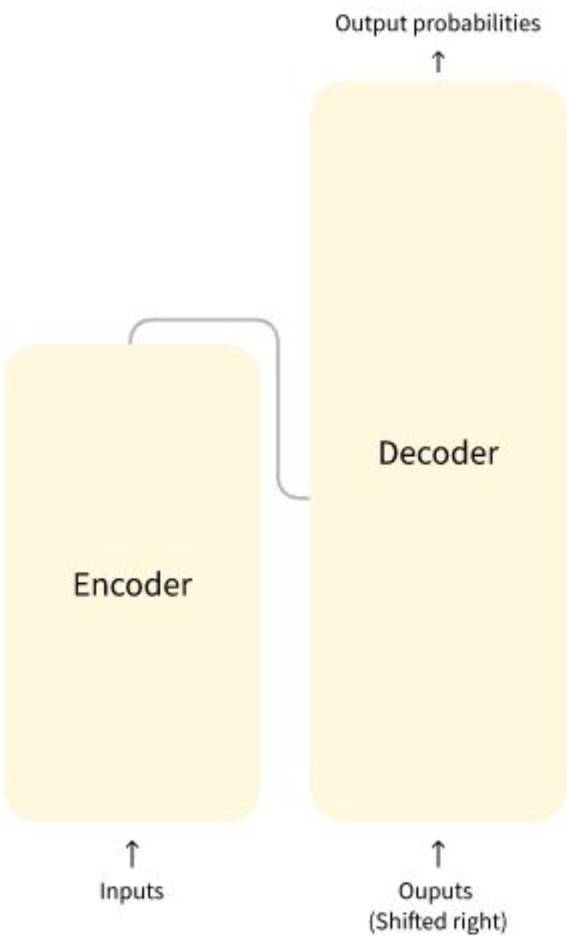


- Solution: learn contextual representations



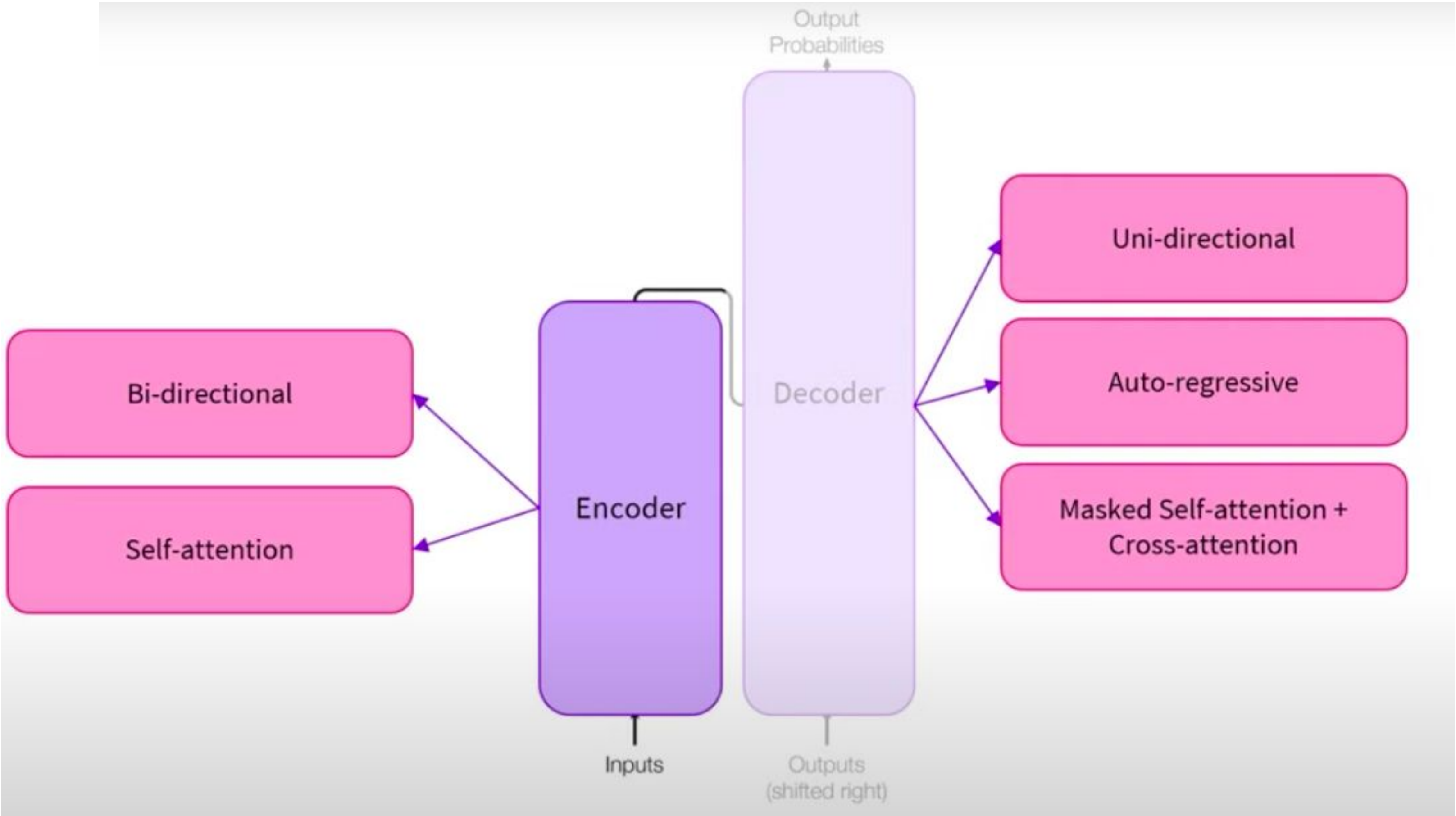
Transformers: Architecture

- Encoder: builds a representation of the input (embeddings)
- Decoder: outputs probabilities based on the encoder's output and other inputs



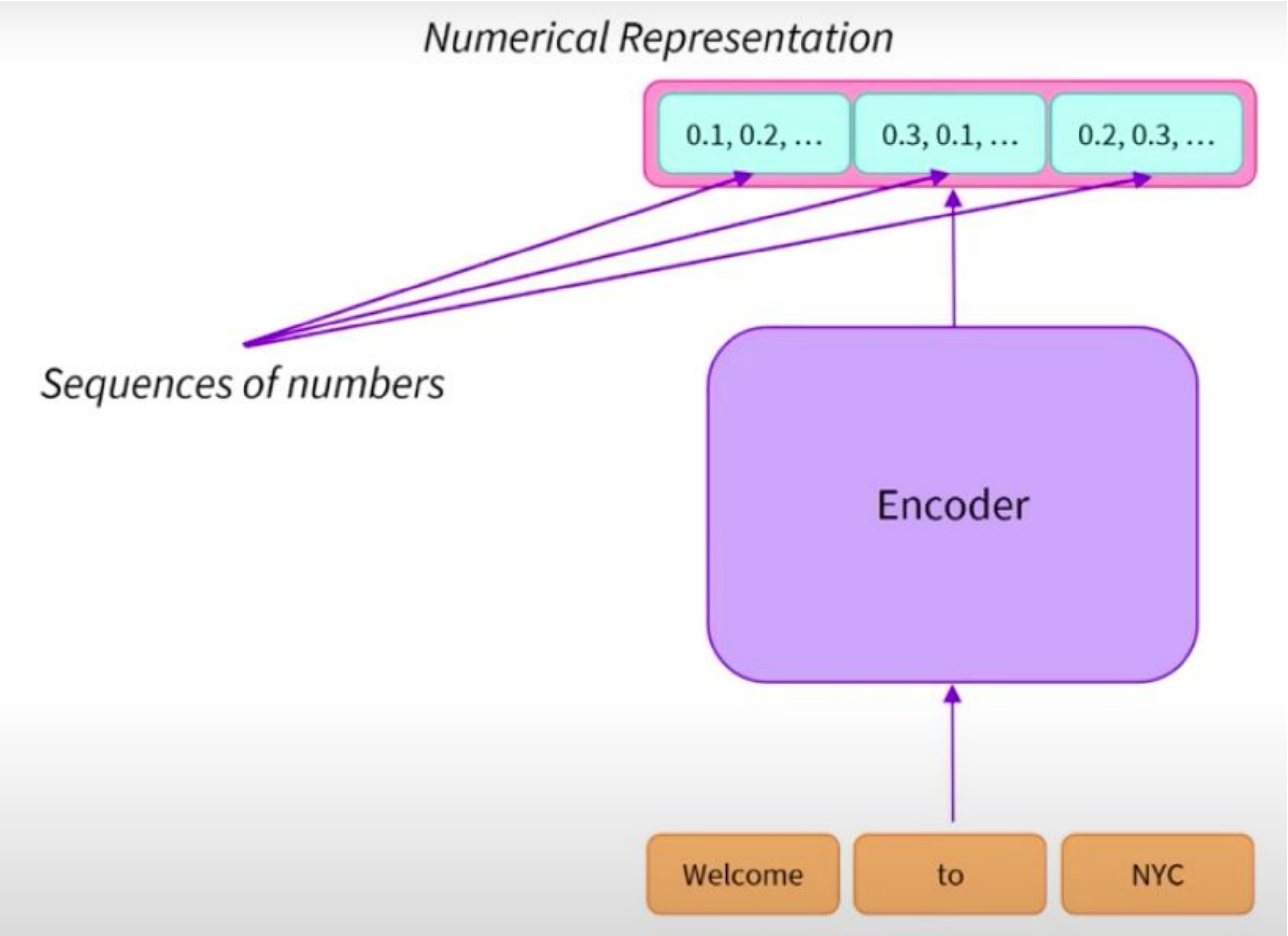
source: https://huggingface.co/course/chapter1/4

