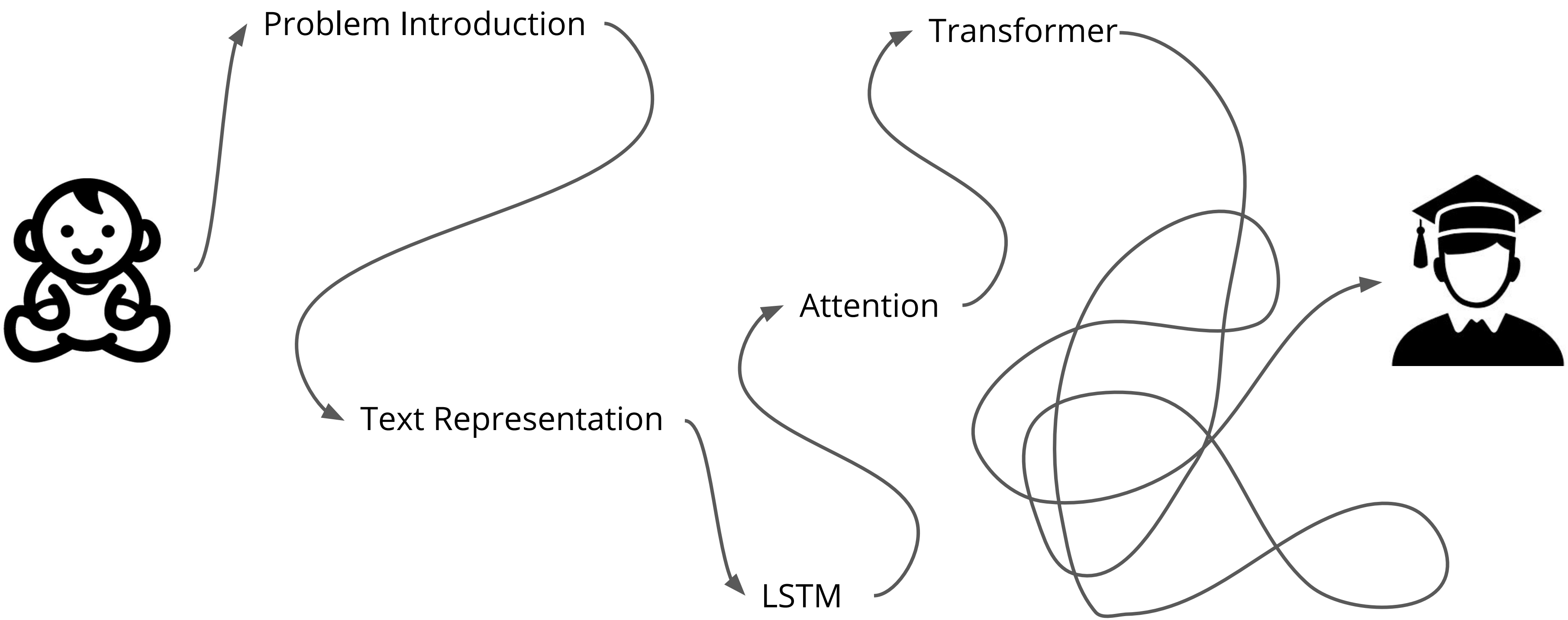
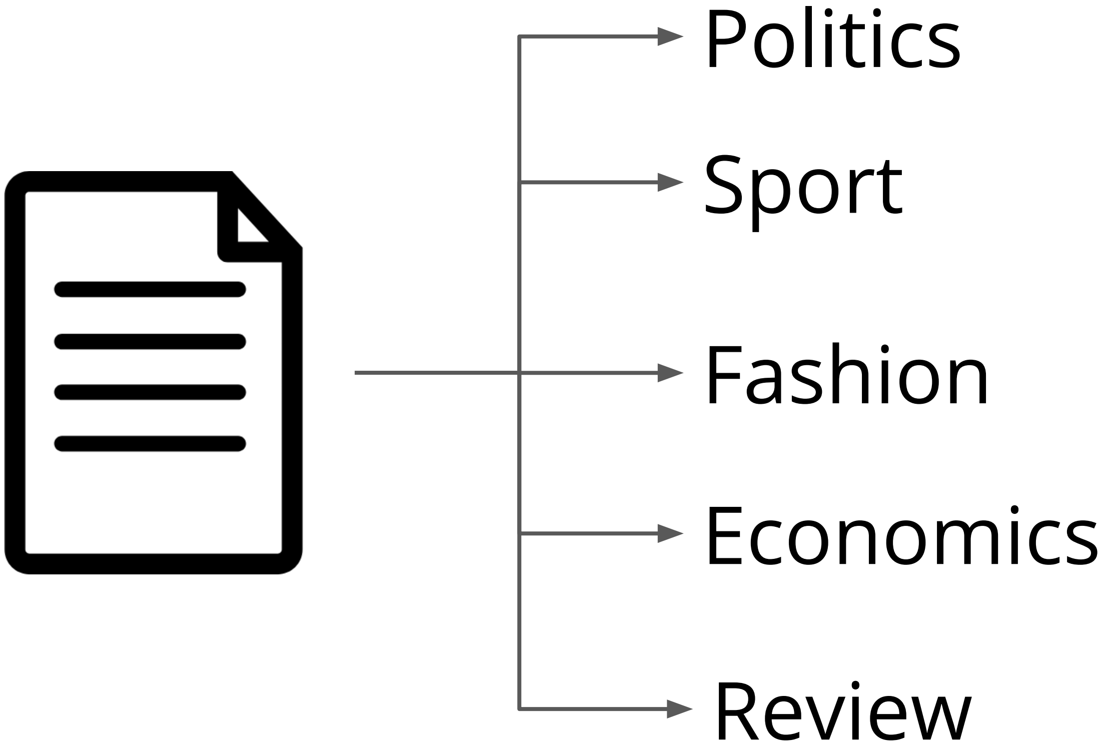
Text Classification

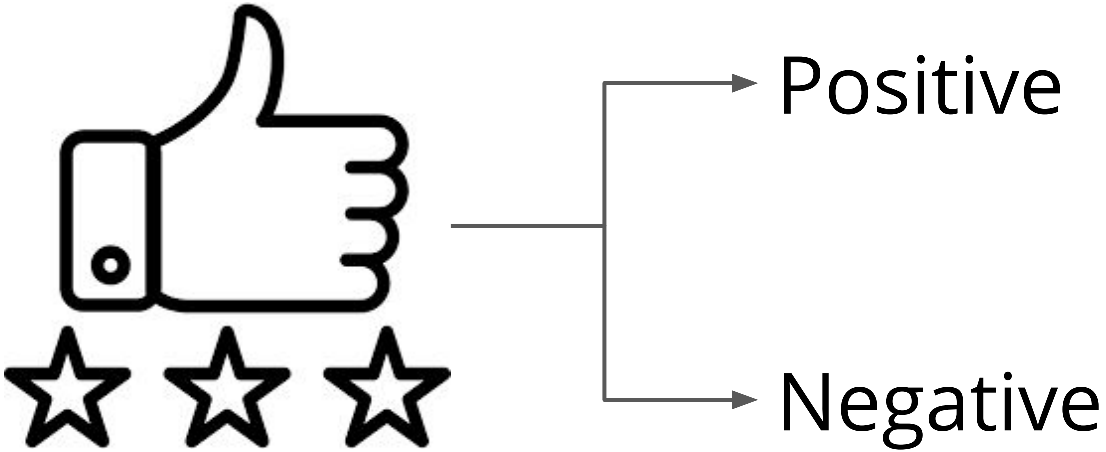
Petr Marek

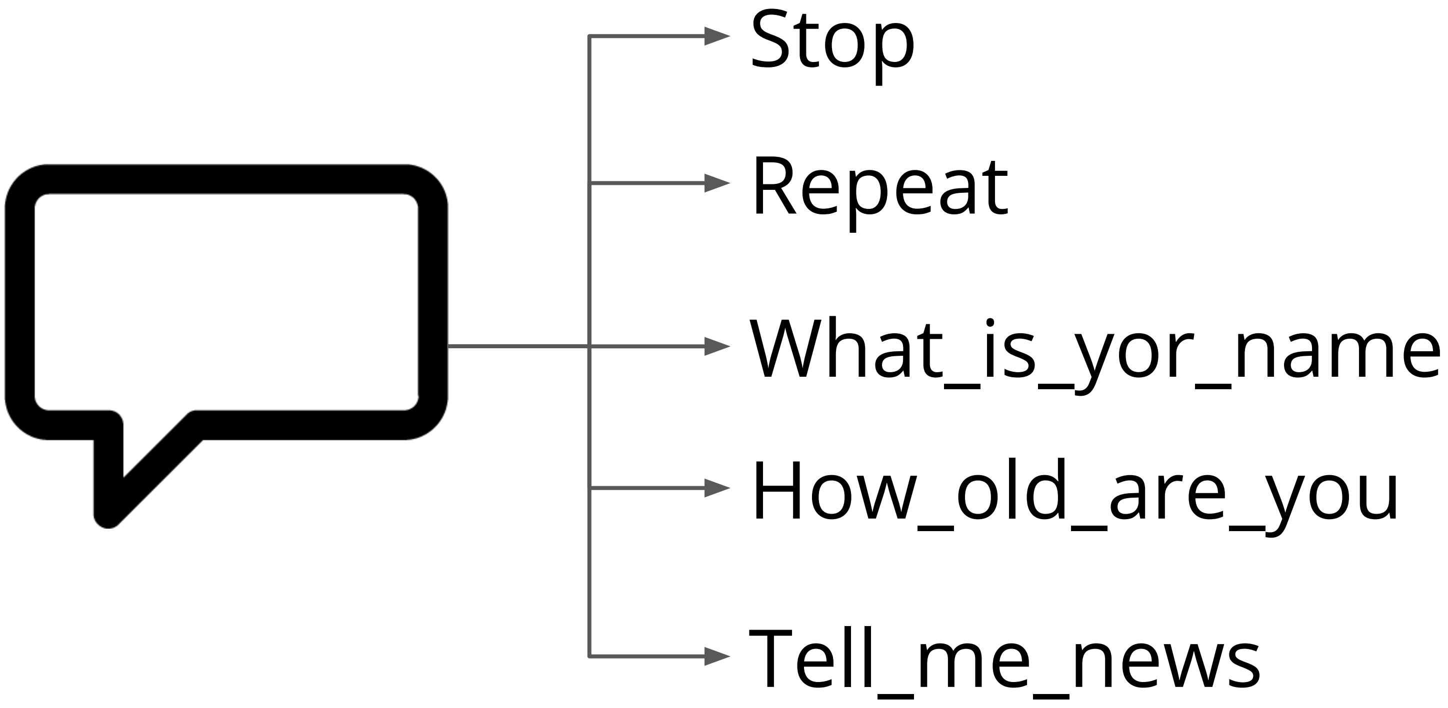
What is on today’s schedule?


Problem Introduction
Text Representation
LSTM
Attention
Transformer

What is the problem?


Politics
Sport
Fashion
Economics
Review


Positive
Negative


Stop
Repeat
What_is_yor_name
How_old_are_you
Tell_me_news

Set of Documents

$D = \{d_1, d_2, d_3, \ldots d_n\}$

Set of classes

$C = \{c_1, c_2, c_3, \ldots c_m\}$

Classifier

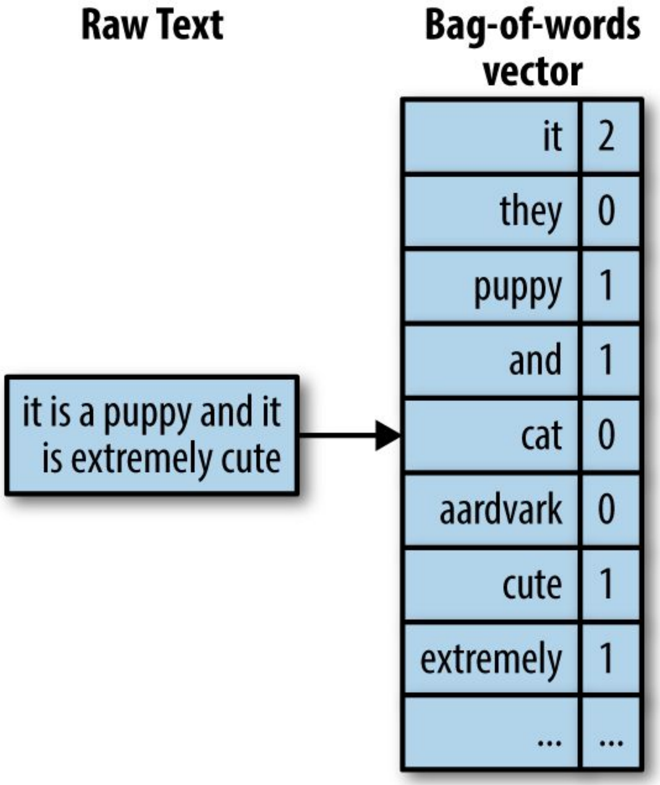
$c_j = f(d_i)$

How to represent text?

