A FEDERATED PURE VISION TRANSFORMER ALGORITHM FOR COMPUTER VISION USING DYNAMIC AGGREGATION MODEL

Hatem Osama Ismail[1], Mohamed Waleed Fakhr[2] andMohamed A. Abo Rezka[3]

[1] College of Engineering and Technology, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt.
[2] Professor, College of Engineering and Technology, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt.
[3]Professor and Dean of Faculty of Computers, and Information, Arab Academy for Science & Technology, Cairo, Egypt.

Email: eng.hatemosama@gmail.com, waleedf@aast.edu, m.aborizka@aast.edu

ABSTRACT

Federated Learning (FL) provides training of global shared model using decentralized data sources on edge nodes while preserving data privacy. However, it's performance in the computer vision applications using Convolution neural network (CNN) considerably behind that of centralized training due to limited communication resources and low processing capability at edge nodes. Alternatively, Pure Vision transformer models (VIT) outperform CNNs by almost four times when it comes to computational efficiency and accuracy. Hence, we propose a new FL model with reconstructive strategy called FED-REV, Illustrates how attention-based structures (pure Vision Transformers) enhance FL accuracy over large and diverse data distributed over edge nodes, in addition to the proposed reconstruction strategy that determines the dimensions influence of each stage of the vision transformer and then reduce its dimension complexity which reduce computation cost of edge devices in addition to preserving accuracy achieved due to using the pure Vision transformer.

KEYWORDS: Federated Learning, Vision Transformer, Model reconstruction.
DOINumber:10.14704/nq.2022.20.10.NQ55730 NeuroQuantology 2022;20(10):7406-7414

1. INTRODUCTION

Recently, Federated Learning (FL) is a novel framework which allows the training of neural network models using private data which spread across several heterogeneous systems [1]. FL maintains the privacy of data on every end-system and trains a global shared model that is updated by transmitted parameters rather

than the transmitting data itself. As a result, it enabled shared machine learning structure throughout many institutions without exposing their private data [2]. This and internet of things[6], where maintaining data privacy is critical.

Lately, vision transformer also gained significant interest and provided novel insights into a variety of computer vision tasks compared to the performance of the convolution neural networks (CNN), including classification tasks [7], object detection[8], and segmentation techniques [9]. Furthermore, most suggested vision transformers require high capacity,

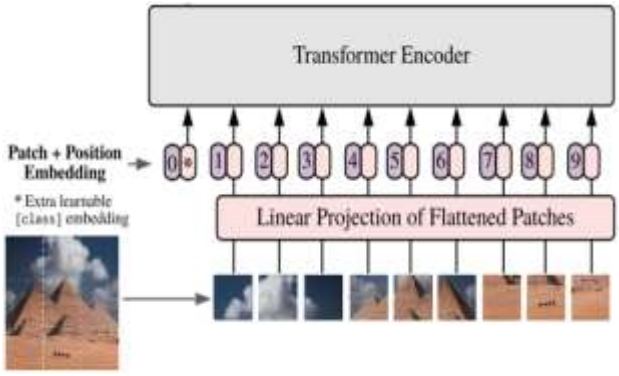

Figure1. Structure of vision transformer

Transformers may be compressed and speeded up to variable degrees by utilizing various designs. ALBERT [14] minimizes network parameterization and accelerates learning curve by decomposing parameters into smaller matrices and allows cross-layer parameter sharing. The Star-Transformer [15] sparingly connects the regular transformer's structure by relocating it to a star-shaped topology. The small networks in [16] learn from the bigger pre-trained master networks using knowledge transfer techniques of teacher-student strategy and the student transformers achieved competitive accuracy of 95% compared to the teacher transformer. There have been several efficient pruning methods developed to minimize the number of attention heads [17] or individual weights [18]. The preceding solutions have concentrated on compressing and speeding up the transformer for text analysis jobs. With the advantage of vision transformers like as VIT [7] a high-performance transformer is needed for image processing applications.

has been particularly beneficial in many sectors like medical [3], blockchain[4], communication networks [5],

run-time memory, and huge computational resource which prevent their wide adoption on edge devices such as self-driven car and smart phones. In CNN several strategies adopted for shrinking its structure and speeding it up, the most efficient ones yet included deep model compression [10], quantization techniques [11], filter pruning strategy[12], and knowledge distillation [13].

