Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/79/20241537

Comparative analysis of transformer and GoogleNet models in image classification based on the CIFAR dataset

Xinran Xie[1,6,*], Qinwen Yan[2,7], Haoye Li[3,8], Sujie Yan[4,9], Zirong Jiang[5,10]

[1]Computer Science and Technology, Zhejiang University, Hangzhou, 310058, China
[2]computer Science, The university of Glasgow, Glasgow, G12 8QQ, Scotland
[3]SWJTU-LEEDS Joint School, Southwest Jiaotong University, Chengdu, 610031, China
[4]Academy of Intelligent Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[5]Computer Science, The University of Liverpool, Liverpool, L69 3BX, England

[6]826678738@qq.com
[7]769423645@qq.com
[8]3064331616@qq.com
[9]Sujie.Yan22@student.xjtlu.edu.cn
[10]523188577@qq.com
*corresponding author

Abstract. Image classification plays a pivotal role in numerous applications, with substantial implications for daily life, including diagnosing disease from medical images and management of images in autonomous vehicles. However, such sort of research in this field continuously challenges scientists in terms of choosing datasets, testing accuracy, and improvement of models, etc. In this paper, we focus on the performance of two prominent models GoogleNet and residual attention network. We construct two models on the Python platform according to available online resources. To assess their capabilities, we employ the CIFAR-100 dataset, a widely used benchmark dataset. Despite the simplicity of our implementations, GoogleNet comprises approximately 75 convolutional layers and inception modules, and the Residual Attention Network incorporates multiple attention modules within its architecture. These characteristics demonstrate the models' potential for achieving exceptional classification results. Through comprehensive testing and visualization, we aim to provide insights into the efficacy of these models in the context of image classification. Our study contributes to a broader and profounder understanding of their suitability for real-world applications. According to our diagrams and analysis, we conclude that although attention56 is suitable to be adopted in image classification concerning its structure since the model is unstable and invalid in a wide range of training image data on dataset SIFAR100 it might not be exploited in practice. However, as to the model GoogleNet, with an increasing number of training, it obviously is prone to robustness and solid capability of noise resistance. Therefore, GoogleNet is a suitable one to be employed in image classification.

Keywords: GoogleNet, CIFAR100, image classification, Transformer model.

DOI: 10.54254/2755-2721/79/20241537

1. Introduction

Image classification plays a critical role in Computer Vision (CV), Its ability to automatically categorize images has prompted significant social and economic progress. In fields such as healthcare, image classification aids in diagnosing diseases from medical images, improving patient care. Moreover, autonomous vehicles, enable safer navigation and accident prevention. Its impact on diverse sectors underscores its importance as a transformative technology, shaping our modern world.

However, so far, researchers have encountered so many technique issues puzzling them so much in the realm of CV, since it has grappled with challenges that underlines the complexity of interpreting visual data in term of CV data structure itself to conventional CV technique. According to Lai's research [1], one of the primary challenges lies in image recognition, traditional machine learning such as SVM can only applied to deal with one-dimensional data while the images are represented in matrix form. Thereby, researchers have to first stretch the image matrix to a one-dimensional vector and in the process of transformation, it might abandon adjacent information which potentially loses some crucial features as a negative result.

Such low efficiency of extraction of features from images poses huge hurdles for traditional AI to develop further. Fortunately, in recent decades, deep learning has gradually found a more adaptive form in CV. Deep learning such as CNN in CV first represents the base color in the image as a matrix of values and then condenses them into 3D tensor storing stacks of features maps tied to images. To output a 3D tensor, it needs to pass images through a series of convolutional and pooling layers. In this repeated process, images' relevant data and segments will be stored in a smaller representative matrix. Finally, extracted features are sent to a fully connected layer, which generates accurate prediction [2].

On the other hand, the transformer models are another solution that originates from Natural Language Processing (NLP) first introduced in the paper "Attention Is All You Need" by Vaswani et al [3], which has several advantages in NLP tasks. The first one is their scalability, Transformers can be scaled up to handle large datasets and complex tasks effectively. Secondly, the attention mechanism is another merit, the self-attention mechanism allows the model to focus on relevant parts of the input, making it highly interpretable. While transformers were initially designed for sequential data, they have been adapted for image classification tasks. Their advantages make them increasingly outstanding in this realm of CV. Their strong scalability provides scientists with the possibility to further develop models. As highlighted in "Five Reasons to Embrace Transformers in Computer Vision" [4], Vision Transformers, including models like Google's ViT-MoE with 15 billion parameters, have set new records in ImageNet-1K classification. Furthermore, similar to NLP, transformers applied to images provide interpretable attention maps, which can be useful for understanding where the model focuses when making classification decisions. These whole reasons are motivations for scientists to introduce transformers into CV.

