Deep Learning for Natural Language Processing

Pre-trained Transformer models

UNIVERSITY OF GOTHENBURG

CHALMERS

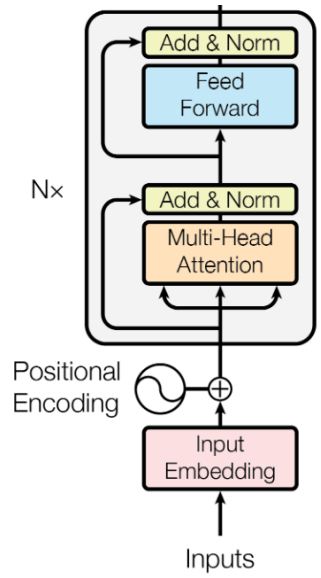
WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Richard Johansson

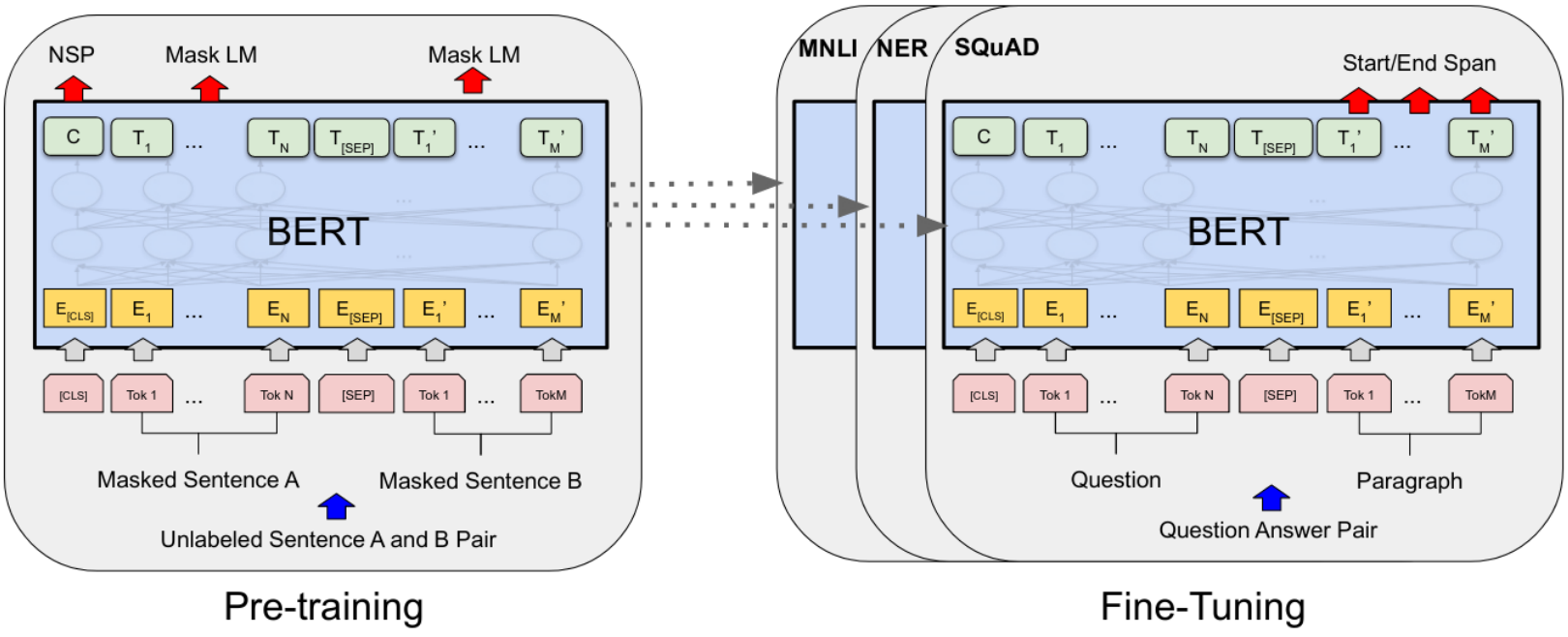
richard.johansson@gu.se

BERT

- **BERT** (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) is an architecture for transfer learning (Devlin et al., 2019)
- pushed the state of the art for several tasks
- uses the encoder part of the Transformer (originally 12 or 24 layers)



transfer learning with BERT



pre-training tasks for BERT (1)

He bought two cans of fish soup .

pre-training tasks for BERT (1)

He	**bought**	**two**	**cans**	**of**	**fish**	**soup**	**.**
He	**[MASK]**	**two**	**cans**	**of**	**fish**	**[MASK]**	**.**

pre-training tasks for BERT (1)

He	bought	two	cans	of	fish	soup	.
He	[MASK]	two	cans	of	fish	[MASK]	.
	⇓					⇓	
	bought					soup	

pre-training tasks for BERT (2)

The man went to the store. He bought a bottle of milk.

