A COMPREHENSIVE SURVEY ON APPLICATIONS OF TRANSFORMERS FOR DEEP LEARNING TASKS

Saidul Islam[1], **Hanae Elmekki**[1], **Ahmed Elsebai**[1], **Jamal Bentahar**[1,2,*], **Nagat Drawel** [1], **Gaith Rjoub**[3,1], **Witold Pedrycz**[4,5,6]
[1]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada
[2]Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE
[3]Faculty of Information Technology, Aqaba University of Technology, Jordan
[4]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
[5]Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[6]Department of Computer Engineering, Istinye University, Sariyer/Istanbul, Turkiye

[*]**Corresponding Author's Email:** jamal.bentahar@concordia.ca
Contributing Authors' Emails: saidul.islam@concordia.ca; hanae.elmekki@mail.concordia.ca; ahmed.elsebai@outlook.com; n_drawe@encs.concordia.ca; grjoub@aut.edu.jo; wpedrycz@ualberta.ca

The authors contributed equally to this work.

ABSTRACT

Transformers are Deep Neural Networks (DNN) that utilize a self-attention mechanism to capture contextual relationships within sequential data. Unlike traditional neural networks and variants of Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM), Transformer models excel at managing long dependencies among input sequence elements and facilitate parallel processing. Consequently, Transformer-based models have garnered significant attention from researchers in the field of artificial intelligence. This is due to their tremendous potential and impressive accomplishments, which extend beyond Natural Language Processing (NLP) tasks to encompass various domains, including Computer Vision (CV), audio and speech processing, healthcare, and the Internet of Things (IoT). Although several survey papers have been published, spotlighting the Transformer's contributions in specific fields, architectural disparities, or performance assessments, there remains a notable absence of a comprehensive survey paper that encompasses its major applications across diverse domains. Therefore, this paper addresses this gap by conducting an extensive survey of proposed Transformer models spanning from 2017 to 2022. Our survey encompasses the identification of the top five application domains for Transformer-based models, namely: NLP, CV, multi-modality, audio and speech processing, and signal processing. We analyze the influence of highly impactful Transformer-based models within these domains and subsequently categorize them according to their respective tasks, employing a novel taxonomy. Our primary objective is to illuminate the existing potential and future prospects of Transformers for researchers who are passionate about this area, thereby contributing to a more comprehensive understanding of this groundbreaking technology.

Keywords: Transformer; Self-Attention; Deep Learning; Natural Language Processing (NLP); Computer Vision (CV); Multi-Modality.

1 INTRODUCTION

Deep Neural Networks (DNNs) have emerged as the predominant infrastructure and state-of-the-art solution for the majority of learning-based machine intelligence tasks in the field of artificial intelligence. Although various types of DNNs are utilized for specific tasks, the Multi-Layer Perceptron (MLP) represents the classic form of neural network which is characterized by multiple linear layers and nonlinear activation functions (Murtagh, 1990). For instance, in Computer Vision (CV), Convolutional Neural Networks (CNNs) incorporate convolutional layers to process images, while Recurrent Neural Networks (RNNs) employ recurrent cells to process sequential data, particularly in Natural Language Processing (NLP) (O'Shea & Nash, 2015, Mikolov et al., 2010). Despite the wide use of RNNs, they exhibit certain limitations. One of the major issues with conventional networks is that they have short-term dependencies associated with exploding and vanishing gradients. In contrast, to achieve good results in NLP, long-term dependencies must be captured. Additionally, RNNs are slow to train due to their sequential data processing and computational approach (Giles et al., 1995). To address these issues, the Long-Short-Term Memory (LSTM) version of recurrent networks was developed, which improves the gradient descent problem of RNNs and increases the memory range of NLP tasks (Hochreiter & Schmidhuber, 1997). However, LSTMs still struggle with the problem of sequential processing, which hinders the extraction of the actual meaning of the context. To tackle this challenge, bidirectional LSTMs were introduced, which process natural language from both directions, i.e., left to right and right to left,

and then concatenate the outcomes to obtain the context's actual meaning. Nevertheless, this technique still results in a slight loss of the true meaning of the context (Graves & Schmidhuber, 2005, Li et al., 2020b).

Transformers are a type of DNNs that offer a solution to the limitations of sequence-to-sequence (seq-2-seq) architectures, including short-term dependency of sequence inputs and the sequential processing of input, which hinders parallel training of networks. Transformers leverage the multi-head self-attention mechanism to extract features, and they exhibit great potential for application in NLP. Unlike traditional recurrence methods, Transformers utilize attention to learn from an entire segment of a sequence, using encoding and decoding blocks. One key advantage of Transformers over LSTM and RNNs is their ability to capture the true meaning of the context, owing to their attention mechanism. Moreover, Transformers are faster since they can work in parallel, unlike recurrent networks, and can be calculated using Graphic Processing Units (GPUs), allowing for faster computation of tasks with large inputs (Niu et al., 2021, Vaswani et al., 2017, Zheng et al., 2020b). The advantages of the Transformer model have inspired deep learning researchers to explore its potential for various tasks in different fields of application (Ren et al., 2023), leading to numerous research papers and the development of Transformer-based models for a range of tasks in the field of artificial intelligence (Yeh et al., 2019, Wang et al., 2019, Reza et al., 2022).

In the research community, the importance of survey papers in providing a productive analysis, comparison, and contribution of progressive topics is widely recognized. Numerous survey papers on the topic of Transformers can be found in the literature. Most of them are addressing specific fields of application (Khan et al., 2022, Wang et al., 2020a, Shamshad et al., 2023), compare the performance of different model (Tay et al., 2023, Fournier et al., 2023, Selva et al., 2023), or conduct architecture-based analysis (Lin et al., 2022). Nevertheless, a well-defined structure that comprehensively focuses on the top application fields and systematically analyzes the contribution of Transformer-based models in the execution of various deep learning tasks within those fields is still widely needed.

Indeed, conducting a survey on Transformer applications would serve as a valuable reference source for enthusiastic deep-learning researchers seeking to gain a better understanding of the contributions of Transformer models in diverse fields. Such a survey would enable the identification and discussion of potential models, their characteristics, and working methodology, thus promoting the refinement of existing Transformer models and the discovery of novel Transformer models or applications. To address the absence of such a survey, this paper presents a comprehensive analysis of all Transformer-based models, and identifies the top five application fields, namely NLP, CV, Multi-Modality, Audio & Speech, and Signal Processing, and proposes a taxonomy of Transformer models, with significant models being classified and analyzed based on their task execution within these fields. Furthermore, the top-performing and significant models are analyzed within the application fields, and based on this analysis, we discuss the future prospects and challenges of Transformer models.

1.1 Contributions and Motivations

Although several survey articles on the topic of Transformers already exist in the literature, our motivations for conducting this survey stem from two essential observations. First, most of these studies have focused on Transformer architecture, model efficiency, and specific artificial intelligence fields, such as NLP, CV, multi-modality, audio & speech, and signal processing. They have often neglected other crucial aspects, such as the Transformer-based model's execution in deep learning tasks across multiple application domains. We aim in this survey to cover all major fields of application and present significant models for different task executions. The second motivation is the absence of a comprehensive and methodical analysis encompassing various prevalent application domains, and their corresponding utilization of Transformer-based models, in relation to diverse deep learning tasks within distinct fields of application. We propose a high-level classification framework for Transformer models, which is based on their most prominent fields of application. The prominent models are categorized and evaluated based on their task performance within the respective fields. In this survey, we highlight the application domains of Transformers that have received comparatively greater or lesser attention from researchers. To the best of our knowledge, this is the first review paper that presents a high-level classification scheme for the Transformer-based models and provides a collection of criteria that aim to achieve two objectives: (1) assessing the effectiveness of Transformer models in various applications; and (2) assisting researchers interested in exploring and extending the capabilities of Transformer-based models to new domains. Moreover, the paper provides valuable insights into potential future applications and highlights unresolved challenges within this field.

The paper follows the organization depicted in the visual abstract shown in Figure 1. To provide context, the motivation behind this paper has been discussed in the current section, and Section 2 explains the preliminary concepts essential for the rest of the paper. A comprehensive account of the systematic methodology used to search for relevant research articles is detailed in Section 3. Section 4 presents a review of related papers, highlighting similarities and differences with our paper. Section 5 employs a pie chart to illustrate the distribution of proposed Transformer models across various fields. Importantly, Section 6 introduces a taxonomy of significant Transformer models within different application domains. This section also extends into various fields, providing an in-depth analysis of these models and their related tasks. Section 7 outlines potential directions for future research and challenges, while Section 8 serves as the conclusion, summarizing the key findings and contributions of the study.

2 Preliminaries

Before delving into the literature of Transformers, let us describe some concepts that will be used throughout this article.

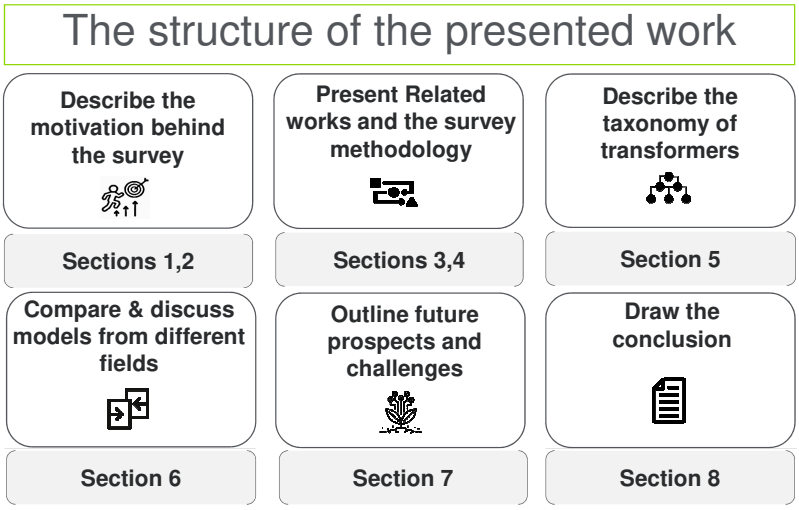

Figure 1: Visual abstract of the survey paper

2.1 Transformer Architecture

The development of the Transformer model was mainly motivated by the inclusion of the attention mechanism (Vaswani et al., 2017). The main goal was to improve how the model processes sequences of data. Since then, numerous models have emerged, drawing inspiration from the original Transformer model, with the aim of addressing a wide array of tasks across various fields. The exceptional performance exhibited by Transformers, especially in achieving state-of-the-art benchmarks in the domain of NLP translation models, has propelled their widespread adoption. While some models have adhered to the unaltered Vanilla Transformer architecture, others have selectively employed either the encoder or decoder module of the Transformer model. Consequently, the nature of the task and the performance of Transformer-based models can exhibit considerable variability depending on the specific architectural choices made. Nonetheless, a pivotal and extensively integrated component within Transformer models is the self-attention mechanism, which constitutes an indispensable element of their fundamental functionality. All Transformer-based models incorporate either the self-attention mechanism or various iterations of Multi-Head Attention (MHA), typically serving as the foundational learning layer in the architecture. In light of the paramount importance of self-attention in the architecture of the Transformer model, a comprehensive examination of this mechanism becomes imperative.

2.1.1 Attention Mechanism

In the late 1980s, early attempts to incorporate attention mechanisms into neural networks emerged. One pioneering study enhanced the Neocognitron by introducing selective attention (Fukushima, 1987). Subsequently, research focused on tasks such as target detection and object recognition (Schmidhuber & Huber, 1991). These foundational works laid the groundwork for the exploration of attention mechanisms in neural networks. Throughout the 2000s, efforts persisted to refine attention mechanisms in neural networks. Studies included the development of a computational model simulating human eye movements for object class detection (Zhang et al., 2005), attention-based systems for object identification and tracking in videos (Gould et al., 2007), and a computational model of visual selective attention for detecting relevant portions of images (Meur et al., 2006). The early 2010s witnessed further endeavors to enhance the utility of attention mechanisms in neural networks. Researchers introduced models that seamlessly integrated attentional orienting and object recognition (Miau & Itti, 2001) and devised models for visual pattern recognition with selective attention (Salah et al., 2002). These studies explored a wide range of applications of attention in visual perception and recognition. In 2015, attention mechanisms gained significant prominence with the introduction of a groundbreaking approach to Neural Machine Translation (NMT) (Bahdanau et al., 2015). This approach extended conventional NMT models by incorporating a Bidirectional RNN (BiRNN) encoder to capture semantic details more effectively. This marked the inception of attention mechanisms becoming a central and indispensable concept in the fields of NLP and deep learning (Soydaner, 2022). In the year 2017, a significant milestone was reached with the introduction of a pioneering attention-based neural network, named the "Transformer". This innovative architecture was developed to address the limitations observed in other neural networks, such as RNNs, particularly in the context of encoding long-range dependencies within sequences, especially in language translation tasks (Vaswani et al., 2017). The Transformer architecture is characterized by its exclusive reliance on self-attention mechanisms, devoid of any recurrence or convolution. It garnered remarkable acclaim for its performance in machine translation tasks, achieved by replacing conventional recurrent layers with self-attention mechanisms (Soydaner, 2022). The incorporation of self-attention within the Transformer model marked a significant advancement in attention mechanisms. It facilitated enhanced capture of local features while diminishing the reliance on external information, thereby contributing to the model's improved performance.

In the original Transformer architecture, the attention mechanism is implemented using the "Scaled Dot Product Attention" technique, which relies on three fundamental parameter matrices: Query (Q), Key (K), and Value (V). Each of these matrices

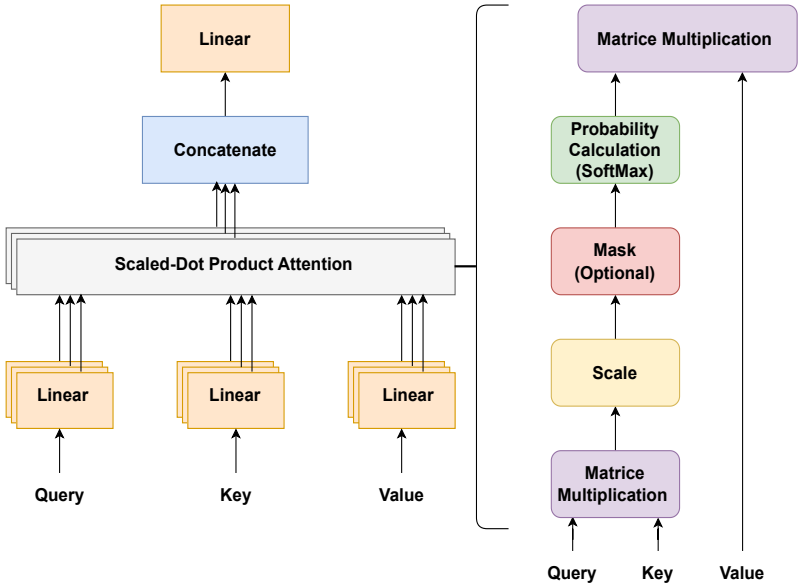

Figure 2: Multi-Head Attention & Scaled Dot-Product Attention(Vaswani et al., 2017)

encapsulates an encoded representation of every input element within the sequence (Vaswani et al., 2017). To derive the ultimate output of the attention process, the SoftMax function is applied, resulting in a probability score computed from the weighted combination of these three matrices, as illustrated in Figure 2.
Mathematically, the scaled dot product attention function is computed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V$$

The matrices Q and K represent the Query and Key vectors respectively, both having a dimension of dk, while the matrix V represents the values vectors.

2.1.2 Multi-Head Attention (MHA)

The application of the scaled dot-product attention function in parallel within the MHA module is essential for extracting the maximum dependencies among different segments in the input sequence. Each head denoted by k performs the attention mechanism based on its own learnable weights W^{kQ}, W^{kK}, and W^{kv}. The attention outputs calculated by each head are subsequently concatenated and linearly transformed into a single matrix with the expected dimension (Vaswani et al., 2017).

$$headk = Attention(QW^{kQ}, KW^{kK}, VW^{kV})$$
$$MultiHead(Q, K, V) = Concat(head1, head2, ...headH)W^0$$

The utilization of MHA facilitates the neural network in learning and capturing diverse characteristics of the input sequential data. Consequently, this enhances the representation of the input contexts, as it merges information from distinct features of the attention mechanism within a specific range, which could be either short or long. This approach allows the attention mechanism to jointly function, which results in better network performance (Vaswani et al., 2017).
The initial Transformer architecture was developed based on the auto-regressive sequence transduction model, comprising two primary modules, namely Encoder and Decoder. These modules are executed multiple times, as required by the task at hand. Each module comprises several layers that integrate the attention mechanism. Particularly, the attention mechanism is executed in parallel multiple times within the Transformer architecture, which explains the presence of multiple "Attention Heads" (Vaswani et al., 2017).

2.1.3 Encoder Module

The stacked module within the Transformer architecture comprises two fundamental layers, namely the Feed-Forward Layer and MHA Layer. In addition, it incorporates Residual connections around both layers, as well as two Add and Norm layers, which play a pivotal role (Vaswani et al., 2017). In the case of text translation, the Encoder module receives an embedding input that is generated based on the input's meaning and position information via the Embedding and Position Encoding

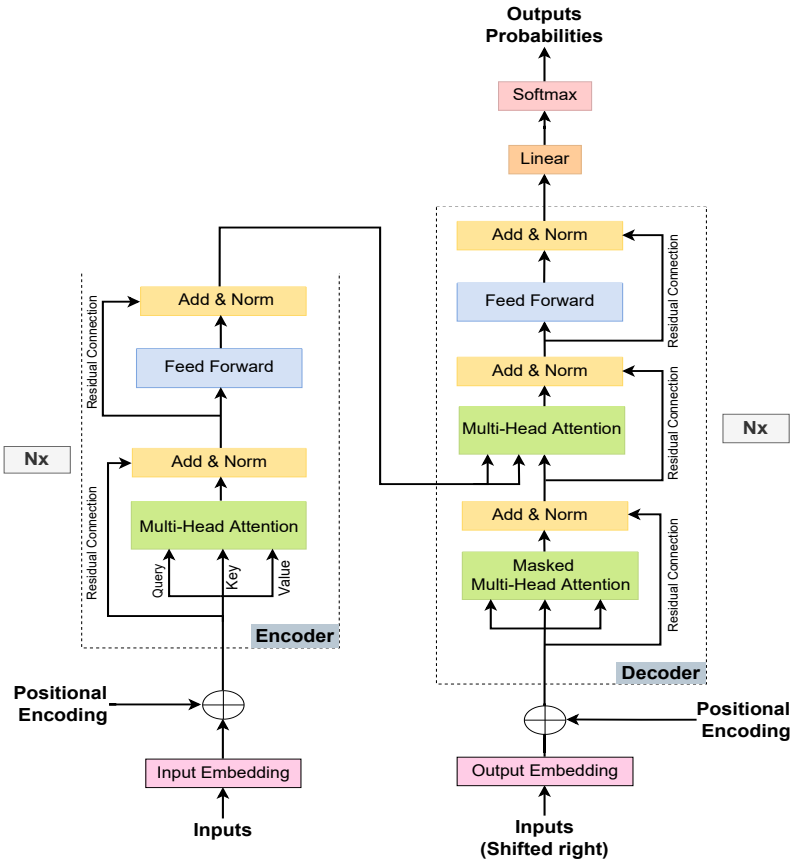

Figure 3: Transformer Architecture (Vaswani et al., 2017)

layers. From the embedding input, three parameter matrices are created, namely the Query (Q), Key (K), and Value (V) matrices, along with positional information, which are passed through the MHA layer. Following this step, the Feed-Forward layer addresses the issue of rank collapse that can arise during the computation process. Additionally, a normalization layer is applied to each step, which reduces the dependencies between layers by normalizing the weights used in gradient computation within each layer. To address the issue of vanishing gradients, the Residual Connection is applied to every output of both the attention and feed-forward layers, as illustrated in Figure 3.

2.1.4 Decoder module

The Decoder module in the Transformer architecture is similar to the Encoder module, with the inclusion of additional layers such as masked MHA. In addition to the Feed-Forward, MHA, Residual connection, and Add and Norm layers, the Decoder also contains masked MHA layers. These layers use the scaled dot product and mask operations to exclude future predictions and consider only previous outputs. The attention mechanism is applied twice in the Decoder: one for computing attention between elements of the targeted output and another for finding attention between the encoding inputs and targeted output. Each attention vector is then passed through the feed-forward unit to make the output more comprehensible to the layers. The generated decoding result is then caught by Linear and SoftMax layers at the top of the Decoder to compute the final output of the Transformer architecture. This process is repeated multiple times until the last token of a sentence is found (Vaswani et al., 2017), as illustrated in Figure 3.

3 Research Methodology

In this survey, we collect and analyze the most recent surveys on Transformers that have been published in refereed journals and conferences with the aim of studying their contributions and limitations. To gather the relevant papers, we employed a two-fold strategy: (1) searching using several established search engines and selected papers based on the keywords “survey”, “review”, “Transformer”, “attention”, “self-attention”, “artificial intelligence”, and “deep learning; and (2) evaluating the selected papers and eliminating those that were deemed irrelevant for our study. A detailed organization of our survey is depicted in Figure 4.

Indeed, by means of a comprehensive examination of survey papers and expert discussions on deep learning, we have identified the top five domains of application for Transformer-based models, these are: (i) NLP, (ii) CV, (iii) multi-modality, (iv) audio/speech, and (v) signal processing. Subsequently, we performed a systematic search for journal and conference papers

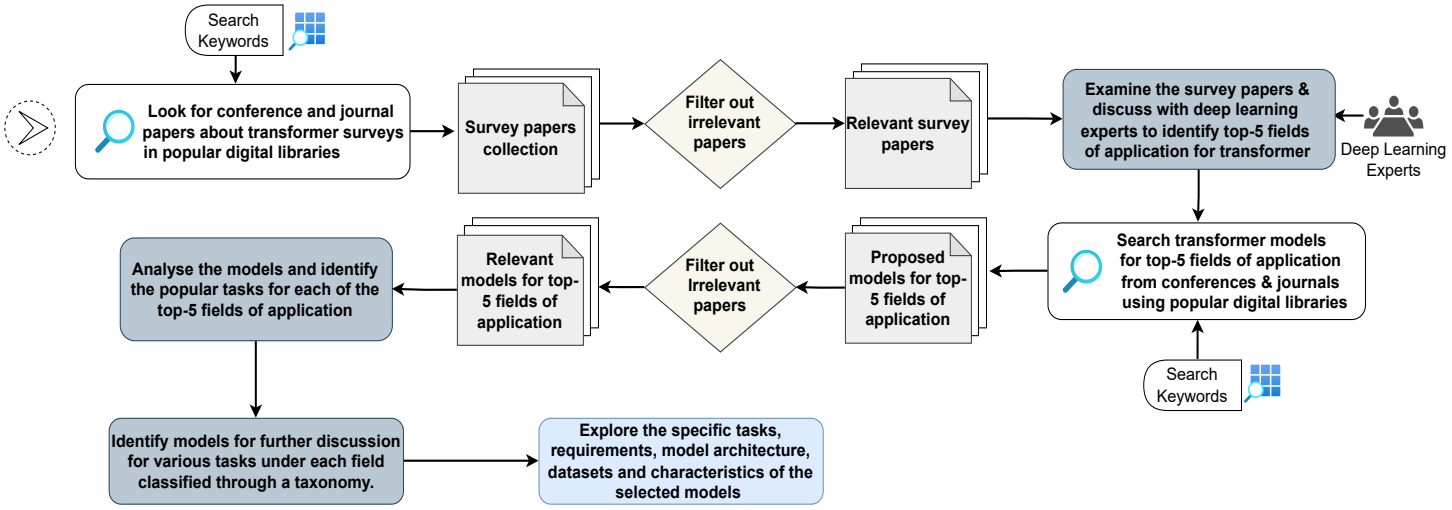

Figure 4: Survey Methodology

that presented Transformer-based models in each of the aforementioned fields of application, utilizing the keywords presented in Table 1. Our search yielded a substantial number of papers for each field, which we thoroughly reviewed and evaluated. We selected papers that proposed novel Transformer-based or Transformer-inspired models for deep learning tasks, while disregarding others. Through our examination of this extensive collection of models, we have identified prevalent deep-learning tasks associated with each field of application. As we have examined more than 600 Transformer models during this process, it has become exceedingly difficult to classify such a large number of models and conduct thorough analyses of each task within every field of application. Therefore, we have opted to perform a more comprehensive analysis of a number of Transformer models for each task within every field of application. These models were selected based on specific criteria and an in-depth analysis was carried out accordingly. The selected models are as follows:

1. The Transformer-based models that have been proposed to execute a deep learning task for the first time and opened up new path for research in the field of Transformer applications.
2. The models that have proposed alternative or novel approaches to implementing the Transformer's attention mechanism, as compared to the vanilla architecture, such as introducing a new attention mechanism or enhancing the position encoding module.
3. The Transformer models have had a significant impact in the field, with higher citation rates, and have been widely accepted by the scientific community. Models that have also contributed to breakthroughs in the advancement of Transformer applications.
4. The models and their variants have been proposed for the purpose of applying the Transformer technology to real-world applications, with the aim of achieving superior performance results in comparison to other deep learning methods.
5. The Transformer-based models generated a significant buzz within the theoretical and applied artificial intelligence community.

In the field of application, we have classified the selected models based on their task execution and developed a taxonomy of Transformer applications. Our analysis involved a comprehensive examination of the models, including their structures, characteristics, operational methods, and datasets, among others. Based on this investigation, we provide an in-depth discussion of the potential future applications of Transformers. To conduct this research, we consulted various prominent research repositories, such as "AAAI", "ACM", "ACL", "CVPR", "ICML", "ICLR", "ICCV", "NeurIPS", "LREC", "IEEE", "PMLR", "National Library of medicine","SCOPUS", "MDPI", "ScienceDirect", and "Cornell University-arxiv library". Table 1 depicts the category of the selected models.

Fields of Application	**Keywords for Paper Search**	**Tasks of Application**	**Number of Papers**	
			Relevant Models using Keywords	**Selected Models for Taxonomy**
Natural Language Processing (NLP)	NLP, Text, Text Processing, Transformer, Attention, Self-attention, Multi-head attention, Language model	Language Translation	257	25
		Text Classification & Segmentation		
		Question Answering		
		Text Summarization		

Continued on next page

Table 1 – continued from previous page

Fields of Application	Keywords for Paper Search	Tasks of Application		Number of Papers	
				Relevant Models using Keywords	Selected Models for Taxonomy
		Text Generation			
		Natural Language Reasoning			
		Automated Symbolic Reasoning			
Computer Vision	Transformer, Attention, Image, Self-attention, Natural image, medical image, Biomedical, Health, Image processing, Computer vision, Vision	Natural Image Processing	Recognition & Object Detection	197	27
			Image Classification		
			Image Segmentation		
			Image Generation		
		Medical Image Processing	Image Segmentation		
			Image Classification		
			Image Translation		
Multi-modal	Transformer, Attention, Self-attention, Multi-head attention, Language-vision, Multi-modality, Text-image, Image-text, Image-audio-text, Text-audio, Audio-text, Vision-language, Multi-modal	Classification & Segmentation		94	21
		Visual Question Answering			
		Visual Captioning			
		Visual Commonsense Reasoning			
		Text/Image/Speech Generation			
		Cloud Task Computing			
Audio & Speech	Transformer, Attention, Self-attention, Audio, Speech, Audio processing, Speech processing	Audio & Speech Recognition		70	15
		Audio & Speech Separation			
		Audio & Speech Classification			
Signal Processing	Transformer, Attention, Self-attention, Signal, Signal processing, Wireless, Wireless signal, Wireless network, Biosignal, Medical signal	Wireless network Signal processing		23	11
		Medical Signal Processing			

Table 1: Illustration of Transformer Model Applications, Search Keywords, Common Deep Learning Tasks, Retrieved Relevant Papers, and Models Selected for Taxonomy and In-Depth Analysis.

4 Related Work

Transformers have been the subject of numerous surveys in recent years due to their effectiveness and wide range of applications. We identified and collected over 50 survey papers related to Transformers from various digital libraries. Subsequently, we conducted a meticulous examination of these surveys to identify the most significant ones. Our selection process involved considering surveys that had been published in reputable conferences and journals and had garnered a high number of citations. We excluded papers that were not yet published or did not meet our criteria. Out of the initial pool, we selected 17 significant survey papers for an in-depth analysis of their content. During this analysis, we thoroughly explored the topics covered in these papers and investigated the specific fields of work and applications they focused on. We placed particular emphasis on examining the similarities and differences between these existing surveys and our own paper. Our investigation uncovered a notable pattern in the existing surveys. Many of them predominantly concentrated on aspects related to the architecture and efficiency of Transformers. In contrast, some surveys were dedicated solely to exploring the applications of Transformers in NLP and CV. However, only a few surveys delved into the utilization of Transformers in multi-modal scenarios, where both text and image data are involved. We have summarized these findings, along with supporting details, in Table 2.

Several review papers have prominently focused on conducting analyses of Transformers, particularly in terms of their architecture and performance. One noteworthy survey paper, titled "A Survey of Transformers," offers a comprehensive examina-

tion of various X-formers and introduces a taxonomy based on architecture, pre-training, and application (Lin et al., 2022). Additionally, another survey paper on Transformers, titled "Efficient Transformers: A survey," is dedicated to comparing the computational power and memory efficiency of X-formers (Tay et al., 2023). Furthermore, there is a paper that specifically explores light and fast Transformers, investigating different efficient alternatives to the standard Transformers (Fournier et al., 2023). In the field of NLP, there is a noteworthy survey paper titled "Visualizing Transformers for NLP: A Brief Survey" (Brasoveanu & Andonie, 2020). This study primarily focuses on exploring the various aspects of Transformers that can be effectively analyzed and comprehended through the application of visual analytics techniques. In a related context, another survey paper delves into the domain of pre-trained Transformer-based models for NLP (Subramanyam et al., 2021). This particular study extensively discusses the pretraining methods and tasks employed in these models and introduces a taxonomy that effectively categorizes the wide range of Transformer-based Pre-Trained Language Models (T-PTLMs) found in the literature Furthermore, there is a paper titled "Survey on Automatic Text Summarization and Transformer Models Applicability" that concentrates on the utilization of Transformers for text summarization tasks and presents a Transformer model designed to address the challenge of handling long sequences as input (Wang et al., 2020a). In a different context, another survey paper explores the application of Bidirectional Encoder Representations from Transformers (BERT) as a word-embedding tool, using multi-layer BERT for this purpose (Kaliyar, 2020). Additionally, the utilization of Transformers for detecting various levels of emotions from text-based data is investigated in a paper titled "Transformer models for text-based emotion detection: a review of BERT-based approaches" (Acheampong et al., 2021). Lastly, another paper delves into the use of Transformer language models within different information systems (Gruetzemacher & Paradice, 2022). This study focuses on harnessing Transformers as text miners to extract valuable data from the vast repositories of large organizations.

Due to significant advancements in image processing tasks and their remarkable applications in CV using Transformer models in recent years, these models have garnered significant attention among CV researchers. For example, the paper titled "Transformers in Vision: A Survey" provides a comprehensive overview of the existing Transformer models in the field of CV and categorizes these models based on popular recognition tasks (Khan et al., 2022). Furthermore, a meticulous survey was conducted to thoroughly analyze the strengths and weaknesses of the leading "Vision Transformers." This study placed considerable emphasis on examining the training and testing datasets associated with these top-performing models, providing valuable insights into their performance and suitability for various applications (Han et al., 2023). Another survey paper conducted a comparative analysis of Transformer models designed for image and video data, focusing on their performance in classification tasks (Selva et al., 2023). Recent advancements in CV and multi-modality are highlighted in a separate survey paper (Xu et al., 2022). This survey compares the performance of various Transformer models and offers insights into their pre-training methods. Additionally, an existing survey provides a comprehensive description of several Transformer models developed specifically for medical images; however, it does not cover information related to medical signals (Li et al., 2023). Another paper offers an overview of Transformer models developed within the medical field, but it primarily addresses medical images, excluding medical signals from its scope (Shamshad et al., 2023).

The growing popularity of multi-modality in deep learning tasks has led to the emergence of several surveys focusing on Transformers in the multi-modal domain. One paper aimed to categorize Transformer vision-language models based on tasks, providing summaries of their corresponding advantages and disadvantages. Furthermore, this survey covered video-language pre-trained models, classifying them into single-stream and multi-stream structures while comparing their performance (Ruan & Jin, 2022). In another survey titled "Perspectives and Prospects on Transformer Architecture for Cross-Modal Tasks with Language and Vision," researchers explored the application of Transformers in multi-modal visual-linguistic tasks (Shin et al., 2022). Beyond NLP, CV, and multi-modality, Transformers are gaining significant attention from researchers in fields like time series analysis and reasoning tasks, although many of these surveys are currently unpublished.

