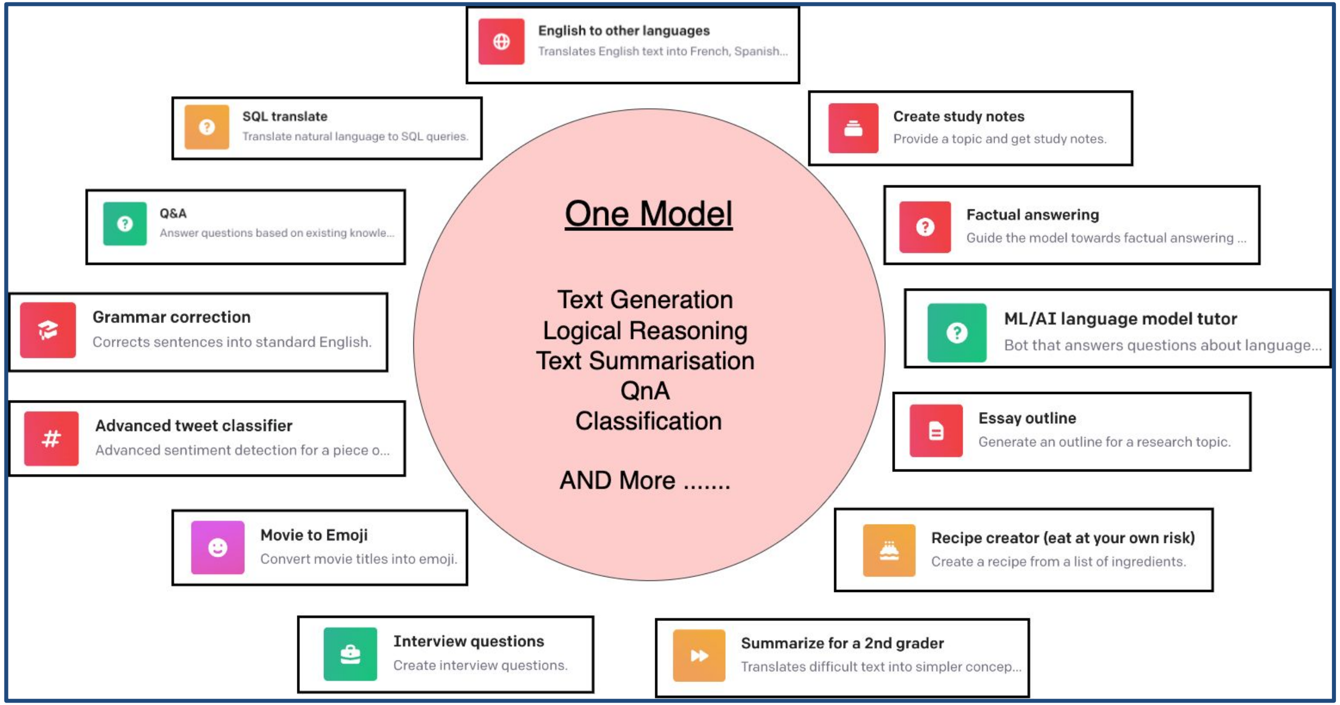
CSCI-B 659 Advanced Natural Language Processing

Language Models are Few-Shot Learners

Hardik Asnani
Anisha Bajaj
Umesh Kumar
Shreyas Vaidya
Arpita Welling

INDIANA UNIVERSITY BLOOMINGTON

One ring to rule them all



SECTION 1

Let’s Play!

- Render HTML Layout
- Arpita's Cooking Recipe

Examples

- **Render HTML output**

https://twitter.com/i/status/1282676454690451457

- **Arpita’s Cooking Recipe**

Give recipe that includes following ingredients:
spaghetti, basil, olive oil

Basil Spaghetti with Olive Oil
Ingredients:

-1 pound spaghetti
-1/2 cup olive oil
-1/4 cup chopped basil
-1/4 cup grated Parmesan cheese
-1/4 teaspoon salt
-1/4 teaspoon black pepper

Directions:

1. Cook spaghetti according to package instructions.

2. In a large bowl, combine cooked spaghetti, olive oil, basil, Parmesan cheese, salt, and black pepper.

Examples

Hardik’s Dilemma:

I missed an important person in the email. What should I do?

