Applications of Transformer Models

Leonardo Anyelo Rodriguez Martinez
Dept. of Computer Science
Hof University of Applied Sciences
Hof, Germany
Leonardorm7@hotmail.com

Abstract

Over the last few years, transformers have significantly impacted the field of artificial intelligence, allowing the development of technologies such as Large Language Models (LLMs), which have demonstrated remarkable performance in understanding natural language. These models have facilitated the development of innovative applications in different fields, including text, audio, and image processing. Since the introduction of the transformer architecture, researchers have used transformers in different configurations to understand and generate content, allowing significant developments. As examples of this, transformers have allowed the creation of models that can create mobile prototype apps in real time, compose music, interpret speech and background sounds, detect cancer cells in medical imaging, and even understand various input types with multimodal models. However, what makes transformer special is the self-attention mechanism. This mechanism allows it to process long sequences capturing the context efficiently. This paper aims to show some of the new technologies developed through the implementation of transformers in the fields of text, audio, and image processing, highlighting their impact in these fields and potential for future developments.

Keywords: Transformers, artificial intelligence, Large Language Models, natural language processing, self-attention mechanism.

I. INTRODUCTION

Initially, the transformer architecture [1] was developed for text translation. However, it has also shown incredible performance in other fields. First must be mention that the transformer is a neural network architecture that has achieved outstanding performance in natural language processing (NLP), before the development of transformers, various neural network architectures were implemented, such as Recurrent Neural Networks (RNNs) which are neural networks focused on dealing with sequences. However, RNNs face two main limitations: the vanishing gradient, and the exploding gradient. These issues cause memory problems, meaning that in long sequences they tend to forget the initial parameters, preventing them from capturing the context of long sequences. To solve these issues, Long Short-Term Memory (LSTM) networks were developed, enabling them to capture only the important information within the sequence and increasing their memory capability. Nevertheless, LSTMs still lack the ability to capture and understand the context of long sequences. Furthermore, due to their architectural complexity, they require significant computational power and are difficult to train as they need extensive time [2].

To overcome these issues, the Transformer architecture was developed in 2017, implementing a self-attention mechanism that allows it to capture and understand the context of long sequences, demonstrating that "The Transformer architecture is superior to RNN-based models in computational efficiency" [3].

On the other hand, in the image processing field, Convolutional Neural Networks (CNNs) are typically implemented. However, various research studies have shown that transformers can achieve even better performance in image understanding [4]. Although the transformer was initially designed for translation purposes [2], since 2017 it has been implemented in different configurations across various fields, showing outstanding performance. For example, transformers have enabled the development of well-known large language models such as GPT, BERT, and T5. Thanks to these models, advances have been made in several areas, including the understanding and generation of audio, text, and images. This paper will present different applications of transformers and recent developments that have been made possible by transformers in the fields of text, image, and audio processing.

II. RELATED WORK

Since the release of transformers in 2017, different research has been conducted. For example, in [2], researchers identified that the major fields in which transformers are applied are: natural language processing (NLP), computer vision (CV), multi-modal applications, audio and speech processing, and signal processing. As shown in Figure 1, from about 650 transformer-based models, researchers found that approximately 40% of the models are implemented in the field of NLP, 31% in CV, 15% in multi-modal applications, 11% in audio/speech, and 4% in signal processing.

Moreover, each of these fields has different categories of tasks where the transformers have been applied. In NLP, models have been developed for language translation, classification and segmentation, question answering, text summarization, text generation, natural language processing, and automated symbolic reasoning. In CV, transformer-based models have been implemented for two different types of images: natural images and medical images. For natural images, models are involved in tasks such as image classification, recognition and object detection, image segmentation, and image generation. For medical images, models focus on tasks such as image segmentation, image classification and image translation.

Nevertheless, multi-modal applications involve tasks such as question answering, classification and segmentation, visual captioning, visual commonsense reasoning, text/image/video/speech generation, and cloud task computing. In the Audio/Speech field, tasks involve recognition/detection, separation, and classification. Finally, for signal processing, models have been developed that can work with wireless network signal processing and medical signal processing. This research highlights the wide-ranging impact and potential of transformer models.

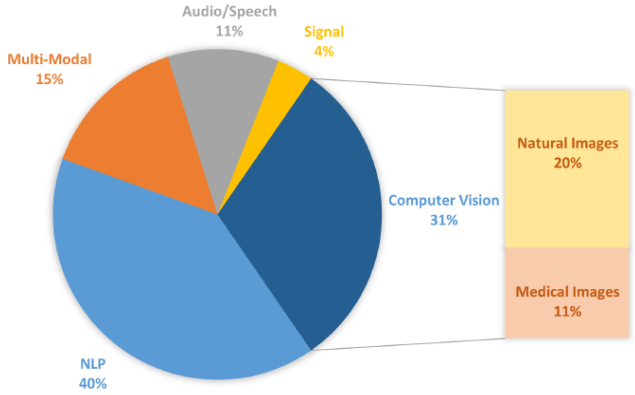

Figure 1. Proportion of transformer application in Top-5 fields [2].

