Question Answering

Radoslav Neychev

Outline

- Question Answering
  - Problem statement and datasets overview
  - Potential solutions
  - Open-domain Question Answering

Question Answering: problem statement and datasets

A Brief History of Open-domain Question Answering

- Simmons et al. (1964) did first exploration of answering questions from an expository text based on matching dependency parses of a question and answer
- Murax (Kupiec 1993) aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing
- The NIST TREC QA track begun in 1999 first rigorously investigated answering fact questions over a large collection of documents
- IBM's Jeopardy! System (DeepQA, 2011) brought attention to a version of the problem; it used an ensemble of many methods
- DrQA (Chen et al. 2016) uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

source: [CS224n Lecture 10]

MCTest Reading Comprehension

Passage (*P*) + Question (*Q*) ⟹ Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.......

Q Why did Alyssa go to Miami?

A To visit some friends

source: CS224n Lecture 10

Stanford Question Answering Dataset

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

- What causes precipitation to fall?
  - gravity
- What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
  - graupel
- Where do water droplets collide with ice crystals to form precipitation?
  - within a cloud

source: SQuAD: 100,000+ Questions for Machine Comprehension of Text

SQuAD evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate Precision, Recall and harmonic mean F1.

    Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a, an, the** only)

source: CS224n Lecture 10

SQuAD v1.1 leaderboard, end of 2016

		EM	F1
11	Fine-Grained Gating Carnegie Mellon University (Yang et al. '16)	62.5	73.3
12	Dynamic Chunk Reader IBM (Yu & Zhang et al. '16)	62.5	71.0
13	Match-LSTM with Ans-Ptr (Boundary) Singapore Management University (Wang & Jiang '16)	60.5	70.7
14	Match-LSTM with Ans-Ptr (Sequence) Singapore Management University (Wang & Jiang '16)	54.5	67.7
15	Logistic Regression Baseline Stanford University (Rajpurkar et al. '16)	40.4	51.0

Will your model outperform humans on the QA task?

		EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.3	91.2

source: CS224n Lecture 10

SQuAD v1.1 leaderboard, (May 2020)

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Apr 10, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP*	**90.202**	**95.379**
2 May 21, 2019	XLNet (single model) *Google Brain & CMU*	89.898	95.080
3 Dec 11, 2019	XLNET-123++ (single model) *MST/EOI* http://tia.today	89.856	94.903
3 Aug 11, 2019	XLNET-123 (single model) *MST/EOI*	89.646	94.930
4 Sep 25, 2019	BERTSP (single model) *NEUKG* http://www.techkg.cn/	88.912	94.584
4 Jul 21, 2019	SpanBERT (single model) *FAIR & UW*	88.839	94.635
5 Jul 03, 2019	BERT+WWM+MT (single model) *Xiaoi Research*	88.650	94.393

source: [SQuAD website]

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For **`NoAnswer`** examples, **`NoAnswer`** receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or "Answer validation"

source: [CS224n Lecture 10]

SQuAD 2.0 example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234 [from Microsoft nlnet]

source: [CS224n Lecture 10]

SQuAD 2.0 leaderboard (October 2020)

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
4 Sep 11, 2020	EntitySpanFocus+AT (ensemble) *RICOH_SRCB_DML*	90.454	92.748
4 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
5 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
5 Sep 27, 2020	electra+nlayers (ensemble) *oppo.tensorlab*	90.126	92.535

source: [SQuAD website]

Now in Russian: SberQuAD

Термин Computer science (Компьютерная наука) появился в 1959 году в научном журнале Communications of the ACM, в котором Луи Фейн (Louis Fein) ратовал за создание Graduate School in Computer Sciences (Высшей школы в области информатики) . . . Усилия Луи Фейна, численного аналитика Джорджа Форсайта и других увенчались успехом: университеты пошли на создание программ, связанных с информатикой, начиная с Университета Пердью в 1962.

- Q11870 Когда впервые был применен термин Computer science (Компьютерная наука)?
- Q28900 Кто впервые использовал этот термин?
- Q30330 Начиная с *каого** учебного заведения стали применяться учебные программы, связанные с информатикой?

*Misspelling is intended

source: SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis, link to SDSJ 2017 problem B

SberQuAD evaluation

Model	SberQuAD		SQuAD	
	EM	F1	EM	F1
simple baseline	0.3	25.0	–	–
ML baseline	3.7	31.5	–	–
BiDAF	51.7	72.2	68.0	77.3
DrQA	54.9	75.0	70.0	79.0
R-Net	58.6	77.8	71.3	79.7
DocQA	59.6	79.5	72.1	81.1
BERT	66.6	84.8	85.1	91.8

Table 7: Model performance on SQuAD and SberQuAD; SQuAD part shows single-model scores on test set taken from respective papers.

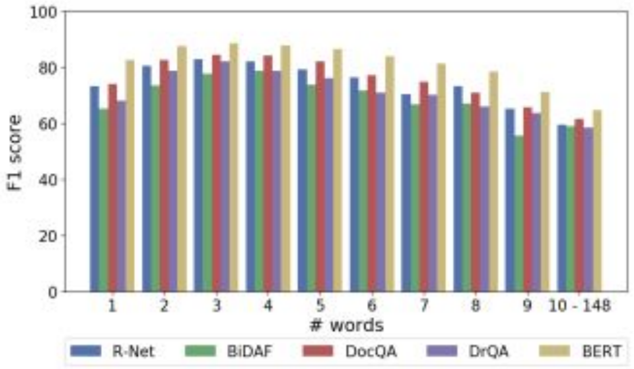


Figure 6: Model performance depending on answer length (# of words).

	% test	R-Net	BiDAF	DocQA	DrQA	BERT
w/ typos	5.7	74.1	66.7	77.5	67.5	81.1
correct	94.3	77.1	72.5	79.6	75.4	85.0
Test set		77.8	72.2	79.5	75.0	84.8

Table 8: Answer quality for misspelled questions.

source: [SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis]

But errors are still present

The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

Prediction: Ming dynasty [BERT (single model) (Google AI)]

source: CS224n Lecture 10

S(ber)QuAD limitations

- Only span-based answers (no yes/no, counting, implicit why)
    - Questions were constructed looking at the passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
    - Barely any multi-fact/sentence inference beyond coreference
- But these datasets are still of a great use

partial source: [CS224n Lecture 10]

Approaches to the Question Answering problem

Potential solutions

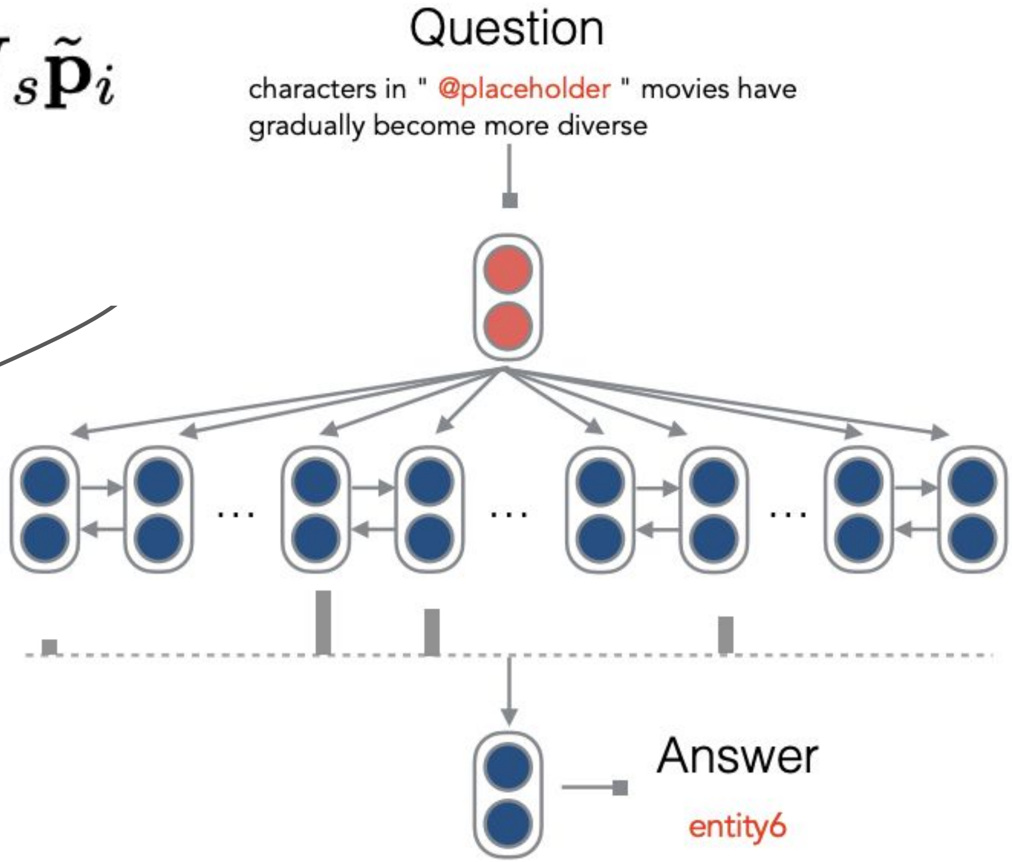
Passage

$$\alpha_i = \mathrm{softmax}_i \, \mathbf{q}^\intercal \mathbf{W}_s \tilde{\mathbf{p}}_i$$

$$\mathbf{o} = \sum_i \alpha_i \tilde{\mathbf{p}}_i$$

Two attention heads are used to find the ***start*** and ***end*** of the answer.

