Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/16/20230898

Applications of transformers in computer vision

Yanxiu Jin[1,3,†], Rulin Ma[2,†]

[1]Beijing-Dublin International College, Beijing University of Technology, Beijing, 100000, China.
[2]College of Arts & Science, New York University, New York, NJ 07302, USA.

[3]corresponding author: yanxiu.jin@ucdconnect.ie
[†]These authors contributed equally.

Abstract. Recently, research based on transformers has become a hot topic. Owing to their ability to capture long-range dependencies, transformers have been rapidly adopted in the field of computer vision for processing image and video data. Despite their widespread adoption, the application of transformer in computer vision such as semantic segmentation, image generation and image repair are still lacking. To address this gap, this paper provides a thorough review and summary of the latest research findings on the applications of transformers in these areas, with a focus on the mechanism of transformers and using ViT (Vision Transformer) as an example. The paper further highlights recent or popular discoveries of transformers in medical scenarios, image generation, and image inpainting. Based on the research, this work also provides insights on future developments and expectations.

Keywords: transformer, semantic segmentation, image processing.

1. Introduction

Transformer is a mainstream and revolutionary deep learning model in NLP tasks. Recent success includes the GPT (Generative Pre-trained Transformer) families that allows auto-regressive modeling and BERT which utilizes the Transformer decoders and encoders, respectively. The core structure in Transformers is the self-attention module that evaluates the importance of a feature on the final results by assisting the model to learn the global contexts and model the long-range dependencies. Inspired by the great achievements in the implementation of transformers in NLP tasks and the emerging capacity of visual attention-based models, researchers have begun to employ Transformer architecture in CV applications, where CNN were regarded as an essential component and a prominent choice [1].

CNN models including ResNet, EfficientNet, AlexNet, VGG, DenseNet etc. possess remarkable performance in feature extraction. CNN is designed domain-specific that helps models to capture visual semantics concerning the locality of pixels. The progressively enlarging receptive field in CNN enables representation of image hierarchical structure in form of semantics. However, the performance of CNN is restricted under some biases in variance and restricted receptive field that limits its ability to recognize the long-range correlations and contextual features, while Vision Transformers relies on minimal inductive biases, making it easier to observe attention links between two input tokens and obtain global sets of connections that models the structural dependencies between input features [2]. Moreover, CNN is domain-specific and lacks scalability to be domain agnostic since it utilizes pixel array, in which each

DOI: 10.54254/2755-2721/16/20230898

pixel passes varying importance, increasing computation and representation complexity. In contrast, ViTs processes images into fixed-sized patches and further reconstructed them into visual tokens through to feed in the standard transformers as words. In this case, the transformers capture the inter relation between different patches within the image instead of words within the text. Additionally, Vision Transformers displays stronger robustness when the foreground objects are under severe occlusion, background regions are not salient and regarding random patch locations in comparison with CNNs. Such robustness can as well be witnessed when dealing with significant and obvious distribution shifts and some nuisance factors taking spatial patch-level permutations, adversarial perturbations and common natural corruptions as examples.

ViT outperforms the SOTA CNN benchmarks not only in computational efficiency but also accuracy. In addition, Transformers displays remarkable scalability against extremely large capacity networks and big datasets and enables processing of many modalities, such as photos and movies, utilizing similar processing blocks [3]. As a result, Vision Transformers have been successfully used for semantic segmentation, image generation and image repair etc. Although ViT has been widely used in computer vision and other fields, but its related review analysis data is not sufficient, so this paper based on ViT, introduced its extension algorithms in different fields of application.

2. Vision transformer

A general structure of Vision Transformers includes a Transformer encoder and a decoder specialized for a down-stream task. Unlike traditional convolutional neural networks (CNNs), which operate on spatially local features, ViTs process images as sequences of fixed-size patches, and apply the Transformer self-attention mechanism to model their interactions (Figure 1). The general pipeline of ViTs involves the steps as follows.

2.1. Patch embedding

An input image x with dimensions H× W × C (height, width, and channels) is first divided into a sequence of 2D non-overlapping patches of size C× P × P, hence the total patches number is HW/P^2. Each patch is then vectorized and linearly embedded into a lower-dimensional space using a learnable linear projection matrix, where D is the embedding dimension.

