



University of Ljubljana
Faculty of Computer and
Information Science

Introduction to state of the art NLP

Andrej Miščič, Luka Vranješ



Agenda

00 Theoretical introduction (~45 min):

- motivation for Transformers
- Transformer architecture
- language representation with GPT and BERT

01 Practical part:

- Named Entity Recognition with BERT
- Sentiment analysis with BERT
- Text generation with GPT-2

02 Hands-on part:

- using obtained knowledge on new datasets

Motivation

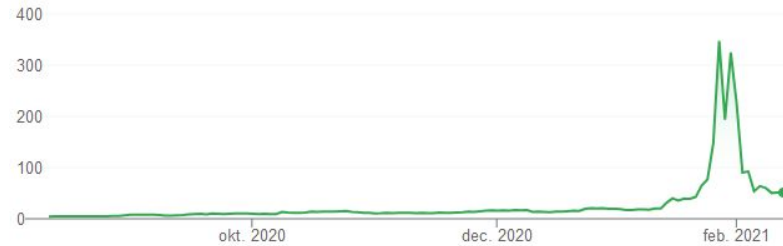
- how to deal with sequential data?

image caption generation



"man in black shirt is playing guitar."

time series forecasting



text classification

"I love this movie.
I've seen it many times
and it's still awesome."

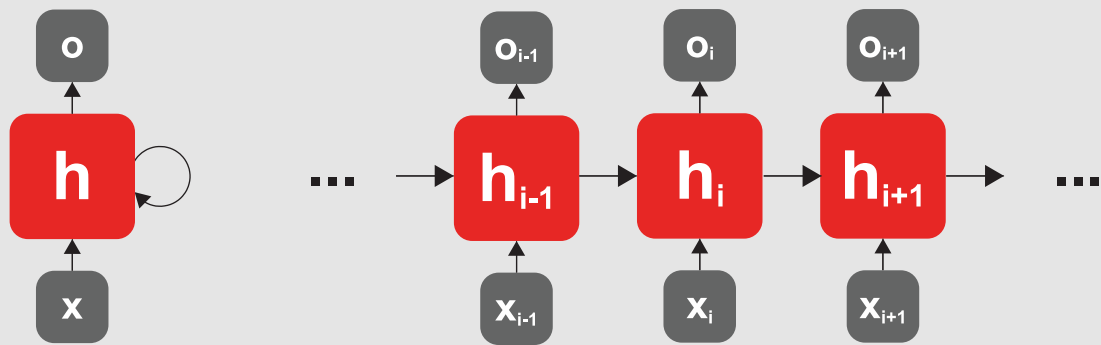


"This movie is bad.
I don't like it at all.
It's terrible."



RNNs

- extension of NNs for sequential data;
- information persists in hidden state h_i .



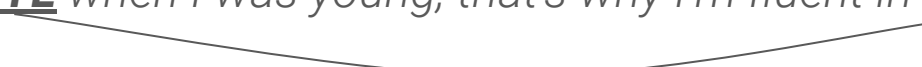
$$h_i = \begin{cases} 0, & \text{if } i = 0 \\ \sigma(W_{in}x_i + W_{hid}h_{i-1}), & \text{otherwise} \end{cases}$$
$$o_i = \sigma(W_{out}h_i)$$

Problems with RNNs

- difficult to train to capture long-term dependencies [1]:

I watched Spongebob on RTL when I was young, that's why I'm fluent in German.

French.
English.



Solution - Gating:

- LSTM [2] - controls information flow with gates
- mitigates vanishing gradients and somewhat better deals with long-term dependencies

[1] [Bengio et al.: Learning long-term dependencies with gradient descent is difficult, 1994](#)

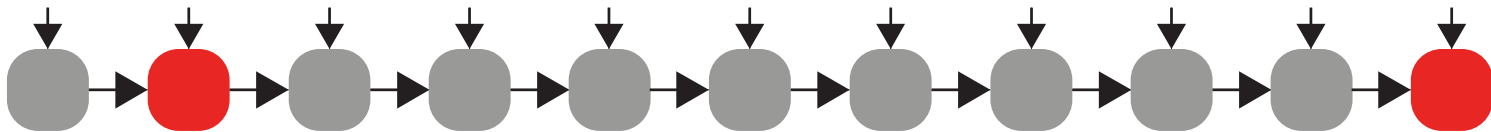
[2] [Hochreiter, Schmidhuber: Long Short-Term Memory, 1997](#)

Transformers



Motivation for Transformers

- RNNs are inherently sequential which prevents parallelization;
- the problem of long-term dependencies:
 - gating somewhat mitigates this problem, however, the path length between any two dependant words is still $O(n)$