While most research efforts focus on improving the optimization process in FL, we propose a federated reconstrued vision transformer based on learnable significance scores to overcome the mentioned problems,resulting in a convenient dynamic structure.

The suggested algorithm's efficacy is demonstrated by experimental results on the benchmark datasets. Our approach of federated vision transformer with reconstruction approach minimizes the complexity toaccelerates learning curve and speedsup the original federated learning models significantly in vision-based tasks. This paper establishes the first dynamic reconstruction approach for federated vision transformers, laying the groundwork for future research.

2. Related work

2.1. Federated Learning (FL):

The Federated framework aims to train machine learning models on private data across massively distributed devices.Adaptive Federated Optimization methods[19] could be used to improve non-guaranteed convergence and model weight divergencein parallel FL methods such as FedAVG [1], whereas serial FL methods such as Split learning[20] train each client serially and produce a significant catastrophic forgetting problem[21],[22]is the first work that applies federated learning to a real-world image dataset, Google Landmark [23], which has now become the standard image dataset for federated learning research. [24], [25]apply federated learning on medical image segmentation tasks, which aims at solving the issue in which the training data may not

be available at a single medical institution due to data privacy regulations, FL Most commonly applied in vision-based applications using convolution neural network (CNN) [26]–[28]or recently with hybrid of CNN and vision transformers [29]–[31]which Transformers can overcome catastrophic forgetting problem which happens in case of CNN.

2.2. Vision Transformers:

Recently, Vision Transformer frameworks have outperformed CNN in visual classification tasks[32]. Their design is like language processing's Transformers but requires separating the input photos into fixed-size patches, adding a "Classification Token" to each pattern, and putting the generated series into the regular Transformer Encoder. VIT's attention is represented in Figure 1, It has many heads that process a chunk of Q, K, and V in order to extract various properties between patches and provide attention scores inside each head using theEquation of the three parameters :Query(Q),Key(k) and Value (V) as following $Q \cdot K^T / \sqrt{d}$which defined as Attention (A) and its dimensions as $H \times L \times L$, the attention of each head is considered to be the result of the inner product of the rows in Q as well as the columns in K^T, which could be interpreted as the mutual dependence of different patches in head (h). Following that, a SoftMax with row-wise conversion is used to convert attention scores to attention probability. SoftMax can discriminate between patches with varying degrees of information content and amplify the significant variation between patches with various degrees of informativeness. Following the computation of A.V of each head, each patch obtains the properties of neighboring patches. The output of the attention mechanism of all the heads is combined and transformed into linear projection (L) dimension.

2.3. Vision transformer reconstruction:

Pruning is well-known for its effectiveness in lowering the cost of deep neural network inference. It can be classified into two types: (i) non-structured pruning: this method entails deleting unnecessary input components based on a set of criteria such as the input's weight relevance, however it is hardware incompatible[33]. , and (ii) structured pruning, that includes removing model sub-structures such as enhanced the result accuracy but still have catastrophic forgetting problem , However, In this paper we applied FL to a pure vision transformer modelshowing that channels [34]and attention heads [35], which are frequently quite matched with hardware performance. VIT has achieved competitive accuracy in a range of computer vision tasks by training the whole complex model first, typically for numerous train-prune-retrain rounds. Their memory and computing requirements, on the other hand, limit their deployment on some applications. MHSA and FFN are quite slow on some devices due to their high computational load, which significantly reduces efficiency. There have been several proposals for transformer pruning approaches, including attention head pruning, weight pruning and patch pruning [36].

