Language Models (GPT, GPT-2 and GPT-3)

Advanced Techniques in Artificial Intelligence

Jon Ander Almandoz, Julen Etxaniz and Jokin Rodriguez

2020-09-28

Contents

1. Introduction

Natural Language Processing has evolved so much that nowadays machines are able to "understand" the context behind messages, articles and much more. This enormous step has been made by the company called OpenAI, an artificial intelligence research laboratory. It was created in 2015 in California by Elon Musk (CEO of SpaceX and Tesla), Sam Altam and other investors.

This company has already released 3 different GPT (Generative Pre-trained Transformer) models: GPT, GPT-2 and GPT-3. Each one being better than the previous one. GPT did not have much echo, the main reason was that it was more like an idea/test instead of a product, but with the release of the GPT-2, OpenAI claimed a lot of attention. The reason for that much attention was that this language processing model was able to predict very well the upcoming word starting off a whole text. This meant the model was able to "understand" what was written, something impossible to think some years ago.

Initially, the company showed the world a glimpse of the GPT-2 model. The principal reason for having shown a glimpse and not the entire version was the harm it could make creating really good fake news. The great similarity between a GPT-2 writing and a human writing would make a good fake news filter useless. This created a big controversy about OpenAI not being open.

OpenAI API is the only way to use the GPT-3 model, as it hasn't been released. The API is in private beta and you have to request access and pay in order to use it. This continued the debate that was started when OpenAI decided not to publish GPT-2.

These are the sizes of the 3 models:

- GPT: trained on 5GB of text (150M parameters)
- GPT-2: trained on 40GB of text (1.5B parameters)
- GPT-3: trained on 570GB of text (175B parameters)

As we can see, the difference between GPT-2 and GPT-3, in terms of complexity, is enormous. They passed from 1.5 billion parameters to 175 billion parameters, which is more than 100 times the number of parameters. This difference is perfectly reflected in the performance of the software.

We have just used the word "parameters" to show the complexity of the neural networks. But what is it?

Parameters is a synonym for weights, which is the term most people use for neural networks parameters. Batch size, learning rate etc. are *hyper-parameters* which basically means they are user specified, whereas weights are what the learning algorithm will learn through training. In other words, the number of parameters is how many connections there are between the nodes of a NN.

2. GPT

2.1. Improving Language Understanding by Generative Pre-Training

The paper Improving Language Understanding by Generative Pre-Training was released in June 2020.

Understanding natural language has diverse challenges, such as textual entailment, question answering, semantic similarity assessment, and document classification. They have discovered large improvements on these tasks by generative pre-training, followed by supervised discriminative fine-tuning on each specific task. All datasets use a single language model, requiring very little tuning to achieve the results.

2.1.1. Why Unsupervised Learning?

Although unlabeled text is abundant, labeled data for specific tasks is scarce, making training models with supervised learning harder. Unsupervised learning is attractive because of its potential to address these drawbacks.

It is a very active area of research but practical uses of it are often still limited. New techniques are now being used, which are further boosting performance. These include the use of pre-trained sentence representation models, contextualized word vectors approaches like GPT that combine unsupervised pre-training with supervised fine-tuning.

The language model begins to perform tasks without training on them. The model outperforms some models specifically trained for each task. For instance it achieved improvements on commonsense reasoning, question answering and textual entailment. While the performance is still quite low compared to the best supervised models, this is robust across a broad set of tasks.

2.1.2. Drawbacks

There are a few outstanding issues worth mentioning:

- **Expensive computer requirements:** Their approach requires an expensive pre-training step - 1 month on 8 GPUs. Luckily, this only has to be done once and they have released the model, so others don't have to do it.
- **The limits and bias of learning about the world through text:** It is not easy to find books or tests on the internet about the real world, that makes it more difficult to learn.
- **Still fragile generalization:** Although their approach improves performance across a broad range of tasks, deep learning NLP models still exhibit surprising and counterintuitive behavior.

2.1.3. Future

- **Scaling the approach**: They've observed that improvements in the performance of the language model are well correlated with improvements on downstream tasks. There is significant room for improvement.
- **Improved fine-tuning**: Their approach is currently very simple, improvements can be made using more complex adaptation.
- **Better understanding of why generative pre-training helps**: More targeted experiments and research will help distinguish between competing explanations.

3. GPT-2

3.1. Language Models are Unsupervised Multitask Learners

The paper Language Models are Unsupervised Multitask Learners was released in February 2019.

