University of Ljubljana, Faculty of Computer and Information Science

BERT and T5 models

Prof Dr Marko Robnik-Šikonja

Natural Language Processing, Edition 2024

Contents

- BERT models
- Cross-lingual transfer
- T5 models

- some slides by Jay Alammar, Jacob Devlin and Andrej Miščič

BERT

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- State-of-the-art pretrained LM based on transformer architecture (only the encoder part)
- Idea:
  - use large unlabeled corpora and an auxiliary task to pretrain a model for a general language representation
  - fine-tune the model on a (possibly small) dataset for a specific downstream task (typically classification)

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, pp. 4171-4186.

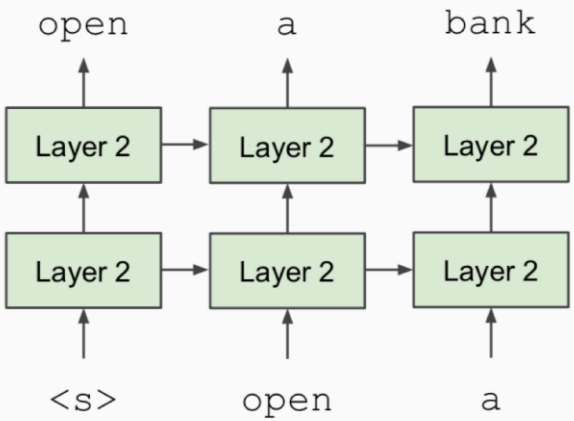
BERT: motivation 1/3

- **Problem**: Classical language models only use the left or right context, but language understanding is bidirectional.
- Why are LMs unidirectional?
  - Reason 1: Directionality is needed to generate a well-formed probability distribution.
    - We don’t care about this.
  - Reason 2: Words can “see themselves” in a bidirectional encoder.

BERT: motivation 2/3

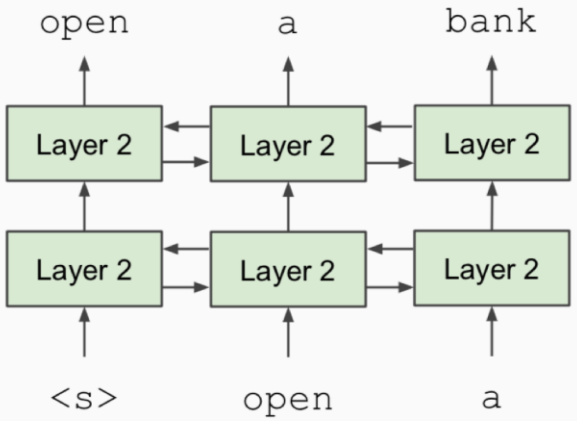
Unidirectional context

Build representation incrementally



Bidirectional context

Words can “see themselves”



BERT: motivation 3/3

- **Solution**: Mask out *k*% of the input words, and then predict the masked words
- BERT uses *k* = 15%

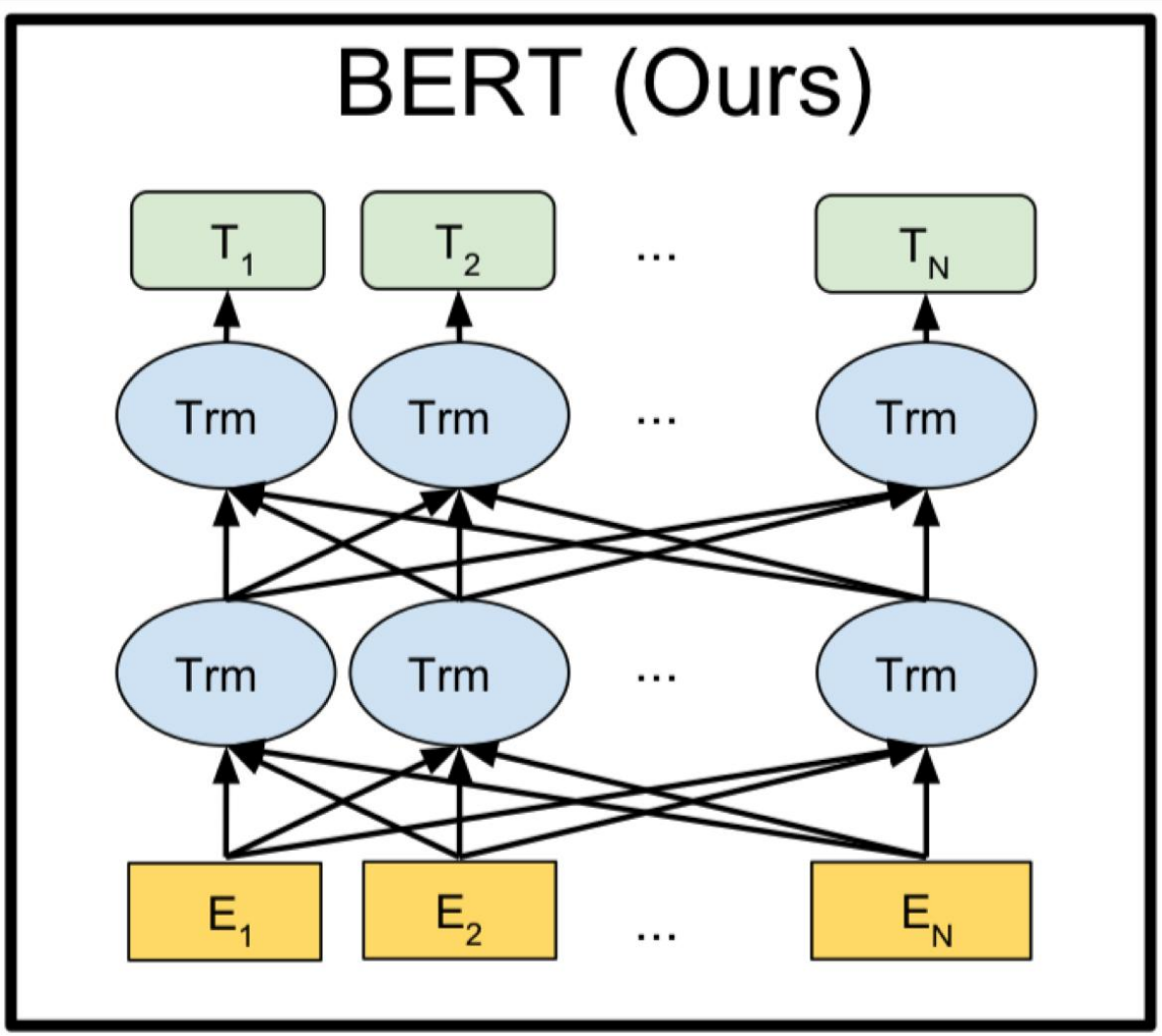
store gallon

↑ ↑

the man went to the [MASK] to buy a [MASK] of milk

- Too little masking: Too expensive to train (not enough masks)
- Too much masking: Not enough context

BERT architecture



BERT uses several tasks

- besides masked LM, BERT learns relationships between sentences
- predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

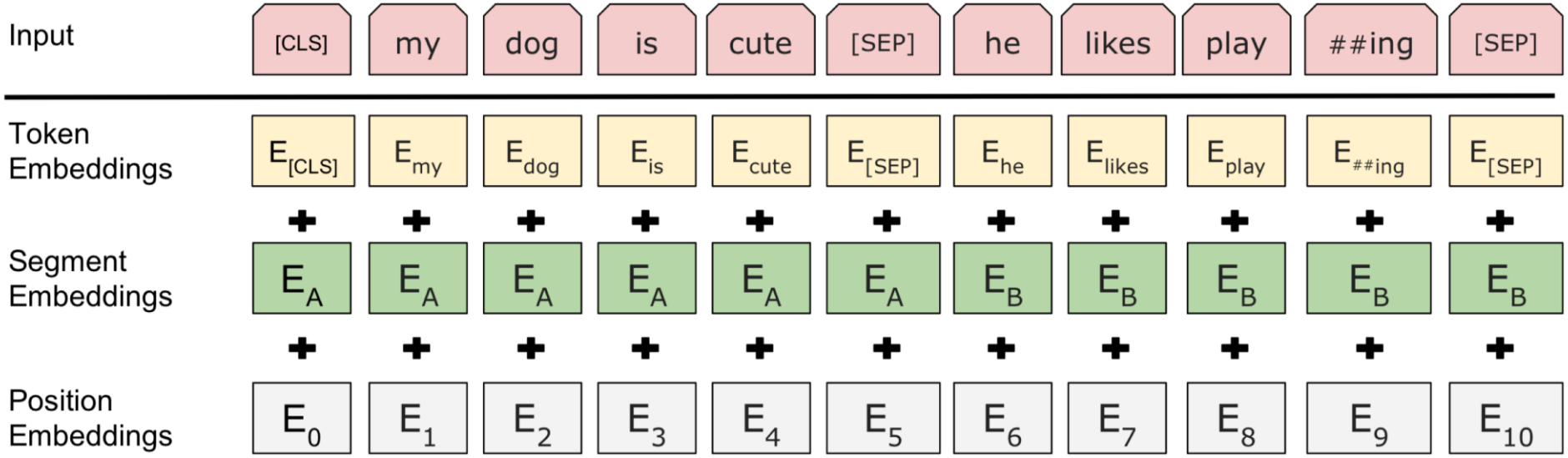
Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

- some follow-up BERT-like models, e.g., RoBERTa, drop this task and claim better performance on downstream tasks

Subword input encoding for BERT

- Token embeddings are word pieces (sub-word encoding called sentence-pair encoding, a variant of BPE)
- (Relatively) common words are in the vocabulary: *at, fairfax, 1910s*
- Other words are built from word-pieces: *hypatia = h ##yp ##ati ##a*
- Learned segmented embeddings represent each sentence
- Positional embedding is the same as for other transformer architectures

