Introduction to NLP

758b.

BERT

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

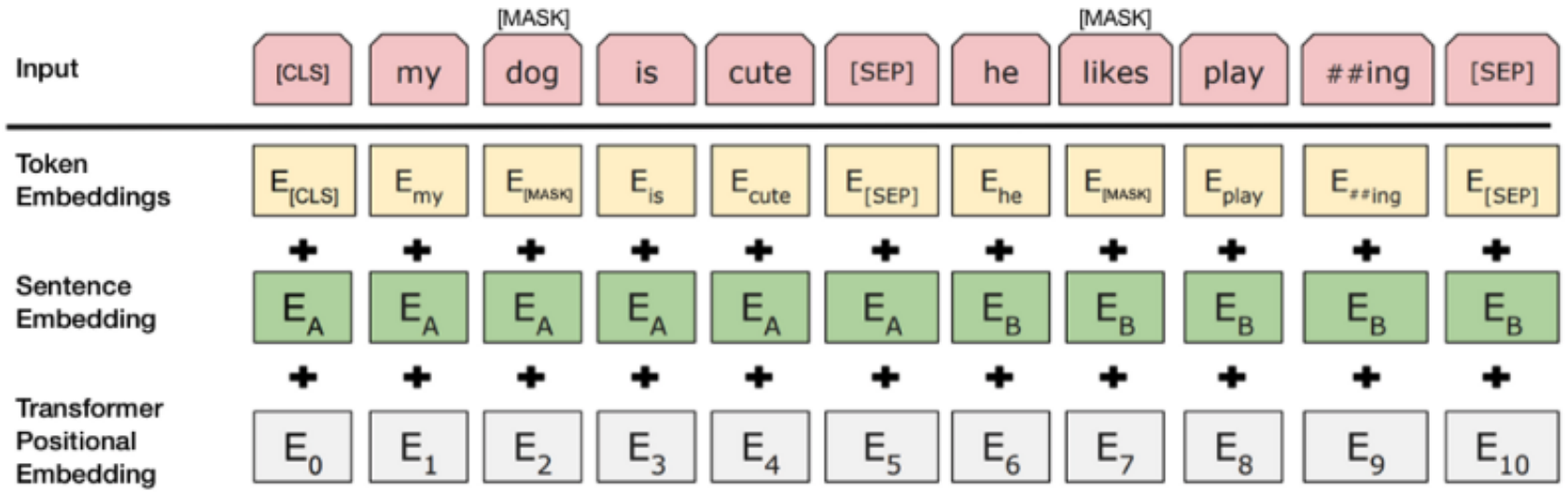
We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

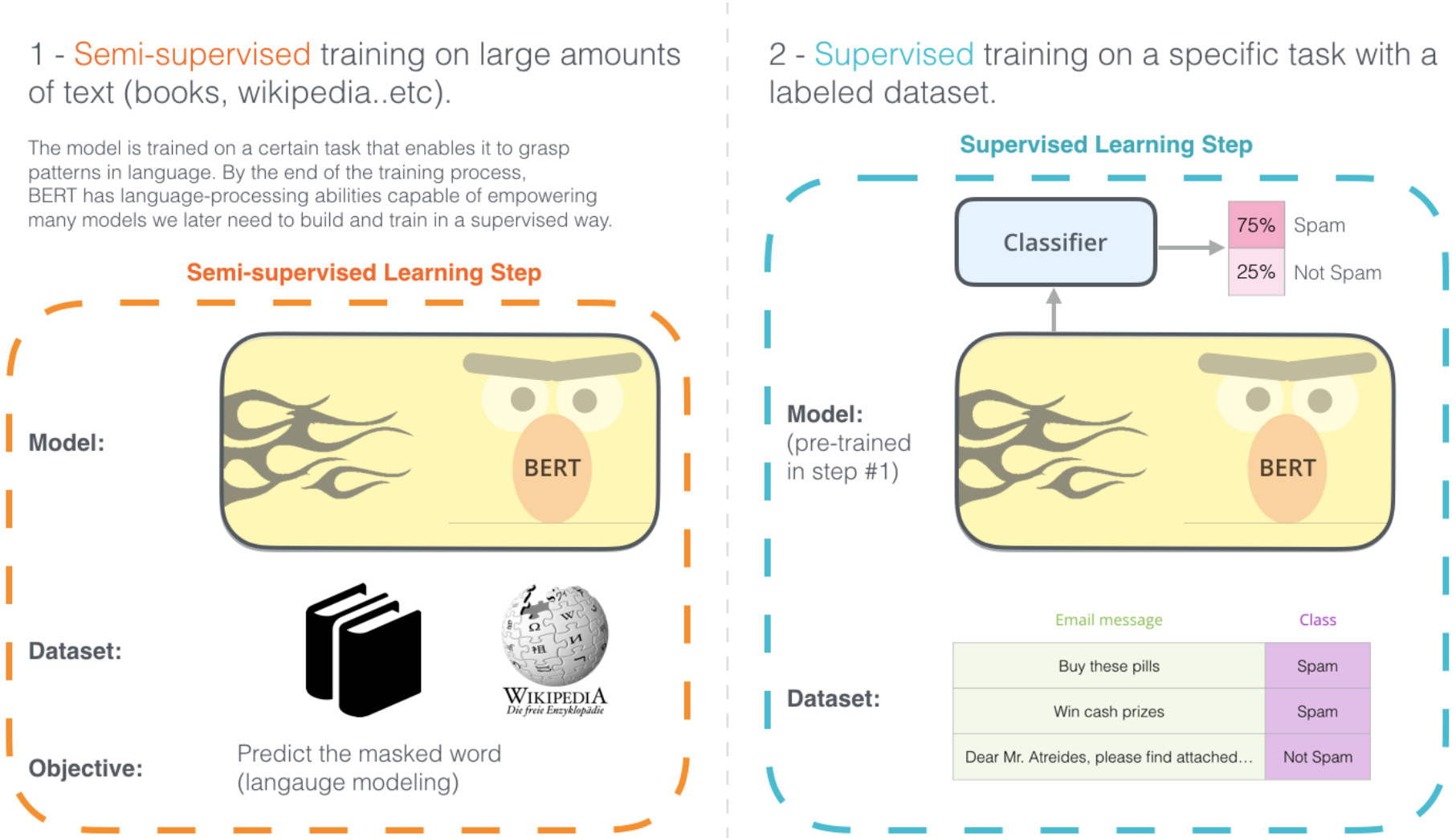
Bert: Pre-training of deep bidirectional transformers for language understanding 4679 2018
J Devlin, MW Chang, K Lee, K Toutanova
arXiv preprint arXiv:1810.04805

What is Bert?

- Bidireectional (actually non-directional)
- Multi-layer self-attention
- Input: one or two sentences
- Start symbol: [CLS]
- Separated by [SEP]



