WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Use of Vision Transformers in Deep Learning Applications

Perumali Sreevatsav

Student, Rajam, Vizianagaram, 532127, India.

ABSTRACT

Nevertheless, the practical problems today are being solved by the data science approaches in efficient mining and analyzing them exploratorily, the implementation of these algorithms is highly scarce in extremely sensitive applications such as in medicine and nuclear simulations. The foundations of Transformers laid thoughts by researchers to rethink if the CNNs are State-Of-The-Art keys in producing trailblazing solutions. Despite the fact at the outset that Vision Transformers (ViT's) were employed for machine translation. The extensively owing flexible architecture of Transformers helps to overcome the anomalous results of the works furnished using techniques of CNN's, RNN's, NLP and advanced deep learning techniques. The odyssey of transformer variants of ViT likely Bidirectional Encoder Representations from Transformers (BERT), Generative Pretrained transformer (GPT), Data-Efficient Image Transformer (DeiT) and Pyramid Vision Transformer (PVT) were capable of accomplishing exceptionally in the fields of NLP, CV, Audio and Video applications, medical imaging, and Security by dint of low/mid/high level vision and self-attention mechanism-based encoder-decoder models for classification and detection. In this paper, we outlined an area of undeveloped but highly crucial topic of study namely multi-sensory data stream handling and current challenges that could incite research.

Keywords: Vision Transformers, CNN, RNN, GPT, NLP, Medical Imaging, Self-Attention, Encoder-Decoder

1. Introduction

- These days CNN, RNN are the fundamental infrastructure to solve the problems of the novel world by considering them as shift-invariant data or sequential time series data.
- Transformer is a new type of neural network. It mainly utilizes the self-attention mechanism to extract intrinsic features and shows great potential for extensive use in AI applications.
- GPT-3 based BERT algorithm formed the primitive concepts of transformers and, with their strong representation capacity, have achieved significant breakthroughs in NLP.
- Eg: BERT, which pre-trains a transformer on unlabeled text taking into account the context of each word (entity) as it is bidirectional.

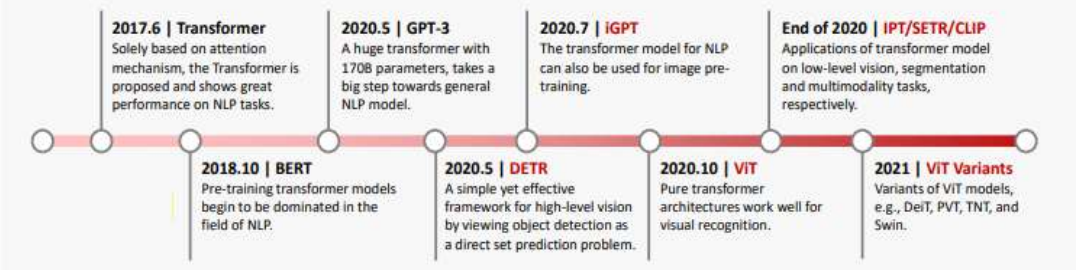

Figure 1: The Generation of Transformers

- A Transformer is a type of neural network that is built on self-attention mechanism wherein an architecture of Encoders and Decoders are involved to produce a vector output and has various applications in the fields of CV, NLP and Medical Imaging.
- This consists of an encoder and a decoder (encoder-decoder attention layer) with several transformer blocks of the same architecture. The encoder generates encodings of inputs, while the decoder takes all the encodings and using their incorporated contextual information to generate the output sequence.

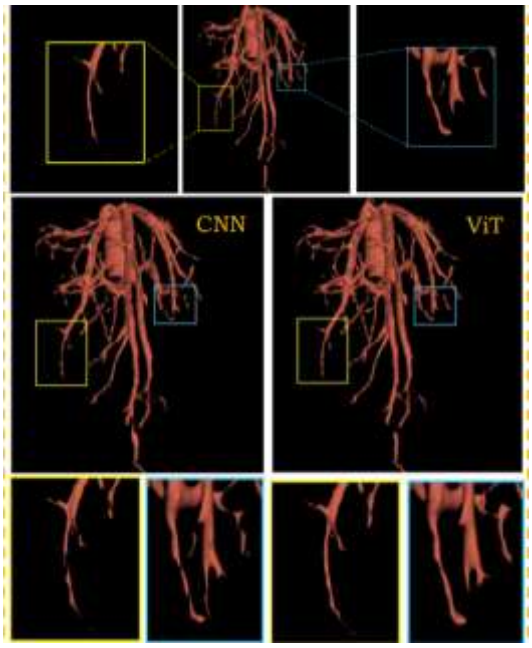

Figure 2: Hepatic Vessel Segmentation using CNN and ViT

- Each transformer block is composed of a multi-head attention layer, a feed-forward neural network, shortcut connection and layer normalization.
- In the self-attention layer, the input vector is first transformed into three different vectors: the query vector q, the key vector k and the value vector v, which are then put in an attention matrix and applied with SoftMax function.
- Accurate medical image segmentation is a crucial step in computer-aided diagnosis, image-guided surgery, and treatment planning. The modeling capability of Transformers is crucial for accurate image segmentation because the organs spread over a large receptive field can be effectively encoded by modeling the relationships between spatially distant pixels.