After conducting a thorough analysis of existing survey papers, we identified a gap in the literature—a lack of a comprehensive survey on Transformers that delves into their wide-ranging applications across various deep learning tasks in specific fields. In response to this gap, our paper aims to fill this void. We started by extensively researching and cataloging all available Transformer-based models. Subsequently, we identified the top five fields where Transformer models have made significant contributions: NLP, CV, Multi-modal, Audio and Speech, and Signal Processing. To enhance understanding of Transformer models' impact, we proposed a taxonomy that categorizes these models based on these top five fields. Within this taxonomy, we classify and analyze the top-performing models based on their task executions within their respective fields. Our survey illuminates various aspects of Transformer-based models, their tasks, and their applications across different fields. Furthermore, it highlights fields of Transformer applications that have received varying levels of attention from researchers. Through our analysis, we provide insights into the future prospects and possibilities of Transformer applications across diverse fields. A primary objective of this survey is to serve as a comprehensive reference source, aiding researchers in better understanding the contributions of Transformer models in various domains. It also offers valuable insights into the characteristics and execution methods of models that have significantly improved task performance within their domains. This paper serves as a valuable resource for researchers seeking to explore and expand the application of Transformers in innovative ways

Approach	Fields of Application	Similarities	Differences
Fournier et al. (2023)	Performance /Architecture	Classification based on attention mechanism or architecture modification.	Surveyed efficient alternatives of standard Transformers, categorized by modifications in attention mechanism or architecture. Our classification is application-driven.
Lin et al. (2022)	Performance /Architecture	Proposed taxonomy of X-formers covering several fields.	Compared X-formers from architectural, pretraining, and application perspectives. Our survey includes tasks in wireless/medical signal processing and cloud computing.
Tay et al. (2023)	Performance /Architecture	Taxonomy for Transformer models in language and vision domains.	Focused on computational power and memory efficiency. Our survey expands to five fields including NLP, computer vision, multimodality, audio/speech, and signal processing.
Brasoveanu & Andonie (2020)	NLP	Explanation of Transformer architecture and features.	Focused on visualization techniques for Transformer architectures. Our survey covers five application fields and various tasks.
Wang et al. (2020a)	NLP	Survey on text summarization using Transformers.	Proposed a Transformer-based summarizer for long text input. Our survey is more comprehensive, covering various applications.
Kaliyar (2020)	NLP	Discussion on NLP tasks that BERT performs.	Focused on using BERT for embedding text. Our survey is broader in scope.
Acheampong et al. (2021)	NLP	Survey on Transformers for emotion detection.	Detailed survey on emotion detection. Our survey includes sentiment analysis but is more varied in tasks.
Gruetzemacher & Paradice (2022)	NLP	Survey on Transformers in text-mining.	Focused on text mining for organizations. Our survey is broader, covering various tasks.
Selva et al. (2023)	Computer Vision	Overview of Transformers for images and video data.	Focused solely on image and video data. Our survey is more comprehensive.
Kalyan et al. (2022)	Medical NLP	Overview of Transformer-based BPLMs for various NLP tasks.	Focused on biomedical NLP. Our survey is broader and not restricted to a specific field.
Han et al. (2023)	Computer Vision	Categorized vision Transformer models based on tasks.	Focused on computer vision tasks. Our survey covers five fields of applications.
Xu et al. (2022)	Computer Vision	Covers computer vision and multimodal fields.	Primarily focused on advancements in computer vision. Our survey is more varied.
Li et al. (2023)	Computer Vision	Comparative analysis of Transformer models for medical vision.	Detailed on Transformer models for medical images. Our survey is more comprehensive.
Shamshad et al. (2023)	Medical Computer Vision	Review of Transformer models for medical images.	Focused on medical images. Our survey includes various modalities and is more comprehensive.
Khan et al. (2022)	Computer Vision	Overview of Transformer computer vision models.	Focused on computer vision tasks. Our survey covers five fields of applications.

Continued on next page

Table 2 – Continued from previous page

Approach	Fields of Application	Similarities	Differences
Ruan & Jin (2022)	Multi-modal (NLP-CV)	Categorization of Transformer vision-language models based on tasks.	Focused on multi-modal (NLP-CV) tasks. Our survey is more varied.
Shin et al. (2022)	Multi-modal (Performance /Architecture)	Survey on Transformers for multi-modal tasks.	Detailed on multimodal visual-linguistic tasks. Our survey is broader.

Table 2: Comparative Summary: Our Survey vs. Existing Surveys

5 Transformer Applications

Since 2017, the Transformer model has emerged as a highly attractive research model in the field of deep learning. Originally developed for processing long-range textual sentences. However, its scope has expanded to a variety of applications beyond NLP tasks. In fact, after a series of successes in NLP, researchers turned their attention to CV, exploring the potential of Transformer models' global attention capability, while CNNs were adept at tracking local features. The Transformer model has also been tested and applied in various other fields and for various tasks. To gain a deeper understanding of Transformer applications, we conducted a comprehensive search of various research libraries and reviewed the Transformer models available from 2017 to the present day. Our search yielded approximately more than 650 Transformer-based models that are being applied in various fields.
We identified the major fields in which Transformer models are being used, including NLP, CV, Multi-modal applications, Audio and Speech processing, and Signal Processing. Our analysis provides an overview of the Transformer models available in each field, their applications, and their impact on their respective industries.

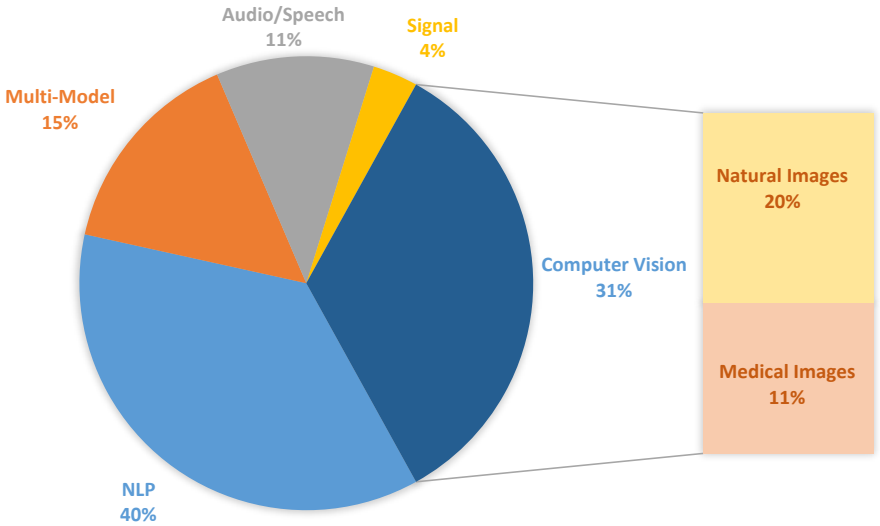

Figure 5: Proportion of Transformer Applications in Top 5 Fields

Figure 5 shows the percentage breakdown of the Transformer models proposed so far across different application fields. Our analysis revealed approximately 250 Transformer-based models for NLP, representing around 40% of the total Transformer models collected. Moreover, we accounted for approximately 200 models for CV. Due to the different processing of natural and medical images, and the extensive growth of both fields, we segmented CV into two categories: (i) Natural Image Processing and (ii) Medical Image Processing. As per this categorization, Natural Image Processing accounted for 20% of Transformer-based models, medical image processing accounted for 11%, and combinedly they accounted for 31% of Transformer-based models. Additionally, our analysis identified approximately 90 Transformer models for multi-modal applications, representing 15% of the total, and around 70 models for audio and speech processing, representing 11% of total Transformer models. Finally, only 4% of Transformer models were recorded for signal processing.

Our analysis provides a clear understanding of the proportion of attention received by Transformer applications in each field, facilitating the identification of further research areas and tasks where Transformer models can be implemented.

6 Application-Based Classification Taxonomy Of Transformers

As a result of conducting a thorough comprehensive analysis of all selected articles following the methodology explained in Section 3, we noticed that the existing categorizations did not fully capture the wide range of Transformer-based models and their diverse applications across different fields. Hence, in this study, we aimed to propose a more comprehensive taxonomy of Transformers that would reflect their practical applications. To achieve this, we carefully reviewed a large number of Transformer models and classified them based on their tasks within their respective fields of application. Our analysis identified several highly impactful and significant Transformer-based models that have been successfully applied in a variety of fields. We then organized these models into five different application areas: NLP, CV, Multi-modality, Audio and Speech, and Signal Processing. The proposed taxonomy in Figure 6 provides a more nuanced and comprehensive framework for understanding the diverse applications of Transformers. We believe that this taxonomy would be beneficial for researchers and practitioners working on Transformer-based models, as it would help them to identify the most relevant models and techniques for their specific applications.

6.1 Natural Language Processing (NLP)

Transformers have become a vital tool in NLP, and various NLP tasks have largely benefited from these models. Our proposed taxonomy focuses on NLP and organizes Transformer models into seven popular NLP tasks, including Translation, Summarization, Classification and Segmentation, Question Answering, Text Generation, Natural Language Reasoning, and Automated Symbolic Reasoning. To ensure a comprehensive analysis, we only considered Transformer models that have significantly impacted the NLP field and improved its performance. Our analysis included an in-depth discussion of each NLP task, along with essential information about each model presented in Table 3. We also highlighted the significance and working methods of each model. This taxonomy provides a valuable framework for understanding the different Transformer models used in NLP and their practical applications. It can help researchers and practitioners select the most appropriate Transformer model for their specific NLP task.

6.1.1 Language Translation

Language translation is a fundamental task in NLP, aimed at converting input text from one language to another. Its primary objective is to produce an output text that accurately reflects the meaning of the source text in the desired language. For example, given an English sentence as input text, the task aims to produce its equivalent in French or any other desired language. The original Transformer model was developed explicitly for the purpose of language translation, highlighting the significance of this task in the NLP field. Table 3 identifies the Transformer-based models that have demonstrated significant performance in the Language Translation task. These models play a vital role in facilitating effective communication and collaboration across different languages, enabling more efficient information exchange and knowledge sharing. Overall, the language translation task represents a crucial area of research in NLP, with significant implications for diverse applications, including business, science, education, and social interactions. The Transformer-based models presented in the table offer promising solutions for advancing the state-of-the-art in this field, paving the way for new and innovative approaches to language translation (Chowdhary & Chowdhary, 2020, Monroe, 2017, Hirschberg & Manning, 2015).

Transformer Models	**Architecture (Encoder/ Decoder)**	**Lingual Capabilities**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**	**BLEU Score**
Transformer (Vaswani et al., 2017)	Encoder & Decoder	Monolingual	NA	WMT 2014 English-German,WMT 2014 English-French Wall Street Journal (WSJ) BerkleyParser corpora	36.8
XLM (Conneau & Lample, 2019)	Encoder & Decoder	Monolingual & Multilingual	NA	XNLI, WMT'16 Romanian-English, WMT'14 English-French, WMT'16 English-German, WMT'16 English-Romanian, Wikipedias of the XNLI languages, MultiUN, IIT Bombay corpus, EUbookshop corpus, OpenSubtitles 2018, GlobalVoices	38.5

Continued on next page

Table 3 – Continued from previous page

Transformer Models	Architecture (Encoder/ Decoder)	Language	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)	BLEU Score
BART (Lewis et al., 2020)	Encoder & Decoder	Monolingual	160Gb of news, books, stories,and web text(Liu et al., 2019)	WMT16 Romanian-English	37.96

Table 3: Transformer Models for NLP: Language Translation Task

- **Transformer:** The first Transformer model, introduced by Vaswani et al. (2017) marked a significant milestone in the field of NLP, revolutionizing the way we approach language translation. This pioneering model, known as the Vanilla Transformer, was purpose-built for this task. The Transformer model comprises two key components: an encoder and a decoder module, each utilizing MHA and masked MHA mechanisms. The encoder module plays a vital role in analyzing contextual information within the input language, while the decoder module generates the output in the target language, leveraging the encoder's output and masked MHA. The Transformer model owes much of its success to its remarkable ability to execute parallel computations, enabling it to process words concurrently while retaining positional information. This parallel processing capability not only enhances its efficiency in handling extensive text data but also empowers it to effectively manage long-range dependencies—an indispensable trait for accurate language translation.
- **XLM:** It is a cross-lingual language pretraining (XLM) model developed to support multiple languages. The model is built using two methods: a supervised method and an unsupervised method. The unsupervised method utilizes Masked Language Modeling (MLM) and Casual Language Modeling (CLM) techniques and has shown remarkable effectiveness in translation tasks. On the other hand, the supervised method has further improved the translation tasks (Conneau & Lample, 2019). This combination of supervised and unsupervised learning has made the XLM model a powerful tool for cross-lingual applications, making it possible to perform NLP tasks in multiple languages. The effectiveness of the XLM model in translation tasks has made it a popular choice among researchers in the field of NLP.
- **BART:** Bidirectional and Auto-Regressive Transformers is an advanced pre-trained model primarily aimed at cleaning up the corrupt text. It features two pre-training stages: the first stage corrupts the text with noise, while the second stage focuses on recovering the original text from the noisy version. BART employs a Transformer translation model that integrates both the encoder and decoder modules, allowing it to perform various tasks such as text generation, translation, and comprehension with impressive accuracy (Lewis et al., 2020). Its bi-directional approach enables it to learn from the past and future tokens, while its auto-regressive properties make it suitable for generating output tokens sequentially. These features make BART an incredibly versatile model for various NLP tasks.

6.1.2 Classification & Segmentation

Text classification and segmentation are fundamental tasks in NLP that enable the automatic organization and analysis of large volumes of textual data. Text classification involves assigning tags or labels to text based on its contents, such as sentiment, topic, or intent, among others. This process helps to categorize textual documents from different sources and can be useful in a variety of applications, such as recommendation systems, information retrieval, and content filtering. On the other hand, text segmentation involves dividing the text into meaningful units, such as sentences, words, or topics, to facilitate further analysis or processing. This task is crucial for various NLP applications, including language understanding, summarization, and question answering, among others (Chowdhary & Chowdhary, 2020, Kuhn, 2014, Hu et al., 2016).
Transformer-based models have been shown to achieve state-of-the-art performance in text classification and segmentation tasks. These models are characterized by their ability to capture long-range dependencies and contextual information in text, making them well-suited for complex NLP tasks. Table 4 highlights some of the most prominent Transformer-based models that have demonstrated significant performance in text classification and segmentation tasks.

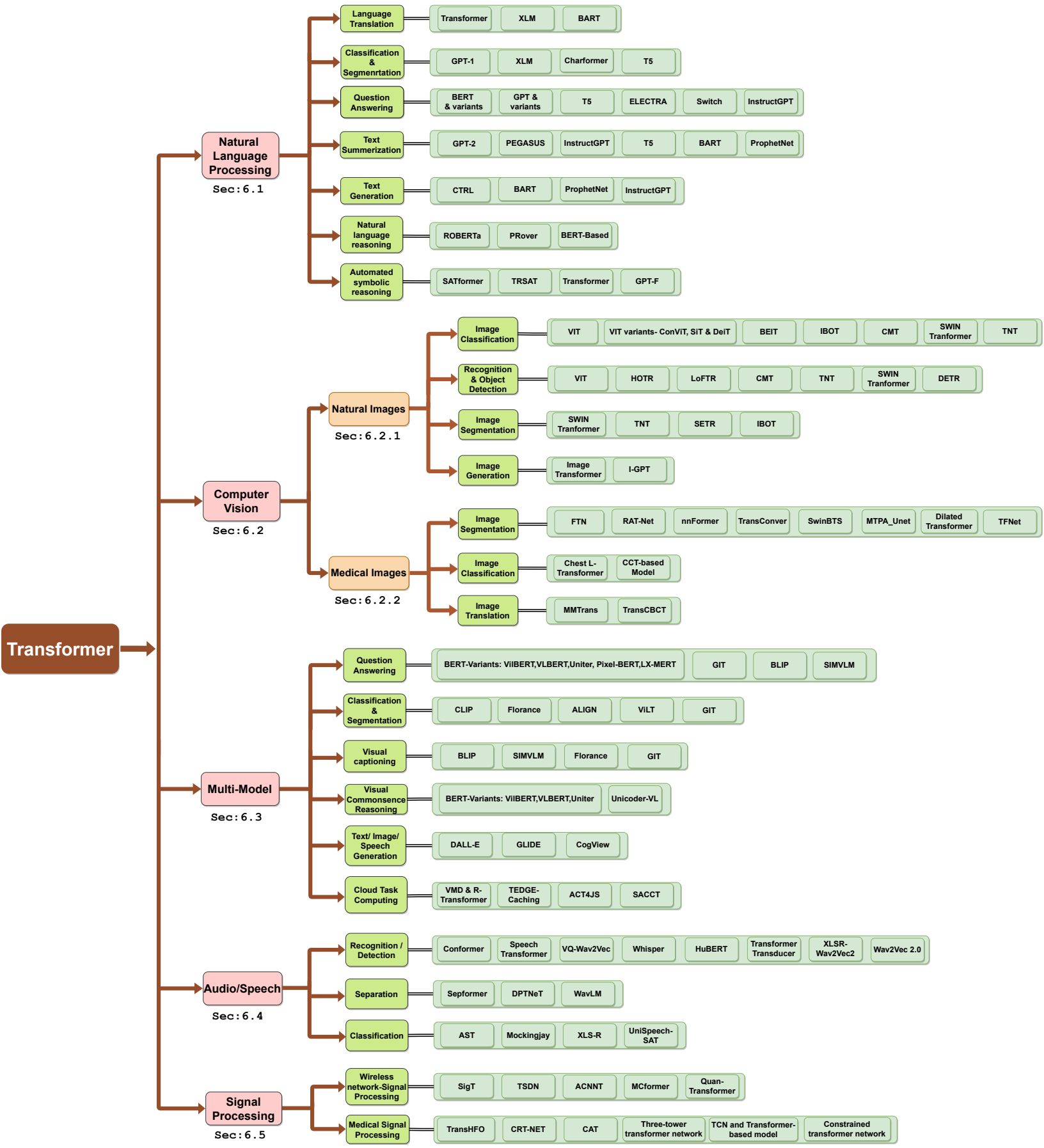

Figure 6: Application-based Taxonomy of Transformer Models

Transformer Models	Architecture (Encoder/ Decoder)	Lingual Capabilities	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)	GLUE Accuracy
GPT-1 (Radford et al., 2018)	Decoder	Monolingual	Book corpus 1B Word Benchmark	Stanford Sentiment Treebank-2 (SST-2), The Corpus of Linguistic Acceptability (CoLA)	91.3
XLM (Conneau & Lample, 2019)	Encoder & Decoder	Monolingual & Multilingual	NA	XNLI	NA
T5 (Raffel et al., 2020)	Encoder & Decoder	Monolingual	Colossal Clean Crawled Corpus (C4)	GLUE and SuperGLUE (including SST-2)	92.7
Charformer (Tay et al., 2022)	Encoder & Decoder	Monolingual & Multilingual	C4 (Colossal Clean Crawled Corpus)	Stanford Sentiment Treebank-2 (SST-2),AGNews	91.6

Table 4: Transformer Models for NLP: Language Classification& Segmentation Tasks

- **Charformer:** It is a Transformer-based model that introduces Gradient-Based Subword Tokenization (GBST), a lightweight approach to learning latent subwords directly from characters at the byte level. Unlike traditional word-level models, Charformer operates at the character level, treating each character in a text as a discrete unit. During training, Charformer learns to predict the next character in a sequence based on the context of preceding characters. The model has both English and multi-lingual variants and has demonstrated outstanding performance on language understanding tasks, such as the classification of long text documents (Tay et al., 2022).
- **GPT-1:** GPT stands for Generative Pre-Trained Transformer. GPT-1 undergoes a two-stage training process. Initially, it acquires knowledge from an extensive dataset through unsupervised learning, with a primary emphasis on language comprehension. Subsequently, it undergoes fine-tuning using smaller datasets tailored for specific tasks. This methodology highlights the potential of a single model, requiring only minimal adjustments, to perform exceptionally well across a range of tasks. GPT-1 leverages the advantages of unsupervised learning, offering the prospect of expanded language capabilities without the necessity for extensive labeled data.

  XLM & T5: These two models are versatile and capable of performing a range of NLP tasks. For instance, in addition to classification and segmentation tasks, both XLM and Text-To-Text Transfer Transformer (T5) models are proficient in language translation and text summarization. You can find a detailed description of the XLM model in Section 6.1.1, while information about the T5 model is available in Section 6.1.4.

6.1.3 Question Answering

Question Answering is a classical NLP task. It involves matching a text query to the most relevant answer in the form of text, based on the relevance of the text to the query. This task is challenging, as finding a concise and accurate answer to a given query can be difficult (Chowdhary & Chowdhary, 2020, Hirschman & Gaizauskas, 2001). Recent research has focused on this task, leading to the development of several Transformer-based models that leverage deep learning techniques to improve the accuracy and efficiency of this task. A detailed overview of these models is provided in Table 5.

Transformer Models	Architecture (Encoder/ Decoder)	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)	GLUE Accuracy
BERT (Devlin et al., 2019)	Encoder	BookCorpus, English Wikipedia	SQuAD v1.1, SQuAD v2.0, TriviaQA, GLUE(QNLI)	93.5
ELECTRA (Clark et al., 2020a)	Encoder	Wikipedia, BookCorpus, ClueWeb, CommonCrawl, Gigaword	SQuAD 1.1, SQuAD 2.0, GLUE	95

Continued on next page

Table 5 – Continued from previous page

Transformer Models	Architecture (Encoder/ Decoder)	Pre-training Dataset	Dataset(Fine-tuning, Training, Testing)	GLUE Accuracy
GPT-1(Radford et al., 2018)	Decoder	BookCorpus,1B Word Benchmark	GLUE(QNLI), RACE, Story Cloze, CoLA	88.1
GPT-2 (Radford et al., 2019)	Decoder	WebText,WikiText(Did not share)	WebText,WikiText(Did not share)	NA
GPT-3 (Brown et al., 2020)	Decoder	Pretraining Similar to GPT-2	Natural Questions, Web Questions, TriviaQA, ARC Reasoning Challenge, CoQA, DROP dataset & CoLA	NA
Switch Transformer (Fedus et al., 2022)	Encoder & Decoder	Colossal Clean Crawled Corpus (C4)	Linear Probe: SQuAD v1.0, ARC Reasoning Challenge, Glue, SuperGlue. Fine-tuning: Natural Questions, Web Questions, TriviaQA	NA
T5 (Raffel et al., 2020)	Encoder & Decoder	Colossal Clean Crawled Corpus (C4)	GLUE and SuperGLUE benchmarks, SQuAD v1.0	94.8
InstructGPT (Ouyang et al., 2022)	Decoder	Based on the pre-training model GPT-3	SQuAD v2.0	NA

Table 5: Transformer Models for NLP: Question Answering Task

- **BERT & BERT Variants:** BERT was introduced by the Google AI team and is embedded within the encoder module of the Transformer. BERT employs a bidirectional approach, allowing it to pre-train a Transformer on unannotated text by considering the context of each word. As a result, BERT has achieved remarkable performance on various NLP tasks (Devlin et al., 2019). A variety of BERT-based models have been developed with different characteristics. For instance, some are optimized for fast computation, while others produce superior results with a small number of parameters. Some are also tailored to specific tasks, such as **RoBERTa**, which is designed for masked language modeling and next sentence prediction (Liu et al., 2019). **FlueBERT** is another model that can be used for tasks such as text classification, paraphrasing, natural language inference, parsing, and word sense disambiguation (Le et al., 2020). Additionally, **DistilBERT** is suitable for question answering and other specific tasks. These models have significantly improved pre-trained Transformer models (Sanh et al., 2019).
- **GPT-3:** GPT models are exclusively based on the decoder block of Transformers, significantly advancing the field of NLP. In 2020, OpenAI unveiled one of the largest pre-trained GPT models, known as GPT-3, which boasts an impressive 175 billion parameters. This model is ten times larger than its predecessors, marking a substantial leap in language modeling capabilities. One of the most remarkable features of GPT-3 is its ability to perform exceptionally well across a diverse range of tasks without the need for gradient updates or fine-tuning, a requirement for models like BERT. GPT-3 operates through a two-step process: pre-training and fine-tuning. During the pre-training phase, GPT-3 is exposed to an extensive corpus of text data, learning to predict the next word in a sentence. When generating text, GPT-3 begins with an initial prompt or context and proceeds to predict subsequent words. It leverages self-attention mechanisms to capture contextual information and dependencies in the input. As argued by Brown et al. (2020), this versatility, coupled with its large-scale pre-training, empowers GPT-3 to excel in a wide spectrum of language understanding and generation tasks.
- **Switch Transformer:** The use of pre-trained models such as BERT and GPT, trained on large datasets, has gained popularity in the field of NLP. However, there are concerns about the economic and environmental costs of training such models. To address these concerns, the Switch Transformer was introduced, which offers a larger model size without a significant increase in computational cost. The Switch Transformer replaces the Feed-Forward Neural Network (FFN) with a switch layer that contains multiple FFNs, resulting in a model with trillions of parameters. Despite the increase in model size, the computational cost of the Switch Transformer remains comparable to that of other models. In fact, the Switch Transformer has been evaluated on 11 different tasks and has shown significant improvement in tasks such as translation, question-answering, classification, and summarization (Fedus et al., 2022).
- **ELECTRA:** An acronym for "Efficiently Learning an Encoder that Classifies Token Replacements Accurately", utilizes a distinct pre-training method compared to other pre-trained models. Electra deploys a "Masked Language Modeling"

approach that masks certain words and trains the model to predict them. Additionally, Electra incorporates a "Discriminator" network that aids in comprehending language without the need to memorize the training data. This unique approach enables Electra to generate superior text and surpass the performance of BERT (Clark et al., 2020a).

- In addition to question-answering tasks, InstructGPT has the capability to generate text, while the **T5** model is particularly noteworthy for text summarization tasks. Detailed descriptions of these models can be found in Section 6.1.5 and Section 6.1.4, respectively. Furthermore, the versatile **GPT-1 & GPT-2** models excel in multiple tasks, including classification, segmentation, and text summarization, in addition to question answering. Detailed descriptions of both GPT-1 & 2 models can be found in Sections 6.1.2 and 6.1.4, respectively.

6.1.4 TEXT SUMMARIZATION

Text summarization is a NLP task that involves breaking down lengthy texts into shorter versions while retaining essential and valuable information and preserving the meaning of the text. Text summarization is particularly useful in comprehending lengthy textual documents, and it also helps to reduce computational resources and time (Chowdhary & Chowdhary, 2020, Tas & Kiyani, 2007). Transformer-based models have shown exceptional performance in text summarization tasks. The Transformer-based models in text summarization are listed in Table 6.

Transformer Models	**Architecture (Encoder/ Decoder)**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**	**ROUGE-L Score**
GPT-2 (Radford et al., 2019)	Decoder	BookCorpus	SNLI, MNLI, QNLI, SciTail, RTE, RACE, CNN, SQuaD, MRPC, QQP, STS-B, SST2 & CoLA	NA
PEGASUS (Zhang et al., 2020a)	Encoder & Decoder	Colossal Clean Crawled Corpus (C4), HugeNews	CNN/DailyMail summarization, XSum, NEWSROOM, Multi-News, Gigaword, arXiv, PubMed, BIGPATENT, WikiHow, Reddit TIFU, AESLC, BillSum	40.76
T5 (Raffel et al., 2020)	Encoder & Decoder	Colossal Clean Crawled Corpus (C4)	CNN/DailyMail summarization	40.69
InstructGPT (Ouyang et al., 2022)	Decoder	Based on the pre-training model GPT-3	SFT dataset, RM dataset, PPO dataset, a dataset of prompts and completions Winogender, CrowS-Pairs, Real Toxicity Prompts, TruthfulQA, DROP, QuAC, SquadV2, Hellaswag, SST, RTE, and WSC, WMT 15 Fr ! En, CNN/Daily Mail Summarization, Reddit TLDR Summarization datasets	NA
BART (Lewis et al., 2020)	Encoder & Decoder	160Gb of news, books, stories,and web text (Liu et al., 2019)	CNN/DailyMail summarization, XSum	40.9
ProphetNet (Qi et al., 2020)	Encoder & Decoder	BookCorpus, English Wikipedia, News, Books, Stories, and Web text	CNN/DailyMail summarization, Gigaword	40.72

Table 6: Transformer Models for NLP - Text Summarization Task

- **GPT-2:** a language model built on the Transformer architecture. Its operation involves pre-training on an extensive corpus of text data, where it learns to predict the next word in a sentence. GPT-2 excels in generating coherent and contextually relevant text. During the generation process, given an initial prompt, the model continues to predict subsequent words, expanding the text in a meaningful and context-aware manner. To achieve this, it employs self-attention mechanisms that capture long-range dependencies within the input text. GPT-2's versatility makes it adept at generating text for various natural language understanding and generation tasks. These tasks include text completion, text generation, translation, and question answering, rendering it a valuable tool for a wide spectrum of NLP applications (Radford et al., 2019).

- **PEGASUS:** It is an exemplary model for generative text summarization that employs both the encoder and decoder modules of the Transformer. While models based on masked language modeling only mask a small portion of text, PEGASUS masks entire multiple sentences, selecting the masked sentences based on their significance and importance, and generating them as the output. In other words, Pegasus utilizes its learned knowledge to capture the main ideas and content of the input text, effectively producing abstractions in its own words. The model has exhibited significant performance on unknown summarization datasets (Zhang et al., 2020a).
- **T5:** It introduced a dataset named "Colossal Clean Crawled Corpus (C4)" that improved the performance in various downstream NLP tasks. It employs the Transformer architecture and is pre-trained on a vast amount of text data with a denoising autoencoder objective. During training, the input text is turned into a "prefix," specifying the task, and the target text is the "suffix," indicating the expected output. T5 is a multi-task model that can be trained to perform a range of NLP tasks using the same set of parameters. Following pre-training, the model can be fine-tuned for different tasks and achieves comparable performance to several task-specific models (Raffel et al., 2020).
- **InstructGPT and ProphetNet:** These models will be discussed in Section 6.1.5. Additionally, apart from text summarization, both of these models are capable of performing text generation tasks. Furthermore, **BART** specializes in language translation tasks, and its description is available in Section 6.1.1 above.

6.1.5 Text Generation

The task of text generation has gained immense popularity in the field of NLP due to its usefulness in generating long-form documentation, among other applications. Text generation models attempt to derive meaning from trained text data and create a connection between the text that has been previously outputted. These models typically operate on the basis of this connection (Chowdhary & Chowdhary, 2020, Reiter & Dale, 1997). The use of Transformer-based models has led to significant advancements in the task of text generation. Please refer to Table 7.

Transformer Models	Architecture (Encoder/ Decoder)	Pre-trained (Yes/No)	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
CTRL (Keskar et al., 2019)	Encoder & Decoder	Yes	Project Gutenberg, subreddits, News Data, Amazon Review, open WebText, WMT Translation date, question-answer pairs, MRQA	Multilingual Wikipedia and Open WebText.
BART (Lewis et al., 2020)	Encoder & Decoder	Yes	Corrupting documents, 1M steps on a combination of books and Wikipedia data, news, stories, and web text (Training)	SQuAD, MNLI, ELI5, XSum, ConvAI2, CNN/DM, CNN/DailyMail, WMT16 Romanian-English, augmented with back-translation data from Sennrich et al. (2016).
ProphetNET (Qi et al., 2020)	Encoder & Decoder	Yes	Bookcorpus, English Wikipedia news, books, stories, and web text	CNN/dailymail, Giga-word Corpus, SQuAD dataset.
InstructGPT (Ouyang et al., 2022)	Decoder	Yes	Based on the pre-training model GPT-3	SFT dataset, RM dataset, PPO dataset, dataset of prompts and completions Winogender, CrowS-Pairs, Real Toxicity Prompts, TruthfulQA, DROP, QuAC, SquadV2, Hellaswag, SST, RTE and WSC, WMT 15 Fr ! En, CNN/Daily Mail Summarization, Reddit TLDR Summarization datasets.

Table 7: Transformer Models for NLP: Text Generation Task

- **CTRL:** The acronym CTRL denotes the Conditional Transformer Language model, which excels in generating realistic text resembling human language, contingent on a given condition. It enables controlled text generation by conditioning on control codes or prompts, offering flexibility and customization in generating text content. In addition, CTRL can produce text in multiple languages. This model is large-scale, boasting 1.63 billion parameters, and can be fine-tuned for various generative tasks, such as question answering and text summarization (Keskar et al., 2019).
- **ProphetNET:** It is a sequence-to-sequence model that utilizes future n-gram prediction to facilitate text generation by predicting n-grams ahead. The model adheres to the Transformer architecture, comprising encoder and decoder modules.

It distinguishes itself by employing an n-stream self-attention mechanism and employs pre-training with masked language modeling that introduces span masking for efficient token prediction. Users can provide prompts to guide text generation, making it versatile for various natural language generation tasks. ProphetNET demonstrates remarkable performance in summarization and is also competent in question generation tasks (Qi et al., 2020).

- **InstructGPT:** It was proposed as a solution to the problem of language generative models failing to produce realistic and truthful results. It achieves this by incorporating human feedback during fine-tuning and reinforcement learning from the feedback. The GPT-3 model was fine-tuned for this purpose. As a result, InstructGPT can generate more realistic and natural output that is useful in real-life applications. ChatGPT, which follows a similar methodology as InstructGPT, has gained significant attention in the field of NLP at the end of 2022 (Ouyang et al., 2022).
- **BART** model's description is provided above in Section 6.1.1. This model is also capable of performing language translation tasks.

6.1.6 Natural Language Reasoning

The pursuit of natural language reasoning is a field of study that is distinct from that of question-answering. Question-answering focuses on finding the answer to a specific query within a given text passage. On the other hand, natural language reasoning involves the application of deductive reasoning to derive a conclusion from the given premises and rules that are represented in natural language. Neural network architectures aim to learn how to utilize these premises and rules to infer new conclusions. Previously, a similar task was traditionally tackled by systems equipped with the knowledge represented in a formal format and rules to be applied for the derivation of new knowledge. However, the use of formal representation has posed a significant challenge to this line of research (Mark A Musen, 1988). With the advent of Transformers and their remarkable performance in numerous NLP tasks, it is now possible to circumvent formal representation and allow Transformers to engage in reasoning directly using natural language. Table 8 highlights some of the significant Transformer models for natural language reasoning tasks.