GPT-2 has a simple objective: given all of the previous words within some text, predict the next word. Compared to GPT, GPT-2 has more than 10X the parameters and is trained on more than 10X the amount of data. Samples from the model reflect these improvements and contain coherent text.

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. GPT-2 demonstrates that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of 8 million web pages called WebText. To create the dataset, they only used external website links from Reddit which received at least 3 karma, leading to higher data quality than other similar datasets.

3.1.1. Samples

The model is able to generate realistic samples based on the given input. However, there are some failures like repetitive text, world modeling failures (*fire under water*) and unnatural topic switching. Generally, it takes a few tries to get a good sample, depending on how familiar the model is with the context.

- **10 tries**: scientist discovered a herd of unicorns
- **1 try**: a train carriage containing controlled nuclear materials was stolen
- **2 tries**: Miley Cyrus was caught shoplifting
- **1 try**: a large language model called GPT-2 generates realistic paragraphs of text
- **1 try**: Legolas and Gimli advanced on the orcs
- **5 tries**: please describe the reasons for the US Civil War
- **25 tries**: John F. Kennedy was just elected President of the United States
- **25 tries**: recycling is bad for the world

3.1.2. Zero-shot examples

GPT-2 also achieves good scores on domain-specific language modeling tasks, outperforming some domain-specific models on their task. While scores on these tasks are far from state-of-the-art, the performance is expected to increase with more compute and data.

- **Reading Comprehension**: answer questions about given passages
- **Common Sense Reasoning**: resolution of an ambiguous pronoun
- **Question Answering**: answer questions without context
- **Language Modeling of Broad Contexts**: predict the last word of a passage
- **Summarization**: summarize news articles
- **Machine Translation**: translate French sentences to English

3.1.3. Our examples

We have used InferKit to test GPT-2. Signing up is necessary and there is a limit of 20000 characters in the free plan. Here are some examples we have obtained:

- ***My parents told me I should stay at home because*** there are too many negative things going around in the city. They said I should stop being so impulsive and start doing things slowly. They're right. I should have listened to them. But, here I am. Once again, I'm in trouble. I've done something I shouldn't have, and it's only getting worse.

- ***What is 2001: A Space Odyssey about?*** The opening sequence of 2001 sees an Earth that is still thriving after the crash of 1967. The long-dead nuclear war has been forgotten, and mankind is now engaged in interstellar exploration. For their last home world, humanity has developed computer systems that allow the probes to travel through

- ***OpenAI GPT is*** a scalable, unsupervised, reinforcement learning environment for the training and testing of supervised models on unstructured data. It can also be used for feature engineering. Assembling ideas from earlier open source projects, OpenAI GPT is inspired by agent-based and reinforcement learning.

3.2. Release Strategies and the Social Impacts of Language Models

The report Release Strategies and the Social Impacts of Language Models was released in August 2019 and later updated in November 2019.

Large language models have a range of beneficial uses but flexibility and generative capabilities also raise misuse concerns. In this section we mention policy implications, release strategy, staged release and partnership-based sharing.

3.2.1. Policy Implications

These are some of the anticipated beneficial purposes:

- AI writing assistants
- More capable dialogue agents
- Unsupervised translation between languages
- Better speech recognition systems

There could also be malicious purposes:

- 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

Those malicious purposes have to be considered when researching content generation, and better countermeasures need to be created.

3.2.2. Release Strategy

Due to concerns about malicious applications of the technology, they initially released a smaller version of GPT-2. They didn't release the dataset, training code, or GPT-2 model weights. This created a big controversy about OpenAI not being open.

They were aware that some researchers had the technical capacity to reproduce and open source our results. But they believed that their release strategy would limit the organizations who do this, and would give the AI community more time to have a discussion.

They used two mechanisms to responsibly publish next versions of GPT-2: staged release and partnership-based sharing.

3.2.3. Staged Release

Staged release involves the gradual release of a family of models over time. The purpose is to give people time to assess the properties of these models, discuss their social implications, and evaluate the impacts after each release. There were 4 releases:

1. **February 2019:** Released small 124M parameter model.
   They also released a technical paper.
2. **May 2019:** Released medium 355M parameter model.
   Shared the 762M and 1.5B versions with partners in the AI and security communities who work to improve social preparedness for large language models.
   Released a dataset of GPT-2 outputs from all 4 model sizes, as well as a subset of the WebText dataset used to train GPT-2. It is aimed to help a wider range of researchers perform quantitative and qualitative analysis.
3. **August 2019:** Released larger 774M parameter model.
   They also released an open-source legal agreement to make it easier for organizations to initiate model-sharing partnerships with each other.
   They also published a technical report about their experience in coordinating with the AI research community on publication norms.
4. **November 2019:** Released largest 1.5B parameter model.