2. Literature Survey

In paper [1] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A Survey on Vision Transformer, they said that the input vector is first transformed into three different vectors: the query vector 'q', the key vector 'k' and the value vector 'v' with dimension $d_q = d_k = d_v = d_{model}$=512.

- **Step 1:** Compute scores between different input vectors with $\mathbf{S} = \mathbf{Q} \cdot \mathbf{K}^\top$;
- **Step 2:** Normalize the scores for the stability of gradient with $\mathbf{S}_n = \mathbf{S}/\sqrt{d_k}$;
- **Step 3:** Translate the scores into probabilities with softmax function $\mathbf{P} = \mathrm{softmax}(\mathbf{S}_n)$;
- **Step 4:** Obtain the weighted value matrix with $\mathbf{Z} = \mathbf{V} \cdot \mathbf{P}$.

The process can be unified into a single function:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d_k}}) \cdot \mathbf{V}. \qquad (1)$$

- The above equation (1) shall produce vectors with large probability values but lacks capturing the positional information of words in a sentence which is a must. To address this issue and allow the final input vector of the word to be obtained, a positional encoding with dimension dmodel is added to the original input embedding.

In paper [2] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2022). Transformers in medical imaging

- To perform computer-aided diagnosis or image guided surgery, proper segmentation of medical images is a highly crucial step.
- Medical image segmentation area is heavily impacted by transformer-based models as the DNN's could not showcase the required efficiencies. Transformers rely on aggregating information from the entire input sequence.

- The ViT's in medical image segmentation are performed for organ-specific and Multi- organ segmentation, where in-turn they perform 2D and 3D segmentation based on True transformers and Hybrid transformers.
- TransUNet architecture was proposed for multi-organ segmentation that happened to be one of the first transformer-based architecture proposed for medical image segmentation and merits both transformer and UNet. It employs a hybrid CNN-Transformer architecture for encoder, followed by multiple upsampling layers in decoder to output final segmentation mask.
- ViT-based Black-box models for COVID-19 imaging classification generally focus on improving accuracy by designing novel and efficient ViT architectures but still these models could not be interpreted and thus did not gain trust for their implementation for sputum analysis to detect CoViD-19.

In the paper [3] Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., ... & He, Z. (2021). A survey of visual transformers, it dealt with the evolving of transformer categories namely Original Visual Transformer, Transformer Enhanced CNN's, CNN Enhanced Transformer, Local attention Enhanced Transformer, Hierarchical and Deep Transformers that blend the concepts of Self-Supervised Learning.

- A few examples of transformer models that address these high/mid-level vision tasks include DETR, deformable DETR for object detection, and Max-Deep Lab for segmentation. Low-level image processing mainly deals with extracting descriptions from images. Typical applications of low-level image processing include super-resolution, image denoising, and style transfer. At present, only a few works, in low-level vision use transformers.
- Image Generative Pre-training Transformer (iGPT) for self-supervised visual learning. Different from the patch embedding of ViT , iGPT directly resizes and flattens the image to a lower resolution sequence. The resized sequences are then input into a GPT-2 for auto-regressive pixel prediction.

3. Data Collection

Kaggle internet source was used by me to get into the datasets that were used by my reference writers in bringing up their results. A division of Google LLC is Kaggle. Users of Kaggle can discover and share data sets, study and develop models in a web-based data science environment, collaborate with other data scientists and machine learning experts, and participate in competitions to address data science challenges.

In addition to providing machine learning competitions when it first began in 2010, Kaggle currently provides a public data platform, a cloud-based workbench for data science, and artificial intelligence courses. Anthony Goldbloom and Jeremy Howard were among its important employees. Max Levchin took over as founding chair after Nicholas Gruen. The company was valued at $25.2 million when equity was raised in 2011.

4. Methodology

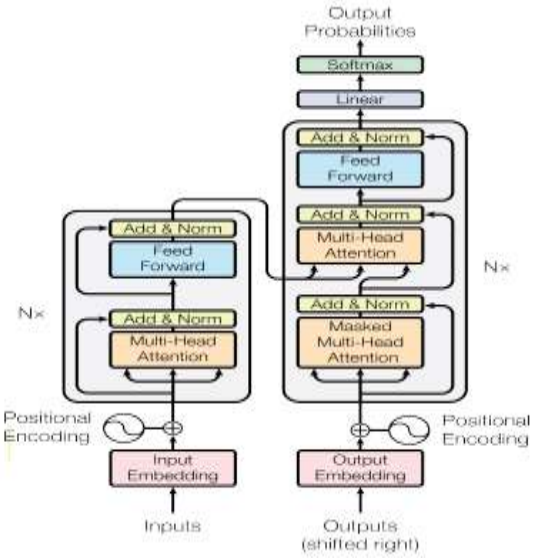

Figure 3: Original Transformer Structure

The image shown above is the structure of an original transformer that my reference writers and I are referring to. It has an input and an output embedding function. The input embedding function is followed by the positional encoding unit which then passes the input to a multi-head attention unit and a normalization unit, then it is fed forward for further normalization and enters into the output embedding unit's multi-head attention unit and a normalizer. It joins the positional encoding part and feeds forward to the linear and the SoftMax functions to get the final output probabilities.

