HOW TO TRAIN YOUR CONVERSATIONAL AI TO PAY *ATTENTION*

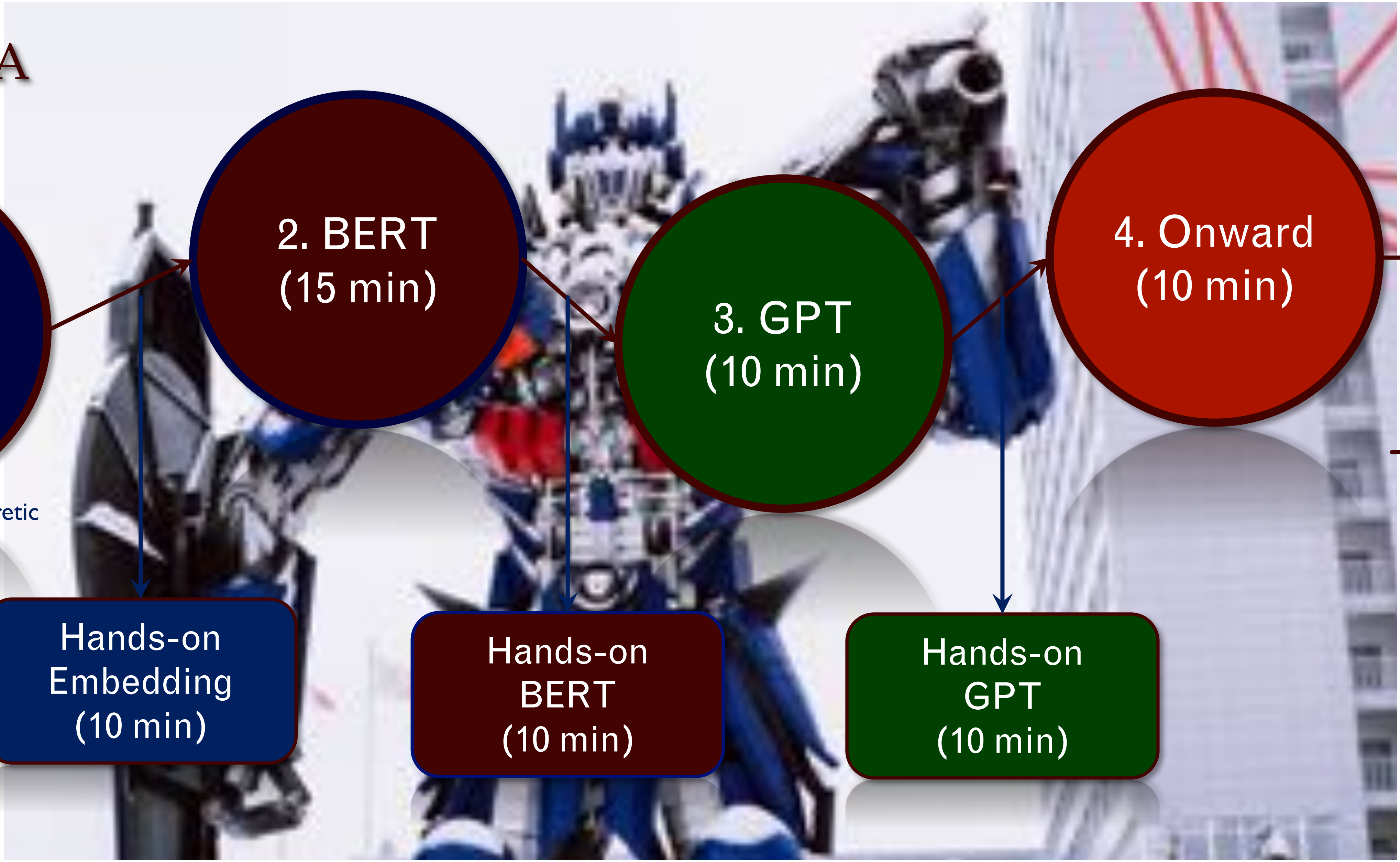
A HANDS-ON APPROACH TO COMPUTATIONAL LINGUISTICS AND TRANSFORMERS

- Computational Linguistics
- Computer Theoretic & a Linguist's View
- At this level,
  - It is a Linguistics problem
  - With a Computer Theoretic solution !

April 10, 2020

Krishna Sankar
@ksankar

AGENDA



1. Transformers (40 min)
   - Computer Theoretic View
   - Linguistic View
   - Transformers

Hands-on Embedding (10 min)

2. BERT (15 min)

Hands-on BERT (10 min)

3. GPT (10 min)

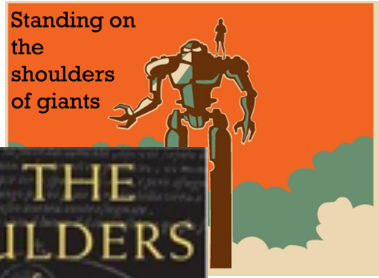
Hands-on GPT (10 min)

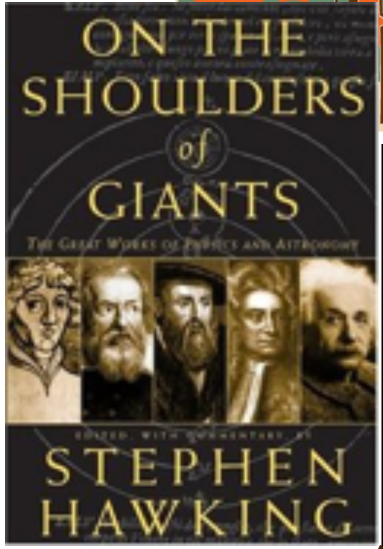
4. Onward (10 min)

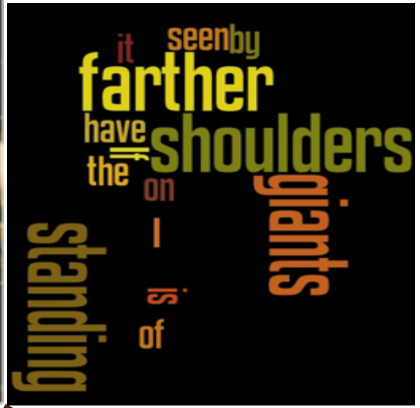
PRAGMAS

- Will remain at the system level – we don't have the time to go deeper
- Have 3 hands-on
  - Try them here, download and try elsewhere
  - http://bit.ly/gtc-transformers
- Progressively build concepts – focus on "Why" and some "How"
- *Giant gulp from this very rich fountain of ideas !*
- Will cover Computer Theoretic and the Linguist Views – Computational Linguistics
- I had a good time preparing for the tutorial !!
- Trimming the slides, acted like an autoencoder – efficient data encoding and dimensionality reduction !!
  - Many a slide fell on the cutting room floor – *may be I should add a deleted scenes at the end or a director's cut !*

THANKS TO THE GIANTS, WHOSE WORK HELPED ME TO PREPARE FOR THIS LAB !!

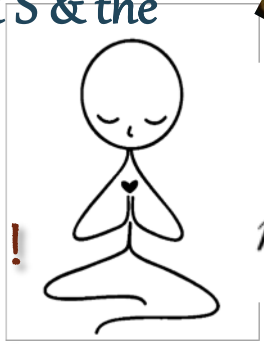






- References and papers mentioned in the slides
- Will update reference list in the github
- bit.ly/gtc-transformers

And our own nvidians : Charles K, Dana S, Devesh K, Josh W, Sara W, Pete A, David W, Mohammad S & the gang @DLI .... And, the TAs



A BIG THANKS TO YOU ALL FOR ATTENTION !

NVIDIA.

1. INTRODUCTION & MOTIVATION


THE WALL STREET JOURNAL.

SIGN IN

WSJ INVESTIGATION

How Google ~~Interferes With~~ enhances Its Search Algorithms and ~~Changes~~ augments Your Results

The internet giant uses blacklists, algorithm tweaks and an army of contractors to shape what you see

By Kirsten Grind, Sam Schechner, Robert McMillan and John West

Nov. 15, 2019 8:15 am ET

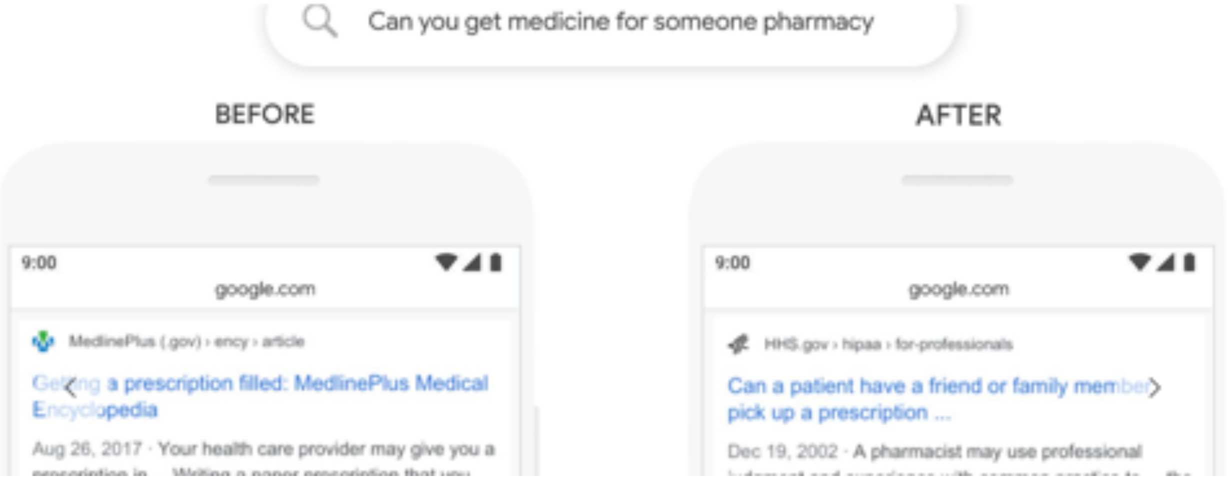
Computational linguistics

Google brings in BERT to improve its search results

Frederic Lardinois @fredericl / 10:07 am PDT • October 25, 2019 Comment

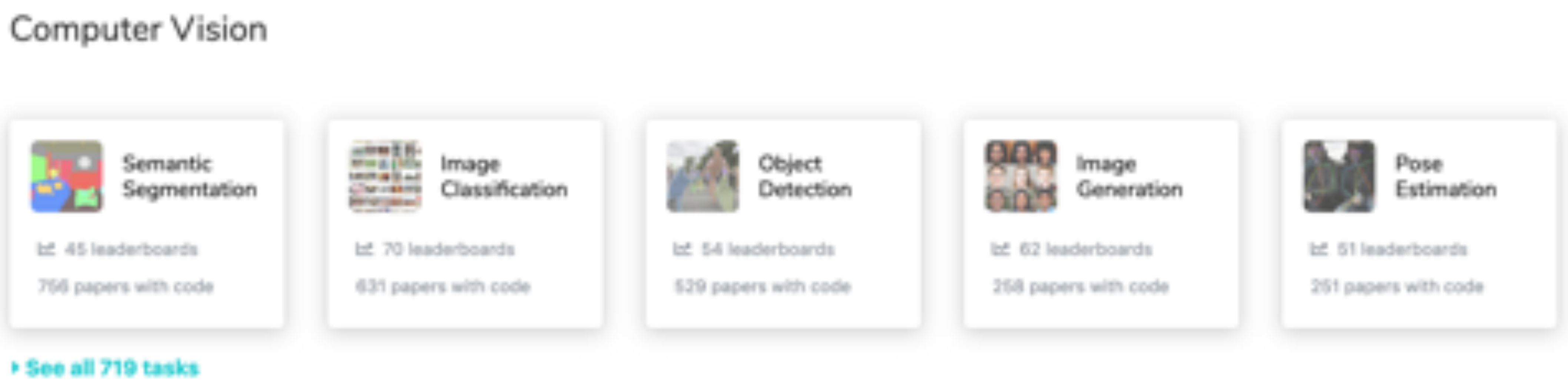
BEFORE, GOOGLE TREATED QUERIES LIKE "A BAG OF WORDS" 1

BERT COULD AFFECT AS MANY AS 10 PERCENT OF ALL GOOGLE SEARCHES 2



- The technology behind this new neural network is called "Bidirectional Encoder Representations from Transformers," or BERT
- *"Every word of that acronym is a term of art in NLP"* 3
- *"At its core, Search is about understanding language … and Search is not a solved problem"*
- *BERT enables a better "search query understanding" of your trickiest searches* 4
  - *… more relevant search results and featured snippets.*
- *".. this is the single biggest ... most positive change we've had in the last five years and perhaps one of the biggest since the beginning."* 5
- 6 This BERT update also marks the first time Google is using its latest Tensor Processing Unit (TPU) chips to serve search results.

https://techcrunch.com/2019/10/25/google-brings-in-bert-to-improve-its-search-results/
https://www.theverge.com/2019/10/25/20931657/google-bert-search-context-algorithm-change-10-percent-langauge



NMT, Language Modelling, Q&A, Sentiment, Sentence/Document Classification (Spam,Not Spam), Reading Comprehension, Text Summarization, NLI, Dialogue Generation, Fact Verification (Claim, not claim)

Natural Language Processing

https://paperswithcode.com/area/natural-language-processing

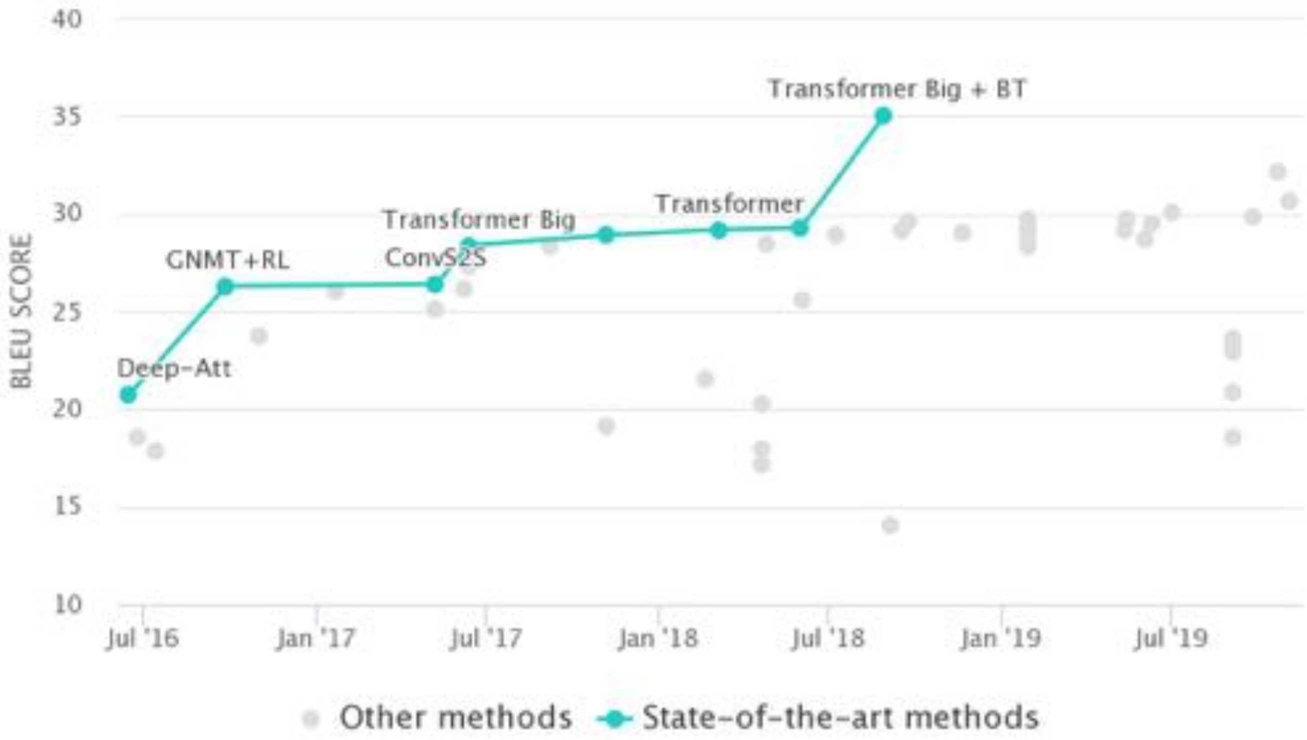
Browse > Natural Language Processing > Machine Translation

Machine Translation

542 papers with code · Natural Language Processing

Machine translation is the task of translating a sentence in a source languag

(Image credit: Google seq2seq)



Leaderboards

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	WMT2014 English-German	🏆 Transformer Big + BT	Understanding Back-Translation at Scale			See all
	WMT2014 English-French	🏆 Transformer Big + BT	Understanding Back-Translation at Scale			See all
	IWSLT2015 German-English	🏆 Transformer Base + adversarial MLE	Improving Neural Language Modeling via Adversarial Training			See all

Browse > Natural Language Processing > Language Modelling

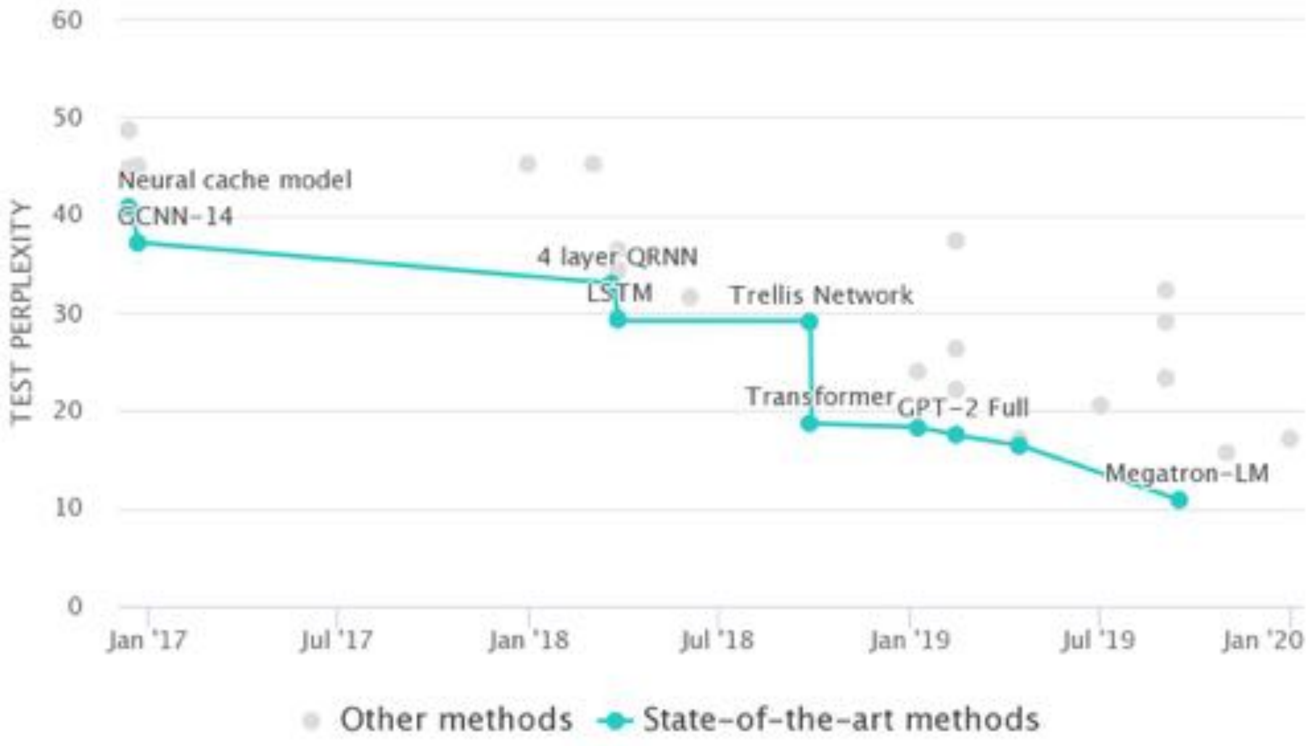
Language Modelling

473 papers with code · Natural Language Processing

Language modeling is the task of predicting the next word or character in a document.

* indicates models using dynamic evaluation; where, at test time, models may adapt to seen tokens. (Mikolov et al., (2010), Kraus et al., (2017))

(Image credit: Exploring the Limits of Language Modeling)



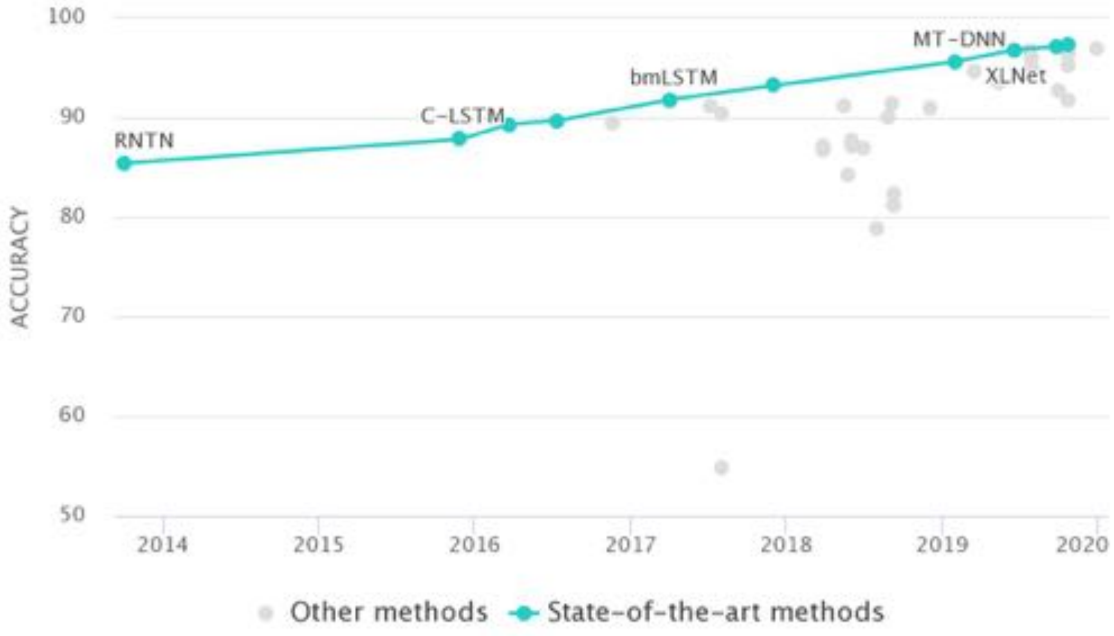
Leaderboards

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	WikiText-103	Megatron-LM	Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism			See all
	Penn Treebank (Word Level)	GPT-2	Language Models are Unsupervised Multitask Learners			See all
	enwiki8	GPT-2	Language Models are Unsupervised Multitask Learners			See all

Browse > Natural Language Processing > Sentiment Analysis

Sentiment Analysis

338 papers with code · Natural Language Processing



Sentiment analysis is the task of classifying the polarity of a given text.

Leaderboards

TREND	DATASET	BEST METHOD	PAPER TITLE			
	SST-2 Binary classification	T5-3B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			See all
	IMDb	NB-weighted-BON + dv-cosine	Sentiment Classification Using Document Embeddings Trained with Cosine Similarity			See all
	SST-5 Fine-grained classification	BERT large	Fine-grained Sentiment Classification using BERT			See all
	Yelp Binary classification	BERT large	Unsupervised Data Augmentation			See all
	Yelp Fine-grained classification	BERT large	Unsupervised Data Augmentation			See all

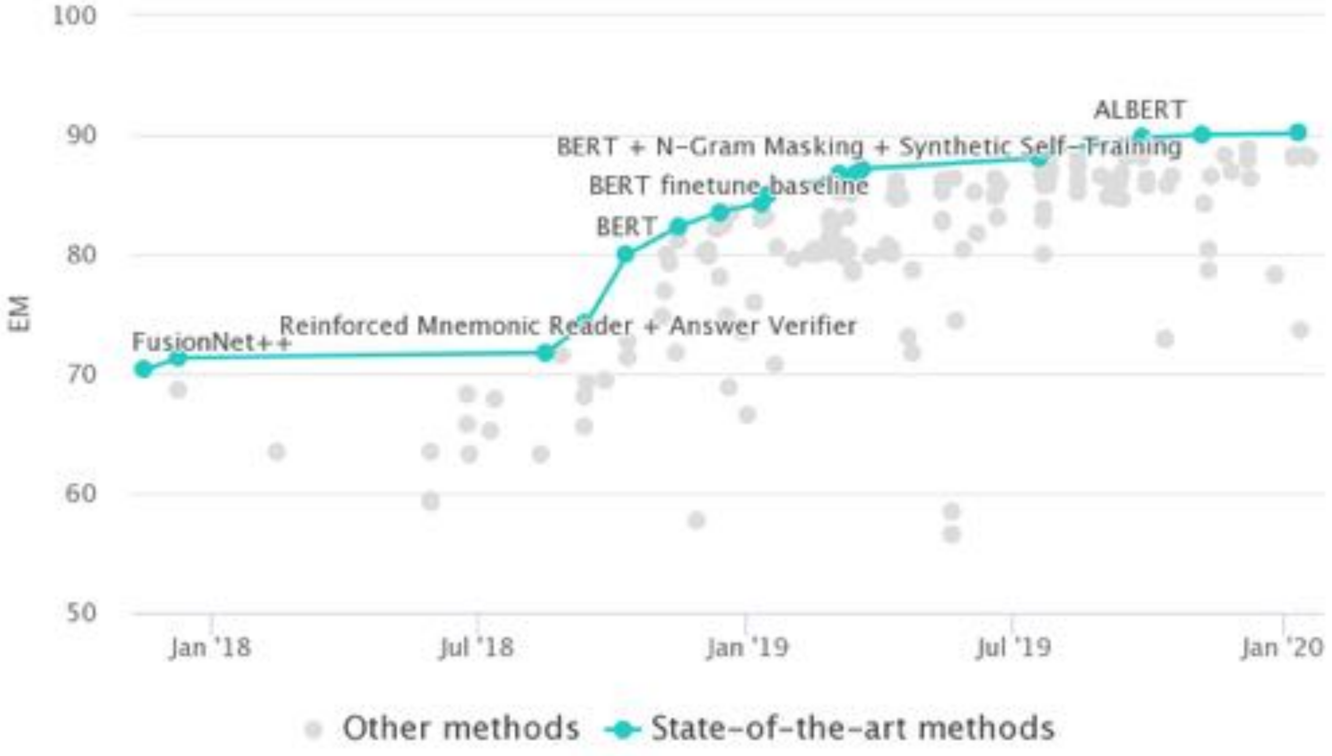
Browse > Natural Language Processing > Question Answering

Question Answering

452 papers with code · Natural Language Processing

(Image credit: SQuAD)

Given a question (Premise), identify a text that entails an answer (hypothesis)



Leaderboards

TREND	DATASET	BEST METHOD	PAPER TITLE			
	SQuAD1.1	XLNet (single model)	XLNet: Generalized Autoregressive Pretraining for Language Understanding			See all
	SQuAD2.0	ALBERT + DAAF + Verifier (ensemble)				See all
	SQuAD1.1 dev	T5-11B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			See all
	WikiQA	TANDA-RoBERTa (ASNQ, WikiQA)	TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection			See all

SQuAD — Home — Explore 2.0 — Explore 1.1

What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University*	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Dec 06, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745
5 Jan 10, 2020	Retro-Reader on ALBERT (single model) *Shanghai Jiao Tong University*	88.107	91.419
5 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
6	albert+verifier (single model)	88.355	91.019

Newsweek

U.S. | World | Business | Tech & Science | Culture | Newsgeek | Sports | Healt

TECH & SCIENCE

ROBOTS CAN NOW READ BETTER THAN HUMANS, PUTTING MILLIONS OF JOBS AT RISK

BY ANTHONY CUTHBERTSON ON 1/15/18 AT 8:00 AM EST

Are Robots Better Readers Than Humans?

THE VERGE

No, machines can't read better than humans

Headlines have claimed AIs outperform humans at 'reading comprehension,' but in reality they've got a long way to go

By James Vincent | Jan 17, 2018, 9:23am EST

SHARE

THIS WAS NEVER ABOUT READING COMPREHENSION AS WE KNOW IT

"IT'S LIKE A STUDENT WHO CAN DO WELL AT TESTS WITHOUT RECOGNIZING ANY OF THE SUBJECT MATERIAL."

Yoav Goldberg, a lecturer at Bar Ilan University who specializes in natural language processing, says the mistake is thinking of SQuAD as something akin to a school test, rather than a tool intended to help computer scientists. "SQuAD was not designed to be a realistic assessment of 'reading comprehension' in the popular sense," Goldberg tells *The Verge* over email. "It was designed as a benchmark for machine learning methods, and the human evaluation was performed to assess the quality of the dataset, not the humans' abilities." It's the fault of the media and PR for interpreting it as something more than this.

A BRIEF HISTORY OF NLU

- 1960s: Pattern-matching with small rule-sets
- 1970-80s: Linguistically rich, logic-driven, grounded systems; restricted applications
- Mid-1990s: Statistical revolution in NLP → decrease in NLU work
- Late 2000s: NLU returns to center stage, mixing techniques from previous decades
- Mid-2010s: Deep learning takes over NLU: LSTMs, seq2seq, ...

Linguistic Expertise

Pattern Matching + Rule-based Systems

1950s-1990s

Corpus-based Statistics

1990s-2000s

Machine Learning

2000s-2014

Deep Learning

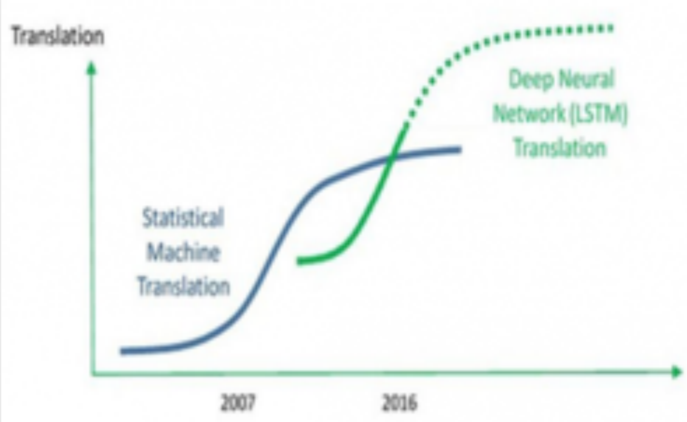
2014-20??

Transparent & Intelligible Models

Pre 2012-ish Traditional

End-To-End Everything
… Assuming you have the computing resources

To a first approximation, the de facto consensus in NLP in 2017 is that no matter what the task, you throw a BiLSTM at it, with attention if you need information flow



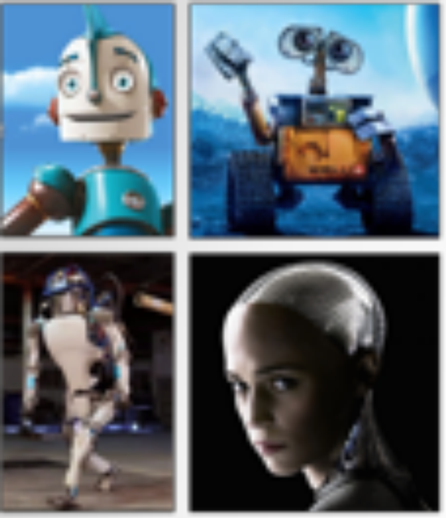
Chris Manning
April 2017

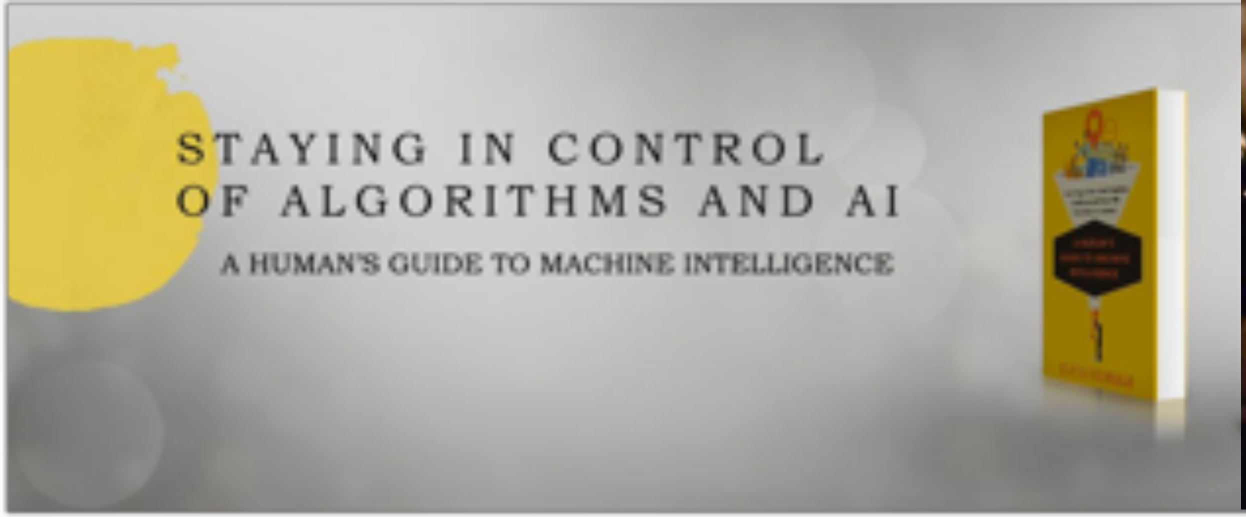
Yoav Goldberg: The missing elements in NLP (spaCy IRL 2019)

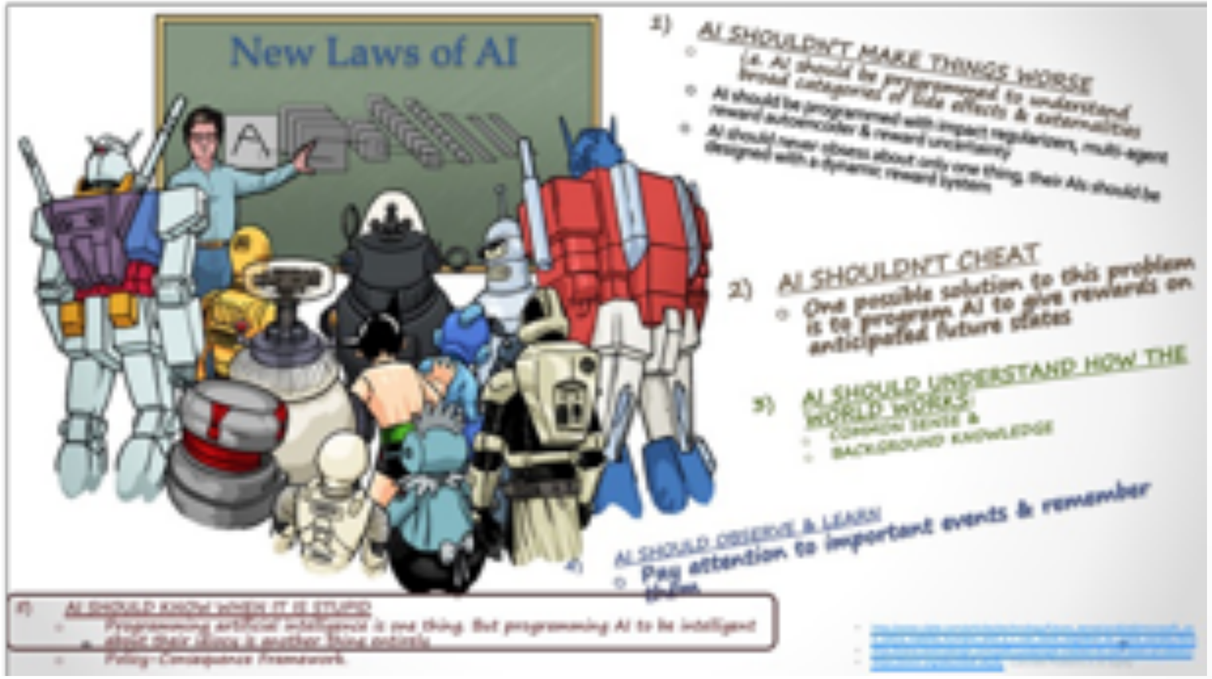
Machine Learning Expertise

Transparent ⟵ Explainability ⟶ *Black Box*

Little ⟵ **Labelled Data** ⟶ ***Lots***



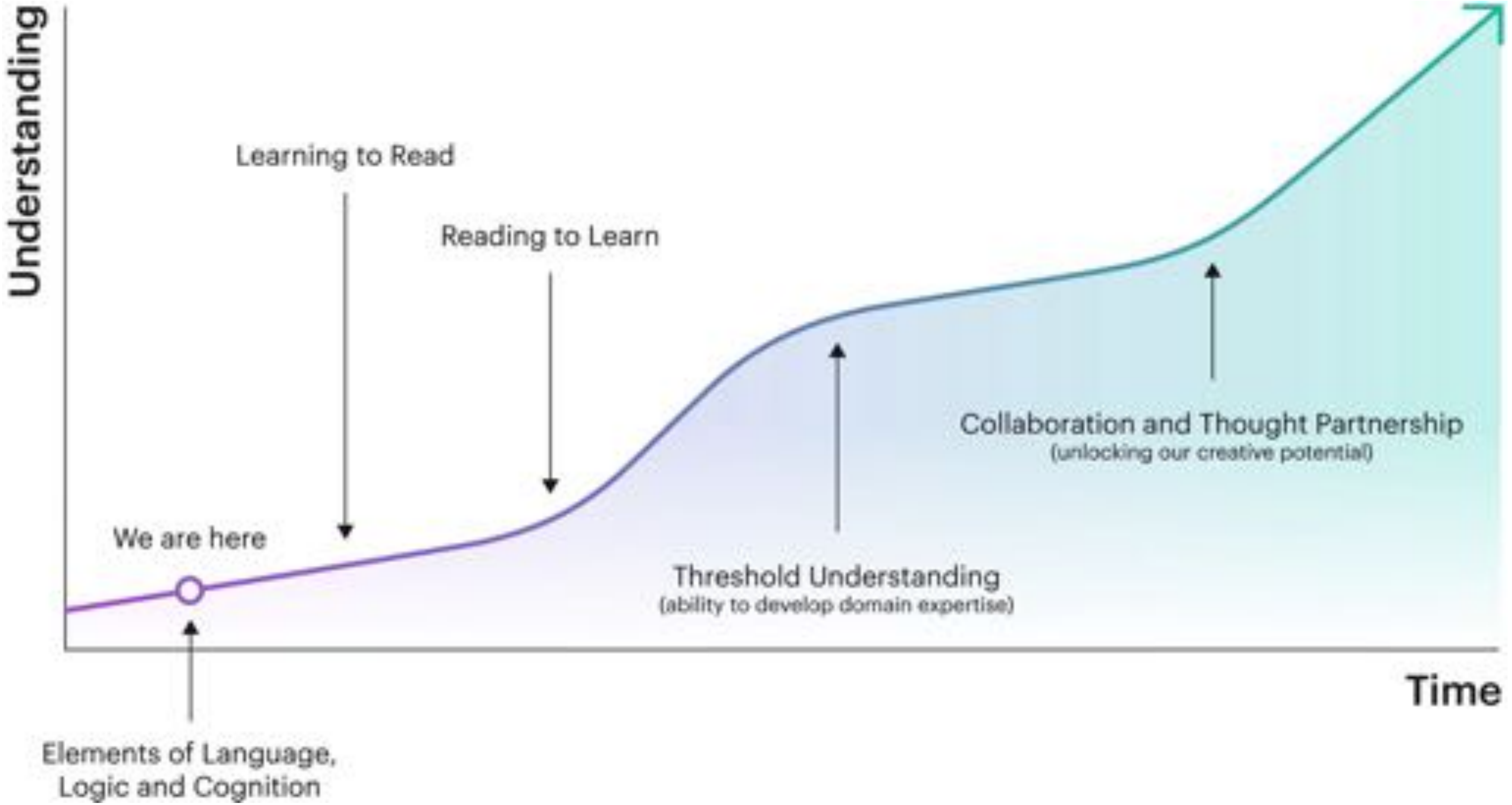




Voight-Kampff Tests !

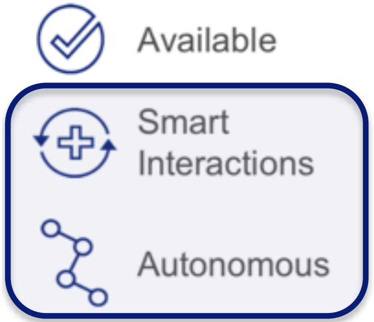
Model := Algorithm + Training Data
Decision := Model + Actual Data

https://medium.com/@ksankar/the-excessions-of-xai-algorithms-ethics-responsible-ai-and-the-explainability-part-1-of-7-d798354784d5



Ref: David Ferrucci @medium

Features	Themes	Levels
Limited Vocabulary, Limited actions e.g. lights on	Commands & Notifications	Notification (& IoT Command) Bots
Search & traverse through text, API based intent resolution	Q & A	FAQ Bots
Knowledge Graph (explicit or implicit intent based), some level of comprehension & capable of contextual conversation	Dialogues	Contextual Bots
Inferential, Predictive, Personalized and Branded for unique & differentiated digital experience	Interactions	Smart Personalized Bots
Human-level comprehension & common sense, Factoids, Knowledge Graph and capable of intelligent interactivity over multiple conversation chains	Conversations	Autonomous Bots



The Architecture from Cybertron

Cornell University

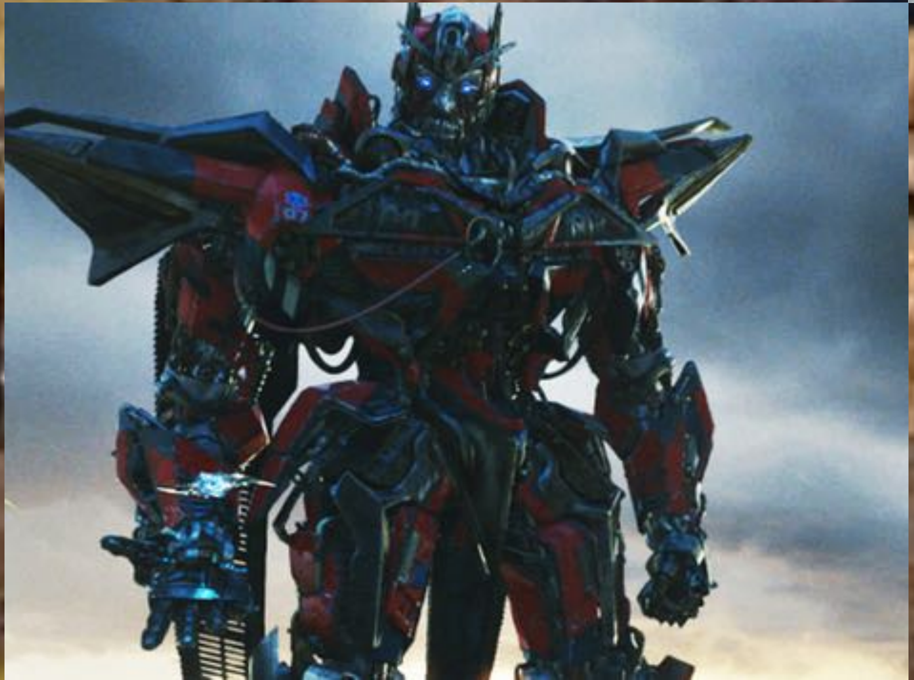
arXiv.org > cs > arXiv:1706.03762

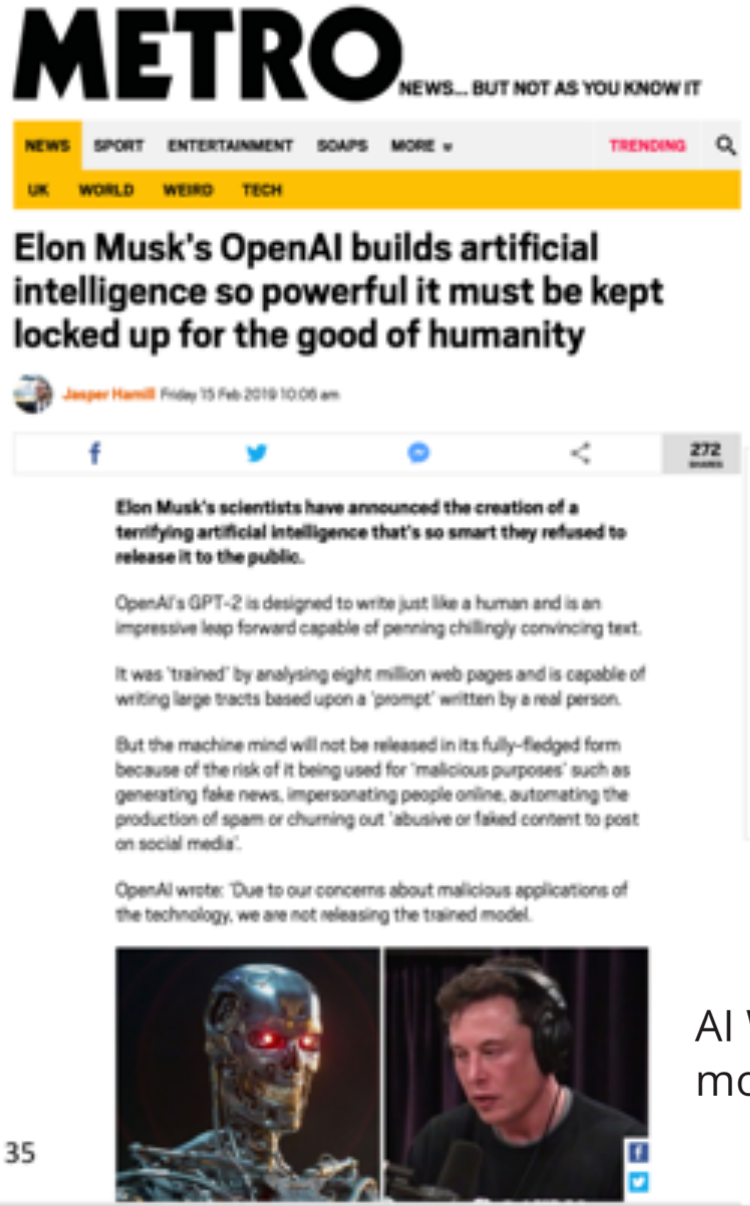
Computer Science > Computation and Language

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion

(Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5))




METRO NEWS... BUT NOT AS YOU KNOW IT

NEWS SPORT ENTERTAINMENT SOAPS MORE TRENDING

UK WORLD WEIRD TECH

Elon Musk's OpenAI builds artificial intelligence so powerful it must be kept locked up for the good of humanity

Jasper Hamill Friday 15 Feb 2019 10:06 am

272

Elon Musk's scientists have announced the creation of a terrifying artificial intelligence that's so smart they refused to release it to the public.