2.2. Positional encoding

Additional positional encoding vectors are incorporated into the input embeddings to preserve spatial/sequential information which is ignored by the simultaneous and disordered nature of Transformers' operation mechanism. One typical solution of the positional encoding vector for patch and embedding dimension is to apply cosine functions under different frequencies, namely, sinusoidal positional embedding. This creates a matrix of positional encoding vectors, where each row corresponds to a patch and each column corresponds to an embedding dimension.

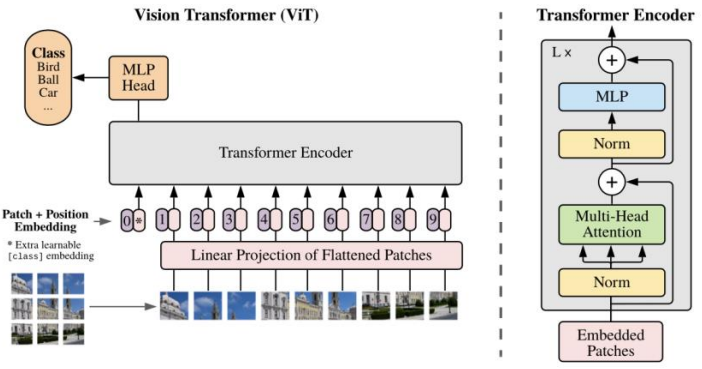

Figure 1. Vision transformer.

DOI: 10.54254/2755-2721/16/20230898

2.3. Encoder

A stack of Transformer encoder layers, each of which is made up of a multi-head self-attention mechanism and a feed-forward neural network, process the augmented patch embeddings. The model can focus on various areas of the image thanks to the self-attention mechanism, and the feed-forward network aids in identifying higher-level features Residual connections are implemented at the end of every block while Layernorm before every block.

2.4. Multi-head self-attention

As displayed in the figure 2, where Q, K, and V are the query, key, and value matrices, respectively, with dimensions N × d_k, N × d_k, and N × d_v. The softmax function computes the weights for each key-value pair, and the resulting weighted sum is multiplied by the value matrix to obtain the output. The MHSA mechanism can be expressed mathematically as follows:

$$Z_i = Attention(Q \times W_i^q, K \times W_i^q, V \times W_i^v) \quad (1)$$

$$MSA(Q, K, V) = concat[Z_1, \ldots, Z_h] \times W^o \quad (2)$$

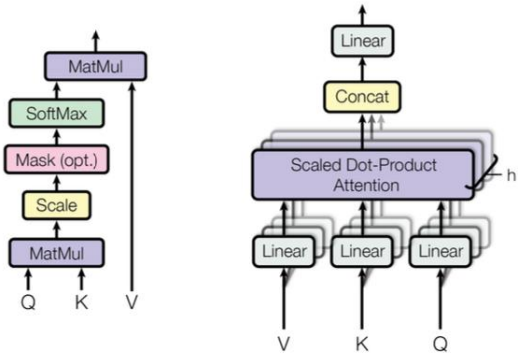

Figure 2. Multi-head self-attention.

Vision transformer-based image segmentation tasks can be categorized into semantic segmentation, medical scene, image generation and image repair. The following details the application of these aspects

3. Medical segmentation

In semantic segmentation, the initialization of the query mechanism is randomized and the subsequent interaction of features using the attention mechanism captures important information and outputs a category information for each pixel, achieving semantic segmentation. FCN (Fully Convolutional Networks) is the first successful trail on end-to-end semantic segmentation. U-Net (U-Shaped net-works) proposed by Ronneberger et al., a variant of FCN is the first dominant architecture in medical image segmentation [1-2]. It managed to provide more detailed hierarchical information with well-modified skip-connections. However, due to the similar limitation of CNNs as described in the section introduction. It is necessary to introduce Vision Transformers to learn contextual dependencies to further improve accuracy of semantic segmentation.

