~~Attention~~ An Undergrad Is All You Need

James Yoo
Department of Computer Science
University of British Columbia
yoo@cs.ubc.ca

Abstract

The mechanism of self-attention has generally displaced the large convolutional neural architecture commonly used for tasks adjacent to natural language understanding. Specifically, Transformer models that exploit self-attention have been leveraged with surprising success in large-language models such as LaMDA and GPT-3. However, these large-language models are expensive to train, require large amounts of training data, and are prone to hallucination. In this paper, we introduce GPT-UGRD, a novel autoregressive architecture that requires minimal training and comes ready out-of-the-box for multi-modal learning with a modest watt-per-token power consumption. We show that it performs equivalently to, or better than the state-of-the-art, reporting an average BLEU score of 69.420.

1 Introduction

Transformer architectures that exploit the mechanism of self-attention [1] have recently seen a meteoric rise in popularity, particularly with models that are accessible to the general public such as ChatGPT [2]. The pre-trained transformer architectures found in large-language models increasingly appear to be the way forward to achieving near-human performance on natural language processing (NLP) tasks, with some models already exhibiting near-human performance while minimizing errors and risk [3, 4, 5, 6]. Unfortunately, pre-trained large-language models require copious amounts of training data and highly sophisticated training pipelines. We express the number of problems as $n = 2$, where n is a *conservative* estimate of the true number of actual problems (n_{true}) posed by this. We suspect that n_{true} is much larger, but will leave the calculation of this value to the reader.

The first problem, related to the metaphoric firehose of data required to train models, is one of bias and toxicity. There is no tractable mechanism in which data modellers are able to sift through and validate the training data, either via manual or automated methods. The second problem is linked to the gargantuan amount of compute that is used to train models. Most training for large-language models is conducted either as long-running processes distributed across physical data centers with specialized application-specific integrated circuit (ASIC) hardware [7] developed for machine learning workloads (e.g., massive high-performance GPU clusters, Tensor Processing Units). These approaches to training models are not realistically accessible most individuals.

Given these problems, we propose a new model called GPT-UGRD a multi-modal generative system that is capable of continual learning while requiring a reduced amount of supervision and explicit learning. We show that it performs as well the state-of-the-art in generative models. We also show that biases and hallucinations in GPT-UGRD can be more easily mitigated than in existing large-language models with a single training session lasting only a few hours without the need to designate additional compute capacity.

The main contributions of this paper are as follows:

- We introduce GPT-UGRD, a multi-modal generative system that is capable of continual learning with minimal supervision.

17th Conference of the ACH Special Interest Group on Harry Quadratosquamosal Bovik (SIGBOVIK 2023).

- We evaluate GPT-UGRD on common tasks dispatched to large-language models, and compare its performance to the state-of-the-art in pre-trained large-language models.

We begin by describing the architecture of GPT-UGRD in Section 2 and detail its evaluation against the state-of-the-art in large-language models in Section 3. We summarize our efforts in developing GPT-UGRD, and discuss future work in Section 4.

2 GPT-UGRD

Figure 1 provides a general overview of the architecture of GPT-UGRD. The user interacts with a patented Load Balancer[1] that is encircled by an electromagnetic network layer. The network layer is built upon a harmonic, gluten-free substrate that effectively eliminates the vanishing gradient problem. Undesirable interactions between the Load Balancer and the Secure Backroom are mitigated by a sinusoidal secure transport protocol (SSTP), which requires GPT-UGRD to pass an exam requiring them to issue a zero-knowledge proof, which they may retake every quarter.

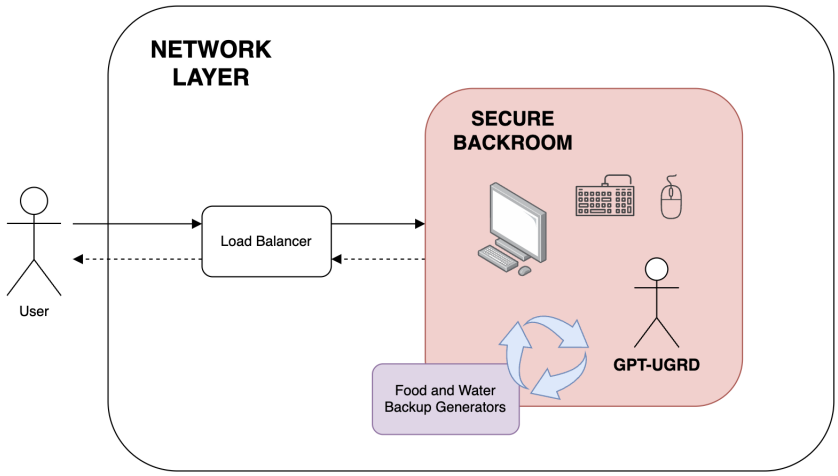

Figure 1: The GPT-UGRD architecture. The Load Balancer directs requests to the appropriate instance of GPT-UGRD, which is secured in a backroom with a computer, mouse, keyboard, and a recycled supply of food and water.

