Bertology

Part 2

Katya Artemova

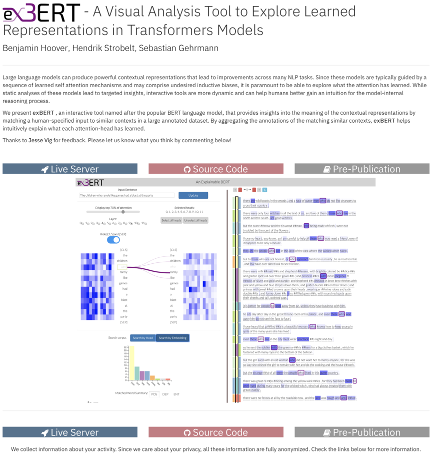
Computational Pragmatics Lab, HSE

October 30, 2019

BERT

- input: $[CLS], x_1, ..., x_N, [SEP], y_1, ..., y_M, [EOS]$
- objectives:
  1. masked language modeling
  2. next sentence prediction
- base architecture: $(L = 12, H = 768, A = 12, 110 Mparams)$

ExBERT: A Visual Analysis Tool to Explore Learned Representations in Transformers Models [1]



exBERT - A Visual Analysis Tool to Explore Learned Representations in Transformers Models

Benjamin Hoover, Hendrik Strobelt, Sebastian Gehrmann

Large language models can produce powerful contextual representations that lead to improvements across many NLP tasks. Since these models are typically guided by a sequence of learned self attention mechanisms and may comprise undesired inductive biases, it is paramount to be able to explore what the attention has learned. While static analyses of these models lead to targeted insights, interactive tools are more dynamic and can help humans better gain an intuition for the model-internal reasoning process.

We present **exBERT** , an interactive tool named after the popular BERT language model, that provides insights into the meaning of the contextual representations by matching a human-specified input to similar contexts in a large annotated dataset. By aggregating the annotations of the matching similar contexts, **exBERT** helps intuitively explain what each attention-head has learned.

Thanks to Jesse Vig for feedback. Please let us know what you think by commenting below!

We collect information about your activity. Since we care about your privacy, all these information are fully anonymized. Check the links below for more information.

Image source: exbert.net

Today

1 Probing BERT

2 BERT & Family

3 BERT & Friends

4 Latest news

What does BERT learn about the structure of language?[2]

Obervation 1: BERT mostly captures phrase-level information in the lower layers and this information gets gradually diluted in higher layers

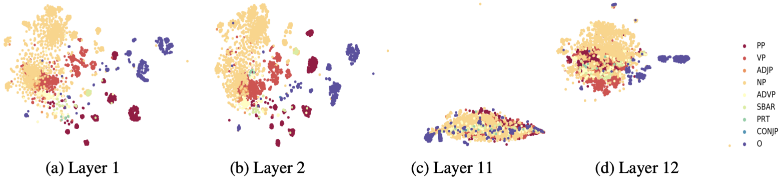
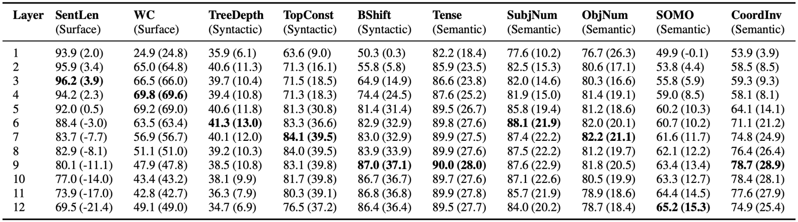


Figure: 2D t-SNE plot of span embeddings computed from the first and last two layers of BERT

What does BERT learn about the structure of language?[2]

Obervation 2: BERT embeds a rich hierarchy of linguistic signals: surface information at the bottom, syntactic information in the middle, semantic information at the top



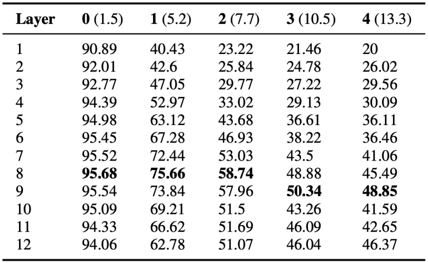
Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	**96.2 (3.9)**	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	**69.8 (69.6)**	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

Figure: Probing task performance for each BERT layer

Probing tasks: predict sentence length, presence of words, test for sensitivity to word order, the depth of the syntactic tree, top level constituents in the syntax tree, check for the tense, the subject number, the sensitivity to random replacement of a noun/verb.