To put it into a nutshell, in the realm of CV which empower social development profoundly and prompts economic progress, researchers come across some fatal technique hurdles that conventional machine learning can't deal with perfectly and thus they come up with new method deep learning such as CNN, and transformer to resolve. Owing to their whole new structure, the new advantages they bring make scientists gradually prompt CV development further.

2. Background and Relative Research

The process of image classification involves categorizing input images into predetermined classes, which is a significant undertaking within the realm of computer vision. Conventional approaches to picture classification typically depend on manually crafted feature extraction techniques. However, the emergence of deep learning has led to significant advancements in image classification, particularly through the utilization of Convolutional Neural Networks (CNNs).

By employing multi-layer convolution and pooling operations, Convolutional Neural Networks (CNNs) have the capability to autonomously acquire high-level information from the initial image. By employing deep learning techniques, Convolutional Neural Networks (CNNs) have the capability to effectively extract and analyze various features such as texture, shape, and edges present in an image.

DOI: 10.54254/2755-2721/79/20241537

This enables CNNs to attain enhanced accuracy in the task of image classification. Prominent deep learning models, like AlexNet, VGG, and ResNet, have consistently achieved exceptional outcomes in picture classification competitions, thereby substantiating the superior efficacy of deep learning in the domain of image classification.

The advancement of deep learning has led to the accumulation of extensive research in the field of image classification within AI models.

In the subsequent essay, we shall provide four distinct models.

The ResMLP architecture is a novel approach to picture categorization that is constructed exclusively using multi-layer perceptions. There are numerous perks associated with it. Firstly, it is important to note that in ResMLP, the self-attention layer is substituted by a linear layer. The ResMLP architecture exhibits enhanced stability during training, which can likely be attributed to the presence of linear layers. Notably, linear layers have the advantage of visualizability, enabling the visualization of interactions across patch embeddings [5].

The Vision Transformer (ViT) has demonstrated considerable success in picture classification tasks when compared to traditional Convolutional Neural Networks (CNNs). The transformer architecture offers a flexible and modular framework for constructing and customizing models to meet diverse needs. The attention mechanism of the model facilitates the acquisition of knowledge on the connections between patches, hence enabling the model to gather information at both the local and global levels [6].

MobileNet is a lightweight deep neural network that exhibits a reduced parameter count while achieving superior classification accuracy. These models are characterized by their tiny size, low latency, and low power consumption, making them suitable for a wide range of use cases with specific resource limitations. These techniques have the potential to be further developed and utilized in various applications such as classification, detection, embeddings, and segmentation [7].

Multi-layer perceptron (MLP) models are capable of addressing intricate nonlinear situations. The system effectively manages substantial volumes of input data. Generates rapid forecasts subsequent to the completion of training. It is possible to acquire the same levels of accuracy even when working with smaller sample sizes [8].

In summary, each of these models possesses distinct advantages over the others and can be effectively employed in various contexts.

Transformer models have achieved immense success in natural language processing. BERT (Bidirectional Encoder Representations from Transformers) is an example of an impressive pre-trained Transformer model that achieved significant improvements across many natural language processing tasks - text classification, named entity recognition, and question answering among them [9]. Through pre-training on massive text datasets it achieved impressive performance gains [10-11]. GPT-3 (Generative Pre-trained Transformer 3), another powerful Transformer model, gained widespread attention for its extraordinary text-generating abilities. Transformers have proven themselves capable of creating high-quality articles, conversations, and code to demonstrate their immense power as text generators [10]. Furthermore, these models have demonstrated notable achievements in cross-domain transferability [11]. Baidu introduced the ERNIE model (Enhanced Representation through Knowledge Integration), which is a Transformer-based pre-trained language model trained on extensive multi-domain data for superior performance across various natural language processing tasks. ERNIE excels at cross-domain transfer. In particular, in the medical domain, ERNIE has demonstrated impressive cross-domain results by pre-training on general domain data before fine-tuning medical text tasks later on. ERNIE stands out for its capacity for knowledge transfer across different domains [11]. Facebook recently unveiled their Cross-Lingual Language Model with Recalibrated Cross-Langual Training, also built upon Transformer architecture, known as XLM-R (Cross-Lingual Language Model with Re-calibrated Cross-Langual Training). Pre-training on multilingual data enables XLM-R to effectively engage in cross-domain tasks involving different languages, most notably cross-lingual sentiment analysis tasks; it excels particularly in this regard due to being capable of moving these tasks even with limited data in target languages [12].