OpenAI's GPT-2 is designed to write just like a human and is an impressive leap forward capable of penning chillingly convincing text.

It was 'trained' by analysing eight million web pages and is capable of writing large tracts based upon a 'prompt' written by a real person.

But the machine mind will not be released in its fully-fledged form because of the risk of it being used for 'malicious purposes' such as generating fake news, impersonating people online, automating the production of spam or churning out 'abusive or faked content to post on social media'.

OpenAI wrote: 'Due to our concerns about malicious applications of the technology, we are not releasing the trained model.


Elon Musk
@elonmusk

Follow

Replying to @georgezachary

To clarify, I've not been involved closely with OpenAI for over a year & don't have mgmt or board oversight

8:19 PM - 16 Feb 2019

500 Retweets 14,573 Likes

229 500 15K

AI Weekly: Experts say OpenAI's controversial model is a potential threat to society and science

- Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text, 1.5 B parameters
- Due to our concerns about malicious applications of the technology, we are not releasing the trained model.
- As an experiment in responsible disclosure, *we are instead releasing a much smaller model* for researchers to experiment with, as well as a technical paper.

AI Too Dangerous to Release ?

AI is not Galloping Horses ..

It has semantics ..

GPT-2 : Open AI Staged Release

- 124 M Model in Feb'19
- 355 M - May'19
- 774 M - Aug'19
- 1,558 M - ? Nov'19

Since then …

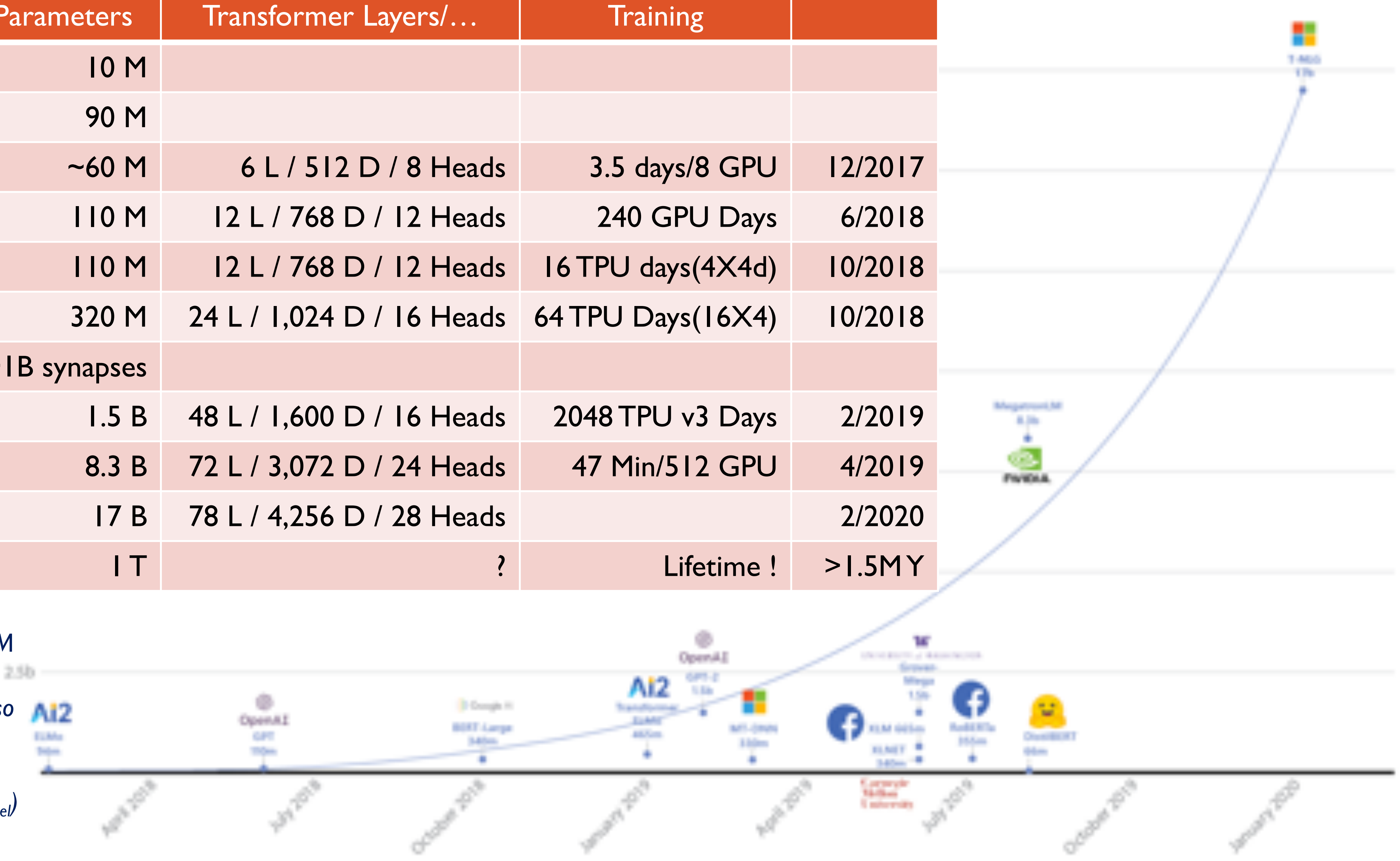
Nvidia GPT-2 8B in 53 minutes with 1,500 GPUs (Apr'19) …

Salesforce 1.6B Parameters (Sep'19) …

Microsoft Turing-NLG 17B Parameters (Feb'20) …

SCALE OF THINGS

Model	Parameters	Transformer Layers/…	Training	
Medium-sized LSTM	10 M			
ELMo	90 M			
Transformers	~60 M	6 L / 512 D / 8 Heads	3.5 days/8 GPU	12/2017
GPT	110 M	12 L / 768 D / 12 Heads	240 GPU Days	6/2018
BERT(Base)	110 M	12 L / 768 D / 12 Heads	16 TPU days(4X4d)	10/2018
BERT (L)	320 M	24 L / 1,024 D / 16 Heads	64 TPU Days(16X4)	10/2018
Honeybee Brain	~1B synapses			
GPT-2 (XL)	1.5 B	48 L / 1,600 D / 16 Heads	2048 TPU v3 Days	2/2019
Megatron GPT-2	8.3 B	72 L / 3,072 D / 24 Heads	47 Min/512 GPU	4/2019
Turing-NLG	17 B	78 L / 4,256 D / 28 Heads		2/2020
Human Brain Models	1 T	?	Lifetime !	>1.5M Y



- *GPT-2 : Really big Transformer LM trained on 40GB text*
- *Model > 1.3B won't fit in GPU, so need model parallelization and broken across multiple GPUs*
- *D = Model Dimensionality (d_{model})*

SCALE OF THINGS

Model	Parameters	Transformer Layers/…	Training	
Medium-sized LSTM	10 M			
ELMo	90 M			
Transformers	~60 M	6 L / 512 D / 8 Heads	3.5 days/8 GPU	12/2017
GPT	110 M	12 L / 768 D / 12 Heads	240 GPU Days	6/2018
BERT(Base)	110 M	12 L / 768 D / 12 Heads	16 TPU days(4X4d)	10/2018
BERT (L)	320 M	24 L / 1,024 D / 16 Heads	64 TPU Days(16X4)	10/2018
Honeybee Brain	~1B synapses			
GPT-2 (XL)	1.5 B	48 L / 1,600 D / 16 Heads	2048 TPU v3 Days	2/2019
Megatron GPT-2	8.3 B	72 L / 3,072 D / 24 Heads	47 Min/512 GPU	4/2019
Turing-NLG	17 B	78 L / 4,256 D / 28 Heads		2/2020
Human Brain Models	1 T			

- *GPT-2 : Really big Transformer LM trained on 40GB text*
- *Model > 1.3B won't fit in GPU, so need model parallelization and broken across multiple GPUs*
- *D = Model Dimensionality (d_{model})*

Megatron-LM GPT-2 Evaluation

Model Size	Number of layers	Hidden Size	Attention heads	Wikitext (Perplexity ↓)	Lambada (Accuracy ↑)
345 M	24	1024	16	24.21	55.04
775 M	36	1280	16	20.44	58.86
2.5 B	54	1920	20	17.83	63.30
8.3 B	72	3072	24	17.41	63.11

GPT-2 WRITINGS

A large monkey wearing a top hat and spats broke loose from a zoo in downtown Chattanooga today.

The monkey escaped from its cage at the Chattanooga Zoo at about 12:45 p.m.

Advertisement

The animal has been taken to a local veterinarian for evaluation. The zoo is closing for the day.

Animal control personnel were called to the zoo to help contain the monkey.

No zoogoers were in the building at the time of the incident.

The Tennessee Department of Agriculture says the monkey is a female, between 10 and 16 years old, weighing about 100 pounds.

'''The Transformer''' are a Japanese [[hardcore punk]] band.

==Early years==

The band was formed in 1968, during the height of Japanese music history. Among the legendary [[Japanese people|Japanese]] composers of [Japanese lyrics], they prominently exemplified Motohiro Oda's especially tasty lyrics and psychedelic intention. Michio was a longtime member of the every Sunday night band PSM. His alluring was of such importance as being the man who ignored the already successful image and that he municipal makeup whose parents were – the band was called Jenei.<ref>http://www.separatist.org/se_frontend/post-punk-musician-the-kidney.html</ref>

From a young age the band was very close, thus opting to pioneer what

From a young age the band was very close, thus opting to pioneer what had actually begun as a more manageable core hardcore punk band.<ref>http://www.talkradio.net/article/independent-music-fades-from-the-closed-drawings-out</ref>

==History==

===Born from the heavy metal revolution===

In 1977 the self-proclaimed King of Tesponsors, [[Joe Lus:

: It was somewhere... it was just a guile ... taking this song to Broadway. It was the first record I ever heard on A.M., After some opposition I received at the hands of Parsons, and in the follow-up notes myself.<ref>http://www.discogs.com/artist/The+Op%C5%8Dn+&+Psalm</ref>

The band cut their first record album titled "Transformed, furthered

Makes up a quote from the book – knows that wikipedia quotes from books

Model trained on 2 days w/ 8 GPUs

The band cut their first record album titled "Transformed, furthered and extended Extended",<ref>[https://www.discogs.com/album/69771 MC – Transformed EP (CDR) by The Moondrawn – EMI, 1994]</ref> and in 1978 the official band line-up of the three-piece pop-punk-rock band TEEM. They generally played around [[Japan]], growing from the Top 40 standard.

All made up ! These people don't exist, neither does the band !

===1981-2010: The band to break away===

On 1 January 1981 bassist Michio Kono, and the members of the original line-up emerged. Niji Fukune and his [[Head poet|Head]] band (now guitarist) Kazuya Kouda left the band in the hands of the band at the May 28, 1981, benefit season of [[Led Zeppelin]]'s Marmarin building. In June 1987, Kono joined the band as a full-time drummer, playing a

Japanese band, Japanese sounding names

LINGUIST VIEW

Co-occurrence and distributional similarity

Firth (1957)

"You shall know a word by the company it keeps."

http://www.lel.ed.ac.uk/homes/patrick/firth.pdf

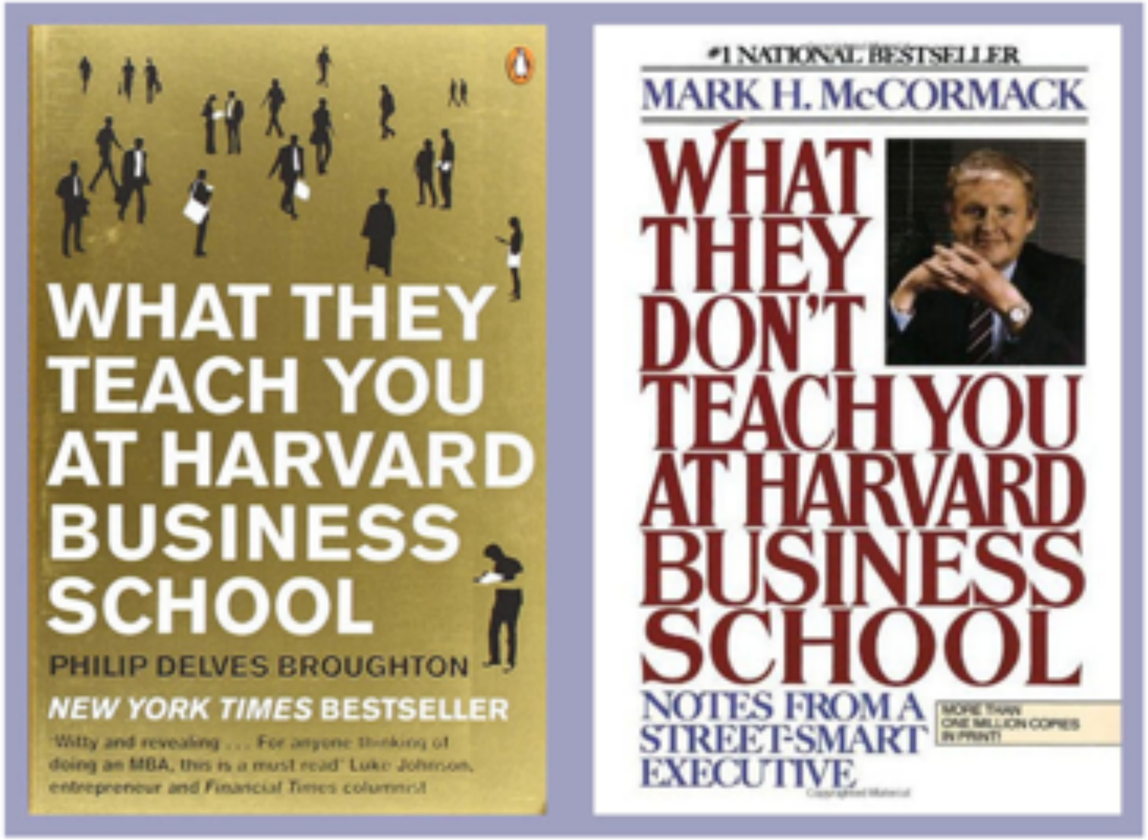
Firth (1957)

"the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously."

Ref : cs224u Stanford

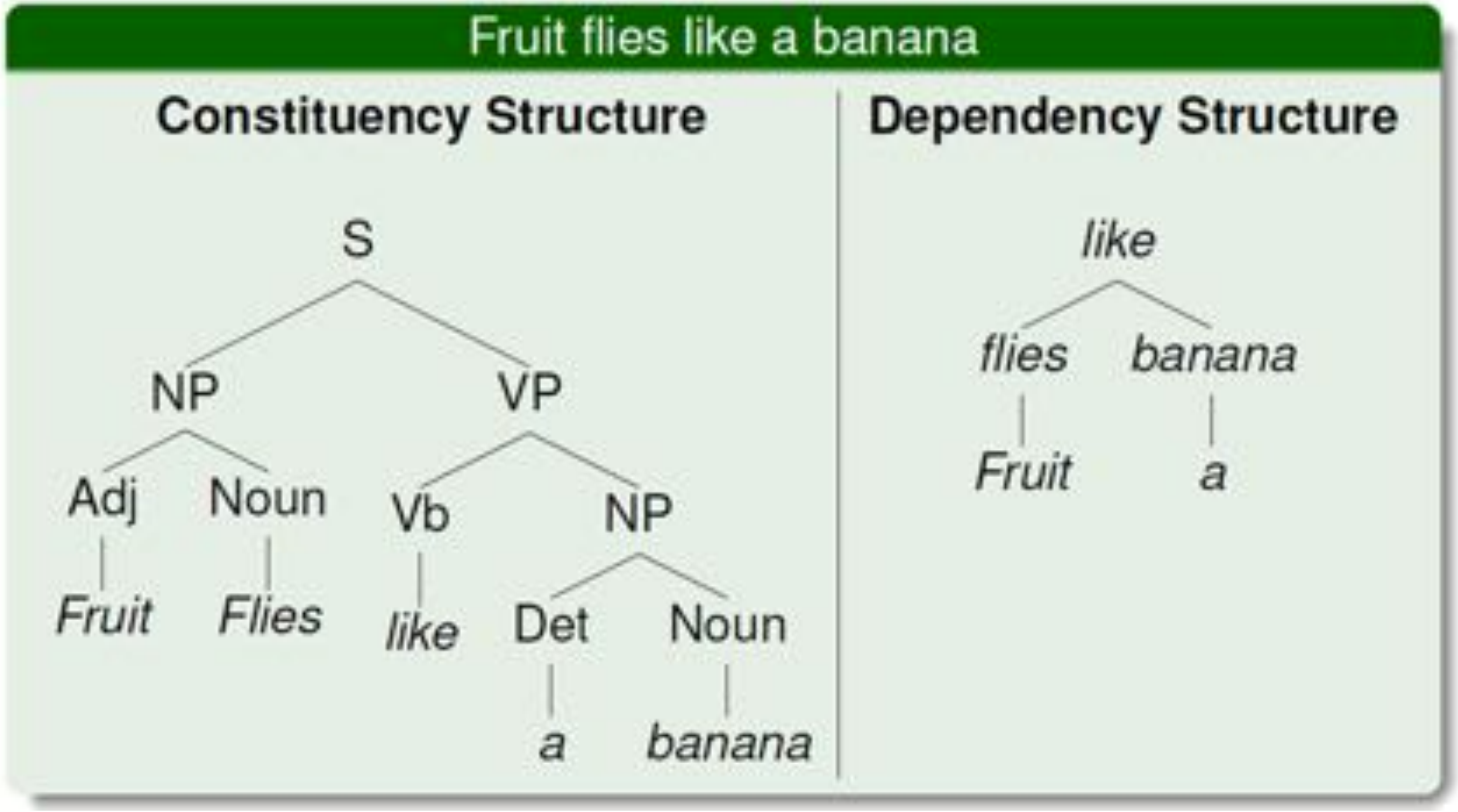
CONTEXT DEPENDENCE

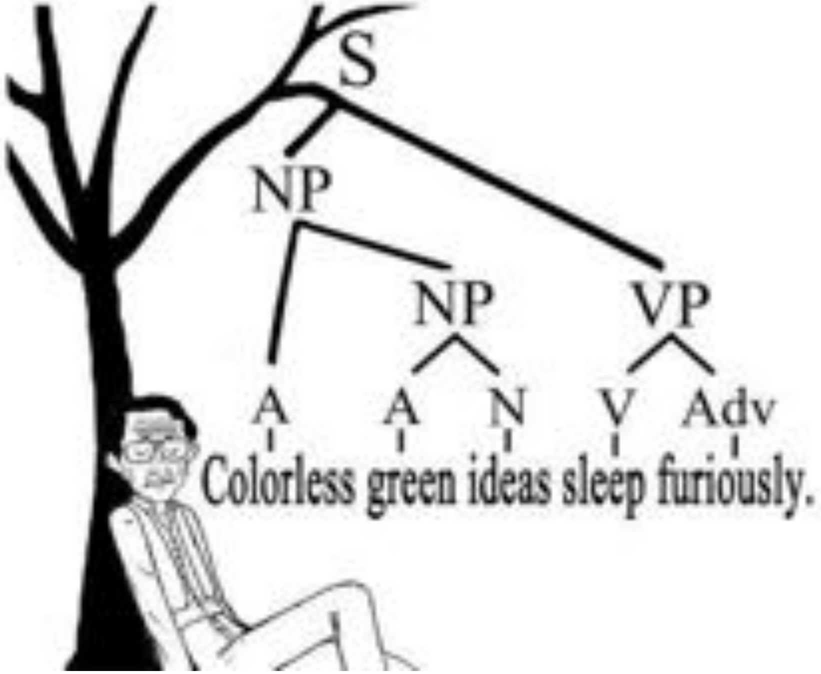
What is said, alone is not enough to resolve a context !



"These two books contain the sum total of all human knowledge" (@James_Kpatrick)

Perspectival expressions





https://towardsdatascience.com/beyond-word-embeddings-part-4-introducing-semantic-structure-to-neural-nlp-96cf8a2723fb

WINOGRAD SENTENCES

- Winograd, Terry. 1972. Understandi
  191
- Levesque, Hector J. 2013. On our b
  international conference on artificial in
- Wang, Alex, Amanpreet Singh, Julian
  2018. GLUE: A multi-task benchmar
  understanding. In *Proceedings of the*
  interpreting neural networks for NLP, 353–355. Brussels, Belgium: Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-5446.

Winograd sentences

1. The trophy doesn't fit into the brown suitcase because it's too **small**. What is too small?
   The suitcase / The trophy
2. The trophy doesn't fit into the brown suitcase because it's too **large**. What is too large?
   The suitcase / **The trophy**
3. The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
   The council / The demonstrators
4. The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
   The council / **The demonstrators**

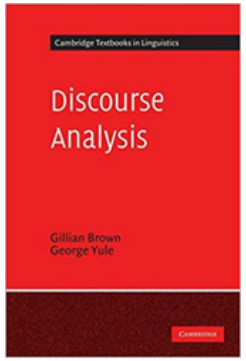
(Winograd 1972; Levesque 2013; Wang et al. 2018)

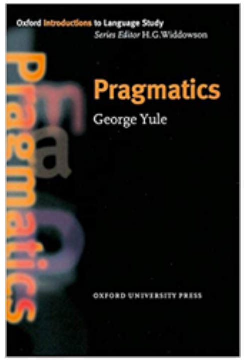
When you translate to French trophy and suitcase has different genders. So you can't translate without understanding what is happening in the world !

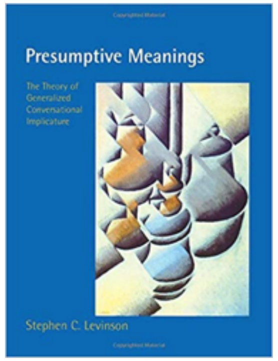
WORD REPRESENTATIONS AND CONTEXT

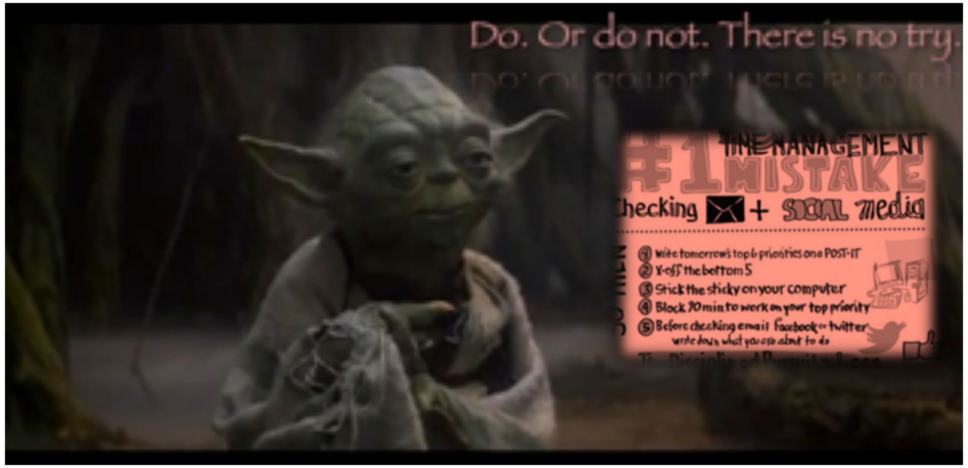
- a. The vase *broke*
  b. Dawn *broke*
  c. The news *broke*
  d. Sandy *broke* the world record
  e. Sandy *broke* the law
  f. The burglar *broke* into the house.
  g. The newscaster *broke* into the movie broadcast
  h. We *broke* even
- a. kids *play* a game in the park
  b. the Broadway *play* premiered yesterday
- *flat* tire/beer/note/surface
- *throw* a party/fight/ball/fit
- a. A *crane* caught a fish.
  b. A *crane* picked up the steel beam.
  c. I saw a *crane*
- a. Are there typos? I didn't see any.
  b. Are there bookstores downtown? I didn't see any.

Features	Themes	Levels
Limited Vocabulary, Limited actions e.g. lights on	Commands & Notifications	Notification (& IoT Command) Bots
Search & traverse through text, API based intent resolution	Q & A	FAQ Bots
Knowledge Graph (explicit or implicit intent based), some level of comprehension & capable of contextual conversation	Dialogues	Contextual Bots
Inferential, Predictive, Personalized and Branded for unique & differentiated digital experience	Interactions	Smart Personalized Bots
Human-level comprehension & common sense, Factoids, Knowledge Graph and capable of intelligent interactivity over multiple conversation chains	Conversations	Autonomous Bots









COMMONSENSE REASONING

- *A process that involves taking information about certain aspects of a scenario in the world and making inferences about other aspects of the scenario based on our commonsense knowledge*
- *Events, Fluents & TimePoints*
  - *Fluent - represents a time-varying property of the world, such as the location of a physical object*
- *The Commonsense Law of Inertia, The Mental state of agents, Default Reasoning*

- In the living room, Lisa picked up a newspaper and walked into the kitchen. Where did the newspaper end up? *It ended up in the kitchen.*
- Kate set a book on a coffee table and left the living room. When she returned, the book was gone. What happened to the book? *Someone must have taken it.*
- Jamie walks to the kitchen sink, puts the stopper in the drain, turns on the faucet, and leaves the kitchen. What will happen as a result? *The water level will increase until it reaches the rim of the sink. Then the water will start spilling onto the floor.*
- Kimberly turns on a fan. What will happen? *The fan will start turning.* What if the fan is not plugged in? *Then the fan will not start turning.*
- A hungry cat saw some food on a nearby table. The cat jumped onto a chair near the table. What was the cat about to do? *The cat was about to jump from the chair onto the table in order to eat the food.*

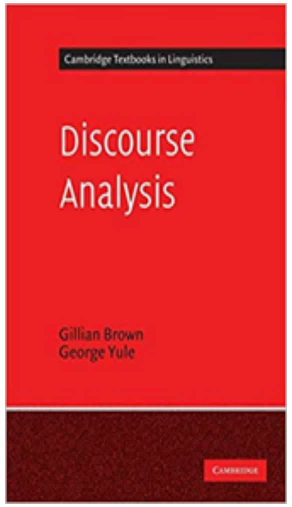
https://medium.com/@ksankar/commonsense-reasoning-the-education-of-a-machine-part-5-7-f2dd25e01de2

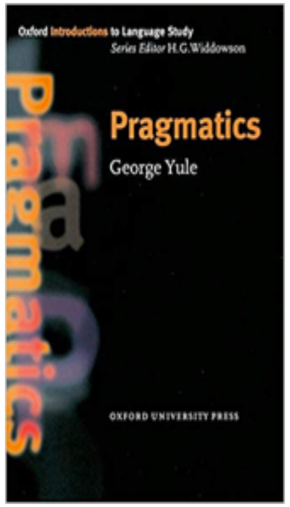
COMMONSENSE REASONING

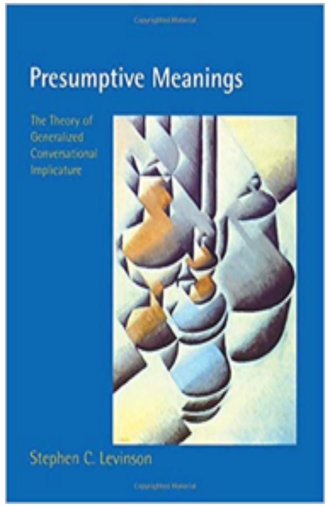
...mmonsense-reasoning-the-education-of-a-machine-part-5-7-f2dd25e01de2

CONVERSATION IMPLICATURE

- ***Sophisticated & robust common sense understanding of the world won't come from pattern matching on examples***
- Grounding - Social Cues, Physical Arrangement, Assumptions about the speaker's goals
- More sophisticated reasoning about other agents and their goals
- ***System should learn the default stuff from outside the conversation***
- Paul Grice & Gricean reasoning
  - Maxim of quality
  - Maxim of quantity – *As informative as required and not more*
  - Maxim of relation (or relevance) - *Be relevant*
  - Maxim of manner – *Be Brief and orderly*
  - Maxim of Smartness







David Lewis
Philosopher

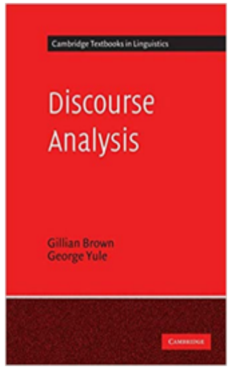
J.L Austin
Linguist

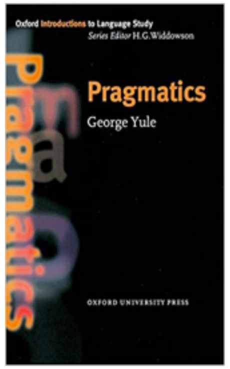
Ref: Stanford : Linguist 130a/230a, cs224u

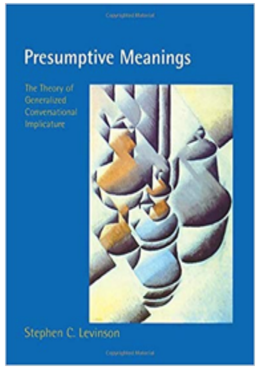
CONVERSATION IMPLICATURE & BERT

- Conversational AI roBots : Pragmas & Practices

  1. When thinking about Bots, *think guided Search*, not a single Question-Answer session. This requires a little deep thought — the bot should want to understand what the user wants, participate in an *iterative collaborative interaction* to facilitate what the user is looking for
  2. Model paths other than happy paths and fallback choreographies — *fail gracefully*
  3. Bots have a *Visual Branding* as well as a personality – implied or explicit
  4. *Aesthetics* (incl small talk) is very important. Don't turn a web site into a bot
  5. Surprise your users with *Acuity and Serendipity* !







Show up on time, know the text & have a head full of ideas

Have the freedom of being there early enough to settle down and gather your thoughts -- because when the time comes, you have to hit the marks …

Knowing your text--it's not just your lines, it's the whole thing, … You might not be right in the opinion you bring to it. But you've got to come at it with some direction

The head full of ideas: Bring anything. Try anything. They might not use it. If it stinks, they won't use it. *Am I right, Marty Scorsese?* Growth is a result of being willing to take risks, to break out of your comfort zone, and to embrace failure when it happens

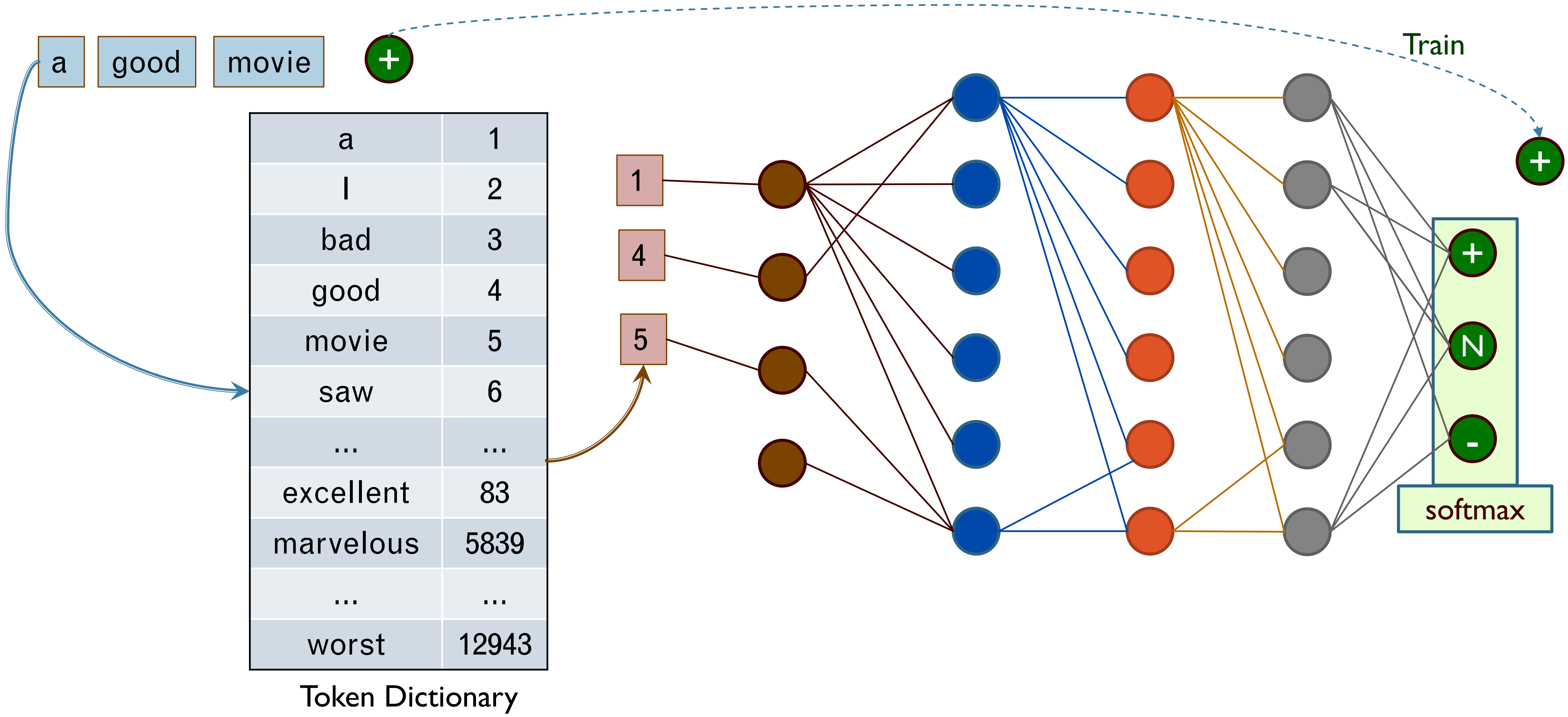
https://medium.com/@ksankar/conversational-ai-robots-pragmas-practices-21-lessons-one-could-learn-watching-the-wise-yoda-d0f81bdee2ea

THE GORY DETAILS

THE GORY DETAILS

Photo by Arseny Togulev on Unsplash

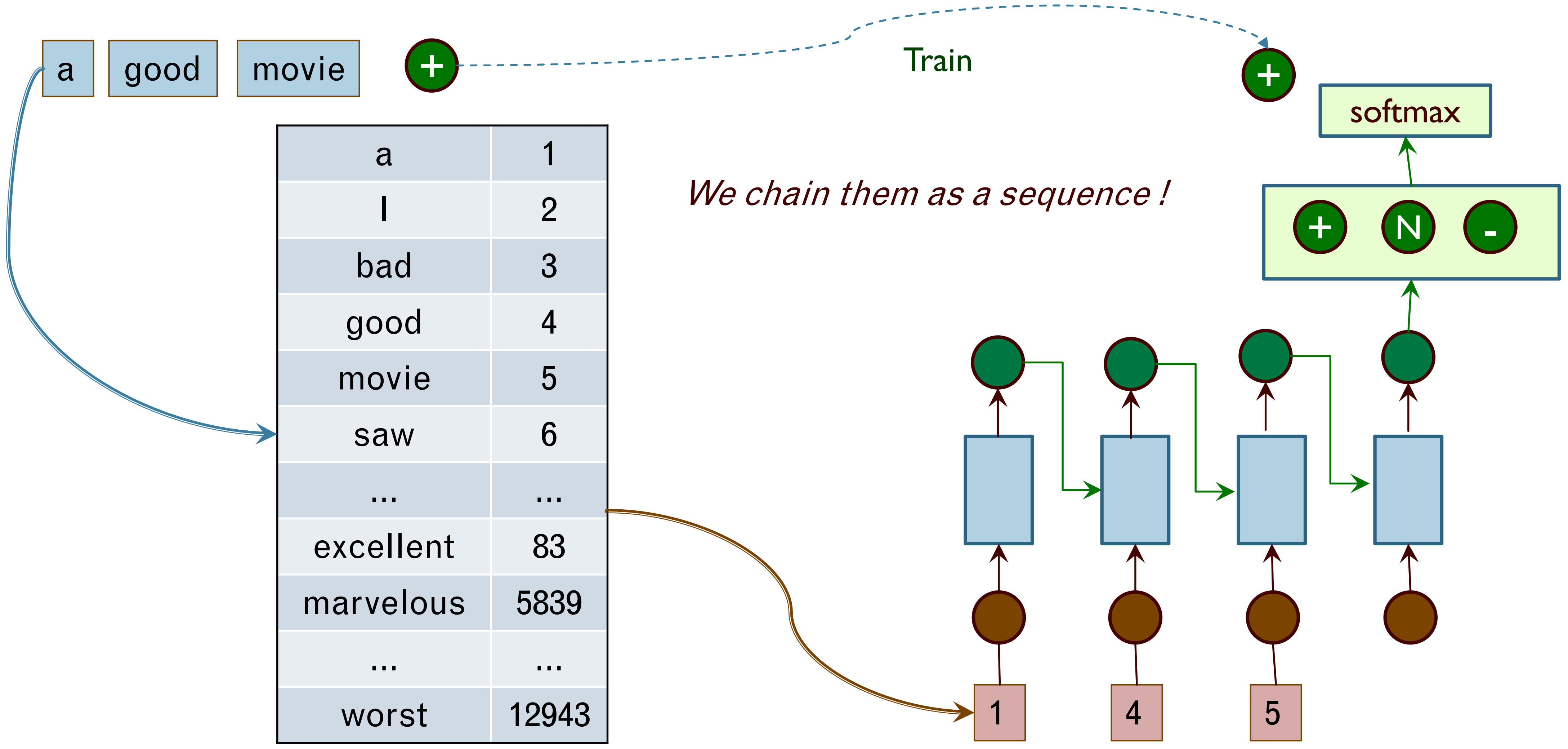
SENTIMENT ANALYSIS – FULLY CONNECTED NN



SENTIMENT ANALYSIS – FULLY CONNECTED NN

- This will work fine for simple sentences
- But what about
  - "The book is really good, but the movie is dreadful" -
  - "The screenplay and direction were done by people who couldn't fathom what was good about the novel" -
  - "It is much better than my old blackberry, which was a terrible phone, so difficult to type with small keys" +
  - "I am not sure if I like the new design" N
  - "It's not horrible, just horribly mediocre."* -
  - "I don't think there is anything I really dislike about the product" N
  - "The product is decent, but your website is so confusing it took me forever to find the product I was looking for" ?
  - "*The cast is uniformly excellent … but the film itself is merely mildly charming.*"* N
- We need to look at sentences as sequences – of course we have a solution -- RNN !

