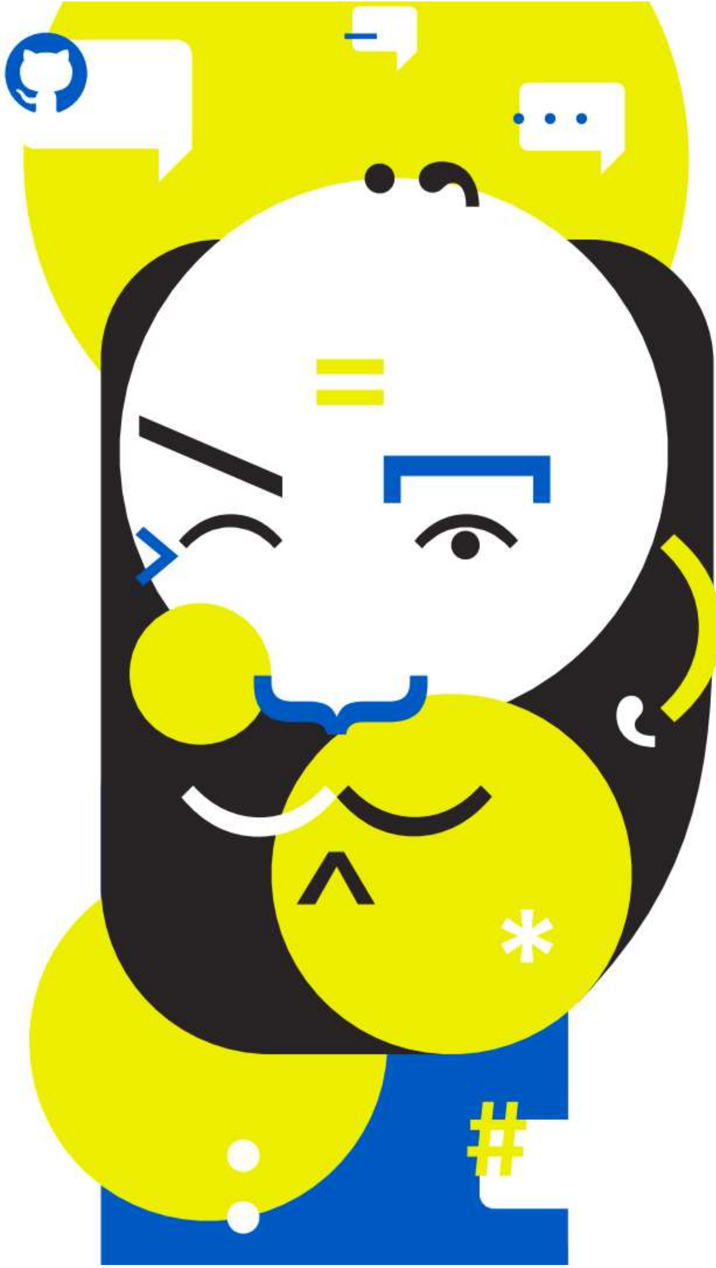
NLP Trends 2020

Mikhail Burtsev,

Neural networks and deep learning lab

MIPT
MOSCOW INSTITUTE
OF PHYSICS AND TECHNOLOGY



NLP tasks

- Sequence classification

Classes: AddToPlaylist BookRestaurant GetWeather PlayMusic RateBook SearchCreativeWork SearchScreeningEvent

BookRestaurant

Find a cosy place for 30 people to celebrated anniversary of the first DeepPavlov release.