What does BERT learn about the structure of language?[2]

Obervation 3: BERT requires deeper layers to model long-range dependency information



Layer	**0** (1.5)	**1** (5.2)	**2** (7.7)	**3** (10.5)	**4** (13.3)
1	90.89	40.43	23.22	21.46	20
2	92.01	42.6	25.84	24.78	26.02
3	92.77	47.05	29.77	27.22	29.56
4	94.39	52.97	33.02	29.13	30.09
5	94.98	63.12	43.68	36.61	36.11
6	95.45	67.28	46.93	38.22	36.46
7	95.52	72.44	53.03	43.5	41.06
8	**95.68**	**75.66**	**58.74**	48.88	45.49
9	95.54	73.84	57.96	**50.34**	**48.85**
10	95.09	69.21	51.5	43.26	41.59
11	94.33	66.62	51.69	46.09	42.65
12	94.06	62.78	51.07	46.04	46.37

Figure: Subject-verb agreement scores for each BERT

The task of predicting the verb number becomes harder when there are more nouns with opposite number intervening between the subject and the verb.

BERT Rediscovers the Classical NLP Pipeline [3]

A probing classifier receives spans $s_1 = [i_1, j_1)$ and (optionally) $s_2 = [i_2, j_2)$ and must predict a label such as a constituent or relation type

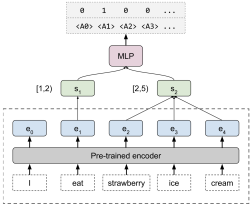
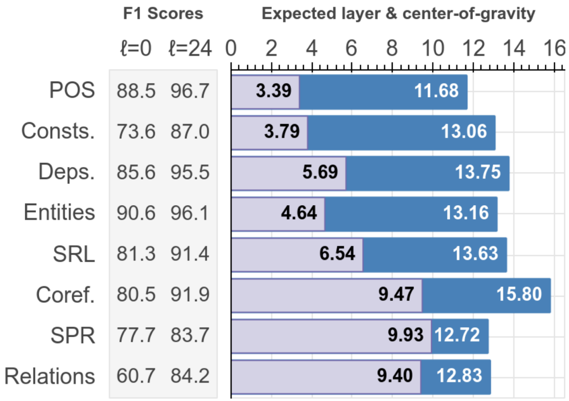


Figure: Probing model architecture

Eight labeling tasks: part-of-speech (POS), constituents (Consts.), dependencies (Deps.), entities, semantic role labeling (SRL), coreference (Coref.), semantic proto-roles (SPR; Reisinger et al., 2015), and relation classification (SemEval)

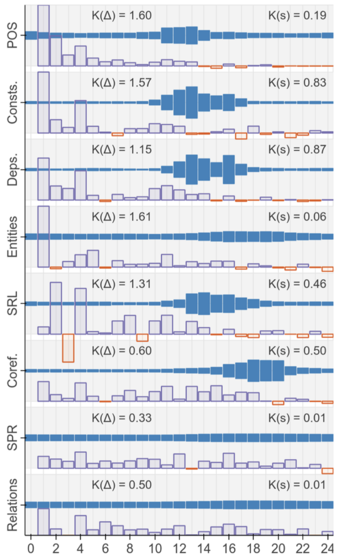
BERT Rediscovers the Classical NLP Pipeline [3]



