A Brief Survey for GPT Series
Zetong Wu
Yandao Street High School, Chengdu 610000, China

Abstract. Since the invention of GPT, the GPT language model has become an increasingly popular topic. ChatGPT as a recent hit point is built on the basis of GPT models, but so far there has been no comprehensive survey for GPT series. In this paper, we would like to present a brief survey of the GPT series and expound on the principle of GPT. Moreover, based on the development history and application scenarios of GPT, we report on the future prospects and applications of the GPT series.

Keywords: ChatGPT, Pre-training Models, AIGC, Transformer.

1. Introduction

ChatGPT[1], an Artificial Intelligence Generated Content (AIGC) developed by OpenAI to handle and solve some complex language understanding and sentence generation tasks, as well as imitate how human speak. It is based on a Transformer architecture (the detailed principle will be referred to in Section 2), a neural network specifically designed for natural language processing tasks. This paper briefly explains the development history, principles, future prospects and applications of GPT, and more specifically, this paper discusses the future development trend of GPT by analyzing past data.

![Figure 1. Displaying development period from GPT-1 to ChatGPT.][2]

Based on the illustration given by OpenAI, we could divide its development period into following stages, as shown in Figure 1.

1. GPT-1[3] has some generalization ability and can be used in NLP tasks.
2. GPT-2[4] builds on the comprehension capabilities of its predecessor and enhances generative functions such as story writing and dialog writing.
3. GPT-3[5] is a self-supervised model that has been able to perform complex and seemingly impossible tasks such as transforming web page descriptions into corresponding code and writing poems to mimic the speech of deceased greats.

![Figure 2. A brief explanation of the evolution of GPT series.][6]
As is shown in Figure 2, We mainly draw this diagram based on OpenAI papers, blog articles, and official APIs. Here, solid lines indicate clear evidence of the evolutionary path between two models (e.g. official statement that the new model was developed based on the base model), while dashed lines indicate relatively weak evolutionary relationships.

Naturally, this very popular project is studied by many researchers, and Figure 3 shows the number of papers on large language models and language models from 2019 to 2023.

Figure 3. Visualizing the number of papers on large language models and language models from 2019 to 2023.

As we can see from the chart, from the 2019 GPT invention dependence, the number of papers on GPT principles (large language models) has increased exponentially, which has led to the growth of the number of papers on language models (including large language models), which all reflect researchers' enthusiasm for GPT.

2. Principle GPT method

About the architecture of GPT: GPT adopts Transformer[7] architecture for NLP tasks, the Transformer model uses a self-attention mechanism to process a variety of inputs, and GPT replaces the encoder-decoder with a decoder to simplify the architecture and make GPT suitable for NLP tasks. To use GPT, you have to pre-train in advance, which is usually language modeling (i.e. GPT predicting the next word in a single time), and then fine-tuning it to complete some complex tasks.

The components of the GPT architecture: there are eight main components, namely the input embedding layer, position encoding, linear and Softmax functions, Transformer module, pre-training, fine-tuning, language modeling, and unsupervised learning.

(1) Input embedding layer: The input can be mapped to the vector, and the vector expression can then be modularly processed by the transformer.

(2) Position encoding: The expression processed by the transformer does not have the concept of position, so the position encoding plays a role in expressing its relative position.

(3) Transformer Module: Specially designed for NLP, it is widely used in machine translation, text classification, etc.

(4) Linear and Softmax Functions: Softmax function is commonly used for classification tasks. The Softmax function is specifically converted into a set of logits before being normalized. The normalized value obtained from the model can be interpreted as the probability or probability that a particular input belongs to each output class. The query, key, and value vectors for each marker in the input sequence are typically computed using a linear function in the attention mechanism. In the feedforward layer, the output of the multi-head attention layer is used for transformation.

(5) Pre-training, fine-tuning and language modeling: We put these three components together, pre-training is an important component of GPT architecture, pre-training must be carried out before the task, when one task is transferred to another, in order to make GPT more suitable for the task to fine-
tune on the basis of the existing database, language modeling: input a large number of languages for GPT processing during pre-training, thus enabling GPT to predict the next word from a large number of previous word sequences.

(6) Unsupervised learning: A machine learning algorithm that enables a model to learn without human intervention.

As is displayed in Figure 4, the principle of GPT operation can be roughly divided into two stages: pre-training and application.

In the unsupervised pre-training phase, GPT uses a large amount of text data for training, including Wikipedia, Gutenberg, etc. The training data is divided into different sequences, each of which is treated as a task in which the model needs to predict the probability distribution of the next word in the sequence. After unsupervised pre-training is complete, GPT often requires supervised fine-tuning in order to achieve better performance on specific tasks. The process of fine-tuning typically consists of two steps. Then we can test the fine-tuned model to evaluate its performance on a specific task. Due to GPT's excellence in natural language processing, it has been widely used in various tasks, including:

(1) Language modeling: GPT is excellent at language modeling, predicting the next word, the next sentence, and even the next paragraph.

(2) Machine translation: GPT can translate from one language to another, especially when working with long sentences.

(3) Text summary: GPT can generate text summary, so that users can quickly understand the main content of the text.

(4) Sentiment analysis: GPT can identify the emotion in the text and help users understand the emotional color of the text.

