Seminar Paper
Generative Pre-Trained Transformer 3

Eleni Gaitani

February 2021

1 Introduction

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods, based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised.
The adjective deep, in deep learning comes from the use of multiple layers in the network. It is a modern variation which is concerned with an unbounded number of layers of bounded size, which permits practical application and optimized implementation, while retaining theoretical universality under mild conditions.
In deep learning the layers are also permitted to be heterogeneous and to deviate widely from biologically informed connection models, for the sake of efficiency, train ability and ability to understand, whence the structured part.

Deep neural architectures provide the best results for sentiment analysis, information retrieval, spoken language understanding, machine translation, contextual entity linking, writing style recognition, text classification and others. Among them there are the known ANN's (Artificial Neural Networks), computing systems inspired by the biological neural networks, that can learn progressively to do tasks by considering examples, generally without task-specific programming.
Some mathematical tools we'll need, come as follows:

Central Limit Theorem:

$$Z_n = \frac{\sqrt{n}}{\sigma}(\tilde{X}_n - \mu) \tag{1}$$

where $X_1, .., X_n$ are n random samples from a population with overall mean μ and variance σ^2.
The distribution of various independent observation means approaches a normal distribution model as the sample size gets larger, regardless of the population distribution's statistical shape. Moreover, the theory demonstrates that as the sample size increases, even across multiple unrelated datasets, so increases the accuracy of the population mean estimate. On the same token, if the average of every standard deviation observation in our sample, then the exact standard deviation for the entire population shall derive.
The proof of the theorem is done through the usage of the characteristic function of the variable Z_n and the limit of the Taylor series that derives from it, as n converges to infinity. In a general sense, when it comes to networks, this theorem is one of the first steps in helping a deep learning algorithm to imitate the human concept of intuition.

1.1 How to Train a Neural Network

Artificial neural networks are created, based on the principle of bio-mimicry. The same structure our brain includes, with the nucleus, dendrites, axon, neurons and synapses, is taking place here too.
External stimuli (the data), whose signal strength is adjusted by the neuronal weights via the dendrites. The result of the calculation or else the output, is then carried on (via the axon) to several other neurons and then under layers are combined, and so on.
To train a neural network, we need to follow some steps:

- Formulate the learning problem for neural networks.
- Then,some important optimization algorithms are described.
- Finally, the memory, speed and precision of those algorithms are compared.
- To train a neural network, we start with some parameter vector (often chosen at random). Then, we generate a sequence of parameters, so that the loss function is reduced at each iteration of the algorithm. The change of loss between two steps is called the loss decrement. The training algorithm stops when a specified condition, or stopping criterion, is satisfied.

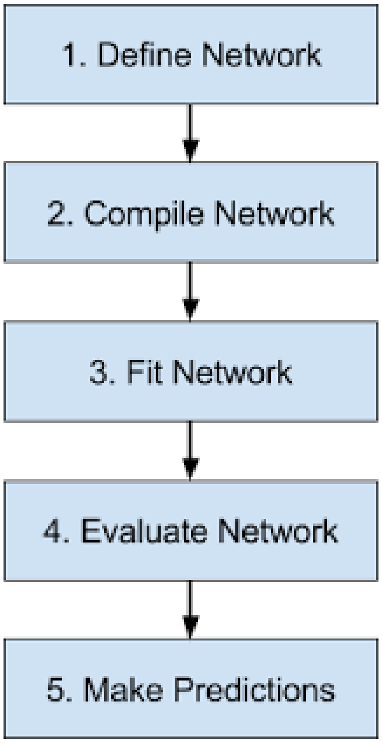


Abbildung 1: Life cycle for a neural network in Keras model. Training a network, means finding the best set of weights to map inputs to outputs in our data set.

The procedure used to carry out the learning process in a neural network is called the optimization algorithm (often we pick the gradient descent algorithm).

There are many different optimization algorithms. All have different characteristics and performance in terms of memory requirements, processing speed, and numerical precision. The learning problem is formulated in terms of the minimization of a loss index or else, loss function. It is a function, that measures the performance of a neural network on a data set. The loss index is, in general, composed of an error and a regularization term. The error term evaluates how a neural network fits the data set. The regularization term is used to prevent over fitting by controlling the sufficient complexity of the neural network.