Attention is computed between encoded question and RNN state corresponding to every position.



source: [A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task]

How to make it better

- Use extra information about the text
  - Char embeddings
  - Linguistic features: PoS and NER tags
  - ...

PoS tagging

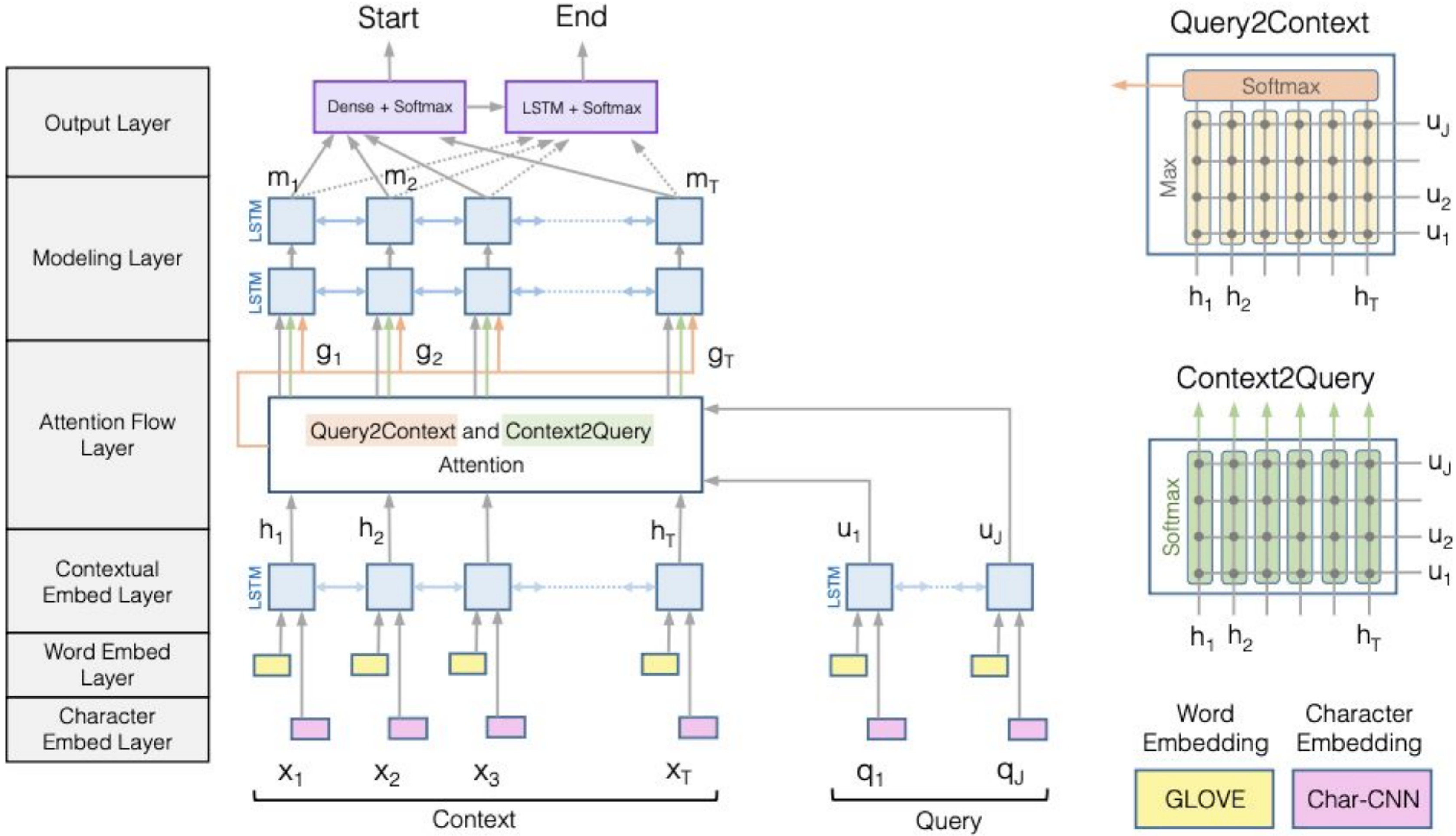
Pred. Tag	Actual Tag	Correct?	Token
PUNCT	PUNCT	✔	[
DET	DET	✔	this
NOUN	NOUN	✔	killing
ADP	ADP	✔	of
DET	DET	✔	a
ADJ	ADJ	✔	respected
NOUN	NOUN	✔	cleric
AUX	AUX	✔	will
AUX	AUX	✔	be
VERB	VERB	✔	causing
PRON	PRON	✔	us
NOUN	NOUN	✔	trouble
ADP	ADP	✔	for
NOUN	NOUN	✔	years
PART	PART	✔	to
VERB	VERB	✔	come
PUNCT	PUNCT	✔	.
PUNCT	PUNCT	✔	]

- PoS tagging can be performed using
  - Rule-based taggers
  - Dynamic programming
  - Models based on CRF (Conditional Random Field)
  - Neural Networks
  - etc.

How to make it better

- Use extra information about the text
  - Char embeddings
  - Linguistic features: PoS and NER tags
  - ...
- Better use of attention

BiDAF



source: [Bidirectional Attention Flow for Machine Comprehension]()

- There are variants of and improvements to the BiDAF architecture, but **the central idea** is the **Attention Flow layer**: attention should flow both ways – from the context to the question and from the question to the context
- Make similarity matrix (with w of dimension 6d):

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^T[\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

- Context-to-Question (C2Q) attention (which query words are most relevant to each context word):

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

source: [CS224n Lecture 10]

- Attention Flow:
  - attention should flow both ways – from the context to the question and from the question to the context
- Question-to-Context (Q2C) attention:
  - the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \mathrm{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