Transformer Models	**Task Accomplished**	**Architecture (Encoder/ Decoder)**	**Pre-trained (Yes/No)**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**
RoBERTa (Clark et al., 2020b)	Binary Classification	Encoder	Yes	RACE	Fine-tuning: RACE Test: DU0-DU5, Birds, Electricity, ParaRules
PRover (Saha et al., 2020)	Binary Classification, Sequence Generation	[RoBERTa-based](Encoder)	Yes	The model is based on the pre-trained RoBERTa-large	Test: DU0-DU5, Birds, Electricity, ParaRules
BERT-Based Model (Picco et al., 2021)	Binary Classification	Encoder	Yes	RACE	Dataset generated by (Clark et al., 2020b)

Table 8: Transformer Models for Natural Language Reasoning

- **RoBERTa:** In a study conducted by (Clark et al., 2020b), a binary classification task was assigned to the Transformer, which aimed to determine whether a given statement can be inferred from a provided set of premises and rules represented in natural language. The architecture utilized for the Transformer was RoBERTa-large, which was pre-trained on a dataset of high school exam questions that required reasoning skills. This pre-training enabled the Transformer to achieve a high accuracy of 98% on the test dataset. The dataset contained theories that were randomly sampled and constructed using sets of names and attributes. The task required the Transformer to classify whether the given statement (Statement) followed from the provided premises and rules (Context) (Clark et al., 2020b)
- **PRover:** In a related study, (Saha et al., 2020) proposed a model called PRover, which is an interpretable joint Transformer capable of generating a corresponding proof with an accuracy of 87%. The task addressed by PRover is the same as that in the study by (Clark et al., 2020b) and (Richardson & Sabharwal, 2022), where the aim is to determine whether a given conclusion follows from the provided premises and rules. The proof generated by PRover is represented as a directed graph, where the nodes represent statements and rules, and the edges indicate which new statements follow from applying rules on the previous statements. Overall, the proposed approach by (Saha et al., 2020) provides a promising direction towards achieving interpretable and accurate reasoning models.

- **BERT-based:** In (Picco et al., 2021), a BERT-based architecture called "neural unifier" was proposed to improve the generalization performance of the model on the RuleTaker dataset. The authors aimed to mimic some elements of the backward-chaining reasoning procedure to enhance the model's ability to handle queries that require multiple steps to answer, even when trained on shallow queries only. The neural unifier consists of two standard BERT Transformers, namely the fact-checking unit and the unification unit. The fact-checking unit is trained to classify whether a query of depth 0, represented by the embedding vector q-0, follows from a given knowledge base represented by the embedding vector C. The unification unit takes as input the embedding vector q-n of a depth-n query and the embedding vector of the knowledge base, vector C, and tries to predict an embedding vector q0, thereby performing backward-chaining.

6.1.7 Automated Symbolic Reasoning

Automated symbolic reasoning is a subfield of computer science that deals with solving logical problems such as SAT solving and automated theorem proving. These problems are traditionally addressed using search techniques with heuristics. However, recent studies have explored the use of learning-based techniques to improve the efficiency and effectiveness of these methods. One approach is to learn the selection of efficient heuristics used by traditional algorithms. Alternatively, an end-to-end learning-based solution can be employed for the problem. Both approaches have shown promising results and offer the potential for further advancements in automated symbolic reasoning (Kurin et al., 2020, Selsam et al., 2019). In this regard, a number of Transformer based models have shown significant performance in automated symbolic reasoning tasks. For those models, please look at Table 9.

Transformer Models	Task Accomplished	Architecture	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
SATFormer (Shi et al., 2022b)	Binary Classification	Combination of Graph Neural Network and a hierarchical Transformer (Encoder & Decoder)	NA	Training: Generated dataset (100K satisfiable instances and 100K unsatisfiable instances) Evaluation: Generated dataset (50 satisfiable instances and 50 unsatisfiable instances)
TRSAT (Shi et al., 2021)	Binary Classification	Combination of Graph Neural Network and a hierarchical Transformer (Encoder & Decoder)	NA	Training: SATLIB benchmark library, 4000 instances from each of the distributions from (Yolcu & Póczos, 2019) Evaluation: Circuit datasets
Transformer (Hahn et al., 2021)	Sequence Generation	Transformer-based model(Encoder & Decoder)	NA	A dataset generated through the application of formulas based on 55 LTL : Training: LTLRandom35, LTLPattern126, PropRandom35 Testing: LTLRandom35, LTLRandom50, LTLPattern126, LTLUnsolved254, PropRandom35, PropRandom50
GPT-f (Polu & Sutskever, 2020)	Sequence Generation	Transformers similar to GPT-2 and GPT-3(Decoder)	CommonCrawl, Mix of Github, arXiv and Math StackExchange	Metamath's set.mm

Table 9: Transformer Models for Automated Symbolic Reasoning

- **SATformer:** the SAT-solving problem for boolean formulas was addressed by Shi et al. in 2022 (Shi et al., 2021) through the introduction of SATformer, a hierarchical Transformer architecture that offers an end-to-end learning-based solution for solving the problem. Traditionally, in the context of SAT-solving, a boolean formula is transformed into its Conjunctive Normal Form (CNF), which serves as an input for the SAT solver. The CNF formula is a conjunction of boolean variables and their negations, known as literals, organized into clauses where each clause is a disjunction of these literals. For example, a CNF formula utilizing boolean variables would be represented as (A OR B) AND (NOT A OR C), where each clause (A OR B) and (NOT A OR C) is made up of literals. The authors employ a Graph Neural Network (GNN) to obtain the embeddings of the clauses in the CNF formula. SATformer then operates on these clause embeddings to capture the interdependencies among clauses, with the self-attention weight being trained to be high when groups of clauses that could

potentially lead to an unsatisfiable formula are attended together, and low otherwise. Through this approach, SATformer efficiently learns the correlations between clauses, resulting in improved SAT prediction capabilities (Shi et al., 2022b).

- **TRSAT:** another research endeavor conducted by Shi et al. (2021) investigated a variant of the boolean SAT problem known as MaxSAT and introduced a Transformer model named TRSAT, which serves as an end-to-end learning-based SAT solver. A comparable problem to the boolean SAT is the satisfiability of a linear temporal formula (Pnueli, 1977), where a satisfying symbolic trace to the formula is sought after given a linear temporal formula.
- **Transformer:** in a study conducted by Hahn et al. (2021), the authors addressed the boolean SAT problem and the temporal satisfiability problem, both of which are more complex than binary classification tasks that were tackled in previous studies. In these problems, the task is to generate a satisfying sequence assignment for a given formula, rather than simply classifying whether the formula is satisfied or not. The authors constructed their datasets by using classical solvers to generate linear temporal formulas with their corresponding satisfying symbolic traces, and boolean formulas with their corresponding satisfying partial assignments. The authors employed a standard Transformer architecture to solve the sequence-to-sequence task. The Transformer was able to generate satisfying traces, some of which were not observed during training, demonstrating its capability to solve the problem and not merely mimic the behavior of the classical solvers used in the dataset generation process.
- **GPT-f:** in their work, (Polu & Sutskever, 2020) presented GPT-F, an automated prover and proof assistant that utilizes a decoder-only Transformers architecture similar to GPT-2 and GPT-3. GPT-F was trained on a dataset called set.mm, which contains approximately 38,000 proofs. The largest model investigated by the authors consists of 36 layers and 774 million trainable parameters. This deep learning network has generated novel proofs that have been accepted and incorporated into mathematical proof libraries and communities.

Discussion:

NLP stands as the most successful and widely applied domain for Transformer-based models. Given this prominence, a diverse array of models has emerged to cater to various NLP tasks. However, comparing these models presents a significant challenge. Each model exhibits distinct characteristics, and its performance is often assessed using task-specific metrics. Following an exhaustive analysis, we have endeavored to establish a framework for model comparisons. In this context, we have taken into account various parameters tailored to specific tasks to facilitate meaningful comparisons.

For the language translation task, we conducted a comparative analysis of three models, while the XLM models (Conneau & Lample, 2019) did not exhibit performance based on the BLEU-BiLingual Evaluation Understudy accuracy metric. Therefore, we evaluated the performance of these models using their respective BLEU scores, as depicted in Figure 7. Notably, the XLM model (Conneau & Lample, 2019) achieved a BLEU score of 38.5, the Vanilla Transformer version (Vaswani et al., 2017) attained a score of 36.8, and BART (Lewis et al., 2020) demonstrated commendable language translation performance with a BLEU score of 37.96. It is worth noting that the performance of these models is closely aligned, with XLM outperforming the others in the context of the language translation task.

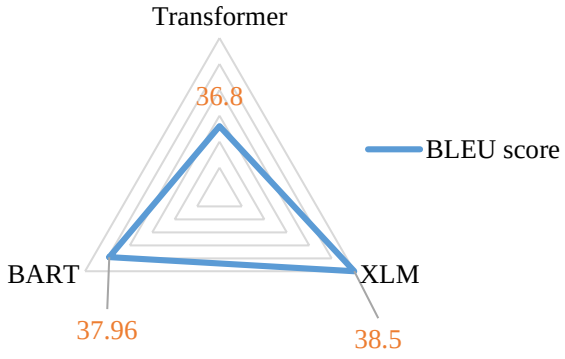

Figure 7: Comparison of Transformer Models for Language Translation Based on the BLEU Score

In the context of language classification and segmentation tasks, the evaluation of model performance can be based on their respective General Language Understanding Evaluation-GLUE scores. Notably, the three models under consideration exhibit closely aligned performance, ranging between 91 and 93 according to the GLUE Score metric. However, it is worth highlighting that T5 (Raffel et al., 2020) achieved the highest performance, with a GLUE score of 92.7. Additionally, GPT-1 (Radford et al., 2018) and Charformer (Tay et al., 2022) demonstrated substantial performance with GLUE scores of 91.3 and 91.6, respectively.

Furthermore, for the question-answering tasks, we conducted an analysis of multiple models to identify factors for comparison. In this regard, we computed the Glue accuracy scores for four of the models as depicted in Figure 8. However, it's worth

noting that we could not obtain Glue accuracy scores for GPT-1 (Radford et al., 2018), GPT-3 (Brown et al., 2020), Switch Transformer (Fedus et al., 2022), and InstructGPT (Ouyang et al., 2022) in the context of question-answering tasks. Among the models assessed, Electra (Clark et al., 2020a) demonstrated the highest performance, achieving a Glue accuracy score of 95. Conversely, GPT-2 (Radford et al., 2019) exhibited the lowest Glue score at 88.1, in comparison to the other models for question-answering tasks. Additionally, BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) showcased substantial performance, recording Glue scores of 93.5 and 94.8, respectively.

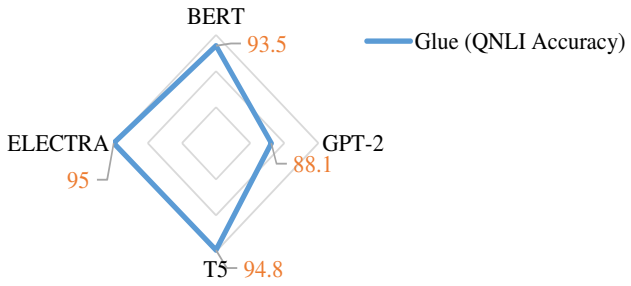

Figure 8: Comparison of Transformer Models for NLP Question Answering Task Based on the ROUGE-L Score

In the case of text summarization, we conducted comparisons using the Rouge-L score as a metric. It is worth noting that GPT-2 (Radford et al., 2019) and InstructGPT (Ouyang et al., 2022) did not provide Rouge-L scores for the summarization task. Conversely, the Rouge-L scores for the other four models were quite similar. Specifically, Pegasus (Zhang et al., 2020a) and ProphetNet (Qi et al., 2020) achieved Rouge-L scores of 40.76 and 40.72, respectively, for text summarization. Similarly, T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) exhibited commendable performance in summarization, both recording Rouge-L scores of 40.69 and 40.90, respectively. A visual representation of these model comparisons is presented in Figure 9.

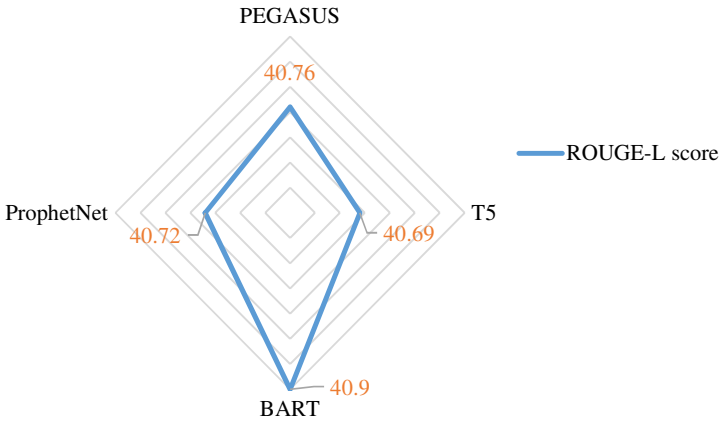

Figure 9: Comparison od Transformer Models for text Summarization based on the ROUGE-L Score.

In the case of language reasoning tasks, upon comprehensive examination of the models, we encountered a lack of a universally applicable metric for direct model comparison. Nevertheless, there are noteworthy factors that merit consideration. It's important to highlight that all models designed for natural language reasoning tasks are pre-trained and employ the supervised learning paradigm. In the context of automated symbolic reasoning, the choice of architectural components assumes particular significance. For instance, the Transformer model (Hahn et al., 2021) exclusively employs the Transformer neural network for this purpose. Conversely, SATformar (Shi et al., 2022b) and TRSAT (Shi et al., 2021) incorporate both Graph Neural Networks and the Transformer architecture to tackle automated symbolic reasoning tasks. Moreover, GPT-F (Polu & Sutskever, 2020) adopts a configuration similar to other GPT models, utilizing only the decoder module of the Transformer architecture.

The choice of an NLP model for a particular task depends on various factors, including the languages involved in translation, the availability of data, computational resources, and the trade-off between translation quality and efficiency. Each model possesses unique strengths, and the optimal selection hinges on the specific requirements of your project. When a significant data and computational resources are at your disposal, the Transformer model (Vaswani et al., 2017) tends to excel in language translation tasks. However, if you are constrained by computational resources and need to work with smaller models, the Switch Transformer (Fedus et al., 2022) can be a viable solution, given its ability to efficiently manage sparsity and token selection. Additionally, XLM (Conneau & Lample, 2019) exhibits commendable performance even with limited data. Partic-

ularly in scenarios where translation occurs across multiple languages, especially when data availability for specific language pairs is limited, XLM's cross-lingual capabilities prove advantageous. For those prioritizing natural and fluent translations, BART (Lewis et al., 2020) is a viable option, although it tends to be slower in generating translations during inference compared to non-auto-regressive models. Lastly, if your NLP tasks encompass a wide range and you have access to a substantial amount of data and computational resources, T5 (Raffel et al., 2020) typically delivers superior performance.

When dealing with NLP tasks involving multiple languages, both XLM (Conneau & Lample, 2019) and GPT models (Brown et al., 2020, Radford et al., 2018; 2019) demonstrate strong performance. However, XLM stands out due to its exceptional cross-lingual capabilities. For tasks requiring the classification and segmentation of languages with complex characters, Charformer's (Tay et al., 2022) character-level modeling provides a significant advantage, particularly when resources are limited. If achieving robust performance in question-answering tasks is a priority and sufficient computational resources are available, BERT (Devlin et al., 2019) is a compelling choice, leveraging its bi-directional capabilities and extensive pre-training. In contrast, InstructGPT (Ouyang et al., 2022) excels when fine-tuned for specific domain-specific QA tasks using dedicated datasets. For text summarization tasks, Pegasus (Zhang et al., 2020a) is purpose-built and has demonstrated outstanding performance in summarizing lengthy documents. On the other hand, ProphetNET (Qi et al., 2020) introduces an innovative causal masking scheme, enabling efficient and coherent text generation. Lastly, CTRL (Keskar et al., 2019) is tailored for conditional language-specific text generation, making it particularly well-suited for generating text with specific attributes or styles.

When there is a need for a versatile and widely-applied model suitable for a broad spectrum of natural language reasoning tasks, particularly in cases of limited computational resources, RoBERTa (Liu et al., 2019) stands out, offering a compelling combination of strong performance and efficiency. For tasks specifically involving probabilistic reasoning over text, uncertainty modeling, or scenarios where capturing probabilities and uncertainties is paramount, PRover (Saha et al., 2020) is tailor-made for such specialized tasks. In the context of general-purpose natural language reasoning tasks spanning a wide range of applications, BERT-based models (Picco et al., 2021) serve as a robust foundational choice, offering versatility and high performance. Highly specialized for SAT-solving tasks, SATFormer (Shi et al., 2022b) and TRSAT (Shi et al., 2021) excel in scenarios where the primary objective is to solve Boolean satisfiability problems (SAT) or similar symbolic reasoning tasks. In contrast, when automated symbolic reasoning tasks require a more generic approach, Transformer (Hahn et al., 2021) can be considered, particularly when tasks exhibit diversity and extend beyond SAT solving. Lastly, GPT-F (Polu & Sutskever, 2020) emerges as a commendable model for various formal reasoning tasks, especially when tasks encompass a broader scope than SAT solving and leverage the generative capabilities inherent in the GPT architecture.

6.2 Computer Vision (CV)

Motivated by the success of Transformers in NLP, researchers have explored the application of the Transformer concept in CV tasks. Traditionally, CNNs have been considered the fundamental component for processing visual data. However, different types of images require different processing techniques, with natural images and medical images being two prime examples. Furthermore, research in CV for natural images and medical images is vast and distinct.

Natural images, intended for human perception, differ from medical images designed for diagnostics. Natural images are diverse, often complex, and larger in scale, while medical images prioritize consistency and focus on specific structures. Noise and artifacts vary in nature, requiring specialized preprocessing in medical images. Annotations are simpler in natural images but demand expertise in medical imaging. Data availability favors natural images, while ethical concerns are more pronounced in medical imaging due to patient privacy and regulations. As a result, Transformer models for CV can be broadly classified into two categories: (i) those designed for natural image processing, and (ii) those designed for medical image processing.

6.2.1 Natural Image processing

In the domain of CV, natural image processing is a primary focus as compared to medical image processing, owing to the greater availability of natural image data. Furthermore, CV with natural images has wide-ranging applications in various domains. Among the numerous tasks associated with CV and natural images, we have identified four of the most common and popular tasks: (i) classification and segmentation, (ii) recognition and feature extraction, (iii) mask modeling prediction, and (iv) image generation. In this context, we have provided a comprehensive discussion of each of these CV tasks with natural images. Additionally, we have presented a table that provides crucial information about each Transformer-based model and have highlighted their working methods and significance.

A. Image Classification

Image classification is a crucial and popular task in the field of CV, which aims to analyze and categorize images based on their features, type, genre, or objects. This task is considered as a primary stage for many other image processing tasks. For example, if we have a set of images of different animals, we can classify them into different animal categories such as cat, dog, horse, etc., based on their characteristics and features (Szummer & Picard, 1998, Lu & Weng, 2007). Due to its significance, many Transformer-based models have been developed to address image classification tasks. Table 10 highlights some of the significant Transformer models for image classification tasks and discusses their important features and working methodologies.

Transformer Models	Architecture	Accuracy	Supervision	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
VIT (Dosovitskiy et al., 2021)	The same architecture as the original Transformer (Encoder)	77.9	Self-supervised	JFT-300M, ILSVRC-2012 ImageNet, ImageNet-21k	CIFAR-10/100, Oxford Flowers-102, Oxford-IIIT Pets, VTAB
Convit (d'Ascoli et al., 2021)	ViT architecture with convolutional conductive bias (Encoder)	82,4	Supervised	ILSVRC-2012 ImageNet (Based on DeiT)	ILSVRC-2012 ImageNet,CIFAR100
SiT (Ahmed et al., 2021)	ViT with self-supervision Using ViT-S (Encoder)	NA	Supervised (The model is pretrained on unsupervised dataset)	STL-10, CUB200, CIFAR10, CIFAR100, ILSVRC-2012 ImageNet, Pascal VOC, MS-COCO, Visual-Genome	CIFAR-10,CIFAR-100 , STL-10, CUB200, ILSVRC-2012 ImageNet, Pascal VOC, MS-COCO, Visual-Genome
DeiT (Touvron et al., 2021)	The same architecture as ViT, but with no convolutions(Encoder)	81,8	Supervised	ILSVRC-2012 ImageNet	ILSVRC-2012 ImageNet, iNaturalist 2018, iNaturalist 2019, Flowers-102, Stanford Cars, CIFAR-100, CIFAR-10
BEIT (Bao et al., 2022)	Based on ViT architecture (Encoder)	83.2	Self-supervised	ILSVRC-2012 ImageNet, ImageNet-22k	ILSVRC-2012 ILSVRC-2012 ImageNet, ADE20K, CIFAR-100
IBOT (Zhou et al., 2021b)	Architecture is based on ViT and Swin Transformer (Encoder)	84	Self-supervised	ILSVRC-2012 ImageNet, ViT-L/16, ImageNet-22K	MS-COCO, ADE20K
CMT (Guo et al., 2022a)	Hybrid architecture between Transformer and CNN	83.1	Supervised	ILSVRC-2012 ImageNet	ILSVRC-2012 ImageNet,CIFAR10, CIFAR100, Standford Cars, Flowers, Oxford-IIIT Pets
SWIN (Liu et al., 2021)	Architecture based on transfoemr with new attention layer SW- MSA (Shifted Window MSA)	83.5	Supervised	ImageNet22K	ILSVRC-2012 ImageNet
TNT (Han et al., 2021)	The architecture combines 2 Transformer. The architecture is inspired from ViT and DEiT	82.9	Supervised	ILSVRC-2012 ImageNet	Oxford 102 Flowers, Oxford-IIIT Pets, iNaturalist 2019, CIFAR-10, CIFAR-100

Table 10: Transformer Models for Natural Image Processing - Image Classification

- **ConViT:** Convolutional Vision Transformer (ConViT) represents a hybrid model designed for vision-related tasks. It commences with an image input and employs a convolutional backbone to extract crucial visual features. Subsequently, the image is segmented into patches, with each patch undergoing transformation into embeddings, and it incorporates positional information. These embeddings are then subject to processing through a Transformer encoder, facilitating the capture of global context and establishing relationships between patches. Task-specific heads are responsible for generating predictions

tailored to various vision tasks. By combining the strengths of CNNs and Transformers, ConViT emerges as a versatile and efficient solution applicable to a broad spectrum of vision-related applications (d'Ascoli et al., 2021).

- **SiT:** the Self-supervised Vision Transformer (SiT) architecture offers the flexibility to function as an autoencoder while seamlessly accommodating multiple self-supervised tasks. SiT serves as an extension of the Vision Transformer (ViT) architecture and leverages self-supervised learning through pretext tasks to instill an understanding of visual context and relationships within the model. SiT's primary function involves extracting valuable visual features from images and employing a contrastive loss mechanism to facilitate the acquisition of semantically meaningful representations. These learned representations can subsequently be fine-tuned for various downstream CV tasks. This adaptability renders SiT a versatile model suitable for both unsupervised and supervised visual learning endeavors (Ahmed et al., 2021).
- **DeiT:** the Data-efficient image Transformer (DeiT) represents a data-efficient image classification model founded on the ViT architecture. This designation implies that DeiT necessitates a smaller volume of training data to achieve proficiency. To enhance its performance, DeiT employs knowledge distillation from a larger, pre-trained teacher model to a more compact student model. Additionally, the utilization of data augmentation and a two-step training process further bolsters the model's capacity to effectively classify images. DeiT finds particular suitability in scenarios characterized by a scarcity of labeled data, rendering it a valuable asset for efficient image classification tasks (Touvron et al., 2021).
- **BEIT:** Bidirectional Encoder Representation from Image Transformers (BEIT) is a Transformer-based model that draws inspiration from BERT and introduces a new pre-training task called Masked Image Modeling (MIM) for vision Transformers. In MIM, a portion of the image is randomly masked, and the corrupted image is passed through the architecture, which then recovers the original image tokens. BEIT has shown competitive performance on image classification and segmentation tasks, demonstrating its effectiveness for a variety of CV applications (Bao et al., 2022).
- **CMT:** CNNs Meet Transformer (CMT) is a model that combines both CNNs and ViT. CNNs are better suited to capturing local features, while Transformers excel at capturing global context. CMT takes advantage of the strengths of both these models and performs well in image classification tasks as well as object detection and recognition tasks. The integration of CNN and Transformer allows CMT to handle both spatial and sequential data effectively, making it a powerful tool for CV tasks (Guo et al., 2022b).
- **IBOT:** It represents Image BERT Pre-training with Online Tokennizer (IBOT) which is a self-supervised model. This model studied masked image modeling using an online tokenizer and it learns to distill features using a tokenizer. This online tokenizer helps this model to improve the feature representation capability. Besides, the image classification task, this model shows significant performance in object detection and segmentation tasks.
- **Swin Transformer**, **TNT** models will be described in the Section 6.2.1 (C). Both of these models are capable of performing the task of Image Segmentation besides image Recognition & Object Detection and object detection tasks. Furthermore, **ViT** model which will be discussed in detail in Section 6.2.1 (B), is also capable of performing recognition and object detection tasks.

B. Image Recognition & Object Detection

Image recognition & Object detection is often considered as nearly similar and related task in CV. It is the capability of detecting or recognizing any object, person, or feature in an image or video. An image or video contains a number of objects & features; by extracting the features from the image, a model tries to capture the features of an object through training. By understanding these useful features, a model can recognize the specific object from the other available object in the image or video (Zhao et al., 2019, Jiao et al., 2019, Hénaff, 2020, Chen et al., 2019a). Here we highlight and discuss the significant Transformer models for image/object recognition tasks (see Table 11).

Transformer Models	Architecture	Data Augmentation	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
VIT (Dosovitskiy et al., 2021)	The same architecture than the original Transformer (Encoder)	No	JFT-300M, ILSVRC-2012 ImageNet, ImageNet-21k	CIFAR-10/100, Oxford Flowers-102, Oxford-IIIT Pets, VTAB
LoFTR (Sun et al., 2021a)	Encoder & Decoder	No	NA	the indoor model of LoFTR is trained on the ScanNet dataset and the outdoor model on the MegaDepth dataset

Continued on next page

Table 11 – continued from previous page

Transformer Models	Architecture	Data Augmentation	Pre-training Dataset	Dataset(Fine-tuning, Training, Testing)
CMT (Guo et al., 2022b)	Hybrid architecture between Transformer and CNN (Encoder)	No	NA	MS-COCO
TNT (Han et al., 2021)	The architecture combines 2 Transformers. The architecture is inspired from ViT and DEiT (Encoder & Decoder)	Yes	ImageNet ILSVRC 2012	MS-COCO
SWIN (Liu et al., 2021)	Architecture based on transfoemr with the attention layer SW-MSA (Encoder)	No	ImageNet-22k	MS-COCO
DETR (Carion et al., 2020)	Architecture is composed of ImageNet pretrained backbone ResNet-50. DETR combines CNN with a Transformer architecture (Encoder & Decoder)	Yes	ImageNet pretrained backbone ResNet-50	MS-COCO
HOTR (Kim et al., 2021a)	Architecture inspired from DETR (Encoder & Decoder)	Yes	MS-COCO	V-COCO HICO-DET

Table 11: Transformer Models for Natural Image Processing - Image Recognition & Object Detection

- **ViT:** the Vision Transformer (ViT) is one of the earliest Transformer-based models that has been applied to CV. ViT views an image as a sequence of patches and processes it using only the encoder module of the Transformer. ViT performs very well for classification tasks and can also be applied to image recognition tasks. It demonstrates that a Transformer-based model can serve as an alternative to CNNs (Dosovitskiy et al., 2021).
- **LoFTR:** it stands for Local Feature Matching with Transformer, is a CV model that is capable of learning feature representations directly from raw images, as opposed to relying on hand-crafted feature detectors for feature matching. This model employs both the encoder and decoder modules of the Transformer. The encoder takes features from the image, while the decoder works to create a feature map. By leveraging the Transformer's ability to capture global context and long-range dependencies, LoFTR can achieve high performance in visual recognition tasks.
- **DETR:** the Detection Transformer (DETR) represents a new approach to object detection or recognition, which performs the object detection task as a direct set of prediction problems (Carion et al., 2020). In contrast, other models accomplish this task in two stages. DETR uses an encoder to generate object queries, a self-attention mechanism to capture the relationship between the queries and objects in the image, and creates an object detection scheme. This model has been shown to be effective for object detection and recognition tasks and represents a significant advancement in the field.
- **HOTR:** the Human-Object Interaction Transformer model (HOTR), is a Transformer-based model designed for predicting Human-Object Interaction. It is the first Transformer-based Human-Object Interaction (HOI) detection prediction model that employs both the encoder and decoder modules of the Transformer. Unlike conventional hand-crafted post-processing schemes, HOTR uses a prediction set to extract the semantic relationship of the image, making it one of the fastest human-object interaction detection models available (Kim et al., 2021a).
- **TNT** & **SWIN Transformer** models will be described in Section 6.2.1 (C). Both of these models are capable of performing the task of Image Segmentation besides image Recognition & Object Detection. Moreover, **CMT** model is discussed in the image classification section 6.2.1 (A).

C. Image Segmentation

Segmentation is the process of partitioning an image based on objects and creating boundaries between them, requiring pixel-level information extraction. There are two popular types of image segmentation tasks in CV: (i) Semantic Segmentation, which aims to identify and color similar objects belonging to the same class among all other objects in an image, and (ii) Instance Segmentation, which aims to detect instances of objects and their boundaries (Minaee et al., 2022, Haralick & Shapiro, 1985). In this section, we will discuss some Transformer-based models that have shown exceptional performance in image segmentation tasks (refer to Table 12 for more details).

Transformer Models	Architecture	Supervision	Data Augmentation	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)	mIoU
SWIN (Liu et al., 2021)	Encoder	Supervised	Yes	ImageNet-22k	ADE20K	44,1
SETR (Zheng et al., 2021)	Encoder & Decoder	Supervised	NA	LSVRC-2012 ImageNet, pre-trained weights provided by ViT or DeiT	ADE20K	50,28
IBOT (Zhou et al., 2021b)	Encoder	self-supervised	NA	LSVRC-2012 ImageNet, ViT-L/16 ImageNet-22K	ADE20K	45,4
TNT (Han et al., 2021)	Encoder & Decoder	Supervision	Yes	LSVRC-2012 ImageNet	ADE20K	43,6

Table 12: Transformer Models for Natural Image Processing - Image Segmentation

- **SWIN Transformer:** it is short for Scaled WINdowed Transformer (SWIN) (Liu et al., 2021), is a Transformer-based model that is capable of handling large images by dividing them into small patches, or windows, and processing them through its architecture. By using shifted windows, the model requires a smaller number of parameters and less computational power, making it useful for real-life image applications. SWIN Transformer can perform image classification, segmentation, and object detection tasks with exceptional accuracy and efficiency (Zidan et al., 2023, Yang & Yang, 2023).
- **SETR:** a SEgmentation TRansformer (SETR), is a Transformer-based model used for image segmentation tasks. It uses sequence-to-sequence prediction methods and removes the dependency of fully convolutional network with vanilla Transformer architecture. Before feeding the image into the Transformer architecture, it divides the image into a sequence of patches and the flattened pixel of each patch. There are three variants of SETR models available with different model sizes and performance levels (Zheng et al., 2021).
- **TNT:** Transformer in Transformers (TNT) is a Transformer-based CV model that uses a Transformer model inside another Transformer model to capture features inside local patches of an image (Han et al., 2021). The image is divided into local patches, which are further divided into smaller patches to capture more detailed information through attention mechanisms. TNT shows promising results in visual recognition tasks and offers an alternative to CNNs for CV tasks.
- **IBOT** model described in Section 6.2.1 (A), as it is also capable of performing the Image Classification task besides the segmentation task.

D. Image Generation

Image generation is a challenging task in CV, and Transformer-based models have shown promising results in this area due to their parallel computational capability. This task involves generating new images using existing image pixels as input. It can be used for object reconstruction and data augmentation (van den Oord et al., 2016, Liu et al., 2017). we focus on image generation models that use image pixels without any other type of data. In Table 13, we discuss some Transformer-based models that have demonstrated exceptional performance in image generation tasks.

Transformer Models	Architecture	Auto-regressive Model	Data Augmentation	Conditional Generation	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
Image Transformer (Parmar et al., 2018)	Encoder & Decoder	Yes	No	Unconditional Generation usign ImageNet. Conditional generation using CIFAR-10	N/A	LSVRC-2012 ImageNet,CIFAR-10

Continued on next page

Table 13 – continued from previous page

Transformer Models	Architecture	Auto-regressive Model	Data Augmentation	Conditional Generation	Pre-training Dataset	Dataset(Fine-tuning, Training, Testing)
I-GPT (Chen et al., 2020b)	Decoder	Yes	Yes	Unconditional generation	ImageNet ILSVRC 2012	Fine-tuning/linear probe: CIFAR-10, CIFAR-100, STL-10 Linear probe: ImageNet ILSVRC 2012

Table 13: Transformer Models for Natural Image Processing - Image Generation

- **Image Transformer:** Image Transformer is an autoregressive sequence generative model that uses the self-attention mechanism for image generation. This model generates new pixels and increases the size of the image by utilizing the attention mechanism on local pixels. It uses both the encoder and decoder module of the Transformer, but does not use masking in the encoder. The encoder layer is used less than the decoder for better performance on image generation. Image Transformer is a remarkable model in the field of image generation (Parmar et al., 2018).
- **I-GPT:** Image GPT is an image generative model that utilizes the GPT-2 model for training to auto-regressively predict pixels by learning image representation, without using the 2D image. BERT motifs can also be used during pre-training. I-GPT has four variants based on the number of parameters: IGPT-S (76M parameters), IGPT-M (455M parameters), IGPT-L (1.4B parameters), and IGPT-XL (6.8M parameters), where models with higher parameters have more validation losses (Chen et al., 2020b).

