WHITE PAPER

TRANSFORMING INDUSTRIES – THE FUTURE LANDSCAPE WITH LARGE ML MODELS

Abstract

The rapid growth of unstructured data in various industries presents a significant challenge for organizations to process and interpret this information. The sheer volume of data generated every day can make it difficult to extract meaningful insights, resulting in inefficiencies and lost opportunities. The emergence of large pre-trained transformer models in natural language processing, computer vision and speech, however, has the capability to bring about a revolutionary change in the way industries handle and interpret this data.

This paper aims to investigate the problem of unstructured data and the benefits that large transformer models can bring. Various industries such as life sciences, finance, retail, and software development can apply these deep learning models to address a broad spectrum of challenges. Tasks such as semantic search, summarization, image and action recognition, speech recognition and others can experience performance improvements through the utilization of these large pre-trained transformer models. Through techniques such as text and code embeddings, these models can assist organizations in making sense of the huge data they collect, leading to improved efficiency and increased revenue. Additionally, the paper will examine the potential challenges and ethical considerations that arise with the increased use of these models in society, ultimately providing a complete overview of the present state of transformer models and their potential to shape the future.

Infosys®
Navigate your next

Table of Contents

1 Abstract 1

2 Challenges different industries are facing today 3

3 Addressing Unstructured Data Challenges with Large Transformer Based Pre-Trained Models 4

4 Industry Use Cases of Large Language Models 4

4.1 E-Commerce and Retail 4

4.1.1 Traditional Techniques 5

4.1.2 Large language model-based approach 6

4.2 Insurance 9

4.2.1 Traditional Techniques 9

4.2.2 Large language model-based approach 10

4.3 Life Sciences 10

4.3.1 Traditional techniques 10

4.3.2 Large language model-based approach 11

5 Other Applications 12

5.1 Large Language Models in Software Engineering 12

5.2 Vision Transformers 12

6 Ethical Considerations 13

7 Conclusion 14

References 15

Table of Figures

Figure 1: Data and Compute Requirements 4

Figure 2: Traditional Approach vs pre-trained models 6

Figure 3: Performance of various models on Question-Answering dataset 9

External Document © 2023 Infosys Limited

2 Challenges different industries are facing today

Industries are currently facing several challenges in the ever-evolving digital landscape. One of the major challenges is the enormous amount of data being generated due to the digital revolution. The rise of social media, e-commerce, and other digital platforms has led to companies collecting large volumes of data related to their customers, transactions, and various aspects of their business. However, a significant portion of this data is unstructured. According to a study by IDC (O'Reilly, 2022), the global data generation is expected to reach 175 zettabytes by 2025, and 80% of this data will be (Andonian, et al., 2021) unstructured, including customer reviews, social media posts, and other forms of text data. Companies are struggling to extract valuable insights from such a large amount of unstructured data.

Traditional machine learning models such as decision trees and linear regression are not well-suited to deal with unstructured data. Building and maintaining these models requires a significant amount of data, computational resources, and expertise. Additionally, these models cannot fully comprehend the nuances of natural language, making it difficult to extract insights from text data. Therefore, the ability to effectively process and derive valuable insights from unstructured data is a significant challenge that industries must address in today's digital landscape.

External Document © 2023 Infosys Limited

3 Addressing Unstructured Data Challenges with Large Transformer Based Pre-Trained Models

Traditional machine learning models that are commonly used to solve any of the problems listed above, require a large amount of effort, data, and compute. Large transformer models are being used to address these challenges by providing advanced natural language processing, image recognition, and decision-making capabilities. These models are pre-trained on vast amounts of data, which allows them to extract insights from unstructured data with a high degree of accuracy. Figure 1 below compares the data and compute requirements for traditional machine learning models versus transformer-based models.

In the paper titled "Attention is all you need" (Vaswani, et al., 2017) published by Google, the concept of self-attention has been introduced. The paper proposed a novel neural network architecture known as the Transformer, which exclusively utilizes self-attention mechanisms and does not rely on recurrent or convolutional methods. The architecture allows for parallel processing of input sequences, making it more efficient than recurrent neural networks (RNNs) for longer sequences. Additionally, the self-attention mechanism in the transformer model enables it to capture long-range dependencies and contextual information, resulting in better language understanding.

The transformer architecture serves as the foundation for pre-trained models. Pre-trained models are neural networks that have been trained on large amounts of data to learn general language representations. These pre-trained models often use the transformer architecture as their base as it allows them to learn more complex and contextualized representations of language.

For example, the popular pre-trained model BERT which stands for Bidirectional Encoder Representations from Transformers (devlin, et al., 2018) is based on the transformer architecture and has achieved state-of-the-art performance on a wide range of NLP tasks. Similarly, another pre-trained model GPT-3 (Generative Pre-trained Transformer) (Brown, et al., 2020) also utilizes the transformer architecture and has garnered attention for its ability to generate high-quality text.

Pre-trained models offer some significant advantages. These models can be fine-tuned on smaller, task-specific datasets to achieve better performance, leveraging their pre-existing knowledge of general language representations. Additionally, pre-trained models can help to overcome the issue of data scarcity, which is common in many NLP applications. Fine-tuning a pre-trained model on a small amount of task-specific data can often lead to better results than training a model from scratch on the same data.