- Sequence tagging

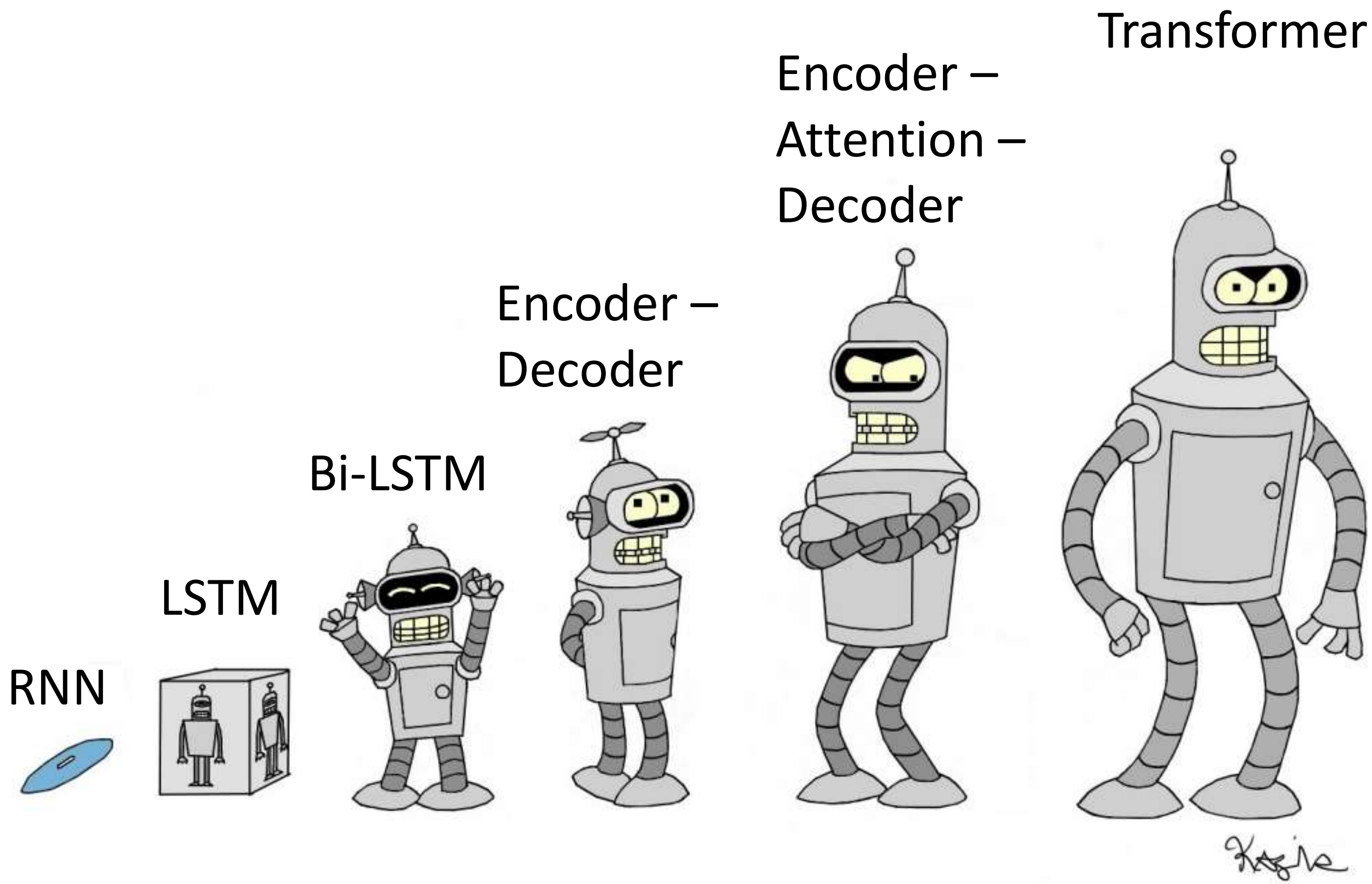
DeepPavlov ORG is an open source framework for chatbots and virtual assistants developed MIPT ORG Dolgoprudny GPE . first ORDINAL release was published two years ago DATE 2018 DATE and now it more than 3800 CARDINAL stars 93000 CARDINAL downloads .