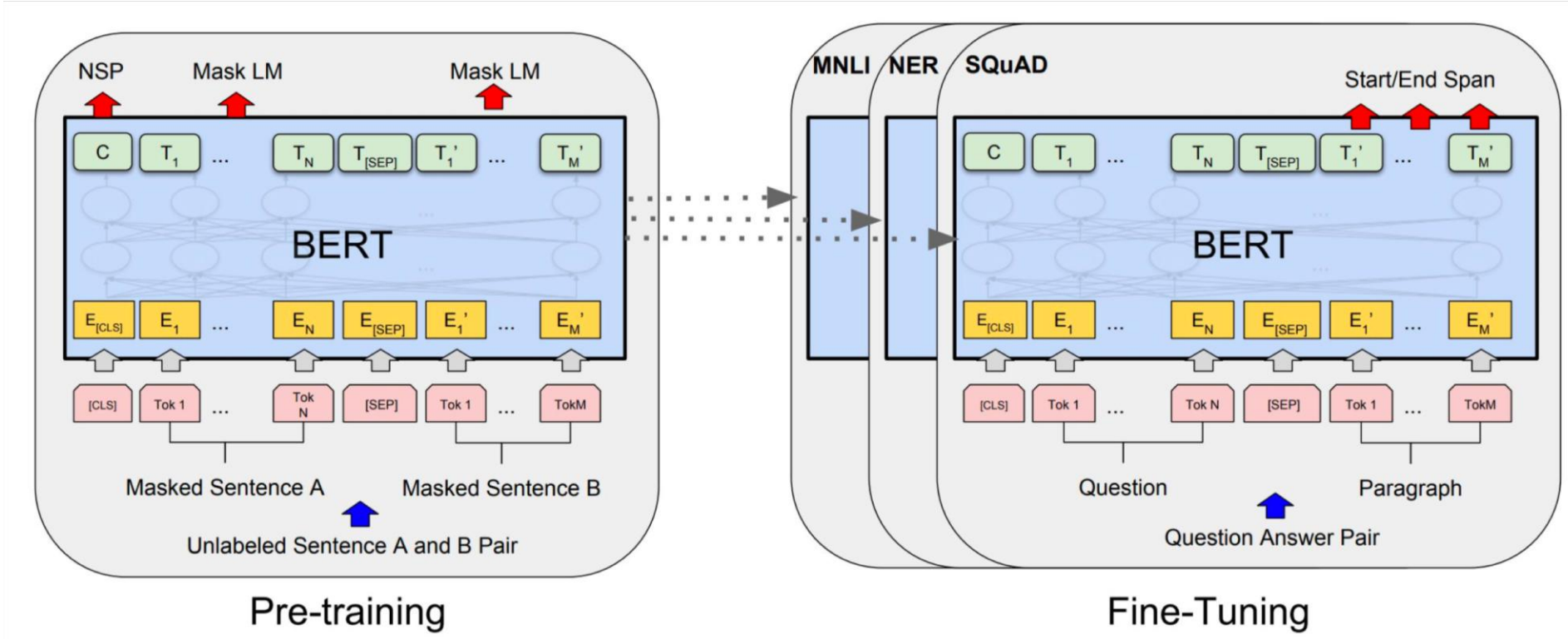


BERT training

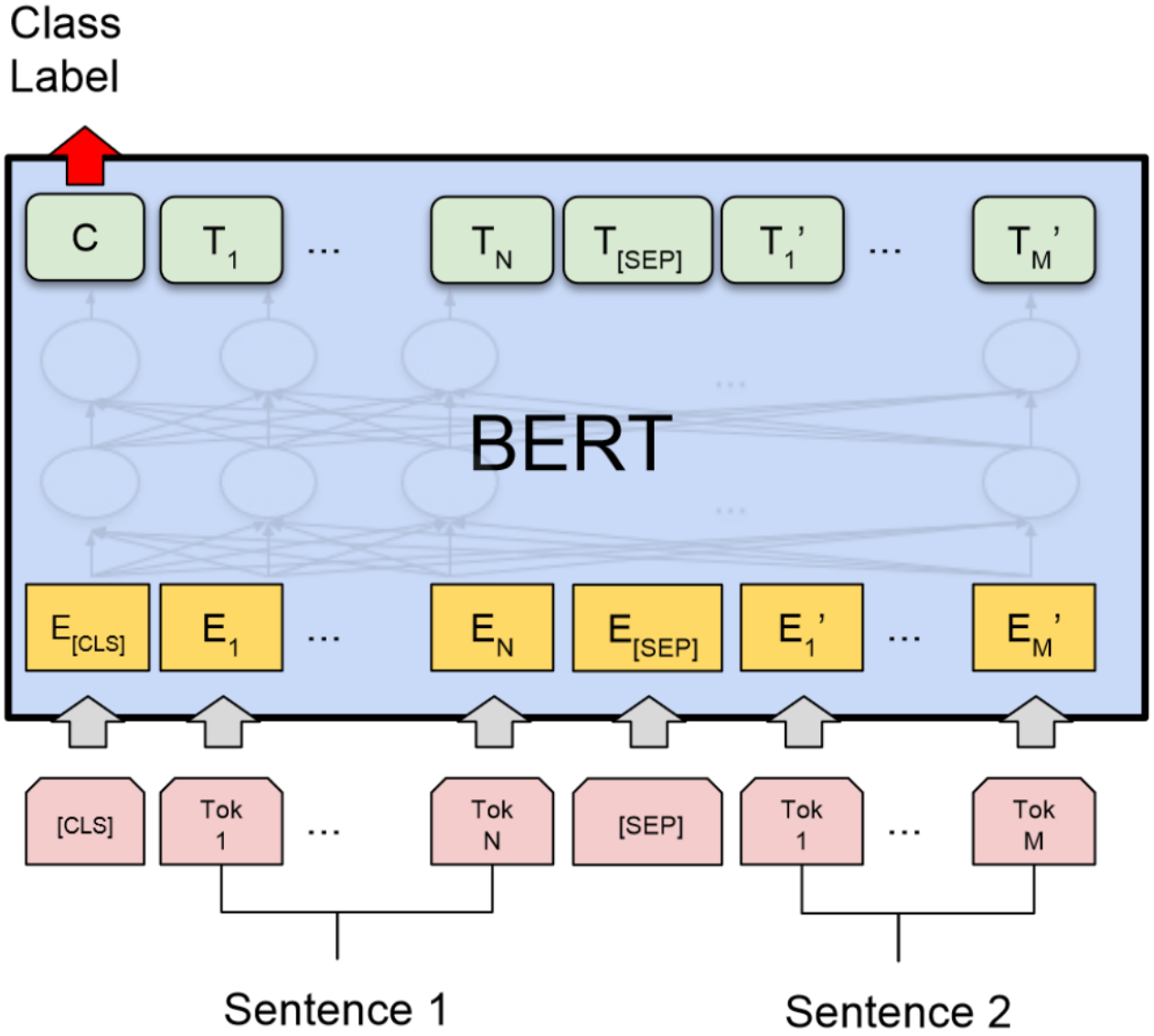
- Transformer encoder
- Self-attention ⇒ no locality bias
- Long-distance context has “equal opportunity”
- Single multiplication per layer ⇒ efficiency on GPU/TPU
- Trained on Wikipedia + BookCorpus
- English BERT was trained with 2 model sizes:
  - BERT-Base: 12-layer, 768-hidden parameters, 12-head, 110M parameters
  - BERT-Large: 24-layer, 1024-hidden parameters, 16-head, 340M parameters
- Trained on 4x4 or 8x8 TPU slice for 4 days

Use of BERT

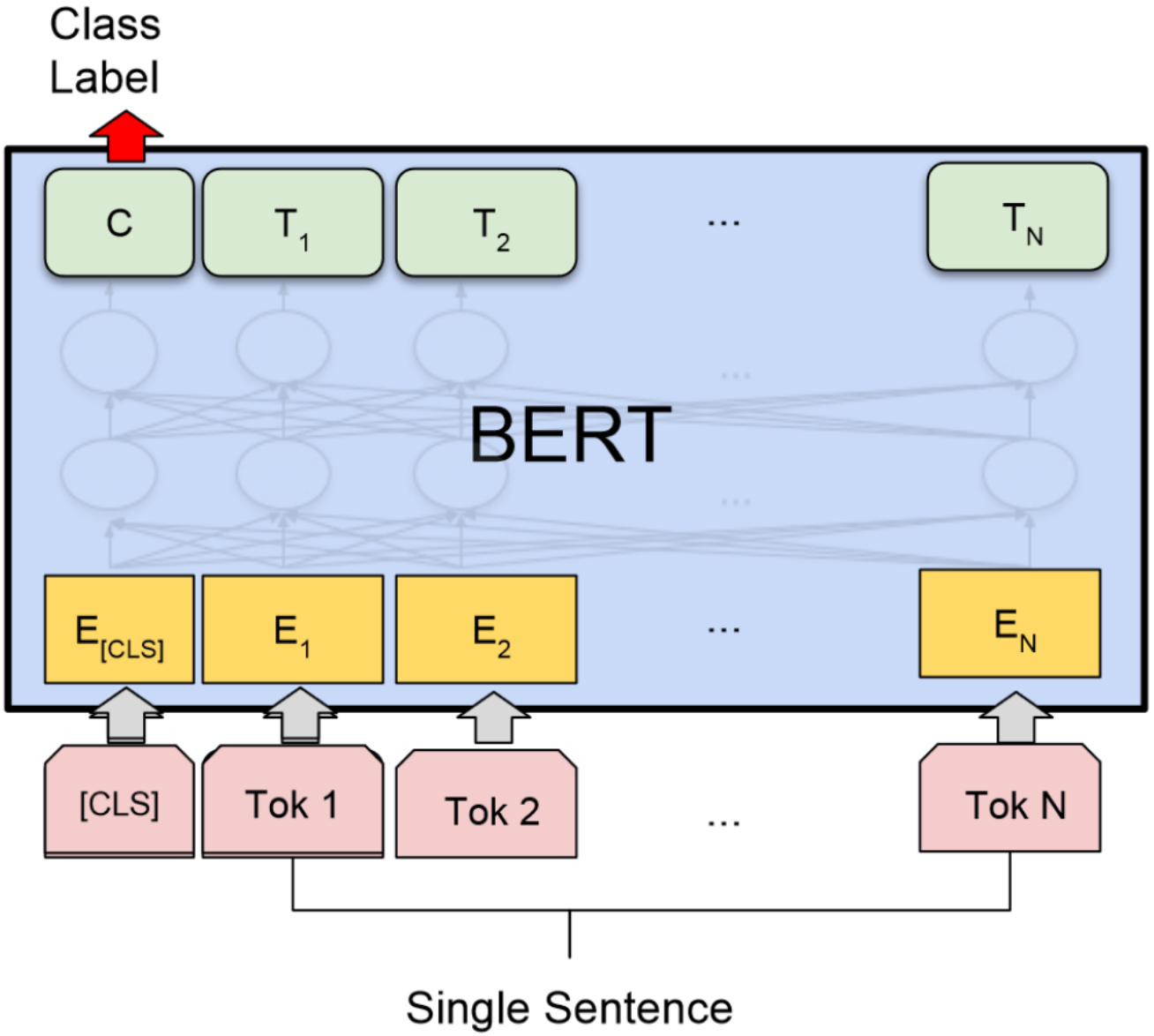
- train a classifier built on the top layer for each task that you fine-tune for, e.g., Q&A, NER, inference
- achieved state-of-the-art results for many tasks
- GLUE and SuperGLUE tasks for natural language understanding