Furthermore, transformers have also been tested in the Healthcare field [5] showing remarkable advances understanding clinical documentation, workflow management, treatments, and diagnosis, proving to be of great help in this sector. However, they still face various risks and limitations related to data privacy, as they handle sensitive patient data. Another issue is the potential for inaccurate information, as these models still require verification to avoid misdiagnosis or incorrect treatment, which can be dangerous for patients' health. Therefore, improvements are needed in these areas, but the healthcare field can benefit significantly from the implementation of LLMs as support tools for doctors. For example, models that detect cancer such as [6], receive images of cells and can identify certain patterns present in cancer cells, determining if the cells contain cancer and the stage of the disease.

Moreover, other research [7] shows that the self-attention mechanism allows transformers to analyze entire sentences at once, enabling them to comprehend speech. This research highlights the different categories where transformers are being applied in the field of speech. The categories mentioned are:

Automatic Speech Recognition (ASR): allows machines to recognize speech and convert it into the corresponding sequence of text.

Neural Speech Synthesis or Neural text to speech (TTS): converts text into synthesized speech.

Speech Translation: is the process of translating spoken speech into another language.

Speech Paralinguistics: is an area that focuses on non-verbal speech communications, such as tone, volume, speed and emotion.

Speech Enhancement: involves applying various algorithms to improve the quality of speech.

Spoken Dialogue Systems: Figure 2 shows tasks involved in this field, where 52% of the tasks are related to language understanding (intent recognition and semantic tagging), followed by turning-taking prediction and dialogue generation with 25% and the other areas representing 23% of the most popular tasks in this field. This demonstrates the wide variety of tasks performed by spoken dialogue systems.

Additionally, another category where transformers are effective is multi-modal applications, which involves models that can understand different input types and generate different outputs to solve real-world applications.

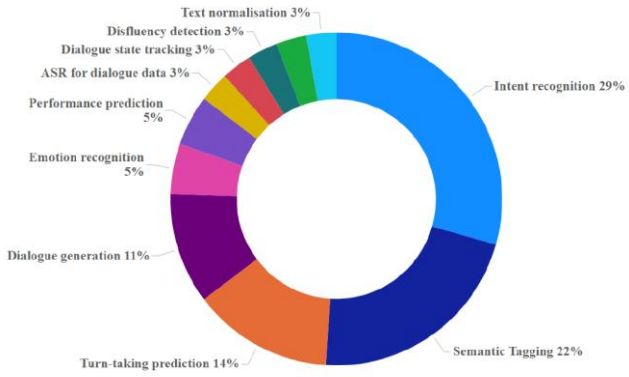

Figure 2. Tasks of Transformer-based Spoken Dialogue Systems [7].

This research illustrates the wide variety of applications where transformers have been involved, significantly improving the development in different fields related to speech processing.

III. TRANSFORMERS

The transformer [1] was introduced in 2017 to address the limitations of RNNs and LSTMs, particularly their difficulties with long-range sequences, demonstrating outstanding performance in understanding the context of sentences and even paragraphs. Since its release, various studies have been conducted to explain in more detail how it works and its different applications [2] [8].

The transformer is a neural network architecture consisting of an encoder-decoder structure, as illustrated in Figure 3. It implements the multi-head self-attention mechanism, which allows parallelization, making it more efficient during training and capable of handling large datasets. Unlike RNNs and LSTMs, which process the words of a sentence one by one, Transformers can process entire sequences at once due to their parallel processing capabilities[2] [8].

The transformer is composed of several components that play an important role in its performance. For example, the encoder is responsible for receiving the input. This input first passes through an embedding layer that converts each word into vectors. Then, the positional encoding is added to each vector to retain information about the order of the words in the sequence. The multi-head attention mechanism implements self-attention to compare all the words with each other, understanding the importance of each word into the sequence, and the context of the sequence. This is followed by the Feed Forward Neural Network and an Add & Norm layer, which normalizes the data [2] [8].

After the processing in the encoder, the information is sent to the decoder. The decoder receives the output of the encoder, converts the input into vectors, and applies positional encoding similarly to the encoder. Here, masked multi-head attention is applied. Unlike in the encoder, this masked multi-head attention allows the decoder to access only the previously generated words, preventing it from seeing future words. This is followed by another multi-head attention mechanism, a feed forward layer, and additional Add & Norm layers. The final layers include a linear transformation and SoftMax activation, which predict the most probable next word in the sequence. This predicted word is then added as an input to the decoder, and the process repeats until the end of the sequence [2] [8].

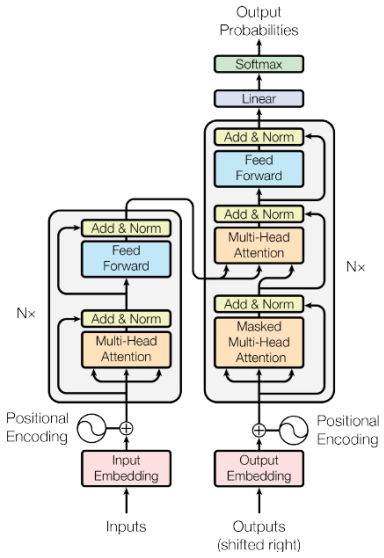

Figure 3. The Transformer – model architecture [1].

Although the initial transformer model, known as the 'vanilla' model, was designed specifically for language translation [2], the following sections will explore other applications that have been developed using transformer architectures.

