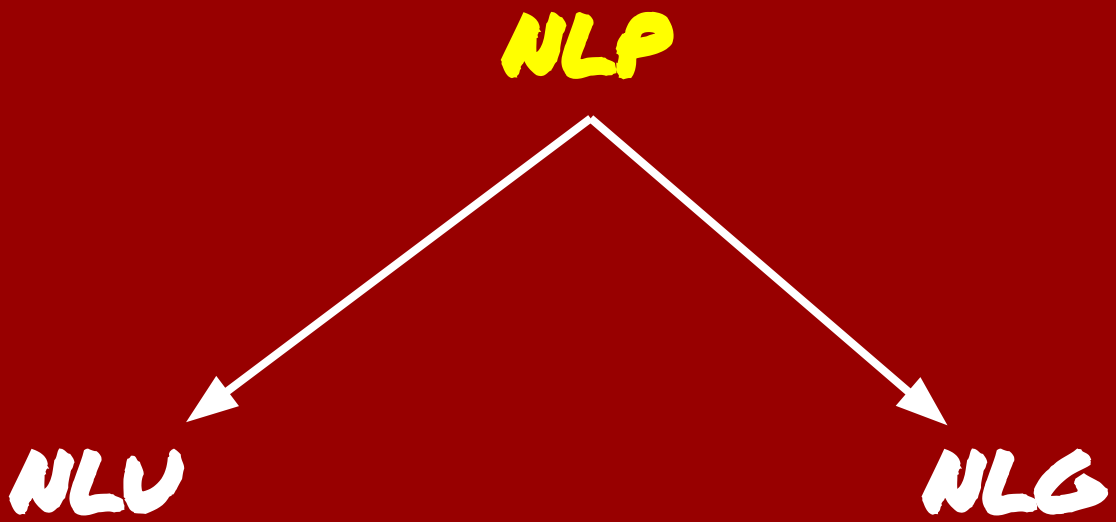
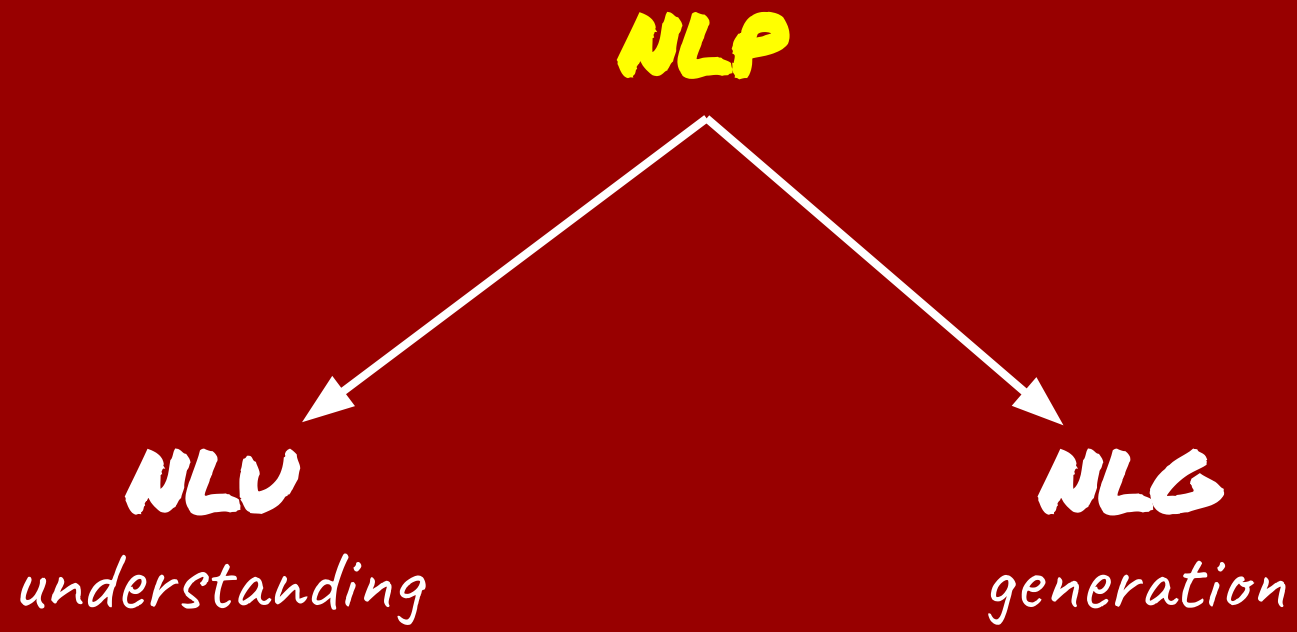
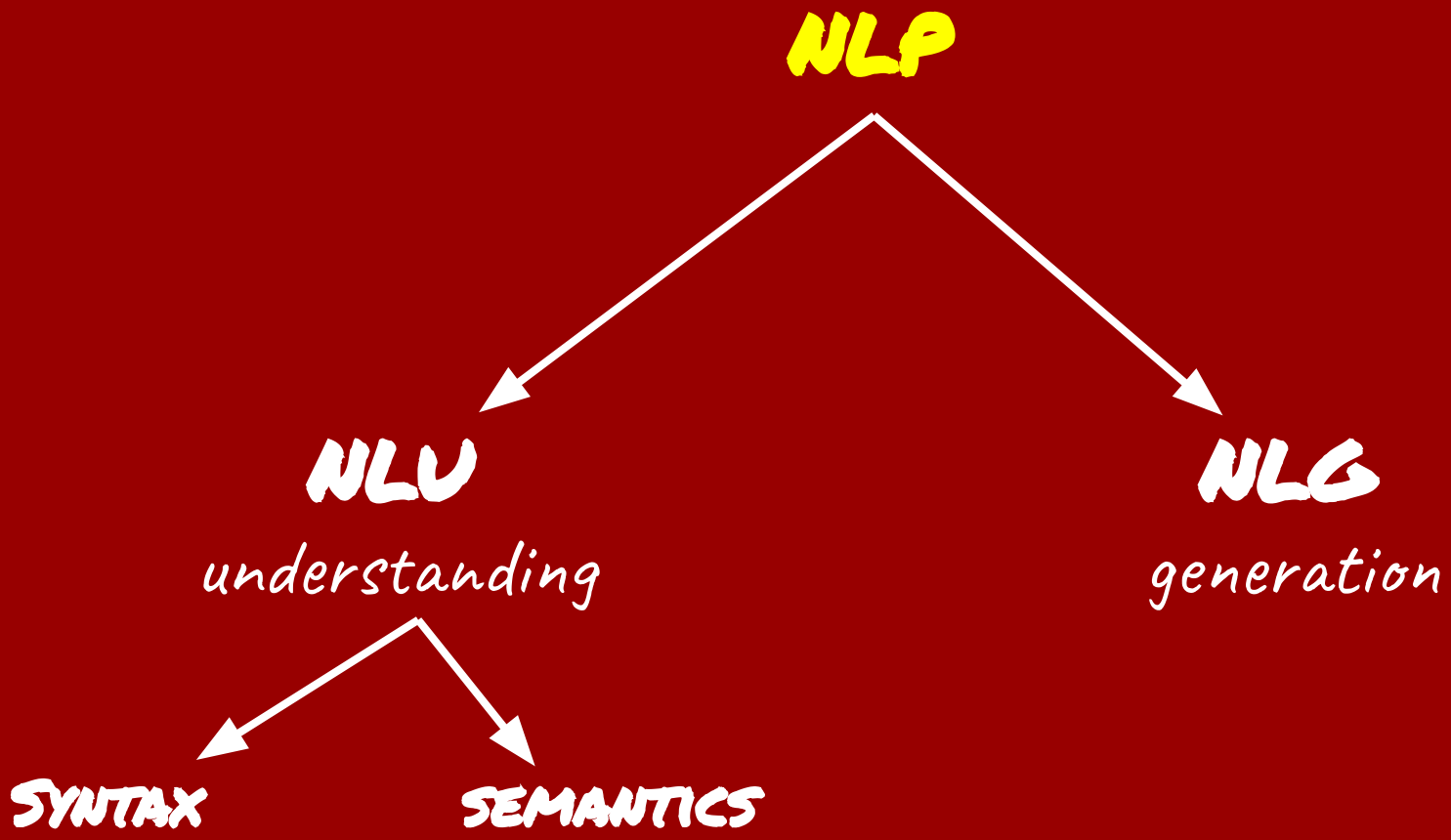
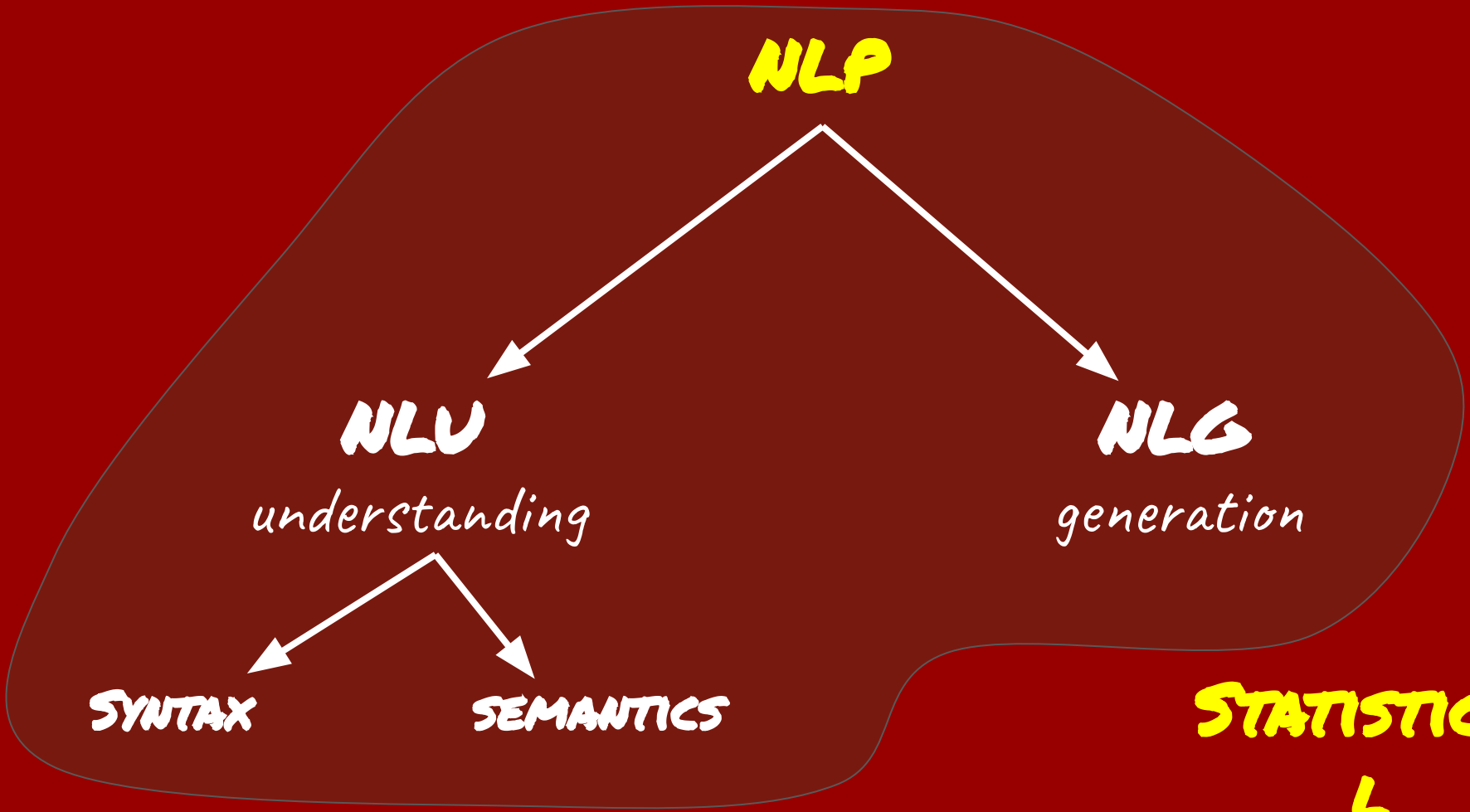
Language Models
&
The transformer