Transformers: Encoder



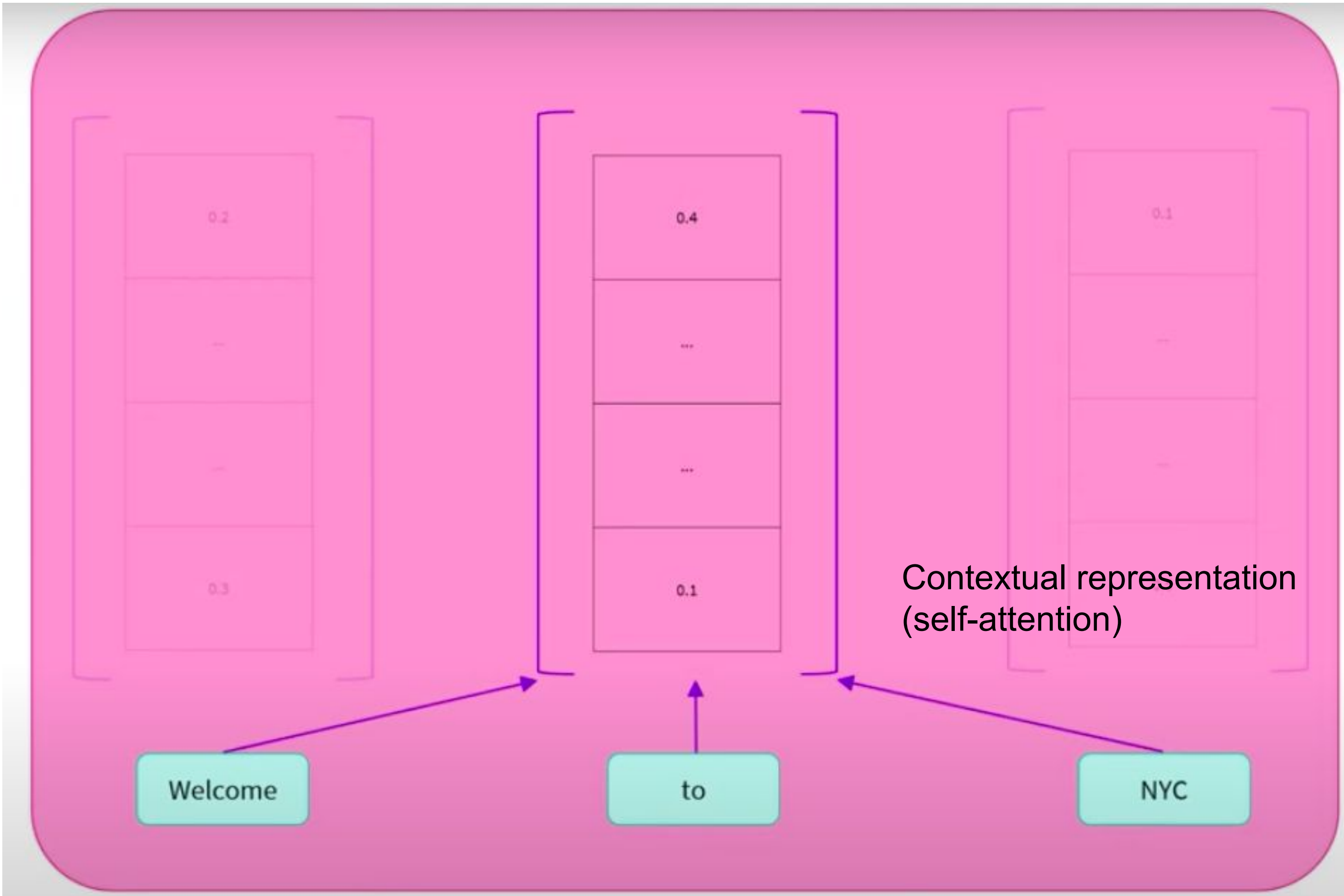
source: https://huggingface.co/course/chapter1/5