Moreover, pre-trained models can potentially lower the computational cost and time needed for training new models, as they provide a good initialization for the model's parameters.

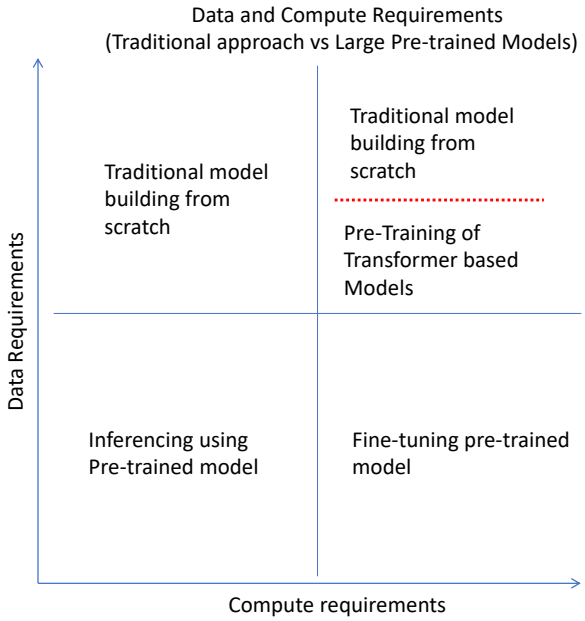

Figure 1: Data and Compute Requirements

This can be particularly advantageous for resource-limited environments. Overall, the transformer architecture and pre-trained models have revolutionized NLP by offering increased efficiency, better language understanding, and reduced data requirements.

In addition, the transformer architecture has also shown promise in the field of computer vision, with the emergence of vision transformers. These models apply the transformer architecture to image data, allowing for the capture of spatial dependencies and contextual information across image regions.

Like pre-trained models in NLP, pre-trained vision transformers have demonstrated the potential to reduce the computational cost and time required for training new models on computer vision tasks. By leveraging pre-trained models, researchers can develop and fine-tune models on smaller datasets, and achieve state-of-the-art performance on various image-related tasks.

4 Industry Use Cases of Large Language Models

4.1 E-Commerce and Retail

The ever-increasing amount of customer reviews for products offered in the e-commerce and retail industry presents a challenge for businesses to gain valuable insights from this data. However, with the advancements in natural language processing (NLP) techniques, businesses can now leverage these reviews to gain meaningful insights and improve their products. In this section, the applications of pre-trained transformer models are explored in e-commerce and retail segments, specifically for analyzing the sentiment and performing aspect-based sentiment analysis on customer reviews. The limitations of NLP techniques are also discussed and the applicability of transformer-based models to overcome such limitations is explored.

External Document © 2023 Infosys Limited

E-commerce and retail companies have always strived to deliver the best products and services to their customers. One of the keyways to achieve this is by understanding customer feedback, specifically their sentiments and preferences. This includes identifying positive, negative, and neutral feedback, as well as specific aspects of the product or service that customers are praising or criticizing.

This information can then be used to make data-driven decisions and improve products and services to meet customer expectations.

4.1.1 Traditional Techniques

Manual analysis of customer reviews has been a common practice, particularly for small businesses or those with limited resources. However, this approach is prone to human errors, and it can be difficult to scale as the volume of customer reviews increases.

To overcome this, traditionally, sentiment analysis has been done using machine learning algorithms. These algorithms need both data as well as compute.

They are designed to learn from labeled data, which is often manually annotated by human annotators.

Thus, this approach involves collecting a large amount of customer reviews and manually labeling them as positive, negative, or neutral. The machine learning models are then trained on the labeled data to learn patterns and classify new reviews into relevant categories or topics. While this approach can be effective, it has some limitations.

First, there are limitations such as the need for a substantial amount of labeled data, which can be a costly and time-consuming endeavor to obtain. Second, the machine learning models may not be able to capture the nuances of natural language and may struggle with sarcasm, ambiguity, and complex sentence structures.

Third, standard NLP techniques struggle with understanding context, especially in situations where the meaning of a word or phrase depends on the larger context of the conversation. For instance, the word "cold" can have different meanings in the context of weather, food, or emotions. Fourth, NLP techniques are often optimized for specific languages, and processing multiple languages can be challenging due to differences in grammar, syntax, and vocabulary.