NLP
NLU
NLG


NLP
NLU
understanding
NLG
generation


NLP
NLU
understanding
NLG
generation
Syntax
semantics

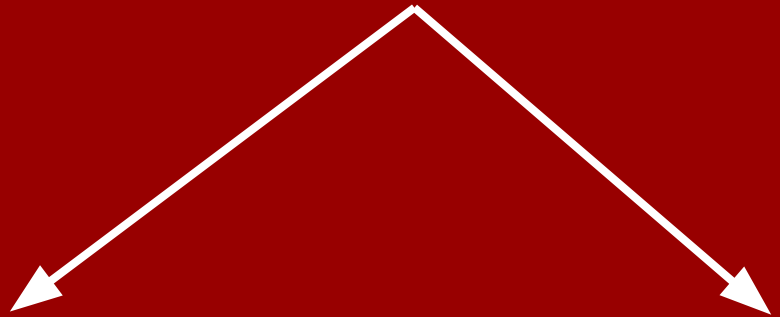


STATISTICA
L

What exactly is a language model??

LANGUAGE REPRESENTATION

LANGUAGE REPRESENTATION

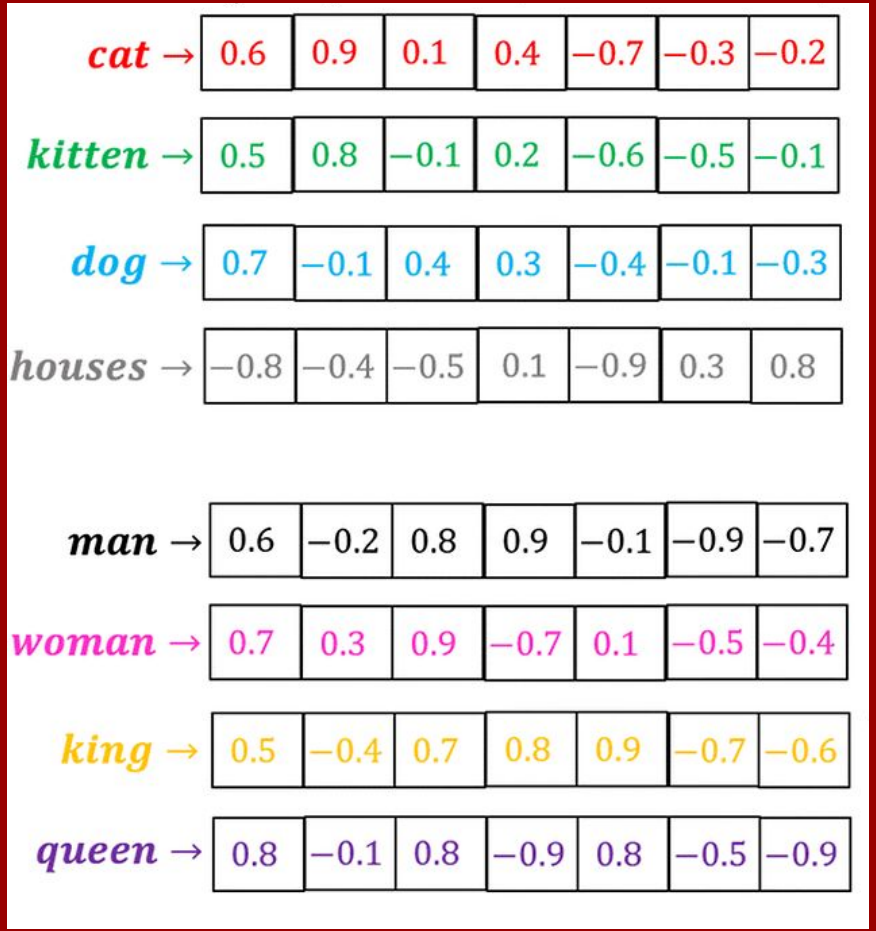


FEATURE EXTRACTIO

Classical ML

EMBEDDING

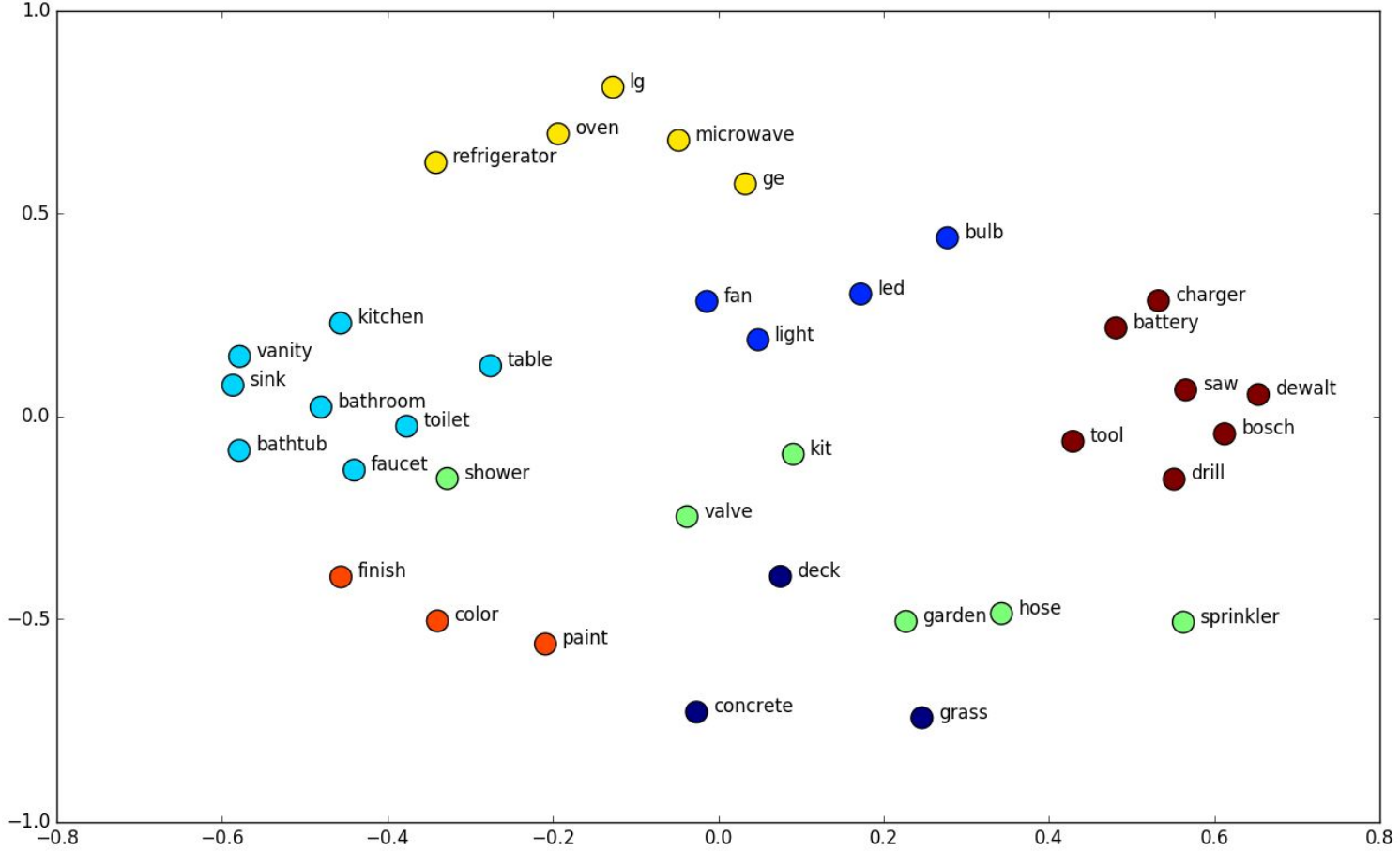
Deep & LLMs



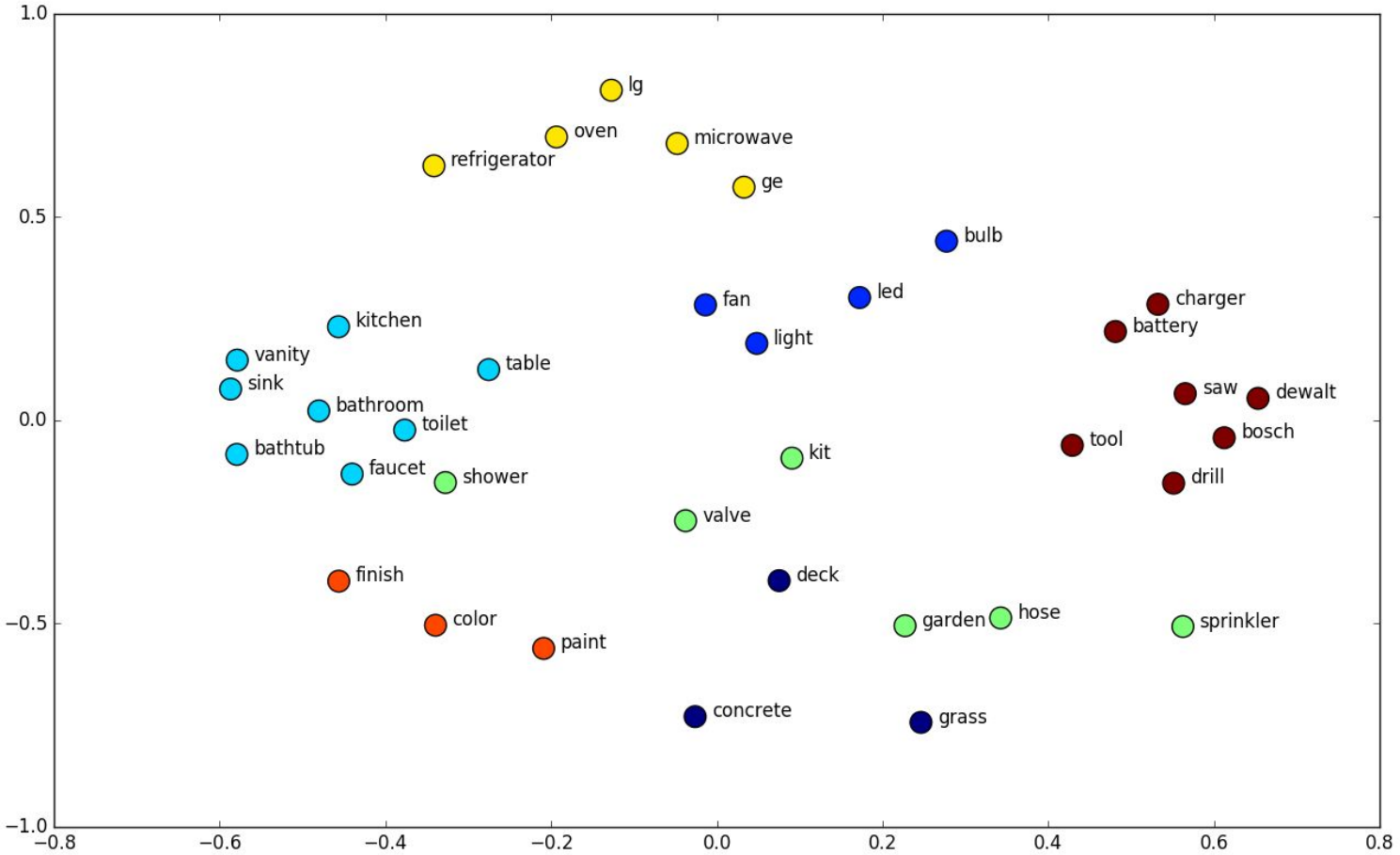
WORD EMBEDDINGS

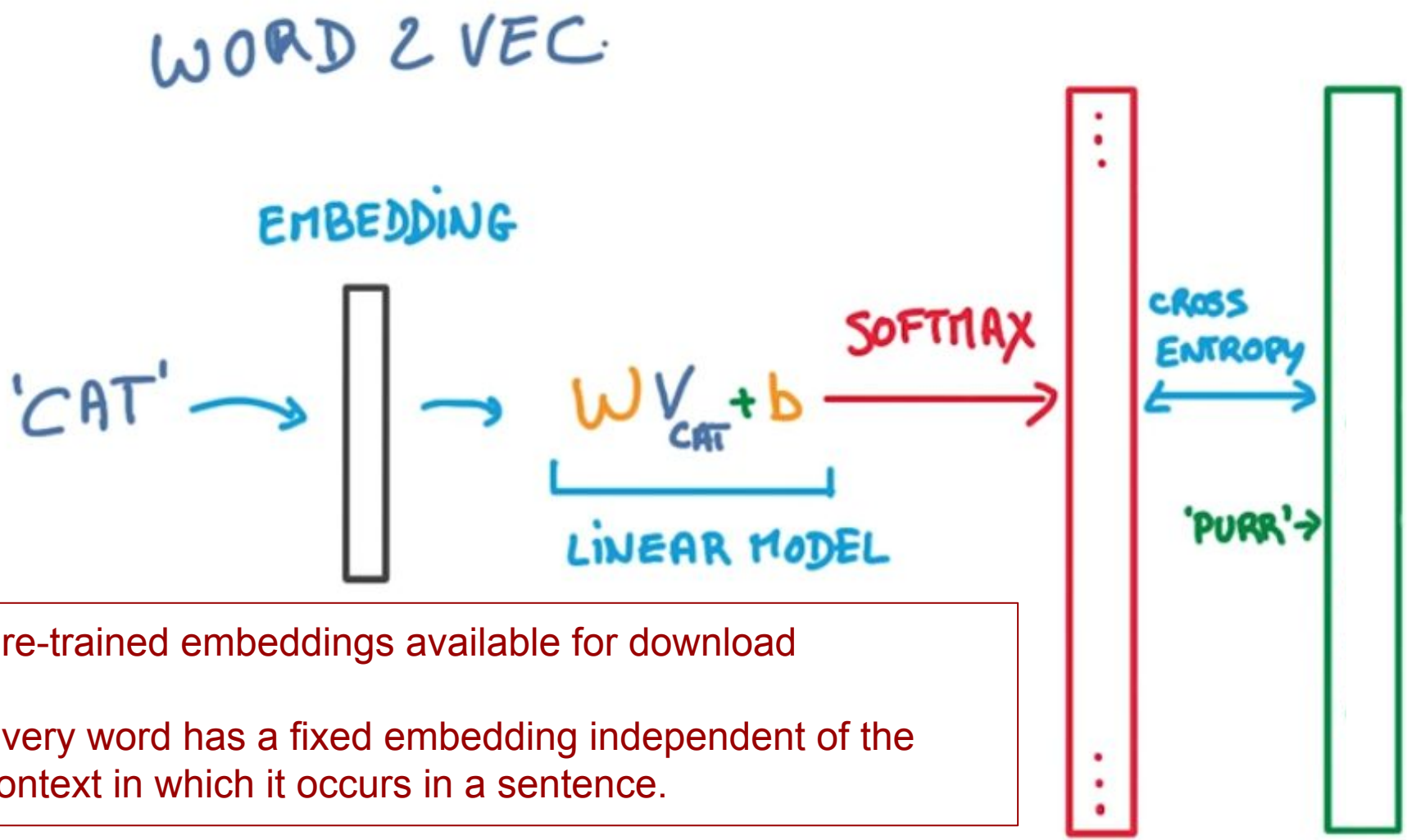
Representing words by a vector of numbers

How long should these vectors be?