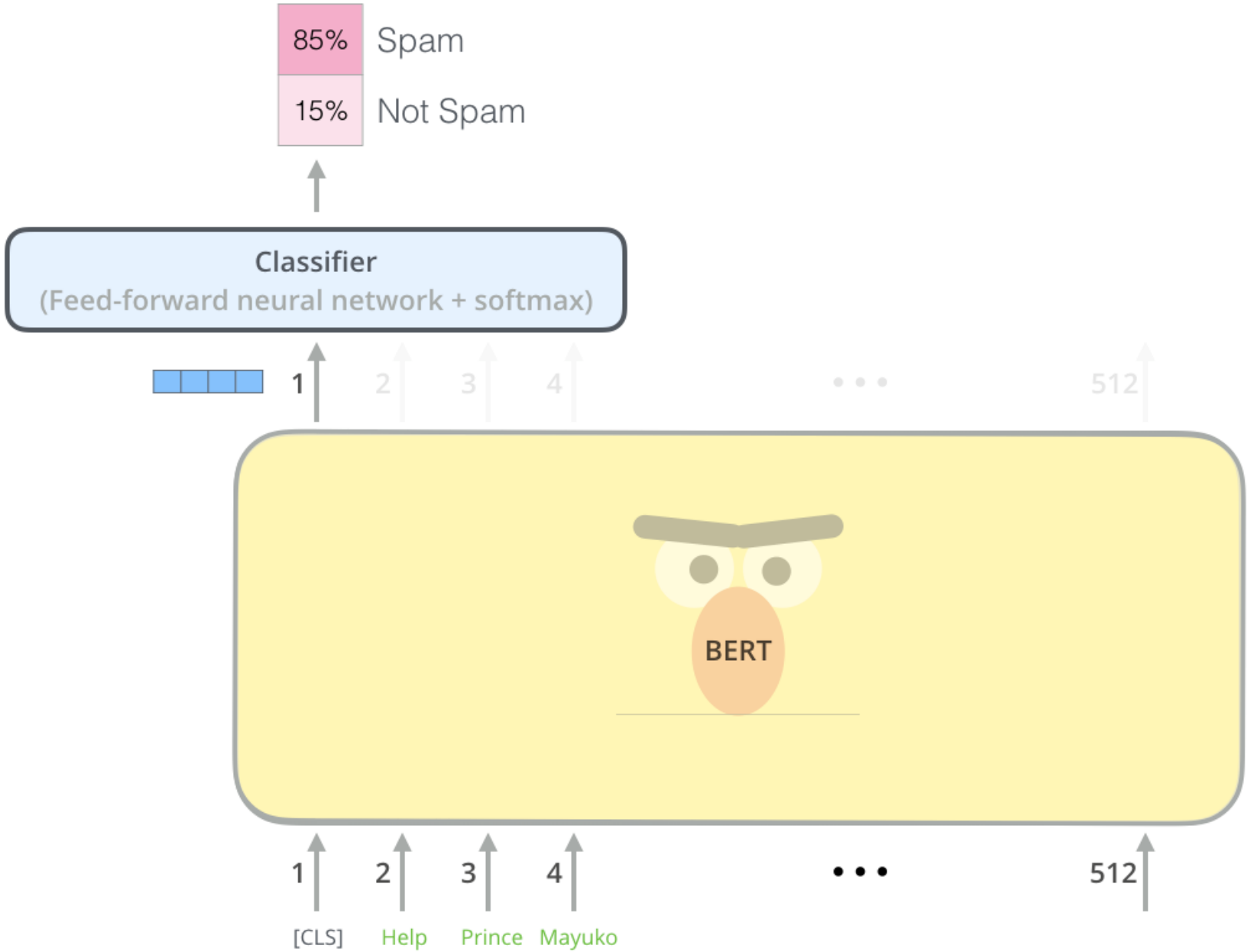
Pre-training



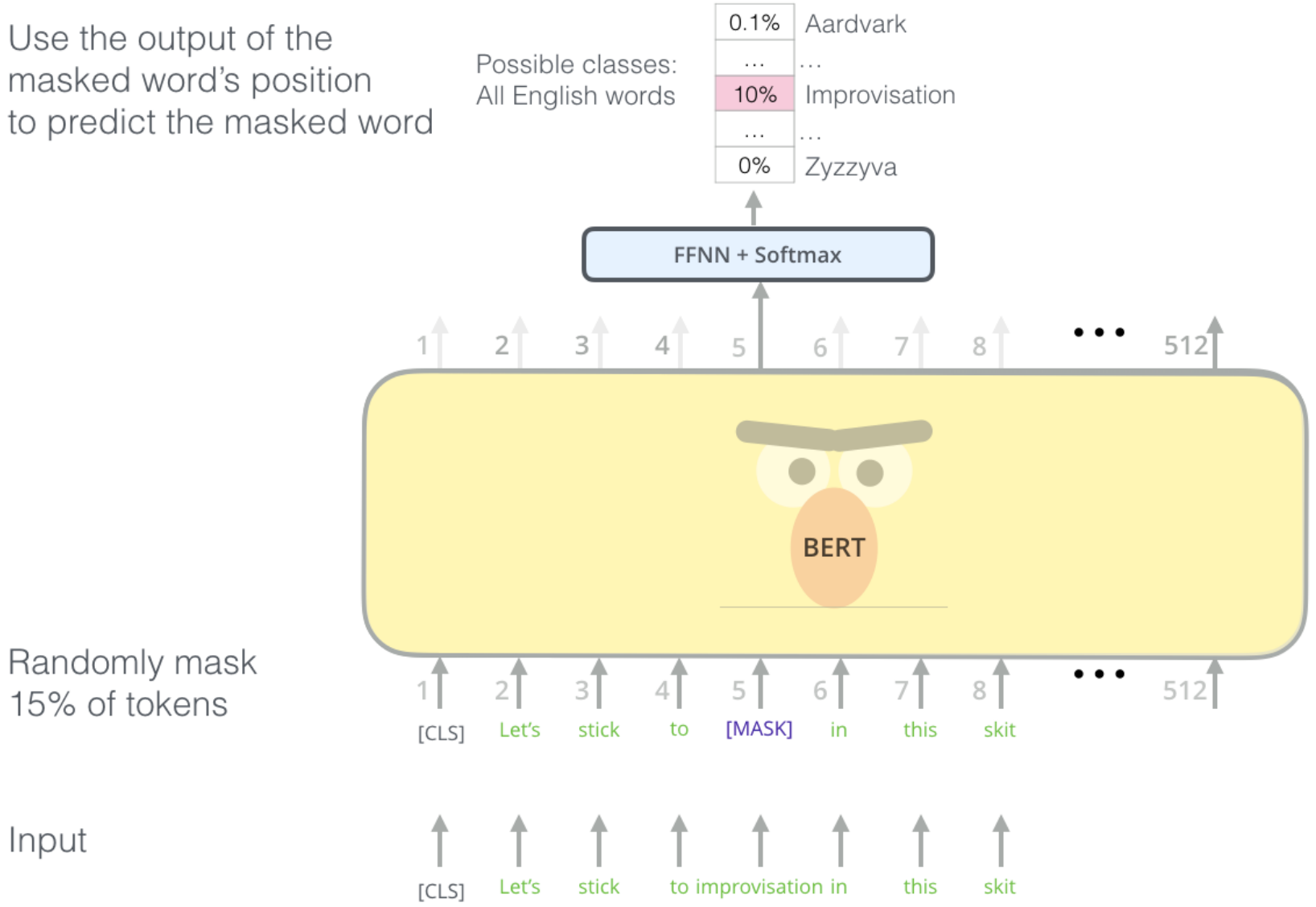
https://jalammar.github.io/illustrated-bert/

Bert Objective #1

- Masked word prediction (15% of the input)
  - 80% replace word with [MASK]
  - 10% replace with random word
  - 10% keep as is



https://jalammar.github.io/illustrated-bert/



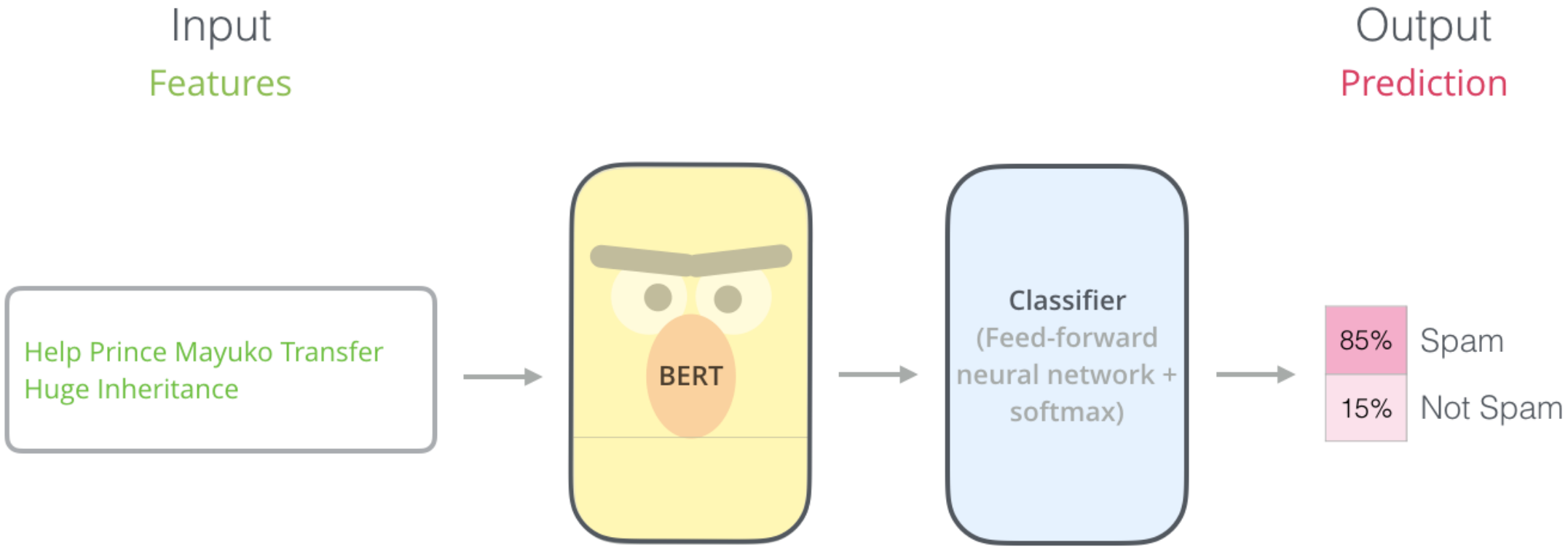
BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

