EXHIBIT J

Private channel #explainlikeimfive - 2022-02-07 (UTC)

Private

February 7th 2022, 7:21 PM UTC

Alethea Power, Ben Wang, Boris Power and 22 others

E

Eli5-Bot

Explain like I'm 5: A few related questions:
- How is it reasonable that our models only do two passes over even the highest quality educational content?
- Or alternatively, that they do the same number of passes on https://en.wikipedia.org/wiki/Solar_eclipse_of_April_5,_1837 as they do on https://en.wikipedia.org/wiki/Bayes%27_theorem?
- Is backpropagation really not able to absorb more knowledge from many repeated passes on dense subjects?
- What are some of our most promising ideas for spending more compute to dig deeper into profound but sparsely-covered topics?

2/7/2022, 7:21 PM UTC

2 1

YB

Yura Burda (OpenAI)

There is no fundamental reason why seeing the same thing twice has to make our models learn more from it - if the model has enough capacity to memorize the answers, it can do so on the first pass, and any generalization that we get is a little miracle (that we should try to understand better, but currently don't understand very well). The paper "Understanding deep learning requires rethinking generalization" has some nice evidence that in practical situations our networks really do have enough capacity to just memorize the training data without generalizing.

With 5 year-olds like yourself, we often try exposing the child to the same concept in a myriad different ways until we believe the child has achieved non-shallow understanding, as well as structure the learning along a carefully crafted curriculum where important ideas are introduced incrementally - we don't try to explain the meaning of simple words by reading texts on calculus to the child, but rather wait until the only new information in the calculus book relative to the child's knowledge is the ideas of calculus.

However we haven't yet gotten to a similar level of data efficiency with neural nets as we have with children, so this strategy seems not viable with neural nets of today.

One day we will have more data efficient architectures that can ponder the data presented to them for longer and generalize much further - perhaps something larger, deeper, more recurrent, regularized, or otherwise imbued with inductive biases would do the trick. We then can ask the networks to observe their own learning process, generalize from past observations, and ask them to create the data for themselves or look up relevant information if it exists somewhere. At that point we will likely have AGI and our main concern will be whether we can guide its learning process towards human-compatible concepts, motivations, and behaviors.

2/7/2022, 7:39 PM UTC

8 H 1

MB

Miles Brundage

Lol at "with 5-year olds like yourself". Always amazed how many super smart 5 year olds work here

2/7/2022, 7:54 PM UTC

12

JT

Jerry Tworek

It's a topic I think about it a lot. Thinking about it this way - human can read a dense math book many times and finally understand the topic of the book. Large overparametrized neural networks can see a book a handful of times and they'll just memorize it, without much understanding and generalizable knowledge. To build any reasonably "robust" understanding, the NN would have to read ~millions of such books, each one phrased differently, using different words, examples etc.

It's clearly a weakness of deep learning. Why is that? We train our networks to **memorize the training data** - that's their objective. If we train large enough model on any dataset large enough number of times - it will perfectly memorize it.

Usually gaining some generalizable knowledge along the way, but less than a human would do, reading through the same amount of text.

Where does that generalization in networks comes from if we just train to memorize? Based on occam's razor and information theory, a generalizable knowledge is a relatively small representation/part of a neural network that explains well a large part of dataset. Because our networks have lots of parameters and we train them to memorize the dataset, naturally some of those small generalizable subnetworks will emerge throughout training. Moreover, because their efficiency is high (high explanation power of the dataset/number of patameters) they'll mostly keep getting amplified at least until the rest of the network doesn't memorize the dataset well enough to explain the dataset better from the memory.

It's interesting to draw a parallel to humans here because humans also just read books to learn. Would humans with perfect memory stop learning from books? Hard to say. Probably a human with perfect memory seeing a book written in a language that they've never seen wouldn't have learned from it. But generally when reading a book, because of our weak and fallible memory we just build a small representation of its contents in our heads as fully memorizing it is exhausting and impractical.

Coming back to neural networks, how are able at all to go over dataset multiple times? It comes from another interesting property of deep learning - **catastrophic forgetting**. It does happen that whenever we try and teach our models to memorize something, they forget a little about everything else. So whenever a model has seen something at some point, and memorized some bits of it, it largely has forgotten specific bits of it but some small generalizable subnetworks have survived because they explained well other datapoints as well. Then, the model can read a particular book again, building deeper general understanding without memorizing too much.

These topics are very central to what we do at OpenAI and long training, repeating data, regurgitation, scaling laws and network architecture all in some way touch upon that. We study those things because we want to train the best generalizable neural networks using the limited data that we have, but there are still much more open questions than definitive answers.

To summarize, there are clearly some similarities in how networks learn to how humans learn, but they generally have much better memory than humans, understand the text much worse than humans and lack the bits of internal monologue that we humans do when reading, that constrain us with how many times we can show them the same bit of data.

2/7/2022, 8:31 PM UTC

5 o 1

DS

Daniel Selsam (OpenAI)

I'll share an experience that has some connections to @Jerry Tworek's comment. Many years ago, I was reading Baby Rudin (undergrad analysis textbook) and was struggling to follow many of the proofs, which would read like "... Thus the result follows trivially by 3.12, 3.14, 3.26, 4.11 and 4.26". By the time I reminded myself of all these statements, I would forget enough that I still struggled to see why the original result followed from them. So, I decided to memorize the entire textbook using spaced repetition software. For each theorem number, I would force myself to remember the statement and the proof. I found that the first time I

did this for each theorem, it would be gruelling, and would require many glances and restarts before I could make it through from memory. When I did finally make it through, what I recited would be basically verbatim to the textbook. But, each time the card was scheduled (exponentially increasing number of days between repeats), I found it took a little less energy to remember, and my recalled explanation of the proof was a little more abstract and concise. After about a month, the proofs in my head would become essentially sub-verbal and would seem too trivial to bother stating in words.

My interpretation is: the brain is an energy-minimizer, and memorizing text (especially proofs) requires a ton of energy for humans. So, just as the body will grow muscle in response to forceful exertion, the brain will synthesize more compact and efficient (and hence generalizing) representations in response to the need to recall. @Ilya Sutskever's doc "Gradient Descent is Compressing" (<https://drive.google.com/file/d/1EzCTJAm0dr0jDRAwXMEc2Sm64HxYuGdD/view>) made me cautiously hopeful that perhaps backpropagation might do something similar under the right regimes. 2/7/2022, 10:14 PM UTC

sgd_is_compressing.pdf 2/7/2022, 10:14 PM UTC

IS

Ilya Sutskever

solving "slow learning", i.e., finding some procedure for consuming data very slowly, while spending lots of compute on each word/sentence to understand it more deeply, will be incredibly impactful. The VLB approach pursued by the GPTZero-VLB project (gpt-zero-vlb-chatter) may give us that. Other strategies (naive multiepoch training) may also give us some of that. 2/7/2022, 10:21 PM UTC

IS

Ilya Sutskever

solving "slow learning", i.e., finding some procedure for consuming data very slowly, while spending lots of compute on each word/sentence to understand it more deeply, will be incredibly impactful. The VLB approach pursued by the GPTZero-VLB project (#gpt-zero-vlb-chatter) may give us that. Other strategies (naive multiepoch training) may also give us some of that. 2/7/2022, 10:21 PM UTC

RG

Ryan Greene

@Ryan Greene has joined the channel 2/7/2022, 10:54 PM UTC