lg
oven
microwave
refrigerator
ge
bulb
fan
led
charger
kitchen
battery
light
vanity
table
sink
saw
dewalt
bathroom
toilet
tool
bosch
bathtub
kit
faucet
shower
drill
valve
finish
deck
color
garden
hose
sprinkler
paint
concrete
grass
1.0
0.5
0.0
−0.5
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

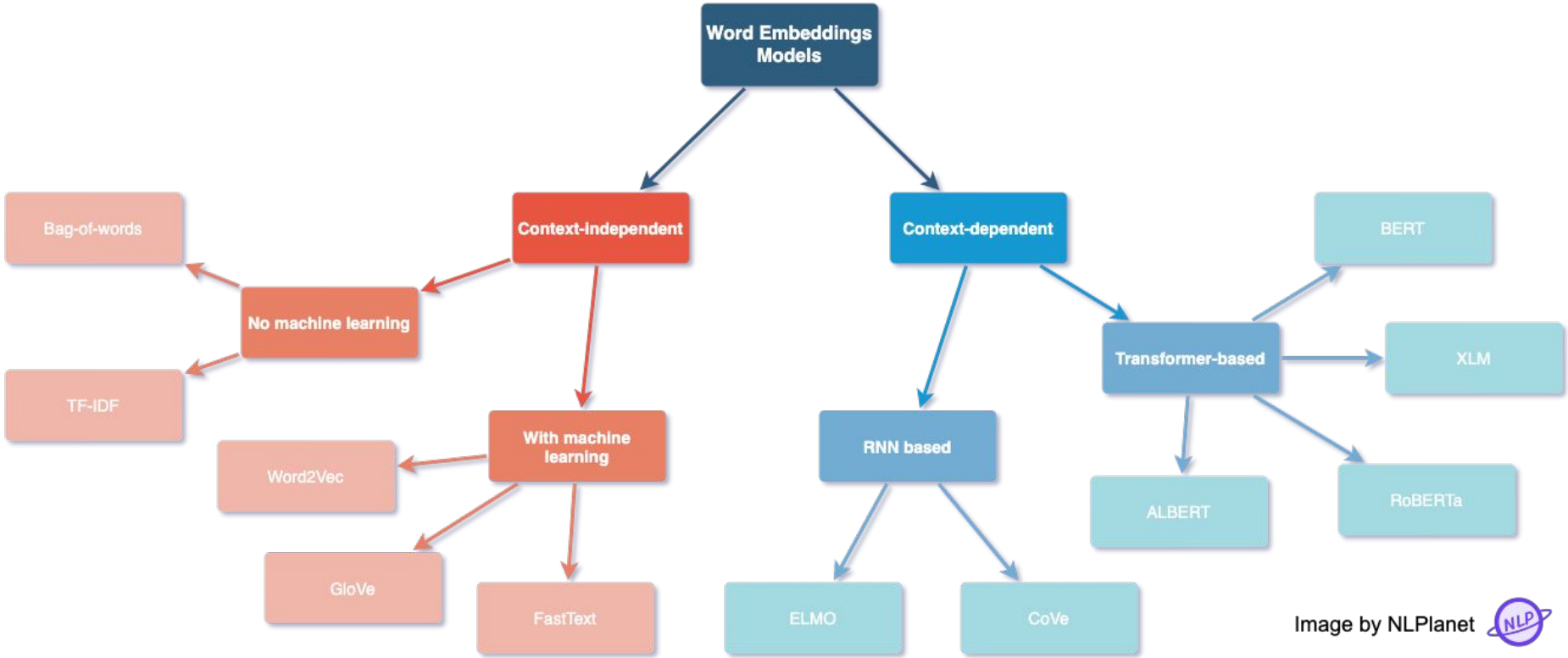
How do we know that the vector representations are correct?





Pre-trained embeddings available for download

Every word has a fixed embedding independent of the context in which it occurs in a sentence.