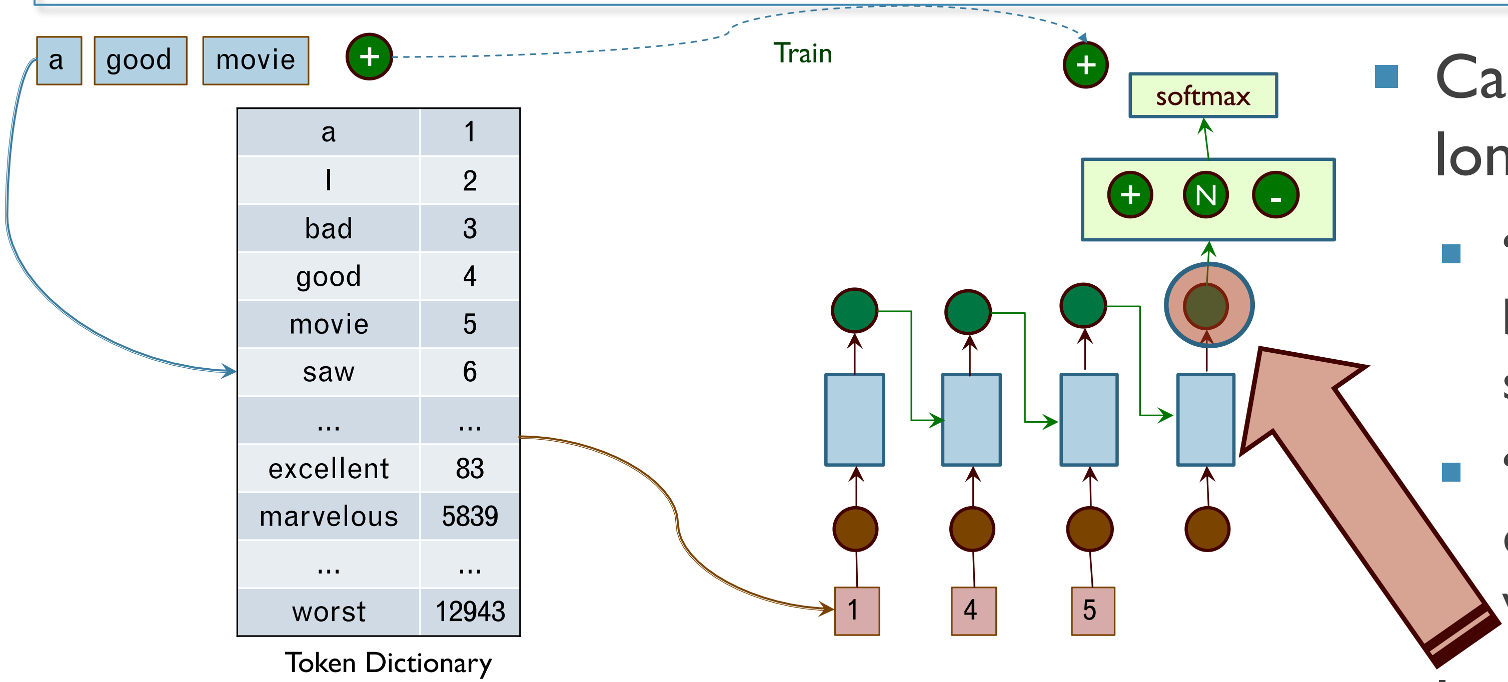
RECURRENT NEURAL NETWORKS



a	1
I	2
bad	3
good	4
movie	5
saw	6
...	...
excellent	83
marvelous	5839
...	...
worst	12943

Token Dictionary

CHALLENGES WITH RNNS



- Can't carry context/relevance over long sentences
  - "It is much better than my old blackberry, which was a terrible phone, so difficult to type with small keys"
  - "The screenplay and direction were done by people who couldn't fathom what was good about the novel"
- It is done by one vector - *bottleneck*
- Solution :
  - Use LSTM/GRU to select contexts
  - Stack the RNNs
- Can't Parallelize due to sequential nature
  - Speed limit - "Unrolling forces sequential calculation"
- It is unidirectional
  - For some tasks, it might be OK, but for others bi directionality helps (as BERT shows)

“Without forgetting it is quite impossible to live at all.”— Friedrich Nietzsche

LSTM

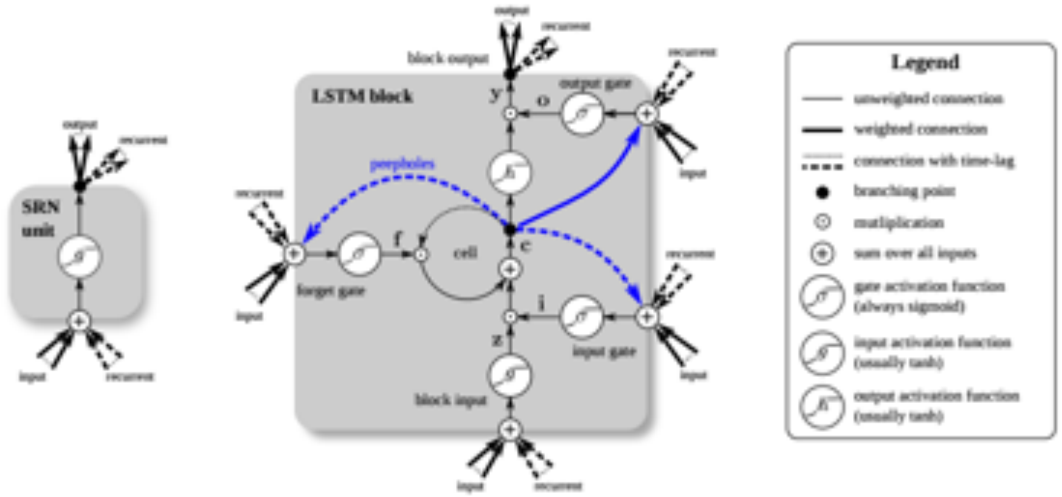
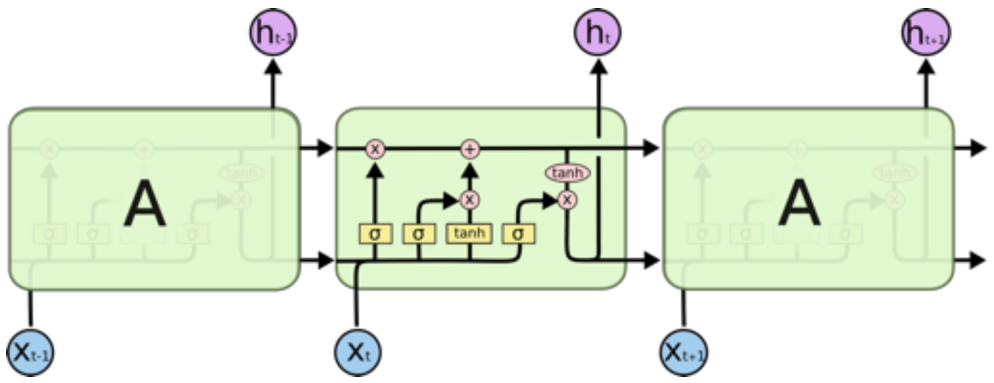
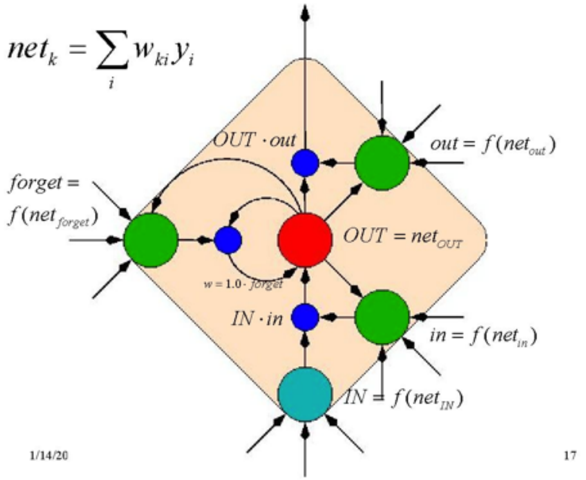


Figure 1. Detailed schematic of the Simple Recurrent Network (SRN) unit (left) and a Long Short-Term Memory block (right) as used in the hidden layers of a recurrent neural network.



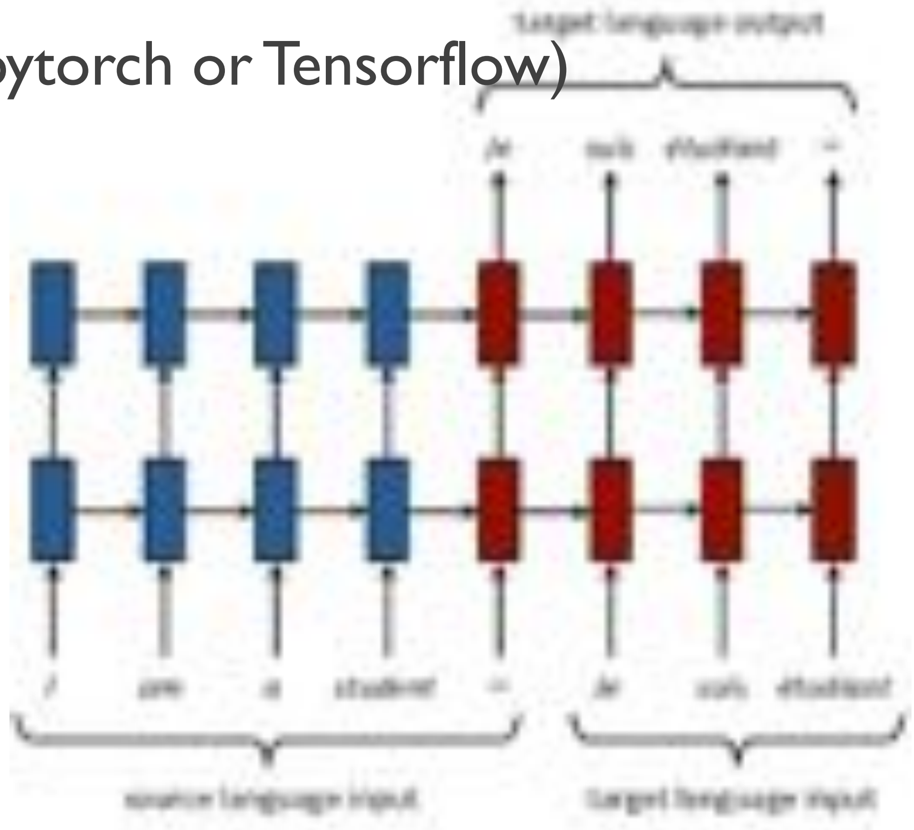


- LSTMs keep an internal context
- They can choose to forget, choose the degree of output based on the internal context and choose to keep (or forget) portions as internal context (kind of memory)
- GRUs are simpler versions of the LSTM with the same advantages

A Beginner's Guide to LSTMs and Recurrent Neural Networks https://pathmind.com/wiki/lstm
https://colah.github.io/posts/2015-08-Understanding-LSTMs/
http://people.idsia.ch/~juergen/lstm/sld017.htm

AND FINALLY, WE HAVE A WORKING ARCHITECTURE !!

- That is how we get to a very familiar diagram that we see in papers
- Stacked RNNs (may be LSTM or GRU – easier with pytorch or Tensorflow)
- Encoder - Decoder architecture
- Applicable across a larger set of NLP tasks
  - Sentiment Analysis
  - Translation (NMT)
  - Generation (NLG)
  - Understanding (NLU)
- We are not done yet !
  - We have the pesky question of the word representation i.e. the token dictionary



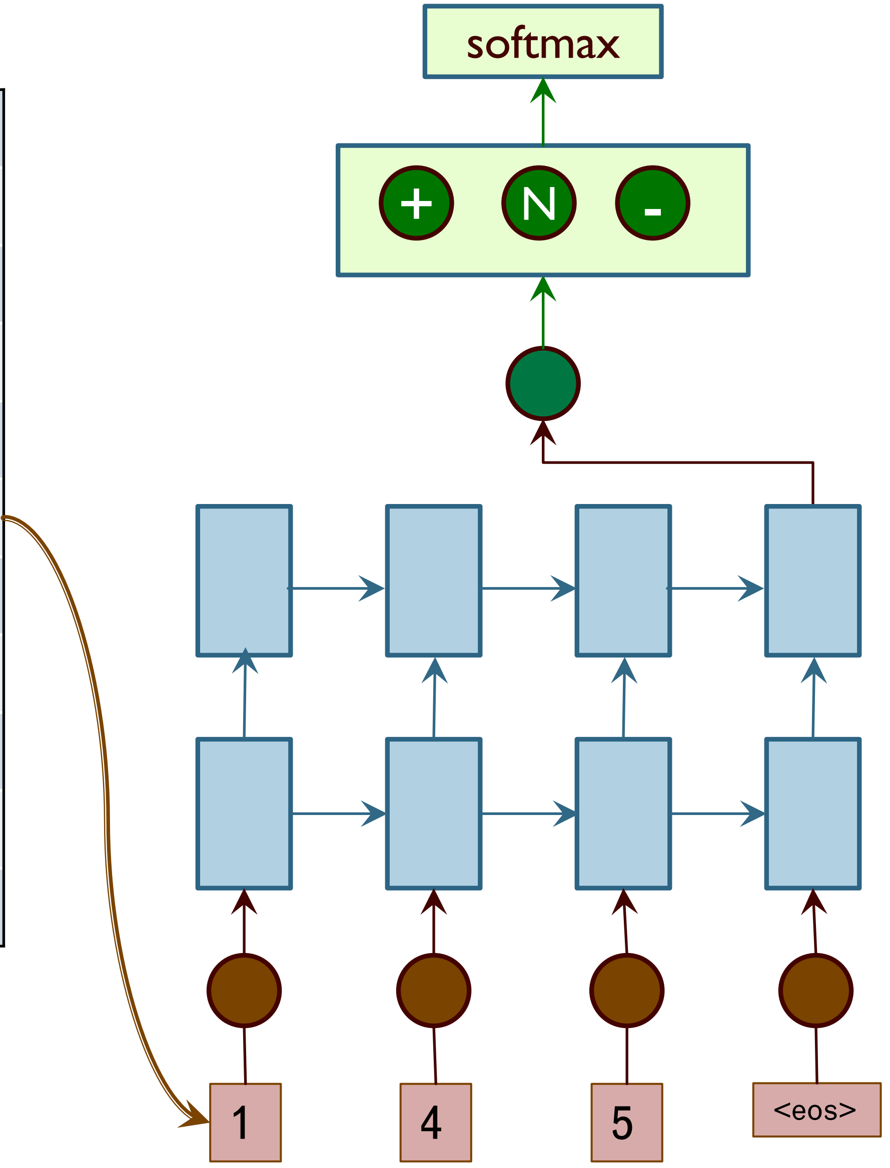
https://www.researchgate.net/figure/A-simplified-diagram-of-NMT_fig1_306093550

CAN WE DO BETTER ?

- The word representation is just a number
- It doesn't capture any semantic meaning of the words
- Solution : Word embeddings !

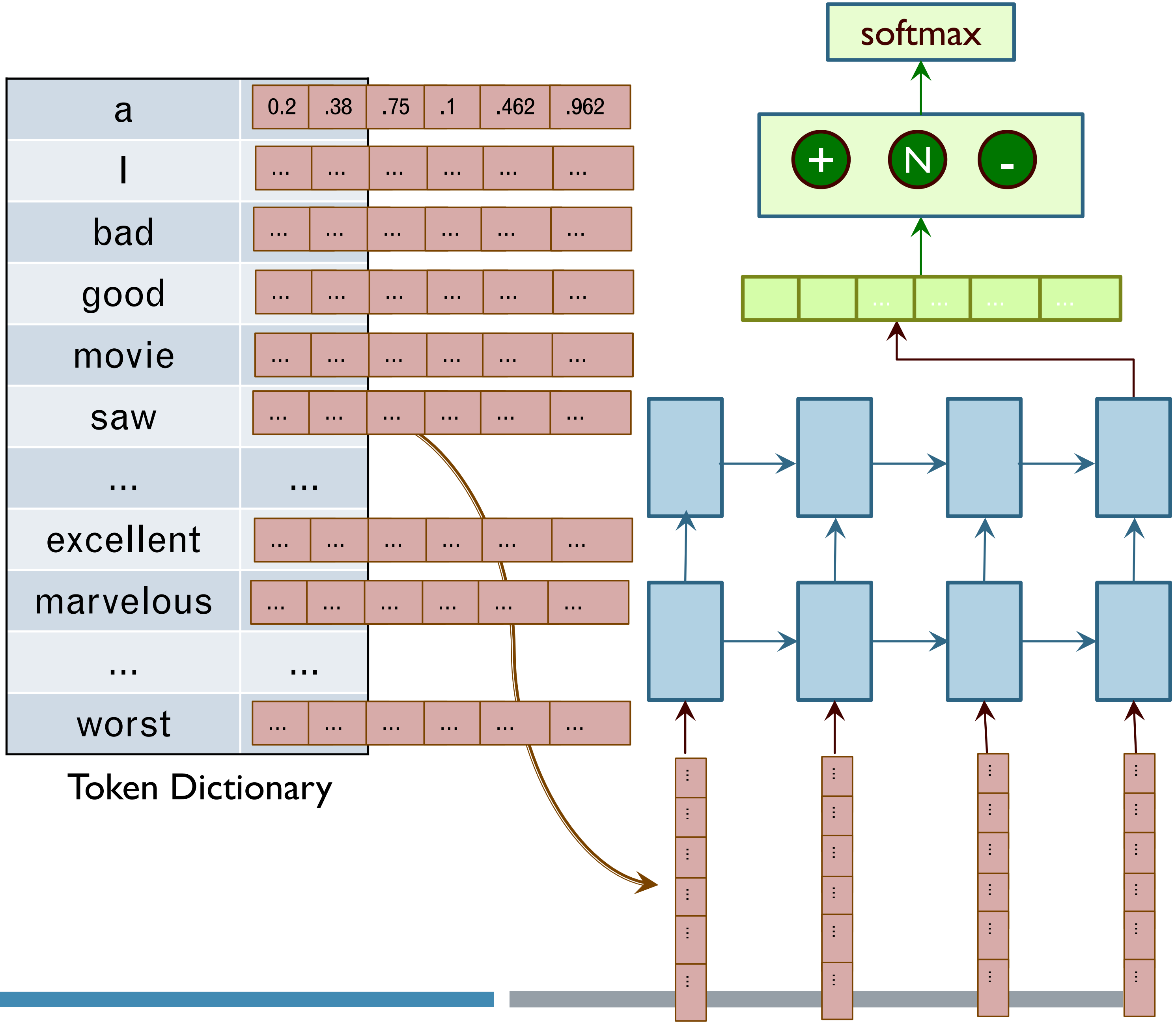
a	1
I	2
bad	3
good	4
movie	5
saw	6
...	...
excellent	83
marvelous	5839
...	...
worst	12943

Token Dictionary



WORD EMBEDDINGS

- Instead of a token dictionary, we use an embedding vector for each word
- Usually 300 long, could be 1000 or 3000
- Which means we get 1000 parameters to represent a word
- And, the embedding can be learned from the huge corpus we have, without any labels !
- And, reused everywhere !!



LINGUISTIC VIEW

- Word embeddings are a mathematical way to identify and cluster in a mathematical space, words which "live" nearby each other in a real-world collection of text
- Chris Potts cs224u
  - Pluck average linguist off the street and say birds could be represented by a set of vectors, it would be tough going; but after some backend forth, they could be convinced, as this is how they think word meaning viz. high dimensional objects with lots of connections
  - But would lose them if we talk about words in isolation, word senses are shaped by their morphosyntactic environment, and usage by the discourse context; can get them on board with ELMO-BERT

WORD EMBEDDINGS - INTUITION

- Suppose you were asked to classify words W1, W2, W3 and W4
- You are also given a class Map
- This is an impossible task !

Class	Word
0	awful
0	terrible
0	lame
0	worst
0	disappointing
1	nice
1	amazing
1	wonderful
1	good
1	awesome

Pr(Class = 1)	Word
?	w_1
?	w_2
?	w_3
?	w_4

Ref: CS224u Stanford Spring 2019

WORD EMBEDDINGS - INTUITION

- What if you are also given a 2-Dimensional embedding ?
- Now the task becomes solvable
- Even without knowing what the words are !
- You just need the embeddings !!

Class	Word	excellent	terrible
0	awful	−0.69	1.13
0	terrible	−0.13	3.09
0	lame	−1.00	0.69
0	worst	−0.94	1.04
0	disappointing	0.19	0.09
1	nice	0.08	−0.07
1	amazing	0.71	−0.06
1	wonderful	0.66	−0.76
1	good	0.21	0.11
1	awesome	0.67	0.26

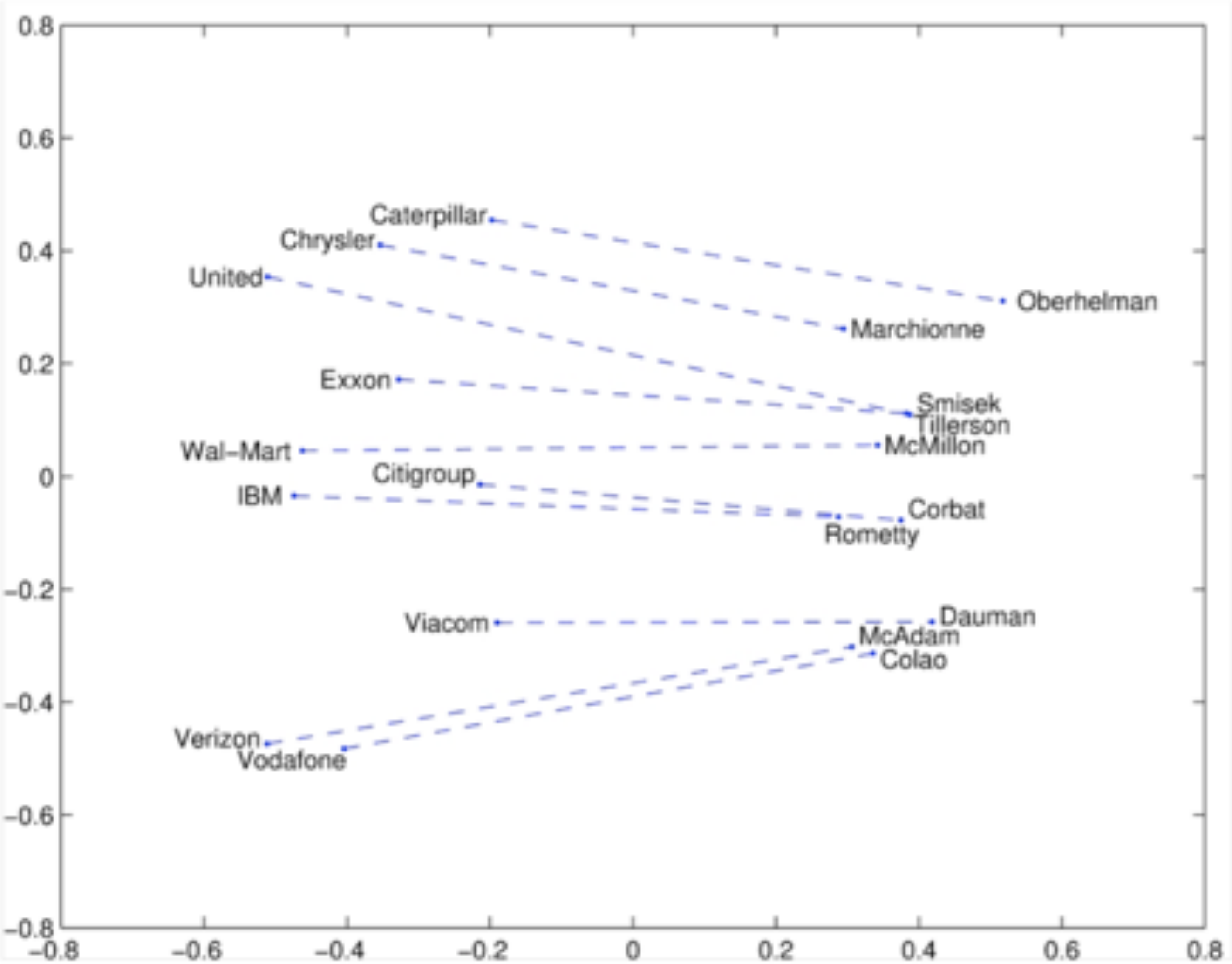
Pr(Class=1)	Word	excellent	terrible
≈ 0	w_1	−0.47	0.82
≈ 0	w_2	−0.55	0.84
≈ 1	w_3	0.49	−0.13
≈ 1	w_4	0.41	−0.11

Ref: CS224u Stanford Spring 2019

Interesting semantic patterns emerge in the vectors

COALS model from
An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence
Rohde et al. ms., 2005

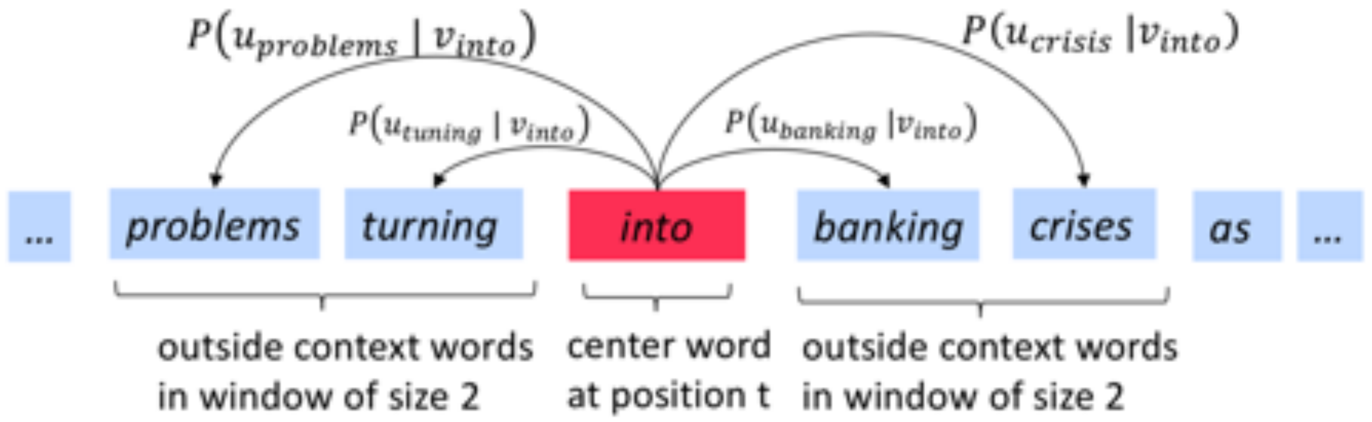
Glove Visualizations: Company - CEO



Ref: CS224u Stanford Spring 2019

TRAINING EMBEDDINGS

- It is a complex task, this is just a simple illustration



- "You shall know a word by the company it keeps"
- Assign random vector of size say 300 or 768 or 3000 to each work
- Use text corpus, sliding window over the text, word vectors are nudged around to minimize surprise
- Iterate until "good enough"
- The vector space of words self-organizes

Ref: CS224u Stanford Spring 2019
https://redcatlabs.com/2018-11-21_TFandDL_BERT/#/3/2

Standard RNN dataset preparation

Embedding			
1	−0.42	0.10	0.12
2	−0.16	−0.21	0.29
3	−0.26	0.31	0.37

Examples

[a, b, a]
[b, c]

$\Downarrow$

Indices

[1, 2, 1]
[2, 3]

$\Downarrow$

Vectors

$\left[[-0.42\ 0.10\ 0.12], [-0.16\ -0.21\ 0.29], [-0.42\ 0.10\ 0.12]\right]$

$\left[[-0.16\ -0.21\ 0.29], [-0.26\ 0.31\ 0.37]\right]$

```
[7]: print(wv.doesnt_match(['fire', 'water', 'land', 'sea', 'air', 'car']))
```

```
car
```

```
[8]: pairs = [
         ('car', 'minivan'),   # a minivan is a kind of car
         ('car', 'bicycle'),   # still a wheeled vehicle
         ('car', 'airplane'),  # ok, no wheels, but still a vehicle
         ('car', 'cereal'),    # ... and so on
         ('car', 'communism'),
     ]
     for w1, w2 in pairs:
         print('%r\t%r\t%.2f' % (w1, w2, wv.similarity(w1, w2)))
```

```
'car'	'minivan'	0.69
'car'	'bicycle'	0.54
'car'	'airplane'	0.42
'car'	'cereal'	0.14
'car'	'communism'	0.06
```

WORD2VECTOR SEMANTICS

- King - Man + Woman = ?
  - wv.most_similar(positive=['woman', 'king'], negative=['man'])
  - ('queen', 0.7118192911148071)
- USA : Pizza :: Japan : ?
  - wv.most_similar(positive=['Japan','Pizza'], negative=['USA'])
  - ('Sushi', 0.5657287836074829)
- Water : Ice :: Liquid: ?
  - wv.most_similar(positive=['liquid', 'ice'], negative=['Water'])
  - ('unmelted', 0.5092144608497620),
    ('Methane_hydrate', 0.4641897082328796),
    ('Francies_tossed', 0.4573051035404205),
    ('ice_crystals', 0.4563573598861694),
    ('starch_granules', 0.4429359138011932),

What exactly is the Methane Hydrate ?
It took ice literally, couldn't abstract it

Try word distances

[15]:
```
# To try
w1 = "happy"
w2 = "cheerful"
w3 = "sad"
w1_w2_dist = wv.distance(w1, w2)
w1_w3_dist = wv.distance(w1, w3)
print("Synonyms {}, {} have cosine distance: {}".format(w1, w2, w1_w2_dist))
print("Antonyms {}, {} have cosine distance: {}".format(w1, w3, w1_w3_dist))
```

```
Synonyms happy, cheerful have cosine distance: 0.6162261664867401
Antonyms happy, sad have cosine distance: 0.46453857421875
```

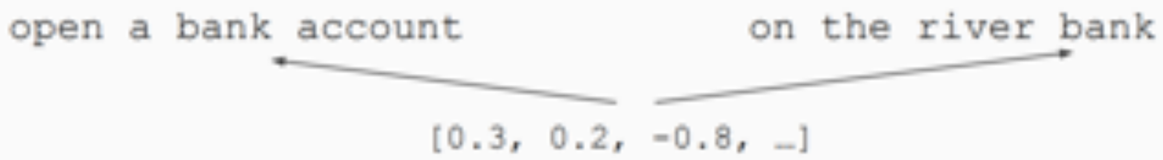
- This is an interesting side effect.
- Most probably "happy" and "sad" occur (in the corpus) near each other in sentences, more than "happy" and "cheerful".
- Also, because happy and cheerful are very close, probably most of the sentences use happy rather than cheerful

WORD REPRESENTATIONS AND CONTEXT

- a. The vase *broke*
  b. Dawn *broke*
  c. The news *broke*
  d. Sandy *broke* the world record
  e. Sandy *broke* the law
  f. The burglar *broke* into the house.
  g. The newscaster *broke* into the movie broadcast
  h. We *broke* even
- a. kids *play* a game in the park
  b. the Broadway *play* premiered yesterday
- *flat* tire/beer/note/surface
- *throw* a party/fight/ball/fit
- a. A *crane* caught a fish.
  b. A *crane* picked up the steel beam.
  c. I saw a *crane*
- a. Are there typos? I didn't see any.
  b. Are there bookstores downtown? I didn't see any.

CONTEXTUAL REPRESENTATIONS

- **Problem:** Word embeddings are applied in a context free manner



- **Solution:** Train *contextual* representations on text corpus

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

CONTEXTUAL EMBEDDING

- "After robbing the bank vault, the bank robber was seen fishing on the Mississippi river bank”
- BERT Representations (5 of 768)
  - bank (vault) tensor([0.8436, -0.4816, -0.0840, 0.4035, 0.6408])
  - bank (robber) tensor([0.8196, -0.4100, -0.1249, 0.3517, 0.5315])
  - river (bank) tensor([-0.3711, -0.6972, -0.6805, -0.1639, 0.4114])
- Vocabulary Size

Model	Vocab	Dimensionality
Spacy (W2V)	1,340,242	300
Gensym	3,000,000	300
GPT-2	50,257	768
BERT	30,522	768

```
0 (101, '[CLS]')
1 (2044, 'after')
2 (26211, 'robb')
3 (2075, '##ing')
4 (1996, 'the')
5 (2924, 'bank')
6 (11632, 'vault')
7 (1010, ',')
8 (1996, 'the')
9 (2924, 'bank')
10 (27307, 'robber')
11 (2001, 'was')
12 (2464, 'seen')
13 (5645, 'fishing')
14 (2006, 'on')
15 (1996, 'the')
16 (5900, 'mississippi')
17 (2314, 'river')
18 (2924, 'bank')
19 (102, '[SEP]')
```

- *BERT uses WordPiece tokens, where the non-word-initial pieces start with ###*
- *GPT2, RoBERTa use the BPE (Byte-Pair Encoding), \u0120 as the special signaling character*

CONTEXTUAL EMBEDDING

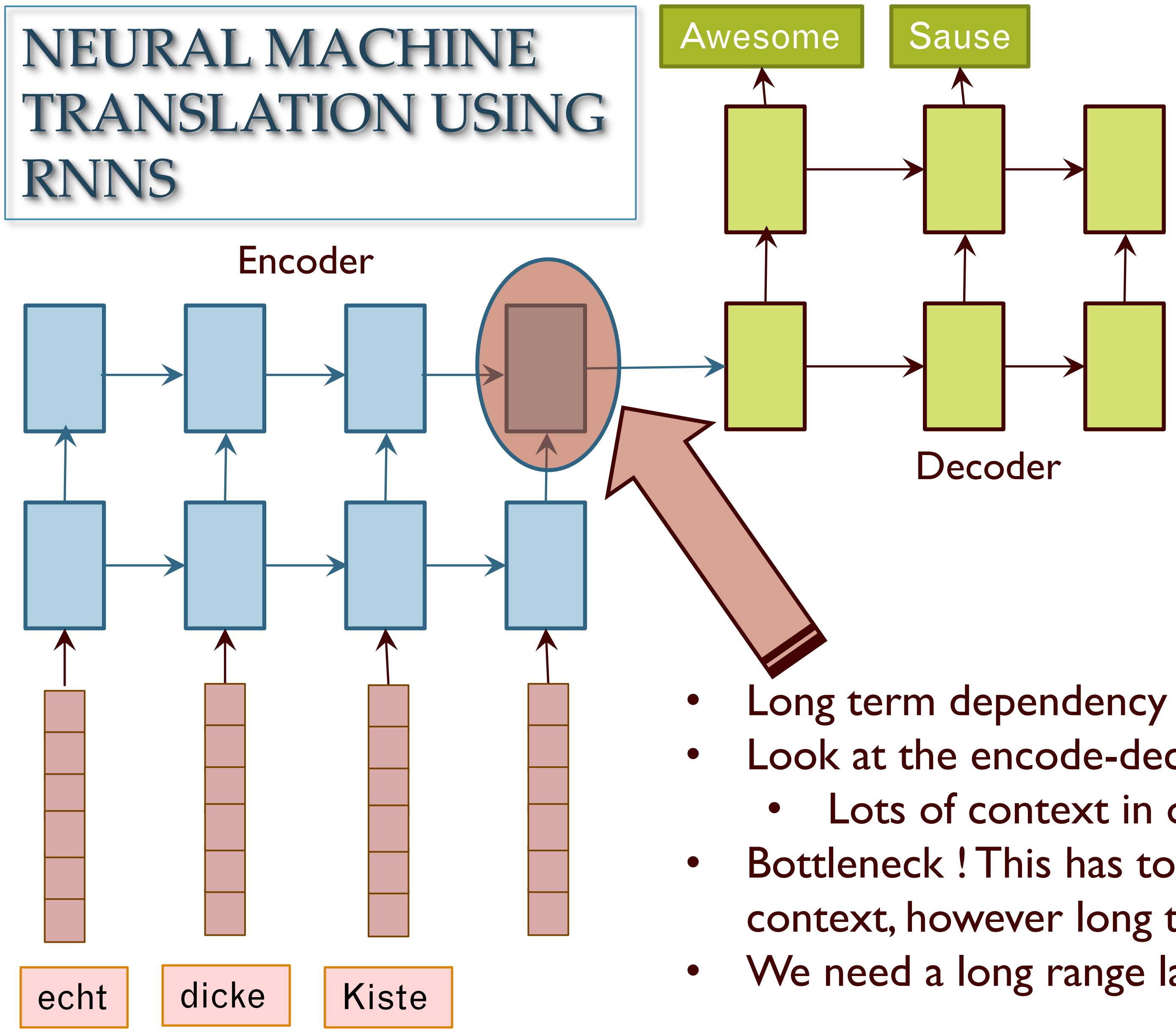
- Vector Similarity

Polarity	Words	Score
Similar	bank (robber) vs bank (vault)	0.95
Similar	robb vs robber	0.70
Different	bank (robber) vs (river) bank	0.40
Different	bank (vault) vs (river) bank	0.40
Similar	fishing vs river	0.43
Different	fishing vs bank (vault)	0.27
Similar	fishing vs (river) bank	0.43

- *bank (robber) vs (river) bank or bank (vault) vs (river) bank are less similar*
- *Interestingly fishing vs river and fishing vs (river) bank have very similar scores*
- *While fishing vs bank (vault) is dissimilar, fishing vs (river) bank has some similarity*

```
0 (101, '[CLS]')
1 (2044, 'after')
2 (26211, 'robb')
3 (2075, '##ing')
4 (1996, 'the')
5 (2924, 'bank')
6 (11632, 'vault')
7 (1010, ',')
8 (1996, 'the')
9 (2924, 'bank')
10 (27307, 'robber')
11 (2001, 'was')
12 (2464, 'seen')
13 (5645, 'fishing')
14 (2006, 'on')
15 (1996, 'the')
16 (5900, 'mississippi')
17 (2314, 'river')
18 (2924, 'bank')
19 (102, '[SEP]')
```

NEURAL MACHINE TRANSLATION USING RNNS



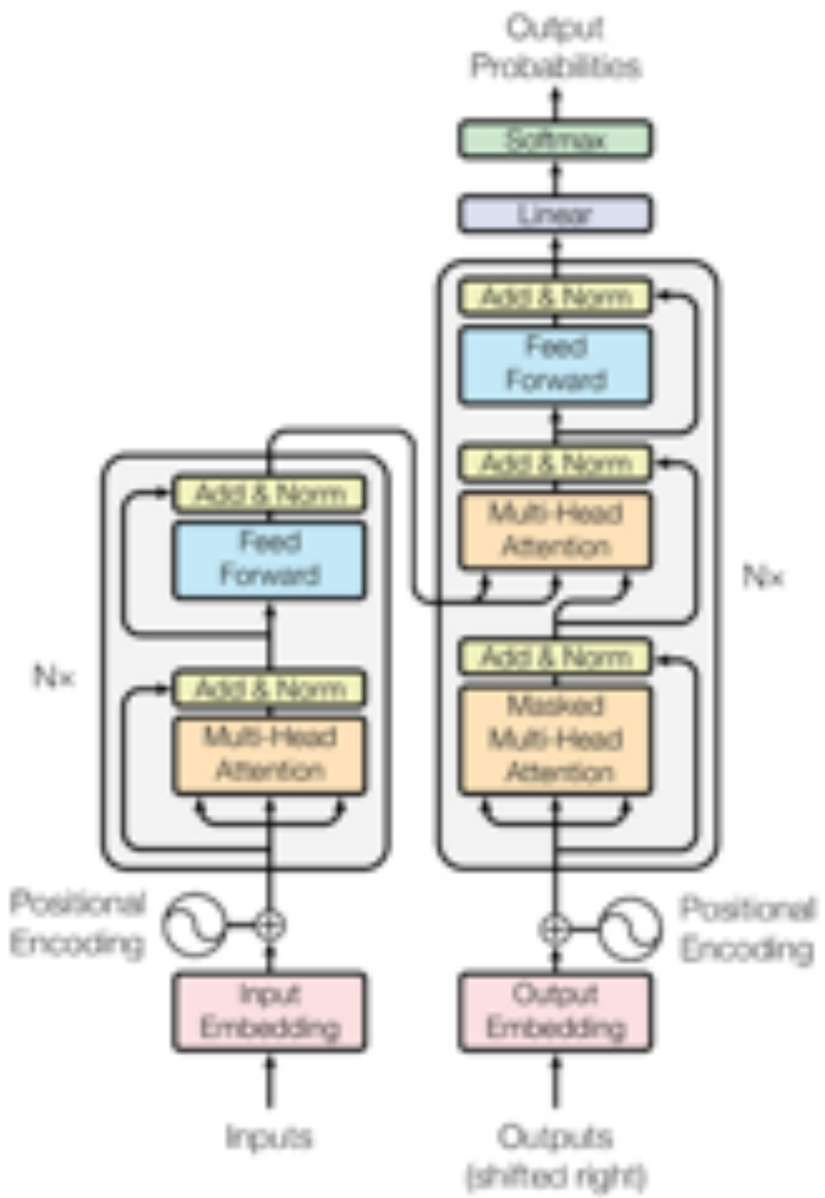
- We haven't solved the bottleneck problem
- The embeddings are still static
- "River <u>Bank</u>" and "<u>Bank</u> Account" still get the same embeddings for Bank

- Long term dependency is a challenge
- Look at the encode-decoder handoff
  - Lots of context in one vector
- Bottleneck ! This has to capture all the context, however long the sentence is !
- We need a long range language model !