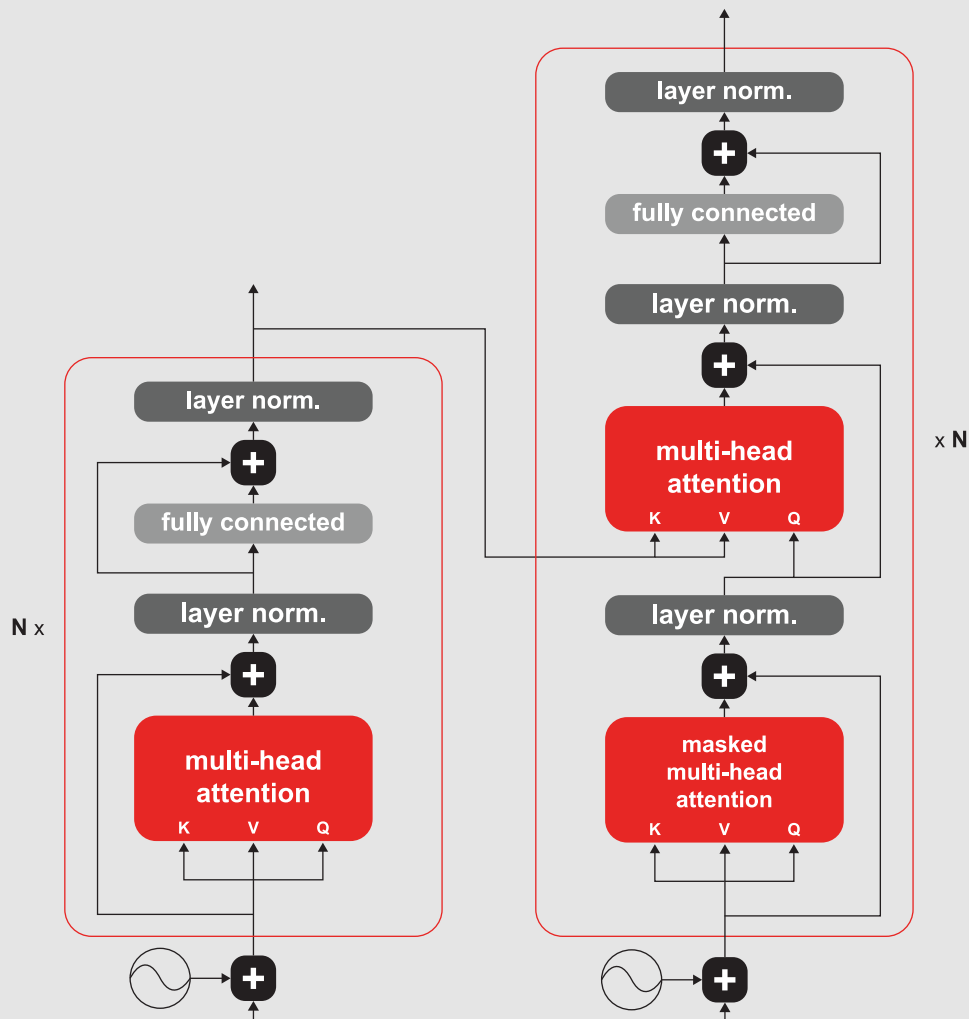


- Can we get rid of recurrence? What to replace it with?

Transformer^[1]

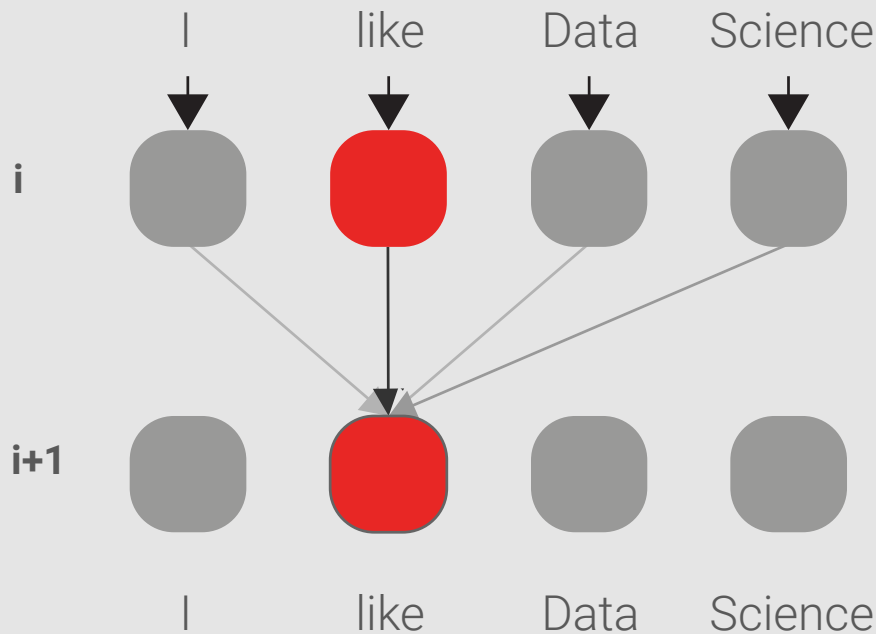
- introduced for Neural Machine Translation;

- uses **self-attention** in place of recurrence.

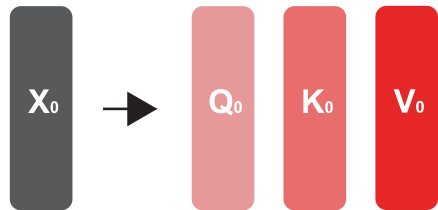


Self-attention

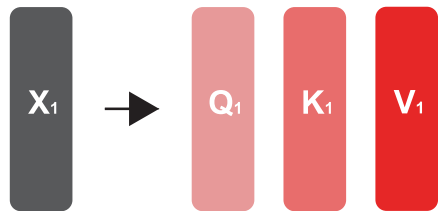
- it allows words to interact directly with other words
- degree of interaction/attention is based on similarity



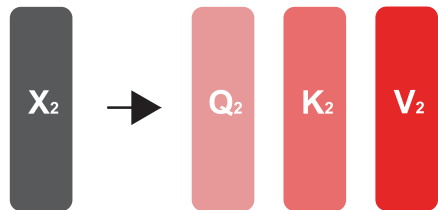
I



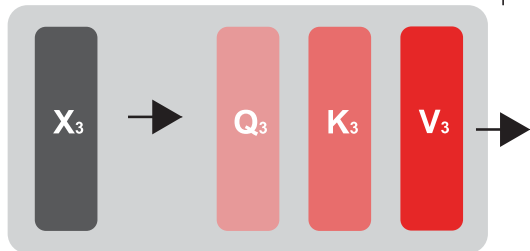
like



Data



Science



$$Q_3 \quad K_0 = 0.2$$

$$Q_3 \quad K_1 = 0.2$$

$$Q_3 \quad K_2 = 1.5$$

$$Q_3 \quad K_3 = 1.9$$

key-query similarity

~ 0.1

~ 0.1

~ 0.3

~ 0.5

softmax weights

$$\begin{aligned} & 0.1 * V_0 \\ & + \\ & 0.1 * V_1 \\ & + \\ & 0.3 * V_2 \\ & + \\ & 0.5 * V_3 \\ & = O_3 \end{aligned}$$