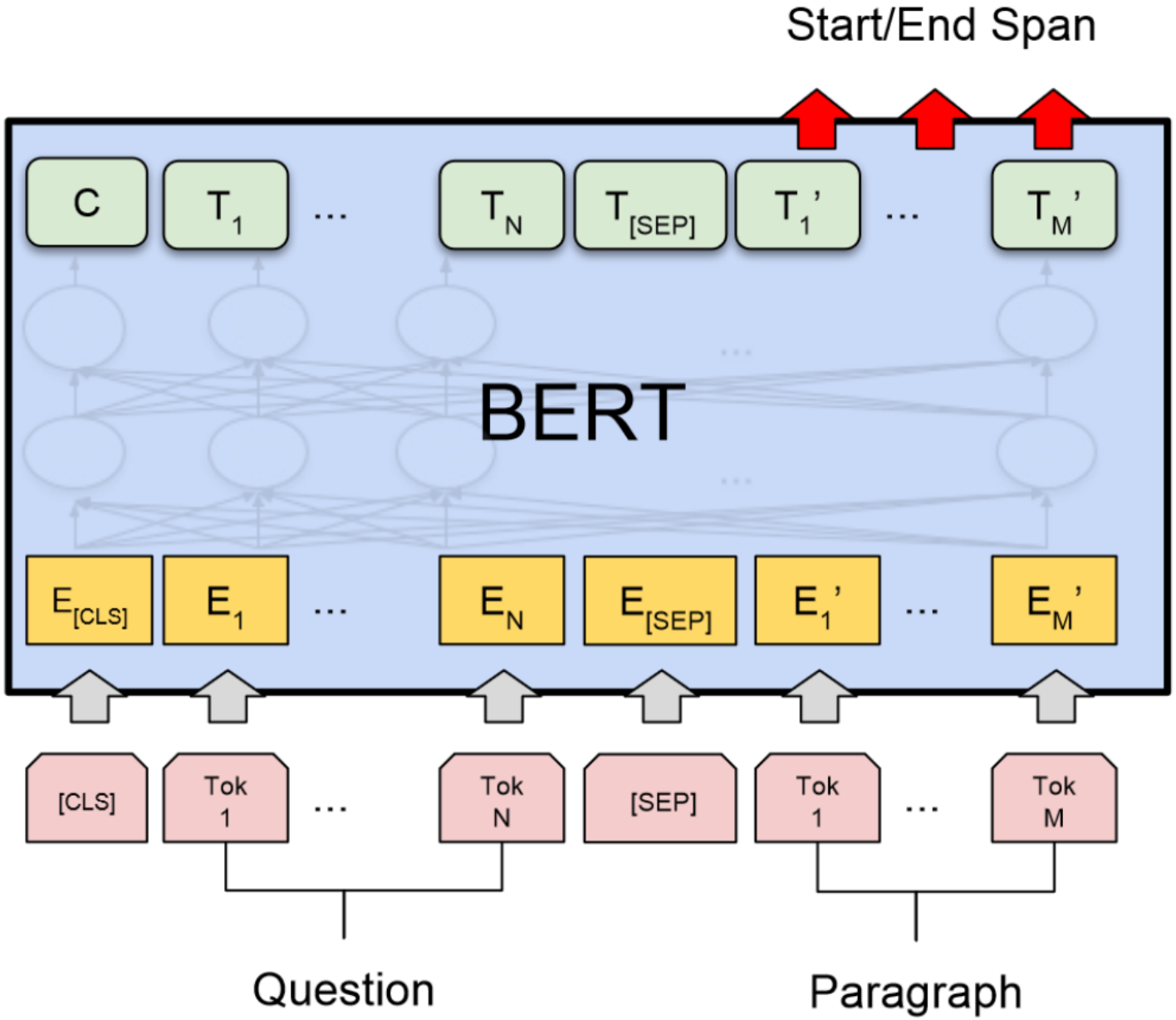
Two sentence classification using BERT- e.g., natural language inference



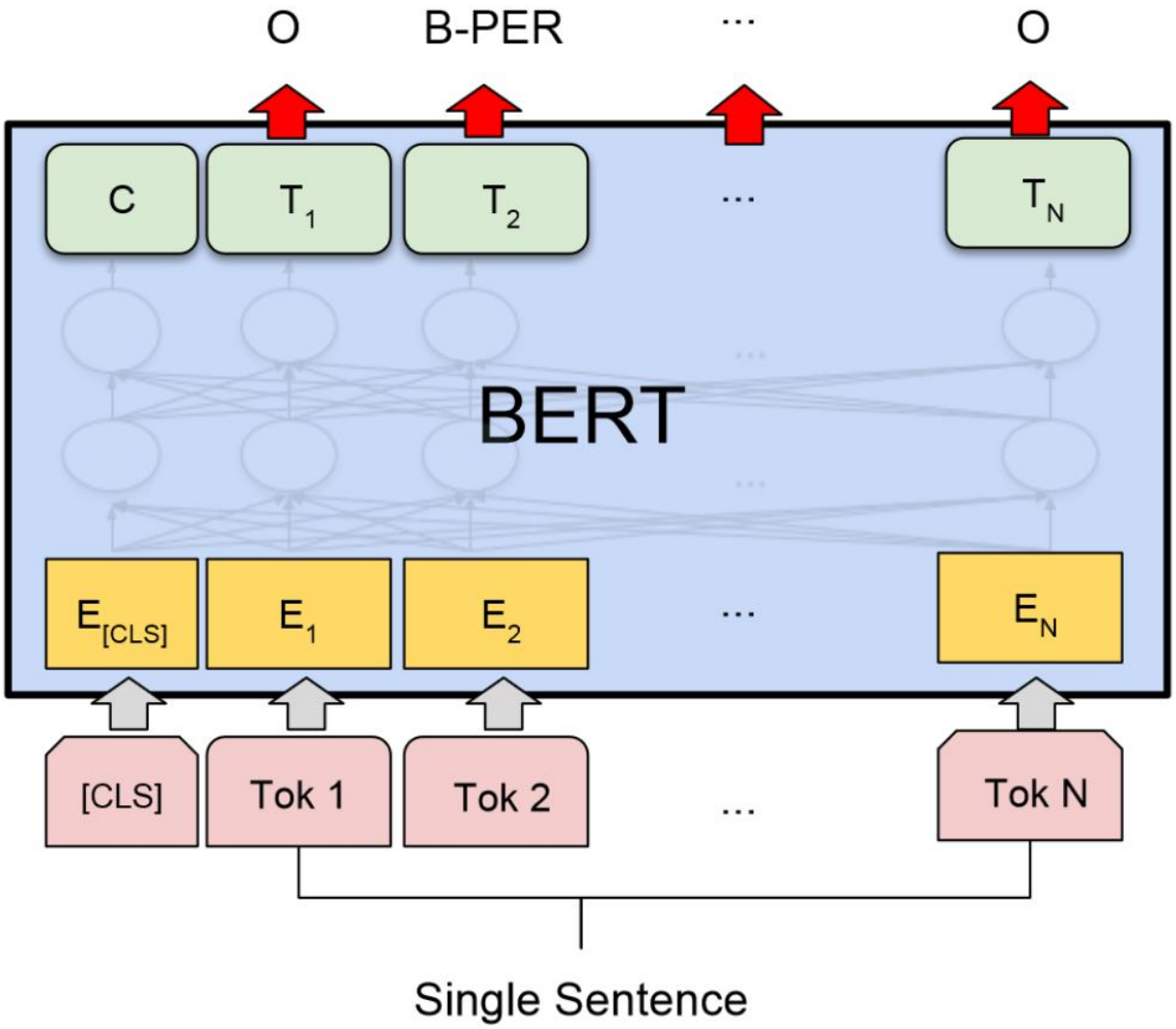
Sentence classification using BERT – sentiment, grammatical correctness