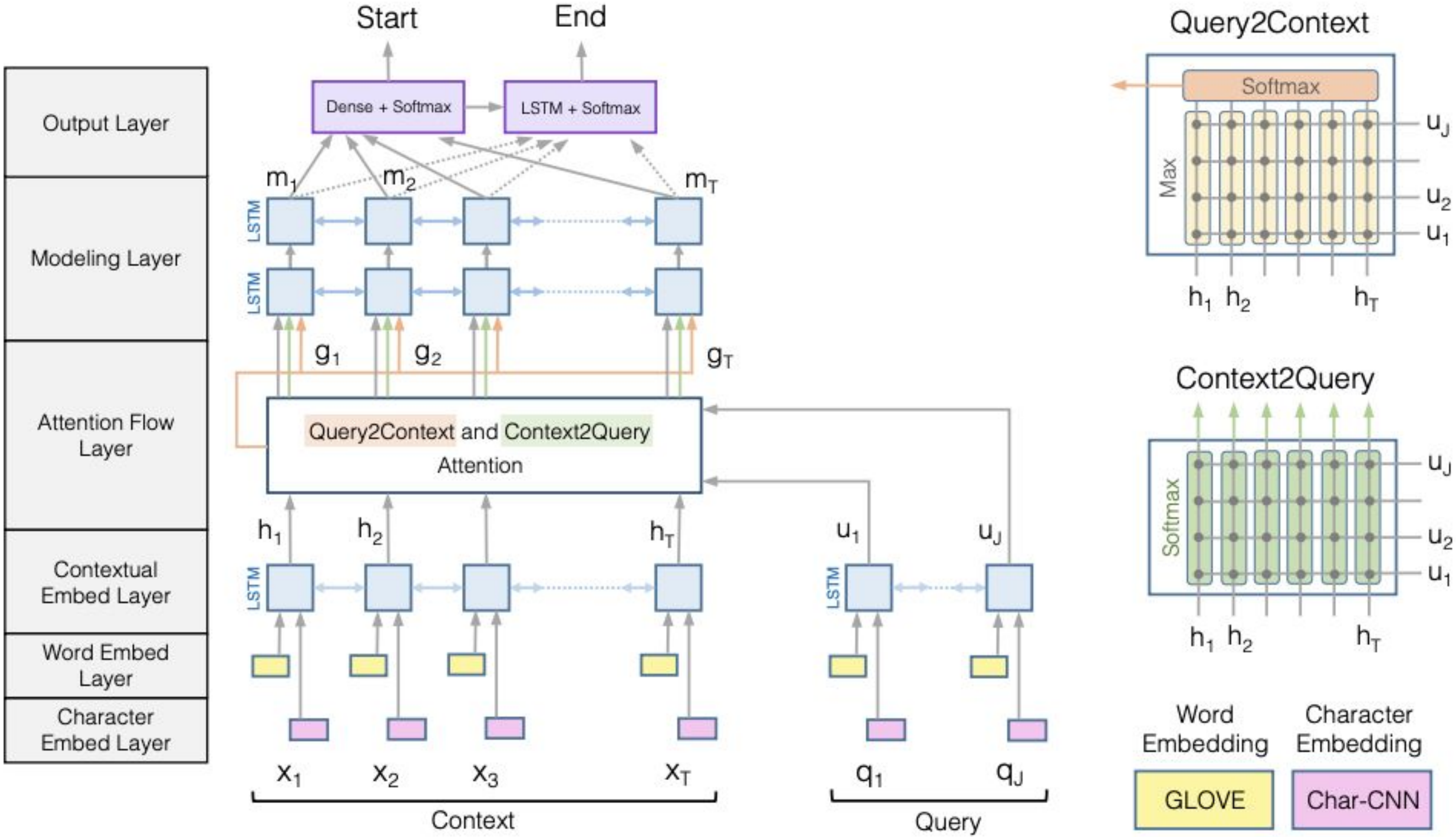
$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^2$$

- For each passage position, output of BiDAF layer is:

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h} \quad \forall i \in \{1, \ldots, N\}$$

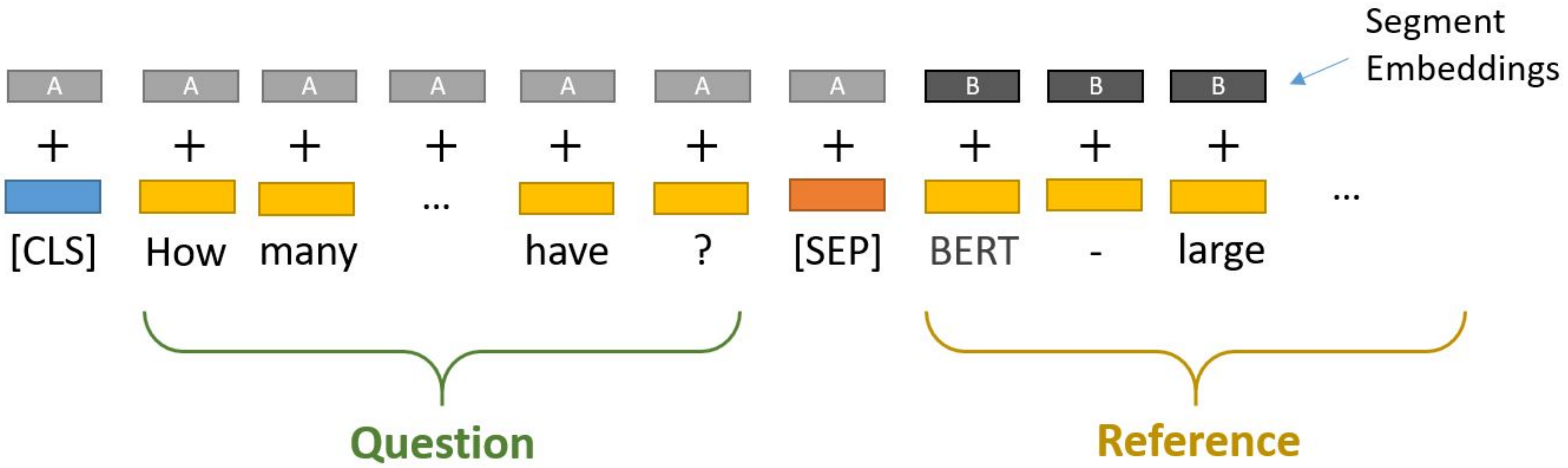
source: [CS224n Lecture 10]

BiDAF



source: [Bidirectional Attention Flow for Machine Comprehension]

BERT for Question Answering

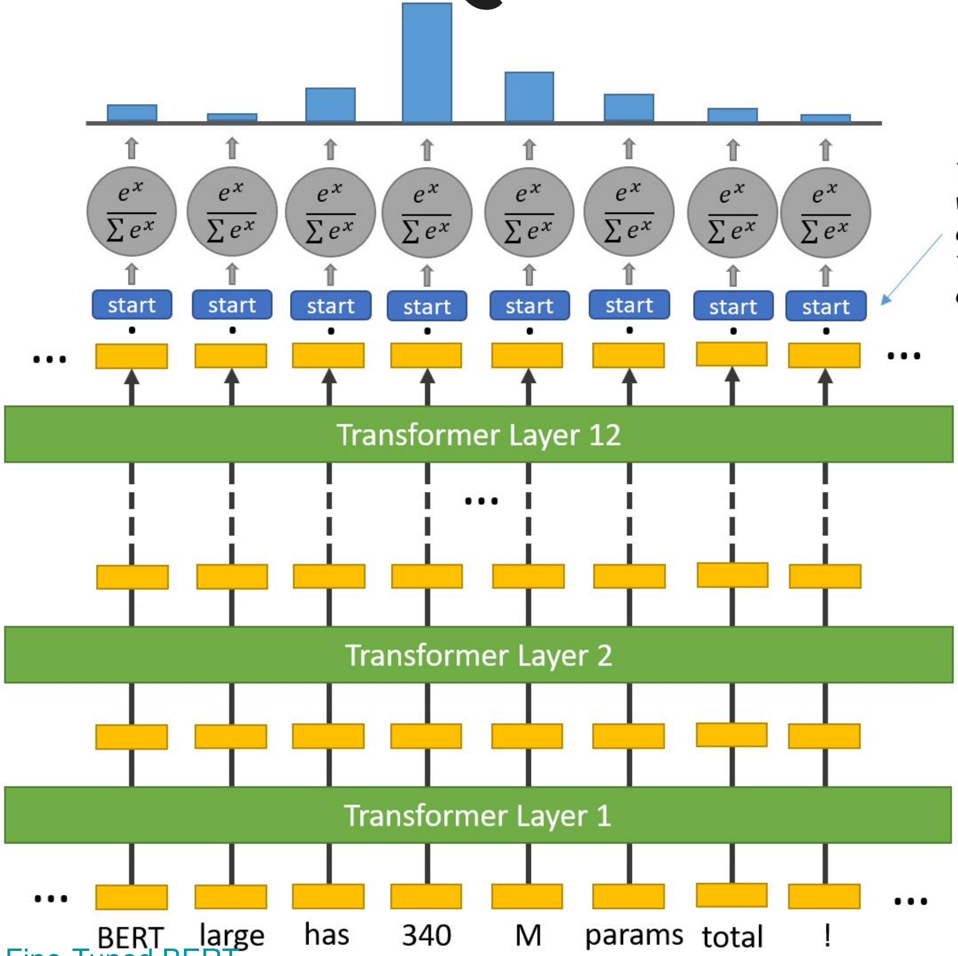


Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

source: [Question Answering with a Fine-Tuned BERT]

