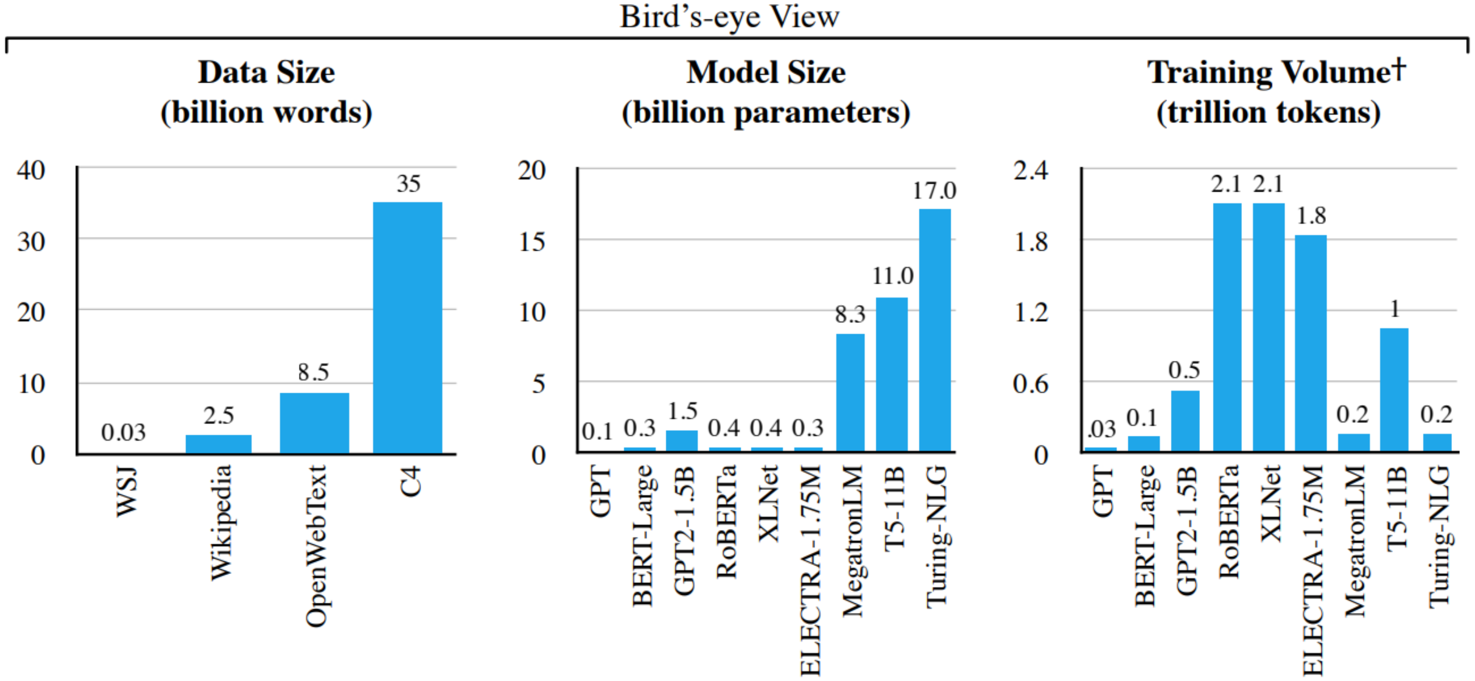
Big Models & More

Yuxiao Hu

May, 2021

In NLP, Everything is Big and Getting Bigger

credit: AI21 labs

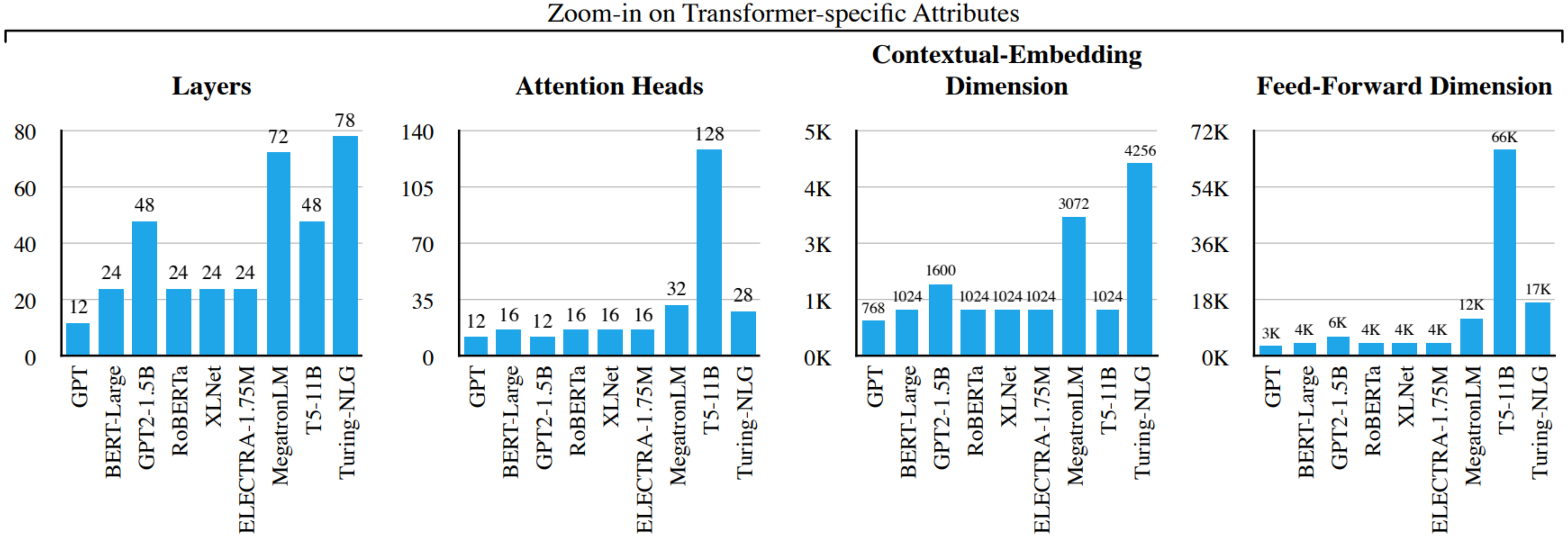


Bird's-eye View

Data Size (billion words)

Model Size (billion parameters)

Training Volume† (trillion tokens)



