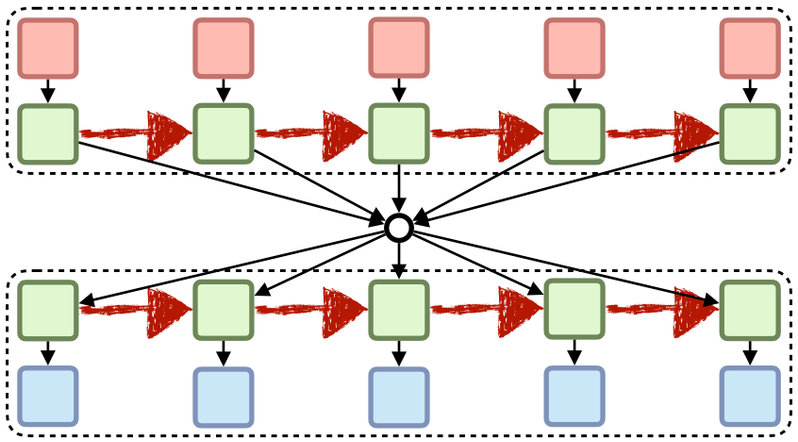
Deep neural language models

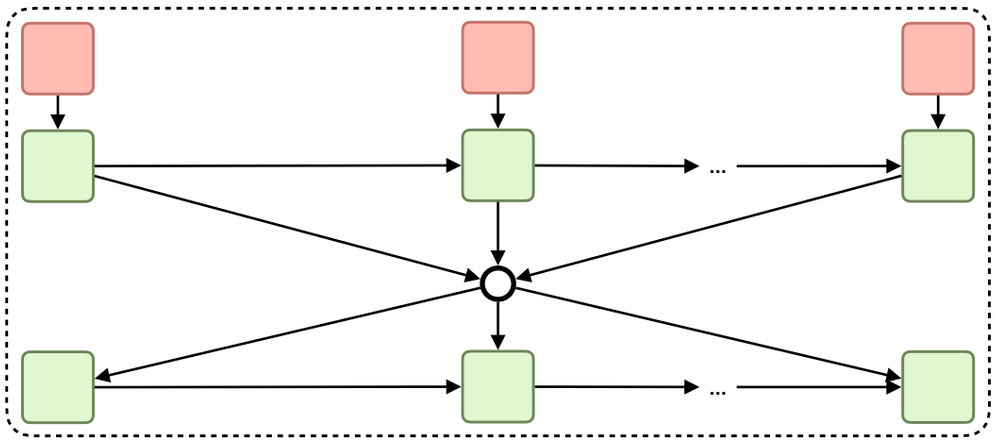
Recap

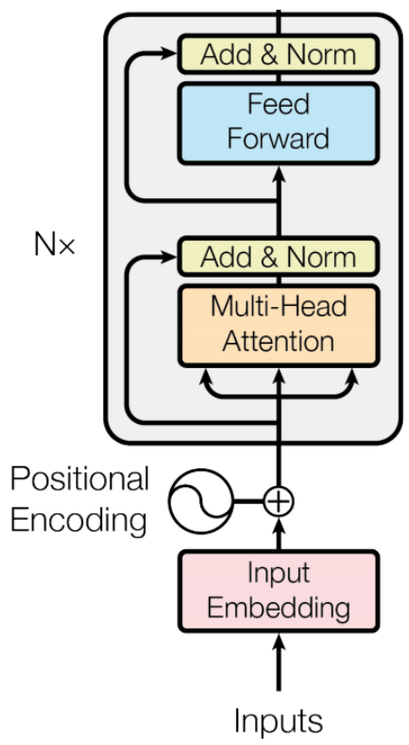
RNN models with encoder-decoder attention have **long-range dependencies** within the encoder and decoder

In self-attention, a model attends to its own representations
→ **constant-length** connections, **contextual representations**

Transformer uses dot-product **self-attention** w/o recurrence (efficient), **positional encodings** for sequence information







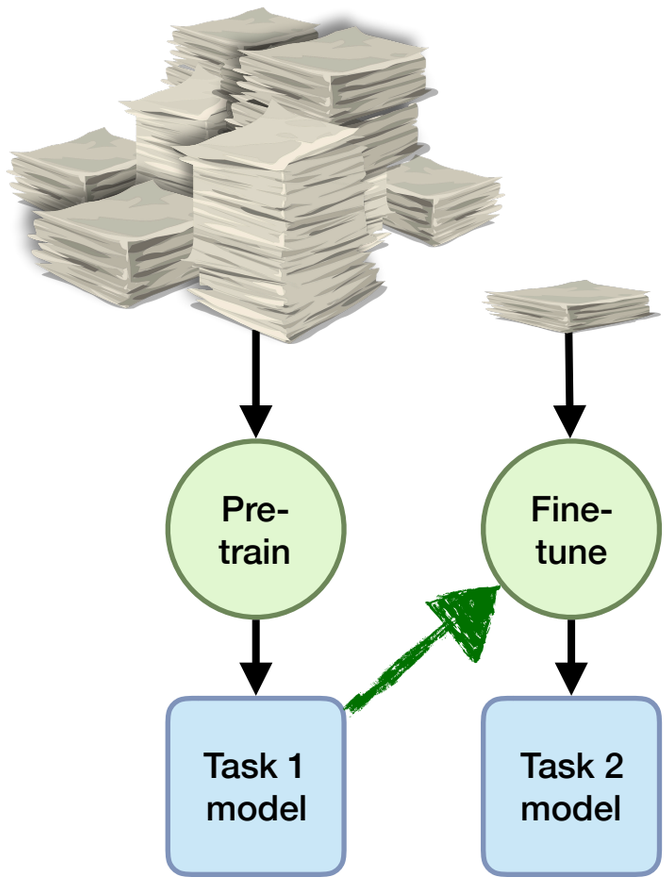
Transformer figure from Vaswani *et al*. (2017) *Attention Is All You Need*

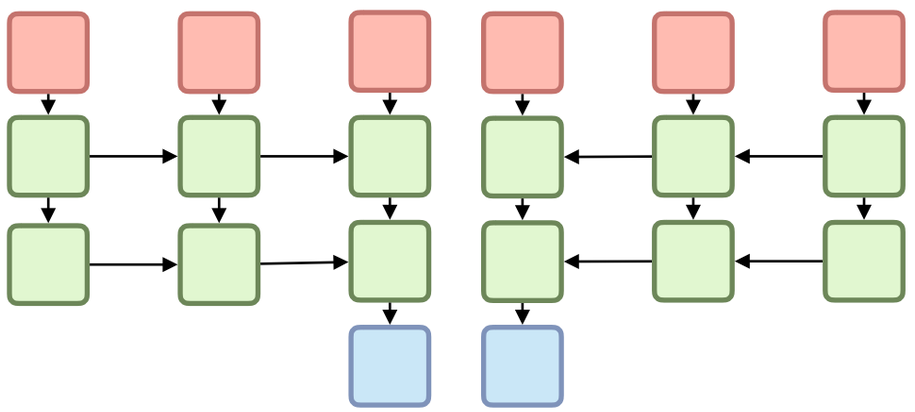
Recap

In **transfer learning**, knowledge from one task is used when learning another (e.g. weight initialization with word embeddings)

Pre-training on unsupervised task (lots of data) and then fine-tuning on task-specific annotated data particularly effective

Neural language models trained on large unannotated corpora can be used to create **contextualized representations** of meaning





ELMo

Embeddings from Language Models

Deep language model using forward and backward LSTMs

Forward and backward RNN states concatenated to create contextual word representation

Pre-trained on 1B word unannotated corpus of English

In combination with supervised NLP tasks, advanced SOTA in tasks including question answering, NER and sentiment analysis

Peters *et al*. (2018) *Deep contextualized word representations*

GPT

Generative Pre-training Transformer

Deep *forward* language model using transformer decoding

→ no need for “future” inputs, emphasis on generation

Pre-trained on ~1B word BooksCorpus (English)

Fine-tuning for supervised NLP tasks, advanced SOTA

Larger follow-up model GPT-2 was model “too dangerous to release”

Radford *et al*. (2018) *Improving Language Understanding by Generative Pre-Training*

BERT

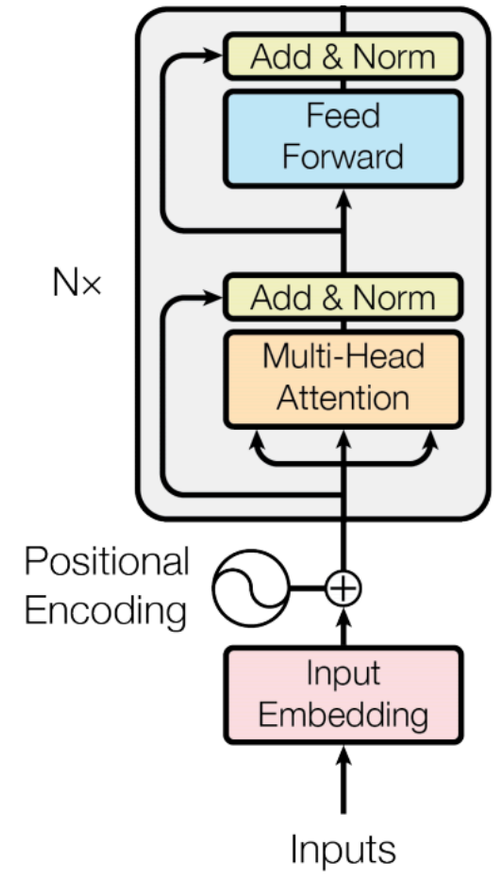
Bidirectional Encoder Representations from Transformers

Deep *bidirectional* language model using transformer encoder

Trained using **masked language modeling** and **next sentence prediction** objectives

English model pretrained on 3B words (Wikipedia + BooksCorpus)

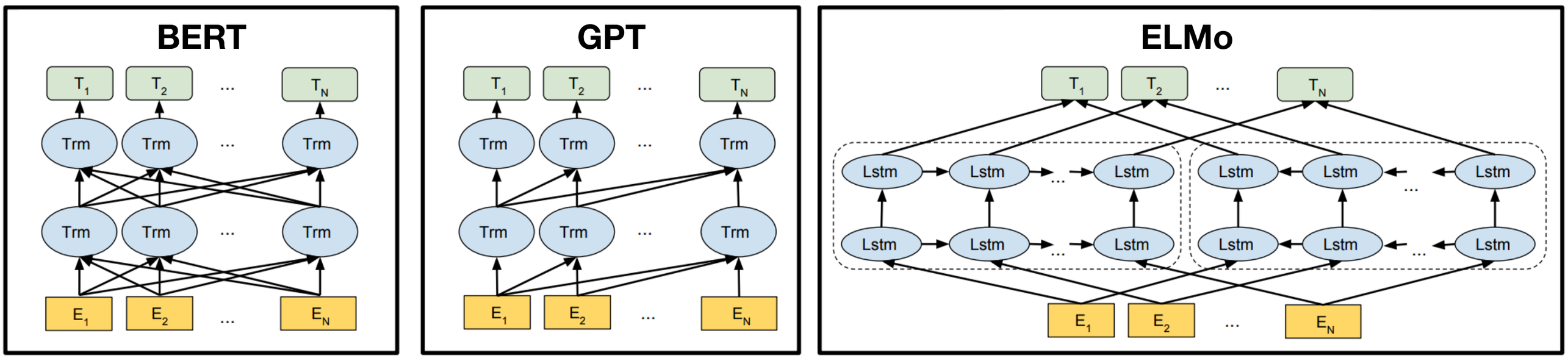
Fine-tuning for various language understanding tasks showed results *surpassing human performance*