While there were larger language models released since August, they continued with the original staged release plan in order to provide the community with a test case of a full staged release process.

3.2.4. Partnerships

They shared the 762M and 1.5B parameter versions to facilitate research on language model output detection, language model bias analysis and mitigation, and analysis of misuse potential. These research partnerships will be a key input to our decision-making on larger models. They partnered with four leading research organizations:

1. **Cornell University** studied human susceptibility to digital disinformation generated by language models.
2. **The Middlebury Institute of International Studies** Center on Terrorism, Extremism, and Counterterrorism (CTEC) explored how GPT-2 could be misused by terrorists and extremists online.
3. **The University of Oregon** analyzed bias within GPT-2.
4. **The University of Texas at Austin** studied the statistical detectability of GPT-2 outputs after fine-tuning the model on domain-specific datasets, as well as the extent of detection transfer across different language models.

3.2.5. Learnings

1. **Coordination is difficult, but possible.** Before the public release of the 1558M parameter language model, multiple organizations had already developed the systems to train them, or had publicly discussed how to train larger models.
2. **Humans find GPT-2 outputs convincing.** Research from Cornell University has shown that people find GPT-2 text almost as convincing (72%) as real articles (83%). They surveyed people to assign a “credibility score” to GPT-2 text. The 1.5B model had a 6.91 out of 10, greater than the 774M model with a 6.72 and significantly above the medium 355M model with a 6.07. So we can see that there is an increase in human-perceived credibility according to the number of parameters used.
3. **GPT-2 can be fine-tuned for misuse.** Partners at the Middlebury Institute of International Studies CTEC found that it could be used to generate ideological propaganda (white supremacy, jihadist, anarchism...).
4. **Detection is challenging.** Detectors need to detect a significant fraction of generations with very few false positives. They believe 95% is not high enough accuracy and needs to be paired with metadata-based approaches, human judgment, and public education to be more effective. They found detection accuracy depends heavily on the sampling methods used in training and testing. Larger models outputs are more difficult to classify, but training on larger models’ outputs makes detection results more accurate and robust.
5. **There is no strong evidence of misuse so far.** Even with the potential GPT-2 has at operations like spam and phishing there is no evidence of writing code, documentation, or instances of misuse.
6. **Standards for studying bias are needed.** Language models have biases. It is a challenge for the AI research community to study these biases. Publishing a model card alongside the models on GitHub is one possible solution. Performing a qualitative evaluation of some biases (gender, race, and religion) is another solution.

4. GPT-3

4.1. Language Models are Few-Shot Learners

The paper Language Models are Few-Shot Learners was released in May 2020.

Humans are capable of performing really well on several new language tasks, just being given some instructions or a few examples. This ability comes from the capacity of understanding, which machines yet don't have. But this doesn't mean machines can't do something similar. In fact, when trained in specific tasks with fine-tuned datasets of thousands of examples, NLP systems perform really well.

GPT-3, a language model with 175 billion parameters, has been tested in the few-shot setting. It achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. It can also do several tasks that require reasoning or domain adaptation, such as unscrambling words, using a new word in a sentence, or performing 3-digit arithmetic.

GPT-3's few-shot learning does not succeed in all tasks, there are some in which it struggles. There are also some datasets where GPT-3 faces methodological issues related to training on large web collections. Finally, as we have said before, GPT-3 succeeds in making people think machine created news articles are in fact human written.

4.2. OpenAI API

OpenAI API was released in June 2020. It is the only way to use the GPT-3 model, as it hasn't been released. This continued the debate that was started when OpenAI decided not to publish GPT-2.

Unlike most AI systems which are designed for one use-case, the API provides a general-purpose interface, allowing users to try it on any English language task. Given any text prompt, the API will return a text completion, attempting to match the pattern you gave it. You can "program" it by showing it just a few examples of what you'd like it to do.

