Journal of Scientific Research and Development (2026) Vol. 25 104-113

A bi-annual journal published by the Faculty of Science, University of Lagos, Nigeria
http://jsrd.unilag.edu.ng/index.php/jsrd

A Systematic Review of Transformers

Adewole Usman Rufai and Samson Ajose
Department of Computer Sciences,
University of Lagos, Lagos, Nigeria
arufai@unilag.edu.ng,
samsonajose@gmail.com

Corresponding author: arufai@unilag.edu.ng

(Received 23 November 2025/ Revised 12 April 2026/Accepted 20 May 2026)

Abstract

In our world today, the way we live and interact has continued to experience rapid development. This feat has largely been enabled by Artificial Intelligence (AI), machine learning, and Neural networks. The advent of transformer model in deep learning has proven to be very revolutionary. This is largely traceable to the self-attention mechanism it adopts. Unnoticeably, we interact with transformer easily today for example, Google uses BERT to enhance its search engine by better understanding users' search queries. Equally, it has been repeatedly mentioned on different media platform that transformers of the GPT family from openAI because of its capability to generate human-like characters and images. These successes and many others have attracted plenty to interest from academic researchers and the industry. In this study, the basic architecture of a transformer was examined including various literatures that surveyed transformers. This study showed the application of transformers in the machine translation, document summarization, document generation, named entity recognition, biological sequence analysis, market intelligence, character recognition. This study also proffers some solutions to some of the challenges with transformers.

Keywords—BERT Bidirectional Encoder Representations from Transformers, A.I Artificial Intelligence

1 INTRODUCTION

In our world today, the way and manner we live and interact has continued to experience rapid development. This rapid development is largely due to the adoption of technology virtually in every work of life. Annika, and Seppänen (2021) described this feat as techceleration which is a new term used to describe the situation where change is happening at an increasingly fast pace due to technological advancements. This feat has largely been enabled by advances in Artificial Intelligence (AI), machine learning and artificial neural networks. Interestingly, the recent covid-19 pandemic also further assisted the adoption of advanced technology in the areas of remote work and replacement of some human activities with the use of intelligent systems. Notably, Natural Language Processing (NLP) which is a component of artificial intelligence has also played a key role

in the evolution of artificial intelligence. Ben (2023) stated that Natural Language Processing is when a computer program recognises the regular human language in the exact way is either written or spoken. Interestingly, since the invention of transformers, it has revolutionized Natural Language Processing (NLP) so much so that it resulted in a great interest in the field of computer vision globally which has resulted to the adaptation of it models for tasks involving vision and multi-modal learning. According to Khan, Naseer, Hayat, Zamir, Khan and Shah (2021), this adaption has led to the successful use of transformer models in areas such as image recognition, detection of objects segmentation, image super-resolution, video understanding, image generation and visual question answering among several other use.

Ilya, Oriol and Qouc (2014), explained that transformer is a prominent deep learning model that is been used widely in different fields of computer science such as natural language processing, computer vision as well as speech processing. Michael (2022) described transformers as a model of deep learning using self-attention as it mechanism which in turn weighs the importance of each part of input data differently.

The huge enhancement shown by transformers for different tasks in computer vision shows clearly that computer vision and natural language processing modelling could potentially be unified under the architecture of Transformers. No doubt, this will be highly valuable for computer vision and natural language process. First, it will be very valuable because it can foster a joint modelling of visual and textual signals. Second, it presents the enablement of better sharing of the knowledge of modelling amongst the two fields, this enablement would therefore lead to the accelerative progress in both fields.

By and large, it is worthy of note that transformers have brought significant success in many aspects of artificial intelligence which includes NLP (natural language processing), computer vision (CV), as well as in audio processing. Therefore, this success has led to a constant natural attraction of various interest from researchers both for academic and industry usage.

This paper aims to first provide a substantive review of literature on transformer variants, it will also look at the architectural modifications and applications of these transformer variants and finally, some potential direction for future work will be outlined.