Discussion:
In the domain of CV tasks, comparing models can be challenging due to their distinct characteristics and varying performance metrics tailored to specific tasks. Nevertheless, after conducting an extensive analysis, we have endeavored to establish a framework for model comparisons, taking into account different parameters tailored to specific tasks. For image classification tasks, we evaluate model performance based on their respective accuracy scores, as illustrated in Figure 10. Notably, Conformer (Peng et al., 2021) and IBOT (Zhou et al., 2021b) emerge as top performers in classification, achieving accuracy scores of 84.1% and 84%, respectively. These models demonstrate a robust capacity to accurately classify images. Following closely are BEIT (Bao et al., 2022) and Swin (Liu et al., 2021), both showcasing strong performance with accuracy scores of 83.2% and 83.5%, respectively, excelling in capturing intricate image features. CMT (Guo et al., 2022b) maintains a competitive stance at 83.1%, underscoring its proficiency in pattern and object recognition. DeiT (Touvron et al., 2021), with an accuracy score of 81.8%, excels in data-efficient image classification. ViT (Dosovitskiy et al., 2021) and TNT (Han et al., 2021), while performing well with scores of 77.9% and 82.9%, respectively, exhibit slightly lower performance levels. The accuracy scores for these models are visually presented in the figure below.

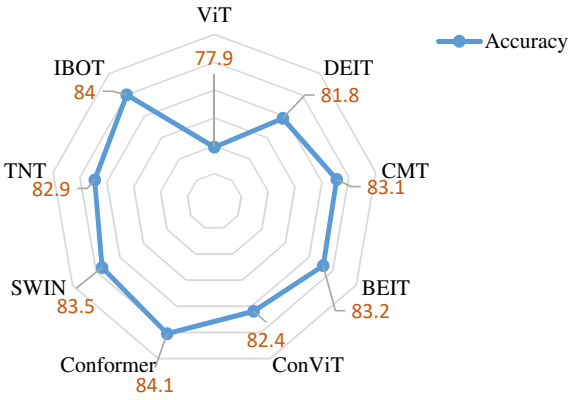

Figure 10: Comparison of Transformer Models Developed for Image Classification Based on Accuracy

When it comes to models for image recognition and object detection tasks, our in-depth analysis revealed a lack of common points for direct comparison. These models vary significantly, and their performance is measured differently. However, some noteworthy aspects stand out. For example, models like HOTR (Kim et al., 2021a), DETR (Carion et al., 2020), and TNT

(Han et al., 2021) employ scale augmentation methods, even if not explicitly mentioned in the papers. Additionally, many of these models heavily rely on various versions of the COCO dataset for fine-tuning in image recognition and object detection tasks.

In the realm of image segmentation tasks, model performance is typically evaluated using the mIoU (mean Intersection over Union) score, which measures pixel-level segmentation accuracy. We observed that these models are typically pretrained with the ImageNet dataset and then fine-tuned on the ADE20K dataset. Among the models listed, SETR (Zheng et al., 2021) leads with a robust mIoU score of 50.28%, demonstrating its excellence in precise object segmentation. IBOT (Zhou et al., 2021b) also performs competitively with an mIoU of 45.4%, showcasing its effectiveness in capturing object boundaries. Swin (Liu et al., 2021) closely follows with 44.1%, while TNT (Han et al., 2021) trails slightly with a mIoU of 43.6%. While all these models exhibit strong segmentation capabilities, SETR (Zheng et al., 2021) stands out as the top performer in terms of mIoU score, making it the preferred choice for tasks demanding highly accurate and detailed image segmentation.

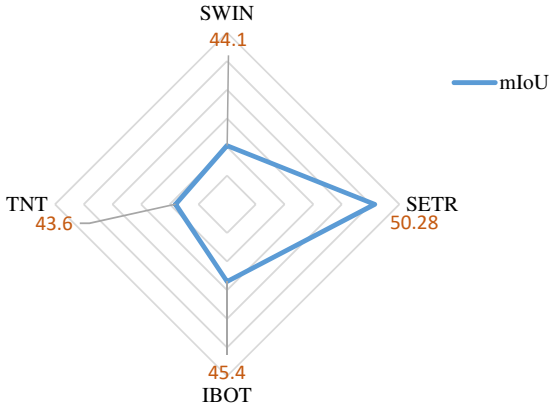

Figure 11: Comparison of Transformer Models for Image Segmentation Based on mIoU Metric

Comparing two image generation models can be a challenging task due to their distinctiveness and the diversity of metrics used to measure their accuracy. It is often difficult to find common points of comparison among them. Nevertheless, we can identify some key characteristics of these models. For instance, the Image Transformer model, as presented in (Parmar et al., 2018), is capable of both unconditional generation using the ImageNet dataset and conditional generation using CIFAR-10. In contrast, I-GPT, as described in (Chen et al., 2020b), focuses on unconditional generation, predicting pixels without prior knowledge of the underlying 2D image structure. Further details regarding the generation conditions and supervision can be found in the provided table.

The effectiveness of CV models in various tasks hinges on several factors, including resource availability, task objectives, data accessibility, and more. These factors contribute to variations in model performance. In this context, we provide recommendations tailored to different scenarios and considerations.

In scenarios with limited data, DeiT (Touvron et al., 2021) excels by specializing in data efficiency, making it a valuable choice for image classification tasks with restricted training data. It performs admirably even in data-scarce environments. Conversely, when abundant data and computational resources are available, ViT (Dosovitskiy et al., 2021) stands out as an effective option for various CV tasks through versatile fine-tuning methods. BEIT (Bao et al., 2022) introduces Transformers with bottlenecks, enhancing efficiency. It's a wise selection when balancing model size and performance is essential. For tasks extending beyond image classification, Conformer (Peng et al., 2021) is recommended. With a combination of convolutional and Transformer layers, it excels in handling tasks requiring both spatial and sequential data processing. SiT (Ahmed et al., 2021) shines in scenarios where precise spatial object arrangement is vital, demanding focused attention on specific image regions. IBOT (Zhou et al., 2021b) proves valuable when classifying objects within videos or tackling complex image segmentation scenarios. LoFTR (Sun et al., 2021a) specializes in local feature matching and is designed for matching and correspondence tasks rather than image recognition. For video-based recognition tasks, CMT (Guo et al., 2022b) is well-suited for object tracking within object detection pipelines. TNT (Han et al., 2021) is a valuable choice for instance segmentation or object detection tasks, where individual object detection and segmentation are essential. It utilizes tokenization effectively for object detection. Swin Transformer (Liu et al., 2021) serves as a versatile model, accommodating a wide range of CV tasks, including image recognition and segmentation. It excels at feature extraction, capturing both local and global features effectively. When tasks demand capturing long-range dependencies in object detection, HOTR (Kim et al., 2021a) utilizes a higher-order Transformer effectively. SETR (Zheng et al., 2021) is the preferred choice when understanding and capturing spatial context are vital for your segmentation task, particularly in cases where spatial relationships within images are crucial. Image Transformer (Parmar et al., 2018) is ideal for generating images based on prompts or descriptions, particularly useful for image synthesis tasks. I-GPT (Chen et al., 2020b) is versatile and suitable for creative artwork generation and other image synthesis tasks, enabling the generation of creative images

6.2.2 MEDICAL IMAGE PROCESSING

The diagnosis of pathologies based on medical images is often criticized as complicated, time-consuming, error-prone, and subjective (López-Linares et al., 2020). To overcome these challenges, alternative solutions such as deep learning approaches have been explored. Deep learning has made great progress in many other applications, such as NLP and CV. Although Transformers have been successfully applied in various domains, their application to medical images is still relatively new. Other deep learning approaches such as CNNs, RNNs, and Generative Adversarial Networks (GAN) are commonly used. This survey aims to provide a comprehensive overview of the various Transformer models developed for processing medical images.

A. Medical Image Segmentation

Image segmentation refers to the task of grouping parts of the image that belong to the same category. In general, encoder-decoder architectures are commonly used for image segmentation (López-Linares et al., 2020). In some cases, image segmentation is performed upstream of the classification task to improve the accuracy of the classification results (Wang et al., 2022c). The most frequently used loss functions in image segmentation are Pixel-wise cross-entropy loss and Dice Loss (López-Linares et al., 2020). Common applications of medical image segmentation include detecting lesions, identifying cancer as benign or malignant, and predicting disease risk. This paper presents a comprehensive overview of relevant models used in medical image segmentation. Table 14 provides a summary of these models.

Transformer Name	Field of Application	Fully Transformer Architecture	Architecture Type	Attention Type	Image Type	Dataset
FTN (He et al., 2022)	Skin lesion	Yes	Serial	Multi-head Spatial Pyramid Attention (MSPA)	2D	ISIC 2018 dataset
RAT-Net (Zhu et al., 2022)	Oncology (breast cancer)	No	Serial	Region Aware Component (RAC)	2D/3D ultrasound	The dataset consists of ABUS (Automatic Breast Ultrasound) images acquired at the Cancer Hospital of the University of Chinese Academy of Sciences, encompassing imagery from 256 subjects, with a total of 330 images for each subject
nnFormer (Zhou et al., 2021a)	Brain tumor multi-organ cardiac diagnosis	Yes	Serial	3D Local Volume-based Multi-head Self-attention (LV-MSA), Shifted 3D Local Volume-based Multi-head Self-attention (SLV-MSA), 3D global Volume-based Multi-head Self-attention (GV-MSA)	3D	Medical Segmentation Decathlon (MSD), Synapse multiorgan segmentation, Automatic Cardiac Diagnosis Challenge (ACDC)
Transconver (Liang et al., 2022)	Brain tumor	No	Parallel	Cross-attention fusion, SWIN attention meachisms	2D/3D	MICCAI BraTS2019, MICCAI BraTS2018
SwinBTS (Jiang et al., 2022b)	Brain tumor	No	Serial	SWIN attention mechanism, Enhanced Self-attention (ESA)	3D	BraTS 2019, BraTS 2020, BraTS 2021

Continued on next page

Table 14 – Continued from previous page

Transformer Name	Field of Application	Fully Transformer Architecture	Architecture Type	Attention Type	Image Type	Dataset
MTPA_Unet (Jiang et al., 2022a)	Retinal vessel	No	Serial	Transformer-Position Attention (TPA)	2D	DRIVE, CHASE DB1, and STARE Datasets
Dilated Transformer (Shen et al., 2022b)	Oncology (Breast Cancer)	No	Serial	Multi-head Attention (MHA), Residual axial attention layer	2D Ultra-sound	Dataset A: a public dataset including 562 breast US images(Xian et al., 2018). Dataset B: a private dataset including 878 breast US images
TFNet (Wang et al., 2021b)	Oncology (Breast lesion)	No	Serial	MultiHead Self-Attention, MultiHead Channel-Attention	2D Ultra-sound	BUSI Dataset, DDTI Dataset, Private collected dataset

Table 14: Transformer Models for Medical Image Segmentation

- **FTN:** a Transformer-based architecture developed specifically for segmenting and classifying 2D images of skin lesions. The model is composed of five layers, which consist of three encoders and two decoders. In each of these layers, there is a tokenization module referred to as "SWT" (Sliding Window Tokenization) and an "SPT Transformer" (Spatial Pyramid Transformer). The model is segregated into encoders and decoders for the segmentation task, while only an encoder is needed for classification tasks. To improve computational efficiency and storage optimization, MSPA "Multi-head Spatial Pyramid Attention" is utilized in the "Transformer" module instead of the traditional masked MHA. In comparison to CNN, FTN has demonstrated superior performance on 10,025 images extracted from the publicly available ISIC 2018 dataset. However, this model has some limitations. FTN yields poor segmentation outcomes when there is minimal contrast between the background and foreground. Furthermore, although the model was designed for classification, it doesn't perform notably well in this task compared to its performance in segmentation (He et al., 2022).

- **RAT-Net:** The primary objective of RAT-Net (Region Aware Transformer Network) is to replace the laborious and time-consuming manual task of detecting lesion contours in 3D ABUS (Automatic Breast Ultrasound) images. RAT-Net is an architecture that combines UNet and Transformer components, functioning as both an encoder and a decoder. It extends the capabilities of the SegFormer Transformer model, which encodes input images and identifies regions of particular importance for lesion segmentation. This model introduces RAC (Region Attention Constraints), which limits self-attention to suspicious and high-probability regions exclusively. This approach leads to significant computational savings compared to processing the entire image. When compared to other advanced models designed for medical image segmentation, RAT-Net has demonstrated outstanding performance in terms of accuracy (ACC), the 95th percentile of the asymmetric Hausdorff Distance (HD95), and the Dice Similarity Coefficient (DSC) metrics (Zhu et al., 2022).

- **nnFormer:** this model employs the Transformer architecture for the segmentation of 3D medical images. The experiments were conducted using 484 brain tumor images, 30 multi-organ scans, and 100 cardiac diagnosis images. Instead of relying on the traditional attention mechanism, nnFormer introduced two novel techniques: LV-MSA (Volume-based Multi-head Self-attention) and GV-MSA (Global Volume-based Multi-head Self-attention) to reduce computational complexity. Similar to most models in this domain, nnFormer follows a U-shape architecture. Its design combines convolutional layers and self-attention mechanisms and consists of an encoder, a bottleneck module, and a decoder. The encoder comprises 2 Transformer blocks, the decoder consists of 2 Transformer blocks, and the bottleneck module integrates 3 Transformer blocks. In order to connect the encoder and the decoder, nnFormer presents Skip attention as a substitute for the conventional Skip connections. Skip attention utilizes attention mechanisms to integrate additional information, as opposed to relying on simple summation and concatenation techniques. Furthermore, nnFormer utilizes multiple convolution layers with smaller kernels in the encoder, departing from the use of larger convolution kernels seen in other visual Transformers. (Zhou et al., 2021a).

- **Transconver:** is an encoder-decoder architecture that uniquely combines a CNN module and a SWIN Transformer in parallel, departing from the conventional serial approach employed by most models. It introduces a novel method called "skip Connection with Cross-Attention Fusion" (SCCAF), where the cross-attention mechanism is applied within the skip connection. This effectively merges global features extracted by the Transformer with local features extracted by convolution

modules, recognizing the distinct semantic characteristics of these features. This versatile network is designed to process both 2D and 3D brain tumor images, utilizing the 2D Swin Transformer and 3D Swin Transformer. Its training process involves a dataset comprising 335 cases from the MICCAI BraTS2019 dataset, followed by an evaluation phase on 66 cases from MICCAI BraTS2018 and 125 cases from MICCAI BraTS2019 (Liang et al., 2022).

- **SwinBTS:** is a model specifically designed for 3D medical image segmentation. It achieves this by combining the Swin Transformer with CNN. The use of SWIN is advantageous as it reduces the number of parameters, leading to faster learning and lower memory usage. SwinBTS adopts an encoder-decoder architecture that incorporates the Swin Transformer in both the encoder and decoder components. It introduces a bottleneck module between the encoder and decoder, known as the Enhanced Transformer (Enhanced Self-Attention ETrans). This module is implemented using convolution operations and HADAMARD products, with the primary objective of capturing deep-level features that may not be extracted by the encoder alone. While SwinBTS generally performs well, it does have limitations, particularly in accurately segmenting image edges (Jiang et al., 2022b).

- **MTPA_Unet:** the Multi-scale Transformer-Position Attention Unet is a model that has been thoroughly evaluated on multiple well-established retinal datasets to enhance the performance of the retinal image segmentation task. This model effectively combines CNN and Transformer architectures in a sequential manner, allowing it to accurately capture both local and global image information. To capture long-term dependencies between pixels and to glean contextual information about each pixel's location, this model makes use of TPA (Transformer Position Attention), which is a fusion of MSA (Multi-headed Self-Attention) and the "Position Attention Module". Furthermore, to optimize the model's feature extraction capabilities, it incorporates feature map inputs of varying resolutions, thereby leveraging the detailed information inherent in retinal images (Jiang et al., 2022a).

- **TFNet:** this model is designed to segment 2D ultrasound images of breast lesions by combining CNN with a Transformer architecture. To effectively address the challenge of lesions with varying scales and variable intensities, CNN serves as the backbone to extract features from the images, resulting in three high-level features containing semantic information and one low-level feature. These high-level features are fused through a Transformer Fuse Module (TFM), while the low-level features are integrated via skip connections. The Transformer module consists of two key components: Vanilla Multi-Head Self-Attention, which captures long-range dependencies between sequences, and Multi-Head Channel-Attention (MCA), designed to detect dependencies between channels. To further enhance the model's performance, novel loss functions have been introduced, resulting in superior segmentation performance when compared to other models. This approach has undergone evaluation on a variety of ultrasound image datasets, consistently delivering excellent segmentation results (Wang et al., 2021b).

- **Dilated Transformer:** the DT model has been specifically developed for the segmentation of 2D ultrasound images from small breast cancer datasets, utilizing the Transformer architecture. Traditional Transformer models typically require large pre-training datasets to achieve high-quality segmentation results. However, DT addresses this challenge by incorporating the "Residual Axial Attention" mechanism, tailored for segmenting images from smaller breast ultrasound datasets. This approach applies attention to individual axes, specifically the height and width axes, rather than the entire feature map, resulting in time-saving benefits and enhanced computational efficiency. Additionally, to prevent the omission of information along the diagonal direction, the architectural design integrates the dilated convolution module (Shen et al., 2022b).

B. Medical Image Classification

Image classification refers to the process of recognizing, extracting, and selecting different types of features from an image for classification using labels (Wang et al., 2020c). Features in an image can be categorized into three types: low-level features, mid-level features, and high-level features (Wang et al., 2020c). Deep learning networks are designed to extract high-level features. Common applications of medical image classification include the detection of lesions, the identification of cancers as benign or malignant, and the prediction of disease risk (Khan & Lee, 2023, Jungiewicz et al., 2023). Table 15 provides an overview of several relevant examples of Transformers used in medical image classification.

Transformer Name	Field of Application	Fully Transformer Architecture	Image Type	Supervision	Dataset
CCT-based Model (Islam et al., 2022)	Malaria Disease	No	2D images	Supervised	National Library of Medicine malaria dataset

Table 15 – continued from previous page

Transformer Name	Field of Application	Fully Transformer Architecture	Image Type	Supervision	Dataset
Chest L-Transformer (Gu et al., 2022)	Chest radiograph / Thoracic diseases	No	2D images	Weakly supervised	SIIM-ACR Pneumothorax Segmentation dataset

Table 15: Transformer Models for Medical Image Classification

- **CCT-based Model (Islam et al., 2022):** the model presented in this work is designed for classifying red blood cell (RBC) images as containing malaria parasites or not, by using Compact Convolutional Transformers (CCTs). The model input consists of image patches generated through convolutional operations and preprocessed by reshaping them to a fixed size. Unlike other vision Transformer models, this model performs classification using sequence pooling instead of class tokens. Compared to other deep learning models such as CNN, this model shows good performance in classifying RBC images. This satisfactory result was achieved by implementing a Transformer architecture, using GRAD-CAM techniques to validate the learning process, and fine-tuning hyperparameters.
- **Chest L-Transformer (Gu et al., 2022):** this model is designed for the classification of chest radiograph images. It employs a CNN backbone, specifically a Restricted ResNeXt, to extract local features from 2D images. Additionally, it incorporates a Transformer block to apply attention mechanisms, enhancing its ability to detect lesion locations within the images. By introducing Transformers into the analysis of chest radiograph images, this model can effectively focus on areas where diseases are more likely to be present, a departure from traditional CNNs that treat all areas equally. In terms of performance, the model exhibits strong sensitivity and achieves a commendable F1 score when compared to other models. However, it lags behind in terms of AUC (Area Under the Curve) and specificity metrics. Nevertheless, this model holds potential value for the development of datasets used in chest radiograph segmentation tasks.

C. Medical Image Translation

The field of research that involves altering the context (or domain) of an image without changing its original content is gaining traction. One example of this involves applying cartoon-style effects to images to change their appearance (Pang et al., 2022). Image-to-image translation is a promising technique that can be utilized to synthesize medical images from non-corrupted sources with less cost and time, and it is also helpful for preparing medical images for registration or segmentation. Some of the most popular deep learning models developed for this area include "Pix2Pix" and "cyclic-consistency generative adversarial network" (GAN) (Yan et al., 2022b). Table 16 provides an overview of some relevant examples of "Transformers" designed for medical image-to-image translation.

Transformer Name	Field of Application	Fully Transformer Architecture	Image Type	Supervision	Dataset
MMTrans (Yan et al., 2022b)	Magnetic resonance imaging (MRI)	No	2D MRI images	Supervised	BraTs2018, fastMRI, The clinical brain MRI dataset
TransCBCT (Chen et al., 2022c)	Oncology (prostate Cancer)	No	2D CBCT images	Supervised	91 patients with prostate cancer

Table 16: Transformer Models for Medical Image Translation

- **MMTrans (Yan et al., 2022b):** the Multi-Modal Medical Image Translation (MMtrans) model is proposed based on the GAN architecture and Swin Transformer structure for performing medical image-to-image translation on Magnetic Resonance Imaging (MRI). Unlike other image-to-image translation frameworks, MMtrans utilizes the Transformer to model long global dependencies to ensure accurate translation results. Moreover, MMtrans does not require images to be paired

and pixel-aligned since it employs SWIN as a registration module adapted for paired and unpaired images, which makes it different from other architectures like Pix2Pix. For the remaining components of its architecture, MMtrans utilizes SwinIR as the generator module and employs CNN as the discriminator module.

- **TransCBCT (Chen et al., 2022c):** A new architecture called TransCBT is proposed for the purpose of performing accurate radiotherapy by improving the quality of 2D images, specifically cone-beam computed tomography (CBCT), and generating synthetic 2D images (sCT) without damaging their structures. TransCBT integrates pure-Transformer modeling and convolution approaches to facilitate the extraction of global information and enhances performance by introducing the multi-head self-attention method (SW-MSA). Another model that can improve the quality of CT images reconstructed via sinograms is the CCTR (Shi et al., 2022a). In comparison to TransCBCT, CCTR experiments utilized a lung image database with 1010 patients, rather than the 91 patients used in TransCBCT.

Discussion:
Medical image processing plays a crucial and pivotal role in the advancement of the biomedical field. However, comparing models in this domain can be challenging due to differences in their purposes, objectives, available resources, and other factors. In this context, we conducted an in-depth analysis of these models to facilitate meaningful comparisons.
In the context of medical image segmentation tasks, one of the key aspects to consider when comparing models is the types of images they are designed to segment, which can be either 2D, 3D, or a combination of both 2D and 3D images. In this regard, nnformer (Zhou et al., 2021a) and swinBTS (Jiang et al., 2022b) specialize in segmenting 2D images, while FTN (He et al., 2022), Dilated Transformer (Shen et al., 2022b), TFNeT (Wang et al., 2021b), and MTPA-Unet (Wang et al., 2021b) are tailored for processing 3D images for segmentation. Conversely, RAT-NeT (Zhu et al., 2022) and Transconver (Liang et al., 2022) are noteworthy models in the realm of medical image processing, as they have the capability to segment both 2D and 3D images. Another notable aspect to consider is the attention mechanisms employed by these models. For example, FTN utilizes MSPA (Multi-head Special Pyramid Attention), RAT-NeT applies RAC (Region Aware Component), and nnFormer relies on several local volume-based MHA mechanisms. Furthermore, models like Transconver utilize cross-attention fusion, while MTPA Unet leverages ESA (Enhanced Self-Attention). It's worth mentioning that among all the medical image segmentation models, Transconver stands out for its parallel execution approach, whereas all other models opt for a serial method.

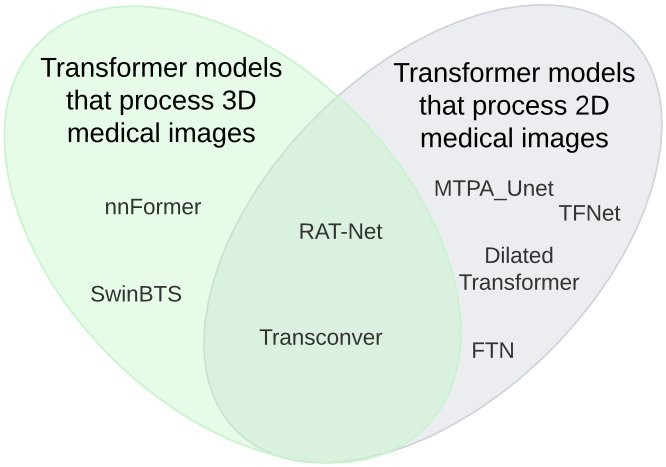

Figure 12: Comparing Transformer Models for the Segmentation of Medical Images Based on Input Image Types

Furthermore, we conducted an analysis of both models for the medical image classification task and observed several similarities and differences between them. One significant point of comparison pertains to the method of supervision and the input image size. Chest L-Transformer (Gu et al., 2022) operates under weak supervision with a 39X39 input image size, while the CCT-based model (Islam et al., 2022) is strongly supervised, employing a 96X96 image size. Additionally, the CCT-based model uses a kernel size of 3X3, whereas Chest L-Transformer utilizes a 1X1 kernel size. Aside from these factors, we did not find any explicit metrics scores to facilitate a direct performance comparison between these models. After conducting a thorough analysis of medical image translation models, we did not identify any significant factors that would facilitate a direct performance comparison between them. Nonetheless, there are noteworthy findings worth mentioning. There are architectural differences between the MMTrans and TransCBCT models. MMTrans (Yan et al., 2022b) follows a GAN-based architecture, incorporating three key modules (Generator, registration, and discriminator). In contrast, TransCBCT (Chen et al., 2022c) is constructed using four encoders and four decoders. Additionally, it's worth noting that MMTrans is trained using 2D MRI images, while TransCBCT utilizes a smaller dataset consisting of 2D CBCT images. Both of these models operate under supervised settings
The choice of a model in medical image processing should align with the specific requirements of your task, taking into consideration factors such as the nature of the images, task complexity, available computational resources, the task's goals,

and the size of your data. We have suggested models based on various factors and situations. For precise segmentation tasks that require multiple iterations to capture fine details in medical images, FTN's feedback loops (He et al., 2022) can be beneficial. When fine-grained medical image segmentation is necessary and you have sufficient computational resources and data, RAT-NeT's recurrent attention mechanisms (Zhu et al., 2022) are valuable for detailed localization. nnFormer (Zhou et al., 2021a) also performs well in such cases. Moreover, when working with large-scale biomedical images that demand efficient segmentation, SwinBTS (Jiang et al., 2022b) is known for its effective handling of such images. If you need to execute multiple segmentation tasks simultaneously, MTPA-Unet's multitasking capabilities (Jiang et al., 2022a) can streamline your workflow. Dilated Transformer and TFNet (Wang et al., 2021b) are preferred when you want to capture both local and global features in medical images. Dilated Transformer (Shen et al., 2022b) is suitable for feature aggregation and refinement, and TFNeT excels at improving feature representation and aggregation in your medical image segmentation tasks. CCT-based models (Islam et al., 2022) are specialized for medical image classification tasks, especially when dealing with computed tomography (CT) images. Chest L-Transformer (Gu et al., 2022) is tailored for chest image classification tasks, particularly those related to lung conditions and chest radiology. For medical image translation tasks that involve converting between different modalities, such as CT to MRI or vice versa, MMTrans (Yan et al., 2022b) is specialized in multi-modality translation tasks. In scenarios where versatile medical image translation capabilities are required, including cross-source and cross-target translation tasks, TransCSCT (Shi et al., 2022a) is suitable for a broad range of translation scenarios.

6.3 Multi-Modality

The Transformer has demonstrated its potential in multi-modality, which stems from the human ability to perceive and process information from various senses such as vision, hearing, and language. Multi-modality machine learning models are capable of processing and combining different types of data simultaneously. Natural language, vision, and speech are among the most common types of data handled by multi-modal models. Several popular tasks in multi-modality include visual question answering, classification and segmentation, visual captioning, commonsense reasoning, and text/image/video/speech generation. In this section, we present a selection of Transformer-based multi-modal models for each of these tasks providing an overview of their key features and working methods.

6.3.1 Visual Question Answering

Visual question answering is a popular task that can be accomplished using multi-modal models. It involves combining NLP and CV to answer questions about an image or video. The goal is to understand the features of both textual and visual information and provide the correct answer. Typically, the models take an image or video and text as input and deliver text as output answers (Antol et al., 2015, Shih et al., 2016). In this context, we have identified and discussed the significant Transformer models for visual question-answering tasks in Table 17.

Transformer Models	**Processed Data Type (I/O)**	**Architecture (Encoder/ Decoder)**	**Supervision**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**	**Test-Std**
Pixel-BERT (Huang et al., 2020)	Text and Image	Encoder	Self-supervised	Visual Genome, MS-COCO	VQA2.0	74.55
LX-MERT (Tan & Bansal, 2019)	Text and Image	Encoder	Supervised	Visual Genome, MS-COCO, VQA v2.0, GQA, VG-QA	VQA v2.0, GQA, NLVR	72.54
ViLBERT (Lu et al., 2019)	Text and Image	Encoder	Supervised	Conceptual Captions dataset, BookCorpus English Wikipedia	VQA 2.0	70.92
VL-BERT (Su et al., 2020)	Text and Image	Encoder	Supervised	Conceptual Captions dataset, BookCorpus English Wikipedia	VQA 2.0	72.22

Continued on next page

Table 17 – Continued from previous page

Transformer Models	Processed Data Type (I/O)	Architecture (Encoder/ Decoder)	Supervision	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)	Test-Std
UNITER (Chen et al., 2020c)	Text and Image	Encoder	Supervised	MS-COCO, Visual Genome, Conceptual Captions dataset, SBU Captions	VQA 2.0,MS-COCO, Flickr30K, VG, Conceptual Captions dataset, SBU Captions	72.46
GIT (Wang et al., 2022a)	Image and Text	Encoder & Decoder	Supervised	Combination of MS-COCO, SBU Captions, CC-3M, Visual Genome, ALT200M, CC12M	VQAv2, TextVQA, VizWiz-VQA, ST-VQA, OCR-VQA, Visual Genome QA, GQA, OK-VQA	78.81
SIMVLM (Wang et al., 2022d)	Image and Text	Encoder & Decoder	Weakly supervised	ALIGN, Colossal Clean Crawled Corpus (C4)	VQA v2	80.34
BLIP (Li et al., 2022)	Image, Video and Text	Encoder & Decoder	Supervised	MS-COCO, Visual Genome, CC12M, SBU Captions, LAION	VQA v2	78.32

Table 17: Transformer Models for Multi-Modality - Visual Question Answering Task

- **Pixel-BERT:** it is constructed using a combination of a CNN for image pixel extraction and an encoder for text token extraction, with the BERT-based Transformer serving as the cross-modality module. While Pixel-BERT takes the entire image as input to capture all spatial information, other models focus on extracting image features from specific regions. This design enables Pixel-BERT to learn and fuse information from both images and text effectively, facilitating the capture of intricate interactions and dependencies between the two modalities and the generation of meaningful representations that blend them seamlessly (Huang et al., 2020)
- **LX-MERT:** stands for Learning Cross-Modality Encoder Representations from Transformers. It involves processing both images and text using two distinct modules and consists of three encoders. These encoders are merged into a single representation, which is further processed through a Transformer-based architecture. Task-specific heads are then employed to make predictions for various multimodal tasks. LXMERT is known for its versatility, having been pre-trained on large datasets and fine-tuned for specific tasks, which makes it highly adept at handling both textual and image inputs across diverse applications. This pre-trained model utilizes masked modeling and cross-modality for pre-training, resulting in improved capture of relationships between text and images (Tan & Bansal, 2019).
- **VL-BERT:** Visual-Linguistic BERT is a single-stream model that is pre-trained and takes both image and text embedding features as input, rendering it a simple yet powerful model. It combines textual embeddings with image features, prioritizing cross-modal interactions through cross-attention mechanisms. This fused information is processed through a Transformer-based architecture, and task-specific heads are employed for various vision and language tasks. VL-BERT's versatility shines as it effectively comprehends and generates multimodal content, making it well-suited for tasks demanding a deep understanding of both textual and visual information (Su et al., 2020).
- **GIT:** The Generative Image-to-Text Transformer is a multi-modal model designed to generate textual descriptions from visual images. This model employs both the encoder and decoder modules of the Transformer architecture, using the image for encoding and decoding the text. To train the model, a large dataset of images paired with textual descriptions is used, allowing GIT to generate textual descriptions for previously unseen images. This approach has shown promising results in generating high-quality textual descriptions of images (Wang et al., 2022a).
- **SimVLM** & **BLIP** Both of these models are capable of performing the task of visual captioning, which involves generating textual descriptions of visual images. The key features of these models are detailed in Section 6.3.3. The **ViLBERT** model will be further discussed in Section 6.3.4, as it is also proficient in performing visual commonsense reasoning tasks in addition to classification and segmentation tasks.

6.3.2 Classification & Segmentation

Multi-modal classification and segmentation are often considered related tasks that involve classifying or segmenting data based on multiple modalities, such as text, image/video, and speech. As segmentation often helps to classify the image,

text, or speech. In multi-modal classification, the task is to classify data based on its similarity and features using multiple modalities. This can involve taking text, image/video, or speech as input and using all of these modalities to classify the data more accurately. Similarly, in multi-modal segmentation, the task is to segment data based on its features and use multiple modalities to achieve a more accurate segmentation. Both of these tasks require a deep understanding of the different forms of data and how they can be used together to achieve better classification or segmentation performance (Liu et al., 2022c, Mahesh & Renjit, 2020, Wu et al., 2016, Menze et al., 2015). In recent years, Transformer models have shown promising results in multi-modal classification and segmentation tasks. In Table 18, we highlight some of the significant Transformer models that have been developed for these tasks.