https://jalammar.github.io/illustrated-bert/

Bert Objective #2

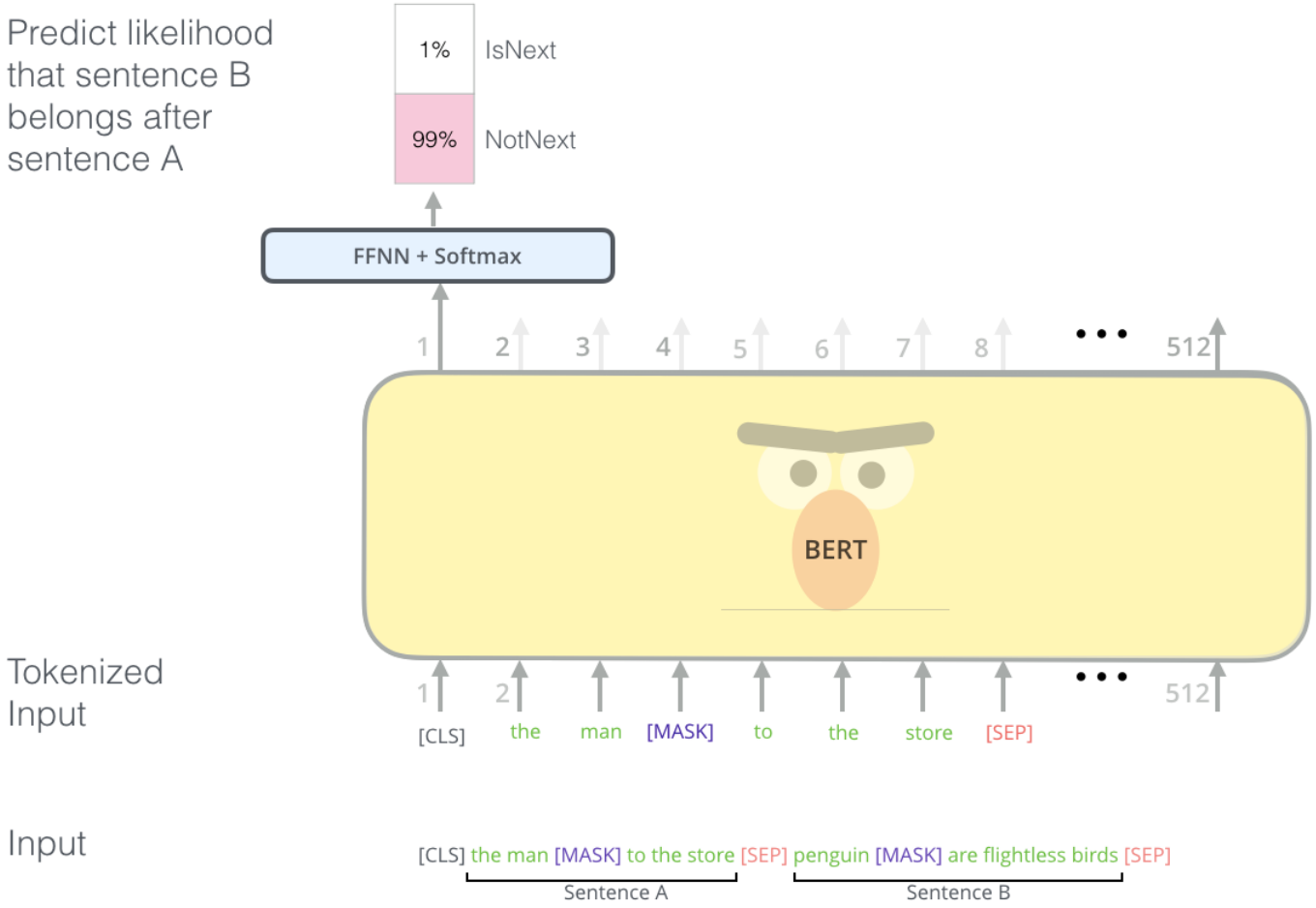
- Next sentence prediction
- Trained on a large book corpus
  - ½ of the examples consisting of consecutive sentences
- Example:
  - [CLS] My dog got sick .
  - [SEP] I called the veterinarian .

Sentence Classification



https://jalammar.github.io/illustrated-bert/

Two-sentence tasks



The second task BERT is pre-trained on is a two-sentence classification task. The tokenization is oversimplified in this graphic as BERT actually uses WordPieces as tokens rather than words --- so some words are broken down into smaller chunks.

https://jalammar.github.io/illustrated-bert/

Different uses

- Supervised
  - Encode, Decode, Classify, Translate
- Unsupervised
  - Predict words

Using Bert as a pre-trained model

- Use the Bert model as the first (pre-trained) layer of a network
- Then train (fine-tune) on the actual task

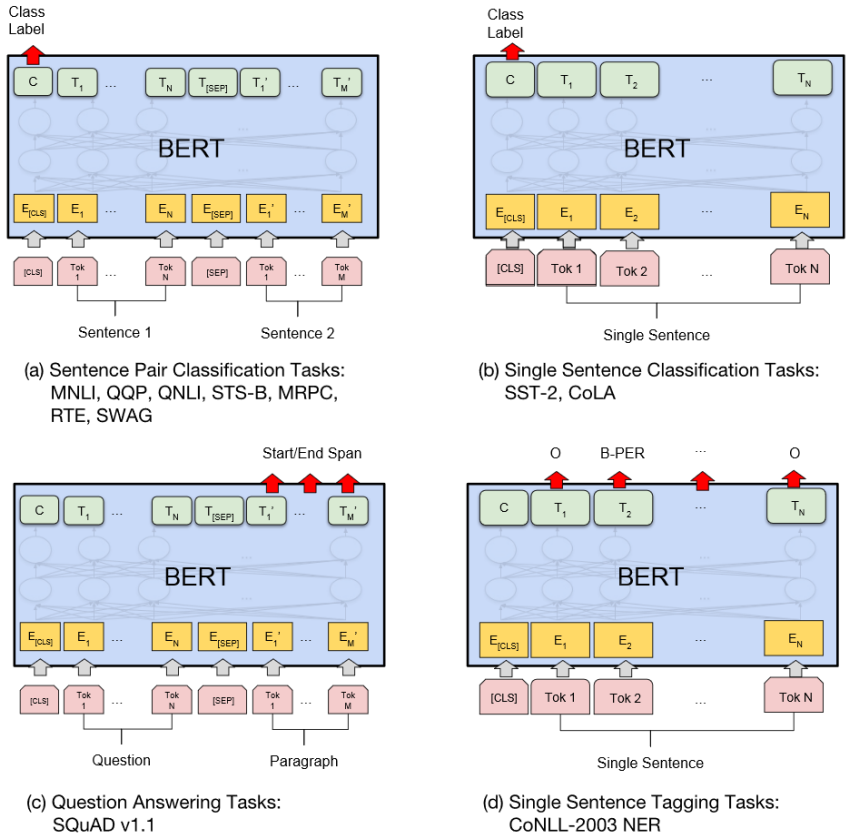


Figure 4: Illustrations of Fine-tuning BERT on Different Tasks.