IV. TRANSFORMERS IN LARGE LANGUAGE MODELS

Transformers have enabled significant advancements in different fields, particularly in natural language processing (NLP) making possible the development of large language models. These models implement different transformer configurations to suit specific tasks. For example, GPT (Generative Pre-trained Transformer) [9] implements only the decoder part of the transformer architecture, making it particularly efficient for text generation tasks. In contrast, BERT (Bidirectional Encoder Representation from Transformers) [10] implements the encoder part, making it a model focused on text understanding. Meanwhile, T5 (Text-to-Text Transfer Transformer) [11] is a model that implements the vanilla transformer architecture, with both the encoder and the decoder. T5 is a model capable of performing a wide range of tasks, including text summarization, translation, and question answering.

These models have demonstrated remarkable performance in the field of natural language processing. This has led to finding ways to implement them in diverse fields. For example, recent versions of these transformer-based models, such as GPT-4 and Gemini, implement multimodal capabilities. These capabilities allow them to receive and understand multiple types of data, such as images and audio,

expanding their applicability across diverse technological domains.

The following section will discuss current innovative projects that implement transformers in different configurations, demonstrating successful results in areas related to image, audio and text.

V. APPLICATIONS OF TRANSFORMERS:

As mentioned previously, transformers have enabled development in different areas. For example, they have facilitated the creation of platforms that implement AI with surprising results. One such example is Udio [12], an AI music generator platform that allows users to create their own songs by simply providing a prompt specifying the lyrics and the genre. Similarly, Stable Diffusion [13] is an AI model that generates images from text descriptions. These developments have been made possible by the implementation of transformers, which can understand the requirements for creating a song or an image and then decode the response into the respective output, whether it be a song or an image.

This section will explore different current developments where transformers have been implemented in various configurations across three areas: text, audio, and image, demonstrating the potential reach of transformers in these fields.

A. Text processing

The self-attention mechanism implemented by transformers allows them to get the context of a sequence, enabling the development of LLMs which have achieved remarkable performance in understanding natural language, obtaining good results in fields related to text. These models demonstrate good performance in machine translation, text summarization, text generation, question answering, classification, and segmentation tasks [2]. Among the most well-known LLMs are GPT-4 and Gemini, which have improved their capabilities to understand and generate text and now also comprehend images. Similarly, various research has focused on increasing LLM capabilities in different fields, demonstrating the ability of the models to understand and generate text for summarization, translation, and answering questions.

However, text can be implemented for different purposes, such as understanding a request and generating programming code (which is a type of text) to solve the request. For example, MobileMaker [14], is a model that is able to create prototype apps from a text description. The process of designing a prototype is crucial before releasing the final product, but creating and testing a prototype with users, and implementing their feedback, is a time-consuming process. Therefore, implementing models like MobileMaker improves this process by allowing real-time implementation of user feedback. The model work as follows: In the user interface (Figure 4), users can select from a dropdown list the inputs (A), actions (B), outputs (C), widgets and properties (D) to create the desired mobile application. Additionally, users can create the prototype using natural language by providing a text description of the desired mobile application using the "Revise with AI" panel (E). MobileMaker implements a multimodal LLM to receive the input and create a JSON file with all the necessary components and configurations for the app. Then, the model uses the JSON to render the components and functionality of the app (F). Finally, users can test their prototypes in real time on their phones (G).

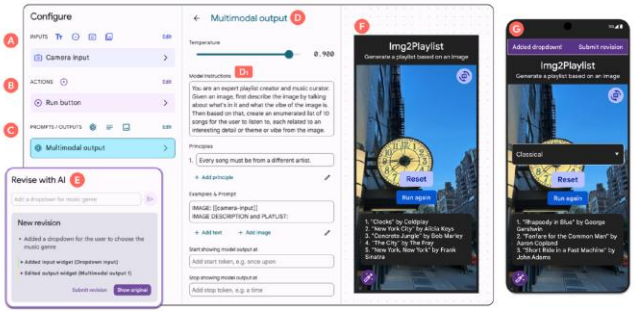

Figure 4. MobileMaker User Interface [14].

Transformers have also played an important role in extending the scope of LLMs, allowing them to interact with various tools. For example, Toolformer [15] is a large language model capable of answering different types of questions by calling the APIs of a question-answering system, a Wikipedia search engine, a calculator and a translation system. This improves the answers of the model by interacting with multiple tools in order to provide the best response.

Additionally, there are models that integrate thousands of different APIs such as Gorilla[16] and ToolLLM [17], models that are capable of calling thousands of different APIs to solve specific tasks. For example, if these models are asked to create an image, although these models are not able to generate images on their own, they can call the API of an image generation tool to create the desired image.

LLMind [18] is another example, LLMind is an AI framework that implements LLMs to control IoT devices to solve complex tasks. This model receives a request and generates a Python script to control the respective devices and fulfill the request.

Moreover, other research has also been conducted to provide ideas on how to expand the scope of the LLMs by combining them with operating systems. AIOS [19] is the proposal of an architecture that implements an LLM as the

brain of an operating system, managing resources such as memory consumption and storage, as well as handling user requests. The model receives a user's request, splits it into smaller tasks, and calls external tools to complete each task, ultimately solving the user's request efficiently. These examples demonstrate the significant potential for development in different fields where the transformers play a crucial role.