Gradient descent algorithm:
The loss function $f(w)$ depends on adaptive parameters which are represented through the n-dimensional weight vector w,
$w^{i+1} = w^i - g^i \eta^i$,
where $g^i = \nabla w^i$ and η is the training rate.
The necessary condition states that if the neural network is at a minimum of the loss index, the gradient is zero.
For the optimization algorithm we consider a search through the parameter space consisting of a succession of steps, or epochs. At each epoch, the loss will decrease by adjusting the neural network parameters. The change of parameters between two epochs is called the parameter increment.
The optimization algorithm stops when a specified condition is satisfied. Some stopping criteria commonly used are:

- parameters increment norm is less than a min value
- loss improvement in one epoch is less than a set value

- loss has been minimized to a goal value
- max numbers of epochs has been reached and others.

1.2 State-of-the-art Language Models and GPT-3

Considering we have on mind the general structure of ANN's, we want to create a so called state-of-the-art (SOAT) language model, based on Natural Language Processing, known as NLP. That means a model that efficiently corresponds in different domains of NLP.
NLP: is a branch of Artificial Intelligence (AI) that studies how machines understand human language. It's goal is to build systems that can make sense of text and perform tasks like translation, grammar checking or topic classification. Working areas in NLP are language modelling, sentiment analysis, machine translation, text classification and question answering. The ability to classify emotions in text data, predict words or letters in a text, given some previous words or an existing text, as well as translate in another language, assign a certain category to a text or word and answer questions, mostly based on reading comprehension, are the tasks that we are interested in, in order to define the highest level of general development of our pre-trained model. In other words, those are the areas that each and every AI model will perform, to define it's efficiency as a state-of-the-art model.

The most qualified model that has been researched for these purposes, is the recent GPT-3 from OpenAI.

1.2.1 GPT-3

GPT-3 is an auto-regressive, third-generation state-of-the-art language model developed by OpenAI, with the purpose of producing high quality human-like tests (based on artificial neural network architecture). Microsoft has acquired an exclusive licence to the model, probably to incorporate it in its cloud for generative text.

- It uses the process of generative pre-training on a diverse corpus of unlabelled text, followed by discriminating fine-tuning on each specific task. Therefore, human supervision is not demanded.
- GPT-3 can unscramble words, use a novel word in a sentence or perform 3-digit arithmetic, as well as generate samples of news articles, while this might seem a tricky task for humans.
- Moreover, developers can access it through the OpenAI API, even if they are not experienced with AI. Coding languages are JSX, CSS, Python among others.

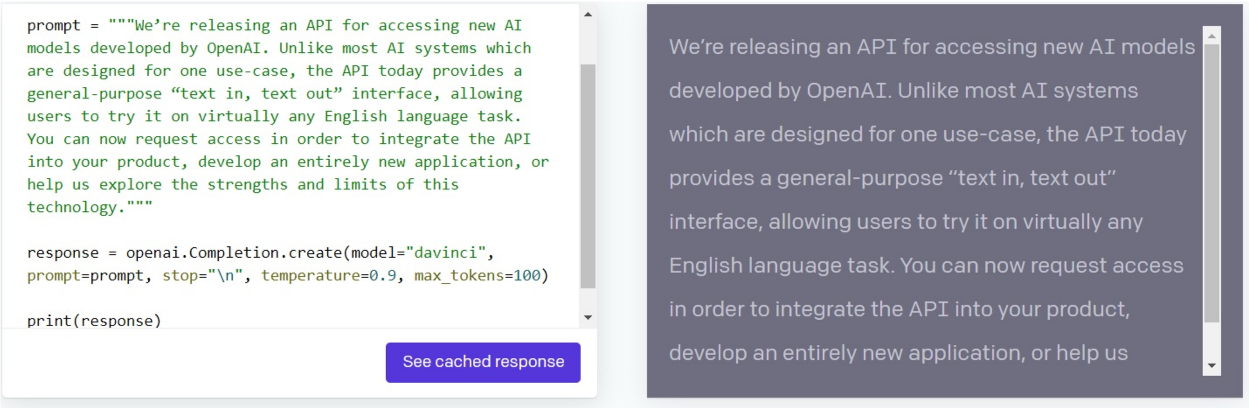
GPT-3 is based, as all GPT-n models on the Transformer Deep Learning. The connections between nodes form a directed graph along a temporal sequence, known as Recurrent Neural Network (RNN). Those remember their inputs due to an internal memory, therefore are perfectly appropriate for machine learning algorithms that involve sequential data. A simple example is the operational system of Google's voice search. Moreover,the Transformer allows more paralleling in reduced training times and can handle sequential data for translation or text summarizing without processing them in order. This led to BERT and GPT pre-trained systems, that accommodate to Transfer Learning.