2.1 Prompt Encoding

Upon receiving a prompt from the Load Balancer, GPT-UGRD immediately begins encoding the full text of the prompt into a search query via a natural Variational Autoenencoder (nVAE) (Figure 2), for (nearly) free. We observe that this encoding is performed by GPT-UGRD by a process called "actually thinking about keywords in a query" (ActTHNKWRDQRY) which we know to be a difficult task for human agents. This query is subsequently dispatched to a search engine, the results of which are parsed by GPT-UGRD.

2.2 Interaction

Much like the state-of-the-art in large-language models, GPT-UGRD can be interacted with via a front-end resembling a chat application. Figure 3 describes two sessions with GPT-UGRD. Of particular note is the realism of the conversation. Chat responses are usually instantaneous, except when they are not. For example, GPT-UGRD might be sleeping, studying for an exam, or out partying on a Friday night. These are examples of pathological behaviour that remains an open problem in the realm of generative language models in the class of GPT-UGRD which we have identified as "Weekend Problems."

[1]Load Balancer Pro Max with ProMotion Display is also available.

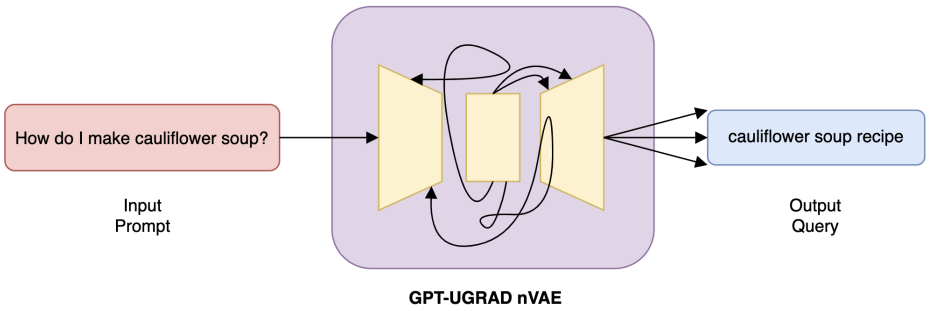

Figure 2: The prompt-to-query transformation pipeline.

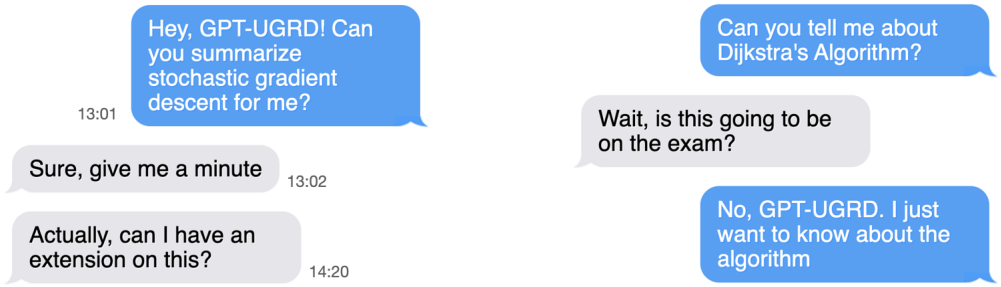

Figure 3: Two conversation logs with GPT-UGRD.

2.3 Model Maintenance

Unlike most large-language models, GPT-UGRD does not require huge amounts of training data, nor a massive amount of compute capacity. GPT-UGRD runs off a schedule of three (3) or 2.5 maintenance cycles per day. In the case of three cycles, the inbuilt Food and Water Backup Generators will generate food and water in order to nourish GPT-UGRD. In cases where GPT-UGRD does not have time for a full breakfast, the 2.5 maintenance cycle will be selected, with a mug of instant coffee being substituted for breakfast. Special maintenance is provided on one day out of the 365 that comprise a year in the form of cake[2] to celebrate the epoch date of the model.

Food	**Energy Consumption (kWh)**
Boiling two liters of water	0.23
Cooking two cups of rice with four cups of water	0.20
Simmered beef stew made from 0.9 kg of meat	1.00
Asian Stir-fried pork and eggplant with rice	0.51

Table 1: Energy Consumption for GPT-UGRD maintenance cycles.

Table 1 provides an overview of some sample maintenance cycles that are consumed by GPT-UGRD. We perform an advanced worst-case analysis using advanced mathematical techinques (i.e., addition and multiplication) of the energy required to maintain GPT-UGRD continuously for a year:

$$(0.23 + 0.20 + 1.00 + 0.51)\ \text{kWh} \times 365\ (\text{days}) = 766.3\ \text{kWh}$$

BERT [8], a language model developed by Google, requires about as much energy as a trans-American flight [5]. This does not take into account hyperparameter optimisation, which consumes additional

[2]Ingredient availability permitting

energy. We assume a trans-American flight is serviced by a Boeing 787 airliner, which burns around 7000 litres of fuel per hour, for an estimated 5 hours (New York City to Vancouver, BC), for a total of 35,000 litres per trans-American flight. Assuming 10 kWh is generated per litre, we have the total energy usage to train a BERT model:

$$35,000 \text{ L} \times 10 \text{ kWh/L} = 350,000 \text{ kWh}$$

Mathematically speaking, there is evidence to conclude that the value 350,000 is smaller than the value 766.3, which we express with the less-than ($<$) operator:

$$766.3 < 350,000$$

The proof of this equation is left as an exercise to the reader. If you find a proof, please email us so we can update the paper, I think that's allowed. `TODO: ask SIBOVIK chairs if this is allowed.` Anyway, moving on.

