I.Introduction

In the report, we will explore the vision transformers architecture that utilizes a multi-head attention mechanism in natural language processing. Our analysis will cover the different approaches that have been developed after ViTs, as well as the unsolved problems and areas of further research.

II. Structure Explanation and Analysis

A. Encoder Input

Tokenize: We first split the 2D image into fixed-sized patches, flatten them, and create a sequence of patches (x_p). Each element of the sequence is a token.

Linear projection: We map each flattened patch to a D-dimension vector using trainable linear projection. So each flattened patch is passed as an input of a linear dense layer and its output is $x_p^i * w + b$ (w and b are learnable parameters, w and b are the same for each flattened patch, w is the same as the input embedding = E)

Adding positional encoding: We add the standard learnable 1-D position embeddings to the previous output to carry positional information. ($x_p^i * E + E_{pos}^i$). Here we add x_{class} as our index 0 argument to the sequence. x_{class} is a learnable embedding to the sequence of embedded patches. The self-attention allows it to interact with all other patch tokens as the tokens pass through the transformer's layers. So it takes global contextual information across the entire image and aggregates important features for classification.

So at the end of this step, our input sequence (z_0 or embedded patches) is like this:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \cdots\ ;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \qquad (1)$$

B.Transformer Encoder

The embedded patches are passed into a norm layer and 3 copies of the result are passed as Query, Key, and Value to the Multi-Head Attention module. Then the output of this step is added to the embedded patches by a skip connection. This output is passed to another Norm layer and an MLP unit and added by itself again and produces an Encoder output sequence.

Norm: Suppose we have a batch of N embeddings so each of them is a vector of numbers. For each item, we calculate the mean and variance independently. We replace each number with the below formula. So all numbers are now between zero and one. We also multiply each number by gamma and add beta to it. Gamma and beta are learnable parameters of the model, the model should learn to multiply by gammas and add betas to amplify the values it wants to be amplified and not amplify the values it doesn't want to be amplified.

$$\hat{x}_j = \frac{x_j - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}} \qquad (2)$$

Multi-Head Attention: We take three copies of the encoder input and pass it to the Multi-head attention model as Query, Key, and Value. For each of them, there is a corresponding parameter matrix called W_Q, W_K, and W_V which has the dimensionality of d_{model} x d_{model}. We multiply each Query, Key, and Value matrix with their corresponding parameter which results Q', K', and V' matrices. We split each Q', K', and V' matrices along the d_{model} dimension so that every head will see the whole input but a smaller part of the embedding of each token ($d_k = d_{model}$ /h, h = number of heads). So to calculate each head we apply the attention formula on the corresponding splits of Q', K', and V'.In the next step, we concatenate multiple heads to a single head along the $d_v = d_k$ dimension and we get back to our initial dimension. After that, we multiply the head by a new matrix called W_O (parameter matrix) which has the dimensionality of ($h * d_v$, d_{model}). The result of this multiplication is our multi-head attention matrix with the same dimensionality as the encoder input. So here each head is watching the full input but the different aspects of the embedding of each token because we want each head to watch different aspects of the same token. We can visualize attention. After applying softmax on each split, we can see the score matrix. We can see relations of one token with multiple tokens by different heads and sometimes there a relation of two tokens is not shown by all heads, this is because each head can only see a part of the token embedding (an aspect) so it is possible that other heads cannot find the relation since they didn't see that aspect of the token embedding.

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad MultiHead(Q,K,V) = Concat(head_1\ \ldots head_h)W^O \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \qquad (3)$$

MLP: It is a multi-layer perceptron module with one hidden layer at pre-training time and a single linear layer at fine-tuning time.

C. MLP head and Layer normalization

From all the elements in the encoder output sequence, only the first one (index = 0 or z^0_L) is important for us since it has the information for doing the classification task (embedding of the x_{class} token). So we take it and apply layer normalization to it and then we have the final result which is the model prediction for the class of the given input. Both during pre-training and fine-tuning, a classification head is attached to z^0_L. The classification head is implemented by an MLP.

III. Related Publications Discussion

A.Transformer-Enhanced CNNs

Transformers can reduce MHSA's bias through skip connections and FFNs. To enhance image processing capabilities, some researchers are integrating transformers into convolutional neural networks. They use MHSA and FFNs to reduce bias and improve semantic modeling. VT-blocks [4] and BoTNet [5] are two novel approaches to substitute or reconfigure convolution stages with attention mechanisms, resulting in significant improvements in representation learning. These hybrid models have demonstrated superior performance on benchmarks like ImageNet.

B.CNN-Enhanced Transformer

Another trending method is enhancing Transformer models with CNN inductive biases to improve image processing capabilities. The idea is to help CNNs process images more efficiently by incorporating their inherent assumptions about data distribution. This enables them to handle locality and translation invariance. CNN biases may limit their potential when there is ample data available. So some researchers are incorporating CNN biases into Transformers to enhance data efficiency and reduce their dependence on large datasets. The DeiT [6] method softens the transfer of inductive biases through teacher-student distillation and outperforms its CNN teacher. Other methods like ConViT [7], CeiT [8], and LocalViT [9] are introducing convolution directly into Transformer structures to embed locality and inductive biases softly. They are significantly improving accuracy and model performance.