Transformer Models	**Processed Data Type (I/O)**	**Architecture (Encoder/ Decoder)**	**Supervision**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**	**Accuracy**
CLIP (Radford et al., 2021)	Image and Text	Image encoder & Text encoder	Supervised	Pre-training dataset from internet for CLIP	LSVRC-2012 ImageNet , ImageNet-V2, ImageNet Rendition, ObjectNet, ImageNet Sketch, ImageNet Adversarial	85.4
ViLT(Kim et al., 2021b)	Image and Text	Encoder	Supervised	MS-COCO,Visual Genome, SBU Captions, Google Conceptual Captions	VQA 2.0, NLVR2, MS-COCO, Flickr30K	NA
ALIGN (Jia et al., 2021)	Image and Text	Image encoder (EfficientNet) & Text encoder (BERT)	Supervised	Conceptual Captions dataset	LSVRC-2012 ImageNet, ImageNet-R, ImageNet-A, ImageNet-V2	88.64
Florence (Yuan et al., 2021)	Image and Text	Text encoder (Transformer) & Image encoder (Vision Transformer)	Supervised	FLD-900M, ImageNet (Swin Transformer & CLIP)	LSVRC-2012 ImageNet	90.05
GIT (Wang et al., 2022a)	Image and Text	Image encoder & Text decoder (Transformer)	Supervised	Combination of MS-COCO, SBU Captions, CC-3M, Visual Genome, ALT200M and CC12M	LSVRC-2012 ImageNet	88.79

Table 18: Multi-Modal Transformer Models - Classification & Segmentation Tasks

- **CLIP:** stands for Constructive Language-Image Pre-Training, is a multi-modal model that is trained in a supervised way with text and image data. This model simultaneously trains both the text and image through an encoder and predicts proper batches of text and image. CLIP is capable of understanding the relationship between text and images, and can generate images based on input text as well as generate text based on input images. CLIP has been shown to perform well on several benchmark datasets and is considered a state-of-the-art model for multi-modal tasks involving text and images (Radford et al., 2021).
- **ALIGN:** it stands for Large-scale ImaGe and Noise-text embedding. This model is a large-scale model that uses vision-language representational learning with noisy text annotations. ALIGN is a pre-trained model which uses a dual-encoder and is trained on huge-sized noisy image-text pair datasets. The dataset scale is able to adjust for noise, eliminating the need for pre-processing. ALIGN uses the contrastive loss to train the model, considering both image-to-text and text-to-image classification losses (Jia et al., 2021).
- **Florence:** is a visual-language representation model that is capable of handling multiple tasks. It is an encoder-based pre-trained model trained on web-scale image-text data, and it can handle high-resolution images. This model shows strong

performance on classification tasks, as well as other tasks like object/action detection and question answering (Yuan et al., 2021).

- **ViLT:** Vision-and-Language Transformer is a multi-modal architecture based on the ViT (Vision Transformer) model, utilizing a free-convolution approach. Unlike other VLP (Vision-and-Language Pre-training) models, ViLT performs data augmentation during the execution of downstream tasks of classification and retrievals, which improves the model's performance. Inspired by Pixel-BERT, ViLT takes the entire image as input instead of just using selected regions. By omitting convolutional visual embedders, ViLT reduces the model size and achieves remarkable performance compared to other VLP models (Kim et al., 2021b).
- **GIT** model is also capable of visual question-answering tasks and the characteristics of the GIT model can be found in Section 6.3.1.

6.3.3 Visual Captioning

Visual captioning is a multi-modal task that involves both CV and NLP. The task aims to generate a textual description of an image, which requires a deep understanding of the relationship between image features and text. The visual captioning process usually involves several steps, starting with image processing, followed by encoding the features into vectors that can be used by the NLP model. These encoded vectors are then decoded into text, typically through generative NLP models. Although it is a complex process, visual captioning has a wide range of applications (Yu et al., 2020, Hossain et al., 2019). In this section, we discuss significant Transformer models for visual captioning tasks (see Table 19).

Transformer Models	Processed Data Type (I/O)	Architecture (Encoder/ Decoder)	Supervision	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
BLIP (Li et al., 2022)	Noisy Image-Text Pair	Text encoder (BERT) & Image encoder (VIT)	Supervised	MS-COCO, Visual Genome, Conceptual Captions, Conceptual 12M, SBU Captions, LAION	MS-COCO,nocaps
SIMVLM (Wang et al., 2022d)	Image and Text	Single Transformer to process both images and text (Encoder & Decoder)	Supervised	ALIGN, Colossal Clean Crawled Corpus (C4)	SNLI-VE, SNLI, MNLI, Multi30k, 10% ALIGN , CC-3M
Florence (Yuan et al., 2021)	Image-Text Pair	Text encoder (Transformer) & Image encoder (Vision Transformer)	Supervised	FLD-900M, ImageNet (Swin Transformer & CLIP)	ILSVRC-2012 ImageNet, MS-COCO, Kinetics-600, Flickr30k, MSR-VTT
GIT (Wang et al., 2022a)	Image-Text Pair	Image encoder & Text decoder (Transformer)	Supervised	Combination of COCO, SBU Captions, CC-3M, Visual Genome, ALT200M and CC12M	Karpathy split of MS-COCO, Flickr30K, nocaps, TextCaps, VizWiz-Captions

Table 19: Multi-modal Transformer Models - Visual Captioning Task

- **BLIP:** Bootstrapping Language-Image Pre-training (BLIP) is a pre-trained model designed to enhance performance on various tasks through fine-tuning for specific tasks. This model utilizes a VLP (Vision and Language Pre-training) framework with an encoder-decoder module of the Transformer architecture, which uses noisy data with captions and is trained to remove noisy captions. BLIP is capable of performing a range of downstream tasks, including image captioning, question answering, image-text retrieval, and more (Li et al., 2022).
- **SimVLM:** a SIMple Visual Language Model (SimVLM) is a pre-trained model that uses weak supervision methods for training. This approach provides the model with greater flexibility and scalability. Instead of using pixel patch projection, this model uses the full image as patches and is trained with a language model. As a result of these methods, SimVLM is capable of performing various tasks, with question answering being one of its significant strengths (Wang et al., 2022d).

- **Florence:** It is a visual-language representation model that can perform multiple tasks. It is an encoder-based pre-trained model trained on web-scale image-text data, which enables it to handle high-resolution images. In addition to tasks such as object/action detection and question answering, Florence also shows strong performance in classification tasks (Yuan et al., 2021).
- The description of the **GIT** model has already been provided in Section 6.3.1. This model also excels at the Question Answering task with high performance.

6.3.4 VISUAL COMMONSENSE REASONING

Visual commonsense reasoning is a challenging task that requires a model with a deep understanding of visualization and different images or videos containing objects and scenes, inspired by how humans see and visualize things. These models capture information from different sub-tasks like object recognition and feature extraction. This information is then transformed into a vector to be used for reasoning. The reasoning module understands the relationship between the objects in the image and the output of the inferencing step provides a prediction about the interaction and relationship between the objects. Visual commonsense reasoning helps to improve the performance of various tasks like classification, image captioning, and other deep understanding-related tasks (Zellers et al., 2019, Xing et al., 2021). In this section, we highlight and discuss significant Transformer models for visual commonsense reasoning tasks that are summarized in Table 20.

Transformer Models	**Processed Data Type (I/O)**	**Architecture (Encoder/ Decoder)**	**Supervision**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**	**Q/AR**
ViLBERT (Lu et al., 2019, Su et al., 2020, Chen et al., 2020c)	Text and Image	Two parallel BERT (Encoder)	Supervised	Conceptual Captions dataset, BookCorpus, English Wikipedia	Visual Commonsense Reasoning (VCR)	54.8
VL-BERT (Su et al., 2020)	Text and Image	BERT-based model (Encoder)	Supervised	Conceptual Captions dataset, BookCorpus, English Wikipedia	Visual Commonsense Reasoning (VCR)	59.7
UNITER (Chen et al., 2020c)	Text and Image	BERT-based model (Encoder)	Supervised	MS-COCO, Visual Genome, Conceptual Captions, SBU Captions, VCR	Visual Commonsense Reasoning (VCR)	62.8
Unicode-VL (Li et al., 2020a)	Image and Text	The structure includes numerous Transformer encoders	Supervised	Conceptual Captions-3M, SBU Captions	Visual Commonsense Reasoning (VCR)	54.5

Table 20: Multi-Modal Transformer Models - Visual Commonsense Reasoning Task

- **ViLBERT:** is a two-stream model trained on text-image pairs, utilizing co-attention to identify crucial features in both text and images. It processes this amalgamated information through a Transformer-based architecture, effectively capturing intricate relationships between text and images. The model employs task-specific heads to make predictions for a range of vision and language tasks. ViLBERT undergoes pre-training on extensive datasets and subsequent fine-tuning for specific tasks, equipping it with proficiency in comprehending and generating multimodal content (Lu et al., 2019).
- **UNITER:** short for Universal Image-Text Representation, is a large-scale pre-trained model constructed through masking techniques. It excels at combining textual embeddings and image features, emphasizing cross-modal interactions with bidirectional attention mechanisms. The amalgamated information is then processed through a Transformer-based architecture, enabling the capture of intricate relationships between text and images. Task-specific heads are utilized for a range of vision and language tasks. Uniter's versatility renders it well-suited for tasks demanding a comprehensive understanding of both textual and visual information (Chen et al., 2020c).
- **Unicoder-VL:** Unicoder-VL is a large-scale pre-trained encoder-based model that utilizes cross-modeling to build a strong understanding of the relationship between image and language. The model employs a masking scheme for pre-training on a large corpus of data. These methods enhance the model's performance on visual commonsense reasoning tasks in addition to visual classification tasks (Li et al., 2020a).

- **VL-BERT** model has been previously described in Section 6.3.2. It is important to note that this model is also capable of performing multi-modal classification and segmentation tasks.

6.3.5 IMAGE/VIDEO/SPEECH GENERATION

Multi-modal generation tasks have gained a lot of attention in the field of artificial intelligence. These tasks involve generating images, text, or speech from inputs of different modalities of input. In recent times, several generative models have demonstrated outstanding performance, making this field of research even more attractive (Suzuki & Matsuo, 2022). In this section, we discuss some significant Transformer models that have been used for multi-modal generation tasks. These models are summarized in Table 21.

Transformer Models	**Processed Data type (I/O)**	**Architecture (Encoder/ Decoder)**	**Data Augmentation**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**	**FID**
DALL-E (Ramesh et al., 2021)	Image and Text	dVAE encoder/decoder & Sparse-Transformer decoder	Yes	Not disclosed	Training: Conceptual Captions, A collection of 250 million pairs of text and images sourced from the internet Evaluation: CUB, MS-COCO	28
GLIDE (Nichol et al., 2022)	Image and Text	Diffusion model (Encoder & Decoder)	No	NA	The model is trained on the same datasets identical to those utilized for DALL-E	12.89
CogView (Ding et al., 2021)	Image and Text	GPT-based model (Decoder)	No	VQ-VAE	MS-COCO, Project WudaoCorpora	27.1

Table 21: Multi-Modal Transformer Models - Image/Video/Speech Generation Task

- **DALL-E:** DALL-E (Ramesh et al., 2021) is a popular Transformer-based model for generating images from text. It is trained on a large and diverse dataset of both text and images, utilizing 12 billion parameters from the GPT-3 architecture. To reduce memory consumption, DALL-E compresses images without compromising their visual quality. An updated version of DALL-E, known as DALL-E 2, has been introduced with a higher number of parameters (175 billion) which allows for the generation of higher resolution images. Additionally, DALL-E 2 is capable of generating a wider range of images.
- **CogView:** CogView (Ding et al., 2021) is an image generation model that generates images based on input text descriptions, making it a challenging task that requires a deep understanding of the contextual relationship between text and image. It utilizes a Transformer-GPT-based architecture to encode the text into a vector and decode it into an image. This model outperforms DALL-E in some cases, which also generates images from text descriptions, but uses text-image pairs for training the model.
- **GLIDE:** short for Guided Language to Image Diffusion for Generation and Editing, GLIDE is a diffusion model that is distinct from conventional models in that it can both generate and edit images. Unlike other models, diffusion models are sequentially injected with random noise and trained to remove that noise to construct the original data. GLIDE takes textual information as input and generates an output image conditioned on that information. In some instances, the images generated by GLIDE are more impressive than those generated by DALL-E (Nichol et al., 2022).

6.3.6 CLOUD COMPUTING

Cloud computing is a crucial element of modern technology, particularly with regard to the Internet of Things (IoT). It encompasses a wide variety of cloud-based tasks, including server computing, task scheduling, storage, networking, and more. In wireless networks, cloud computing aims to improve scalability, flexibility, and adaptability, thereby providing seamless connectivity. To achieve this, data or information is retrieved from the network for computation, with various types of data being processed, including text, images, speech, and digits. Due to this multi-modal approach, cloud computing is classified in this category (Li et al., 2017, Jauro et al., 2020, Yu et al., 2017). In this article, we focus on the significant Transformer models used in cloud computing tasks, which are presented in Table 22.

Transformer Models	Processed Data type (I/O)	Task Accomplished	Year	Architecture (Encoder/ Decoder)	Pre-trained (Yes/No)	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
VMD & R-Transformer (Zhou et al., 2020)	workload sequence	cloud workload forecasting	2020	Encoder & Decoder	No	NA	Google cluster trace, Alibaba cluster trace
ACT4JS (Xu & Zhao, 2022)	Cloud jobs/task	Cloud computing resource job scheduling.	2022	Encoder	No	NA	Alibaba Cluster-V2018
TEDGE-Catching (Hajiakhondi-Meybodi et al., 2022)	Sequential request pattern of the contents (Ex: video, image, websites, etc)	Predict the content popularity in proactive caching schemes.	2021	Encoder	No	NA	MovieLens
SACCT (Wang et al., 2022b)	Network status (Bandwidth, storage, and etc)	Optimize network based on network status.	2021	Encoder	No	NA	Not mentioned clearly

Table 22: Multi-Modal Transformer Models - Cloud Computing Task

- **VMD & R-TRANSFORMER:** Workload forecasting is a critical task for the cloud, and previous research has focused on using RNNs for this purpose. However, due to the highly complex and dynamic nature of workloads, RNN-based models struggle to provide accurate forecasts because of the problem of vanishing gradients. In this context, the proposed Variational Mode Decomposition-VMD and R-Transformer model offers a more accurate solution by capturing long-term dependencies using MHA and local non-linear relationships of workload sequences with local techniques (Zhou et al., 2020). Therefore, this model is capable of executing the workload forecasting task with greater precision than existing RNN-based models.
- **TEDGE-Caching:** It is an acronym for Transformer-based Edge Caching, which is a critical component of the 6G wireless network as it provides a high-bandwidth, low-latency connection. Edge caching stores multimedia content to deliver it to users with low latency. To achieve this, it is essential to proactively predict popular content. However, conventional models are limited by long-term dependencies, computational complexity, and the inability to compute in parallel. In this context, TEDGE caching framework incorporates a ViT to overcome these limitations, without requiring data preprocessing or additional contextual information to predict popular content at the Mobile Edge. This is the first model to apply a Transformer-based approach to execute this task, resulting in superior performance (Hajiakhondi-Meybodi et al., 2022).
- **ACT4JS:** The Actor-Critic Transformer for Job Scheduling (ACT4JS) is a Transformer-based model designed to allocate cloud computing resources to different tasks in cloud computing. The model consists of an Actor and Critic network, where the Actor-network selects the best action to take at each step, and the Critic network evaluates the action taken by the Actor-network and provides feedback to improve future steps. This approach allows for a better understanding of the complex relationship between cloud jobs and enables prediction or scheduling of jobs based on different features such as job priority, network conditions, resource availability, and more (Xu & Zhao, 2022).
- **SACCT:** It refers to the Soft Actor-Critic framework with a Communication Transformer, which combines the Transformer, reinforcement learning, and convex optimization techniques. This model introduces the Communication Transformer (CT), which works with reinforcement learning to adapt to different challenges, such as bandwidth limitations, storage constraints, and more, in the wireless edge network during live streaming. Adapting to changing network conditions is critical during live streaming, and SACCT provides a system that adjusts resources to improve the quality of service based on user demand and network conditions. The SACCT model's ability to adapt to changing conditions and optimize resources makes it an important contribution to the field of wireless edge network technology (Wang et al., 2022b).

Discussion:

Multi-modality is an emerging field within deep learning, involving tasks that encompass various types of data. Comparing models for multi-modal tasks presents unique challenges. We conducted a comprehensive analysis of models for each task and identified key factors for comparison. In the realm of visual question-answering tasks, we identified a common factor for model comparison: all models were fine-tuned on the VAQ 2.0 dataset. In terms of the test-std accuracy score, SIMVLM leads the pack with an impressive accuracy score of 80.34%, demonstrating its robust performance. GIT follows closely with an accuracy of 78.81%, while BLIP also performs strongly at 78.32%. These three models emerge as top choices for accurate multi-modal question answering. Pixel-BERT exhibits competence with an accuracy score of 74.55%, making it a reliable option for tasks that require a balance between accuracy and computational efficiency. LX-MERT and Uniter achieve similar scores at 72.54% and 72.46%, respectively, indicating their suitability for mid-level performance tasks. VL-BERT and ViLBERT, with scores of 72.22% and 70.92%, respectively, provide viable options but may benefit from further tuning or customization to match the accuracy of the leading models

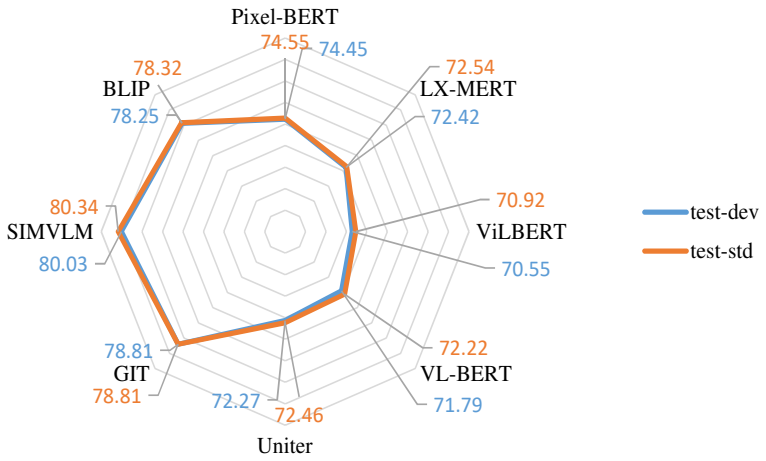

Figure 13: Comparing Transformer Models for Visual Question Answering: Test-Dev and Test-Std Metrics

When comparing models for multi-modal classification and segmentation tasks, our analysis revealed that each of these models is fine-tuned with the ImageNet dataset. In this context, Florance (Yuan et al., 2021) emerges as the top performer, achieving an impressive top-1 score of 90.05%, demonstrating excellence in both domains. GIT (Wang et al., 2022a) closely follows with an 88.79% top-1 score, particularly excelling in classification. Align (Jia et al., 2021), with an 88.64% score, stands out in segmentation while slightly trailing in classification compared to GIT and Florance. VATT (Akbari et al., 2021), with an 85.4% score, although not reaching the same heights as the top three models, remains a reliable choice for both tasks.

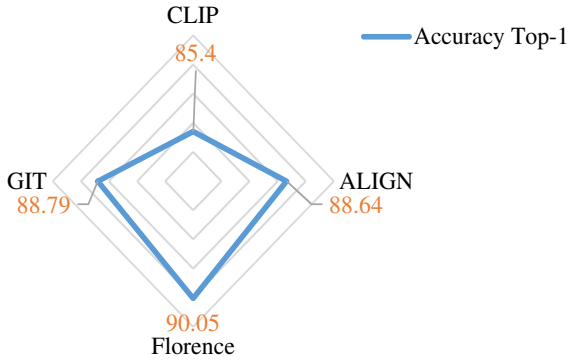

Figure 14: Comparison of Transformer Models for Classification & Segmentation Tasks Based on the Top-1 Accuracy Metric

Furthermore, following our thorough and detailed analysis of models for visual captioning tasks, we found no common points or direct relevance for comparison between these models. These models exhibit varying performance metrics for visual captioning tasks, making direct comparisons challenging. However, we can still highlight some key insights about these models: Supervised Training: Most of these models are trained in a supervised manner. BLIP (Li et al., 2022) and Florence (Yuan et al., 2021) are built on an encoder module. GIT (Wang et al., 2022a) employs an image encoder and a single text decoder. SimVLM (Wang et al., 2022d) utilizes both encoder and decoder modules of the Transformer architecture. While

these insights provide context, it's important to note that the models excel in visual captioning tasks with varying performance metrics and unique approaches

In our examination of models for visual commonsense reasoning tasks, we discovered that most of these models utilize masked language modeling. Additionally, all of them undergo fine-tuning on the VCR dataset. To evaluate their performance, we consider scores for Q!A, QA!R, and Q!AR. While a basic Q!A pair provides an answer to a question, QA!R and Q!AR go further by providing reasoning or rationale behind the answer. When we focus on QA!R performance, Uniter (Chen et al., 2020c) stands out as the top-performing model with an impressive score of 80.0%. ViLBERT (Lu et al., 2019) and Unicoder-VL (Li et al., 2020a) perform quite similarly, achieving scores of 74.6% and 74.5%, respectively. On the other hand, VL-BERT (Su et al., 2020) attains a score of 78.4% in visual reasoning tasks, as viewed from the perspective of QA!R. The models' performance for Q!A and Q!RA is visually presented below

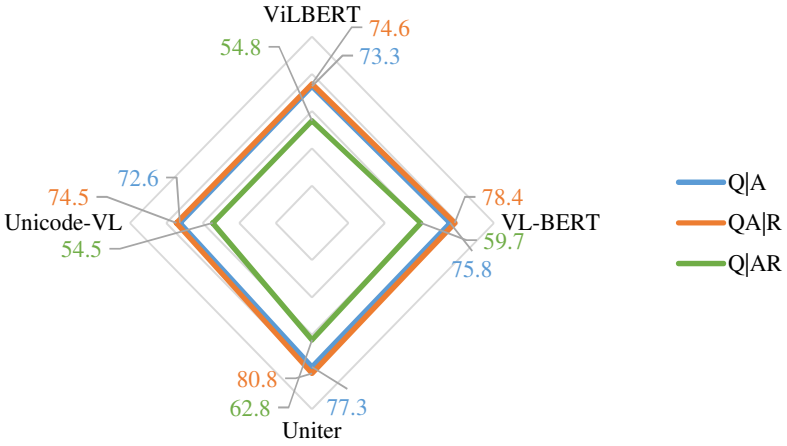

Figure 15: Comparison of Transformer Models Developed for Visual Commonsense Reasoning Based on the Q!A, QA!R, and Q!AR Metrics

When assessing multi-modal generative models, we found that the FID (Fréchet Inception Distance) performance metric serves as an option for comparing these models. In this context, DALL-E (Ramesh et al., 2021) achieved the highest FID score of 28 among these models. On the other hand, GLIDE (Nichol et al., 2022) obtained a FID score of 12.89, which is lower than DALL-E but still performs admirably in specific cases. Furthermore, CogView (Ding et al., 2021) demonstrated a solid performance, with a FID score of 27.1, which is very close to DALL-E's score. Detailed performance scores are provided in the table below.

Cloud computing tasks encompass a wide range of considerations, including task nature, content generation requirements, available data, computational resources, and the degree of fine-tuning needed for optimal performance. When the goal is to accurately forecast cloud workload with a focus on handling long-range dependencies, VMD & R-Transformer (Zhou et al., 2020) excels. This model effectively tracks nonlinear relationships within workload sequences. On the other hand, for advanced edge caching in wireless networks, such as 6G, TEDGE-Caching (Hajiakhondi-Meybodi et al., 2022) is a valuable choice. It can predict popular content without the need for extensive data preprocessing or additional contextual information. For workload scheduling in the cloud, the ACT4JS (Xu & Zhao, 2022) model offers a viable solution. It demonstrates an exceptional ability to comprehend complex relationships among cloud jobs. Using an actor and critic approach, it enables precise job prediction and scheduling based on various job features. Lastly, in scenarios where enhancing service quality through resource adjustments and adapting to cloud changes with optimization is crucial, the SACCT (Wang et al., 2022b) model stands out as a suitable option.

The choice of a model for a given task is of paramount importance, as the selection should align with the available resources and the level of specialization needed for the specific use case. It is crucial to consider the strengths and characteristics of each model in relation to the project's objectives and constraints. When substantial computational resources are at hand, and versatility is required for various multimodal tasks, including Visual Question Answering (VQA), BERT-based models (Huang et al., 2020, Tan & Bansal, 2019, Lu et al., 2019, Su et al., 2020, Chen et al., 2020c) can be fine-tuned to excel in VQA and a wide array of tasks, including commonsense reasoning. For tasks that involve both text and image generation alongside VQA, GIT (Wang et al., 2022a) proves to be a versatile choice, capable of generating textual and visual content conditioned on each other. In scenarios where simplicity and efficiency are valued while maintaining competitive VQA performance, SIMVLM (Wang et al., 2022d) focuses on essential visual and textual interactions for VQA. Conversely, when the objective is to enhance VQA performance by emphasizing fine-grained linguistic and visual reasoning, BLIP (Li et al., 2022) concentrates on improving language-image pre-training for VQA, particularly leveraging bootstrapping data to achieve valuable results. Therefore, the choice of model should be guided by a thoughtful assessment of project goals and available

resources. When the need arises for a versatile multi-modal model, one that has demonstrated exceptional performance across a wide spectrum of tasks, CLIP, with its contrastive learning approach as demonstrated in (Radford et al., 2021), becomes a fitting choice. It particularly shines in tasks like classification and segmentation, where aligning text and images is crucial. In scenarios where the primary task revolves around text-to-text generation, encompassing the generation of textual descriptions or labels for images, VATT (Akbari et al., 2021) proves invaluable. Its focus on visual attention makes it well-suited for such endeavors, especially when the goal is to capture complex relationships between vision and language. For situations where bridging the gap between different modalities is paramount, Unicoder-VL (Li et al., 2020a) excels in this endeavor. When the objective is to enhance cross-modal pre-training for multi-modal tasks, including classification and segmentation, ViLT specializes in improving the alignment between vision and language, as described in (Kim et al., 2021b). Moreover, if there is a desire to extend BERT's capabilities to encompass multi-modal tasks, ALIGN, as described in (Jia et al., 2021), is tailored to capturing correlations between text and images, rendering it suitable for a diverse array of multi-modal tasks. Furthermore, when tasks entail intricate vision-language interactions and demand a model equipped with a fusion mechanism to seamlessly integrate information from different modalities, Florance, introduced in (Yuan et al., 2021), is meticulously crafted for such vision-language tasks. Its primary goal is to enhance language-image pre-training for visual captioning by enabling improved fine-grained linguistic and visual reasoning. For specialized expertise in vision-language tasks and an exemplary ability to capture intricate connections between modalities, VILT (Kim et al., 2021b) stands out as a valuable choice. If your primary objective is to create visually coherent and imaginative content based on textual descriptions, DALL-E (Ramesh et al., 2021) proves to be an ideal choice for tasks involving the generation of images from textual prompts. Conversely, when it comes to image enhancement and denoising tasks, GLIDE (Nichol et al., 2022) excels. It offers significant value in improving the visual quality of multi-modal content and possesses the versatility to generate both textual and visual content. For tasks that necessitate the generation of diverse multi-modal content, with the capability to create both text and images, Chimera, presented in (Li & Hoefler, 2021), stands as a valuable option, providing flexibility in content generation. In scenarios where concept-driven image generation is of paramount importance, such as tasks involving the creation of images based on specific concepts or ideas, CogView, as specialized in concept-driven image synthesis per (Ding et al., 2021), is aptly suited for concept-driven multi-modal content generation tasks.

6.4 Audio & Speech

Audio and speech processing is one of the most essential tasks in the field of deep learning. Along with NLP, speech processing has also gained attention from researchers, leading to the application of DNN methods. As Transformers have achieved great success in the field of NLP, researchers have also had significant success when applying Transformer-based models to speech processing.

6.4.1 Speech Recognition

Speech Recognition is one of the most popular tasks in the field of artificial intelligence. It is the ability of a model to identify human speech and convert it into a textual or written format. This process is also known as speech-to-text, automatic speech recognition, or computer-assisted transcription. Speech recognition technology has advanced significantly in the last few years. It involves two types of models, namely the acoustic model and the language model. Several features contribute to the effectiveness of speech recognition models, including language weighting, speaker labeling, acoustics training, and profanity filtering. Despite significant advancements in speech recognition technology, there is still room for further improvement (Yu & Deng, 2016, Nassif et al., 2019, Deng et al., 2013). One promising development in this area is the use of Transformer-based models, which have shown significant improvement in various stages of the speech recognition task. The majority of Transformer-based audio/speech processing models have focused on speech recognition tasks, and among these, several models have made exceptional contributions, exhibited high levels of accuracy, introduced new effective ideas, or created buzz in the AI field. In Table 23, we highlight the most significant Transformer models for speech recognition tasks.

Transformer Models	Architecture (Encoder/ Decoder)	Lingual Capabilities	Supervision	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)
Conformer (Peng et al., 2021)	Encoder & Decoder	Monolingual	Supervised	NA	Librispeech, test/testother
Speech Transformer (Dong et al., 2018)	Encoder & Decoder	Monolingual	Supervised	NA	Wall Street Journal (WSJ)

Continued on next page

Table 23 – Continued from previous page

Transformer Models	Architecture (Encoder/ Decoder)	Lingual Capabilities	Supervision	Pre-training Dataset	Dataset(Fine-tuning, Training, Testing)
VQ-Wav2vec (Baevski et al., 2020a)	Encoder	Monolingual	Self-supervised	Librispeech	TIMIT, Wall Street Journal (WSJ)
Wav2vec 2.0 (Baevski et al., 2020b)	Encoder	Monolingual	Self-supervised	Librispeech, LibriVox	Librispeech, LibriVox, TIMIT
HuBERT (Hsu et al., 2021)	Encoder	Monolingual	Self-supervised	Librispeech, Libri-light	Librispeech, Libri-light
Whisper (Radford et al., 2022)	Encoder & Decoder	Multilingual	Weakly-supervised	NA	VoxLingua107, LibriSpeeech, CoVoST2, Fleurs, Kincaid46
Transformer Transducer (Zhang et al., 2020b)	Encoder	Monolingual	Supervised	NA	LibriSpeeech
XLSR-Wav2Vec2 (Conneau et al., 2021)	Encoder	Multilingual	Unsupervised	53 languages datasets	CommonVoice, BABEL Benchmark,Multilingual LibriSpeech (MLS)

Table 23: Transformer Models for Audio & Speech Recognition Task

- **Conformer:** The Conformer architecture is a model that combines the advantages of both the Transformer and CNN for automatic speech recognition tasks. While the Transformer is proficient at capturing global features, and CNN excels at capturing local features, the Conformer architecture leverages the strengths of both to achieve superior performance. Recent studies have shown that the Conformer architecture outperforms both CNN and Transformer models individually, thereby setting a new state-of-the-art in automatic speech recognition performance (Gulati et al., 2020).
- **Speech Transformer:** It is a speech recognition model published in 2018 which is one of the earliest Transformer inspired speech models. It eliminates the conventional RNN-based approach in speech processing and instead applies an attention mechanism. The introduction of the Transformer model into speech recognition has led to a number of benefits. For instance, the training time and memory usage become lower, allowing for better scalability. This is especially helpful for tasks that require a long-term dependency due to the elimination of recurrence sequence-to-sequence processing (Dong et al., 2018).
- **VQ-Wav2vec:** VQ-Wav2Vec is a Transformer based model that enables Vector Quantization-VQ tasks to be executed in a self-supervised way. It is built on the Wav2Vec model, which is an effective way of compressing continuous signals into discrete symbols. Unlike other models, VQ-Wav2Vec trains without the need of unlabeled data. Instead, corrupted speech is used, and the model learns by predicting the missing parts of the speech. This process of training has been proven to be highly effective, and the model is capable of achieving a higher accuracy when compared to other models (Baevski et al., 2020a).
- **Wav2vec 2.0:** Wav2Vec 2.0 is a self-supervised model which uses discretization algorithms to capture the vocabulary from raw speech representation. The learned vocabulary is passed through an architecture consisting of multi-layer convolutional feature encoder. This encoder has multiple convolution layers, layer normalization and an activation function, with the audio representations being masked during the training process. Wav2Vec 2.0 offers the advantage of performing well on speech recognition tasks with a small amount of supervised data (Baevski et al., 2020b).
- **HuBERT:** HuBERT, short for Hidden-Unit BERT, is a self-supervised speech representation model. Its approach involves offline clustering for feature representation, with the loss calculation restricted to the masked regions. This emphasis allows the model to effectively learn a combination of acoustic and language models over the input data. HuBERT consists of a convolutional waveform encoder, a projection layer, a BERT encoder, and a code embedding layer. The CNN component generates feature representations, which are then subjected to random masking. These masked representations are subsequently passed through the BERT encoder, yielding another set of feature representations. HuBERT's functioning resembles

that of a mask language model and has demonstrated notable performance in speech representation tasks, particularly speech recognition (Hsu et al., 2021).

- **Whisper:** It is a noteworthy speech recognition model that emerged in late 2022, specifically designed to address the challenging task of recognizing speech with low volume. The uniqueness of this model lies in its dedicated efforts to improve low-volume speech recognition. Whisper adopts a training approach that incorporates a lower level of speech data and leverages weak supervision methods, enabling training on a larger corpus of data. This strategic approach has proven instrumental in enhancing the performance of the Whisper model, enabling it to effectively capture and comprehend low-level speech phenomena (Radford et al., 2022).
- **Transformer Transducer:** It is a speech recognition model that capitalizes on the strengths of both the self-attention mechanism of the Transformer and the RNN. This model is constructed by integrating the encoder module of the Transformer with the RNN-T loss function. The encoder module is responsible for extracting speech representations, while the RNN-T component utilizes this information to make real-time predictions of the transcript, facilitating a swift response—an essential requirement for speech recognition tasks (Transformer-Transducer (Zhang et al., 2020b).
- **XLSR-Wav2Vec2:** This model demonstrates the capability to recognize speech across multiple languages, eliminating the need for extensive labeled data in each language for training. By learning the relationships and shared characteristics among different languages, this model surpasses the requirement of training on specific language-labeled speech data. Consequently, the XLSR-Wav2Vec2 model offers an efficient solution for multiple-language speech recognition, requiring significantly less data for training while adhering to the architectural principles of Wav2Vec2 (Conneau et al., 2021).