Zoom-in on Transformer-specific Attributes

Layers

Attention Heads

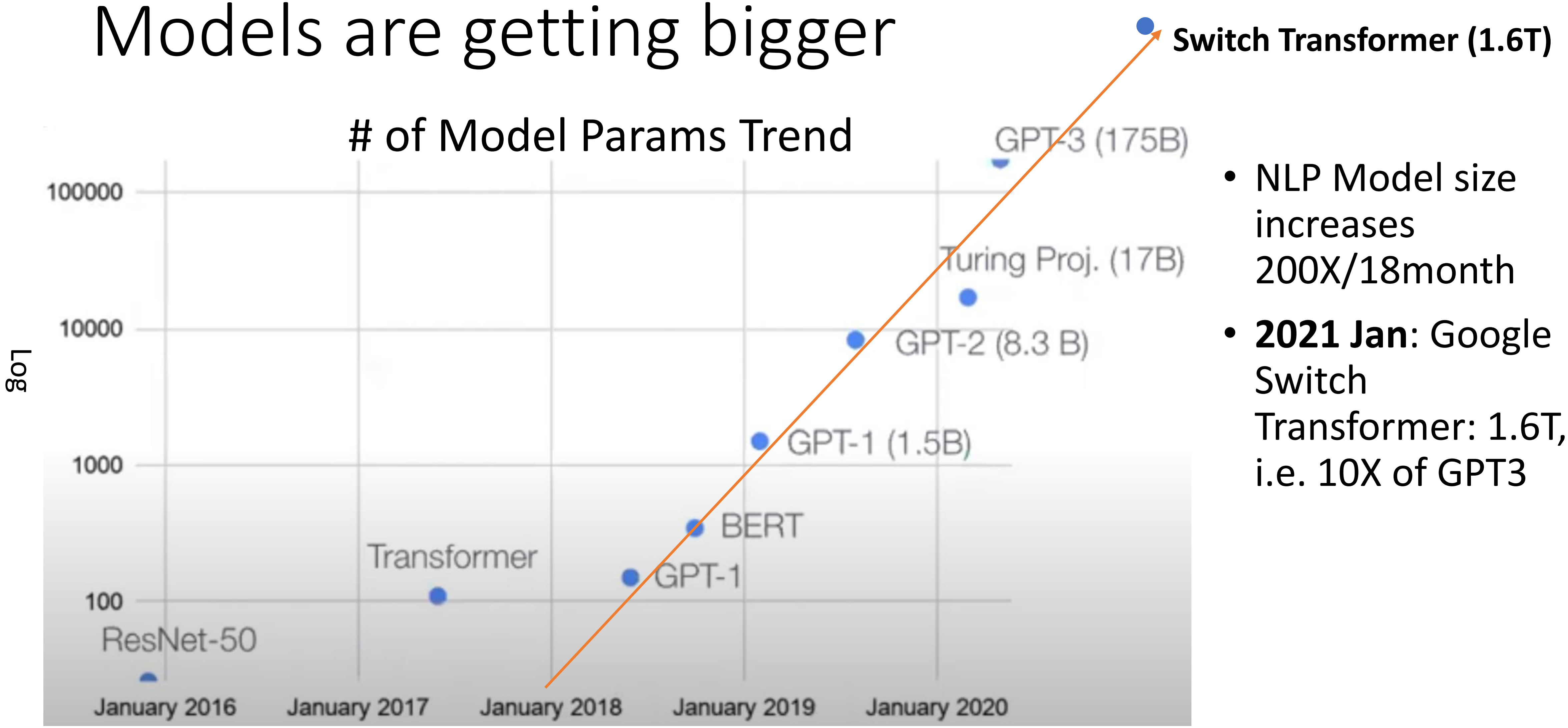
Contextual-Embedding Dimension

Feed-Forward Dimension

Outline

- Big Model:
  - what
  - who
  - why
  - how
- Current Progress
- Future Directions