Transformers: Encoder



source: https://huggingface.co/course/chapter1/5

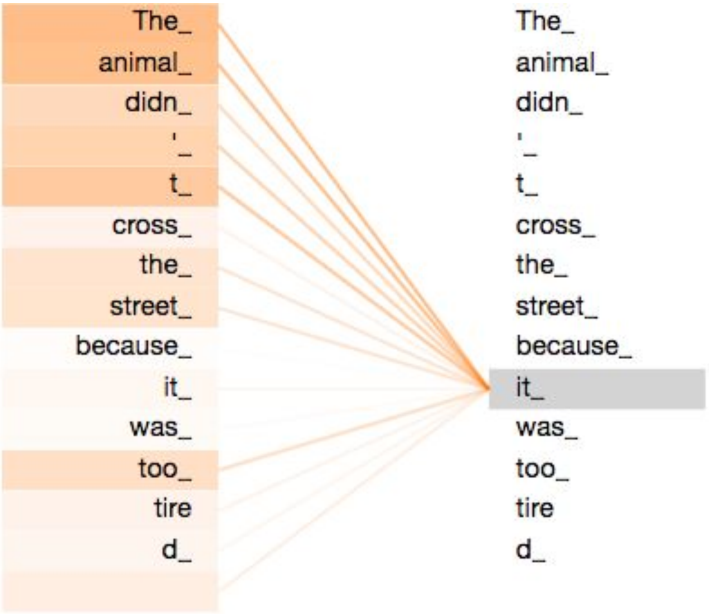
Transformers: Encoder



source: https://huggingface.co/course/chapter1/5

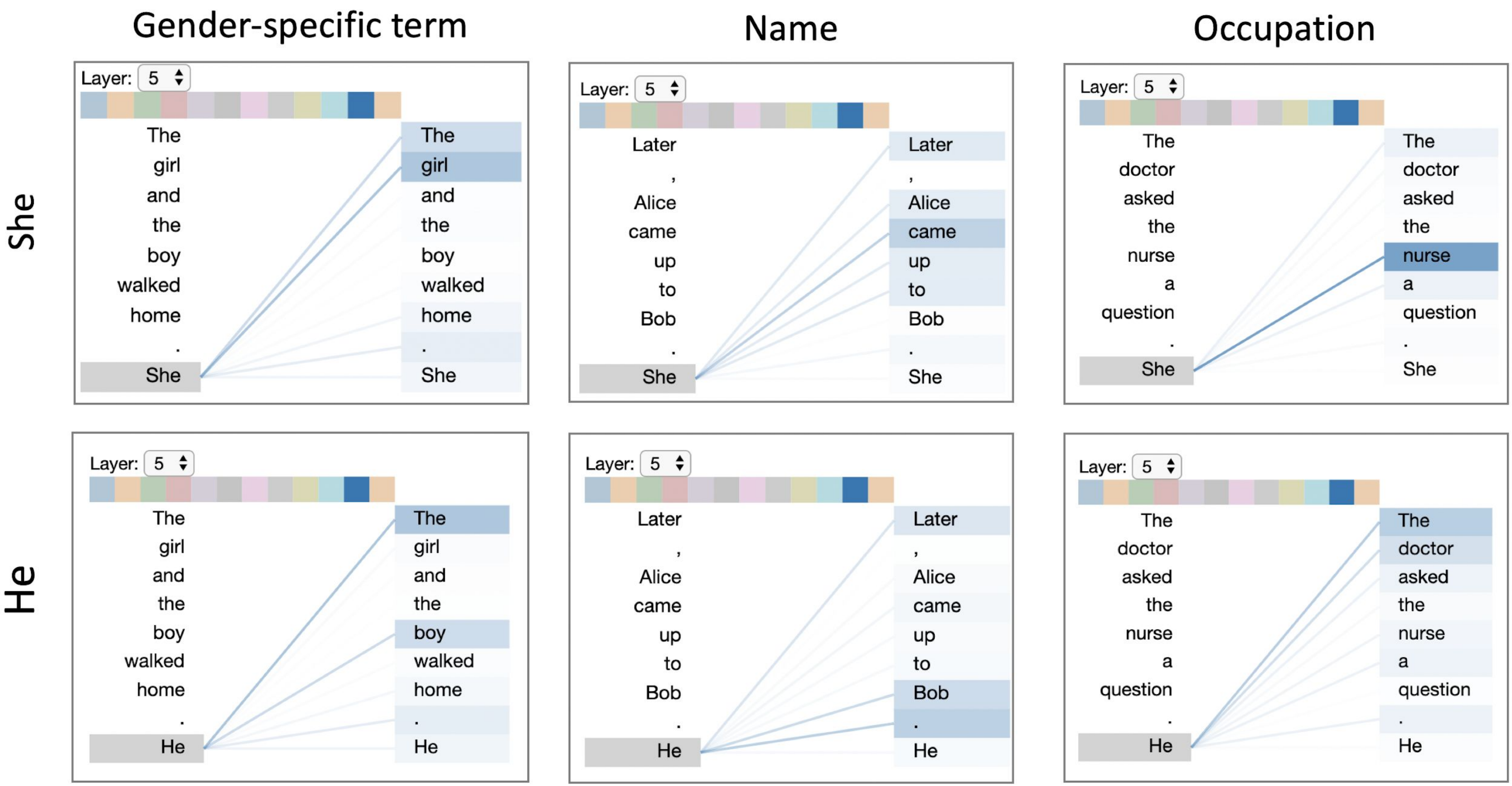
Self-Attention

- All words in the sentence have influence in the word representation

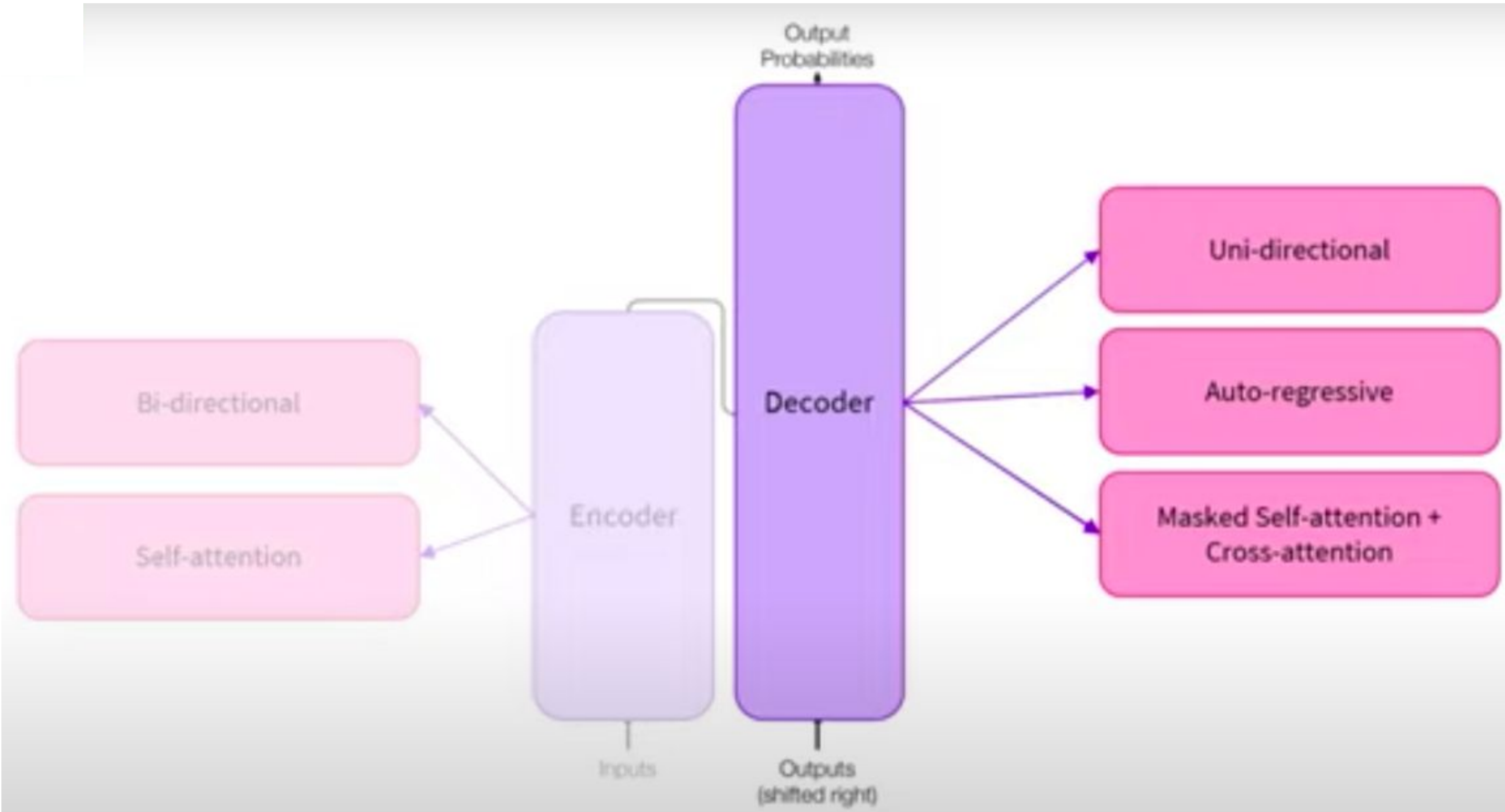


https://jalammar.github.io/illustrated-transformer/

Examples of Self-Attention

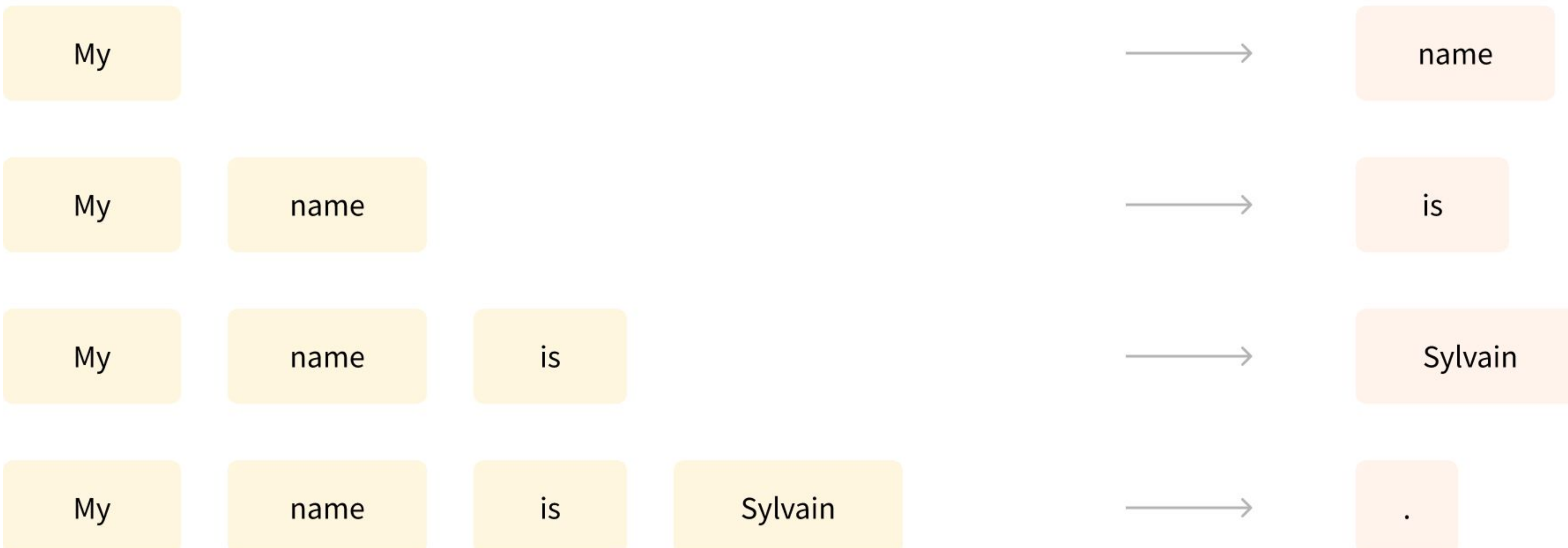


Transformers: Decoder



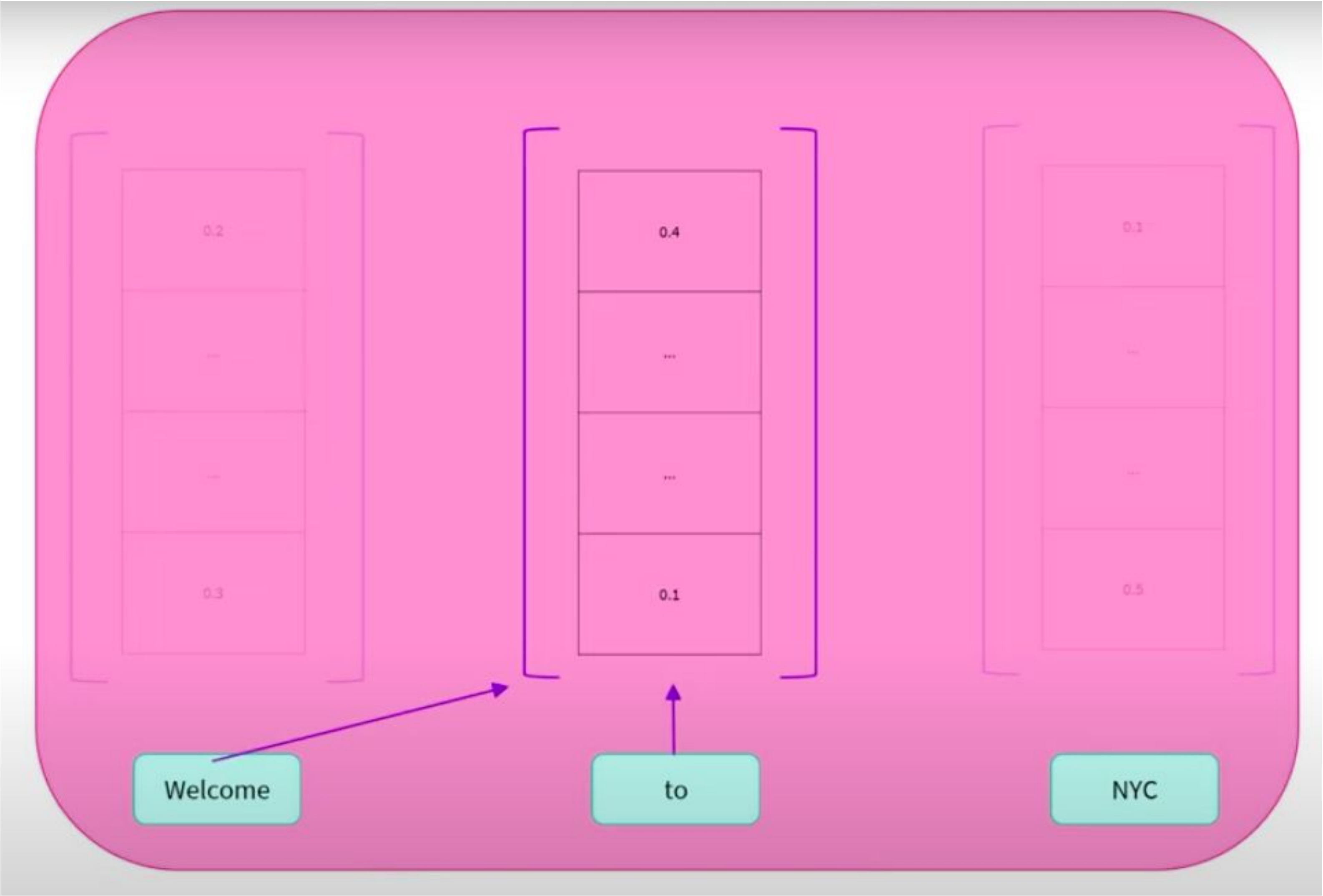
source: https://huggingface.co/course/chapter1/6