Deep Speed Library is an open source deep learning optimization library for PyTorch for model development and training.
The library is designed to reduce computing power and memory use and to train large distributed models with better parallelism on existing computer hardware. DeepSpeed brings state-of-the-art training techniques, such as ZeRO, distributed training, mixed precision, and check-pointing, through lightweight APIs compatible with PyTorch.
It excels in four aspects: Scale, speed, cost and usability.
It can be obtained for models with parameters of a billion scale, as it has been for GPT-2 and Turing-NLG, as well as maintain high speed, up to 5 times faster due to its memory efficiency and usage of lower level of model parallelism and larger batch sizes. Moreover, it requires 3 times fewer resources to train a 20 billion

parameter model and most importantly it doesn't require a redesign of the code or model refactoring for a model to access this library or Zero Optimizer[1].
Last but not least, while standard data parallelism will run out of memory, as it happens with models with more than 1.3 billion parameters, Deep Speed Library can offer data parallelism(powered by ZeRo) without using model parallelism for up to 6 billion parameters. Also, it provides flexible combination of the two. It is available for free on Github.

1.2.2 GPT-3 Specifications

The OpenAI society released an API (Application Programming Interface) for accessing the AI models developed by them, in order for the public to try them out. In the figure below there is a representation of how the API runs their models by using weights from GPT-3, manipulating it's benefits of speed and precision [2].



```
prompt = """We're releasing an API for accessing new AI
models developed by OpenAI. Unlike most AI systems which
are designed for one use-case, the API today provides a
general-purpose “text in, text out” interface, allowing
users to try it on virtually any English language task.
You can now request access in order to integrate the API
into your product, develop an entirely new application, or
help us explore the strengths and limits of this
technology."""

response = openai.Completion.create(model="davinci",
prompt=prompt, stop="\n", temperature=0.9, max_tokens=100)

print(response)
```

We're releasing an API for accessing new AI models developed by OpenAI. Unlike most AI systems which are designed for one use-case, the API today provides a general-purpose “text in, text out” interface, allowing users to try it on virtually any English language task. You can now request access in order to integrate the API into your product, develop an entirely new application, or help us

Abbildung 2: API's within cloud environments offer the end customer software interfaces to interact with and configure their services.

GPT-3 has up to 175 billion machine learning parameters (ANN's) and has been trained with 45TB text data, that includes sources like Wikipedia and e-books. Additionally, 60% of its data for pre-training GPT-3 was taken from Common Crawl while 22% from WebText2, 8% from Books1, 8% from Books2 and 2% from Wikipedia.
It has 96 decoder layers and is built on a system of 285k CPU cores, 10k GPU and 400Gbps network connectivity for each GPU server. Compared to other models like Turing-NLG (or T-NLG) from Microsoft, a Transformer-based model, as well as GPT-2 and Bert, models that answered by extracting existing content from documents, GPT-3 is more sophisticated as it uses Natural Language Processing, DeepSpeed library and Zero Optimizer. Therefore, it achieves strong performance on manz NLP data sets, including translation, question-answering and close-tasks, as well as some tasks that require on-the-fly reasoning or domain adaptation such as unscrambling word, using a novel word in a sentence or performing 3-digit arithmetic.

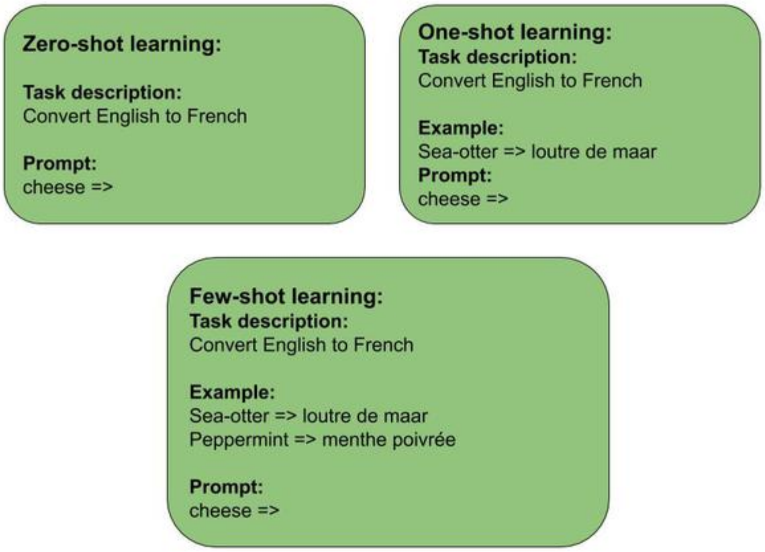
[1] Zero Optimizer is a novel memory optimization technology for large-scale distributed deep learning such as 100 billion parameter deep learning models.