Transformers are crucial especially in medical segmentation tasks since the organs are not restricted within a small receptive field, the backgrounds of the medical scans like ultrasound scans are scattered and that the structure of segmentation target varies between different patients. Medical image semantic segmentation (MISS) includes 2D and 3D segmentation in terms of input data dimensions [3]. 2D segmentation includes but not limited to cells segmentation, kidney tumor segmentation, skin lesion segmentation and polyp segmentation. 3D segmentation is widely applied in fields such as breast tumor and brain tumor segmentation where 3D medical images like MRI and CTs are required for further analysis. Based on the model architectures, the segmentation models can also be divided into hybrid models (ViT+CNN) where the vision transformer is placed in encoder or decoder of both encoder and decoder or in between encoder and decoder, and pure transformers.

DOI: 10.54254/2755-2721/16/20230898

Karimi et al.'s article "Convolution-Free Medical Image Segmentation Using Transformers" from 2021 makes the first pure-transformer medical image segmentation model suggestion. The network takes a 3D block as input and perform non-overlapping patch embedding on it. After adding positional encoding to the flattened and embedded patches, the prediction segmentation of the central patch is derived through utilization of the information in all patches in the input block. Pure ViT can only create low-resolution, single-scale representations, whereas semantic segmentation demands fine-grained, highly position-sensitive image information. MISSFormer provides multi-scale representations (Figure 3).

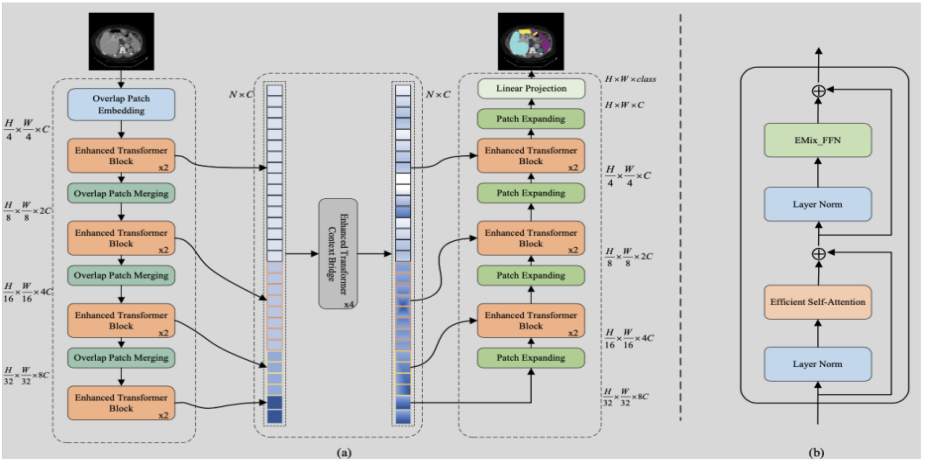

Figure 3. The overall structure of the MISSFormer.

MISSFormer is pure transformer with a U-shaped network consists of encoder, decoder, bridge and skip-connection built on a redesigned enhanced transformer block, rethinking the skip-connection design. The hierarchical encoder contains enhanced transformer blocks to model long-range dependencies and local contexts with better feature consistency. The multi-scale hierarchical features will then be flattened, reshaped and concatenated and get passed through the enhanced transformer context bridge to fuse the global and the local correlations. The output is eventually split and restored to gain the discriminative hierarchical multi-scale information. The segmentation accuracy of MISSFormer outperforms TransUNet and Swin-Unet over Synapse and ACDC datasets.

Another path to obtain multi-scale feature is to model hybrid transformers which possess the advantages of both CNNs and Transformers.