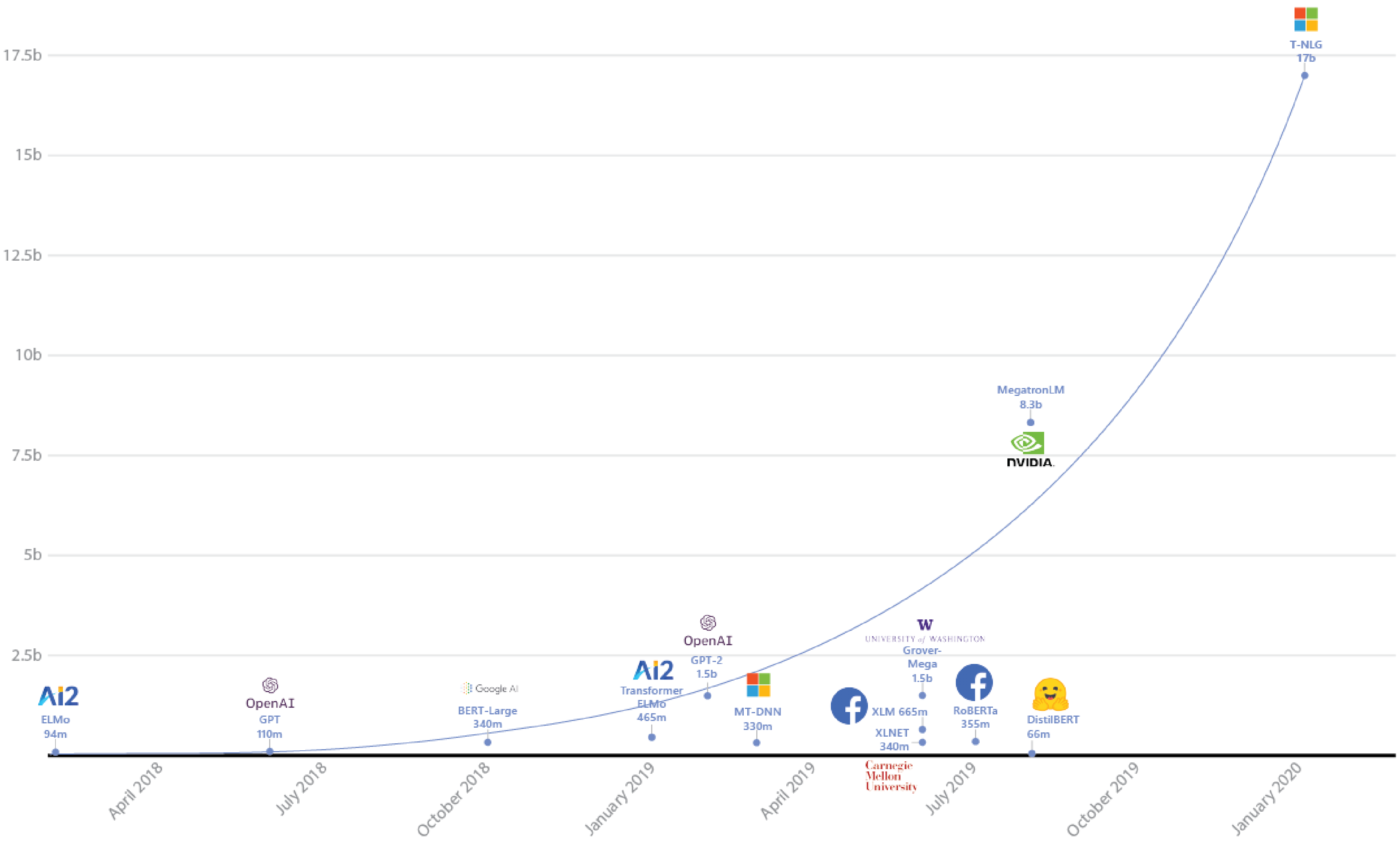
Models are getting bigger



- NLP Model size increases 200X/18month
- **2021 Jan**: Google Switch Transformer: 1.6T, i.e. 10X of GPT3

https://developer.nvidia.com/blog/training-bert-with-gpus/
https://arxiv.org/pdf/2101.03961.pdf

Who are the players?



What is BigModel, e.g. GPT?

- Algorithm : [2005.14165] Language Models are Few-Shot Learners (arxiv.org)
- Model: Generative Pre-trained Transformer 3 (GPT-3) is a new language model created by OpenAI that is able to generate written text of such quality that is often difficult to differentiate from text written by a human.