Self-attention

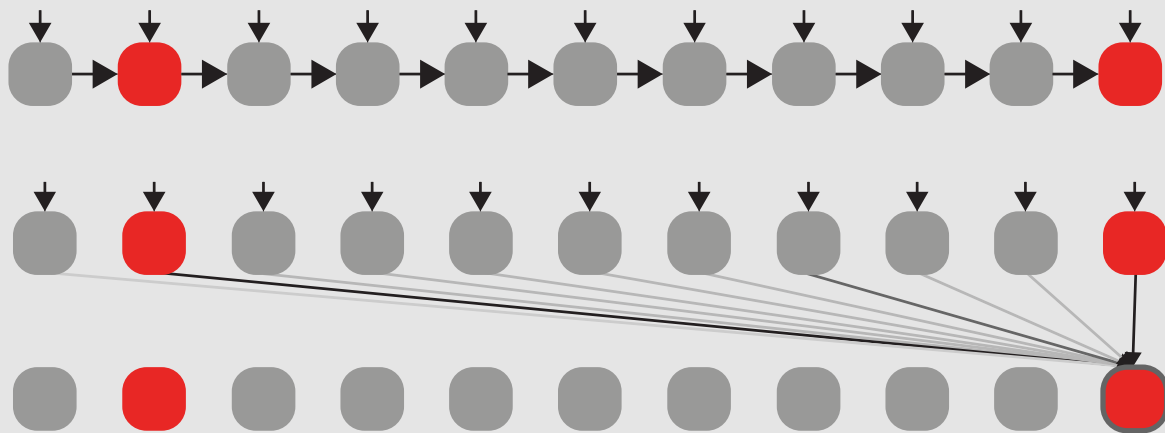
- *power of self-attention:*

- RNN: path length
between two words is
 $O(n)$;

- in self-attention the
path length is $O(1)$.

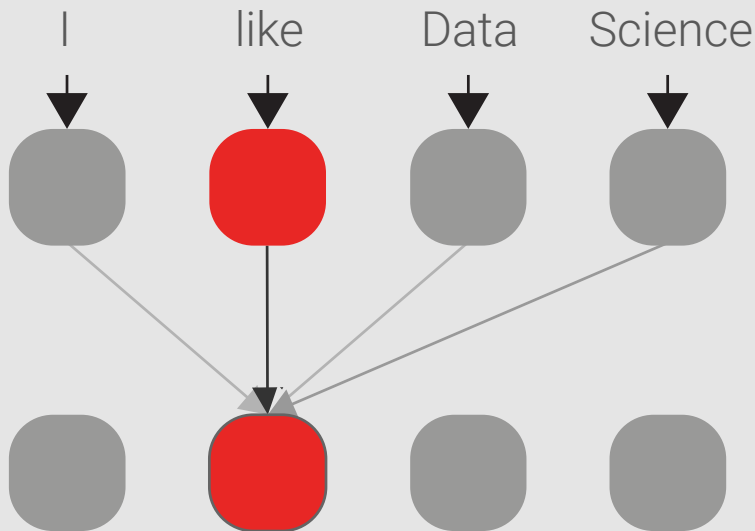
The animal didn't cross the **street** because it was too **wide**.

The **animal** didn't cross the street because it was too **tired**.



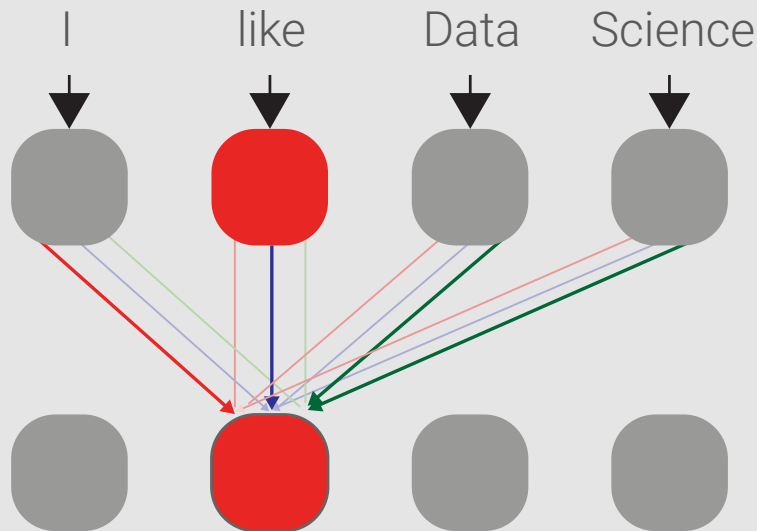
Problem

- single self-attention can be a bottleneck;
- cannot capture multiple interactions between words;
- in our example we want to know for word *like*:
 - who likes?
 - does what? (attend to itself)
 - likes what?



Solution

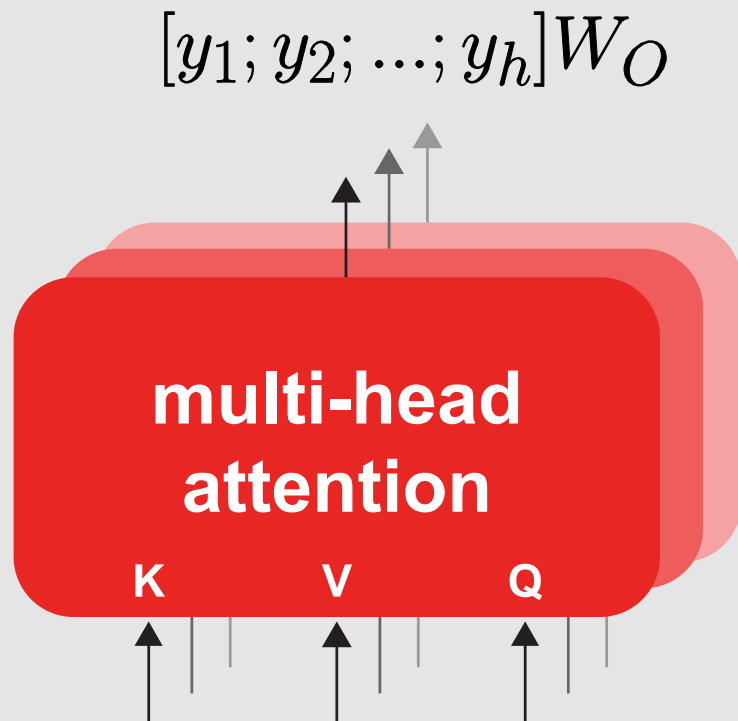
- multiple parallel copies of attention - Multi-Head attention;
- different attention heads can now pick up different interactions.