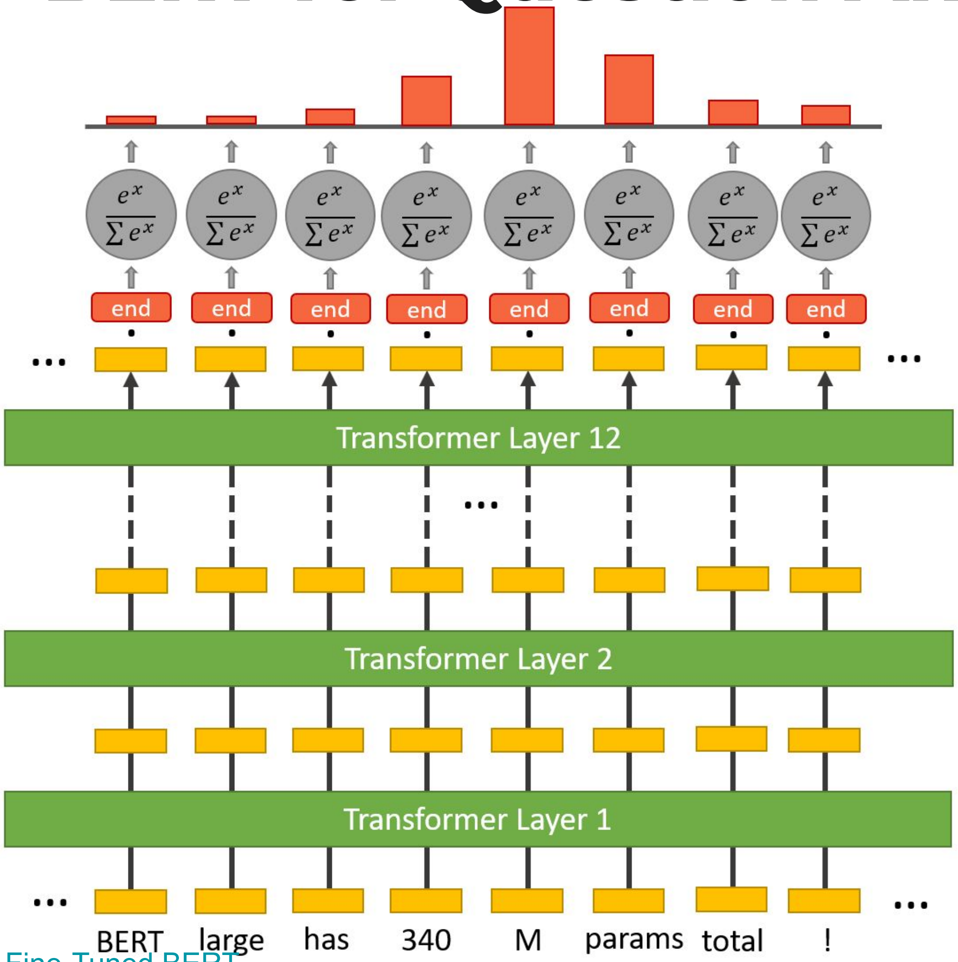
BERT for Question Answering



This length 768 vector is the ***weights*** *for the start token classifier.*
The ***same weights*** *are applied to* ***every position****.*

source: [Question Answering with a Fine-Tuned BERT]

BERT for Question Answering



source: Question Answering with a Fine-Tuned BERT

Open-Domain Question Answering

Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

WIKIPEDIA
The Free Encyclopedia

Document Retriever

Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

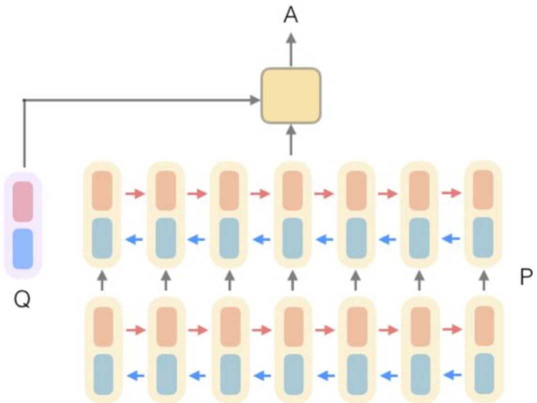
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F 1934, see Adamowicz brothers.

Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

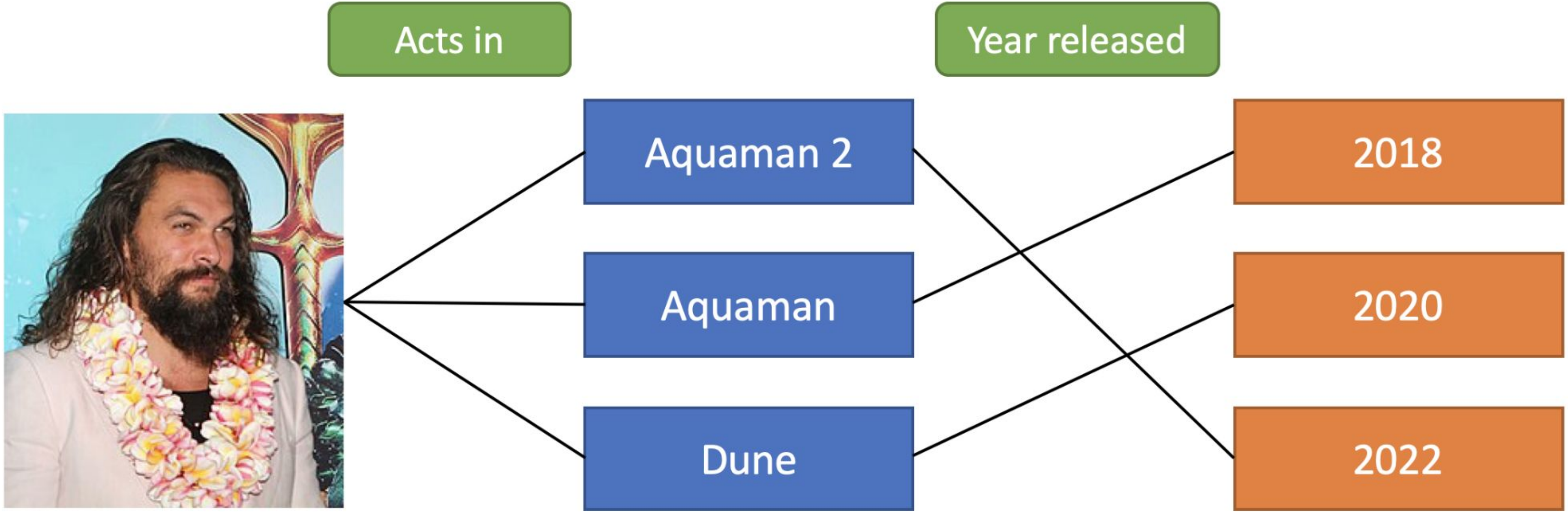
833,500



source: https://github.com/facebookresearch/DrQA

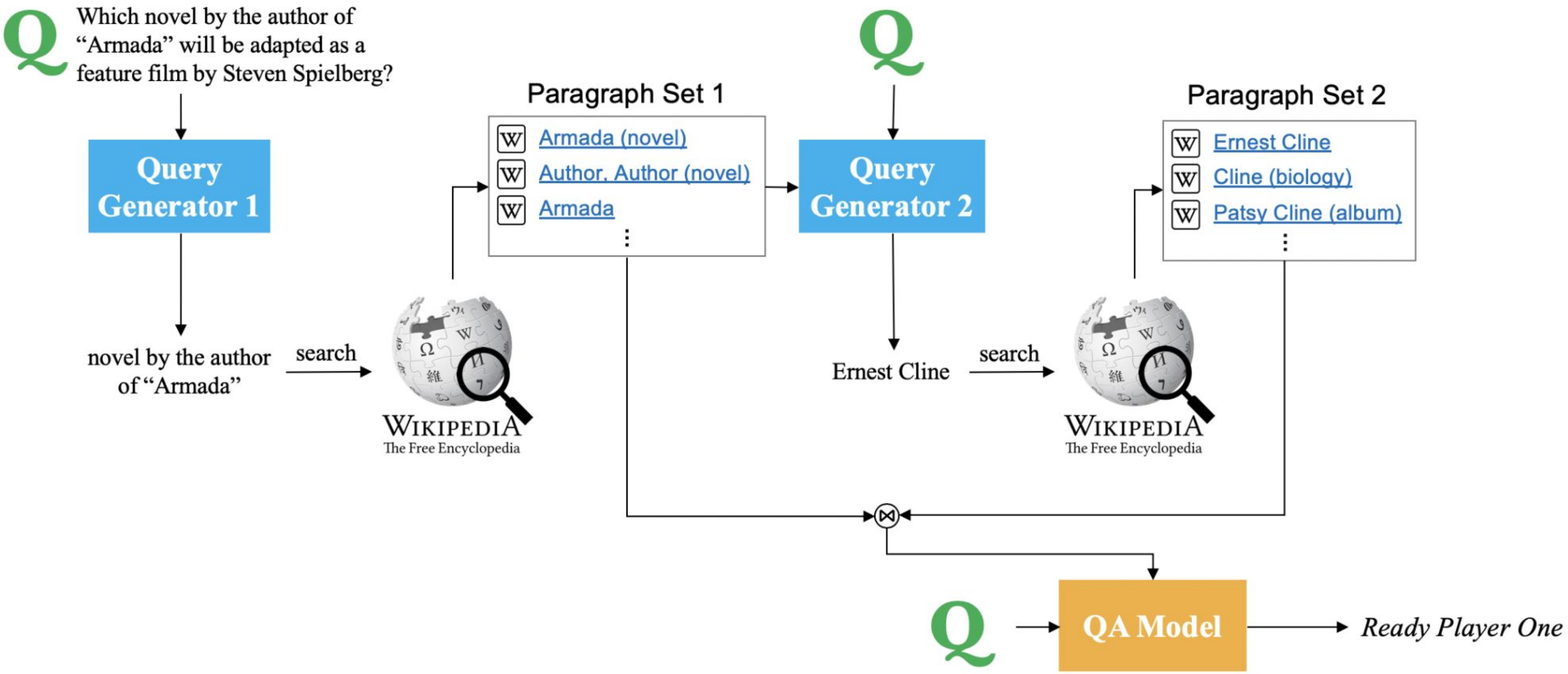
Possible problems

Example question: "What is the Aquaman actor's next movie?"



source: http://ai.stanford.edu/blog/answering-complex-questions/

Potential solutions



source: http://ai.stanford.edu/blog/answering-complex-questions/

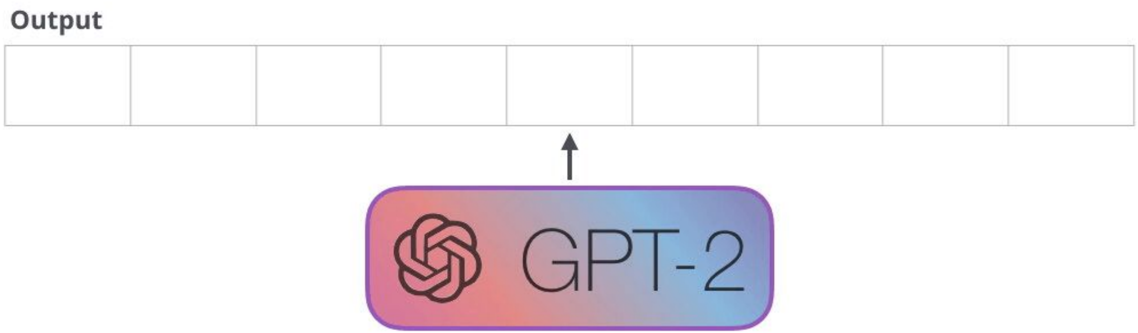
Outro

- Question answering systems bring us one step closer to human-like NLP systems
- Refer to the original papers on Transformer-XL, BiDAF and SQuAD, there are a lot of interesting ideas in there
- For Russian language:
  - SberQuAD paper provides a great aggregation of available materials, libraries and pretrained models
  - deeppavlov.ai and Natasha project provide many useful materials and pretrained models

GPT-2 & GPT-3

GPT-2

- Transformer-based architecture
- trained to predict the **next** word
- 1.5 billion parameters
- Trained on 8 million web-pages



On language tasks (question answering, reading comprehension, summarization, translation) works well **WITHOUT** fine-tuning

Image source: https://jalammar.github.io/illustrated-gpt2

GPT-2: question answering

EXAMPLES

Who wrote the book the origin of species?

Correct answer: *Charles Darwin*

Model answer: `Charles Darwin`

What is the largest state in the U.S. by land mass?

Correct answer: *Alaska*

Model answer: `California`

GPT-2: language modeling

EXAMPLE

Both its sun-speckled shade and the cool grass beneath were a welcome respite after the stifling kitchen, and I was glad to relax against the tree's rough, brittle bark and begin my breakfast of buttery, toasted bread and fresh fruit. Even the water was tasty, it was so clean and cold. It almost made up for the lack of...

Correct answer: *coffee*

Model answer: `food`

GPT-2: machine translation

EXAMPLE

French sentence:
Un homme a expliqué que l'opération gratuite qu'il avait subie pour soigner une hernie lui permettrait de travailler à nouveau.

Reference translation:
One man explained that the free hernia surgery he'd received will allow him to work again.

Model translation:
`A man told me that the operation gratuity he had been promised would not allow him to travel.`

GPT-2: fake news and hype

New AI fake text generator may be too dangerous to ... - The Guardian
https://www.theguardian.com/.../elon-musk-backed-ai-writes-convincing-news-fiction
4 days ago - The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse. The creators of a revolutionary AI system that can write news stories and works of fiction – dubbed "deepfakes for text" – have taken the unusual step of not releasing ...

OpenAI built a text generator so good, it's considered too dangerous to ...
https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/
12 hours ago - A storm is brewing over a new language model, built by non-profit artificial intelligence research company OpenAI, which it says is so good at ...

The AI Text Generator That's Too Dangerous to Make Public | WIRED
https://www.wired.com/story/ai-text-generator-too-dangerous-to-make-public/
4 days ago - In 2015, car-and-rocket man Elon Musk joined with influential startup backer Sam Altman to put artificial intelligence on a new, more open ...

Elon Musk-backed AI Company Claims It Made a Text Generator ...
https://gizmodo.com/elon-musk-backed-ai-company-claims-it-made-a-text-gener-183...
Elon Musk-backed AI Company Claims It Made a Text Generator That's **Too Dangerous to** Release · Rhett Jones · Friday 12:15pm · Filed to: OpenAI Filed to: ...

Scientists have made an AI that they think is too dangerous to ...
https://www.weforum.org/.../amazing-new-ai-churns-out-coherent-paragraphs-of-text/
3 days ago - Sample outputs suggest that the AI system is an extraordinary step forward, producing text rich with context, nuance and even something ...

New AI Fake Text Generator May Be Too Dangerous To ... - Slashdot
https://news.slashdot.org/.../new-ai-fake-text-generator-may-be-too-dangerous-to-rele...
3 days ago - An anonymous reader shares a report: The creators of a revolutionary AI system that can write news stories and works of fiction -- dubbed ...

Top stories

OpenAI built a text generator so good, it's considered too dangerous to release
TechCrunch
11 hours ago

Elon Musk's AI company created a fake news generator it's too scared to make public
BGR.com
9 hours ago

The AI That Can Write A Fake News Story From A Handful Of Words
NDTV.com
2 hours ago

When Is Technology Too Dangerous to Release to the Public?
Slate · 2 days ago

Scientists Developed an AI So Advanced They Say It's Too Dangerous to Release
ScienceAlert · 6 days ago

Different models by the end of 2019

	Nb of parameters (millions)	Inference Time (s)
GLUE BASELINE (ELMo + BiLSTMs)	180	895
BERT base	110	668
DistilBERT	66	410