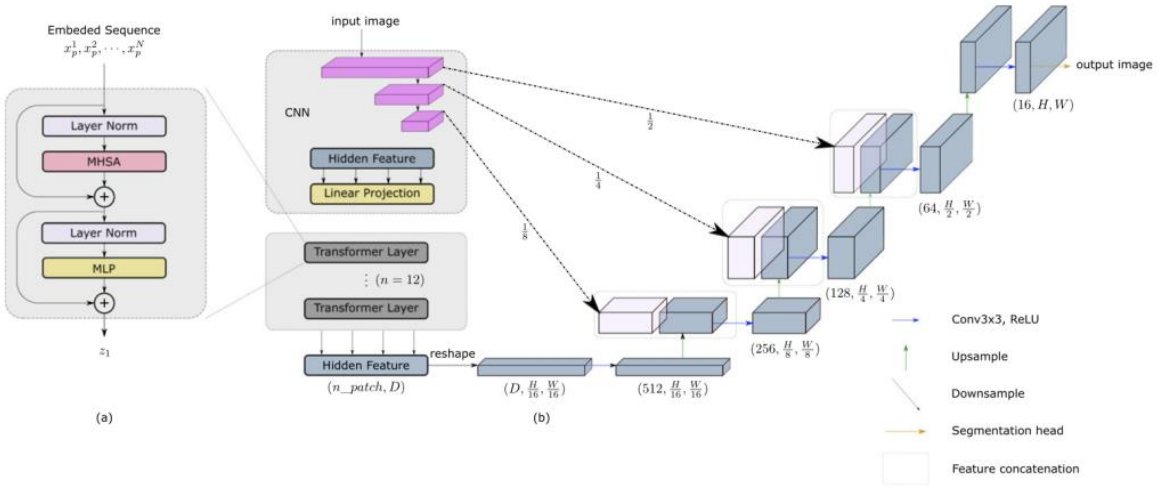

Figure 4. Overview of TransUNet for medical image segmentation.

DOI: 10.54254/2755-2721/16/20230898

Considering the great achievement of U-Net, most researchers have combined Transformers with the U-shape liked architecture to make up its limitation in modeling global and long-range dependencies. TransUNet is the first study of such architecture, using Transformers to encode global and long-range dependencies [4]. It deals with the resolution loss brought about by Transformer by exploiting spatial information by CNN block (Figure 4). Such various high-resolution features are then combined with the up-sampled attention features and enables precise localization.

The design of the dilated convolution module is used by the D-Former, which enlarges the convolution kernel by creating holes between its nearby continuous pieces [5]. It contains a 3D U-shaped hierarchical encoder-decoder architecture and newly built D-former blocks that are made up of LSMs and GSMs in various sequences to provide local and global contextual features, respectively. The figure 5 shows that GSM employs global scope multi-head self-attention while LSM uses local scope multi-head self-attention. In order to perform self-attention inside each unit, LSM separates a 3D feature map into non-overlapping units with a certain number of 3D patches. The implementation of GS-MSA to model interactions through several units in a dilated manner seeks to make up for the lack of global information and long-range dependencies. The receptive field of self-attention is enlarged without extra computational costs since the patch number remain unchanged.

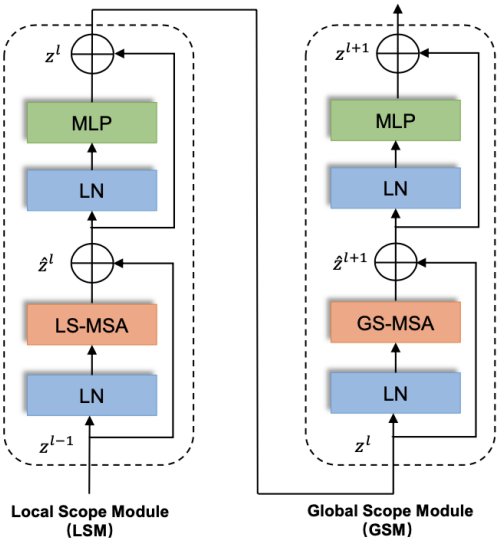

Figure 5. The local scope module and global scope module.

In addition, relative and absolute position information is embedded via dynamic position encoding by D-Former. D-Former is a SOTA method that outperforms MISSFormer in average accuracy.

4. Image inpainting

Transformers have been widely employed in image reconstruction, especially have shown promise in image inpainting, super resolution, tomography reconstruction and compressed sensing. In image inpainting, in particular, the standard deep learning method such as using a convolutional filter response (CNN) conditioned on valid pixels and also with the mean value in holes has led to color discrepancy and blurriness. Thus, given the rise of transformers, researchers employed transformers to meet the need of filling in holes in an image to improve its quality and completeness. One of the methods included transformers is Deep Image Prior, an image inpainting method that uses a combination of convolutional and transformer layers [6]. The method is based on the idea that the structure of natural images itself can be used as a prior for image reconstruction, without the need for any training data.