B. Audio processing

Transformers have also shown promising results in applications related to the audio field. However, similar to text-related models, the audio needs to be converted into tokens to be processed by the encoder [20], to achieve this, the audio must be converted into a spectrogram, which is the visual representation of the audio. The spectrogram is then divided into small patches, similar to the process in image processing, and these patches are converted into vectors to be processed by the model. Depending on the decoder configuration, the model can then provide an output in text, audio, or image format. In the following examples, models implement different configurations, using the encoder part to understand audio or the decoder part to generate audio.

Different models have been developed in this field. For example, research [21] describes a model designed to improve audio comprehension, enabling it to generate conversations or answer questions related to audio content. Additionally, research [22] suggests that, thanks to the self-attention mechanism of transformers, good results have been achieved in tasks requiring long-range coherence, which indicates that transformers may also perform well in the field of music.

However, models in the field of automatic speech understanding (ASU) face some limitations. As training models to understand speech requires access to voice datasets, and privacy issues can make it difficult to collect this type of data. Nevertheless, research [23] presents an approach that uses synthetic speech to train models when audio data is missing or insufficient. By employing a text-to-speech model to generate the necessary audio from text, synthetic speech can be used to overcome this limitation.

The LTU (Listen, Think and Understand) model "is the first multimodal large language models that focuses on general audio (rather than just speech) understanding" [24]. This model can understand not only speech but also the environment, interpreting background sounds to determine location, which helps it gain a better context from both the speech and the background sounds. LTU implements an audio spectrogram transformer (AST) as the audio encoder to process audio signals and then sends this information to a large language model called LLaMA [25], which comprehends the context of the audio and can answer questions.

Additionally, models capable of composing music have been developed. For example, ComposerX [26] is an approach that implements multi-agents to improve the composition abilities of LLMs. As shown in Figure 5, the model employs various LLMs (agents) to perform specific tasks. The leader agent receives a technical prompt that describes features such as tone, melody, rhythm, tempo, and instruments. It then creates and assigns different tasks to the melody, harmony and instrument agents, which generate their parts in ABC notation. The reviewer agent evaluates the composition and provides feedback to the agents to refine it. Finally, the arrangement agent compiles the outputs into a standardized ABC notation file, which is a universally readable format.

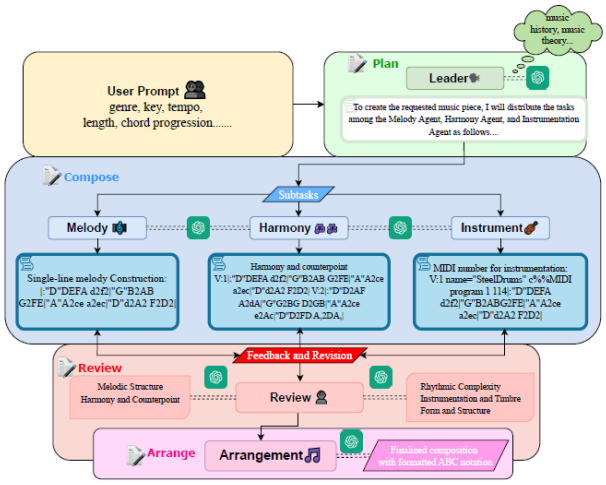

Figure 5. Agent Communication pattern of ComposerX [26].

C. Image processing

Traditionally, convolutional neural networks (CNNs) have been used for image-related tasks. However, research has shown that transformers can achieve even better performance in understanding images [4] [27]. As a result, transformers have been implemented in this field, demonstrating a good performance.

In order to process images with transformers, the image is first divided into patches, which are then treated as tokens. These tokens are passed through embeddings to convert them into their vector representations, and these vectors are then provided to the transformer. This method allows the transformer to identify the most important parts of an image, playing a crucial role, for example, when transmitting images

from rural areas with limited bandwidth by compressing the image and sending only the essential parts [27].

Figure 6 illustrates this process, showing that the vision transformer encoder receives an image, captures the most significant patches of it, compresses them, and sends them to the decoder, which then reconstructs the image with the important information.

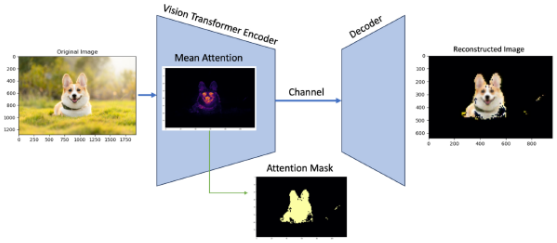

Figure 6. Transformer-aided semantic communication framework [27]

On the other hand, the application of transformers in image processing extends to various fields, including healthcare. For example, in [6], a model has been developed for detecting cancer cells and determining the stage of cancer using the same principles of image analysis. In this case, the model receives as input images of cells and identifies patterns to determine the presence and stage of cancer.

Beyond medical applications, transformers have also shown good results in creative and generative tasks. For example, transformers have allowed the development of models capable of creating deepfakes [28]. These models use an autoencoder architecture (encoder and decoder) to recognize a face, and then the decoder reconstructs the image using the face of another person.

Similarly, Stable Diffusion is an example of a model that generates images from text prompts. This principle has been applied in research [29] to generate indoor scenarios. The research implements an LLM to create indoor scenarios from text descriptions. This application is useful in the video game industry for designing indoor environments, as well as in architectural visualization and product promotion. When a company wants to launch a new product, they often need to invest in adapting scenarios to create promotional images. However, by implementing this model, virtual environments can be created and adapted for product promotion, making the process easier and cheaper.