	ℓ=0	ℓ=24		
POS	88.5	96.7	3.39	11.68
Consts.	73.6	87.0	3.79	13.06
Deps.	85.6	95.5	5.69	13.75
Entities	90.6	96.1	4.64	13.16
SRL	81.3	91.4	6.54	13.63
Coref.	80.5	91.9	9.47	15.80
SPR	77.7	83.7	9.93	12.72
Relations	60.7	84.2	9.40	12.83

Figure: Probing model architecture

BERT Rediscovers the Classical NLP Pipeline [3]



Today

1 Probing BERT

2 BERT & Family

3 BERT & Friends

4 Latest news

RoBERTa: A Robustly Optimized BERT Pretraining Approach [4]

RoBERTa is an improved pretraining procedure for BERT. It is trained with:

1. dynamic masking
2. on full sentences that may cross document boundaries (a special separator token is added) without NSP loss
3. with larger batches (2K, 8K)
4. a larger byte-level BPE

on approx. 160GB uncompressed unannotated texts.
RoBERTa achieves state-of-the-art results on GLUE, RACE and SQuAD, without multi-task finetuning for GLUE or additional data for SQuAD.

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations [5]

1. factorized embedding parameterization: in orginal model, $E = H$. Here the one-hot vectors are projected into a lower dimensional embedding space of size E, and then project it to the hidden space H
2. cross-layer parameter sharing: FFN parameters and Attention parameters are shared
3. inter-sentence coherence loss: sentence order predictions $(< s_1, s_2, 1 >, < s_2, s_1, 0 >)$

ALBERT has about 18x fewer parameters compared to BERT.
ALBERT-xxlarge (1M params) outperforms both BERT and RoBERTa.

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [6]

Knowledge distillation is a compression technique in which a compact model - the student, DistilBERT - is trained to reproduce the behaviour of a larger model - the teacher, BERT.

1. a linear combination of the distillation loss $L_{ce} = t_i \log(s_i)$, L_{mlm} and cosine embedding loss L_{cos}
2. student architecture: same architecture, but $H/2$ hidden layers
3. the student is initialized from the teacher by taking one layer out of two

DistilBERT compares surprisingly well to BERT, retaining 97% of the performance with 40% fewer parameters on GLUE

BERT & Family

1. More data and even more GPU! ... this approach is heavily criticized
2. GLUE benchmarking leads to ensemble training and a lot of fine tuning
3. BERT ancestors share a lot: data sources, NSP loss is omitted
4. MLM remains the core objective
5. BUT all of these models lack of common sense! understanding
6. BERT separately reconstructs all masked tokens
7. Pretrain-finetune discrepancy: the input contains artificial symbols like [MASK] that never occur in downstream tasks

Pretrained language model based on BERT

1. BioBERT [7]
2. ClinicalBERT [8]
3. SciBERT [9]

There is no surprise, all these models show significant improvement over current SoTA.
Experiment design:

1. to create a new vocabulary or not
2. to train the model from scratch or to fine-tune BERT

Today

1 Probing BERT

2 BERT & Family

3 BERT & Friends

4 Latest news

AR vs AE LMs

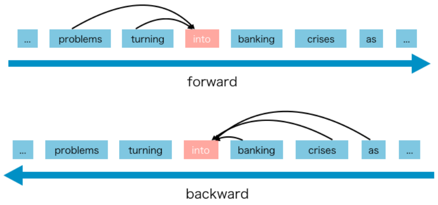


Figure: Autoregressive language model predicts the word based on its left (or right) context

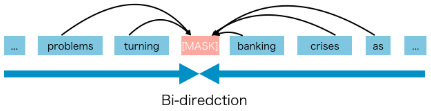
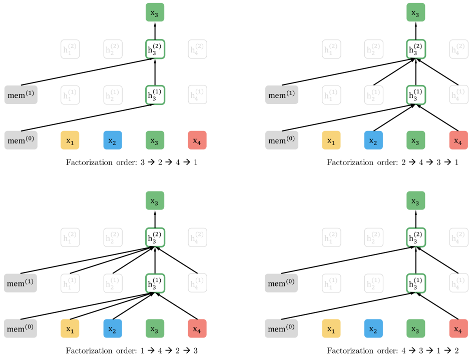