Devlin *et al*. (2019) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

Deep neural language models

Comparison of BERT, GPT and ELMo



Transformer, Bidirectional

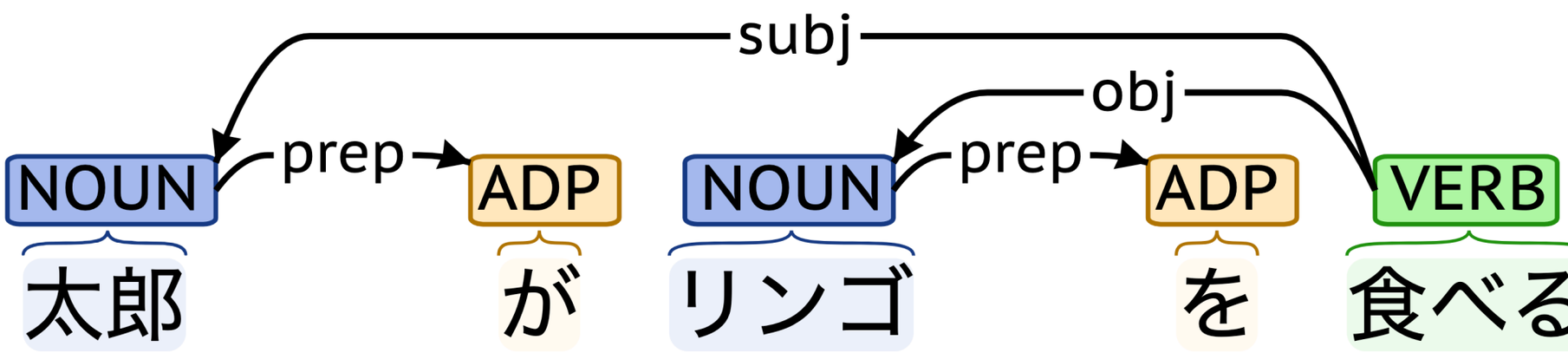
Transformer, Forward only

LSTMs, Bidirectional

Figure from Devlin *et al*. (2019) *BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding*

Deep neural language models

Why language models? (i.e. why try to learn P(w|context)?)



Intuitively: to do NLP tasks, it help if you know the language first

(e.g. easier to learn to annotate syntax for a language you know)

“know a language” ~ “have a good language model”

Radford *et al*. (2019) *Language Models are Unsupervised Multitask Learners*

Super-human NLP?

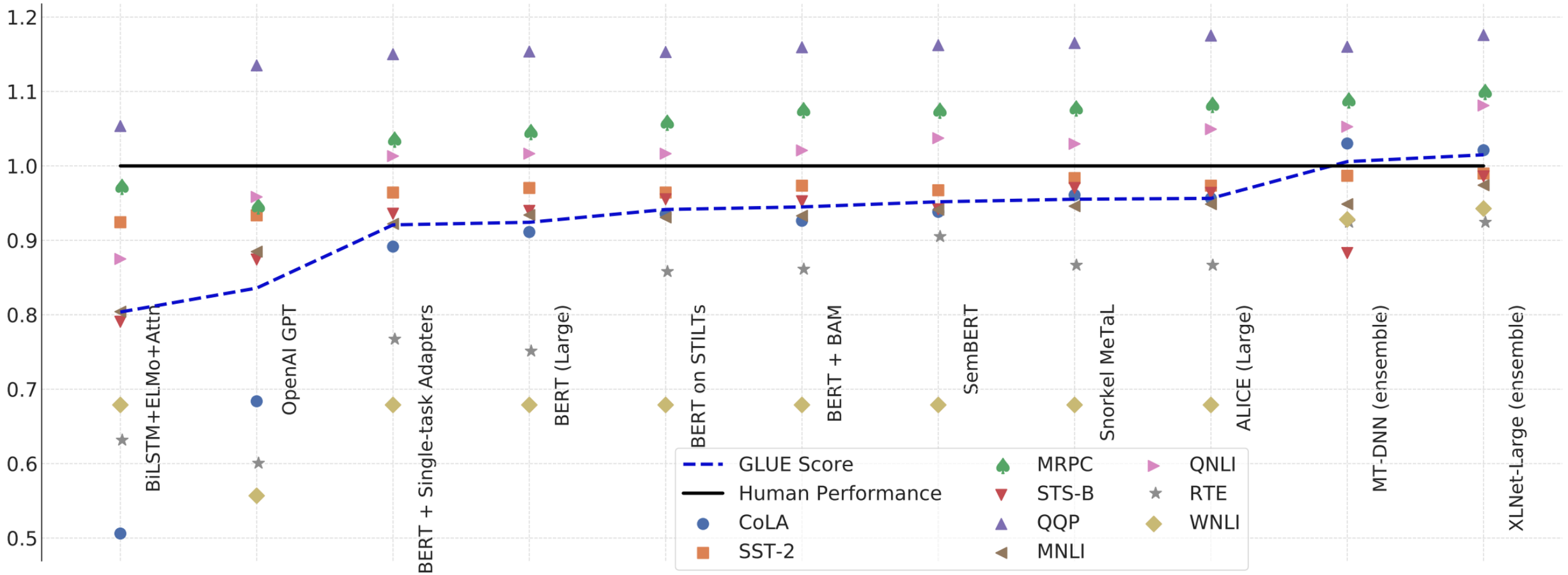
https://gluebenchmark.com/leaderboard

9	Junjie Yang	HIRE-RoBERTa		88.3
10	Facebook AI	RoBERTa		88.1
11	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
12	GLUE Human Baselines	GLUE Human Baselines		87.1
13	Stanford Hazy Research	Snorkel MeTaL		83.2
14	XLM Systems	XLM (English only)		83.1
15	Zhuosheng Zhang	SemBERT		82.9
16	Danqi Chen	SpanBERT (single-task training)		82.8

Wang *et al*. (2019) *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*

Super-human NLP?

"the current state of the art GLUE Score as of early July 2019 […] surpasses human performance […] by 1.3 points, and in fact exceeds this human performance estimate on four tasks." (Wang *et al*. 2019)

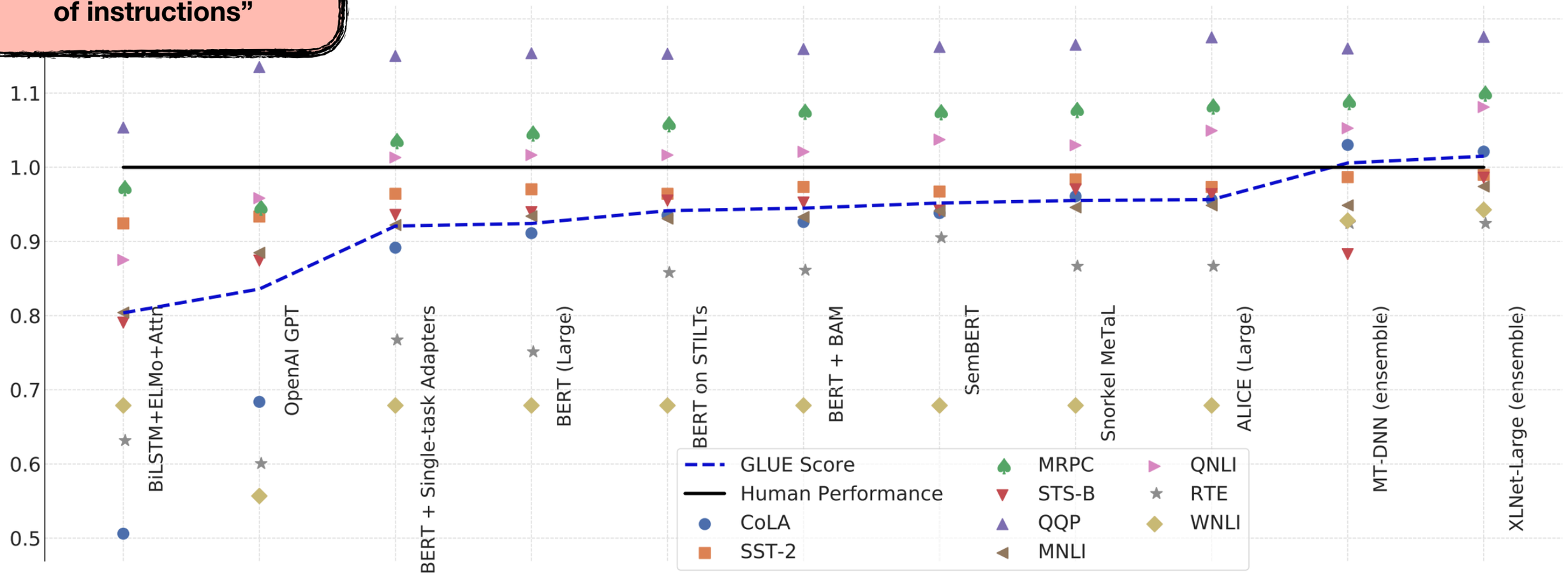


Wang *et al*. (2019) *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*

Super-human NLP?

crowdsourcing: "annotators are non-experts who must learn each task from a brief set of instructions"

"the current state of the art GLUE Score as of early July 2019 […] surpasses human performance […] by 1.3 points, and in fact exceeds this human performance estimate on four tasks." (Wang *et al*. 2019)