Transformers: Decoder



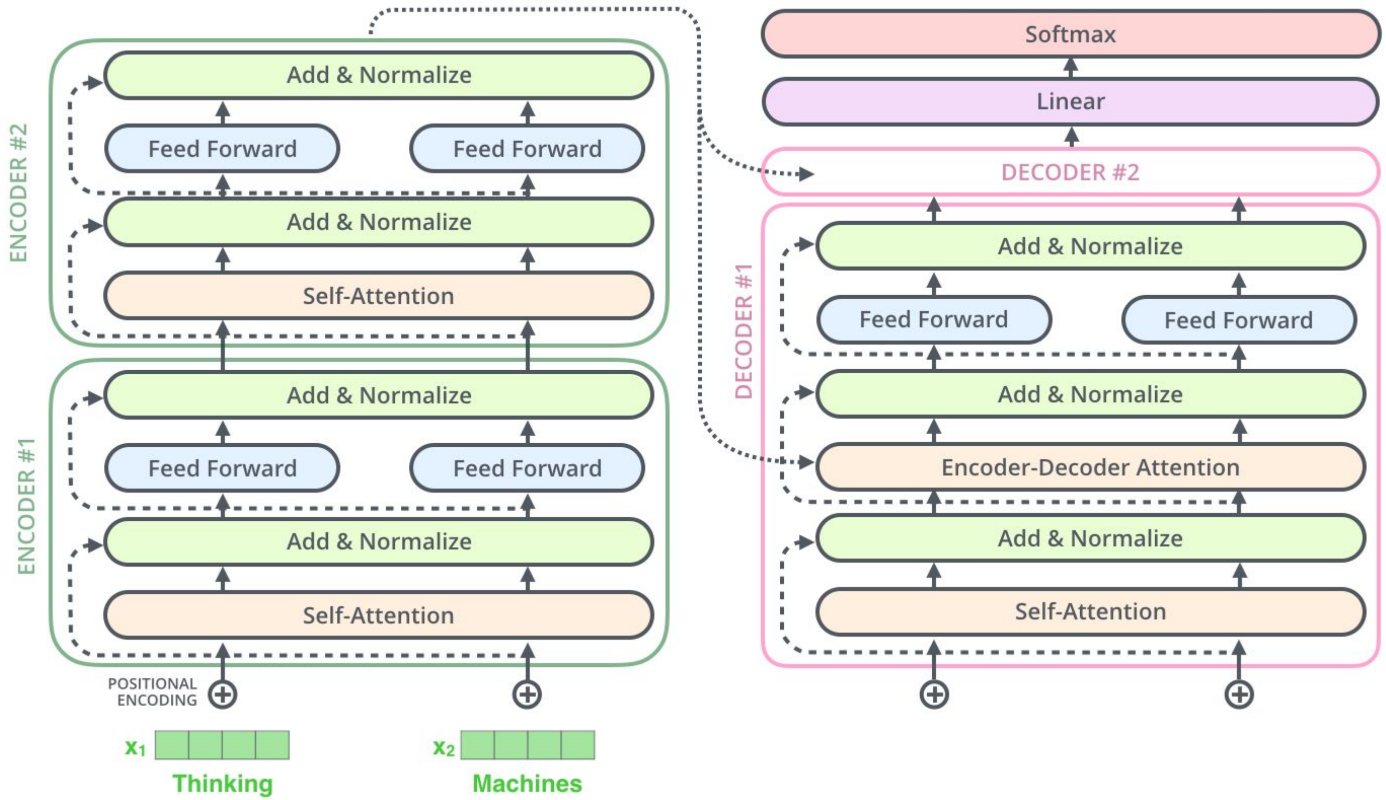
source: https://huggingface.co/course/chapter1/4

Transformers: Decoder



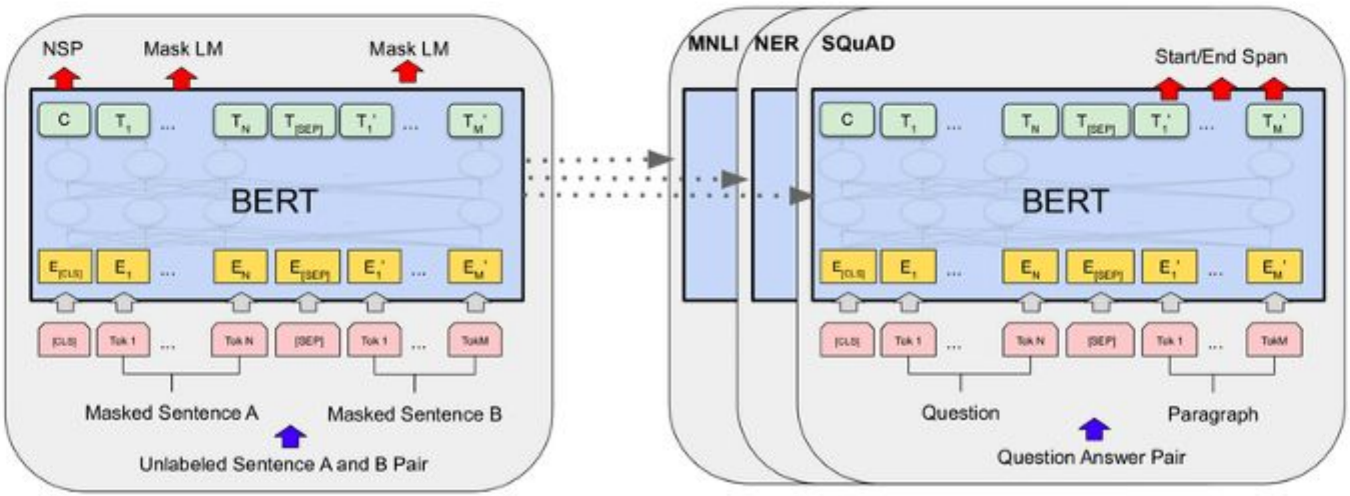
source: https://huggingface.co/course/chapter1/6