Figure: Aautoencoder language model aims to reconstruct the original data from corrupted input

XLNet: Generalized Autoregressive Pretraining for Language Understanding [10]



Factorization order: 3 → 2 → 4 → 1
Factorization order: 2 → 4 → 3 → 1
Factorization order: 1 → 4 → 2 → 3
Factorization order: 4 → 3 → 1 → 2
mem(0)
mem(1)
x1
x2
x3
x4

Figure: Permutation language model objective

XLNet: Generalized Autoregressive Pretraining for Language Understanding [10]

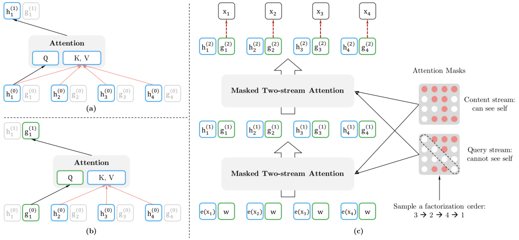
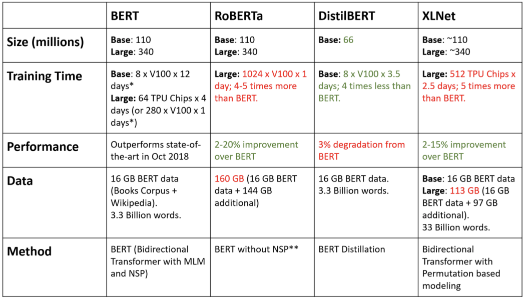

Attention
Q
K, V
(a)
(b)
(c)
Masked Two-stream Attention
Attention Masks
Content stream: can see self
Query stream: cannot see self
Sample a factorization order: 3 → 2 → 4 → 1

Figure: Architecture overview

XLNet: Generalized Autoregressive Pretraining for Language Understanding [10]

- XLNet significantly improves upon BERT on 20 tasks
- It costs $245,000 to train the XLNet model



	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

Image source: towardsdatascience

Today

1 Probing BERT

2 BERT & Family

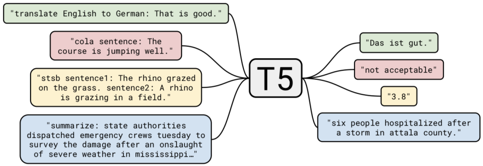
3 BERT & Friends

4 Latest news

T5: Text-To-Text Transfer Transformer [11]

In short:

1. NMT Transfromer architecture
2. new SoTA on GLUE
3. an extensive study of transfer learning techniques (a must read!)


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Reference I

B. Hoover, H. Strobelt, and S. Gehrmann, *Exbert: A visual analysis tool to explore learned representations in transformers models*, 2019. arXiv: 1910.05276 [cs.CL].

G. Jawahar, B. Sagot, D. Seddah, S. Unicomb, G. Iñiguez, M. Karsai, Y. Léo, M. Karsai, C. Sarraute, É. Fleury, *et al.*, "What does bert learn about the structure of language?," in *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*, 2019.

I. Tenney, D. Das, and E. Pavlick, *Bert rediscovers the classical nlp pipeline*, 2019. arXiv: 1905.05950 [cs.CL].

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

Reference II

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*, 2019. arXiv: `1910.01108 [cs.CL]`.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: Pre-trained biomedical language representation model for biomedical text mining," *arXiv preprint arXiv:1901.08746*, 2019.

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

I. Beltagy, K. Lo, and A. Cohan, *Scibert: A pretrained language model for scientific text*, 2019. arXiv: `1903.10676 [cs.CL]`.

Reference III

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, *Xlnet: Generalized autoregressive pretraining for language understanding*, 2019. arXiv: 1906.08237 [cs.CL].

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer*, 2019. arXiv: 1910.10683 [cs.LG].