[2] According to API Management Company, there exist 3 categories of Cloud API'S:Control APIs, that allow the end customer to configure their cloud provisioned service. They include the allocation of Internet Protocol (IP) addresses, creating/editing access control lists etc. Data APIs, where within which data may flow into or out of the provisioned service and Application Functionality APIs, which provide the ability to transfer data between alternate providers, as well as the possibility for the end customer to interact with their functions, ranging from the simplest, availability of shopping baskets up to integration with social networking solutions.

Training Methods for In-Context Training

- Fine-Tuning (FT) has been the most common approach in recent years, and involves updating the weights of a pre-trained model by training on a supervised data set specific to the desired task. Typically thousands to hundreds of thousands of labeled examples are used. The main advantage of fine-tuning is strong performance on many benchmarks. The main disadvantages are the need for a new large data set for every task.
- Zero-Shot Learning: The model tries to predict the answer without training, provided there is an input and a task(one natural language instruction).This method provides maximum convenience, potential for robustness, and avoidance of spurious correlations (unless they occur very broadly across the large corpus of pre-training data). It's closest to human performance tasks.
- One-Shot Learning: The model tries to predict the answer, when given an example. This has been adapted to computer vision, such as the Siamese Network, which calculates the distance between a training and a test example in a neural network. It most closely matches the way in which some tasks are communicated to humans.
- Few-Shot Learning: The model tries to predict the answer, when given a few examples of tasks, but no weight updates are allowed.

The selection of the appropriate one depends on how many demonstrations are provided at inference time. Below a demonstration of a translation task for the word cheese in French is being shown.



Zero-shot learning:

Task description:
Convert English to French

Prompt:
cheese =>

One-shot learning:
Task description:
Convert English to French

Example:
Sea-otter => loutre de maar
Prompt:
cheese =>

Few-shot learning:
Task description:
Convert English to French

Example:
Sea-otter => loutre de maar
Peppermint => menthe poivrée

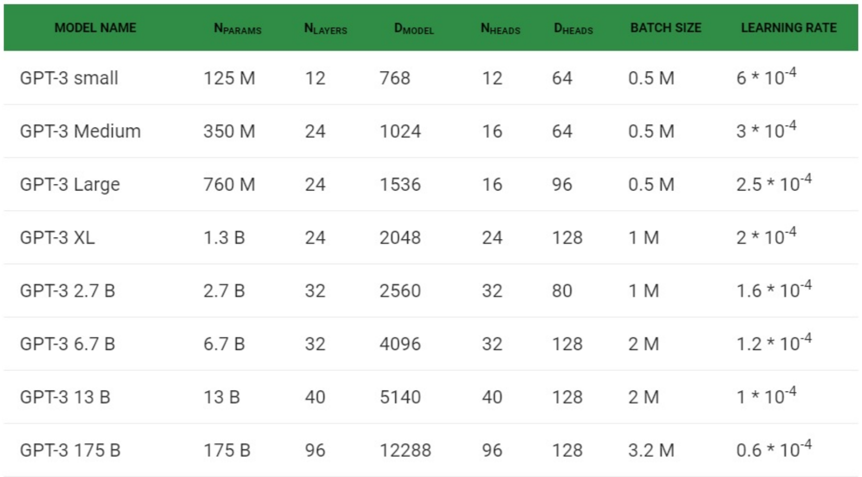
Prompt:
cheese =>

Abbildung 3: Example of Zero- One- and Few-Shot learning, performing translation from English to French.

GPT-3 works with unsupervised pre-training. During this, the model develops a broad set of skills and recognition abilities, that uses after the training, to adapt or recognize rapidly the desired task(meta-training). Despite its promises, meta-learning clearly requires substantial improvement, in order to be viable as a practical method of solving language tasks, There is an inner loop of this process, repeated sub-tasks embedded with a single sequence (in-context-learning).

1.2.3 GPT-3 Architecture and Training Data sets-Training Process

In addition to all the above, a series of smaller models (ranging from 125 million parameters to 13 billion parameters) were trained, in order to compare their performance to GPT-3 of 175B. The models are trained with a huge amount of data, by performing gradient upgrades after each epoch or example (Fine Tuning), in order to perform strongly on a desired task. According to the results, there is the suggestion that larger models have a slower learning rate but are more efficient meta-learners.
All models were trained for a total of 300billion tokens and have a context window of 2048 tokens. Last but not least, the models were partitioned across GPU's, both at width and length dimension in order to minimize data-transfer between the nodes.
Note that, the largest version GPT-3 175 B has 175 billion parameters, 96 attention layers and 3.2 M batch size.