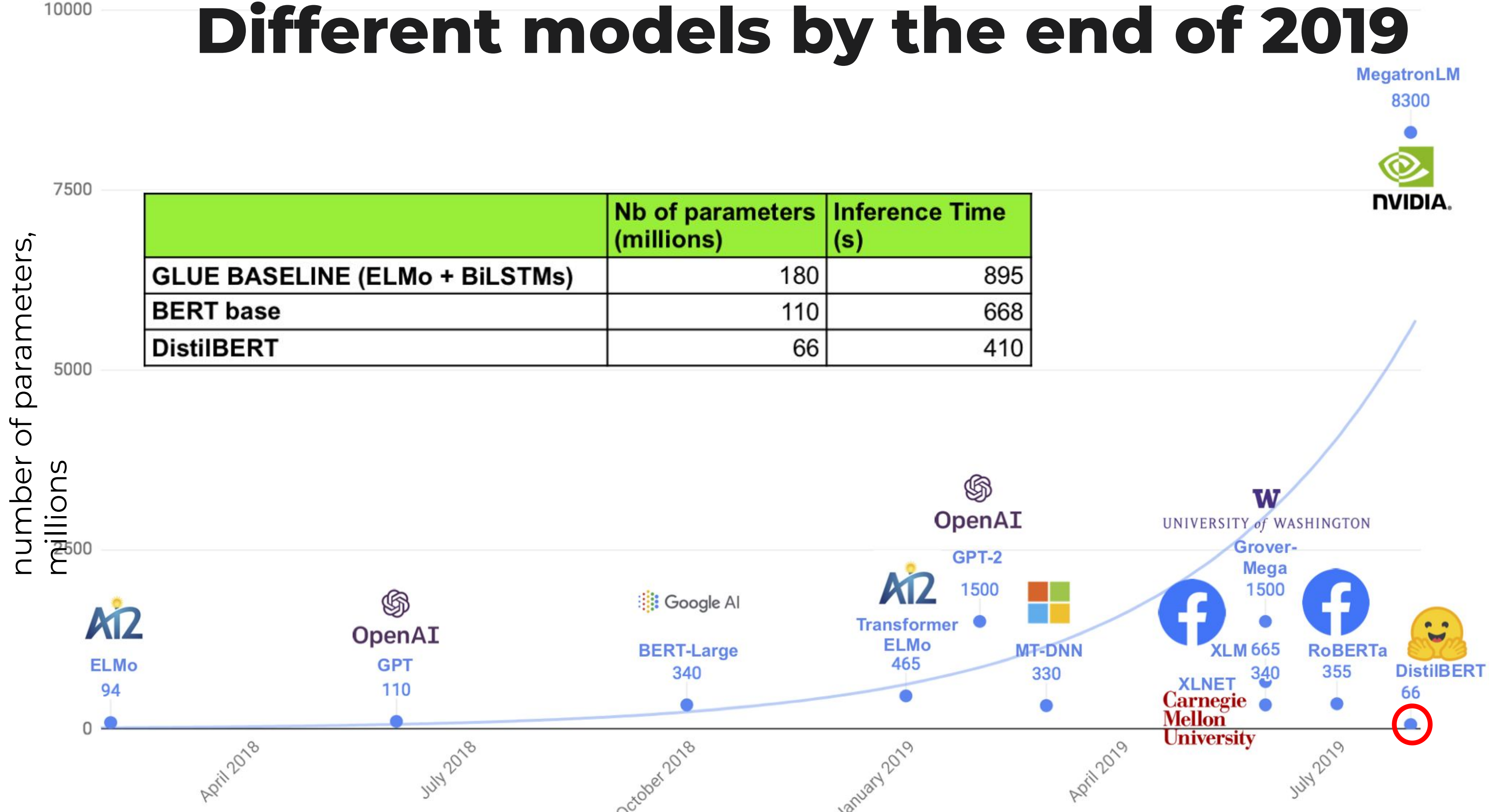
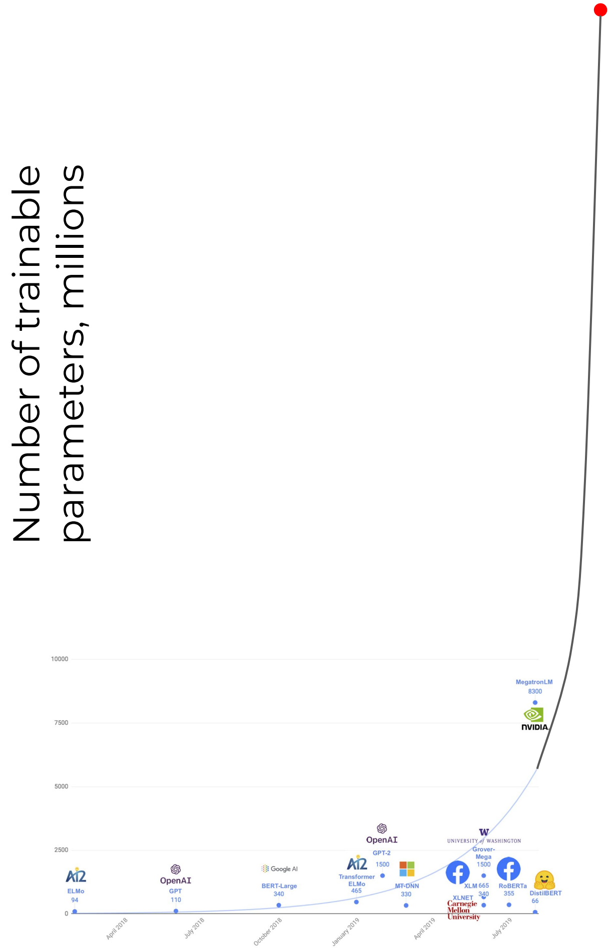


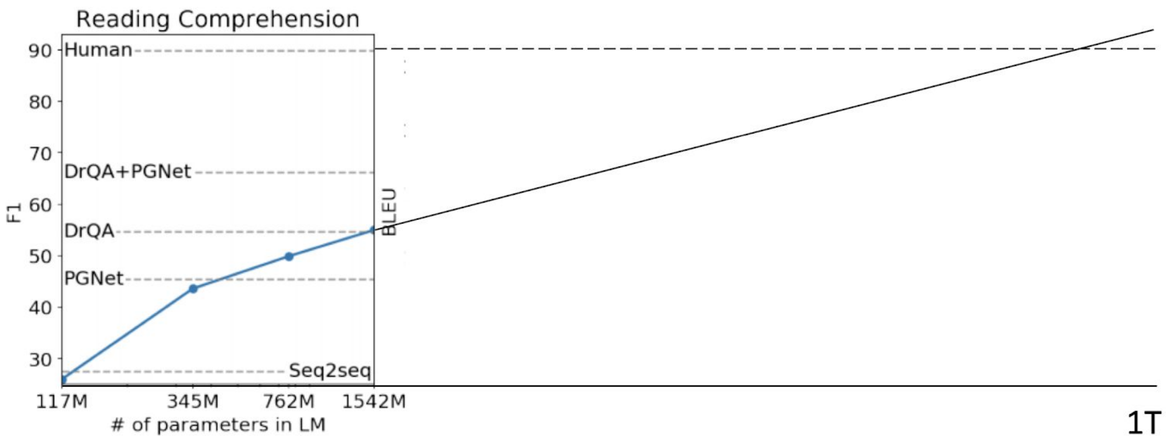
Image source: [Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT]

GPT-3



GPT-3, May 2020

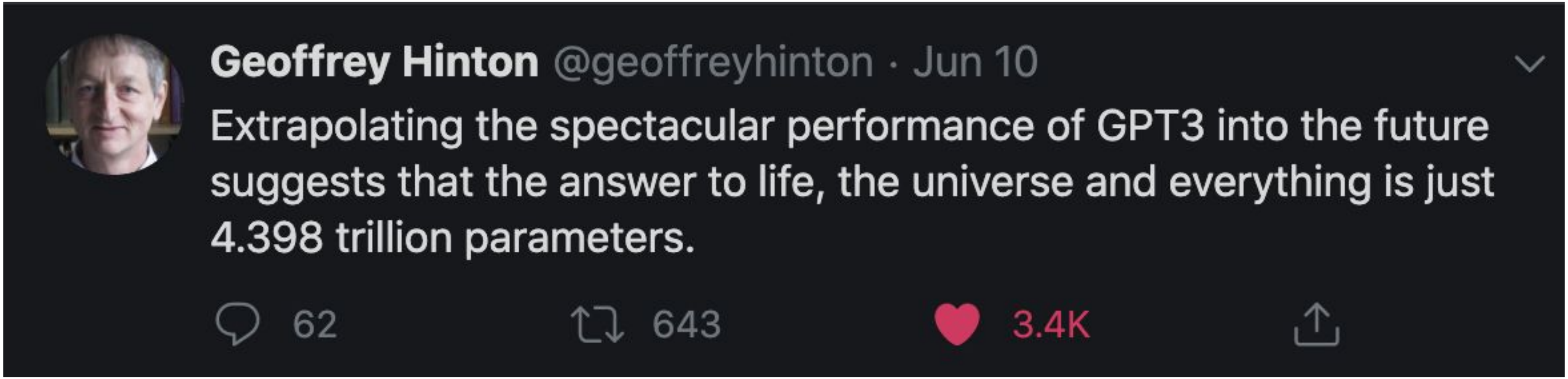
Proportions are not preserved for visual sake