The approach involves first initializing a randomly generated image and then optimizing it using a deep convolutional-transformer network to reconstruct the missing or damaged parts of the input image. The network consists of a series of convolutional layers followed by self-attention transformer layers, which capture both local and global features of the image.

DOI: 10.54254/2755-2721/16/20230898

The Deep Image Prior approach has been shown to achieve state-of-the-art results on several image inpainting benchmarks, including the CelebA and Places datasets. One of the key advantages of this approach is that it does not require any training data, making it useful for inpainting tasks where little or no data is available. Additionally, the use of transformer layers allows the network to capture long-range dependencies and global context, which can improve the quality of the reconstructed image. Built on those merits of transformers, many researchers seek to construct models that combine the long-range interactions advantage of transformers and the efficiency of processing and computing of convolutions. One example was the MAT (Mask-Aware Transformer for Large Hole Image Inpainting) introduced by Wenbo and his fellow researchers (Figure 6) [7]. MAT targeted large-hole image inpainting and high-quality image production. Researchers proposed a multihead contextual attention (MCA) to quickly fill out the holes; this facilitates the efficiency boost. However, this model only treats 512x512 images, which is not ideal enough for industrial practice. Another worth-noting model is the bidirectional autoregressive transformer (BAT) [8]. To enhance precision and efficiency problems unaddressed by unidirectional transformers, BAT sorts the valid and missing pixels and then models from the first missing pixel. This process makes all context available for more precise image inpainting. BAT provides more diversity and accuracy when it comes to image inpainting.

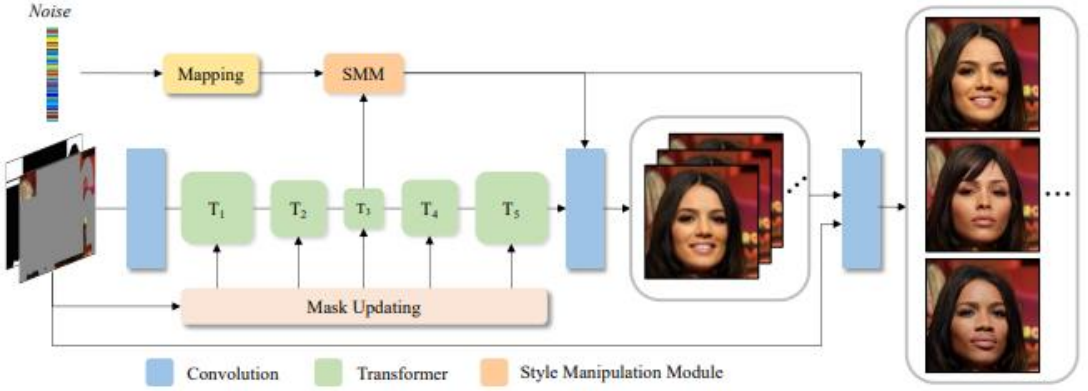

Figure 6. MAT framework.

5. Image generation

Image generation refers to the process of computers generating new images from existing dataset, which is also called image synthesis. Before the arrival of transformers, Generative Adversarial Networks (GANs) were one of the most popular methods for image generation [9]. GANs were first introduced by Ian Goodfellow. et al. It displayed great results when researchers trained their models based on MNIST Handwritten Digit Dataset, the Toronto Face Database (TFD), and CIFAR-10. One advantage of GANs is that they can generate highly realistic images. However, GANs require relatively harsh conditions to train successfully and may suffer from issues such as mode collapse or instability.

The rise of transformers offers the industry an option for image generation. In 2021, Kwonjoon Lee and his colleagues in UCSD optimized GANs by combining GANs and Vision Transformer(ViTs) and including novel regularization techniques, which led to the method called ViTGANs in figure 7 [10]. One advantage of this method is that unlike the CNN based GAN. such as StyleGAN2, ViTGANs require no convolution or pooling and thus have smaller computational budgets. At the same time, ViTGANs maintains comparable accuracy with styleGAN2.