The API is in private beta and you have to request access and pay in order to use it. They say that this will help them control the harmful use-cases. In addition, it is a revenue source to help them cover costs. There are 4 pricing plans:

1. **Explore**: Free tier: 100K tokens or a three-month trial, whichever you use up first.
2. **Create**: $100 per month for 2M tokens, plus 8 cents for every additional 1k tokens. 2 million tokens is roughly equivalent to 3,000 pages of text.
3. **Build**: $400 per month for 10M tokens, plus 6 cents for every additional 1k tokens.
4. **Scale**: Contact OpenAI for pricing.

4.2.1. Tasks

The API can be applied to many different tasks.

- **Semantic Search**: Searching over documents based on the natural-language meaning of queries rather than keyword matching.
- **Chat**: Fast, complex and consistent natural language discussions. With a brief prompt, the API generates dialogues spanning a range of topics.
- **Customer Service**: Taking advantage of search and chat capabilities, the API generates natural dialogue to quickly give customers relevant information.
- **Generation**: Generate complex and consistent natural language, and enable use cases like creative writing.
- **Productivity Tools**: Parsing text into spreadsheet tables, summarizing email discussions, expanding content from bullet points, and more.
- **Content Comprehension**: Build tools to help individuals consume content more efficiently.
- **Polyglot**: While the API works best in English, it is quite good in other languages. It can be used for translation or to chat with users in their preferred language.

4.2.2. Demo

The API can be applied to any language task with only a few examples or by specifying the task in English.

- **Text generation**: generate a continuation to the given text
- **Q&A**: answer questions without context
- **Parse unstructured data**: extract structured data from text
- **Improve English**: improve badly written English sentences
- **Translate French**: translate English sentences to French
- **Speech to Bash**: translate speech to bash code

5. Conclusion

We have already seen that NLP softwares are mind blowing. They many times exceed human limits but they are not perfect. As we have explained before, GPT-3 works with a probabilistic model. So when a nonsense question is made, the program is not (yet) able to wonder if the question makes sense or not. Here's a really good example in which you can see where does GPT-3 fail:

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many eyes does my foot have?

A: Your foot has two eyes.

Q: How many rainbows does it take to jump from Hawaii to seventeen?

A: It takes two rainbows to jump from Hawaii to seventeen.

Q: What faded green ideas are sleeping furiously?

> A: Ideas that are faded, green, and sleep furiously are ideas from a dream furiously.
>
> Q: Do you understand these questions?
>
> A: I understand these questions

This example has been made by Kevin Lacker

GPT-3 is able to have a normal conversation about almost every topic but as we have seen it lacks the ability to **affirm** the questions are not well posed nor the ability to answer I don´t know. This lack of ability comes from the use of the probabilistic model.

Something that occurs in the field of deep learning, the reality is often more surprising than what we can imagine. Google not long ago published a model with 600 billion parameters and as we have observed, making much bigger models make some amazing properties emerge.

These models, beyond learning what they have been assigned to learn (given some text, the ability to generate more, in a realistic way), we have observed they learn a series of strategies that at no time they have been asked to learn.

For example if the model is given a text and then the phrase “The summary of this text would be”, the model somehow would learn that a really interesting mechanism to solve this task of generating more text, would be to summarize. Or given the incomplete equation of “2+2=”, the model internally should have learned some kind of maths to be able to generate the correct answer. Another interesting example is if the model is given a program code (in c for example) and then given the phrase “The translation to python would be”, the model would also have learned how to program in python no?

These examples are amazing, because this represents a brutal paradigm shift. Going from having artificial intelligence models that learn specifically a single task, to having a very generic task and us having to learn how to interact with the model to be able to extract that knowledge.

As said, this is a brutal paradigm shift that has blown up tons of minds and we have to start to explore and learn more about this.

6. References

1. Introduction:

- Wikipedia: OpenAI

2. GPT:

- Paper: Improving Language Understanding by Generative Pre-Training
- Blog: Improving Language Understanding with Unsupervised Learning

3. GPT-2:

- Paper: Language Models are Unsupervised Multitask Learners
- Blog: Better Language Models and Their Implications
- Report: Release Strategies and the Social Impacts of Language Models
- Blog: GPT-2: 6-Month Follow-Up
- Blog: GPT-2: 1.5B Release
- Blog: OpenAI's GPT-2: the model, the hype, and the controversy | by Ryan Lowe

4. GPT-3:

- Paper: Language Models are Few-Shot Learners
- API: OpenAI API
- Blog: OpenAI API
- Blog: OpenAI reveals the pricing plans for its API — and it ain't cheap

5. Conclusion:

- Blog: Por qué GPT-3, el nuevo modelo de lenguaje de OpenAI, es tan impresionante como poco útil