Gmail

LA Annie <anne.hong@gmail.com>

U.S. Blocks AI Chip Sales to China, Joe Rogan Meets Steve Jobs (Virtually), Massively Multilingual Translation, Smart Farms
1 message

The Batch @ DeepLearning.AI <thebatch@deeplearning.ai>
Reply-To: thebatch@deeplearning.ai
To: anne.hong@gmail.com

Wed, Oct 19, 2022 at 1:20 PM

View in browser

DeepLearning.AI
THE BATCH
October 19, 2022 What Matters in AI Right Now
Subscribe Tips

Dear friends,

Is prompt engineering — the art of writing text prompts to get an AI system to generate the output you want — going to be a dominant user interface for AI? WIth the rise of text generators like GPT-3 and AI21's Jurassic and image generators such as DALL·E, Midjourney, and Stable Diffusion, which take text input and produce output to match, there has been growing interest in how to craft prompts to get the output you want. For example, when generating an image of a panda, how does adding an adjective such as "beautiful" or a phrase like "trending on artstation" influence the output? The response to a particular prompt can be hard to predict and varies from system to system.

So is prompt engineering an important direction for AI, or is it a hack?

Here's how we got to this point:

- *Scraping large amounts of text or text-image data from the web enabled researchers to train text-to-text or text-to-image models.*
- *Because of this, our models expect text as input.*
- *So many people have started experimenting with more sophisticated prompts.*

Some people have predicted that prompt engineering jobs would be plentiful in the future. I do believe that text prompts will be an important way to tell machines what we want — after all, they're a dominant way to tell other humans what we want. But I think that prompt engineering will be only a small piece of the puzzle, and breathless predictions about the rise of professional prompt engineers are missing the full picture.

Just as a TV has switches that allow you to precisely control the brightness and contrast of the image — which is more convenient than trying to use language to describe the image quality you want — I look forward to user interfaces (UIs) that enable us to tell computers what we want in a more intuitive and controllable way.

Take speech synthesis (also called text-to-speech). Researchers have developed systems that allow users to specify which part of a sentence should be spoken with what emotion. Virtual knobs allow you to dial up or down the degree of different emotions. This provides fine control over the output that would be difficult to express in language. Further, by examining an output and then fine-tuning the controls, you can iteratively improve the output until you get the effect you want.

So, while I expect text prompts to remain an important part of how we communicate with image generators, I look forward to more efficient and understandable ways for us to control their output. For example, could a set of virtual knobs enable you to generate an image that is 30 percent in the style of Studio Ghibli and 70 percent the style of Disney? Drawing sketches is another good way to communicate, and I'm excited by img-to-img UIs that help turn a sketch into a drawing.

Likewise, controlling large language models remains an important problem. If you want to generate empathetic, concise, or some other type of prose, is there an easier way than searching (sometimes haphazardly) among different prompts until you chance upon a good one?

When I'm just playing with these models, I find prompt engineering a creative and fun activity; but when I'm trying to get to a specific result, I find it frustratingly opaque. Text prompts are good at specifying a loose concept such as "a picture of a panda eating bamboo," but new UIs will make it easier to get the results we want. And this will help open up generative algorithms to even more applications; say, text editors that can adjust a piece of writing to a specific style, or graphics editors that can make images that look a certain way.

Lots of exciting research lies ahead! I look forward to UIs that complement writing text prompts.

Keep learning!

Andrew

News

AI Chips Spark International Tension

New U.S. restrictions on chip sales aim to hamper China's AI efforts.

What's new: The U.S. government published sweeping limits on sales of processors that involve U.S. designs and technology to Chinese businesses. U.S. officials stated that the restrictions are meant to prevent China from militarizing AI.

New rules: The rules block sales of certain processors as well as U.S.-made equipment used to design and manufacture them. This includes high-end graphics processing units (GPUs) and other processors optimized for machine learning.

- The rules apply to chips capable of processing and interconnection speeds on par with Nvidia's flagship A100 GPU, which is designed to be used in data centers. (Nvidia supplies 95 percent of China's AI chips.) The less-capable chips typically found in personal computers and video game consoles are not restricted.
- The restrictions prohibit sales to Chinese companies of advanced chips produced using U.S.-made software and hardware as well as sales of the equipment itself. This goes for companies anywhere in the world.
- They also bar U.S. citizens and permanent residents from supporting development or manufacturing of advanced chips without permission from the U.S. government.

China's response: A spokesperson for China's foreign ministry accused the U.S. of abusing export-control measures to target Chinese firms, stating that it would hinder global cooperation and supply chains.
Behind the news: The restrictions initially came to light in September, when Nvidia and AMD independently alerted shareholders that the U.S. had imposed controls on their most advanced products. However, their details became publicly available only last week. They represent a significant escalation of earlier U.S. efforts to thwart China's ambitions in advanced technology.

