A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks

Saidul Islam[1], **Hanae Elmekki**[1], **Ahmed Elsebai**[1], **Jamal Bentahar**[1,2,*], **Najat Drawel** [1], **Gaith Rjoub**[3,1], **Witold Pedrycz**[4,5,6,7]
[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]King Hussein School of Computing Sciences, Princess Sumaya University for 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
[7]Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, Saudi Arabia

[*]**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; g.rjoub@psut.edu.jo; wpedrycz@ualberta.ca

The authors contributed equally to this work.

Abstract

Transformer is a deep neural network that employs a self-attention mechanism to comprehend the contextual relationships within sequential data. Unlike conventional neural networks or updated versions of Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM), transformer models excel in handling long dependencies between input sequence elements and enable parallel processing. As a result, transformer-based models have attracted substantial interest among researchers in the field of artificial intelligence. This can be attributed to their immense potential and remarkable achievements, not only in Natural Language Processing (NLP) tasks but also in a wide range of domains, including computer vision, audio and speech processing, healthcare, and the Internet of Things (IoT). Although several survey papers have been published highlighting the transformer's contributions in specific fields, architectural differences, or performance evaluations, there is still a significant absence of a comprehensive survey paper encompassing its major applications across various domains. Therefore, we undertook the task of filling this gap by conducting an extensive survey of proposed transformer models from 2017 to 2022. Our survey encompasses the identification of the top five application domains for transformer-based models, namely: NLP, Computer Vision, Multi-Modality, Audio and Speech Processing, and Signal Processing. We analyze the impact of highly influential transformer-based models in these domains and subsequently classify them based on their respective tasks using a proposed taxonomy. Our aim is to shed light on the existing potential and future possibilities of transformers for enthusiastic researchers, thus contributing to the broader understanding of this groundbreaking technology.

Keywords: Self-attention; Transformer; Deep learning, Recurrent networks; Long short-term memory-LSTM; 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 multilayer 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, convolutional neural networks incorporate convolutional layers to process images, while recurrent neural networks 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 recurrent neural networks, 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, recurrent neural networks 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 recurrent neural networks 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 deep neural network (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 recurrent neural networks 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., 2020). 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., 2021, 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, Computer Vision, 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, computer vision, 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 remainder of the paper is organized as follows. Preliminary concepts important for the rest of the paper are explained in Section 2. A detailed description of the systematic methodology used to search for relevant research articles is provided in Section 3. Section 4 presents related review papers and discusses similarities and differences with the current survey paper, which helps us identify the unique characteristics and the value added of our survey. Section 5 identifies the the transformer models proposed so far across different fields of application. A Classification of the selected scientific articles on section 6. Section 7 outlines potential directions for future work. Finally, Section 8 concludes the paper and summarizes 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.

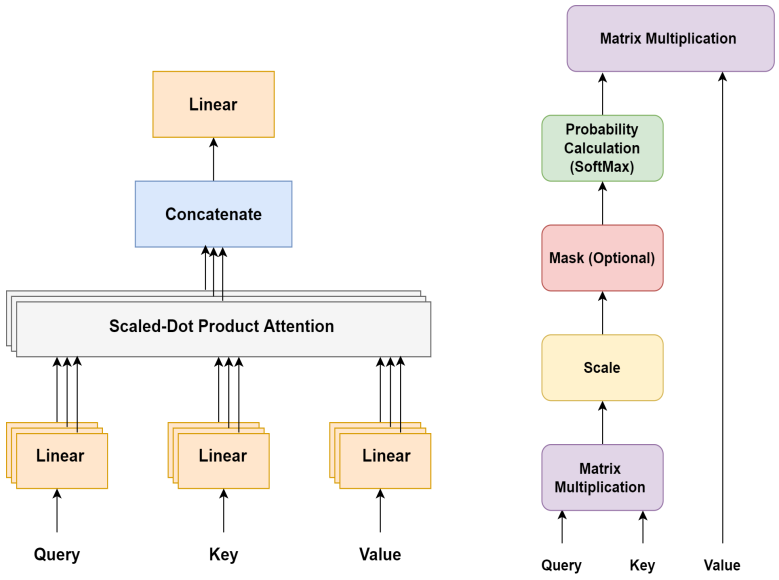

Figure 1: Multi-head attention & scaled dot product attention (Vaswani et al., 2017)