6.4.2 Speech Separation

It poses a considerable challenge within the field of audio signal processing. It involves the task of separating the desired speech signal, which may include various sources such as different speakers or human voices, from additional sounds such as background noise or interfering sources. In the domain of speech separation, three commonly employed methods are followed: (i) Blind source separation, (ii) Beamforming, and (iii) Single-channel speech separation. The significance of speech separation has grown with the increasing popularity of automatic speech recognition (ASR) systems. It is often employed as a preprocessing step for speech recognition tasks. The accurate distinction between the desired speech signal and unwanted noise is crucial to ensure precise speech recognition results. Failure to properly segregate the desired speech from interfering noise can lead to erroneous speech recognition outcomes (Wang & Chen, 2018, Huang et al., 2014). In this context, we present several Transformer-based models that have showcased noteworthy advancements in audio and speech separation tasks. The details of these models are presented in Table 24.

Transformer Models	Architecture	Input	Pre-training Dataset	Dataset (Fine-tuning, Training, Testing)	SDR
DPTNeT (Chen et al., 2020a)	Hybrid of Transformer and RNN (Encoder & Decoder)	Multi-speaker monaural speech	NA	WSJ0-2mix, LS-2mix	20.6
Sepformer (Subakan et al., 2021)	RNN-free Transformer-based architecture(Encoder & Decoder)	Multi-speaker speech	NA	WSJ0-2mix, WSJ0-3mix	20.5
WavLM (Chen et al., 2022a)	Transformer & convolutional encoders	Multi-speaker noisy speech	GigaSpeech, VoxPopuli,Libri-light	CSR-II (WSJ1) Complete,LibriCSS	NA

Table 24: Transformer Models for Audio & Speech Separation Task

- **DPTNeT:** DPTNeT stands for Dual-Path Transformer Network for monaural speech separation tasks. This model trained directly minimizes the error between estimated and target value which is called end-to-end processing. This model uses dual-path architecture, replacing positional encoding with the RNN in the Transformer architecture which helps to capture complex features of the signal and improves the performance of the speech separation from the overlapped speech (Chen et al., 2020a).
- **Sepformer:** The sepformer model was published in a paper titled "Attention is all you need in speech separation" which uses an attention mechanism to separate speeches that are overlapped. This model does not contain any kind of recurrence scheme and it follows the self-attention mechanisms. Sepformer uses a binary mask prediction scheme for training while this masking network captures both short and long-term dependencies and provides higher accuracy in performance (Subakan et al., 2021).

- **WavLM:** WavLM is a large-scale pre-trained model that can execute a range of tasks for speech. WavLM follows BERT-inspired speech processing model-HuBERT, whereas, with the help of mask speech prediction, the model predicts the actual speech by removing the noise from the corrupted speech. By this way, this model is trained for a variety of tasks besides automatic speech recognition-ASR task (Chen et al., 2022a).

6.4.3 SPEECH CLASSIFICATION

The speech classification task refers to the ability to categorize input speech or audio into distinct categories based on various features, including speaker, words, phrases, language, and more. There exist several speech classifiers, such as voice activity detection (binary/multi-class), speech detection (multi-class), language identification, speech enhancement, and speaker identification. Speech classification plays a crucial role in identifying important speech signals, enabling the extraction of relevant information from large speech datasets (Livezey et al., 2019, Gu et al., 2017). In this context, we present a compilation of Transformer-based models, which have demonstrated superior accuracy in speech classification tasks compared to conventional models. The details of these models are depicted in Table 25.

Transformer Models	**Architecture (Encoder/ Decoder)**	**Lingual Capabilities**	**Supervision**	**Data augmentation**	**Pre-training Dataset**	**Dataset (Fine-tuning, Training, Testing)**
AST (Gong et al., 2021)	Encoder	Monolingual	Supervised	Yes	ILSVRC-2012 ImageNet	AudioSet, ESC-50, Speech Commands
Mockingjay (Liu et al., 2020)	Encoder	Monolingual	Unsupervised	No	LibriSpeech	LibriSpeech train-clean-360 subset
XLS-R (Babu et al., 2022)	Encoder & Decoder	Multilingual	Self-supervised	No	VoxPopuli, Multilingual LibriSpeech (MLS), Common Voice, BABEL Benchmark, VoxLingua107	CoVoST-2
UniSpeech-SAT (Chen et al., 2022b)	Encoder	Monolingual	Self-supervised	Yes	LibriSpeech, LibriVox, GigaSpeech, VoxPopuli	SUPERB

Table 25: Transformer Models for Audio & Speech Classification Task

- **AST:** AST - Audio Spectrogram Transformer is a Transformer-based model which is applied to an audio spectrogram. AST is the first audio classification model where the convolution was not used and it is capable of capturing long-range frames context. It used a Transformer encoder to capture the features in the audio spectrogram, a linear projection layer, and a sigmoid activation function to capture the audio spectrogram representation for audio classification. As the attention mechanism is renowned for capturing global features so it shows significant performance in audio/speech classification tasks (Gong et al., 2021).
- **Mockingjay:** Mockingjay is an unsupervised speech representation model that uses multiple layers of bidirectional Transformer pre-trained encoders. It uses both past and future features for speech representation rather than only past information, which helps it to gather more information about the speech context. Mockingjay also can improve the performance of the supervised learning tasks as well where the amount of labeled data is low. Capturing more information helped to improve several speech representational tasks like speech classification and recognition (Liu et al., 2020).
- **XLS-R:** XLS-R is a Transformer-based self-supervised large-scale speech representation model that is trained with a large amount of data. It is built on the wav2vec discretization algorithm, whereas it uses Wav2Vec 2.0 model that is pre-trained with multiple languages. The architecture contains multiple convolution encoders to map raw speech and the output from this stage is transferred to the Transformer model(encoder module) as input which provides better audio representation. A large amount of training data is crucial for this model where a range of public speech is used and it performed well for multiple downstream multilingual speech tasks (Babu et al., 2022).
- **UniSpeech:** UniSpeech is a semi-supervised unified pre-trained model for speech representation. This model follows the Wav2Vec 2.0 architecture where it contains convolutional feature encoders that convert the raw audio to a higher-dimensional representation and this output is fed into the Transformer. This model is capable of learning to multitask while a quantizer is used in its architecture which helps to capture specific speech recognition information (Chen et al., 2022b).

Discussion:
Audio or speech processing is among the most prominent deep-learning tasks. Comparing models in this field is challenging, as each model exhibits distinct strengths and qualities. In our assessment of audio and speech processing models, we focused on supervision, as depicted in Figure 16. For audio and speech recognition tasks, we observed that the type of supervision and the number of languages supported are significant factors for comparison. Whisper (Radford et al., 2022) and XLSR-Wav2Vec2 (Conneau et al., 2021) are weakly supervised and unsupervised models, respectively, while all other models fall under the category of supervised models. Additionally, the majority of models are monolingual, with the exception of Whisper and XLSR-Wav2Vec2, which can handle audio recognition for multiple languages.
In the case of audio separation tasks, the Source-to-Distortion Ratio (SDR) serves as a crucial performance metric. DPTNeT and Sepformer both utilize SDR metrics, achieving similar performance with DPTNeT (Chen et al., 2020a) scoring 20.6 SDR and Sepformer (Subakan et al., 2021) scoring 20.5 SDR. However, we did not find an SDR score for the WavLM model (Chen et al., 2022a). It's worth noting that DPTNeT combines Transformer and RNN components, while Sepformer is an RNN-free model, and WavLM employs only a Transformer encoder module in conjunction with a convolution encoder.
Audio and speech classification is another significant task in this domain. In our examination of these models, we considered two key factors for comparison: the level of supervision and language support. Monkingjay (Liu et al., 2020) is the sole model trained with unlabeled data, making it unsupervised, while all others are supervised models. Furthermore, AST (Gong et al., 2021), Mockingjay (Liu et al., 2020), and Unispeech-SAT (Chen et al., 2022b) are monolingual audio classification models, whereas XLS-R (Babu et al., 2022) stands out as a significant model capable of classifying audio in multiple languages.

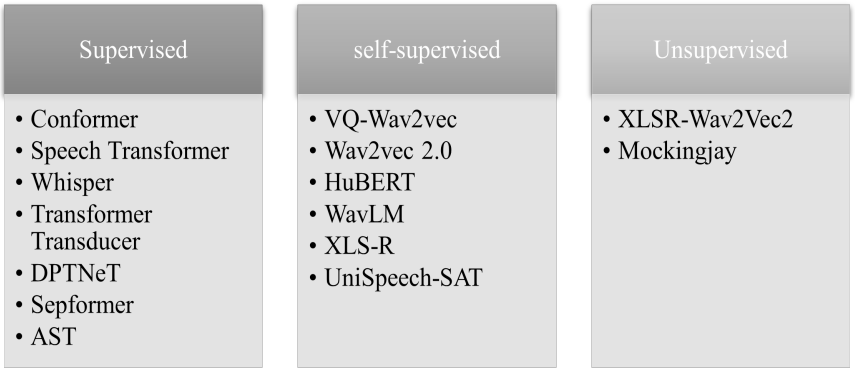

Figure 16: Comparison of Audio and Speech Models Based on Type of Supervision

Audio and speech tasks encompass various levels and types, each model tailored to specific goals. Performance depends on factors such as resource availability and data. For robust Automatic Speech Recognition (ASR) with long-range dependencies, choose Conformer (Peng et al., 2021). For real-time ASR with limited resources, opt for Speech Transformer (Dong et al., 2018). In cases of unsupervised or self-supervised learning with limited labeled data but ample unlabeled data, consider VQ-Wav2vec (Baevski et al., 2020a) or HuBERT (Hsu et al., 2021). Whisper (Radford et al., 2022) is efficient for low-resource ASR. For multilingual ASR, XLSR-Wav2Vec2 (Conneau et al., 2021) is designed for versatility. Building end-to-end ASR systems, Transformer Transducer (Zhang et al., 2020b) combines Transformer and sequence-to-sequence models effectively. For permutation-invariant speech separation with unknown or variable source orders, choose DPTNeT (Chen et al., 2020a). Moreover, for high-quality speech separation capturing long-range dependencies, consider Sepformer (Subakan et al., 2021). For complex waveform-level speech separation, WavLM (Chen et al., 2022a) is a viable option. AST (Gong et al., 2021) is suitable in the case of speech classification with complex temporal dependencies and long-term context. In scenarios requiring end-to-end ASR tasks that include speech classification components, such as classifying spoken words or phrases, Mockingjay (Liu et al., 2020) is a valuable option. For cross-lingual speech classification, XLS-R (Babu et al., 2022) excels. For data augmentation to enhance classification tasks, UniSpeech-SAT (Chen et al., 2022b) is effective, improving robustness and accuracy through data augmentation.

6.5 Signal Processing

With the growing recognition of the usability of Transformer-based models across various sectors, researchers have started exploring their application in signal processing. This recent development of utilizing Transformer-based models in signal processing represents a novel approach that outperforms conventional methods in terms of performance. Signal processing involves the manipulation and analysis of various types of data, including signal status, information, frequency, amplitude, and more. While audio and speech are considered forms of signals, we have segregated the audio and speech sections to highlight the specific applications of Transformer-based models in those domains. Within the signal processing domain, we have focused on two distinct areas: wireless network signal processing and medical signal processing. These two fields exhibit distinct processing methods and functionalities due to their inherent differences. Here, we delve into both of these

tasks and provide an overview of significant Transformer-based models that have demonstrated effectiveness in these specific domains.

6.5.1 Wireless Network & Signal Processing

In the current era of the 21st century, wireless network communication has emerged as a prominent technology. However, the application of Transformers in wireless network signal processing has not received substantial attention thus far. Consequently, the number of Transformer-inspired models developed for this field remains limited. Wireless network signal processing encompasses various tasks, including signal denoising, signal interface detection, wireless signal channel estimation, interface identification, signal classification, and more. DNNs offer great potential for tackling these tasks effectively, and Transformer-based models have introduced significant advancements in this domain (Sun et al., 2017b, Clerckx et al., 2021, Zhang et al., 2019, Chen et al., 2019b). In this section, we present several models that have made notable contributions to the enhancements in wireless communication networks and signal processing. The details of these models are provided in Table 26.

Transformer Models	Task Accomplished	Input	Architecture	Type of Attention	Dataset (Fine-tuning, Training, Testing)
SigT (Ren et al., 2022)	Signal detection, channel estimation, interference suppression, and data decoding in MIMO-OFDM	Signal	Encoder	Multi-head self-attention (MHA)	Peng Cheng Laboratory(PCL), local area data
TSDN (Liu et al., 2022b)	Remove interference and nose from wireless signal	Noisy signal	Encoder & Decoder	Multi-head self-attention (MHA)	Wall NLoS, Foil NLOS
ACNNT (Wang et al., 2021a)	Wireless interface identification	Interference signal	Encoder	Multi-head self-attention (MHA) Channel Attention (CA)	Datasets of signals (ST, BPSK, AM, NAM, SFM, LFM, 4FSK, 2FSK) generated in Matlab
MCformer (Hamidi-Rad & Jain, 2021)	Automatic modulation classification complex raw radio signals	Raw radio signal	Encoder	Multi-head self-attention (MHA)	RadioML2016.10b
Quan-Transformer (Xie et al., 2022)	Compress & recover channel state information	Signal	Encoder & Decoder	Multi-head self-attention (MHA)	NA(Customer data)

Table 26: Transformer Models for Wireless Networks & Signal Processing

- **SigT:** SigT is a wireless communication network signal receiver designed with a Transformer architecture, capable of handling Multiple-input Multiple-output (MIMO-OFDM) signals. Leveraging the Transformer's encoder module, this innovative framework enables parallel data processing and performs essential tasks such as signal detection, channel estimation, and data decoding. Unlike traditional receivers that rely on distinct modules for each task, SigT seamlessly integrates these functions, providing a unified solution (Ren et al., 2022).
- **TSDN:** It is an abbreviation for Transformer-based Signal Denoising Network. It refers to a signal denoising model based on Transformers that aims to estimate the Angle-of-Arrival (AoA) of signals transmitted by users within a wireless communication network. This Transformer-based model significantly enhances the accuracy of AoA estimation, especially in challenging non-line-of-sight (NLoS) environments where conventional methods often fall short in delivering the desired precision (Liu et al., 2022b).
- **ACNNT:** The Augmented Convolution Neural Network with Transformer (ACNNT) is an architectural framework specifically designed for identifying interference within wireless networks. This model combines the power of CNNs and Transformer architectures. The multiple CNN layers in ACNNT extract localized features from the input signal, while the Transformer component captures global relationships between various elements of the input sequence. By exploiting the strengths of both CNN and Transformer, this model has demonstrated superior accuracy in the identification of wireless interference compared to conventional approaches (Wang et al., 2021a).

- **MCformer:** MCformer, short for Modulation Classification Transformer, refers to a model architecture based on Transformers that performs feature extraction from input signals and subsequently classifies them based on modulation. This architectural design combines CNN and self-attention layers, enabling the processing of intricate features within the signal and achieving superior accuracy in comparison to conventional approaches. The introduction of this model has brought about noteworthy advancements in a wireless network and communication signals, particularly in the realm of Automatic Modulation Classification, thereby enhancing system security and performance (Hamidi-Rad & Jain, 2021).
- **Quan-Transformer:** It refers to a Transformer-based model specifically designed to perform quantization in wireless communication systems. Quantization is the vital process of converting a continuous signal into a discrete signal, and it plays a crucial role in network channel feedback processing. Channel feedback processing is essential for estimating channel state information, which in turn aids in adjusting signal transmission parameters. This feedback mechanism holds particular significance in the context of Reconfigurable Intelligent Surface (RIS)-aided wireless networks, a critical component of the 6th-generation communication system (Xie et al., 2022).

6.5.2 MEDICAL SIGNAL PROCESSING

The rise of healthcare data has resulted in the rapid growth of deep learning applications, enabling the automatic detection of pathologies, enhanced medical diagnosis, and improved healthcare services. These data can be categorized into three distinct forms: relational data (symptoms, examinations, and laboratory tests), medical images, and biomedical signals (consisting of raw electronic and sound signals). While the application of deep learning models, particularly Transformers, in the context of medical images has gained considerable attention and yielded promising results, the application of Transformers to biomedical signals is still in its early stages. The majority of relevant studies have been published between the years 2021 and 2022, with a particular focus on the task of signal classification. We have summarized our findings regarding the application of Transformers to biomedical signals in Table 27.

- **Epilepsy disease case:**
  Epilepsy is a serious and debilitating condition for those it affects. Typically, its symptoms are detected by analyzing electrical signals, such as electroencephalograms (EEGs) and magnetoencephalograms (MEGs). In recent years, deep learning has become a powerful tool for detecting and predicting epilepsy occurrences. Among the models proposed to assess electrical signals to predict and categorize epilepsy cases is the Transformer model. The application of Transformer is still in its early stages, which accounts for the limited number of models found in the literature. The two recently published models are as follows:

  - **Three-tower Transformer network (Yan et al., 2022a):** This model has been specifically designed to forecast epileptic seizures by analyzing EEG signals. It employs a Transformer architecture to conduct binary classification on EEG signals, considering three crucial EEG characteristics: time, frequency, and channel. This model processes the entire EEG signal in a unified manner, utilizing a Transformer model composed of three encoders: a time encoder, a frequency encoder, and a channel encoder. Remarkably, this model demonstrates superior performance in comparison to alternative approaches such as CNN, all while eliminating the need for manually crafted features. However, the author acknowledges the potential for further improvements in both the model's architecture and the labeling of the training samples.
  - **TransHFO (Guo et al., 2022c):** The (Transformer-based HFO) model is a deep learning approach that leverages the BERT architecture to automatically detect High-Frequency Oscillation (HFO) patterns within magnetoencephalography (MEG) data, aiding in the identification of epileptic regions. The model's classification performance is evaluated using k-fold cross-validation, and it demonstrates high performance in terms of accuracy. Given the limited size of the dataset, the authors suggest employing the data augmentation technique known as "ADASYN" to address this constraint. However, even with augmented data, the dataset remains relatively small, which leads to the observation that a shallow Transformer with fewer layers is more efficient than a deeper one. Nonetheless, it is noted that increasing the number of layers may actually result in reduced performance. The paper points out certain limitations, such as the constrained duration of signals, which may lead to the loss of significant information. Additionally, the segmentation approach utilized in the study is considered suboptimal because it treats segments as independent units, failing to account for their interrelationships, potentially decreasing the model's effectiveness.

Transformer Name	Field of Application	Data Augmentation	Fully Transformer Architecture	Signal Type	Transformer Task	Dataset
Three-tower Transformer network (Yan et al., 2022a)	Epilepsy	No	Yes	EEG	Binary classification	CHB-MIT dataset (Shoeb, 2009)
TransHFO (Guo et al., 2022c)	Epilepsy	Yes	Yes	MEG	Binary classification	MEG data collected from 20 clinical patients
TCN and Transformer-based model (Casal et al., 2022)	Sleep pathologies	No	No (using TCN which is based on CNN)	Cardiac signals (Heart Rate)	Binary classification Multi-classification	Sleep Heart Health Study dataset (Redline et al., 1998)
Constrained Transformer network (Che et al., 2021)	Heart disease	No	No (Using CNN)	ECG	Multi-classification	Data collected from 6877 patients (Liu et al., 2018)
CRT-NET (Liu et al., 2022a)	Heart Disease	Yes	No (Using CNN and Bi-directional GRU)	ECG	Binary classification Multi-classification	MIT-BIH (Mark & Moody, 1997), CPSC arrhythmia dataset (Liu et al., 2018), Private clinical data
CAT (Yang et al., 2022)	Atrial Fibrillation	Yes	No (Using MLP)	ECG	Binary classification	The Lobachevsky University electrocardiography database (LUDB) (Kalyakulina et al., 2020), Shaoxing database (Zheng et al., 2020a)

Table 27: Transformer Models for Medical Signal Processing

- **Cardiac diseases cases:** heart diseases are among the areas in which researchers are interested in applying Transformers. Using ECG signals, a Transformer model can detect long-range dependencies and identify heart disease types based on their characteristics.

  - **Constrained Transformer network (Che et al., 2021):** This model blends the CNN and Transformer architectures to perform multi-class classification of heart arrhythmia diseases using temporal information extracted from ECG (electrocardiogram) signals. Additionally, it incorporates a link constraint module to mitigate the impact of data imbalance by extracting relevant features and ensuring that embedding vectors exhibit high quality by identifying similarities among them, thereby potentially boosting the model's performance. In a comparative analysis against many state-of-the-art deep-learning architectures, this model outperforms all of them, achieving a notably high F1 score.

  - **CRT-NET (Liu et al., 2022a):** is a model capable of tackling both binary and multi-class classification tasks, with its primary objective being the identification of cardiovascular diseases through the analysis of annotated ECG data. It also presents an approach for extracting ECG signals from 2D ECG images. CRT-NET is composed of three main elements: convolutional blocks, a bidirectional GRU (Gated Recurrent Unit) module, and four Transformer encoders.
    The VGG-Net architecture is employed to capture the ECG data's morphological characteristics. To mitigate the risk of overfitting, a concern amplified by small training dataset constraints, the bidirectional GRU module is employed to capture temporal features. Following this, Transformer encoders are used to consolidate the extracted time-domain features and construct a unified representation for subsequent classification.

To further enhance its performance, the model incorporates data augmentation techniques. The model's performance was thoroughly evaluated on three distinct datasets, employing different convolutional blocks. Across all these datasets, CRT-NET demonstrated strong performance.
CRT-NET showcases its effectiveness not only in the accurate recognition and differentiation of various cardiovascular diseases but also in its capability to perform well in the identification of other medical conditions such as chronic kidney disease (CKD) and Type-2 Diabetes (T2DM).

- **CAT (Yang et al., 2022):** a Component-Aware Transformer (CAT) model is specifically designed to automate the binary classification of Atrial Fibrillation (AF) using ECG signals. The classification process begins with ECG signal segmentation using the U-NET network (Moskalenko et al., 2020). Subsequently, Transformer encoders are employed to capture both spatial and temporal features, enabling the MLP head to distinguish between cases of AF and non-AF. The CAT model outperforms other models in classifying AF due to its use of the attention mechanism for capturing extended dependencies and its access to a sizable dataset. However, the authors have identified potential areas for future enhancements. They propose a reevaluation of the tokenization algorithm to prevent data loss and suggest revisiting the segmentation mechanism to encompass various aspects of ECG signals.

- **TCN and Transformer-based model (Casal et al., 2022):** An automatic sleep stage classification system based on 1-dimensional cardiac signals (Heart Rate). The classification is conducted in two steps: extracting features from signals using Temporal Convolution Network (TCN) (Bai et al., 2018), and modeling signal sequence dependencies using the standard Transformer architecture consisting of two stacks of encoders and a simplified decoder module. Based on a dataset of 5000 different participants, this study demonstrated that this new model outperforms other networks, such as CNN and RNN, which consume more memory and reduce process efficiency.

Discussion:
Wireless network and signal processing tasks have recently seen the emergence of Transformer models as powerful tools. In our analysis of these models for wireless signal processing, we identified several key factors for comparison. The models exhibit variations in their attention layers, encoder/decoder module usage, and input data types. Notably, all models employ Multi-head Self Attention (MSA), except for ACNNT (Wang et al., 2021a), which combines channel attention with MSA. Additionally, SigT (Ren et al., 2022), ACNNT (Wang et al., 2021a), and McFormer (Hamidi-Rad & Jain, 2021) exclusively utilize the encoder module, while TSDN (Liu et al., 2022b) and Quan-Transformer (Xie et al., 2022) incorporate both encoder and decoder modules. Furthermore, the choice of input data type varies: SigT and Quan-Transformer work with normal signal data, TSDN with noisy signal data, ACNNT with interference signal data, and McFormer with radio signal data. When selecting among these models for a specific wireless network signal processing task, it's essential to consider their specializations. For tasks like signal decomposition, SigT (Ren et al., 2022) excels, making it suitable for time-series data analysis and signal decomposition. TSDN (Liu et al., 2022b) is specialized in signal decomposition, particularly valuable for separating mixed signals in wireless communication. ACNNT (Wang et al., 2021a) is tailored for radar applications, offering accurate detection and localization of targets within radar signals. If you're working with multi-carrier modulation systems, MCformer (Hamidi-Rad & Jain, 2021) is designed to assist in signal analysis and modulation-related tasks. Finally, when quantization is a critical aspect of signal processing, as in reducing signal precision while maintaining quality, Quan-Transformer (Xie et al., 2022) can be a valuable choice for optimizing quantization-related tasks.

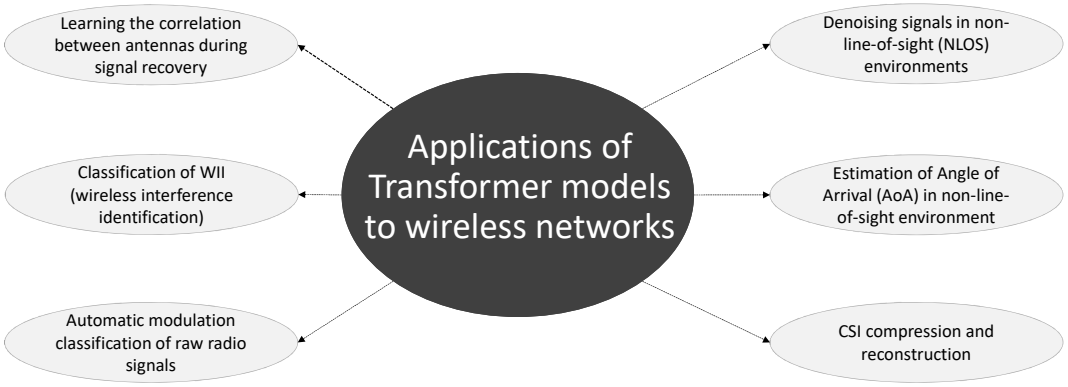

Figure 17: Applications of Transformer Models to Wireless Networks

Medical signal processing is an emerging field for applying Transformer models. We compared these models based on their classification approaches. Three-tower Transformer (Yan et al., 2022a), CAT (Yang et al., 2022), and TransHFO (Guo et al., 2022c) employ binary classification for signal processing. In contrast, the Constrained Transformer (Che et al., 2021) is the

only model that exclusively utilizes multi-class classification. TCN and Transformer-based models (Casal et al., 2022) and CRT-NeT (Liu et al., 2022a) incorporate both binary and multi-class classification for signal processing. Moreover, these models are designed with specific medical use cases in mind. However, it's crucial to align the choice of model with the specific requirements and characteristics of your medical signal processing task. Consider factors such as the type of signals, task complexity, available computational resources, and the need for domain-specific features.
When dealing with complex, multi-modal medical signal processing tasks that require effective integration of multiple modalities or features, the Three-tower Transformer (Yan et al., 2022a) proves valuable. For tasks related to detecting high-frequency oscillations in neurophysiological data, TransHFO (Guo et al., 2022c) excels, especially in capturing subtle patterns and features in signal data. TCN (Casal et al., 2022) is a versatile choice for a wide range of medical signal processing tasks that demand both temporal and spatial processing capabilities. Furthermore, when your medical signal processing task involves specific constraints or prior knowledge that can enhance signal quality or feature extraction, the Constrained Transformer (Che et al., 2021) is well-suited. In scenarios where real-world constraints significantly impact signal quality, CRT-NeT (Liu et al., 2022a), tailored to these constraints, becomes essential for effective signal processing

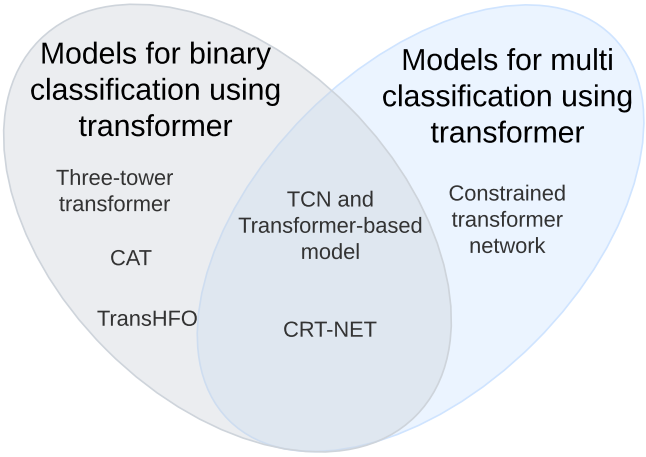

Figure 18: Comparative Analysis of Transformer Models for Medical Signal Classification

7 Future Prospects and Challenges

One of the primary objectives of this survey is to identify and highlight potential research directions for Transformer applications, with the goal of expanding their range of applications beyond the currently popular fields of NLP and CV. Despite considerable research attention in these areas, there are still a number of areas that remain relatively unexplored with the potential for significant improvements in the future. In order to expand the application areas of Transformers, we have identified several potential directions for future research. These directions include but are not limited to the exploration of Transformer-based models for speech recognition, recommendation systems, and natural language generation. In addition, further exploration of Transformer-based approaches for multimodal tasks, such as combining audio and visual inputs, would be an interesting direction for future research. By pursuing these research directions, we hope to continue the advancement of Transformer-based models and their utility in a broader range of applications.

7.1 Transformers in Wireless Network and Cloud Computing

While the majority of Transformer applications have been in NLP and CV, there is an exciting potential for Transformers in the wireless communication and cloud computing domains. Although there have been relatively fewer studies in this area, the ones that exist have demonstrated the enormous potential of Transformers for improving various aspects of wireless signal communication and cloud workload computing. In this section, we discuss some of the Transformer models that have been developed for wireless signal communication and cloud computing. These models have shown promising results in areas such as wireless interference recognition, wireless signal communication mitigation, and cloud workload forecasting. Moving forward, there are several potential directions for future research in both the wireless network and cloud domains. Some of the possible areas of focus for wireless communication include improving network security, enhancing the efficiency of wireless communication, and developing more accurate interference recognition models. On the other hand, for cloud computing, future work could focus on improving resource allocation and workload management, optimizing cloud performance, and enhancing data privacy and security.

Scope of future work for the wireless Signal communication:

- Detection of Wireless Interference: The global attention capabilities of Transformers present an exciting avenue for future research in the field of wireless signal communication. By leveraging the power of Transformers, researchers

can explore and develop more widely used applications for detecting wireless interference in communication systems. This can involve experimenting with various successive Transformer models to reduce complexity and enhance the efficiency of wireless interference recognition. This research can lead to improved communication systems that are more resilient to interference and provide better overall performance.

- Enhancing 5G & 6G Networks: As 5G and 6G networks gain popularity, there is significant potential for Transformers to contribute to this field. Advanced networking architectures, such as Reconfigurable Intelligent Surfaces (RIS) and Multiple-Output and Orthogonal Frequency-Division Multiplexing (MIMO-OFDM), play a crucial role in these networks. Transformers have shown promise in improving performance in these areas. Additionally, signal state feedback, which is essential for adjusting and updating networks based on signal state changes, can benefit from the parallel computational capability of Transformers. The ability of Transformers to handle multiple tasks simultaneously, including signal detection, channel estimation, and data decoding, makes them an effective alternative to conventional methods.
- Integration of Transformers with Advanced Communication Technologies: Transformers can be integrated with other advanced communication technologies to further improve wireless signal communication. For example, combining Transformers with technologies like Massive MIMO, millimeter-wave communication, and cognitive radio can enhance the performance, capacity, and spectrum efficiency of wireless networks. Future research can focus on exploring these synergies and developing innovative solutions that leverage the unique capabilities of Transformers in conjunction with other cutting-edge communication technologies.

Future Possibilities for Cloud Computing:

- Advancements in Cloud Computing: With the increasing application of the IoT, the cloud plays a crucial role in supporting and managing IoT devices. Transformers offer exciting possibilities for advancing cloud capabilities in various tasks, such as early attack and anomaly detection. By leveraging different Transformer approaches, the cloud can learn and adapt to its behavior, bringing more stability and security. Additionally, Transformers can be applied to cloud computing tasks like task scheduling and memory allocation. The MHA and long-range attention features of the Transformers model make it well-suited for optimizing resource allocation and improving overall performance in cloud environments.
- Transformation in Mobile Edge Computing (MEC) and Mobile Edge Caching (MEC): In the context of advanced 6G networking systems, Mobile Edge Computing (MEC) and Mobile Edge Caching (MEC) play vital roles in reducing communication latency. Transformers have demonstrated significant potential in enhancing MEC and MEC through their parallel computational capabilities. Transformers can be applied to predict popular content, improve content management, optimize resource allocation, and enhance data transmission in MEC systems. By leveraging Transformers, the mobile cloud can respond and process user requests faster, resulting in reduced network response times and faster data transmission.
- Intelligent Resource Management in the Cloud: Transformers offer opportunities for intelligent resource management in cloud environments. By applying Transformers to tasks like workload prediction, resource allocation, and load balancing, cloud systems can optimize resource utilization and enhance performance. Transformers' ability to capture long-range dependencies and handle complex patterns makes them well-suited for efficiently managing cloud resources and improving overall system efficiency.
- Security and Privacy in the Cloud: Transformers can contribute to enhancing security and privacy in the cloud by enabling advanced threat detection, anomaly detection, and data privacy protection mechanisms. Transformers can analyze large volumes of data, identify patterns, and detect potential security breaches or anomalies in real-time. Additionally, Transformers can be utilized for data anonymization and privacy-preserving computations, ensuring that sensitive information remains protected in cloud-based systems.