String + Machine Learning = 😟

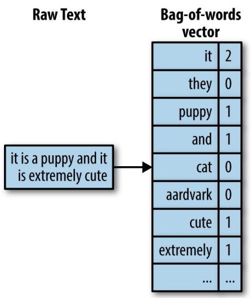
Vector + Machine Learning = 💗

Bag of Words


Raw Text
Bag-of-words vector
it is a puppy and it is extremely cute
it 2
they 0
puppy 1
and 1
cat 0
aardvark 0
cute 1
extremely 1
... ...

Bag of Words

\+ Simple to create



\- Sparse

\- Huge Dimension

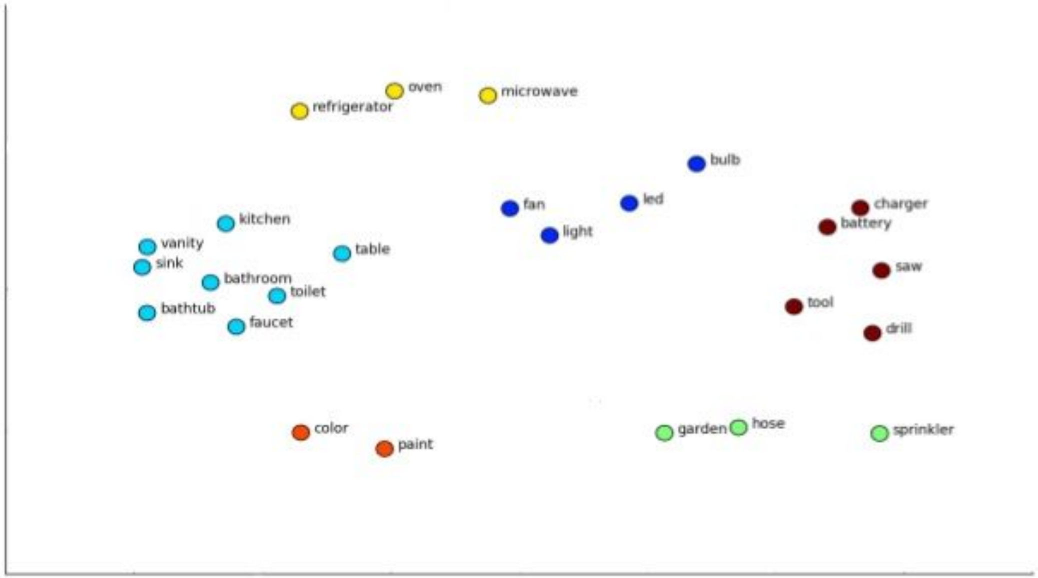
\- No order of words

\- No meaning of words

Embeddings

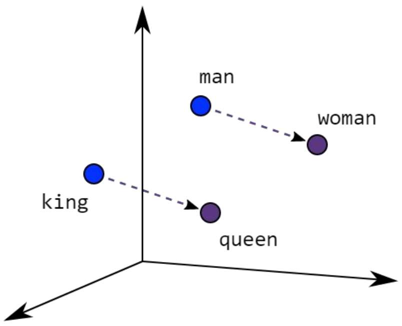
Word vectors	Dimensions			
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

Embeddings

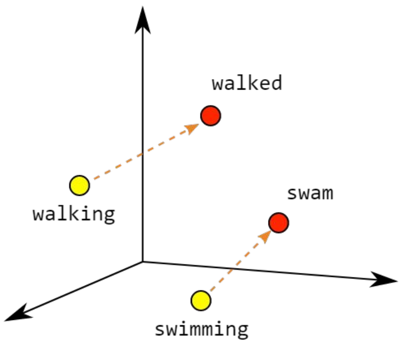


oven
microwave
refrigerator
bulb
fan
led
charger
kitchen
light
battery
vanity
table
sink
saw
bathroom
toilet
tool
bathtub
faucet
drill
color
garden
hose
sprinkler
paint