Multi-Head Attention

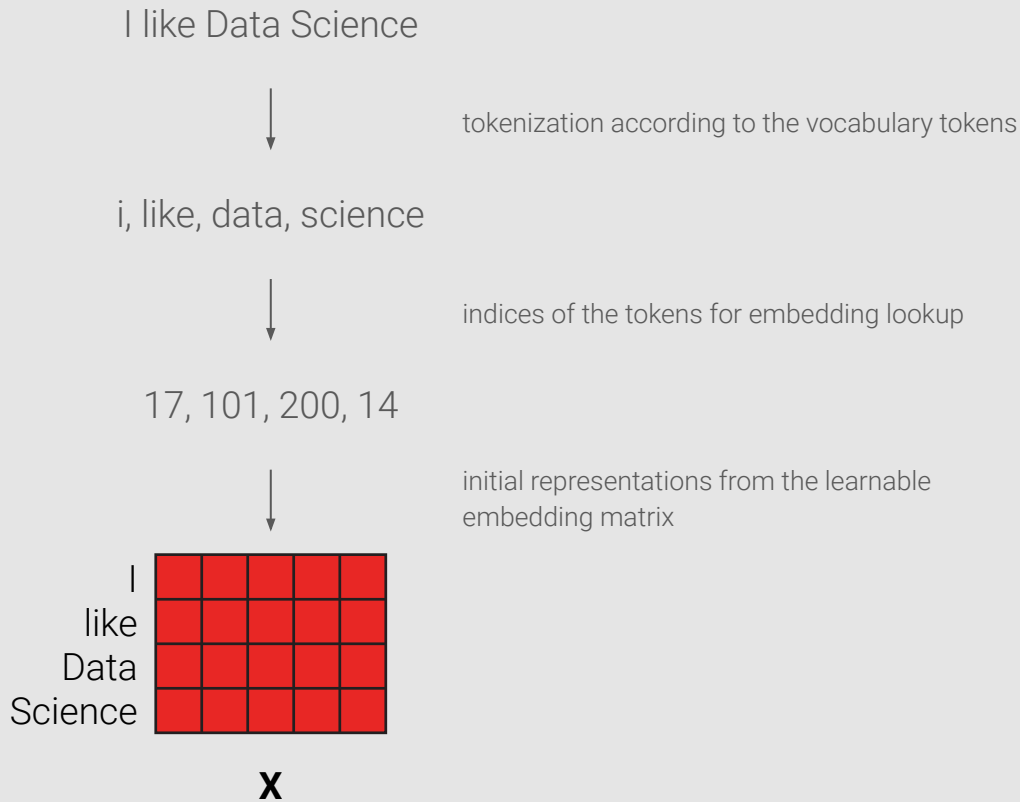
- each attention head has different projection matrices W_K , W_V , W_Q , therefore different K , V , Q

- y_i - output of head i



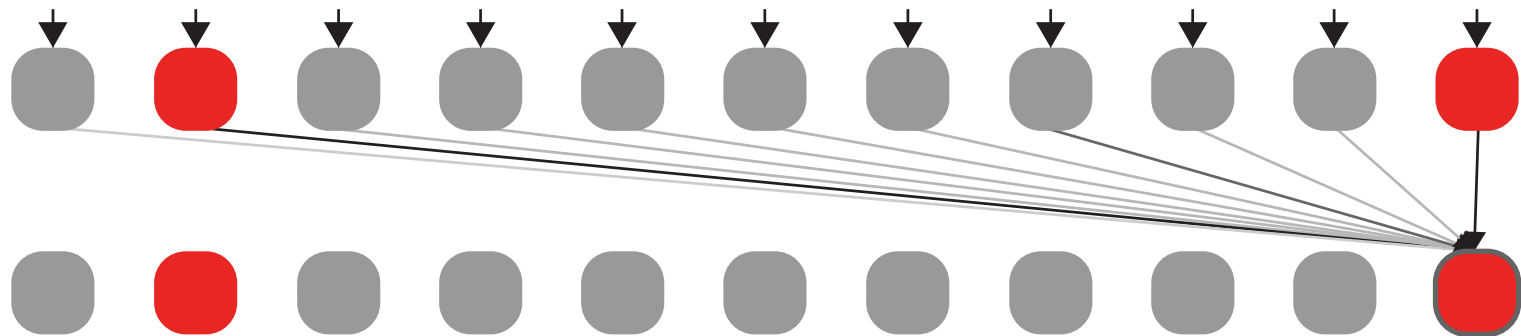
Inputs

- we need a predefined
fixed-sized vocabulary
(usually order of 10^4 tokens);



Problem

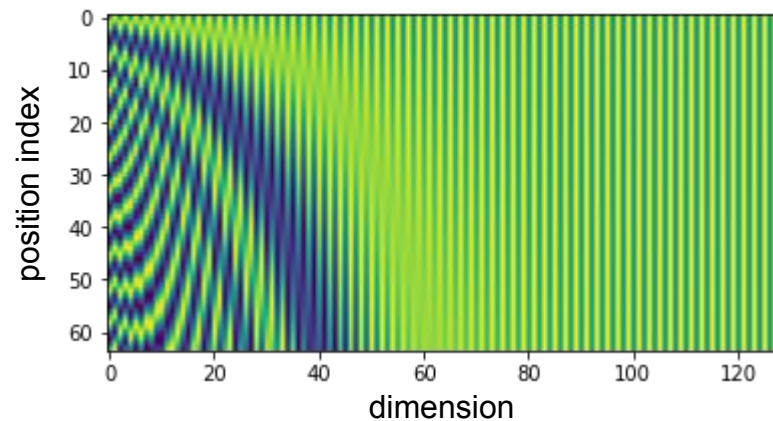
- by getting rid of recurrence we lose positional information which is important as our data is sequential;
- self-attention is permutation invariant, i.e. no matter the order of the inputs, the output will be the same.