$\Downarrow$

`Adjacent`

pre-training tasks for BERT (2)

The man went to the store. He bought a bottle of milk.

$\Downarrow$

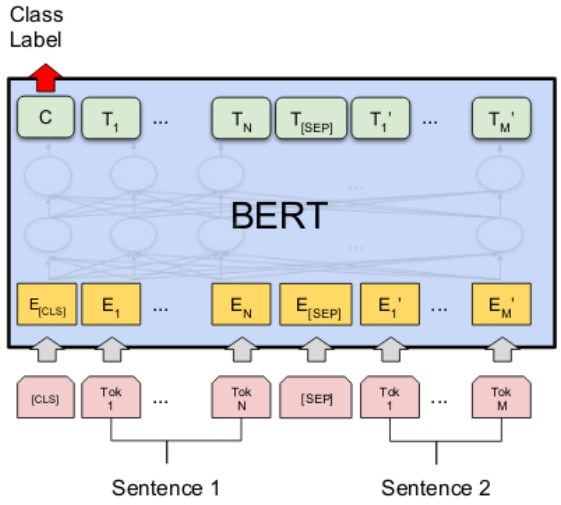
`Adjacent`

The man went to the store. Penguins are flightless birds.

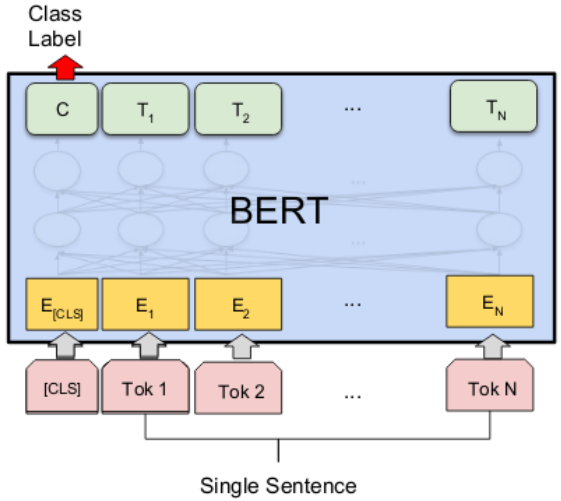
$\Downarrow$

`NonAdjacent`

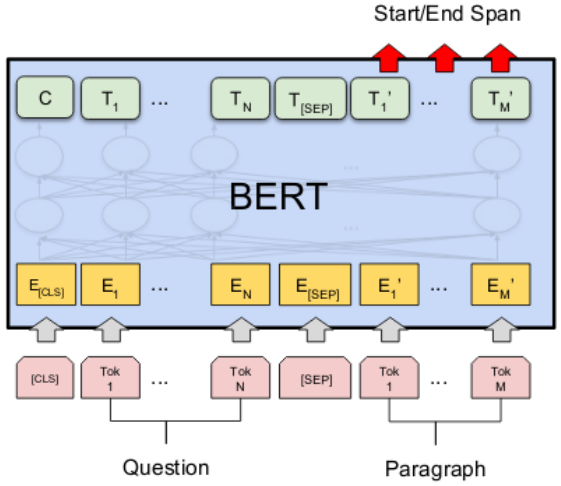
how to use BERT in different types of tasks



