Proceedings of Machine Learning Research vol 245:1–9, 2024 2024 International Conference on Machine Learning and Intelligent Computing

Research on Features Extraction and Classification for Images based on Transformer Learning

Chao Wang* APK66578@163.COM
Artificial Intelligence, Shanghai Normal University, Shanghai, 200233, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

Image processing and analysis have become an essential method in many areas including medical impact, facial recognition, and social media analysis. With the rapid development of big data and artificial intelligence technology, especially the emergence of Transformer learning models, new methods have been brought to image feature extraction and classification. However, the existing transformer model limits the ability to handle variable-length sequences and understand complex sequence relationships. In this work, we propose a novel transformer-based framework that combines a self-attention mechanism and a multi-head attention technique to efficiently extract features from complex image data. In addition, we introduce an improved classifier that enables efficient image classification using extracted features. Our method takes into account not only the local features of the image but also the global relationships between different regions to achieve a more accurate representation of the features. We simulate our model with existing convolutional neural networks and other traditional machine learning methods in the public datasets including CIFAR-10 and MNIST. From our experimental results, we can observe that our transformer-learning-based framework shows significant performance improvement in image feature extraction and classification tasks.

Keywords: Features extraction, Image classification, Transformer learning, Self-attention mechanism

1. Introduction

Image feature extraction and classification constitute the core and cornerstone of the field of computer vision and image processing, and carry high research value and practicability. These processes are not only the fate of many computer vision and image processing problems, but also their performance directly determines the success of subsequent computer vision tasks. In view of the fundamental and critical nature of this field, image feature extraction and classification technology is shown through almost all practical applications of computer vision and multimedia systems, reflecting its extensive influence and indispensable position in modern digital life (Chen et al., 2021). Image feature extraction occupies a fundamental and critical position in the field of computer vision and image processing, and is the core technology that supports any image or video-related intelligent vision system. Across the computer vision field, the quality of image features is generally considered to be a key factor affecting the upper-performance limit of computer vision systems (Lanchantin et al., 2021). Based on the precisely extracted image features, the role of subsequent learning, clustering, or classification algorithms is to gradually approach the theoretical upper limit of this performance.

As a rich and intuitive information carrier, image data plays an increasingly important role in various fields. From image sharing on social media to geographic analysis of satellite imagery to

disease diagnosis with medical images, the applications of image data are wide and far-reaching (Wang et al., 2021). With the rapid development of computer vision technology, how to effectively extract valuable information from these images, and accurately classify and identify them, has become a major challenge for scientific research and industry. Feature extraction and classification are key technologies to achieve this.

Feature extraction, that is, extracting key information from the original image that can represent the content and attributes of the image, is the first step in image processing. These features may include colors, shapes, textures, or more complex patterns that provide the basis for further analysis and understanding of the image (Wang et al., 2020). This is followed by the classification process, which uses the extracted features to group images into predefined categories. This process involves complex algorithms and models, such as support vector machines, neural networks, and more recently, the popular deep learning method. However, with the rapid increase in the amount of image data and the complexity of image content, traditional feature extraction and classification methods are facing great challenges (Sarwinda et al., 2021). How to design efficient and robust algorithms that not only accurately identify and classify images, but also be able to process large-scale datasets and adapt to changing application needs has become a hot topic of research.

One of the core challenges in the broad field of computer vision and image processing is image classification. For example, object detection is essentially a binary classification problem, while face recognition and gait recognition fall under the category of image classification of biometric features (Rao et al., 2021). Scene understanding can be understood as a classification task for scene images, and even some more advanced image feature extraction challenges, such as attribute learning, can also be regarded as multi-label classification problems (Hong et al., 2021). The key to solving the problem of image classification lies in the accurate extraction of image features and the selection of appropriate classifiers.