DOI: 10.54254/2755-2721/79/20241537

Transformers may have initially been developed to perform natural language processing tasks; however, their unique model architecture and self-attention mechanism have found widespread applications within image processing as well. Studies demonstrate how transformer models can effectively capture global dependencies within image data, improving tasks such as recognition and segmentation. Transformer models' self-attention mechanism enables them to model relationships among positions in input data--something immensely helpful for object recognition and feature extraction in images. By making necessary modifications, Transformer models can also be adjusted specifically to image processing applications for faster extraction of local or global features in images thereby improving the performance of image-related tasks [13].

Transformer models' unique encoder-decoder structure gives them an advantage when dealing with sequential data, where conventional models require manual configuration of feature extractors and decoders whereas Transformer models automatically learn key features of input data through self-attention mechanisms to generate output sequences during decoding - providing exceptional results across a broad spectrum of tasks across many domains.

Transformer models have achieved extraordinary success in natural language processing, while their distinctive encoder-decoder structure and self-attention mechanism open them up for applications in image processing. Transformer models excel at extracting both local and global features from images efficiently - driving advancement in image processing technologies.

3. Methodology and Technology

The architecture of the Transformer model can be categorized into four distinct modules: the Input module, the Coding module, the Decoding module, and the Output module. The construction of the input module comprises two main components: the source text embedding layer and its corresponding position encoder, as well as the target text embedding layer and its corresponding position encoder. The architecture of the encoder module is as follows: The structure consists of a series of N encoder layers that are layered together. Each layer of the encoder is composed of two sublayers that are interconnected. The initial sub-layer connection architecture comprises a multi-head self-attention sub-layer, a normalization layer, and a residual connection. Similarly, the subsequent sub-layer connection architecture encompasses a feedforward fully connected sub-layer, a normalized layer, and a residual connection. The subsequent component is the decoder module, which is comprised of a series of N-stacked decoder layers. Every decoder layer is composed of three interconnected sublayer structures. The initial sub-layer connection structure comprises a multi-head self-attention sub-layer, a normalization layer, and a residual connection. The second sub-layer connection structure comprises a multi-head attention sub-layer, a normalization layer, and a residual connection. The third sublayer connection structure comprises a feedforward fully connected sublayer, a normalized layer, and a residual connection. The final component of the module structure is the output module, which consists of a linear sheaf and a SoftMax layer.

The notable aspect of this approach lies in the effective integration of a highly parallelizable decomposable attention mechanism with a feedforward network. This finding suggests that attention mechanisms possess inherent strength and that the sequential recurrent processing of data is not a must for attaining the quality improvements observed in recurrent neural networks (RNNs) with attention. Jakob Uszkoreit, a researcher from Google, then suggested the substitution of Recurrent Neural Networks (RNNs) with self-attention mechanisms, so initiating the endeavor to assess the viability of this concept. Transformers employ an attention method to simultaneously process all tokens, wherein they compute "soft" weights between tokens in consecutive levels. The attention mechanism exclusively relies on information from lower layers to compute its output. Consequently, it may be efficiently computed for all tokens simultaneously, resulting in enhanced training speed.