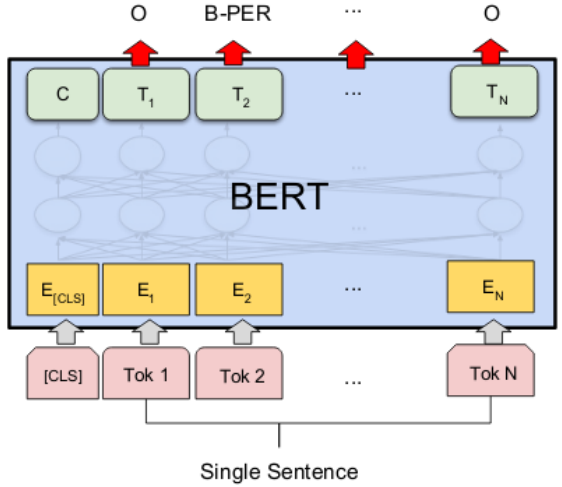
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

working with BERT in PyTorch

- the `transformers` library implements several types of pre-trained Transformer-based models (BERT and derivatives)
  - `https://github.com/huggingface/transformers`
  - needs to be installed separately
- it implements classes for the standard use cases
  - such as `BertForSequenceClassification`: just a linear layer on top of the final Transformer layer

tokenization in BERT

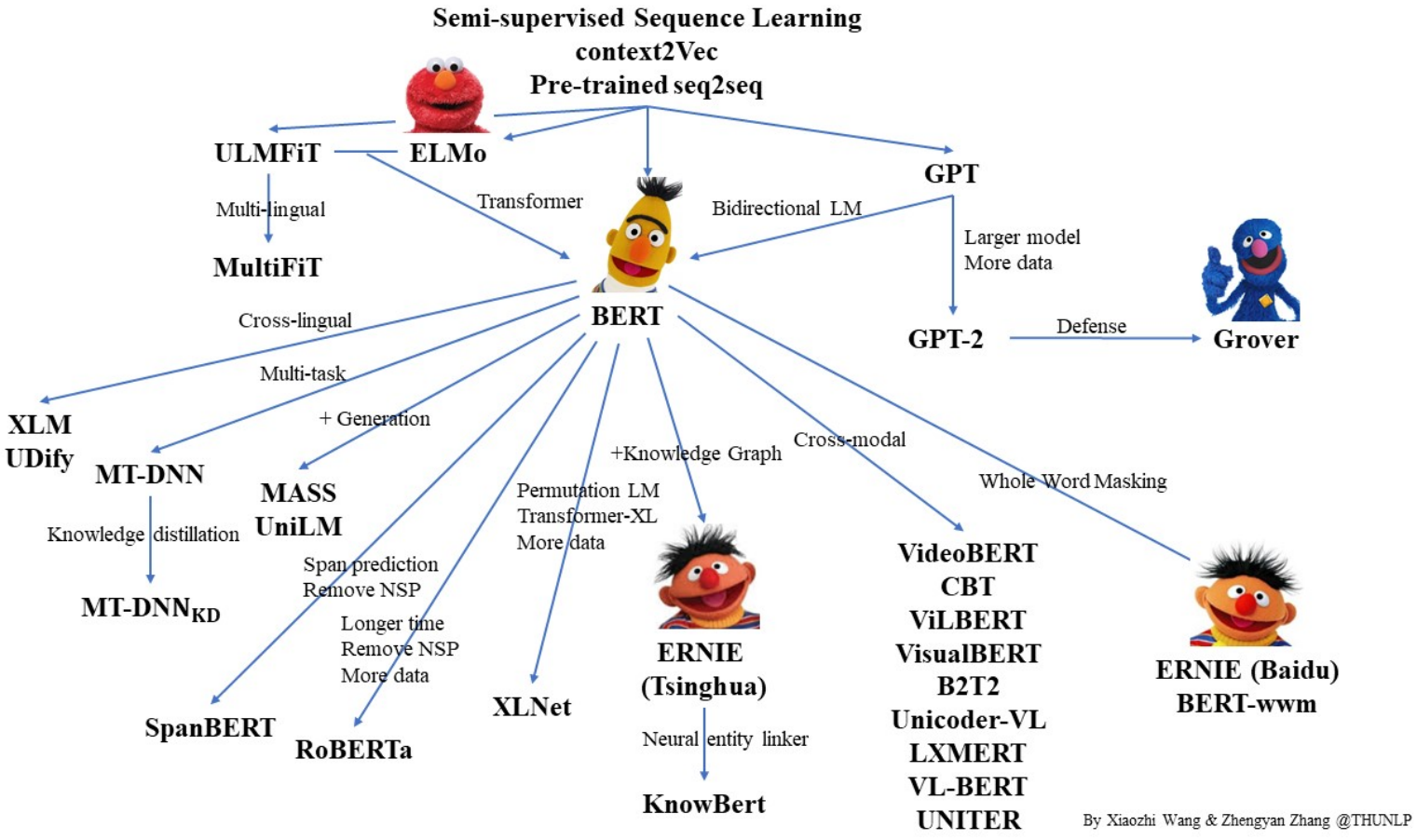
- BERT comes with a built-in tokenizer
- it uses **WordPiece** tokenization to avoid out-of-vocabulary situations
- BERT can handle documents of up to 512 WordPiece tokens

```
tokenizer.tokenize('Rolf lives in Gothenburg.')
```

```
['rolf', 'lives', 'in', 'gothenburg', '.']
```

```
tokenizer.tokenize('Margareta lives in Jonsered.')
```

```
['margaret', '##a', 'lives', 'in', 'jon', '##ser', '##ed', '.']
```