Traditional single-image feature representation methods including image grayscale or color information, edge contours, image gradient features, scale-invariant features, and dictionary models, have been widely studied (Zhang et al., 2020). However, in real-world application scenarios, face recognition technology that relies on a single image often performs poorly when dealing with complex variables such as lighting, posture, expression, age, and gender due to factors such as imperfection of the image acquisition process, image diversity, and noise interference (He et al., 2021). Although researchers have proposed many algorithms to deal with the challenges of face recognition under these complex conditions, and have achieved rich research results, most of these studies are still limited to the specific constraints of the laboratory, and are difficult to meet the needs of practical applications. Especially in the field of video surveillance, due to the complexity of the environment and the unsatisfactory quality of the collected images, the recognition efficiency of face recognition technology based on a single image is significantly reduced.

2. Related work

Researchers effectively used the reverse training strategy to extend the binary classifier and proposed an innovative multi-class image set classification method. Considering that most studies convert image set data into vectors, which seriously destroys the intrinsic spatial structure information of image samples, researchers proposed a tensor-based discrimination dictionary learning method. In this method, the image set is regarded as a tensor, containing two spatial modes and one set number

mode, and seeks to obtain three dictionaries of the corresponding modes, and finally classifies them by minimizing the reconstruction error.

Subsequently, researchers effectively used the compensation information between datasets by combining forward and backward sparse representations, so as to achieve excellent classification performance. In view of the huge storage and time overhead of directly using the original image for modeling and computation when processing large-scale image set data, researchers introduced hash encoding for image set classification. They learned a compact hash code representing the image set to eliminate the redundant characteristics of the data, and the model further considered the discrimination characteristics of the data inside and outside the class and finally achieved a good classification effect and fast operation efficiency.

Additionally, a convolutional neural network (CNN) is a deep learning model designed to process data with a grid structure, such as images, that automatically learns spatial features from the data through its unique hierarchical structure—including convolutional, activated, pooling, and fully connected layers. These layers automatically extract key features of the image, such as edges and textures, through sliding filters, and help the network understand and classify complex patterns in the image by introducing nonlinearity and reducing feature dimensions. Convolutional neural network can automatically identify and abstract important information in images, so as to achieve efficient and accurate results in image classification, facial recognition, and other fields, greatly simplifying the feature extraction and classification process in traditional image processing.

3. Methodologies

3.1. Notions

Above all, we conclude the primary parameters of our proposed methods in following Table 1.

Table 1: Primary notions.

Symbols	Explanations
$I(x, y)$	Input images
$m(x, y)$	Magnitude of the gradient
$\theta(x, y)$	Direction
f	Descriptor vector
α	Gaussian blurs

3.2. Feature extraction mechanism

The Scale Invariant Feature Transformation (SIFT) algorithm is a feature extraction tool widely used in image processing and computer vision, especially in tasks such as object recognition, image matching, and image retrieval. First proposed by David Lowe in 1999, SIFT was designed to extract scale-invariant feature points from an image and generate a unique descriptor for each feature point, which could be used to match between different images and remain stable even with changes in scale, rotation, brightness, and even some degree of viewing angle.

The following items describe the detailed procedures of feature extraction.

- Extreme value detection in scale space: The difference in the Gaussian function is used to find potential key points in different scale-spaces. The purpose of this step is to identify key point locations that remain constant at all scales.
- Key Point Positioning: Precisely locate the position and scale of key points around potential key points detected in the first step, and remove key points with low contrast and key points with large edge response to enhance matching stability and noise immunity.
- Direction Assignment: Assign one or more directions to each key, based on the direction of the local image gradient. This step is to increase the rotation invariance of the key descriptor.
- Key descriptor generation: In the neighborhood around each key, according to the scale of the key, select an area of appropriate size, and calculate the direction and size of the local gradient in this area, and then this information is used to construct the feature descriptor of the key. Descriptors are high-dimensional vector representations of gradient information around key points, which are designed with good invariance properties.