Word Embeddings Models
Context-independent
Context-dependent
Bag-of-words
No machine learning
TF-IDF
With machine learning
Word2Vec
GloVe
FastText
RNN based
ELMO
CoVe
Transformer-based
BERT
XLM
ALBERT
RoBERTa
Image by NLPlanet
NLP

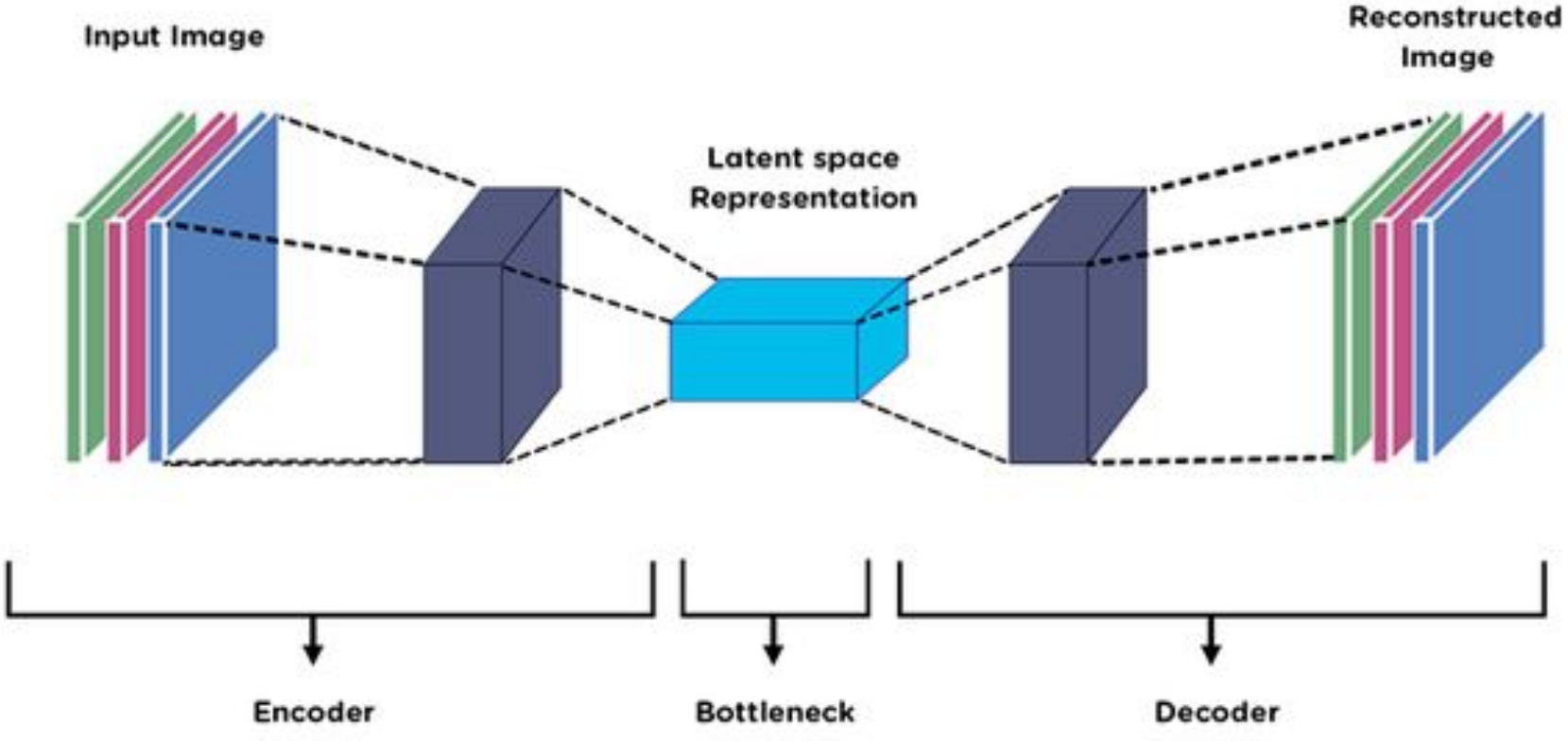
THE TRANSFORMER MODEL

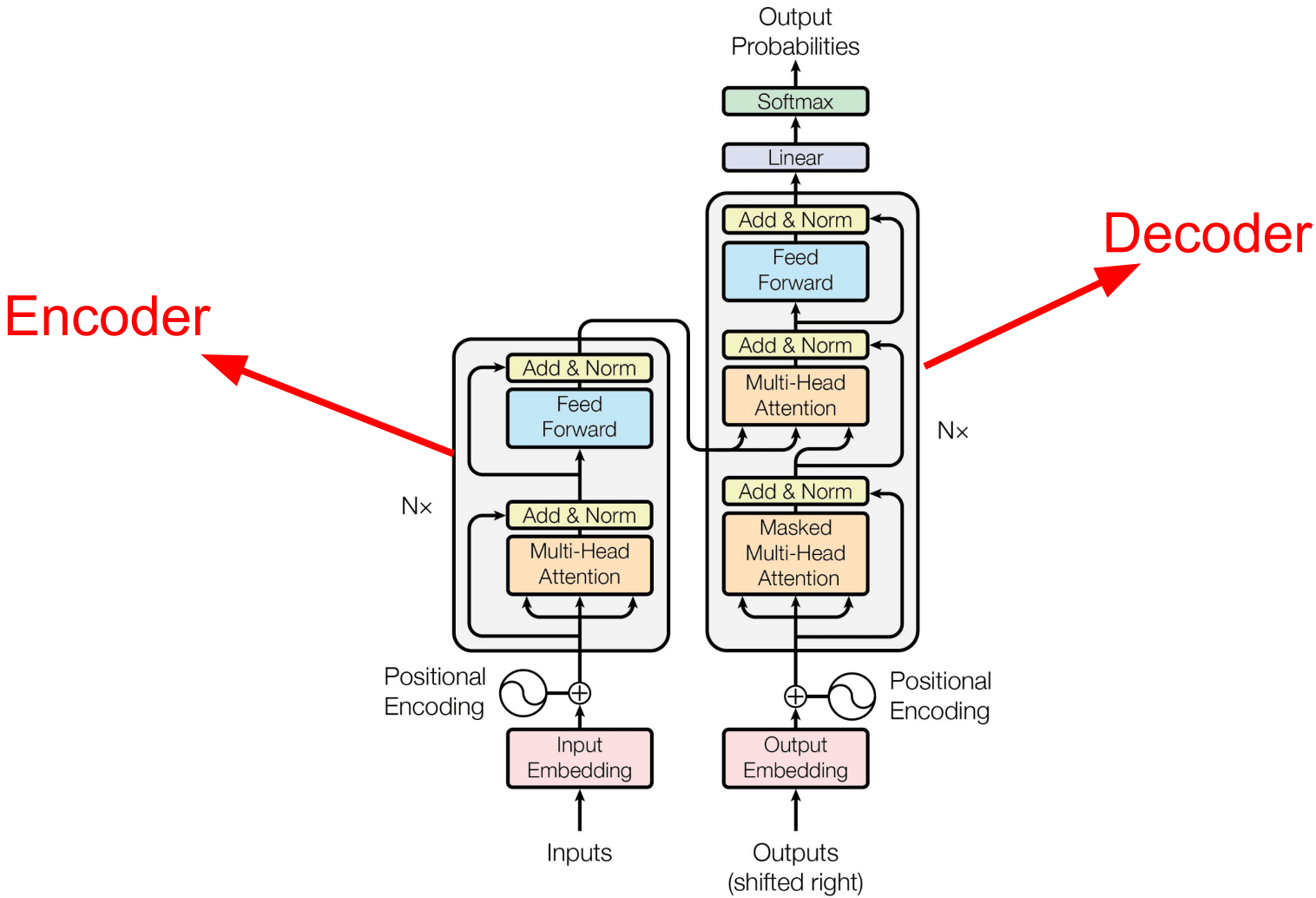
What's the difference between these two devices
in terms of how they treat the incoming information and data?

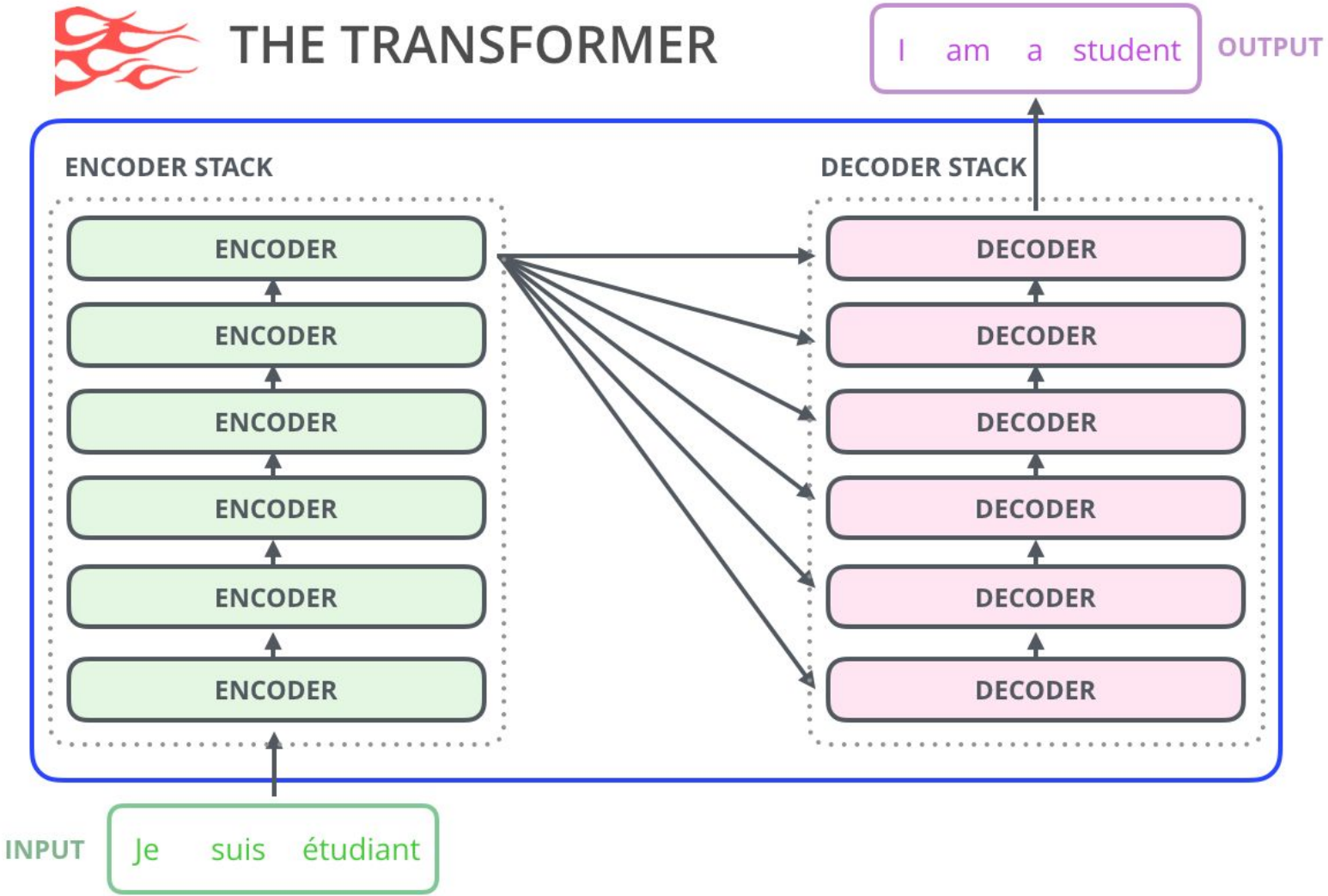
Why is the one on the left considered to be intelligent,
and the one on the right considered to be dumb?