7.2 Medical Image & Signal Processing

Two types of medical data are discussed in this paper: images and signals. According to our literature review, segmentation, and classification are the most Transformer-based medical applications, followed by image translation (Yan et al., 2022b, Chen et al., 2022c). In the context of medical images, we commonly see the reuse of existing Transformers, such as BERT, ViT, and SWIN, regardless of how the original model was modified. Further, we observe that various types of medical images are used to conduct Transformer-based medical applications, such as 2D images (He et al., 2022, Gu et al., 2022), 3D images (Jiang et al., 2022b, Liang et al., 2022, Zhou et al., 2021a, Zhu et al., 2022) and multi-mode images (Sun et al., 2021b).
The selected research papers for this survey span the years 2021 to 2022, indicating that the use of Transformer architecture in the medical field is still in its nascent stages. Despite its early adoption, there has been a remarkable influx of excellent publications exploring Transformer applications in the analysis of medical images within this relatively short period. However, several challenges persist in applying Transformers to medical images that need to be addressed and overcome:

- **Limited focus on 3D images**: There is a scarcity of studies that specifically address the application of Transformers to 3D medical images. Most research has been concentrated on 2D images, indicating the need for further exploration and development in this area.
- **Small and private medical image databases**: Medical image databases are often small and privately owned due to legal and ethical concerns regarding patient data privacy (López-Linares et al., 2020). This limits the availability of large-scale datasets necessary for training Transformer models effectively.
- **Computational complexity in high-resolution imaging**: Transformer-based architectures encounter computational challenges when dealing with high-resolution medical images. The self-attention mechanism, which is integral to Transformers, becomes computationally demanding for large images. However, some models, like DI-UNET (Wu et al., 2022), have introduced enhanced self-attention mechanisms to handle higher resolution images effectively.
- **Limited number of fully developed Transformer-based models**: The development of Transformer-based models for processing medical images is still relatively nascent. Due to the computational complexity and parameter requirements of Transformers, existing architectures often combine deep learning techniques like CNNs and GANs with Transformers (Ma et al., 2022). Knowledge distillation techniques may offer a viable solution for training Transformer models with limited computational and storage resources (Leng et al., 2022).

Moreover, the application of Transformers to bio-signals is relatively limited compared to medical images. There are two main challenges that Transformers face in the domain of biomedical signals:

- **Small bio-signal databases**: Bio-signal databases often have limited sizes, which poses challenges for training and validating Transformer models effectively. For instance, in a study mentioned by (Guo et al., 2022c), only 20 patients were included, which is considered insufficient to establish the effectiveness of a model. To mitigate the limitations of small databases, some studies have proposed the use of virtual sample generation techniques like ADASYN (He et al., 2008) to augment the dataset.
- **Limited availability of Transformer-based models**: Currently, there is a scarcity of models that are exclusively based on Transformers for processing biomedical signals. The application of Transformers in this context is still relatively unexplored, and more research is needed to develop dedicated Transformer architectures for bio-signal analysis and processing.

In the context of this study, our primary focus was on medical signals and medical imagery. However, it's essential to acknowledge that within the medical field, Electronic Health Records (EHR) data plays a significant role. EHR data comprises a digital record of a patient's medical history and is typically private, not stored centrally. A notable research paper, referenced as (Shoham & Rappoport, 2023), delves into the application of Transformers for predicting patient visits using EHR data. They employ the BEHRT model, utilizing a federated learning approach, and simulate the use of the publicly accessible MIMIC-IV dataset as a representation of multi-central data. This effectively addresses the challenge of centralizing data in healthcare.

7.3 Reinforcement Learning

The integration of Transformers with deep reinforcement learning (RL) methods has emerged as a promising approach for enhancing sequential decision-making processes. Within this domain, two main research categories can be identified: architecture enhancement and trajectory optimization. In the architecture enhancement category, Transformers are applied to RL problems based on traditional RL paradigms. This involves leveraging the capabilities of Transformers to improve the representation and processing of RL states and actions. On the other hand, the trajectory optimization approach treats RL problems as sequence modeling tasks. It involves training a joint state-action model over entire trajectories, utilizing Transformers to learn policies from static datasets, and leveraging the Transformers' ability to model long sequences.

Deep RL heavily relies on interactions with the environment to collect data dynamically (Rjoub et al., 2019; 2021). However, in certain scenarios such as expensive environments like robotic applications or autonomous vehicles, collecting sufficient training data through real-time interaction may be challenging. To address this, offline RL techniques have been developed, which leverage deep networks to learn optimal policies from static datasets without direct environment interaction. In deep RL settings, Transformers are often used to replace traditional components like CNNs or LSTM networks (Rjoub et al., 2022), providing memory-awareness and improved modeling capabilities to the agent network. However, standard Transformer structures applied directly to decision-making tasks may suffer from stability issues. To overcome this limitation, researchers have proposed modified Transformer architectures, such as GtrX (Parisotto et al., 2020), as an alternative solution.

In summary, approaches like Decision Transformer and Trajectory Transformer have addressed RL problems as sequence modeling tasks, harnessing the power of Transformer architectures to model sequential trajectories (Chen et al., 2021, Janner et al., 2021). While these methods show promise in RL tasks, there is still significant room for improvement. Treating RL as sequence modeling simplifies certain limitations of traditional RL algorithms but may also overlook their advantages. Therefore, an interesting direction for further exploration is the integration of traditional RL algorithms with sequence modeling using Transformers, combining the strengths of both approaches.

7.4 Other Prospects

The successful application of Transformers in the field of NLP has sparked interest and exploration in various other domains. Researchers have been inspired to apply Transformer models to diverse areas, leading to promising developments. For instance, the Transformer model BERT has been utilized to model proteins, which, similar to natural language, can be considered as sequential data (Vig et al., 2021). Additionally, the Transformer model GPT-2 has been employed to automatically fix JavaScript software bugs and generate patches without human intervention (Lajkó et al., 2022).

Beyond its impact in traditional machine learning and deep learning domains, Transformers have found applications in industrial studies as well. They have demonstrated impressive performance in various tasks, ranging from predicting the state-of-charge of lithium batteries (Shen et al., 2022a) to classifying vibration signals in mechanical structures (Jin & Chen, 2021). Notably, Transformers have showcased superior capabilities compared to Graph Neural Networks (GNNs) in constructing meta-paths from different types of edges in heterogeneous graphs (Yun et al., 2019). This highlights the potential of Transformers in handling complex and diverse data structures.

In our review, we have demonstrated how Transformers are utilized for object detection in 2D images. However, it is crucial to recognize that the application of Transformers extends beyond the realm of 2D, encompassing the domain of 3D object recognition using point cloud data. An exemplary academic publication that showcases remarkable performance in terms of both robustness and efficiency is (Guan et al., 2022). This publication introduces a structural integration of Transformer encoders, referred to as the M3 Transformer, designed to comprehend interactions among features at both inter- and intra-levels. Subsequently, the model is integrated with a PV-RCNN-based approach to enhance the precision of 3D object detection, localization, and orientation estimation.

Another intriguing future application of Transformers lies in the field of Generative Art, where intelligent systems are leveraged for automated artistic creation, including images, music, and poetry. While image generation is a well-explored application area for Transformers, often focused on natural or medical images, the domain of artistic image generation remains relatively unexplored. However, there have been some initial models based on Transformers, such as AffectGAN, which generates images based on semantic text and emotional expressions using Transformer models (Galanos et al., 2021). The exploration of Transformers in generative art has significant untapped potential for further advancements and creative outputs.

Overall, the application of Transformers extends beyond NLP and showcases immense potential in various domains, ranging from scientific research to industrial applications and artistic creativity. Continued exploration and innovation in these areas will further expand the possibilities and impact of Transformers in the future.

8 Conclusion

The Transformer, as a DNN, has demonstrated superior performance compared to traditional recurrence-based models in processing sequential data. Its ability to capture long-term dependencies and leverage parallel computation has made it a dominant force in various fields such as NLP, CV, and more. Despite the presence of numerous survey papers that have explored the Transformer model's impacts in distinct domains, its architectural variances, and performance assessments, there remains a notable gap in the literature for a comprehensive survey paper that encompasses its applications across diverse domains. In this survey, we conducted a comprehensive overview of Transformer models' applications in different deep learning tasks and proposed a new taxonomy based on the top five fields and respective tasks: NLP, CV, Multi-Modality, Audio & Speech, and Signal Processing. By examining the advancements in each field, we provided insights into the current research focus and progress of Transformer models. This survey serves as a valuable reference for researchers seeking a deeper understanding of Transformer applications and aims to inspire further exploration of Transformers across various tasks. Additionally, we plan to extend our investigation to emerging fields like wireless networks, cloud computing, reinforcement learning, and others, to uncover new possibilities for Transformer utilization. The rapid expansion of Transformer applications in diverse domains showcases its versatility and potential for continued growth. With ongoing advancements and novel use cases, Transformers are poised to shape the future of deep learning and contribute to advancements in fields beyond the traditional realms of NLP and CV.

References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of bert-based approaches. *Artif. Intell. Rev.*, *54*, 5789–5829.

Ahmed, S. A. A., Awais, M., & Kittler, J. (2021). Sit: Self-supervised vision transformer. *CoRR*, *abs/2104.03602*. URL: `https://arxiv.org/abs/2104.03602`. `arXiv:2104.03602`.

Akbari, H., Yuan, L., Qian, R., Chuang, W., Chang, S., Cui, Y., & Gong, B. (2021). VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, December 6-14, 2021, virtual* (pp. 24206–24221).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: visual question answering. In *IEEE International Conference on Computer Vision, ICCV, Santiago, Chile, December 7-13* (pp. 2425–2433). IEEE Computer Society.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2022). XLS-R: self-supervised cross-lingual speech representation learning at scale. In H. Ko, & J. H. L. Hansen (Eds.), *Interspeech, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September* (pp. 2278–2282). ISCA.

Baevski, A., Schneider, S., & Auli, M. (2020a). vq-wav2vec: Self-supervised learning of discrete speech representations. In *8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30*. OpenReview.net.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, December 6-12, virtual*.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio, & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL: http://arxiv.org/abs/1409.0473.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, *abs/1803.01271*. URL: http://arxiv.org/abs/1803.01271. arXiv:1803.01271.

Bao, H., Dong, L., Piao, S., & Wei, F. (2022). Beit: BERT pre-training of image transformers. In *The Tenth International Conference on Learning Representations, ICLR Virtual Event, April 25-29*. OpenReview.net.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015* (pp. 632–642). The Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/d15-1075. doi:10.18653/v1/d15-1075.

Brasoveanu, A. M. P., & Andonie, R. (2020). Visualizing transformers for NLP: A brief survey. In *24th International Conference on Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020* (pp. 270–279). IEEE.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS, December 6-12, virtual*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, & J. Frahm (Eds.), *Computer Vision - ECCV - 16th European Conference, Glasgow, UK, August 23-28, Proceedings, Part I* (pp. 213–229). Springer volume 12346 of *Lecture Notes in Computer Science*.

Casal, R., Persia, L. E. D., & Schlotthauer, G. (2022). Temporal convolutional networks and transformers for classifying the sleep stage in awake or asleep using pulse oximetry signals. *J. Comput. Sci.*, *59*, 101544.

Che, C., Zhang, P., Zhu, M., Qu, Y., & Jin, B. (2021). Constrained transformer network for ECG signal processing and arrhythmia classification. *BMC Medical Informatics Decis. Mak.*, *21*, 184.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. (2019a). This looks like that: Deep learning for interpretable image recognition. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 8928–8939).

Chen, J., Mao, Q., & Liu, D. (2020a). Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. In H. Meng, B. Xu, & T. F. Zheng (Eds.), *Interspeech, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October* (pp. 2642–2646). ISCA.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, December 6-14, virtual* (pp. 15084–15097).

Chen, M., Challita, U., Saad, W., Yin, C., & Debbah, M. (2019b). Artificial neural networks-based machine learning for wireless networks: A tutorial. *IEEE Commun. Surv. Tutorials*, *21*, 3039–3071.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020b). Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning, ICML, 13-18 July, Virtual Event* (pp. 1691–1703). PMLR volume 119 of *Proceedings of Machine Learning Research*.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022a). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, *16*, 1505–1518.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., & Yu, X. (2022b). Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Virtual and Singapore, 23-27 May* (pp. 6152–6156). IEEE.

Chen, X., Liu, Y., Yang, B., Zhu, J., Yuan, S., Xie, X., Liu, Y., Dai, J., & Men, K. (2022c). A more effective ct synthesizer using transformers for cone-beam ct-guided adaptive radiotherapy. *Frontiers in Oncology*, *12*.

Chen, Y., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020c). UNITER: universal image-text representation learning. In *Computer Vision - ECCV - 16th European Conference, Glasgow, UK, August 23-28* (pp. 104–120). Springer volume 12375 of *Lecture Notes in Computer Science*.

Chowdhary, K., & Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, (pp. 603–649).

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020a). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30*. OpenReview.net.

Clark, P., Tafjord, O., & Richardson, K. (2020b). Transformers as soft reasoners over language. In C. Bessiere (Ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI* (pp. 3882–3890). ijcai.org.

Clerckx, B., Huang, K., Varshney, L. R., Ulukus, S., & Alouini, M. (2021). Wireless power transfer for future networks: Signal processing, machine learning, computing, and sensing. *IEEE J. Sel. Top. Signal Process.*, *15*, 1060–1094.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, & P. Motlícek (Eds.), *Interspeech, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September* (pp. 2426–2430). ISCA.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, Vancouver, BC, Canada* (pp. 7057–7067).

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., & Sagun, L. (2021). Convit: Improving vision transformers with soft convolutional inductive biases. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML, 18-24 July, Virtual Event* (pp. 2286–2296). PMLR volume 139 of *Proceedings of Machine Learning Research*.

Deng, L., Hinton, G. E., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: an overview. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Vancouver, BC, Canada, May 26-31* (pp. 8599–8603). IEEE.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., & Tang, J. (2021). Cogview: Mastering text-to-image generation via transformers. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, December 6-14, virtual* (pp. 19822–19835).

Dong, L., Xu, S., & Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Calgary, AB, Canada, April 15-20* (pp. 5884–5888). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR, Virtual Event, Austria, May 3-7*. OpenReview.net.

Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics. URL: https://doi.org/10.18653/v1/w16-3210. doi:10.18653/v1/w16-3210.

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, *23*, 120:1–120:39. URL: http://jmlr.org/papers/v23/21-0998.html.

Fournier, Q., Caron, G. M., & Aloise, D. (2023). A practical survey on faster and lighter transformers. *ACM Computing Surveys*, *55*, 1–40.

Fukushima, K. (1987). A neural network model for the mechanism of selective attention in visual pattern recognition. *Syst. Comput. Jpn.*, *18*, 102–113. URL: https://doi.org/10.1002/scj.4690180110. doi:10.1002/scj.4690180110.

Galanos, T., Liapis, A., & Yannakakis, G. N. (2021). Affectgan: Affect-based generative art driven by semantics. In *9th International Conference on Affective Computing and Intelligent Interaction, ACII - Workshops and Demos, Nara, Japan, September 28 - Oct. 1* (pp. 1–7). IEEE.

Giles, C. L., Chen, D., Sun, G., Chen, H., Lee, Y., & Goudreau, M. W. (1995). Constructive learning of recurrent neural networks: limitations of recurrent cascade correlation and a simple solution. *IEEE Trans. Neural Networks*, *6*, 829–836.

Gong, Y., Chung, Y., & Glass, J. R. (2021). AST: audio spectrogram transformer. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, & P. Motlícek (Eds.), *Interspeech, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September* (pp. 571–575). ISCA.

Gould, S., Arfvidsson, J., Kaehler, A., Sapp, B., Messner, M., Bradski, G. R., Baumstarck, P., Chung, S., & Ng, A. Y. (2007). Peripheral-foveal vision for real-time object recognition and tracking in video. In M. M. Veloso (Ed.), *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007* (pp. 2115–2121). URL: http://ijcai.org/Proceedings/07/Papers/341.pdf.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*, 602–610.

Gruetzemacher, R., & Paradice, D. B. (2022). Deep transfer learning & beyond: Transformer language models in information systems research. *ACM Comput. Surv.*, *54*, 204:1–204:35.

Gu, H., Wang, H., Qin, P., & Wang, J. (2022). Chest l-transformer: local features with position attention for weakly supervised chest radiograph segmentation and classification. *Frontiers in Medicine*, (p. 1619).

Gu, Y., Li, X., Chen, S., Zhang, J., & Marsic, I. (2017). Speech intention classification with multimodal deep learning. In M. Mouhoub, & P. Langlais (Eds.), *Advances in Artificial Intelligence - 30th Canadian Conference on Artificial Intelligence, Canadian AI, Edmonton, AB, Canada, May 16-19, Proceedings* (pp. 260–271). volume 10233 of *Lecture Notes in Computer Science*.

Guan, T., Wang, J., Lan, S., Chandra, R., Wu, Z., Davis, L., & Manocha, D. (2022). M3DETR: multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022* (pp. 2293–2303). IEEE. URL: https://doi.org/10.1109/WACV51458.2022.00235. doi:10.1109/WACV51458.2022.00235.

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In H. Meng, B. Xu, & T. F. Zheng (Eds.), *Interspeech, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October* (pp. 5036–5040). ISCA.

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., & Xu, C. (2022a). CMT: convolutional neural networks meet vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022* (pp. 12165–12175). IEEE. URL: https://doi.org/10.1109/CVPR52688.2022.01186. doi:10.1109/CVPR52688.2022.01186.

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., & Xu, C. (2022b). CMT: convolutional neural networks meet vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, New Orleans, LA, USA, June 18-24* (pp. 12165–12175). IEEE.

Guo, J., Xiao, N., Li, H., He, L., Li, Q., Wu, T., He, X., Chen, P., Chen, D., Xiang, J. et al. (2022c). Transformer-based high-frequency oscillation signal detection on magnetoencephalography from epileptic patients. *Frontiers in Molecular Biosciences*, *9*.

Hahn, C., Schmitt, F., Kreber, J. U., Rabe, M. N., & Finkbeiner, B. (2021). Teaching temporal logics to neural networks. In *9th International Conference on Learning Representations, ICLR, Virtual Event, Austria, May 3-7*. OpenReview.net.

Hajiakhondi-Meybodi, Z., Mohammadi, A., Rahimian, E., Heidarian, S., Abouei, J., & Plataniotis, K. N. (2022). Tedge-caching: Transformer-based edge caching towards 6g networks. In *IEEE International Conference on Communications, ICC Seoul, Korea, May 16-20* (pp. 613–618). IEEE.

Hamidi-Rad, S., & Jain, S. (2021). Mcformer: A transformer based deep neural network for automatic modulation classification. In *IEEE Global Communications Conference, GLOBECOM, Madrid, Spain, December 7-11* (pp. 1–6). IEEE.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, *45*, 87–110.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, December 6-14, virtual* (pp. 15908–15919).

Haralick, R. M., & Shapiro, L. G. (1985). Image segmentation techniques. *Comput. Vis. Graph. Image Process.*, *29*, 100–132.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI, Hong Kong, China, June 1-6* (pp. 1322–1328). IEEE.

He, X., Tan, E., Bi, H., Zhang, X., Zhao, S., & Lei, B. (2022). Fully transformer network for skin lesion analysis. *Medical Image Anal.*, *77*, 102357.

Hénaff, O. J. (2020). Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event* (pp. 4182–4192). PMLR volume 119 of *Proceedings of Machine Learning Research*.

Hershey, J. R., Chen, Z., Roux, J. L., & Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016* (pp. 31–35). IEEE. URL: https://doi.org/10.1109/ICASSP.2016.7471631. doi:10.1109/ICASSP.2016.7471631.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*, 261–266.

Hirschman, L., & Gaizauskas, R. J. (2001). Natural language question answering: the view from here. *Nat. Lang. Eng.*, *7*, 275–300.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*, 1735–1780.

Hoos, H. H., & Stützle, T. (2000). Satlib: An online resource for research on sat. *Sat*, *2000*, 283–292.

Hoover, A., Kouznetsova, V., & Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, *19*, 203–210.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, *51*, 118:1–118:36.

Hsu, W., Bolte, B., Tsai, Y. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, *29*, 3451–3460.

Hu, R., Rohrbach, M., & Darrell, T. (2016). Segmentation from natural language expressions. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I* (pp. 108–124). Springer volume 9905 of *Lecture Notes in Computer Science*.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., & Wang, L. (2022). Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17980–17989).

Huang, P., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014). Deep learning for monaural speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Florence, Italy, May 4-9* (pp. 1562–1566). IEEE.

Huang, Z., Zeng, Z., Liu, B., Fu, D., & Fu, J. (2020). Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, *abs/2004.00849*. URL: https://arxiv.org/abs/2004.00849. arXiv:2004.00849.

Islam, M. R., Nahiduzzaman, M., Goni, M. O. F., Sayeed, A., Anower, M. S., Ahsan, M., & Haider, J. (2022). Explainable transformer-based deep learning model for the detection of malaria parasites from blood cell images. *Sensors*, *22*, 4358.

Janner, M., Li, Q., & Levine, S. (2021). Offline reinforcement learning as one big sequence modeling problem. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, December 6-14, virtual* (pp. 1273–1286).

Jauro, F., Chiroma, H., Gital, A. Y., Almutairi, M., Abdulhamid, S. M., & Abawajy, J. H. (2020). Deep learning architectures in emerging cloud computing architectures: Recent development, challenges and next research trend. *Appl. Soft Comput.*, *96*, 106582.

Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML, 18-24 July, Virtual Event* (pp. 4904–4916). PMLR volume 139 of *Proceedings of Machine Learning Research*.

Jiang, Y., Liang, J., Cheng, T., Lin, X., Zhang, Y., & Dong, J. (2022a). Mtpa_unet: Multi-scale transformer-position attention retinal vessel segmentation network joint transformer and CNN. *Sensors*, *22*, 4592.

Jiang, Y., Zhang, Y., Lin, X., Dong, J., Cheng, T., & Liang, J. (2022b). Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. *Brain Sciences*, *12*, 797.

Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., & Qu, R. (2019). A survey of deep learning-based object detection. *IEEE Access*, *7*, 128837–128868.

Jin, C., & Chen, X. (2021). An end-to-end framework combining time-frequency expert knowledge and modified transformer networks for vibration signal classification. *Expert Syst. Appl.*, *171*, 114570.

Jungiewicz, M., Jastrzébski, P., Wawryka, P., Przystalski, K., Sabatowski, K., & Bartuś, S. (2023). Vision transformer in stenosis detection of coronary arteries. *Expert Syst. Appl.*, *228*, 120234.

Kaliyar, R. K. (2020). A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of bert. *2020 10th International Conference on Cloud Computing, Data Science & Engineering*, .

Kalyakulina, A. I., Yusipov, I. I., Moskalenko, V. A., Nikolskiy, A. V., Kosonogov, K. A., Osipov, G. V., Zolotykh, N. Y., & Ivanchenko, M. V. (2020). LUDB: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, *8*, 186181–186190. URL: https://doi.org/10.1109/ACCESS.2020.3029211. doi:10.1109/ACCESS.2020.3029211.

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2022). AMMU: A survey of transformer-based biomedical pretrained language models. *J. Biomed. Informatics*, *126*, 103982. URL: https://doi.org/10.1016/j.jbi.2021.103982. doi:10.1016/j.jbi.2021.103982.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. *CoRR*, *abs/1909.05858*. URL: http://arxiv.org/abs/1909.05858. arXiv:1909.05858.

Khan, A., & Lee, B. (2023). DeepGene transformer: Transformer for the gene expression-based classification of cancer subtypes. *Expert Syst. Appl.*, *226*, 120047.

Khan, S. H., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Comput. Surv.*, *54*, 200:1–200:41.

Kim, B., Lee, J., Kang, J., Kim, E., & Kim, H. J. (2021a). HOTR: end-to-end human-object interaction detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, virtual, June 19-25* (pp. 74–83). Computer Vision Foundation / IEEE.

Kim, W., Son, B., & Kim, I. (2021b). Vilt: Vision-and-language transformer without convolution or region supervision. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML, 18-24 July, Virtual Event* (pp. 5583–5594). PMLR volume 139 of *Proceedings of Machine Learning Research*.

Kuhn, T. (2014). A survey and classification of controlled natural languages. *Comput. Linguistics*, *40*, 121–170.

Kurin, V., Godil, S., Whiteson, S., & Catanzaro, B. (2020). Can q-learning with graph networks learn a generalizable branching heuristic for a SAT solver? In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, December 6-12, virtual*.

Lajkó, M., Csuvik, V., & Vidács, L. (2022). Towards javascript program repair with generative pre-trained transformer (GPT-2). In *3rd IEEE/ACM International Workshop on Automated Program Repair, APR@ICSE, Pittsburgh, PA, USA, May 19* (pp. 61–68). IEEE.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020* (pp. 2479–2490). European Language Resources Association.

Leng, B., Leng, M., Ge, M., & Dong, W. (2022). Knowledge distillation-based deep learning classification network for peripheral blood leukocytes. *Biomed. Signal Process. Control.*, *75*, 103590.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (pp. 7871–7880). Association for Computational Linguistics.

Li, G., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2020a). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI, New York, NY, USA, February 7-12* (pp. 11336–11344). AAAI Press.

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023). Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, (p. 102762).

Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *International Conference on Machine Learning, ICML, 17-23 July, Baltimore, Maryland, USA* (pp. 12888–12900). PMLR volume 162 of *Proceedings of Machine Learning Research*.

Li, P., Fu, T., & Ma, W. (2020b). Why attention? analyze bilstm deficiency and its remedies in the case of NER. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, February 7-12* (pp. 8236–8244). AAAI Press.

Li, P., Li, J., Huang, Z., Li, T., Gao, C., Yiu, S., & Chen, K. (2017). Multi-key privacy-preserving deep learning in cloud computing. *Future Gener. Comput. Syst.*, *74*, 76–85.

Li, S., & Hoefler, T. (2021). Chimera: efficiently training large-scale neural networks with bidirectional pipelines. In B. R. de Supinski, M. W. Hall, & T. Gamblin (Eds.), *International Conference for High Performance Computing, Networking, Storage and Analysis, SC, St. Louis, Missouri, USA, November 14-19* (p. 27). ACM.

Liang, J., Yang, C., Zeng, M., & Wang, X. (2022). Transconver: transformer and convolution parallel network for developing automatic brain tumor segmentation in mri images. *Quantitative Imaging in Medicine and Surgery*, *12*, 2397.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132.

Liu, A. T., Yang, S., Chi, P., Hsu, P., & Lee, H. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Barcelona, Spain, May 4-8* (pp. 6419–6423). IEEE.

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z. et al. (2018). An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, *8*, 1368–1373.

Liu, J., Li, Z., Fan, X., Hu, X., Yan, J., Li, B., Xia, Q., Zhu, J., & Wu, Y. (2022a). Crt-net: A generalized and scalable framework for the computer-aided diagnosis of electrocardiogram signals. *Appl. Soft Comput.*, *128*, 109481.

Liu, J., Wang, T., Li, Y., Li, C., Wang, Y., & Shen, Y. (2022b). A transformer-based signal denoising network for aoa estimation in nlos environments. *IEEE Commun. Lett.*, *26*, 2336–2339.

Liu, M., Breuel, T. M., & Kautz, J. (2017). Unsupervised image-to-image translation networks. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, Long Beach, CA, USA* (pp. 700–708).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

Liu, Y., Qiao, L., Yin, D., Jiang, Z., Jiang, X., Jiang, D., & Ren, B. (2022c). OS-MSL: one stage multimodal sequential link framework for scene segmentation and classification. In J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, & L. Toni (Eds.), *MM: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14* (pp. 6269–6277). ACM.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision ICCV, Montreal, QC, Canada, October 10-17* (pp. 9992–10002). IEEE.

Livezey, J. A., Bouchard, K. E., & Chang, E. F. (2019). Deep learning as a tool for neural data analysis: Speech classification and cross-frequency coupling in human sensorimotor cortex. *PLoS Comput. Biol.*, *15*.

López-Linares, K., García Ocaña, M. I., Lete Urzelai, N., González Ballester, M. Á., & Macía Oliver, I. (2020). Medical image segmentation using deep learning. *Deep Learning in Healthcare: Paradigms and Applications*, (pp. 17–31).

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*, 823–870.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS, December 8-14, Vancouver, BC, Canada* (pp. 13–23).

Ma, M., Xu, Y., Song, L., & Liu, G. (2022). Symmetric transformer-based network for unsupervised image registration. *Knowl. Based Syst.*, *257*, 109959.

Mahesh, K. M., & Renjit, J. A. (2020). Deepjoint segmentation for the classification of severity-levels of glioma tumour using multimodal MRI images. *IET Image Process.*, *14*, 2541–2552.

Mark, R., & Moody, G. (1997). Mit-bih arrhythmia database. *See http://ecg. mit. edu/dbinfo. html*, .

Mark A Musen, J. V. d. L. (1988). Of brittleness and bottlenecks: Challenges in the creation of pattern-recognition and expert-system models. *In Machine Intelligence and Pattern Recognition,*, *7*, 335–352.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J. S., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E. R., Weber, M., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M. T., Sarikaya, D., Schwartz, L. H., Shin, H., Shotton, J., Silva, C. A., Sousa, N. J., Subbanna, N. K., Székely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Ünal, G. B., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., & Leemput, K. V. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Medical Imaging*, *34*, 1993–2024.

Meur, O. L., Callet, P. L., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, *28*, 802–817. URL: https://doi.org/10.1109/TPAMI.2006.86. doi:10.1109/TPAMI.2006.86.

Miau, F., & Itti, L. (2001). A neural model combining attentional orienting to object recognition: preliminary explorations on the interplay between where and what. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 789–792 vol.1). volume 1. doi:10.1109/IEMBS.2001.1019059.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010* (pp. 1045–1048). ISCA.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, *44*, 3523–3542.

Monroe, D. (2017). Deep learning takes on translation. *Commun. ACM*, *60*, 12–14.

Moskalenko, V., Zolotykh, N., & Osipov, G. (2020). Deep learning for ecg segmentation. In *Advances in Neural Computation, Machine Learning, and Cognitive Research III: Selected Papers from the XXI International Conference on Neuroinformatics, October 7-11, 2019, Dolgoprudny, Moscow Region, Russia* (pp. 246–254). Springer.

Murtagh, F. (1990). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*, 183–197.

Nassif, A. B., Shahin, I., Attili, I. B., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, *7*, 19143–19165.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *International Conference on Machine Learning, ICML, 17-23 July, Baltimore, Maryland, USA* (pp. 16784–16804). PMLR volume 162 of *Proceedings of Machine Learning Research*.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62.

van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., & Graves, A. (2016). Conditional image generation with pixelcnn decoders. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, December 5-10, Barcelona, Spain* (pp. 4790–4798).

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *CoRR*, *abs/1511.08458*. URL: http://arxiv.org/abs/1511.08458. arXiv:1511.08458.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *NeurIPS*. URL: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Owen, C. G., Rudnicka, A. R., Mullen, R., Barman, S. A., Monekosso, D., Whincup, P. H., Ng, J., & Paterson, C. (2009). Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. *Investigative ophthalmology & visual science*, *50*, 2004–2010.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015* (pp. 5206–5210). IEEE. URL: https://doi.org/10.1109/ICASSP.2015.7178964. doi:10.1109/ICASSP.2015.7178964.

Pang, Y., Lin, J., Qin, T., & Chen, Z. (2022). Image-to-image translation: Methods and applications. *IEEE Trans. Multim.*, *24*, 3859–3881.

Parisotto, E., Song, H. F., Rae, J. W., Pascanu, R., Gülçehre, Ç., Jayakumar, S. M., Jaderberg, M., Kaufman, R. L., Clark, A., Noury, S., Botvinick, M. M., Heess, N., & Hadsell, R. (2020). Stabilizing transformers for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML, 13-18 July, Virtual Event* (pp. 7487–7498). PMLR volume 119 of *Proceedings of Machine Learning Research*.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In J. G. Dy, & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML, Stockholmsmässan, Stockholm, Sweden, July 10-15* (pp. 4052–4061). PMLR volume 80 of *Proceedings of Machine Learning Research*.

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., & Ye, Q. (2021). Conformer: Local features coupling global representations for visual recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021* (pp. 357–366). IEEE.

Picco, G., Hoang, T. L., Sbodio, M. L., & López, V. (2021). Neural unification for logic reasoning over natural language. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP, Virtual Event / Punta Cana, Dominican Republic, 16-20 November* (pp. 3939–3950). Association for Computational Linguistics.

Pnueli, A. (1977). The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November* (pp. 46–57). IEEE Computer Society.

Polu, S., & Sutskever, I. (2020). Generative language modeling for automated theorem proving. *CoRR*, *abs/2009.03393*. URL: https://arxiv.org/abs/2009.03393. arXiv:2009.03393.

Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., & Zhou, M. (2020). Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November* (pp. 2401–2410). Association for Computational Linguistics volume EMNLP 2020 of *Findings of ACL*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML, 18-24 July, Virtual Event* (pp. 8748–8763). PMLR volume 139 of *Proceedings of Machine Learning Research*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *CoRR*, *abs/2212.04356*. URL: https://doi.org/10.48550/arXiv.2212.04356. doi:10.48550/arXiv.2212.04356. arXiv:2212.04356.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding with unsupervised learning*. Technical Report OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*, 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*, 140:1–140:67.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML, 18-24 July, Virtual Event* (pp. 8821–8831). PMLR volume 139 of *Proceedings of Machine Learning Research*.

Redline, S. H. H. R. G. S., cwru, edu Sanders Mark H, K, L. B., F, Q. S., J, I. C. G. D., H, B. W., M, R. D., L, S. P., & P, K. J. (1998). Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. *Sleep*, *21*, 759–767.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Nat. Lang. Eng.*, *3*, 57–87.

Ren, Q., Li, Y., & Liu, Y. (2023). Transformer-enhanced periodic temporal convolution network for long short-term traffic flow forecasting. *Expert Syst. Appl.*, *227*, 120203.

Ren, Z., Cheng, N., Sun, R., Wang, X., Lu, N., & Xu, W. (2022). Sigt: An efficient end-to-end MIMO-OFDM receiver framework based on transformer. In *5th International Conference on Communications, Signal Processing, and their Applications, ICCSPA, Cairo, Egypt, December 27-29* (pp. 1–6). IEEE.

Reza, S., Ferreira, M. C., Machado, J. J. M., & Tavares, J. M. R. S. (2022). A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert Syst. Appl.*, *202*, 117275.