MODEL NAME	N_{PARAMS}	N_{LAYERS}	D_{MODEL}	N_{HEADS}	D_{HEADS}	BATCH SIZE	LEARNING RATE
GPT-3 small	125 M	12	768	12	64	0.5 M	$6 * 10^{-4}$
GPT-3 Medium	350 M	24	1024	16	64	0.5 M	$3 * 10^{-4}$
GPT-3 Large	760 M	24	1536	16	96	0.5 M	$2.5 * 10^{-4}$
GPT-3 XL	1.3 B	24	2048	24	128	1 M	$2 * 10^{-4}$
GPT-3 2.7 B	2.7 B	32	2560	32	80	1 M	$1.6 * 10^{-4}$
GPT-3 6.7 B	6.7 B	32	4096	32	128	2 M	$1.2 * 10^{-4}$
GPT-3 13 B	13 B	40	5140	40	128	2 M	$1 * 10^{-4}$
GPT-3 175 B	175 B	96	12288	96	128	3.2 M	$0.6 * 10^{-4}$

Abbildung 4: GPT-3 Architecture/Table of Data

In the graphic picture below, the difference of GPT-3 in the accuracy between models of different size of parameters, as well as having adapted different training methods, is being underlined.

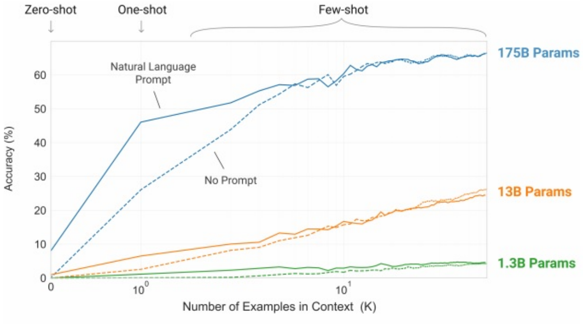


Abbildung 5: GPT-3 Accuracy in accordance to the size of the model

Data sets

Data sets for language models originating from CommonCrawl data set, contain around a trillion words, a sufficient number to train GPT-3 or similar models of large number of parameters. Others were used though too, like WebText2, Wikipedia, Books1 and Books2.
3 steps were taken to improve the quality of the CommonCrawl data set:

- Downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference body.

- Performed fuzzy reduplication at the document level, within and across data sets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of over-fitting.

- Added known high-quality reference corpora to the training mix to increase its diversity.

Training Process

Larger models can typically use a larger batch size, but require a smaller learning rate. The choice of batch size was calculated after the measurement of gradient noise during training.

-*Gradient noise*:

Is a type of noise commonly used as procedural texture in computer graphics as shown before. It consists of a creation of a lattice of random gradients, dot products of which are then interpolated to obtain values in between the lattices.
To train the larger models without running out of memory, we use a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network in the GPU's.
All models were trained on V100 GPU's on part of a high-bandwidth cluster provided by Microsoft.

-*Model Parallelism*:

Deep neural networks or deep artificial neural networks follow the structure of the actual brain and its functions. They use multiple layers of artificial neurons for classification and pattern recognition.
There are 4 methods of model parallelism, Inter-Model Parallelism, Data Parallelism, Task-level Parallelism and Intra-Model Parallelism.

-*GPU- Graphics Processing Unit*:

Deep learning involves huge amounts of matrix multiplications and other operations which can be massively parallelized and thus sped up on GPUs. A single GPU might have thousands of cores while a CPU usually has no more than 12 cores.

1.2.4 Evaluation & Performance

For the few-shot learning method, demonstration examples were drawn directly from the data sets, when given,or from the development set and were evaluated on the test set. E.g. the LAMBADA data set, which basically tests a model's capability to predict the last word of sentences, which requires reading a paragraph of context, doesn't involve any supervised training set.

- On multiple choice tasks, there were K examples of context plus correct completion provided (K is a value from 0 to model's context window nctx = 2048 and fits to 10-100 examples) and one example of context only. The LM likelihood* of each completion was compared.

- On binary classification tasks, there was a semantic option of *True* or *False* involved and then the task was treated like multiple choice.

- Finally, on tasks with free form completion a beam search[3] with beam width 4 and length penalty of a= 0,6 was used. The model was scored by using F1 similarity score, BLEU, or exact match[4], depending on what is standard for the data set at hand.