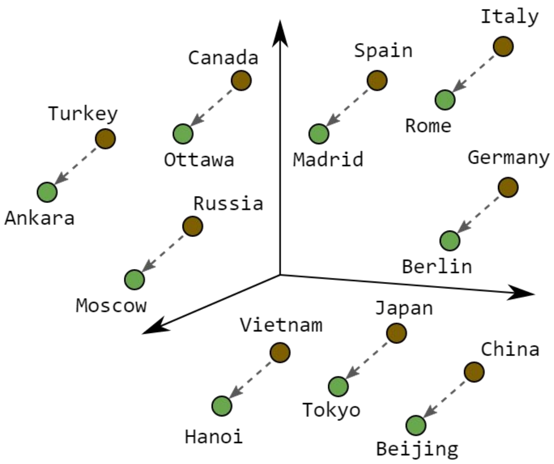
Embeddings



Male-Female

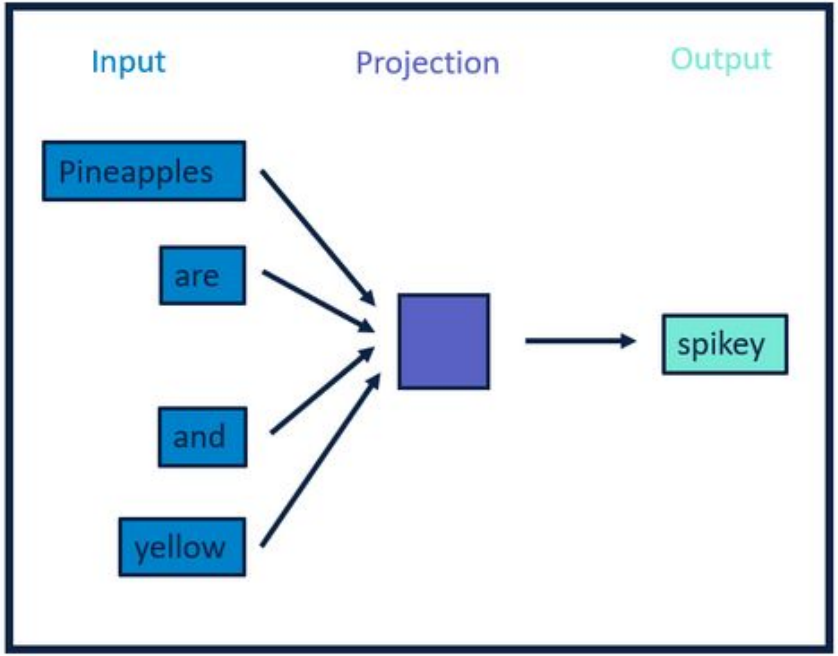


Verb Tense

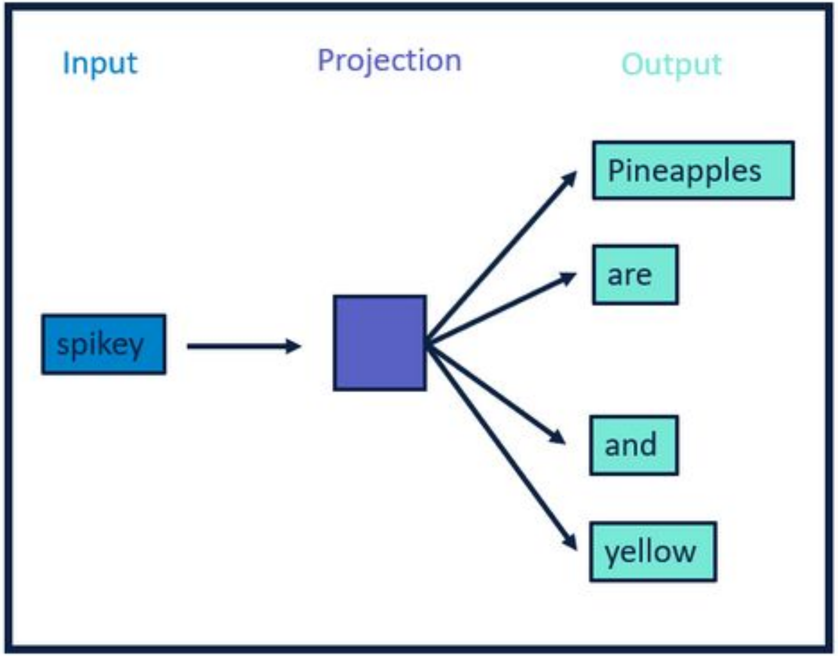


Country-Capital

Embeddings

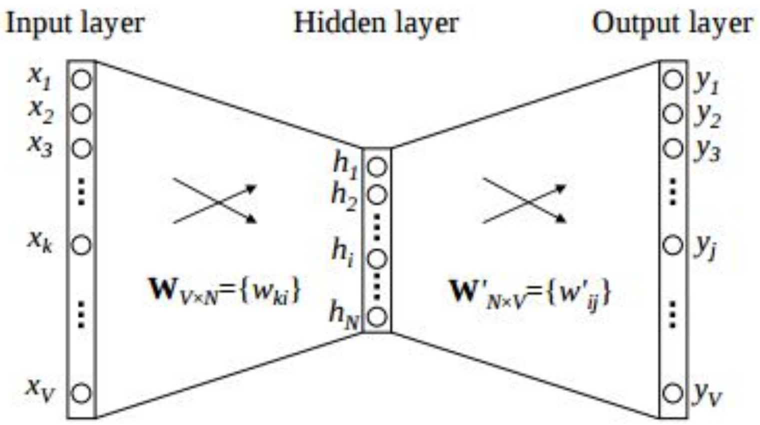


CBOW



Skip-gram

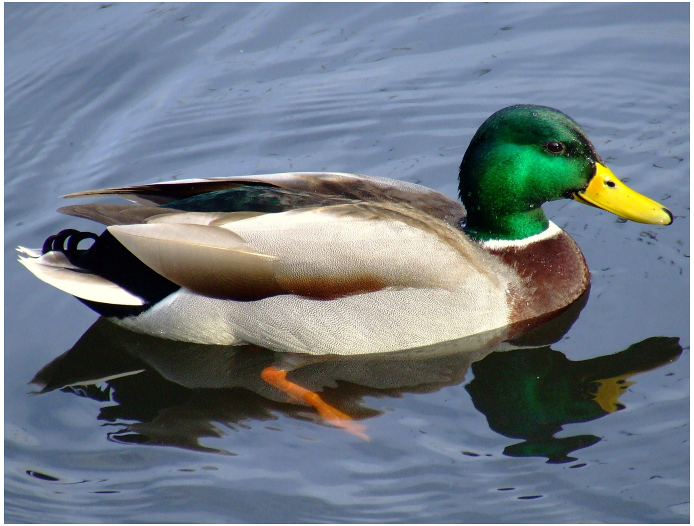
Embeddings

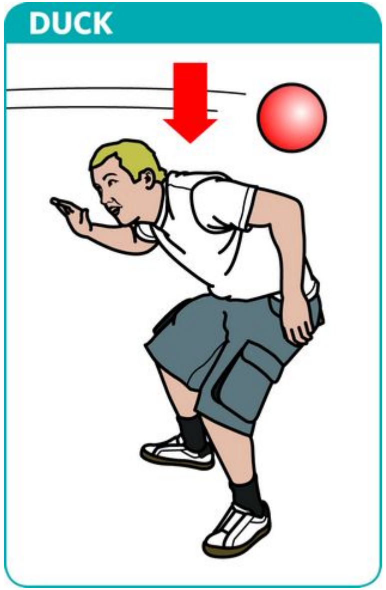


word2vec model architecture

Embeddings

Embeddings





Embeddings

Word2Vec

GloVe

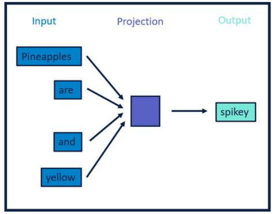
Sent2Vec

fastText

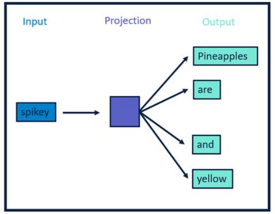
Embeddings

+ Low Dimension
+ Dense representation
+ Similar words has similar meaning

- Some words has multiple meanings



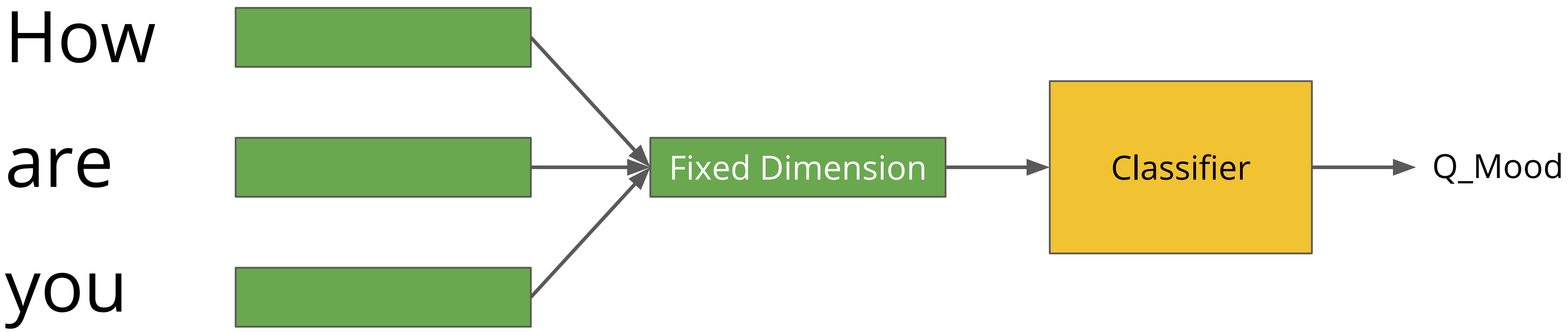
CBOW



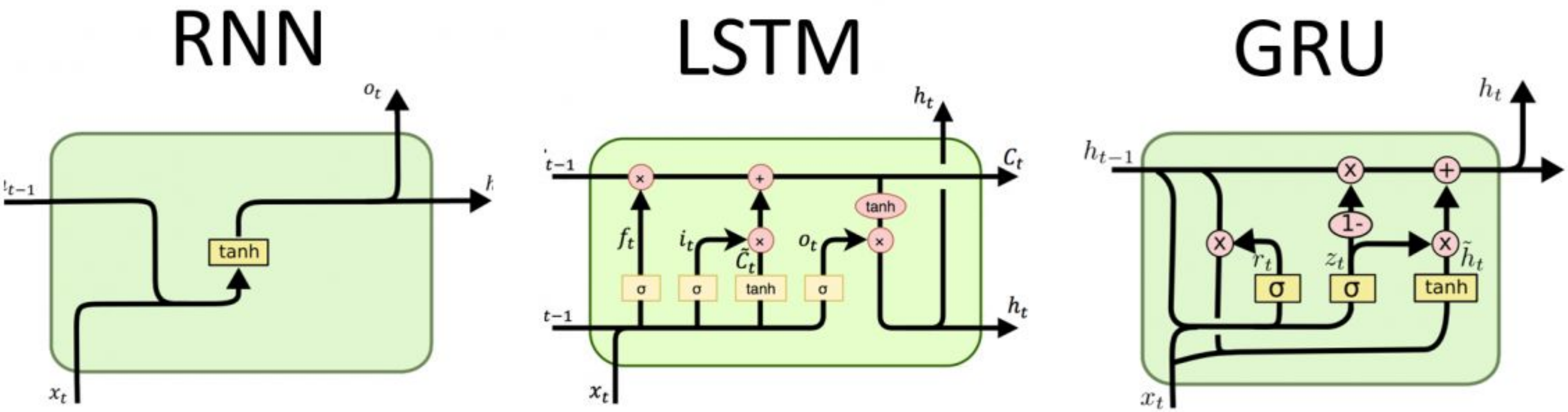
Skip-gram

How to classify?

Average

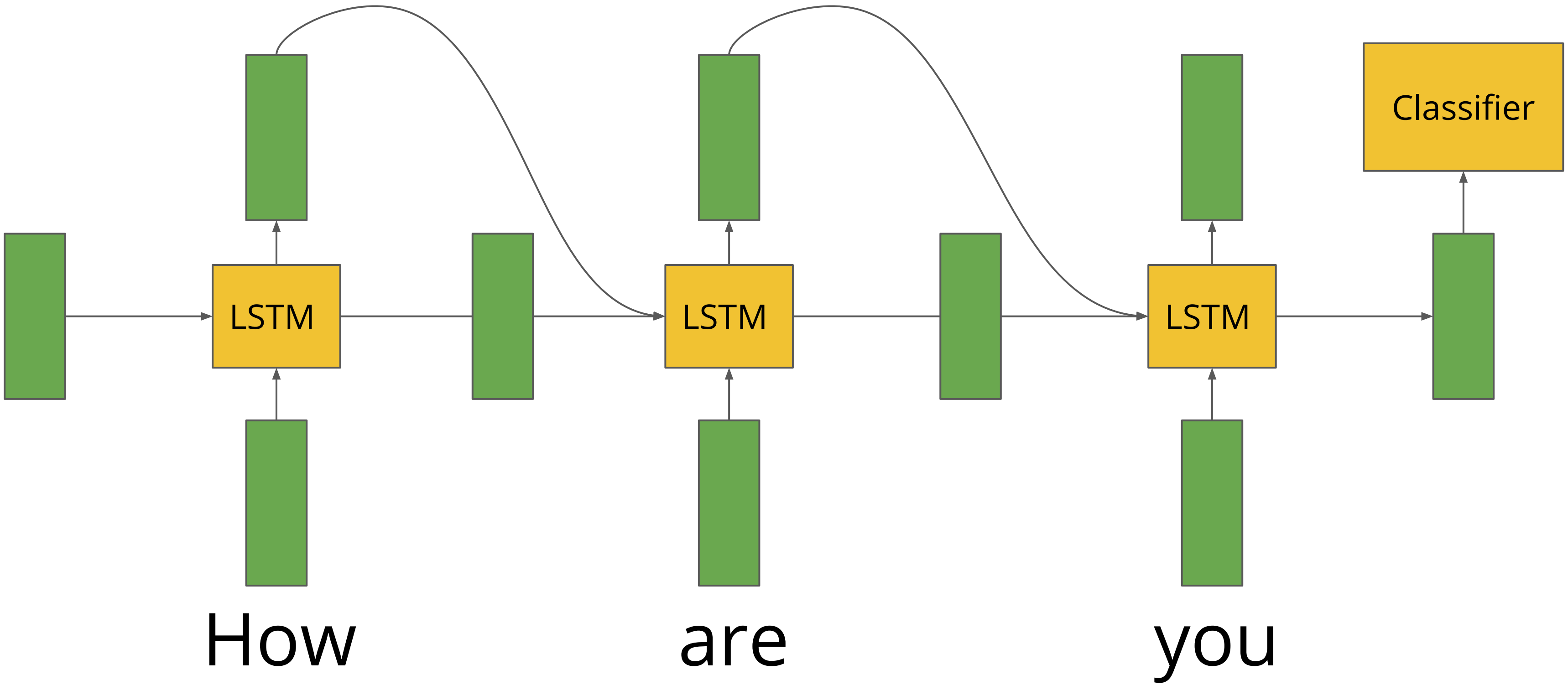


Recurrence

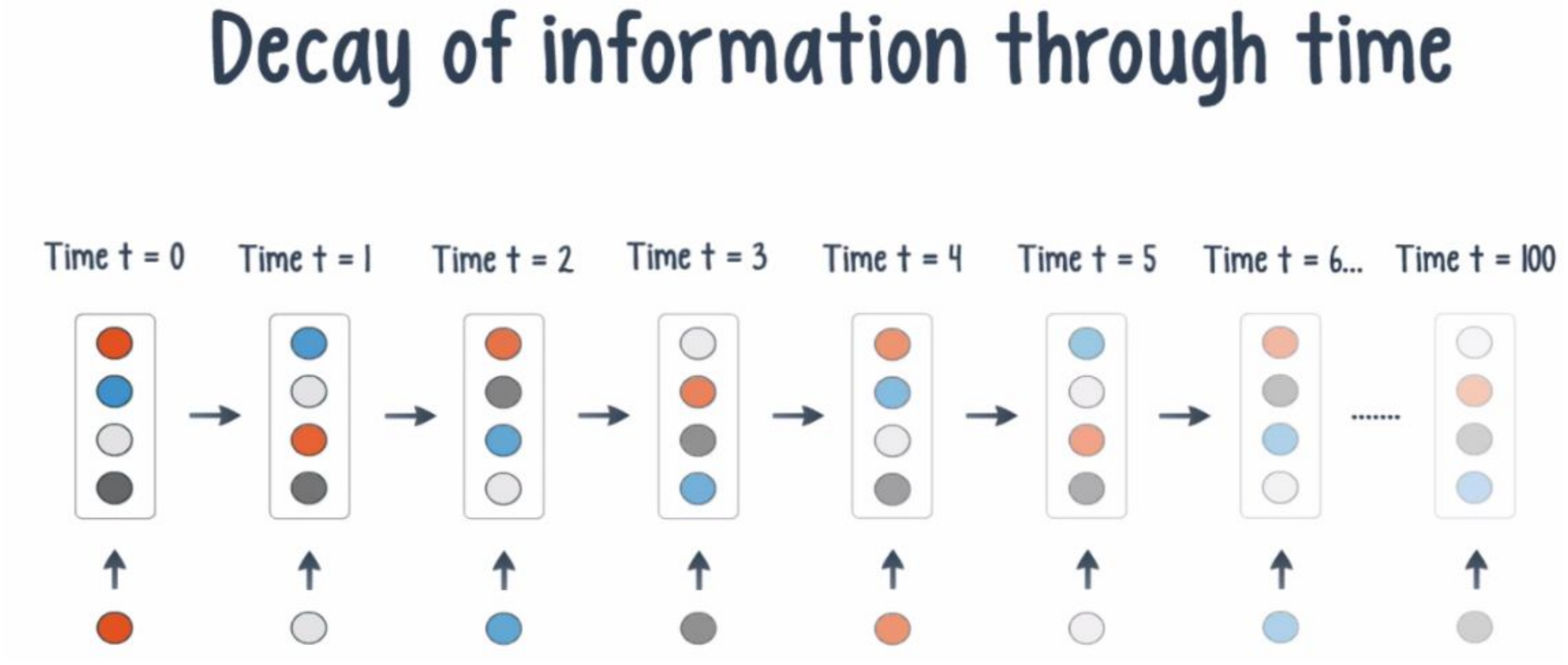


RNN
o_t
tanh
x_t
LSTM
h_t
C_t
f_t
i_t
$\tilde{C}_t$
o_t
tanh
σ
σ
tanh
σ
h_t
x_t
GRU
h_t
h_{t-1}
1-
r_t
z_t
$\tilde{h}_t$
σ
σ
tanh
x_t

Recurrence



Recurrence


Decay of information through time
Time t = 0
Time t = 1
Time t = 2
Time t = 3
Time t = 4
Time t = 5
Time t = 6...
Time t = 100