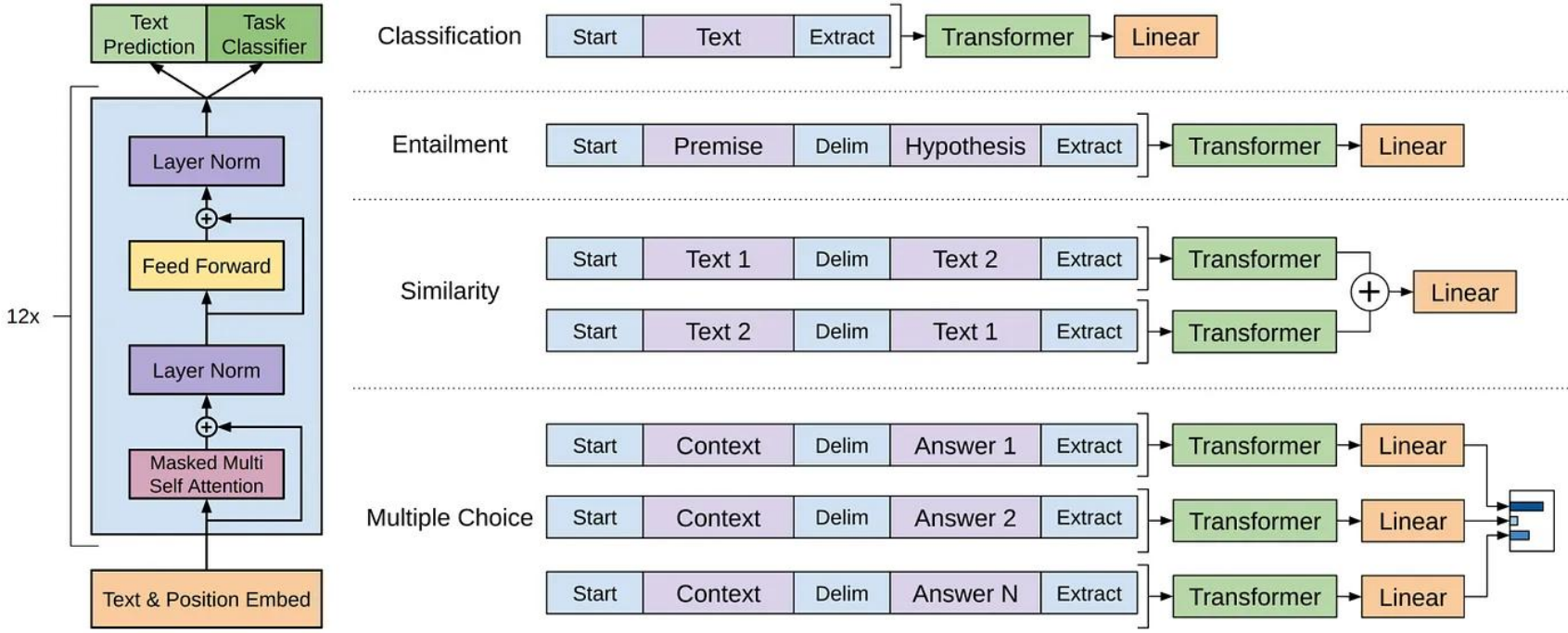
Questions and answers with BERT



Sentence tagging with BERT-NER, POS tagging, SRL



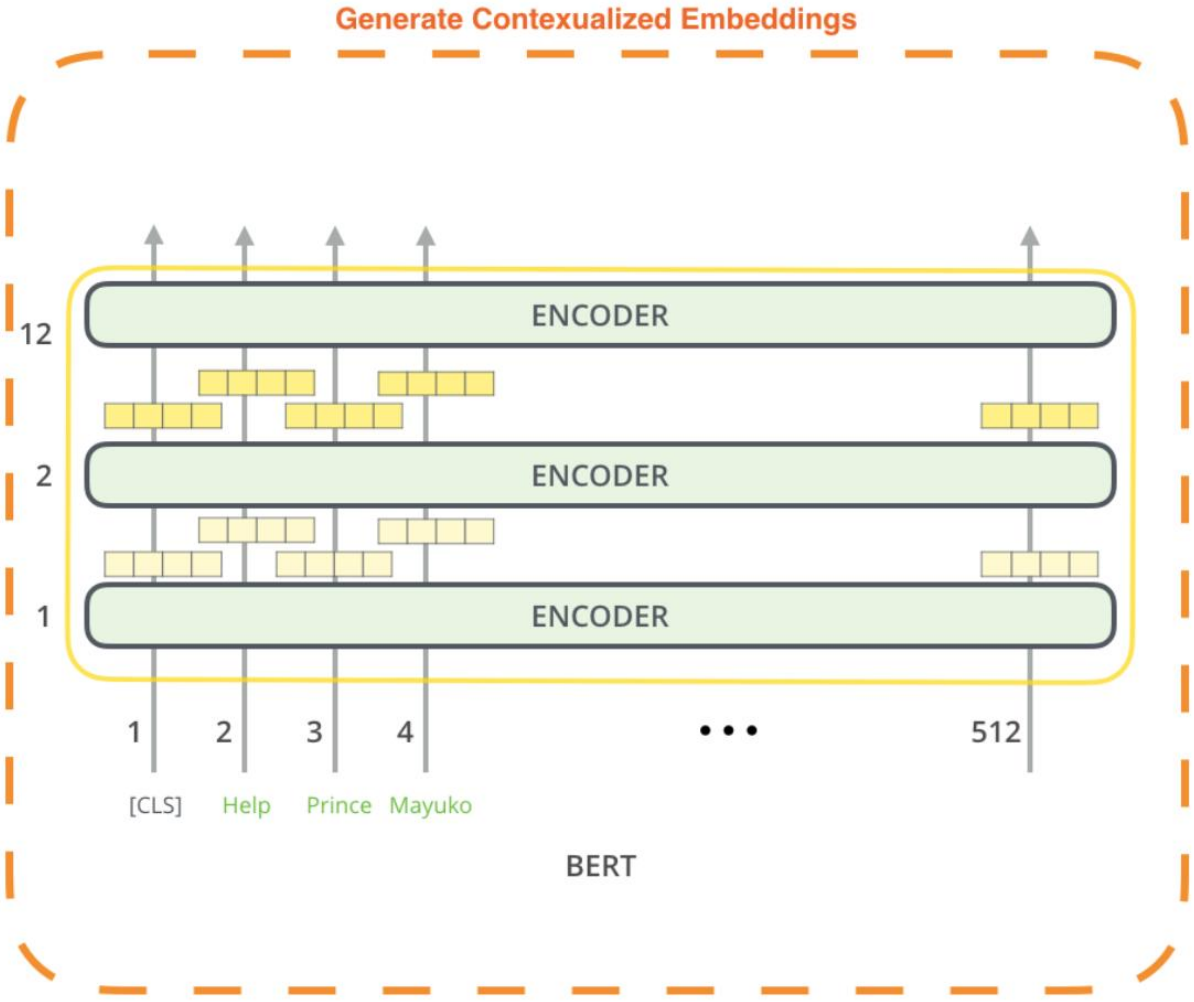
Finetuning


Text Prediction
Task Classifier
Layer Norm
Feed Forward
Layer Norm
Masked Multi Self Attention
12x
Text & Position Embed
Classification
Start
Text
Extract
Transformer
Linear
Entailment
Start
Premise
Delim
Hypothesis
Extract
Transformer
Linear
Similarity
Start
Text 1
Delim
Text 2
Extract
Start
Text 2
Delim
Text 1
Extract
Transformer
Transformer
Linear
Multiple Choice
Start
Context
Delim
Answer 1
Extract
Start
Context
Delim
Answer 2
Extract
Start
Context
Delim
Answer N
Extract
Transformer
Linear

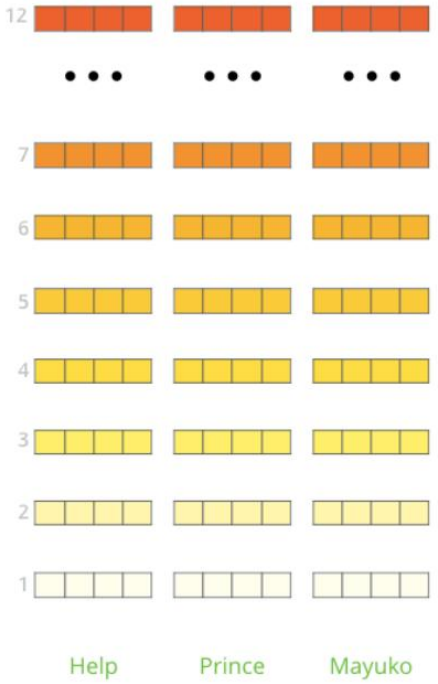
BERT can produce embeddings

- one can extract fixed size contextual vectors from BERT, achieving slightly lower accuracy than using the whole BERT as the first stage model

Layer-wise embeddings



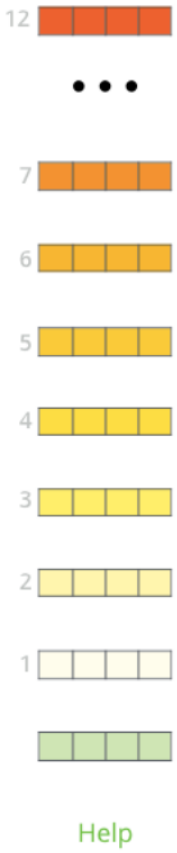
The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

Which layer of BERT to use as embeddings?

What is the best contextualized embedding for "Help" in that context?
For named-entity recognition task CoNLL-2003 NER



		Dev F1 Score
First Layer	Embedding	91.0
Last Hidden Layer	12	94.9
Sum All 12 Layers	12 + ... + 2 + 1 =	95.5
Second-to-Last Hidden Layer	11	95.6
Sum Last Four Hidden	12 + 11 + 10 + 9 =	95.9
Concat Last Four Hidden	9 10 11 12	96.1

Examples of GLUE tasks

- GLUE benchmark is dominated by natural language inference tasks, but also has sentence similarity and sentiment

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.

Hypothesis: Jainism hates nature.

Label: Contradiction

CoLA (Corpus of Linguistic Acceptability)

Sentence: The wagon rumbled down the road. Label: Acceptable

Sentence: The car honked down the road. Label: Unacceptable

SuperGLUE tasks

BoolQ - Boolean Questions
CB – Commitment Bank
COPA - Choice of Plausible Alternatives
MultiRC - Multi-Sentence Reading Comprehension
ReCoRD - Reading Comprehension with Commonsense Reasoning Dataset
RTE - Recognizing Textual Entailment
WiC - Word-in-Context
WSC - Winograd Schema Challeng

Table 2: Development set examples from the tasks in SuperGLUE. **Bold** text represents part of the example format for each task. Text in *italics* is part of the model input. *Underlined* text is specially marked in the input. Text in a `monospaced font` represents the expected model output.