3. Vision Transformers

The usual architecture of a vision transformer[7] is composed of a Multi-Head Self-Attention (MHSA), a Multi-Layer Perceptron (MLP), layer normalization, an activation function, and a bypass connection. MHSA is a transformer component that enables information interaction between tokens. Through fully connected layers, the input $X \in R^{N \times D}$ is translated into query $Q \in R^{N \times D}$, key $K \in R^{N \times D}$, and value $V \in R^{N \times D}$is the dimension of the embedding. To represent the link between patches, the self-attention mechanism is used.

$$\text{Attention (Q,K,V)} = \text{SoftMax} \left(\frac{QK}{\sqrt{d}} \right) V \quad (1)$$

Finally, the MHSA output is generated via a linear transformation.:

$$Y = X + \mathrm{FC}_{\mathrm{out}} \left(\text{Attention } \left(\mathrm{FC_q(X)}, \mathrm{FC_k(X)}, \mathrm{FC_v(X)} \right) \right) \quad (2)$$

If layer normalization and activation parameters are omitted for simplification. the two-layer MLP as result:

$$Z = Y + \mathrm{FC}_2\left(\mathrm{FC}_1(Y)\right) \quad (3)$$

On the top, the significant utilization of fully connected layers by transformers results in higher computing and storage requirements.

3.1. Pure Vision Transformer architecture:

eISSN1303-5150

we intend to use only pure Vision Transformers, which do not employ any standard convolutional layers. Rather than that, the following two stages are added to replace CNN's function of extracting picture features:

3.1.1. Image quantization: As described in [32], we begin by reshaping the input X_p into a sequence of flattened 2-Directional patches $\{X_p^i \in R^{p^2.c} | i = 1, .., N\}$, where each patch has a size of (P × P) and the total number of image patches is $(N = \frac{HW}{P^2})$

3.1.2. **Patch Embedding**: We move the vectorized patches X_p into a latent D-dimensional embedding space using a trainable linear projection. To encode the spatial information, we learn precise position embeddings that are merged with patch embeddings as follows: $Z_0 = [X_{class}; X_p^1 E; X_p^2 E; \dots \dots \dots X_p^N E + X_{pos}$.where X_{class}stands for the class token.$E \in R^{(p^2.c)\times D}$ and $E_{pos} \in R^{N\times D}$are the patchembedding projection and the position embedding, respectively. Transformer encoders contains multiple layers of Multi-head Self-Attention (MHSA) and Multi-Layer Perceptron (MLP) blocks. As a result, the output can be represented as follows:

$$z'_l = MSA\big(LN(z_l - 1)\big) + z_{l-1} \qquad (2)$$

$$z_l = MLP\left(LN\left(z'_l\right)\right) + z'_l \qquad (3)$$

where LN () signifies the layer normalization operator and the L-th layer output's first element Z_o is the corresponding image representation.

4. FED-REV

Our proposed federated Reconstructed vision transformer approach includes two stages presented in Figure 2:

Initial stage: the cloud sends initial vision transformer model to all edge nodes then aggregates all the model weightsafter local training using Adam optimizer for faster convergence in addition to deploying the reconstruction strategy.

Dynamic stage:it starts from the second roundtrip and involves reconstructionof the new vision transformer model at the cloud after each roundtrip based on the aggregated values from the edge nodesby detecting the lower outlier head scores using standard deviation method, mask them in addition to averaging the weights of the remaining heads finally send the new customized model to each edge node as represented in (Algorithm 1).