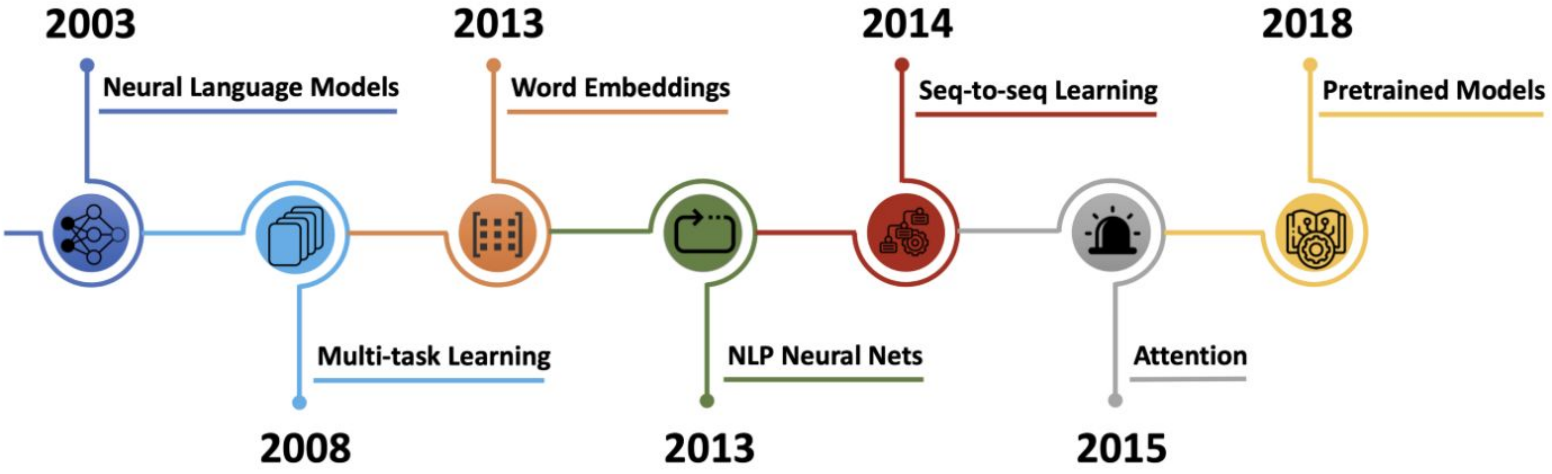
If you realize that you have missed an important person in the email after you have already sent it, you can send a follow-up email to that person apologizing for forgetting to include them in the original email and then include the original email as an attachment.

Activity : Names Meaning Prediction

SECTION 2

Why GPT3?

Language Model Evolution



The big stages of NLP in the deep learning era.

GPT3 v/s the rest

- State of the art architectures are task agnostic but require task-specific data
  - Need Labeled examples – Time intensive hand labeling
  - Difficult to get large supervised train set
  - Repeated for every task
  - Lack of fluidity and scalability
- GPT3 eliminates all of this - few shot learning
- No fine tuning
- Little human supervision

SECTION 3

What’s GPT3?

Some Basics - Language Models

- What - Statistical and probabilistic techniques determine the probability of a given sequence of words
- Where - Linguistic Applications - Speech Recognition, OCR (Optical Character Recognition), code generation, and many more
- Example - Unigrams, Bigrams, Bidirectional Encoder Representation, Transformers Models.

Zero-shot

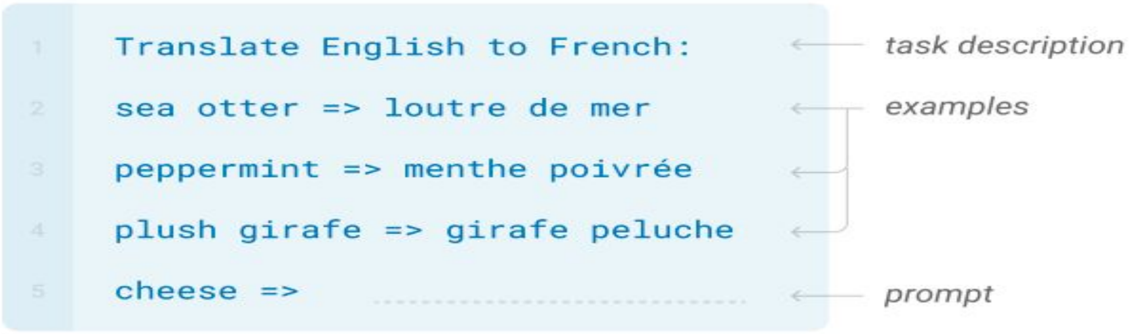
- Requires no examples
- Little Human Intervention
- Models depends on historical trained concepts and data
- Example:

One Shot

- Exactly as the name suggest
- Single example per task description
- Example:

Few Shots

- Provide few examples for the task description to predict the output
- 10 - 100 examples can fit the context window
- Reduction in the need for task-specific data
- Example:



SECTION 4

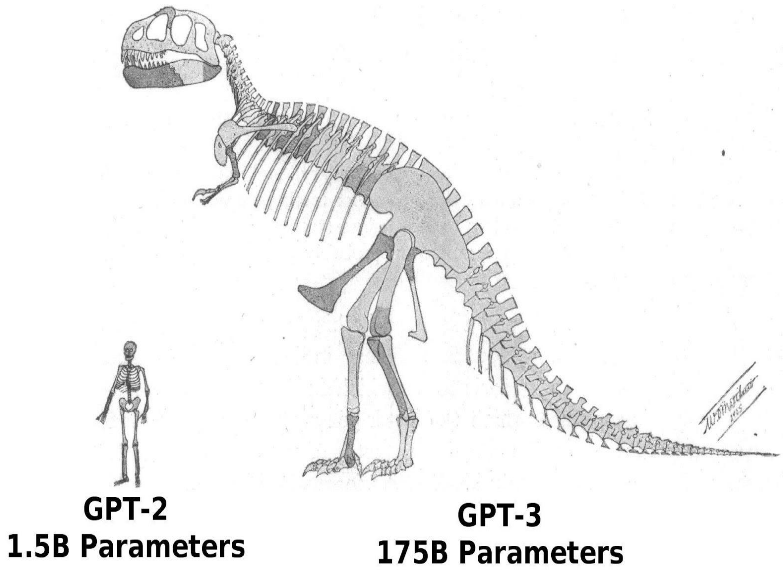
GPT3 Architecture

GPT3

- Generative Pre-trained Transformer 3
- Released in June 2020, claimed to largest one in terms of params and storage
- Autoregressive Language Model
- Meta Learning, In-context Learning
- Uses Deep Learning for Natural Language Tasks (like text classification, machine translation, Q&A)
- Human like Text Output
- Performs tasks using all, Zero Shot, One Shot and Few Shots.

What's 3 doing in GPT3?

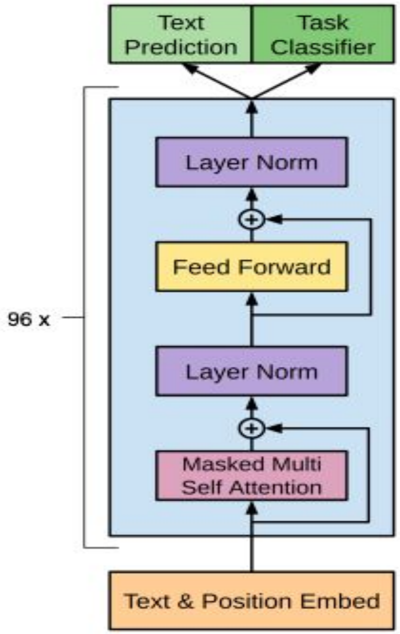
	GPT-1	GPT-2	GPT-3
Parameters	117 Million	1.5 Billion	175 Billion
Decoder Layers	12	48	96
Context Token Size	512	1024	2048
Hidden Layer	768	1600	12288
Batch Size	64	512	3.2M