BoolQ	**Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.* **Question:** *is barq's root beer a pepsi product* **Answer:** `No`
CB	**Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?* **Hypothesis:** *they are setting a trend* **Entailment:** `Unknown`
COPA	**Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?* **Alternative 1:** *The sun was rising.* **Alternative 2:** *The grass was cut.* **Correct Alternative:** `1`

MultiRC

Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*
Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (`T`), *No* (`F`), *Yes* (`T`), *No, she didn't recover* (`F`), *Yes, she was at Susan's party* (`T`)

ReCoRD

Paragraph: *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electorcal Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood*
Query For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** `US`

RTE

Text: *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*
Hypothesis: *Christopher Reeve had an accident.* **Entailment:** `False`

WiC

Context 1: *Room and board.* **Context 2:** *He nailed boards across the windows.*
Sense match: `False`

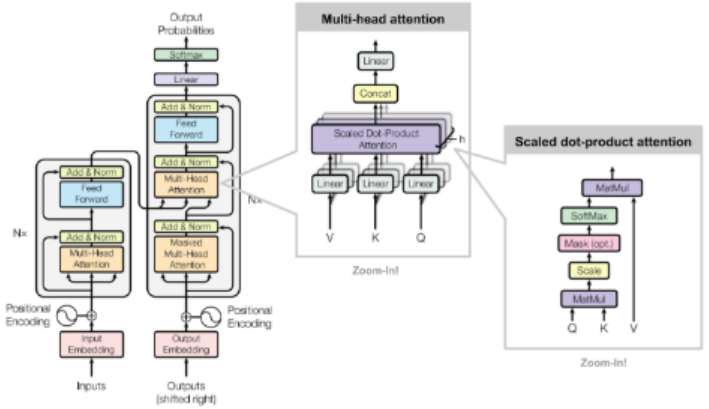
WSC

Text: *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** `False`

University of Ljubljana
Faculty of Computer and Information Science

Pretrained word representations

- large pretrained neural language models
- trained on large text corpora to capture relations in language
- finetuned to specific tasks
- many publicly available
- for Slovene:
  - fastText, ELMo,
  - SloBERTa, CroSloEngual BERT, SlEng BERT,
  - SloT5, Slo GPT
- for Croatian: fastText, ELMo, BERTić, CroSloEngual BERT
- on Clarin.si and HuggingFace
- hundreds of papers investigating BERT-like models in major ML & NLP conferences



- Ulčar, M., & Robnik-Šikonja, M. (2020). High Quality ELMo Embeddings for Seven Less-Resourced Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 4731-4738).
- Ulčar, M. and Robnik-Šikonja, M., 2021. SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of SI-KDD within the Information Society 2021*, pp.17-20.
- Ljubešić, N., & Lauc, D. (2021). BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (pp. 37-42).
- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue* (pp. 104-111).
- Ulčar, M. & Robnik-Šikonja, M. (2023) Sequence-to-sequence pretraining for a less-resourced Slovenian language.
- Frontiers in Artificial Intelligence, Section on Language and Computation, Volume 6 – 2023, https://doi.org/10.3389/frai.2023.932519

SloBERTa

- Currently the best Slovene BERT-like LLM
- Many thousands of downloads from HuggingFace
- Training set: 3.41 B words (corpora Gigafida, KAS, partially Janes, siParl, slWaC)
- Training duration: 4 weeks on Nvidia DGX A100 using 4xGPU
- An example of direct use:
  - <mask> je najlepše mesto na svetu.
  - Odgovori: Ljubljana, Barcelona, London, Madrid, To

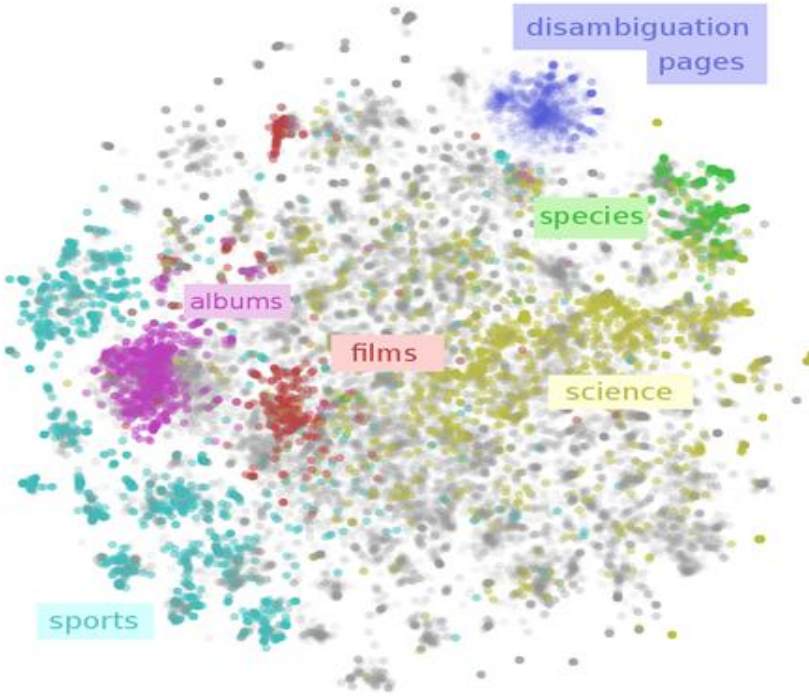
ULČAR, Matej, ROBNIK ŠIKONJA, Marko. (2021) SloBERTa: Slovene monolingual large pretrained masked language model. In Proceedings of Information Science - *Data Mining and Data Warehouses - SiKDD* 2021, pp. 17-20

Classical cross-lingual transfer

Explicit alignment of vector spaces

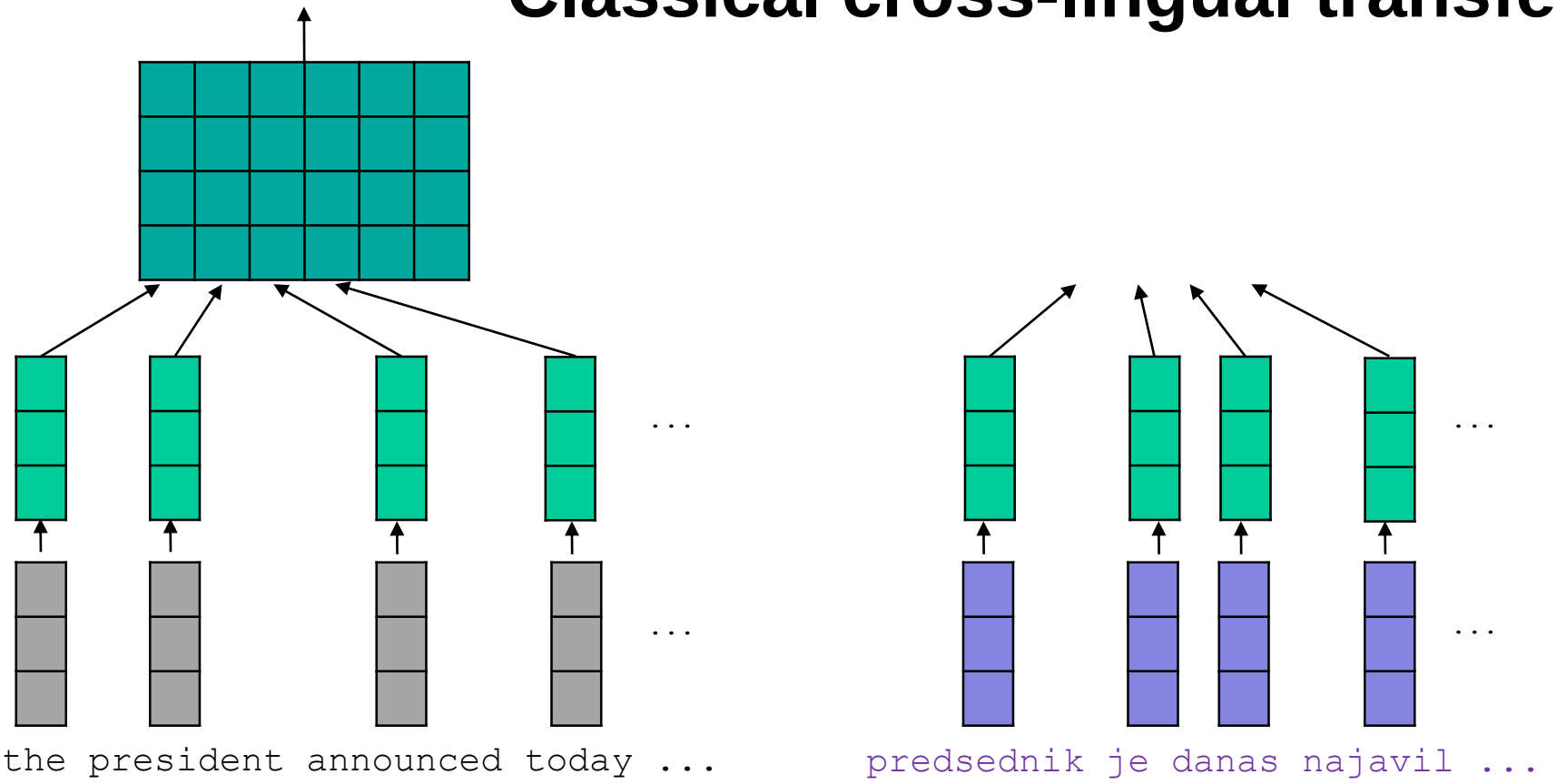
WS ≈ E

Nowadays: use multilingual LLMs directly



Ulčar, M. and Robnik-Šikonja, M., 2022. Cross-lingual alignments of ELMo contextual embeddings. *Neural Computing and Applications*, *34*(15), pp.13043-13061.

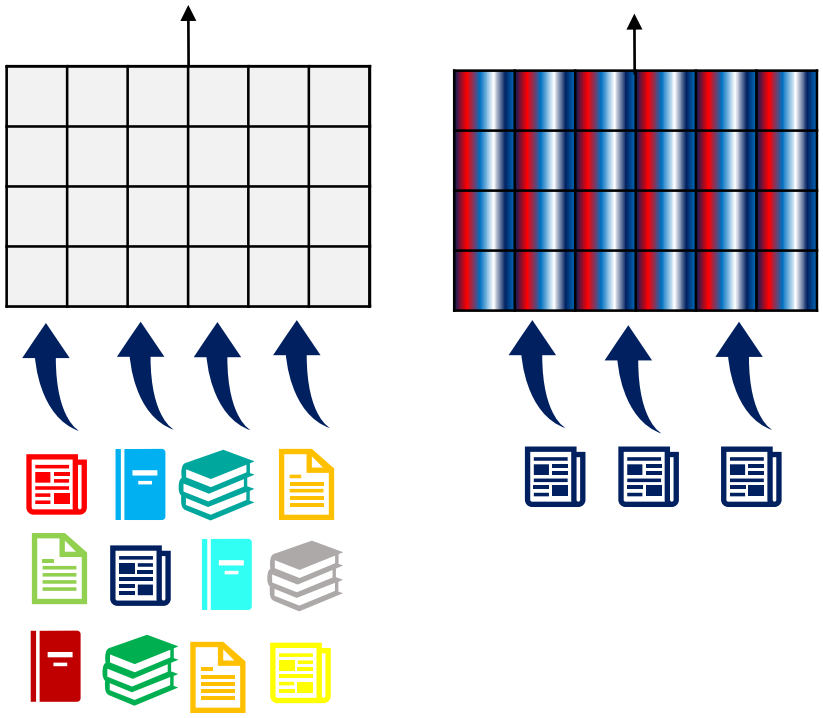
Classical cross-lingual transfer



Multilingual representational LLMs

- Pretrained on multiple languages simultaneously
- multilingual BERT supports 104 languages by training on Wikipedia
- XLM-R was trained on 2.5 TB of texts from 100 languages
- these models allow cross-lingual transfer
- solve the problem of insufficient training resources for less-resourced languages
- zero-shot transfer and few-shot transfer

Using multilingual models



predsednik je danas najavil ...

