History of the forbidden model

Karol Kaczmarek

Adam Mickiewicz University
Poznań

Applica.ai
Warsaw

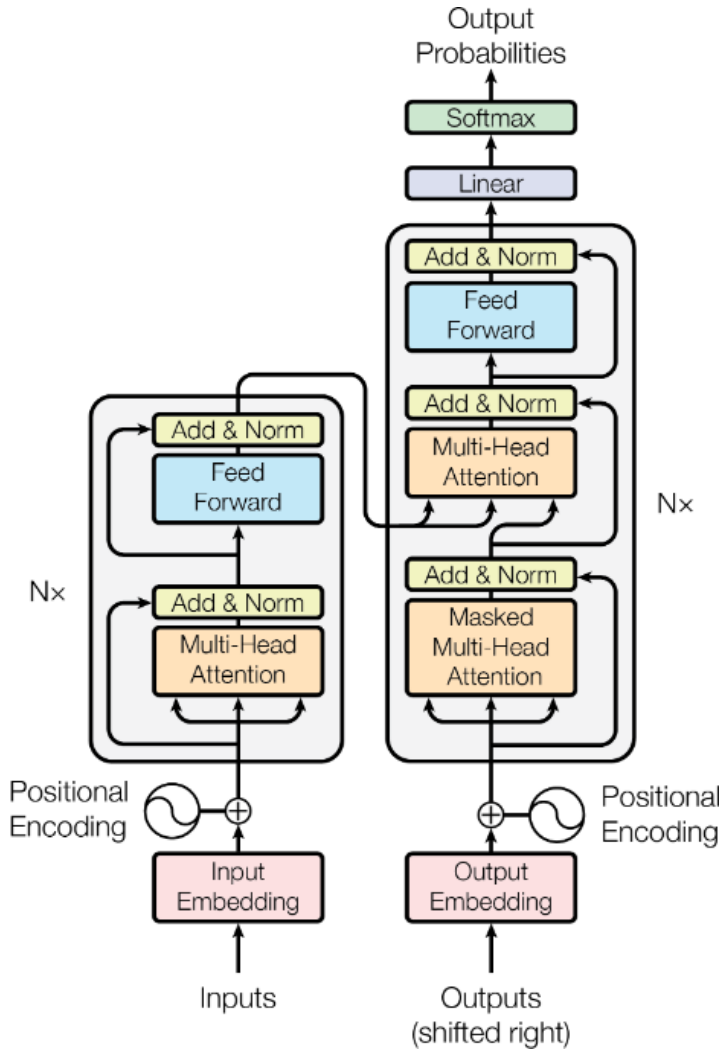
2019

Outline

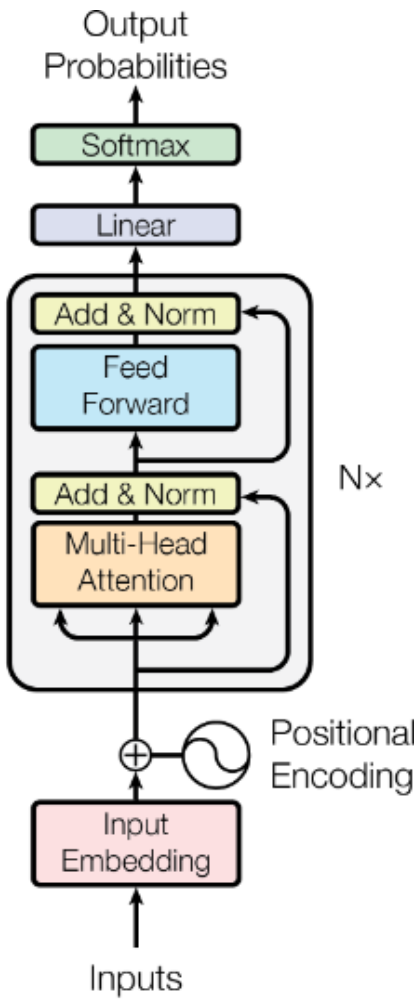
Transformer - Architecture

- publication: "Attention Is All You Need" [1]
- encoder-decoder model
- dispensing with recurrence and convolutions entirely
- attention mechanism