- OpenAI APIs:
  - Classification
    - Tweet sentiment
    - Company categorization
    - Labeling parts of speech
  - Generation
    - Idea generator
  - Conversation
    - Q&A agent
    - Sarcastic chatbot
  - Transformation
    - Summarize text
    - English -> French
    - Movie Titles -> Emoji
  - Completion
    - Generate react components
  - Factual responses
    - Provide factual answers
- License:
  - License: Microsoft exclusively license GPT-3 language model from OpenAI

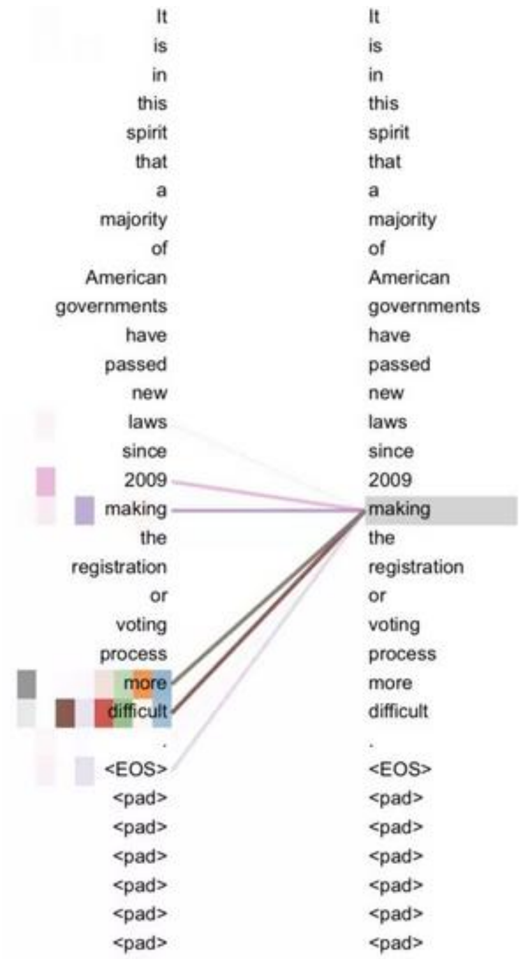
What can big models do?