Notes

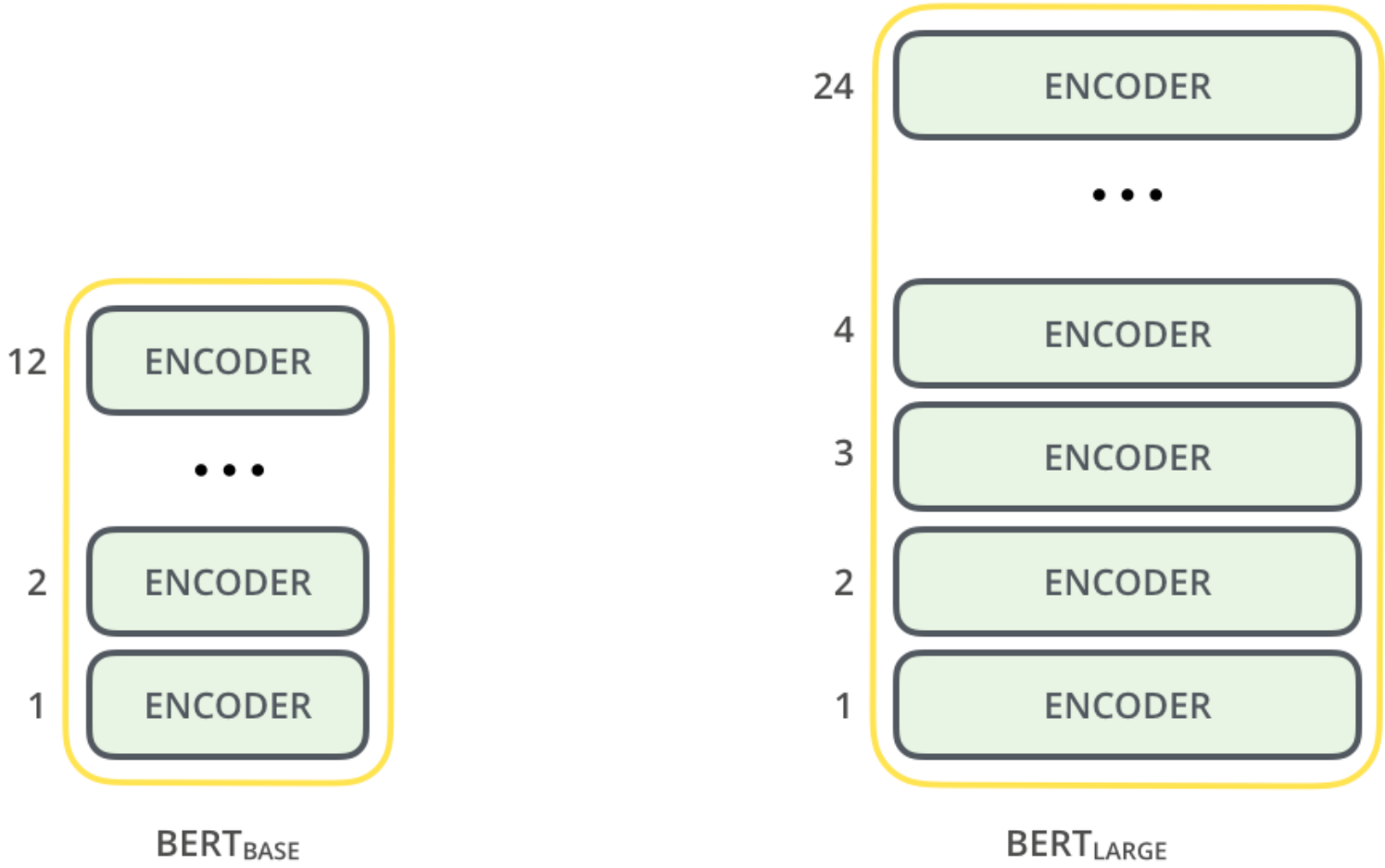
- Bert_large
  - 24 layers
  - 16 heads
  - 1024 dimensions
  - 340M parameters
- Bert_base
  - 12 layers
  - 12 heads
  - 768 dimensions
  - 110M parameters
- Positional embeddings
  - 512 of them
- Wordpieces
  - Vocabulary of 30,000

	Training Compute + Time	Usage Compute
$BERT_{BASE}$	4 Cloud TPUs, 4 days	1 GPU
$BERT_{LARGE}$	16 Cloud TPUs, 4 days	1 TPU

https://github.com/google-research/bert

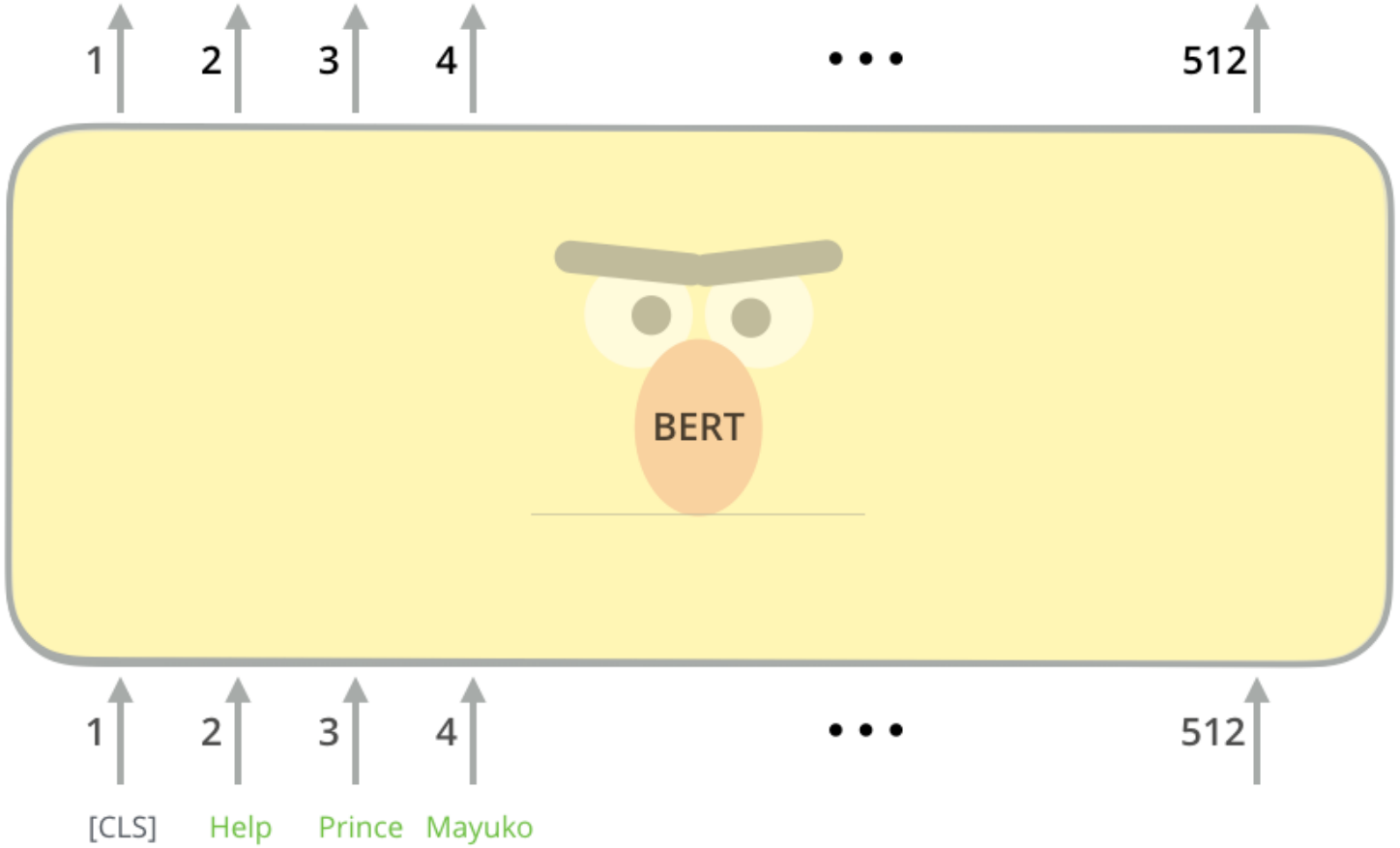
https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1