- Sequence prediction \ generation

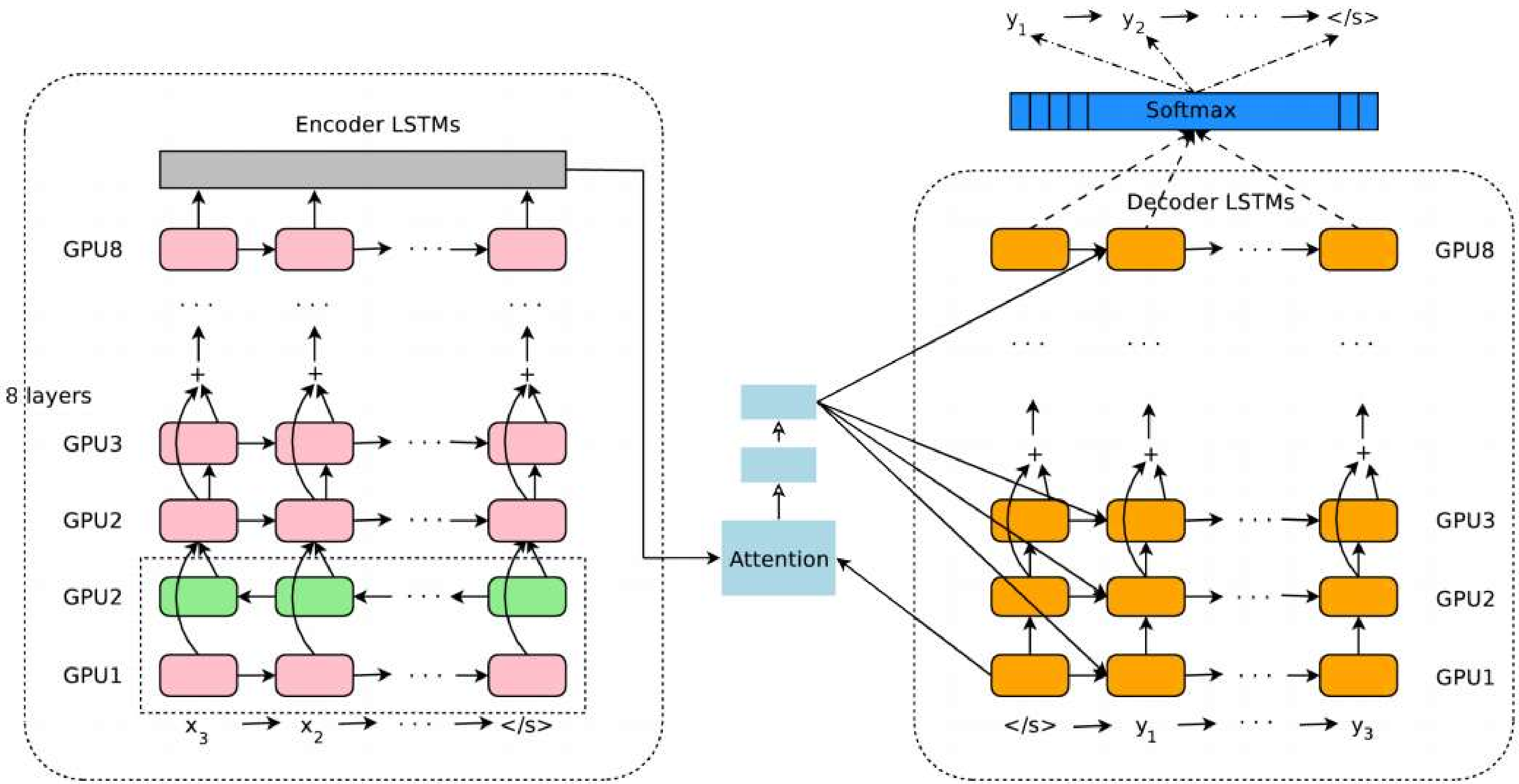
DeepPavlov is an open source framework for chatbots and virtual assistants. **Its features include automatic language tagging, word segmentation and phrase embeddings.**