2 REVIEW OF LITERATURE

Vaswani et al. (2017), showed clearly that a Transformer model relies essentially on the use of self-attention, where the representation of a sentence is calculated by relating different words in the same sequence. Shengzhong et al. (2021), emphasized that when classifying any transformer-based model, the first thing that is computed is the contextual embeddings of every expression present in the sentence inputted via a stack layer of self-attention. Secondly, a prediction is then generated after the weight of the attention has been gathered and given to the contextual embeddings in the first phase.

Dynamically, the predicted output has great importance on the contextual embeddings which is a clear reflection of the weight of the attention. Therefore, the weight of the attention can be regarded as fundamentally explainable which connotes the methods by which transformers are explained naturally.

Shamshad et al (2023) showed how in medical imaging, features which includes the detection, segmentation, reconstruction, registration, and medical reports are generated using transformer models. During this study, not less than 35 survey papers about transformers from conferences and journals in various digital libraries were studied. After a careful consideration of work done, the number of works with more significance to the work at hand was further delved into. We studied their investigations in the different fields of work and the applications they considered. Our findings indicating the strengths and limitations of the studies are shown in Table 1.

Table 1: Summary of Strengths and Limitations of studies

Author	Year	Methodology	Strength	Limitation
Guo, et al.,	2023	A comparation of ChatGPT and humans in conversation quality using a data set and three metric	The work Compared ChatGPT's performance against experts (human). A method to spot chatbot impersonation was also presented.	The performance of the ChatGPT evaluated is inadequate. Limited metric was their focus.
Yelysei, et al.,	2021	Exploration of quantization for BERT-like transformers.	They achieved a 4-bit weight and 2-bit token embedding quantization with less than 0.8% drop leading to significant memory and compute savings.	It is an early work on BERT-like transformer quantization. Hence, the method presented is not exclusive to BERT and is easily applicable to other pre-trained transformer models.
Jiao et al.,	2023	Investigated the performance of machine translation transformers	Showed insights, strategies, and limitations in evaluating the performance of translation.	Coverage of test data was limited. No detailed analysis of translation performance.
Salman et al.,	2021	Focused on the transformer and bi-directional encoding architectures that are built on the principle of self-attention.	Essentially considered self-supervision and self-attention in computer vision.	The study did not consider reducing carbon footprint.
Yang et al.,	2023	Investigated performance on query-based summarization tasks	They identified the research gaps.	Sole reliance on rouge scores for it evaluation.
Antaki et al.,	2023	Provision of an accurate and personalized ophthalmology consultations at a low cost	Provides precise diagnoses and treatment options	No comparison with similar systems
Koco´n et al.,	2023	About 25 NLP tasks was evaluated.	Ethical implications were address. The process of query was automated for efficiency.	No detailed analysis of biases acknowledged.
Frieder et al.,	2019	How transformer model is helping in answering questions in mathematics as well as providing mathematical solution with	Intelligent teaching system.	It only provided a theoretical approach without evidence.

		comparison to other models.		
Chen et al.,	2021	They provided an improved reference library service. Improve and transform library reference service	Analysis for the impact of transformer model.	No detail on and implementation.
Susnjak	2022	How transformer model poses a potential threat to the integrity of exams done online. This includes some strategies for maintaining integrity for the academic.	Online exam threats.	The work didn't provide a detailed analysis of the likely threat to online exams.
Prieto et al.,	2023	Transformer model as used in scheduling of construction project.	Provided solutions for transformer models in both time saving and data mining solution.	Data reliability was not guaranteed. Zero details for applications on other projects.

3.0 ARCHITECTURE OF A TRANSFORMER

In discussing the architecture of transformers, we must re-emphasis that transformers drew it motivation from the encoder-decoder architecture originated in Recurrent Neural Networks owing to the attention mechanism present in them. The motivation of attention mechanisms is that, instead of producing a singular concealed input sentence state, it (the encoder) presents an output of a concealed state at along the line that the decoder can access easily. Interestingly, when transformers were originally published, it was described as a neural machine translation model. This means that, a transformer can be trained to translate English into Yoruba sentences.