DOI: 10.54254/2755-2721/79/20241537

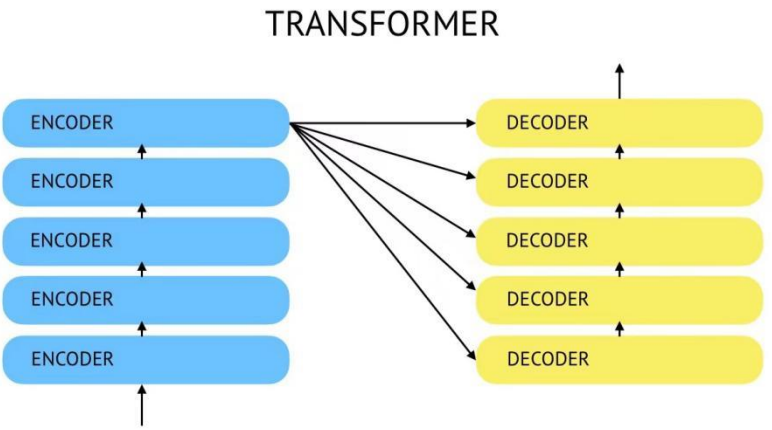

Figure 1. Transformer structure

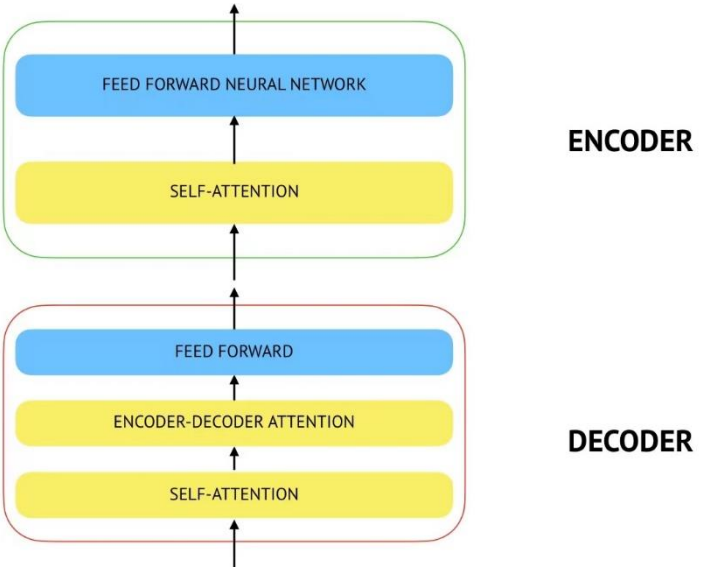

Figure 2. Encoder-decoder structure

GoogleNet stands out as an innovative deep convolutional neural network architecture thanks to its Inception structure, designed to increase performance during image classification tasks by simultaneously capturing features at various levels using multi-scale convolutional kernels.

GoogleNet's Inception Structure offers an efficient sparse feature representation. By extracting information at different scales via multiple convolutional kernels and then fusing them together, a more complete representation of an image can be reached. Furthermore, this structure introduces various convolutional kernels of differing scales parallelly capturing distinct features within images, further improving accuracy in classification accuracy.

Inception Module is the basic building block, which is composed of four components: 1x1 convolution, 3x3 convolution, 5x5 convolution, and 3x3 max pooling. In the end, however, the results are concatenated along the channel dimension, which is the core concept of the Inception module. The basic structure of Inception is realized by stacking multiple convolution layers, pooling layers, and 1x1 convolution layers. Different features in the image are captured by using convolution kernels of different sizes at each layer and then connected along the channel dimension. Through this structure, the network can learn features at different scales at the same time to achieve a more comprehensive representation of image content.

DOI: 10.54254/2755-2721/79/20241537

GoogleNet exhibits several unique characteristics when used for image classification tasks:

1. Multi-scale Feature Extraction: Inception employs convolutional kernels of differing sizes to capture both local and global features in images simultaneously, providing the model with enough power to detect objects of various sizes within them and provide accurate identification capabilities for object identification purposes.

2. Parameter Efficiency: With 1x1 convolutional layers, Inception's structure helps minimize network parameters by mitigating the risk of overfitting while speeding up both training and inference processes.

3. Information-Rich Feature Representation: By concatenating features across scales along a channel dimension, networks are able to learn a more information-rich feature representation which enhances their expressive capacity by better distinguishing among image categories.

4. Mitigating Gradient Vanishing: Gradient vanishing can have detrimental effects on training stability and convergence speeds for traditional deep networks, but Inception structures provide intermediate layers with varied depths and scales, enabling gradients to flow more freely while mitigating gradient vanishing issues.

To sum up, GoogleNet's Inception structure has the characteristics of high parametric efficiency and multi-scale feature extraction information, which can enhance the effect of image classification tasks. Meanwhile, its parallel structural network can further improve the performance of image classification by capturing features at different levels.