Transformer - Architecture



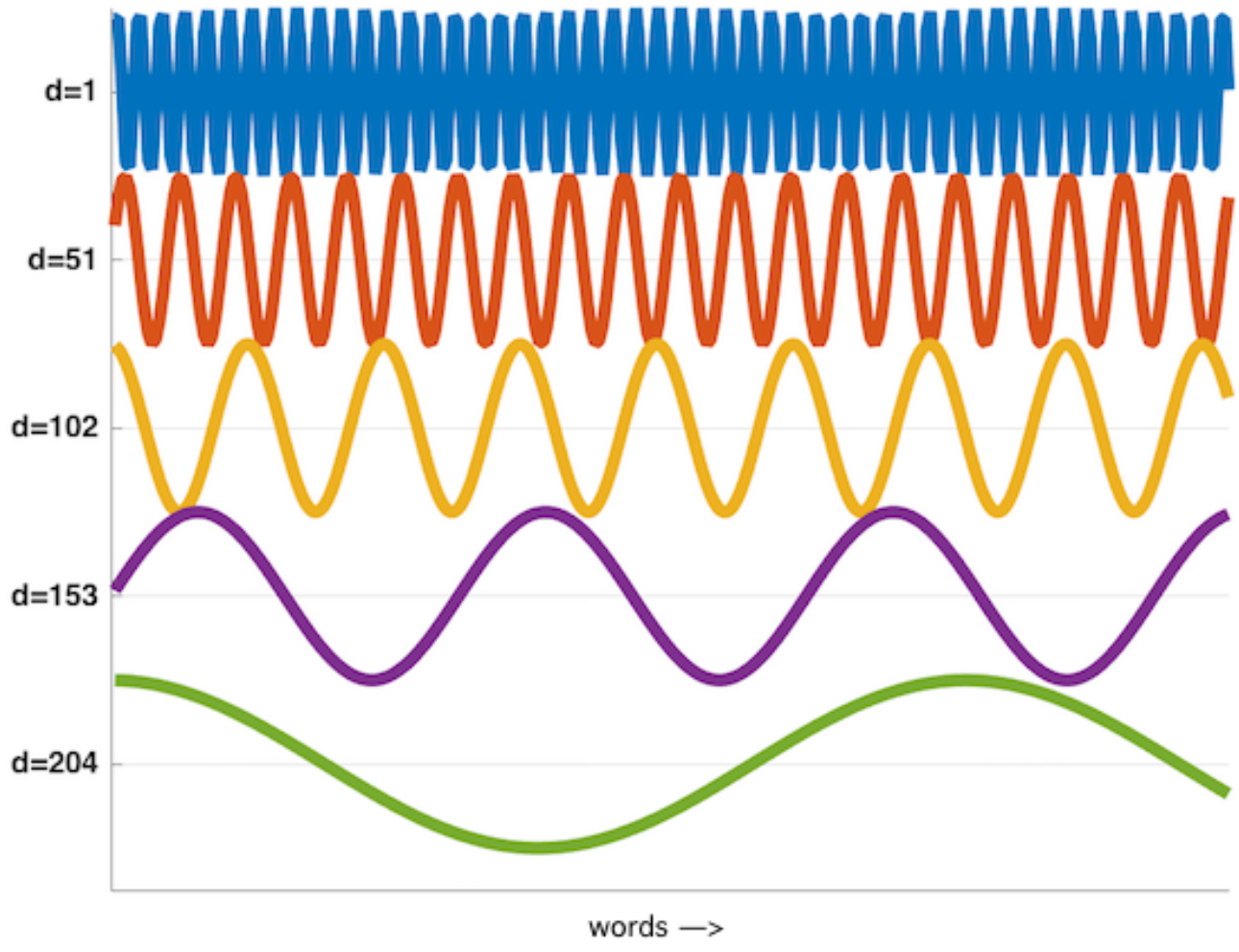
Transformer - Architecture



Transformer - Positional Encoding

- position of the word in the text (at the beginning or at the end)
- add *positional encodings* to the *input embeddings*
- $PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$
- $PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$
  - pos - position
  - i - dimension
  - d_{model} - dimension embeddings