Figure 1: Translation done by a transformer. Source (image by author)

With the use of the encoder-decoder architecture in transformers, the encoder abstracts sequences from an input sentence, while on the other hand the decoder uses the sequences to produce an output sentence (translation).

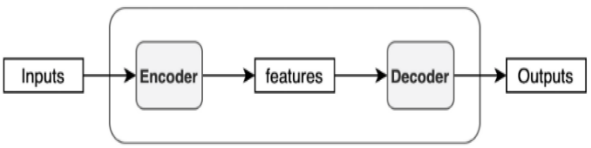

Figure 2: Translation done by a transformer. Source Kheirie (2022)

Unfortunately, using all the states at a time will result in an input that is very huge for the decoder, therefore, there arise a need of an instrument to order the states to be used. Hence, the need for attention comes to play here. This makes the decoder give a size or give attention to the specific states in the past (and the context length can be very long - several thousand words in the past for recent models like GPT or reformers) which are most relevant for producing the next element in the output sequence. The best part is that this process is differentiable, so the process of "paying attention" can be learned during training (Lewis et al 2022). It is worthy of note that a transformer does not make use of sequential order to carry out its data processing. This feature allows for a higher parallelization and quicker training.

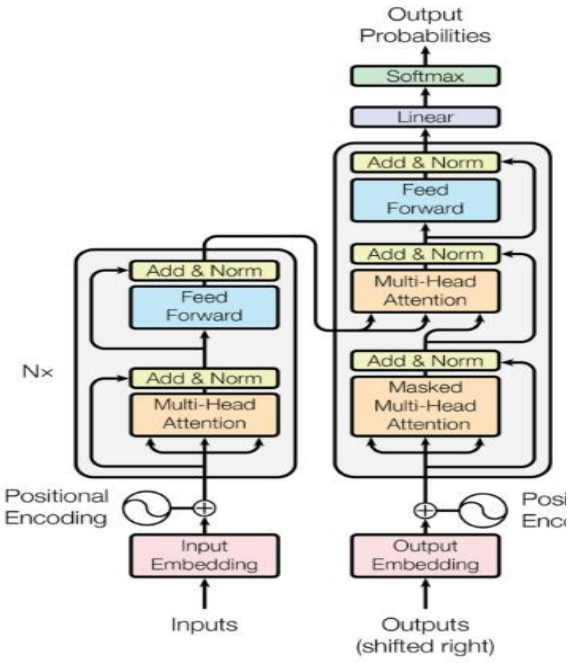

Figure 3: The Transformer model architecture Source: Vaswani et al (2017).

From figure three above and according to Vaswani et al. (2017), 'Most competitive neural sequence transduction models have an encoder-decoder structure'. The figure shows (on the left and the right) the complete architecture made up of pointwise and self-attention, all layers of both the encoder and decoder connected completely.

3.1 Encoder in Transformer

This is also known as the transformer encoder. This is a critical part of the transformer encoder-decoder architecture. It is primarily responsible for the analysis and representing the input sequence in a manner that can be understood by the model. It starts by processing the sequence inputs received and goes on to produce an embedding of the input. Finally, it passes the embedded sequence to the other side of the transformer (the decoder) for the generation of the resulting output sequence.

From the figure 3 above, a transformer architecture generally is made up of different layers with each layer having a self-attention instrument (mechanism) and having a feed-forward neural network (Navaneeth 2023). The self-attention mechanism which is also referred to as the multi-head attention provides room for the model to evaluate the importance of the various sequence inputs received. With the help of the feed-forward network, the model can extract higher-level features from the input sequence received. Additionally, it helps the model to get innate meaning from data received as input for a useful representation of the input. This profound ability of the encoder to process and understand input sequences and generate correct output sequences in a very effective manner has made the transformer encoder-decoder architecture to be a very prevalent architectures in natural language processing.