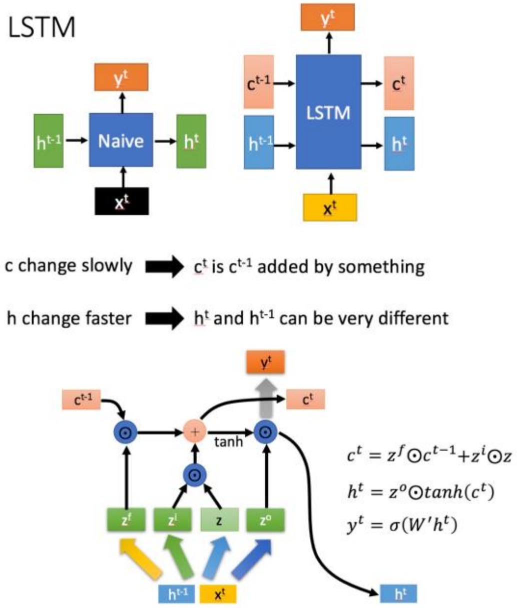
- Benchmarks
  - XTREME, SuperGLUE, GEM, SQuAD, SWAG, …
- Applications:
  - Search engine
  - Voice assistant
  - Office/Productivity
  - Software development
  - Research
  - Media (news/documents/books/etc.)
- Demos
  - GPT-3 playground
  - Debuild.co: describe what your app should do in plain English, then start using it
  - GPT-3 Examples

Why Big Models? A little bit history

- Image/NLP/Speech
- CNN, RNN, LSTM, Transformer
- Deep, Deeper, Wider, Complex
    - 2012, AlexNet, VGG, Inception, etc.
    - 2015, ResNet 18, 34, 50, 101, 152, 1001,…
    - 2017, Attention: LSTM, GRU, Transformer (6 Layers)
    - 2018, Pretraining : BERT(24 Layers, 340M)
    - 2020, Turing(78 Layers, 17B):
    - 2020,July, GPT(3: 96Layers, 170B)