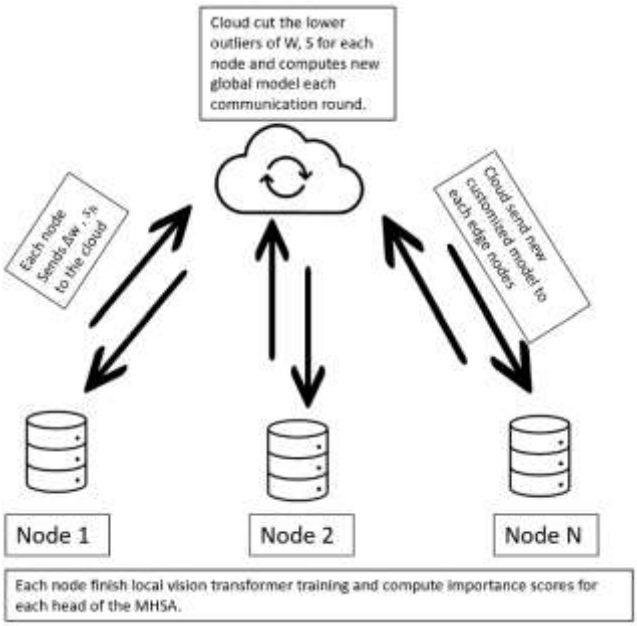

Figure 2. illustration of **FED-REV**

4.1. Calculating the importance score of each head of the MHSA for each client

To optimize the size of the transformer's design, the FLOPs of MHSA and MLP are reduced. By knowing their respective importance. for analyzing the architecture of MHSA, we Consider a Transformer Encoder with attention head h ∈ [1, 12], defining the matrix of $Q \cdot K^T / \sqrt{D}$ as A, For a patch p_0:

$$A_h[p_0, :] = \sum_p Q_h[p_0, p] \,.\, K_h^T[p, :]$$

and

$$A_h[:, p_0] = \sum_p Q_h[:, p] \,.\, K_h^T[p, p_0]$$

thatrepresent the interdependency between p_0 and other input patches. To be specific$A_h[p_0, p]$ is computed as the weighted sum of$K_h[p, :]$, which can be viewed as the impact from p to p_0,in head h.

The total informativeness of patch p_0, in head h can be defined as:

$$\alpha \sum_i A_h[p_0, i] + \beta \sum_i A_h[j, p_0] \qquad (1)$$

where α and β are two parameters indicating the difference between the impact of p_0, on other patches

and the impact of other patches on p_0,. Further to say, we can obtain the total informativeness of patch p_0, to the whole layer as:

$$\sum_{h}\left(\alpha\sum_{i}A_h[p_0,i]+\beta\sum_{i}A_h[j,p_0]\right) \quad (2)$$

To define the importance score of head h, the formula is:

$$\boldsymbol{s_h}=\sum_{i}\sum_{j}A_h[i,j] \quad (3)$$

.

After calculating the importance score of each head of all the MHSA for all nodes, the cloud will collect all the heads scores beside the weight of each dataset from each node and detect the lower outliers using the standard deviation method, generate mask for these heads thensend the new customizedglobal model to each nodeas represented in (algorithm 1).

*e*ISSN1303-5150

Algorithm 1: Dynamic stage of FED-REV

Input: Communication round t;
visiontransformermodel V;
No of nodes n;
No of iterations k;
Number of heads: h;
Model MHSA weights w;
head importance scores s;
1for each t=1,2..T **do**
cloud sends initial V model to all nodes
calculateweights of each edge node dataset
3 **for** each n in **parallel do**
4 **for k= 0,1….K do**
5/*Adam optimizerprocess */

$$w_{t+1} = w_t - \widehat{m_t}\left(\frac{\alpha}{\sqrt{\widehat{v_t}} + \mathcal{E}}\right)$$

Node UpdateV (w_k)
End for
8Compute local V changesΔW_v
Compute importance scores$s_h$as follows:
Attention probability = SoftMax($\frac{Qk^T}{\sqrt{d}}$)
Initializeaccumulation process
13for head = 0 ← h **do**
| **for** patch = 0 ← L **do**
| | s_p+ = max (attention probability [head,
| | , patch])
| **end**
| s_h+ = sum (s_p)
End
return Δw and$s_h$values to the cloud.
End for
Cloud aggregates all Δwand $s_h$from all nodes.
/* **Initialize**model reconfiguration process follow: */
Compute σ, μ for standard deviation method:
$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$,$\sigma_{threshold}$ = 1,2,3
/* compute Lower outliers */
ifs_h<$s_{h-threshold}$then generate mask $\widehat{M_h}$ for its head, average theweights of the remaining heads.
End if

$$w_{t+1} = w_t - \sum_{n=1}^{N} \frac{1}{N} w_{t+1}^n$$

Update*Normalization* and *SoftMax*
Send the new customized model to all nodes
End for

5. EXPERIMENTS

In this section, we present the datasetsand experimental setup.