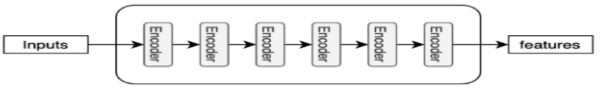

Figure 4: The Transformer Encoder model architecture. Source: Vaswani et al (2017).

3.1.1 **Internal Workings of the Blocks in the Encoder**

1. The input

When input is keyed in, the input data is in English, since the transformer just like every other model does not understand the English language, therefore, the input text is processed, and every word is converted into a unique identification in numeric format. This process is completed with the use of a specific dictionary to vocabulary which is generated from the training data which maps each word to a numeric index.

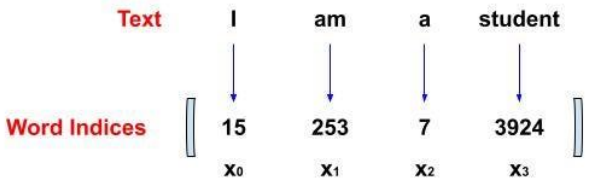

Figure 5 showing a Numerical Representation of the Raw Text. Source Kheirie (2022)

2. **Embedding Layer**

The transformation of inputted tokens into vectors of dimension d=512 is done by the embeddings learned by the transformer. The input tokens are represented with the help of the update numbers in the vector which is done by the model during training.

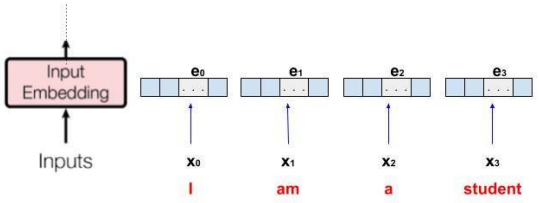

Figure 6: Embeddings of d=512 by The Embedding Layer. Source Kheirie (2022)

3. Positional Encoding

One notable difference between the transformer from other sequence models is that a transformer does take its input embeddings all at once. Notably, this decreases the training time meaningfully which enables parallelization. The disadvantage with this is that it has potential to lose very important information that relates to the order of words. Positional encoding is added to the input embeddings for the model to ensure that words order is preserved.

3.2 **Decoder in Transformer**

The transformer decoder typically just like the encoder. The notable difference between the decoder and the encoder are that the decoder takes in two inputs and applies multi-head attention twice with one of them being "masked" (Kheirie 2022). The decoder generates output sequence by using the produced embeddings received from the encoder with a combination with its (decoder') internal states.

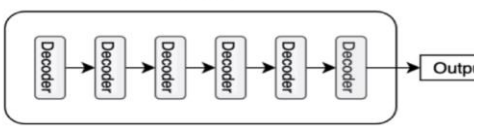

Figure 7: The Transformer Encoder model architecture. Source: Vaswani et al (2017).

Notably, the decoder takes in two inputs. Firstly, the decoder takes the output of the encoder which are the keys (k) and the values (v) that the decoder then goes ahead to perform multi-head attention. In this multi-head attention layer, the query (Q) is the output of the masked multi-head attention (Kheirie 2022).

4.0 APPLICATIONS

Since their introduction, transformers have been used in various applications. However, it been used primarily in areas like artificial intelligence, natural language processing as well as computer vision (CV). Interestingly, transformers were presented initially as a sequence-to-sequence machine translation in natural language processing. Some areas of application of transformers will be discussed.

1. Translation

In context of this study, translation implies the conversion of text from one language to another. For example, Google translator is a very common translator mostly used for text conversion into multiple languages. Transformers essentially uses internal library application models to make the language translations. In order to get this done, the first step is to carryout pipeline initialization using an original language identifier and the identifier of the language to be translated. For instance, to perform a translation from English to Yoruba, we can therefore use : translation_en_to_yr.

2. Document summarization

Document summarization can also be referred to as text summarization. This is because there is an uprooting summary often generated from the planned text. As a result, it is imperative that we make available both the description of the job and the identifier for summarization to ensure the initialization of the text summarization pipeline. Three samples of the arguments should be passed are text, minimum sequence, and maximum sequence.