Why Big Models? Current Techniques





c change slowly ➡ c^t is c^{t-1} added by something

h change faster ➡ h^t and h^{t-1} can be very different

$$c^t = z^f \odot c^{t-1} + z^i \odot z$$

$$h^t = z^o \odot tanh(c^t)$$

$$y^t = \sigma(W'h^t)$$

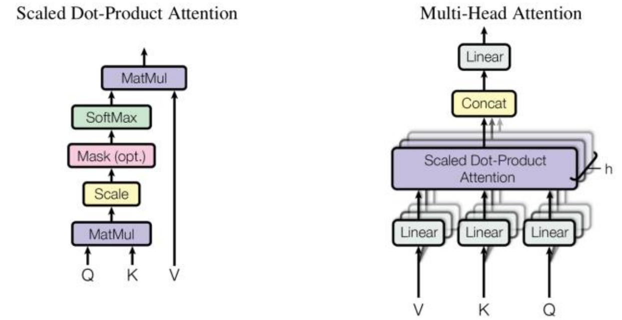


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

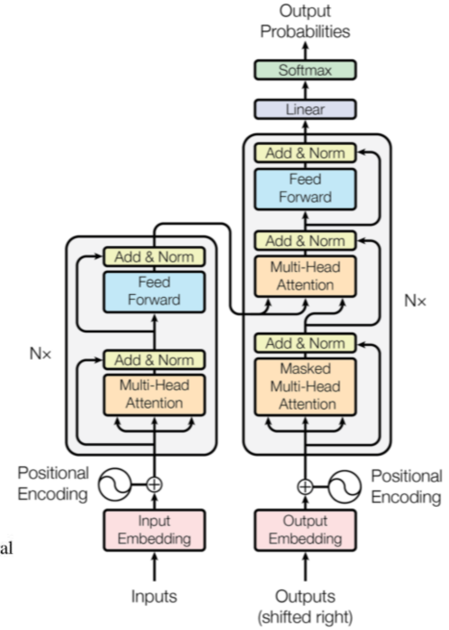
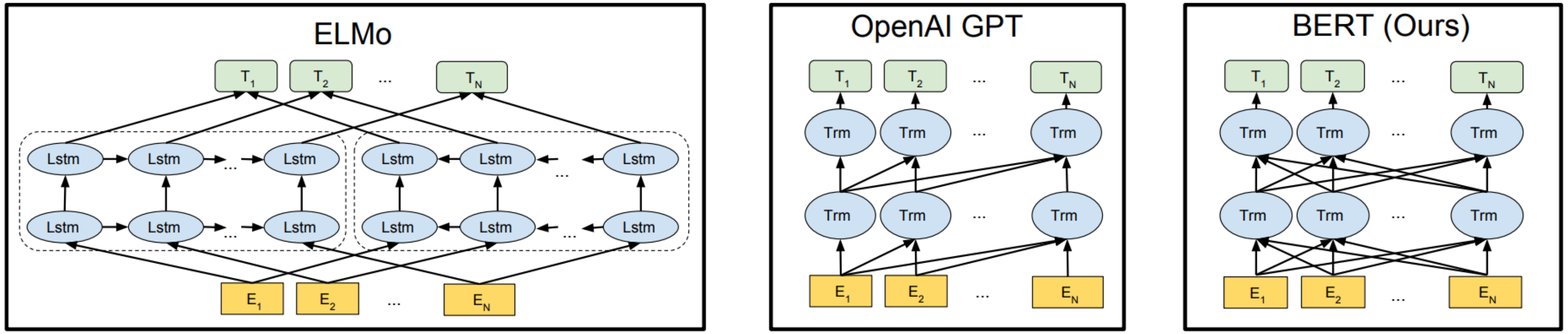
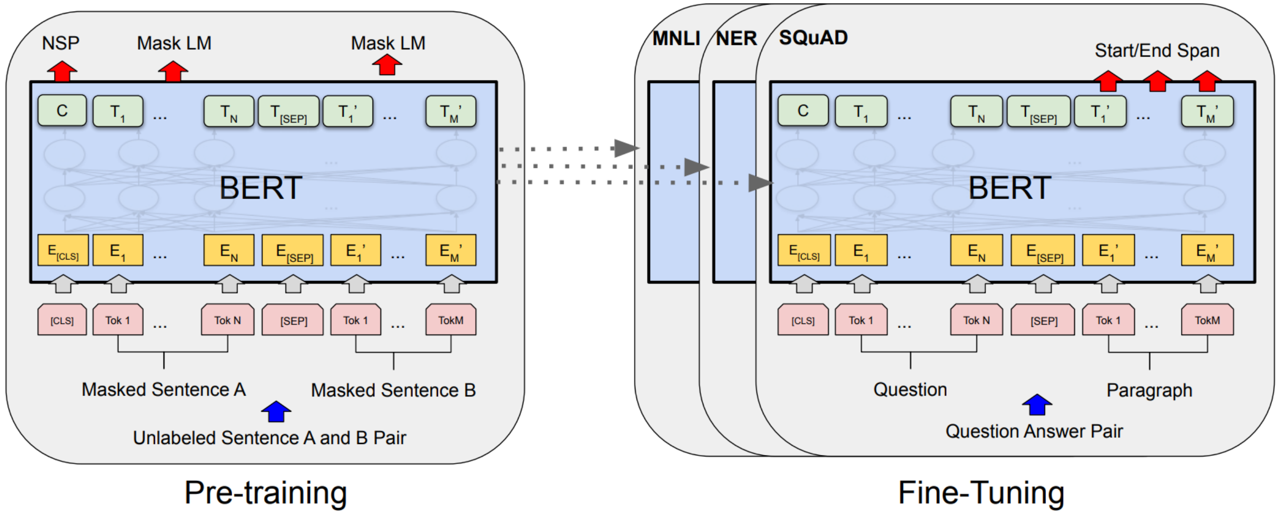


Figure 1: The Transformer - model architecture.



How to Train/Serve Such Big Models?