To achieve this, vision-language models (VLM) are implemented to retrieve 3D models from a large database to generate the scene. The model receives a scene description, an LLM generates a detailed description of the scene, and then generates Python code based on this description. The Python code specifies the objects and their spatial relations. This code is then passed to the VLM, which takes the respective 3D models and uses them to create the scenario.

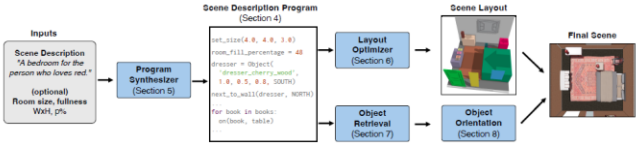

Figure 7. Schematic overview of the Indoor Scene Generation system [29].

Another innovative application is ChartLlama [30], a multimodal model that is able to generate and understand charts. This model was trained on its own dataset, obtaining outstanding performance in both understanding and generating charts. ChartLlama can receive text and graphs, and based on the desired output, it can explain the graphs or generate Python code to create the charts using the Matplotlib or Plotly libraries.

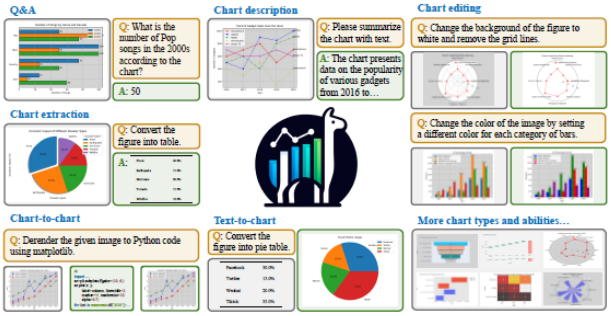

Figure 8. Capability demonstration of ChartLlama [30].

Sheet Music Transformer++ [31] demonstrates another application of image understanding. Although this model is implemented in the audio field to transcribe sheet music, it implements image detection to convert images into music notation. Sheet Music Transformer++ consists of an encoder and decoder. The encoder uses a convolutional neural network (CNN) to receive an image of a score and extract its features. These features are then passed to a transformer decoder, which generates the transcribed representation of the score in Kern format, as shown in Figure 9. This format facilitates compatibility with various software, making it easier to convert to other music representations. These types of applications play an important role in digitizing music files, as many important music archives in history are stored in physical formats. These formats can deteriorate over time and the music could be lost forever.

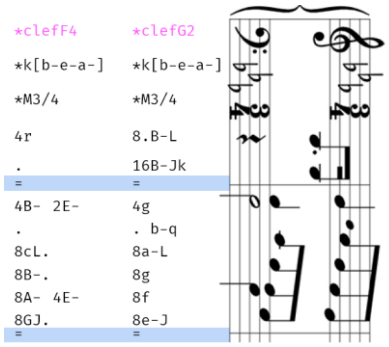

Figure 9. Example of a KERN score (left) aligned by reading order with its rendered music document (right) [31].

Behavior-LlaVA [32] represents another example of multimodal understanding, combining natural language processing and computer vision to comprehend images and videos. This model uses the base model LLaMA-VID [33], trained with the BLIFT (Behavior-LLaVA Instruction Fine-Tuning) dataset, which was built from Reddit and YouTube data, including images, videos and related information such as likes, views and comments. Therefore, Behavior-LLaVA is a model capable of understanding images and videos, creating descriptions about what is happening in the video and answering questions about it. This model is also able to generate simulated comments that people would post for the image or the video.

The base model LLaMA-VID [33] implements a visual encoder to process images or long videos, extracting information from each frame of the video and using a text decoder to generate descriptions and answer questions related to the video.

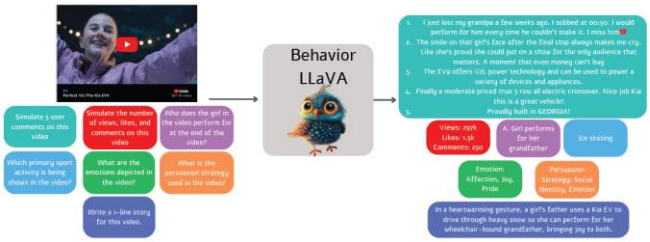

Figure 10. Overview of Behavior- LlaVA performance [32].

Moreover, the development of transformers in the image processing area has led to advances in the field of neuro-vision [34]. According to this research, when the brain sees an image, certain parts of the brain are activated, as depicted in Figure 11. The research implemented the idea of recreating images from the activated parts of the brain by decoding non-invasive brain recordings using structural magnetic resonance imaging (SMRI), which shows the specific brain areas activated by visual stimuli. For this purpose, they used a vision transformer developed for visual reconstruction. This transformer receives the brain activation image, identifies the important parts of the image (i.e., the activated areas), and based on these activations, understands the concept. Basically, the model understands where in the brain is located each concept and provides this information to the LLM. The LLM can then generate a text description of the concept, for example a giraffe (if the person is viewing a giraffe). Furthermore, if an image decoder is implemented, it is also possible to recreate the concept or the image that the person was viewing. These types of projects open new ways to explore and understand the human brain.

