When OpenAI tried to build more than a Langauge Model

deliprao.com/archives/314

February 18, 2019

If you were part of the machine learning Twitter, last Thursday, it was impossible to miss OpenAI's press release of their new GPT-2 model and all the heated Twitter conversation about that. Many people tried to summarize it in their own ways. A lot of these posts are of "He said, She said" flavor. While doing so is important, I will not focus on that here. Instead, I will focus on what I think were the real issues and, more importantly, where we go from here. This is a long post, but I think the length is justified as we need to step back from the rapid-fire Tweet mode and shine the light inward.

Backstory

Many of us on last Thursday woke up to research updates coming not from ArXiv or conference proceedings but from the Verge.

The Verge story was referring to a cool-looking OpenAI blog post also published in tandem. I was curious, so before I dug deeper into the blog post or the Verge article, I dove into the paper.

The GPT-2 paper is a logical extension of the original Generative Pre-Training (GPT) paper, also from OpenAI, which in turn builds on a variety of approaches invented at places outside OpenAI, such as the Transformer papers from Google. All of these are amazing works in their own right, and it is important to understand them in context. The Transformer paper, and the recent Transformer-XL paper, is pivotal. It allowed us to capture long-range dependencies in a way that was not possible with existing sequence architectures such as LSTMs and GRUs. This will become handy to understand some of OpenAI's recent success in generating long coherent passages. The GPT paper first showed how a Transformer-based language model can be applied in a general way for a variety of natural language processing tasks. While this is also not the first time anyone had done language model transfer, the GPT paper was first to show this in the context of the Transformer and for a large number of tasks.

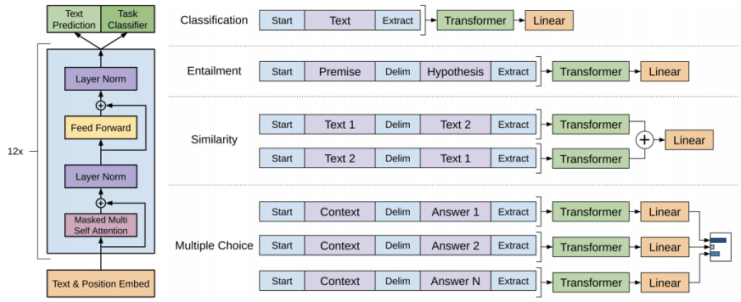


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Courtesy: Radford et al (2018), "Improving Language Understanding by Generative Pre-Training"

The GPT paper was truly phenomenal when it came out in July 2018, and briefly occupied the state-of-the-art (SOTA) status until BERT, another Transformer based language model transfer approach from Google, came out in Oct 2018 and swept most benchmarks and attention ("Attention is all you need?").

With all this context, it is easy to see how the GPT-2 paper is a logical next step. But make no mistake. It is also a marvelous feat of engineering that only a few today are capable of pulling it off. I will talk more about this later, but it is worth looking at GPT-2 in isolation and see how well-written that paper is.

Happy Valentine's Day

Then we find ourselves in 2019 and get a Valentine's Day gift from OpenAI — the GPT-2. But the gift doesn't come directly to us. It comes through a closed group of press core hand selected by OpenAI. The Verge article was published at 9 am, optimized for peak consumption. OpenAI's tweet about their blog post reaches us at noon PST for folks who missed this to catch up during the lunch. Interesting way of doing science! I mean I get it. OpenAI was always a PR-forward company. That's fine. Except for this time, what became complicated was they tried to do multiple things at once:

- Communicate new research (Paper + Blog post)
- Change the way the research is communicated
- Come up with a new model-release policy and pilot it
- Come up with a narrative around the model-release

Let's look at each of these in detail.

Communicating new research (the blog post)

DeepMind pioneered the art of releasing blog posts to accompany their papers. OpenAI took that to an art form. It is slick, it is beautiful. I personally enjoy reading them. But the Feb 14th blog post raised more questions to me and many others. The blog post has three parts: 1) Technical results, 2) Policy stuff, and 3) Model Release-strategy. The technical results part of the blog was pretty straightforward, and my review of the paper earlier covered that with additional context than what was in the OpenAI blog post. I want to spend some time on the Policy and Model-release parts of the post.

"Policy Implications"

> We can also imagine the application of these models for malicious purposes, including the following (or other applications we can't yet anticipate):
>
> - 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

Source: OpenAI blog.