- In May 2020, the U.S. required foreign chipmakers that use U.S. equipment to obtain permission to do business with the Chinese tech giant Huawei.
- In 2019, the government blocked U.S. firms from selling equipment to Huawei and 114 of its affiliates.
- In 2015, the country barred Intel from selling high-end chips to the Chinese military.

Why it matters: China has announced its ambition to become the global leader in AI by 2030, and this requires access to cutting-edge processing power. The most advanced chips are manufactured in Taiwan and South Korea using chip-fabrication equipment made by U.S. companies, and the leading chip designers and makers of chip-design software reside in the U.S. This gives U.S. authorities a tight grip on other countries' ability to buy and make chips. China's effort to build domestic capacity to produce advanced semiconductors — which are hampered by the sheer difficulty and expense of etching features on silicon measured in nanometers — now faces additional hardware, software, business, and talent hurdles.

We're thinking: International cooperation has been essential to recent progress in AI. As barriers rise between the U.S. and China, the AI community must navigate a world where geography will have a much bigger impact on access to ideas and resources.

Smarts for Farms

The next green revolution may be happening in the server room.

What's new: Microsoft open-sourced AI tools designed to help farmers cut costs and improve yields.

How it works: FarmVibes-AI includes systems that analyze overhead imagery and sensor data to guide farm operations.

- AsyncFusion uses drone imagery, satellite imagery, and data from soil sensors to map soil conditions in real time. Farmers can use the output to plan where and when they should plant their fields.
- DeepMC is a neural network that combines data from soil sensors, climate sensors, and weather predictions to forecast field temperature, precipitation, and soil moisture up to 120 hours ahead. Its output can enable farmers to prepare for extreme temperatures and other events.
- SpaceEye, another neural network, filters clouds from satellite imagery for use by AsyncFusion and DeepMC. Microsoft engineers trained the network via an adversarial method using infrared and visible-light images partly covered with synthetic clouds.

Behind the news: Nonprofits and academic institutions provide other open-source AI systems to increase food production in collaboration with large agribusiness firms, independent farmers, and rural communities.

- Last year, the Linux Foundation launched Agstack, a partnership among universities, nonprofits, and IBM. The effort provides code, data, and frameworks to developers of open-source AI projects for agriculture.
- MIT's now-defunct OpenAg included models that predicted how plants would grow under various environmental conditions.

Why it matters: The emerging practice of precision agriculture, which seeks to take into account not only entire fields but also local conditions down to the level of individual plants, could help farmers sow seeds, grow crops, fight pests, and harvest produce more efficiently. Off-the-shelf systems may not serve farmers who work in different parts of the world or grow niche crops. Open-source projects can expand their options effectively and inexpensively.

We're thinking: Farmers tend to welcome innovations that improve yields and cut costs. They're also famously self-sufficient, performing repairs and installing upgrades to their equipment. As self-driving tractors and precision-ag systems take root, they're great candidates to become early adopters of industry-focused platforms that make it easy for anyone to build useful AI applications.

All Synthetic, All the Time

Joe Rogan meets Steve Jobs in an AI-generated podcast.

What's new: For the debut episode of a new podcast series, Play.ht synthesized a 19-minute interview between the rock-star podcaster and late Apple CEO. You can hear it here and propose computer-generated participants in future episodes here.

How it works: The Dubai-based text-to-speech startup created the podcast using text generation and voice cloning.

- Play.ht generated the script using an unnamed natural language model that it fine-tuned on Jobs' biography, interviews, and other sources.
- It rendered the transcript into audio using proprietary synthetic voices trained on audio recordings of each speaker. Play.ht's voice editor synthesizes voices in over 120 languages with phonetic control over pronunciation.
- The production is the first in a series called Podcast.ai. The public can propose meetings of the virtual minds for future episodes.

Behind the news: Rogan was also the subject of an early experiment in voice cloning. In 2019, Toronto-based Dessa released ersatz Rogan audio clips — the first of a parade of fake celebrity voices.

- Earlier this year, James Earl Jones, the voice of Darth Vader, signed a deal that permits Disney to recreate the Star Wars villain's speech using technology from Ukrainian startup ReSpeecher.
- Two documentary filmmakers separately generated vocal facsimiles of deceased celebrity chef Anthony Bourdain and iconic artist Andy Warhol. The Bourdain imitation sparked controversy when his widow revealed that she had not given the filmmaker permission to recreate her husband's voice.

Why it matters: The declamation is occasionally stilted and the script meandering (with occasional lapses into incoherence), but the rapid progress of generative audio combined with the entertainment world's appetite for novelty suggests that satisfying synthetic productions may not be far off.

We're thinking: How long before we can produce *Heroes of Deep Learning* without actually talking with any of the heroes of deep learning?

Massively Multilingual Translation

Sentence pairs that have equivalent meanings in different languages — typically used to train machine translation systems — have been available in sufficient quantities for only around 100 languages. New work doubled that number and produced a more capable model.

What's new: Marta R. Costa-jussà and colleagues at Meta, Johns Hopkins, and UC Berkeley developed an automated process for scraping multilingual sentence pairs from the web. They released No Language Left Behind (NLLB-200), a machine translation model that handles 200 languages. They also released the models, code, and data used to build it.