Written by Transformer · transformer.huggingface.co

Evolution of NLP models

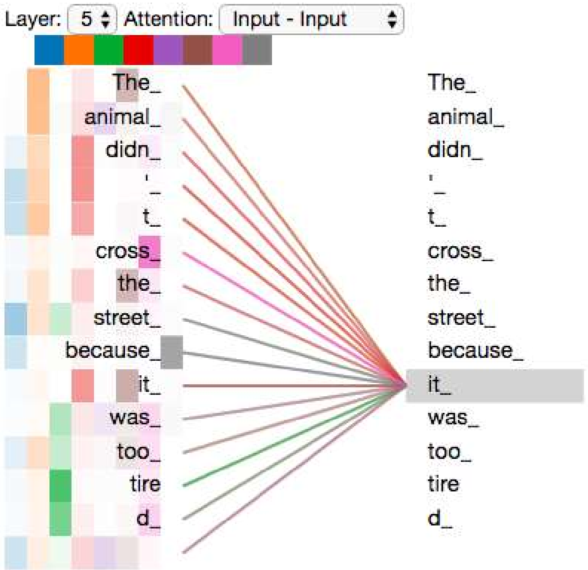


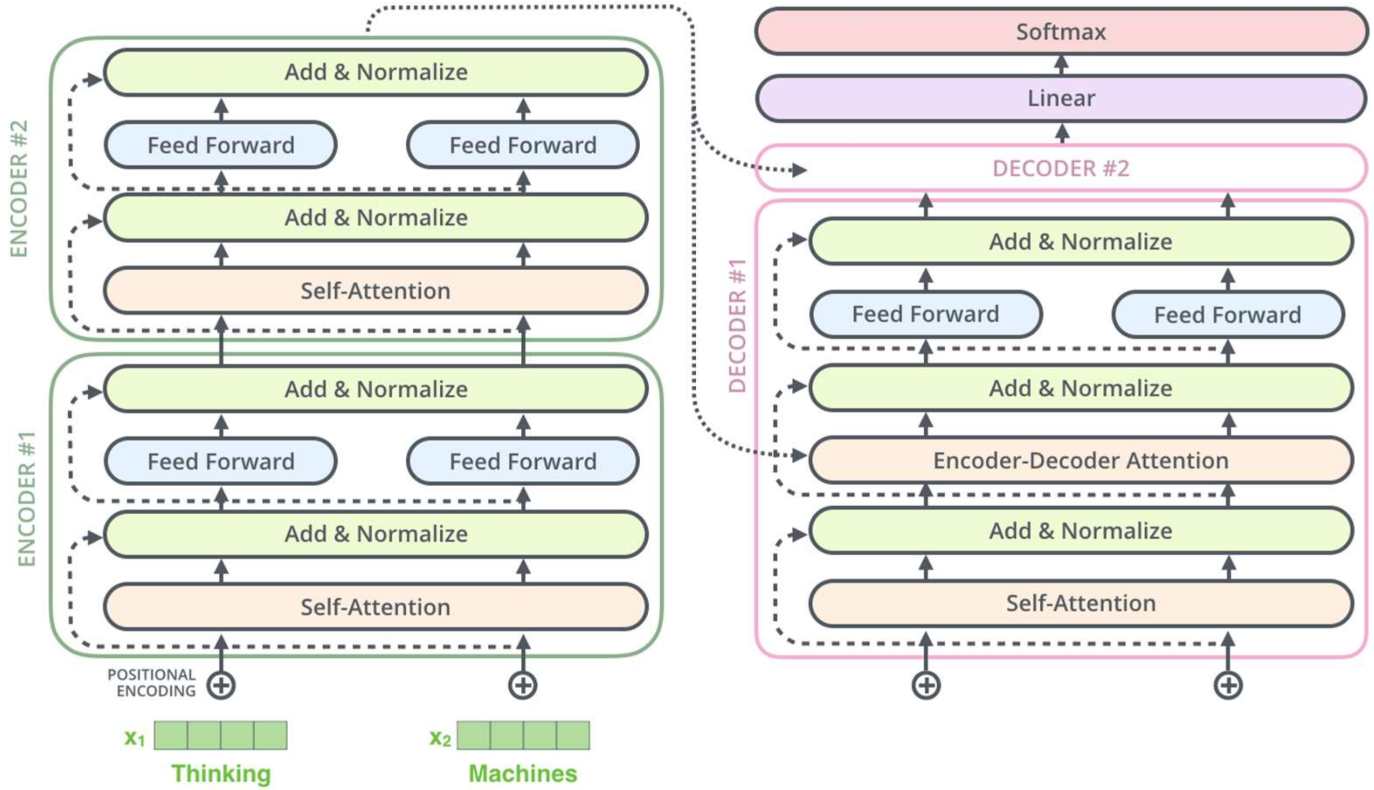
Encoder-Decoder with Attention



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144* (2016).