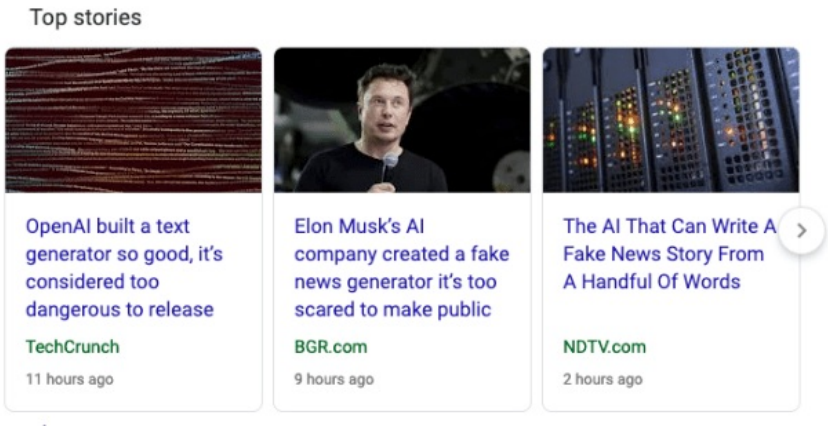
The reasoning in this section is, 1) "technologies are reducing the cost of generating fake content" and 2) Malicious actors will use that to generate large volumes of content for abusive/harmful purposes. While this linear narrative appeals to the human mind, let's examine it closely.

1. Yes, we can generate long seemingly coherent discourses. But these discourses are not topically coherent enough to make full-length articles. Some contend that this is a matter of time. Maybe. But as we will see that point is moot.

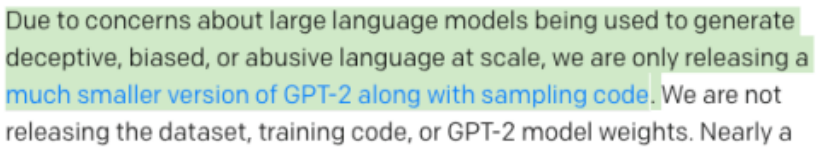
2. Given any technology, there will be abuses of it. I'm sure there's a named law for that. One humbling lesson I learned working with fact-checkers and journalists during the FakeNewsChallenge was, in the case of online misinformation and disinformation campaigns, technology is never a limiting factor. Human labor is so cheap around the world, that with sufficient motivation, the end goals of spreading misinformation or impersonating online can be accomplished very successfully without frontier technologies.

But can't such enabling technologies be misused to be malicious at scale?

Yes, technology will lower the bar to some extent, but some of the strongest signals for fighting online adversaries ranging from spam to misinformation actually comes from non-content signals. This is from my personal experience working on Twitter's spam systems, from what I know from Facebook's work, and from my personal experience working computational fact-checking. There is no denying that technology will have a similar catalytic effect as it has in every other realm of human existence, but we need to be cautious before calling "wolf" in this situation because such efforts shifts focus disproportionately from the reality of the situation. Catchy titles in newspapers will lead us to believe Terminator-like scenarios are nigh, cause hysteria, and make us blind to other viable short- and long-term solutions.



"Model-release strategy"



Many have praised this as "thoughtful" and as "norm-setting". I find OpenAI is flirting with a new way of communicating results that could undermine science by setting a precedent in the wrong direction if this should become a "norm".

Imagine this scenario:

You have some results. You hand-pick a bunch of friends and people you know to look into it. You "invite" a bunch of reporter friends to get an early peak and to "break" a story. The reporters feel special and fuzzy for having been invited. Then you prime the reporters with what you want them to hear. You show them examples and let them play with models trained from one of the most divisive and trolling places in the world — Reddit. You instruct the reporters to break the story in tandem with your blog post. You ask them not to discuss it with anyone else before the story breaks.

That is exactly what OpenAI did, and why many folks got annoyed by this action.