Gaussian difference function: This is achieved by using Gaussian blurs α different scales on the same image, and then calculating the difference between these blurred images. For both scales α and $k\alpha$ from image $L(x, y, k\alpha)$ and $L(x, y, \alpha)$, the gaussian difference function can be described as following Equation 1.

$$D(x, y, \alpha) = L(x, y, k\alpha) - L(x, y, \alpha) \tag{1}$$

Where the $L(x, y, \alpha)$ is equal to $G(x, y, \alpha) * I(x, y)$. The calculation of $G(x, y, \alpha)$ is a two-dimensional Gaussian function with a scale of α, and $*$ represents a convolution operation, and $I(x, y)$ is the original image.

Additionally, the principal orientation of a key is determined by calculating the gradient direction and size of each pixel within the key's neighborhood. The magnitude of the gradient $m(x, y)$, and the direction $\theta(x, y)$ can be calculated by the following Equation 2.

$$\begin{aligned} m(x, y) &= \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \\ \theta(x, y) &= \arctan \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \end{aligned} \tag{2}$$

Subsequently, according to the gradient direction, the gradient size is added to the direction histogram, and the direction corresponding to the peak of the histogram is the main direction of the key. For added stability, the local maximum in the histogram is often chosen as the orientation of the keys, which allows each key to have multiple directions, resulting in a descriptor for each direction.

In order to construct a descriptor of a key, first take a scale-dependent neighborhood around the key, then divide this neighborhood into several small blocks with a 4x4 grid, and for each small block, calculate the gradient direction histogram of the pixels inside it by using 8 directions, so that each small block can produce an 8-dimensional vector. Therefore, for each key, its descriptor will be a collection of these vectors, forming a 128-dimensional eigenvector. This eigenvector is normalized to provide robustness to lighting changes.

The construction formula of the descriptor can be summarized as follows: for each pixel in a small block, calculate its gradient direction θ' and amplitude m' relative to the direction of the key point, and then accumulate it into the corresponding direction histogram according to θ'. The parameters θ' and m' are the adjusted gradient direction and amplitude that take into account the direction of the key points to ensure rotation invariance. The normalization of descriptor vectors is an important step to reduce the impact of lighting changes on feature matching. This can be achieved by dividing the eigenvector by its modulus length.

$$f' = \frac{f}{||f||} \tag{3}$$

Where parameter f is the descriptor vector, $||f||$ is the modulo length of f, and f' is the normalized descriptor vector. This series of steps of the SIFT algorithm enables it to extract feature points and descriptors with high invariance, which is suitable for stable feature matching between different images. These feature points not only remain unchanged to scale and rotation changes, but are also robust to changes in viewing angles, affine transformations, and lighting changes to a certain extent.

3.3. Transformer framework

In this section, we delve into the intricacies of our proposed transformer model, designed to address the unique challenges of image feature extraction and classification. Our model leverages the transformer architecture, renowned for its ability to process sequential data, and adapts it for the complex domain of image analysis. The core components of our model include:

3.3.1. SELF-ATTENTION MECHANISM

At the heart of Transformer is the self-attention mechanism, which allows the model to take into account all the patches in the image when processing each patch. The self-attention mechanism dynamically adjusts the representation of patches by calculating attention scores between patches, allowing the model to focus on key parts of the image.

The cornerstone of our transformer model is the self-attention mechanism. This innovative approach allows the model to consider every patch of the input image simultaneously, unlike traditional models that process parts of the image sequentially. By calculating attention scores that measure the importance of each patch in relation to others, the model dynamically adjusts the representation of each patch. This process enables the model to focus selectively on the most informative parts of the image, enhancing its ability to extract relevant features for classification tasks.

To provide a clearer understanding of the self-attention mechanism, we define the attention score calculation as follows:

- Each image patch is represented by a vector of features. For a given patch, we compute its query vector (Q), and for every other patch, we compute key (K) and value (V) vectors.
- The attention score between a query and a key is determined by the dot product of their vectors, followed by a scaling factor to control the variance of the scores.
- A softmax function is applied to the attention scores to ensure they sum up to one, turning them into probabilities that determine the significance of each patch's features.
- The final output for each patch is a weighted sum of all value vectors, weighted by their attention scores, allowing the model to aggregate information across the entire image.