Encoder Stack

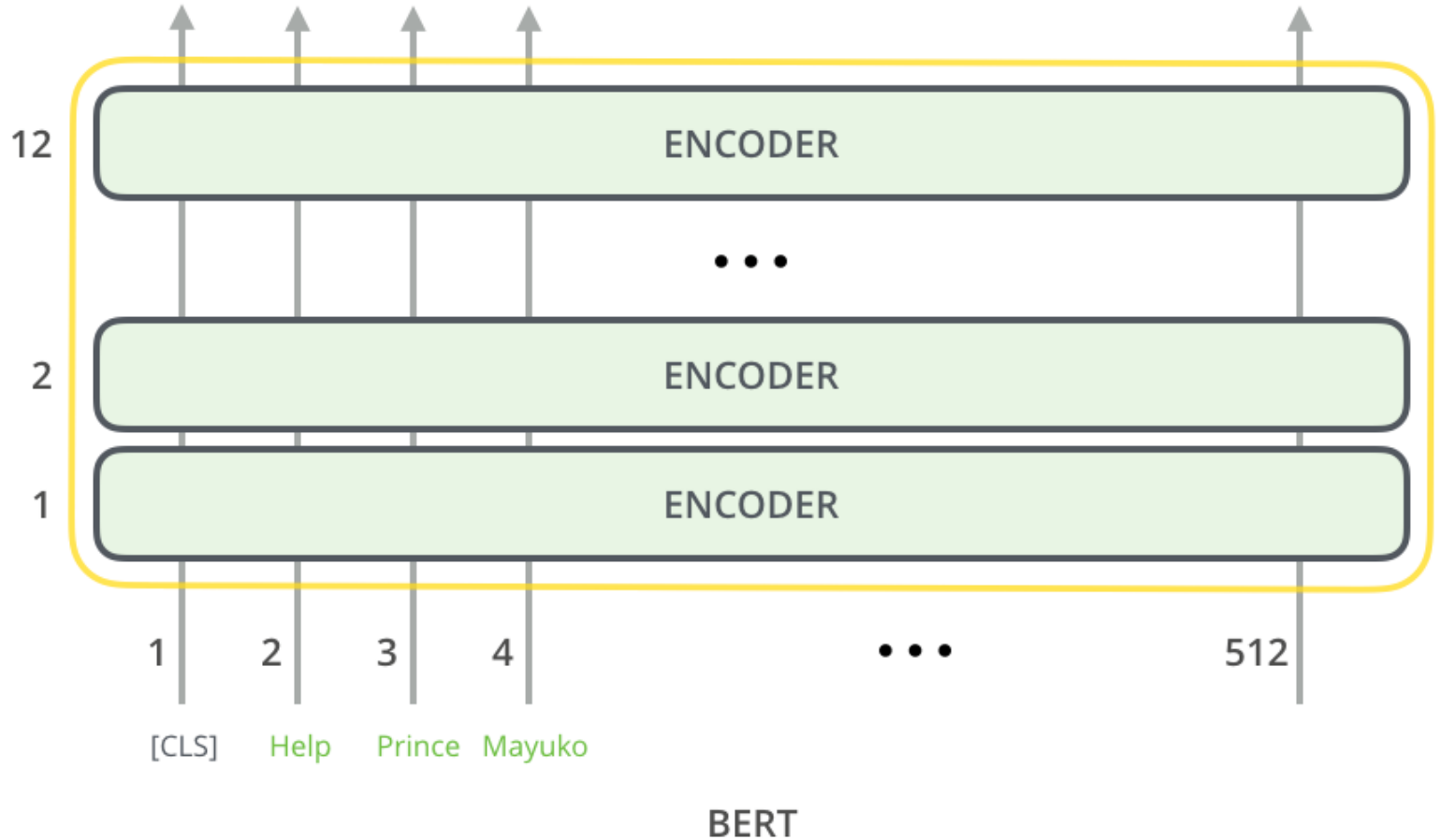


https://jalammar.github.io/illustrated-bert/

Model Inputs

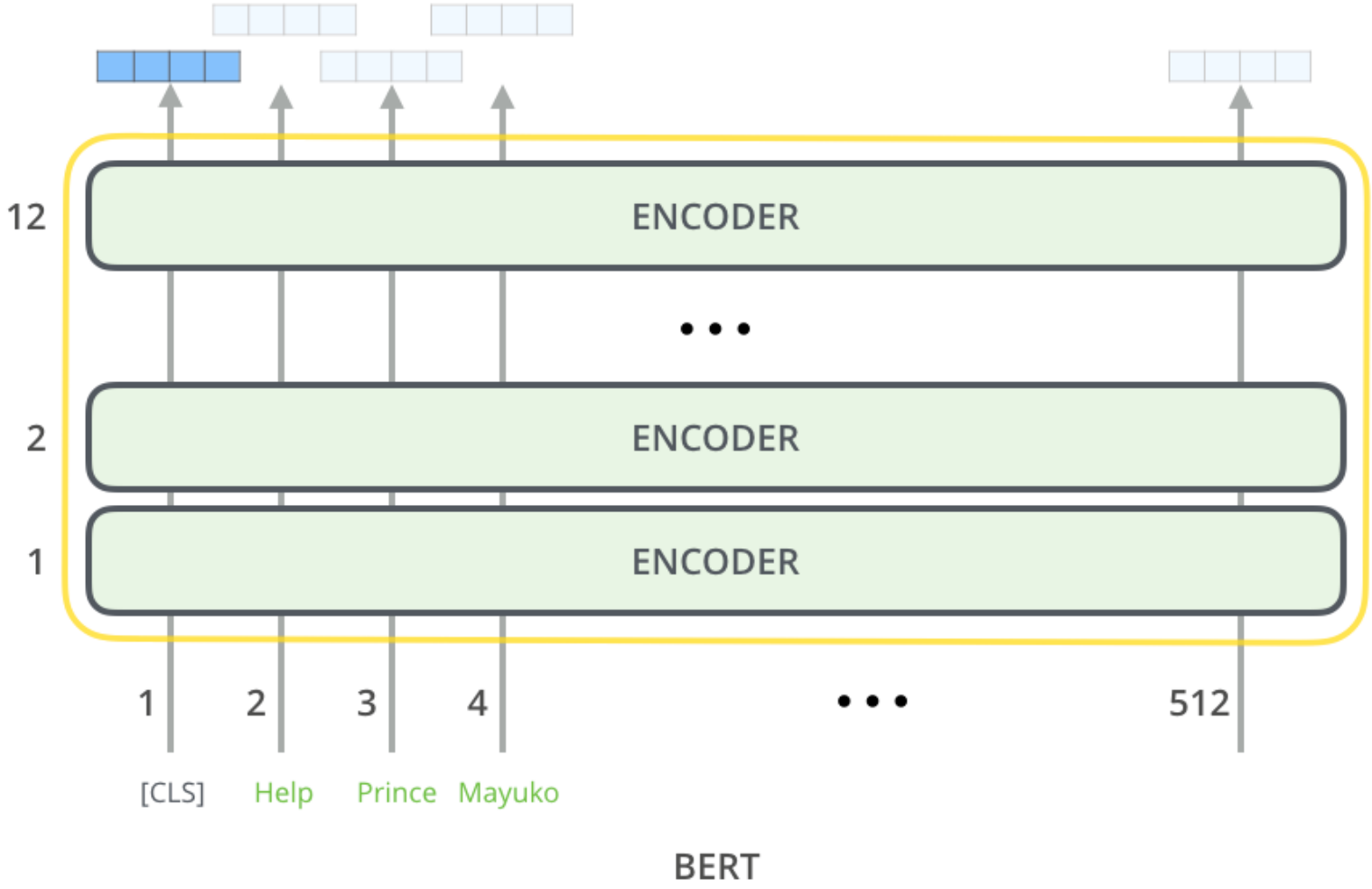


https://jalammar.github.io/illustrated-bert/



https://jalammar.github.io/illustrated-bert/

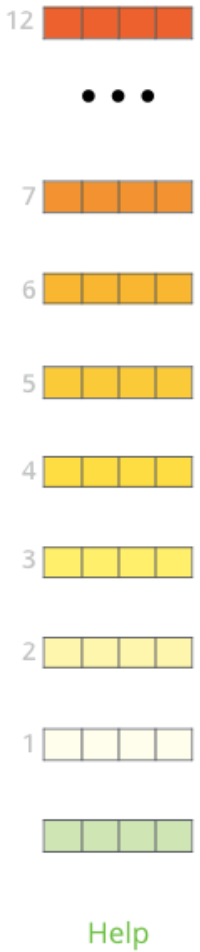
Model Outputs

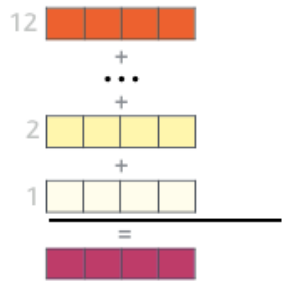
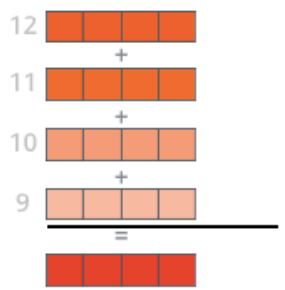
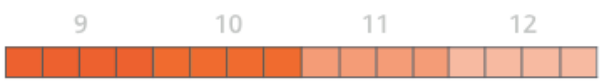


https://jalammar.github.io/illustrated-bert/

What is the best contextualized embedding for "Help" in that context?