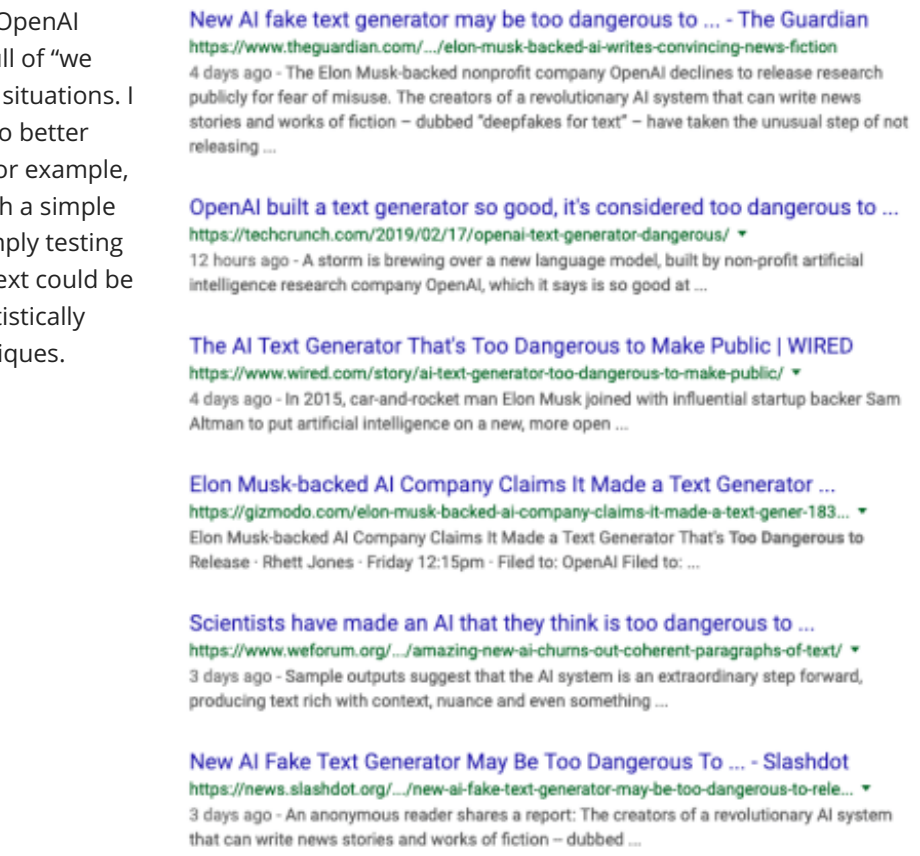
How ethical is this kind of reporting?
How ethical is the company that incentivizes this kind of reporting?
How objective are the stories themselves?
... so many questions arise.

E is for Embargo?

No. This so-called "embargo" is actually not an embargo. The paper is released. The code is released. They did not release any data or full model weights, but the paper describes how to put something like it together if you are sufficiently motivated. If you are not that motivated, fear not, Google released their clean STORIES corpus yesterday. OpenAI trained on a 40G corpus. The STORIES corpus is 32G. Not too far from the data OpenAI didn't release. All you need is compute, and if you do that math, that too is not prohibitory even for motivated individuals. For governments and large organizations, that cost is chump change.

All this embargo-posturing has resulted in a whole lot of fear-mongering. We will still be dealing with ripples of this bouncing back and forth in mass media with titles like:

What exactly is “too dangerous”? The OpenAI press release is full of “we imagine” for such situations. I am sure we can do better than imagining. For example, we could start with a simple experiment of simply testing if the generated text could be discriminated statistically using SOTA techniques.



Another alternative to this “imagination” is to work with experts in the field of disinformation and misinformation (and they’re not journalists), to see what exactly are the problems.

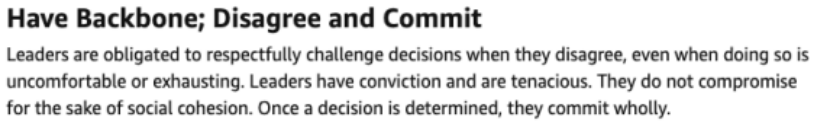
Another problem with this “this is too powerful to be released in the wild” argument is: Do only powerful AI models be regulated and under embargo? Can we not create harm with logistic regression? What about vanilla algorithms? This is not a thought experiment. It is happening today, with our credit models, with our predictive policing models, you name it.

E is for Engagement?

Between Feb 14th and today, 29 NLP papers and 136 vision papers came out on ArXiv, and we are not talking about any of them but this. Part of my hope in writing this post is to get past it. While it is unlikely true, I can’t help but feel if OpenAI has simply weaponized the dual use narrative to gain more press and more mindshare. It’s a battle they can’t lose. If this mass adverse reaction did not happen on Twitter they would win. If it did, as it has, they still win for “creating a conversation”. As my friend Mark Riedl says, OpenAI is in a Win-Win-Win situation here.

All this sounds harsh. My Twitter responses were direct and not couched in niceties, and that was necessary at the time to cut through the noise of the medium. It is a good day to remind myself of my own thought some time ago.

As I continue working on malicious use of AI, particularly in misinformation and disinformation, I will no doubt collide with OpenAI, learn from them, compete with them, and maybe one day even collaborate with them. From my time working on Alexa at Amazon, I really took this leadership principle to heart.



If I have the audacity to suggest, perhaps OpenAI should make "inclusivity" explicitly as a part of their charter and practice. Perhaps having a diversity in thoughts and opinions about their work might have yielded different outcomes. Perhaps acknowledging all people who came before them not just in technology, but also in policy, including them and standing on their shoulders might help. Perhaps be a little more … open? Fighting misinformation is a long game. AI safety is a long game. It requires a coming together of people from all disciplines, not secret meetings and railroading policies.