External Document © 2023 Infosys Limited

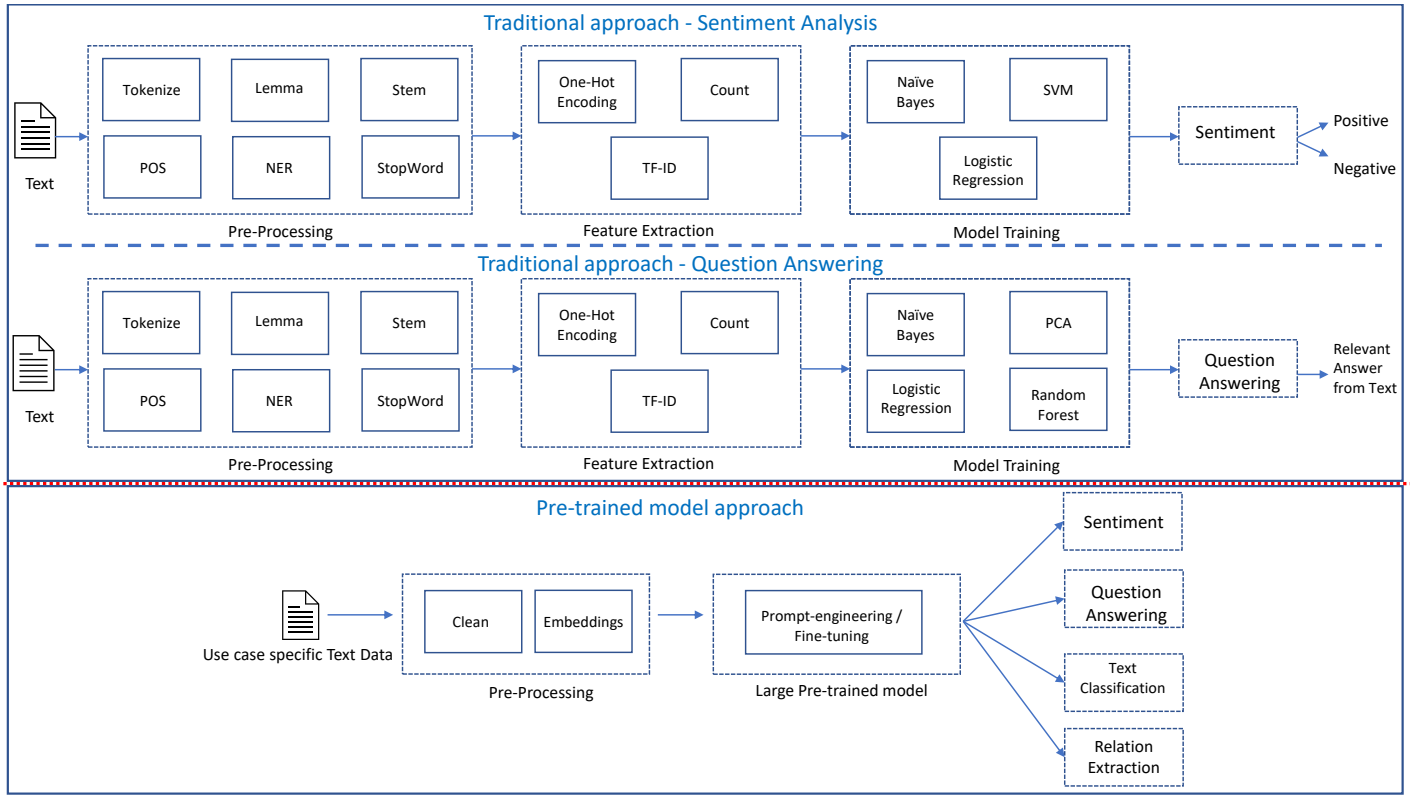

Figure 2: Traditional Approach vs pre-trained models

4.1.2 Large language model-based approach

With the recent advances in NLP and machine learning, large pre-trained models such as GPT-3 have become a game-changer for sentiment analysis and aspect-based sentiment analysis in e-commerce and retail. These models can understand the context of the sentences and effectively capture the relationships between words. They can identify the sentiment and classify the type of defect in the product from reviews with a high degree of accuracy. Moreover, these models can be fine-tuned on specific tasks and domains with minimal labeled data, making it much more accessible for businesses of all sizes to leverage the power of NLP. Figure 2 below highlights the complexity in building a model using traditional approach versus leveraging large pre-trained models. There are different ways to use large pre-trained models for sentiment and aspect-based sentiment analysis. Zero-shot learning, few-shot learning, and fine-tuning are the most popular methods. Zero-shot learning involves providing the review and instructions to the pre-trained model to perform sentiment analysis and, if negative, classify the type of defect in the product. Few-shot learning involves providing a few review-defect in the product pairs and a tweet in the end to predict its class to the pre-trained model. In this case, the pre-trained model will try to predict the defect in the product review based on the samples provided. Finetuning involves finetuning the pre-trained model with review-defect in the product pairs and creating a finetuned model. Lastly, embeddings involve calculating embeddings for the tweet-label pairs dataset and training a machine learning model.

The advantages of large pre-trained models are numerous. First, they require less labeled data and compute resources than traditional machine learning models (Ozcift, et al., 2021). Second, they can capture the nuances of natural language and are better equipped to handle sarcasm, ambiguity, and complex sentence structures (Parameswaran, et al., 2021). Lastly, they are easier to fine-tune and can provide more accurate results.

Thus, large pre-trained models offer an efficient and accurate solution to sentiment and aspect-based sentiment analysis in e-commerce and retail. They can help companies understand customer feedback and make data-driven decisions to improve their products and services.

With the continued advancements in natural language processing, the use of large pre-trained models will become even more prevalent in the e-commerce and retail industry. Here are some additional ways in which large-language models (LLM) can be beneficial:

- Competitive advantage: By leveraging LLM based NLP techniques, e-commerce companies can gain a competitive advantage over their rivals by delivering better products and services that meet customer expectations.
- Personalized responses: LLM like GPT-3 can be tuned to generate personalized responses to customer reviews, addressing their concerns and providing solutions (The founder, 2021).

External Document © 2023 Infosys Limited

This can help improve customer satisfaction and loyalty, leading to increased sales and revenue.