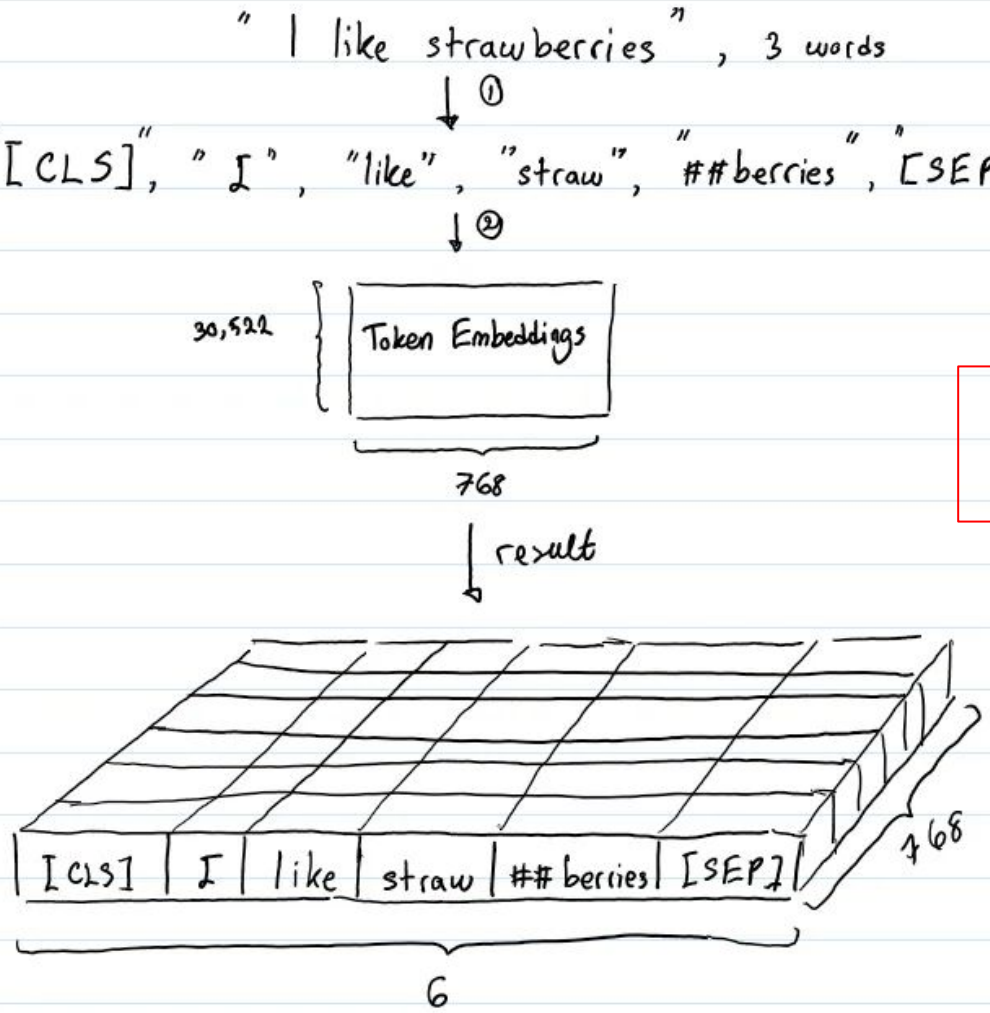
Intelligence is about being able to figure out the **essence** of a topic, and not just memorizing facts.

AUTO-ENCODERS
[lossy compression]

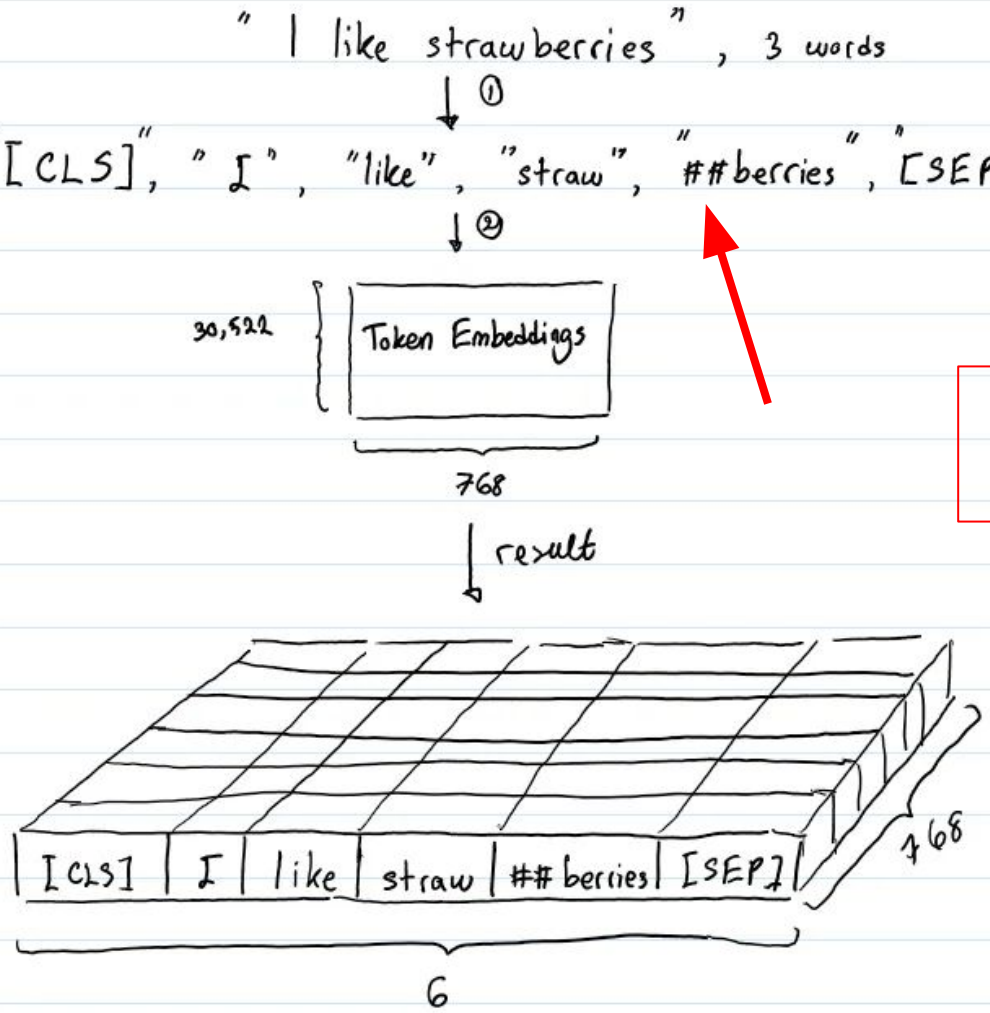



Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Decoder
Encoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)


THE TRANSFORMER
I am a student
OUTPUT
ENCODER STACK
DECODER STACK
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant



Non-contextual Token Embeddings



Non-contextual Token Embeddings

THE TRANSFORMER

ENCODER BLOCK

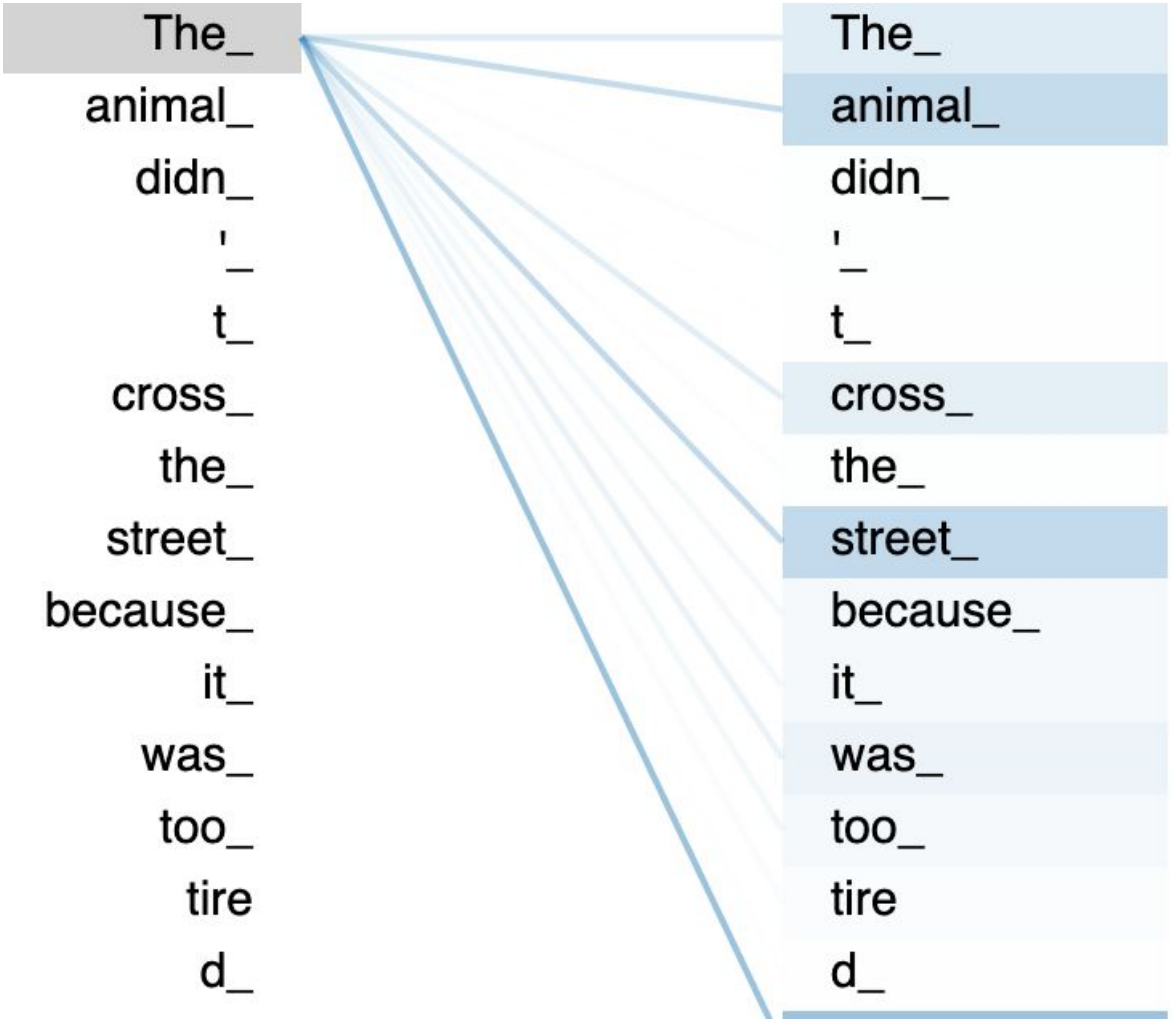
Feed Forward Neural Network

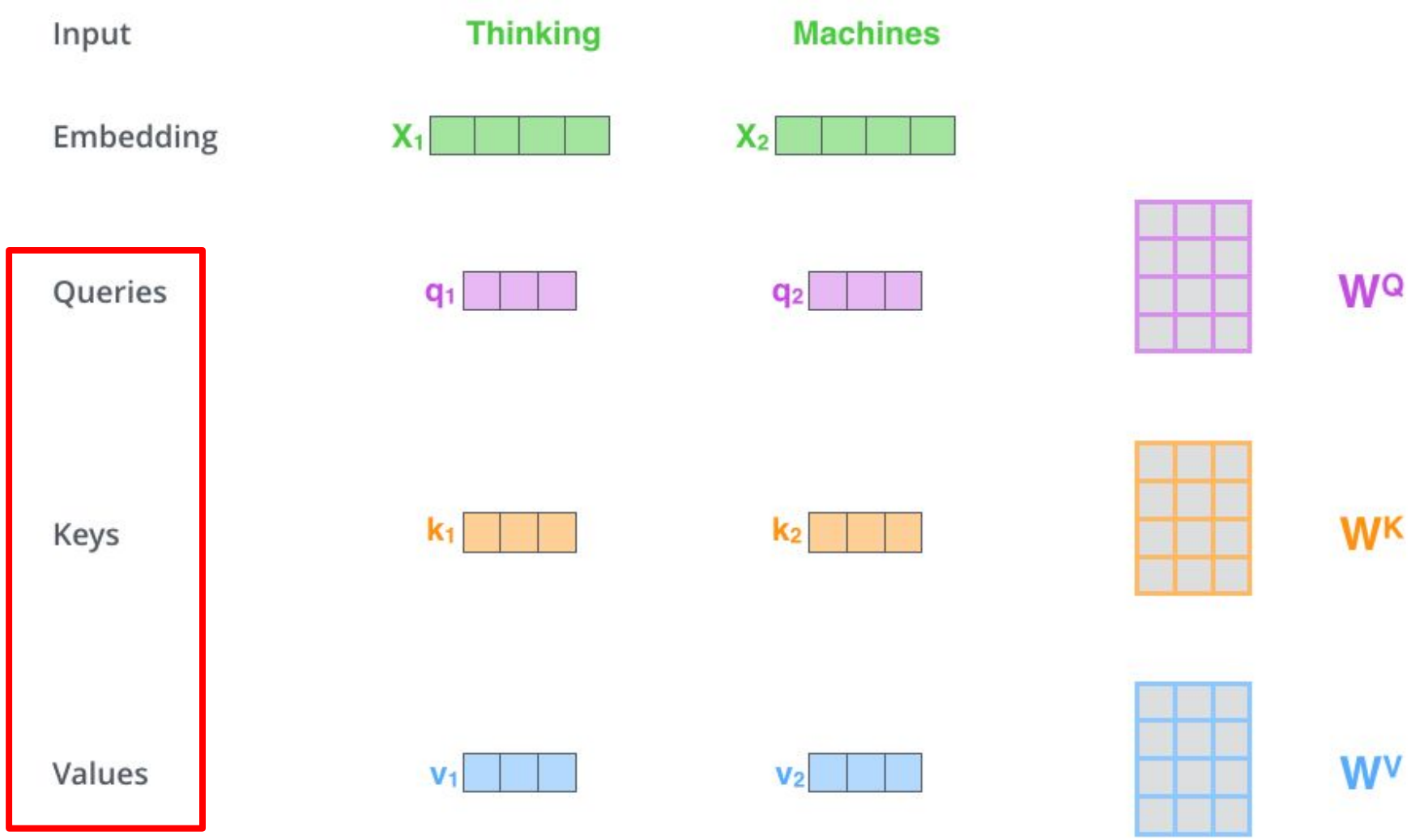
Self-Attention

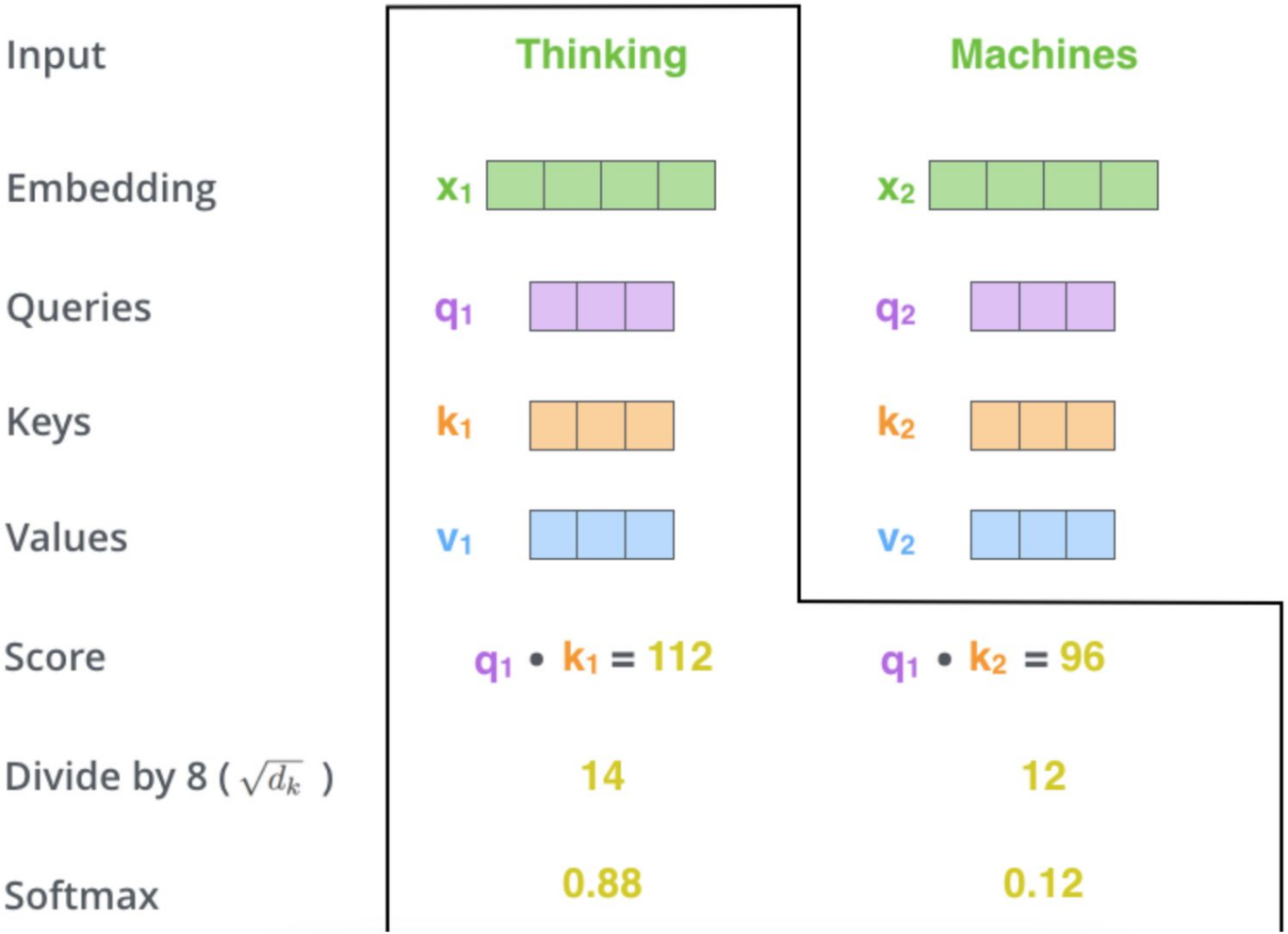
robot	must	obey	orders	<eos>	<pad>	…	<pad>
1	2	3	4	5	6		512

