




The Design of Drawing Layout for Young Children in the Context of Artificial Intelligence

Hua Xie 

Academy of Fine Arts, Zhengzhou Preschool Education College, Zhengzhou 450000, China,
xiehua@zzpec.edu.cn

Corresponding author: Hua Xie, xiehua@zzpec.edu.cn

Abstract. Image feature extraction is one of the current research hotspots in computer vision and a key tool for aiding computer-aided drawing in artificial intelligence, which has potential applications in helping children learn drawing layouts. Despite this, few studies have investigated the application of image feature extraction to children's drawing layouts. Using computer-aided technology science to support early childhood drawing education is challenging. Consequently, this paper presents a detailed review of the critical technologies of image feature extraction and proposes E-Attention based on linear attention and token pruning. It is implemented in the Transformer model, and experimental results have demonstrated that the method can be used to design an efficient image feature extraction network based on Transformer. According to the experimental results, the technique is capable of designing an efficient Transformer-based image feature extraction network, reducing the computation effort by 60%-70%, as well as the overall model performance by about 1.5%-3%. The proposed method can accurately extract the drawing features of young children and assist, thereby optimizing the drawing teaching patterns.

Keywords: drawing layout, computer-aided technology, young children, image feature extraction.

DOI: <https://doi.org/10.14733/cadaps.2023.S5.180-189>

1 INTRODUCTION

Computer vision technology uses computers to simulate human visual intelligence to make meaningful judgments about the image targets and scenes it perceives, including image target recognition, scene classification, image classification, image emotion semantic analysis, etc. The key and core of the computer that can make correct judgments are to select the features that best express the target and essence of the image, which in turn can better understand the image itself and its connotation. The pictorial image is an essential artistic creation and national cultural heritage, which is a unique and wonderful way for human beings to understand the world by using beautiful forms of representation, reflecting the face of human culture, and conveying human

thoughts and emotions. For thousands of years, human beings have created a large number of painting images. For the massive amount of paintings with different artistic styles, research and analysis using advanced artificial intelligence algorithms, machine learning algorithms under big data, etc., have become a hot research topic in image processing and computer vision. Current research results show that by using image analysis and feature extraction techniques, computers can learn a large amount of knowledge from paintings, extract effective features of paintings, and realize the evaluation of painting value, painting protection, painting author identification, painting style comparison and semantic emotion analysis of paintings [1]. More importantly, image and feature extraction techniques can analyze pictures' layouts and compositional features.

Layout is essential for painting because it is through layout that the painter combines images, colors, etc., to express his emotions and themes. Since ancient times, artists have been devoted to how to express subtle spatial relationships through the spatial composition of paintings. Thus, Hassanat et al. [2] believe that as the central aspect of painting, the layout has been highly valued in art. As the image has become increasingly sophisticated, many art educators have begun to focus on how young children express space through painting. However, young children's physical and psychological developmental characteristics differ significantly from those of adults, especially in the layout of drawing, making it challenging for human educational resources to guide young children's drawing layouts, thus making it difficult to provide scientifically sound and relevant teaching. Therefore, the primary purpose of this paper is to try to introduce computer vision technology into young children's drawing layout, and realize the digital analysis of drawing layout and composition characteristics through human-computer interaction, to provide a basis for the integration of artificial intelligence and artistic expression, as well as scientific guidance of young children's drawing composition and layout.

2 IMAGE FEATURE EXTRACTION METHOD

Computer vision and artificial intelligence have made it possible to analyze and manage paintings more efficiently and effectively. As a class of image resources, researchers have used a variety of representative image feature extraction and analysis methods, including support vector machines, neural networks, decision trees, K-nearest neighbor classification, hidden Markov models, and many others.

2.1 Support Vector Machines

Support vector machines are a supervised learning method based on statistical learning theory and classification boundaries to improve their generalization ability by minimizing structural risk, mainly for binary classification tasks, to find a hyperplane to split two classes of training sample points so that classification errors are minimized. If the training samples are linearly divisible, there exist one or more hyperplanes that can divide them completely. For each data class, Wei and Zhang [3] argue that SVM aims to determine the optimal hyperplane, ensuring that the distance between the nearest vector and the hyperplane is the largest. If the training samples are nonlinearly separable, a nonlinear mapping algorithm (i.e., kernel function) is necessary to transform linearly indistinguishable sample points in the low-dimensional input space into the high-dimensional feature space so that they become linearly separable and then determine the classification boundary by linear division.