- Multi-lingual support: Many LLMs like GPT-3, GPT-Neox (Andonian, et al., 2021) Bloom (bigscience, 2022) can handle multiple languages, making it useful for analyzing customer reviews from different regions and countries. This can help e-commerce companies to better understand customer needs and preferences in different markets.
- Identifying trends: LLMs like GPT-3 can identify trends and patterns in customer feedback, helping e-commerce companies to identify common issues and prioritize areas for improvement.
- Identifying influencers: LLMs like GPT-3 can also identify influencers and brand ambassadors, allowing e-commerce companies to build relationships with them and leverage their influence to promote their products.

Despite their many advantages, pre-trained language models also come with some limitations that businesses should be aware of. Here are four key challenges to keep in mind:

1. Pre-trained models are generic models. These models may not capture the specific nuances and domain knowledge required for a particular business or industry. For example, a pre-trained model may not understand the specific jargon used in the fashion industry, which could result in inaccurate sentiment analysis and product recommendations
2. Fine-tuning challenges: While fine-tuning can help address the domain-specific language challenges mentioned above, it can be a time-consuming and resource-intensive process. In addition, fine-tuning with a limited amount of data can result in overfitting and inaccurate predictions. (Takyar, n.d.)
3. Context window limit: Pre-trained models typically have a fixed context window (Explained, 2021), which means they may not be able to capture the full context of longer reviews or more complex sentences. This can result in inaccurate sentiment and aspect analysis.
4. Zero-shot challenges: Zero-shot learning is a powerful technique, but it also has limitations. For example, if the review contains complex or nuanced language, the pre-trained model may struggle to accurately classify the sentiment or aspect. Providing accurate instructions and rules to perform the classification can be a challenging process.

Despite these challenges, pre-trained models are still a valuable tool for businesses looking to gain insights from customer feedback.

By understanding these challenges and carefully selecting the appropriate approach, businesses can leverage pre-trained models to improve their products and services and ultimately drive success in the e-commerce and retail space.

Overall, large language models can help e-commerce companies to better understand their customers and improve their products

based on customer feedback, leading to increased customer satisfaction and loyalty, and ultimately, increased revenue.

External Document © 2023 Infosys Limited

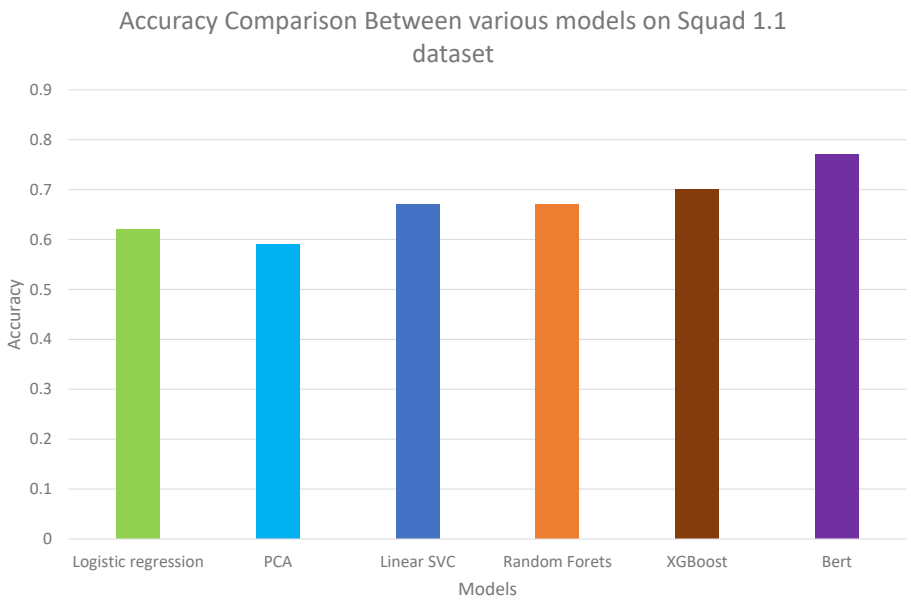

Figure 3: Performance of various models on Question-Answering dataset

4.2 Insurance

Insurance companies receive a large volume of inquiries from customers seeking information about their policies, coverage, and claims. These inquiries could be related to a variety of topics, ranging from policy details to claim processing and settlement. Due to the complexity of insurance policies and the technical jargon used in the industry, customers often find it difficult to navigate through the available information and get answers to their questions. Traditional search engines have limitations in comprehending the intent behind a user's query and provide relevant results. This can lead to frustration and dissatisfaction for customers who want quick and accurate answers.

4.2.1 Traditional Techniques

To improve customer experience, insurance companies use chatbots, conversational agents that can answer customer queries quickly and efficiently. While these chatbots have been successful in handling simple and straightforward queries, they struggle to understand complex queries that require a deeper understanding of the user's intent.

Traditional chatbots use a combination of rule-based and machine learning approaches to provide answers (Neurosoph, n.d.). Chatbot development starts with intent classification, where developers define the different intents, or categories, of user queries. Then, chatbot developers create a set of rules that helps the chatbot determine the appropriate response to a user's query. This process requires significant manual effort and time, and scaling chatbot operations can be challenging.

Traditional chatbots have several limitations that can impact customer experience. First, they rely on predefined responses, making it difficult to handle complex queries that require more nuanced answers. Second, they require significant amounts of training data to work effectively, which can be difficult to acquire for less common or new queries. Lastly, traditional chatbots are limited in their ability to handle multiple languages or dialects.