3. Named Entity Recognition

In the event of name entity recognition, some entity name is assigned (e.g I-MISc, I-PER, I-ORG, I-LOC,) to the tokens contained in the sequence of text. For the initialized pipeline to carry out this name entity recognition, the first thing to be done is to assign to the pipeline a task identifier. Subsequently, the object for the single stream of

text can then be passed. Let us now learn and see what these entities mean.

4. Biological Sequence Analysis

In molecular biology, some examples of fundamental representations of the application of computational methods includes DNA, RNA and protein sequence analysis. These are examples of biological sequence analysis. For example, transformer has helped as a key technique for considering different aspects of proteomics data analysis. Cao and Shen (2021) showed the use of transformer in the annotation function of protein which is a very critical step to identify the distribution of different proteins.

5. Market Intelligence

Essentially, market intelligence can be said to be the process of collecting relevant data with respect to a specific business and then using the data collected to train the NLP model in predicting the dynamics of the current market space using the result of the trained model. The use of market intelligence is to analyse and come up with decisions considering patterns of competing products or markets or needs of specific customers. Marketing intelligence is used to analyse and make decisions using the behaviour of competitors, products, markets, customer needs, etc.

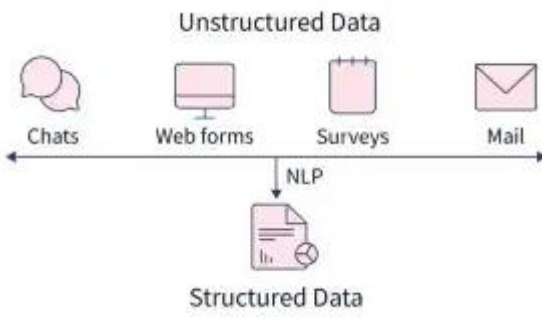

Figure 8 showing NLP in marketing Intelligence source https://www.scaler.com/topics/nlp/nlp-transformers-application/

6. Character Recognition

Computers generally reads and analyses text to recognize the contents which can be characters, numbers, letters, special characters e.t.c which is known as character recognition. Many NLP models have been trained using characters, number and symbols from different languages, therefore, these models can be easily integrated with different applications to recognize characters contained in the input text.

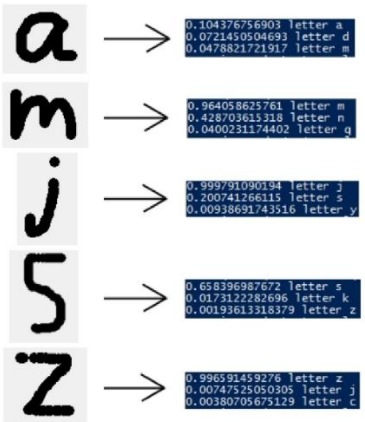

Figure 9 showing character recognition source https://www.scaler.com/topics/nlp/nlp-transformers-application/

4.1 Some Challenges of Transformers

No doubt the use of transformers has been proven to be of immense usefulness, nevertheless, transformers are not magical therefore, it has its own some challenges. Firstly, there exist a barrier encountered by languages because transformer research is largely dominated by the English language. Secondly, the use of transformers works very well on paragraph-long text, however, it is quite expensive to move to longer texts contained in a whole document. Another drawback with transformer is it high computational demand. This is traceable to the complexity and size of models based on transformers that requires a vast amount of computing training time and resources. Additionally, transformers are quite sensitive to the corresponding quality and quantity of the training data. A compromised training data will lead to an adversely affected performance of the model. Notably, this challenge exists in situations where obtaining data is rare of inadequate to find.

High economic cost that comes with training poses another challenge to transformers. This is attributed to the level of computational resources that is required. The continuous rise in the cost of training model even in the face of decrease in the cost of individual floating-point operation is traceable to reasons like growth in the size of training dataset, the number of model parameters in view as well as the amount of operations in the process of training (Shariret al.,2020). Furthermore, the process of training also comes with a notable influence on the environment which can lead to an increase in the consumption of energy and emission of greenhouse gas according to (Strubell et al.,2020). All these challenges therefore make it necessary that the implications resulting from this environmental and economic challenges are addressed promptly.