Wang *et al*. (2019) *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*

BERT

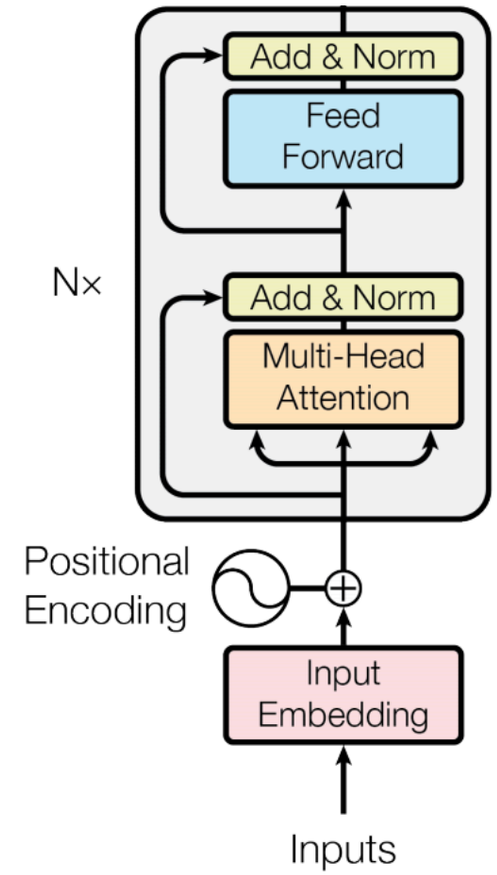
Bidirectional Encoder Representations from Transformers

Deep *bidirectional* language model using transformer encoder

Trained using **masked language modeling** and **next sentence prediction** objectives

English model pretrained on 3B words (Wikipedia + BooksCorpus)

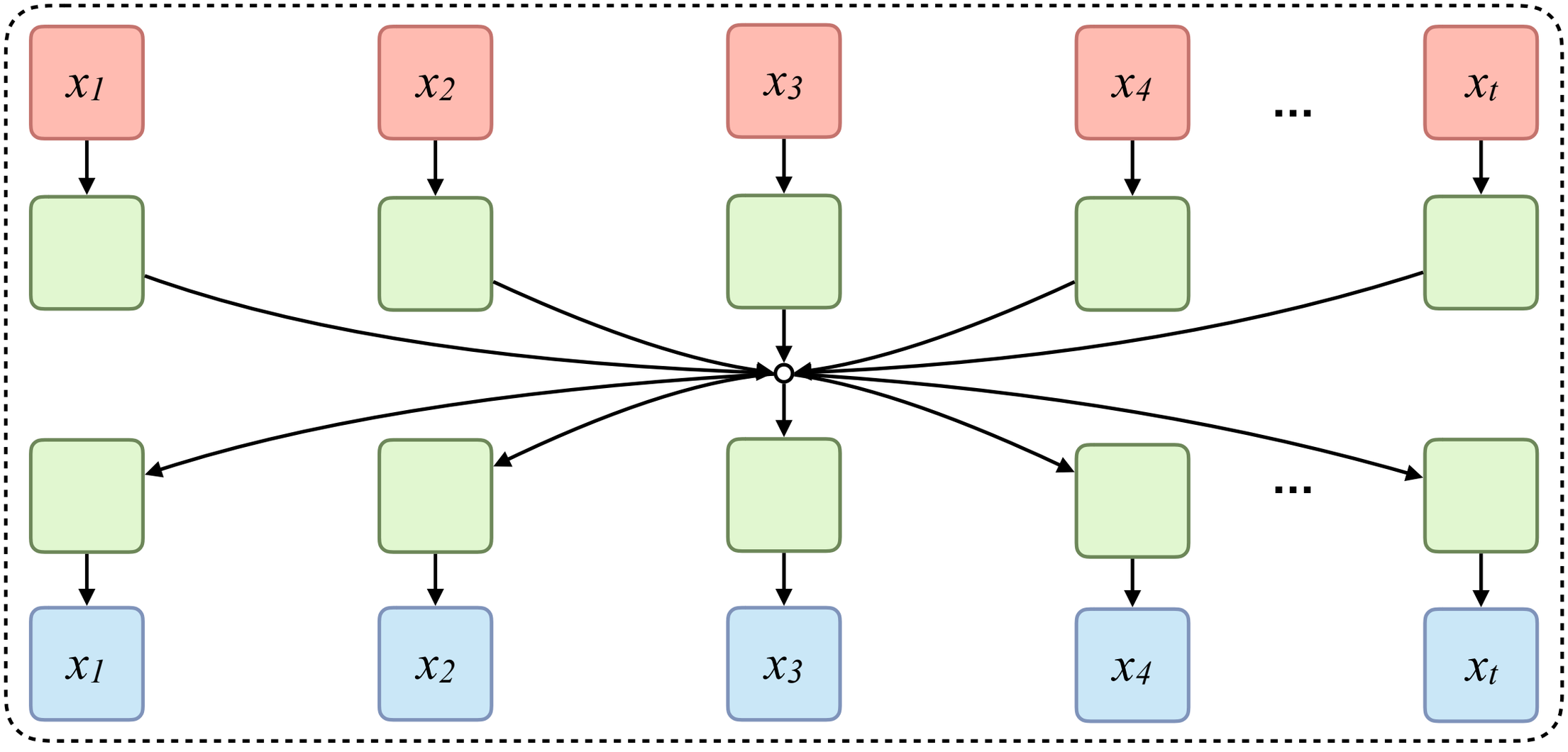
Fine-tuning for various language understanding tasks showed results *surpassing human performance*



Devlin *et al*. (2019) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

BERT

Challenge: words can “see themselves” in bidirectional LM training

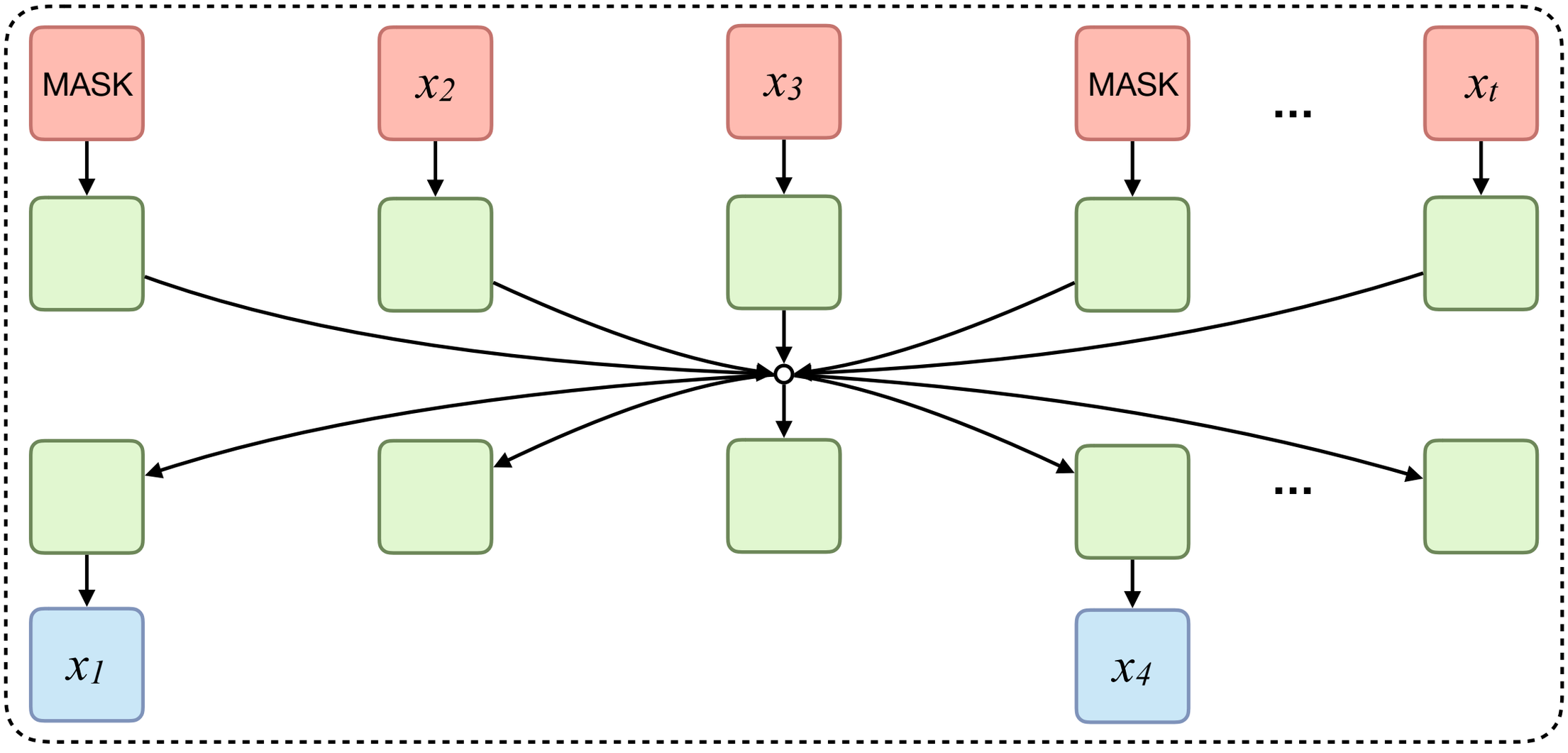


Devlin *et al*. (2019) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

BERT

Challenge: words can “see themselves” in bidirectional LM training

Solution: randomly mask some % of words and predict those



Devlin *et al*. (2019) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

BERT

Challenge: learning relationships between sentences

Solution: next sentence prediction

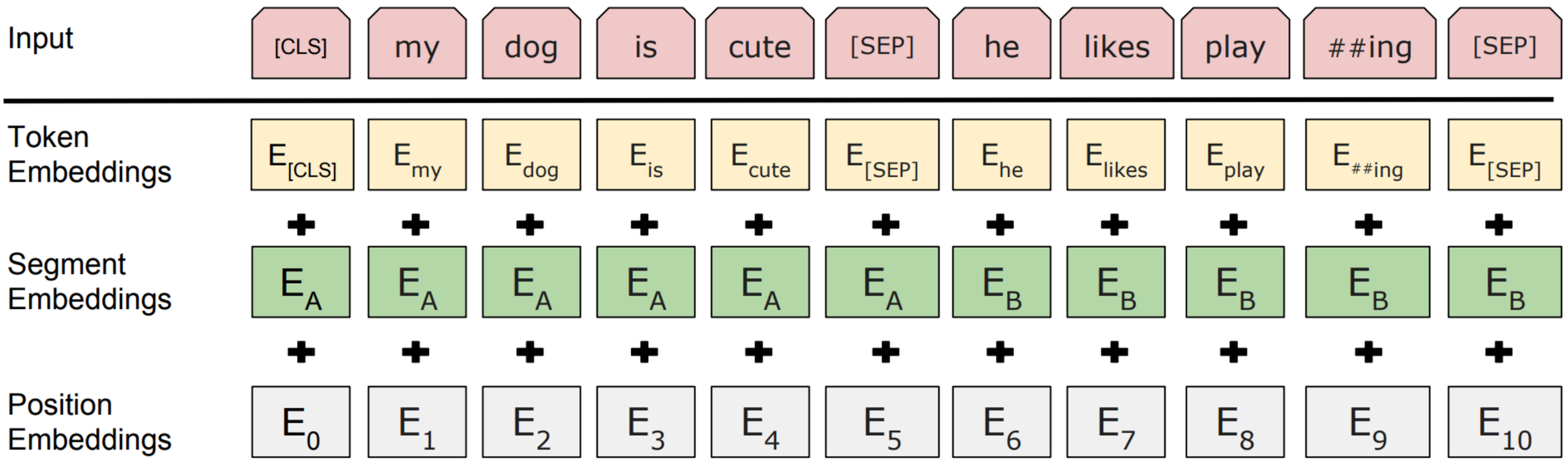


Figure from Devlin *et al*. (2019) *BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding*

BERT

Challenge: rare and unknown words

Solution: subword tokenization

Byte-pair encoding (Gage 1994): find most common pair of consecutive bytes in corpus, replace with new byte, repeat

```
onerously dogmatic iconoclast
```

⬇

```
one ##rous ##ly dog ##matic icon ##oc ##last
```

Gage (1994) *A New Algorithm for Data Compression*
Sennrich *et al*. (2016) *Neural Machine Translation of Rare Words with Subword Units*

BERT models

BERT models released by Google (max sequence length 512):

Base: 12 layers, 768-dimensional state, 110M parameters, English, Chinese, and multilingual (104 languages!)