Training large-scale samples with support vector machines are challenging. To solve multi-classification tasks, multiple two-class support vector machines must be combined. Even so, support vector machines offer advantages when dealing with small samples, linearly indistinguishable samples, and a large number of dimensions.

2.2 Decision Tree

The decision tree algorithm is a tree-like structure to build a decision model based on the attribute characteristics of the data and is often used to solve classification and regression problems. C4.5

and Random Forest are standard decision tree algorithms. The ID3 algorithm is at the core of C4.5, which is an improvement on ID3 (iterative Dichotomiser 3). Using the ID3 algorithm, which selects attributes based on the information gain of each attribute, entropy and information gain are the measures used in information theory. Following splitting, the attribute with the highest information gain is split, and a decision tree is constructed by traversing the space of possible decision trees using a greedy top-down search.

To overcome this problem, Tu et al. [4] proposed that the C4.5 algorithm selects attributes according to the information gain rate, divides the data set at the splitting point with the highest information gain rate, and prunes the tree during construction to prevent uncontrolled growth of the tree height and overfitting of the data. By assigning common values to their corresponding examples or assigning probability values to each possible value, the algorithm converts incomplete data into complete available data. The C4.5 algorithm requires multiple sequential scans and sequencing of the dataset, so it is inefficient and only works with memory-resident datasets. Despite this, the resulting classification rules are easy to understand and relatively accurate. Multiple decision trees are created in a randomized manner in a random forest, with no connection between them. This is an improvement over the C4.5 decision tree. Each decision tree in the forest makes a separate judgment about the sample and predicts its class when new data is available.

2.3 Artificial Neural Networks and Deep Learning

Artificial neural networks are pattern-matching algorithms that simulate biological neural networks to perform classification or regression tasks. Artificial neural networks commonly use backpropagation, perceptron, and deep learning algorithms [5].

The BP algorithm consists of input training samples, a backpropagation algorithm that adjusts the network weights and deviations continuously, and a backpropagation algorithm that ensures the output vector is as close to the desired vector as possible when the network error sum of squares is less than the specified error. Weights and deviations of the network are saved. Although its training time is long, and it is easy to fall into local minima, genetic algorithms can optimize it to be globally optimal, fast, and efficient.

Binary classification is learned using the perceptron linear classification model. The basic idea is to divide the feature vectors of instances in the input space into positive and negative classes, i.e., the output class. An MLP consists of an input layer, an output layer, and one or more hidden layers, each with multiple nodes fully connected to each other. MLPs solve nonlinear global partitioning and achieve high levels of parallel computation, with strong adaptive and self-learning capabilities. There is still a challenge in selecting the number of nodes in the hidden layer of the network.

A deep learning algorithm is a further development of an artificial neural network, which is a learning method that uses unsupervised feature learning and feature hierarchy. It involves building a model with multiple hidden layers and a large amount of training data to learn more valuable features adaptively and ultimately improve classification accuracy. A convolutional neural network (CNN) is a multi-layer perceptron deep learning algorithm derived from a biological mechanism of perception. In the C layer, the input image is convolved with a filter and a bias, and these feature maps are summed, weighted, and biased, followed by an activation function (commonly used Sigmoid function). In general, CNNs are displacement, scale, and distortion-invariant because of three structural properties: local perceptual field, weight sharing, and temporal or spatial down sampling. In pattern recognition, image processing, and computer vision, CNNs are widely used.

2.4 K-Nearest Neighbor

The classification idea of the KNN algorithm is to select K samples from the training samples closest to the current input sample and then determine the class of most of the K samples, which is the class of the recent input sample. The algorithm often uses "Euclidean distance" as the classification model, and the selected nearest neighbor samples are correctly classified. When the

number of samples is unbalanced, such as a large number of samples from one class and a minimal number of samples from other classes, it is possible that a new input sample has a large number of samples from the large-capacity class among its K nearest neighbors. Thus the input sample class may be misclassified. However, Chen et al. [6] pointed out that this algorithm is relatively simple, easy to understand and implement, and especially suitable for multi-classification tasks.