Solution: Positional encodings

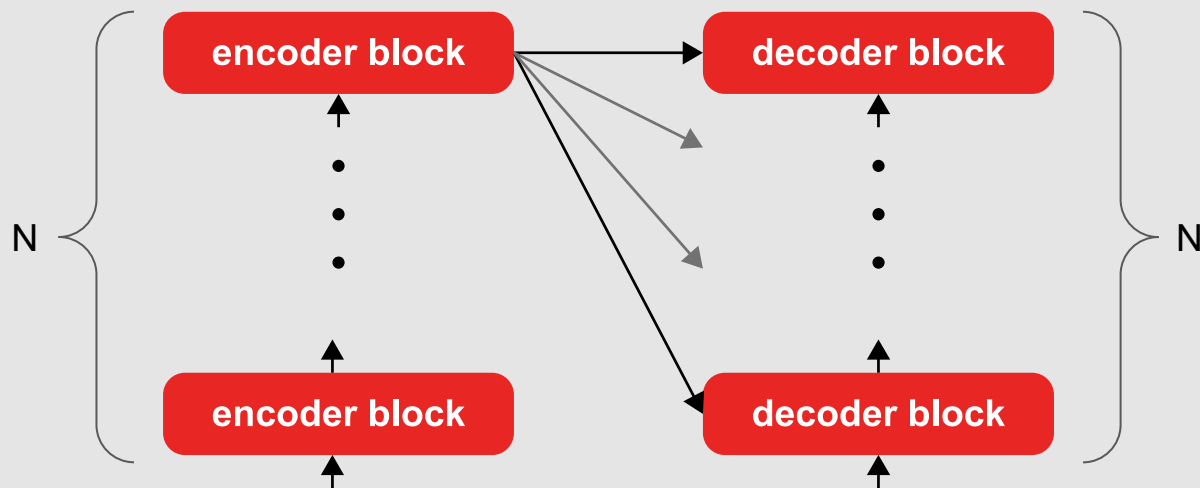
- positional encodings have the same dimension as input embeddings and are added to them before the first self-attention layer;
- they can be either:
 - LEARNED: use an embedding layer to learn a pos. embedding for each position in the sequence;
 - FIXED: set before training, used in original paper.

$$PE_{ij} = \begin{cases} \sin(i/10000^{\frac{j}{dm}}) & \text{if } j \text{ is even} \\ \cos(i/10000^{\frac{j-1}{dm}}) & \text{if } j \text{ is odd} \end{cases}$$

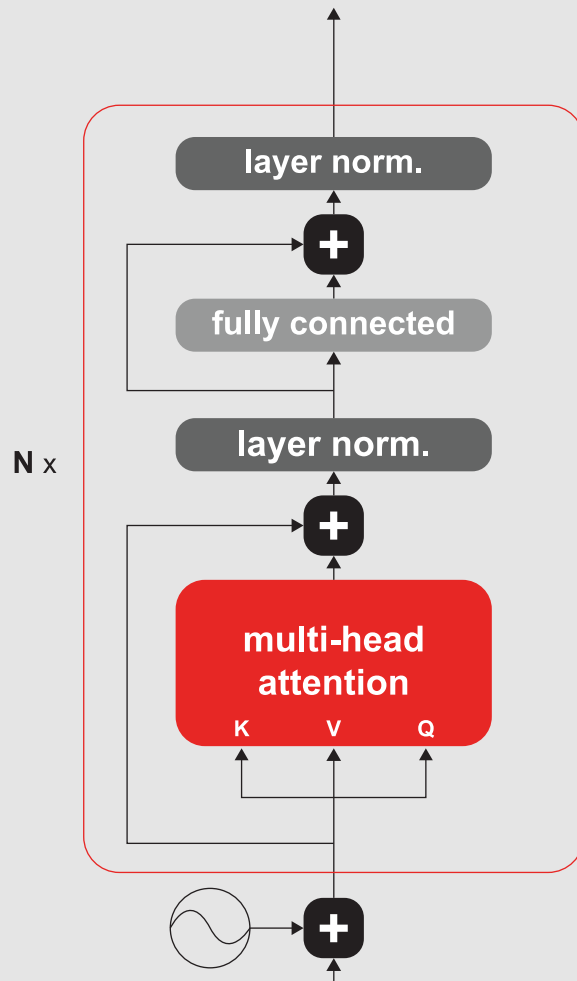


Architecture

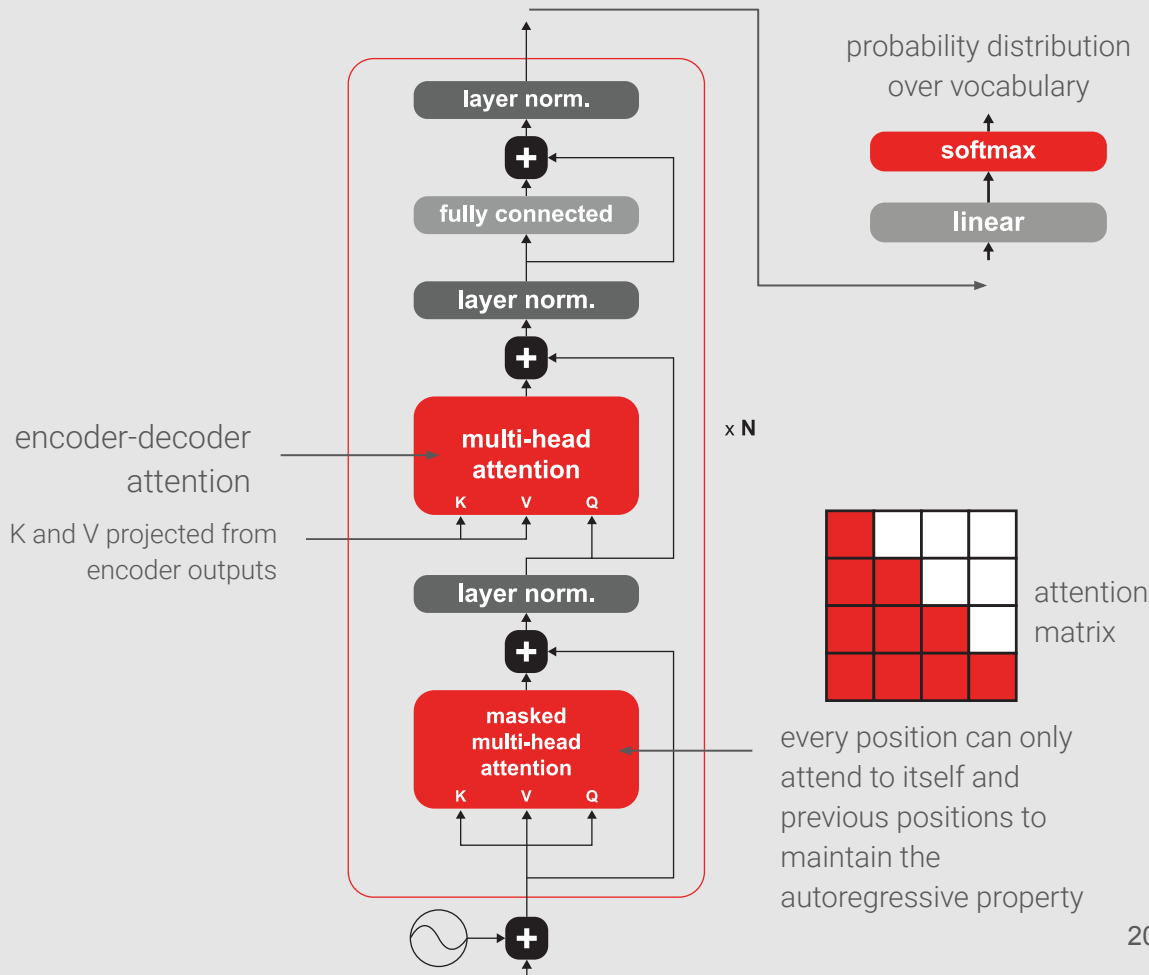
- N encoder and N decoder blocks;
- inputs to an encoder block are either network inputs or previous encoder block outputs;
- inputs to a decoder block are encoder outputs, previous decoder block outputs or previous outputs.