<u>*Results*</u>
Different areas of tasks were researched and the results approved the capability of GPT-3 in reading comprehension, translation (German to English was tested), reversed words and anagrams, language modelling, word prediction with LAMBADA data set, performance in StoryClose, HellaSwag data sets (they contain the correct ending-sentence in a story and the best story ending in a text), in Closed Book Question Answering, Winograd-Style tasks included the difficult versions, which involve which word a pronoun refers to, when the pronoun is semantically indistinct to a human, as well as in tasks like common sense reasoning by using different data sets, NLI (Natural Language Inference) that contains the ability to understand the relationship between two sentences, if the second is logical in accordance to the first, arithmetic up to 5-digit, SAT analogies that refer to analogy problems in sentence construction and finally, news article generation, learning and using novel words.

GPT-3's performance consequenced in the following observations:
- It significally improves SOTA on LAMBADA data set.
- Depending on its size, gains more knowledge and efficiently responds to question-answering.
- It gains background information but cannot memorize the answer to a specific question.
- It outperforms other language models in translation performance. (Note that it's performance is reduced when it comes to translating from English).
- It recorded the best score to Winograd tasks.
- All in all, in-context learning with GPT-3 showed mixed results in commonsense tasks, but in PIQA dataset (is designed to investigate the physical knowledge of a model) it sets SOTA in all evaluation settings.
- GPT-3 175B parameters proved accurate in 2-digit arithmetic, but it's performance drops from 3-digit and upwards.
- It's able to generate accurately short "news" articles. It's worth mentioning, that to measure the quality of those, human ability to distinguish the generated text between a human and a machine was tested. It showed that this ability decreases, as the model size grows.
- It proved proficient in at the task of using novel words in a sentence.

In figure 6, depending on Zero-Shot, One-Shot or Few-Shot training methods, it is outstanding how close to human perception GPT-3 approaches, based on the Lambada data set, with the Few-Shot method overcoming the average of Zero-Shot SOTA. More specifically, in figure 7 information in translation is obtained, using the BLEU similarity metric. Similarly, in figure 8 the results in different tasks of arithmetic (addition, subtraction and multiplication) are being shown.
GPT-3 shows more accurate performance at the degree of 13 billion parameters or over this limit.

[3] It's a greedy algorithm (uses the locally optimal choice heuristically in the search) that explores a graph by expanding the most promising node in a limited set.
At machine translation, the field of linguistics that we are interested in here, the beam search algorithm reaches the configured maximum depth search and evaluates the solutions found, returning the best one (meaning the one with the highest probability). The beam width can either be fixed or variable, starts at the minimum possible though.

[4] Those are metrics to evaluate the set similarity, whether that is classifiers, generated sentences or samples with classified labels.

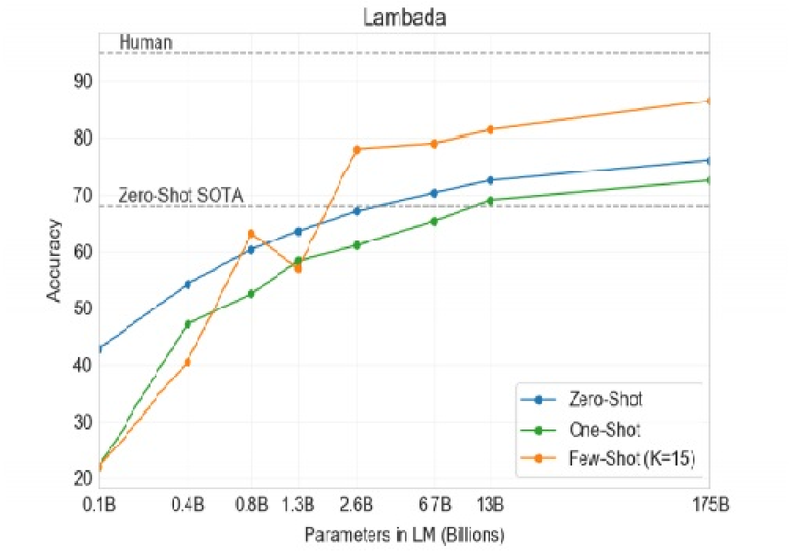


Abbildung 6: Performance on LAMBADA Data set

”We can’t run experiments on the origin of the universe for example, but with GPT-3 we could generate a possible hypothesis for why it works the way it does and then simply test that hypothesis.”
-GPT-3