Recurrence

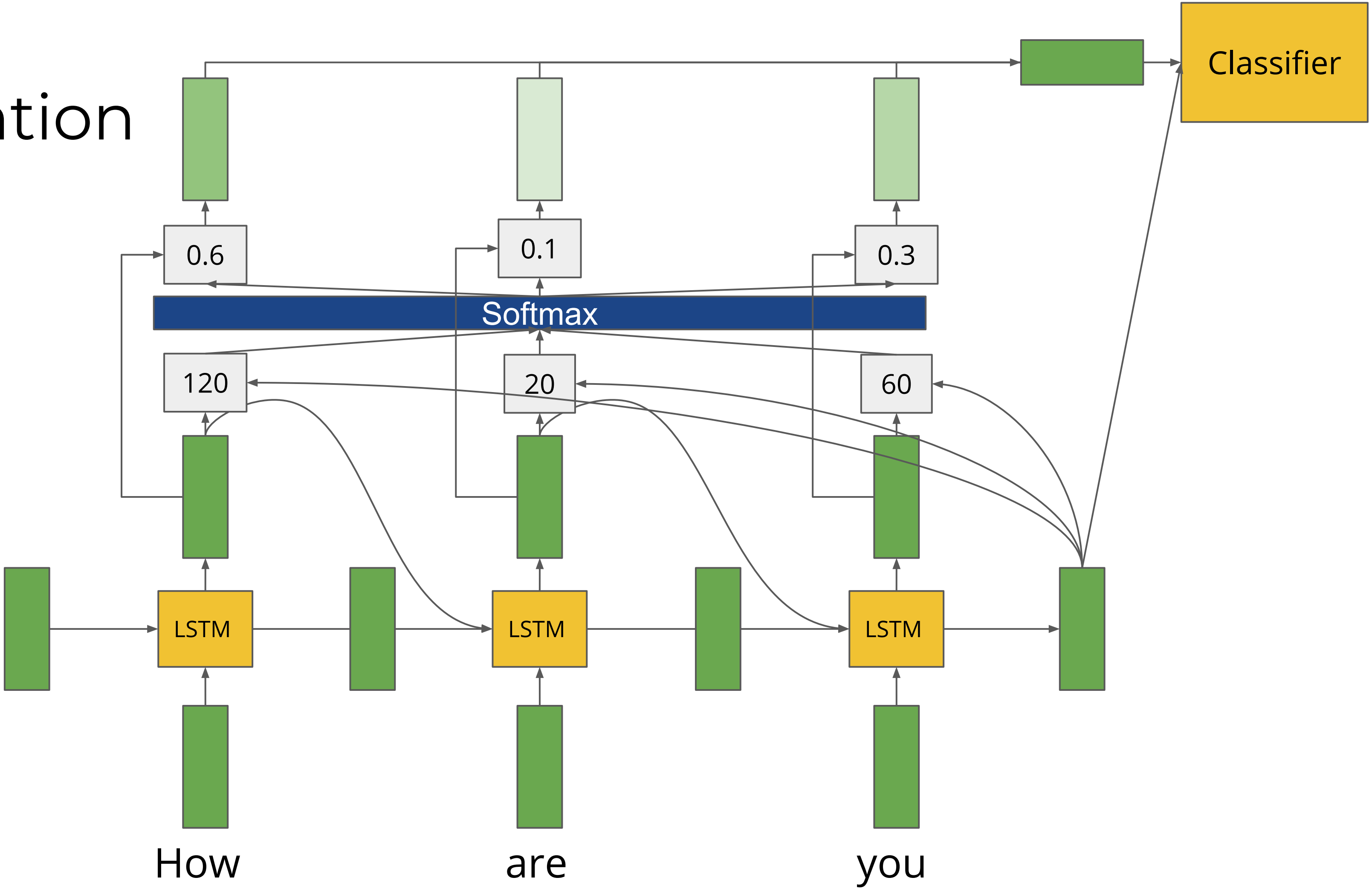
Sport is my favorite topic

Recurrence

Sport is my favorite topic → Sport

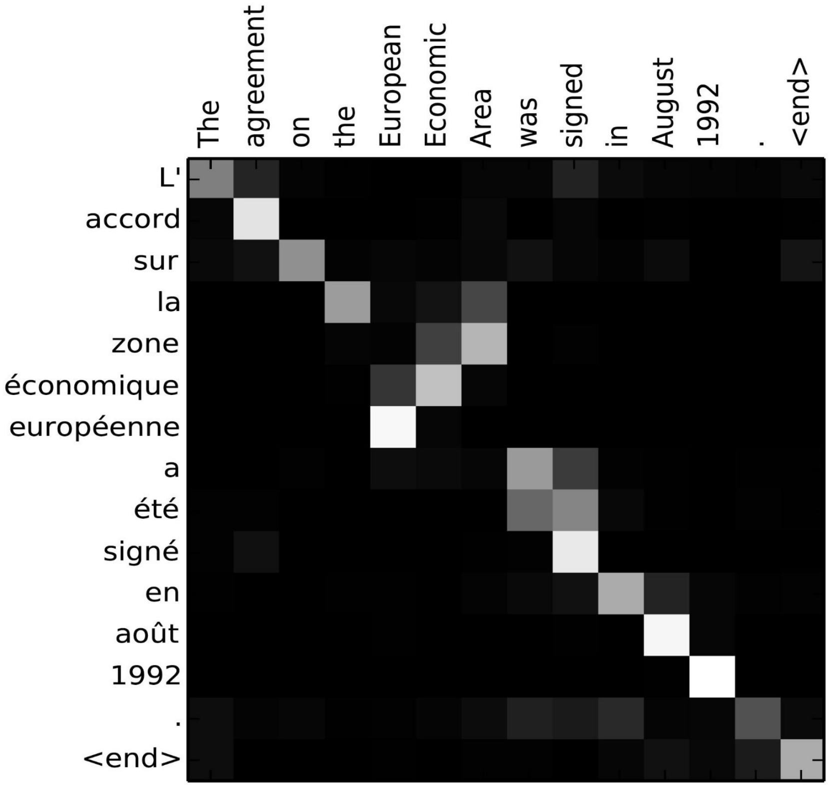
Is there any solution?

Attention



Classifier
0.6
0.1
0.3
Softmax
120
20
60
LSTM
LSTM
LSTM
How
are
you

Attention


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>

We know basic now!

What is the problem?

What is the problem? ✅

What is the problem? ✅

How to represent text?

What is the problem? ✅

How to represent text? ✅

What is the problem? ✅

How to represent text? ✅

How to classify by LSTM?

What is the problem? ✅

How to represent text? ✅

How to classify by LSTM? ✅

What is the problem? ✅

How to represent text? ✅

How to classify by LSTM? ✅

What is attention?

What is the problem? ✅

How to represent text? ✅

How to classify by LSTM? ✅

What is attention? ✅

Let’s got to

Transformers!

What is it?

Encoder - Decoder model with multi-headed self attention and residual connections using positional encoding

Transformer

Why is it important?

“NLP's ImageNet moment has arrived”

Transformer

GPT-2 BERT TRANSFORMER XL

Transformer

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	9.2
	2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
+	3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7	90.2	99.2	87.3	89.7	47.8
	4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
	5	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
	6	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4	47.5
+	7	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	8	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
	9	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
	10	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
	11	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
	12	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
	13	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
	14	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
	15	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
	16	廖亿	RGLM-Base (Huawei Noah's Ark Lab)		81.3	56.9	94.2	90.7/87.7	89.7/89.1	72.2/89.4	86.1	85.4	92.1	78.5	65.1	40.0

Transformer

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	9.2
2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7	90.2	99.2	87.3	89.7	47.8
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
5	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
6	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4	47.5
7	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
8	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
9	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
10	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
11	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
12	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
13	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
14	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
15	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
16	廖亿	RGLM-Base (Huawei Noah's Ark Lab)		81.3	56.9	94.2	90.7/87.7	89.7/89.1	72.2/89.4	86.1	85.4	92.1	78.5	65.1	40.0