Large: 24 layers, 1024-dimensional state, 340M parameters, English

(**Tiny**: 2L/128D; **Mini**: 4L/256D; **Small**: 4L/512D, **Med**: 8L/512D, …)

Models released by others:

BERTje (Dutch), BETO (Spanish), CamemBERT/FlauBERT (French), FinBERT (Finnish) RuBERT (Russian), Arabic, German and Swedish BERTs, …

“We do not plan to release more single-language models”
https://github.com/google-research/bert/blob/master/multilingual.md

BERT for Finnish: FinBERT

TurkuNLP trained BERT from scratch for Finnish to make FinBERT

Pre-trained on 3B words of Finnish (web crawl, news, social media)

Trained for 12 days on 8 V100 GPUs

Outperforms Google's multilingual BERT and previous SOTA on range of NLP tasks

http://turkunlp.org/FinBERT/ (open!)

TURKUNLP .ORG

CSC

Virtanen *et al*. (2019) *Multilingual is not enough: BERT for Finnish*

Applications

Neural LMs such as BERT applicable to a range of downstream tasks

Sentence and sentence pair classification, e.g. sentiment, entailment, paraphrase detection

Sequence tagging, e.g. NER, part of speech

More broadly: any task that represents words as vectors

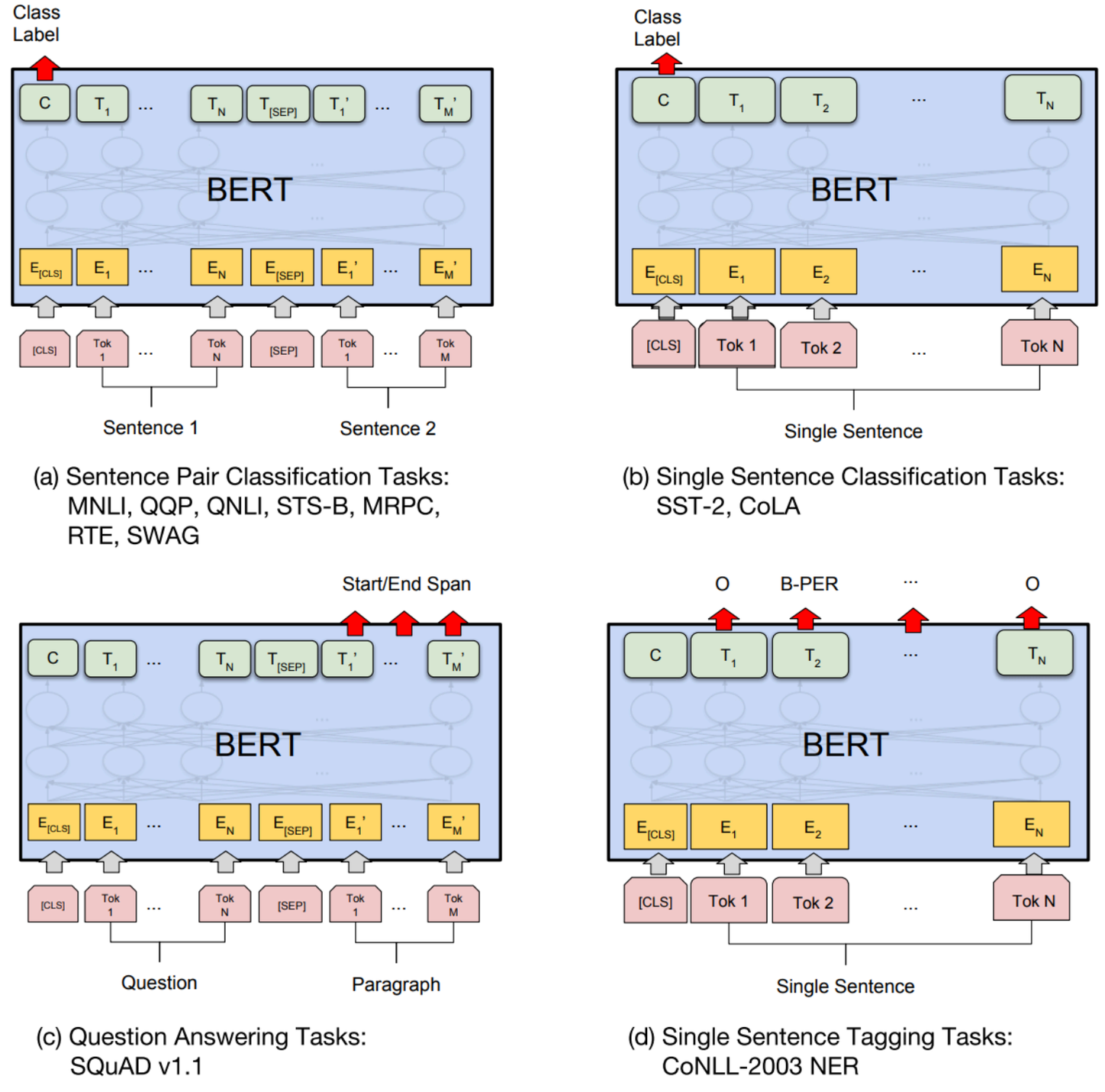


Figure from Devlin *et al*. (2019) *BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding*

Applications: MT

Deep transfer learning models, many based on transformers, have continued to advance the SOTA in machine translation

(Table: En-De MT progress)

Transformer models are used e.g. by Google Translate and Microsoft Translator

Model	BLEU	Paper / Source
Transformer Big + BT (Edunov et al., 2018)	35.0	Understanding Back-Translation at Scale
DeepL	33.3	DeepL Press release
MUSE (Zhao et al., 2019)	29.9	MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning
DynamicConv (Wu et al., 2019)	29.7	Pay Less Attention With Lightweight and Dynamic Convolutions
AdvSoft + Transformer Big (Wang et al., 2019)	29.52	Improving Neural Language Modeling via Adversarial Training
Transformer Big (Ott et al., 2018)	29.3	Scaling Neural Machine Translation
RNMT+ (Chen et al., 2018)	28.5*	The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation
Transformer Big (Vaswani et al., 2017)	28.4	Attention Is All You Need
Transformer Base (Vaswani et al., 2017)	27.3	Attention Is All You Need

Table from http://nlpprogress.com/english/machine_translation.html

Applications: web search

BERT used to interpret and disambiguate Google search queries

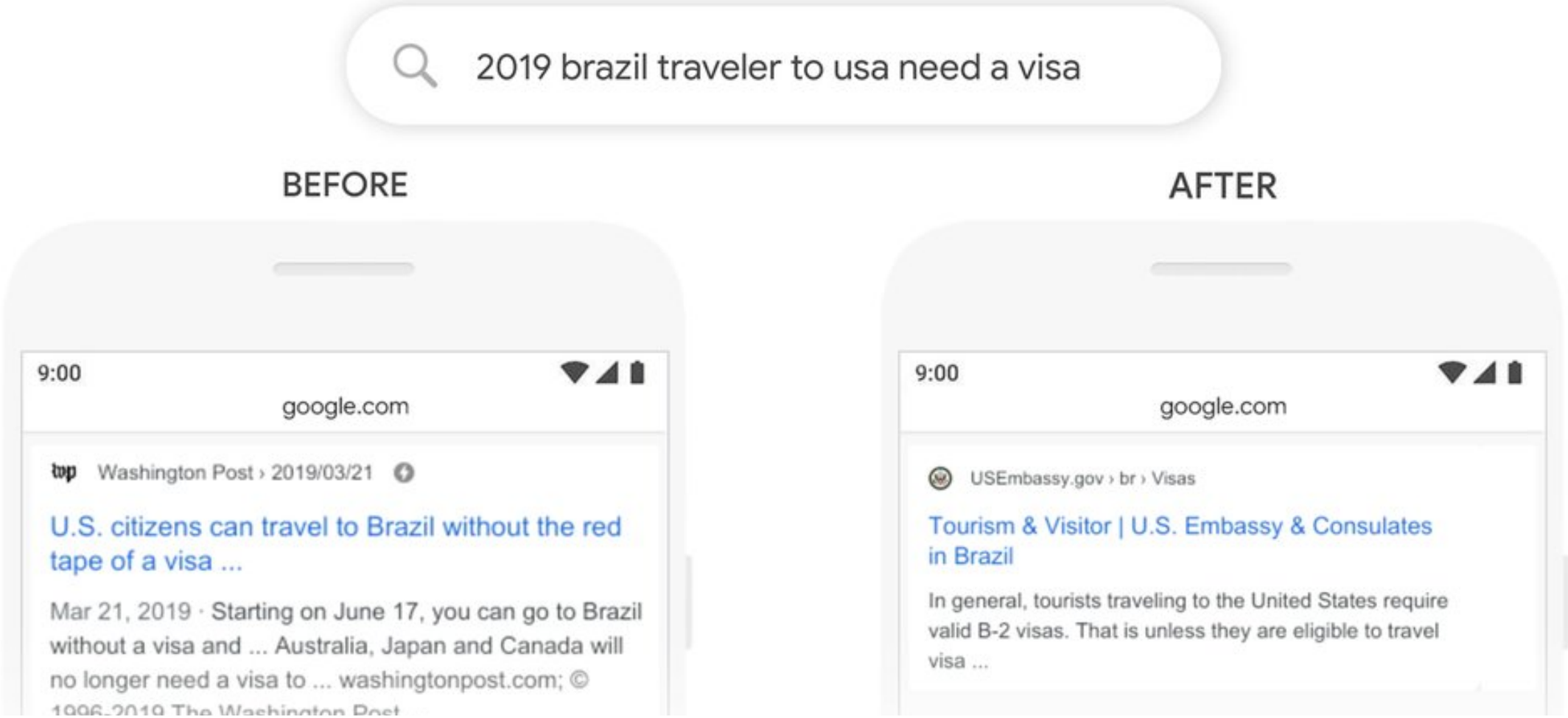


Figure from https://www.blog.google/products/search/search-language-understanding-bert/

Applications: generation

PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

…

From GPT-2, the model (previously) "too dangerous to release"

Example from https://openai.com/blog/better-language-models/

Applications: generation

PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns

Now, after almost tw is finally solved.