Transformer - Positional Encoding

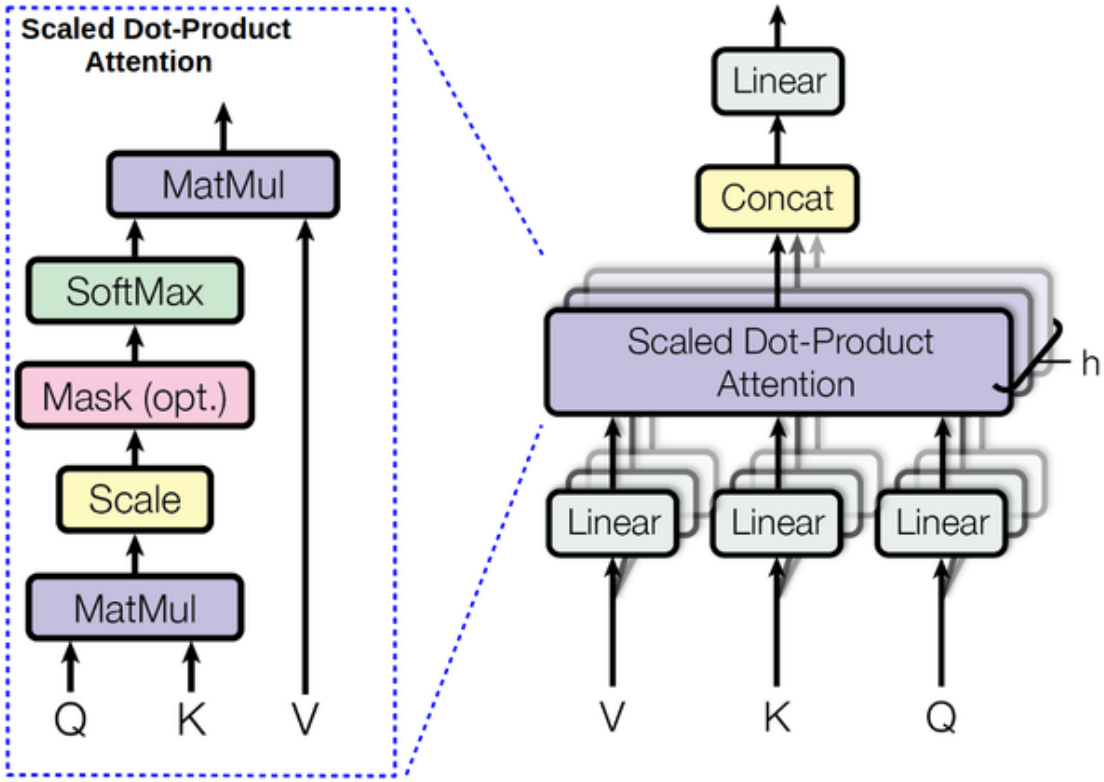


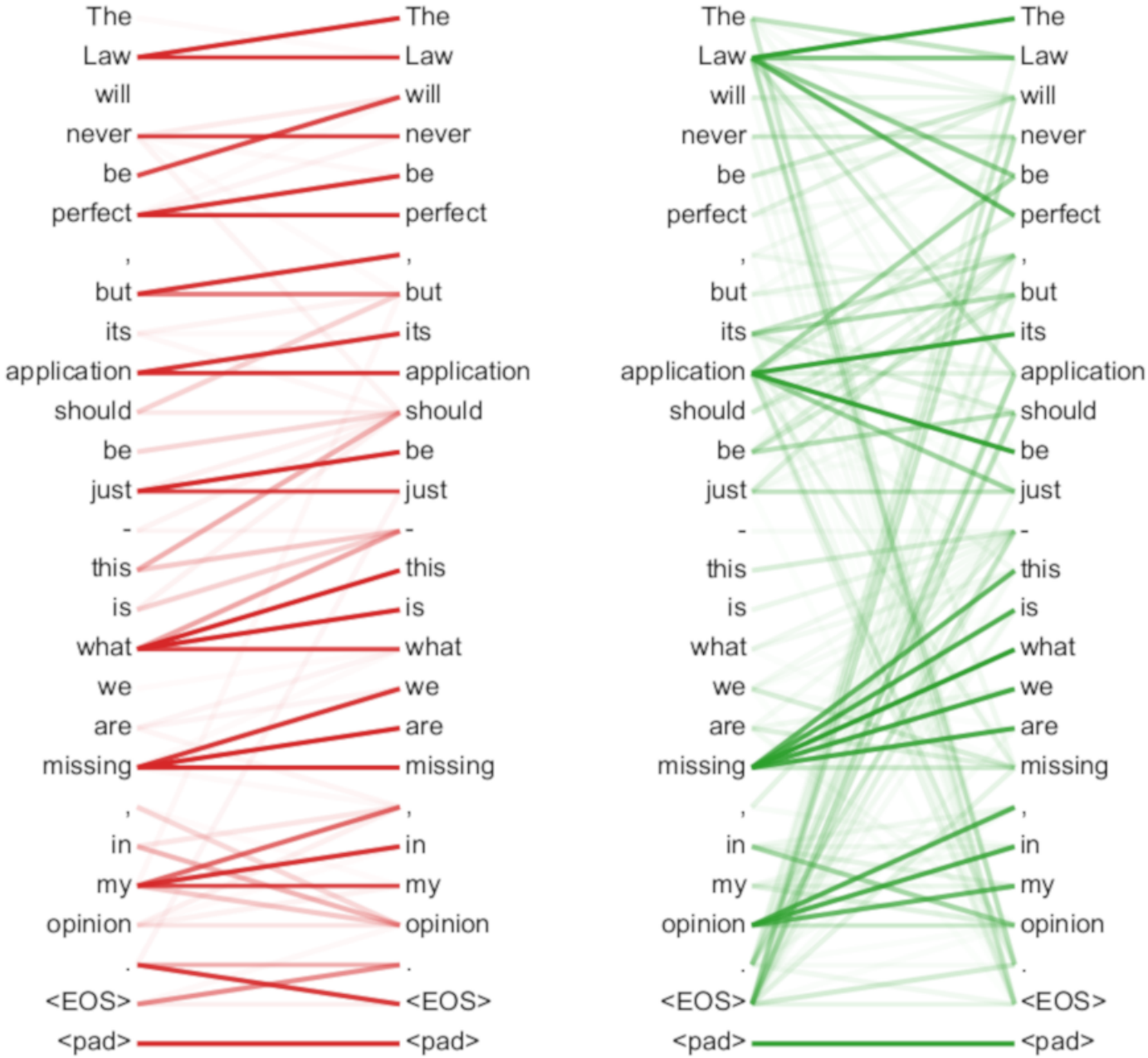
Transformer - Attention

- attention function can be described as mapping a *query* and a set of *key-value* pairs to an *output*, where:
  - Q - queries matrix
  - K - keys matrix
  - V - values matrix

Transformer - Multi-head attention

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$




The
Law
will
never
be
perfect
,
but
its
application
should
be
just
-
this
is
what
we
are
missing
,
in
my
opinion
.
<EOS>
<pad>
The
Law
will
never
be
perfect
,
but
its
application
should
be
just
-
this
is
what
we
are
missing
,
in
my
opinion
.
<EOS>
<pad>

Transformer - Visualization

Transformer - Visualization

Outline

GLUE Benchmark [2]

- **GLUE** – **G**eneral **L**anguage **U**nderstanding **E**valuation
- **N**atural **L**anguage **U**nderstanding tasks (NLU):
  - single-sentence tasks: CoLA, SST-2
  - similarity and paraphrase tasks: MRPC, STS-B, QQP
  - inference tasks: MNLI, QNLI, RTE, WNLI

GLUE Benchmark - Leaderboard

Name	**Score**
Human	87.1
ALICE large (Alibaba DAMO NLP)	82.9
MT-DNNv2 (Microsoft D365 AI)	82.9
BERT	$82.0 - 80.2$
Transformer	72.8
ELMO	70.5
Baseline	~ 70.0