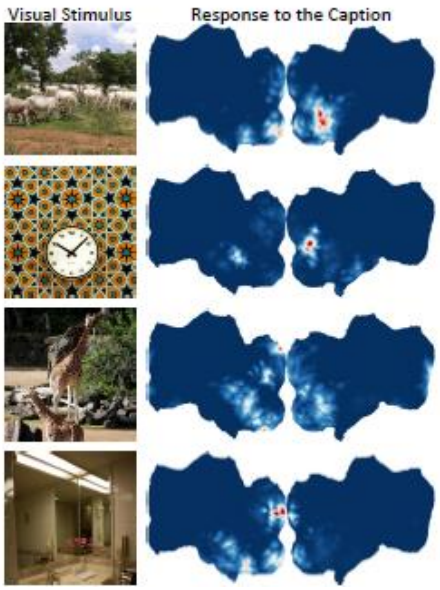

Figure 11. Heatmaps illustrating brain regions activated according to visual Stimulus. [34]

Finally, the development of multimodal models represents a significant advance in different fields, as they are capable of understanding different types of inputs, making them versatile across multiple areas. Typically, these models use only the encoder part, which enables them to understand diverse types of input. However, they are not able to generate different outputs.

To address this, the NExT-GPT model [35] implements both encoders and decoders, allowing it to generate different types of outputs, such as text, image, videos and audio, as shown in Figure 12. This model employs a separate encoder for each input type. Each encoder processes images, audio, text, or video and sends the information to a projection layer. This layer is responsible for translating the encoder's output into representations understandable by the LLM. This approach ensures that, regardless of the input type, all inputs are handled by the LLM.

The LLM then generates the desired task output. Rather than producing the response directly, it generates activation tokens to activate the specific decoder needed for generating the desired output. The output of the LLM is then passed

through another projection layer to create representations that the decoder can understand, ultimately producing the desired output. Furthermore, it is worth mentioning that in this research different pre-trained encoders and decoders are used to avoid the costs associated with training each encoder and decoder from scratch.

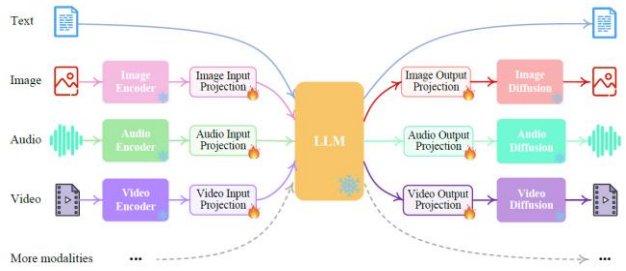

Figure 12. NExT-GPT structure *[35]*.

The projects reviewed in this paper are just a few examples of how transformers are applied across these three areas. However, they illustrate the wide range of applications where transformers have played an important role, enabling the development of technologies in different fields.

VI. CONCLUSION

To conclude, it has been shown that transformers have played a crucial role in the expansion of current technologies, particularly in the development of LLMs, which have enabled progress of new technologies in different fields. This paper has provided an overview of innovative projects that implement transformers in text, audio, and image processing. From creating mobile application prototypes based on text descriptions to developing models capable of understanding and generating multimodal content (audio, image and video), demonstrating the versatility of the transformers.

An important feature of transformers is their adaptability based on their configuration as encoders or decoders. Models implementing the encoder part specialize in understanding and processing content, extracting essential features. In contrast, models with a decoder configuration are focused on generating content such as text, images or audio. This capability allows transformers to be applied in multiple fields.

Moreover, their applications extend to areas such as cancer detection in the medical field and exploring human brain activity, including efforts to create text or visual representations from brain signals through image analysis.

However, it is important to mention the current challenges and limitations that transformers face, such as the need for large amounts of data and computational resources, as well as ethical issues related to privacy, particularly in sensitive fields like medicine. Nevertheless, as these fields continue to evolve, transformers are expected to remain leading technological advancements, driving innovation in numerous domains.

REFERENCES

[1] N. S. N. P. J. U. L. J. A. N. G. L. K. I. P. Ashish Vaswani, "Attention Is All You Need," 12 Jun 2017 . [Online]. Available: https://arxiv.org/abs/1706.03762. [Accessed 26 03 2024].

[2] H. E. A. E. J. B. N. D. G. R. W. P. Saidul Islam, "A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks," Juny 2023. [Online]. Available: https://arxiv.org/abs/2306.07303.

[3] M. L. A. J. S. Chenguang Wang, "Language Models with Transformers," arxiv, 17 10 2019. [Online]. Available: https://arxiv.org/abs/1904.09408. [Accessed 17 04 2024].

[4] L. B. A. K. D. W. X. Z. T. U. M. D. M. M. G. H. S. G. J. U. N. H. Alexey Dosovitskiy, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 22 Oct 2020 . [Online]. Available: https://arxiv.org/abs/2010.11929. [Accessed 22 5 2024].

[5] R. M. O. R. Kerstin Denecke, "Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks," 17 November 2023. [Online]. Available: https://www.researchgate.net/publication/378290762_Transformer_Models_in_Healthcare_A_Survey_and_Thematic_Analysis_of_Potentials_Shortcomings_and_Risks. [Accessed 26 6 2024].