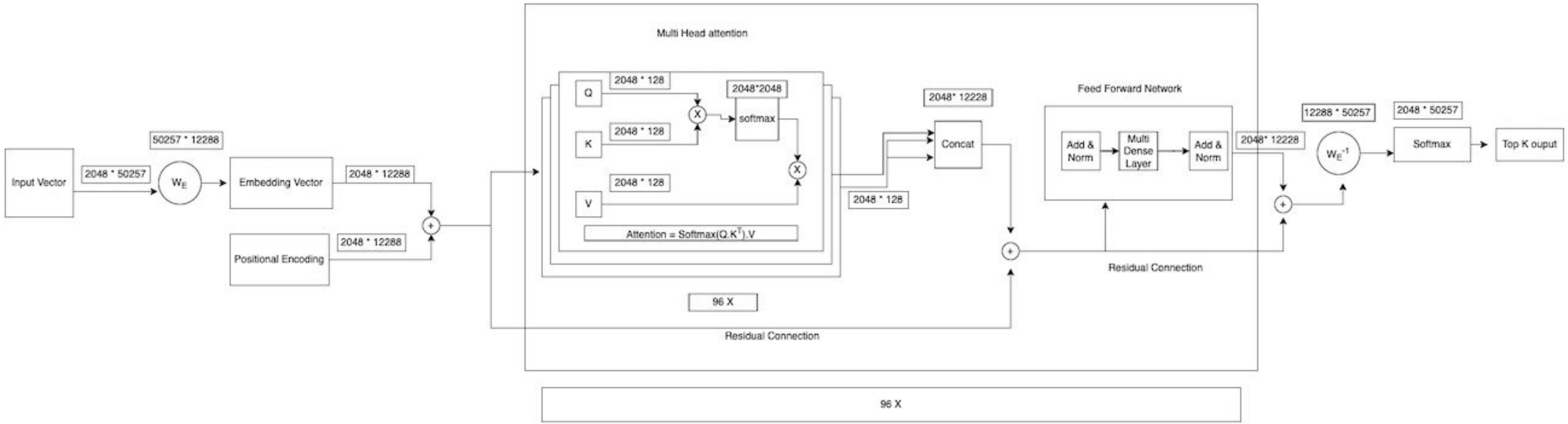
Architecture

GPT 3 Model consists of following layers

- Input Embeddings
- Positional Embeddings
- MultiHead Attention
- Normalisation
- Feed Forward Layers
- Softmax



GPT3 Model Architecture



GPT3 Model

- GPT3 model was trained on vocabulary of 50257 words
- The input sequence is actually fixed to 2048 words.
- If the input is less than 2048 words, it will pad with 0’s.
- The output sequence is also fixed to 2048 words.
- GPT3 uses the embedding dimensions of 12288 dimensions.
- Embedding dimensions are nothing but the features of the input like “softness”, “color”, “past tense”, “numerical” etc..

SECTION 5

Approach

Data

1. Common Crawl Dataset
2. Downloaded from 41 shards of monthly covering from 2016-19
3. 45TB of plain text before filtering, 570GB after that
4. 400 billion byte-paired encoded tokens

Data Preprocessing

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- Filtered based on similarity to range of high-quality corpora
- Fuzzy de-duplication at document level to prevent redundancy
- Sampling and Augmentation - Addition of other datasets – increased diversity
- Higher quality datasets sampled more frequently

Training

- Large models might go out of memory - Parallelism in matrix multiplication
- Used Gradient Noise Scale to determine batch size
- V100 GPU's - high bandwidth cluster by Microsoft
- Adam with cosine decay for Learning Rate

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Evaluation

- Randomly draw K samples from training as conditioning for few-shots
  - Maximum size of k depends on matrix window (2048)
  - Typically fits 0 to 100 examples
- Evaluation is Task specific
- For example, for MCQ’s, K samples of context + completion with 1 of context only. Then compare LM likelihood of each completion
- Another eg, for free-form completion tasks, beam search is used with score using F1 similarity score, BLEU, or exact match based on data
- Evaluated 8 models - 175B GPT3 + 7 smaller models

Results

- Results of various datasets grouped into 9 task-specific categories
- Examples:
  - Language Models
  - Translation
  - Reading Comprehension