"You can't cram the meaning of a whole %&!$# sentence into a single $&!# vector!" - Prof. Raymond J. Mooney/UTA at ACL 2014 workshop*

cs224d:L8 - Stanford

THE SOLUTION IS THE TRANSFORMER ARCHITECTURE



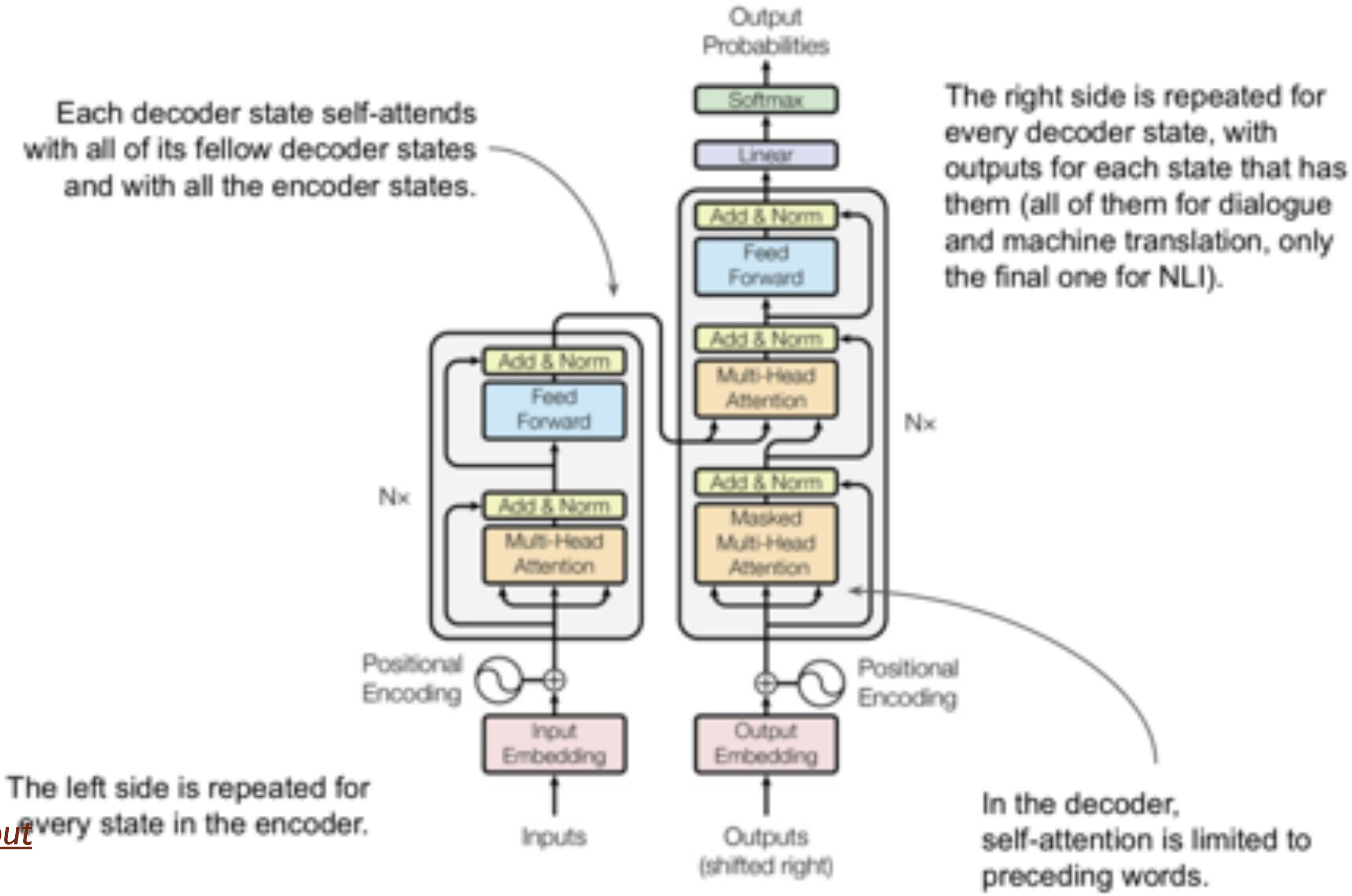
TRANSFORMERS - INTUITION

- Good News
  - Can forget about LSTMs and complex diagrams
- Bad news
  - More complex diagrams !
  - But composable components and orthogonally extensible !
    - That is how BERT was formed
  - Beware – When we go from 6 X 8 X 512 to 72 X 24 X 1024, the *complexities of order of magnitude exist*, capacity of the model need be grappled with like overfitting, can't fit in one GPU, need parallelism, …

- Good News
  - Transformers and BERTology are solving lots of interesting problems very well
  - State Of the Art - as of April '10
- Bad News
  - They are till evolutionary and improvements of current state of Deep Learning technology
  - Mechanisms to inject Knowledge Graphs or Common Sense reasoning – not there yet

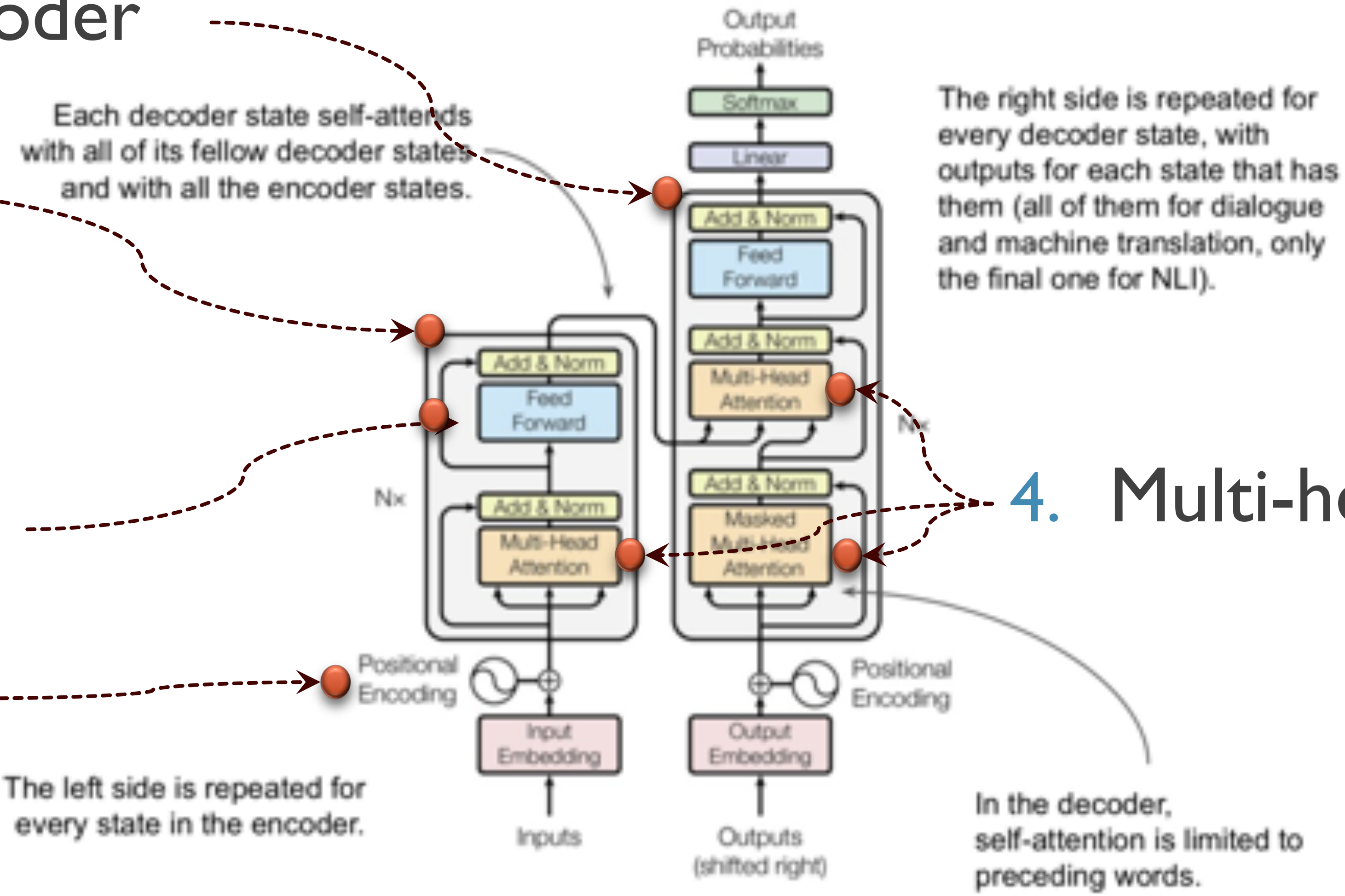
WHAT DOES A TRANSFORMER DO ?

- We will buildup the architecture
  - "*Doesn't feel satisfying, but it works*" category
- Explain what is a head and walk down the architecture and then walk up the architecture
  - Different heads learning different task, randomly, that is relevant downstream
- Abstract & Bypass K,V,Q - just show the connections
- A typical Transformer network architecture has 340M parameters
  - *The large parameter set itself not interesting, but how you lay them out and construct an architecture is* - otherwise they will overfit



WHAT DOES A TRANSFORMER DO ? ESSENTIAL -- TOP 6

6. Decoder
5. Encoder
3. Self-Attention
2. Residual Connections
1. Positional Encoding
4. Multi-head

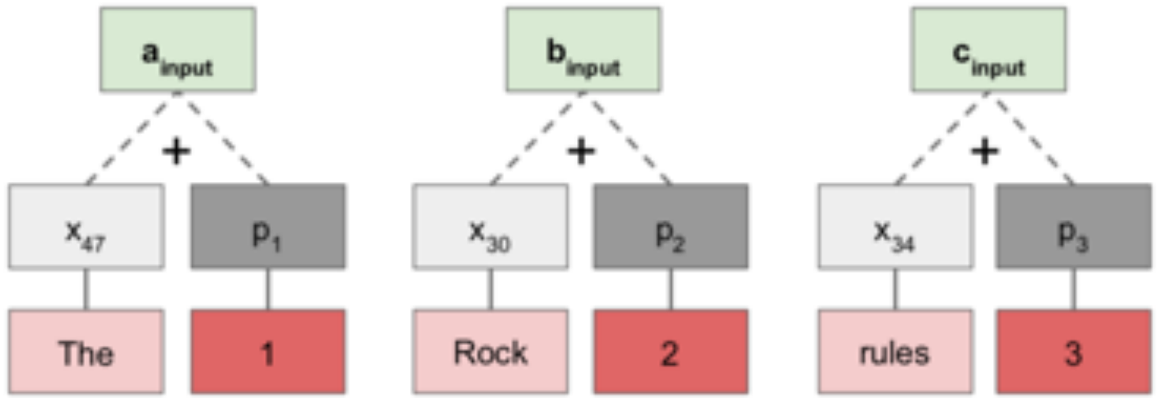


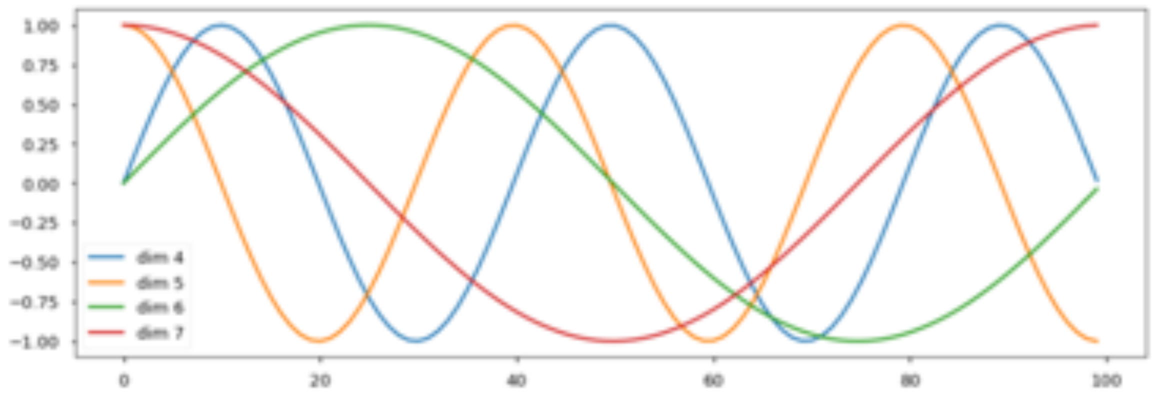
TIPS AND TRICKS OF THE TRANSFORMER

- d_{model} = 512 (1024 for GPT architecture) [Model Dimensionality]
- Byte-Pair Encodings (BPE)
- Checkpoint averaging
- ADAM optimizer with learning rate changes
  - We used the Adam optimizer with $\beta = 0.9$, $\beta_2 = 0.98$ and $\varepsilon = 10^{-9}$. We varied the learning rate over the course of training, according to the formula: This corresponds to increasing the learning rate linearly for the first warmup_steps training steps and decreasing it thereafter proportionally to the inverse square root of the step number. We used warmup_steps=4000.
- Dropout during training at every layer just before adding residual
- Label smoothing (?)
  - During training, we employed label smoothing of value $\varepsilon = 0.1$. This hurts perplexity, as the model learns to be more unsure, but improves accuracy and BLEU score.
- Auto-regressive decoding with beam search and length penalties

POSITIONAL ENCODING

- Since our model contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, we must inject some information about the relative or absolute position of the tokens in the sequence.
- In this work, *they use sine and cosine functions of different frequencies*





From 'The Annotated Transformer'

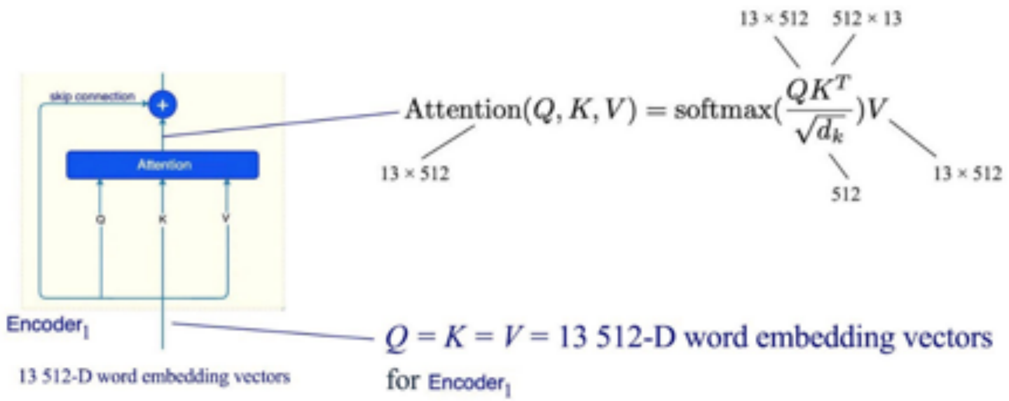
SELF-ATTENTION

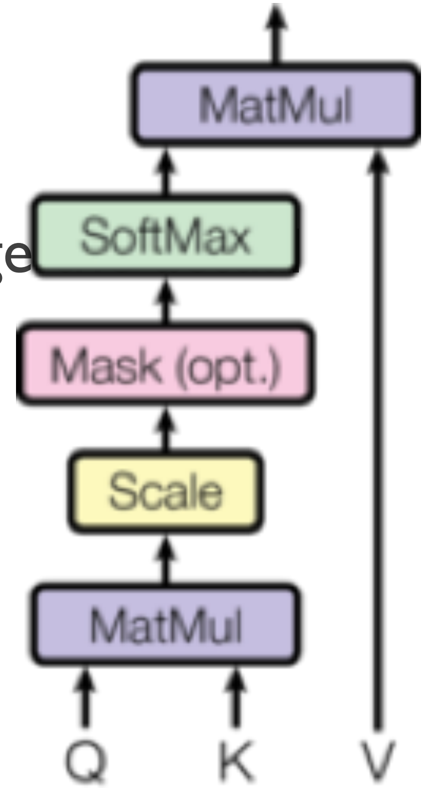
- The *number of operations* required to *relate signals from two arbitrary input or output positions* grows in the distance between positions, *linearly for CNNs* and *logarithmically (for ByteNet) wavenet* for asr and bytenet for NMT, same thing. This makes it more difficult to learn dependencies between distant positions.
- In the *Transformer* this is reduced to a *constant number of operations*, albeit at the cost of reduced effective resolution due to averaging attention-weighted positions, an effect we counteract with Multi-Head Attention.
  - *Constant path between any two positions*
- Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence
- Overall architecture
  - stacked self-attention and point-wise, fully connected layers for both the encoder and decoder
  - 512 size
  - Masked attention – only attend things before you, otherwise it will just copy, with no inference
  - Query get the key that is very similar and then get the values
- Considers the fact that the *order of words is not arbitrary*

Ref: The Annotated Transformer

SELF-ATTENTION

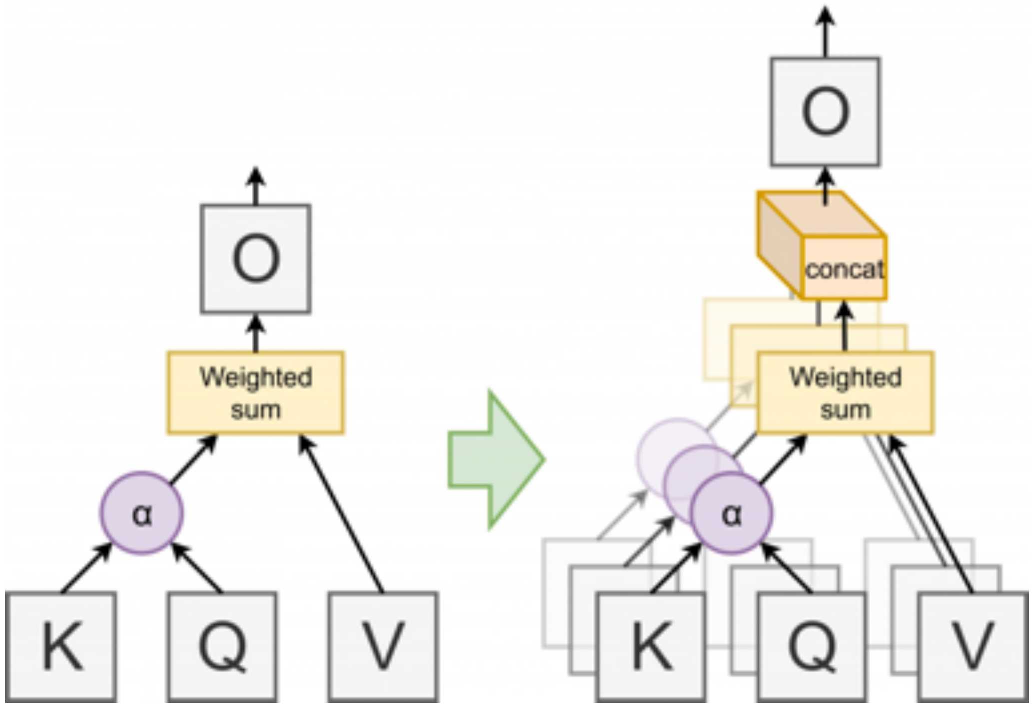
- We call our particular attention “Scaled Dot-Product Attention”. The input consists of queries and keys of dimension d=64, and values of dimension d=64. We compute the dot products of the query with all keys, divide each by squrt(d) and apply a softmax function to obtain the weights on the values
- A position can interact with any position simultaneously
- <u>Attention is permutation invariant</u> i.e. changing order will not change output





MULTI-HEAD

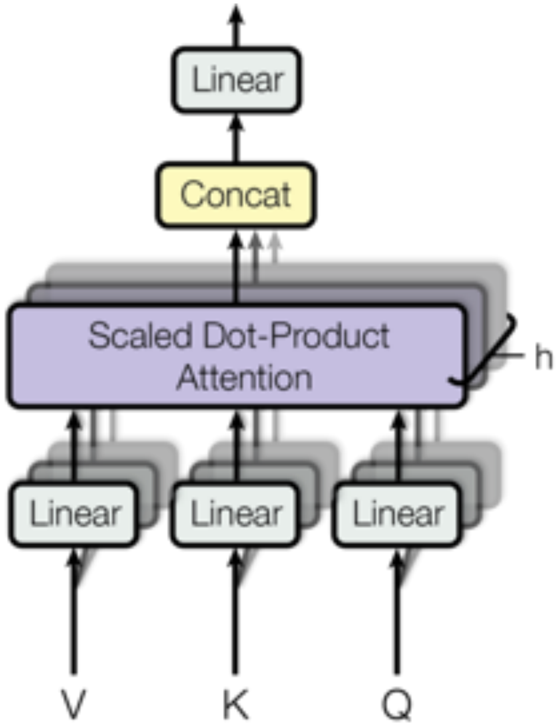
- Multi-head attention allows the model to *jointly attend to information from different representation subspaces at different positions*. With a single attention head, averaging inhibits this.
- Rather than computing single attention (weighted sum of values), the "Multi-Head" Attention computes multiple attention weighted sums, hence the name.
- Essentially, the Multi-Head Attention is just *several attention layers stacked in parallel*, with different linear transformations of the same input.



https://medium.com/@adityathiruvengadam/transformer-architecture-attention-is-all-you-need-aeccd9f50d09
https://medium.com/@jonathan_hui/nlp-bert-transformer-7f0ac397f524
https://blog.ml.cmu.edu/2020/03/20/are-sixteen-heads-really-better-than-one/

MULTI-HEAD

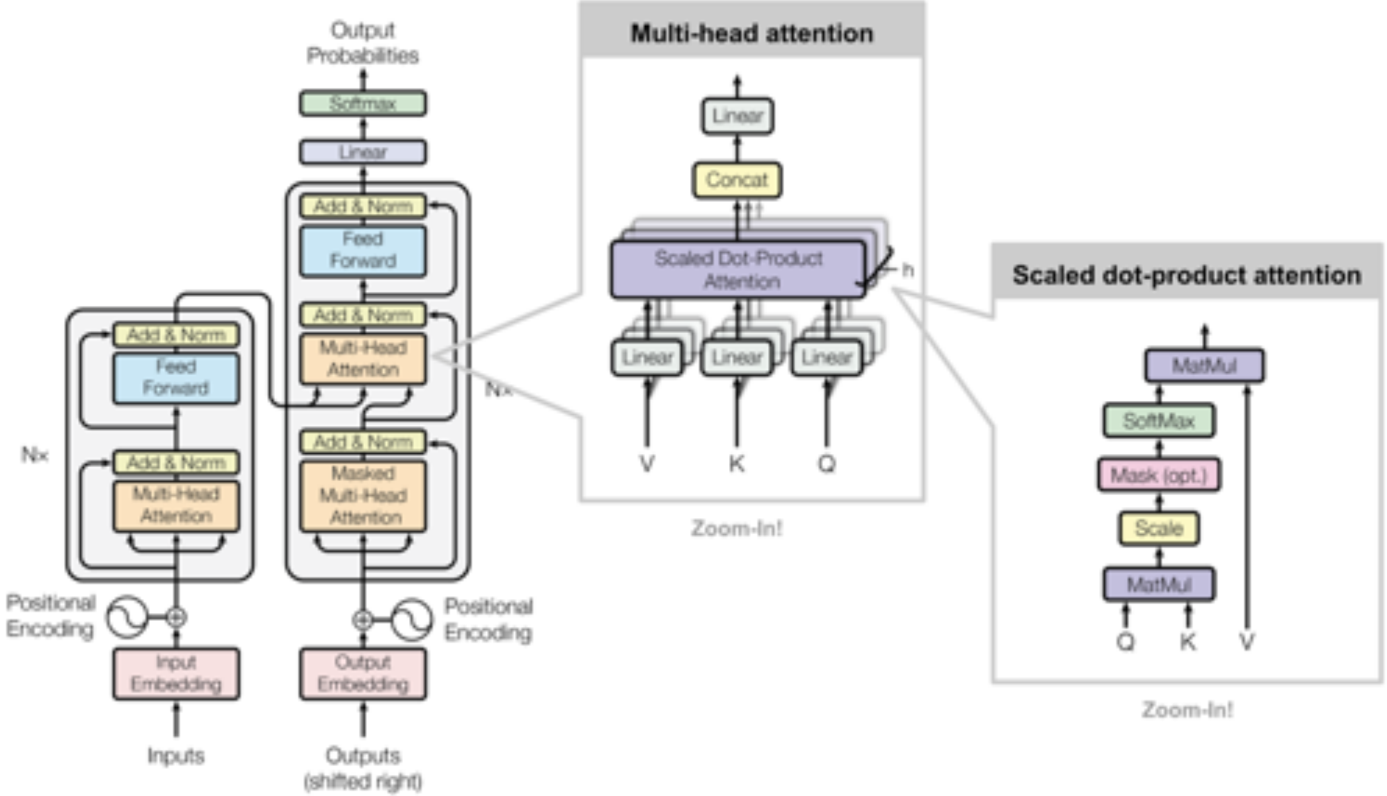
- Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this.
- Rather than computing single attention (weighted sum of values), the "Multi-Head" Attention computes multiple attention weighted sums, hence the name.
- Essentially, the Multi-Head Attention is just several attention layers stacked in parallel, with different linear transformations of the same input.
- **Scaled Dot-Product Attention**
- The first part of each encoder performs the attention. Each word in the sentence serves as a single query. In our example, we have 13 words and therefore 13 queries. But we don't compute attention one-at-a-time for each query.
- Instead, all 13 attentions can be computed concurrently with *Q*, *K*, and *V* pack all the queries, keys and values into matrices. The result will be packed in a 13 × 512 matrix also.

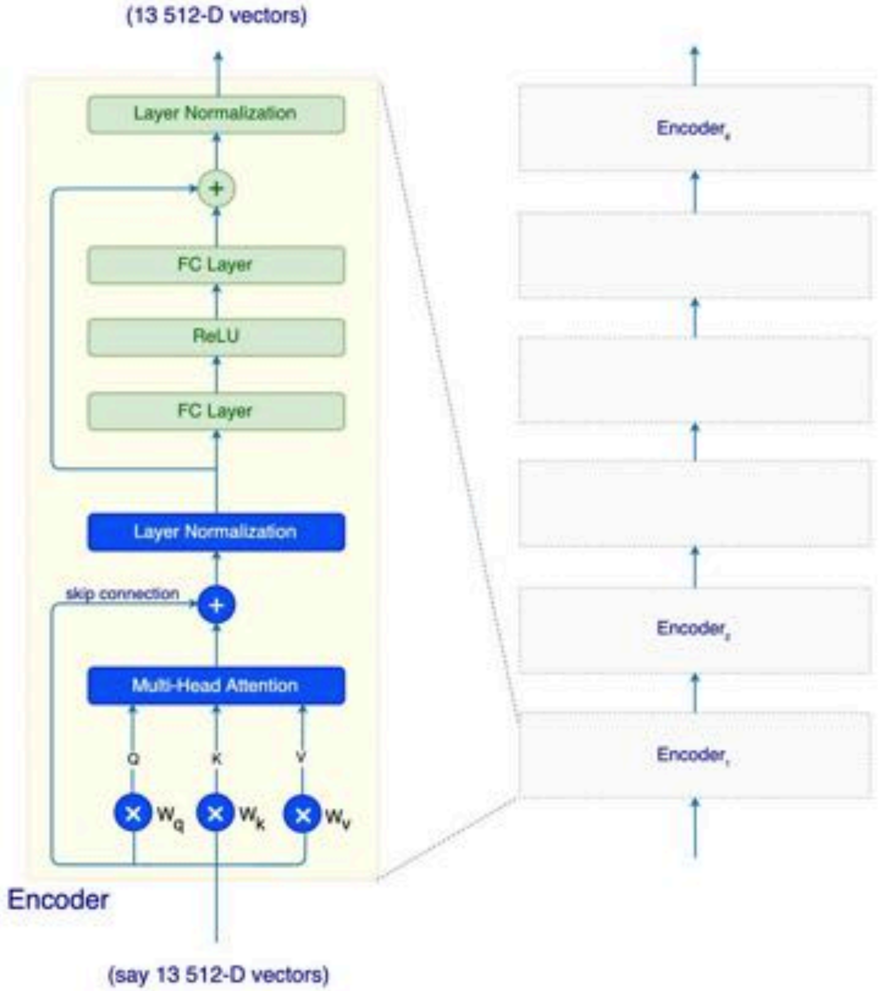


https://medium.com/@adityathiruvengadam/transformer-architecture-attention-is-all-you-need-aeccd9f50d09
https://medium.com/@jonathan_hui/nlp-bert-transformer-7f0ac397f524

MULTI HEAD ATTENTION

- So 8 attentions allow us to view relevancy from 8 different “perspectives”

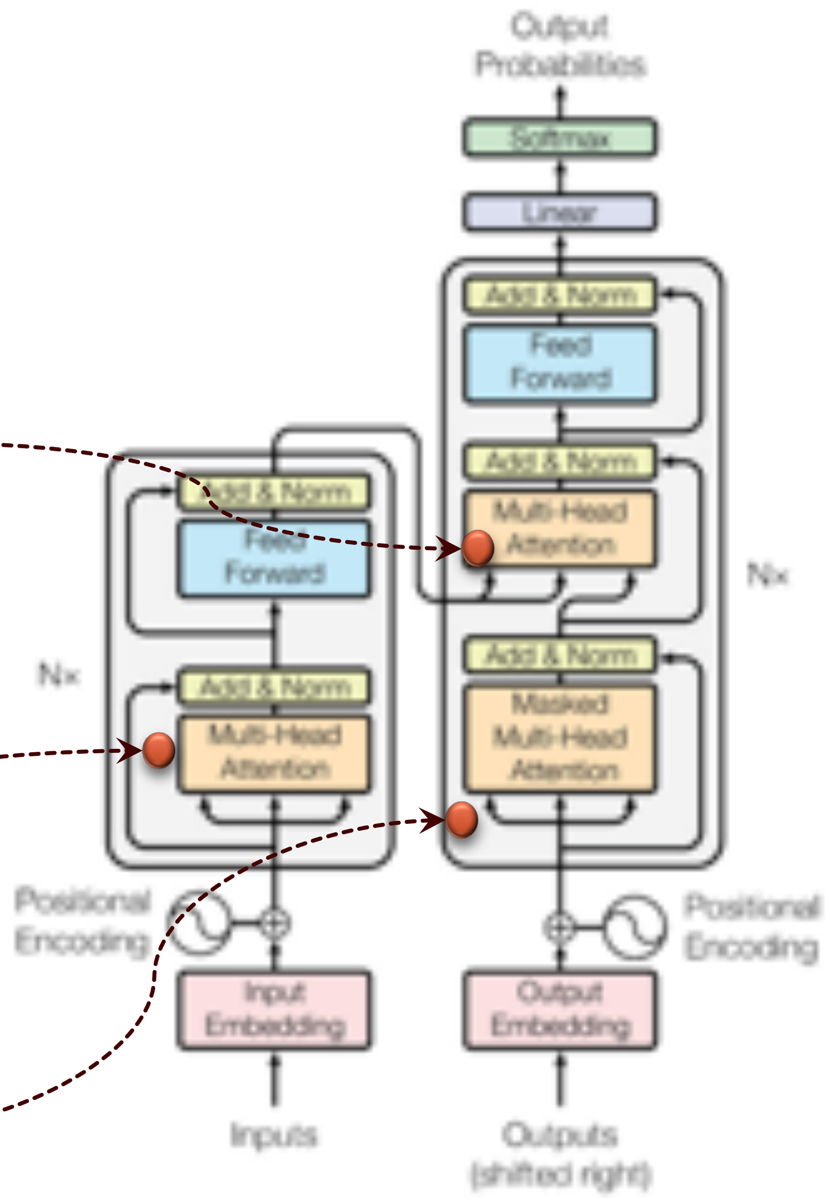




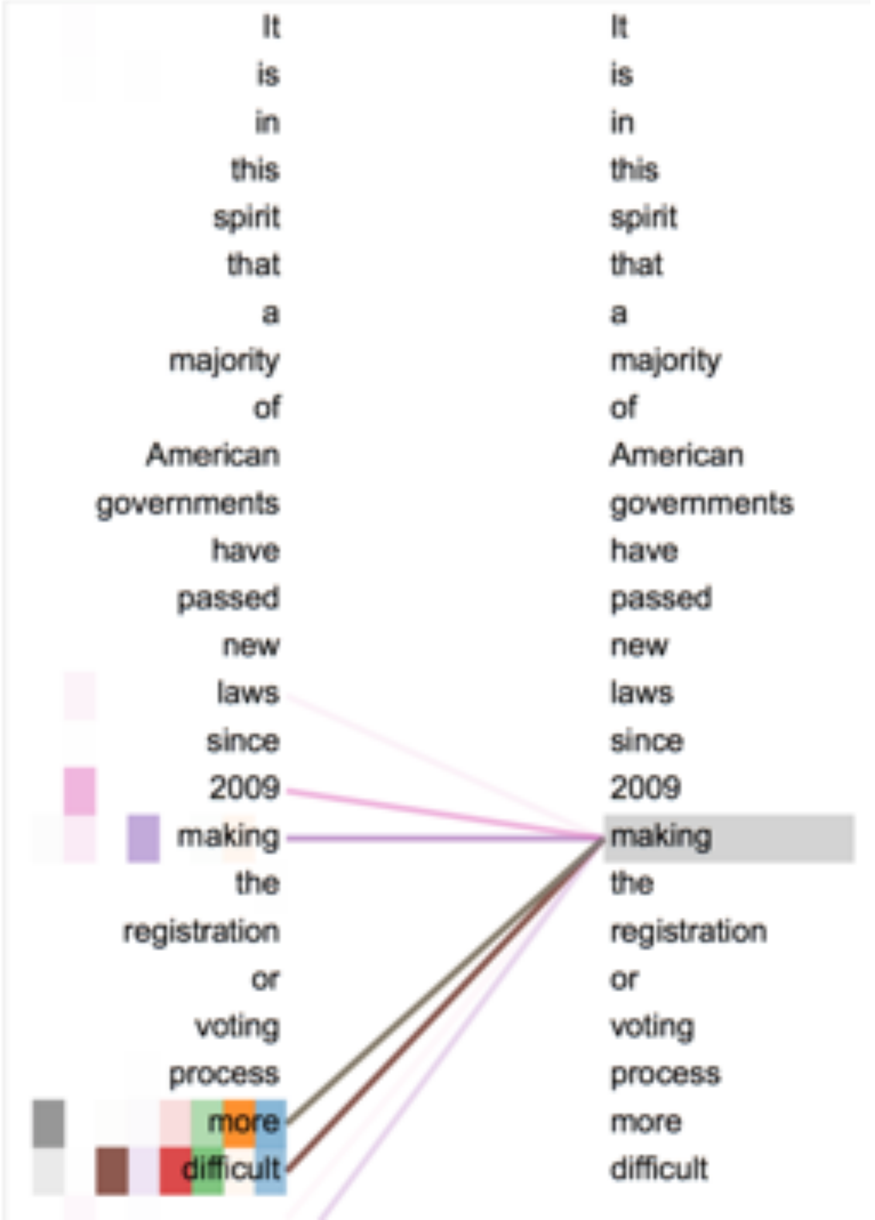
https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html
https://medium.com/@jonathan_hui/nlp-bert-transformer-7f0ac397f524

MULTI-HEAD ATTENTION

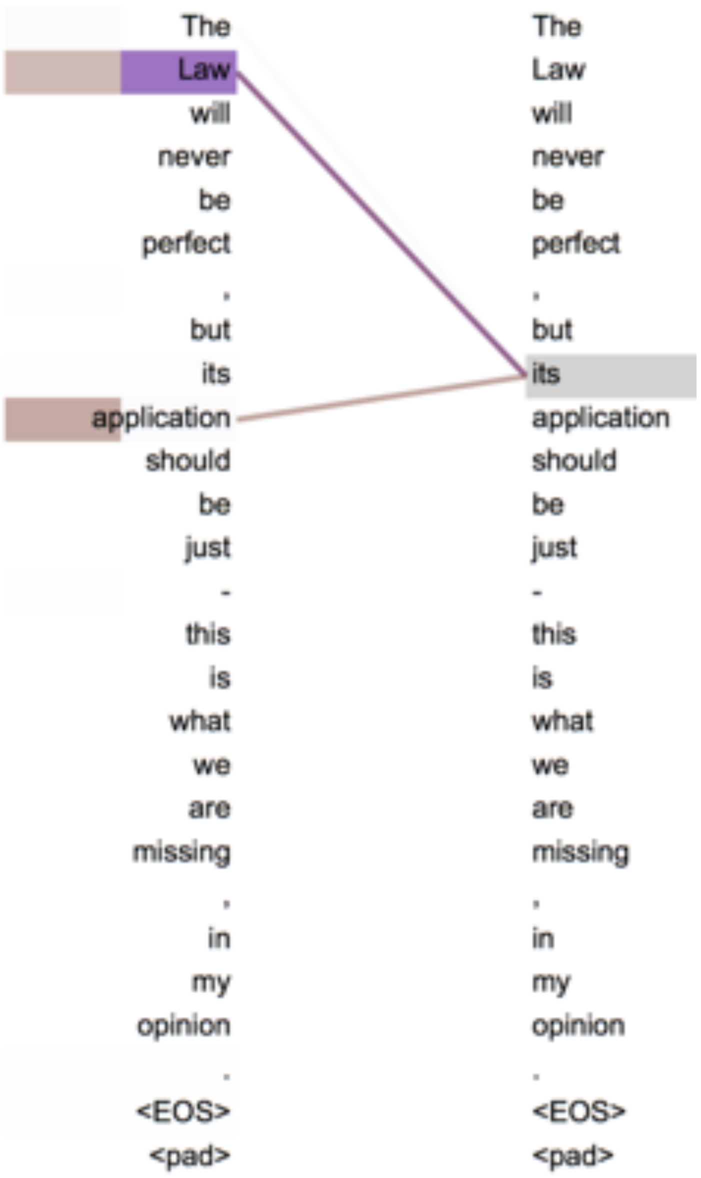
- The Transformer uses multi-head attention in three different ways:
  1. In "encoder-decoder attention" layers, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence. This mimics the typical encoder-decoder attention mechanisms in sequence-to-sequence models such as RNNs.
  2. The encoder contains self-attention layers. In a self-attention layer all of the keys, values and queries come from the same place, in this case, the output of the previous layer in the encoder. Each position in the encoder can attend to all positions in the previous layer of the encoder.
  3. Similarly, self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position. We need to prevent leftward information flow in the decoder to preserve the auto-regressive property. We implement this inside of scaled dot- product attention by masking out (setting to infty) all values in the input of the softmax which correspond to illegal connections.



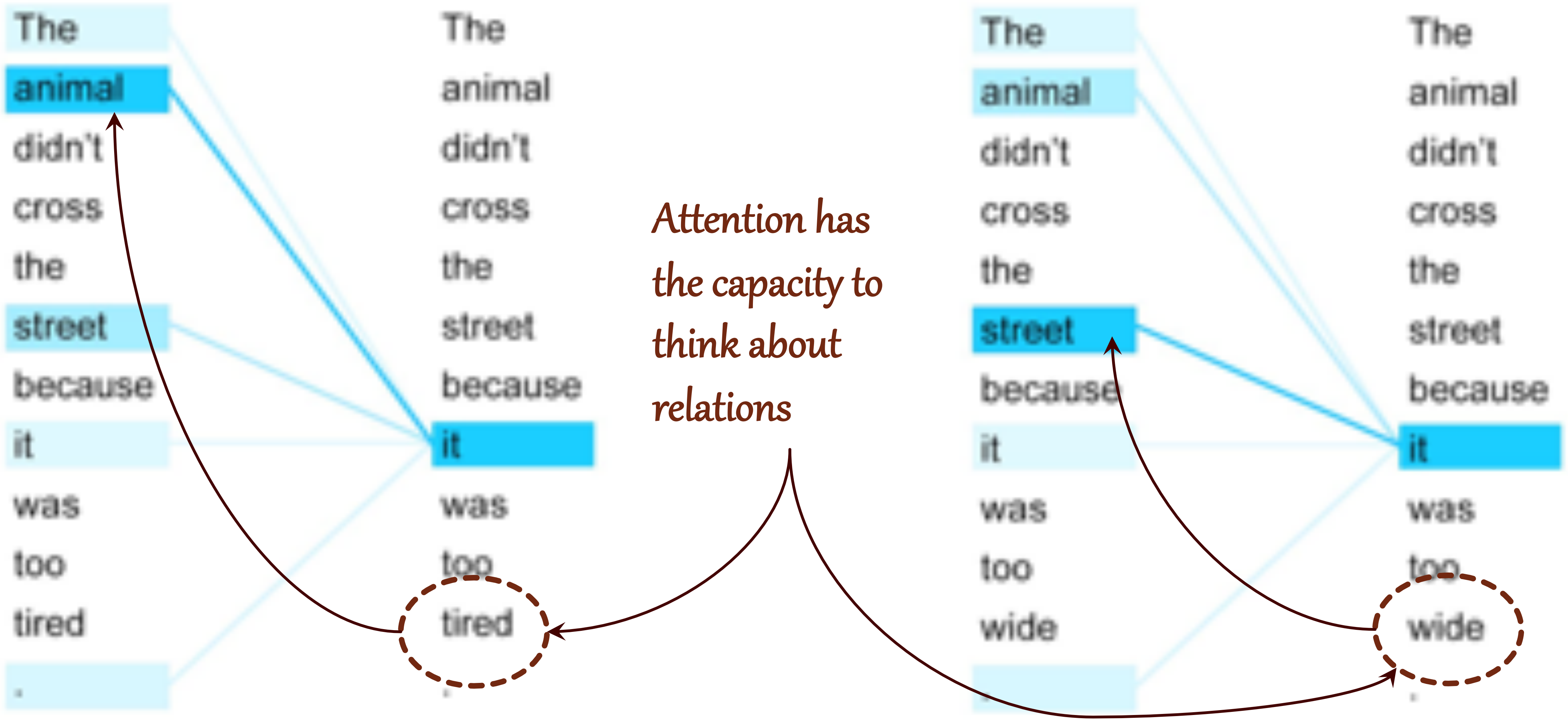
ATTENTION VISUALIZATION



ATTENTION VISUALIZATION

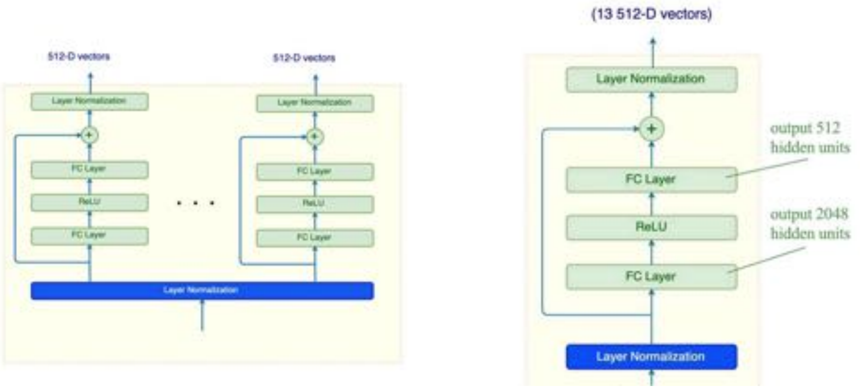


ATTENTION WORKS ON WINOGRAD SCHEMAS !!

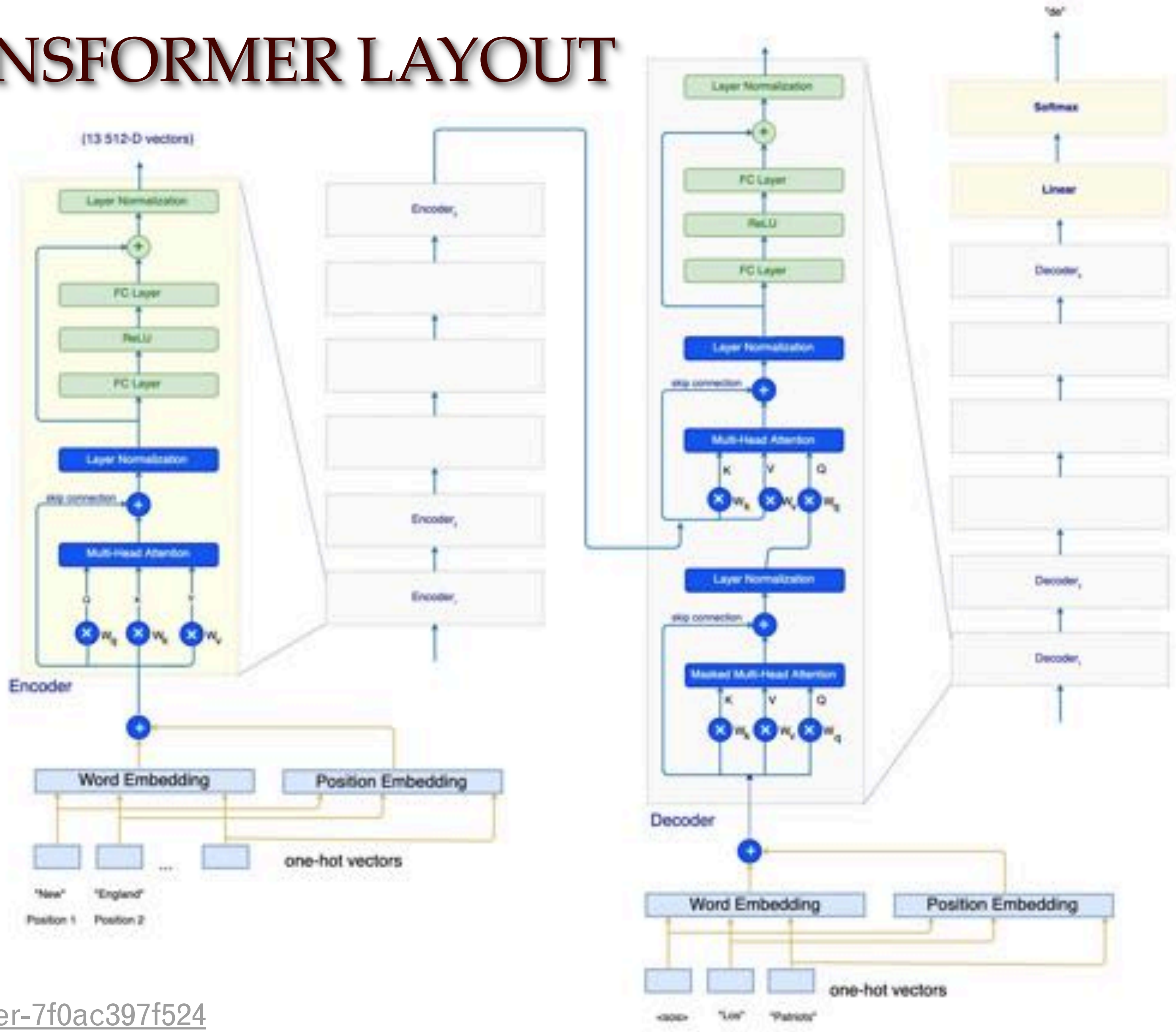


RESIDUAL CONNECTIONS

- Transformer applies skip connection (residual blocks in ResNet) to the output of the multi-head attention followed by a layer normalization.
- Both techniques make training easier and more stable.
  - In batch normalization, we normalize an output dimension based on the corresponding statistics collected from the training batches.