Richardson, K., & Sabharwal, A. (2022). Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI, Virtual Event, February 22 - March 1* (pp. 11209–11219). AAAI Press.

Rjoub, G., Bentahar, J., Abdel Wahab, O., & Saleh Bataineh, A. (2021). Deep and reinforcement learning for automated task scheduling in large-scale cloud computing systems. *Concurrency and Computation: Practice and Experience*, *33*, e5919.

Rjoub, G., Bentahar, J., Wahab, O. A., & Bataineh, A. (2019). Deep smart scheduling: A deep learning approach for automated big data scheduling over the cloud. In *2019 7th International Conference on Future Internet of Things and Cloud (FiCloud)* (pp. 189–196). IEEE.

Rjoub, G., Wahab, O. A., Bentahar, J., & Bataineh, A. (2022). Trust-driven reinforcement selection strategy for federated learning on IoT devices. *Computing*, (pp. 1–23).

Ruan, L., & Jin, Q. (2022). Survey: Transformer based video-language pre-training. *AI Open*, *3*, 1–13.

Saha, S., Ghosh, S., Srivastava, S., & Bansal, M. (2020). Prover: Proof generation for interpretable reasoning over rules. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020* (pp. 122–136). Association for Computational Linguistics.

Salah, A., Alpaydin, E., & Akarun, L. (2002). A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 420–425. doi:10.1109/34.990146.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, *abs/1910.01108*. URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.

Schmidhuber, J., & Huber, R. (1991). Learning to generate artificial fovea trajectories for target detection. *Int. J. Neural Syst.*, *2*, 125–134. URL: https://doi.org/10.1142/S012906579100011X. doi:10.1142/S012906579100011X.

Selsam, D., Lamm, M., Bünz, B., Liang, P., de Moura, L., & Dill, D. L. (2019). Learning a SAT solver from single-bit supervision. In *7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6-9*. OpenReview.net.

Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., & Clapés, A. (2023). Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, .

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, (p. 102802).

Shen, H., Zhou, X., Wang, Z., & Wang, J. (2022a). State of charge estimation for lithium-ion battery using transformer with immersion and invariance adaptive observer. *Journal of Energy Storage*, *45*, 103768.

Shen, X., Wang, L., Zhao, Y., Liu, R., Qian, W., & Ma, H. (2022b). Dilated transformer: residual axial attention for breast ultrasound image segmentation. *Quantitative Imaging in Medicine and Surgery*, *12*, 4512.

Shi, C., Xiao, Y., & Chen, Z. (2022a). Dual-domain sparse-view ct reconstruction with transformers. *Physica Medica*, *101*, 1–7.

Shi, F., Lee, C., Bashar, M. K., Shukla, N., Zhu, S., & Narayanan, V. (2021). Transformer-based machine learning for fast SAT solvers and logic synthesis. *CoRR*, *abs/2107.07116*. URL: https://arxiv.org/abs/2107.07116. arXiv:2107.07116.

Shi, Z., Li, M., Khan, S., Zhen, H., Yuan, M., & Xu, Q. (2022b). Satformer: Transformers for SAT solving. *CoRR*, *abs/2209.00953*. URL: https://doi.org/10.48550/arXiv.2209.00953. doi:10.48550/arXiv.2209.00953. arXiv:2209.00953.

Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA, June 27-30* (pp. 4613–4621). IEEE Computer Society.

Shin, A., Ishii, M., & Narihira, T. (2022). Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision. *Int. J. Comput. Vis.*, *130*, 435–454.

Shoeb, A. H. (2009). *Application of machine learning to epileptic seizure onset detection and treatment*. Ph.D. thesis Massachusetts Institute of Technology.

Shoham, O. B., & Rappoport, N. (2023). Federated learning of medical concepts embedding using BEHRT. *CoRR*, *abs/2305.13052*. URL: https://doi.org/10.48550/arXiv.2305.13052. doi:10.48550/arXiv.2305.13052. arXiv:2305.13052.

Soydaner, D. (2022). Attention mechanism in neural networks: where it comes and where it goes. *Neural Comput. Appl.*, *34*, 13371–13385. URL: https://doi.org/10.1007/s00521-022-07366-3. doi:10.1007/s00521-022-07366-3.

Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., & van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Medical Imaging*, *23*, 501–509. URL: https://doi.org/10.1109/TMI.2004.825627. doi:10.1109/TMI.2004.825627.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2020). VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30*. OpenReview.net.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021). Attention is all you need in speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Toronto, ON, Canada, June 6-11* (pp. 21–25). IEEE.

Subramanyam, K., Rajasekharan, A., & Sangeetha, S. (2021). AMMUS : A survey of transformer-based pretrained models in natural language processing. *CoRR*, *abs/2108.05542*. URL: https://arxiv.org/abs/2108.05542. arXiv:2108.05542.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017a). Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017* (pp. 843–852). IEEE Computer Society. URL: https://doi.org/10.1109/ICCV.2017.97. doi:10.1109/ICCV.2017.97.

Sun, H., Chen, X., Shi, Q., Hong, M., Fu, X., & Sidiropoulos, N. D. (2017b). Learning to optimize: Training deep neural networks for wireless resource management. In *18th IEEE International Workshop on Signal Processing Advances in Wireless Communications, SPAWC, Sapporo, Japan, July 3-6* (pp. 1–6). IEEE.

Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021a). Loftr: Detector-free local feature matching with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR virtual, June 19-25* (pp. 8922–8931). Computer Vision Foundation / IEEE.

Sun, Q., Fang, N., Liu, Z., Zhao, L., Wen, Y., Lin, H. et al. (2021b). Hybridctrm: Bridging cnn and transformer for multimodal brain image segmentation. *Journal of Healthcare Engineering*, *2021*.

Suzuki, M., & Matsuo, Y. (2022). A survey of multimodal deep generative models. *Adv. Robotics*, *36*, 261–278.

Szummer, M., & Picard, R. W. (1998). Indoor-outdoor image classification. In *1998 International Workshop on Content-Based Access of Image and Video Databases, CAIVD 1998, Bombay, India, January 3, 1998* (pp. 42–51). IEEE Computer Society.

Tan, H., & Bansal, M. (2019). LXMERT: learning cross-modality encoder representations from transformers. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, Hong Kong, China, November 3-7* (pp. 5099–5110). Association for Computational Linguistics.

Tas, O., & Kiyani, F. (2007). A survey automatic text summarization. *PressAcademia Procedia*, *5*, 205–213.

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2023). Efficient transformers: A survey. *ACM Comput. Surv.*, *55*, 109:1–109:28.

Tay, Y., Tran, V. Q., Ruder, S., Gupta, J. P., Chung, H. W., Bahri, D., Qin, Z., Baumgartner, S., Yu, C., & Metzler, D. (2022). Charformer: Fast character transformers via gradient-based subword tokenization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29*. OpenReview.net.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML, 18-24 July, Virtual Event* (pp. 10347–10357). PMLR volume 139 of *Proceedings of Machine Learning Research*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, December 4-9, Long Beach, CA, USA* (pp. 5998–6008).

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., & Rajani, N. F. (2021). Bertology meets biology: Interpreting attention in protein language models. In *9th International Conference on Learning Representations, ICLR, Virtual Event, Austria, May 3-7*. OpenReview.net.

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., & Hinton, G. E. (2015). Grammar as a foreign language. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada* (pp. 2773–2781). URL: https://proceedings.neurips.cc/paper/2015/hash/277281aada22045c03945dcb2ca6f2ec-Abstract.html.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE ACM Trans. Audio Speech Lang. Process.*, *26*, 1702–1726.

Wang, G., Smetannikov, I., & Man, T. (2020a). Survey on automatic text summarization and transformer models applicability. In *CCRIS: International Conference on Control, Robotics and Intelligent System, Xiamen, China, October 27-29* (pp. 176–184). ACM.

Wang, H., Lu, P., Zhang, H., Yang, M., Bai, X., Xu, Y., He, M., Wang, Y., & Liu, W. (2020b). All you need is boundary: Toward arbitrary-shaped text spotting. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020* (pp. 12160–12167). AAAI Press. URL: https://doi.org/10.1609/aaai.v34i07.6896. doi:10.1609/aaai.v34i07.6896.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., & Wang, L. (2022a). GIT: A generative image-to-text transformer for vision and language. *Trans. Mach. Learn. Res.*, *2022*. URL: https://openreview.net/forum?id=b4tMhpN0JC.

Wang, P., Cheng, Y., & Dong, B. (2021a). Augmented convolutional neural networks with transformer for wireless interference identification. In *IEEE Global Communications Conference, GLOBECOM, Madrid, Spain, December 7-11* (pp. 1–6). IEEE.

Wang, S., Bi, S., & Zhang, Y.-J. A. (2022b). Deep reinforcement learning with communication transformer for adaptive live streaming in wireless edge networks. *IEEE Journal on Selected Areas in Communications*, *40*, 308–322.

Wang, T., Lai, Z., & Kong, H. (2021b). Tfnet: Transformer fusion network for ultrasound image segmentation. In C. Wallraven, Q. Liu, & H. Nagahara (Eds.), *Pattern Recognition - 6th Asian Conference, ACPR, Jeju Island, South Korea, November 9-12, Revised Selected Papers, Part I* (pp. 314–325). Springer volume 13188 of *Lecture Notes in Computer Science*.

Wang, T., Lan, J., Han, Z., Hu, Z., Huang, Y., Deng, Y., Zhang, H., Wang, J., Chen, M., Jiang, H. et al. (2022c). O-net: a novel framework with deep fusion of cnn and transformer for simultaneous segmentation and classification. *Frontiers in Neuroscience*, *16*.

Wang, W., Liang, D., Chen, Q., Iwamoto, Y., Han, X.-H., Zhang, Q., Hu, H., Lin, L., & Chen, Y.-W. (2020c). Medical image classification using deep learning. *Deep learning in healthcare: paradigms and applications*, (pp. 33–51).

Wang, Z., Ma, Y., Liu, Z., & Tang, J. (2019). R-transformer: Recurrent neural network enhanced transformer. *CoRR*, *abs/1907.05572*. URL: http://arxiv.org/abs/1907.05572. arXiv:1907.05572.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2022d). Simvlm: Simple visual language model pretraining with weak supervision. In *The Tenth International Conference on Learning Representations, ICLR, Virtual Event, April 25-29*. OpenReview.net.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In M. A. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)* (pp. 1112–1122). Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/n18-1101. doi:10.18653/v1/n18-1101.

Wu, D., Pigou, L., Kindermans, P., Le, N. D., Shao, L., Dambre, J., & Odobez, J. (2016). Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, *38*, 1583–1597.

Wu, Y., Wang, G., Wang, Z., Wang, H., & Li, Y. (2022). Di-unet: Dimensional interaction self-attention for medical image segmentation. *Biomed. Signal Process. Control.*, *78*, 103896.

Xian, M., Zhang, Y., Cheng, H.-D., Xu, F., Huang, K., Zhang, B., Ding, J., Ning, C., & Wang, Y. (2018). *A benchmark for breast ultrasound image segmentation (BUSIS)*. Infinite Study.

Xie, N., Lai, F., Doran, D., & Kadav, A. (2019). Visual entailment: A novel task for fine-grained image understanding. *CoRR*, *abs/1901.06706*. URL: http://arxiv.org/abs/1901.06706. arXiv:1901.06706.

Xie, W., Zou, J., Xiao, J., Li, M., & Peng, X. (2022). Quan-transformer based channel feedback for ris-aided wireless communication systems. *IEEE Commun. Lett.*, *26*, 2631–2635.

Xing, Y., Shi, Z., Meng, Z., Lakemeyer, G., Ma, Y., & Wattenhofer, R. (2021). KM-BART: knowledge enhanced multimodal BART for visual commonsense generation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP, (Volume 1: Long Papers), Virtual Event, August 1-6* (pp. 525–535). Association for Computational Linguistics.

Xu, Y., Wei, H., Lin, M., Deng, Y., Sheng, K., Zhang, M., Tang, F., Dong, W., Huang, F., & Xu, C. (2022). Transformers in computational visual media: A survey. *Computational Visual Media*, *8*, 33–62.

Xu, Y., & Zhao, J. (2022). Actor-critic with transformer for cloud computing resource three stage job scheduling. In *7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 22-24 April* (pp. 33–37).

Yan, J., Li, J., Xu, H., Yu, Y., & Xu, T. (2022a). Seizure prediction based on transformer using scalp electroencephalogram. *Applied Sciences*, *12*, 4158.

Yan, S., Wang, C., Chen, W., & Lyu, J. (2022b). Swin transformer-based GAN for multi-modal medical image translation. *Frontiers in Oncology*, *12*.

Yang, H., & Yang, D. (2023). Cswin-pnet: A cnn-swin transformer combined pyramid network for breast lesion segmentation in ultrasound images. *Expert Syst. Appl.*, *Volume 213, Part B*, 119024.

Yang, M., Lee, D., & Park, S. (2022). Automated diagnosis of atrial fibrillation using ECG component-aware transformer. *Comput. Biol. Medicine*, *150*, 106115.

Yeh, C., Mahadeokar, J., Kalgaonkar, K., Wang, Y., Le, D., Jain, M., Schubert, K., Fuegen, C., & Seltzer, M. L. (2019). Transformer-transducer: End-to-end speech recognition with self-attention. *CoRR*, *abs/1910.12977*. URL: http://arxiv.org/abs/1910.12977. arXiv:1910.12977.

Yolcu, E., & Póczos, B. (2019). Learning local search heuristics for boolean satisfiability. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 7990–8001). URL: https://proceedings.neurips.cc/paper/2019/hash/12e59a33dea1bf0630f46edfe13d6ea2-Abstract.html.

Yu, D., & Deng, L. (2016). *Automatic speech recognition* volume 1. Springer.

Yu, J., Li, J., Yu, Z., & Huang, Q. (2020). Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circuits Syst. Video Technol.*, *30*, 4467–4480.

Yu, S., Wang, X., & Langar, R. (2017). Computation offloading for mobile edge computing: A deep learning approach. In *28th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, PIMRC, Montreal, QC, Canada, October 8-13* (pp. 1–6). IEEE.

Yuan, L., Chen, D., Chen, Y., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., & Zhang, P. (2021). Florence: A new foundation model for computer vision. *CoRR*, *abs/2111.11432*. URL: https://arxiv.org/abs/2111.11432. arXiv:2111.11432.

Yun, S., Jeong, M., Kim, R., Kang, J., & Kim, H. J. (2019). Graph transformer networks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS, December 8-14, Vancouver, BC, Canada* (pp. 11960–11970).

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA, June 16-20* (pp. 6720–6731). Computer Vision Foundation / IEEE.

Zhang, C., Patras, P., & Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. *IEEE Commun. Surv. Tutorials*, *21*, 2224–2287.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020a). PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July, Virtual Event* (pp. 11328–11339). PMLR volume 119 of *Proceedings of Machine Learning Research*.

Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., & Kumar, S. (2020b). Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Barcelona, Spain, May 4-8* (pp. 7829–7833). IEEE.

Zhang, W., Yang, H., Samaras, D., & Zelinsky, G. J. (2005). A computational model of eye movements during object class detection. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]* (pp. 1609–1616). URL: https://proceedings.neurips.cc/paper/2005/hash/e07bceab69529b0f0b43625953fbf2a0-Abstract.html.

Zhao, Z., Zheng, P., Xu, S., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Trans. Neural Networks Learn. Syst.*, *30*, 3212–3232.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., & Rakovski, C. (2020a). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, *7*, 48.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., & Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, virtual, June 19-25* (pp. 6881–6890). Computer Vision Foundation / IEEE.

Zheng, Y., Li, X., Xie, F., & Lu, L. (2020b). Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020* (pp. 6734–6738). IEEE.

Zhou, H., Guo, J., Zhang, Y., Yu, L., Wang, L., & Yu, Y. (2021a). nnformer: Interleaved transformer for volumetric segmentation. *CoRR*, *abs/2109.03201*. URL: https://arxiv.org/abs/2109.03201. arXiv:2109.03201.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A. L., & Kong, T. (2021b). iBOT: Image BERT pre-training with online tokenizer. *CoRR*, *abs/2111.07832*. URL: https://arxiv.org/abs/2111.07832. arXiv:2111.07832.

Zhou, S., Li, J., Zhang, K., Wen, M., & Guan, Q. (2020). An accurate ensemble forecasting approach for highly dynamic cloud workload with VMD and r-transformer. *IEEE Access*, *8*, 115992–116003.

Zhu, X., Hu, H., Wang, H., Yao, J., Li, W., Ou, D., & Xu, D. (2022). Region aware transformer for automatic breast ultrasound tumor segmentation. In E. Konukoglu, B. H. Menze, A. Venkataraman, C. F. Baumgartner, Q. Dou, & S. Albarqouni (Eds.), *International Conference on Medical Imaging with Deep Learning, MIDL, 6-8 July, Zurich, Switzerland* (pp. 1523–1537). PMLR volume 172 of *Proceedings of Machine Learning Research*.

Zidan, U., Gaber, M. M., & Abdelsamea, M. M. (2023). Swincup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer. *Expert Syst. Appl.*, *216*, 119452.

A Datasets and Data Sources

Field of Application	Dataset	Dataset Link
Natural Language Processing (NLP)	WMT 2014	https://www.statmt.org/wmt14/index.html
	Wall Street Journal (WSJ) portion of the Penn Treebank	https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html
	BerkleyParser corpora	(Vinyals et al., 2015)
	XNLI	https://github.com/facebookresearch/XNLI
	WMT 2016	https://www.statmt.org/wmt16/index.html
	MultiUN	https://conferences.unite.un.org/UNCORPUS/
	IIT Bombay corpus	https://www.cfilt.iitb.ac.in/iitb_parallel/
	EUbookshop corpus	https://opus.nlpl.eu/EUbookshop.php
	OpenSubtitles 2018	https://opus.nlpl.eu/OpenSubtitles-v2018.php
	GlobalVoices	https://paperswithcode.com/dataset/global-voices
	Stanford Sentiment Treebank-2 (SST-2)	https://nlp.stanford.edu/sentiment/
	1B Word Benchmark	https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark
	The Corpus of Linguistic Acceptability (CoLA)	https://nyu-mll.github.io/CoLA/
	BookCorpus dataset	https://yknzhu.wixsite.com/mbweb https://github.com/soskek/bookcorpus
	Colossal Clean Crawled Corpus (C4)	https://github.com/google-research/text-to-text-transfer-Transformer#c4
	AGNews	http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
	The Stanford Question Answering Dataset (SQuAD)	https://rajpurkar.github.io/SQuAD-explorer/
	TriviaQA	http://nlp.cs.washington.edu/triviaqa/
	Natural Questions	https://ai.google.com/research/NaturalQuestions
	Web Questions	https://nlp.stanford.edu/software/sempre/
	ARC Reasoning Challenge	http://data.allenai.org/arc
	CoQA	https://stanfordnlp.github.io/coqa/
	DROP	https://allenai.org/data/drop

Continued on next page

Table 28 – Continued from previous page

Field of Application	Dataset	Dataset Link
	RACE	https://www.cs.cmu.edu/∼glai1/data/race/
	Story Cloze	https://cs.rochester.edu/nlp/rocstories/
	GLUE benchmark	https://gluebenchmark.com/
	SuperGlue benchmark	https://super.gluebenchmark.com/
	English Wikipedia	https://dumps.wikimedia.org/
	ClueWeb	https://lemurproject.org/clueweb09.php/
	CommonCrawl	https://catalog.ldc.upenn.edu/LDC2011T07
	Gigaword	https://catalog.ldc.upenn.edu/LDC2011T07
	CNN/DailyMail summarization	https://github.com/google-deepmind/rc-data
	DU0-DU5	https://allenai.org/data/ruletaker
	Birds	https://allenai.org/data/ruletaker
	Electricity	https://allenai.org/data/ruletaker
	ParaRules	https://allenai.org/data/ruletaker
	Mix of Github, arXiv and Math StackExchange	Private dataset
	SATLIB benchmark library	(Hoos & Stützle, 2000)
	Circuit datasets	Private dataset
	LTLRandom35, LTLRandom50, LTLPattern126, LTLUnsolved254, PropRandom35, PropRandom50	Private dataset
	Metamath's set.mm	https://github.com/metamath/set.mm
	XSum	https://github.com/EdinburghNLP/XSum/tree/master /XSum-Dataset
	NEWSROOM	https://lil.nlp.cornell.edu/newsroom/
	Multi-News	https://github.com/Alex-Fabbri/Multi-News
	Gigaword	https://github.com/harvardnlp/sent-summary https://github.com/microsoft/unilm/
	arXiv	https://github.com/armancohan/long-summarization
	PubMed	https://github.com/armancohan/long-summarization
	BIGPATENT	https://evasharma.github.io/bigpatent/

Continued on next page

Table 28 – Continued from previous page

Field of Application	Dataset	Dataset Link
	WikiHow	https://github.com/mahnazkoupaee/WikiHow-Dataset
	Reddit TIFU	https://github.com/ctr4si/MMN
	AESLC	https://github.com/ryanzhumich/AESLC
	BillSum	https://github.com/FiscalNote/BillSum
	HugeNews	https://github.com/google-research/pegasus
Computer Vision (Natural Images)	JFT-300M	(Sun et al., 2017a)
	ILSVRC-2012 ImageNet	https://image-net.org/index.php
	ImageNet-21k	https://github.com/Alibaba-MIIL/ImageNet21K
	ImageNet-22k	https://github.com/microsoft/Swin-Transformer
	CIFAR-10/100	https://www.cs.toronto.edu/ kriz/cifar.html
	Oxford Flowers-102	https://www.robots.ox.ac.uk/ vgg/data/flowers/102/
	Oxford-IIIT Pets	https://www.robots.ox.ac.uk/ vgg/data/pets/
	VTAB benchmark	github.com/google-research/task_adaptation
	STL-10	https://cs.stanford.edu/ acoates/stl10/
	CUB200	http://www.vision.caltech.edu/datasets/
	Pascal VOC	http://host.robots.ox.ac.uk/pascal/VOC/
	MS-COCO	https://cocodataset.org/#home
	Visual-Genome	https://homes.cs.washington.edu/ ∼ranjay/visualgenome/index.html
	ADE20K	https://groups.csail.mit.edu/vision/datasets/ADE20K/
	iNaturalist	https://github.com/visipedia/inat_comp
	Stanford Cars	https://www.kaggle.com/datasets/jessicali9530/stanford-cars-dataset
	MegaDepth	http://www.cs.cornell.edu/projects/megadepth/
	ScanNet	http://www.scan-net.org/
	Hpatches	https://github.com/hpatches/hpatches-dataset
	VisLoc benchmark	https://www.visuallocalization.net/
	HICO-DET	http://www-personal.umich.edu/∼ywchao/hico/

Continued on next page

Table 28 – Continued from previous page

Field of Application	Dataset	Dataset Link
	Pascal Context	https://cs.stanford.edu/∼roozbeh/pascal-context/
	CityScapes	https://www.cityscapes-dataset.com/dataset-overview/
	CelebA	http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
Computer Vision (Medical Images)	ISIC 2018 dataset	https://challenge.isic-archive.com/landing/2018/
	Medical Segmentation Decathlon (MSD)	http://medicaldecathlon.com/
	Synapse multiorgan segmentation	https://www.synapse.org/#!Synapse: syn3193805/wiki/89480
	Automatic Cardiac Diagnosis Challenge (ACDC)	https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html
	MICCAI BraTS	http://braintumorsegmentation.org/
	DRIVE	(Staal et al., 2004)
	CHASE DB1	(Hoover et al., 2000)
	STARE	(Owen et al., 2009)
	BUSI	https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset
	DDTI	https://www.kaggle.com/datasets/dasmehdixtr/ddti-thyroid-ultrasound-images
	Dataset A	(Xian et al., 2018)
	National Library of Medicine malaria dataset	https://lhncbc.nlm.nih.gov/ LHC-publications/pubs/MalariaDatasets.html
	SIIM-ACR Pneumothorax Segmentation dataset	https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation
	fastMRI	https://fastmri.med.nyu.edu/
	Visual Genome	https://homes.cs.washington.edu/∼ranjay/visualgenome/index.html
	MS-COCO	https://cocodataset.org/#home
	VG-QA	https://ai.stanford.edu/∼yukez/visual7w/
	Training set fo ALIGN	(Jia et al., 2021)
	ImageNet-V2	https://github.com/modestyachts/ImageNetV2
	ObjectNet	https://objectnet.dev/
	ImageNet Sketch	(Wang et al., 2020b)

Continued on next page

Table 28 – Continued from previous page

Field of Application	Dataset	Dataset Link
	ImageNet Rendition (ImageNet-R)	https://github.com/hendrycks/imagenet-r
	ImageNet Adversarial (ImageNet-A)	https://github.com/hendrycks/natural-adv-examples
	FLD-900M	Private dataset
	SNLI-VE	(Xie et al., 2019)
	SNLI	(Bowman et al., 2015)
	MNLI	(Williams et al., 2018)
	Multi30k	(Elliott et al., 2016)
	CC-3M	https://github.com/google-research-datasets/conceptualcaptions
	Kinetics-600	https://www.deepmind.com/open-source/kinetics
	MSR-VTT	https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/
	Conceptual Captions dataset	https://github.com/google-research-datasets/conceptual-captions
	BookCorpus	https://yknzhu.wixsite.com/mbweb https://github.com/soskek/bookcorpus
	English Wikipedia	https://dumps.wikimedia.org/
	SBU Captions	https://www.cs.rice.edu/∼vo9/sbucaptions/
	Conceptual Captions (CC12M)	https://github.com/google-research-datasets/conceptual-12m
	ALT200M	(Hu et al., 2022)
	Colossal Clean Crawled Corpus (C4)	https://github.com/google-research/text-to-text-transfer-Transformer#c4
	LAION	https://laion.ai/blog/laion-400-open-dataset/
	VQA v2.0	https://visualqa.org/
	GQA	https://cs.stanford.edu/people/dorarad/gqa/
	NLVR	https://lil.nlp.cornell.edu/nlvr/
	TextVQA	https://textvqa.org/
	VizWiz-VQA	https://vizwiz.org/tasks-and-datasets/vqa/
	ST-VQA	https://rrc.cvc.uab.es/?ch=11

Continued on next page

Table 28 – Continued from previous page

Field of Application	Dataset	Dataset Link
	OCR-VQA	https://ocr-vqa.github.io/
	OK-VQA	https://okvqa.allenai.org/
	AudioSet	https://research.google.com/audioset/index.html
	HowTo100M	https://www.di.ens.fr/willow/research/howto100m/
	ILSVRC-2012 ImageNet	https://image-net.org/index.php
	Karpathy split	https://www.kaggle.com/datasets/shtvkumar/karpathy-splits
	Flickr30K	https://shannon.cs.illinois.edu/DenotationGraph/
	nocaps	https://nocaps.org/
	TextCaps	https://textvqa.org/textcaps/
	VizWiz-Captions	https://vizwiz.org/tasks-and-datasets/image-captioning/
	Visual Commonsense Reasoning (VCR)	https://visualcommonsense.com/
	Project WudaoCorpora	https://data.wudaoai.cn and https://github.com/THUDM/ Chinese-Transformer-XL
Audio & Speech	LibriSpeech	http://www.openslr.org/12
	SMS-WSJ	https://github.com/fgnt/sms_wsj
	test/testother	I didn't find this dataset. Should we put private dataset
	Kincaid46	https://github.com/openai/whisper/blob/main/data/README.md
	TIMIT	https://catalog.ldc.upenn.edu/LDC93S1
	Wall Street Journal (WSJ)	https://catalog.ldc.upenn.edu/LDC93s6a
	LibriVox	https://librivox.org/
	BABEL Benchmark	https://babel.is.tue.mpg.de/
	CommonVoice	https://commonvoice.mozilla.org/en/datasets
	Multilingual LibriSpeech (MLS)	https://www.openslr.org/94/
	YouTube-8M	https://research.google.com/youtube8m/
	LS-2mix	(Panayotov et al., 2015)
	WSJ0-3mix	https://www.merl.com/demos/deep-clustering (Hershey et al., 2016)

Continued on next page

Table 28 – Continued from previous page

Field of Application	Dataset	Dataset Link
	LibriSpeech train-clean-360 subset	Private dataset
	WSJ0-2mix	https://www.merl.com/demos/deep-clustering (Hershey et al., 2016)
	GigaSpeech	https://github.com/SpeechColab/GigaSpeech
	VoxPopuli	https://github.com/facebookresearch/voxpopuli
	LibriCSS	https://github.com/chenzhuo1011/libri_css
	CSR-II (WSJ1) Complete	https://catalog.ldc.upenn.edu/LDC94S13A
	AudioSet	https://research.google.com/audioset/index.html
	ESC-50	https://github.com/karolpiczak/ESC-50
	Speech Commands	http://download.tensorflow.org/data/speech_commands_v0.02.tar.gz
	ILSVRC-2012 ImageNet	https://image-net.org/index.php
	VoxLingua107	https://github.com/alumae/torch-xvectors-wav
	CoVoST-2	https://github.com/facebookresearch/covost
	Libri-light	https://github.com/facebookresearch/libri-light
	Fleurs	https://github.com/openai/whisper/blob/main/data/README.md
Signal Processing (Wireless)	Peng Cheng Laboratory (PCL) dataset	https://naic.pcl.ac.cn/contest/10/34
	Foil NLoS	Private dataset
	wall NLoS	Private dataset
	RadioML2016.10b	https://www.deepsig.ai/datasets
Medical Signal Processing	CHB-MIT dataset	https://physionet.org/content/chbmit/1.0.0/
	Sleep Heart Health Study dataset	(Redline et al., 1998)
	MIT-BIH dataset	https://physionet.org/content/mitdb/1.0.0/
	Shaoxing database	(Zheng et al., 2020a)
	The Lobachevsky University electrocardiography database(LUDB)	(Kalyakulina et al., 2020)
	CPSC arrhythmia dataset	http://2021.icbeb.org/CPSC2021

Table 28: References to the Datasets

B REFERENCE TO THE MODELS' IMPLEMENTATION

Field Name	Model Name	Code Link
Natural Language Processing (NLP)	XLM (Conneau & Lample, 2019)	https://github.com/facebookresearch/XLM
	BART(Lewis et al., 2020)	https://github.com/pytorch/fairseq
	Charformer(Tay et al., 2022)	https://github.com/tensorflow/mesh
	T5(Raffel et al., 2020)	https://github.com/google-research/text-to-text-transfer-Transformer
	BERT(Devlin et al., 2019)	https://github.com/google-research/bert
	ELECTRA(Clark et al., 2020a)	https://github.com/google-research/electra
	InstructGPT (Samples)(Ouyang et al., 2022)	https://github.com/openai/following-instructions-human-feedback
	CTRL(Keskar et al., 2019)	https://github.com/salesforce/ctrl
	RoBERTa (Clark et al., 2020b)	https://rule-reasoning.apps.allenai.org/ https://allenai.org/data/ruletaker
	BERT-Based Model (Picco et al., 2021)	https://github.com/IBM/Neural_Unification_for_Logic _Reasoning_over_Language
	PRover (Saha et al., 2020)	https://github.com/swarnaHub/PRover
	PEGASUS (Zhang et al., 2020a)	https://github.com/google-research/pegasus
	Transformer (Vaswani et al., 2017)	https://github.com/tensorflow/tensor2tensor
Computer Vision	ConViT(d'Ascoli et al., 2021)	https://github.com/facebookresearch/convit
	TNT(Han et al., 2021)	https://github.com/huawei-noah/CV-Backbones
	IBOT(Zhou et al., 2021b)	https://github.com/bytedance/ibot
	DETR(Carion et al., 2020)	https://github.com/facebookresearch/detr
	LoFTR(Sun et al., 2021a)	https://zju3dv.github.io/loftr/
	FTN (He et al., 2022)	https://github.com/Novestars/Fully- Transformer- Network
	RAT-Net(Zhu et al., 2022)	https://github.com/zhenxiner/RAT-Net
	nnFormer(Zhou et al., 2021a)	https://git.io/JSf3i

Continued on next page

Table 29 – Continued from previous page

Field Name	Model Name	Code Link
	SwinBTS(Jiang et al., 2022b)	https://github.com/langwangdezhexue/Swin_BTS
Multi-Modality	VL-BERT(Lu et al., 2019)	https://github.com/jackroos/VL-BERT
	UNITER(Chen et al., 2020c)	https://github.com/ChenRocks/UNITER
	BLIP(Li et al., 2022)	https://github.com/salesforce/BLIP
	VATT(Akbari et al., 2021)	https://github.com/google-research/google-research/tree/master/vatt
	DALL-E(Ramesh et al., 2021)	https://github.com/openai/DALL-E
	GLIDE(Nichol et al., 2022)	https://github.com/openai/glide-text2im
	CogView(Ding et al., 2021)	https://github.com/THUDM/CogView
	CogView (Image generation website)(Ding et al., 2021)	https://models.aminer.cn/CogView/index.html
Audio & Speech	VQ-Wav2vec(Baevski et al., 2020a)	http://github.com/pytorch/fairseq
	Wav2vec(Baevski et al., 2020b)	http://github.com/pytorch/fairseq
	XLSR-Wav2Vec2(Conneau et al., 2021)	https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
	Whisper(Radford et al., 2022)	https://github.com/openai/whisper
	HuBERT(Hsu et al., 2021)	https://github.com/pytorch/fairseq/tree/main/examples/hubert
	AST(Gong et al., 2021)	https://github.com/YuanGongND/ast
	Mockingjay(Liu et al., 2020)	https://github.com/andi611/Mockingjay-Speech-Representation
	XLS-R(Babu et al., 2022)	www.github.com/pytorch/fairseq/tree/master/examples/wav2vec/xlsr
	UniSpeech-SAT(Chen et al., 2022b)	https://github.com/microsoft/UniSpeech
Signal Processing	SigT(Ren et al., 2022)	https://github.com/SigTransformer/SigT

Table 29: References to the Code for the Models