Transformer encoder



Transformer decoder



GPT + BERT



Pre-training word representations

- deep learning requires lots of annotated data, which can be scarce;
- on the other hand, we have abundant unlabeled text data;



- leverage this unlabeled data to pre-train word representations in an unsupervised manner and use these embeddings as building blocks of supervised networks.

Neural embeddings

- approaches such as Word2Vec [1], GloVe [2], etc.;
- trained on unlabeled corpus using co-occurrence;
- problem: used in a **context-free** manner.

context-free: using the same word representation regardless of the context and word sense

*We went to see a **play** at the local theater.*

*Children went out to **play** in the park.*

[1] [Mikolov et al.: Efficient Estimation of Word Representations in Vector Space, 2013.](#)

[2] [Pennington et al.: GloVe: Global Vectors for Word Representation, 2014.](#)

Solution

- contextual word representations;
- trained on an unlabeled corpus with some auxiliary task, e.g. language modelling;
- early approach: ELMo [1].

[1] [Peters et al.: Deep contextualized word representations, 2018.](#)

ELMO:

- bi-directional LSTM model
- pre-train for language modelling
- for downstream tasks include pre-trained representations as additional features in task specific architectures

Fine-tuning approaches

- can we develop models that adapt to many NLP tasks with little modification?
- training has two phases:
 - **Pre-training:** using a large unlabeled corpus and an auxiliary task, pre-train a model for general language representation;
 - **Fine-tuning:** using a (possibly smaller) labeled dataset, further train the pre-trained model for a specific downstream task.
- GPT [1], BERT [2]

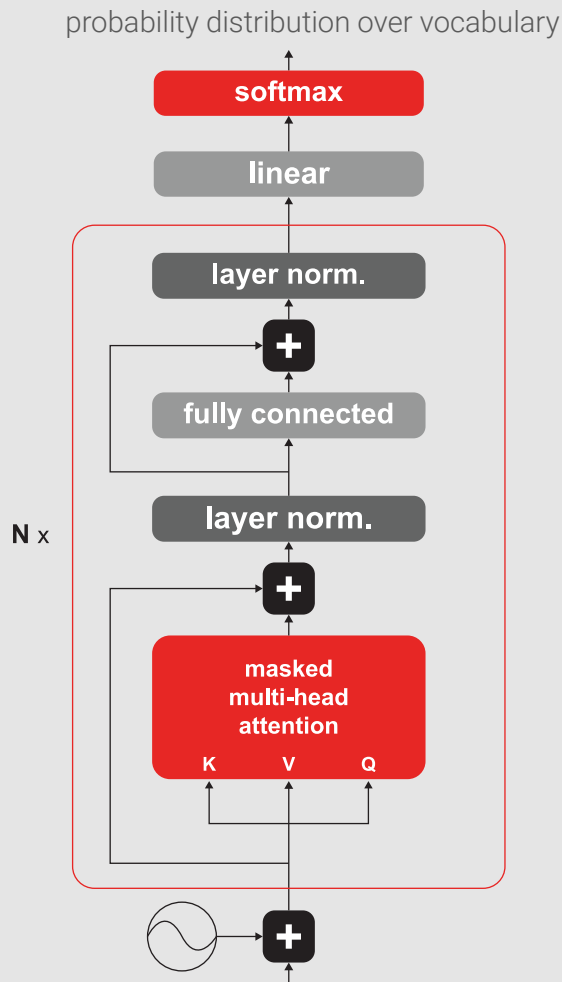
[1] [Radford et al.: Improving Language Understanding by Generative Pre-Training, 2018.](#)

[2] [Devlin et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.](#)

GPT [1]

- Generative Pre-trained Transformer;
- using only the decoder part of Transformer;
- pre-trained for language modelling, i.e. predicting next word given the context.

[1] [Radford et al.: Improving Language Understanding by Generative Pre-Training, 2018.](#)