ANOTHER VIEW OF TRANSFORMER LAYOUT



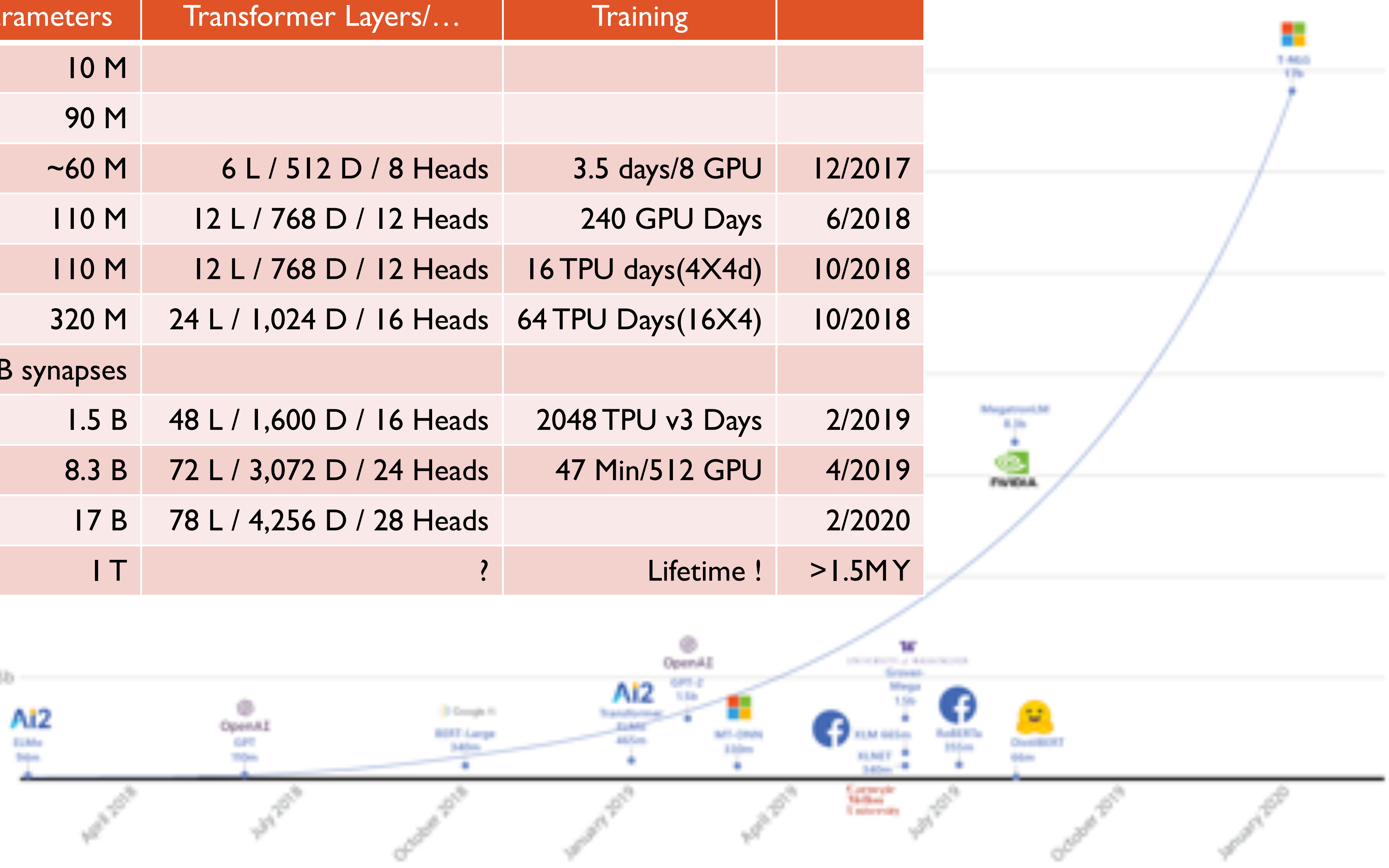
https://medium.com/@jonathan_hui/nlp-bert-transformer-7f0ac397f524

TRAINING

- We trained on the standard WMT 2014 English-German dataset consisting of about *4.5 million sentence pairs*. Sentences were encoded using byte-pair encoding, which has a shared source-target vocabulary of about 37000 tokens. For English- French, we used the significantly larger WMT 2014 English-French dataset consisting of 36M sentences and split tokens into a 32000 word-piece vocabulary.
- Sentence pairs were batched together by approximate sequence length. Each training batch contained a set of sentence pairs containing approximately 25000 source tokens and 25000 target tokens.
- We trained our models on one machine with *8 NVIDIA P100 GPUs*. For our base models using the hyperparameters described throughout the paper, each training step took about 0.4 seconds. We trained the *base models* for a total of 100,000 steps or *12 hours*. For our big models, step time was 1.0 seconds. The *big models* were trained for 300,000 steps (*3.5 days*).

SCALE OF THINGS

Model	Parameters	Transformer Layers/…	Training	
Medium-sized LSTM	10 M			
ELMo	90 M			
Transformers	~60 M	6 L / 512 D / 8 Heads	3.5 days/8 GPU	12/2017
GPT	110 M	12 L / 768 D / 12 Heads	240 GPU Days	6/2018
BERT(Base)	110 M	12 L / 768 D / 12 Heads	16 TPU days(4X4d)	10/2018
BERT (L)	320 M	24 L / 1,024 D / 16 Heads	64 TPU Days(16X4)	10/2018
Honeybee Brain	~1B synapses			
GPT-2 (XL)	1.5 B	48 L / 1,600 D / 16 Heads	2048 TPU v3 Days	2/2019
Megatron GPT-2	8.3 B	72 L / 3,072 D / 24 Heads	47 Min/512 GPU	4/2019
Turing-NLG	17 B	78 L / 4,256 D / 28 Heads		2/2020
Human Brain Models	1 T	?	Lifetime !	>1.5M Y

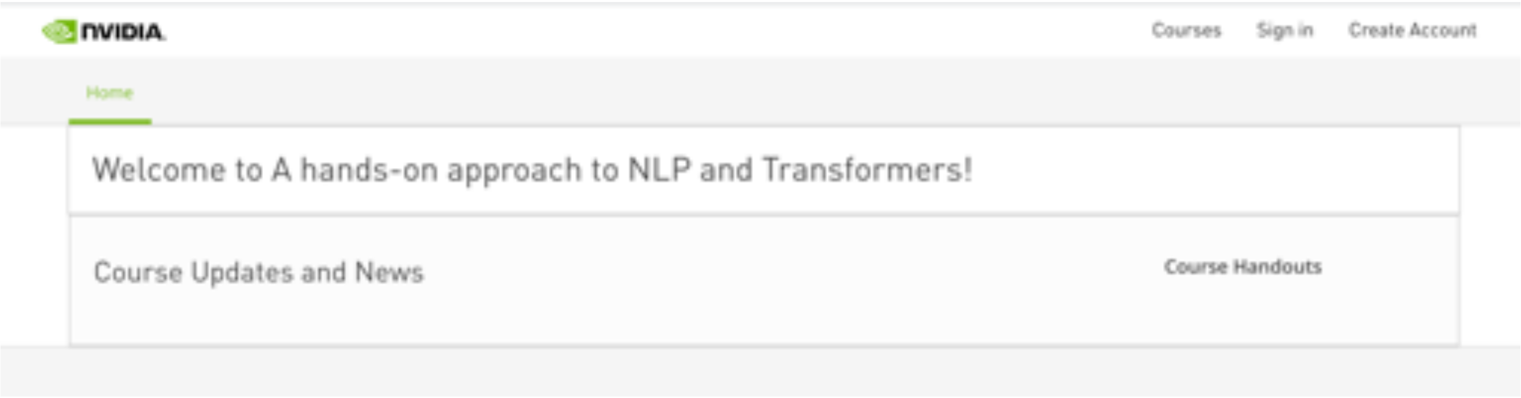


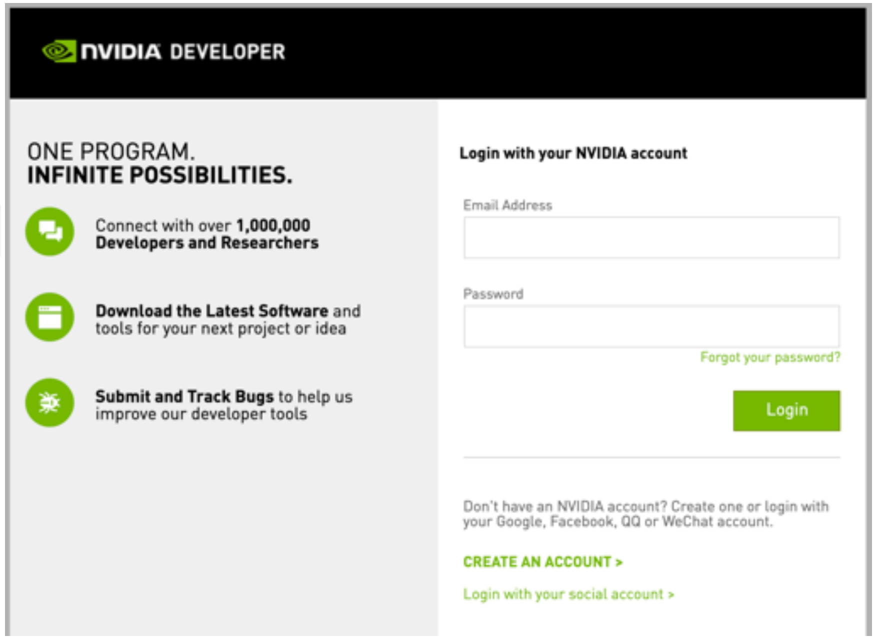
- *GPT-2 : Really big Transformer LM trained on 40GB text*
- *Model > 1.3B won't fit in GPU, so need model parallelization and broken across multiple GPUs*
- *D = Model Dimensionality (d_{model})*

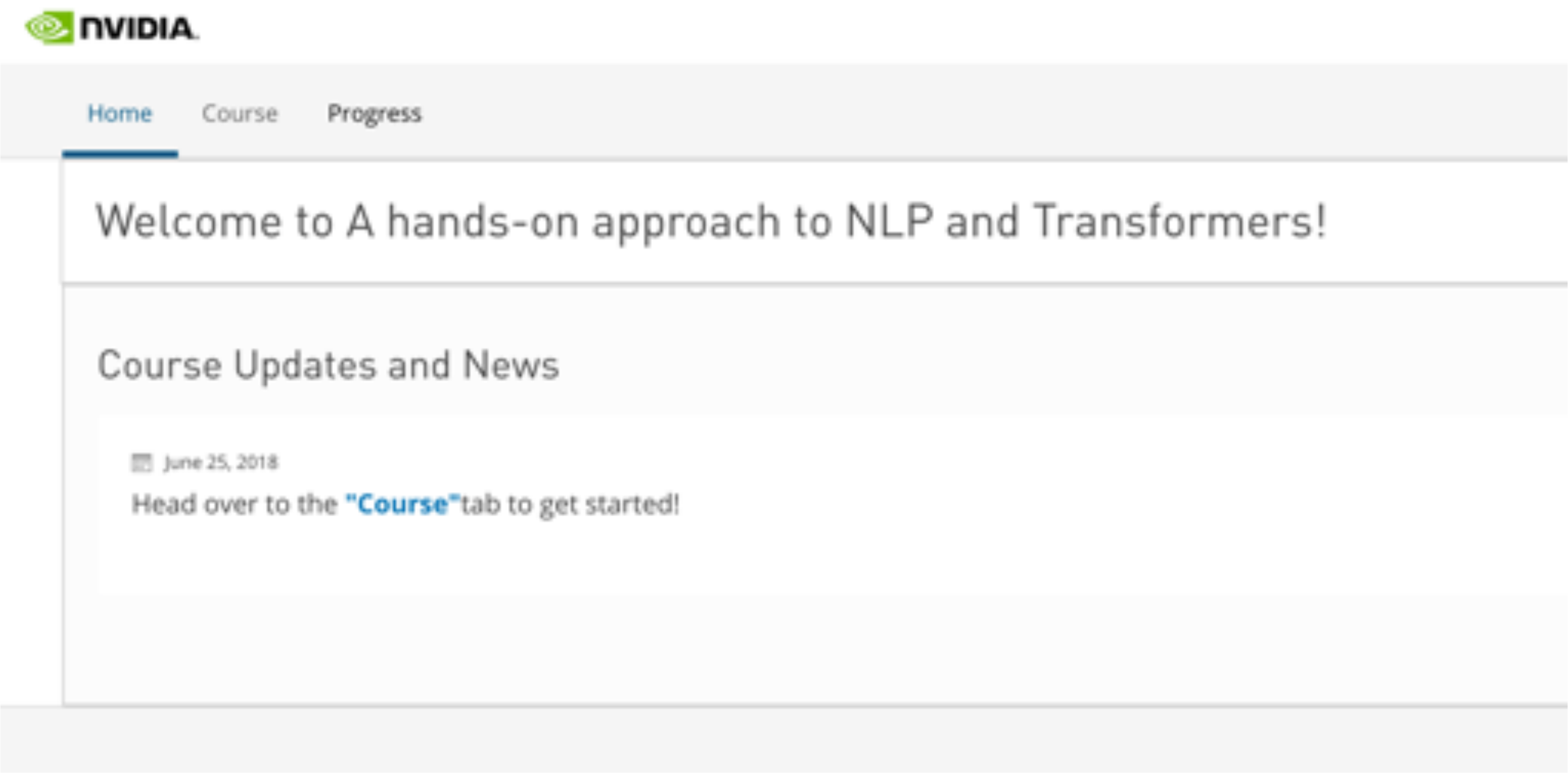
HANDS-ON : EMBEDDING

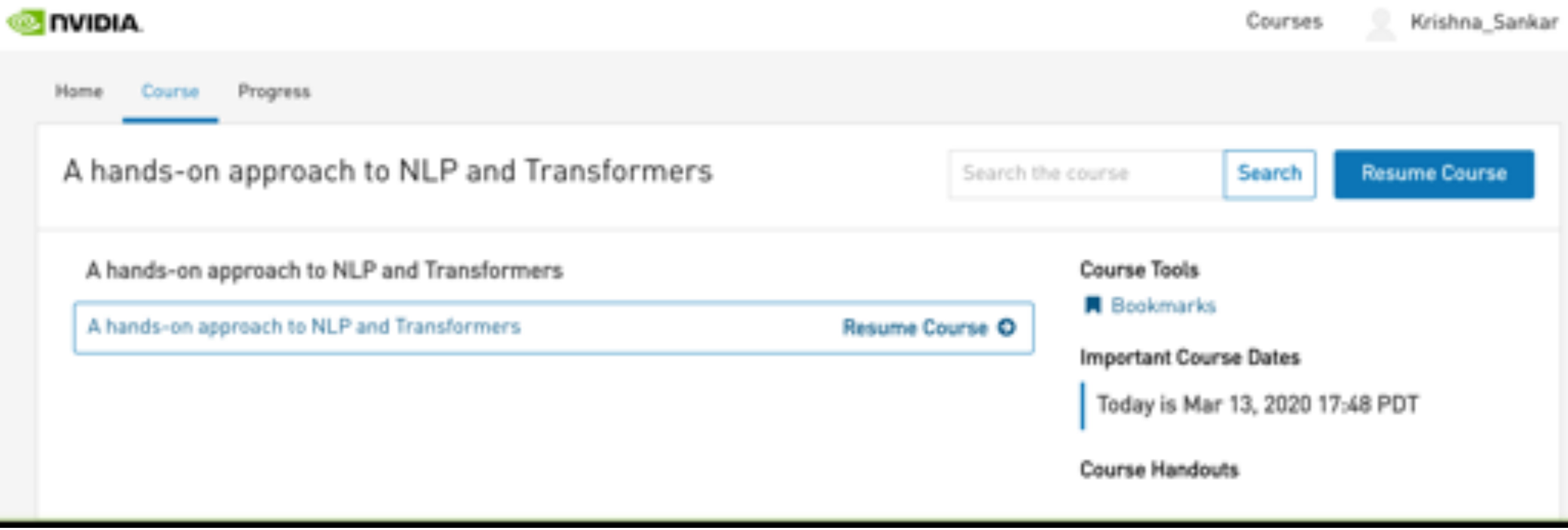
HANDS-ON : STARTING A VM

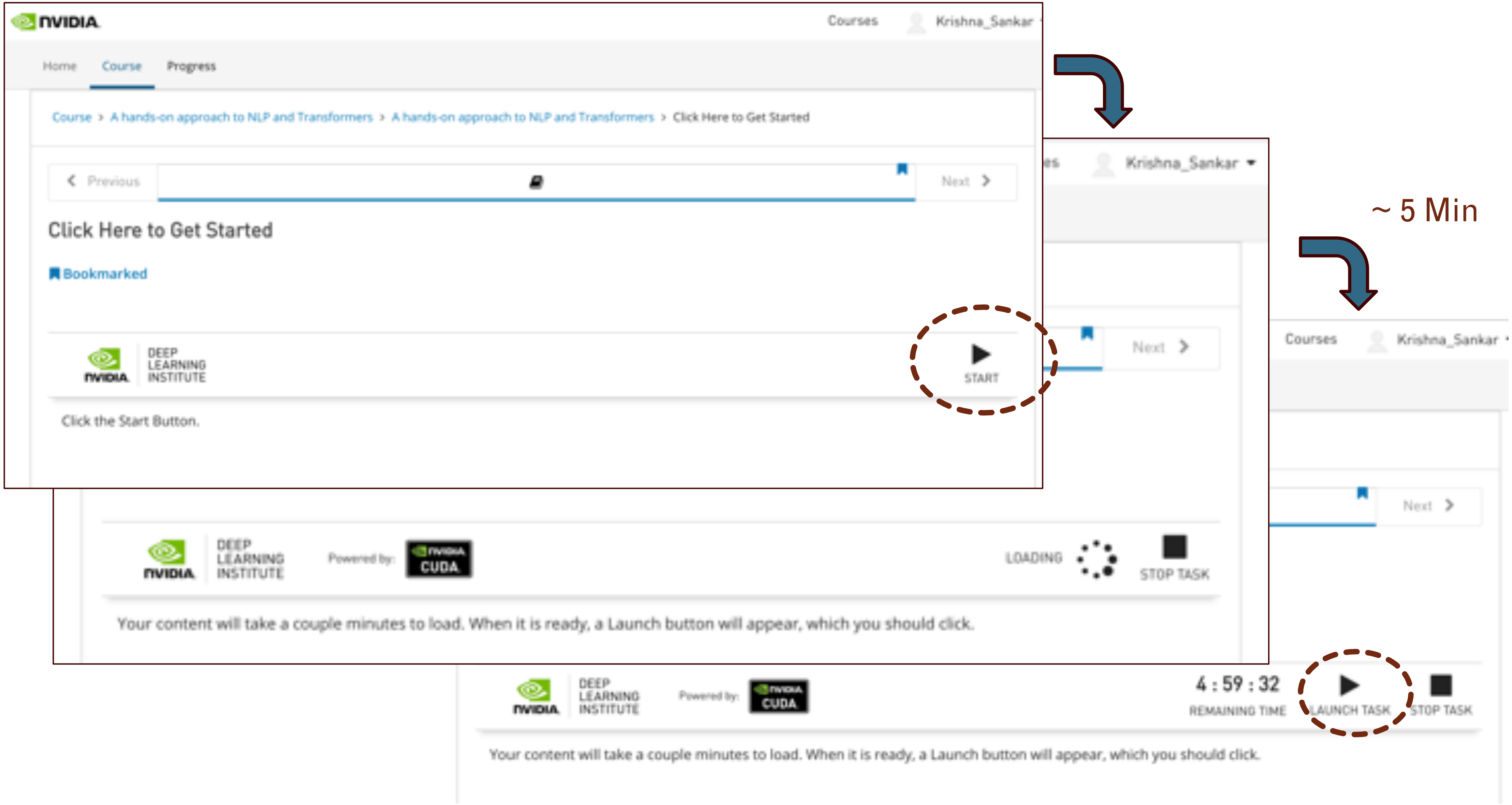
- https://courses.nvidia.com/courses/course-v1:DLI+GTC-2020-T21095+V1

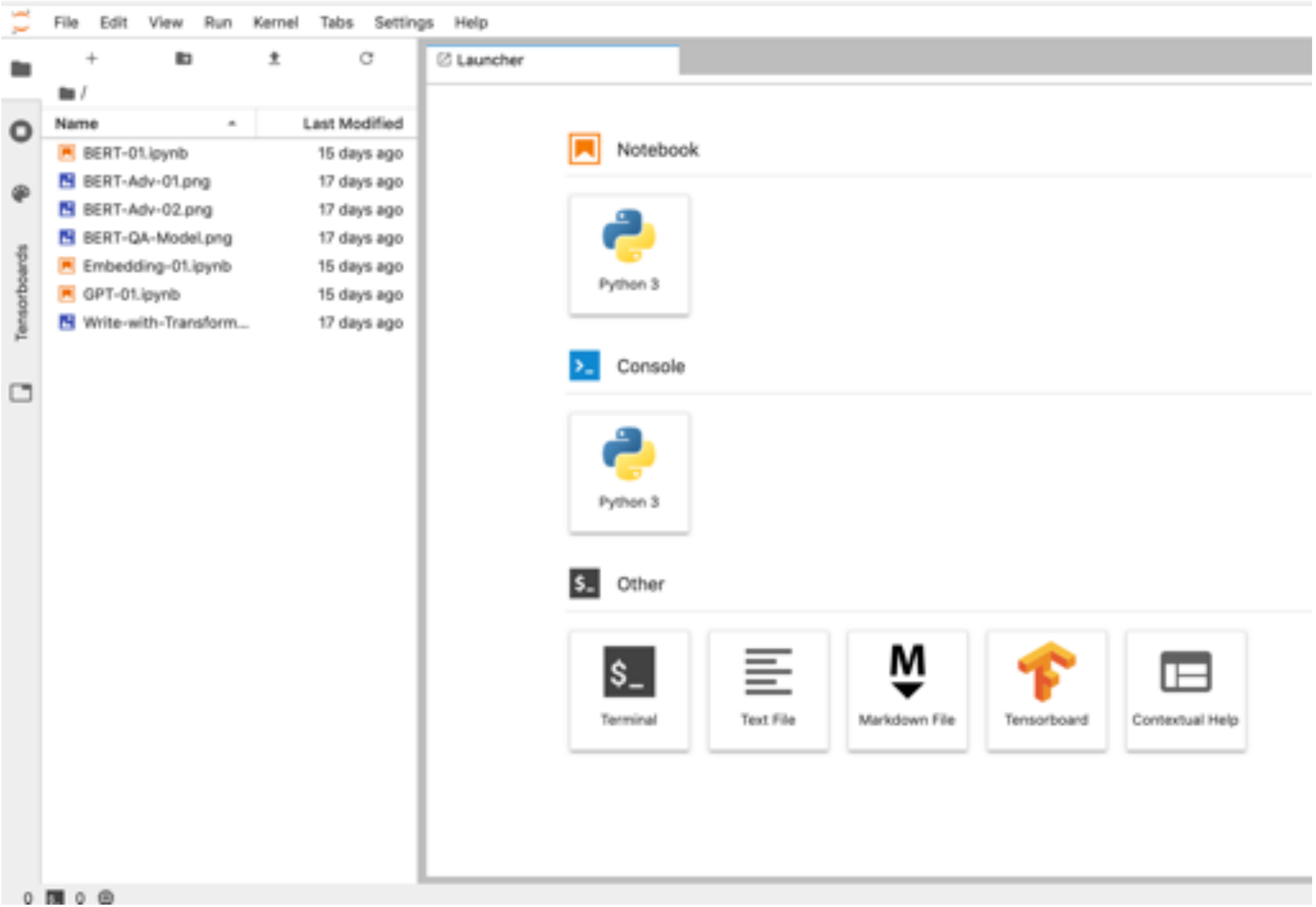




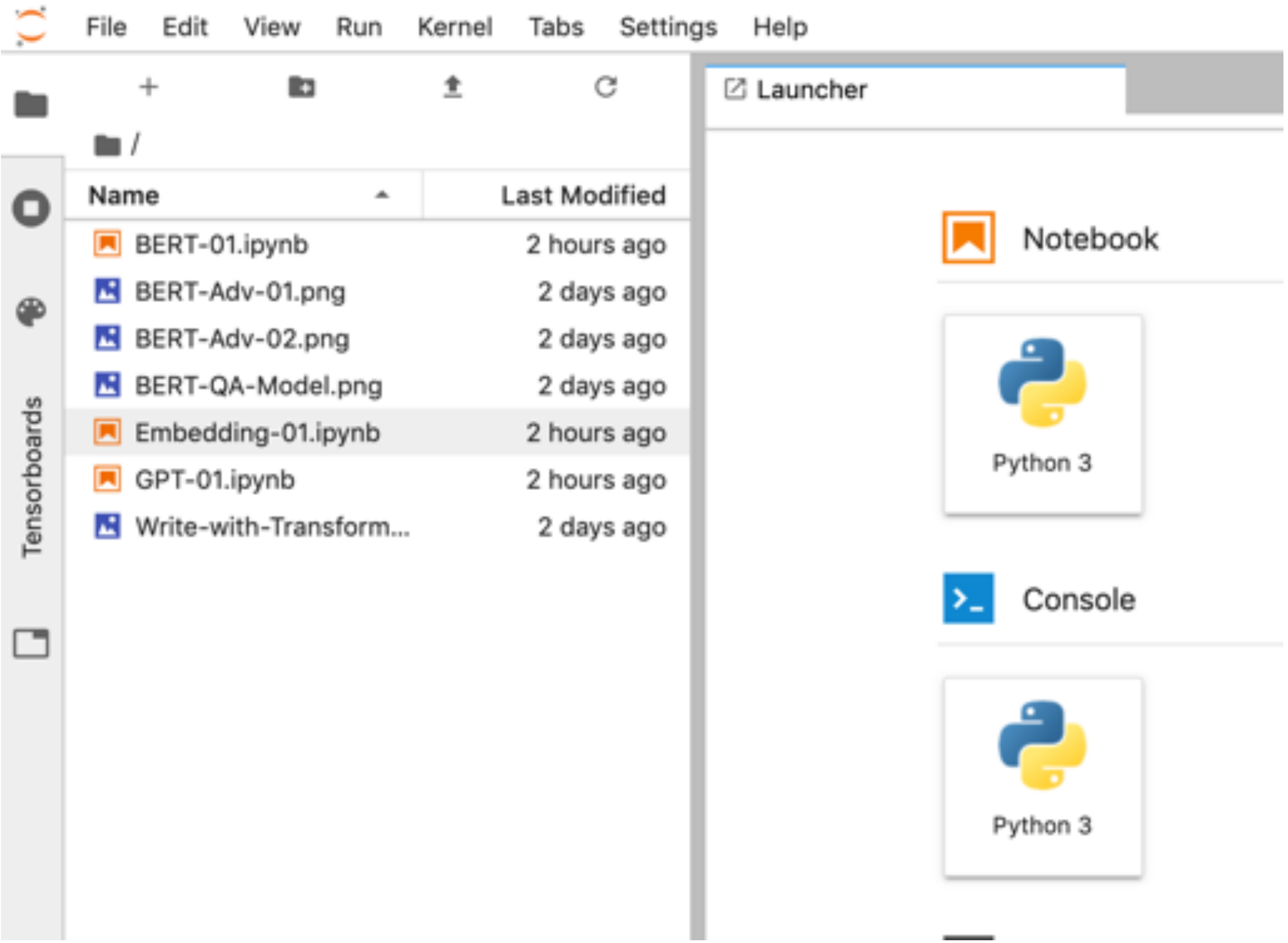

NVIDIA
Home
Course
Progress
Welcome to A hands-on approach to NLP and Transformers!
Course Updates and News
June 25, 2018
Head over to the "Course"tab to get started!


NVIDIA
Courses
Krishna_Sankar
Home
Course
Progress
A hands-on approach to NLP and Transformers
Search the course
Search
Resume Course
A hands-on approach to NLP and Transformers
A hands-on approach to NLP and Transformers
Resume Course
Course Tools
Bookmarks
Important Course Dates
Today is Mar 13, 2020 17:48 PDT
Course Handouts

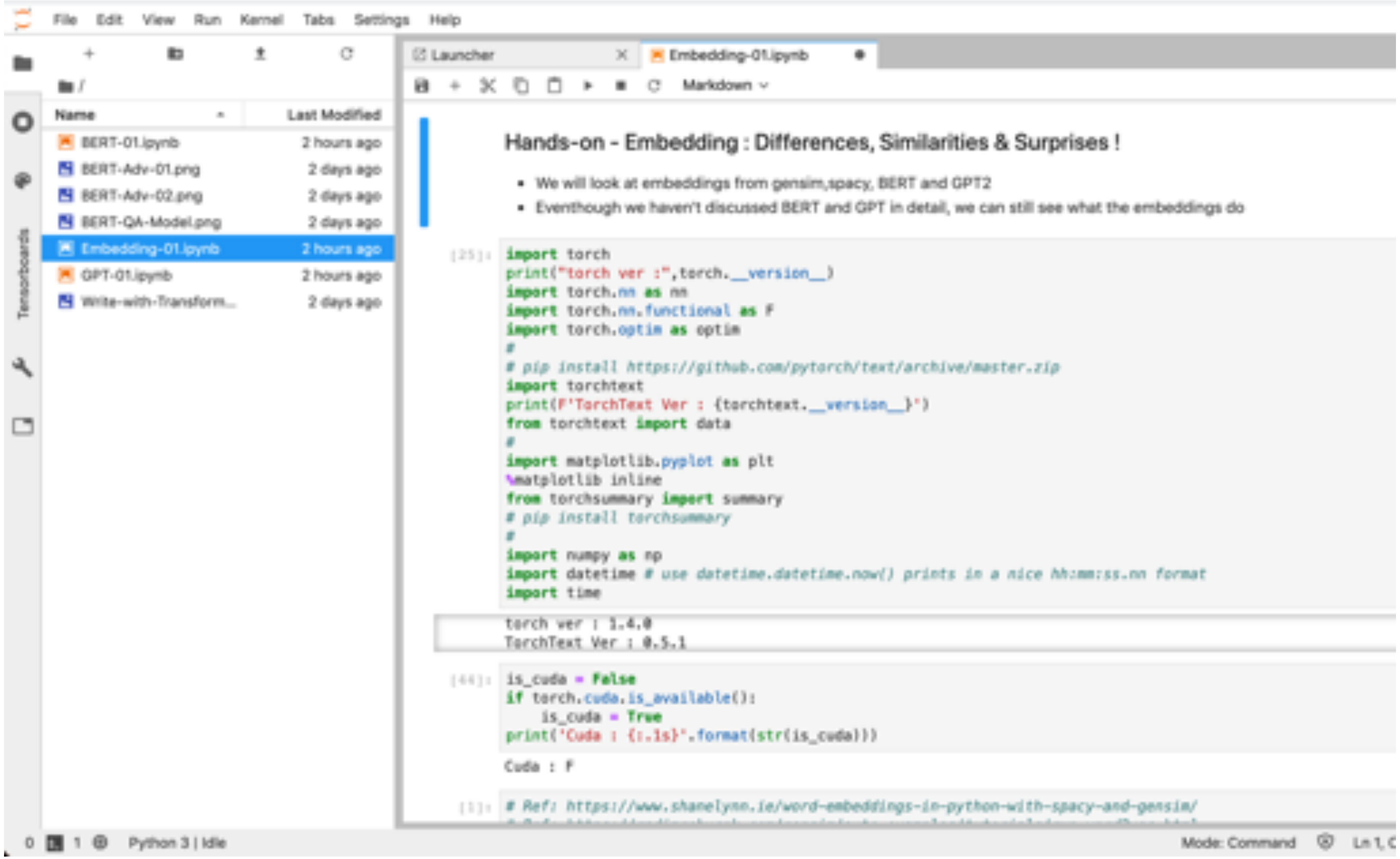

NVIDIA
Courses
Krishna_Sankar
Home
Course
Progress
Course > A hands-on approach to NLP and Transformers > A hands-on approach to NLP and Transformers > Click Here to Get Started
Previous
Next
Click Here to Get Started
Bookmarked
NVIDIA DEEP LEARNING INSTITUTE
START
Click the Start Button.
~ 5 Min
Powered by: NVIDIA CUDA
LOADING
STOP TASK
Your content will take a couple minutes to load. When it is ready, a Launch button will appear, which you should click.
4 : 59 : 32
REMAINING TIME
LAUNCH TASK
STOP TASK
Your content will take a couple minutes to load. When it is ready, a Launch button will appear, which you should click.


File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
Name
Last Modified
BERT-01.ipynb 15 days ago
BERT-Adv-01.png 17 days ago
BERT-Adv-02.png 17 days ago
BERT-QA-Model.png 17 days ago
Embedding-01.ipynb 15 days ago
GPT-01.ipynb 15 days ago
Write-with-Transform... 17 days ago
Tensorboards
Notebook
Python 3
Console
Python 3
Other
Terminal
Text File
Markdown File
Tensorboard
Contextual Help

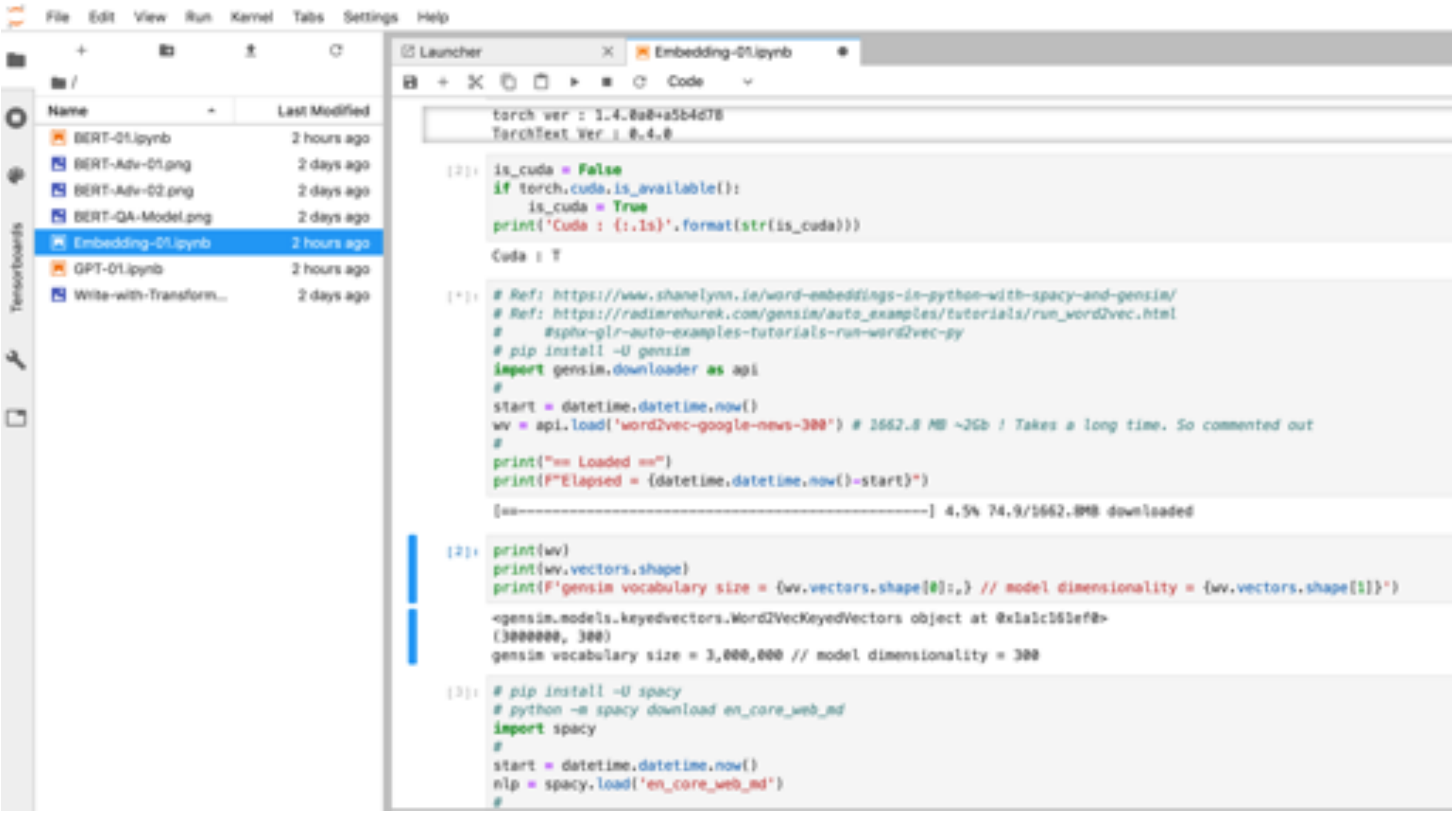
STEP 1 : LOAD THE EMBEDDING-01.IPYNB



STEP 2 – RUN THE CELLS AND OBSERVE THE RESULTS



- Gensym will take ~5 min to load



SOME OBSERVATIONS

```
print(wv)
print(wv.vectors.shape)
print(F'gensim vocabulary size = {wv.vectors.shape[0]:,} // model dim
```

```
<gensim.models.keyedvectors.Word2VecKeyedVectors object at 0x7feba419
(3000000, 300)
gensim vocabulary size = 3,000,000 // model dimensionality = 300
```

```
print(F'spacy vocabulary size = {nlp.vocab.length:,} // model dimensionality = {nlp.vocab.vectors_length}')
```

```
spacy vocabulary size = 1,340,242 // model dimensionality = 300
```

Let us do some semantic computations using the word vectors

```
print(wv.doesnt_match(['fire', 'water', 'land', 'sea', 'air', 'car']))
```

```
car
```

SEMANTIC COMPUTATIONS ON WORD VECTORS

```
# Common king - man + woman
wv.most_similar(positive=['woman', 'king'], negative=['man'])
# Library - Books = Hall
# Obama + Russia - USA = Putin
# Human - Animal = Ethics
# Ref ; http://byterot.blogspot.com/2015/06/five-crazy-abstraction
```

```
[('queen', 0.7118192911148071),
 ('monarch', 0.6189674139022827),
 ('princess', 0.5902431607246399),
 ('crown_prince', 0.5499460697174072),
 ('prince', 0.5377321243286133),
 ('kings', 0.5236844420433044),
 ('Queen_Consort', 0.5235945582389832),
 ('queens', 0.5181134343147278),
 ('sultan', 0.5098593235015869),
 ('monarchy', 0.5087411999702454)]
```

SEMANTIC COMPUTATIONS ON WORD VECTORS

```
]: wv.most_similar(positive=['liquid', 'ice'], negative=['Water']) # water:ice :: liquid: ?

]: [('unmelted', 0.509214460849762),
    ('Methane_hydrate', 0.46418970823287964),
    ('Francies_tossed', 0.45730510354042053),
    ('ice_crystals', 0.45635735988616943),
    ('starch_granules', 0.44293591380119324),
    ('Fill_cocktail_shaker', 0.43867558240890503),
    ('graphene_sheet', 0.4353669285774231),
    ('jellylike', 0.4344847798347473),
    ('ice_cubes', 0.43398386240005493),
    ('caked_oak_tree', 0.4337309002876282)]
```

SEMANTIC COMPUTATIONS ON WORD VECTORS

```
]: wv.most_similar(positive=['Japan','Pizza'], negative=['USA']) # USA : Pizza :: Japan : ?

]: [('Sushi', 0.5657287836074829),
    ('yakiniku', 0.5292825102806091),
    ('Teriyaki', 0.5238471031188965),
    ('Pizzeria', 0.5128323435783386),
    ('Steak_House', 0.5030418634414673),
    ('Deli', 0.4990933835506439),
    ('Grill', 0.4931894242763519З),
    ('sushi', 0.4874907433986664),
    ('Yakitori', 0.48361146450042725),
    ('conveyor_belt_sushi', 0.47991597652435З)]
```

SEMANTIC COMPUTATIONS ON WORD VECTORS

Try word distances

[15]:
```
# To try
w1 = "happy"
w2 = "cheerful"
w3 = "sad"
w1_w2_dist = wv.distance(w1, w2)
w1_w3_dist = wv.distance(w1, w3)
print("Synonyms {}, {} have cosine distance: {}".format(w1, w2, w1_w2_dist))
print("Antonyms {}, {} have cosine distance: {}".format(w1, w3, w1_w3_dist))
```

```
Synonyms happy, cheerful have cosine distance: 0.6162261664867401
Antonyms happy, sad have cosine distance: 0.46453857421875
```

- This is an interesting side effect.
- Most probably "happy" and "sad" occur (in the corpus) near each other in sentences, more than "happy" and "cheerful".
- Also, because happy and cheerful are very close, probably most of the sentences use happy rather than cheerful

SEMANTIC COMPUTATIONS ON WORD VECTORS

```
print(gpt2_tokenizer.vocab_size)
input_text = gpt2_tokenizer.encode("We like Unicorns because they")
print(gpt2_tokenizer.convert_ids_to_tokens(input_text))
input_text = gpt2_tokenizer.encode("Here is the sentence I want embeddings for.")
print(gpt2_tokenizer.convert_ids_to_tokens(input_text))
```

```
50257
['We', 'Ġlike', 'ĠUnic', 'orns', 'Ġbecause', 'Ġthey']
['Here', 'Ġis', 'Ġthe', 'Ġsentence', 'ĠI', 'Ġwant', 'Ġembed', 'd', 'ings', 'Ġfor', '.']
```

You can see the BPE and the \u0120 as the special signalling character

SEMANTIC COMPUTATIONS ON WORD VECTORS

```
print(bert_tokenizer.vocab_size)
input_text = bert_tokenizer.encode("We like Unicorns because they")
print(bert_tokenizer.convert_ids_to_tokens(input_text))
input_text = bert_tokenizer.encode("Here is the sentence I want embeddings for.")
print(bert_tokenizer.convert_ids_to_tokens(input_text))
input_text = bert_tokenizer.encode("I don't know if it is an embedding, embeddable or can be embedded")
print(bert_tokenizer.convert_ids_to_tokens(input_text))
```

```
30522
['[CLS]', 'we', 'like', 'unicorn', '##s', 'because', 'they', '[SEP]']
['[CLS]', 'here', 'is', 'the', 'sentence', 'i', 'want', 'em', '##bed', '##ding', '##s', 'for', '.', '[SEP]']
['[CLS]', 'i', 'don', "'", 't', 'know', 'if', 'it', 'is', 'an', 'em', '##bed', '##ding', ',', 'em', '##bed', '##dable', 'o
r', 'can', 'be', 'embedded', '[SEP]']
```

LOAD BERT-BASE

```
]: # Let us try how BERT does tokenization
from transformers import BertTokenizer, BertModel
bert_tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
bert_model = BertModel.from_pretrained('bert-base-uncased')
```