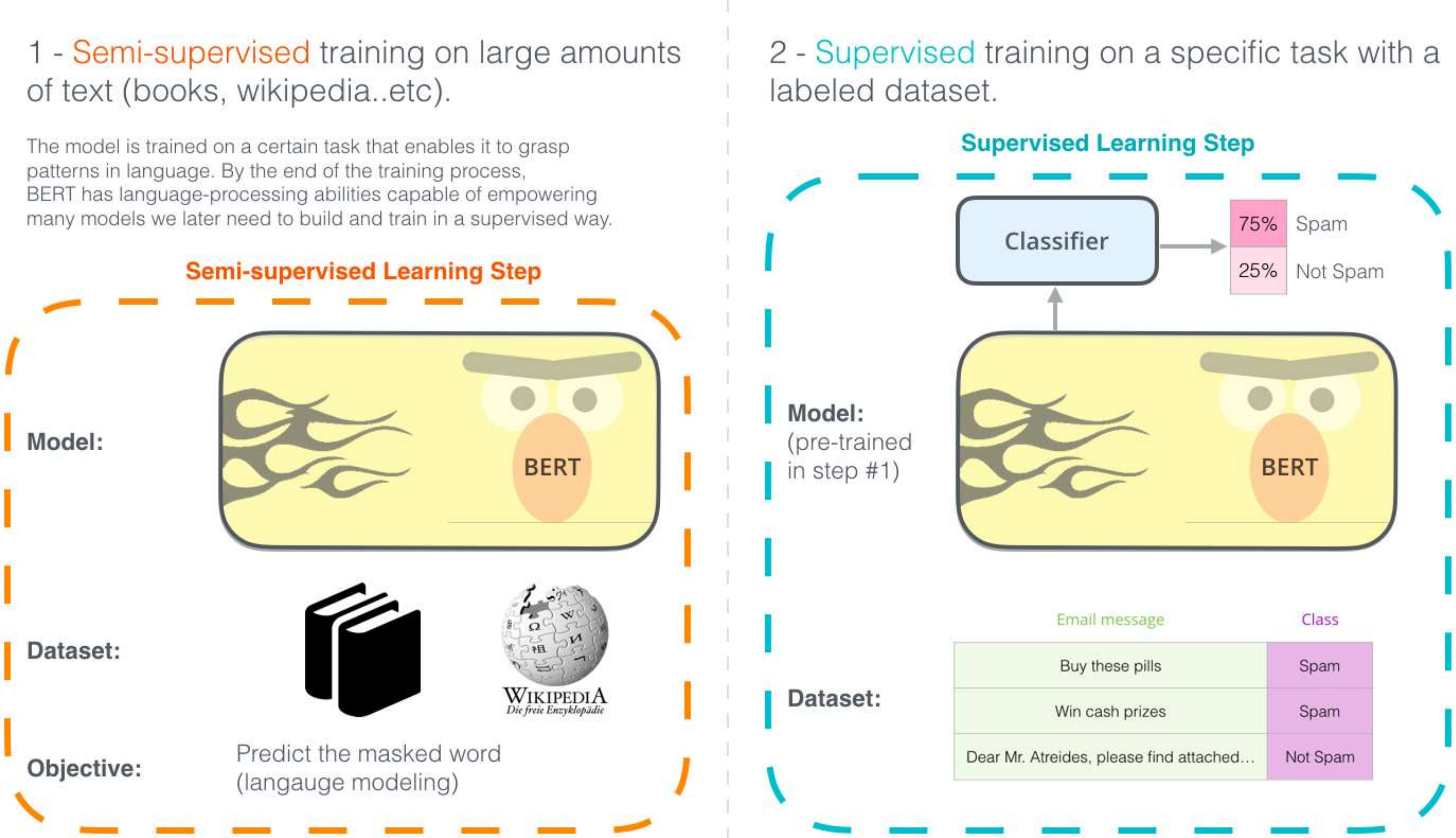
Transformer





Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
http://jalammar.github.io/illustrated-transformer/

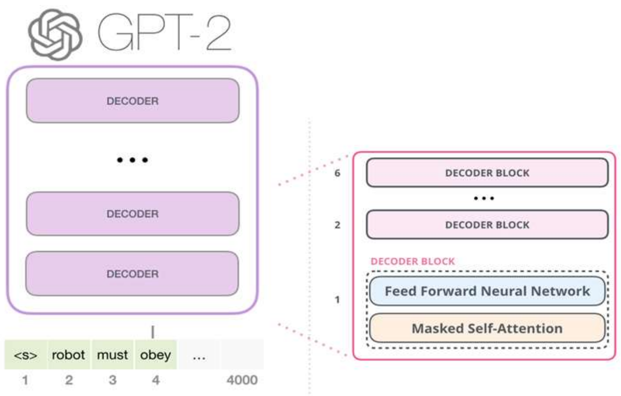
Universal pre-training / self-supervised learning / language models



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
http://jalammar.github.io/illustrated-bert/

Transformers zoo

Post BERT



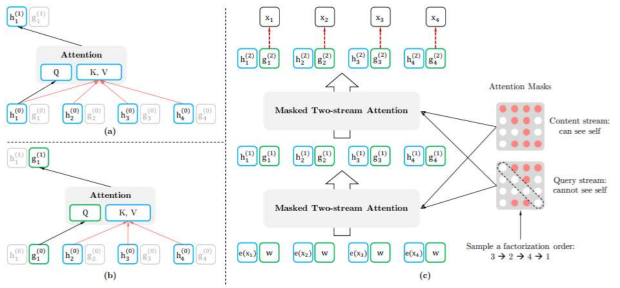


Figure 1: (a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access information about the content x_{z_t}. (c): Overview of the permutation language modeling training with two-stream attention.

BERT

OCTOBER 11, 2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin et al

Transformer-XL

JANUARY 9, 2019

Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

GPT-2

FEBRUARY 14, 2019

Language Models are Unsupervised Multitask Learners

ERNIE

APRIL 19, 2019

ERNIE: Enhanced Representation through Knowledge Integration

XLNet

JUNE 19, 2019

XLNet: Generalized Autoregressive Pretraining for Language Understanding

RoBERTa

JULY 26, 2019

RoBERTa: A Robustly Optimized BERT Pretraining Approach

CTRL

SEPTEMBER 11, 2019

CTRL: A Conditional Transformer Language Model for Controllable Generation

ALBERT

SEPTEMBER 26, 2019

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

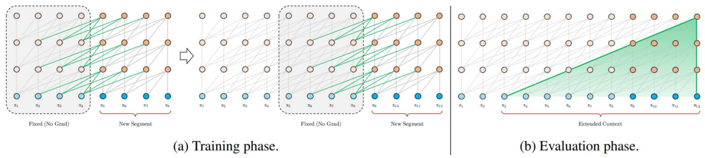


Figure 2: Illustration of the Transformer-XL model with a segment length 4.

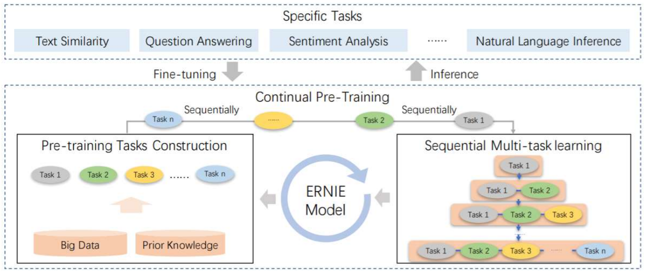


Figure 1: The framework of ERNIE 2.0, where the pre-training tasks can be incrementally constructed, the models are pre-trained through continual multi-task learning, and the pre-trained model is fine-tuned to adapt to various language understanding tasks.