Dr. Jorge Pérez, an e companions, were exploring the Andes s or humans. Pérez noticed that the valley wo peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

…

Demo: https://talktotransformer.com/

From GPT-2, the model (previously) "too dangerous to release"

Example from https://openai.com/blog/better-language-models/

Applications: QA

EXAMPLE

The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of “one world, one dream”. Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the “Journey of Harmony”, lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.

[…]

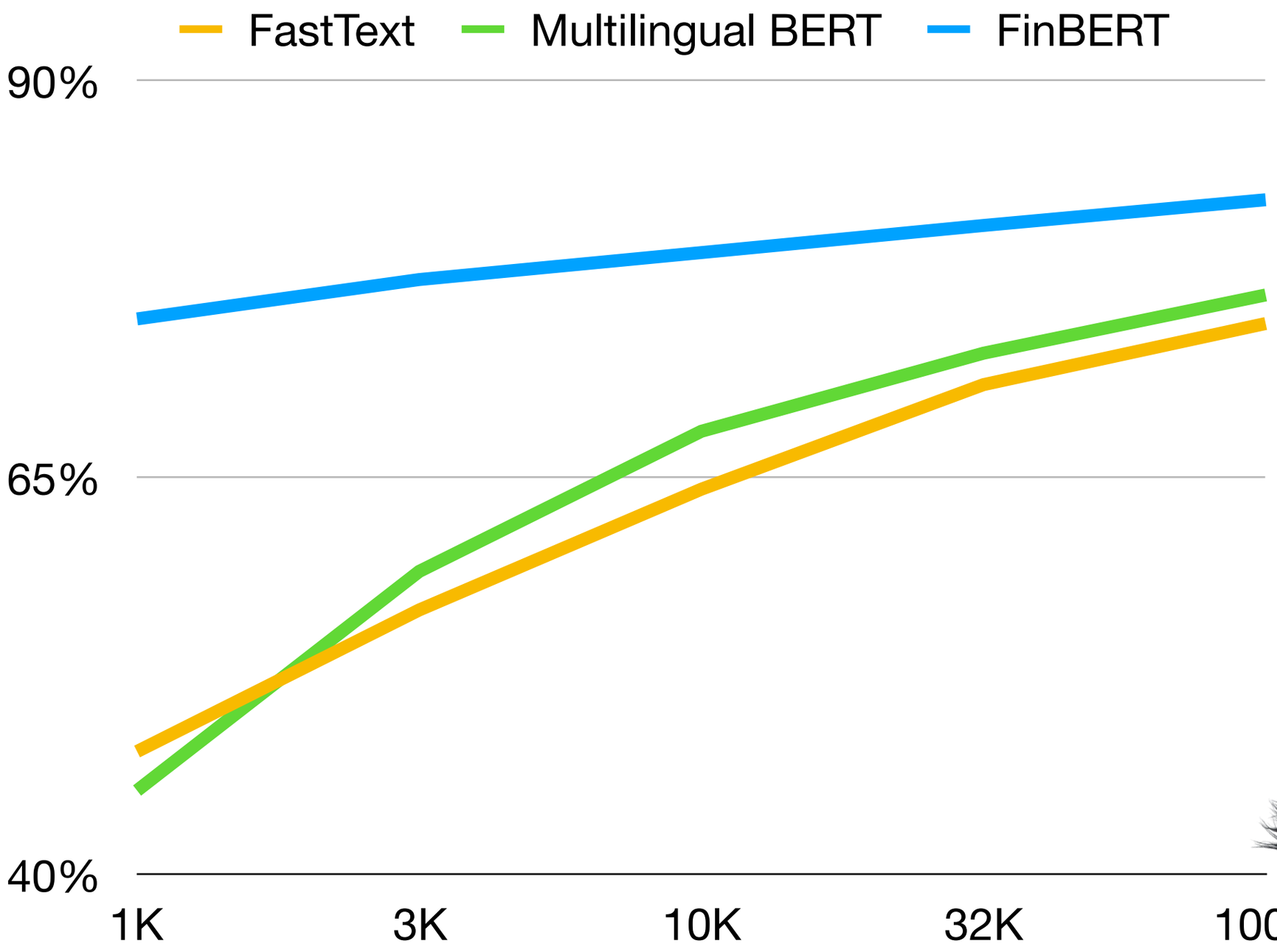
Q: What was the theme?
A: “one world, one dream”.

Q: What was the length of the race?
A: 137,000 km

Q: Was it larger than previous ones?
A: No

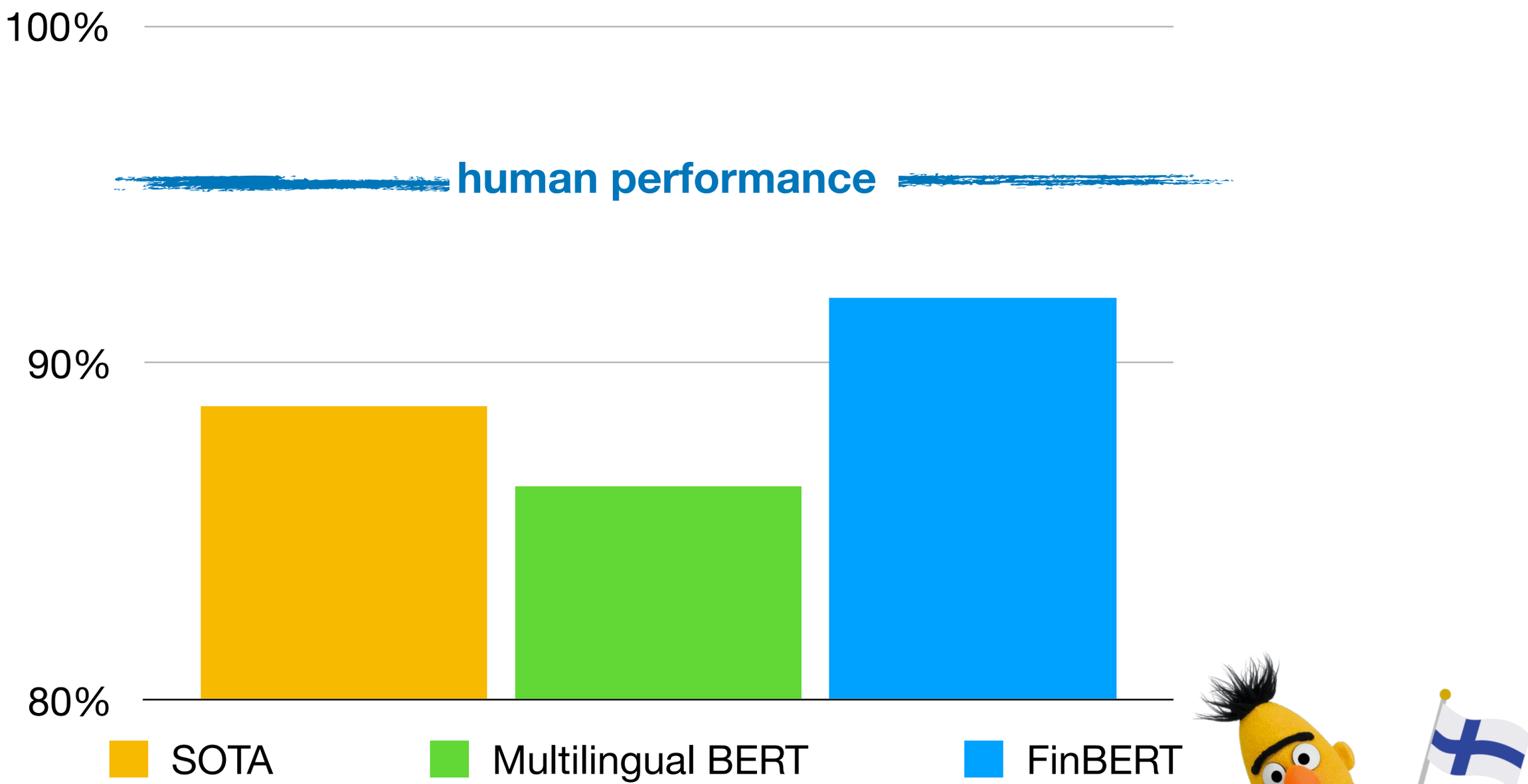
Example from https://openai.com/blog/better-language-models/