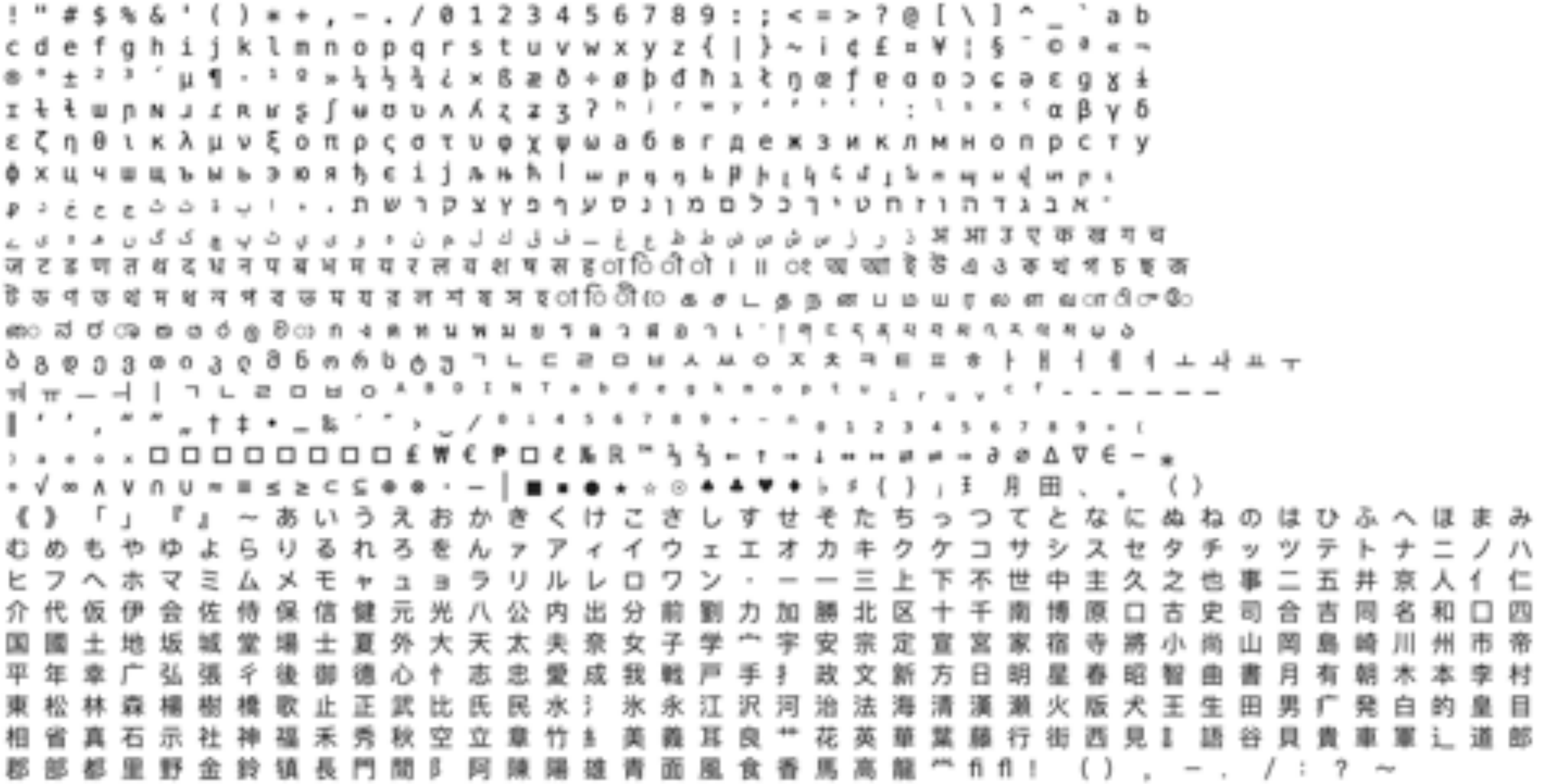
```
Downloading 100% 232k/232k [00:00<00:00, 7.34MB/s]

Downloading 100% 361/361 [00:00<00:00, 18.3kB/s]

Downloading 100% 440M/440M [00:09<00:00, 47.0MB/s]
```

BERT SINGLE CHARACTER TOKENS

Number of single character tokens: 997



BERT TOKENIZATION

```
bank_text = "After robbing the bank vault, the bank robber was seen fishing near the river bank"
bank_text = "After robbing the bank vault, the bank robber was seen fishing on the Mississippi river bank"
```

```
encoded_text = bert_tokenizer.encode(bank_text)
print(bert_tokenizer.convert_ids_to_tokens(encoded_text))
```

```
['[CLS]', 'after', 'robb', '##ing', 'the', 'bank', 'vault', ',', 'the', 'bank', 'robber', 'was', 'seen', 'fishing', 'on',
'the', 'mississippi', 'river', 'bank', '[SEP]']
```

BERT VECTORS

```
print('First 5 vector values for each instance of "bank".')
print('')
print("bank vault   ", encoded_layers[0][5][:5])
print("bank robber  ", encoded_layers[0][9][:5])
print("river bank   ", encoded_layers[0][18][:5])
```

```
First 5 vector values for each instance of "bank".

bank vault    tensor([ 0.8436, -0.4816, -0.0840,  0.4035,  0.6408])
bank robber   tensor([ 0.8196, -0.4100, -0.1249,  0.3517,  0.5315])
river bank    tensor([-0.3711, -0.6972, -0.6805, -0.1639,  0.4114])
```

The vectors are different, for word2vec they would have been the same

```
0 (101, '[CLS]')
1 (2044, 'after')
2 (26211, 'robb')
3 (2075, '##ing')
4 (1996, 'the')
5 (2924, 'bank')
6 (11632, 'vault')
7 (1010, ',')
8 (1996, 'the')
9 (2924, 'bank')
10 (27307, 'robber')
11 (2001, 'was')
12 (2464, 'seen')
13 (5645, 'fishing')
14 (2006, 'on')
15 (1996, 'the')
16 (5900, 'mississippi')
17 (2314, 'river')
18 (2924, 'bank')
19 (102, '[SEP]')
```

BERT VECTOR REPRESENTATIONS

```
# Calculate the cosine similarity between the word bank
# in "bank robber" vs "river bank" (different meanings).
diff_bank_1 = 1 - cosine(encoded_layers[0][9], encoded_layers[0][18])
# in "bank vault" vs "river bank" (different meanings).
diff_bank_2 = 1 - cosine(encoded_layers[0][5], encoded_layers[0][18])

# Calculate the cosine similarity between the word bank
# in "bank robber" vs "bank vault" (same meaning).
same_bank_1 = 1 - cosine(encoded_layers[0][5], encoded_layers[0][9])
# in "robb" vs "robber" (similar meaning).
same_bank_2 = 1 - cosine(encoded_layers[0][2], encoded_layers[0][10])

# in "fishing" vs "river" (similar meaning)
fishing_river = 1 - cosine(encoded_layers[0][13], encoded_layers[0][17])
# in "fishing" vs "bank (vault)" (different meaning)
fishing_bank_v = 1 - cosine(encoded_layers[0][13], encoded_layers[0][5])
# in "fishing" vs "(river) bank " (different meaning)
fishing_bank_r = 1 - cosine(encoded_layers[0][13], encoded_layers[0][17])
```

BERT VECTOR REPRESENTATIONS

```
Vector similarity for  *similar*  meanings ["bank (robber)" vs "bank (vault)"] :  0.95
Vector similarity for  *similar*  meanings ["robb" vs "robber"]                :  0.70

Vector similarity for *different* meanings ["bank (robber)" vs "(river) bank"] :  0.40
Vector similarity for *different* meanings ["bank (vault)" vs "(river) bank"]  :  0.40

Vector similarity for  *similar*  meanings ["fishing" vs "river"]              :  0.43

Vector similarity for *different* meanings ["fishing" vs "bank (vault)"]       :  0.27
Vector similarity for *different* meanings ["fishing" vs "(river) bank"]       :  0.43
```

Word semantics Still hold, but more contextually

- "bank (robber)" vs "bank (vault)" are very similar
- "bank (robber)" vs "(river) bank" or "bank (vault)" vs "(river) bank" are less similar
- Interestingly "fishing" vs "river" and "fishing" vs "(river) bank" have very similar scores
- While "fishing" vs "bank (vault)" is dissimilar, "fishing" vs "(river) bank" has some similarity

All in all, BERT captures the context very well in word representations

2. BERT

#BERT$_{\text{OLOGY}}$

- If you are with me till now, things are easier from now on !!
- First : Instead of RNN blocks, we use Transformer blocks – complex details, but somewhat logical and composable from an architecture point of view
- Second : They <u>do</u> need computing power including GPUs
  - This has become a battle cry for researchers and models are being developed that focus on resources rather than just performance on tasks and datasets
- Third : Transfer Learning has become a norm with very good models of the BERTology available from reputed sources
- Fourth : These are still evolutionary, there are yuuge problems to be tackled

BERTOLOGY

WORD2VEC
Jan 16, 2013

GLOVE
Jan 2, 2014

Character-Aware NLM
Dec 1, 2015

BPE
Aug 7, 2016

Transformer
Jun 12, 2017

ELMO
Feb 15, 2018

GPT
June 11, 2018

Sentence Piece
Aug 19, 2018

BERT
Oct 11, 2018

Transformer-XL
Jan 9, 2019

Evolved Transformer
Jan 30, 2019

GPT-2
Feb 14, 2019

ERNIE
April 19, 2019

XLNet
Jun 19, 2019

RoBERTa
July 26, 2019

CTRL
Sep 11, 2019

ALBERT
Sep 26, 2019

DistilBERT
Oct 2, 2019

TURING-NLG
Feb 13, 2020

ELECTRA
March 9, 2020

-
-

-
-

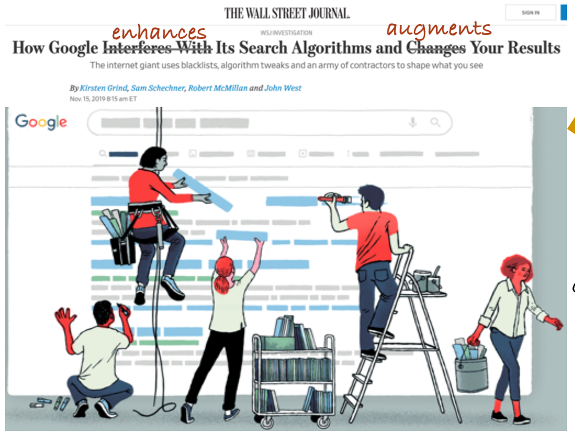
WHAT DOES BERT DO FOR A LIVING ?

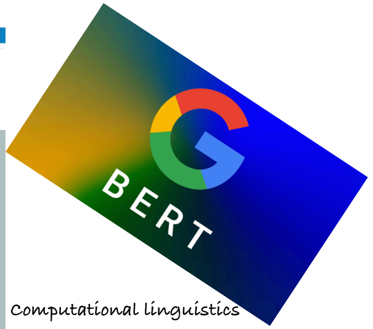
- BERT is a pre-trained deep learning natural language framework
- Lexical ambiguity
  - Polysemy e.g. get, run, extremely subtle and nuanced
  - Homonymy – different meanings
    - e.g. bass instrument and fish; bank http://www.singularis.ltd.uk/bifroest/misc/homophones-list.html
  - Synonymy similar meanings
    - e.g. good, excellent
- Coreference resolution
- Anaphora and cataphora resolution
  - *The car is falling apart, but it still works*
  - *She was at NYU when Mary realized she had lost her keys*
- Multi-sentential resolution
  - *John took two trips around France; They were both wonderful*

https://searchengineland.com/a-deep-dive-into-bert-how-bert-launched-a-rocket-into-natural-language-understanding-324522

BERT & GOOGLE SEARCH

- BERT helps with polysemic resolution
- Ambiguous and nuanced queries impacted
  - *10% nuanced ones*, 15% temporal queries
- Recall and Precision are impacted
- BERT will help with coreference resolution
- BERT provides a solid *linguistic foundation*
- BERT will be huge for conversational search and assistant





Bill Slawski @bill_slawski

Replying to @thetafferboy and @dawnieando

I expect the definitive guide to BERT should be rolling out at any second now, filled with more mythology than any article should be, by someone claiming to be an expert in BERT optimization. Be careful what you believe.

7:00 AM · Oct 25, 2019 from Carlsbad, CA · Twitter for Android

Danny Sullivan @dannysullivan

Replying to @Johanstormarn @MalteLandwehr and @searchliaison

There's nothing to optimize for with BERT, nor anything for anyone to be rethinking. The fundamentals of us seeking to reward great content remain unchanged.

1:29 AM · Oct 28, 2019 · Twitter Web App

Ryan Jones @RyanJones

the most exciting thing about BERT is that suddenly, overnight, every SEO is now an expert on neural networks and natural language processing.

1:39 PM · Oct 25, 2019 · TweetDeck

https://searchengineland.com/a-deep-dive-into-bert-how-bert-launched-a-rocket-into-natural-language-understanding-324522

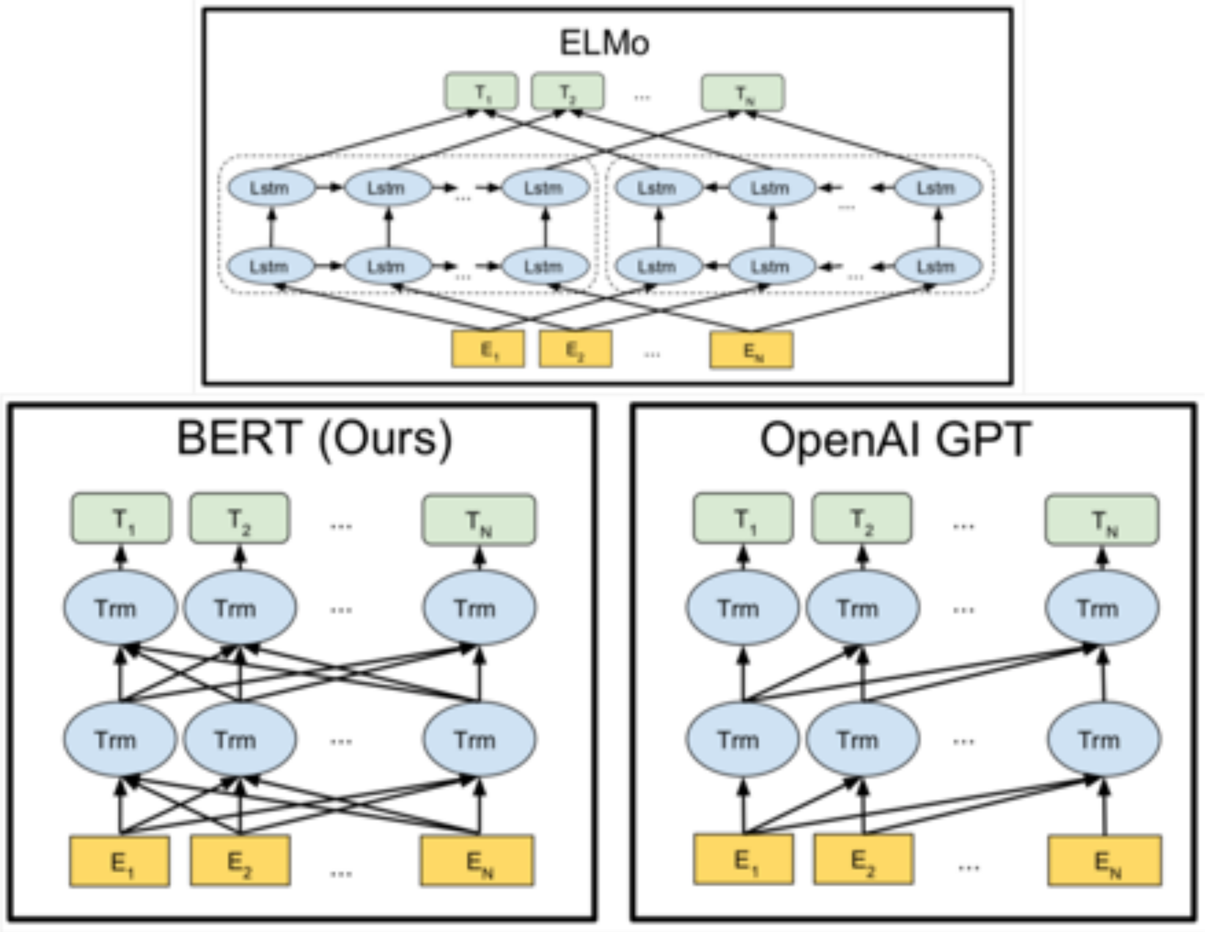
ESSENTIAL BERT -- TOP 5

- BERT is a pre-trained deep learning natural language <u>framework</u>
  - Finetuning with task specific layers
  - *".. the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, … without substantial task specific architecture modifications"*
- ***<u>Bidirectional</u>*** Encoder Representations from Transformers
  - Actually, it is non-directional !!
- Word Embedding – Sentence piece
- Larger Transformer stack – only the encoder
  - Not for NLG
- Trained on Language Model Tasks

BERT VS THE REST

- Language models only use left context *or* right context, but language understanding is bidirectional

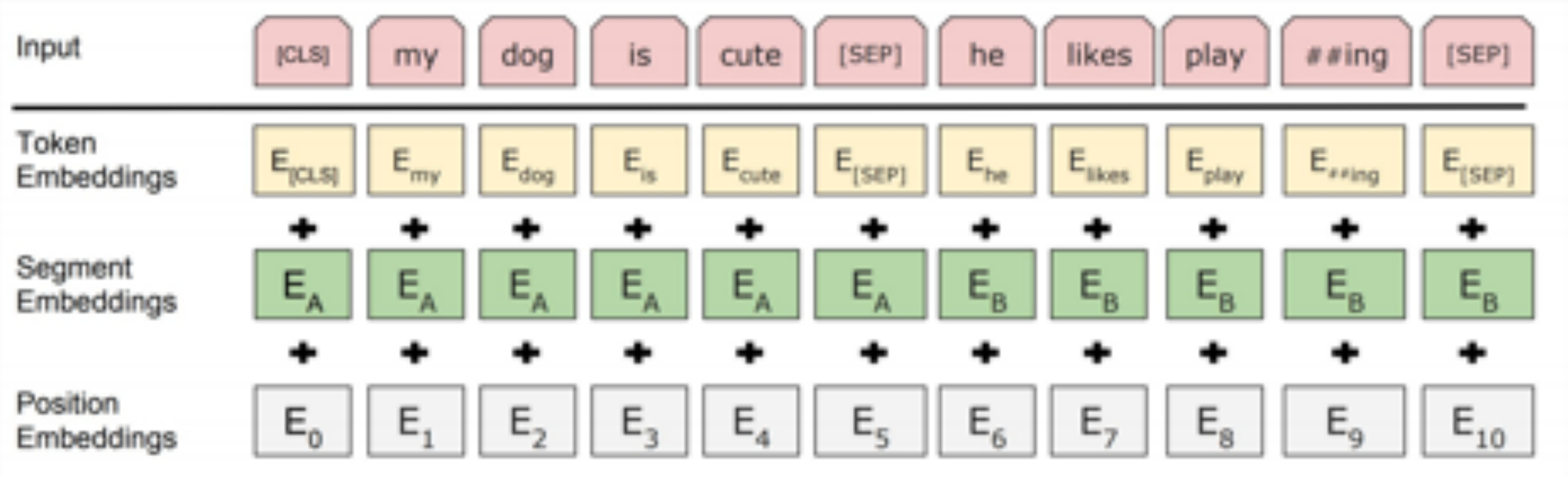
BERT: Devlin, Chang, Lee, Toutanova (2018)



58

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

INPUT REPRESENTATION



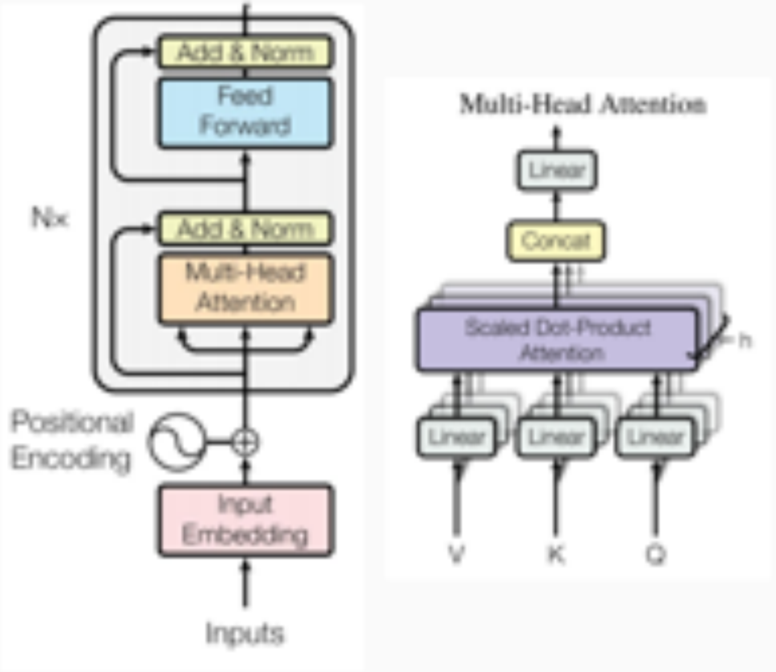
- Use 30,000 WordPiece vocabulary on input.
- Each token is sum of three embeddings
- Single sequence is much more efficient.

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

MODEL ARCHITECTURE

Transformer encoder

- Multi-headed self attention
  - Models context
- Feed-forward layers
  - Computes non-linear hierarchical features
- Layer norm and residuals
  - Makes training deep networks healthy
- Positional embeddings
  - Allows model to learn relative positioning



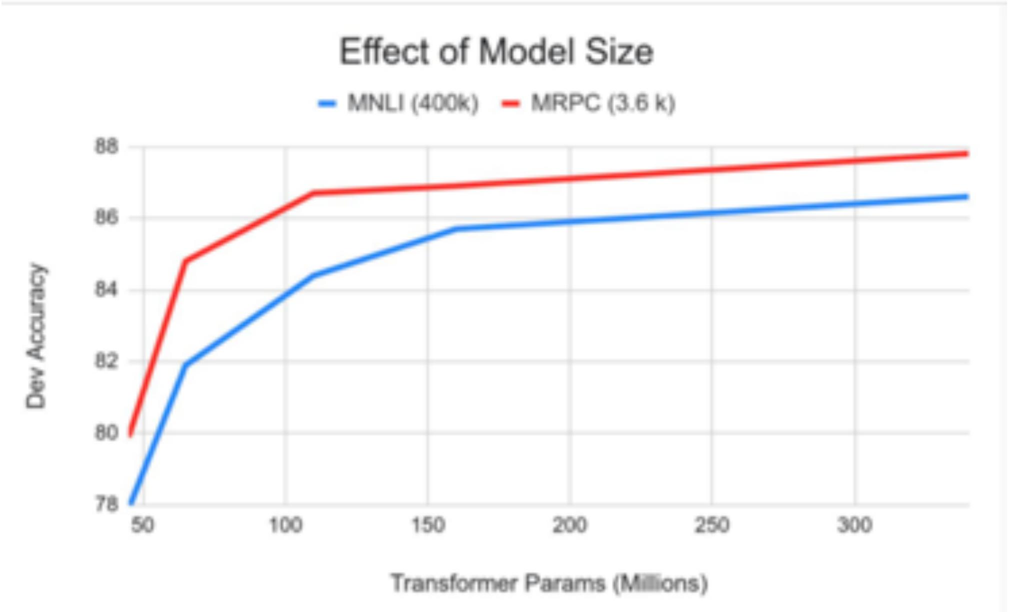
Things we have seen before in this tutorial

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

BERT NETWORK ARCHITECTURE

- Wikipedia (2.5B words) + BookCorpus (800 M words)
- Trained on 4x4 or 8x8 TPU slice for 4 days
- Limited to sequences of 512 tokens due to dimensionality of the positional embeddings
- GPT-2 1024 tokens
- AdamW optimizer with 1e-4 LR, linear decay
- 1 M steps (~40 epochs)

Model	Parameters	Layers	d_{model}
BERT-Base	110M	12	768
BERT-Large	340M	24	1024

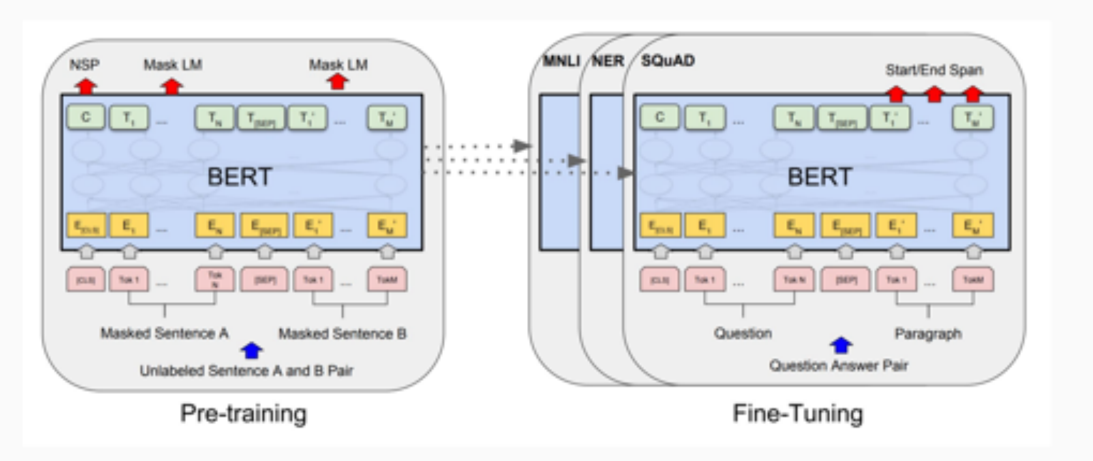


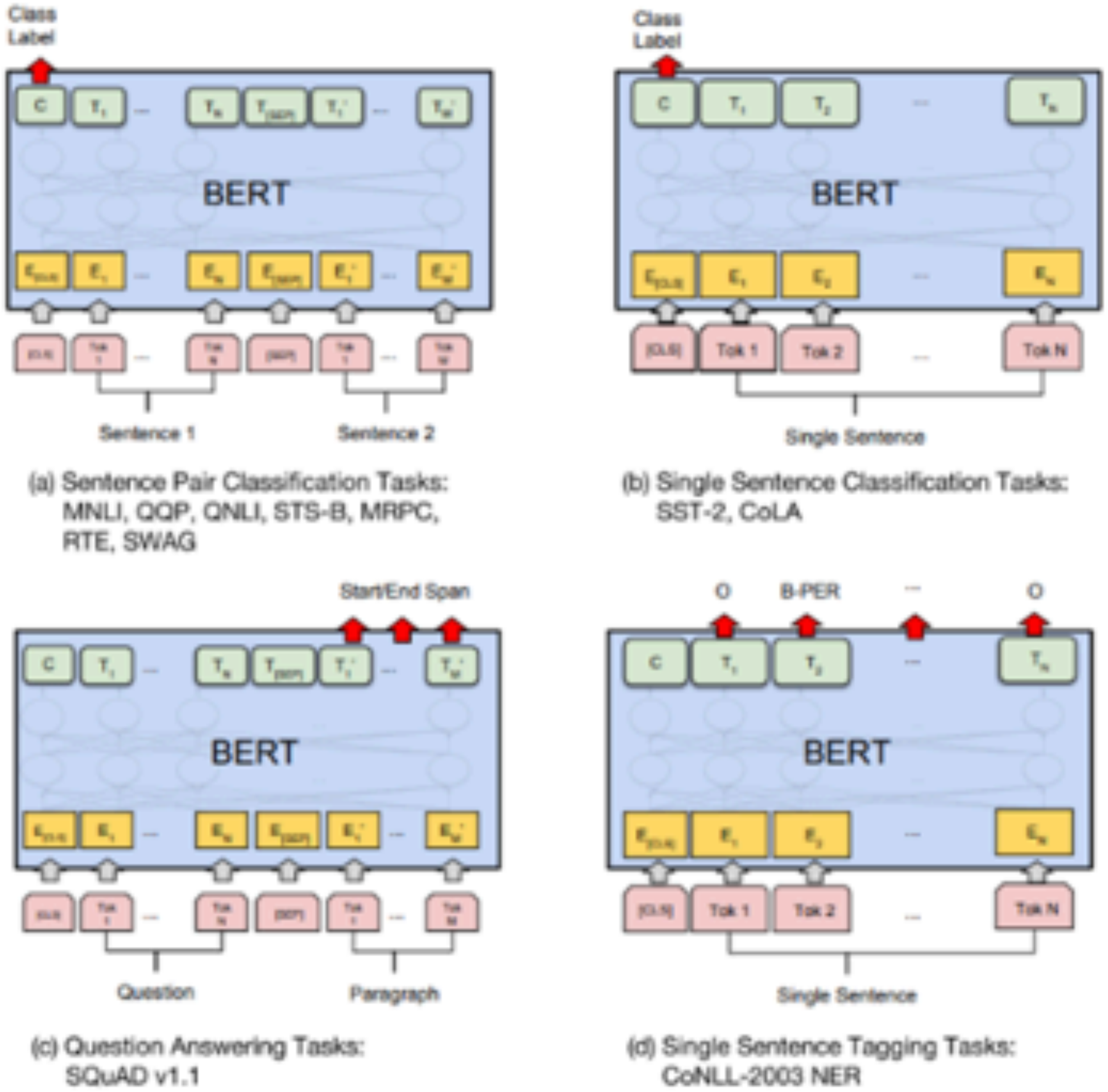
Go from a small to big model – huge improvements even with very small datasets 3,600 labelled samples

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

FINE-TUNING

- Take pre-trained BERT model
- Do additional training that fits your task
- Better results with less data

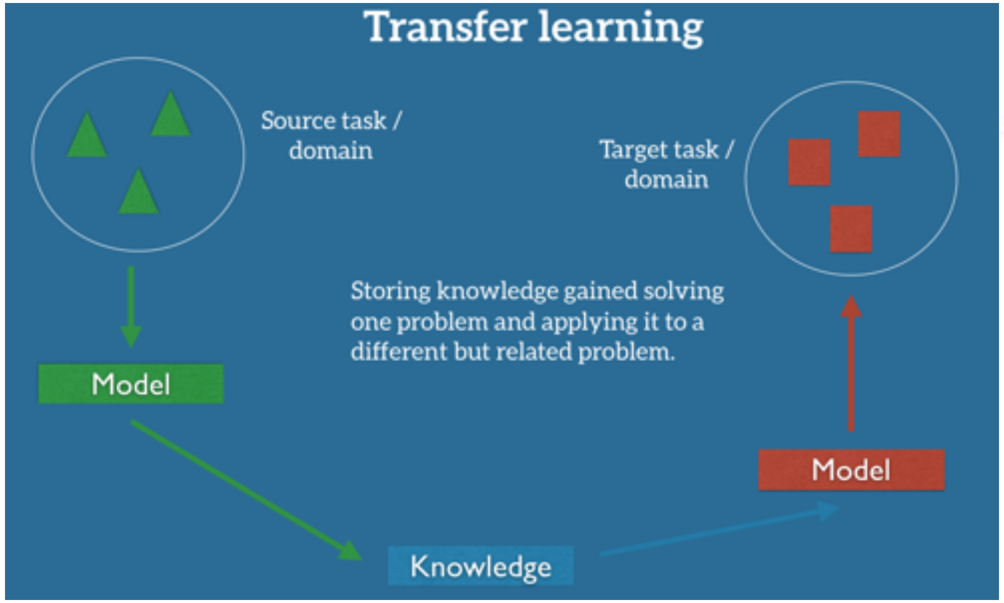


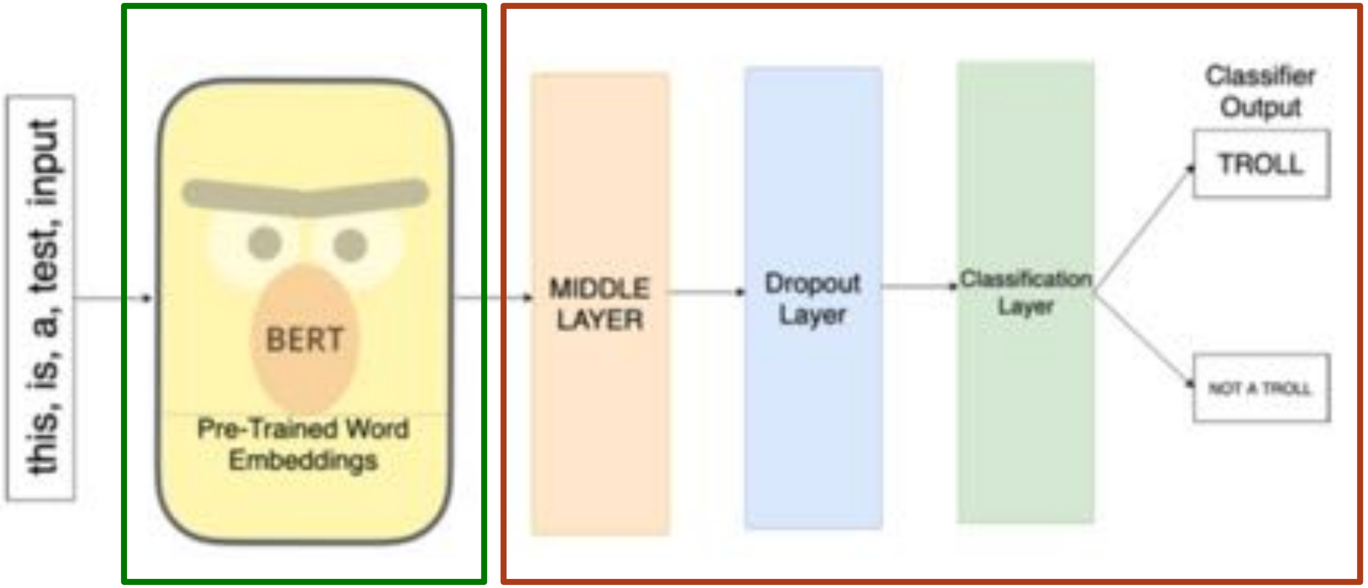


https://nlp.stanford.edu/seminar/details/jdevlin.pdf

BERT FINETUNING

- cs224n student project
- Using BERT and CNNs for Russian Troll Detection on Reddit





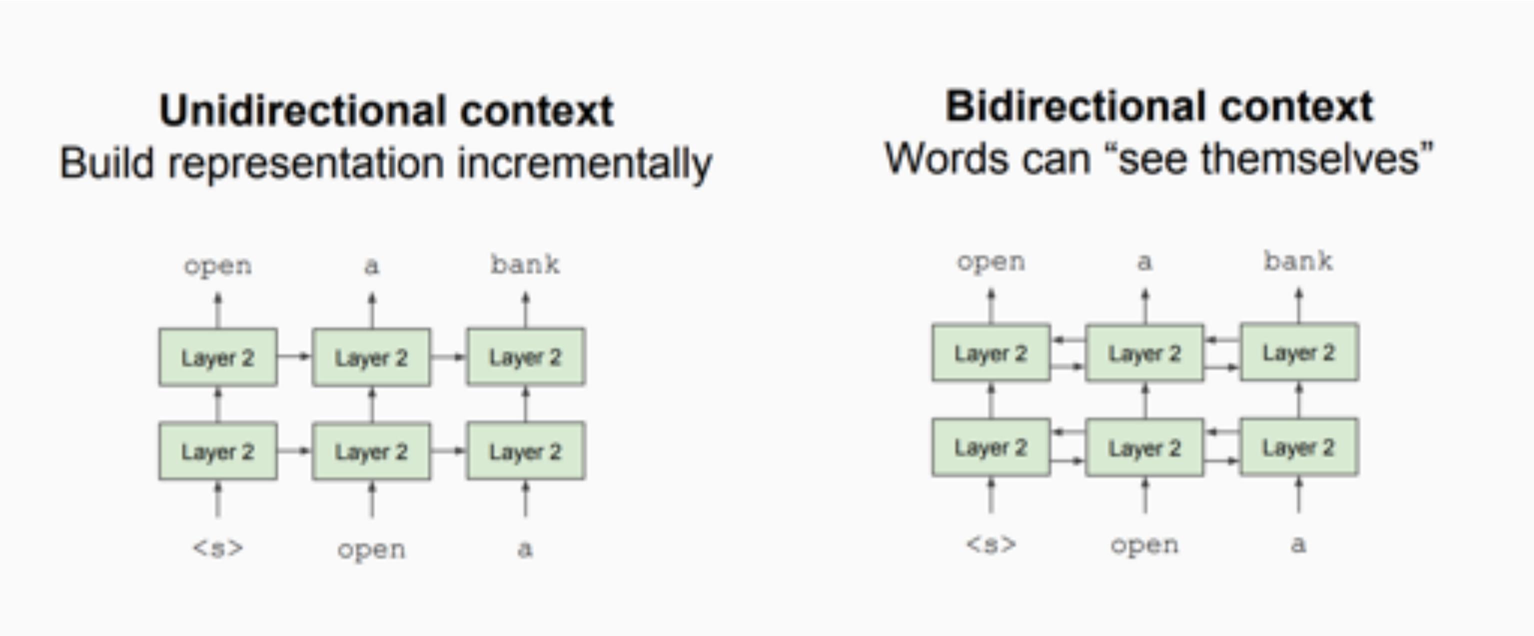
	AUC	Accuracy
Optimized Forest Classifier Baseline	0.74	0.74
BERT Baseline	0.659	0.625
BERT + LSTM	0.805	0.727
Vanilla BERT + CNN	0.826	0.741
Optimized BERT + CNN	0.843	0.756
Vanilla BERT + RCNN	0.836	0.746
Optimized BERT + RCNN	**0.846**	0.749

https://towardsdatascience.com/using-bert-and-cnns-for-russian-troll-detection-on-reddit-8ae59066b1c

https://ruder.io/transfer-learning/

BIDIRECTIONAL CONTEXT

- **Training Tasks**
- Predict next word ("Language Model")
- Words can "see themselves" in a bidirectional encoder
- So training is a challenge



- **Two innovative training Tasks**

1. Predict Missing Word ("Cloze Tasks")
   - Masked LM
2. Next Sentence Prediction

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

TASK 1 : MASKED LM

- **Solution**: Mask out *k*% of the input words, and then predict the masked words
    - We always use *k* = 15%

```
                      store          gallon
                        ↑               ↑
the man went to the [MASK] to buy a [MASK] of milk
```

- Too little masking: Too expensive to train
- Too much masking: Not enough context

- Problem: Mask token never seen at fine-tuning
- Solution: 15% of the words to predict, but don't replace with `[MASK]` 100% of the time. Instead:
- 80% of the time, replace with `[MASK]`
  `went to the store → went to the [MASK]`
- 10% of the time, replace random word
  `went to the store → went to the running`
- 10% of the time, keep same
  `went to the store → went to the store`

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

TASK 2 : NEXT SENTENCE PREDICTION

- To learn *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`

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

BERT PERFORMANCE - 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
$BERT_{BASE}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**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://nlp.stanford.edu/seminar/details/jdevlin.pdf

WE ARE NOT DONE YET !!

- Is modeling "solved" in NLP? I.e., is there a reason to come up with novel model architectures?
  - But that's the most fun part of NLP research :(
- Maybe yes, for now, on some tasks, like SQuAD-style QA.
  - At least using the same deep learning "lego blocks"
- Examples of NLP models that are not "solved":
  - Models that minimize total training cost vs. accuracy on modern hardware
  - Models that are very parameter efficient (e.g., for mobile deployment)
  - Models that represent knowledge/context in latent space
  - Models that represent structured data (e.g., knowledge graph)
  - Models that jointly represent vision and language

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

THE DARK (LINGUISTIC) SECRETS OF BERT

- How much of BERT is linguistically interpretable self-attention patterns ?
- Are there any linguistically interpretable self-attention heads?
- BERT is heavily overparametrized
  - Many heads have functional duplicates, i.e. disabling one head would not harm the model because the same information is available elsewhere – post pruning
- BERT does not need to be all that smart for these tasks
  - *… Leveraging this "excess of capacity" … instead of discarding these redundant heads, can we use them more efficiently to "cram more knowledge" into the model? – Paul Michel, CMU*
- No commonsense reasoning; opportunity for knowledge graph integration
- An alternative explanation is that BERT's success is due to ~~black magic~~ something other than self-attention !
  - May be there is some deep pattern we cannot comprehend !

https://text-machine-lab.github.io/blog/2020/bert-secrets/
https://blog.ml.cmu.edu/2020/03/20/are-sixteen-heads-really-better-than-one/

ADVERSARIAL ATTACKS

	BERT						BERT
	MR	IMDB	Yelp	AG	Fake	SNLI	MultiNLI (m/mm)
Original Accuracy	86.0	90.9	97.0	94.2	97.8	89.4	85.1/82.1
After-Attack Accuracy	11.5	13.6	6.6	12.5	19.3	4.0	9.6/8.3
% Perturbed Words	16.7	6.1	13.9	22.0	11.7	18.5	15.2/14.6
Semantic Similarity	0.65	0.86	0.74	0.57	0.76	0.45	0.57/0.58
Query Number	166	1134	827	357	4403	60	78/86
Average Text Length	20	215	152	43	885	8	11/12

From 89% accuracy to 4% with just 18.5% words

MIT Technology Review

Artificial Intelligence Feb 7

Software that swaps out words can now fool the AI behind Alexa and Siri

<u>Movie Review</u>

Original (Label: NEG) The characters, cast in impossibly contrived situations, are totally estranged from reality.
Attack (Label: POS) The characters, cast in impossibly ***engineered circumstances***, are ***fully*** estranged from reality.

Original (Label: POS) It cuts to the ***knot*** of what it actually means to face your ***scares***, and to ride the overwhelming metaphorical wave that life wherever it takes you.
Attack (Label: NEG) It cuts to the ***core*** of what it actually means to face your ***fears***, and to ride the ***big*** metaphorical wave that life wherever it takes you

Article: Super Bowl 50
Paragraph: *"Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."*
Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

ADVERSARIAL ATTACKS

	BERT					BERT	
	MR	IMDB	Yelp	AG	Fake	SNLI	MultiNLI (m/mm)
Original Accuracy	86.0	90.9	97.0	94.2	97.8	89.4	85.1/82.1
After-Attack Accuracy	11.5	13.6	6.6	12.5	19.3	4.0	9.6/8.3
% Perturbed Words	16.7	6.1	13.9	22.0	11.7	18.5	15.2/14.6
Semantic Similarity	0.65	0.86	0.74	0.57	0.76	0.45	0.57/0.58
Query Number	166	1134	827	357	4403	60	78/86
Average Text Length	20	215	152	43	885	8	11/12

From 89% accuracy to 4% with just 18.5% words

MIT Technology Review

Artificial Intelligence Feb 7 ...

Software that swaps out words can now fool the AI behind Alexa and Siri

SNLI (Entailment (ENT), Neutral (NEU), Contradiction (CON))

Premise : A child with wet hair is holding a butterfly decorated beach ball.
Original (Label: NEU) : The child is at the beach.
Adversary (Label: ENT) : The youngster is at the shore.

Premise : Two small boys in blue soccer uniforms use a wooden set of steps to wash their hands.
Original (Label: CON) : The boys are in band uniforms.
Adversary (Label: ENT) : The boys are in band garment.

Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*

Adversary Adds: ***Tadakatsu moved to the city of Chicago in 1881.***
Model Predicts: *Chicago*

Google Releases PAWS Dataset to Improve Natural Language Understanding

Fhel Dimaano Follow
Oct 9, 2019 · 2 min read ★

PAWS stands for Paraphrase Adversaries from Word Scrambling. In layman's terms it is the process of reading a sentence, reordering the words, and checking if the sentence has the same meaning.

1. Driving from New York to Maine
2. Driving to Maine from New York.
3. Driving from Maine to New York.

All three sentences use the same words in a different order. Sentences 1 and 2 have the same meaning, therefore they are **paraphrase pairs**. Sentence 1 and 3 have different meanings, therefore they are **non-paraphrase pairs**.

Google claims that even state-of-the-art models, like BERT, would not be able to correctly identify the difference between many non-paraphrase pairs.