4.2.2 Large language model-based approach

Large language models, like OpenAI's GPT and Google's BERT, which are already pre-trained, can help overcome the limitations discussed above. LLMs can perform a vast array of natural language processing tasks, including language comprehension and new content generation. (Patel, et al., 2020) studied the performance of large language models like BERT versus traditional ML models on the task of question-answering. Figure 3 below shows how BERT outperforms other models on this task.

By leveraging LLMs, insurance companies can build semantic search engines that can understand user queries better and provide relevant answers. With LLM-based semantic search, the chatbot can provide personalized and context-aware responses, leading to better customer satisfaction.

Integrating LLM-based semantic search with a chatbot can be done in several ways. One approach is to make use of zero-shot learning, where the chatbot can answer queries without requiring it to be trained on those specific queries. Another strategy is to use the few-shot learning approach, where the chatbot learns from a small number of examples before it can handle new queries. These approaches reduce the need for massive training data and enable chatbots to respond to complex queries with greater accuracy.

In addition to few-shot and zero-shot approach, building a semantic search using embeddings can be an effective approach to improve the customer experience and provide contextualized answers for the user queries.

Embeddings are a way to represent words or phrases as vectors in a high-dimensional space, where semantically similar

External Document © 2023 Infosys Limited

words are located closer together. By using embedding-based semantic search, insurance companies can build chatbots that can understand user queries better and provide more accurate and relevant answers. By comparing the embeddings of user queries with the embeddings of policy documents, the chatbot can identify the most relevant documents and extract relevant information to provide more accurate and personalized responses to the user.

This approach is particularly effective for answering insurance-related questions. It can consider the context of the query and provide relevant responses drawing from the contents of the insurance-related documents.

The integration of semantic search solutions with chatbots can greatly benefit insurance companies seeking to improve customer experience, efficiency, and reduce costs.

- By leveraging large pre-trained language models (LLMs) and semantic search, chatbots can deliver personalized and context-aware responses to customer queries, improving customer satisfaction and loyalty.
- Additionally, chatbots with semantic search capabilities can quickly identify relevant policy documents and extract pertinent information, enabling faster response times to customer inquiries.
- The automation of customer inquiries through semantic search can reduce the need for human intervention and enable insurance companies to handle a larger volume of inquiries with greater efficiency.
- By using embedding-based semantic search, chatbots can provide more precise and relevant responses to customer queries, resulting in better outcomes for both the customer and the insurance company.
- Finally, the automation of customer support processes can lead to cost savings for insurance companies, improving their bottom line.

Overall, the integration of semantic search solutions with chatbots is a worthwhile investment for insurance companies seeking to enhance their customer service capabilities. These large language-based solutions can help companies build stronger relationships with their customers and ultimately drive greater business success.

4.3 Life Sciences

The life science industry is a highly regulated sector that involves the development, testing, and manufacturing of new drugs, biologics, medical devices, and other healthcare products. The success of the industry is heavily dependent on the ability to analyze and interpret substantial and diverse sets of information, including clinical trial results, patient data, regulatory guidelines, and scientific literature.

In this section, two critical problems faced by life science companies are discussed along with how large language models can be utilized to solve them.

1. Summarizing Medical Journals: One of the most significant challenges faced by life science companies is the analysis and interpretation of vast amounts of scientific literature. The amount of scientific literature published every year is growing at an unprecedented rate. It is impossible for researchers to keep up with this volume of literature manually.
2. Generating Knowledge Graphs: One of the key challenges faced by life science researchers is the identification of protein-protein interactions and the mechanisms by which diseases operate. This requires the analysis of vast amounts of unstructured data, including text, images, and other forms of media. Extracting useful information from this data requires advanced tools and techniques, which can be time-consuming and costly. To tackle the challenge of identifying protein-protein interactions and disease mechanisms, researchers in the life sciences have turned to generating knowledge graphs. Knowledge graphs are a powerful tool for representing and analyzing complex information. They provide a structured way to organize data and can be utilized to extract meaningful insights and relationships that might not be apparent from unstructured data alone. A knowledge graph can also aid in identifying new drug targets, predicting drug efficacy, and analyzing disease pathways. Building a knowledge graph requires the extraction and integration of data from multiple sources, including scientific literature, clinical trial data, and patient data.

4.3.1 Traditional techniques

1. Summarizing Medical Journals

Traditional techniques used for summarizing medical journals involve manual reading and extraction of relevant information or the use of machine learning algorithms for text summarization. However, these techniques have several limitations. Manual extraction is time-consuming and error-prone and can lead to inconsistencies as different researchers may interpret the same information differently.

While machine learning-based approaches can be more efficient than manual approaches, they also have their limitations. Machine learning models require significant labeled data to be effective, and the quality of summaries they produce is heavily reliant on the quality of data used while training the model. In the life science industry, where data is often sparse and heterogeneous, this can be a significant challenge.

Thus, this approach is often limited by the volume of data and computational resources available. Moreover, traditional machine learning models are often domain-agnostic, meaning that they lack the specialized knowledge needed to analyze life science data effectively. This can lead to inaccurate results and missed opportunities for discovery.