BERT: Bidirectional Encoder Representations from Transformers



Img-Source: https://jalammar.github.io/illustrated-transformer/

BERT: Two Steps

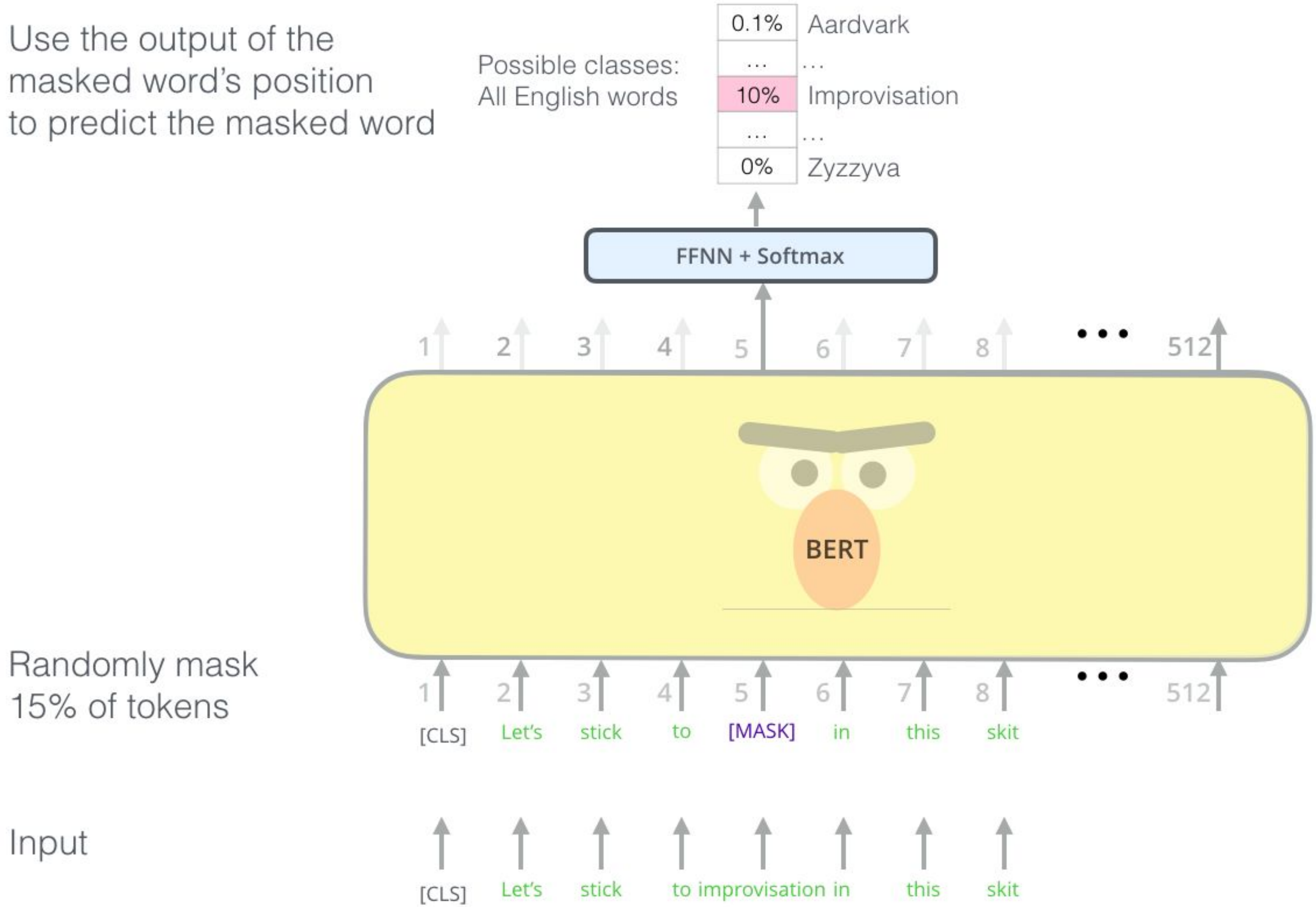


Pre-training

Fine-Tuning

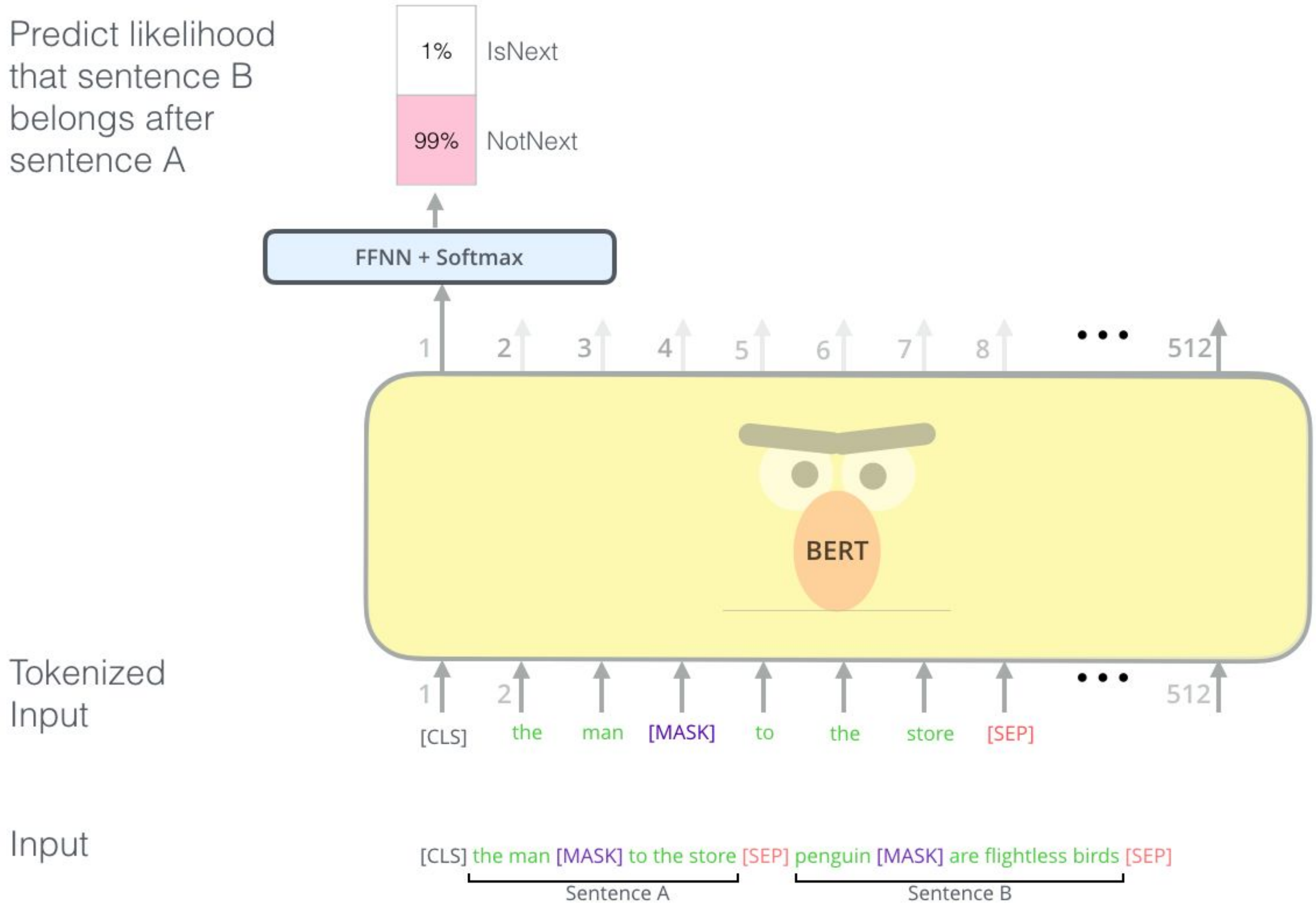
Pre-training: Masked Language Model

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



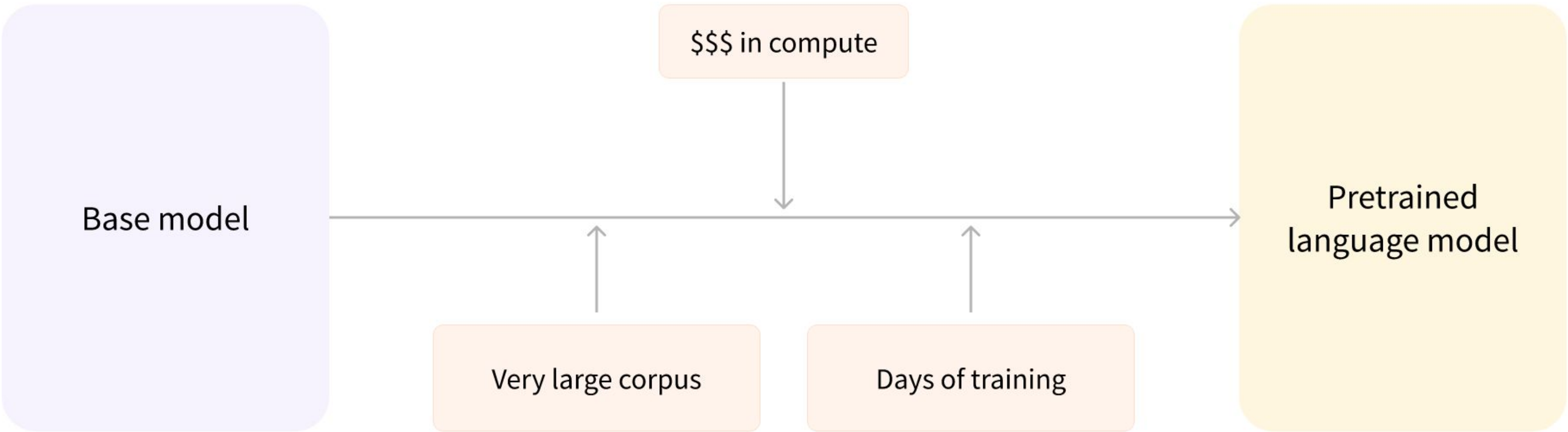
Img-Source: https://jalammar.github.io/illustrated-transformer/

Pre-training: Next-Sentence Prediction



Img-Source: https://jalammar.github.io/illustrated-transformer/

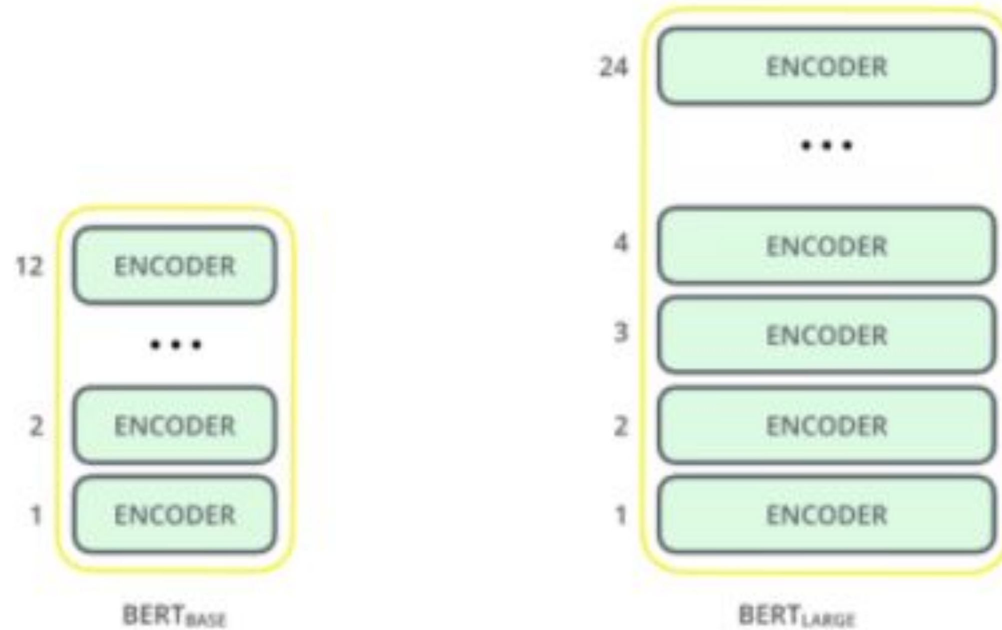
Pre-training



source: https://huggingface.co/course/chapter1/4

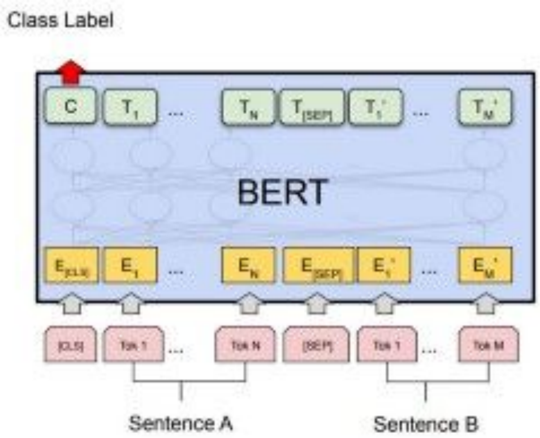
BERT: Details

- Data: Wikipedia (2.5B words) + BookCorpus (800M words)
- BERT-Base: 110M
- BERT-Large: 340M

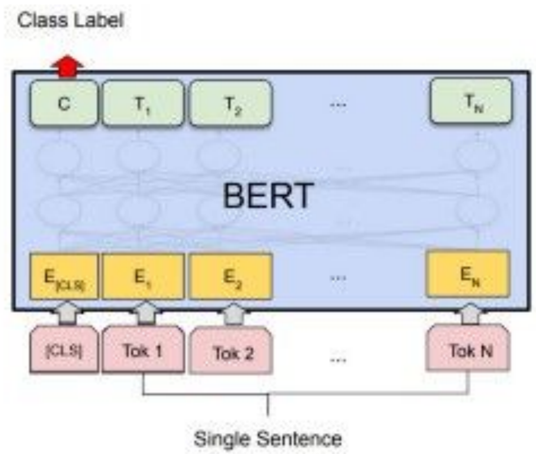


Img-Source: https://jalammar.github.io/illustrated-transformer/

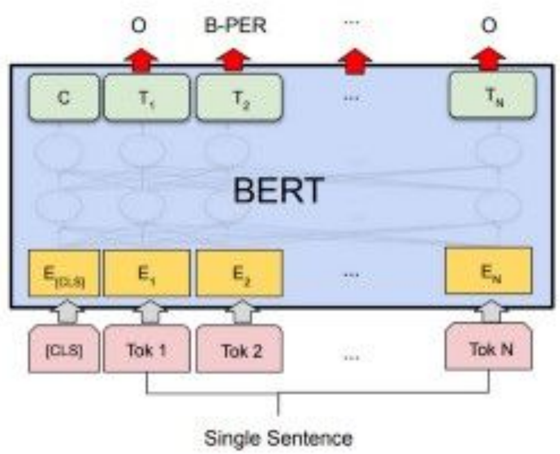
Fine-tuning (Transfer Learning)