Applications: text classification



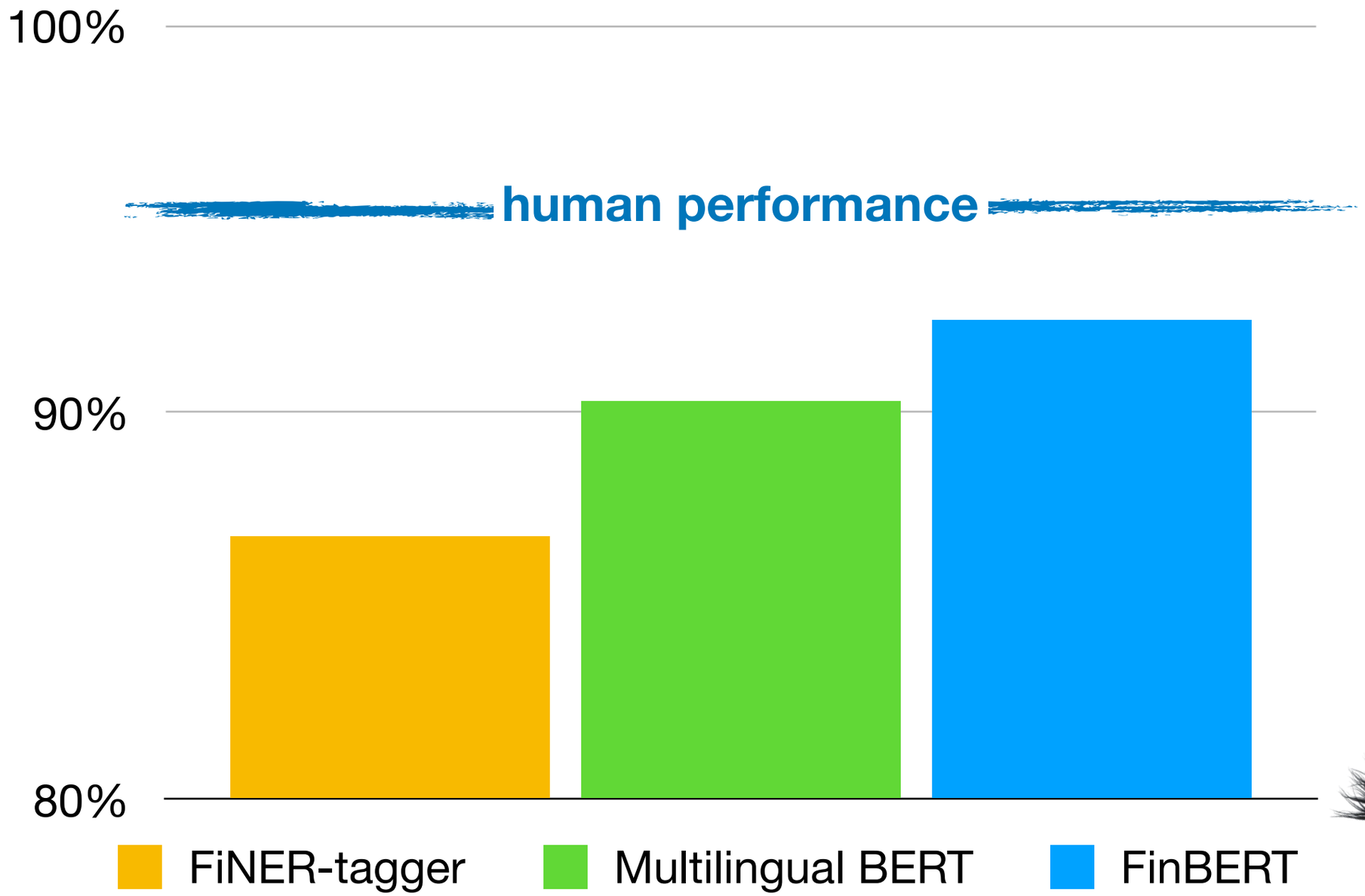
Virtanen *et al*. (2019) *Multilingual is not enough: BERT for Finnish*

Applications: parsing



Virtanen *et al*. (2019) *Multilingual is not enough: BERT for Finnish*

Applications: NER



Luoma *et al*. (2020) *A Broad-coverage Corpus for Finnish Named Entity Recognition*

Applications: images

Image completion (left) and super-resolution (right)

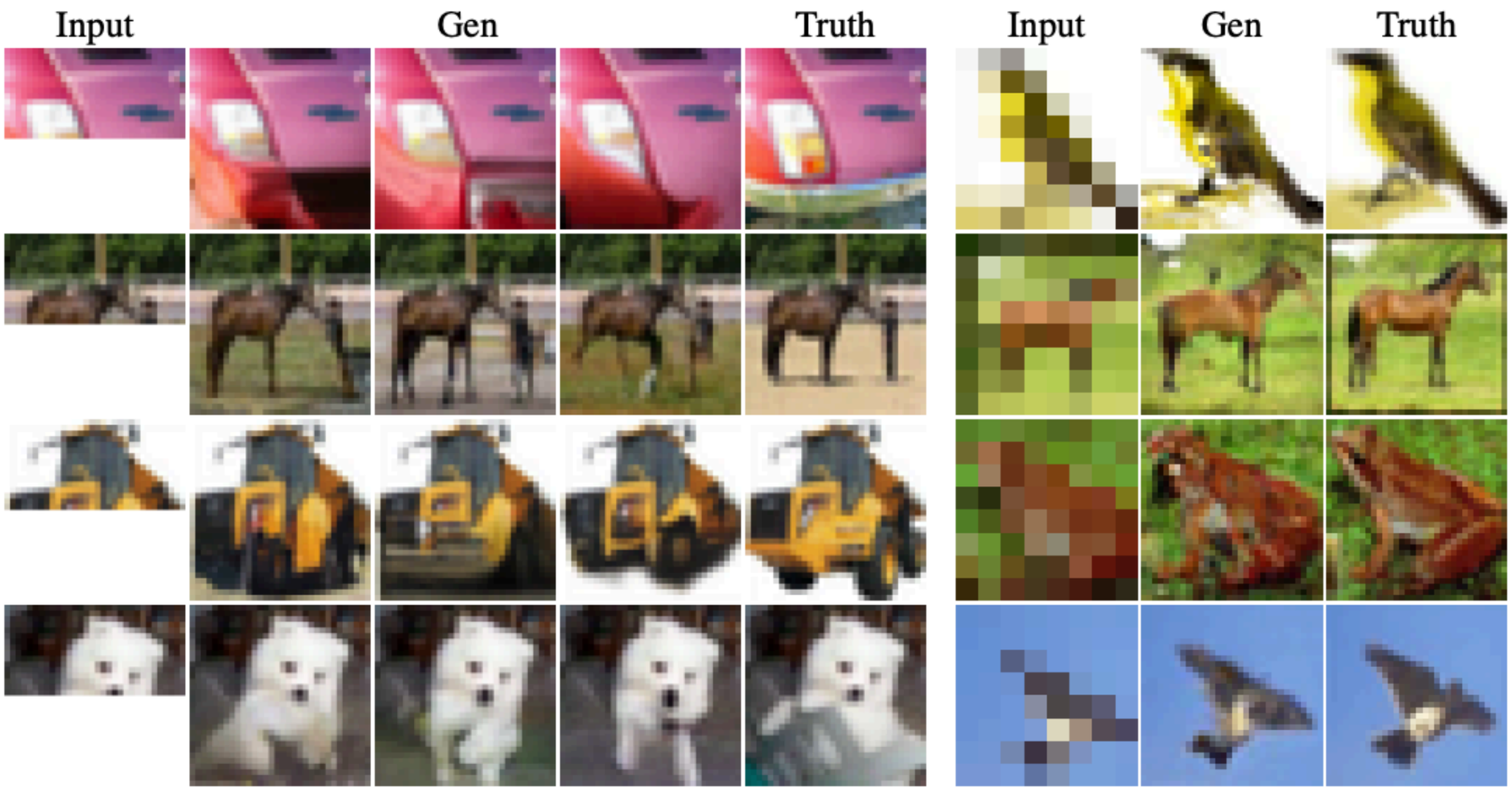
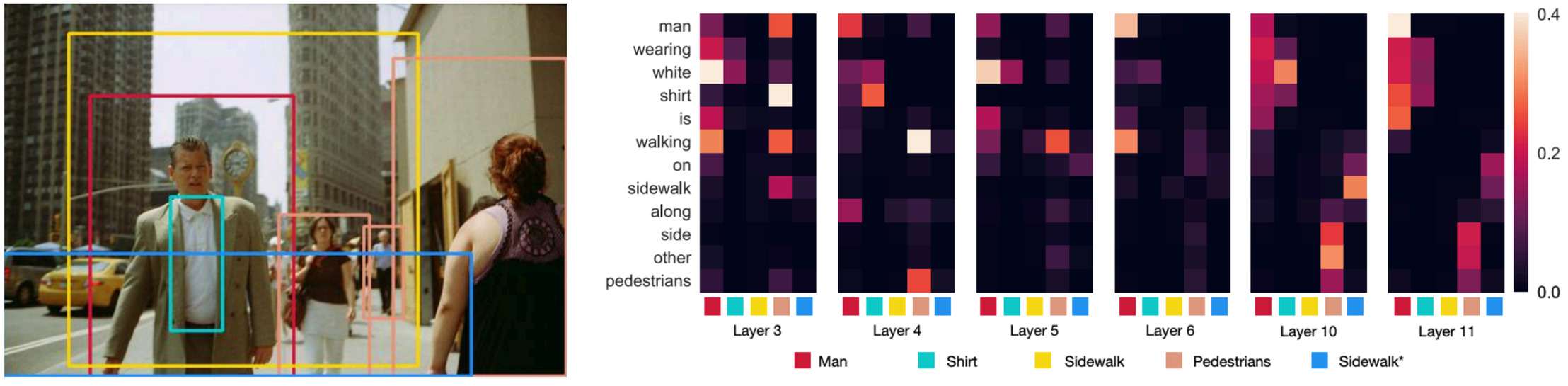
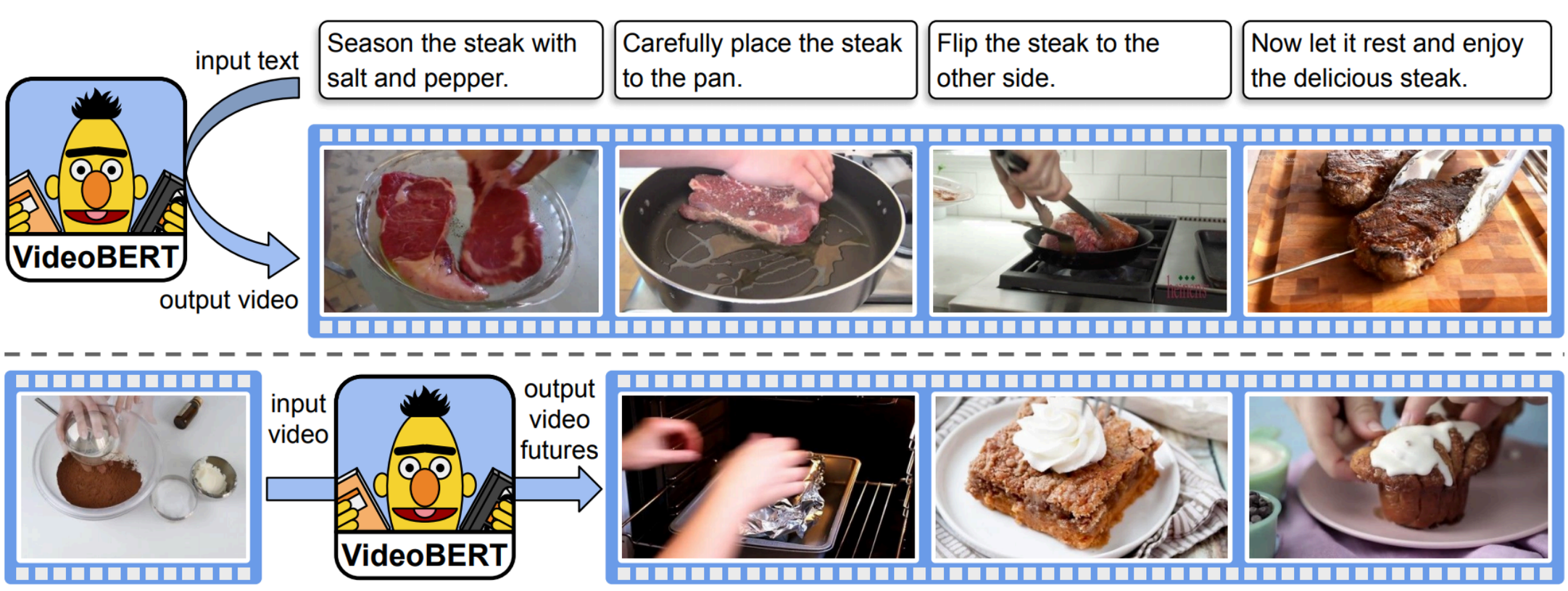