Pretraining Fine-tuning Classification

Zero-shot transfer and few-shot transfer

Why not only multilingual?

- performance on many tasks drops with more languages
- results for a few tasks in Slovene (Named Entity Recognition – NER,Part-of-Speech Tagging – POS, Dependency Parsing – DP, Sentiment Analysis – SA, Word Analogy – WA)

Model	NER	POS	DP	SA	WA
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
SloBERTa	**0.933**	**0.991**	**0.844**	**0.623**	**0.405**

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., & Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *arXiv preprint arXiv:2107.10614*.

University of Ljubljana
Faculty of Computer and Information Science

Vocabulars in LLMs

- Tokenization depends on the vocabular
- The dictionary is constructed statistically (typically BPE or a variant)

- Sentence: “Letenje je bilo predmet precej starodavnih zgodb.”
- SloBERTa:

'▁Le', 'tenje', '▁je', '▁bilo', '▁predmet', '▁precej', '▁staroda', 'vnih', '▁zgodb', '.’

- mBERT:

'Let', '##en', '##je', 'je', 'bilo', 'pred', '##met', 'pre', '##cej', 'star', '##oda', '##vnih', 'z', '##go', '##d', '##b', '.'

Ulčar, M., & Robnik-Šikonja, M. (2021) Training dataset and dictionary sizes matter in BERT models: The case of Baltic languages. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 162-172)

Trade off: trilingual models

- BERT trained with only a few languages
- more data for training
- more specific dictionary
- good for cross-lingual transfer
- Trilingual models
  - CroSloEngual BERT
  - FinEst BERT
  - LitLat BERT
- SlavBERT (ru, pl, cs, bg; DeepPavlov)

Model	NER	POS	DP	SA	WA
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
CSE-BERT	0.928	0.990	**0.854**	0.610	0.195
SloBERTa	**0.933**	**0.991**	0.844	**0.623**	**0.405**

- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In International Conference on Text, Speech, and Dialogue (pp. 104-111).

Cross-lingual transfer in classification

- Excellent cross-lingual transfer between similar languages like Slovene and Croatian
- But: the transfer quality is problem-dependent

sentiment analysis

		LASER		mBERT		CSE BERT		Both target	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Croatian	Slovene	0.53	0.53	0.53	0.54	**0.61**	**0.60**	0.60	0.60
Croatian	English	**0.63**	0.63	**0.63**	**0.66**	0.62	0.64	0.62	0.65
English	Slovene	0.54	0.57	0.50	0.53	**0.59**	**0.57**	0.60	0.60
English	Croatian	0.62	**0.67**	0.67	0.63	**0.73**	**0.67**	0.73	0.68
Slovene	English	0.63	0.64	**0.65**	**0.67**	0.63	0.64	0.62	0.65
Slovene	Croatian	0.70	0.65	0.64	0.63	**0.73**	**0.69**	0.73	0.68
Croatian English	Slovene	0.54	0.54	0.53	0.54	**0.60**	**0.58**	0.60	0.60
Croatian Slovene	English	0.62	0.61	**0.65**	**0.67**	0.63	0.65	0.62	0.65
English Slovene	Croatian	0.64	0.68	0.63	0.63	**0.68**	**0.70**	0.73	0.68
Average performance gap		0.04	0.03	0.04	0.03	0.00	0.01		

idiom detection

Language	Slovene ELMo	mBERT	Default F_1
Slovene	0.8163	0.8359	0.667
Croatian	0.9191	0.8970	0.667
Polish	0.2863	0.6987	0.667

- Robnik-Šikonja, M., Reba, K., & Mozetič, I. (2021). Cross-lingual transfer of sentiment classifiers. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, *9*(1), 1-25.
- Škvorc, T., Gantar, P., & Robnik-Šikonja, M. (2022). MICE: mining idioms with contextual embeddings. *Knowledge-Based Systems*, *235*, 107606.

What LLMs learn?

- We would like to travel to [MASK], ki je najlepši otok v Mediteranu.

SloBERTa: ..., Slovenija, I, Koper, Slovenia

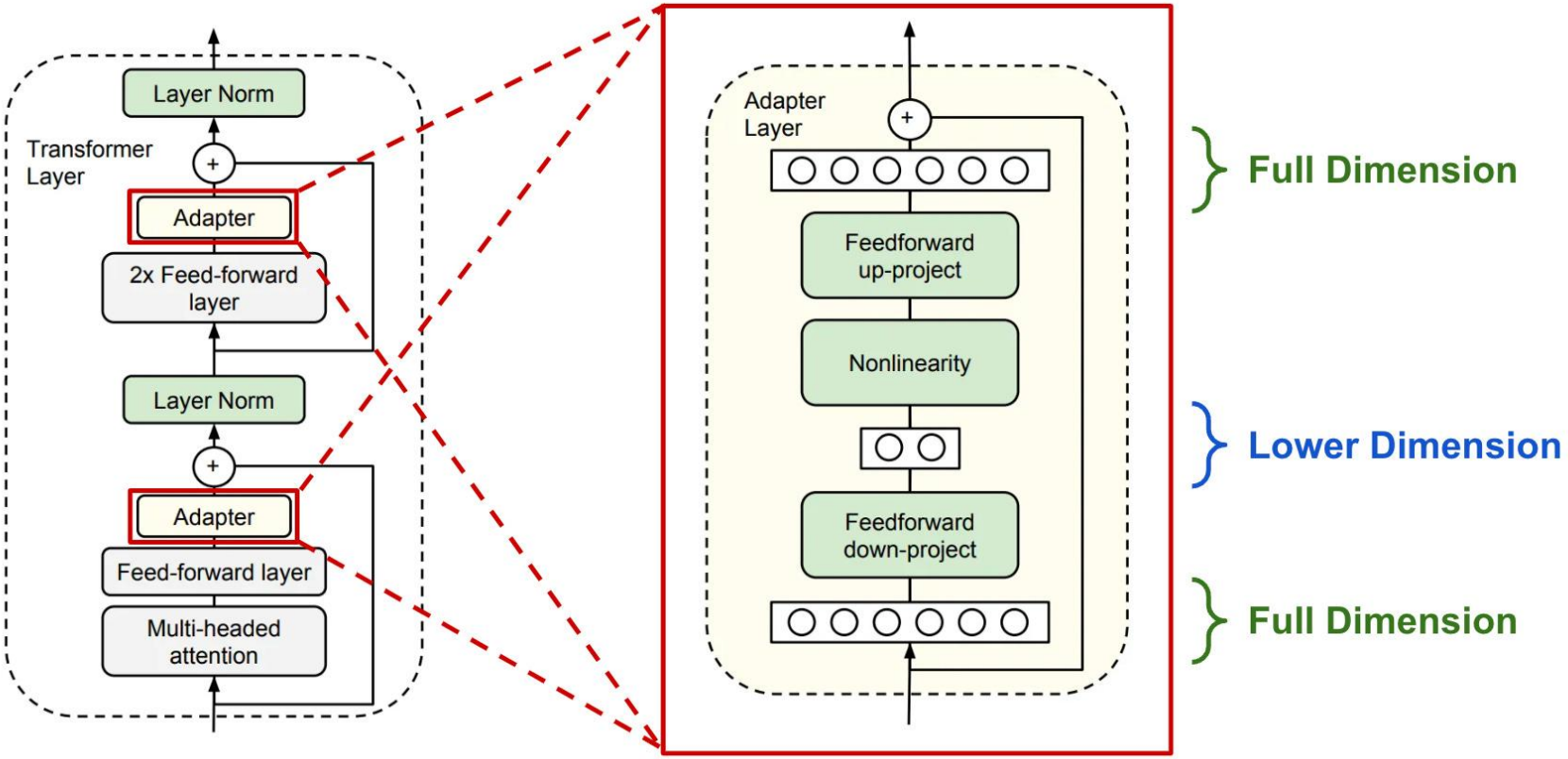
CSE-BERT: Hvar, Rab, Cres, Malta, Brač

XLM-R: Mallorca, Tenerife, otok, Ibiza, Zadar

mBERT: Ibiza, Gibraltar, Tenerife, Mediterranean, Madeira

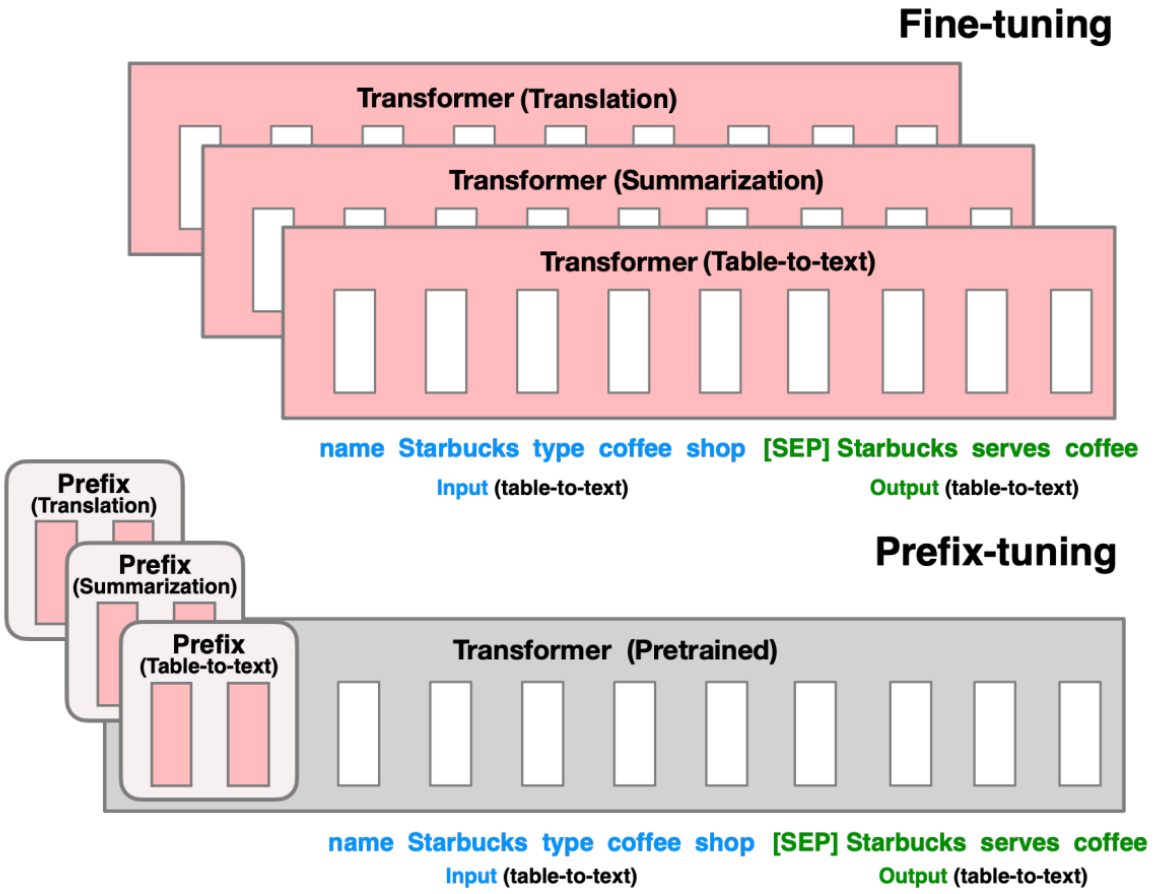
BERT (en): Belgrade, Italy, Serbia, Prague, Sarajevo