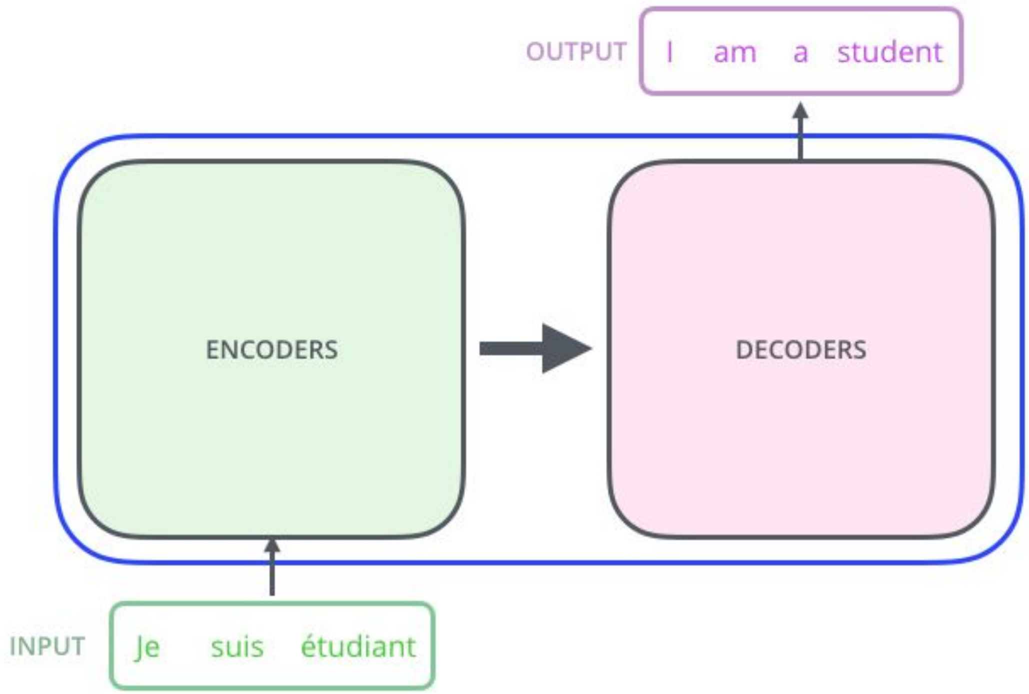
Stacked Encoder - Decoder model

with multi-headed self attention

and residual connections using

positional encoding

Transformer



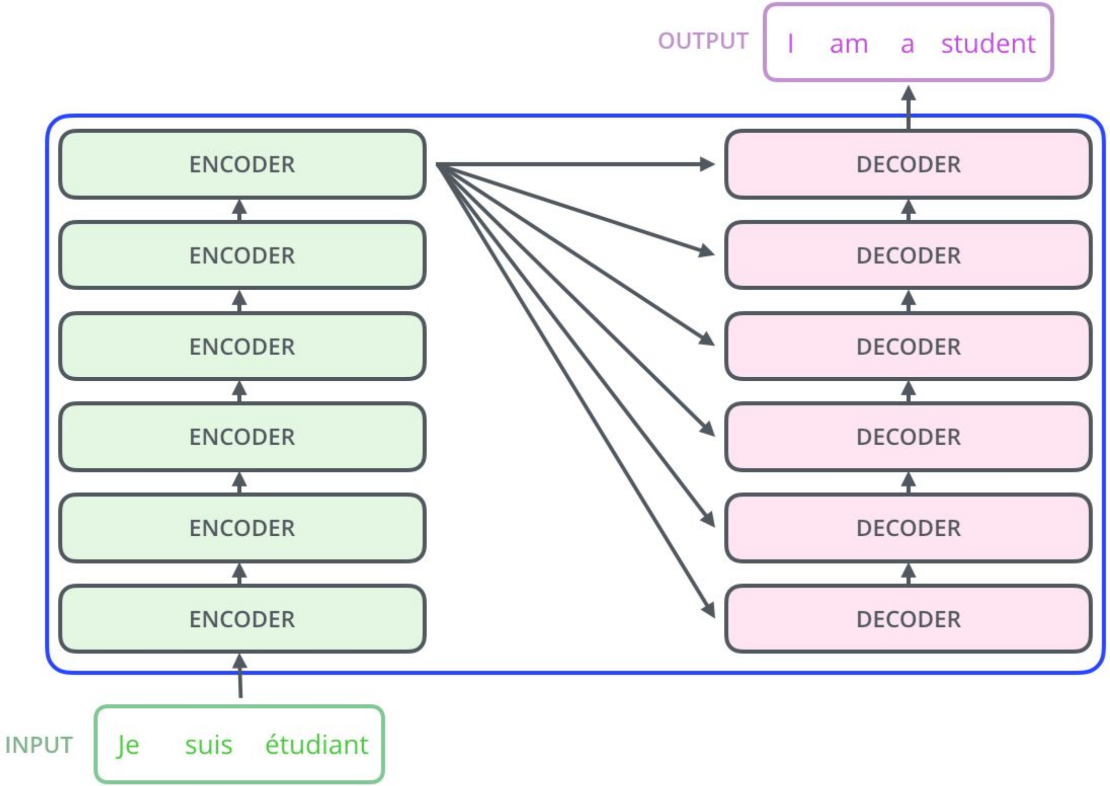
Stacked Encoder - Decoder model

with multi-headed self attention

and residual connections using

positional encoding

Transformer



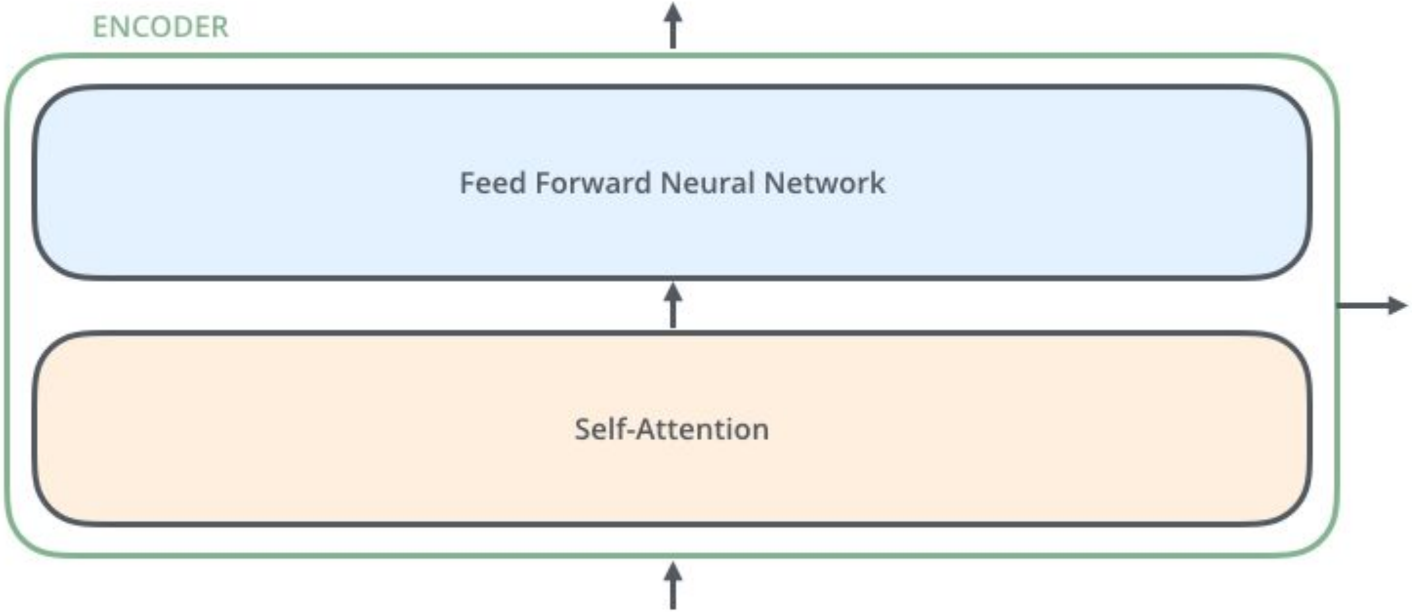
Stacked **Encoder** - Decoder model

with multi-headed self attention

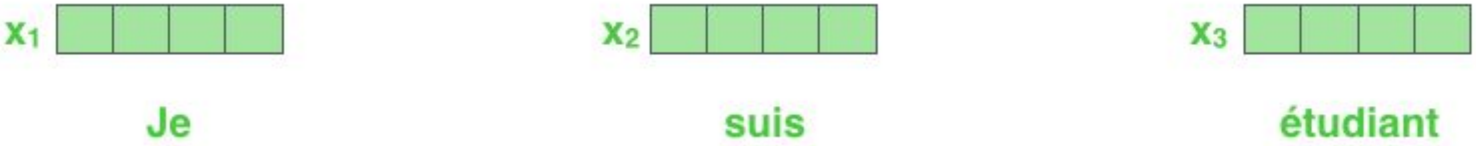
and residual connections using

positional encoding