2.1 Transformer Architecture

The transformer model was first proposed in 2017 for a machine translation task, and since then, numerous models have been developed based on the inspiration of the original transformer model to address a variety of tasks across different fields. While some models have utilized the vanilla transformer architecture as is, others have leveraged only the encoder or decoder module of the transformer model. As a result, the task and performance of transformer-based models can vary depending on the specific architecture employed. Nonetheless, a key and widely used component of transformer models is self-attention, which is essential to their functionality. All transformer-based models employ the self-attention mechanism and multi-head attention, which typically forms the primary learning layer of the architecture. Given the significance of self-attention, the role of the attention mechanism is crucial in transformer models (Vaswani et al., 2017)

2.1.1 Attention Mechanism

The attention mechanism has garnered significant recognition since its introduction in the 1990s, owing to its ability to concentrate on critical pieces of information. In image processing, certain regions of images were found to be more pertinent than others. Consequently, the attention mechanism was introduced as a novel approach in computer vision tasks, aiming to emphasize important parts based on their contextual relevance in the application. This technique yielded significant outcomes when implemented in computer vision, thereby promoting its widespread adoption in various other fields such as language processing.

In 2017, a novel attention-based neural network, named “Transformer”, was introduced to address the limitations of other neural networks (such as A recurrent neural network (RNN)) in encoding long-range dependencies in sequences, particularly in language translation tasks (Vaswani et al., 2017). The incorporation of a self-attention mechanism in the transformer model improved the performance of the attention mechanism by better capturing local features and reducing the reliance on external information. In the original transformer architecture, the attention technique is implemented through the “Scaled Dot Product Attention”, which is based on three primary parameter matrices: Query (Q), Key (K), and Value (V). Each of these matrices carries an encoded representation of each input in the sequence (Vaswani et al., 2017). The SoftMax function is applied to obtain the final output of the attention process, which is a probability score computed from the combination of the weights of the three matrices (see Figure 1). 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

The application of the scaled dot-product attention function in parallel within the multi-head Attention 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

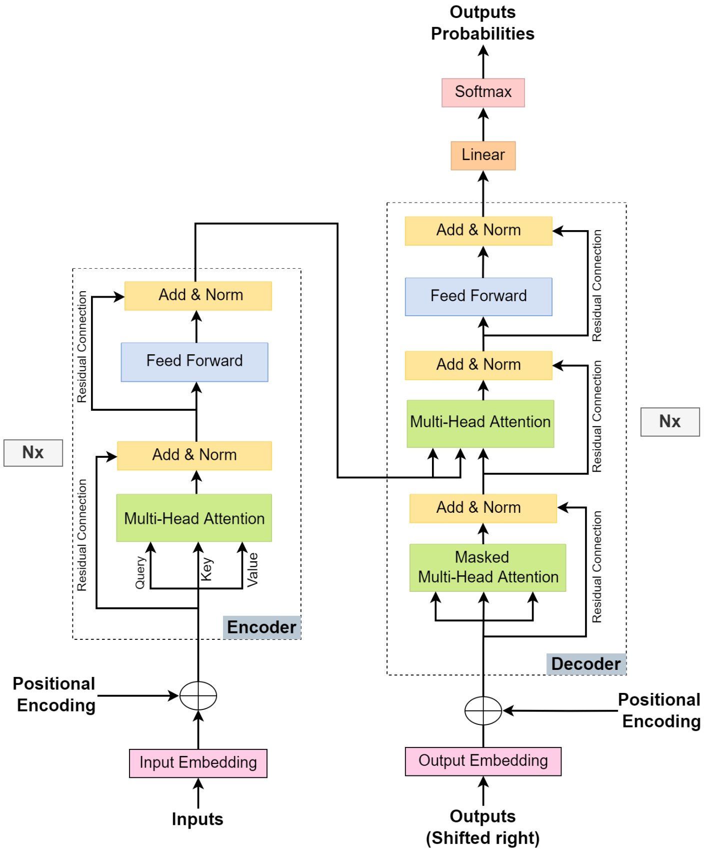

Figure 2: Transformer architecture (Vaswani et al., 2017)

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 multi-head attention 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).

2.2 Architecture of the Transformer Model

The transformer model was primarily developed based on the attention mechanism (Vaswani et al., 2017), with the aim of processing sequential data. Its outstanding performance, especially in achieving state-of-the-art benchmarks for NLP translation models, has led to the widespread use of transformers. As depicted in Figure 2, the overall architecture of the transformer model for sentence translation tasks involves the use of attention mechanisms. However, for different applications, the transformer architecture may be subject to variation, depending on specific requirements.