4.2 Solutions to some challenges

One active area of research presently in solving some challenges with transformer model is on how to make the model more efficient in areas of it computation time and memory. Some of this work has been on the architectural changes to increase the speed of self-attention, this again has proven to be a very costly operation for effective processing of tokens that are of long sequences.

Notably, quantization has proven to be a very useful tool to consider in view of decreasing the consumption of memory and time used in computation. This is done with the application of low-bit representations for weight and tensors. For instance, to transit to 8bits from 32bits, there will be a decrease of four (4) in the memory overhead of storing tensors. On the other hand, there is a quadratic reduction by factor of sixteen (16) in the computational cost for multiplication of the matrix.

5.0 CONCLUSION

The transformer model which is a type of deep learning model has been about the new buzz word now in natural language processing. It has since lead to the birth of top performing models such as BERT which is short form for Bidirectional Encoder Representations from Transformers. The dominance the transformer model has gained in the field of NLP and its increasing penetration in the other areas such as computer vision, there is a need that the architecture of this model is understood properly. In this paper, the architecture of transformer was examined which includes the encoder and the decoder side. It also includes its technical advancements. This study also identified and discussed the challenges and limitations of transformer model, emphasizing areas of improvement and potential research areas. This survey lays the groundwork for a deeper understanding of the buzzing transformer model.

REFERENCES

Abdulazi, A., Amandeep, K., Rao, M., Salman, K., Hisham, C., Gui-Song, X., and Fahad, S. (2022). Transformers in Remote Sensing: A Survey. arXiv:2209.01206v1 [cs.CV] 2 Sep 2022

Ashish V, Noam S, Niki P, Jakob U, Llion J, Aidan N. G, Lukasz K., Illia P. (2017) 'Attention is all you need', Available at: https://doi.org/10.48550/arXiv.1706.03762

Annika, K. and A, Seppänen. (2021) 'Tech-celeration and innovations are blurring borders', Esignals, 09.11.2021. Available at: https://esignals.fi/en/category-en/experience-economy/tech-celeration-and-innovations-are-blurring-borders/#edb8ea4d (Accessed: 3 June, 2023).

Antaki, F., Touma, S., Milad, D., El-Khoury, J., and Duval, R. "Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings," medRxiv, pp. 2023–01, 2023.

Argha, D., Arna, R., Habib, Md. & Akhand, M. A. (2021). Transformer Deep Learning Model for Bangla-English Machine Translation. Conference: nd International Conference on Artificial Intelligence: Advances and Applications (ICAIAA 2021)

Ba, J., Mnih, V., and Kavukcuoglu, K. "Multiple object recognition with visual attention," (2015), arXiv:1412.7755.

Bahdanau, D., Cho, K., and Bengio, Y. "Neural machine translation by jointly learning to align and translate," Tech. Rep., 2016.

Ben, L. (2023) Natural Language Processing, TechTarget, 10 January, 2023. Available at: https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP#:~:text=Natural%20language%20processing%20%28NLP%29%20is%20the%20ability%20of,and%20has%20roots%20in%20the%20field%20of%20linguistics (Accessed: 7 July, 2023).

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.

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.

Fournier, Q., Caron, G. M., & Aloise, D. (2021). A practical survey on faster and lighter transformers. CoRR, abs/2103.14636. URL: https://arxiv.org/abs/2103.14636. arXiv:2103.14636.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Ding, Y., Yue, J., and Wu, Y. (2023) "How close is chatgpt to human experts? comparison corpus,evaluation, and detection," arXiv preprint arXiv:2301.07597, 2023

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In Proceedings of NeurIPS. 3104–3112. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

Jiao, W., Wang. W., Huang, J., Wang, X., and Tu, Z. "Is chatgpt a goodtranslator? a preliminary study," arXiv preprint arXiv:2301.08745, 2023.