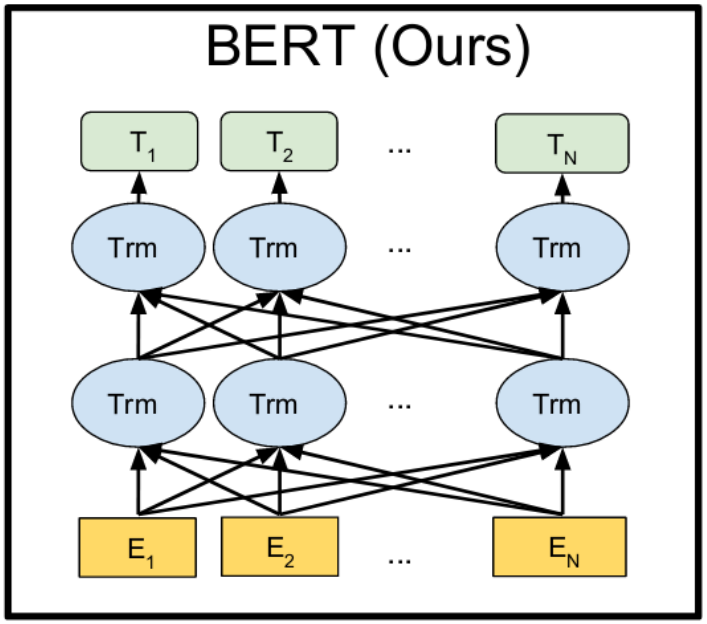
GLUE Benchmark - Leaderboard

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	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	-
+	2	王玮	ALICE large (Alibaba DAMO NLP)		82.9	61.6	95.2	91.1/87.7	89.6/88.6	74.0/90.4	87.9	87.4	95.4	80.9	65.1	40.7
+	3	Microsoft D365 AI & MSR AI	MT-DNNv2 (BigBird)		82.9	62.5	95.6	91.1/88.2	89.5/88.8	72.7/89.6	86.7	86.0	94.9	81.4	65.1	40.3
−	4	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
			GPT on STILTs		76.9	47.2	93.1	87.7/83.7	85.3/84.8	70.1/88.1	80.7	80.6	-	69.1	65.1	29.4
+	5	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hi		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
	6	Neil Houlsby	BERT + Single-task Adapters		80.2	59.2	94.3	88.7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2
	7	Alec Radford	Singletask Pretrain Transformer		72.8	45.4	91.3	82.3/75.7	82.0/80.0	70.3/88.5	82.1	81.4	-	56.0	53.4	29.8
+	8	Samuel Bowman	BiLSTM+ELMo+Attn		70.5	36.0	90.4	84.9/77.9	75.1/73.3	64.8/84.7	76.4	76.1	-	56.8	65.1	26.5
	9	GLUE Baselines	BiLSTM+ELMo+Attn		70.0	33.6	90.4	84.4/78.0	74.2/72.3	63.1/84.3	74.1	74.5	79.8	58.9	65.1	21.7
			BiLSTM+ELMo		67.7	32.1	89.3	84.7/78.0	70.3/67.8	61.1/82.6	67.2	67.9	75.5	57.4	65.1	21.3

Outline

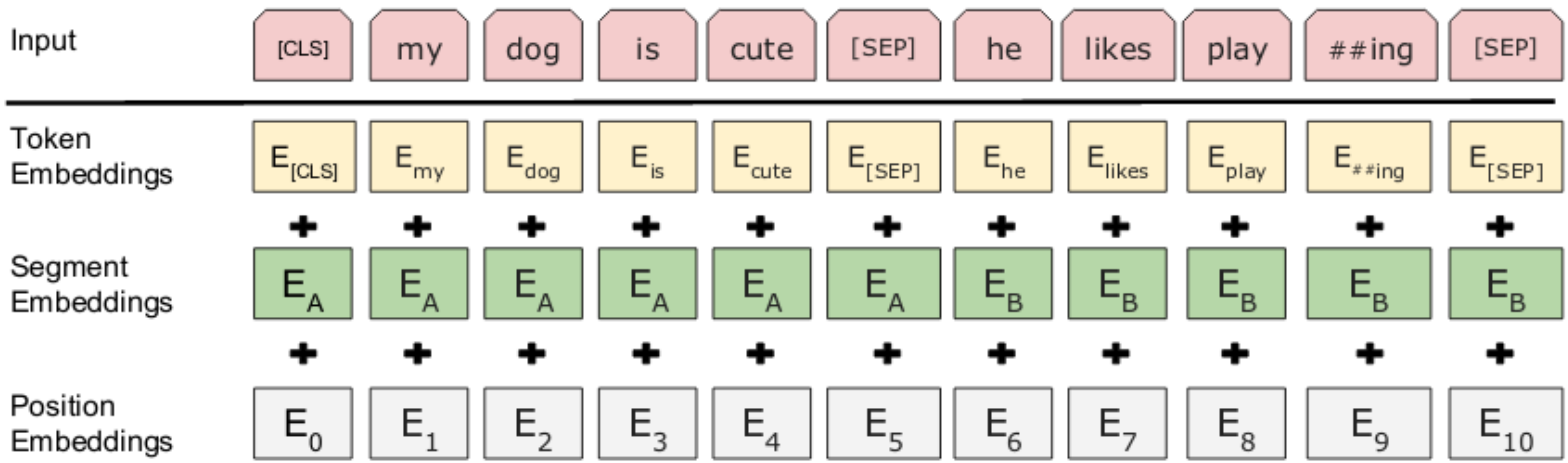
BERT [3]