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).

2.2.1 Encoder module

The stacked module within the transformer architecture comprises two fundamental layers, namely the Feed-Forward Layer and Multi-Head Attention 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 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 "Multi-Head Attention" 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 2.

2.2.2 Decoder module

The Decoder module in the transformer architecture is similar to the Encoder module, with the inclusion of additional layers such as Masked Multi-Head Attention. In addition to the Feed-Forward, Multi-Head Attention, Residual connection, and Add and Norm layers, the Decoder also contains Masked Multi-Head Attention 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 2.

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 eliminated those that were deemed irrelevant for our study. A detailed organization of our survey is depicted in Figure 3.

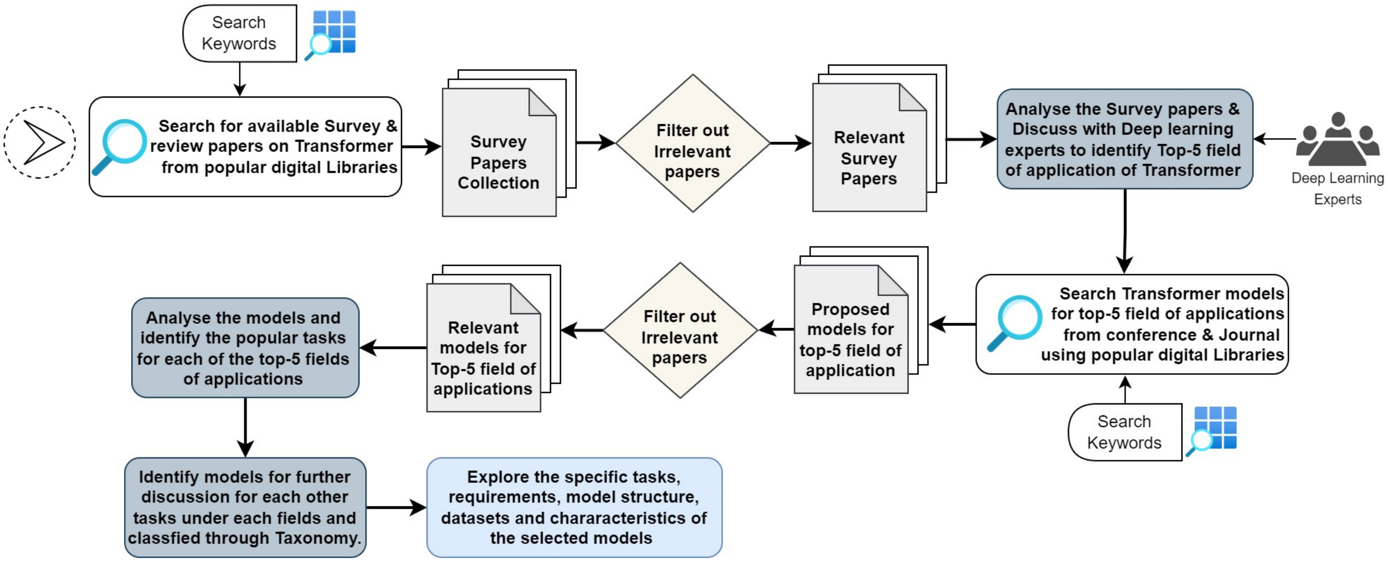
Search Keywords
Search for available Survey & review papers on Transformer from popular digital Libraries
Survey Papers Collection
Filter out Irrelevant papers
Relevant Survey Papers
Analyse the Survey papers & Discuss with Deep learning experts to identify Top-5 field of application of Transformer
Deep Learning Experts
Search Transformer models for top-5 field of applications from conference & Journal using popular digital Libraries
Search Keywords
Proposed models for top-5 field of application
Filter out Irrelevant papers
Relevant models for Top-5 field of applications
Analyse the models and identify the popular tasks for each of the top-5 fields of applications
Identify models for further discussion for each other tasks under each fields and classfied through Taxonomy.
Explore the specific tasks, requirements, model structure, datasets and chararacteristics of the selected models

Figure 3: Methodology of the survey

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) computer vision, (iii) multimodality, (iv) audio/speech, and (v) signal processing. Subsequently, we performed a systematic search for journal and conference papers 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