DOI: 10.54254/2755-2721/16/20230898

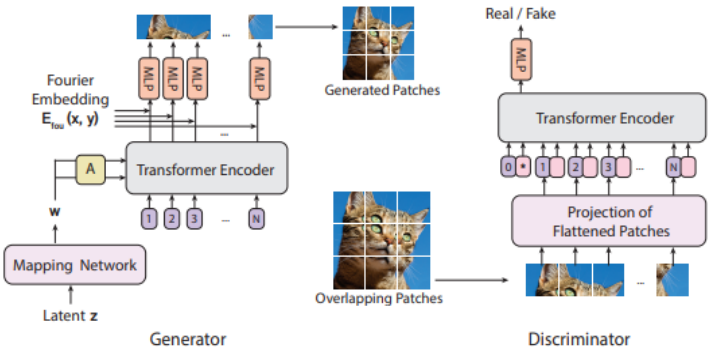

Figure 7. VitGAN framework.

Still, limited by the process of tokenization and autoregressive prediction, which are crucial for transformer-based models, the time-consumption is huge for image generation. Therefore, Huiwen Chang and his colleagues in Google AI proposed a new transformer called MaskGIT that extends on the idea in VQGAN [11]. Researchers in Chang's team focused on improving the autoregressive prediction; they switched the unidirectional transformer for a bidirectional transformer. In other words, their idea is based on the normal painting process of humans. Unlike the sequential decoding used by other transformers, Chang and other researchers start with a "sketch" and refine the sketch (or the tokens) in the following fixed number of iterations. Because of all these novelties, MaskGIT outperforms BigGAN in better coverage (Recall), and better sample quality (Precision) compared to VQVAE-2 and diffusion models (Figure 8). Researchers also observed better objects and global structures completion which allows great potential in Class-conditional Image Editing, Image Inpainting and Image extrapolation.

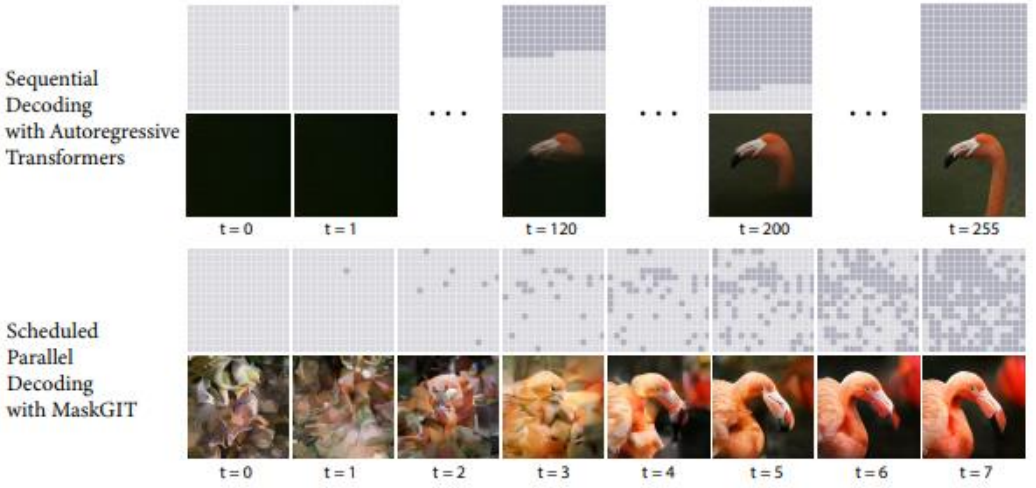

Figure 8. Comparison between sequential decoding.

In 2023, Google AI released a new text-to-image model called Muse [12]. Muse utilizes masked generative transformers trained on discrete token space, in contrast to traditional pixel-space diffusion models (Figure 9). The model incorporates parallel decoding to predict multiple output tokens in a single forward pass, resulting in improved time efficiency. Comparative results reveal that Muse outperforms Imagen-3B and Parti-3B models is 3x faster than Stable Diffusion v1.4, while maintaining high quality and precision.