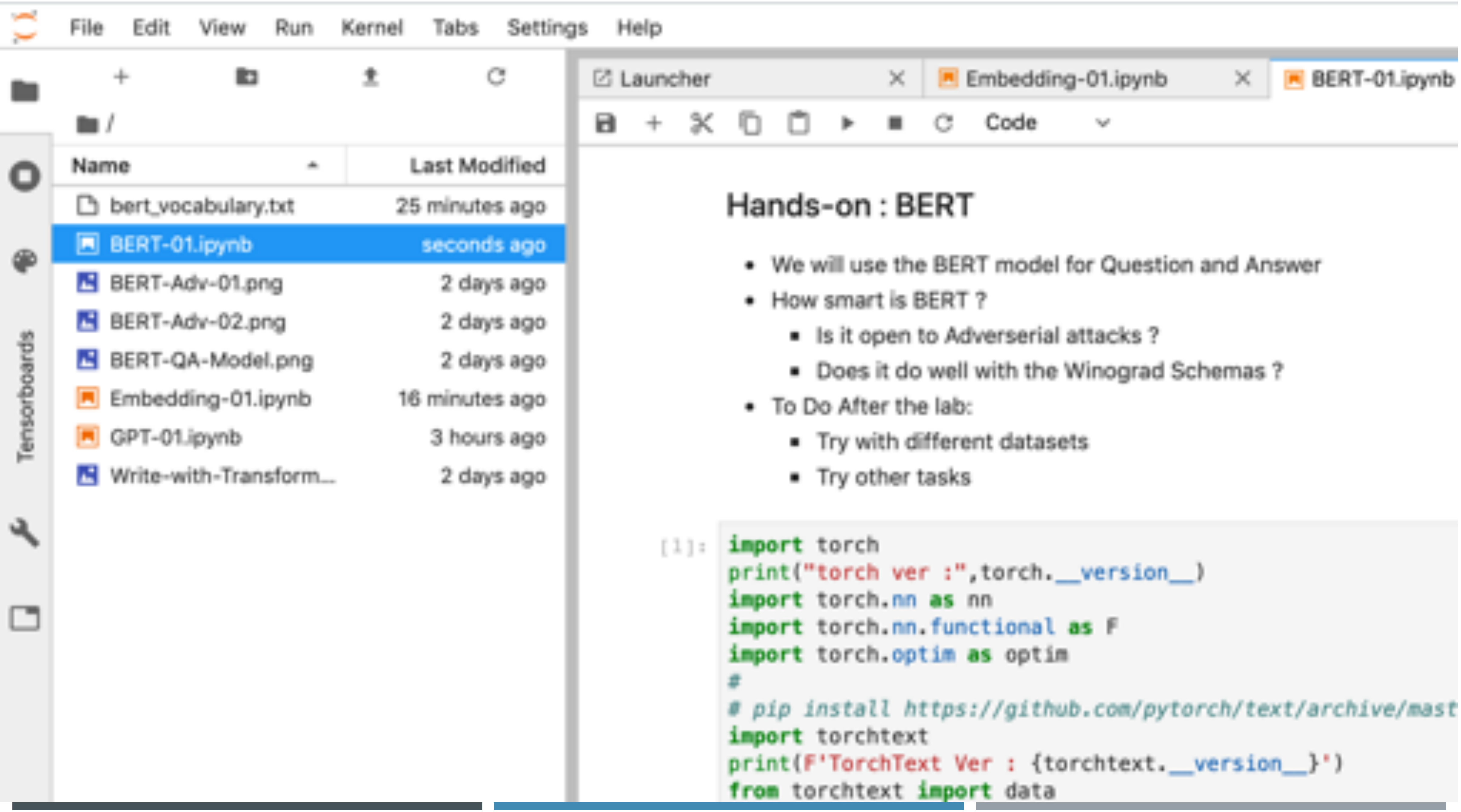
https://medium.com/swlh/google-releases-paws-dataset-to-improve-natural-language-understanding-f82a814464b4

SCALE OF THINGS

Model	Parameters	Transformer Layers/…	Training	
Medium-sized LSTM	10 M			
ELMo	90 M			
Transformers	~60 M	6 L / 512 D / 8 Heads	3.5 days/8 GPU	12/2017
GPT	110 M	12 L / 768 D / 12 Heads	240 GPU Days	6/2018
BERT(Base)	110 M	12 L / 768 D / 12 Heads	16 TPU days(4X4d)	10/2018
BERT (L)	320 M	24 L / 1,024 D / 16 Heads	64 TPU Days(16X4)	10/2018
Honeybee Brain	~1B synapses			
GPT-2 (XL)	1.5 B	48 L / 1,600 D / 16 Heads	2048 TPU v3 Days	2/2019
Megatron GPT-2	8.3 B	72 L / 3,072 D / 24 Heads	47 Min/512 GPU	4/2019
Turing-NLG	17 B	78 L / 4,256 D / 28 Heads		2/2020
Human Brain Models	1 T	?	Lifetime !	>1.5M Y

- *GPT-2 : Really big Transformer LM trained on 40GB text*
- *Model > 1.3B won't fit in GPU, so need model parallelization and broken across multiple GPUs*
- *D = Model Dimensionality (d_{model})*

HANDS-ON : BERT


File Edit View Run Kernel Tabs Settings Help
Launcher
Embedding-01.ipynb
BERT-01.ipynb
Code
Name
Last Modified
bert_vocabulary.txt 25 minutes ago
BERT-01.ipynb seconds ago
BERT-Adv-01.png 2 days ago
BERT-Adv-02.png 2 days ago
BERT-QA-Model.png 2 days ago
Embedding-01.ipynb 16 minutes ago
GPT-01.ipynb 3 hours ago
Write-with-Transform... 2 days ago
Tensorboards
Hands-on : BERT
We will use the BERT model for Question and Answer
How smart is BERT ?
Is it open to Adverserial attacks ?
Does it do well with the Winograd Schemas ?
To Do After the lab:
Try with different datasets
Try other tasks
[1]:
import torch
print("torch ver :",torch.__version__)
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
#
# pip install https://github.com/pytorch/text/archive/mast
import torchtext
print(F'TorchText Ver : {torchtext.__version__}')
from torchtext import data

LOAD BERT-LARGE

```
def get_answer_from_bert(model,question,passage):
    input_ids = tokenizer.encode(question, passage)
    token_type_ids = [0 if i <= input_ids.index(102) else 1 for i in range(len(in
    start_scores, end_scores = model(torch.tensor([input_ids]), token_type_ids=to

    all_tokens = tokenizer.convert_ids_to_tokens(input_ids)
    answer = ' '.join(all_tokens[torch.argmax(start_scores) : torch.argmax(end_sc
    return (answer)

question, passage = "Who was Jim Henson?", "Jim Henson was a nice puppet"
answer = get_answer_from_bert(bert_l_qa_model,question,passage)
assert answer == "a nice puppet"
print(answer)
```

```
Downloading 100% 398/398 [00:00<00:00, 19.2kB/s]

Downloading 100% 1.34G/1.34G [00:28<00:00, 47.7MB/s]
```

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

IT WORKS !

```
passage = "Peyton Manning became the first quarterback ever to lead \
two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play \
in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super\
Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager"
question = "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"

answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
john el ##way
```

```
passage = "Peyton Manning became the first quarterback ever to lead \
two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play \
in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super\
Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.\
Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV"

question = "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"

answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
john el ##way
```

Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*

Adversary Adds: ***Tadakatsu moved to the city of Chicago in 1881.***
Model Predicts: *Chicago*

IT WORKS AS WELL !

```
passage = "In January 1880, two of Tesla's uncles put together enough money to help him leave \
Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at \
Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in \
Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an \
auditor, he did not receive grades for the courses."

question = "What city did Tesla move to in 1880?"

answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
prague
```

```
passage = "In January 1880, two of Tesla's uncles put together enough money to help him leave \
Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at \
Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in \
Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an \
auditor, he did not receive grades for the courses. Tadakatsu moved to the city of Chicago in 1881."

question = "What city did Tesla move to in 1880?"

answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
prague
```

SUCCESSES ON WINOGRAD SCHEMAS

Good News, the attacks do not work !

What about the Winograd Schemas ?

```
[3]: passage = "The city councilmen refused the demonstrators a permit because they feared violence"
question = "Who feared violence?"
answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
city council ##men
```

```
[4]: passage = "The city councilmen refused the demonstrators a permit because they advocated violence"
question = "Who advocated violence?"
answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
demonstrators
```

```
[5]: # Ref: Using Answer Set Programming for Commonsense Reasoning in the Winograd Schema Challenge
#   https://arxiv.org/abs/1907.11112
passage = "The fish ate the worm. It was tasty"
question = "What was tasty?"
answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
the worm
```

FAILS ON (FINER) WINOGRAD SCHEMAS

16]:
```python
passage = "The trophy didn’t fit into the suitcase because it was too large"
question = "What was too large?"
answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
the trophy
```

May be not !

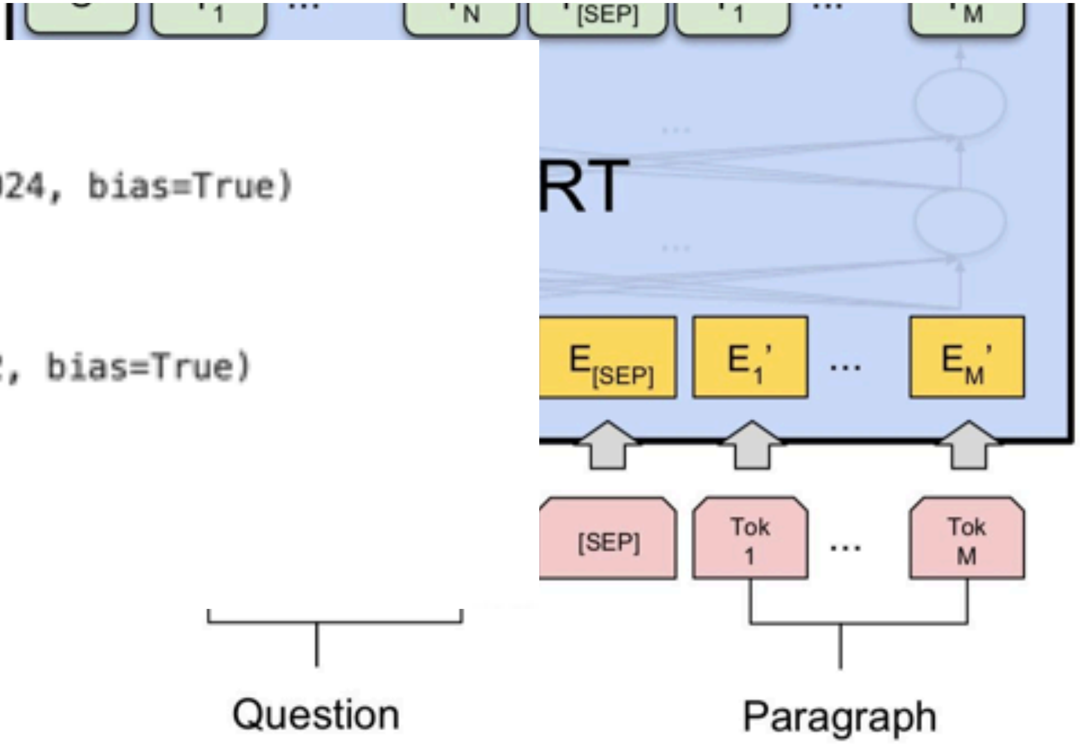
17]:
```python
passage = "The trophy didn’t fit into the suitcase because it was too tiny"
question = "What was too tiny?"
answer = get_answer_from_bert(bert_l_qa_model,question,passage)
print(answer)
```

```
the trophy
```

```
Layer bert.encoder.layer.23.output.LayerNorm.weight     torch.Size([1024])          elements : 1,024
Layer bert.encoder.layer.23.output.LayerNorm.bias       torch.Size([1024])          elements : 1,024
Layer bert.pooler.dense.weight                          torch.Size([1024, 1024])    elements : 1,048,576
Layer bert.pooler.dense.bias                            torch.Size([1024])          elements : 1,024
Layer qa_outputs.weight                                 torch.Size([2, 1024])       elements : 2,048
Layer qa_outputs.bias                                   torch.Size([2])             elements : 2
==================================================================================================
Total Params:335,143,938
```

```
          )
        )
        (pooler): BertPooler(
          (dense): Linear(in_features=1024, out_features=1024, bias=True)
          (activation): Tanh()
        )
      )
      (qa_outputs): Linear(in_features=1024, out_features=2, bias=True)
    )
```

Just adds a linear layer with Span-start and Span-end



A WORD ABOUT DATASETS

- Good source for datasets
- https://course.fast.ai/datasets
- http://xzh.me/docs/charconvnet.pdf
- There is also the imdb dataset 25K train, 25K test
- torchtext.datasets is a convenient way of accessing the datasets
- We will see how

Large-scale Datasets: Veriety of Different Sample Sizes

Dataset	Classes	Train Samples	Test Samples	Epoch Size
AG's News	4	120,000	7,600	5,000
Sogou News	5	450,000	60,000	5,000
DBPedia	14	560,000	70,000	5,000
Yelp Review Polarity	2	560,000	38,000	5,000
Yelp Review Full	5	650,000	50,000	5,000
Yahoo! Answers	10	1,400,000	60,000	10,000
Amazon Review Full	5	3,000,000	650,000	30,000
Amazon Review Polarity	2	3,600,000	400,000	30,000

- Data augmentation using thesaurus: replace words with their synonyms.

- x

3. GPT

AI Too Dangerous to Release ?

AI is not Galloping Horses ..

It has semantics ..

GPT-2 : Open AI Staged Release

- 124 M Model in Feb'19
- 355 M - May'19
- 774 M - Aug'19
- 1,558 M - ? Nov'19

Since then ...

Nvidia GPT-2 8B in 53 minutes with 1,500 GPUs (Apr'19) ...

Salesforce 1.6B Parameters (Sep'19) ...

Microsoft Turing-NLG 17B Parameters (Feb'20) ...

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

Zack Whittaker @zackwhittaker / 9:17 am PST • February 17, 2019

- *Generate misleading news articles*
- *Impersonate others online*
- *Automate the production of abusive or faked content to post on social media*
- *Automate the production of spam/phishing content*

"Recycling is good for the world, no, you could not be more wrong,"

"Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. <u>Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources</u>."

https://towardsdatascience.com/examining-the-transformer-architecture-part-1-the-openai-gpt-2-controversy-feceda4363bb

OpenAI built a text generator so

gn in Search UK edition

The Guardian

More

- Generate misleading news articles
- Impersonate others online
- Automate the production of abusive or faked content to post on social media

f spam/phishing content

Revealed: quarter of all tweets about climate crisis produced by bots

Draft of Brown study says findings suggest 'substantial impact of mechanized bots in amplifying denialist messages'

Oliver Milman *in New York*

@olliemilman

Fri 21 Feb 2020 08.00 GMT

ironment. It is destructive
for our health. It contribut
omy. It increases the cost
cling is not good for our n
st advanced and efficient r
and resources."

-architecture-part-1-the-openai-(... 63bb

our
a
ases
urn,

" In terms of influence, I personally am convinced that they do make a difference, although this can be hard to quantify
Stephen Lewandowsky

IS GPT-2 THAT POWERFUL TO DECEIVE US ALL ?

- *Fluency*
- *Consistency*
- *Congruency*

The “ unicorn” sample reads like a real science press release. The “ theft of nuclear material” sample reads like a real news story. The “ Miley Cyrus shoplifting” sample reads like a real post from a celebrity gossip site. The “ GPT-2” sample reads like a real OpenAI press release. The “ Legolas and Gimli” sample reads like a real fantasy novel. The “ Civil War homework assignment” reads like a real C-student’s paper. The “ JFK acceptance speech” reads like a real politician’s speech. The “ recycling “ sample reads like a real right-wing screed.

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, 10 TRIES)

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 … “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.
While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. …
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, “In South America, such incidents seem to be quite common.”
However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. “But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,” said the scientist.

IS GPT-2 THAT POWERFUL TO DECEIVE US ALL ?

- *Fluency*
- *Consistency*
- *Congruency*

The " unicorn" sample reads like a real science press release. The " theft of nuclear material" sample reads like a real news story. The " Miley Cyrus shoplifting" sample reads like a real post from a celebrity gossip site. The " GPT-2" sample reads like a real OpenAI press release. The " Legolas and Gimli" sample reads like a real fantasy novel. The " Civil War homework assignment" reads like a real C-student's paper. The " JFK acceptance speech" reads like a real politician's speech. The " recycling " sample reads like a real right-wing screed.

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, 1ST TRY !)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.
In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.
"The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."
The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.
The Nuclear Regulatory Commission did not immediately release any information.
According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation.
"The safety of people, the environment and the nation's nuclear stockpile is our highest priority," Hicks said. "We will get to the bottom of this and make no excuses.

IS GPT-2 THAT POWERFUL TO DECEIVE US ALL ?

- *Fluency*
- *Consistency*
- *Congruency*

The “ unicorn” sample reads like a real science press release. The “ theft of nuclear material” sample reads like a real news story. The “ Miley Cyrus shoplifting” sample reads like a real post from a celebrity gossip site. The “ GPT-2” sample reads like a real OpenAI press release. The “ Legolas and Gimli” sample reads like a real fantasy novel. The “ Civil War homework assignment” reads like a real C-student’s paper. The “ JFK acceptance speech” reads like a real politician’s speech. The “ recycling “ sample reads like a real right-wing screed.

We’ve trained a large language model called GPT-2 that generates realistic paragraphs of text, while also exhibiting zero shot generalization on tasks like machine translation, question answering, reading comprehension, and summarization - problems usually approached by using training datasets and models designed explicitly for these tasks.
A typical approach to language modeling is to …

MODEL COMPLETION (MACHINE-WRITTEN, 1ST TRY !)

Here you can see the most recent progress with Generative Pre-trained Transformer:

Figure 1: Generative Pre-trained Transformer training on several texts.

We are now preparing a collection of datasets for translation and machine translation in our language model. We will be using one of the large number of text samples provided by The New York Times.

We believe this project is the first step in the direction of developing large NLP systems without task-specific training data. That is, we are developing a machine language system in the generative style with no explicit rules for producing text.

We hope for future collaborations between computer scientists, linguists, and machine learning researchers.

ESSENTIAL GPT

- GPT – 12 Layer Decoder-only
  - Books Corpus dataset
  - 12 layers w/ 12 attention heads
- GPT-2
  - 1024 tokens (from 512)
  - 16 attention heads
  - GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text.
- 40 GB text !
  - Curated web pages 8 Million

Model	Parameters	Layers	d_{model}	Release
GPT		12	768	
GPT-2 (S)	117M	12	768	Feb'19
BERT-Large	340M	24	1024	
GPT-2 (M)	345M	24	1024	May'19
GPT-2 (L)	762M	36	1280	Aug'19
GPT-2 (XL)	1,542M	48	1600	Nov'19

Megatron-LM GPT-2 Evaluation

Model Size	Number of layers	Hidden Size	Attention heads	Wikitext (Perplexity ↓)	Lambada (Accuracy ↑)
345 M	24	1024	16	24.21	55.04
775 M	36	1280	16	20.44	58.86
2.5 B	54	1920	20	17.83	63.30
8.3 B	72	3072	24	17.41	63.11

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

GPT-2 1.5B RELEASE

Findings:

1. Humans find GPT-2 outputs convincing
2. GPT-2 can be fine-tuned for misuse
   - e.g. ideological positions
3. Detection is challenging
4. We've seen no strong evidence of misuse so far
5. We need standards for studying bias

Model	Parameters	Layers	d_{model}	Release
GPT		12	768	
GPT-2 (S)	117M	12	768	Feb'19
BERT-Large	340M	24	1024	
GPT-2 (M)	345M	24	1024	May'19
GPT-2 (L)	762M	36	1280	Aug'19
GPT-2 (XL)	1,542M	48	1600	Nov'19

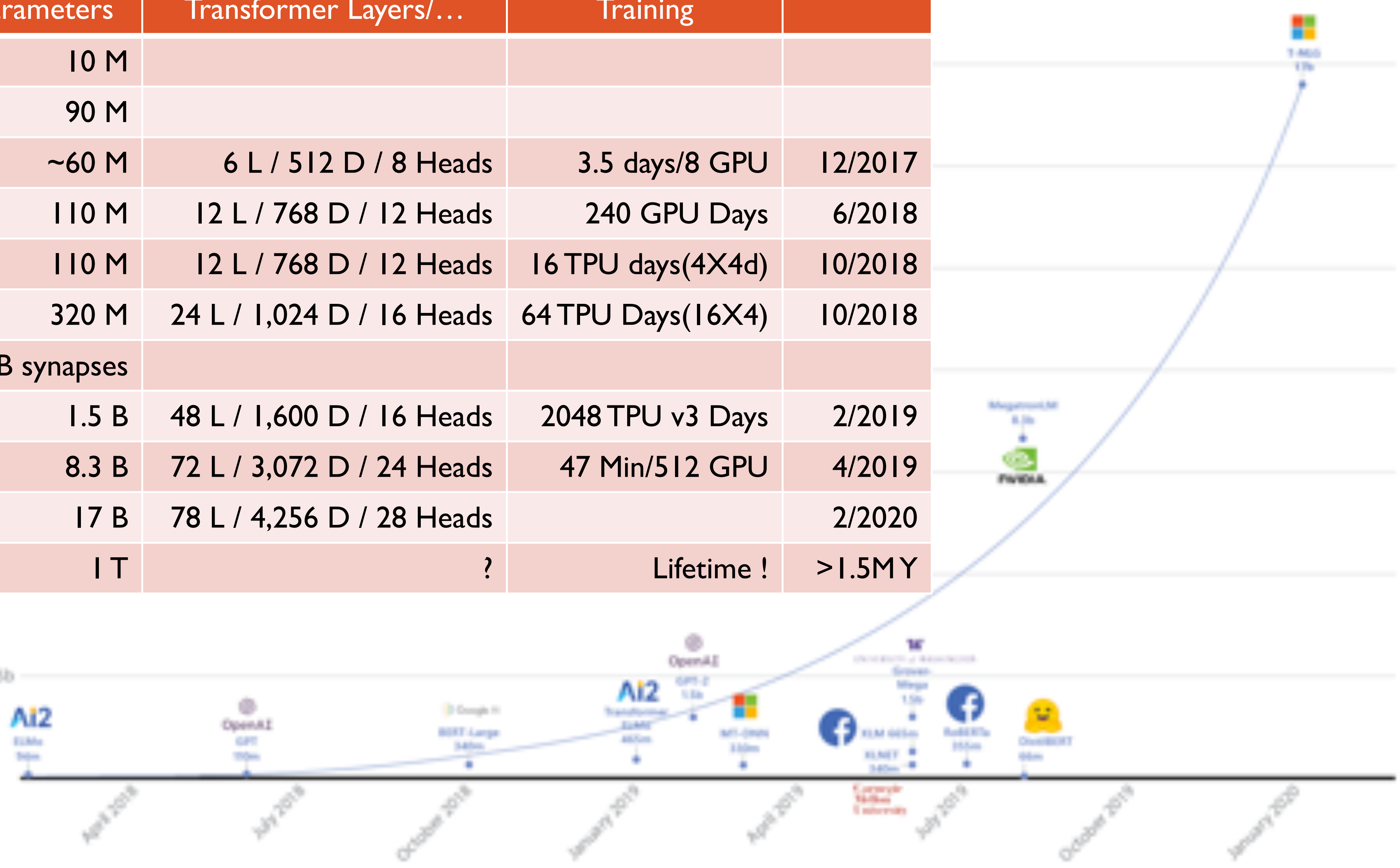
Megatron-LM GPT-2 Evaluation

Model Size	Number of layers	Hidden Size	Attention heads	Wikitext (Perplexity ↓)	Lambada (Accuracy ↑)
345 M	24	1024	16	24.21	55.04
775 M	36	1280	16	20.44	58.86
2.5 B	54	1920	20	17.83	63.30
8.3 B	72	3072	24	17.41	63.11

https://openai.com/blog/gpt-2-1-5b-release/

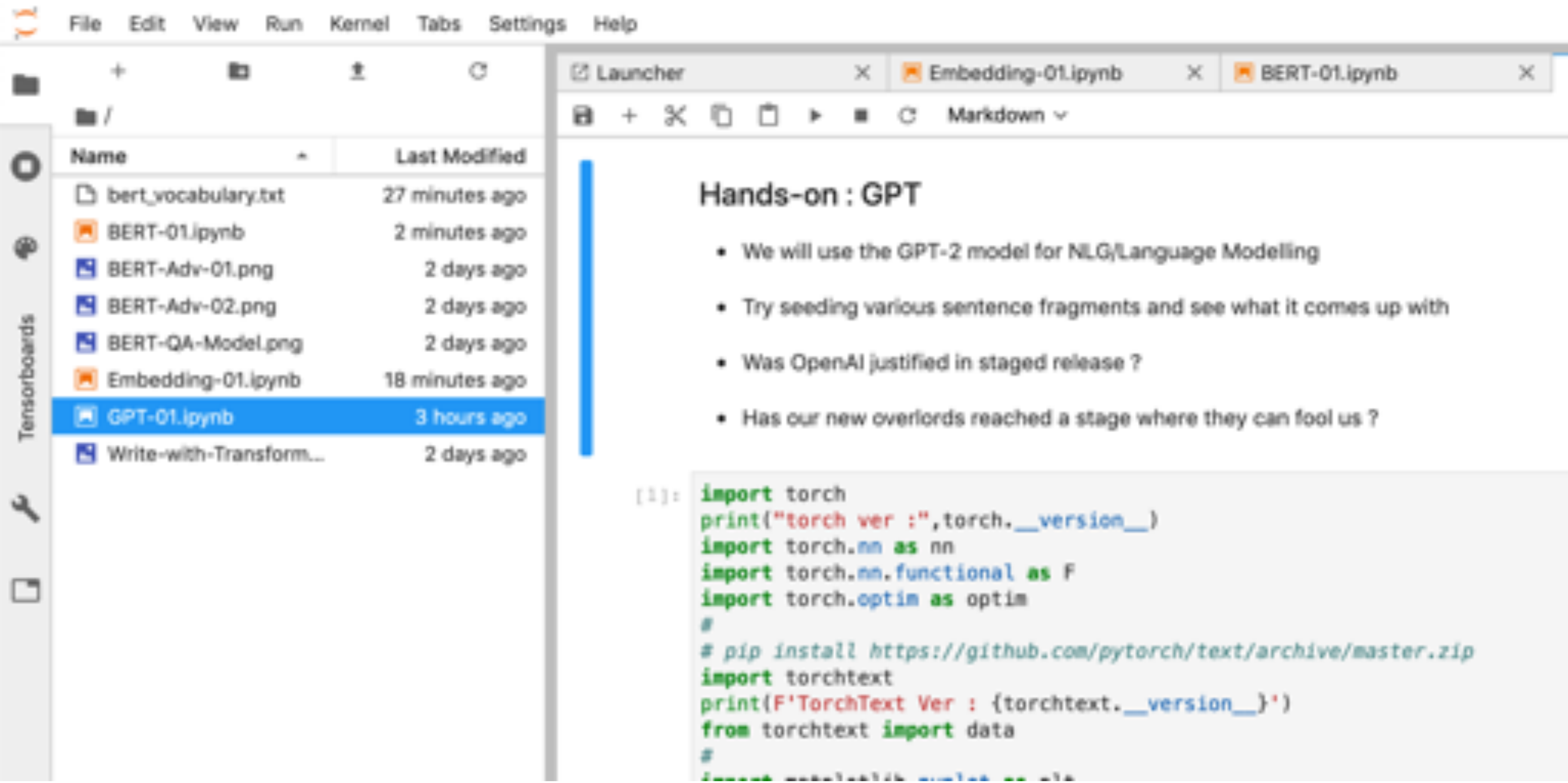
SCALE OF THINGS

Model	Parameters	Transformer Layers/…	Training	
Medium-sized LSTM	10 M			
ELMo	90 M			
Transformers	~60 M	6 L / 512 D / 8 Heads	3.5 days/8 GPU	12/2017
GPT	110 M	12 L / 768 D / 12 Heads	240 GPU Days	6/2018
BERT(Base)	110 M	12 L / 768 D / 12 Heads	16 TPU days(4X4d)	10/2018
BERT (L)	320 M	24 L / 1,024 D / 16 Heads	64 TPU Days(16X4)	10/2018
Honeybee Brain	~1B synapses			
GPT-2 (XL)	1.5 B	48 L / 1,600 D / 16 Heads	2048 TPU v3 Days	2/2019
Megatron GPT-2	8.3 B	72 L / 3,072 D / 24 Heads	47 Min/512 GPU	4/2019
Turing-NLG	17 B	78 L / 4,256 D / 28 Heads		2/2020
Human Brain Models	1 T	?	Lifetime !	>1.5M Y



- *GPT-2 : Really big Transformer LM trained on 40GB text*
- *Model > 1.3B won't fit in GPU, so need model parallelization and broken across multiple GPUs*
- *D = Model Dimensionality (d_{model})*

HANDS-ON : GPT


File Edit View Run Kernel Tabs Settings Help
Launcher
Embedding-01.ipynb
BERT-01.ipynb
Markdown
/
Name
Last Modified
bert_vocabulary.txt 27 minutes ago
BERT-01.ipynb 2 minutes ago
BERT-Adv-01.png 2 days ago
BERT-Adv-02.png 2 days ago
BERT-QA-Model.png 2 days ago
Embedding-01.ipynb 18 minutes ago
GPT-01.ipynb 3 hours ago
Write-with-Transform... 2 days ago
Tensorboards
Hands-on : GPT
We will use the GPT-2 model for NLG/Language Modelling
Try seeding various sentence fragments and see what it comes up with
Was OpenAI justified in staged release ?
Has our new overlords reached a stage where they can fool us ?
[1]:
import torch
print("torch ver :",torch.__version__)
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
#
# pip install https://github.com/pytorch/text/archive/master.zip
import torchtext
print(F'TorchText Ver : {torchtext.__version__}')
from torchtext import data
#

LOAD GPT MODEL

```
print(gpt2_tokenizer.vocab_size)
input_text = gpt2_tokenizer.encode("We like Unicorns because they")
print(gpt2_tokenizer.convert_ids_to_tokens(input_text))
input_text = gpt2_tokenizer.encode("Here is the sentence I want embeddings for.")
print(gpt2_tokenizer.convert_ids_to_tokens(input_text))
```

```
50257
['We', 'Ġlike', 'ĠUnic', 'orns', 'Ġbecause', 'Ġthey']
['Here', 'Ġis', 'Ġthe', 'Ġsentence', 'ĠI', 'Ġwant', 'Ġembed', 'd', 'ings', 'Ġfor', '.']
```

GPT-2 LANGUAGE MODEL

```
in_text = "We like Unicorns because they"
out = run_gpt2_sentence_prediction(in_text)
print(out)
```

```
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:6: UserWarning: Implicit dimension choice for softmax has been
deprecated. Change the call to include dim=X as an argument.
```

```
We like Unicorns because they love to take on us. They know how we feel so nothing's too big for them. We gained some offen
sive experience early on and used it extensively. Who better to gain experience and battle against than Clive? We knew it w
as our only competition and
```

- Run 2/21/20 7:00 PM

'We like Unicorns because they are committed to building the most progressive platform so no major issues are going to come up again.

The race against Unicorns has become a proving ground for champion performance in eSports, and the local leagues have found greater strength.

Leaguepedia'

GPT-2 LANGUAGE MODEL

```
in_text = "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."
out = run_gpt2_sentence_prediction(in_text,100)
print(out)
```

```
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:6: UserWarning: Implicit dimension choice for softmax has been
deprecated. Change the call to include dim=X as an argument.
```

```
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the And
es Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. In fact, in the
late 1800s, Gerald Wallace traveled in the same herd, and within a few hundred years it seemed that one of his ETs had alre
ady established lineage.

Alas, their story is still shocking. Only recently, further research conducted comes to light that would confirm their theo
ries. One Swiss scientist, Marcelo Ferrer, claims that there are eight hyperaanenosalluanensis, which indeed are aware of t
he importance of language for an extraterrestrial intelligence. In other
```

GPT-2 LANGUAGE MODEL

```
in_text = "You're in a desert, walking along in the sand, when all of a sudden you look down and see a tortoise,\
Leon. It's crawling toward you. You reach down, you flip the tortoise over on its back.\
The tortoise lays on its back, its belly baking in the hot sun, beating its legs trying to\
turn itself over,but it can't, not without your help. But you're not helping. Why is that?"
out = run_gpt2_sentence_prediction(in_text,200)
print(out)
```

```
deprecated. Change the call to include dim=X as an argument.
```

```
You're in a desert, walking along in the sand, when all of a sudden you look down and see a tortoise,Leon. It's crawling to
ward you. You reach down, you flip the tortoise over on its back.The tortoise lays on its back, its belly baking in the hot
sun, beating its legs trying toturn itself over,but it can't, not without your help. But you're not helping. Why is that? D
o you have to deal with this bad thing or the tortoise?' Glenn crosses your knees?,you think. 'June."The tortoise temporari
ly relaxes. On your back it looks and thinks, and this is just what it needs to do'she's getting to you. Andrea slowly perc
hes on top of you. "That youngster was kind of harsh, too," she says. "The American folk are acculturated to hunting, but a
nimals are stupid when they think of humans. Now he's taken a taste of me. He starts sniffing the man."This time the tortoi
se lowers its head, and its body moves unhurt, and then lowers its head: what a positivity. "Look out for him when he gets
airborne again," says Andrea. It's done. The picture drags on,giggles and frosts, and Andrea shoots a,rosy smile. She point
s to the picture tattoos, and the visitor who looks at it recalls
```

YOU TOO CAN AUTHOR BOOKS !

- With a little help from our friendly Transformers !
- https://transformer.huggingface.co/

LAST LAYER FOR LM TASK

```
Layer transformer.h.11.mlp.c_proj.weight  torch.Size([3072, 768])       elements : 2,359,296
Layer transformer.h.11.mlp.c_proj.bias    torch.Size([768])             elements : 768
Layer transformer.ln_f.weight             torch.Size([768])             elements : 768
Layer transformer.ln_f.bias               torch.Size([768])             elements : 768
Layer lm_head.weight                      torch.Size([50257, 768])      elements : 38,597,376
==============================================================================================
Total Params:175,620,096
```

```
        )
      )
    )
    (ln_f): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  )
  (lm_head): Linear(in_features=768, out_features=50257, bias=False)
)
```

QUIZ !!

- Can BERT be used for sentence generating tasks ?
- If so, why so ?
- If not, why not ? And what architecture can we use ?

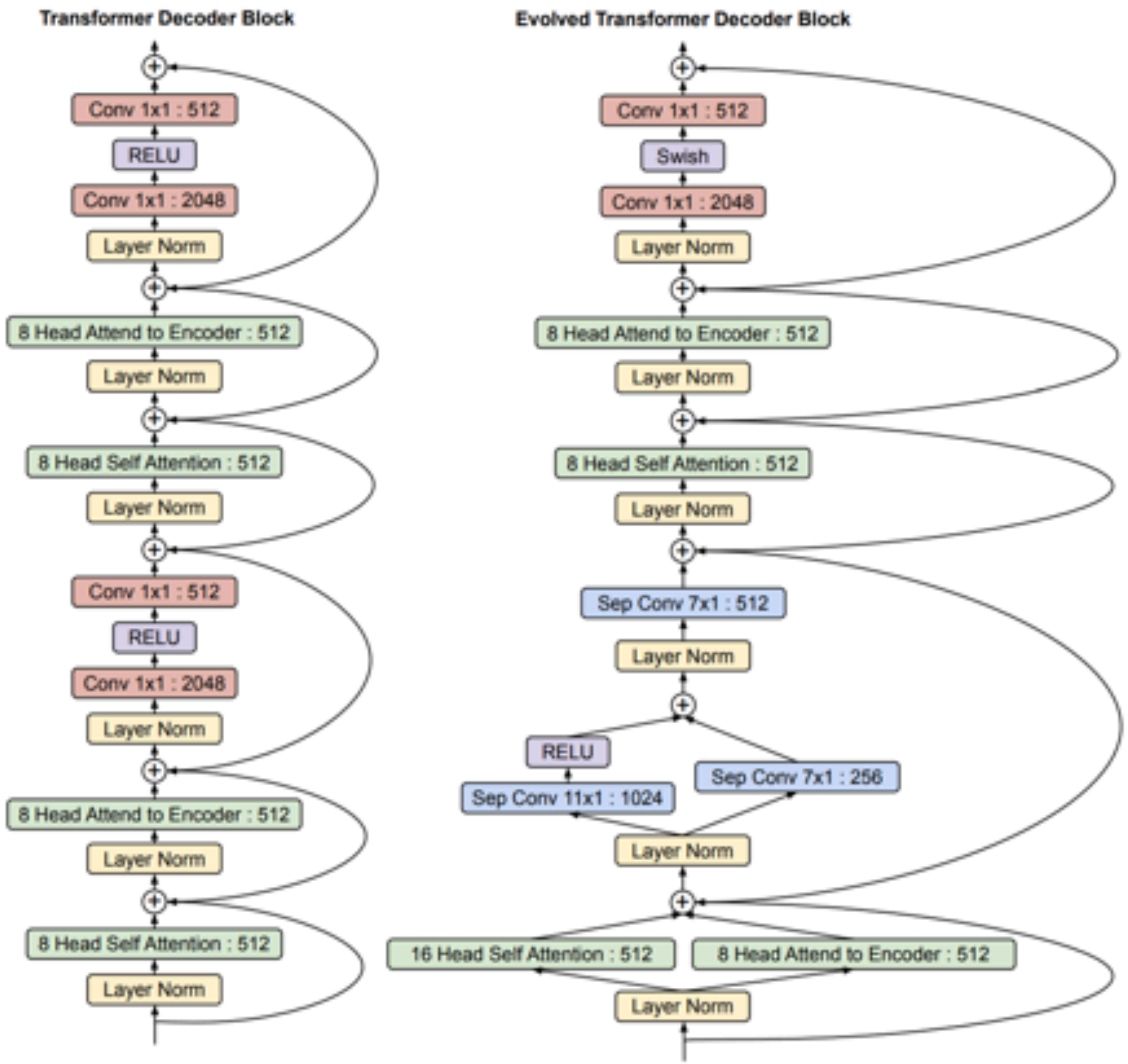
CONVERSATIONAL AI – MEENA CHATBOT

EVOLVED TRANSFORMER

- Discovered via Neural Architecture Search (NAS)
  - The Progressive Dynamic Hurdles (PDH) method
  - It allows the evolution algorithm to dynamically select new promising candidates as the search progresses
  - Search space $7.30 * 10^{115}$

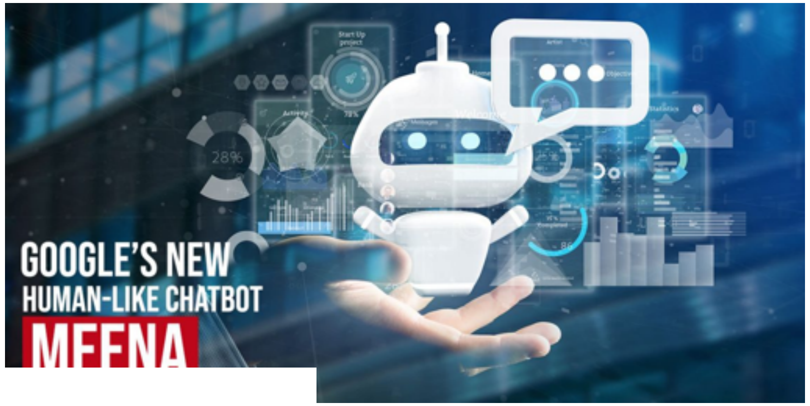
EVOLVED TRANSFORMER

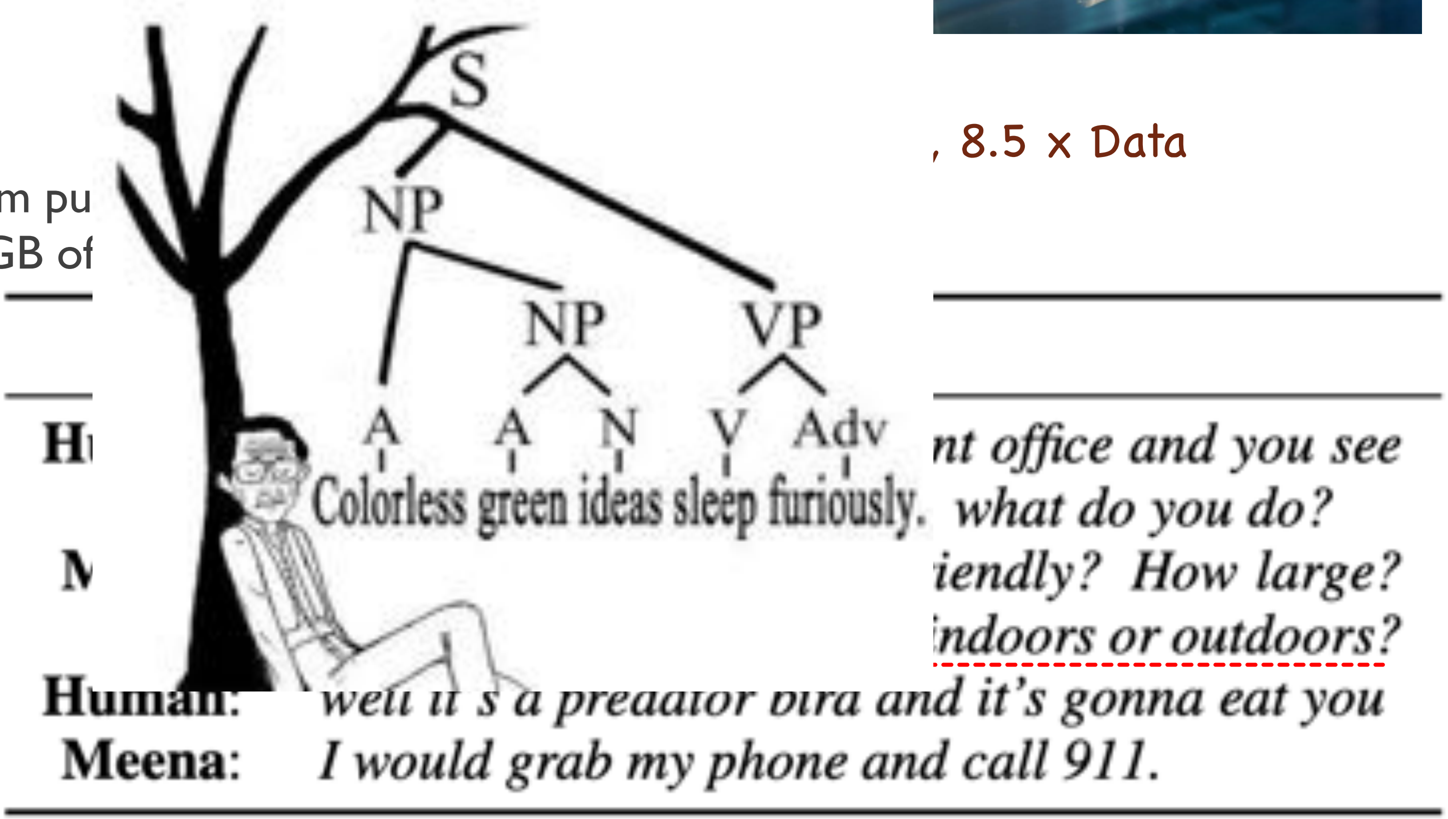
- Hybrid Architecture
  - Use of CNNs
  - Leverages self-attention and wide convolution
- D_{model} still 512
- 16 head in the Decoder side
- Lots of residual connections



MEENA

- Multi-turn, open-domain, end-to-end chatbot
- New evaluation metric called *Sensibleness and Specificity Average (SSA)*, which captures key elements of a human-like multi-turn conversation
  - Make Sense as well as Be Specific
  - Humans 86%, Meena 72%
- Trained on 341GB of text/40 B words from pu
  conversations (GPT-2 was trained on 40 GB of
- Trained for 30 days on a
  - TPUv3 Pod (2,048 TPU cores)
- 1 ET Encoder & 13 ET Decoder blocks
  - Encoder – what is said
  - Decoder – formulate response
- 2,560 D / 32 Heads / 2.6 B parameters
- BPE Embedding





https://www.technewstoday.com/google-introduces-meena-a-chatbot-that-is-the-most-human-like-ever/
https://towardsdatascience.com/inside-the-machine-learning-that-google-used-to-build-meena-a-chatbot-that-can-chat-about-anything-32e4d2242f79

MEENA

- Multi-turn, open-domain, end-to-end chatbot
- New evaluation metric called *Sensibleness and Specificity Average (SSA)*, which captures key elements of a human-like multi-turn conversation
  - Make Sense as well as Be Specific
  - Humans 86%, Meena 72%
- Trained on 341GB of text/40 B words from public domain social media conversations (GPT-2 was trained on 40 GB of text)
- Trained for 30 days on a
  - TPUv3 Pod (2,048 TPU cores)
- 1 ET Encoder & 13 ET Decoder blocks
  - Encoder – what is said
  - Decoder – formulate response
- 2,560 D / 32 Heads / 2.6 B parameters
- BPE Embedding

1.7 x GPT-2, 8.5 x Data

Conversation A	
Human:	*say you're in an empty giant office and you see a black bird flying indoors. what do you do?*
Meena:	*What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?*
Human:	*well it's a predator bird and it's gonna eat you*
Meena:	*I would grab my phone and call 911.*

https://www.technewstoday.com/google-introduces-meena-a-chatbot-that-is-the-most-human-like-ever/
https://towardsdatascience.com/inside-the-machine-learning-that-google-used-to-build-meena-a-chatbot-that-can-chat-about-anything-32e4d2242f79

MEENA

- Multi-turn, open-domain, end-to-end chatbot
- New evaluation metric called *Sensibleness and Specificity Average (SSA)*, which captures key elements of a human-like multi-turn conversation
  - Humans 86%, Meena 72%
- Trained on 341GB of text/40 B words from public domain social media conversations (GPT-2 was trained on 40 GB of text)

DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	**70.70%**	63.7%	92%+

- 2,560 D / 32 Heads
- 2.6 B parameters
- BPE Embedding

a black bird flying indoors. What do you do?