C.Local Attention-Enhanced Transformer

To address the limitation of vision transformers in capturing local image details, we can integrate local attention mechanisms and CNN features. The Swin [10] Transformer uses shifted window techniques to enhance the modeling of global and boundary features. It reduces computational complexity and boosts performance achieving 84.2% accuracy on ImageNet. The TNT [11] model combines patch- and pixel-level representations for comprehensive image understanding. Models like Twins [12] and ViL [13] use new mechanisms for local-global representation and leverage local embeddings for focused attention, respectively. VOLO [14] uses outlook attention, similar to dynamic convolution, that focuses on extracting finer details in images. It achieves state-of-the-art results without external data.

D.Hierarchical Transformer

We can address the limitations of the vision transformer by introducing variable resolution features and reducing computational costs like T2T-ViT [3] and PVT [15] hierarchical transformers. T2T-ViT uses overlapping operations for downsampling which improves resolution but also increases memory and computational demands. PVT technique optimizes non-overlapping patch partitioning and a Spatial Reduction Attention (SRA) layer to effectively reduce computational requirements while adapting to tasks that demand large inputs and detailed features. PiT [16] and CvT [17] use pooling and convolution for downsampling. CvT enhances SRA through convolutional projection and allows adaptation to inputs of any size without positional encodings. These transformer models are more efficient and flexible for complex vision tasks.

E.Deep Transformer

These models face challenges such as attention collapse [3] and patch oversmoothing [3]. These issues can result in less representative deep layer features and indistinguishable latent patch representations. The CaiT [18] model introduces a two-stage class attention mechanism and distillation training can combat these issues. It achieves state-of-the-art results on ImageNet-1k without external data. Deep ViT solves diversity loss by aggregating different head attention maps and introducing a distributed local attention mechanism for enhanced local and global feature modeling. Patch diversity losses methods use techniques such as patchwise cosine loss, contrastive loss, and mixing loss to improve deep Transformer training strategies and outcomes [3].

F.Transformers With Self-Supervised Learning (SSL)

The main idea is to use both generative and discriminative approaches for SSL. Despite high computational costs, the iGPT [19] generative model shows the potential of transformers through autoregressive pixel prediction. BEiT [20] and DINO [21] discriminative models use masked image reconstruction and teacher-student knowledge distillation to stabilize training and enhance feature learning.

IV. Future Research areas

The set prediction problem arises due to the convergence of class tokens caused by identical gradient influences from the loss function. This convergence leads to training instability and accuracy issues. So there is a need for innovative label assignment methods and set prediction loss designs to overcome this issue. SSL problem is methodological discrepancies between convolutional Siamese networks (contrastive learning) and masked autoencoders (NLP). This shows a need for innovative or adapted self-supervised techniques for visual Transformers and research into encoder–decoder frameworks to bridge this gap which is aligned with my interest since it is foundational to many applications across NLP and CV.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[3] Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J. and He, Z., 2023. A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems.

[4] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K. and Vajda, P., 2020. Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677.

[5] Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P. and Vaswani, A., 2021. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16519-16529).

[6] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H., 2021, July. Training data-efficient image transformers & distillation through attention. In International conference on machine learning (pp. 10347-10357). PMLR.

[7] d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G. and Sagun, L., 2021, July. Convit: Improving vision transformers with soft convolutional inductive biases. In International Conference on Machine Learning (pp. 2286-2296). PMLR.

[8] Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F. and Wu, W., 2021. Incorporating convolution designs into visual transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 579-588).

[9] Li, Y., Zhang, K., Cao, J., Timofte, R. and Van Gool, L., 2021. Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707.

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[11] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C. and Wang, Y., 2021. Transformer in transformer. Advances in Neural Information Processing Systems, 34, pp.15908-15919.

[12] Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H. and Shen, C., 2021. Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems, 34, pp.9355-9366.

[13] Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L. and Gao, J., 2021. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2998-3008).

[14] Yuan, L., Hou, Q., Jiang, Z., Feng, J. and Yan, S., 2022. Volo: Vision outlooker for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 45(5), pp.6575-6586.

[15] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P. and Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 568-578).

[16] Heo, B., Yun, S., Han, D., Chun, S., Choe, J. and Oh, S.J., 2021. Rethinking spatial dimensions of vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11936-11945).

[17] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L. and Zhang, L., 2021. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 22-31).

[18] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G. and Jégou, H., 2021. Going deeper with image transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 32-42).

[19] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D. and Sutskever, I., 2020, November. Generative pretraining from pixels. In International conference on machine learning (pp. 1691-1703). PMLR.

[20] Bao, H., Dong, L., Piao, S. and Wei, F., 2021. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

[21] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650-9660).

[22] Attention is all you need (Transformer) - Model explanation (including math), Inference and Training “https://www.youtube.com/watch?v=bCz4OMemCcA&list=WL&index=6”

[23] Vision Transformer for Image Classification “https://www.youtube.com/watch?v=HZ4j_U3FC94&list=WL&index=2&t=707s”

[24] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Paper Explained) “https://www.youtube.com/watch?v=TrdevFK_am4&t=1539s”

[25] BERT for pretraining Transformers “https://www.youtube.com/watch?v=EOmd5sUUA_A”

[26] Attention Is All You Need “https://www.youtube.com/watch?v=iDulhoQ2pro&t=35s”