DOI: 10.54254/2755-2721/16/20230898

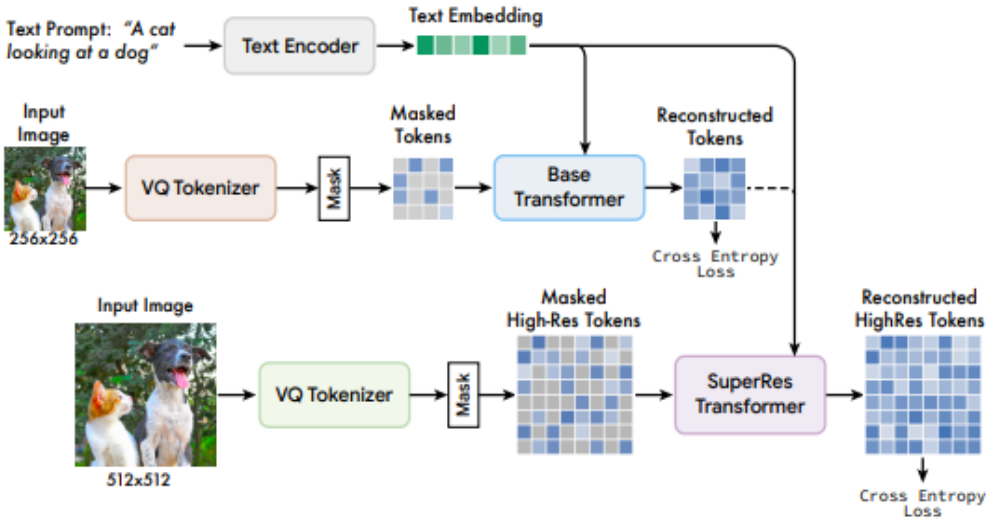

Figure 9. Muse framework.

6. Conclusion

Upon reviewing the working mechanism of transformers and some latest discoveries, we believed that despite transformer's success in various regions, it's important to acknowledge the limitations of transformers. Transformers perform worse than other deep learning algorithms such as CNN in smaller training data, which limits its advantage in fields where data is expensive or hard to obtain. Similarly, the requirement of training large-scale data makes transformers consume large computational resources. This restrains transformers' industry usage as high-performance GPUs or TPUs could be expensive. At the same time, limitations of transformers are actively being addressed in ongoing research, and improvements are constantly being made to overcome these limitations and enhance the capabilities of transformer models for various applications.

References

[1] Alexey Dosovitskiy,.et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021, Conf. Comput. Vis. Patt. Recog. 28 1-10.
[2] Antonelli, et al. The medical segmentation decathlon. 2020, Nature communications, 1-10.
[3] Xiaohong Huang,.et al. MissFormer: An Effective Medical Image Segmentation Transformer, 2021 ArXiv:2109.07162
[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, et al. Transunet: Transformers make strong encoders for medical image segmentation. 2021 arXiv:2102.04306.
[5] Yixuan Wu,. et al. D-former: A U-shaped dilated transformer for 3D medical image segmentation, 2022, arXiv:2201.00462.
[6] Ulyanov,.et al. Deep Image Prior. 2018 Conf. Comput. Vis. Patt. Recog.1-11.
[7] Wenbo Li,.et al, MAT: Mask-Aware Transformer for Large Hole Image Inpainting 2020, arXiv:2203.15270.
[8] Yingchen Yu,.et al, Diverse Image Inpainting with Bidirectional and Autoregressive Transformers, 2021 arXiv:2104.12335.
[9] Ian J. Goodfellow, .et al, Generative Adversarial Networks, arXiv:1406.2661.
[10] Kwonjoon Lee,.et al, ViTGAN: Training GANs with Vision Transformers, 2021 arXiv:2107.0458
[11] Huiwen Chang,.et al, MaskGIT: Masked Generative Image Transformer,2022, arXiv:2202.04200.
[12] Huiwen Chang,.et al, Muse: Text-To-Image Generation via Masked Generative Transformers, 2023, arXiv:2301.00704