Khan, Salman and Naseer, Muzammal & Hayat, Munawar & Zamir, Syed Waqas & Khan, Fahad & Shah, Mubarak. (2021). Transformers in Vision: A Survey. ResearchGate.

Kheirie, E. (2022) *The Transformer Model: A Step-by-Step Breakdown of the Transformer's Encoder-Decoder Architecture.* Available at: https://towardsdatascience.com/attention-is-all-you-need-e498378552f9 (Accessed:10 July 2023).

Koco´n, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., M. "Chatgpt: Jackof all trades, master of none," arXiv preprint arXiv:2302.10724, 2023

Kodialam, R. S., Boiarsky, R., Lim, J., Dixit, N., Sai, A., & Sontag, D. (2020) Deep Contextual Clinical Prediction with Reverse Distillation

Larochelle, H. and Hinton, G. "Learning to combine foveal glimpses with a third-order Boltzmann machine," in Advance Neural Information Processing System, vol. 1, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds. Red Hook, NY, USA: Curran Associates, 2010, pp. 1243–1251.

Lewis, T., Leandro W., and Thomas, W. (2022) Natural Language Processing with Transformers: Building Language Applications with Hugging Face. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020) BEHRT: Transformer for Electronic Health Records. Scientific Reports, 10(1), 7155

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

Ma, L., Liu, Y. , Zhang, X., Ye, Y., Yin, G. & Johnson, B. (2019)"Deep learning in remote sensing applications: A meta-analysis and review," ISPRS journal of photogrammetry and remote sensing.

Michael, S. (2022) *How do Transformers Work in NLP? A Guide to the Latest State-of-the-Art Models*. Available at: https://datasciencelearningcenter.substack.com/p/how-do-transformers-work-in-nlp-a (Accessed:10 July 2023).

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. "Recurrent models of visual attention," in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 1–9.

Navaneeth, M. (2023) *What are Encoder in Transformers*. Available at: https://www.scaler.com/topics/nlp/transformer-encoder-decoder/ (Accessed:17 July

2023).

Rasche, C. (2019) Computer Vision. Available at: https://www.researchgate.net/publication/336460083_Computer_Vision

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).

Shengzhong, L., Franck, L., Supriyo, C., & Tarek, A. (2021). On Exploring Attention-based Explanation for Transformer Models in Text Classification. 1193-1203. 10.1109/BigData52589.2021.9671639.

Stefania, C. (2023) *Implementing the Transformer Encoder from Scratch in TensorFlow and Keras*. Available at: https://machinelearningmastery.com/implementing-the-transformer-encoder-from-scratch-in-tensorflow-and-keras/ (Accessed:17 July 2023).

Stefania, C. (2022) *The Transformer Attention Mechanism*. Available at: https://machinelearningmastery.com/the-transformer-attention-mechanism/ (Accessed:10 July 2023).

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

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.

Tianyang, L., Yuxin, W., Xiangyang, L., & Xipeng, Q. (2021) School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China

Tuia, D., Volpi, M., Copa, L., Kanevski, M., and Munoz-Mari, J. (2011) "A survey of active learning algorithms for supervised remote sensing image classification," IEEE Journal of Selected Topics in Signal Processing, 2011.

Turing (2022) *Understanding Transformer Neural Network Model in Deep Learning and NLP*. Available at: https://www.turing.com/kb/brief-introduction-to-transformers-and-their-power (Accessed: 10 July 2023).

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)

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.

Yang., X., Li, Y., Zhang, X., Chen, H., and Cheng, W. "Exploring the limits of chatgpt for query or aspect-based text summarization," arXiv preprint arXiv:2302.08081, 2023

Zhu, X., Tuia, D., Mou, L., Xia, G., Zhang, L., XU, F., & Fraundorfer, F. "Deep learning in remote sensing: A comprehensive review and list of resources," IEEE Geoscience and Remote Sensing Magazine, 2017.

Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., & Wang, C. (2022). Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing. 60. 1-1. 10.1109/TGRS.2022.3144894.