- Pretrain
  - Loading data batch
  - Forward, Lost, Gradients, Update
- Finetune/Retrain
- Hardware
  - TPU-v4: ~250TFlops, POD(x4096): 1exaFlops
    - Google TPU v4 Puts Supercomputer Power In The Google Cloud
  - GPU A100: ~20TFlops, DGX-2(x8): ~156TFlops, Clusters(x2048)
    - https://blogs.microsoft.com/ai/openai-azure-supercomputer/



How Much: $Cost to train big models

- “Price”:
  - $2.5k - $50k (110 million parameter model)
  - $10k - $200k (340 million parameter model)
  - $80k - $1.6m (1.5 billion parameter model)
- Examples:
  - Google-T5: $1.3M/model, $10M/project
  - OpenAI GPT-3
    - ~**$10 million** in expenses for research on GPT-3 and **training** the final model
    - **Tens of thousands of dollars** in monthly cloud computing or server and electricity costs for **running** the model
    - Possibly **more than a million dollars** in yearly **retraining** costs due to model decay
    - Additional costs of customer support, marketing, IT, security, legal and other requirements of running a product. This could be in the tens of thousands of dollars based on the number and size of customers OpenAI acquires.

Challenges for Infrastructure

- Storage: data/model
- Speed
  - Network/disk: Data/Model Loading
  - Compute: GPU / TPU
- Memory
  - Model Parameters, internal results
- Parrallel
- Reproduction

Even More Memory Needed for Training

- Data:
  - Parameters(Weights/Bias)
  - Gradients
  - Activation
  - Optimizer States
- Precision
  - Float
  - Int

Possible Solutions

- Parallelization
  - Data
  - Model
  - Pipeline
- Offloading
  - GPU→CPU
  - GPU Memory→ CPU Memory → SSD

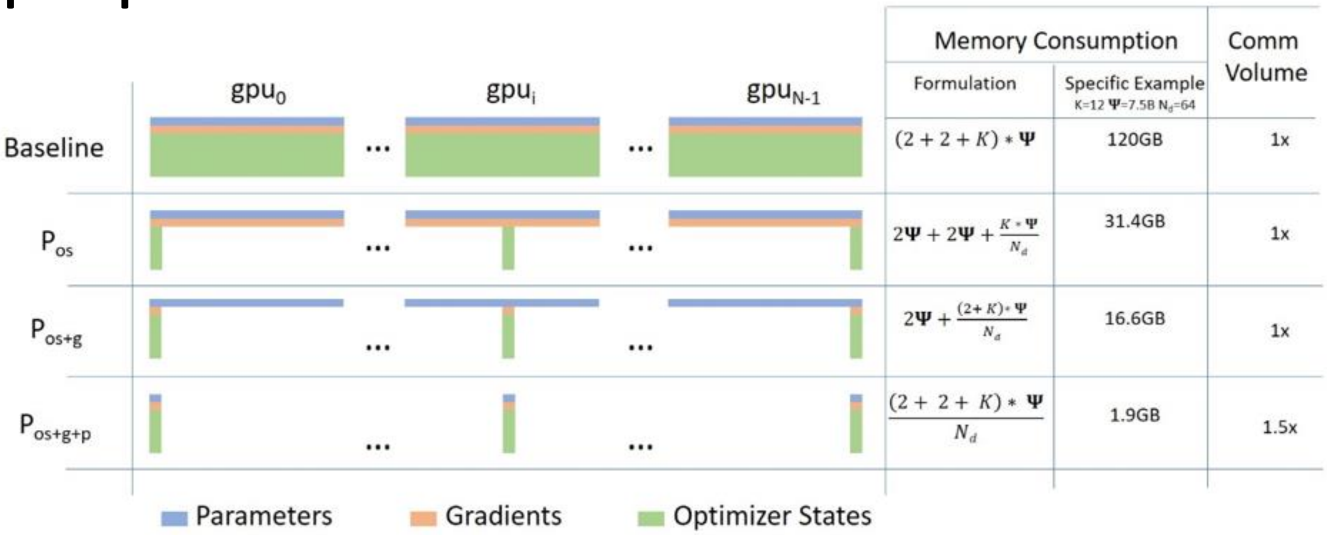
ZeRO + DeepSpeed

ZeRO 4-way data parallel training

Using:

- P_{os} (Optimizer state)
- P_g (Gradient)
- P_p (Parameters)