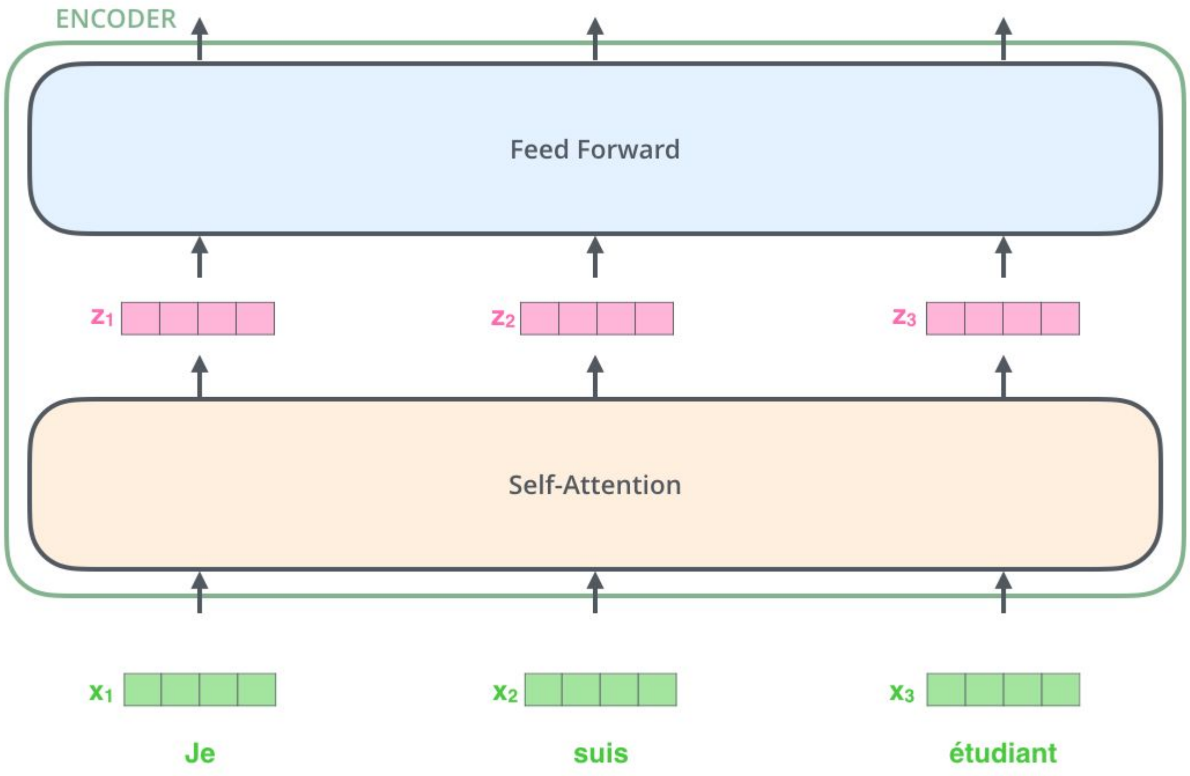
Transformer



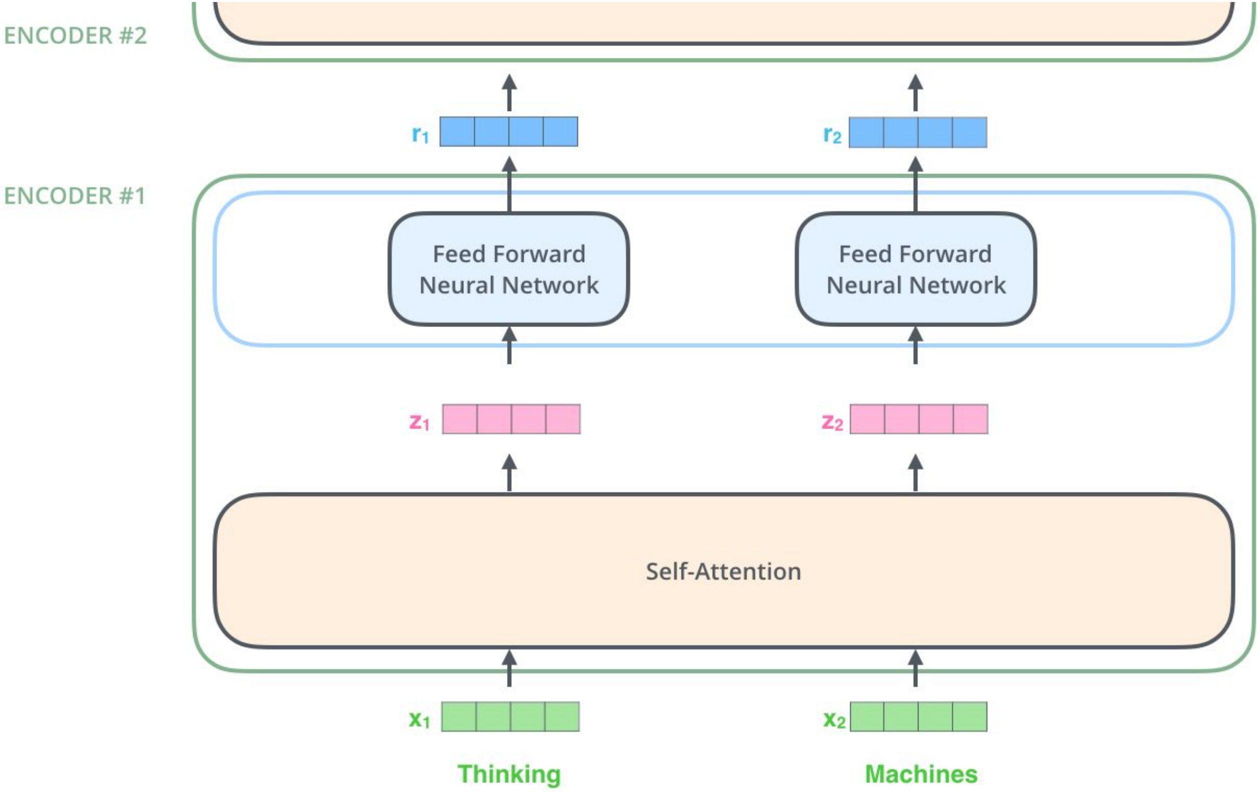
Transformer



Transformer



Transformer



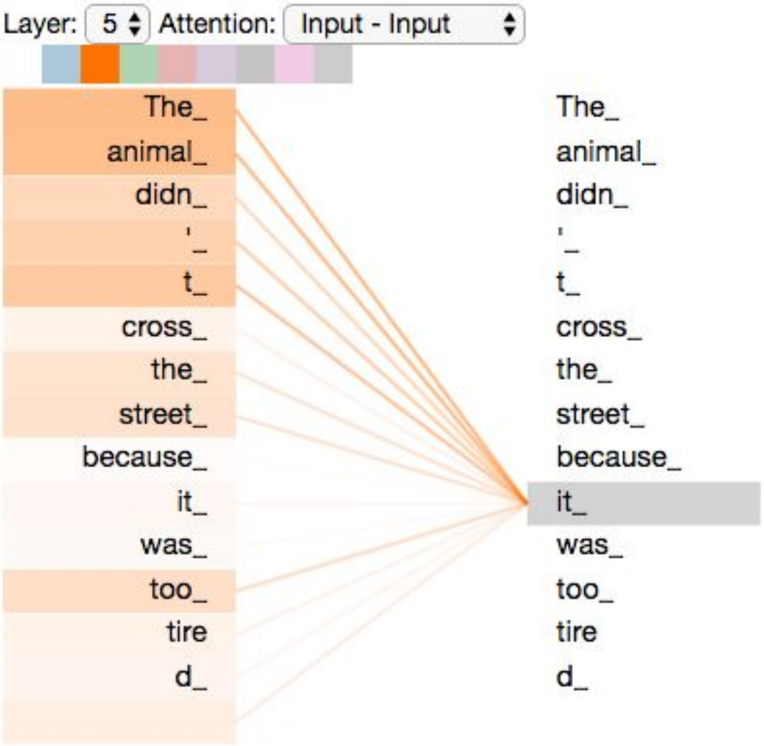
Stacked Encoder - Decoder model

with multi-headed self attention

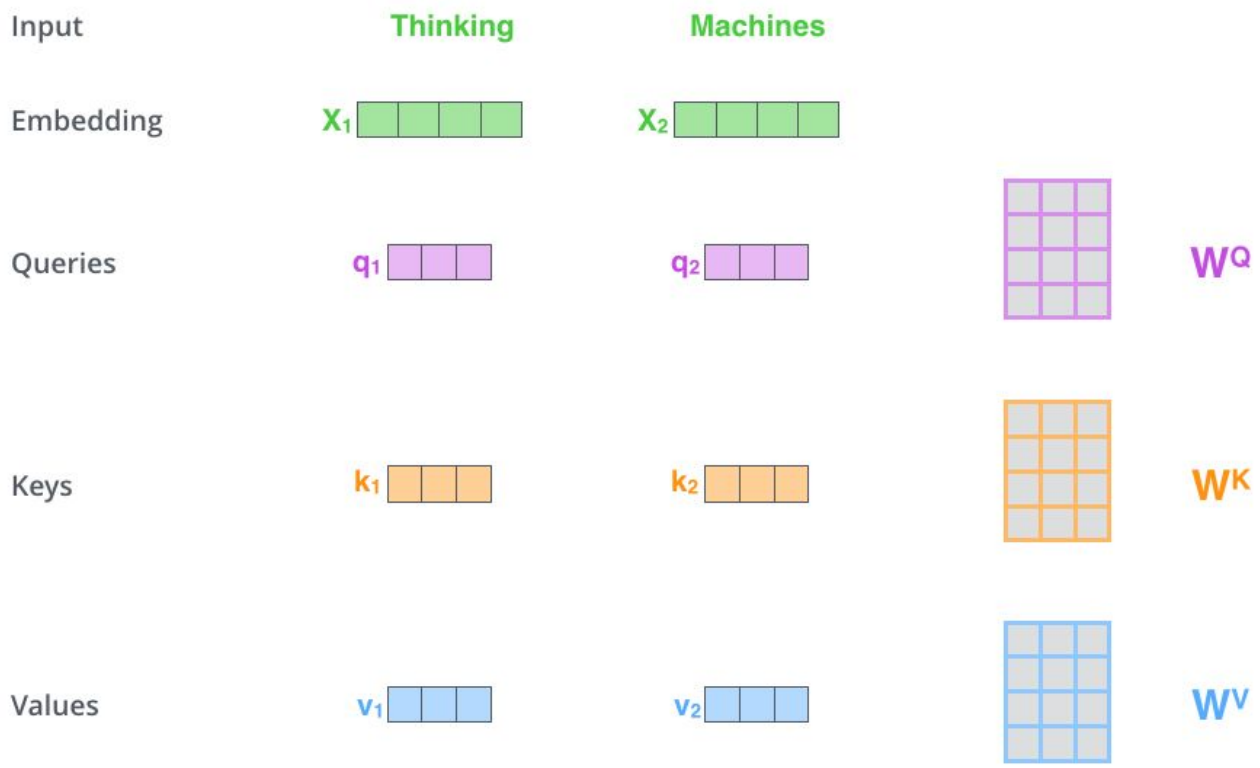
and residual connections using

positional encoding

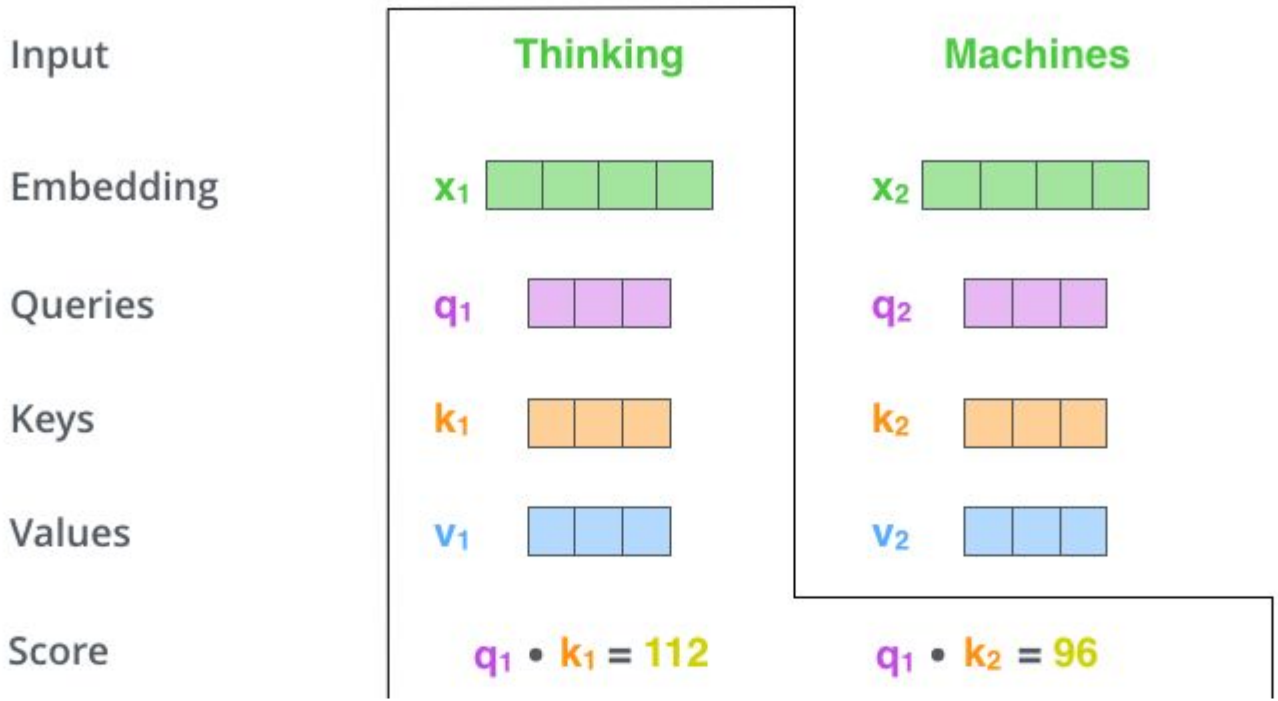
Transformer



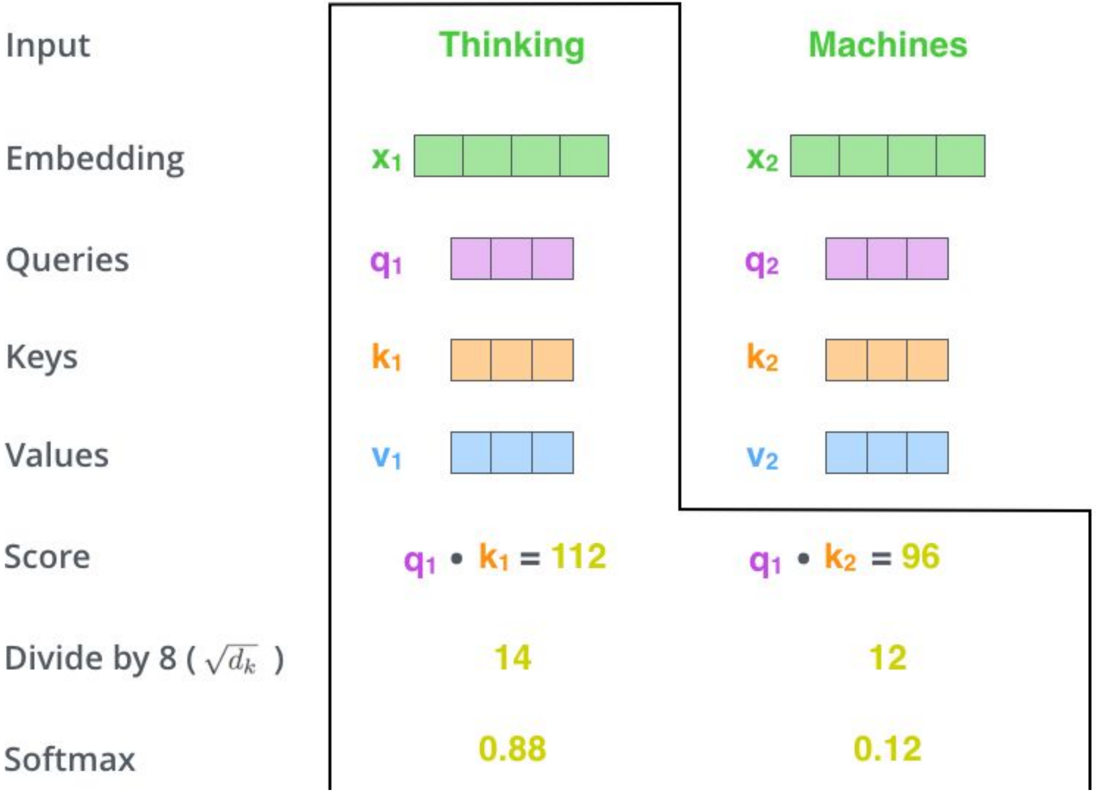
Transformer



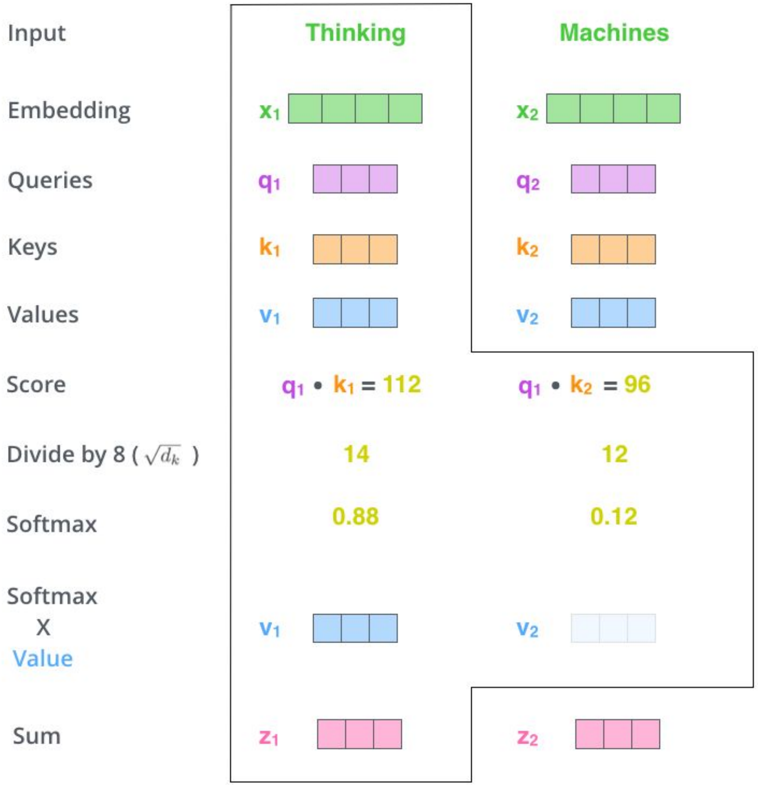
Transformer



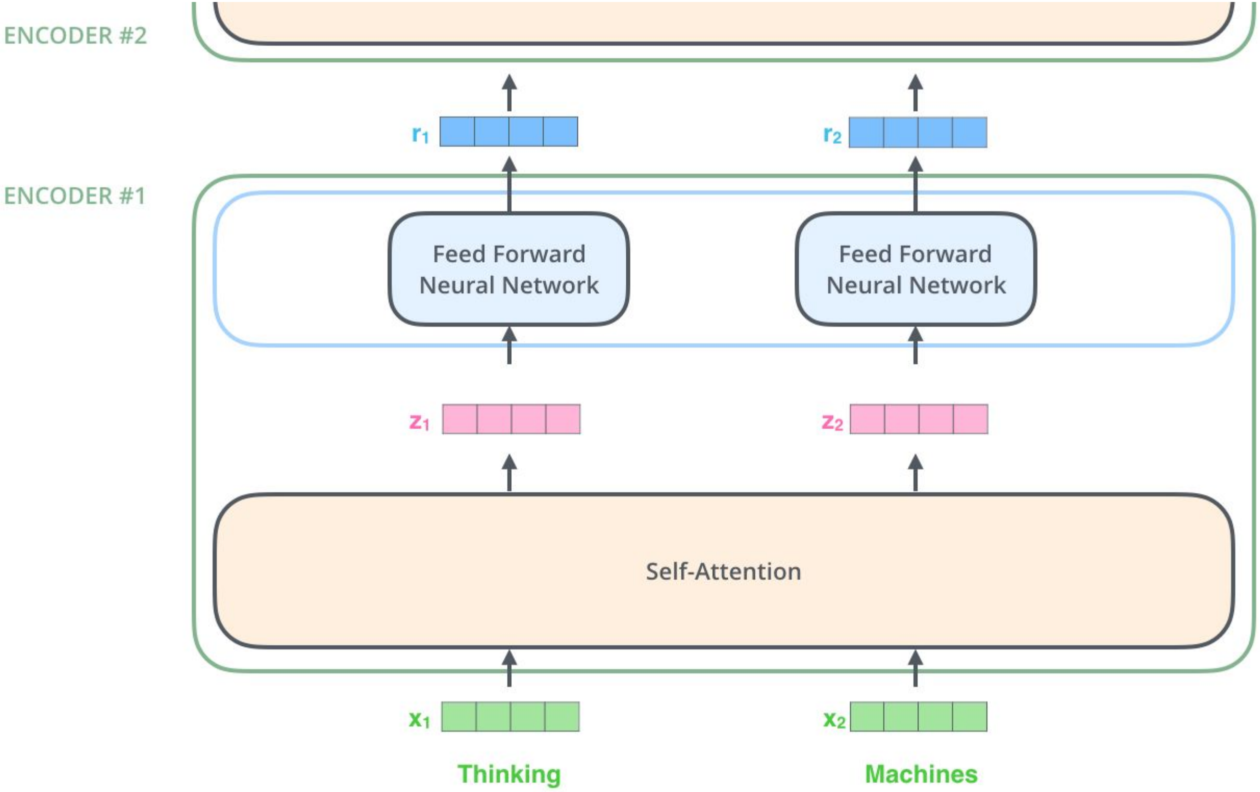
Transformer