3.3.2. MULTI-HEADED ATTENTION

Our model extends the self-attention mechanism through the use of multi-headed attention. This technique involves running several instances of the self-attention mechanism in parallel, each with its own set of parameters and focusing on different subspaces of the feature representation. This parallelism allows the model to capture a richer and more diverse set of relationships between image patches, improving its ability to recognize complex patterns and features within the image data.

Specifically, our model employs N parallel attention heads, where N is a hyperparameter tuned based on experimental validation. Each head computes its own set of attention scores, providing unique contributions to the final output. The outputs of all heads are concatenated and linearly transformed to match the original feature vector size, ensuring a comprehensive representation of the image.

3.3.3. FEEDFORWARD NEURAL NETWORK

Following each multi-headed attention block, our model includes a position-wise feedforward neural network (FFNN). This network consists of two linear transformations with a ReLU activation in between. The FFNN serves to process the output of the attention mechanism at each position (i.e., each patch) independently, allowing for the introduction of non-linearity into the model's representation and enhancing its ability to capture complex relationships between features.

The FFNN is defined as follows:

- The input to the FFNN is the aggregated output from the multi-headed attention mechanism.
- A first linear layer expands the dimensionality of the input, followed by a ReLU activation function to introduce non-linearity.
- A second linear layer then compresses the representation back to its original dimensionality, preparing it for subsequent layers or output.

3.3.4. MODEL ARCHITECTURE OVERVIEW

Figure 1 provides a schematic representation of our transformer model. As illustrated, the model processes input images through a series of layers, each comprising a multi-headed attention mechanism followed by a position-wise feedforward neural network. This design allows the model to iteratively refine its representation of the image, capturing both local and global features essential for accurate feature extraction and classification.

4. Experiments

4.1. Experimental setups

Experimental analysis using CIFAR-10 and MNIST are two widely used benchmark datasets to evaluate the performance of image recognition and handwritten number classification algorithms, respectively. CIFAR-10 contains 60,000 color images of 32x32 pixels divided into 10 categories, while MNIST contains 70,000 grayscale images of 28x28 pixels representing handwritten numbers from 0 to 9.

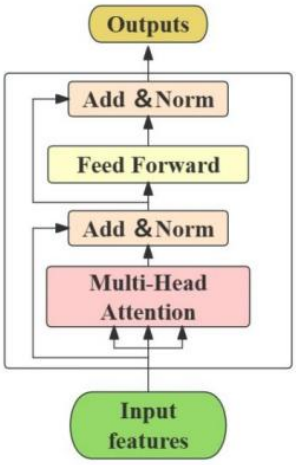

Figure 1: Framework of proposed transformer framework.

4.2. Experimental analysis

Classification accuracy is one of the most intuitive and commonly used metrics to measure the performance of a classification model. It is defined as the ratio of the number of correctly classified samples to the total number of samples, reflecting the model's classification accuracy on the overall dataset. Following Figure 2 depicts the classification comparison results.

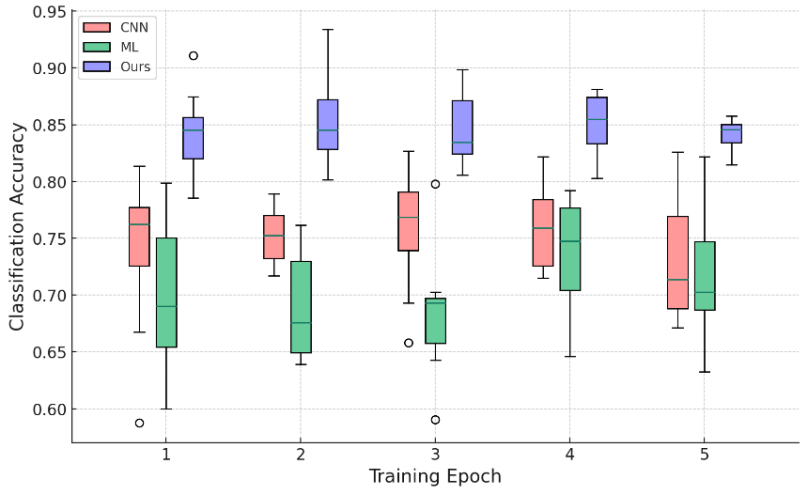

Figure 2: Classification accuracy comparison results.