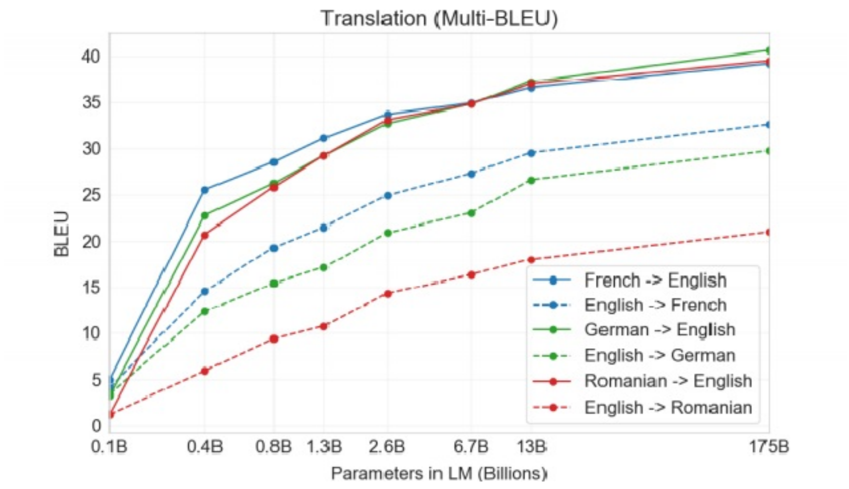


Abbildung 7: Performance of GPT-3 in translation

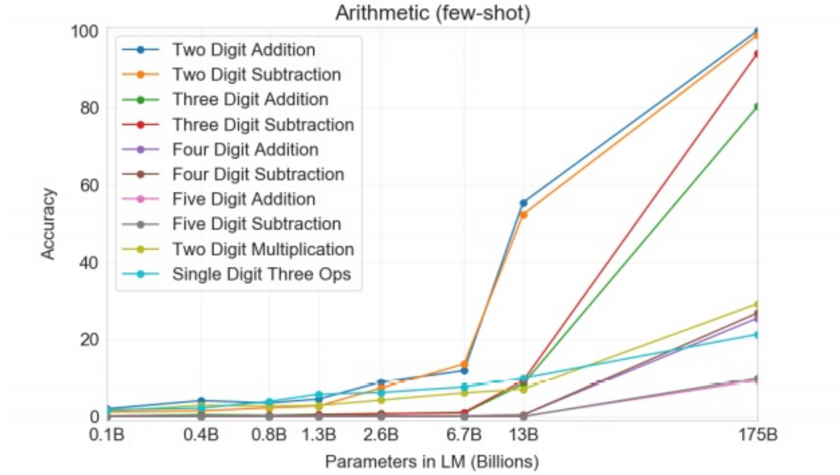


Abbildung 8: Performance of GPT-3 in arithmetic

1.2.5 GPT-3 and other NLP Models

GPT-3 has similar architecture with the original Transformer architecture (see figure), only that it is pretty much larger. While other language models like BERT use the Encoder to generate embeddings from the raw text, which can be used in other machine learning applications, GPT-3 uses the Decoder half, to produce text from the embeddings. Like explained above, GPT-3 can act also on Zero- Shot Learning method. That makes the model Task Agnostic.

Turing NGL

Prior to GPT-3, apart from BERT and GPT-2 which provided an answer by extracting existing content from documents, the most efficient language model was Turing-NGL from Microsoft.
This 17 billion parameter Transformer-based model learns from text publishing on the internet and outstands in:

- question-answering, conversational agents and document understanding

- While its usefulness relied on assisting authors in content composure, summarizing long texts and provide customer service digitally by using DeepSpeed Library and Zero Optimizer.

Basic goal was to create a model that responds like humans in any situation (on a language level).
It is parallelized across many GPU's, uses NVIDIA DGX-2 hardware set up, is compatible with the Megatron-LM framework and PyTorch, it's batch size per node is increased and it's training time is reduced by 3 times.

T5 Model

The Text-To-Text Transfer Transformer (T5) is a smaller parameter model, which proceeds the Turing-NGL and is also an encoder-decoder model like BERT. It is trained on Colossal Clean Crawled Corpus (C4), that is an open source pre-trained dataset.
It can perform a variety of tasks like translation, classification and others, (here it's regression worth mentioning). This is accomplished by directing a variety of NLP tasks to a single input during this pre-training process (text-to-text format).

The disadvantages that appeared were lack of performance in Winograd tasks, translation due to the fact that it failed in suggesting it as a language agnostic approach. Lastly, there is the possibility for the model to output words that are not to be expected in the output and the regression was attended to a classification task since the model is trained to predict values in small ranges.