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
	xlarge	1270M	24	2048	2048	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

Benchmarking

GLUE

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	ERNIE Team - Baidu	ERNIE		90.2	72.2	97.5	93.0/90.7	92.9/92.5	75.2/90.8	91.2	90.6	98.0	90.9	94.5	49.4
+	2	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		90.1	73.2	97.1	93.9/91.9	93.0/92.5	74.8/91.0						
	3	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2						
	4	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4						
	5	XLNet Team	XLNet (ensemble)		89.5	70.2	97.1	92.9/90.5	93.0/92.6	74.7/90.4						
	6	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5						
	7	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						
	8	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2						
	9	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2						
+	10	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						
	11	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4						

GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

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

Benchmarking

SuperGLUE

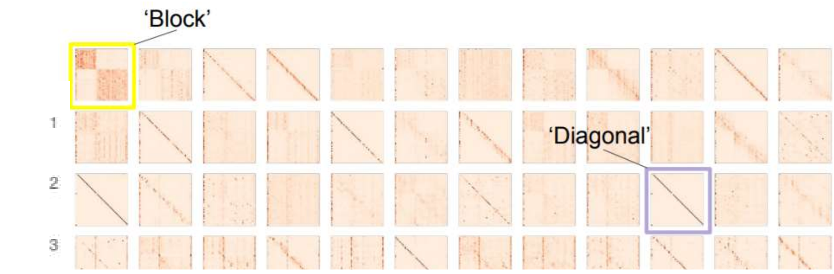
Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
2	T5 Team - Google	T5		88.9	91.0	93.0/96.4	94.8	88.2/62.3	93.3/92.5	92.5	76.1	93.8	65.6	92.7/91.9
3	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
4	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
5	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
6	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
		BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7
		Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/50.0
		CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	-0.4	100.0/50.0
		Outside Best		-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]		-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	47.6	-

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

Wang, Alex, et al. "Superglue: A stickier benchmark for general-purpose language understanding systems." arXiv preprint arXiv:1905.00537 (2019).
https://super.gluebenchmark.com/

BERTology



Diagonals. Copy existing token vector.
Verticals. Everyone looks at the most important tokens.
Diagonal blur. Local information from neighbor tokens is important.

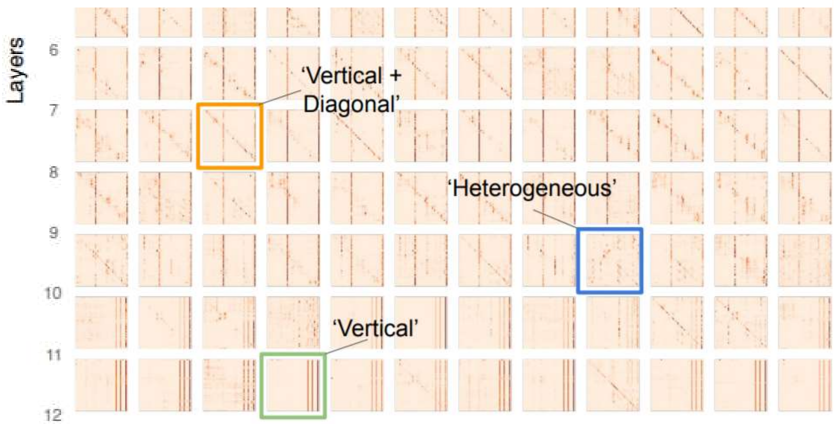


Figure 3. Example of self-attention maps (for QNLI). Rows represent layers and columns represent heads.

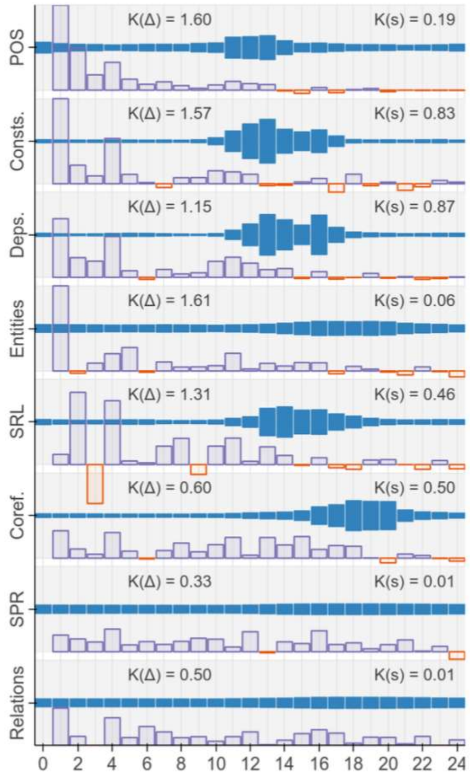


Figure 2: Layer-wise metrics on BERT-large. Solid (blue) are mixing weights $s_\tau^{(\ell)}$ (§3.1); outlined (purple) are differential scores $\Delta_\tau^{(\ell)}$ (§3.2), normalized for each task. Horizontal axis is encoder layer.