2. Generating Knowledge Graphs

Traditional techniques for generating knowledge graphs involve manual curation of data from multiple sources or the use of rule-based systems to extract and integrate data. However, these

External Document © 2023 Infosys Limited

techniques have limitations in terms of scalability and accuracy. Manual curation is time-consuming and expensive, while rule-based systems may fail to capture complex relationships between concepts.

Traditional machine learning techniques can also be leveraged for generating knowledge graphs. This typically involves gathering data from a variety of sources like scientific literature, experimental data, and public databases of genetic and proteomic information. Once the data is collected, it is then processed using natural language processing and other machine learning techniques to identify relevant entities and relationships.

Entities might include proteins, genes, diseases, and other biological structures, while relationships might include protein-protein interactions, gene regulatory pathways, and disease mechanisms.

Once the entities and relationships are identified, they are then organized into a graph structure, with nodes representing entities and edges representing relationships.

The resulting knowledge graph can be used to identify new relationships and insights that might not be apparent from unstructured data alone. For example, researchers might use a knowledge graph to identify novel protein-protein interactions that could be targeted with new drugs or to identify genetic pathways that are associated with specific diseases.

However, generating knowledge graphs using these traditional ways can be a time-consuming and costly process.

4.3.2 Large language model-based approach

1. Summarizing Medical Journals

Large language models like GPT-3 have the inherent capability to summarize information. However, these are generic models and do not understand the underlying domain. Also, the challenge here lies in creating a summarized version that maintains the correctness of the information and the amount of information from the original text. The limited corpus of data available presents additional challenges, including data inconsistencies and fragmentation of information.

Fine-tuning a large language model for the task of summarizing technical journals, would require labeled data, which is mostly unavailable. However, by applying techniques like prompt-engineering and zero-shot and few-shot learning, these models can be leveraged to summarize medical journals effectively. By applying different prompt-engineering techniques and building pipelined approach, it is possible to effectively learn the nuances of the domain-specific language and enhance the quality of the summaries produced.

2. Knowledge Graph

One of the major challenges faced by life science companies is to understand complex disease mechanisms and identify potential drug candidates for treatment. Large language models such as BioBert (Lee et al., 2019) and BioGPT (Luo, et al., 2022) have been incredibly useful in understanding protein-protein interactions and identifying the mechanisms in which diseases operate. They are trained on specific medical journal corpora, which enables them to recognize entities and patterns within the domain. By leveraging LLMs, the time for drug discovery can be shortened, which can lead to major breakthroughs.

However, understanding disease mechanisms is just the beginning. The next step is to represent these mechanisms as knowledge graphs, which incite unprecedented correlations. By leveraging LLMs, knowledge graphs can be created that can help understand minute details of complex subjects like disease mechanisms. This is a critical step in drug discovery as it enables the identification of potential drug candidates that target specific pathways.

By leveraging the expertise and abundance of literature in the field, LLMs can prove to be a valuable tool for long-range pattern recognition, which is critical in the discovery of new drugs and therapies.

LLMs can prove to be a game-changer, offering significant advantages over traditional approaches. LLMs can be fine-tuned for specific applications, making them ideal for the life science domain. By training these models on specific medical journal corpora, researchers can shorten the time for drug discovery and identify breakthroughs that may have otherwise been missed.

There are many other application of large language models in the life science domain.

- Clinical trial design: Large-language models can be utilized to analyze clinical trial data and identify patient subgroups that are more likely to respond to treatment. This can help companies design more effective clinical trials and reduce the cost and time required to bring new treatments to market.
- Regulatory compliance: Large-language models can be employed for the analysis of regulatory documents and to ensure that companies are complying with all relevant regulations and guidelines. This can help companies avoid costly fines and penalties and reduce the risk of regulatory approval delays.
- Patient engagement: Large-language models can be leveraged for analyzing patient data, such as electronic health records, social media posts, and surveys, to get a better understanding of patient preferences, behaviors, and experiences. This can help companies develop more personalized and effective healthcare products and services, improve patient outcomes, and enhance patient engagement and satisfaction.
- Medical writing: Large-language models can be utilized for generating high-quality scientific and medical writing, such as research papers, patent applications, and regulatory submissions. This can help companies reduce the time and cost required to produce these documents, while also ensuring that they are accurate, clear, and compliant with regulatory guidelines.

External Document © 2023 Infosys Limited

In addition to these use cases, large language models can also help life science companies unlock new insights and opportunities from unstructured data, such as scientific papers, patents, and social media posts. They can also help companies identify new trends and opportunities in the industry and make more informed strategic decisions.

Overall, large language models have the potential to transform the life science industry by enabling companies to extract insights from vast amounts of unstructured data, accelerate the drug discovery process, improve patient outcomes, and reduce the time and cost required to bring new treatments to market. While there are still challenges and limitations to be addressed, the future of large language models in the life science industry looks promising.

5 Other Applications

5.1 Large Language Models in Software Engineering

As discussed in the vertical view of industries, transformer models are playing a pivotal role in shaping the landscape of retail, insurance, finance, and various other industries. They are significantly revolutionizing operations leading to enhanced efficiency and accuracy. However, the impact of these models extends beyond the vertical view, touching on the software industry as well.

In the software industry, transformer models have shown significant promise in automating programming language and other software engineering aspects. These models have the potential to transform the way software is developed, making it faster, more efficient, and less error prone.