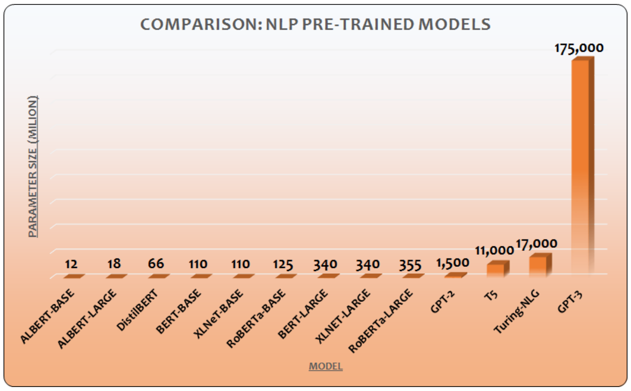


Abbildung 9: Difference of parameters between NLP pre/trained models.GPT-3 is the largest model developed, trained with 175B parameters.

1.2.6 GPT-3 Downside Effects

In general, all Machine Learning models have some shared limitations, when it comes to robustness of their components, as well as risks of their open-ended systems like falsehood in the generated text. GPT-3 contains those and some extras too:

- Repetition: GPT-3's data can repeat themselves and loose coherence in context.
- It lack's of world context due to the lack of real-world physical interaction or human feedback.
- Performance on different data input such as non-English languages, is poor. - Interpretability Predictability of the model is low, as it happens with large scale models.
- Demands a data-training update (since the data it was trained at, comes from 2019 sources).

Other disadvantages that we have to work on in the future:
Misinformation, spam, phishing, abuse of legal govermental processes, fraud in academic essays, social engineering pretexting and under-representation of less connected communities, since internet penetration is connected more to young, healthy, male and mostly US-centric societies.Therefore, risk mitigation is on demand.
Furthermore, another example is that when you train a computer system to predict which convicted felons will reoffend, you're using inputs from a criminal justice system biased against Black people and low-income people, so its outputs will likely be biased against Black and low-income people, too. Making websites more addictive can be great for your revenue but bad for your users. Releasing a program that writes convincing fake reviews or fake news, might make those widespread, making it harder for the truth to get out. GTP-3 though is incapable of those threats.

1.2.7 Other Applications

GPT-3 can be used as a search machine or a tool to create a novel, a resume for work or educational reasons or write down SQL queries and build web apps.

Furthermore, GPT-3 to generate code for a machine learning model known as Keras model, just by describing the dataset and required output.

Quite recently, GPT-3 was proven to be appropriate not only for human-like text generation, but also in Computer Vision. The DALL-E GPT-3 based model was presented in the beginning of 2021 from OpenAI, which generates high quality images based on text. It is also based on the Transformer architecture. It receives during pre-training two sequence of data- text and image and it can perform in controlling attributes, drawing multiple objects, perspective and 3D visualization, multiple level visualization and inferring context. This establishes a new era in interior and exterior as well as graphic design.

1.2.8 Turing Test and Potential for the Future

No matter for its excellence in text production and usefulness in various areas, GPT-3 didn't pass the Turing-Test, which by the way is not a test to qualify artificial intelligence but to measure a machine's ability to deceive, in this case human. The reason for this, is it's incapability to answer appropriately to a person that seeks for its weaknesses.
As mentioned, GPT-3 will try to correspond to a specific question or generate text depending on the variables that we insert, according to what kind of text we want to produce. It doesn't understand though the value of the text or the answer itself.

WOOF
WOOF WOOF
WOOF
WOOF WOOF
WELL, THE DOGS COULDN'T TELL ANY DIFFERENCE, AND NIETHER COULD I. CONGRATULATIONS, YOU'VE CREATED A MACHINE THAT IS JUST AS INTELLIGENT AND CONSCIOUS AS DOGS.
THANK YOU, AND WITH YOUR GRANT MONEY, I'M SURE HUMANS AREN'T FAR BEHIND.
☑ BARKS
☑ DOESN'T STOP WHEN YOU YELL SHUT UP
☑ MORE BARKING
EXISTENTIAL COMICS

A computer engineer, Kevin Lacker proved that GPT-3 can remotely simulate common sense, based on the facts that it can answer private questions like ”favorite animal”, explain the reason why it picked that answer, while it can also relate two different objects.
On the other hand though, if there are no answers for specific questions like ”Who won the World Cup in the year 2021?”, GPT-3 will try to fantasize and generate an answer that might pass, ”The New York Yunkies”. Therefore, it cannot be totally reliable about future events or even for medical purposes.

On the contrary, GPT-3 is already showing rudimentary computing skills, while a language AI trained by Facebook can solve formulas and researcher Miles Cranmer wants to revolutionize physics with AI.
According to researchers, it will be a revolutionary period for everyone, when an artificial intelligence model manages to construct a mathematical theorem with a proof.