5.1. Datasets

1. **MNIST:**It is a dataset of 60,000 square 28×28-pixel grayscale images of handwritten single digits between 0 and 9.
2. **CIFAR-10:**consists of 60000 32×32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.
3. **ImageNet-100:** collected as a subset of ImageNet1K. We first randomly sampled 100 classes which has 1000 training data and 300 validation datafor each class.
4. **CXR images:**Total of 30482 CXR image contains 13992 positive covid-19 and 16490negative covid-19 collected by university of waterloo, Canada on Kaggle

5.2. Implementation Details

We evaluate FED-REV on image classification tasks by Dividing data into training and testing by the ratio of 5:1 and then divide into nth number of edge nodes.

6. Evaluation Configurations and Results

Dataset	MNIST	CIFAR-10	ImageNet-100	CXR images
Learning rate of Adam optimizer	0,25	0,25	0,25	0,25
Number of data samples used in initial round	300	300	500	150
Number of nodes	10	10	10	10
Reconstruction	Every 40 round	Every 40 round	Every 40 round	Every 40 round
Mini-batch size	5	5	5	5
local iterations in each round	20	20	20	20
Total number of FL rounds	10,000	10,000	30,000	5,000
Accuracy percentageof	95.13%	94.82%	81.91%	83.25%

eISSN1303-5150

CVT				[3]
Accuracy percentage ofFed-REV method	**98.13%**	**97.82%**	**85.91%**	**90.25%** [4]

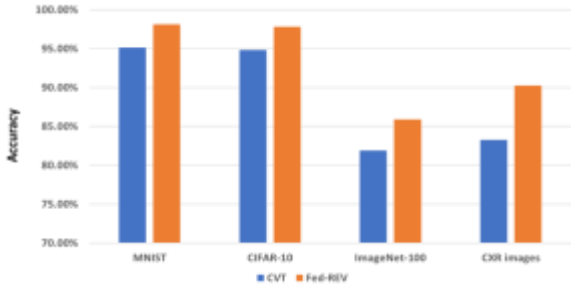

Figure 3. illustration of accuracy (%) for CVT method and FED-REV method.

7. CONCLUSION

In this paper, we introduce FED-REV,a novel dynamic federated vision transformer model with reconstructive strategy.The experiments were conducted on MNIST, ImageNet-100 and CIFAR-10. As a practical demonstration we deployed FED-REVon chest X-ray (CXR) imageswhich used forcentralized neural network training for COVID-19 diagnosis without the need to collect patient CXR data from several hospitals. Thisdemonstratedthat using Fed-REV achieve remarkable accuracy results compared to the centralized vision transformer performance in addition to lowering the computation cost at the edge nodes due to the reconstruction strategy.

8. REFERENCES

[1] H. Brendan McMahan Eider Moore Daniel Ramage Seth Hampson Blaise AgüeraAg and A. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," 2017.

[2] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards Personalized Federated Learning",2021.

D. H. Mahlool and M. H. Abed, "A Comprehensive Survey on Federated Learning: Concept and Applications",2022.

D. C. Nguyen *et al.*, "Federated Learning Meets Blockchain in Edge Computing: Opportunities and Challenges",2021.

[5] Z. Yang, M. Chen, K. K. Wong, H. V. Poor, and S. Cui, "Federated Learning for 6G: Applications, Challenges, and Opportunities," *Engineering*, vol. 8, pp. 33–41, Jan. 2022, doi: 10.1016/J.ENG.2021.12.002.

[6] S. R. Priya M, Q.-V. Pham, K. Dev, P. Kumar Reddy Maddikunta, T. Reddy Gadekallu, and T. Huynh-The, "Fusion of Federated Learning and Industrial Internet of Things: A Survey",2022.

[7] A. Dosovitskiy *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE", Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/

[8] M. Zheng *et al.*, "End-to-End Object Detection with Adaptive Clustering Transformer," 2021, Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/gaopengcuhk/SMCA-DETR/

[9] J. Chen *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation", Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/Beckschen/

[10] S. H. Shabbeer Basha, M. Farazuddin, V. Pulabaigari, S. R. Dubey, and S. Mukherjee, "Deep Model Compression Based on the Training History",2021.