Figure from Parmar *et al*. (2018) *Image Transformer*

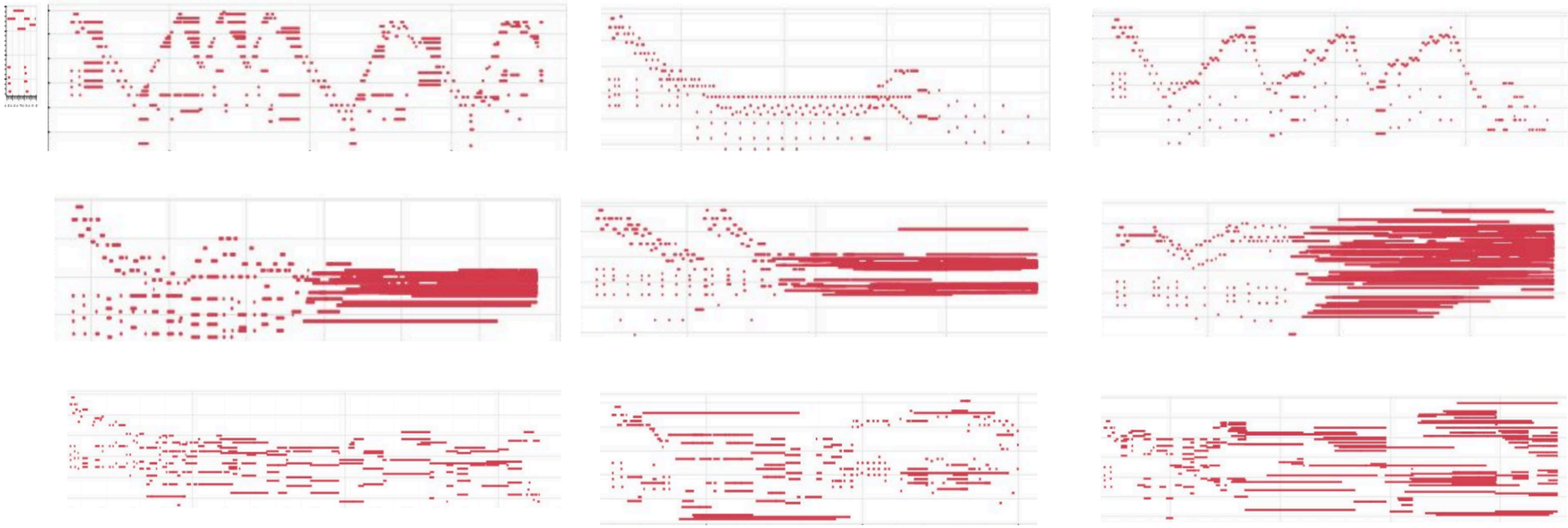
Applications: image/video





Figures from Li *et al*. (2019) *VisualBERT: A Simple and Performant Baseline for Vision and Language* (top) and Sun *et al*. (2019) *VideoBERT: A Joint Model for Video and Language Representation Learning* (bottom)

Applications: music generation



https://magenta.tensorflow.org/music-transformer

Figure from Huang *et al*. (2019) *Music Transformer*

https://github.com/jessevig/bertviz

http://exbert.net/

Visualization

Illustration of attention pattern on one layer of transformer stack

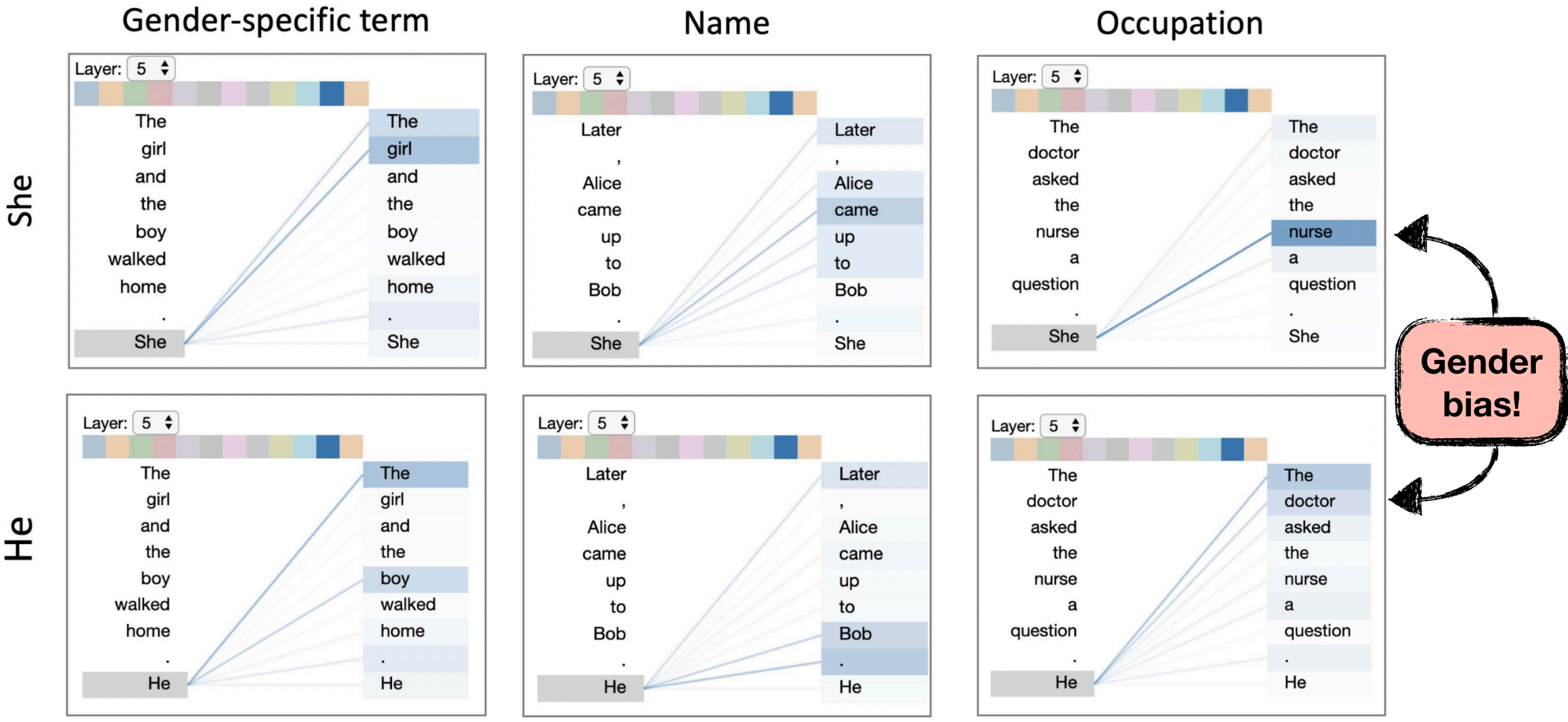


Figure from Vig (2019) *A Multiscale Visualization of Attention in the Transformer Model*

Trends: model parameters

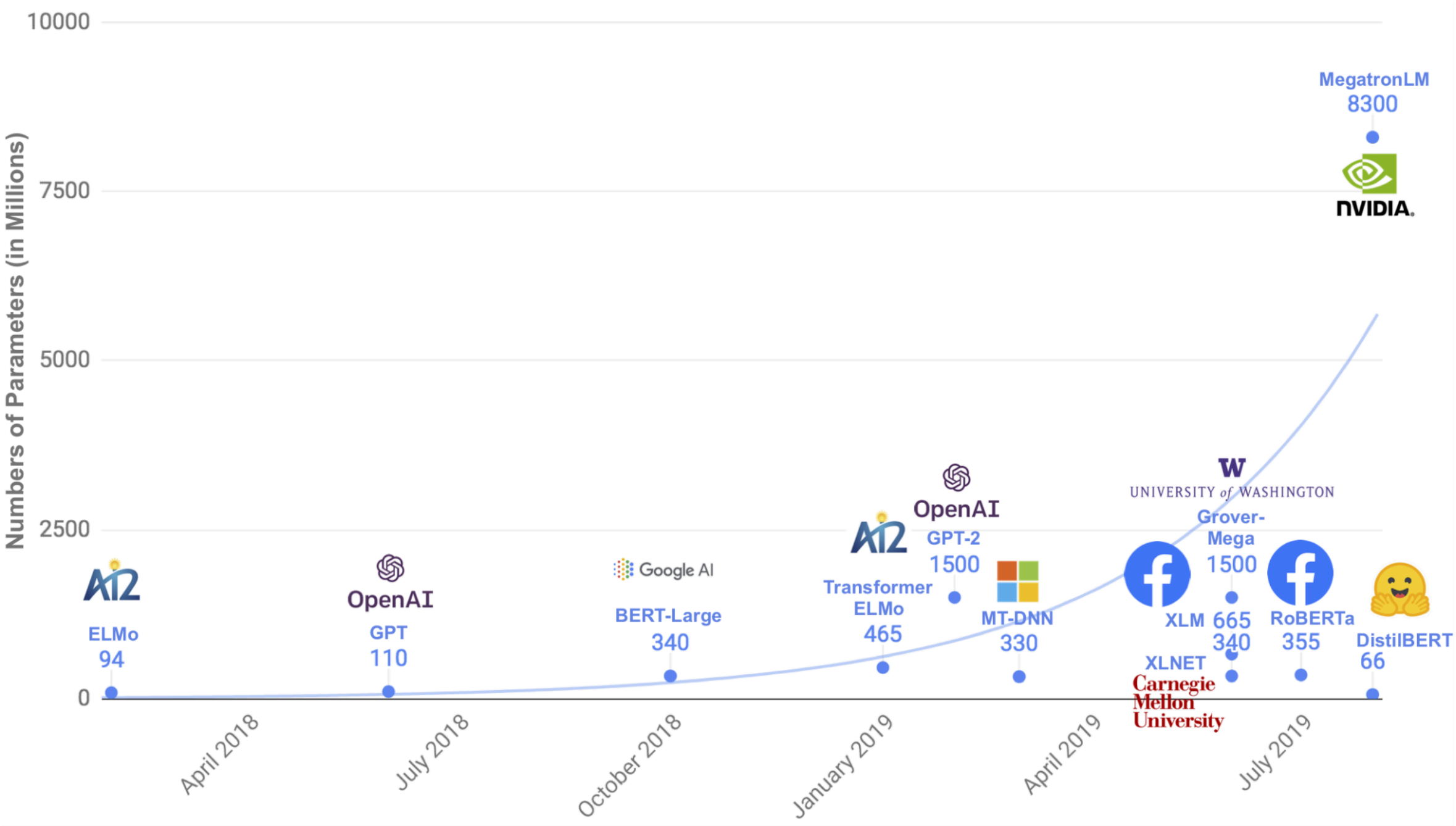


Figure from Sans *et al*. (2020) *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*