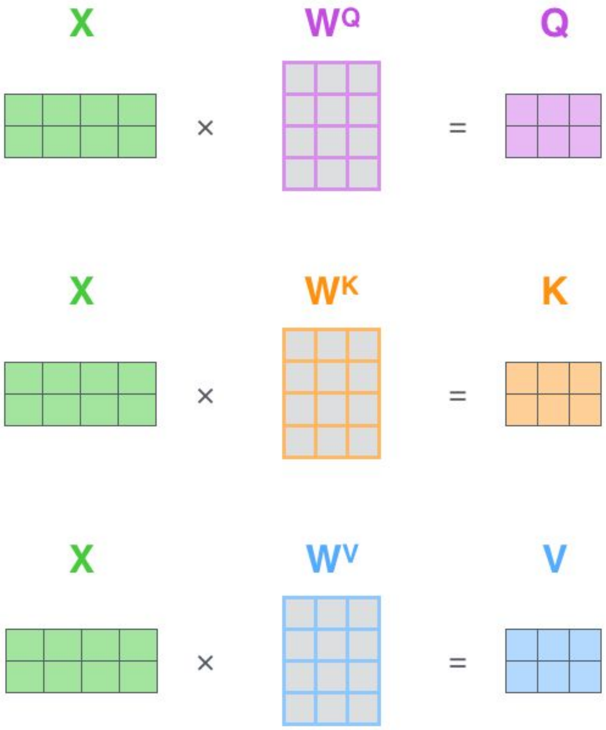
Transformer



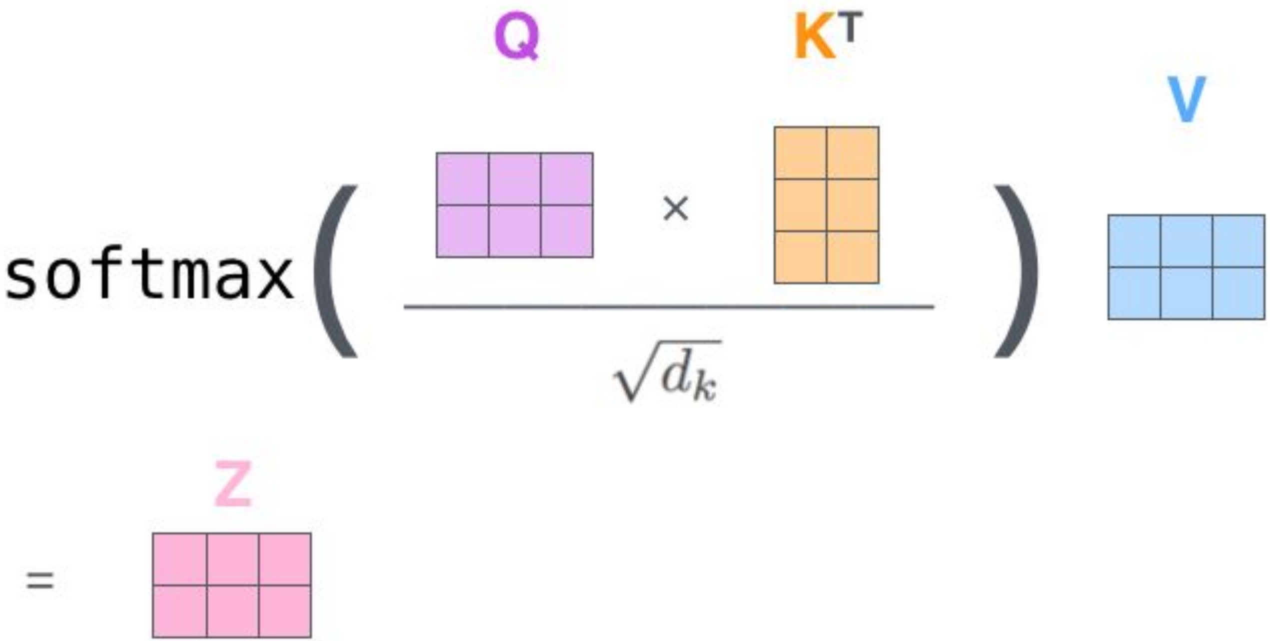
Transformer



Transformer

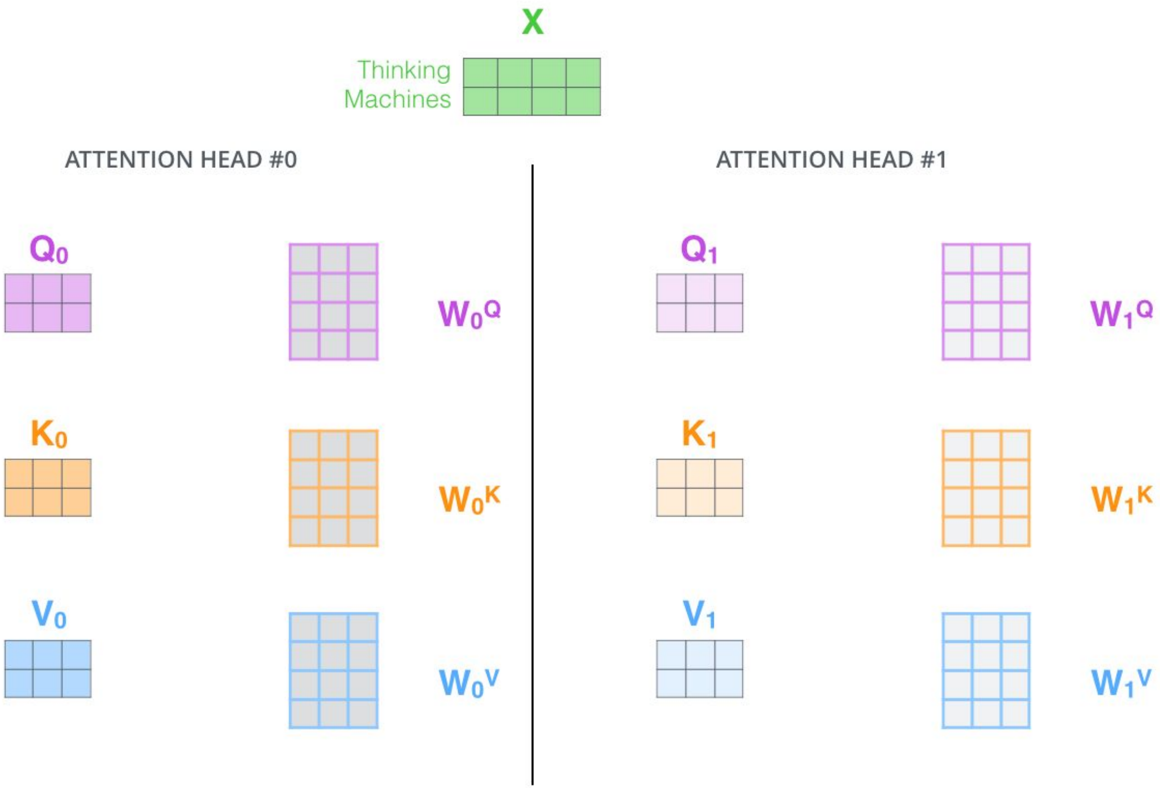


Transformer

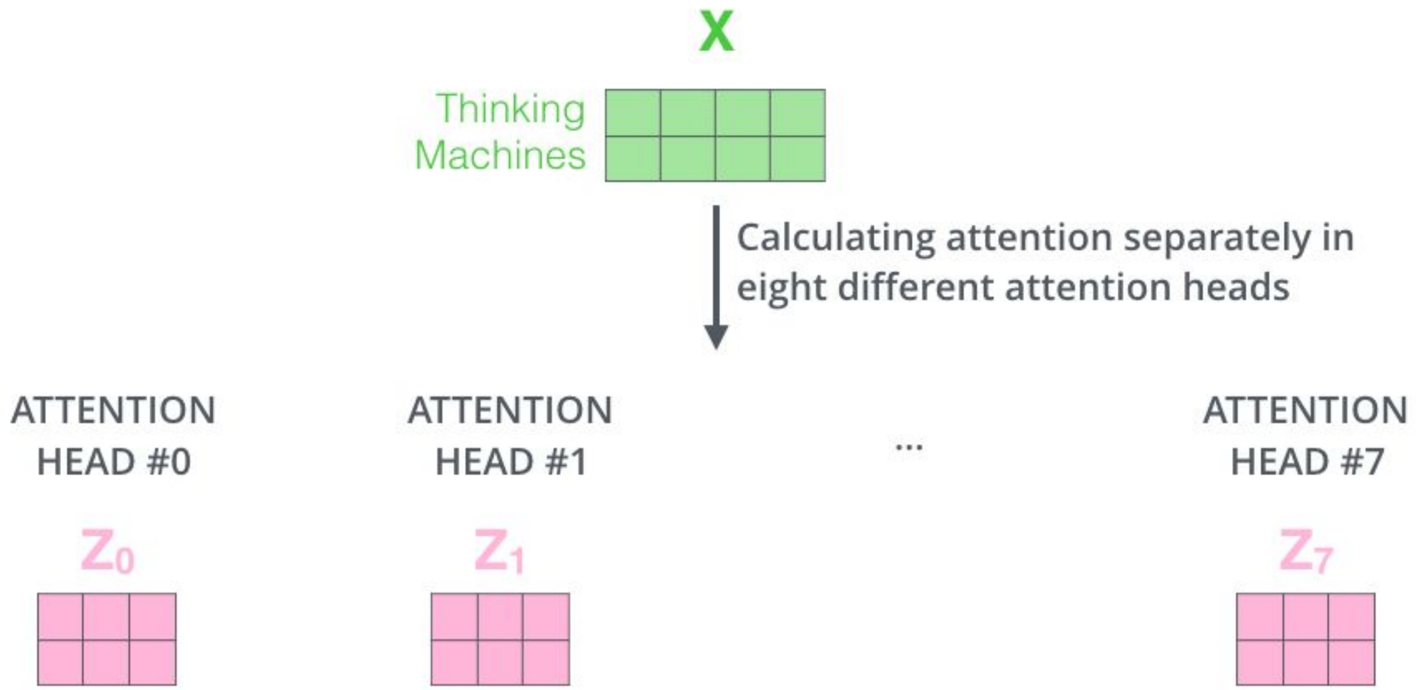


Stacked Encoder - Decoder model with **multi-headed** self attention and residual connections using positional encoding

Transformer



Transformer

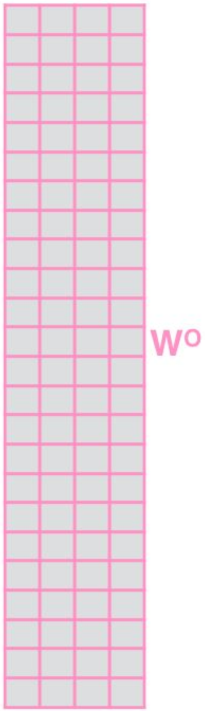


Transformer

1) Concatenate all the attention heads

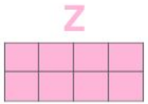
2) Multiply with a weight matrix W^O that was trained jointly with the model