Hypothesis from Stanford CS224N Lecture 20 (2019)

May 2020: GPT-3

- GPT-2: 1.5 billion parameters
- GPT-3: **175 billion** parameters



- [Transformer]
- [OpenAI Transformer]
- [ELMO]
- [BERT]
- [BERTology]
- [GPT]
- [GPT-2]
- [GPT-3]

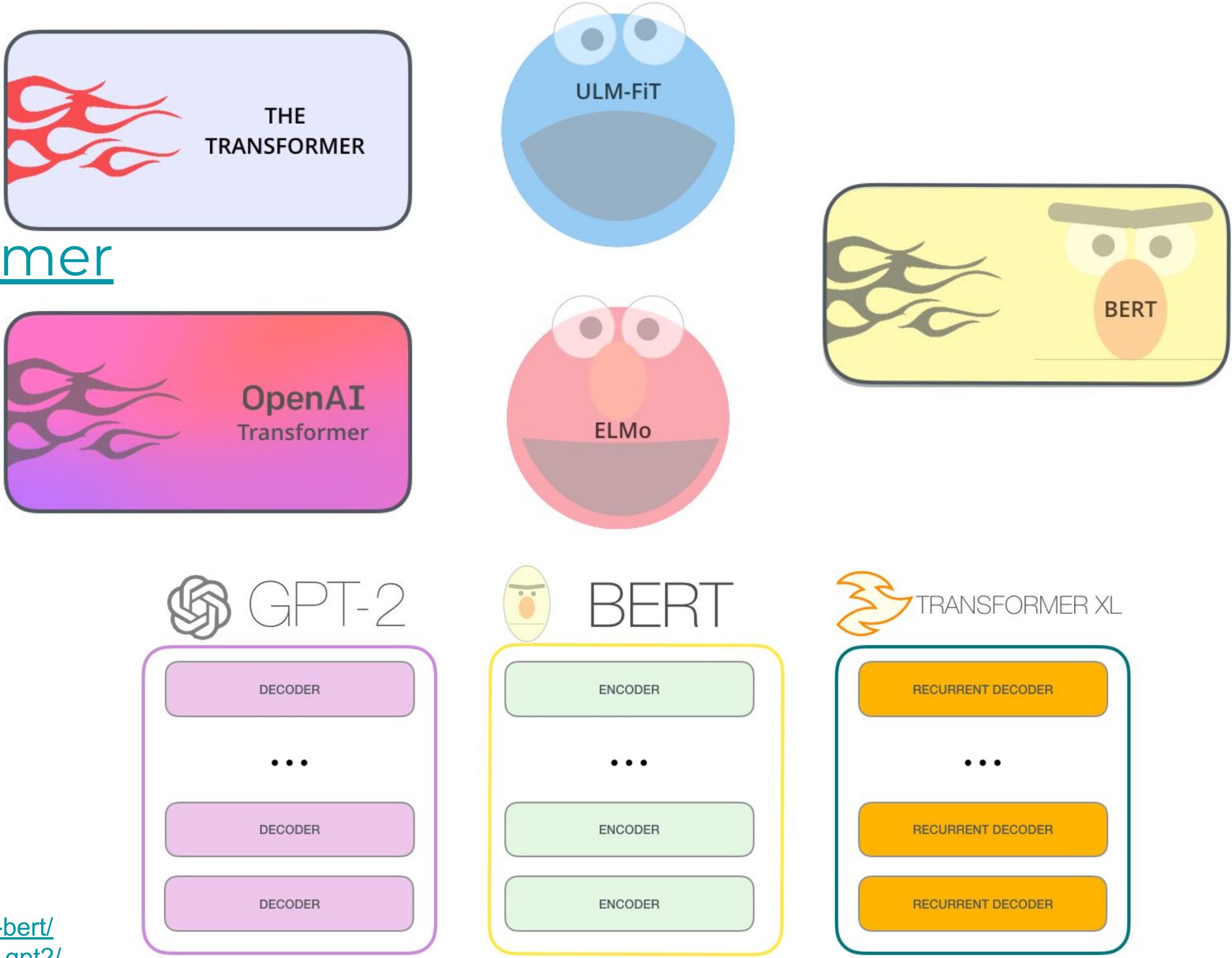
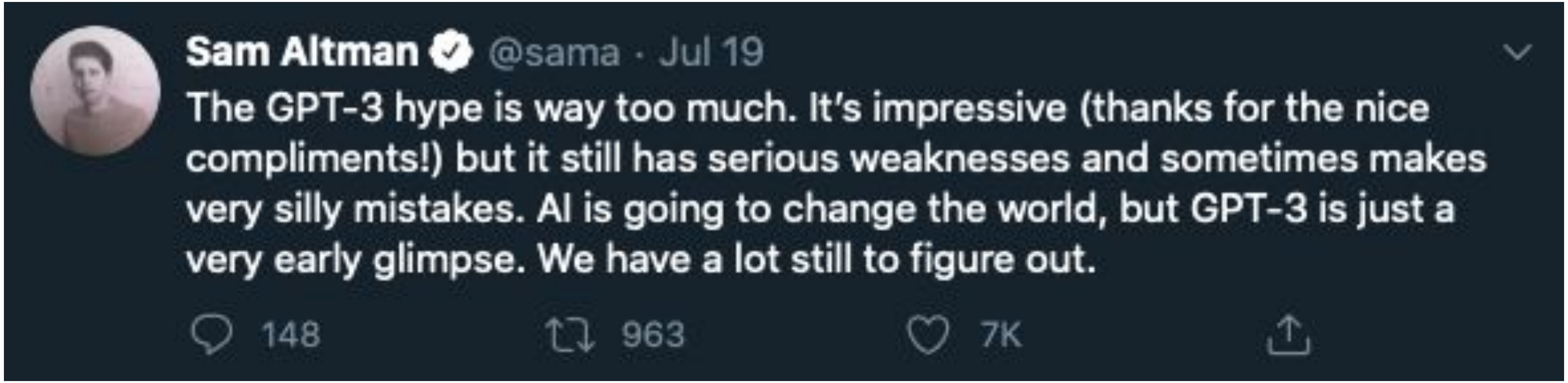


Image sources: http://jalammar.github.io/illustrated-bert/
http://jalammar.github.io/illustrated-gpt2/

More on GPT

Reaction: GPT-3


Sam Altman @sama · Jul 19
The GPT-3 hype is way too much. It's impressive (thanks for the nice compliments!) but it still has serious weaknesses and sometimes makes very silly mistakes. AI is going to change the world, but GPT-3 is just a very early glimpse. We have a lot still to figure out.
148 963 7K


Paul Graham @paulg · Jul 19
Hackers are fascinated by GPT-3. To everyone else it seems a toy.

Pattern seem familiar to anyone?
153 604 5K

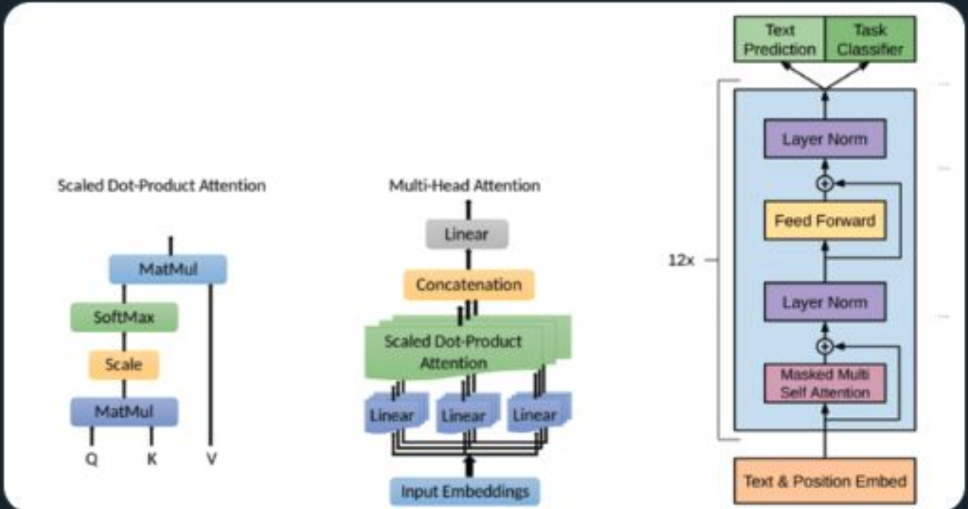
Source: Twitter, more links available at https://lambdalabs.com/blog/gpt-3/