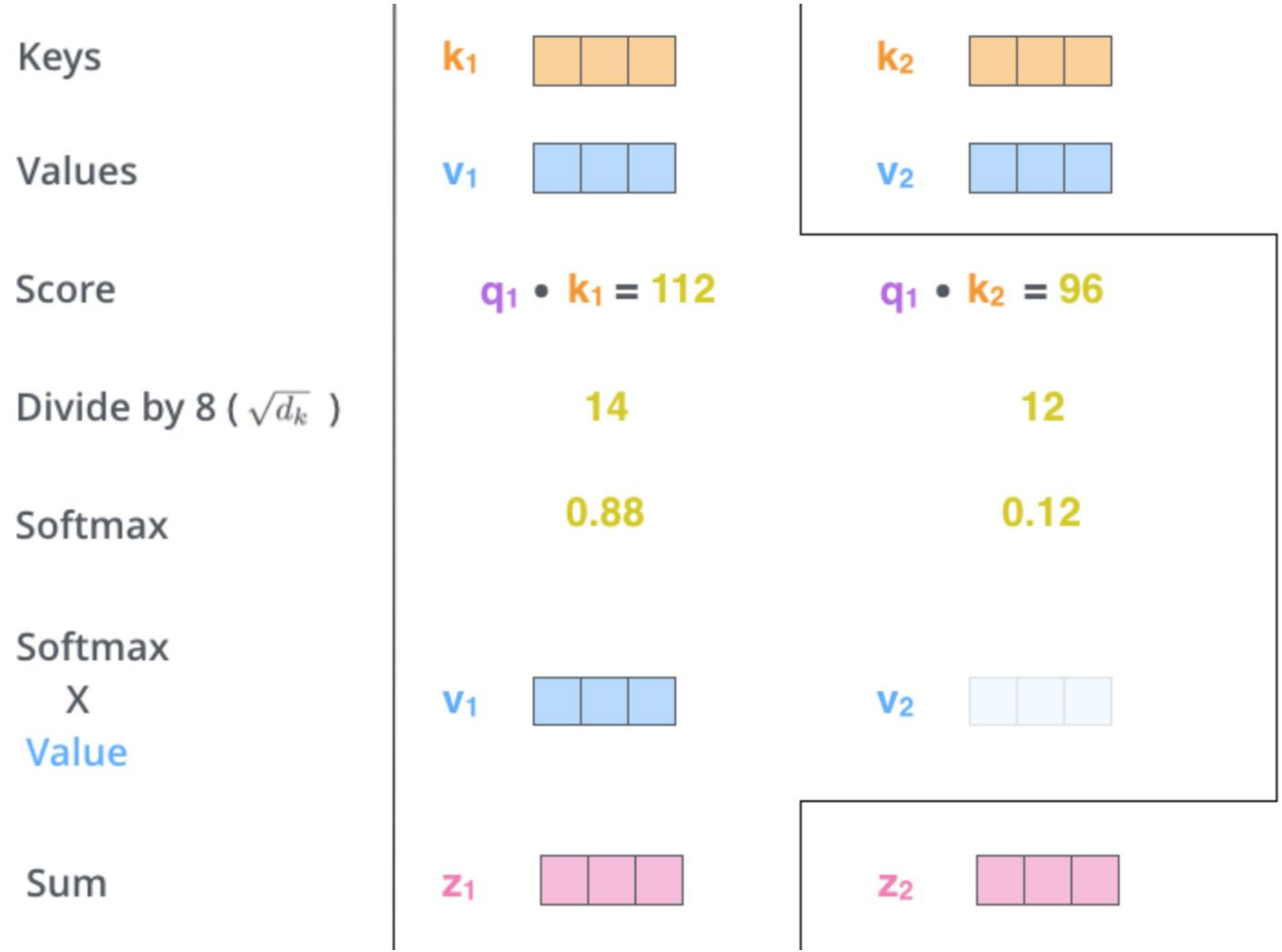
The Attention Mechanism

Taking care of the word sequence is important, but there are also long range dependencies between words.




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
Divide by 8 (√dk)
14
12
Softmax
0.88
0.12


Keys
k_1
k_2
Values
v_1
v_2
Score
$q_1 \cdot k_1 = 112$
$q_1 \cdot k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)
14
12
Softmax
0.88
0.12
Softmax X Value
v_1
v_2
Sum
z_1
z_2

THE TRANSFORMER

DECODER BLOCK

Feed Forward Neural Network

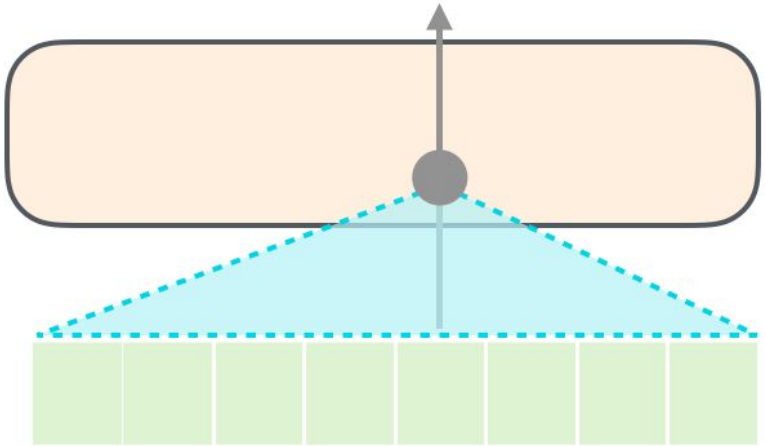
Encoder-Decoder Self-Attention

Masked Self-Attention

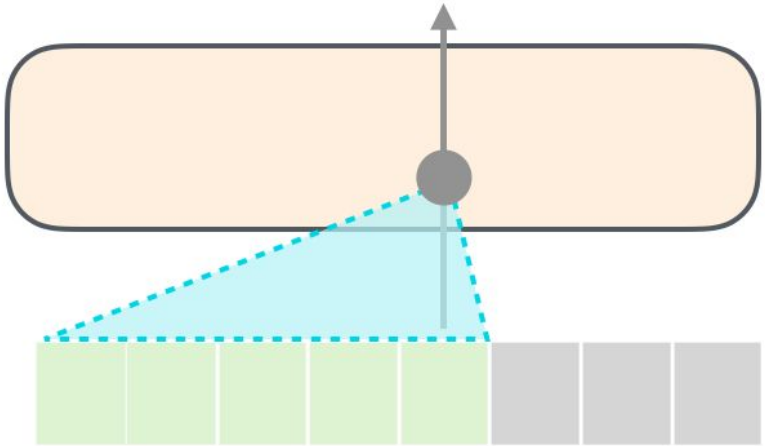
Input

<s>	robot	must	obey				
1	2	3	4	5	6		512

Self-Attention



Masked Self-Attention



THE TRANSFORMER

DECODER BLOCK

Feed Forward Neural Network

Encoder-Decoder Self-Attention

Masked Self-Attention

Input

<s>	robot	must	obey				
1	2	3	4	5	6		512

THE TRANSFORMER

Important for tasks like translation

DECODER BLOCK

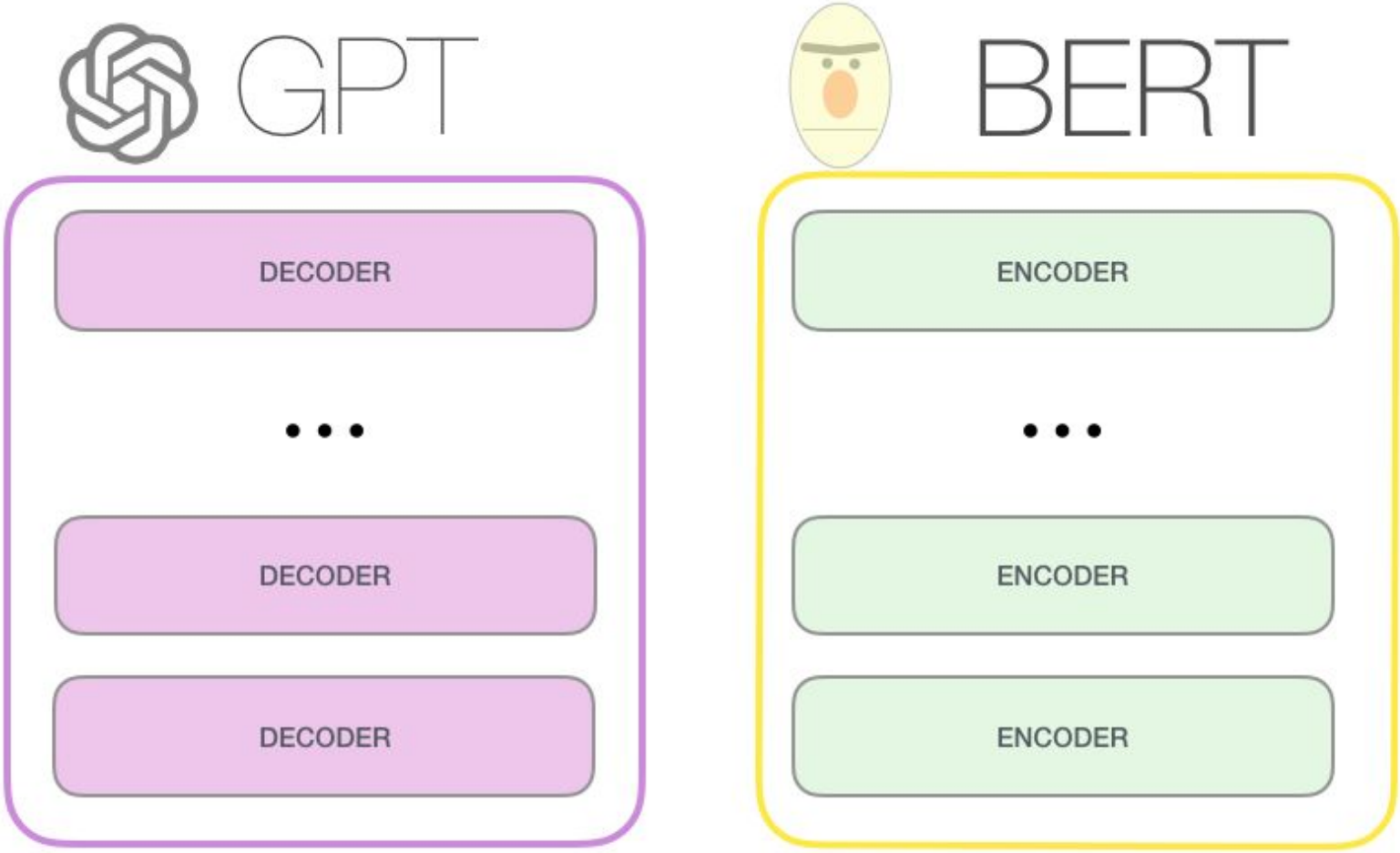
Feed Forward Neural Network

Encoder-Decoder Self-Attention

Masked Self-Attention

Input

<s>	robot	must	obey				
1	2	3	4	5	6		512


GPT
BERT
DECODER
DECODER
DECODER
ENCODER
ENCODER
ENCODER

What exactly is a language model?

How are transformer embeddings different from word2vec?

How are transformer embeddings different from ELMO embeddings?

What exactly is an auto-encoder?

In the transformer model, what does an encoder do?

In the transformer model, what does a decoder do?