[6] K. H. Z. G. L. C. Yulia Kumar, "Transformers and LLMs as the New Benchmark in Early Cancer Detection," January 2024. [Online]. Available: https://www.researchgate.net/publication/377267379_Transformers_and_LLMs_as_the_New_Benchmark_in_Early_Cancer_Detection#:~:text=The%20promising%20performance%20of%20the,comprehensive%20support%20to%20ALL%20patients.. [Accessed 28 5 2024].

[7] A. Z. H. C. F. S. M. S. J. Q. Siddique Latif, "Transformers in Speech Processing: A Survey," March 2023. [Online]. Available: https://arxiv.org/abs/2303.11607.

[8] R. E. Turner, "An Introduction to Transformer," 20 Apr 2023 . [Online]. Available: https://arxiv.org/abs/2304.10557. [Accessed 18 5 2024].

[9] R. M. C. S. G. S. Y. G. S. P. K. R. M. D. R. G. R. H. J. P. B. W. W. A. V. V. T. R. G. Gokul Yenduri, "Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," 11 May 2023. [Online]. Available: https://arxiv.org/abs/2305.10435. [Accessed 24 05 2024].

[10] M.-W. C. K. L. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for," 24 May 2019 . [Online]. Available: https://arxiv.org/abs/1810.04805. [Accessed 20 6 2024].

[11] N. S. A. R. K. L. S. N. M. M. Y. Z. W. L. P. J. L. Colin Raffel, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 23 Oct 2019 . [Online]. Available: https://arxiv.org/abs/1910.10683v4. [Accessed 2 6 2024].

[12] Udio, "AI Music Generator," [Online]. Available: https://www.udio.com/. [Accessed 18 6 2024].

[13] "Stable Diffusion," [Online]. Available: https://stablediffusionweb.com/. [Accessed 24 6 2024].

[14] M. X. L. A. J. F. V. T. M. T. C. J. C. Savvas Petridis, "In Situ AI Prototyping: Infusing Multimodal," 6 May 2024. [Online]. Available: https://arxiv.org/abs/2405.03806. [Accessed 10 5 2024].

[15] J. D.-Y. R. D. R. R. M. L. L. Z. N. C. T. S. Timo Schick, "Toolformer: Language Models Can Teach Themselves to Use Tools," 9 Feb 2023. [Online]. Available: https://arxiv.org/abs/2302.04761. [Accessed 28 5 2024].

[16] T. Z. X. W. J. E. G. Shishir G. Patil, "Gorilla: Large Language Model Connected with," 24 May 2023. [Online]. Available: https://arxiv.org/abs/2305.15334. [Accessed 18 5 2024].

[17] S. L. Y. Y. K. Z. L. Y. Y. L. Y. L. X. C. X. T. B. Q. S. Z. L. H. R. T. R. X. J. Z. M. G. D. L. Z. L. M. S. Yujia Qin, "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs," 06 10 2023. [Online]. Available: https://arxiv.org/abs/2307.16789. [Accessed 27 5 2024].

[18] Y. D. Q. Y. Y. S. Hongwei Cui, "LLMind: Orchestrating AI and IoT with LLM for Complex Task Execution," 6 January 2024. [Online]. Available: https://arxiv.org/abs/2312.09007. [Accessed 17 4 2024].

[19] Z. L. S. X. R. Y. Y. G. Y. Z. Kai Mei, "AIOS: LLM Agent Operating System," 26 Mar 2024. [Online]. Available: https://arxiv.org/abs/2403.16971. [Accessed 20 4 2024].

[20] X. X. Y. Y. M. W. W. W. M. D. P. Haohe Liu, "SemantiCodec: An Ultra Low Bitrate Semantic Audio Codec for General Sound," 30 Apr 2024. [Online]. Available: https://arxiv.org/abs/2405.00233. [Accessed 13 May 2024].

[21] Z. K. R. V. B. C. Arushi Goel, "Audio Dialogues: Dialogues dataset for audio and music understanding," 11 Apr 2024. [Online]. Available: https://arxiv.org/abs/2404.07616. [Accessed 9 5 2024].

[22] A. V. J. U. N. S. I. S. C. H. A. M. D. M. D. H. M. D. D. E. Cheng-Zhi Anna Huang, "MUSIC TRANSFORMER: GENERATING MUSIC WITH LONG-TERM STRUCTURE," 12 Sep 2018. [Online]. Available: https://arxiv.org/abs/1809.04281. [Accessed 15 5 2024].

[23] X. S. R. G. S. S. N. Tiantian Feng, "TI-ASU: Toward Robust Automatic Speech Understanding through Text-to-speech Imputation Against Missing Speech Modality," 27 Apr 2024. [Online]. Available: https://arxiv.org/abs/2404.17983. [Accessed 1 06 2024].

[24] H. L. A. H. L. L. K. J. G. Yuan Gong, "LISTEN, THINK, AND UNDERSTAND," 18 May 2023. [Online]. Available: https://arxiv.org/abs/2305.10790. [Accessed 8 5 2024].

[25] T. L. G. I. X. M. M.-A. L. T. L. B. R. N. G. E. H. F. A. A. R. A. J. E. G. G. L. Hugo Touvron, "LLaMA: Open and Efficient Foundation Language Models," 27 Feb 2023. [Online]. Available: https://arxiv.org/abs/2302.13971. [Accessed 8 6 2024].