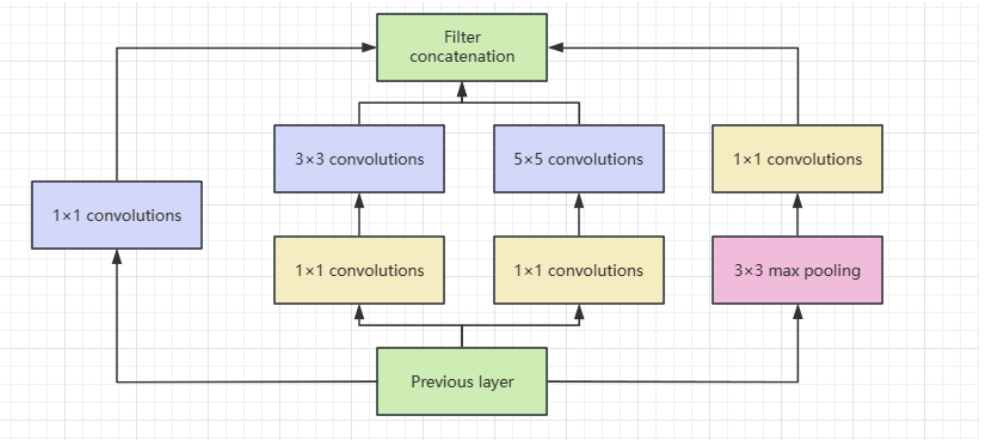

Figure 3. Inception module with dimension reduction

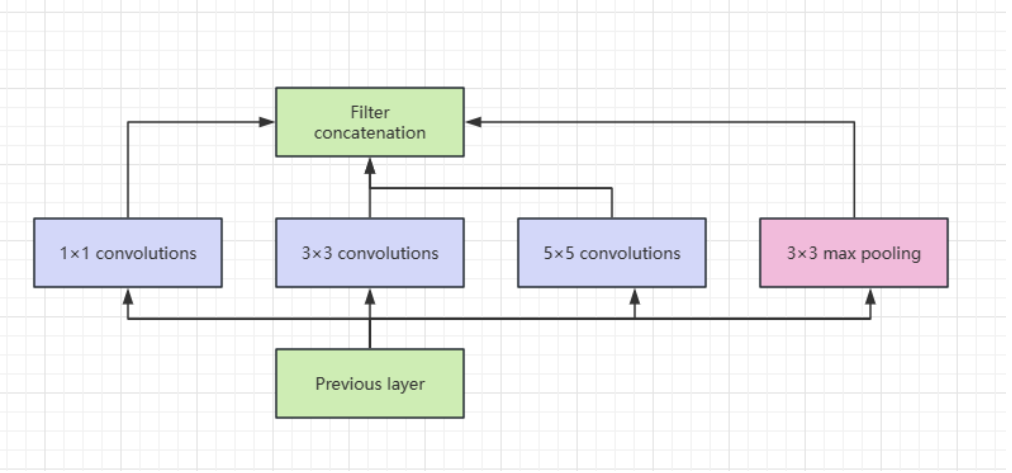

Figure 4. Inception module structure

DOI: 10.54254/2755-2721/79/20241537

4. Experiment

The CIFAR-100 Dataset plays a critical role in our experiment. Standing as a cornerstone in the realm of CV, it encapsulates several key attributes so that we consider it as our experimental dataset. To start with, according to a document from Krizhevsky, A [14], as fundamental structure, it owns 100 classes containing 600 images in each class. With each class, there are 500 training images and 100 testing images. The 100 classes in the CIFAR-100 are further divided into 20 superclasses. Each individual image is associated with a "fine" label (the specific class membership) and a "coarse" label (broader superclass affiliation). Meanwhile, designers offer various versions of CIFAR-100 to groups of users with different demands to import and exploit. Furthermore, inside each individual class, a whole group of images are associated with the real world such as vehicles, animals, and so on. As a result, it can assist users in testing algorithms' real-world applicability. On the other hand, with 20 superclasses and 5 subclasses each, such complexity challenges the model with both robustness and precision in classification with diverse categories of images. These merits make us exploit it as our experimental datasets.

As to the choice of models, we select residual attention networks and GoogleNet as training targets. Considering our research orientation, GoogleNet is expected since its efficiency, power, and memory use prompts it to outcompete other models. Additionally, the residual attention network is an improved convolutional LSTM algorithm, to improve the accuracy of information by extracting salient regions of images efficiently [15]. In other words, their appealing qualities have garnered attention, making them a trending topic in image classification.