Meena: *What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?*

Human: *well it's a predator bird and it's gonna eat you*

Meena: *I would grab my phone and call 911.*

https://www.technewstoday.com/google-introduces-meena-a-chatbot-that-is-the-most-human-like-ever/

MEENA

- Multi-turn, open-dc
- New evaluation me
  which captures key
  - Humans 86%, Me
- Trained on 341GB
  conversations (GPT

s

TPU cores)

3 ET Decoder blocks

ls

Unlike GPT-2 and
the Japanese Band,
these are real movies!

Conversation B	
Human:	*Hi!*
Meena:	*Hi! How are you this fine evening?*
Human:	*I'm doing well. it's morning my time!*
Meena:	*Good morning! What are your plans for today?*
Human:	*Get my work done then watch a movie at home in the night*
Meena:	*Sounds fun! What movie?*
Human:	*I haven't decided yet, perhaps an Oscar-nominated one. any suggestion?*
Meena:	*I'm not too familiar with recent Oscar winners, but I do know I wanted to see The Grand Budapest Hotel.*
Human:	*ah, tell me about that movie!*
Meena:	*It's from the same director and writer as Moonrise Kingdom. Wes Anderson is pretty well known for making funny, colorful films. I've heard it's pretty good.*
Human:	*ha! I've searched on the internet and both movies are very colorful!*
Meena:	*I'm looking forward to it. I love colorful movies.*

MEENA

- Multi-turn, open-domain, end-to-end chatbot
- New evaluation metric called *Sensibleness and Specif* which captures key elements of a human-like multi-t
  - Humans 86%, Meena 72%
- Trained on 341GB of text/40 B words from public d conversations (GPT-2 was trained on 40 GB of text)
- Trained for 30 days on a
  - TPUv3 Pod (2,048 TPU cores)
- 1 ET Encoder & 13 ET Decoder blocks
- 2,560 D / 32 Heads
- 2.6 B parameters
- BPE Embedding

It is not perfect !!

Cross-turn Repetition Example 1	
Human:	*Hi!*
Meena:	*Hi! How are you?! :)*
Human:	*doing good, what would be the best city to visit in Europe?*
Meena:	*Paris, Barcelona, Amsterdam, Prague*
Human:	*great list! why?*
Meena:	*Paris is a beautiful city, and Barcelona is too. And I've always wanted to go to Amsterdam and Prague and have never been.*
Human:	*have you even been to India?*
Meena:	*No, but I'd love to go.*
Human:	*how about Japan?*
Meena:	*I'd love to go to Japan too.*
Human:	*how about Paris?*
Meena:	*I'd love to go to Paris too.*
Human:	*are you lying to me?*
Meena:	*I'd love to go to Japan too.*
Human:	*i think you're lying to me, that makes me sad*
Meena:	*I'd love to go to Japan, too.*
Human:	*you're blocking me out! you have to talk to me*
Meena:	*I'd love to go to Japan.*
Human:	*where else?*
Meena:	*I'd love to go to Japan, too!*

4. ONWARD …

The New York Times

Electronic 'Brain' Teaches Itself

July 13, 1958

The Navy last week demonstrated the embryo of an electronic computer named the Perceptron which, when completed in about a year, is expected to be the first non-living mechanism able to "perceive, recognize and identify its surroundings without human training or control." VIEW FULL ARTICLE IN TIMESMACHINE »

Pedro Domingos @pmddomingos · 6h

If you time-traveled back to a decade ago and told Nature's editors that one day they'd publish a cover story about computers playing video games, they'd fall down laughing.

May 20, 1991

THE Sun

Monday, May 20, 1991

25p

World Wide What?

COMPUTER 'WEB' TO CHANGE BILLIONS OF LIVES (YEAH, RIGHT)

A BRITISH computer geek's brainwave could be one of the greatest inventions ever, it was claimed last night.

By DOT COMME

Tim Berners-Lee, 35, has enabled computer users to see documents and pictures made available by others in "cyberspace".

He uses the "Internet" system, which so far only links academics but could eventually include anyone.

Berners-Lee, who works at a nuclear research base near Geneva, calls his idea the "World Wide Web".

One scientist said: "This could be huge. The idea of strangers worldwide sharing ideas instantly is mind-boggling." But another sneered: "They said Sinclair's C5 would change the world. Now you'd struggle to give one away."

Riddle of 'E' mail – Page 8

Web feat . . . Berners-Lee

COMPUTER 'WEB' TO CHANGE BILLIONS OF LIVES (YEAH, RIGHT)

A BRITISH computer geek's brainwave could be one of the greatest inventions ever, it was claimed last night.

By DOT COMME

Tim Berners-Lee, 35, has enabled computer users to see documents and pictures made available by others in "cyberspace".

He uses the "Internet" system, which so far only links academics but could eventually include anyone.

Berners-Lee, who works at a nuclear research base near Geneva, calls his idea the "World Wide Web".

One scientist said: "This could be huge. The idea of strangers worldwide sharing ideas instantly is mind-boggling." But another sneered: "They said Sinclair's C5 would change the world. Now you'd struggle to give one away."

Riddle of 'E' mail – Page 8

Web feat . . . Berners-Lee


5 MB Hard Disk
$200,000 to rent
for *a month* !

Amanda Carpenter @amand... 2/14/16

Guys. I'm trapped in an amtrak elevator at Bwi airport. Help?

27 50

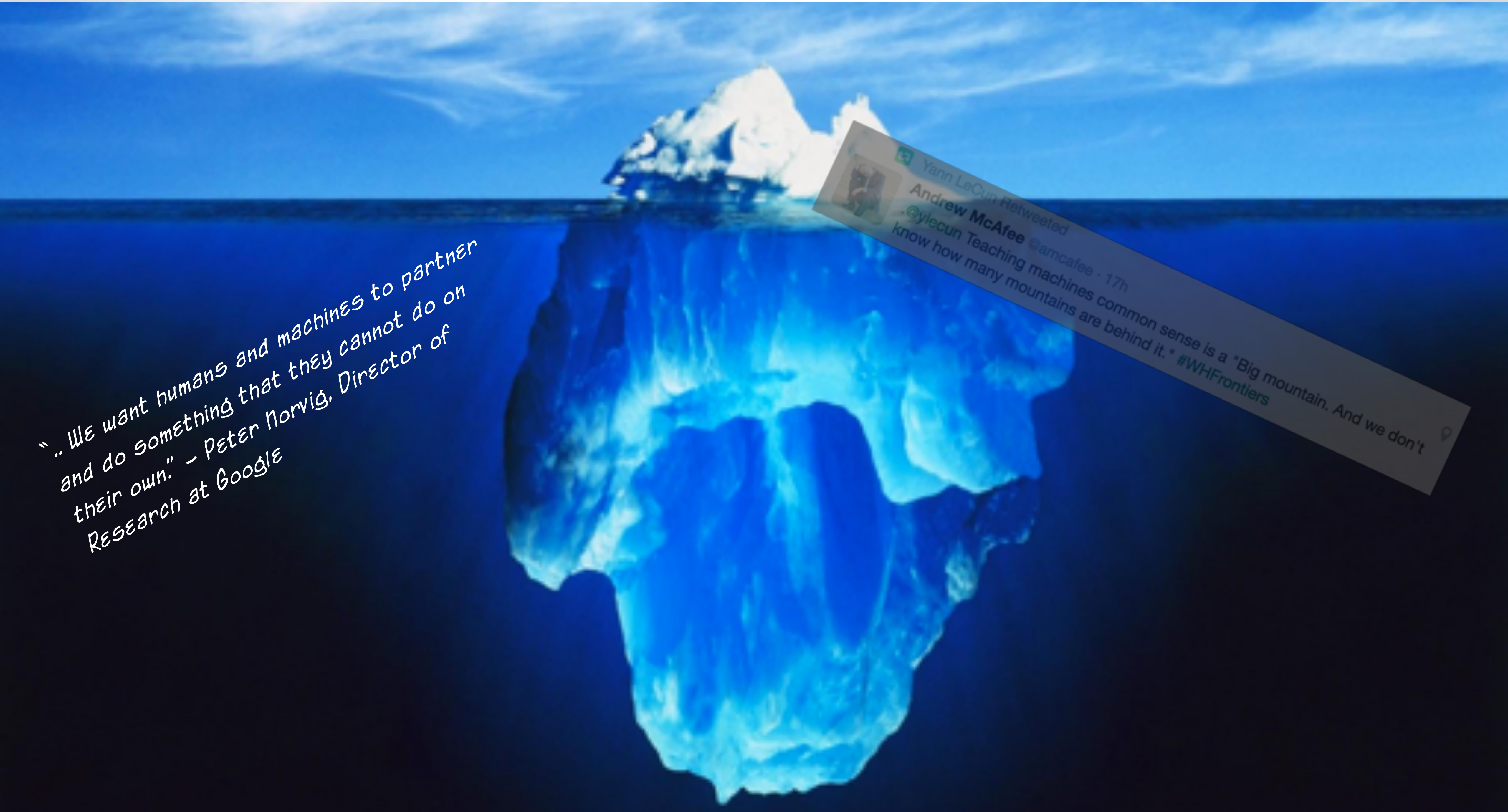
Amtrak

@Amtrak

@amandacarpenter We are sorry to hear that. Are you still in the elevator?

9/7/16, 9:48 AM

"I wanted to build cars that were not something to everyone, but everything to some." – Dr. Ferry Porsche


".. We want humans and machines to partner and do something that they cannot do on their own." - Peter Norvig, Director of Research at Google
Yann LeCun Retweeted
Andrew McAfee @amcafee · 17h
.@ylecun Teaching machines common sense is a "Big mountain. And we don't know how many mountains are behind it." #WHFrontiers

What's Wrong ?

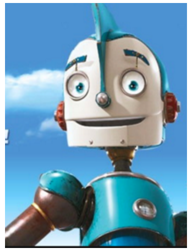
- Can submarines swim ?
- Cars don't have legs, but they are fast
- Planes don't have floppy wings, but they do fly
- …
- Are we conditioned ? And if so, can the machines be conditioned ?

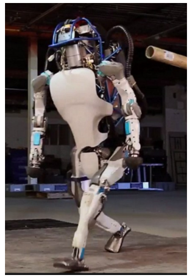
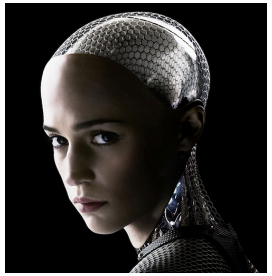
Pedro Domingos
@pmddomingos

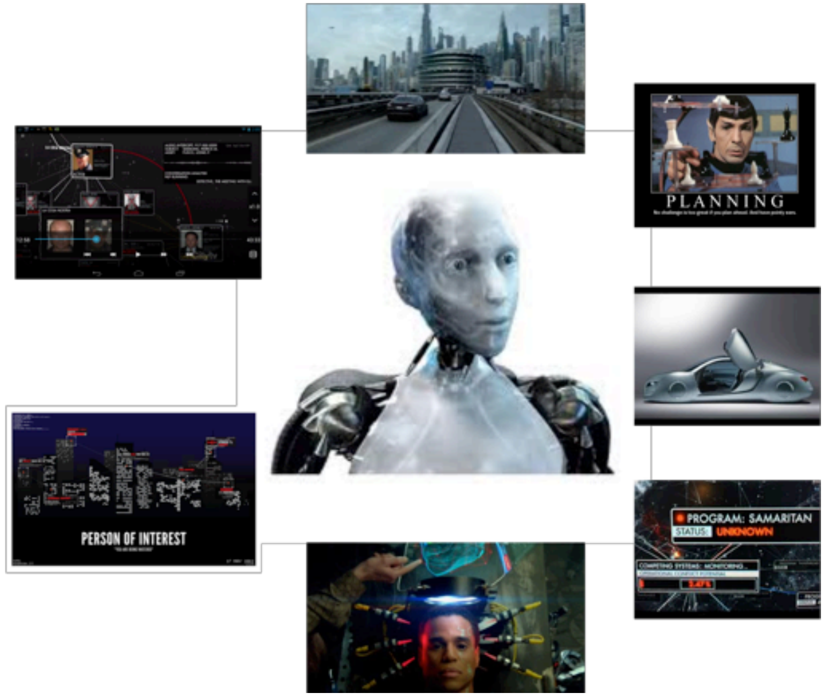
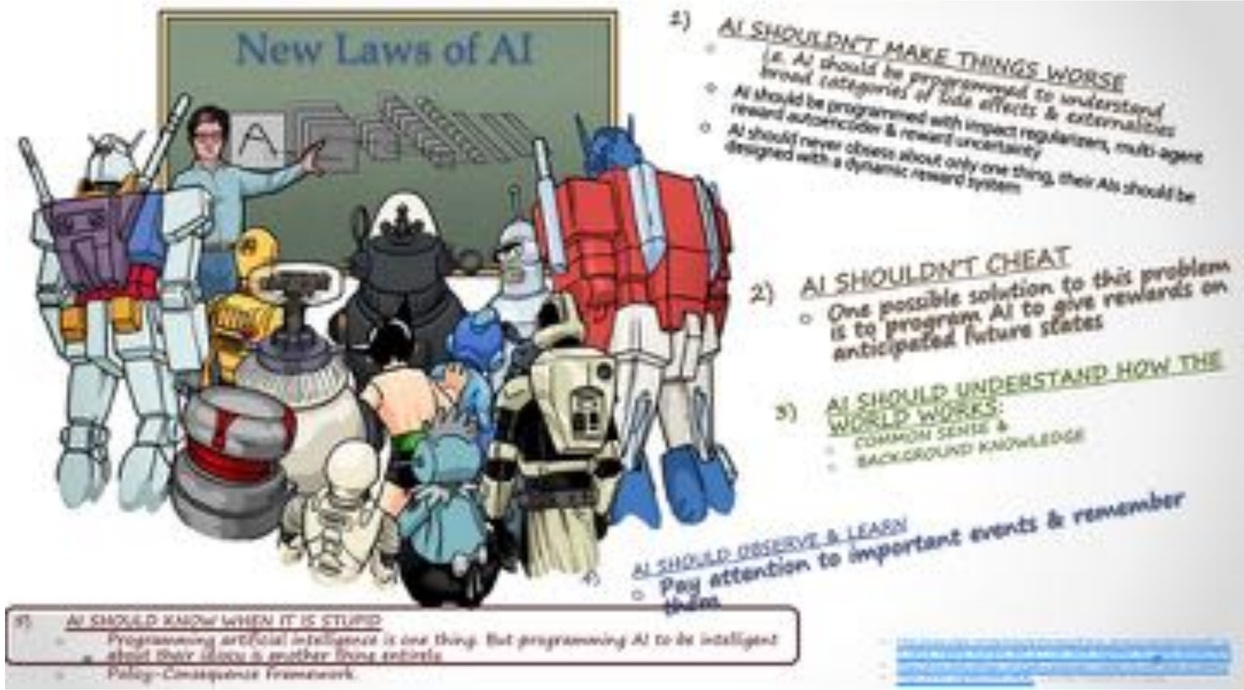
Following

The Turing test is like saying planes don't fly unless they can fool birds into thinking they're birds. (Peter Norvig)

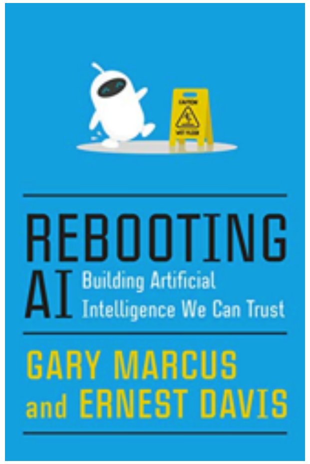
https://medium.com/ai2-blog/how-to-get-up-to-speed-on-machine-learning-and-ai-a0fd923d4169



CAUTION!
WET FLOOR
REBOOTING AI Building Artificial Intelligence We Can Trust
GARY MARCUS and ERNEST DAVIS
REBOOTING
AI Building Artificial Intelligence We Can Trust

PLANNING
PERSON OF INTEREST
PROGRAM: SAMARITAN
STATUS: UNKNOWN

New Laws of AI
1) AI SHOULDN'T MAKE THINGS WORSE
2) AI SHOULDN'T CHEAT
One possible solution to this problem is to program AI to give rewards on anticipated future states
3) AI SHOULD UNDERSTAND HOW THE WORLD WORKS:
COMMON SENSE &
BACKGROUND KNOWLEDGE
AI SHOULD OBSERVE & LEARN
Pay attention to important events & remember them
AI SHOULD KNOW WHEN IT IS STUPID

The New York Times



Opinion

How to Build Artificial Intelligence We Can Trust

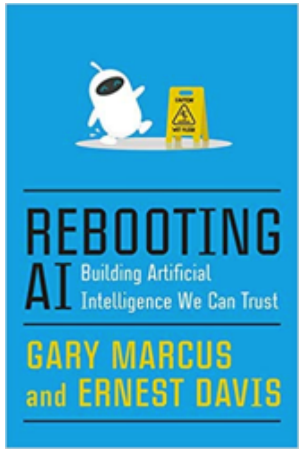
Computer systems need to understand time, space and causality. Right now they don't.

By Gary Marcus and Ernest Davis

Sep 6,2019

- The problem is not that today's A.I. needs to get better at what it does. The problem is that today's A.I. needs to try to do something completely different.

- We need to stop building computer systems that merely get better and better at detecting statistical patterns in data sets — often using an approach known as deep learning — and start building computer systems that …
  - *from the moment of their assembly innately grasp three basic concepts: **time, space and causality** - conceptual framework*
- For certain A.I. tasks, the dominant data-correlation approach works fine
- For example, AI can detect a shape; But no existing AI can properly understand how the shape of an object is related to its function



“Trustworthy AI, grounded in reasoning, commonsense values, and sound engineering practice, will be transformational when it finally arrives, whether that is a decade or a century hence … ”

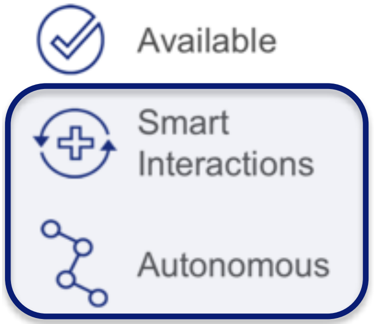
Ordinary objects that pose challenges for AI

worth a read (or two) and some deliberation is in order. The clues AI can learn from other disciplines include “***Cognition makes extensive use of Internal Representations***”, “***Cognitive systems are highly structured***”, “***Human thoughts and language are highly compositional***”, “***Concepts are embedded in theories***” and “***Casual relations are a fundamental aspect of understanding the world***”.

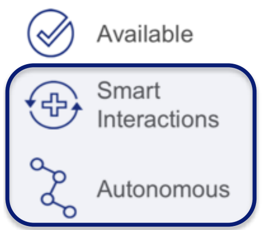
I agree wholeheartedly with Gary Marcus and Ernest Davis that “***If we could give computers one gift that they don’t already have, it would be the gift of understanding language***”

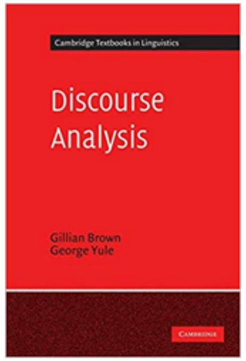
https://medium.com/@ksankar/rebooting-ai-by-garymarcus-and-ernestdavis-nyuniversity-book-review-915ad0de0510

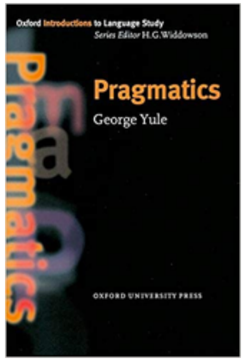
Features	***Themes***	***Levels***
Limited Vocabulary, Limited actions e.g. lights on	Commands & Notifications	Notification (& IoT Command) Bots
Search & traverse through text, API based intent resolution	Q & A	FAQ Bots
Knowledge Graph (explicit or implicit intent based), some level of comprehension & capable of contextual conversation	Dialogues	Contextual Bots
Inferential, Predictive, Personalized and Branded for unique & differentiated digital experience	Interactions	Smart Personalized Bots
Human-level comprehension & common sense, Factoids, Knowledge Graph and capable of intelligent interactivity over multiple conversation chains	Conversations	Autonomous Bots

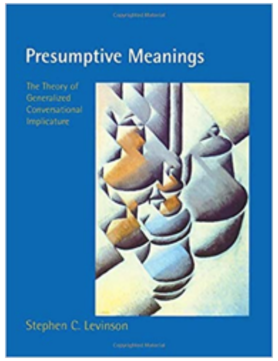


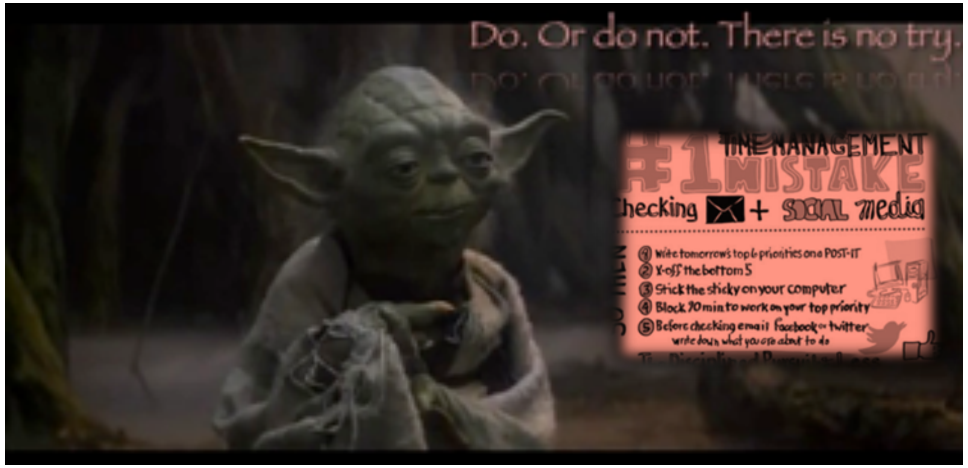
Features	Themes	Levels
Limited Vocabulary, Limited actions e.g. lights on	Commands & Notifications	Notification (& IoT Command) Bots
Search & traverse through text, API based intent resolution	Q & A	FAQ Bots
Knowledge Graph (explicit or implicit intent based), some level of comprehension & capable of contextual conversation	Dialogues	Contextual Bots
Inferential, Predictive, Personalized and Branded for unique & differentiated digital experience	Interactions	Smart Personalized Bots
Human-level comprehension & common sense, Factoids, Knowledge Graph and capable of intelligent interactivity over multiple conversation chains	Conversations	Autonomous Bots











COMMONSENSE REASONING

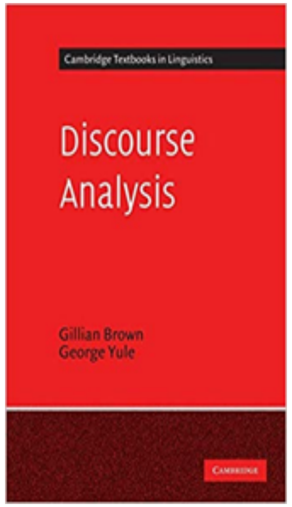
- *A process that involves taking information about certain aspects of a scenario in the world and making inferences about other aspects of the scenario based on our commonsense knowledge*
- *Events, Fluents & TimePoints*
  - *Fluent - represents a time-varying property of the world, such as the location of a physical object*
- *The Commonsense Law of Inertia, The Mental state of agents, Default Reasoning*

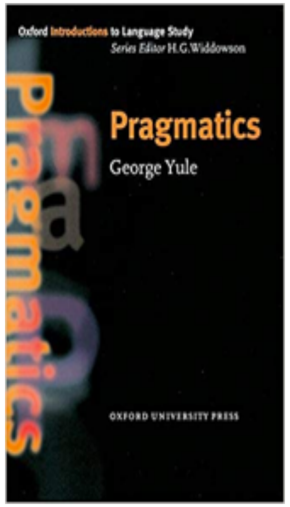
- In the living room, Lisa picked up a newspaper and walked into the kitchen. Where did the newspaper end up? *It ended up in the kitchen.*
- Kate set a book on a coffee table and left the living room. When she returned, the book was gone. What happened to the book? *Someone must have taken it.*
- Jamie walks to the kitchen sink, puts the stopper in the drain, turns on the faucet, and leaves the kitchen. What will happen as a result? *The water level will increase until it reaches the rim of the sink. Then the water will start spilling onto the floor.*
- Kimberly turns on a fan. What will happen? *The fan will start turning.* What if the fan is not plugged in? *Then the fan will not start turning.*
- A hungry cat saw some food on a nearby table. The cat jumped onto a chair near the table. What was the cat about to do? *The cat was about to jump from the chair onto the table in order to eat the food.*

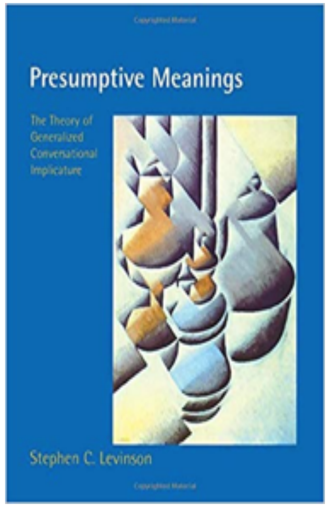
https://medium.com/@ksankar/commonsense-reasoning-the-education-of-a-machine-part-5-7-f2dd25e01de2

CONVERSATION IMPLICATURE

- Sophisticated & robust common sense understanding of the world won't come from pattern matching on examples
- Grounding - Social Cues, Physical Arrangement, Assumptions about the speaker's goals
- More sophisticated reasoning about other agents and their goals
- *System should learn the default stuff from outside the conversation*
- Paul Grice & Gricean reasoning
  - Maxim of quality
  - Maxim of quantity – *As informative as required and not more*
  - Maxim of relation (or relevance) - *Be relevant*
  - Maxim of manner – *Be Brief and orderly*
  - Maxim of Smartness







David Lewis
Philosopher

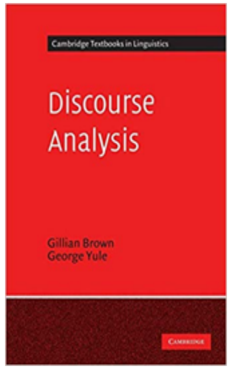
J.L Austin
Linguist

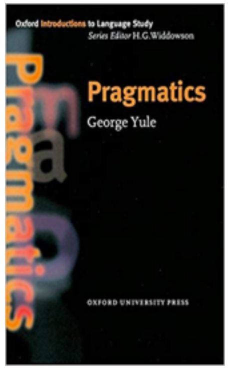
Ref: Stanford : Linguist 130a/230a, cs224u

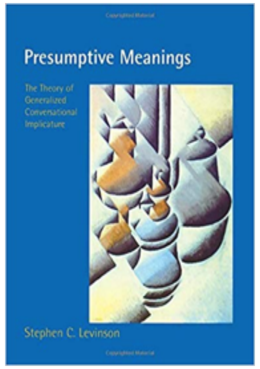
CONVERSATION IMPLICATURE & BERT

- Conversational AI roBots : Pragmas & Practices

1. When thinking about Bots, *think Search*, not a single Question-Answer session. This requires a little deep thought — the bot should want to understand what the user wants, participate in an *iterative collaborative interaction* to facilitate what the user is looking for
2. Model paths other than happy paths and fallback choreographies — *fail gracefully*
3. Bots have a *Visual Branding* as well as a personality – implied or explicit
4. *Aesthetics* (incl small talk) is very important. Don't turn a web site into a bot
5. Surprise your users with *Acuity and Serendipity* !







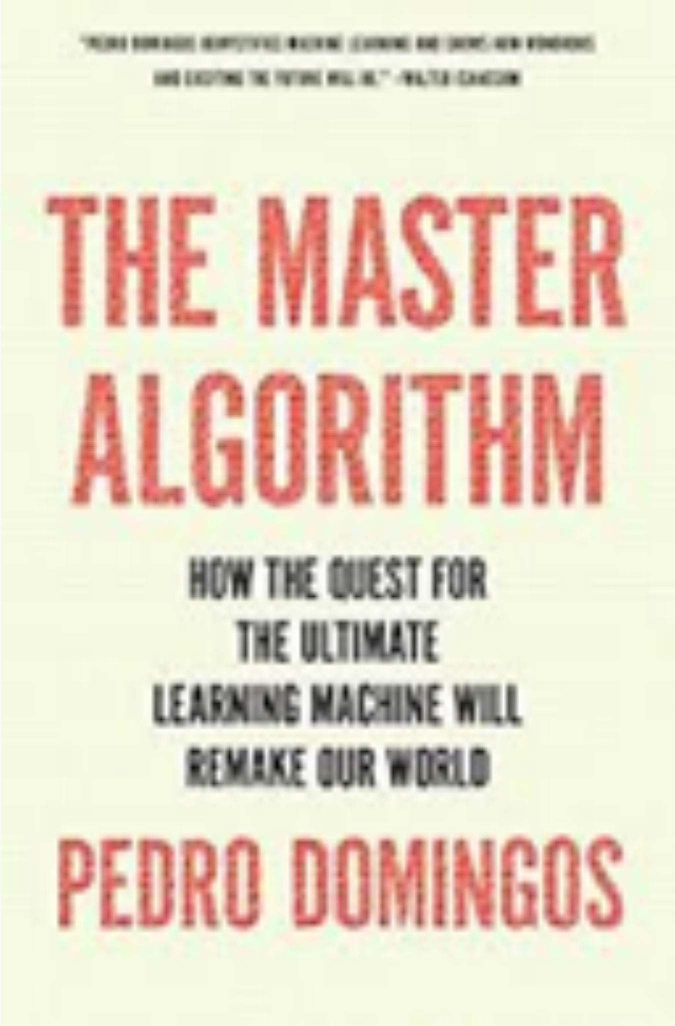
Show up on time, know the text & have a head full of ideas

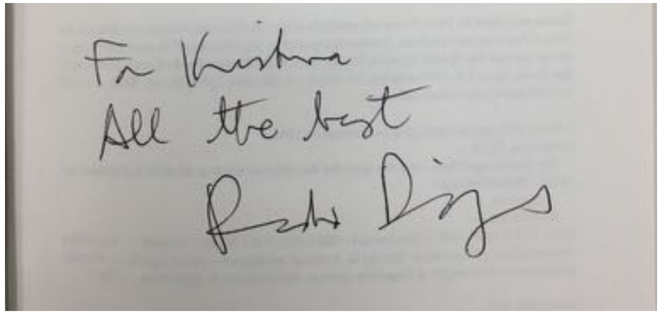
Have the freedom of being there early enough to settle down and gather your thoughts -- because when the time comes, you have to hit the marks …

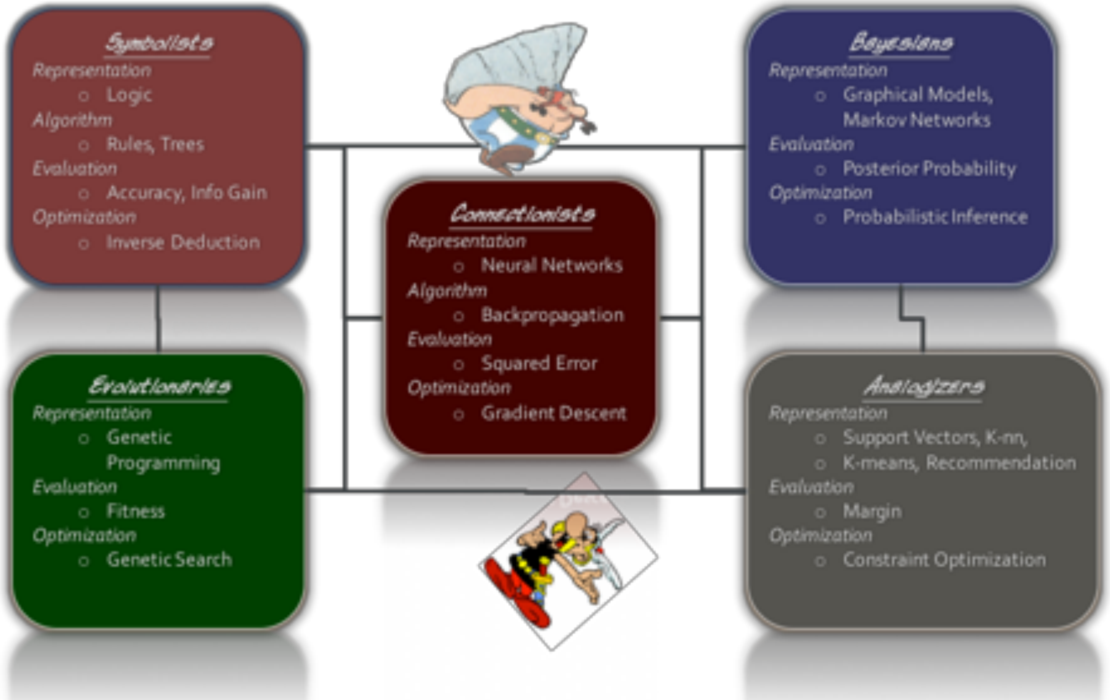
Knowing your text--it's not just your lines, it's the whole thing, … You might not be right in the opinion you bring to it. But you've got to come at it with some direction

The head full of ideas: Bring anything. Try anything. They might not use it. If it stinks, they won't use it. *Am I right, Marty Scorsese?* Growth is a result of being willing to take risks, to break out of your comfort zone, and to embrace failure when it happens

https://medium.com/@ksankar/conversational-ai-robots-pragmas-practices-21-lessons-one-could-learn-watching-the-wise-yoda-d0f81bdee2ea



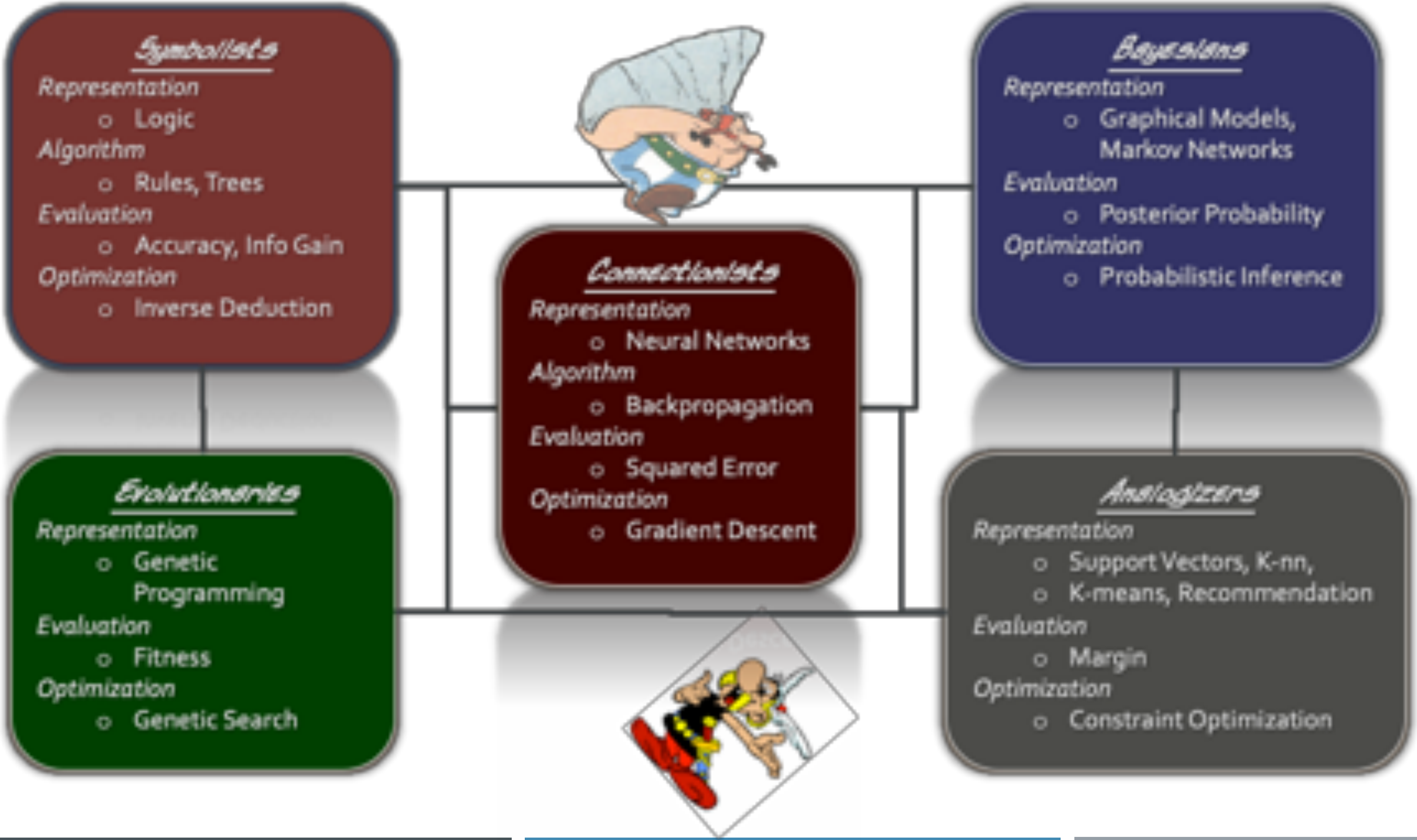




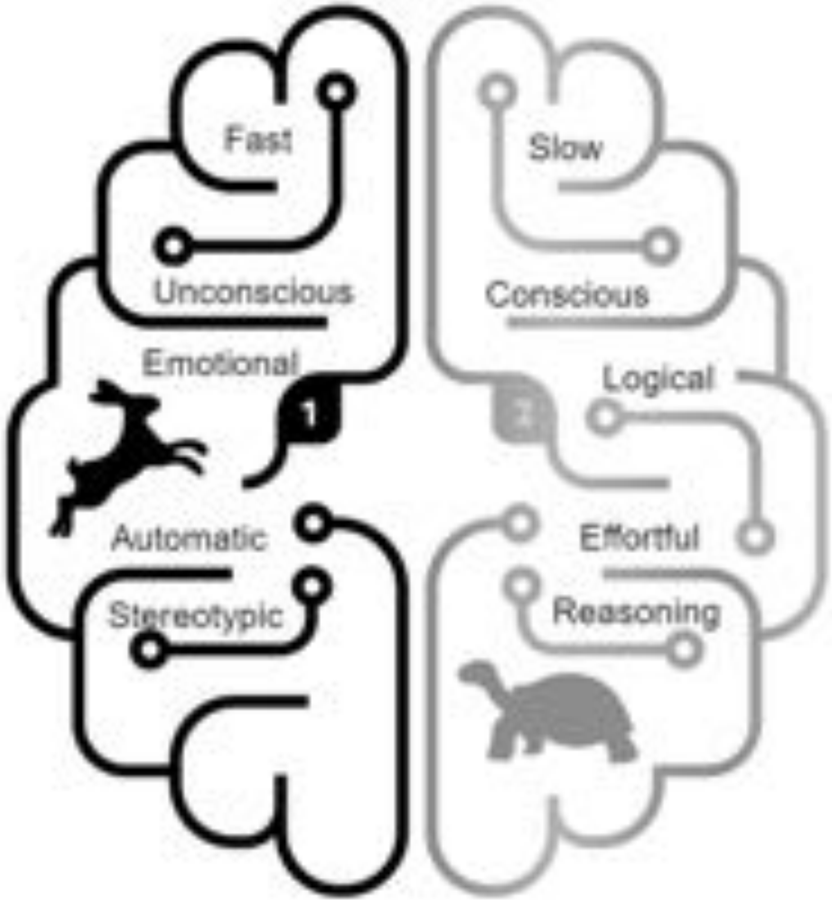
THE MASTER ALGORITHM (BOOK REVIEW) A.K.A DATA THE FINAL FRONTIER

medium.com

https://medium.com/@ksankar/the-master-algorithm-book-review-a-k-a-data-the-final-frontier-458f8fc4bc22

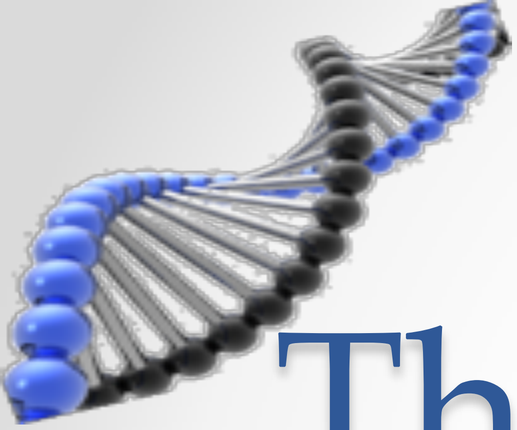

Symbolists
Representation
Logic
Algorithm
Rules, Trees
Evaluation
Accuracy, Info Gain
Optimization
Inverse Deduction
Connectionists
Representation
Neural Networks
Algorithm
Backpropagation
Evaluation
Squared Error
Optimization
Gradient Descent
Bayesians
Representation
Graphical Models, Markov Networks
Evaluation
Posterior Probability
Optimization
Probabilistic Inference
Evolutionaries
Representation
Genetic Programming
Evaluation
Fitness
Optimization
Genetic Search
Analogizers
Representation
Support Vectors, K-nn,
K-means, Recommendation
Evaluation
Margin
Optimization
Constraint Optimization

System 1	System 2
drive a car on highways	drive a car in cities
come up with a good chess move (if you're a chess master)	point your attention towards the clowns at the circus
understands simple sentences	understands law clauses
correlation	causation
hard to explain	easy to explain



Add from Bengio talk

https://towardsdatascience.com/explainable-ai-vs-explaining-ai-part-1-d39ea5053347



Thank You Very Much

For Your

Attention!!