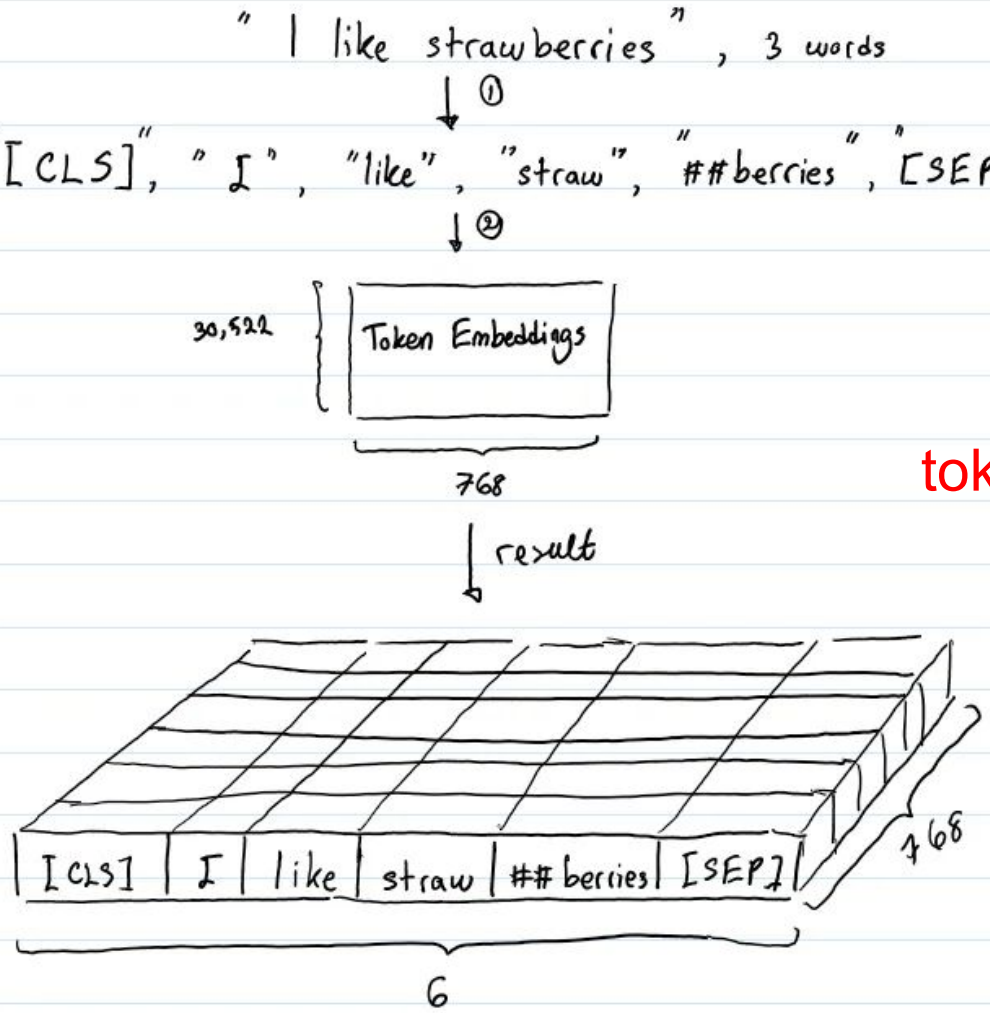
What does the encoder block contain?

What does the decoder block contain?

How is the Self-attention of decoder block different from that of encoder block?

Why does not the GPT model have any encoder block?

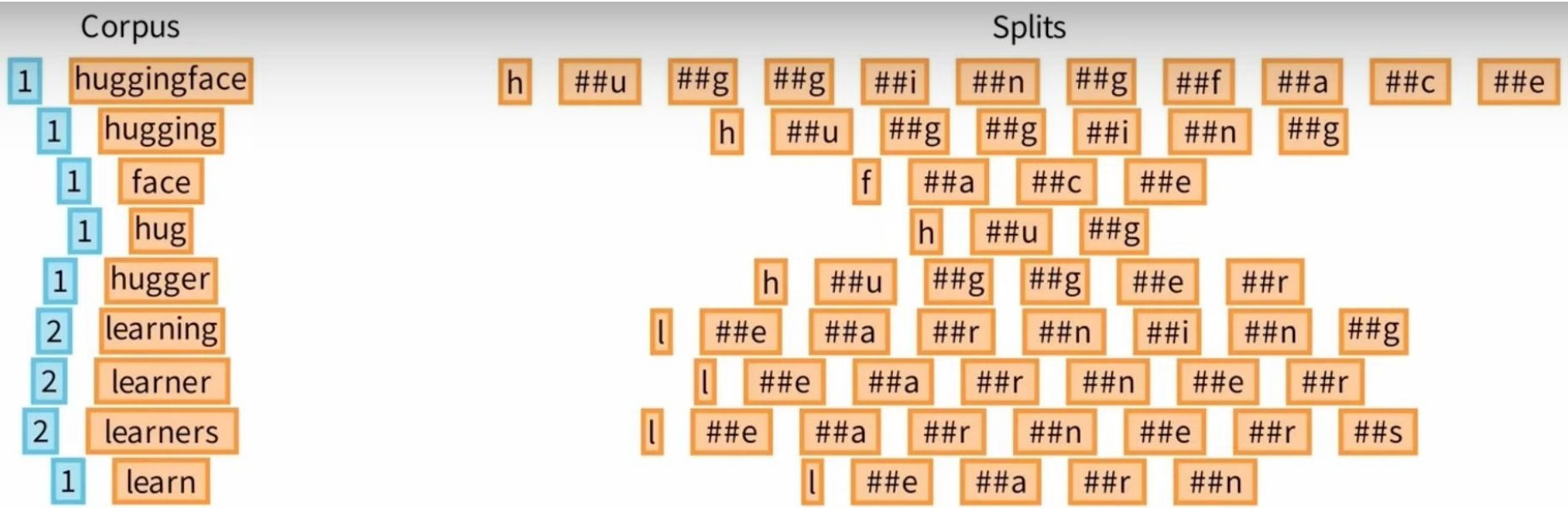
Why does not the BERT model have any decoder block?



How does tokenization work?

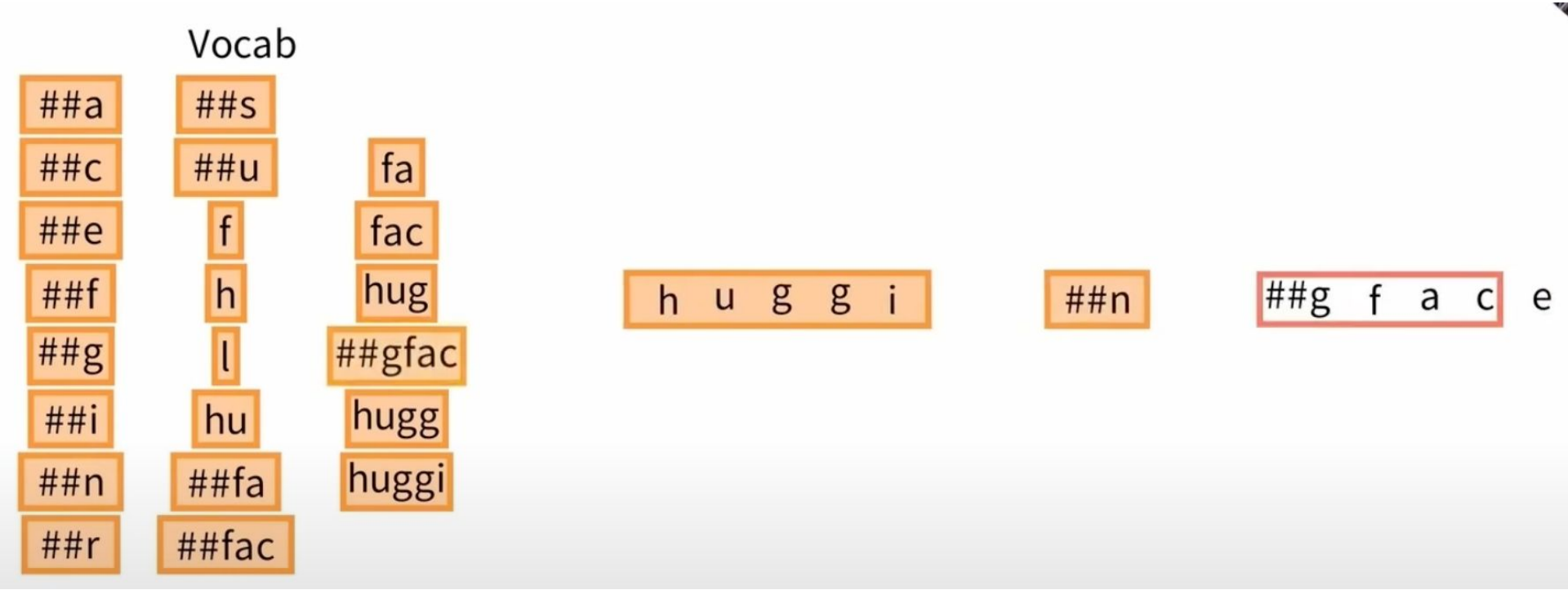
SUB-WORD TOKENIZERS for TRANSFORMERS

Tokenizer	By	Used In	Merge Criteria	Advantage
WordPiece	Google	BERT	Normalized Score	More context
Byte Pair Encoding (BPE)	Philip Gage	GPT	Sub-word frequency	Faster training
SentencePiece	Google	Llama, XLNet, T5, PaLM	Same as BPE	Language independent


Corpus
Splits
1 huggingface
h ##u ##g ##g ##i ##n ##g ##f ##a ##c ##e
1 hugging
h ##u ##g ##g ##i ##n ##g
1 face
f ##a ##c ##e
1 hug
h ##u ##g
1 hugger
h ##u ##g ##g ##e ##r
2 learning
l ##e ##a ##r ##n ##i ##n ##g
2 learner
l ##e ##a ##r ##n ##e ##r
2 learners
l ##e ##a ##r ##n ##e ##r ##s
1 learn
l ##e ##a ##r ##n


Vocab
##a
##c
##e
##f
##g
##i
##n
##r
##s
##u
f
h
l
hu
##fa
##fac
fa
fac
hug
##gfac
hugg
huggi

h u g g i n g f a c e


Vocab
##a
##c
##e
##f
##g
##i
##n
##r
##s
##u
f
h
l
hu
##fa
##fac
fa
fac
hug
##gfac
hugg
huggi
h u g g i
##n
##g f a c e