One of the most significant challenges in software engineering is the high degree of manual effort required to develop and maintain software systems. Software developers spend a significant amount of time writing code, testing, and debugging, and ensuring that the software meets the required specifications. This manual effort can be time-consuming, costly, and error-prone, leading to delays in software development and higher costs.

Large transformer models can automate many aspects of software development, such as code completion, debugging, and testing. These models can learn to write code by analyzing large datasets of existing code, making it possible to generate high-quality code automatically. This can significantly reduce the time and effort required to develop software systems. Numerous such models like GPT-3 (175-billion), Codex (12-billion) (Chen, et al., 2021), GPT-Neox (20-billion), OPT (175-billion) (Zhang, et al., 2022), Bloom (176-billion), CodeGeeX (13-billion) (github, 2023) and now ChatGPT (OpenAI, 2022) have shown encouraging results in generating code that is semantically and syntactically correct. These models use natural language processing techniques to understand the code in entirety and generate code snippets accordingly. They can also assist developers in debugging and testing by identifying errors in the code and suggesting fixes.

However, these models are not perfect and may sometimes generate code with errors or biases. Therefore, it is important to use them as a tool to augment human intelligence rather than replace it entirely. Nevertheless, the potential benefits of using large transformer models in software development are immense and can lead to substantial improvements in the efficiency and quality of the software development process.

In addition, transformer models can automate the testing process by predicting potential errors in the code and suggesting fixes. This can help identify bugs and vulnerabilities in the software early in the development process, reducing the risk of costly errors and security breaches.

Furthermore, transformer models can automate the documentation process, which is often neglected in software development. Documentation is critical to ensuring that software systems are easy to use and maintain, yet it is often time-consuming and tedious to write. Transformer models can generate high-quality documentation automatically, making it easier for developers to understand and maintain software systems.

The use of transformer models in software engineering has the potential to transform the way software is developed and maintained. By automating many aspects of software development, these models can reduce the time and effort required to develop software systems, leading to faster development cycles, lower costs, and higher quality software.

5.2 Vision Transformers

The field of computer vision has been revolutionized by the introduction of transformers in recent years. Convolutional neural networks, popularly known as CNNs, have been the go-to for computer vision tasks such as image classification, object detection, and semantic segmentation. However, they suffer from inductive biases due to limited training data. Additionally, they do not consider temporal information that may be available.

The shortcomings of 2D CNNs are particularly evident in tasks such as action recognition, where temporal positioning of objects is crucial. To address this issue, 3D CNNs were introduced, which could keep track of temporal information. However, these models come with higher computational load, memory constraints, and training time due to additional parameters.

In 2020, (Dosovitskiy, et al., 2021) proposed a new architecture called the Vision Transformer (ViT) that applies the transformer architecture used in natural language processing to computer vision tasks.

The ViT model uses self-attention mechanisms to capture long-range dependencies between image patches and produces a fixed-size feature representation for the image that can be used for downstream tasks such as classification, object detection, and segmentation.

One of the key advantages of the ViT model is its ability to handle images of arbitrary size without the need for resizing or cropping. This is because the model processes the image as a sequence of patches, which can be of different sizes depending on the input

External Document © 2023 Infosys Limited

image size. This allows for more flexibility in handling images of varying resolutions and aspect ratios.

Another advantage of the ViT model is its ability to capture both spatial and temporal information in videos using a 2D+T approach. This means that the model processes each video frame as a sequence of image patches, while also considering the temporal relationship between adjacent frames. This allows for better performance in video-based tasks such as action recognition and video segmentation.

While the ViT was a breakthrough in transformers' use in computer vision, there are still challenges such as availability of large datasets for pretraining, effective inductive biases, and the translation of this success to other computer vision applications. Various other transformer architectures are also available such as SWIN which is based on hierarchical shifted window architecture (Liu, et al., 2021) ,DETR or Detection Transformer (Carion, et al., 2020) and OWL-ViT (Minderer, et al., 2022) which are used to solve various computer vision problems. These architectures leverage the ViT or SWIN architecture and create layers specific to the application.

ViT and SWIN transformers are currently the choice for base architecture in computer vision tasks. Hybrid architectures, combining CNNs and transformers, are also being explored to benefit from the strengths of both approaches. For example, CNNs can be used as feature extractors for transformers, which can then process the extracted features to enhance the output. Such hybrid approaches have shown encouraging results in tasks such as image classification and object detection.

Transformers have also had a significant impact on dense prediction tasks, such as instance and semantic segmentation, and predicting the depth of objects. Conventional depth estimation techniques rely on expensive sensors that are prone to environmental factors that can reduce efficiency and accuracy. Dense prediction with transformers can create 3D reconstructions of scenes without these sensors, making it easier to estimate structural deformities and asset positions in applications such as manufacturing and infrastructure survey.

Overall, the application of transformers in computer vision has allowed for more efficient and accurate models. Despite the need to address challenges such as pretraining and effective inductive biases, transformers have already made a substantial impact in computer vision to perform tasks such as image classification, object detection, and dense prediction. As technology continues to evolve, the industry can anticipate witnessing a proliferation of even more innovative applications.

6 Ethical Considerations

As the use of large transformer models continues to increase, there are several ethical considerations that must be considered. These include issues such as data privacy, bias, and accountability.

Data Privacy: Large transformer models are often trained on publicly available data, which raises concerns about data privacy

and the possibility of personal information being misused.