ZeRO+DeepSpeed



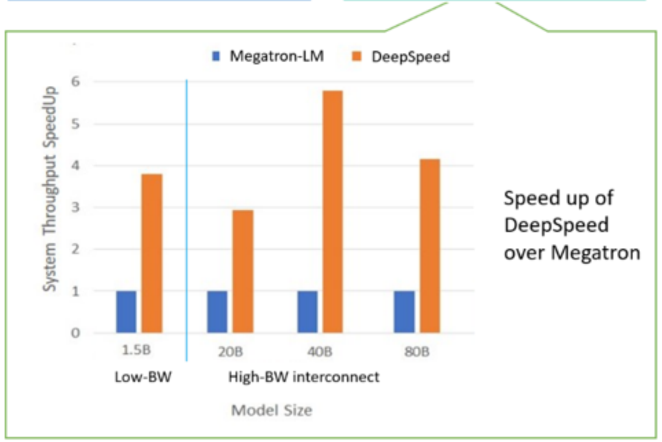
	Memory Consumption		Comm Volume
	Formulation	Specific Example K=12 Ψ=7.5B N_d=64	
Baseline	$(2+2+K)*\Psi$	120GB	1x
P_{os}	$2\Psi+2\Psi+\frac{K*\Psi}{N_d}$	31.4GB	1x
P_{os+g}	$2\Psi+\frac{(2+K)*\Psi}{N_d}$	16.6GB	1x
P_{os+g+p}	$\frac{(2+2+K)*\Psi}{N_d}$	1.9GB	1.5x

Scale
- 100B parameter
- 10X bigger

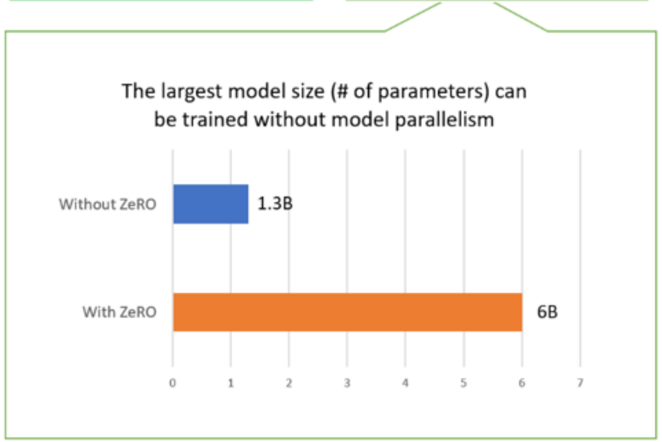
Speed
- Up to 5X faster

Cost
- Up to 5X cheaper

Usability
- Minimal code change



Speed up of DeepSpeed over Megatron


The largest model size (# of parameters) can be trained without model parallelism

Zero-Infinity DeepSpeed

PanGu Big Models

- PanGu-NLP:
  - PanGu-Alpha
    - 首个2000亿参数GPT-3，以中文为核心的预训练生成语言模型
    - 基于80T文本，1T，
    - 2048卡集群“鹏城云脑Ⅱ”
    - MindSpore框架的自动混合并行模式
    - Details: https://zhuanlan.zhihu.com/p/368261642
    - 部分开源：https://git.openi.org.cn/PCL-Platform.Intelligence/PanGu-Alpha
  - PanGu-Beta
    - 1000亿参数Transformer，主打理解类任务
    - 基于40TB文本, 600G
    - MindSpore+千张昇腾910训练1月+
    - Details: https://zhuanlan.zhihu.com/p/370336501
- PanGu-Vision
  - 30亿参数

https://www.huaweicloud.com/product/modelarts/mls.html

Limitations of Existing Big Models

- Disparity between Pre-training and down-stream tasks
- Disparity between Text/Speech/Conversation
- Data/Sample efficiency
- Learning = Understanding or Remembering ?
- Big

Future Directions

- Vision & Multi-Modality
- Faster/Stronger/MoreAccurate: Bigger?
- End the SOTA race: Benchmarks/Leaderboards
- Distillation/Compression
- Less Data/No Data?
- Cost-Reduction
- Non-Transformer/Non-DL?

Summary

- Big models are inspired by NLP, with many potential applications and businesses
- The state-of-the-art technique is Transformer with self-attentions mechanism
- Big model post challenges to training infrastructure, which demand large memory and fast computation
- Parallelization and Offloading can improve training speed and break memory limitation