Trends: pretraining data

Model	Words (billions)
ELMo	1
BERT	3
GPT-2	7
C4	120

C4: Raffel *et al*. (2019) *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*

Trends: compute

since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time
https://openai.com/blog/ai-and-compute/

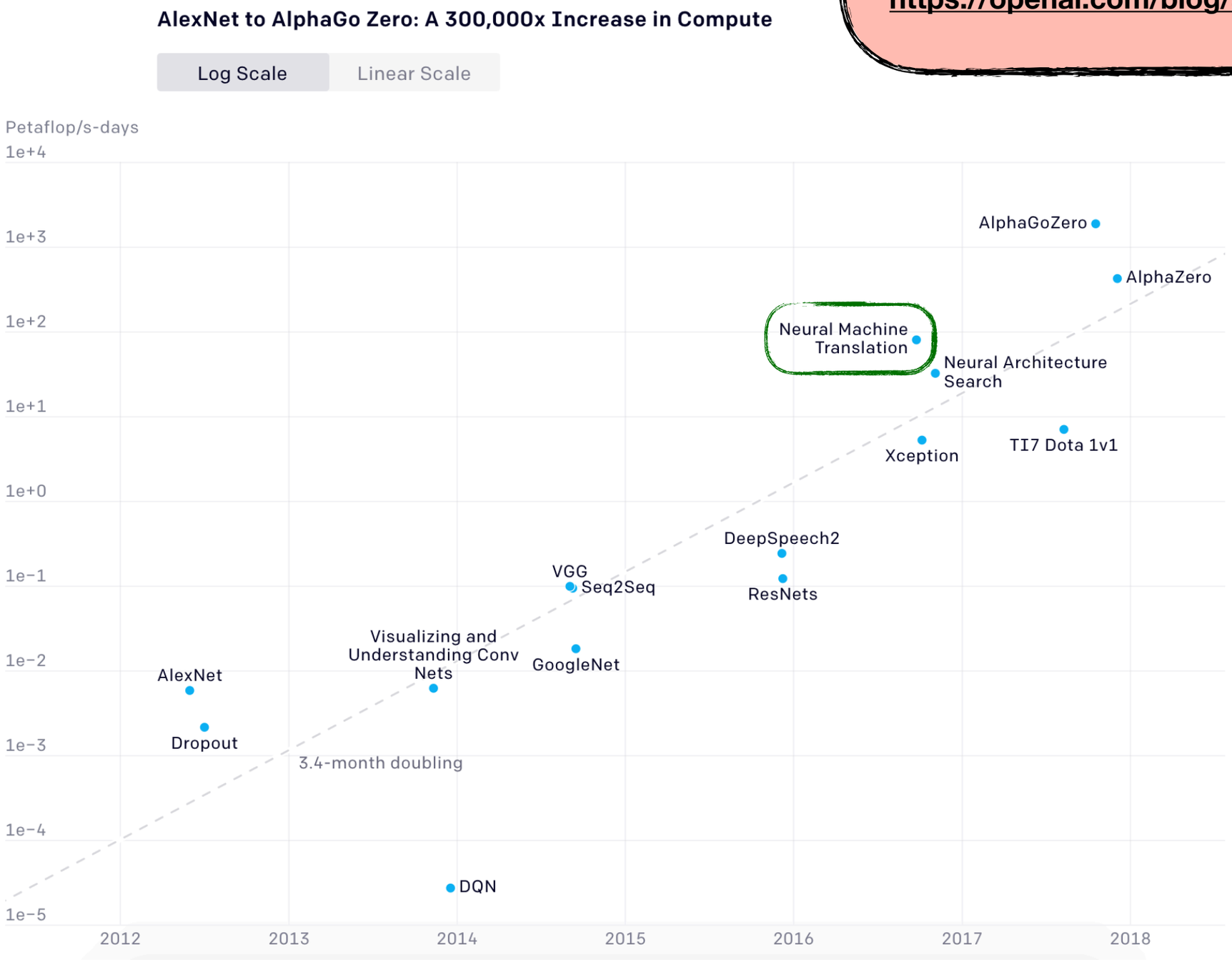
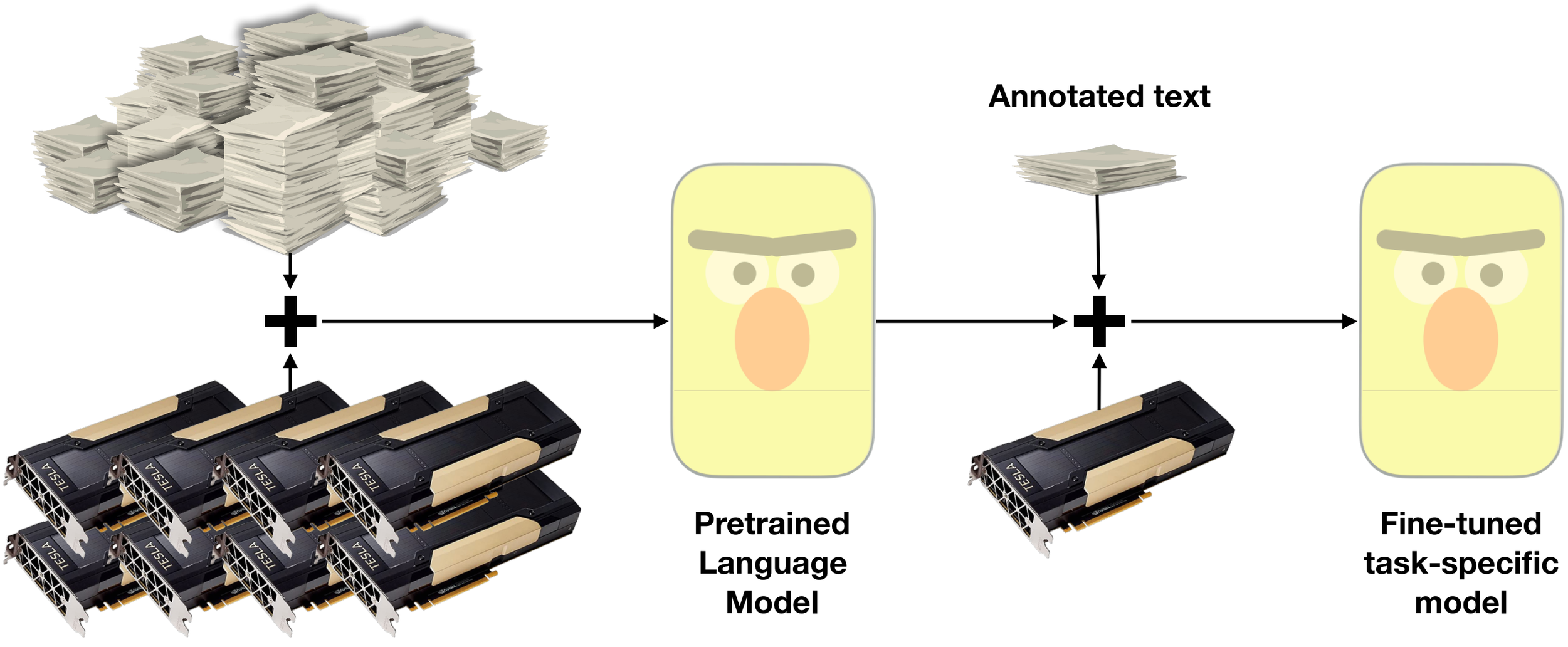


Figure from https://openai.com/blog/ai-and-compute/

Deep transfer learning for NLP

Unannotated text (billions of words)



Bert figure from http://jalammar.github.io/illustrated-bert/

Deep transfer learning for NLP

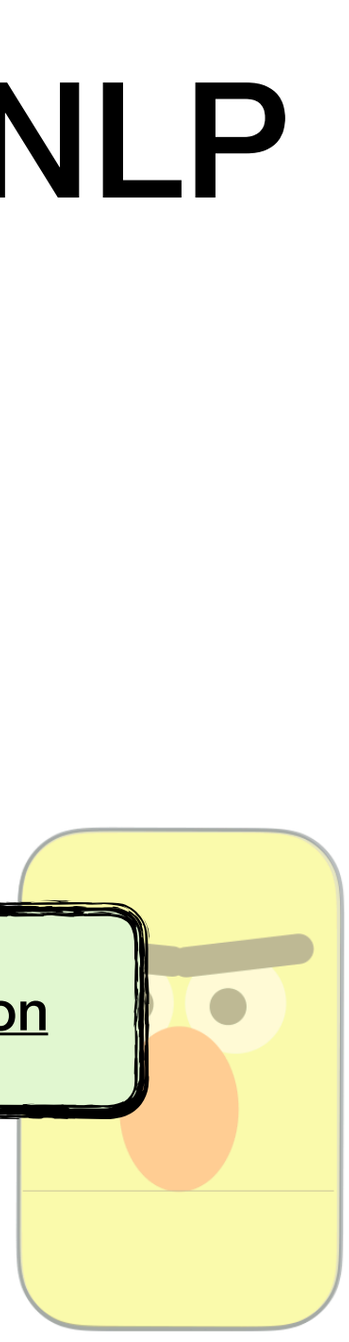
Unannotated text (billions of words)

https://www.csc.fi/-/supertietokone-puhti-on-avattu-tutkijoiden-kayttoon

Pretrained Language Model

Annotated text

Fine-tuned task-specific model



Bert figure from http://jalammar.github.io/illustrated-bert/

Related work

Melamud *et al*. (2016) *context2vec: Learning Generic Context Embedding with Bidirectional LSTM*

McCann *et al*. (2018) Learned in Translation: Contextualized Word Vectors [CoVe]

Howard and Ruder (2018) *Universal Language Model Fine-tuning for Text Classification* [ULMFiT]

Radford *et al*. (2019) *Language Models are Unsupervised Multitask Learners* [GPT-2]

Dai *et al.* (2019) Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

Liu *et al*. (2019) RoBERTa: *A Robustly Optimized BERT Pretraining Approach*

Lan *et al*. (2019) ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Clark *et al*. (2020) *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*

Resources

Google AI Blog: Transformer: A Novel Neural Network Architecture for Language Understanding
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Alexander Rush: The Annotated Transformer
https://nlp.seas.harvard.edu/2018/04/03/attention.html

Jay Alammar: The Illustrated Transformer
http://jalammar.github.io/illustrated-transformer/

Google AI: Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing
https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

OpenAI: Better Language Models and Their Implications
https://openai.com/blog/better-language-models