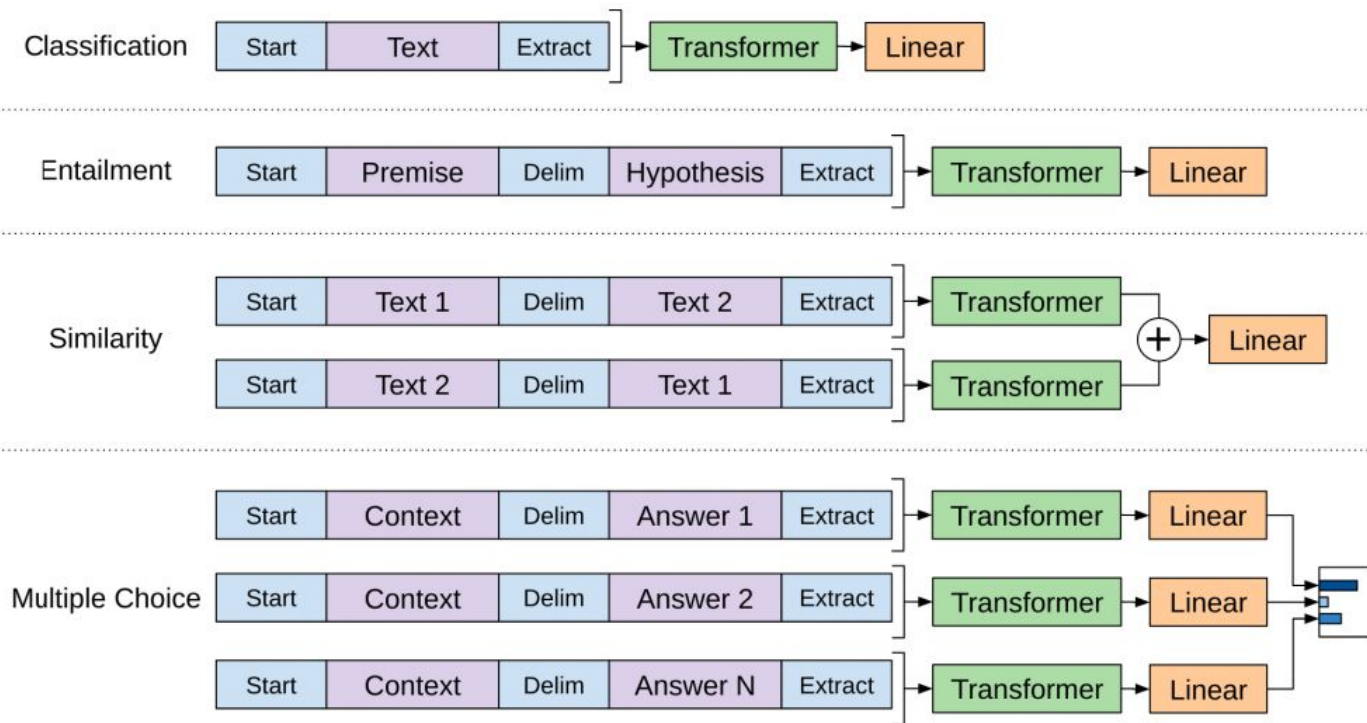


Fine-tuning

$$L(X) = L_{LM}(X) + \lambda L_D(X)$$

language
modelling
loss

downstream
task
loss



GPT shortcomings

- language modelling is an unidirectional task, models predict the next word given the left context:

*'What are those?' he said while looking at my **[?]***

- better language understanding requires incorporating bidirectionality:

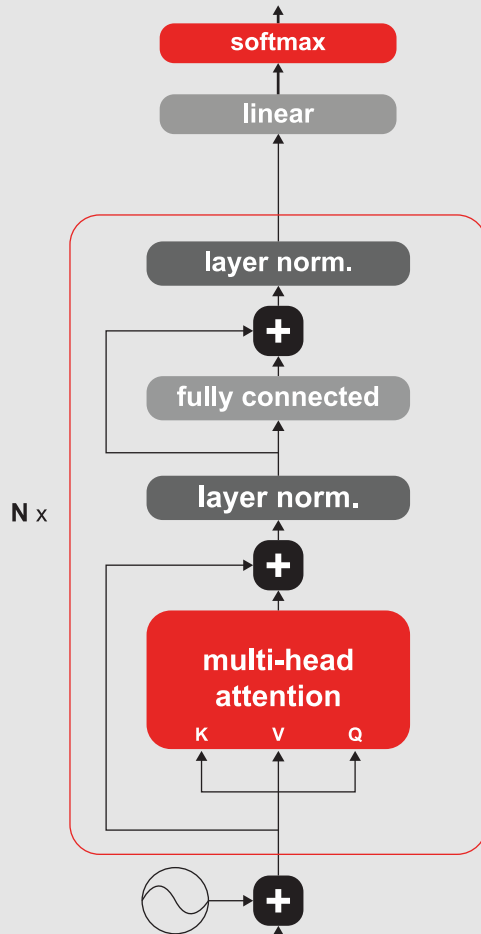
*'What are **those**?' he said while looking at my crocs.*

BERT ^[1]

- Bidirectional Encoder Representations from Transformers;
- using only the encoder part of Transformer.

[1] [Devlin et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.](#)

probability distribution over vocabulary



Task #1: Masked language modelling

- 15% of input words are masked, the model learns to predict the missing words

What looking
↑ ↑
'[MASK] are those?' he said while [MASK] at my crocs.

Too much masking:

Model is not provided with enough context.

Too little masking:

Learning becomes very slow.

Task #2: Next sentence prediction

- given a pair of sentences predict if they follow one another;
- aims to learn sentence relationships that are important is certain downstream tasks (e.g. question answering).

A: 'What are those?' he said while looking at my crocs.

B: My new shoes.