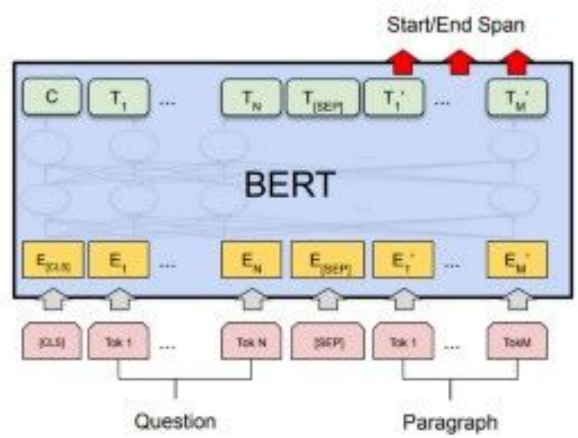
Sentence Pair Classification



Single Sentence Classification

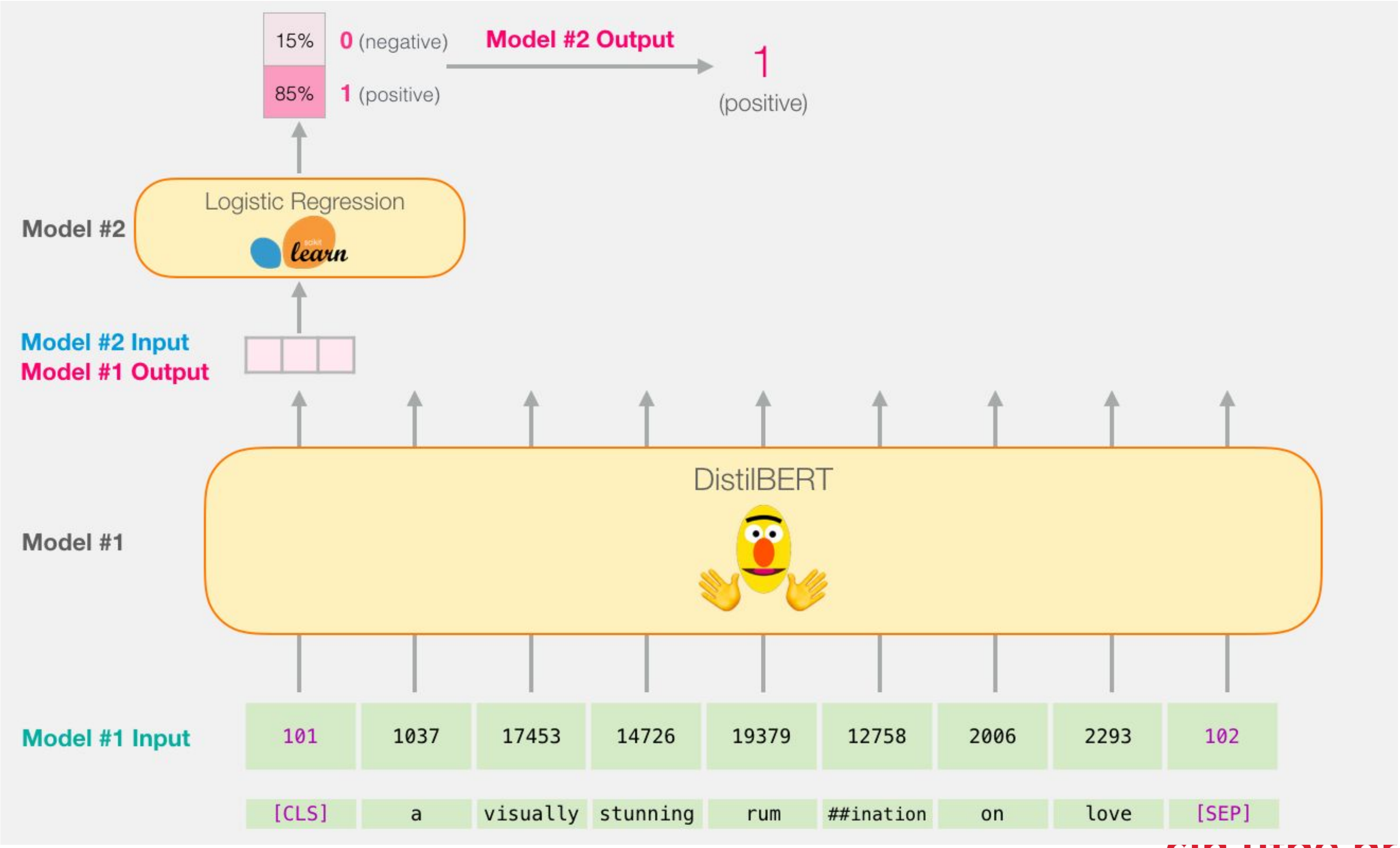


Sequence Tagging



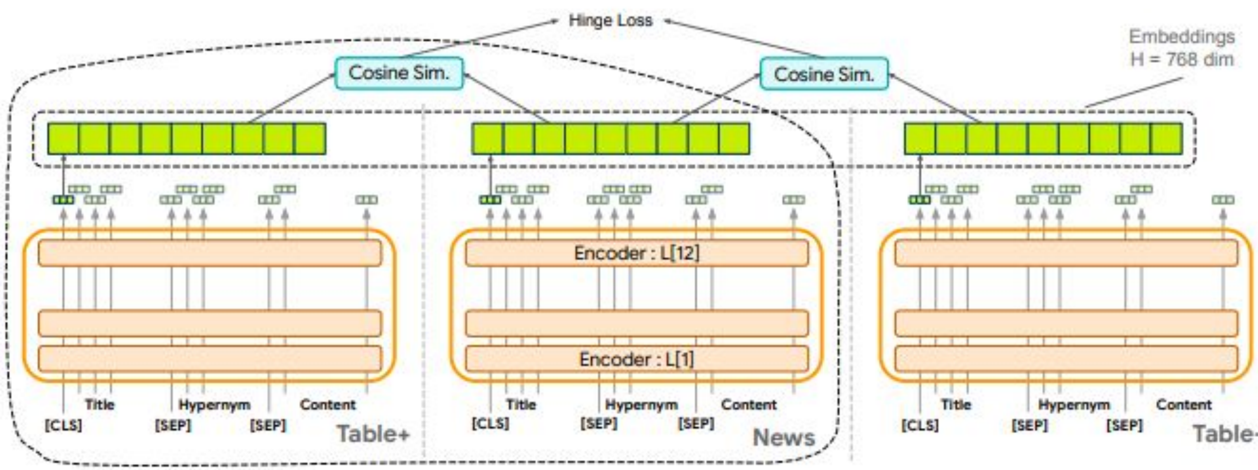
Question Answering

Sentence Classification: Sentiment Analysis



Img-Source: https://jalammar.github.io/illustrated-transformer/

Pair Classification: Article-Table Matching



Model	Table	News	acc.@k=1	5	10	100
BM25	⟨Title,Content⟩	⟨Title,Content⟩	**.426**	.574	.622	.831
	⟨Title,Content,Table Content⟩	⟨Title,Content⟩	.372	**.588**	**.703**	**.838**
TF-IDF	⟨Title,Content,Table Content⟩	⟨Title,Content⟩	**.453**	**.642**	**.730**	**.892**
USE	⟨Title,Content⟩	⟨Title⟩	**.250**	.439	.507	.669
	⟨Title,Content⟩	⟨Content⟩	.237	.419	.520	**.851**
	⟨Title,Content⟩	⟨Title,Content⟩	.243	**.466**	**.561**	.838
Doc2Vec	⟨Title⟩	⟨Title⟩	**.297**	**.507**	**.581**	**.824**
	⟨Title,Content⟩	⟨Title,Content⟩	.223	.378	.487	.737
BERTpublic	⟨Title,Content⟩	⟨Title⟩	**.155**	**.291**	**.372**	**.574**
NewsBERT	⟨Title,Content⟩	⟨Title,Content⟩	**.458**	**.725**	**.779**	**.824**

Model	Table	News	acc.@k=1	5	10	100
NewsBERT	⟨Title⟩	⟨Title⟩	.422	.602	.656	.719
	⟨Title⟩	⟨Title,Content⟩	.438	.677	.746	.8
	⟨Title,Content⟩	⟨Title⟩	.176	.244	.313	.588
	⟨Title,Content⟩	⟨Title,Content⟩	.458	.725	.779	.824
NT-BERT	⟨Title⟩	⟨Title⟩	.545	.773	.833	.939
	⟨Title⟩	⟨Title,Content⟩	.53	.818	.871	.955
	⟨Title,Content⟩	⟨Title⟩	.545	.795	.841	.932
	⟨Title,Content⟩	⟨Title,Content⟩	.553	.856	.879	.947
NTH-BERT	⟨Title,Hypernyms,Content⟩	⟨Title⟩	.575	.811	.858	.961
	⟨Title,Hypernyms,Content⟩	⟨Title,Hypernyms⟩	.543	.803	.858	.953
	⟨Title,Hypernyms,Content⟩	⟨Title,Hypernyms,Content⟩	**.669**	**.898**	**.953**	**.976**