In addition to the above learning models, there are machine learning models such as probabilistic, fusion, clustering, etc. The Hidden Markov Model (HMM) is a time-series probabilistic model that describes the process state by a single discrete random variable. The K means algorithm is a simple clustering algorithm that divides N samples into K partitions ($K < N$) based on their attributes.

3 IMPROVED IMAGE EXTRACTION DESIGN BASED ON TRANSFORMER

The transformer has been used to design image feature extraction networks. Transformer-based models are, however, limited by the fact that, given a sequence of tokens as input, the self-attentive mechanism is required to iteratively learn the feature representation by associating any two tokens from the series, resulting in a model with quadratic time and space complexity. Transformers cannot model high-resolution images due to their quadratic complexity, and their high computational cost makes it difficult to apply to edge devices. As shown in Figure 1, we propose a linear attention and pruning-based Transformer image feature extraction network. 4. Tokens are scored from an external perspective based on their importance to the final Class Token. Tokens are then sampled based on their scores and pruned from the Token dimension based on their scores. In order to obtain the new linear attention matrix, a combination function is used to replace the Softmax algorithm that computes the self-attentive matrix from the internal perspective. In the end, the two methods are combined to create an efficient attention mechanism (E-Attention).

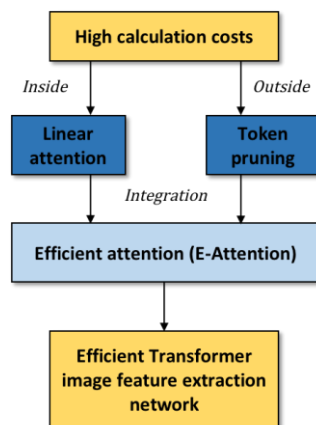


Figure 1: Model framework.

3.1 Linear Attention-based Light Weighting

The key to implementing a linear attention mechanism is to be able to find a decomposable similarity function. Most of the current linear attention mechanisms in academia attempt to approximate the Softmax algorithm. For example, Hernández et al. [7] uses the stochastic Fourier identity theorem, and Liang et al. [8] uses positive random features to approximate the Softmax

operator. However, it has been found empirically that these methods are sensitive to the choice of sampling rate and become unstable if the sampling rate is too high. Since these methods implement the linear attention mechanism by using an effective approximation of the Softmax operator only within a constrained theoretical range, they may not always outperform the common structure when the corresponding assumptions are not satisfied or the approximation error accumulates. Therefore, whether a decomposable similarity function can be used to directly replace the Softmax operator with a decomposable similarity function while ensuring experimental results is considered. This requires identifying the key features of the current attention mechanism and thus designing a decomposable similarity function that satisfies the requirements to achieve linear attention. We propose a combinatorial function for replacing Softmax that satisfies both of these properties and consists of two sub-functions for non-negativity and non-linear reweighting, respectively; the algorithmic flow of the two sub-functions is shown in Figure 2.

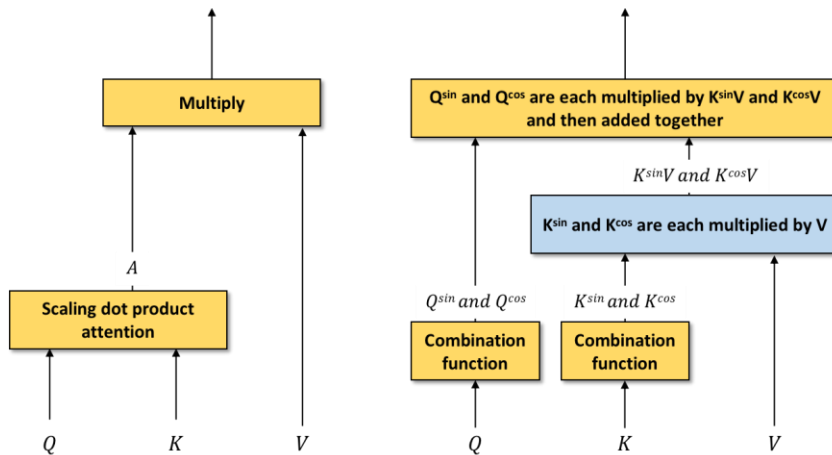


Figure 2: Scaled dot product attention and linear attention.