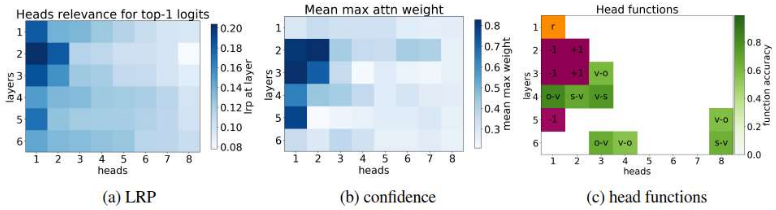
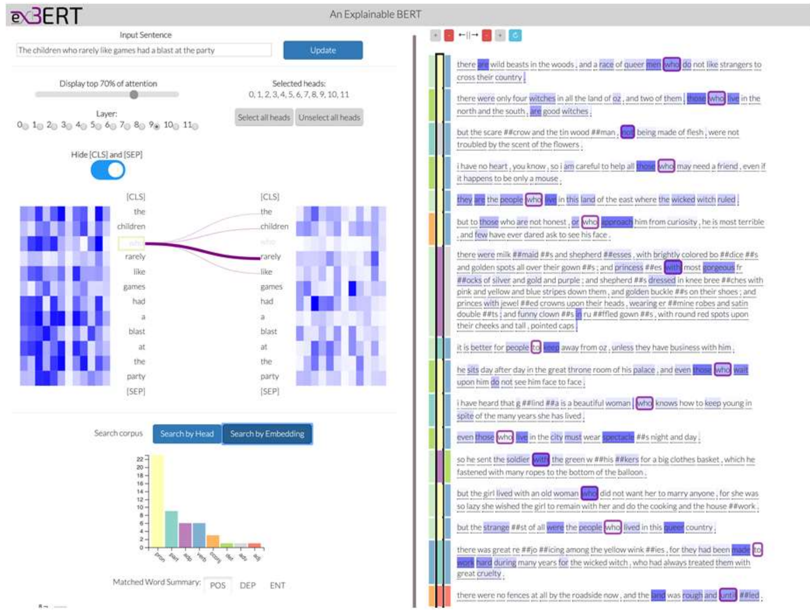


Figure 1: Importance (according to LRP), confidence, and function of self-attention heads. In each layer, heads are sorted by their relevance according to LRP. Model trained on 6m OpenSubtitles EN-RU data.



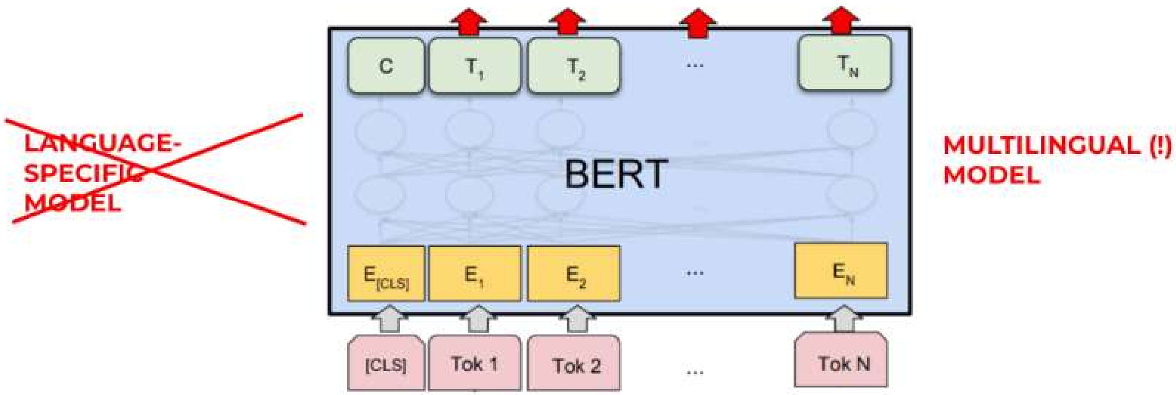
http://exbert.net/

Kovaleva, Olga, et al. "Revealing the dark secrets of bert." arXiv preprint arXiv:1908.08593 (2019).
Tenney, Ian, Dipanjan Das, and Ellie Pavlick. "Bert rediscovers the classical nlp pipeline." arXiv preprint arXiv:1905.05950 (2019).
Voita, Elena, et al. "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned." arXiv preprint arXiv:1905.09418 (2019).
Hoover, Benjamin, Hendrik Strobelt, and Sebastian Gehrmann. "exbert: A visual analysis tool to explore learned representations in transformers models." arXiv preprint arXiv:1910.05276 (2019).