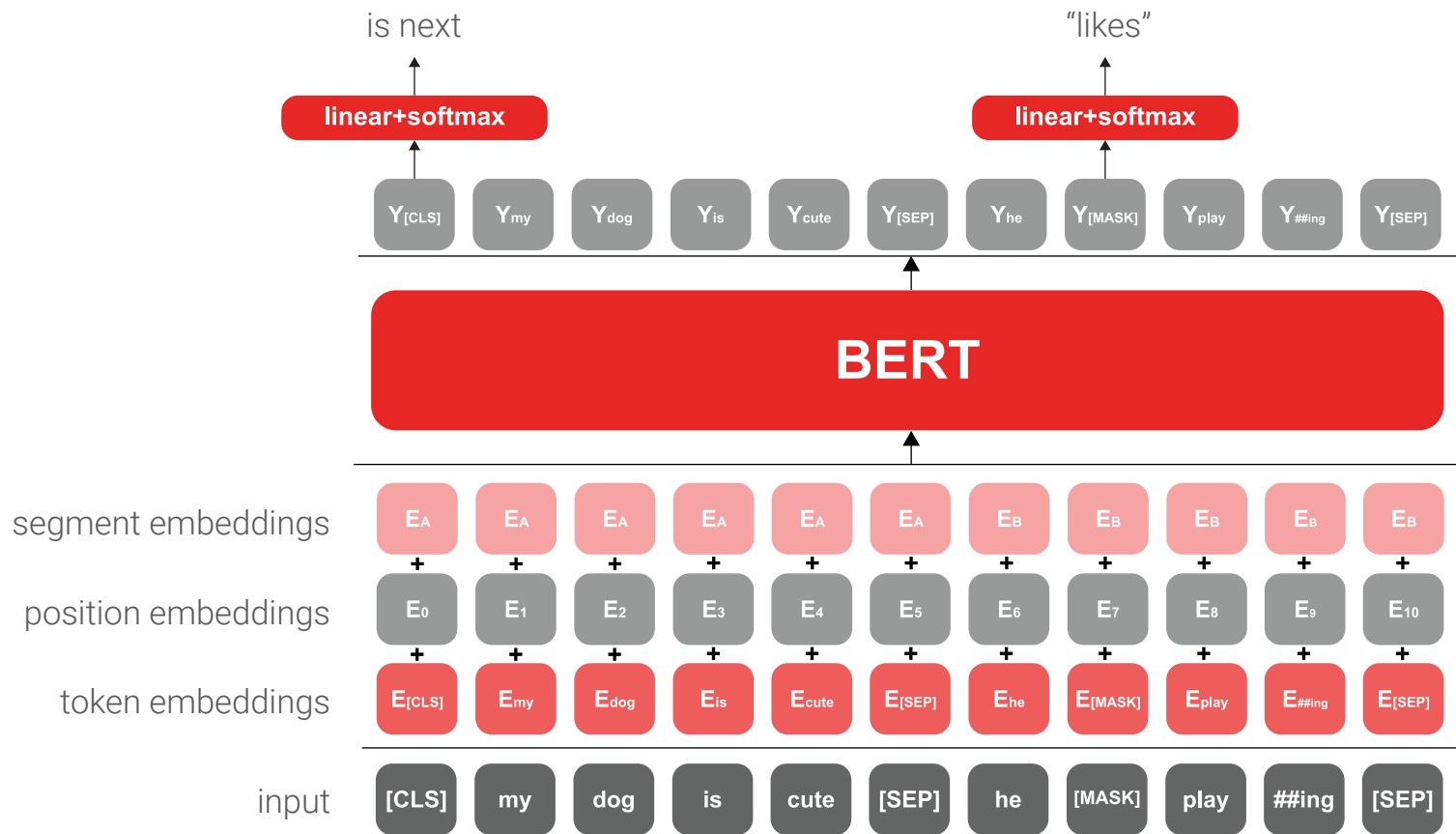
Ground truth: next

A: 'What are those?' he said while looking at my crocs.

B: The sky is blue.

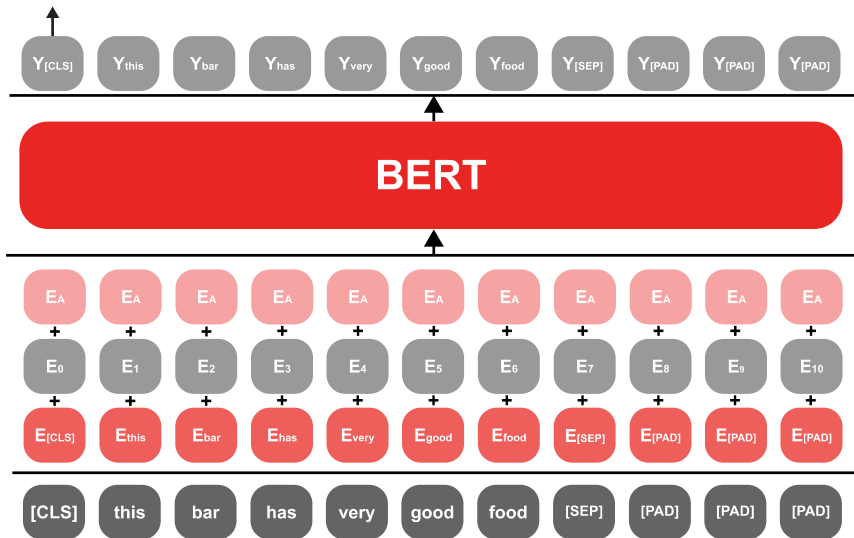
Ground truth: not next

Pre-training

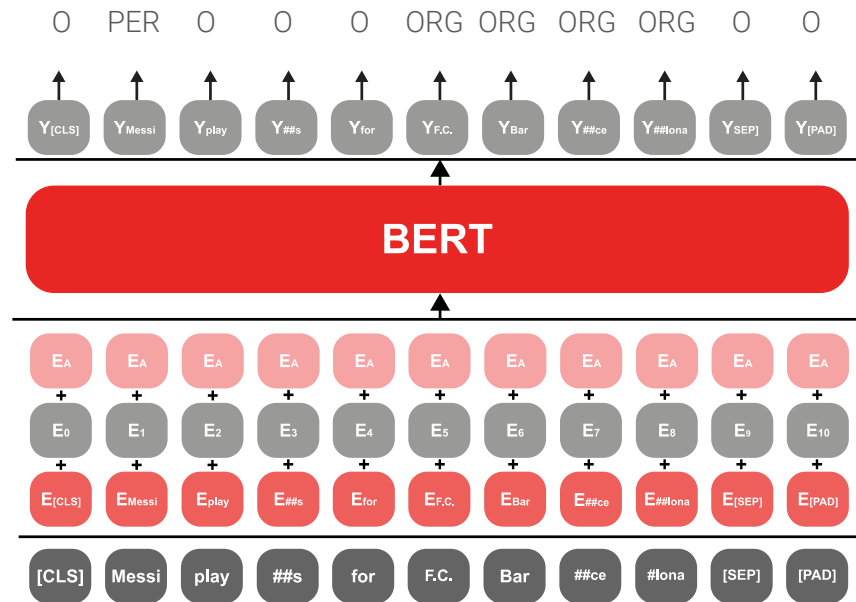


Fine-tuning

positive



classification



named entity recognition

GPT-2 ^[1] / GPT-3 ^[2]

- few architectural changes, layer norm now applied to input of each subblock, GPT-3 also uses some sparse attention layers;
- more data, larger batch sizes (GPT-3 uses batch size of 3.2M);
- the models are scaled:

GPT-2:

48 layers, 25 heads

$d_m = 1600$, $d = 64$

context size = 1024

~ 1.5B parameters

GPT-3:

96 layers, 96 heads

$d_m = 12288$, $d = 128$

context size = 2048

~ 175B parameters

[1] [Radford et al.: Language Models are Unsupervised Multitask Learners, 2019.](#)

[2] [Brown et al.: Language Models are Few-Shot Learners, 2020.](#)

- GPT-3 is still a language model and can be used for text generation

- only 12% of respondents correctly classified this as not written by a human

Title: United Methodists Agree to Historic Split

Subtitle: Those who oppose gay marriage will form their own denomination

Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Conclusion

- attention allows each word to directly access the words it depends on
- transformers achieve high parallelization/scalability by leveraging attention

- GPT and BERT can be applied to many different NLP tasks
- GPT is pre-trained on language modelling, whereas BERT is pre-trained for masked language modelling and next sentence prediction