For named-entity recognition task CoNLL-2003 NER



		Dev F1 Score
First Layer		91.0
Last Hidden Layer		94.9
Sum All 12 Layers		95.5
Second-to-Last Hidden Layer		95.6
Sum Last Four Hidden		95.9
Concat Last Four Hidden		96.1

https://jalammar.github.io/illustrated-bert/

Results on GLUE

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.
Hypothesis: Jainism hates nature.
Label: Contradiction

CoLa

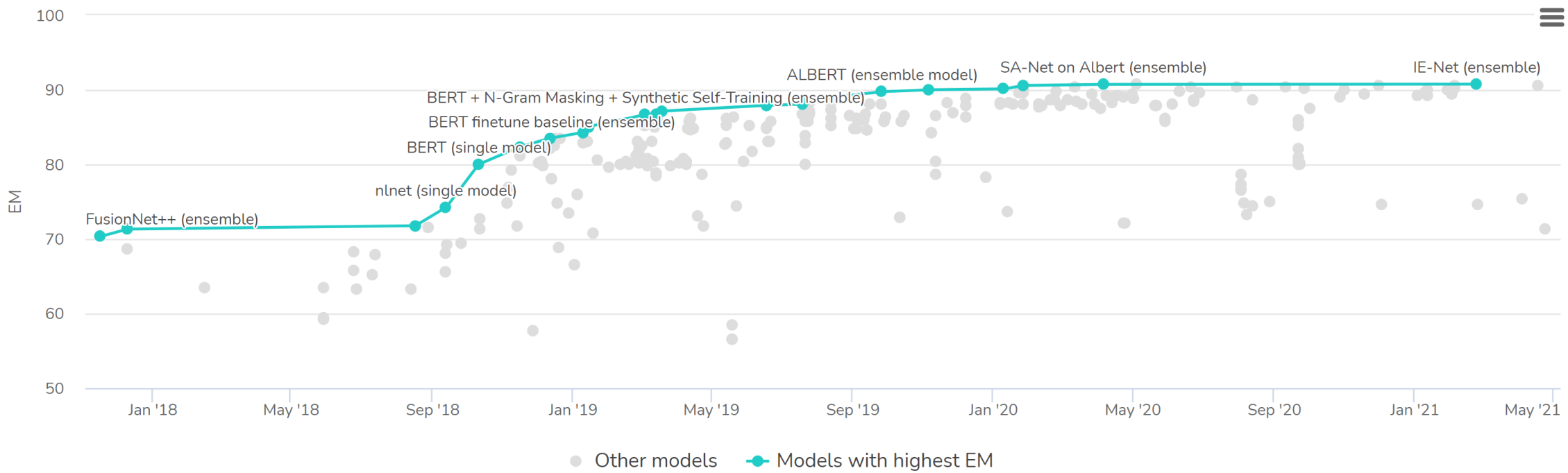
Sentence: The wagon rumbled down the road.
Label: Acceptable

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

https://gluebenchmark.com/tasks

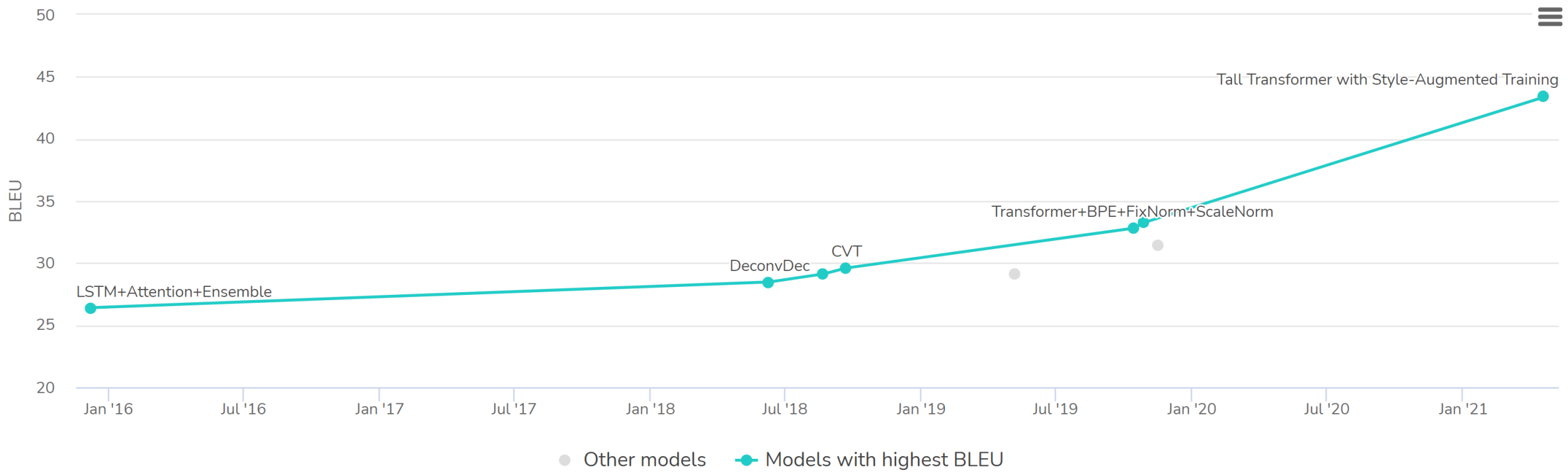
Question Answering on SQuAD2.0

Leaderboard | Dataset | Description



Machine Translation on IWSLT2015 English-Vietnamese

Leaderboard Dataset



Roberta

"Robustly optimized BERT"

160GB of data instead of 16 GB

Dynamic masking: standard BERT uses the same MASK scheme for every epoch, RoBERTa recomputes them

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7

New training + more data = better performance

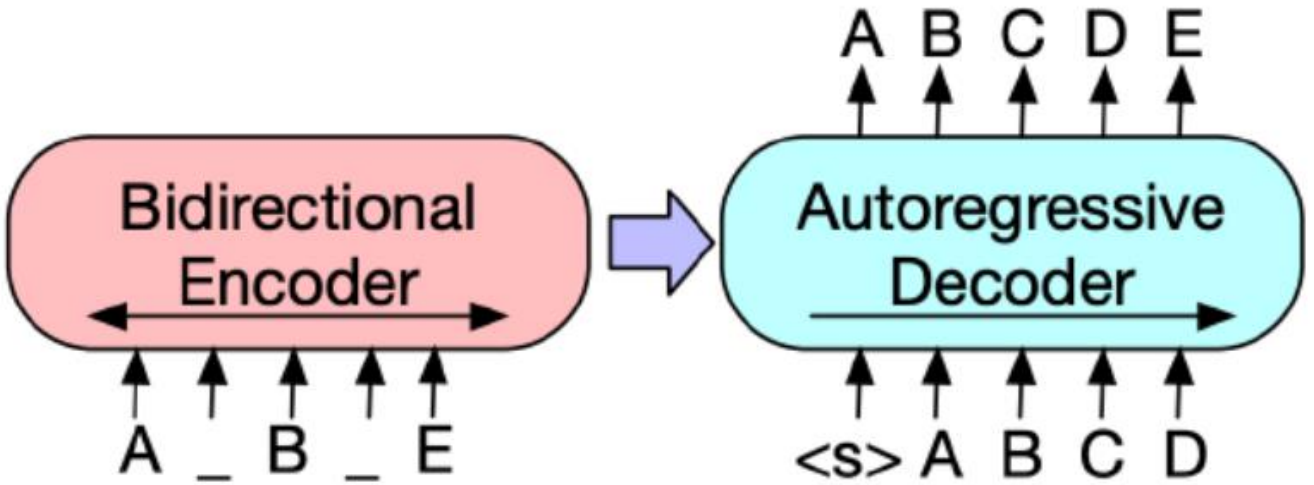
[Liu et al. 2019]