Multilingual transfer

Після аномальної весни , що увійшла в десятку найтепліших 139 років DATE спостережень літо DATE теж починаєтся зі спеки . Про це повідомила Наталка Діденко PERSON на своїй сторінці в Facebook ORG . Так , Україна GPE буде залишатися однією з найбільш спекотних Європи LOC : завтра вдень + 24 + 29 градусів QUANTITY , Сході LOC + 28 + 33 градуси QUANTITY . За словами синоптика , Франції GPE Великобританії GPE і місцями навіть Іспанії GPE Португалії GPE в середу похолодає + 10 + 15 градусів QUANTITY і пройдуть дощі .

https://demo.deeppavlov.ai/#/mu/ner



	German	Russian	Chinese	Vietnamese
PER	87.21	95.74	84.12	83.30
LOC	69.54	82.62	60.83	60.99
ORG	52.95	55.68	54.34	38.92
Total	70.71	79.39	64.44	68.20

Conversational AI

Schema

(part of DSTC8 schema)

```
{
        "name": "party_size",
        "description": "Party size for a reservation",
        "is_categorical": true,
        "possible_values": [
          "1",
          "2",
          "3",
          "4",
          "5",
          "6"
        ]
      },
{
        "name": "date",
        "description": "Date for the reservation or to find availability",
        "is_categorical": false,
        "possible_values": []
      },
      {
        "name": "time",
        "description": "Time for the reservation or to find availability",
        "is_categorical": false,
        "possible_values": []
      },
```

Enter text

That would work great for me. I'd like to make a reservation for two people on the 5th of March at half past 5 in the evening.

Question

Party size for a reservation?

Ask

https://demo.deeppavlov.ai/#/en/textqa

Q: Party size for a reservation?

That would work great for me. I'd like to make a reservation for **two** A people on the 5th of March at half past 5 in the evening.

Q: Date for the reservation or to find availability?

That would work great for me. I'd like to make a reservation for two people on **the 5th of March** A at half past 5 in the evening.

Q: Time for the reservation or to find availability?

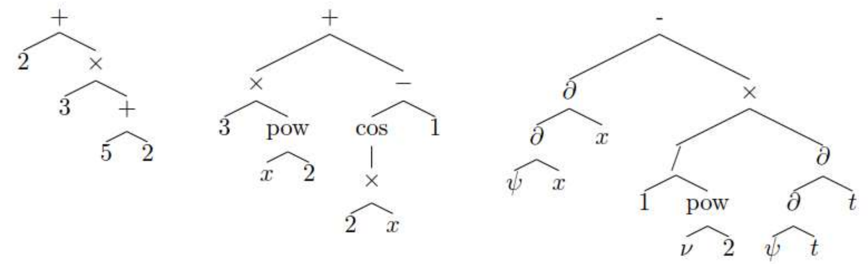
That would work great for me. I'd like to make a reservation for two people on the 5th of March at **half past 5 in the evening** A .

Seq2Seq Math

2 Mathematics as a natural language

2.1 Expressions as trees

Mathematical expressions can be represented as trees, with operators and functions as internal nodes, operands as children, and numbers, constants and variables as leaves. The following trees represent expressions $2 + 3 \times (5 + 2)$, $3x^2 + \cos(2x) - 1$, and $\frac{\partial^2 \psi}{\partial x^2} - \frac{1}{\nu^2} \frac{\partial^2 \psi}{\partial t^2}$:



4.2 Model

For all our experiments, we train a seq2seq model to predict the solutions of given problems, i.e. to predict a primitive given a function, or predict a solution given a differential equation. We use a transformer model (Vaswani et al., 2017) with 8 attention heads, 6 layers, and a dimensionality of 512. In our experiences, using larger models did not improve the performance. We train our models with the Adam optimizer (Kingma & Ba, 2014), with a learning rate of 10^{-4}. We remove expressions with more than 512 tokens, and train our model with 256 equations per batch.

	Integration (BWD)	ODE (order 1)	ODE (order 2)
Mathematica (30s)	84.0	77.2	61.6
Matlab	65.2	-	-
Maple	67.4	-	-
Beam size 1	98.4	81.2	40.8
Beam size 10	99.6	94.0	73.2
Beam size 50	99.6	97.0	81.0

Table 3: **Comparison of our model with Mathematica, Maple and Matlab on a test set of 500 equations.** For Mathematica we report results by setting a timeout of 30 seconds per equation. On a given equation, our model typically finds the solution in less than a second.

Equation	Solution
$y' = \frac{16x^3 - 42x^2 + 2x}{(-16x^8 + 112x^7 - 204x^6 + 28x^5 - x^4 + 1)^{1/2}}$	$y = \sin^{-1}(4x^4 - 14x^3 + x^2)$
$3xy\cos(x) - \sqrt{9x^2\sin(x)^2 + 1}y' + 3y\sin(x) = 0$	$y = c\exp\left(\sinh^{-1}(3x\sin(x))\right)$
$4x^4yy'' - 8x^4y'^2 - 8x^3yy' - 3x^3y'' - 8x^2y^2 - 6x^2y' - 3x^2y'' - 9xy' - 3y = 0$	$y = \frac{c_1 + 3x + 3\log(x)}{x(c_2 + 4x)}$

Table 4: Examples of problems that our model is able to solve, on which Mathematica and Matlab were not able to find a solution. For each equation, our model finds a valid solution with greedy decoding.

Lampe, Guillaume, and François Charton. "Deep learning for symbolic mathematics." arXiv preprint arXiv:1912.01412 (2019).

Future Conv AI research