Fine-tuning LLMs with adapters for transformers



Prefix tuning

- Comparison of classical finetuning (top) with prefix tuning which adds different prefix blocks for each task but other weights are frozen



LoRA –Low Rank Approximation for LLMs

- See the Substack post
- injects AB into each layer
- r can be very small, e.g., 1% of original d

Finetuned Weights — Weight Update

$$W_{\text{ft}} = W_{\text{pt}} + \Delta W$$

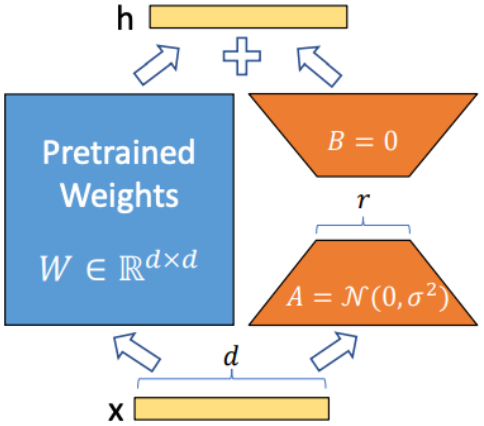
Pretrained Weights

Rank Decomposition Matrix

$$W_{\text{ft}} = W_{\text{pt}} + \underbrace{\Delta W}_{\text{Approximation}} = W_{\text{pt}} + \overbrace{AB}$$

where $W_{\text{ft}}, W_{\text{pt}}, \Delta W, AB \in \mathbb{R}^{d \times d}$

and $\underbrace{A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{r \times d}}_{\text{Low Rank}}$



Scaling Factor

$$W_{\text{ft}} = W_{\text{pt}} + \frac{\overbrace{\alpha}}{r} \underbrace{AB}_{\text{Rank Decomposition Matrix}}$$

Attention efficiency

- time and space complexity of self-attention grows quadratically with n (size of input)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \qquad K, V, Q \in \mathbb{R}^{n \times d}$$

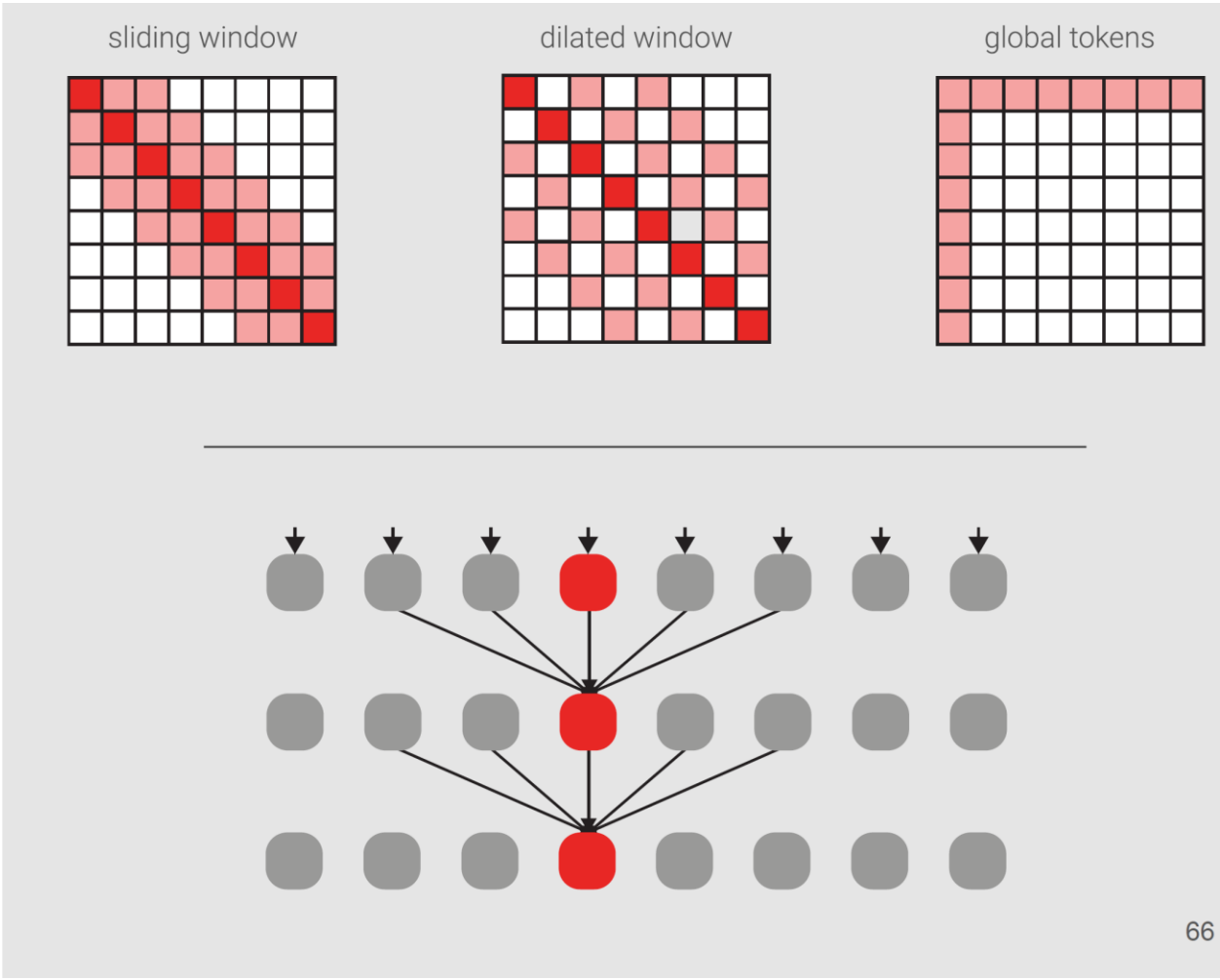
- not suitable for very long sequences like
  - documents
  - character-level language models
  - images (as sequences of pixels);
  - protein sequences.

Longformer [1]

- sliding window attention:
each position can attend to 1/2w tokens on each side - **O(w x n)**

- dilated window attention:
increases the receptive field of the attention layer - **O(w x n)**

- global attention:
k special tokens that aggregate information from whole sequence (e.g. [CLS] as in BERT) - **O(k x n)**



[1] Beltagy et al.: Longformer: The Long-Document Transformer, 2020.

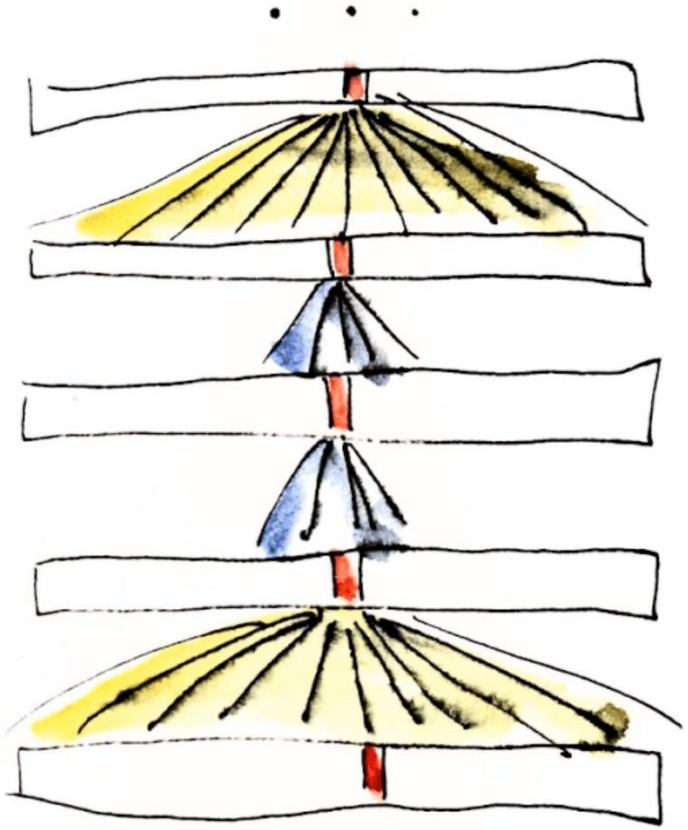
Modern BERT

- sequence length 8192
- **base** (149M params) and **large** (395M params)
- improvement in speed and accuracy over the original (2018) BERT
- includes code in the training set
- a modernized transformer architecture
- better efficiency
- modern data scales & sources
- uses rotary positional encoding (RoPE) instead of absolute positional encoding
- uses continuously derivable GEGLU (Gaussian Error Gated Linear function) activation function
  GeGLU(x) = x *sigmoid(x) + x* 0.5 *(1 + tanh[sqrt(2/pi)* (x + 0.044715 x^3)]), see https://vitalab.github.io/blog/2024/08/20/new_activation_functions.html
- remove unnecessary bias terms
- add an extra normalization layer after embeddings to stabilize training

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., ... & Poli, I. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.
https://huggingface.co/blog/modernbert