In terms of architecture, we first implement the GoogleNet model according to the experiment of Szegedy et al [16]. This architecture mainly consists of three parts: prelayer, inception module, and maxpool. To begin with, the prelayer serves as the initial extraction of features with batch normalization, ReLU activation, and 3 convolutional layers totally. Moreover, the maxpool layer with 2 strides functions as max-pooling operations that downsample the spatial dimensions of the feature maps. Lastly, the main purpose of the inception module is to capture features at various scales and combine them. The module which is the main building of GoogleNet contains multiple parallel convolutional layer pathways along with different sizes of kernels. In our experiment, we contain a total of 8 convolutional layers. To sum up, the whole structure contains one prelayer with 3 convolutional layers, 9 inception modules with 8 layers each, and 75 layers in total.

In terms of the architecture of the transformer model, we construct such a model according to Wang et al [17]. This architecture has two different versions attention56 and attention92, they are both composed of attention networks. Inside the network, there are several attention modules. To implement the module, we first need one preprocessing-residual unit before splitting the trunk branch which functions to perform feature processing, and the mask branch serving as a feature selector during forward inference, and gradient update filter during backpropagation. Meanwhile, the truck branch consists of two residual units and one between two adjacent max-pooling layers with a sigmoid activation function by default. Thereby, the number of them can be altered. On the other hand, the fabric of the residual attention network contains one pre-convolutional layer to process input initially. Moreover, the four stages and first three stages contain three attention modules respectively to capture features and prepare for classification. Besides, there is another global average pooling layer that aggregates spatial information for each channel. Lastly, the linear layer performs the last classification. In general, this architecture is designed to extract features at various scales and levels of abstraction and then transform them to enable effective classification.

After several repeated training along with testing, we can gradually depict a relatively exact diagram with a sea of collected data and we can find some extraordinary traits of the GoogleNet. In a general view of four curves, the GoogleNet gradually convergences to a comparatively high accuracy and low average loss rate in the ladder pattern which proves that the two models are able to prevent error increasingly efficiently. Meanwhile, their fluctuation rates all tend to smoothness with increasing number of training with respect to variance. This reflects the two models' stability and robust noise resistance ability. As to the regional curve of each graph, we can generalize 60 epochs as a period. Every

DOI: 10.54254/2755-2721/79/20241537

time, GoogleNet finishes a period, their accuracy will get hugely enhanced and the loss rate is going to get eliminated obviously. Before a threshold period, the loss rate and accuracy rate will get local peaks and valleys. Such sudden increase or decrease possibly means that this model might learn some useful data and make themselves better at dealing with noise information.