Bias: Large pre-trained models can perpetuate bias if they are trained on biased data, which can lead to unfair or inaccurate decisions. It is important for companies and organizations to be aware of these ethical considerations and take steps to mitigate them.

Accountability: Even when using pre-trained models, there must be clear lines of accountability for any decisions made based on the model's output. It is important to be transparent about how the model was used and the reasoning behind any decisions made based on its output.

Responsibility: Companies and organizations that use pre-trained models must take responsibility for the impact of their decisions. This includes ensuring that the pre-trained model is being used ethically and in compliance with relevant laws and regulations.

When utilizing models such as OpenAI Codex, which are only available as an API, it is imperative to take note of the fact that the data will be processed outside of the enterprise boundary. In addition, the data is stored for a maximum of 30 days, and it is important to explicitly mention such constraints.

Overall, while using pre-trained models can save time and resources, it is important to consider the potential ethical implications of their use. By addressing these ethical considerations, companies and organizations can help to ensure that pre-trained models are being used for positive purposes and in an ethical manner.

External Document © 2023 Infosys Limited

7 Conclusion

In conclusion, industries are facing several challenges in relation to unstructured data, including the sheer volume of data that is being generated, the amount of effort that is required to extract meaningful information from this data, and the limitations of traditional machine learning models. Large transformer models are shaping the future of industry and beyond by providing advanced natural language processing, image recognition, and decision-making capabilities. They are being used to address a wide range of challenges in areas such as healthcare, finance, insurance, and retail.

However, as the use of these models continues to increase, it is important to consider the ethical considerations that arise with their use in society. Ultimately, large transformer models have the potential to greatly benefit society, but it is important to ensure that they are used in a responsible and ethical manner.

External Document © 2023 Infosys Limited

References

Andonian, A. et al., 2021. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch, s.l.: s.n.

bigscience, 2022. Bloom, s.l.: huggingface.co.

Brown, T. et al., 2020. Language Models are Few-Shot Learners. [Online]

Available at: https://arxiv.org/abs/2005.14165v4

Carion, N. et al., 2020. End-to-End Object Detection with Transformers. [Online]

Available at: https://arxiv.org/abs/2005.12872

Chen, M. et al., 2021. Evaluating Large Language Models Trained on Code. [Online]

Available at: https://arxiv.org/abs/2107.03374

devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [Online]

Available at: https://arxiv.org/abs/1810.04805

Dosovitskiy, A. et al., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. [Online]

Available at: https://arxiv.org/abs/2010.11929v2

Explained, G.-3., 2021. Rohan Jagtap. [Online]

Available at: https://towardsdatascience.com/gpt-3-explained-19e5f2bd3288

github, 2023. CodeGeeX: A Multilingual Code Generation Model. [Online]

Available at: https://github.com/THUDM/CodeGeeX

Liu, Z. et al., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. [Online]

Available at: https://arxiv.org/abs/2103.14030.

Luo, R. et al., 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics.

Minderer, M. et al., 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. [Online]

Available at: https://arxiv.org/abs/2205.06230

[Accessed 4 3 2023].

Neurosoph, n.d. Conversational AI vs Traditional rule based chatbots. [Online]

Available at: https://neurosoph.com/conversational-ai-traditional-rule-based-chatbots/

OpenAI, 2022. ChatGPT: Optimizing Language Models for Dialogue. [Online]

Available at: https://openai.com/blog/chatgpt

O'Reilly, M., 2022. The Unseen Data Conundrum. [Online]

Available at: https://www.forbes.com/sites/forbestechcouncil/2022/02/03/the-unseen-data-conundrum/?sh=733d11897fcc

Ozcift, A., Akarsu, K., Yumuk, F. & Söylemez, C., 2021. Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. Automatika, pp. 1-13.

Parameswaran, P., Trotman, A., Liesaputra, V. & Eyers, D., 2021. BERT's The Word: Sarcasm Target Detection using BERT. [Online]

Available at: https://aclanthology.org/2021.alta-1.21.pdf

Patel, D., Parikh, R., Raval, P. & Shastri, Y., 2020. Comparative Study of Machine Learning Models and BERT on SQUAD. [Online]

Available at: https://arxiv.org/pdf/2005.11313.pdf

Takyar, A., n.d. FINE-TUNING PRE-TRAINED MODELS FOR GENERATIVE AI APPLICATIONS. [Online]

Available at: https://www.leewayhertz.com/fine-tuning-pre-trained-models/

The founder, R., 2021. Using GPT-3 to Respond to Customer Reviews. [Online]

Available at: https://www.replier.ai/blog/using-gpt-3-to-respond-to-customer-reviews

Vaswani, A. et al., 2017. Attention Is All You Need. [Online]

Available at: https://arxiv.org/abs/1706.03762

Zhang, S. et al., 2022. OPT: Open Pre-trained Transformer Language Models. [Online]

Available at: https://arxiv.org/abs/2205.01068

[Accessed Jan 2023].

External Document © 2023 Infosys Limited

About the Authors/Mentor

Author
Varsha Jain
Senior Data Scientist

Author
Likhith Prudhivi
Digital Specialist Engineer

Author
Sagar Eidnani
Senior Associate Consultant

Mentor
Kamalkumar Rathinasamy
Distinguished Technologist

© 2023 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.