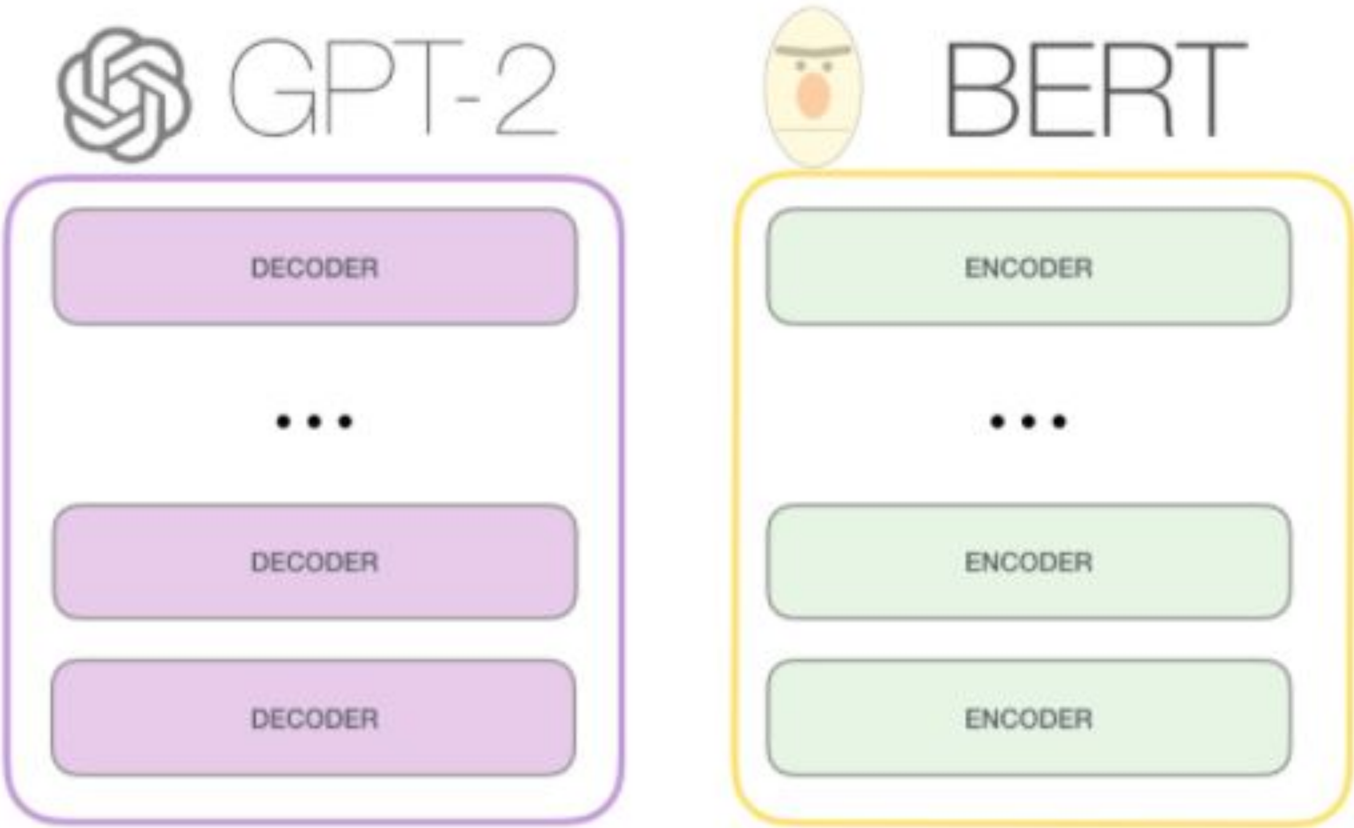
Lees et al. **Collocating News Articles with Structured Web Tables.** International Workshop on News Recommendation and Intelligence, 2021


TernaryBERT
CuBERT
X-BERT
Ta-BERT
Japanese-BERT
KoBERT
R-BERT
SemBERT
Graph-BERT
m-BERT
AraBERT
DeVLBERT
AnchiBERT
AMBERT
PhoBERT
NER-BERT
GANBERT
BERT-QA
RecoBERT
AMBERT
FlauBERT
SpanBERT
TinyBERT
ALBERT
SciBERT
FastBERT
CamemBERT
BERT
RoBERTa
DistilBERT

Generative Pretrained Transformer (GPT)

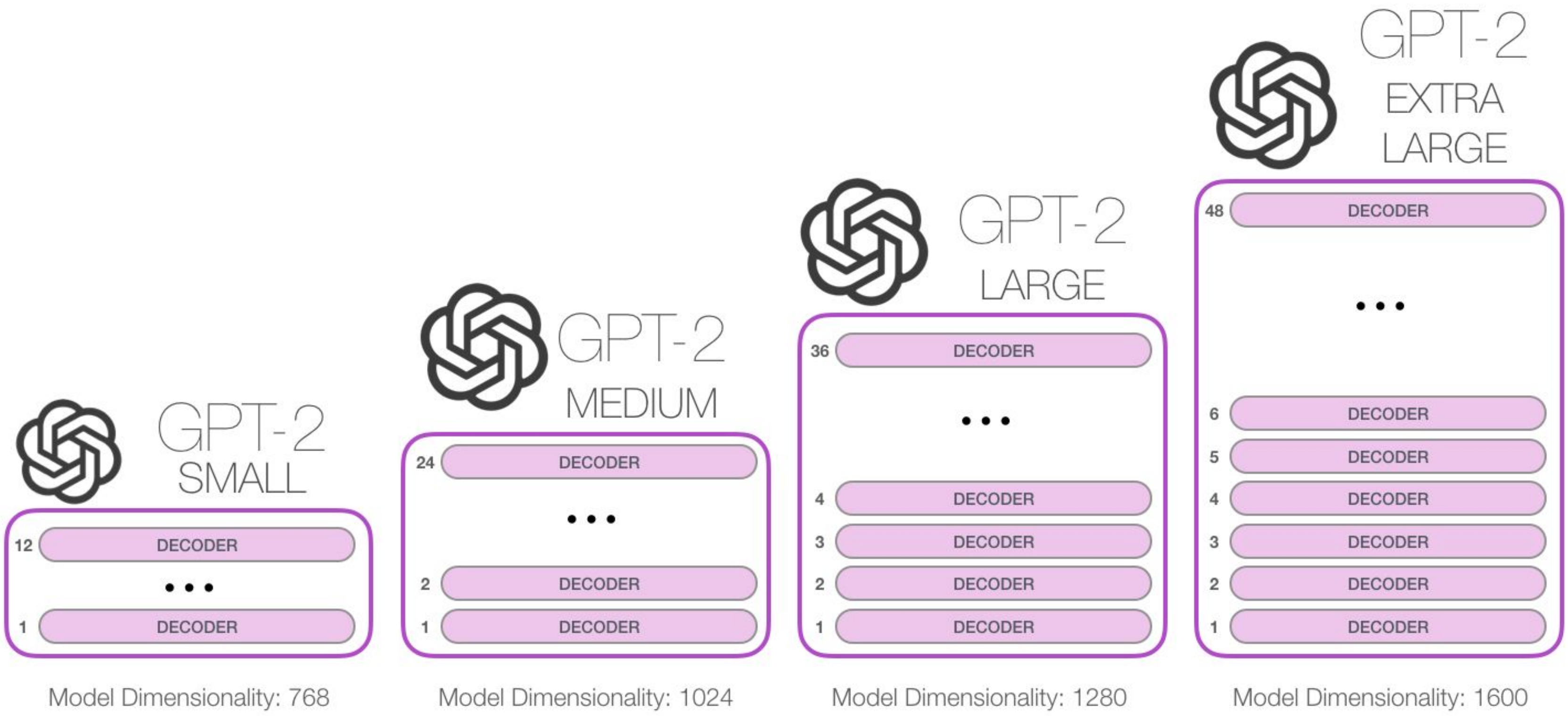
- Transformer decoder com 12 camadas
- 768-dimensional hidden states, 3072-dimensional feed-forward hidden layers
- Byte-pair encoding com 40.000 merges
- Treinado no BooksCorpus: mais de 7.000 livros (sentenças longas)
- Demo: https://demo.allennlp.org/next-token-lm

GPT-2 vs BERT



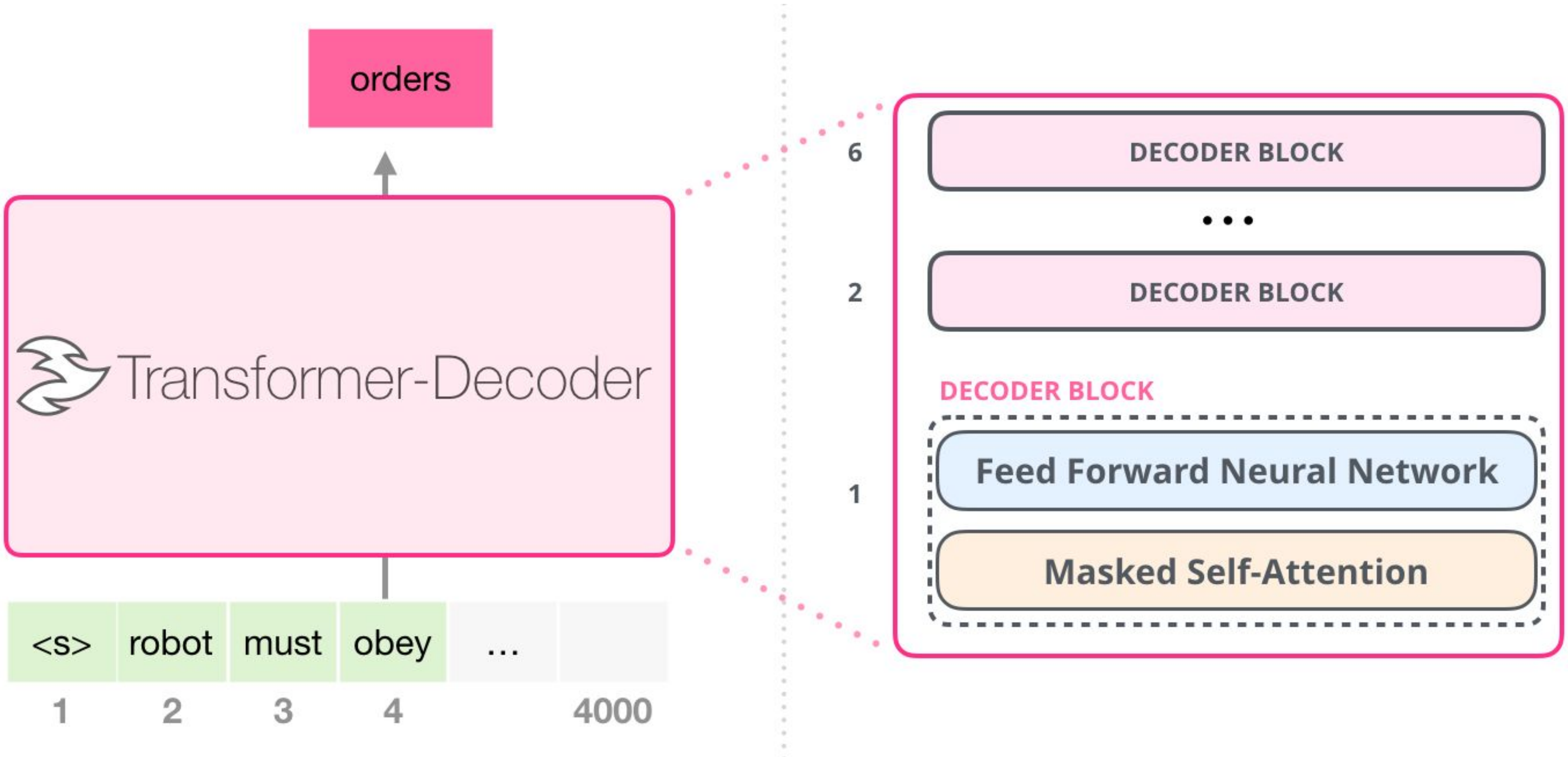
https://jalammar.github.io/illustrated-gpt2/