[11] "Transform Quantization for CNN Compression." https://www.researchgate.net/publication/350709335_Presentation_Transform_Quantization_for_CNN_Compression (accessed Jun. 04, 2022).

[12] J.-H. Luo and J. Wu, "An Entropy-based Pruning Method for CNN Compression",2018.

*e*ISSN1303-5150

[13] Y. Tian, D. Krishnan, G. Research, and P. Isola, "CONTRASTIVE REPRESENTATION DISTILLATION", Accessed: Jun. 04, 2022. [Online]. Available: http://github.com/HobbitLong/RepDistiller.

[14] Z. Lan *et al.*, "ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS", Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/google-research/ALBERT.

[15] Q. Guo, X. Qiu, P. Liu, Y. Shao, X. Xue, and Z. Zhang, "Star-Transformer", Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/dmlc/dgl

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2021.

[17] P. Michel, O. Levy, and G. Neubig, "Are Sixteen Heads Really Better than One?", Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/neulab/compare-mt

[18] M. A. Gordon, K. Duh, and N. Andrews, "Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning", Accessed: Jun. 04, 2022. [Online]. Available: https://www.tensorflow.org/versions/

[19] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "FEDERATED OPTIMIZATION IN HETEROGENEOUS NETWORKS," 2020.

[20] S. Park, G. Kim, J. Kim, B. Kim, and J. C. Ye, "Federated Split Vision Transformer for COVID-19 CXR Diagnosis using Task-Agnostic Training.",2021.

[21] M. J. Sheller *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–12, Jul. 2020, doi: 10.1038/s41598-020-69250-1.

[22] J. Luo *et al.*, "Real-World Image Datasets for Federated Learning", Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/FederatedAI/

[23] Z. Kim *et al.*, "Towards A Fairer Landmark Recognition Dataset", Accessed: Jun. 04, 2022. [Online]. Available: https://www.kaggle.com/c/

[24] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated Learning for Healthcare Informatics," *J Healthc Inform Res*, vol. 5, no. 1, pp. 1–19, Mar. 2021, doi: 10.1007/S41666-020-00082-4/TABLES/2.

[25] R. S. Antunes, C. A. da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated Learning for Healthcare: Systematic Review and Architecture Proposal," *ACM Transactions on Intelligent Systems and Technology (TIST)*, May 2022, doi: 10.1145/3501813.

[26] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-IID data",2020.

[27] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated Learning with Non-IID Data",2018.

[28] C. He, M. Annavaram, and S. Avestimehr, "Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge", Accessed: Jun. 04, 2022. [Online]. Available: https://fedml.ai

[29] "Federated Split Task-Agnostic Vision Transformer for COVID-19 CXR Diagnosis | OpenReview." https://openreview.net/forum?id=Ggikq6Tdxch (accessed Jun. 04, 2022).

[30] A. Dosovitskiy *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE", Accessed: Jun. 04, 2022. [Online]. Available: https://github.com/

[31] B. Yassine and G. Rasool, "Scopeformer: n-CNN-ViT Hybrid Model for Intracranial Hemorrhage Classification," *Medical Imaging with Deep Learning,2021*.

*e*ISSN1303-5150

[32] H. Touvron *et al.*, "Training data-efficient image transformers & distillation through attention",2020.
[33] T. Chen, Y. Cheng, Z. Gan, L. Yuan, L. Zhang, and Z. Wang, "Chasing Sparsity in Vision Transformers: An End-to-End Exploration",2021.
[34] H. Yu and J. Wu, "A Unified Pruning Framework for Vision Transformers",2021.
[35] Z. Song, Y. Xu, Z. He, L. Jiang, N. Jing, and X. Liang, "CP-ViT: Cascade Vision Transformer Pruning via Progressive Sparsity Prediction," 2022, [Online]. Available: http://arxiv.org/abs/2203.04570
[36] Y. Tang *et al.*, "Patch Slimming for Efficient Vision Transformers",2021.