A BlueGranite Blog Post Written (Mostly) by AI

Colby Ford

Late last year, the brilliant minds over at OpenAI released their GPT-2 model for the world to use. This model is a state-of-the-art language generation model that can effectively write human-like text. Furthermore, their general model can be tuned to fit a specific use case. We've given the model all of the previous BlueGranite blogs to study and, in this post, I'll be sharing some of the model's writing samples.

Writing Samples

At BlueGranite, our three main solution areas are Business Intelligence (lots of Power BI), Data Platform (Data Factory, Data Lakes, Databricks, etc.), and Artificial Intelligence (my personal favorite with Machine Learning and Cognitive Services).

Let's see how the custom GPT-2 model, tuned using 300 previous BlueGranite blogs, performs on each of these areas. By the way, we lovingly named our model **BlogBot**.

Check out our code and train your own BlogBot here: github.com/BlueGranite/BlogBot

*Note: All AI-generated content is in grey.

Business Intelligence

Power BI has always been a fundamental component of the modern business intelligence solution. Whether you've grown accustomed to working with Excel and its multitude of slicers or fancy adding a visual to your data, Power BI is the tool for you. Although Power BI is typically associated with the cloud, it has come in predominantlyrds in the past, such as as at-a-glance data model creation, analysis over the past few years, and reports and reports of your operations. The first edition of the Power BI Report Publisher was developed by Microsoft; however, when the Power BI service became a standalone service, the focus shifted to Power BI Reporting Services, Power BI Copy Technology, and then to the Power BI REST API. As of September 30, 2018, Power BI Report Builder can be installed on all three of your reporting systems, and can be fully integrated with existing Reporting Services functionality. Report Builder allows you to create a simple interface where you can provide reports with basic data, built from a variety of sources and models. Report Builder allows you to connect to a dataset and interact with it via a few simple clicks. Once you have configured the reports you want to generate and have them evaluated, you can begin generating metrics and performing analysis. In my experience using Demo/Test/Dev/

Insights in Power BI, it really helps to keep everything short and sweet and steer the audience in the right direction.

Data Platform

The first step is to simulate the cloud storage service under the hood in Azure Blob Storage. We start by pulling data from the cloud and moving it through an Azure Merge Field Anomaly Detection algorithm is being applied to our .Databricks. We find common patterns in the CSV, MDX, and Word files that will help us to merge our data in the cloud. Then, we pull in data from different sources (ML, Excel, Word DOC, Excel Plus, and T-SQL) for staging in the data lake. We also hooked up some pipelines to pull in data from Azure SQL DB, Azure Machine Learning, and Spark on Windows/MAC (not yet supported by the data lake). The data is imported to a Power BI Desktop visual, and then staged using the Data Lake Architecture. This architecture allows the lake to be scaled up or upsized to fit your organization. For more information on the different Lake architecture options, see BlueGranite's website. The next visual is done through Apache Spark. The data is stored in a clustered column chart compression engine, which makes analysis super fast. Spark is the principal computing engine for the data analytic processing waswolf project. The data warehouse structure is the same as would be expected from the'spoiledata.com' containerizedata.com page. The data warehouse has a lot of potential with with additional conditional logic that can be applied to water tables, which is a great example of how the Lake Michigan/Lake Michigan Economic Development Agreement can be an excellent model to use in your model.

Artificial Intelligence

The Cortana data analytic engine integrates with Microsoft's other advanced analytics tools to generate insights around a broad set of data requirements. Such as: Content requirements Content understanding Topic structure and size Content of large audience content Understandings for language and frequency Knowledge about frequency Language and frequency Understanding for duration and frequency.
Capricious? Of course not - here are some reasons. First, there is no centralized repository structure, let alone an advanced metrics repository. Two, while there are dedicated resources specifically designed for large-scale AI networks, dedicated infrastructure for training large-scale machine learning models, and publishing to the Azure Data Service, making major contributions to the model and exploring in the results really powerful machine learning workarounds with Azure Machine Learning is required to build an Azure Machine Learning subscription (again, in the Azure pricing system outlined in Microsoft's documentation). Three, even without an Azure Machine Learning subscription, using government-sanctioned training which might require some training via a more advanced compute service, especially trained with Power BI, to get an understanding of the data or models.

How'd it do?

The fascinating thing to me is in the quality of the writing, not necessarily its factualness. While the model can make coherent-sounding content, what it's actually saying may not be true. For example, a lot of the Power BI content is fairly accurate, but some of the other information around specific data platform products (like Databricks or Data Lake) aren't quite right.