BART

Sequence-to-sequence BERT variant: permute/make/delete tokens, then predict full sequence autoregressively

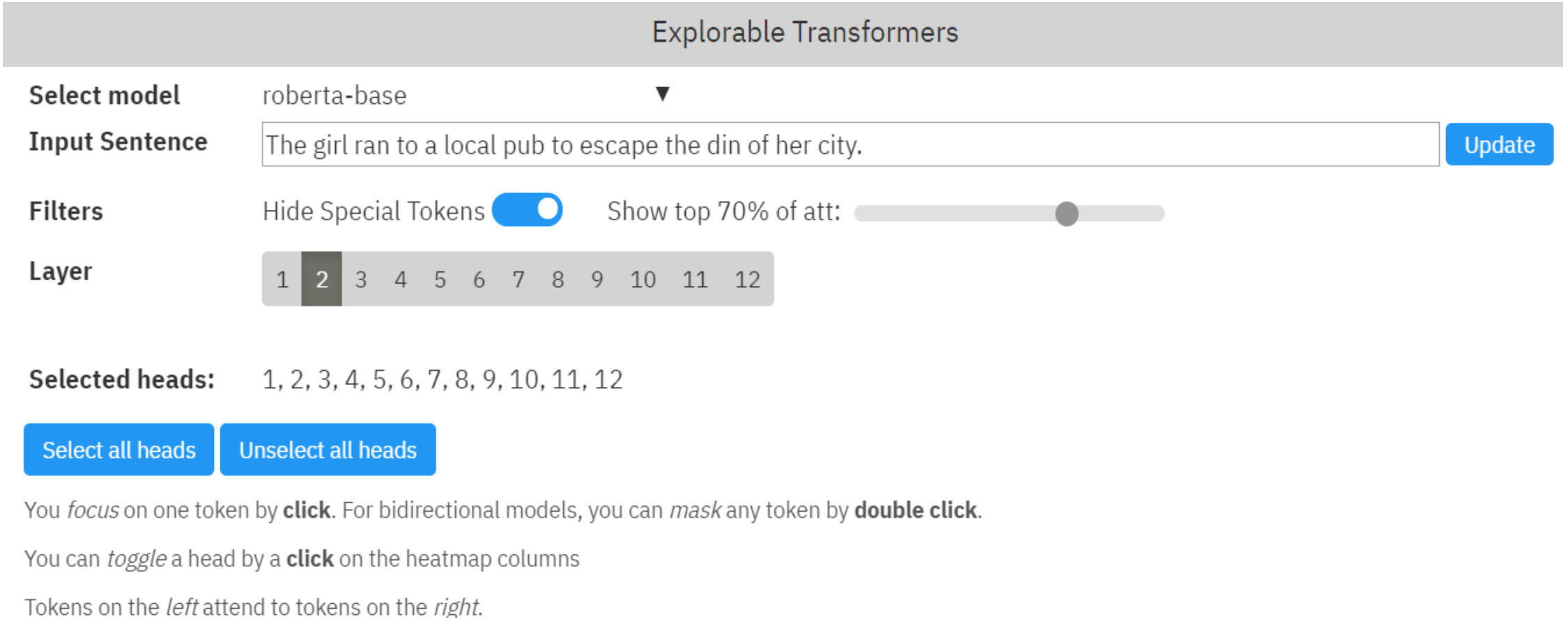
For downstream tasks: feed document into both encoder + decoder, use decoder hidden state as output