3.2 Light Weighting based on Token Pruning

Many current methods for pruning Transformer-based image feature extraction networks from the Token dimension introduce additional neural networks for training to calculate Token scores, which are used to determine further which Tokens are redundant and which Tokens should be retained. However, this class of methods has a fixed rate of Token reduction at each stage. While this does reduce the computational effort of the Transformer-based image feature extraction network from an external perspective, it also introduces an additional computational overhead, i.e., the scoring network itself, which needs to be trained together with the Transformer-based image feature extraction network. Another limitation is that if a fixed pruning Token ratio needs to be changed, the network must be retrained when deployed on different edge devices, which significantly limits the application scenarios of the model [9]. In addition, some of the information in the input image may be redundant, and this information cannot be used as a basis for discriminating the image class. The amount of this information depends on the image itself. Suppose the pruning ratio is fixed at each stage. In that case, to reduce the computation, on the one hand, if the pruning ratio is too large, some vital information may be unconsciously lost, resulting in the reduction of classification accuracy; on the other hand, if the pruning ratio is too small, some redundant Tokens will be retained, which will also lead to the waste of computation resources. On the other hand, if the pruning ratio is too small, some redundant Tokens will be retained, wasting computational resources. The adaptive retention of different Tokens according to additional input images can solve these limitations well. It can dynamically retain the number of remaining Tokens in a

sampling manner according to the actual input images, thus adapting to different practical application scenarios [10].

Therefore, we first compute the scores of the input Token by a parameter-free method to avoid introducing additional parameters that would increase the computational effort; at the same time, due to the excellent feature of parameter-free, our pruning method can be presented as a plug-and-play module into any off-the-shelf Transformer-based image feature extraction network. Our designed method for pruning the Transformer-based image feature extraction network from the Token dimension has two features: parameter-free evaluation of Token scores; and adaptive sampling based on the input image, thus dynamically preserving Token.

3.3 Light Weighting based on Linear Attention and Token Pruning

This paper further tries to combine the linear attention mechanism with the Token pruning module. Due to the fact that the scoring is based on the degree of association between the first Class Token and other Tokens, that is, all the values in the first row except the first one of the attention matrix, or all the values in the first column except the first one of the attention matrix, the two are equivalent. For example, the element in the first column of the second row of the attention matrix represents the degree of association between the second Token and the Class Token, and this element is calculated by multiplying all the elements in the second row of matrix Q by all the elements in the first column of matrix K^T and then adding them up. As a result, we can ensure a linear attention mechanism while simultaneously calculating the degree of association between the Class Token used for classification and other Tokens, and the calculation process is linear, so the computational effort is not increased. A more efficient Attention (E-Attention) mechanism is achieved by combining linear attention and token pruning. Figure 3 shows the flow of computation.

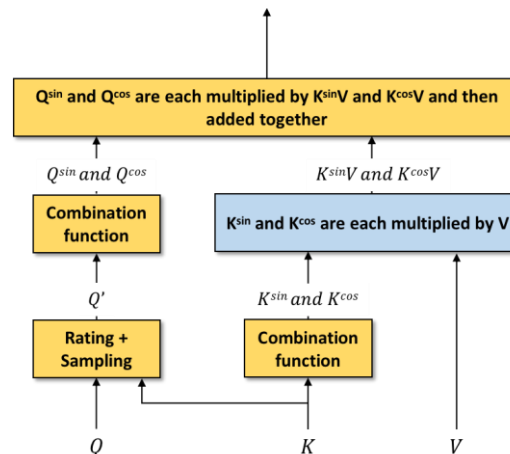


Figure 3: E-Attention.