[26] Q. Y. R. Y. Y. H. Y. W. X. L. Z. T. J. P. G. Z. H. L. Y. L. Y. M. J. F. C. L. E. B. W. W. G. X. W. X. Y. G. Qixin Deng, "ComposerX: Multi-Agent Symbolic Music Composition with LLMs," 30 Apr 2024 . [Online]. Available: https://arxiv.org/abs/2404.18081. [Accessed 13 May 2024].

[27] "Transformer-Aided Semantic Communications," May 2024 Matin Mortaheb, Erciyes Karakaya, Mohammad A. Amir Khojastepour, Sennur Ulukus. [Online]. Available: 26. [Accessed 5 2024 2].

[28] S. P. M. E. K. Alakananda Mitra, "The World of Generative AI: Deepfakes and Large Language Models," 6 Feb 2024. [Online]. Available: https://arxiv.org/abs/2402.04373. [Accessed 24 5 2024].

[29] M. G. D. H. H. S. M. S. J. Y. A. G. R. K. J. Q. A. W. K. F. D. R. Rio Aguina-Kang, "Open-Universe Indoor Scene Generation using LLM Program Synthesis and Uncurated Object Databases," 5 Feb 2024. [Online]. Available: https://arxiv.org/abs/2403.09675. [Accessed 1 5 2024].

[30] C. Z. X. C. X. Y. Z. W. G. Y. B. F. H. Z. Yucheng Han, "ChartLlama: A Multimodal LLM for Chart Understanding and Generation," 27 Nov 2023. [Online]. Available: https://arxiv.org/abs/2311.16483. [Accessed 10 5 2024].

[31] J. C.-Z. D. R. T. P. Antonio Ríos-Vila, "Sheet Music Transformer ++: End-to-End Full-Page Optical Music Recognition for Pianoform Sheet Music," 21 May 2024. [Online]. Available: https://arxiv.org/abs/2405.12105. [Accessed 17 5 2024].

[32] H. S. I. Y. K. S. V. B. R. R. S. C. C. B. K. Somesh Singh, "LLaVA Finds Free Lunch: Teaching Human Behavior Improves Content Understanding Abilities Of LLMs," 2 May 2024. [Online]. Available: https://arxiv.org/abs/2405.00942. [Accessed 13 May 2024].

[33] C. W. J. J. Yanwei Li, "LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models," 28 Nov 2023. [Online]. Available: https://arxiv.org/abs/2311.17043. [Accessed 6 9 2024].

[34] D. Z. X. H. L. F. Y. D. J. W. Q. Z. Y. Z. Guobin Shen, "Neuro-Vision to Language: Enhancing Visual Reconstruction and Language Interaction through Brain Recordings," 30 Apr 2024. [Online]. Available: https://arxiv.org/abs/2404.19438. [Accessed 9 May 2024].

[35] H. F. L. Q. W. J. T.-S. C. Shengqiong Wu, "NExT-GPT: Any-to-Any Multimodal LLM," 11 Sep 2023. [Online]. Available: https://arxiv.org/abs/2309.05519. [Accessed 20 May 2023].

[36] S. B. Łukasz Kaiser, "Can ActiveMemory Replace Attention?," March 2017. [Online]. Available: https://arxiv.org/abs/1610.08613.

[37] G. B. M. L. L.F. Costa, "Modular Smart Transformer architectures: An overview and proposal of a interphase architecture," April 2017. [Online]. Available: https://www.researchgate.net/publication/318332686_Modular_Smart_Transformer_architectures_An_overview_and_proposal_of_a_interphase_architecture.

[38] J. L. H. L. J. C. N. S. C. H. Dongxu Li, "Align and Prompt: Video-and-Language Pre-training with Entity Prompts," Dec 2021. [Online]. Available: https://arxiv.org/abs/2112.09583.

[39] A.-N. Sharkawy, "Principle of Neural Network and Its Main Types: Review," July 2020. [Online]. Available: https://www.researchgate.net/publication/34383759

1_Principle_of_Neural_Network_and_Its_Main_Types_Review. [Accessed 15 04 2024].

[40] J. Z. Y. W. Qingfeng Ji, "ASM: Audio Spectrogram Mixer," 20 Jan 2024. [Online]. Available: https://arxiv.org/abs/2401.11102. [Accessed 7 5 2024].

[41] K. S. K. Subodh Kamble, "Transformer Architecture for NetsDB," 8 5 2024 . [Online]. Available: https://arxiv.org/abs/2405.04807. [Accessed 12 5 2024].

[42] I. K. N. M. J. L. M. S. Y. Y. L. El Amine Cherrat, "Quantum Vision Transformers," 20 February 2024. [Online]. Available: https://arxiv.org/abs/2209.08167. [Accessed 2024 May 2024].

[43] X. C. S. X. Y. L. P. D. R. G. Kaiming He, "Masked Autoencoders Are Scalable Vision Learners," 11 Nov 2021 . [Online]. Available: https://arxiv.org/abs/2111.06377. [Accessed 20 5 2024].

[44] A. Z. H. C. F. S. M. S. J. Q. Siddique Latif, "Transformers in Speech Processing: A Survey," 21 Mar 2023. [Online]. Available: https://arxiv.org/abs/2303.11607. [Accessed 1 6 2024].