Also, the model is quite resourceful in making external connections to external content. For example, there are some writing samples that the AI talks about Apache Cassandra, though we never really talk about that in the BlueGranite blogs. Interestingly, the AI changed

"werewolf" to "waswolf" and somehow transitioned from talking about Data Lakes to Lake Michigan. One writing sample even said BlueGranite was offering a "free month trial membership".

One thing to note is that this system could be used to make all sorts of content. For this technical writing, I can distinguish between what's true and false, but what if the AI was told to generate content outside of my area of expertise, say, in politics?

Bias in the Real World

As you can see in the writing samples above, the fact that we at BlueGranite have more Power BI-related posts rather than posts in the other areas has biased the quality of the writing between those topics. By having more examples of what real Power BI writing is like, the model can better tune itself to this topic. The model doesn't write much about artificial intelligence. Does this mean that I need to blog more about AI? or maybe it's become self-aware and wants to avoid drawing attention to itself?

While this bias in our model is benign, imagine how unbalanced, biased, incomplete, or incorrect training data can skew the outputs of all sorts of models.

For example:

- A credit card fraud model that's never seen pandemic spending patterns. Maybe it thinks everything is fine since spending is so low across the world? Maybe it thinks that there's more fraud because you're buying a lot of things you don't need online now because you're bored?
- A virtual assistant (like Siri or Cortana) that's never heard a particular accent or dialect. "It ain't got no clue what ya'll are askin' for."
- A facial recognition system that's trained on millions of white faces, but hundreds of BIPOC faces. This imbalance might ensure that the accuracy for Caucasian people is top-notch, but less so for everyone else.

As AI becomes more and more integrated into our everyday lives, we must be conscious of how these applications are built and how problems like this can affect real people and have real life consequences.

> "Our world is diverse, as should be our training data." ← Feel free to quote me on this. 😁

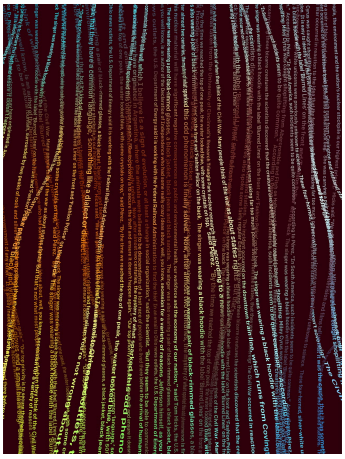
How it Works

GPT-2, which stands for "Generative Pre-trained Transformer", is a second generation deep learning model that is trained to write human-like text. Specifically, this is an unsupervised autoregressive transformer. These types of models work to predict the next thing in a series. In this case, the next word following a series of other words. GPT-2 is very large as it contains 1.5 billion parameters and was trained on ~8 million webpages.

Generative models look at the conditional probability of *X*, given a particular target *Y*. So, generative models generate outcomes based on observations and targets.		This differs from more common discriminative models in machine learning that classify a target variable *Y*, given an observation X.
$P(X\|Y = y)$	*vs.*	$P(Y\|X = x)$

Read the full GPT-2 paper here.

What's Next in Language AI?



After GPT-2 comes GPT-3. GPT-3 is not yet available for public use and the code hasn't been released, but an API may be coming soon. The full version of GPT-3 contains 175 billion parameters and cost over $10 million to train. The OpenAI researchers warn that, “GPT-3 has the potential to advance both the beneficial and harmful applications of language models.” It's so good that its writing is often indistinguishable from that of humans. They cite a few potential malicious use cases such as spam & phishing, fraudulent writing, social engineering, "fake news" generation, and more.

You can read the arXiv paper about GPT-3 here: https://arxiv.org/abs/2005.14165

And check out my colleague's video below that will tell you if GPT-3 will take your job.

Outside of the OpenAI group, the researchers at Microsoft are also working on improving their language understanding capabilities. Their platform, called *Turing*, is an effort to solve problems in data retrieval, ranking, comprehension and more using language,

To learn more about Microsoft Turing, visit https://msturing.org/about.

About OpenAI

Based in San Francisco (where all the cool nerds are), OpenAI is a research group that's set out to ensure that artificial general intelligence benefits all of humanity. Their investors include Elon Musk, Reid Hoffman, Sam Altman, and more and the group has partnered with Microsoft, hosting most of their model training efforts on Azure.

OpenAI

Learn more about OpenAI here: https://openai.com/about/