Good results on dialogue, summarization tasks

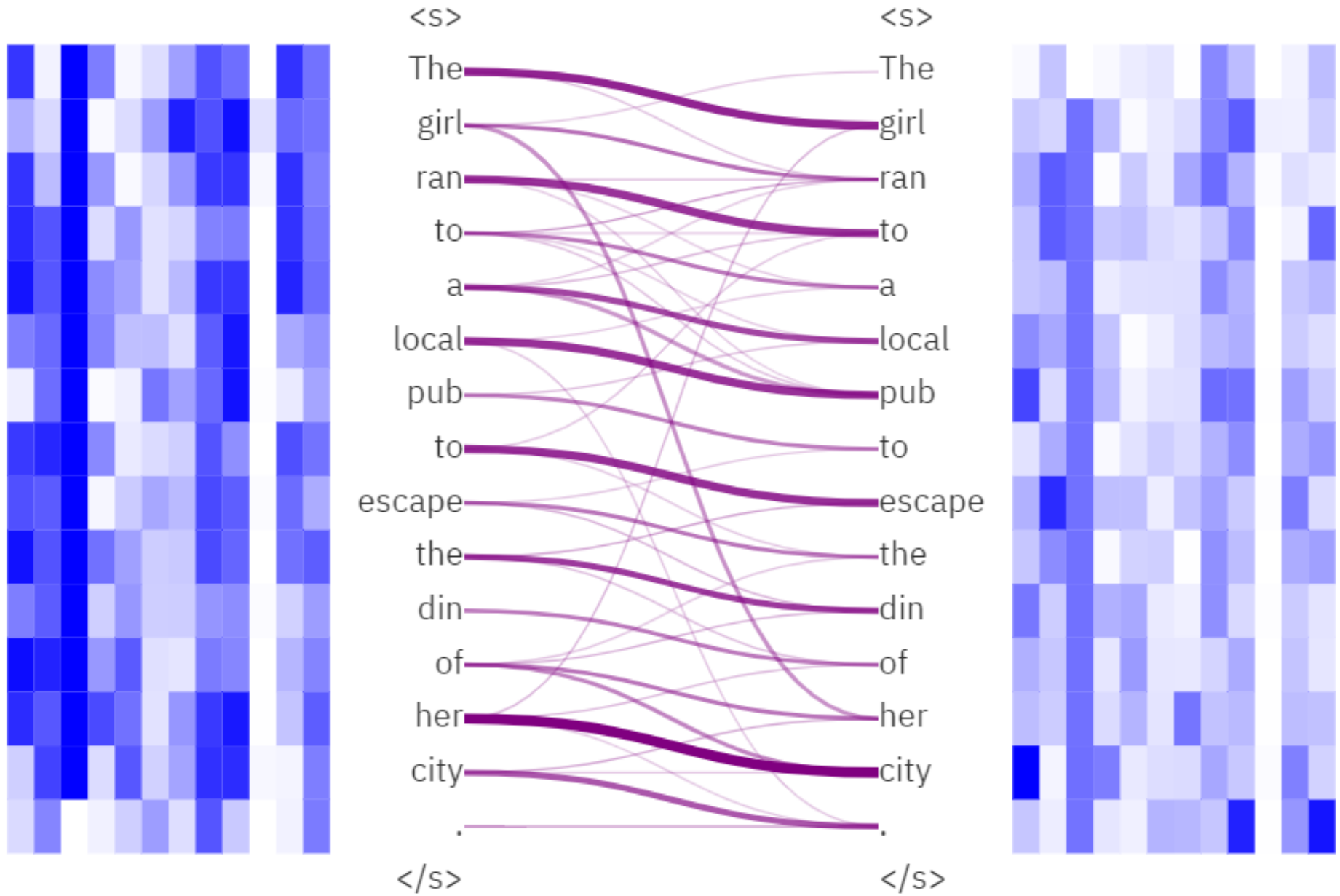


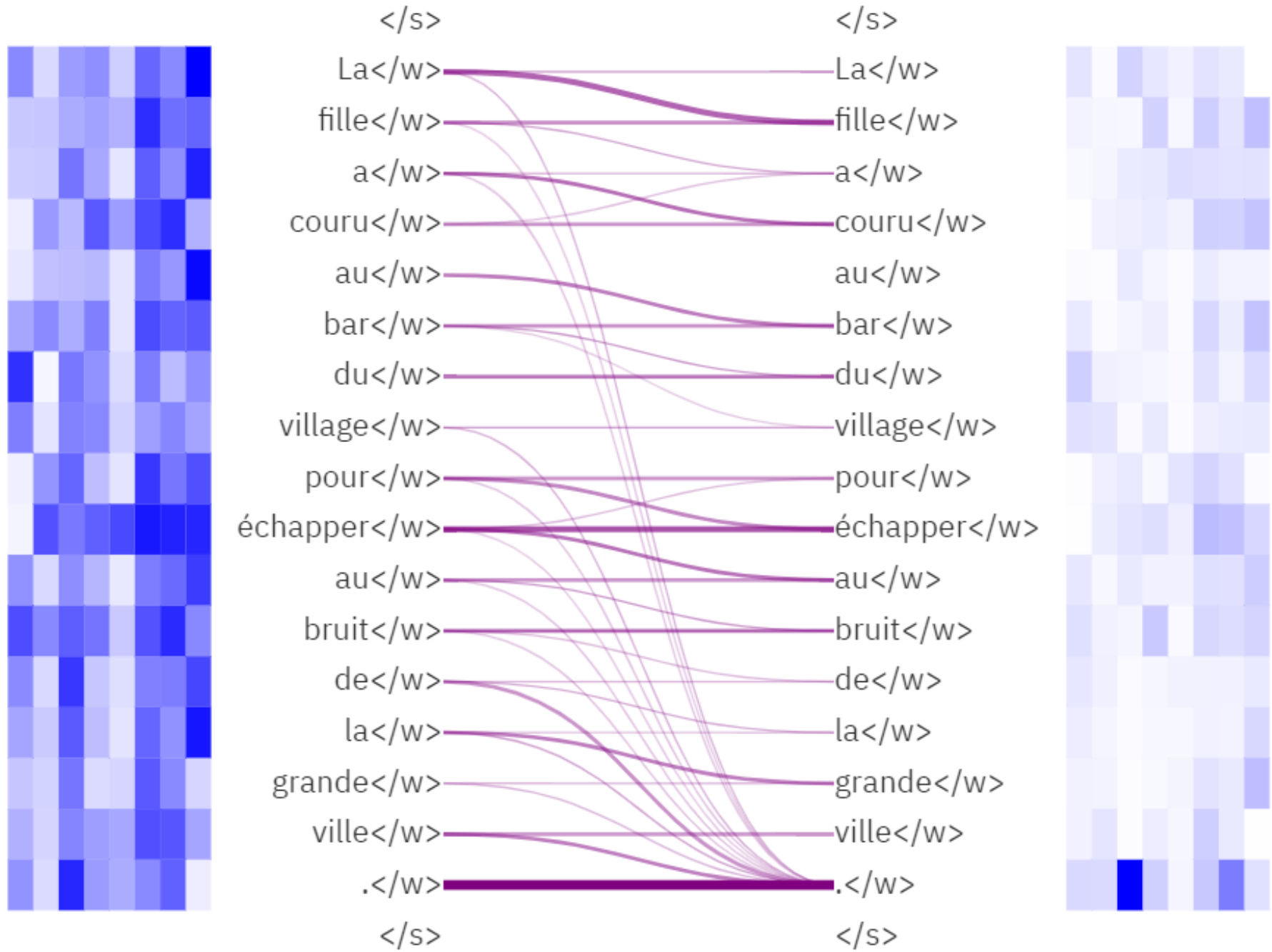
[Lewis et al. 2019]

ExBert - Exploring Transformers



https://huggingface.co/exbert/?model=gpt2&modelKind=bidirectional&sentence=The%20girl%20ran%20to%20a%20local%20pub%20to%20escape%20the%20din%20of%20her%20city.&layer=1&heads=..0,1,2,3,4,5,6,7,8,9,10,11&threshold=0.7&tokenInd=null&tokenSide=null&maskInds=..&hideClsSep=true


<s>
The
girl
ran
to
a
local
pub
to
escape
the
din
of
her
city
.
</s>
<s>
The
girl
ran
to
a
local
pub
to
escape
the
din
of
her
city
.
</s>


</s>
La</w>
fille</w>
a</w>
couru</w>
au</w>
bar</w>
du</w>
village</w>
pour</w>
échapper</w>
au</w>
bruit</w>
de</w>
la</w>
grande</w>
ville</w>
.</w>
</s>
</s>
La</w>
fille</w>
a</w>
couru</w>
au</w>
bar</w>
du</w>
village</w>
pour</w>
échapper</w>
au</w>
bruit</w>
de</w>
la</w>
grande</w>
ville</w>
.</w>
</s>

exBERT

An Explainable BERT

Input Sentence

The children who rarely like games had a blast at the party

Update

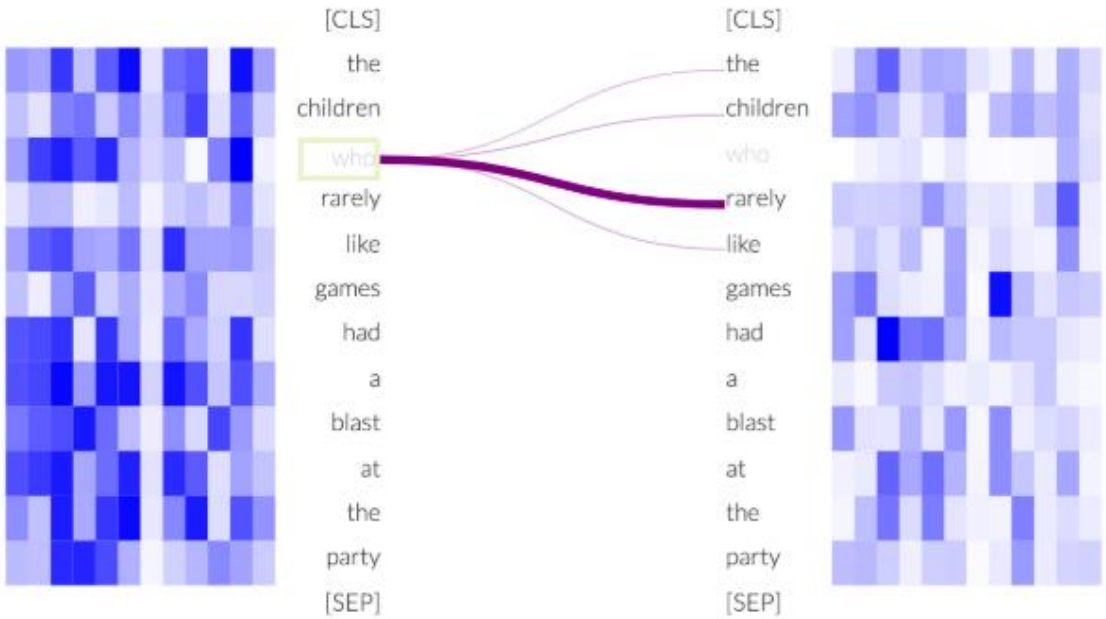
Display top 70% of attention

Selected heads:
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Layer:
0 1 2 3 4 5 6 7 8 9 10 11

Select all heads | Unselect all heads

Hide [CLS] and [SEP]



there are wild beasts in the woods , and a race of queer men who do not like strangers to cross their country .

there were only four witches in all the land of oz , and two of them , those who live in the north and the south , are good witches .

but the scare ##crow and the tin wood ##man , not being made of flesh , were not troubled by the scent of the flowers .

i have no heart , you know , so i am careful to help all those who may need a friend , even if it happens to be only a mouse .

they are the people who live in this land of the east where the wicked witch ruled .

but to those who are not honest , or who approach him from curiosity , he is most terrible , and few have ever dared ask to see his face .

there were milk ##maid ##s and shepherd ##esses , with brightly colored bo ##dice ##s and golden spots all over their gown ##s ; and princess ##es with most gorgeous fr ##ocks of silver and gold and purple ; and shepherd ##s dressed in knee bree ##ches with pink and yellow and blue stripes down them , and golden buckle ##s on their shoes ; and princes with jewel ##ed crowns upon their heads , wearing er ##mine robes and satin double ##ts ; and funny clown ##s in ru ##ffled gown ##s , with round red spots upon their cheeks and tall , pointed caps .

it is better for people to keep away from oz , unless they have business with him .

he sits day after day in the great throne room of his palace , and even those who wait upon him do not see him face to face .

i have heard that g ##lind ##a is a beautiful woman , who knows how to keep young in spite of the many years she has lived .

Other demos

- https://github.com/jessevig/bertviz
- https://home.ttic.edu/~kgimpel/viz-bert/viz-bert.html
- https://colab.research.google.com/drive/1c73DtKNdl66B0_HF7QXuPenraDp0jHRS
- https://colab.research.google.com/drive/1PEHWRHrvxQvYr9NFRC-E_fr3xDq1htCj#scrollTo=fZAXH7hWyt58