X

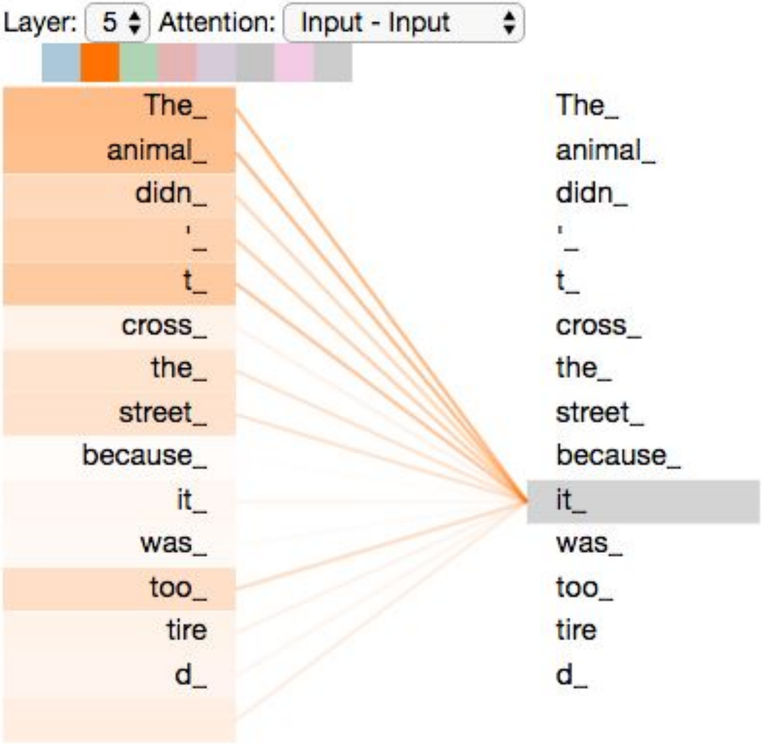


3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN

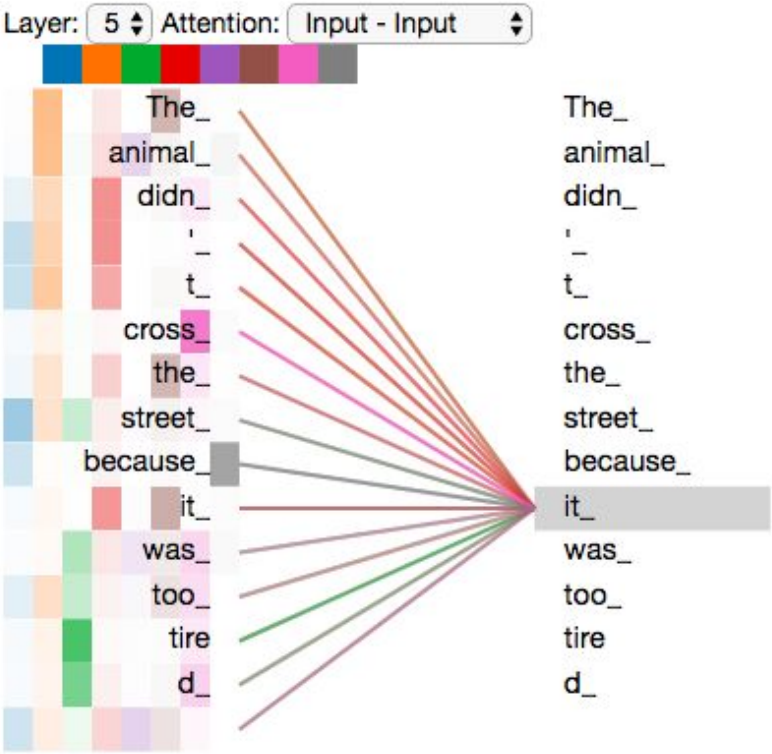
=



Transformer



Transformer



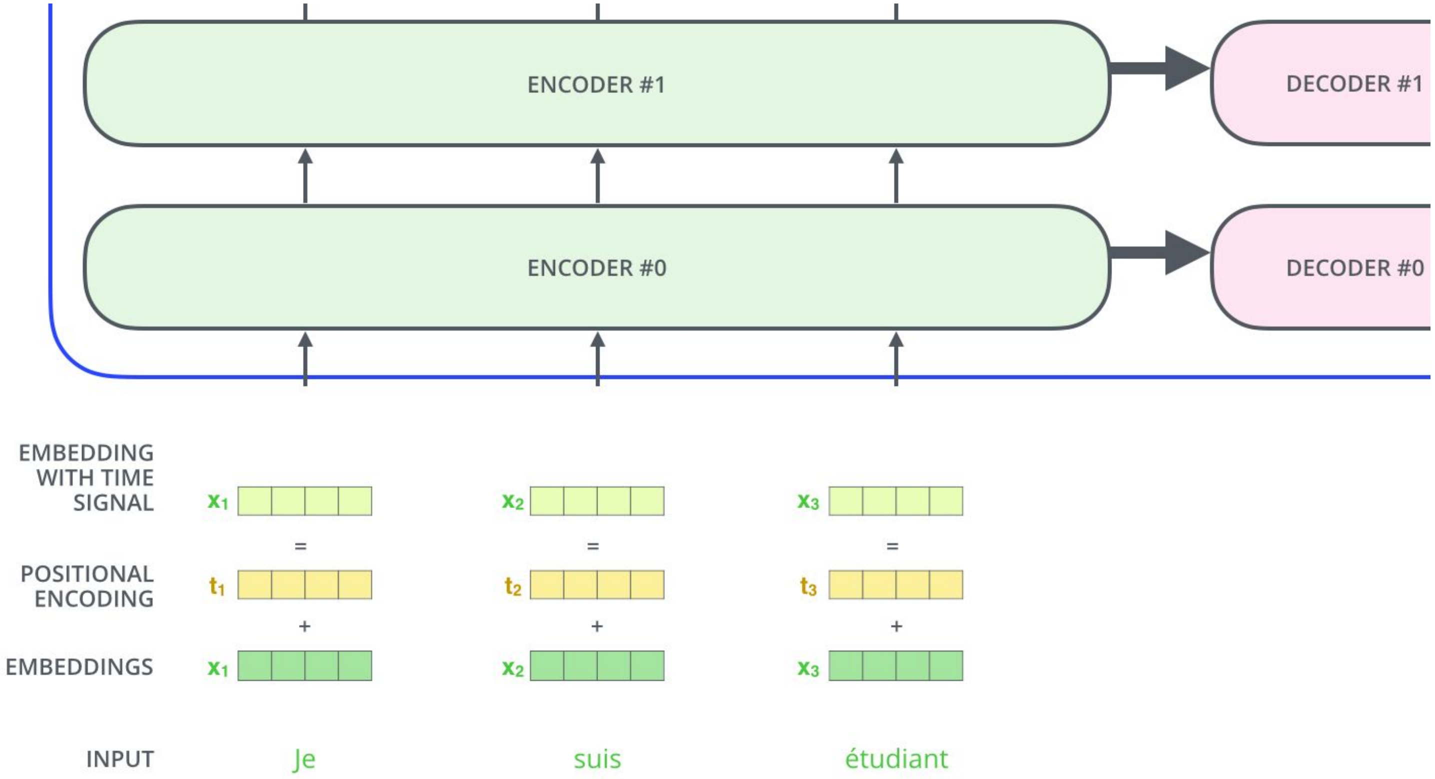
Stacked Encoder - Decoder model

with multi-headed self attention

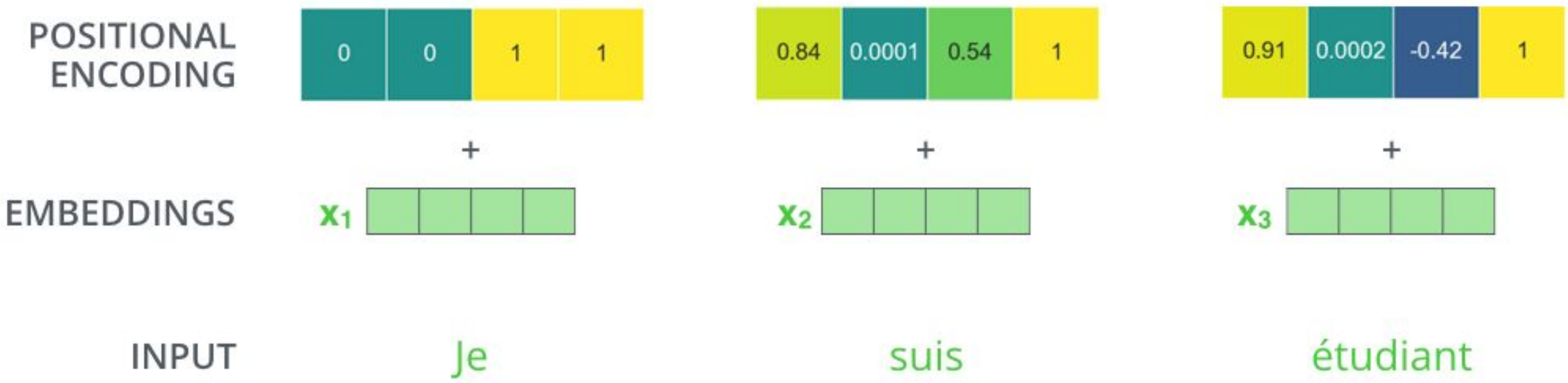
and residual connections using

positional encoding

Transformer



Transformer

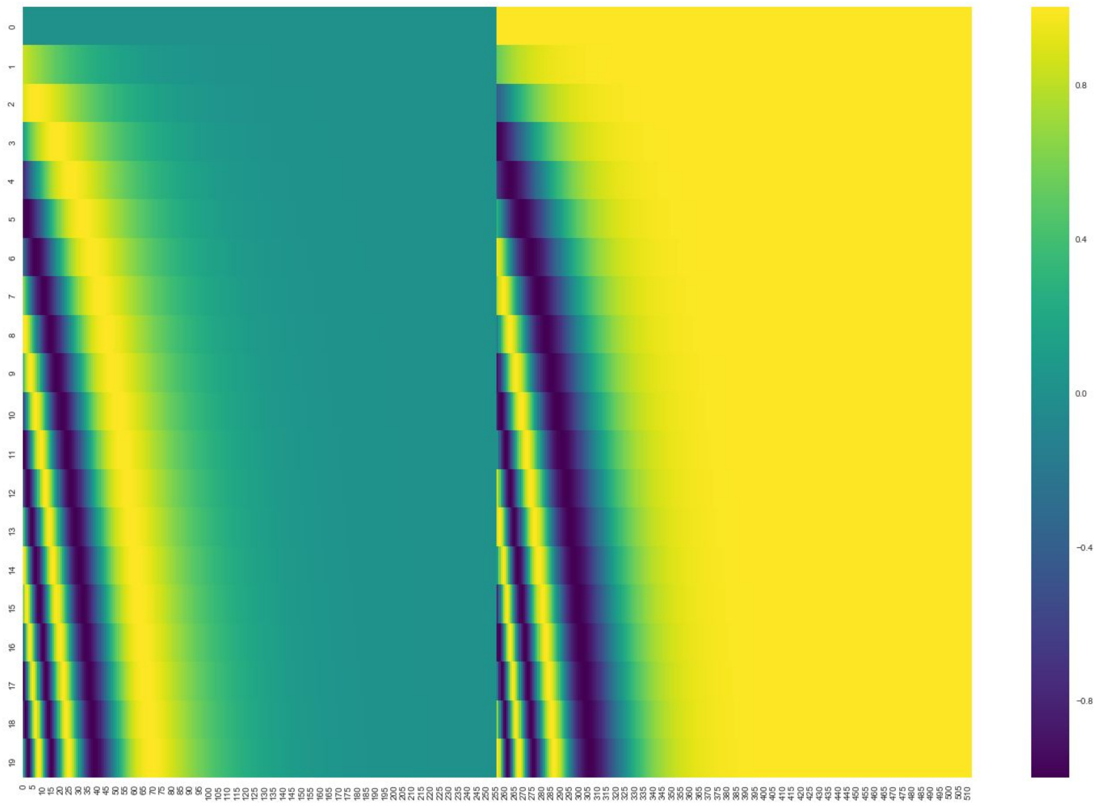


Transformer

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

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

Transformer



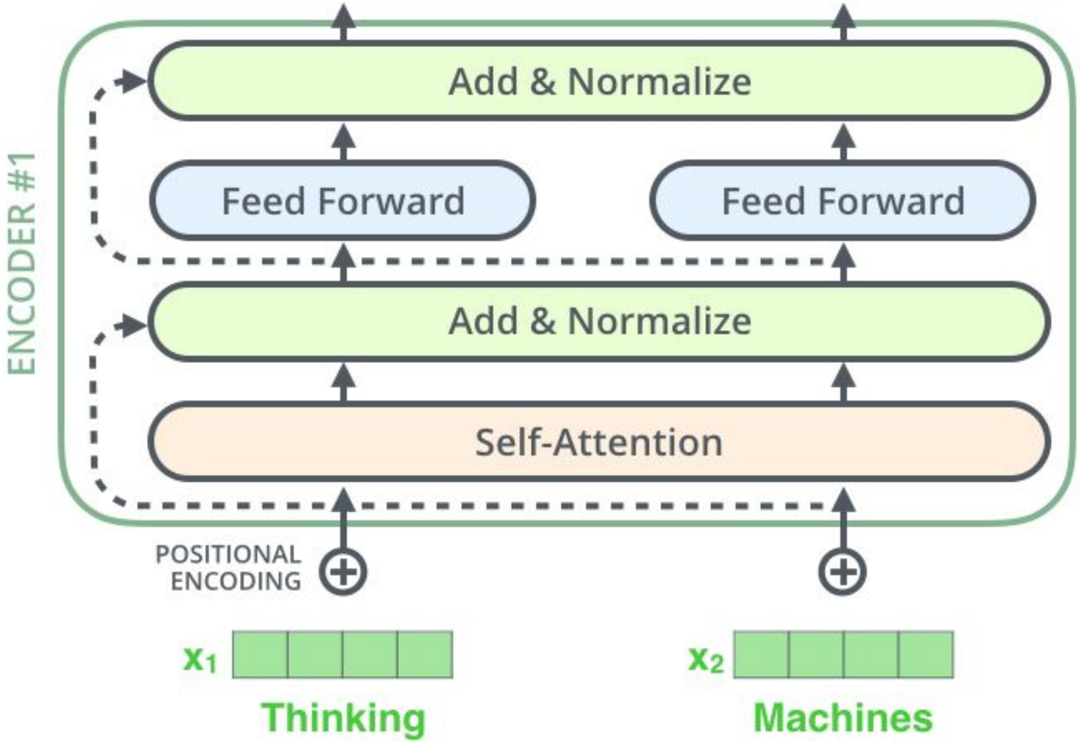
Stacked Encoder - Decoder model

with multi-headed self attention

and residual connections using

positional encoding

Transformer



Stacked Encoder - Decoder model

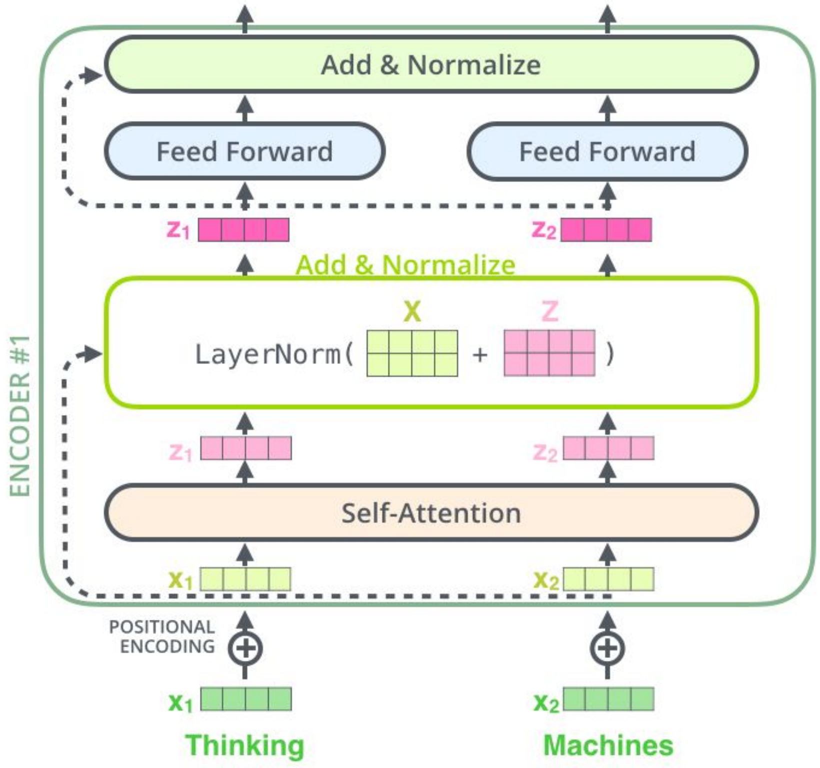
with multi-headed self attention

and residual connections using

positional encoding

and layer normalization

Transformer



References

The Illustrated Transformer, Jay Alammar, http://jalammar.github.io/illustrated-transformer/

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5 (2017): 135-146.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Gupta, Prakhar, Matteo Pagliardini, and Martin Jaggi. "Better Word Embeddings by Disentangling Contextual n-Gram Information." *arXiv preprint arXiv:1904.05033* (2019).

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

Wang, Alex, et al. "Glue: A multi-task benchmark and analysis platform for natural language understanding." *arXiv preprint arXiv:1804.07461* (2018).

NLP's ImageNet moment has arrived, Sebastian Ruder, https://thegradient.pub/nlp-imagenet/

Text Classification

Petr Marek