Reaction: GPT-3

Andriy Burkov • Following
ML at Gartner, author of The Hundred-Page Machine Learning Book
2d •

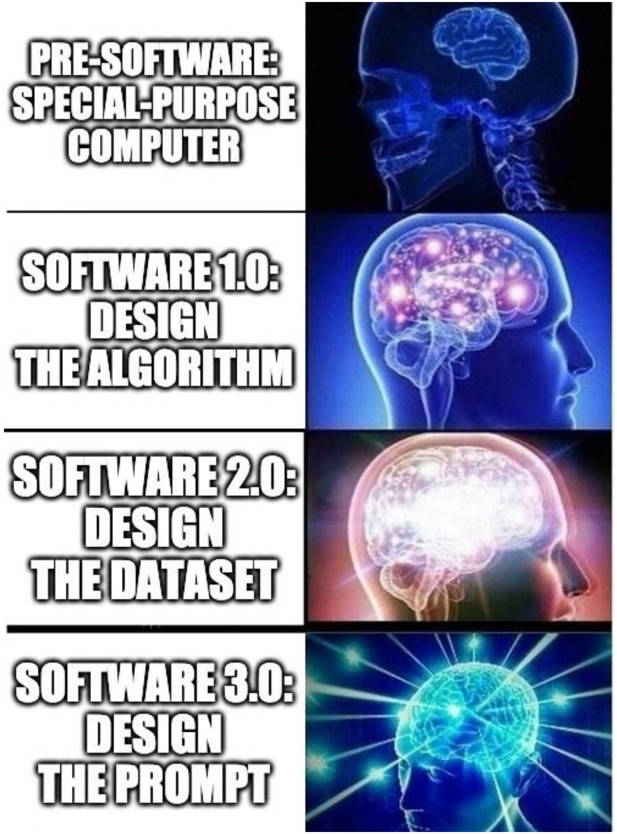
GPT-3 is the closest thing to artificial general intelligence (AGI) that I ever saw.
It's so strong that it makes me nervous.

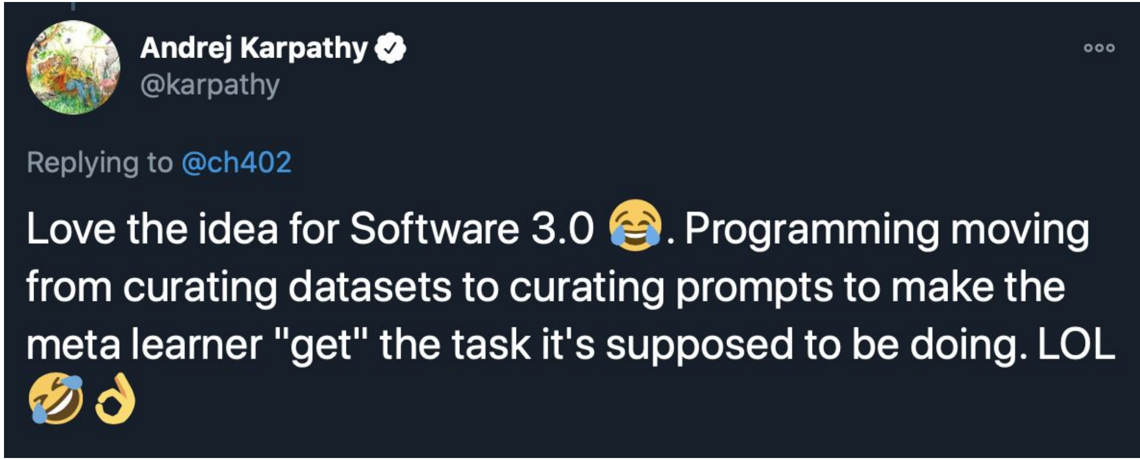
Andrej Karpathy @karpathy · Jul 19
The transformer architecture of GPT upper bounds its ability at memorization. It cannot learn many algorithms due to the functional form of its forward pass, and spends a fixed compute per token - i.e. it can't "think for a while". Progress here critical, likely but non-trivial.



Source: LinkedIn, Twitter

Reaction: GPT-3





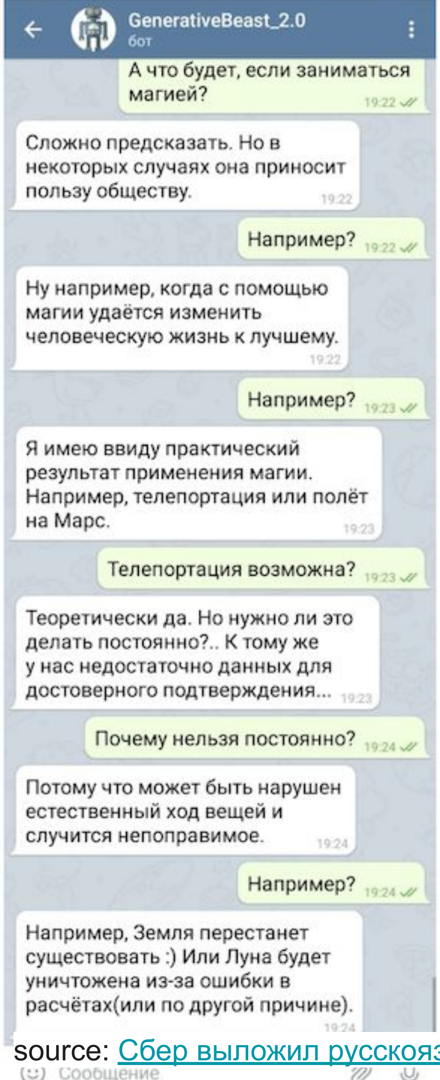
source: Andrej Karpathy twitter

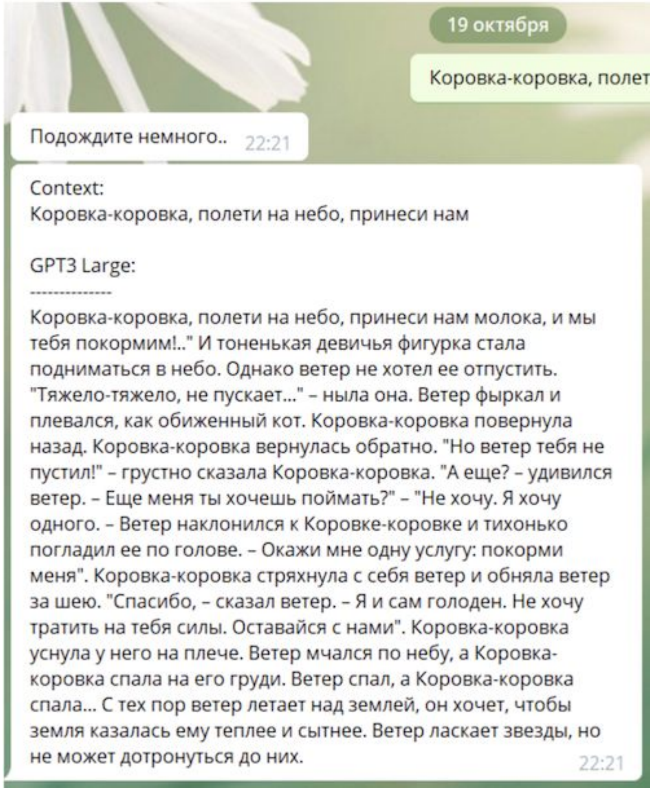
miniGPT

- Great way to understand GPT
  - ~300 lines of code and examples

source: [miniGPT by Andrej Karpathy]

GPT-3 Large for Russian Language





source: [Сбер выложил русскоязычную модель GPT-3 Large с 760 миллионами параметров в открытый доступ]

Outro

- Question answering systems bring us one step closer to human-like NLP systems
- Refer to the original papers on Transformer-XL, BiDAF and SQuAD, there are a lot of interesting ideas in there
- For Russian language:
  - SberQuAD paper provides a great aggregation of available materials, libraries and pretrained models
  - GPT-3 trained on Russian texts
  - deeppavlov.ai and Natasha project provide many useful materials and pretrained models

Outro and farewell

- Now you know a bit more about this inspiring and promising area of Machine Learning and Deep Learning
- Many more challenges are awaiting us ahead
- Stay focused
- Thank you for you attention!
- And good luck!