The multi head attention unit of the output unit is shifted right and is always in a masked condition unlike the input embedding unit that has the multi-head attention part unmasked. The flow of the input is diagrammatically shown in the image shown above.

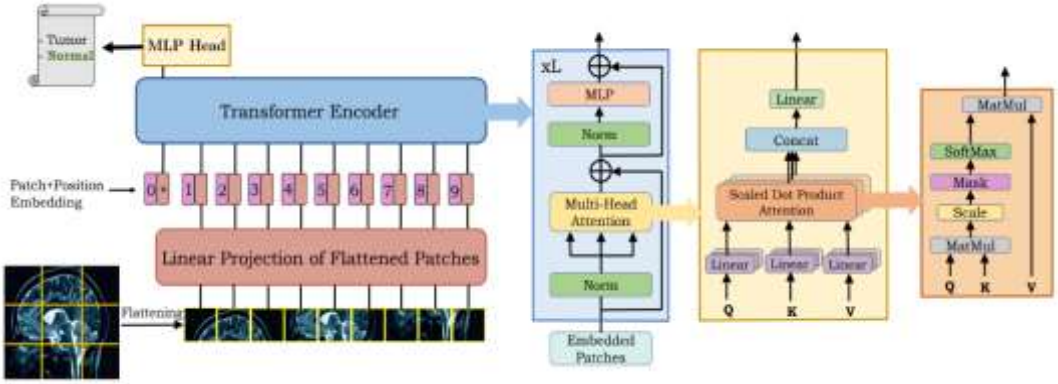

Figure 4: Diagrammatic representation of the working principle

The methodology described above is an algorithm. ViT Operating Principle:

1. Patches of fixed size are formed.
2. Vectorization is accomplished through the flattening process.
3. A linear trainable layer is chosen, and lower dimensional embedding is performed.
4. Position-based encoding is provided.
5. The ViT encoder is given the sequence.
6. Pre-training is performed on a large dataset.
7. The classification task has been fine-tuned.

5. Results and Discussion

Although both global context and sparsely detailed context were critical for precise building segmentation, we designed the BuildFormer with a dual-path structure that could capture both global and spatial information at the same time. Furthermore, we proposed a window-based linear MHSA to reduce the W-MHSA's into O(d N). As a result, the BuildFormer could use large windows to improve global context modelling without requiring a lot of computation. An extensive ablation study assessed the impact of each component of the BuildFormer, and experimental results on the Massachusetts, WHU, and Inria building datasets demonstrated the superiority of the proposed method over existing methods.

S. No.	Without ViTs	With ViTs
1.	Efficiency is low	High efficiency is obtained
2.	Accuracy levels are lesser without ViTs	Good accuracy levels are achieved
3.	No encoding or decoding	Encoding and Decoding is a characteristic
4.	No matrix transformation	Matrix transformation with vectorization
5.	Difficult to analyse the input	Easy to analyse the input
6.	Uses general algorithms	Uses advanced deep learning techniques

6. Conclusion

The use of vision transformers shall be increasing in the future with its outstanding results that it is obtaining due to its special features such as self-attention mechanism and directional encoding with matrix plotting. Further, it needs the support of algorithms to accomplish the process of Computer Vision, but if the input is supplied after the normalization of the transformer, then it would produce a better result than just by the algorithms alone.

Transformer is gaining popularity in the field of computer vision due to its competitive performance and enormous potential when compared to CNNs. A number of methods have been proposed in recent years to discover and use the power of transformers, as summarized in this survey. These methods perform admirably on a variety of visual tasks, including backbone, high/mid-level vision, low-level vision, and video processing.

Nonetheless, the transformer's potential for computer vision has not yet been fully explored, implying that several challenges remain to be overcome. In this section, we discuss these challenges and offer predictions for the future.

REFERENCES

[1]. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A Survey on Vision Transformer. IEEE transactions on pattern analysis and machine intelligence.

[2]. Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2022). Transformers in medical imaging: A survey. arXiv preprint arXiv:2201.09873.

[3]. Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., ... & He, Z. (2021). A survey of visual transformers. arXiv preprint arXiv:2111.06091.

[4]. Wang, L., Fang, S., Li, R., & Meng, X. (2022). Building extraction with vision transformer. IEEE Transactions on Geoscience and Remote Sensing.

[5]. Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A survey of transformers. arXiv preprint arXiv:2106.04554.

[6]. Xu, Y., Wei, H., Lin, M., Deng, Y., Sheng, K., Zhang, M., ... & Xu, C. (2022). Transformers in computational visual media: A survey. Computational Visual Media, 8(1), 33-62.

[7]. Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., & Clapés, A. (2022). Video transformers: A survey. arXiv preprint arXiv:2201.05991.

[8]. De Lima, L. M., & Krohling, R. A. (2022). Exploring Advances in Transformers and CNN for Skin Lesion Diagnosis on Small Datasets. arXiv preprint arXiv:2205.15442.