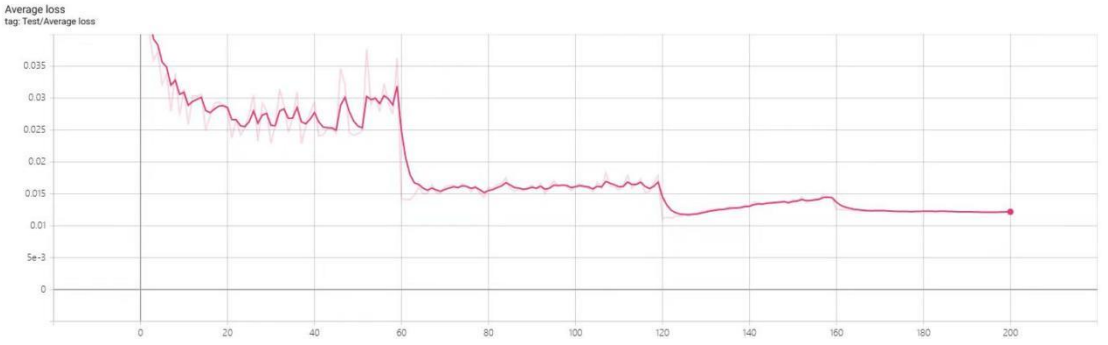

Figure 5. Loss rate of GoogleNet model

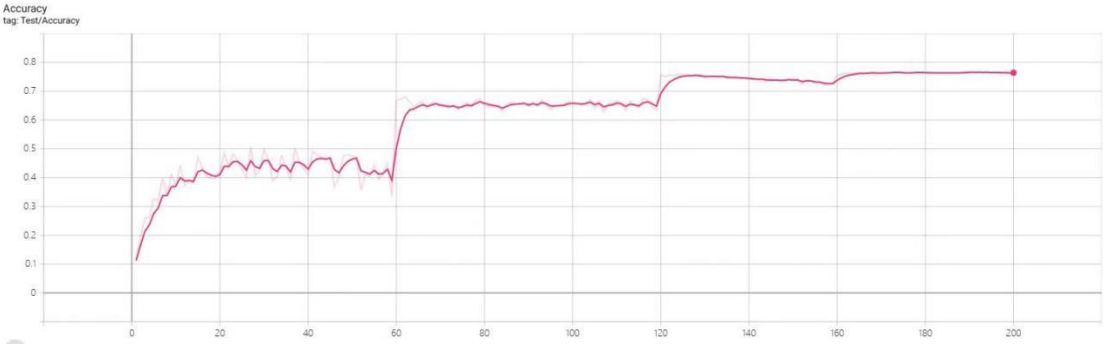

Figure 6. Accuracy of GoogleNet model

However,when we analyze the attention56 curves generated from training data, it has a tremendous difference from the analysis of the GoogleNet model's curve. In general, these two graphs don't have any convergence and spots are likely to distribute randomly. Moreover, the model merely fluctuates unexpectedly and the x coordinates show that the model runs correctly until the 50th epoch finishing which doesn't correspond to our expectation. Although we attempt to resume it with a series of tries, no indication shows the model can operate correctly after the 50th epoch. In a specific view of these two graphs, as to the accuracy and loss graph, its maximum point is close to 0.14 and 0.055 respectively which are still unexpected.

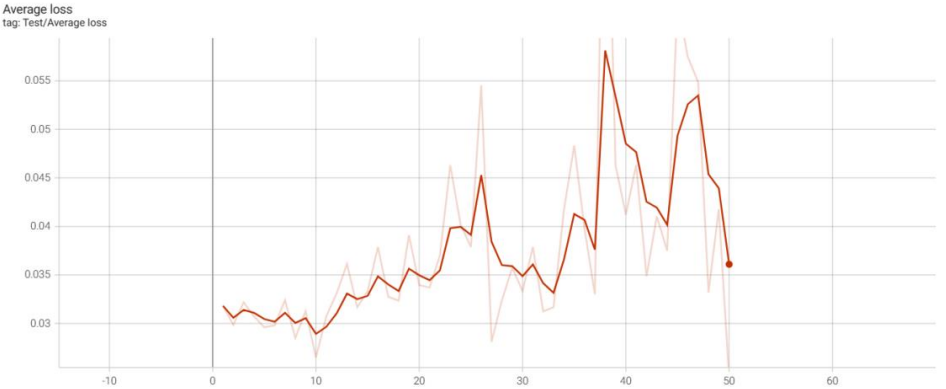

Figure 7. Loss rate of attention56 model

DOI: 10.54254/2755-2721/79/20241537

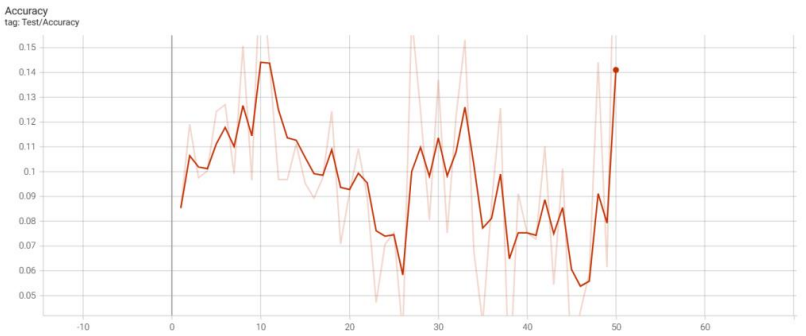

Figure 8. Accuracy of attention56 model

We conducted a comprehensive comparison between our Transformer network (Attention59) and the state-of-the-art GoogleNet using the CIFAR-100 dataset. The table below summarizes the key metrics:

Table 1. Comparisons Transformer with GoogleNet on CIFAR-100

Network	dataset	params	convergence	accuracy (highest)	loss (lowest)
GoogleNet	cifar100	55.7M	11	0.5025	0.02291
Attention59	cifar100	6.2M	-	0.2207	0.025

The convergence refers to the number of iterations when a relatively stable accuracy is achieved. Since the learning rate was changed during the experiment, the first learning rate was the first 60 epoch samples (In fact, it is almost impossible to observe convergence since the performance of the Transformer is so unstable). We can clearly see that model GoogleNet performs much better than Attention 59 in terms of convergence and accuracy of final convergence.