Computational cost is a multi-faceted indicator, which not only reflects the efficiency of the algorithm, but also directly relates to the feasibility and scalability of the algorithm. Reducing computational costs is one of the key factors in improving the usability of algorithms when selecting or designing algorithms. Following Figure 3 shows the computation costs comparison results.

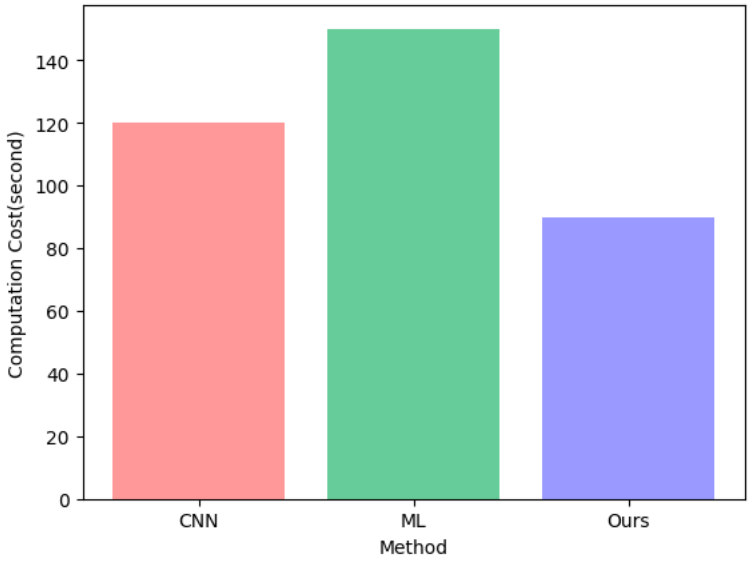

Figure 3: Computation cost comparison results.

5. Conclusion

In conclusion, the proposed Transformer-based model significantly enhances feature extraction by integrating global context information, resulting in superior classification performance in datasets such as CIFAR-10 and MNIST. Despite potential computational cost concerns, our optimized Transformer model demonstrates competitive efficiency and adaptability to a wide range of visual tasks. This research not only highlights the transformative potential of applying Transformer architectures in the field of computer vision, but also lays the groundwork for future research to further optimize these models and extend their applications to broader challenges in machine learning. These promising results provide a wealth of avenues for innovation in image processing technology and its applications.

References

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356, 2021. doi: 10.1109/ICCV48922.2021.00041.

Xin He, Yushi Chen, and Zhouhan Lin. Spatial-spectral transformer for hyperspectral image classification. *Remote Sensing*, 13(3), 2021. ISSN 2072-4292. doi: 10.3390/rs13030498.

Danfeng Hong, Lianru Gao, Jing Yao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5966–5978, 2021. doi: 10.1109/TGRS.2020.3015157.

Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16473–16483, 2021. doi: 10.1109/CVPR46437.2021.01621.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, and Pinkie Anggia. Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. *Procedia Computer Science*, 179:423–431, 2021. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.01.025. 5th International Conference on Computer Science and Computational Intelligence 2020.

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 943–952, 2021. doi: 10.1109/CVPR46437.2021.00100.

Weibin Wang, Dong Liang, Qingqing Chen, Yutaro Iwamoto, Xian Hua Han, Qiaowei Zhang, Hongjie Hu, Lanfen Lin, and Yen Wei Chen. *Medical Image Classification Using Deep Learning*, pages 33–51. Springer International Publishing, Cham, 2020. ISBN 978-3-030-32606-7. doi: 10.1007/978-3-030-32606-7_3.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12200–12210, 2020. doi: 10.1109/CVPR42600.2020.01222.