- **BERT** – **B**idirectional **E**ncoder **R**epresentations from **T**ransformers = bidirectional Transformers



BERT - Input representation

- **[CLS]** – the special classification embedding
- **[SEP]** – sentences separator, sentence pairs are packed together into a single sequence
- segment embeddings



Masked Language Model (MLM)

- masking some percentage of the input tokens at random
- predicting only those masked tokens (**[MASK]**)
- mask 15% of all
- masking procedure:
  - 80% of the time - replace the word with the **[MASK]** token
    - *my dog is hairy* $\rightarrow$ *my dog is [MASK]*
  - 10% of the time - replace the word with a random word
    - *my dog is hairy* $\rightarrow$ *my dog is apple*
  - 10% of the time - keep the word unchanged
    - *my dog is hairy* $\rightarrow$ *my dog is hairy*

Outline

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

theguardian.com

An Elon Musk-backed AI firm is keeping a text generating tool under wraps amid fears it's too dangerous

businessinsider.com

Fake news: OpenAI's 'deepfakes for text', GPT2, may be too dangerous to be released

betanews.com

packtpub.com

OpenAI boi się sztucznej inteligencji, którą stworzyła

rp.pl

GPT-2: algorytm do tworzenia tekstów tak dobry, że aż tajny

polityka.pl

OpenAI nie upubliczni nowej sztucznej inteligencji. Jest zbyt niebezpieczna

ithardware.pl

GPT2: sztuczna inteligencja zbyt niebezpieczna, by udostępnić ją publicznie

giznet.pl

The world now

The world if OpenAI released GPT-2

Outline

GPT-2 [4]

- transformer-based language model
- training dataset (WebText):
  - scrapted web pages which have been curated/filtered by humans
  - links from *reddit*, which received at least 3 karma
  - 45 million links = 8 million documents = 40 GB of text
  - removed all Wikipedia documents
- using Byte Pair Encoding (BPE)
- 1 of 4 models have been published (117M, 345M, 762M, 1542M)
- training the big GPT-2 model costs $43k
- OpenWebText

Byte Pair Encoding (BPE)

- often operate of Unicode code points (vocabulary of over 130000, usually 32000-64000)
- operate of byte-level only required a base vocabulary of size 256
- generating multiple versions of common words (*dog. dog! dog?* for the word *dog*)
- prevents BPE from merging characters across categories (*dog* would not be merged with punctuation characters)
- <UNK> occurring only 26 times in 40 billion bytes
- do not need to worry about pre-processing, tokenization

GPT-2 - Model size

Name	**Parameters**	**Layers** (L)	**Dimensions** (d_{model})
base	117M	12	768
medium	345M	24	1024
large	762M	36	1280
xl	1542M	48	1600

GPT-2 - Score

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Human written: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model completion: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

askskynet.com

Outline

Transformer on Polish CommonCrawl

Jarosław Kaczyński , przewodniczący PiS i kandydat Prawa ręka - mówił szef Biura Rady Miejskiej Robert Miller zapowiedział , że szef PiS złożył na niego list , by minister sprawiedliwości Jacek Giertych o jego kadencji minister właściwy z dniem 11 I i

Outline

References I

[1] A. Vaswani and et al., “Attention is all you need,” 2017.

[2] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding,” 2019.

[3] J. Devlin and et al., “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.

[4] A. Radford, J. Wu, and et al, “Language models are unsupervised multitask learners,” 2019.