Key insight: The web is full of text in various languages, including sentences that have the same meaning in different languages. For instance, unrelated pages in different languages may say the equivalent of, "Manchester United defeated Melbourne in yesterday's match," or "A long time ago in a galaxy far, far away." An automated system can recognize such parallel sentences by learning to produce similar representations of sentences that have similar meaning regardless of their language. A teacher/student arrangement — with a multilingual teacher trained on languages with plentiful data to produce embeddings, and a separate monolingual student for each language scraped from the web — can align representations produced by the students.

How they built the dataset: The authors identified languages in text scraped from the web, trained a teacher model on pre-existing multilingual data, and used it to train a student model to produce similar representations for similar meanings in the web text.

- The authors trained fasttext, a linear classifier, to classify text according to its language. They trained it on publicly available datasets and their own corpus of 6,000 human-translated sentence pairs in 39 languages (released with this paper).

- Fasttext classified the language of individual sentences and full paragraphs in web-text corpora such as Common Crawl and ParaCrawl. The authors discarded sentences if their classification didn't match that of the paragraph and removed sentences in languages for which they already had a lot of parallel data. After deleting duplicates, they had 43.7 billion sentences, each labeled as one of 148 languages.
- They trained a separate transformer — a student — on each language (or several similar languages) to produce similar representations for sentences with similar meanings. To do this, they trained a Bidirectional LSTM — the teacher — to translate between the 93 languages in the OPUS dataset. This model learned similar representations of equivalent sentences in different languages. Using publicly available datasets of parallel sentences, the teacher received a sentence in one language (usually English) while a student received the equivalent sentence in its designated language(s). The students learned to maximize the cosine similarity between the teacher's and students' representations. Simultaneously, the students were trained to fill in missing words of sentences in their designated language(s).
- The authors discarded sentence pairs if their representations' cosine similarities were too different, leaving 1.1 billion parallel sentence pairs. Combined with pre-existing datasets, the parallel sentences represented 202 languages.

How they built the translator: NLLB-200 is a transformer encoder-decoder that comprises 54.5 billion parameters.

- In every fourth transformer layer (made up of a self-attention sublayer and a fully connected sublayer), the authors exchanged the fully connected sublayer with a Sparsely Gated Mixture-of-Experts (MoE) sublayer that activated only a subnetwork of neurons for each input. This enabled the network to learn to activate different portions depending on the language, which may have helped to prevent learning about languages that had many examples from interfering with learning about languages that had few.
- Training proceeded in two stages. In the first stage, NLLB-200 filled in missing words in sentences and translated between pairs of sentences in different languages. In the second, it trained only on translations. In both stages, the paired sentences included human-translated sentence pairs, sentences scraped from the web and paired automatically, and back translations in which the model converted its own translations back to the original language.

Results: The authors' NLLB-200 model achieved 24.0 average spBLEU across all 202 languages, while the earlier DeltaLM achieved a 101-language average 16.7 spBLEU (which measures the overlap of word fragments between machine translations and ground truth, higher is better). A sparse NLLB-200 that used MoE rather than fully connected layers generally performed better than a dense version. For example, evaluated on Akan, a language spoken in Ghana for which little training data was available, the sparse model scored 36.2 chrF, while a dense version scored 35.6 chrF (which measures overlapping groups of consecutive characters between machine translations and ground truth, higher is better). NLLB-200 performed inconsistently compared to bilingual models: It achieved 36.2 chrF compared to an English-to-Akan model's 16.8 chrF, but 51.4 chrF compared to an English-to-Gujarati model's 51.7 chrF. A possible explanation: Languages that are dissimilar to other languages in the training data may not benefit as much from multilingual training.

Why it matters: Faced with an apparent scarcity of data, the authors extracted it from the web. The data didn't need to be perfect: To compensate for flaws such as typographical and grammatical errors, the model learned to convert its own translations — of flawed sentences but presumably many more correct ones — into good sentences.

We're thinking: University of Texas machine learning professor Raymond Mooney said, "You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector." Apparently these researchers did it!

Work With Andrew Ng

Senior Controller: Woebot Health seeks a controller to help drive the development of processes, technology, compliance, and reporting and ensure that data is available for decision-making. The ideal candidate has 10-plus years of experience. MBA and CPA preferred. Apply here

Director of Finance and Accounting: Landing AI seeks a finance leader to oversee and manage all aspects of financial operations and tax planning. The ideal candidate has a finance and accounting background, experience with managing debt and raising equity, and knowledge of product-led growth strategies in business-to-business software as a service. Apply here

Subscribe and view previous issues here.

Thoughts, suggestions, feedback? Please send to thebatch@deeplearning.ai. Avoid our newsletter ending up in your spam folder by adding our email address to your contacts list.

DeepLearning.AI, 195 Page Mill Road, Suite 115, Palo Alto, CA 94306, United States