GPT-2 Models



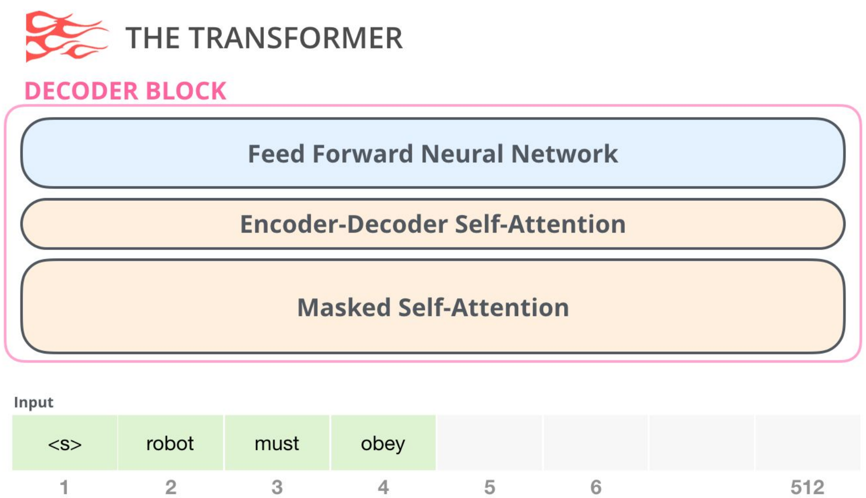
https://jalammar.github.io/illustrated-gpt2/

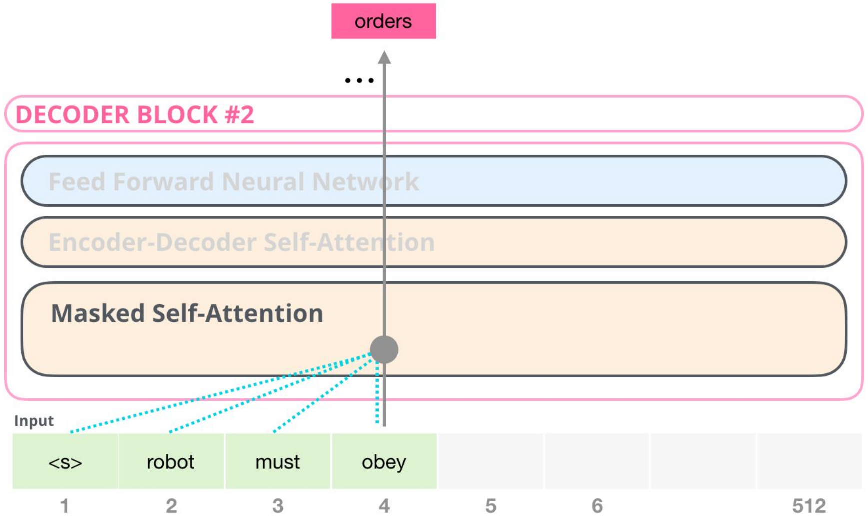
GPT-2



https://jalammar.github.io/illustrated-gpt2/

GPT-2: Masked Self-Attention

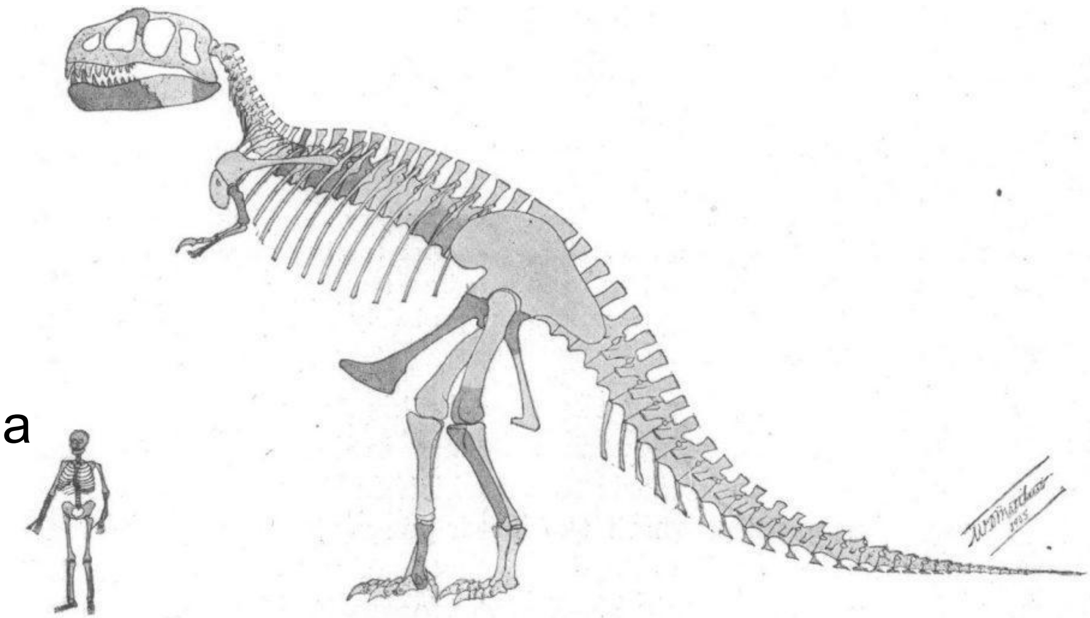




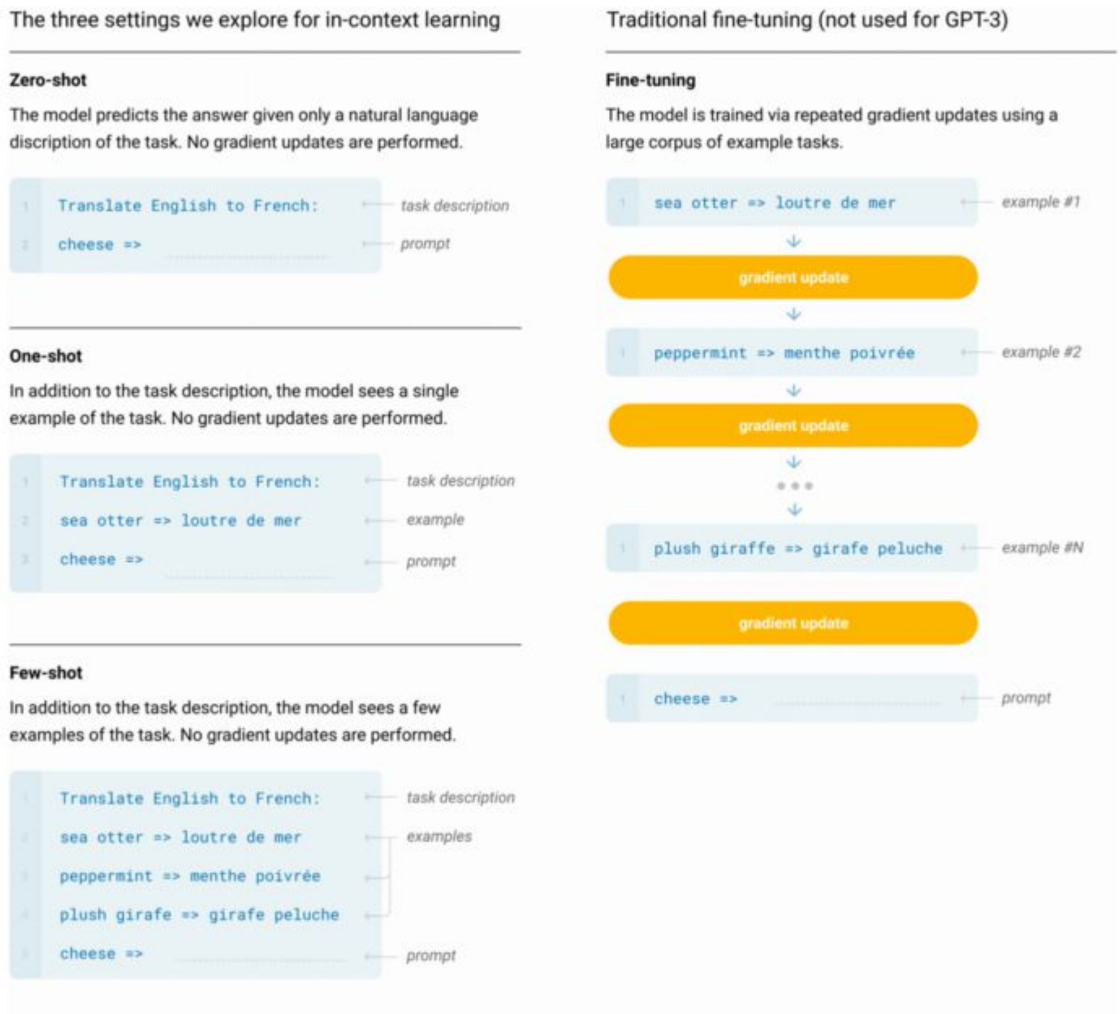
https://jalammar.github.io/illustrated-gpt2/

GPT-3

- 175 billion of parameters
- (100x > GPT-2)
- Corpus:
- 45TB
- Common Crawl, WebText2, Books1, Books2 e Wikipedia
- 96 layers



GPT-3 In-context Learning



Source: https://arxiv.org/pdf/2005.14165.pdf

ChatGPT (GPT-3.5)

- Supervised Fine Tuning (SFT) Model
- Fine-tuning the GPT-3 model on a large labeled dataset
- Inputs and prompts from users of the OpenAI API platform
- Manual labelers needed to create examples of categories not covered by OpenAI API users