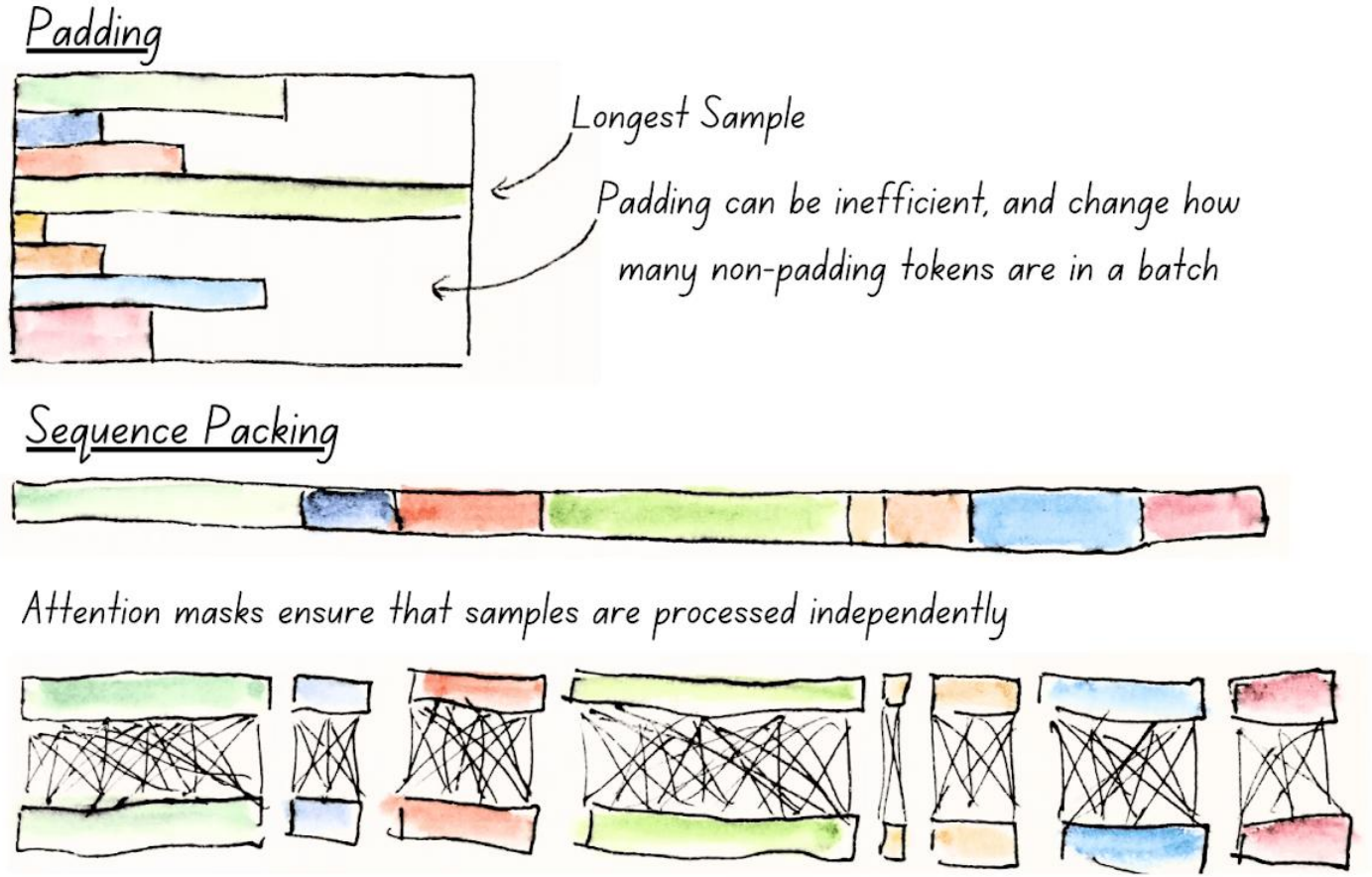
Modern BERT: Alternating attention

- attention mechanism only attends to the full input every 3 layers (**global attention**),
- all other layers use a sliding window where every token only attends to the 128 tokens nearest to itself (**local attention)**.



Alternating global and local attention. In local attention layers, a token only attends to those in a small window around it

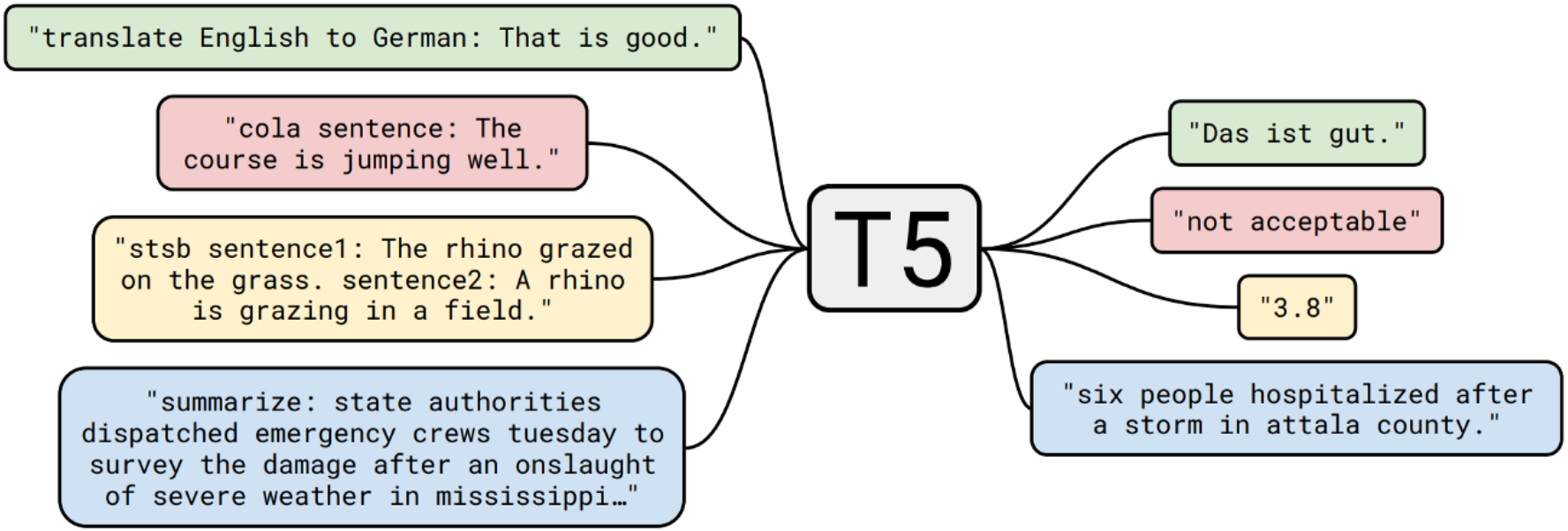
Modern BERT: Unpadding and Sequence Packing



Modern BERT: training

- original BERT data: Wikipedia and Wikibooks
- Modern BERT uses much more diverse data: web documents, code, and scientific articles
- 2 trillion tokens
- increased % of masked tokens from 15% to 30%
- lots of efficiency and hardware utilization tricks for faster training

T5 (Text-To-Text Transfer Transformer) models

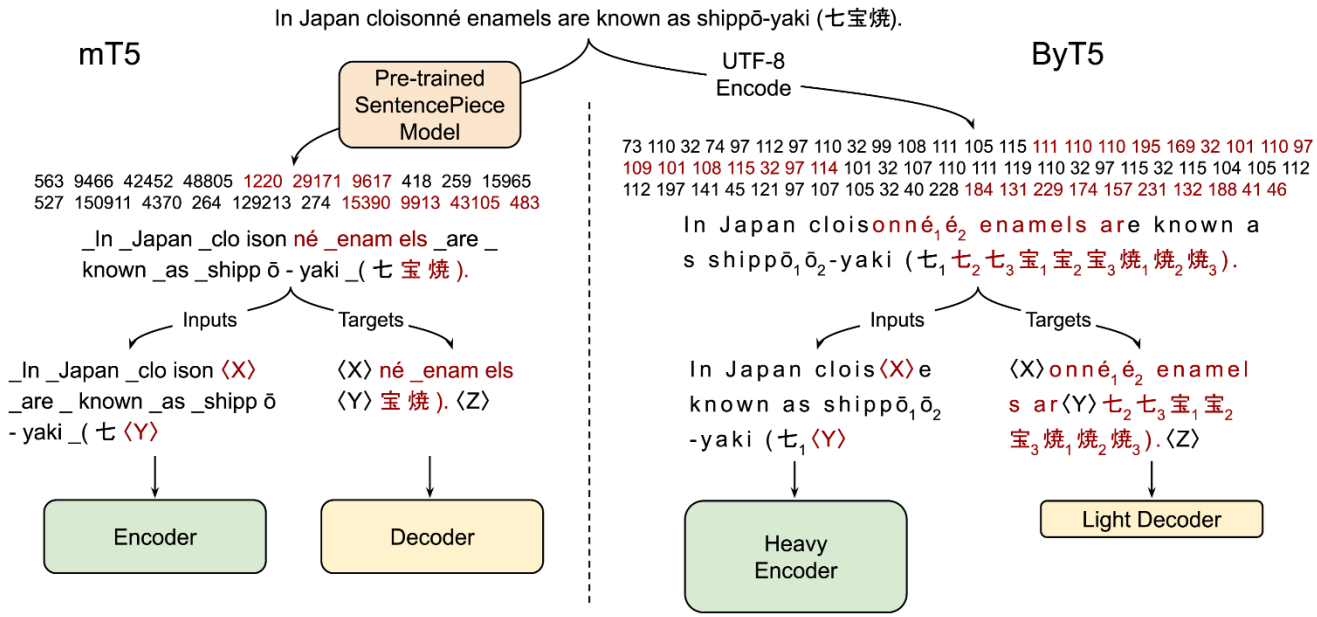


- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y, Li, W. & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(140), 1-67.
- Ulčar, M. & Robnik-Šikonja, M. (2023) Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence,* Section on Language and Computation, Volume 6 – 2023, https://doi.org/10.3389/frai.2023.932519

University of Ljubljana
Faculty of Computer and Information Science

Character-based T5

- Character-based input requires more input layers
- But is very suitable for nonstandard spelling, misspelled words, dialects
- Can process text in any language out of the box
- More robust to noise
- Simplifies text preprocessing



Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, Colin Raffel (2022): ByT5: Towards a Tok… Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*. 2022;10:291-306. doi:10.1162/tacl_a_00461