(5) Automatic Q&A: GPT can answer questions raised by users, thus providing users with convenient information services.

In summary, GPT is one of the most powerful models in the field of natural language processing, and its outstanding performance has made it widely used in various application scenarios.
3. Prospect and Application of GPT

<table>
<thead>
<tr>
<th></th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
<th>2022</th>
<th>2023</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>None</td>
<td>None</td>
<td>T0</td>
<td>BLOOMmT0</td>
<td>StarCoder</td>
</tr>
<tr>
<td>None</td>
<td>GPT-3</td>
<td>WebGPT</td>
<td>Codex</td>
<td>InstructGPT</td>
<td>ChatGPT</td>
</tr>
<tr>
<td>T5</td>
<td>Gshard</td>
<td>FLAN</td>
<td>Fian-PaLM</td>
<td>UL2</td>
<td>Bard</td>
</tr>
<tr>
<td>mT5</td>
<td></td>
<td></td>
<td>PaLM</td>
<td>LaMDA</td>
<td>PaLM2</td>
</tr>
</tbody>
</table>

Figure 5. Mainstream methods for GPT series from 2019 to 2023.

As you can see from Figure 5 from the invention of GPT in 2019 to 2023, the application of GPT has developed rapidly, showing a prosperous scene in 2022. From the chart, we can intuitively see that the development of GPT will also show an upward trend in the future, More and more high-tech companies have participated in the development of GPT, such as Google and OpenAI at the beginning, and Baidu and Huawei later. And a company tends to have multiple GPTs, so GPT has great prospects. More generally, having said that, what else is in the future for GPT? This article is divided into the following points:

1. **Increase in model size and complexity**
   One of the most notable trends is the increasing size and complexity of GPT models. Each generation, from GPT-3 to its successor, has seen a dramatic increase in the number of parameters, improving the ability of the model to understand and generate more detailed and contextualized content. We can expect future models to continue this trend, pushing the boundaries of AI understanding and output.

2. **Increased efficiency and accessibility**
   Despite the growth in model size, future GPT models are likely to become more efficient and accessible. Advances in AI optimization techniques and hardware are expected to reduce the amount of computational resources required, making these powerful tools more accessible to a wider range of users and developers.

3. **Specialization and Customization**
   Future GPT models may offer more specialization and customization options to meet specific industry needs. These models can be fine-tuned for specific languages, terminology, or authoring styles, enhancing their usefulness across multiple domains.

4. **Increased Interactivity and Personalization**
   As technology evolves, the GPT models of the future will provide a deeper level of interactivity and personalization of the experience. This means that they will be able to better understand the unique needs and preferences of users and provide customized responses and recommendations. This enhanced capability will make GPTs even more valuable in the areas of education, health advice, and personalized entertainment.

5. **Multimodal and cross-domain capabilities**
We can also expect GPT models to become more flexible in handling multiple data types (e.g., text, images, and audio), enabling truly multimodal interactions. In addition, they will be able to operate across domains, such as applying findings from scientific research to innovative teaching methods or drawing on cross-industry knowledge when designing new products.

**About its past & future:**

GPT is undoubtedly an interesting progress in the field of technology, and it is also an amazing revolution in the field of artificial intelligence, before machine learning, the development of artificial intelligence was very slow, but with the blessing of machine learning, artificial intelligence on the recognition of pictures, language, etc., perception ushered in a significant enhancement, but even so, these models are still far from human beings in complex logic. For example, in the early days, GPT could generate some short sentences very beautifully, but the context logic was confusing. GPT-3 and GPT-4 have made very surprising progress, and they can even answer further questions by studying the entire book of AP (Advanced Placement) biology, and the answers given are logical and feasible. Although there are still some problems waiting to be perfected for GPT, it is a fundamental change from the emergence of GPT-3 and GPT-4. Natural language can now be used as an interface for human-computer interaction, which means that we can provide more detailed information to AI, so that AI can help us complete more complex tasks. GPT can be applied in many fields in the future, such as education, health, business, etc., and in the future, through GPT, the role of computers in promoting education is no longer just to check information, but also to become a teacher who is available at any time. Regarding medical care, the emergence of GPT allows AI to provide personalized and differentiated healthcare according to the individual needs of patients, and GPT also has infinite potential to develop, and the following are several development directions:

1. GPT will approach or even surpass humans in multiple dimensions means that text-to-text (a technology that GPT-4 is now proficient in) is a thing of the past, text-to-everything is coming, and even more so everything-to-everything.
2. Applications built on the GPT-3/GPT-4 model will flourish, and the application scenarios will also expand from a single chat/writing to all latitudes of life.
3. GPT will be more specific, and will be more inclined to personalized development for specific enterprises, fields, etc.
4. The update of the algorithm will make the algorithm more comprehensive and more concise, thereby reducing the demand for data and the amount of calculation.

4. **Conclusion**

It is true that GPT is thriving and tends to be used in many different fields such as education, agriculture, entertainment, etc., and more and more top technology companies are participating in the research and development of GPT. However, the development of GPT will inevitably be hindered by people questioning whether GPT will make inappropriate statements or be used to forge some important documents. All in all, there is still a long way to go for GPT, and GPT still has a bright future.

**References**