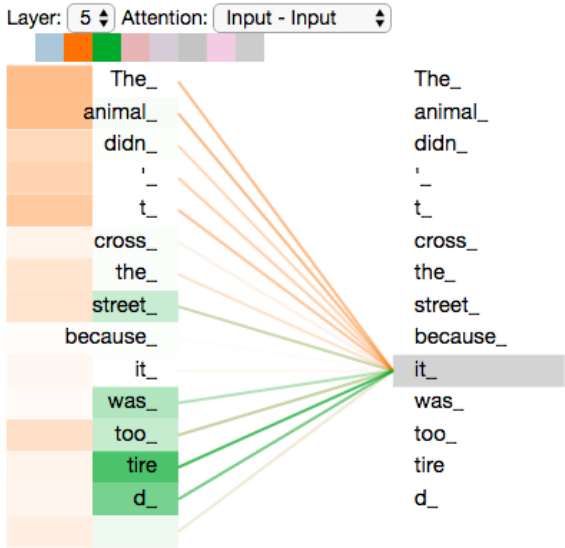
https://github.com/thunlp/PLMpapers/

improved and specialized BERT variants (small sample)

- **RoBERTa** (Liu et al., 2019) uses more robust optimization
- **DistilBERT** (Sanh et al., 2019) "distils" BERT into a slightly less enormous model
- domain-specific BERT models:
  - BioBERT `https://github.com/dmis-lab/biobert`
  - SciBERT `https://github.com/allenai/scibert`
- Swedish BERT by the Royal Library: `https://huggingface.co/KB/bert-base-swedish-cased`
- Multilingual BERT: `https://huggingface.co/bert-base-multilingual-cased`

introduction to BERTology

- there is an increasing interest in trying to understand complex models such as BERT
- several recent papers try to interpret parts of BERT's Transformer layers
  - *"BERT Rediscovers the Classical NLP Pipeline"* (Tenney et al., 2019)
  - *"What Does BERT Look At? An Analysis of BERT's Attention"* (Clark et al., 2019)



- the **BlackboxNLP** workshop publishes many interesting contributions: https://blackboxnlp.github.io/

code example

- in a notebook, we'll see how to use BERT as the encoder for classifying reviews

★☆☆☆☆ **Just plain lame.**, August 14, 2007

By **Gary Smith "Editor, Handgun Hunter Magazine"** (Texas) - See all my reviews

This review is from: Garden & Gun (Magazine)

This magazine has a catchy title and very nice graphics and photography. What the premier issue lacks is anything of any substance about guns or hunting. I wonder if they actually read their own title. In my opinion these guys are nothing more than posers from the guns/hunting standpoint and many of the photographs appear to be staged. In particular, there are a couple pictures of a woman shooting a bow and arrow. Not only is she showing extremely poor form she's using the equipment shown in the photographs incorrectly. This is tantamount to using spinning gear with the reel positioned over the top of the fishing pole. If they want to cover hunting they should at least hire a photo editor that knows what (s)he's looking at. If you want a hunting magazine buy something else…

reading

- tutorials, overviews:
  - Alammar: *"The illustrated BERT, ELMo"*
  - **NAACL 2019 tutorial** on transfer learning in NLP
- the BERT paper (Devlin et al., 2019) is also readable

references

K. Clark, U. Khandelwal, O. Levy, and C. Manning. 2019. What does BERT look at? an analysis of BERT's attention. arXiv:1906.04341.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

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

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

I. Tenney, D. Das, and E. Pavlick. 2019. BERT rediscovers the classical NLP pipeline. arXiv:1905.05950.