In general, according to the behaviors of the two models on dataset CIFAR100, it's more likely that the GoogleNet model is a more suitable one to be adopted in the classification of images since it can be implemented and trained with a wide range of sample data validly. Besides, GoogleNet's capability of noise resistance and robustness is heavily proven. After several periods of training, it evidently convergence to relatively high accuracy and low loss rate, and its fluctuation rate is gradually controllable. However, as to residual attention network, it is unable to accept a sea of image data. Its practice is, thus, not supported.

To sum up, although the chosen models have potential, various models still have their own limitations and merits which means they might be skilled in different areas. As GoogleNet has a flexible scale and level to extract features, it is increasingly ideal for researchers to collect data from wide categories of abstract entities. Furthermore, its multiple-layer structure can enhance the accuracy of extraction in models. Finally, its error calculation and correction of error methods are beneficial to prevent overfitting and underfitting possibilities. The whole traits make researchers find the huge potential of this sort of model in image classification.

5. Conclusion

Numerous factors contribute to variations among different models. These factors include data preprocessing and organization, neural network model selection and construction, choice of loss and optimization functions, and parameter tuning. Data preprocessing is crucial due to common issues like incomplete, noisy, inconsistent, redundant, imbalanced, outlier-prone, and duplicate data. Model architecture significantly impacts performance, and selecting appropriate loss and optimization functions improves training outcomes. Therefore, optimizing these aspects is essential for successful model application.

In conclusion, Transformer models have shown remarkable potential in computer vision, particularly in image classification, object detection, and image generation tasks. Their versatility extends to

DOI: 10.54254/2755-2721/79/20241537

addressing various visual challenges. Future research should focus on optimizing pre-training techniques, enhancing position coding methods, exploring novel adaptation approaches, and improving fine-tuning stability. As technology evolves, innovative architectural designs and attention mechanisms, along with powerful pretraining and data augmentation strategies, will continue to shape the future of Transformer-based models in computer vision. We look forward to further advancements in this field to address real-world problems.

Acknowledgments
Xinran Xie, Qinwen Yan, Haoye Li, Sujie Yan, and Zirong Jiang contributed equally to this work and should be considered co-first authors.

References
[1] Lai, Y. (2019, October). A comparison of traditional machine learning and deep learning in image recognition. In Journal of Physics: Conference Series (Vol. 1314, No. 1, p. 012148). IOP Publishing.
[2] "Deep Learning for Computer Vision." (Year, Month Day). Run: AI. URL: https://www.run.ai/guides/deep-learning-for-computer-vision#Deep-Learning-Architectures-for-Computer-Vision
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systms, 30.
[4] Microsoft Research Asia. (2021, December 5). Five reasons to embrace Transformer in co mputer vision. Microsoft Research. URL:https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/articles/five-reasons-to-embrace-transformer-in-computer-vision/
[5] Touvron, Hugo et al. “ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training.” IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2021): 5314-5321.
[6] Chen, C., Fan, Q., & Panda, R. (2021). CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 347-356.
[7] Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., & Liu, Z. (2021). Mobile-Former: Bridging MobileNet and Transformer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5260-5269.
[8] Pal, S.K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. IEEE transactions on neural networks, 3 5, 683-97.
[9] [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding] (https://arxiv.org/abs/1810.04805)
[10] [Language Models are Few-Shot Learners] (https://arxiv.org/abs/2005.14165)
[11] Zhang, S., Xu, X., Liu, J., Huang, Y., & Zhu, X. (2019). ERNIE: Enhanced Representation through Knowledge Integration. arXiv preprint arXiv:1904.09223.
[12] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Joulin, A. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
[13] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV). [Link: https://arxiv.org/abs/2005.12872]
[14] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.
[15] Ge, H., Yan, Z., Yu, W., & Sun, L. (2019). An attention mechanism based convolutional LSTM network for video action recognition. Multimedia Tools and Applications, 78, 20533-20556.
[16] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

DOI: 10.54254/2755-2721/79/20241537

[17] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., ... & Tang, X. (2017). Residual attention network for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).