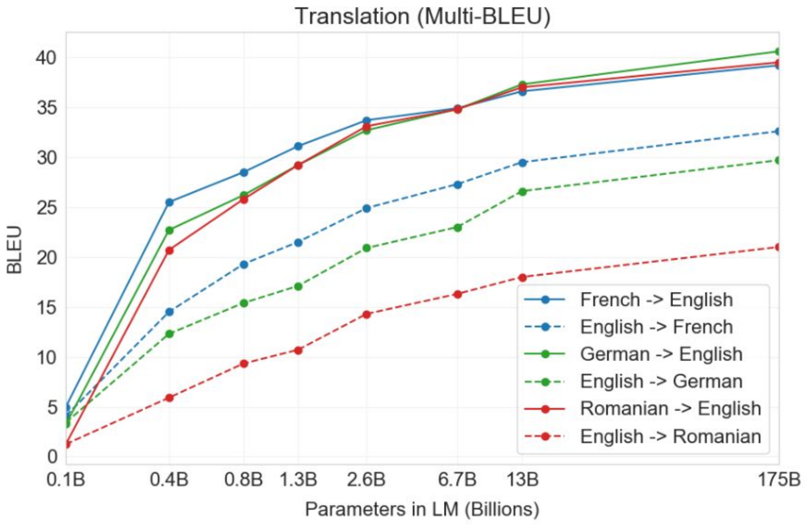
Results - Language Translation

- GPT-3 Few shot gives similar average performance with SOTA while translating to English
- Significant improvement from unsupervised models while translating to English
- Better BLEU score accuracy with increase in number of shots

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Results - Language Translation

- Improvement in BLEU score across different model capacity for all translation tasks



SECTION 6

Nothing's Perfect

Limitations

- Text Synthesis - Loses coherence, predictions contradict themselves
- Limited Input and Output Size
- Underperforms in case of bidirectional use cases
    - Eg: Fill in the blanks, Reading comprehension
- Pre-training limitations with scaling
    - Weighs tokens equally, lack of world context
- Is it really adaptive?
- Bias from data, non-interpretable
- And yes, it’s expensive, large, inconvenient….

Future Work

- Improve pre-training, sample efficiency
- How exactly few shots learning works?
- Distillation to sized specific tasks

GPT4 expected to come by end of this year!

Broader Impacts

- Misinformation, spam, phishing, legal abuse, social engineering pretexting
  - GPT3 can generate text indistinguishable from human-written ones.
- Enabling Threat actors
- Bias in Data
  - Generates prejudiced content, (Gender, Race, Religion)
  - May result in demeaning portrayal
- Energy usage

Fun Facts

In one study, GPT-3 was able to generate “news articles” almost indistinguishable from human-made pieces. Judges barely achieved above-chance accuracy (52%) at correctly classifying GPT-3 texts.

GPT-3 has an artist’s soul! Arram Sabeti told GPT-3 to write a poem about Elon Musk by Dr. Seuss and a rap song about Harry Potter by Lil Wayne.

GPT3 can also generate PowerPoint presentations!
(Though this presentation is solely prepared by us, hope you believe it :))

Questions?

- Ask GPT3!

SECTION 7

Practical

Task 1

- Study the Impact of GPT3 Parameters on Classification Accuracy
  - Few shot learning
  - Twitter Sentiment Classification
  - Plot Accuracy Scores vs Parameters

- Parameters
  - Engine Type
  - Temperature
  - Token Length

Task 2

- Given the abstract of a research paper, generate it's title
    - Feeding examples to GPT-3
    - Zero/Few Shots
    - BLEU Score
    - Plot
    - Learnings
- Dataset

Title and Abstract of Research Papers

	title	abstract
0	On the Cohomological Derivation of Yang-Mills Theory in the Antifield Formalism	We present a brief review of the cohomological solutions of self-coupling interac
1	Regularity of solutions of the isoperimetric problem that are close to a smooth manifold	In this work we consider a question in the calculus of variations motivated by rie
2	Asymptotic theory of least squares estimators for nearly unstable processes under strong dependence	This paper considers the effect of least squares procedures for nearly unstable l
5	Weight Reduction for Mod l Bianchi Modular Forms	Let K be an imaginary quadratic field with class number one and ring of integers
6	Nonequilibrium phase transition in a spreading process on a timeline	We consider a nonequilibrium process on a timeline with discrete sites which ev

Task 3

- By this task we will learn how GPT3 perform better than any SOTA models by comparing the performance of GPT3 with any fine-tuning Language Model.
- For our ease, let's take the reference of our last practical Seq_to_Seq translation use case.
- Let's solve the problem using the Bahdinau_Attention fine-tuning model and solve the same using the GPT3 and compare the metrics and results.

Things to Explore

- Openai Playground - https://beta.openai.com/playground
- GPT3 Github Repo - https://github.com/elyase/awesome-gpt3

References

- GPT3 Paper - https://arxiv.org/pdf/2005.14165.pdf
- https://www.datacamp.com/blog/a-beginners-guide-to-gpt-3
- https://dugas.ch/artificial_curiosity/GPT_architecture.html
- https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd
- https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2

THANK YOU