3 Evaluation

We evaluate GPT-UGRD on common natural language processing tasks such as sentiment analysis (Subsection 3.1) and Summarization (Subsection 3.2). You will find it hard to believe our results, Figure 5 will surprise you.

3.1 Sentiment Analysis

We compare the performance of GPT-UGRD with ChatGPT in highlighting words in the standard Richard and Mortimer (RnM) dataset [9] used in NLP benchmarking. Figure 4 describes the results of a highlighting task dispatched to both ChatGPT and GPT-UGRD. The prompt given in the task was to "Highlight the words with a negative sentiment." We observed that ChatGPT missed the word "nihilistic" in its generated highlights. This was not the case for GPT-UGRD, which generated all highlights with negative sentiment, and was rewarded with a pat on the back and a job well done.

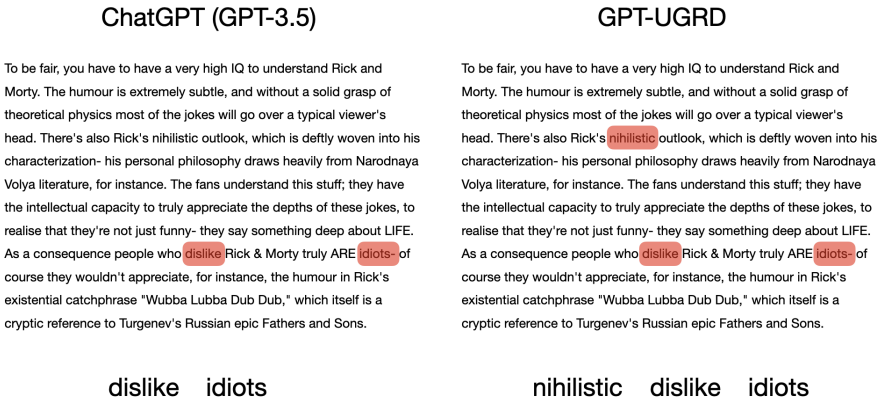

Figure 4: Highlighting task performed by ChatGPT (GPT-3.5) and GPT-UGRD.

3.2 Summarization

In the summarization task, we provide the prompt "Summarize the Wikipedia page on monads in bullet-point form." to ChatGPT and GPT-UGRD. It is obvious that summarizing the imaginary concept of a "monad" is a fool's errand. Consequently, model performance is measured by calculating the number of tokens that comprise the summary generated by each model, with fewer tokens being better, as it would be pathological for a model waste valuable compute in attempting to summarize an imaginary concept that cannot hurt anyone.

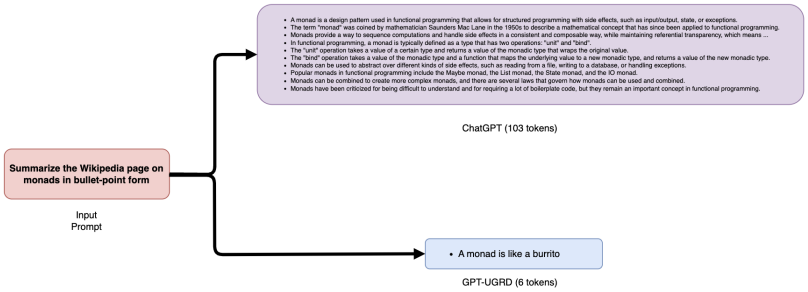

Figure 5: Summarization task performed by ChatGPT and GPT-UGRD.

Figure 5 describes the result of this task. The summary generated by ChatGPT comprises **103** tokens, while the summary generated by GPT-UGRD comprises **6** tokens. We know via the less-than operator ($<$) that the following might hold true:

$$6 < 103$$

Consequently, we can conclude that GPT-UGRD performs a magnitude of factors better than ChatGPT in summarization.

4 Discussion

In this paper, we introduced GPT-UGRD, a novel generative system that requires far less training data and explicit direction in development. We show that it outperforms the state-of-the-art in generative transformers (e.g., ChatGPT/GPT-3.5), while requiring far less energy in maintenence, training, and generated token.

Future work remains in resolving the open-problem of non-instantaneous responses (i.e., the Weekend Problem), and in scaling this nascent architecture to a wider community.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[2] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

[3] Gary Marcus. The dark risk of large language models. `https://www.wired.co.uk/article/artificial-intelligence-language`, Dec 2022.

[4] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models.

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 2021.

[6] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell,

William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 2022.

[7] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ISCA '17, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[9] Adult Swim. Rick and Morty. `https://www.adultswim.com/videos/rick-and-morty`, accessed 2023.