4 MODEL TEST AND RESULT ANALYSIS

We first conduct internal experiments on these lightweight methods and then introduce several Transformer models as the image feature extraction part to train on the image classification and target detection tasks to obtain the original experimental results, and then verify the experimental results on the classification and detection tasks with the introduction of linear attention mechanism, Token pruning or E-Attention alone. Then, the experimental results of the Transformer model with linear attention mechanism, Token pruning or E-Attention alone are verified on the classification

and detection tasks and evaluated and analyzed. This chapter validates the comparison experiments based on the ImageNet1k dataset, widely used in image classification, and the COCO dataset, commonly used in target detection.

The experiments were conducted using the VS-code editor, Python 3.6, Windows 10 Education Edition 64-bit system with 32GB of memory, Pytorch as the neural network library, and Nvidia GeForce RTX 2080Ti 11GB as the graphics processor used in the experimental environment to validate the model. The Intel(R) Core(TM) i79700K was used as the processor. In addition, all the hyperparameters of the experimental model were set using the hyperparameters provided by the model DeiT. The Epochs are 300, the Batch size is 1024, the base learning rate is 0.0005, the optimizer is AdamW, the learning rate decay strategy is Cosine, the weight decay is 0.05, the Dropout is 0.1, the Warmup epochs are 5. The pruning method is The fixed starting parameter is set to 70% of the number of input Token in the initial stage. Six models, DeiT, PVT, Swin Transformer, TNT, T2T-ViT, and CaiT, are selected as the benchmark models to verify the innovations proposed in this chapter, and different sizes of each of these models are compared in the specific experiments.

First, image classification experiments are conducted on the ImageNet1k dataset using different linear attention models to compare the advantages and disadvantages of each linear attention mechanism. Figure 4 compares the performance of the linear attention mechanism proposed in this chapter with the linear attention mechanism currently proposed in academia, which shows that compared with the original Transformer model DeiT-S, the linear Transformer can significantly reduce the computational effort of FLOPs for forwarding inference, and all linear attention mechanisms are improved by about 50%. The linear attention mechanism proposed in this chapter has the highest improvement in FLOPs, which is better than other linear attention mechanisms in academia and is second only to Nyströmformer regarding classification accuracy.

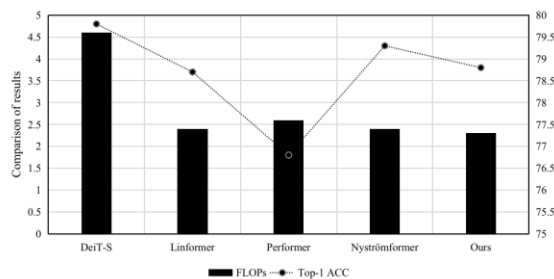


Figure 4: Comparison of linear attention mechanisms.

The improvement of the proposed linear attention mechanism is derived from two key features that affect the performance of Softmax attention: non-negative attention matrix elements and non-linear reweighting. The proposed method is based on the assurance of these two properties, so we verify the improvement of the linear attention mechanism to see how much each of these properties improves the final experimental results. The results are shown in Figure 5. It can be seen that the linear attention mechanism obtained by keeping only one feature is much less accurate than the linear attention mechanism obtained by keeping both features. On the other hand, even if only one feature is retained, the FLOPs are still significantly improved because the attention mechanism is still linear.

This paper calculates the importance score of each candidate Token by referring to the attention weight of the Class Token. To evaluate the effectiveness of this method, we choose two other methods to calculate the importance score of each candidate Token.

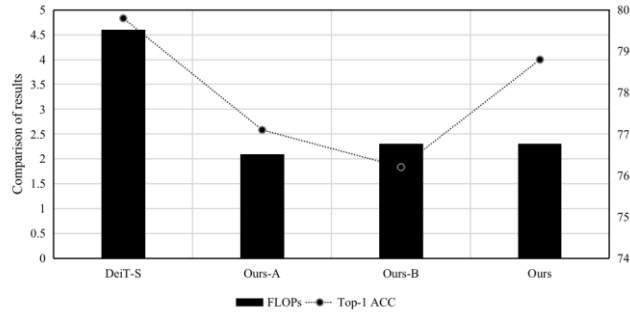


Figure 5: An internal exploration of linear attention mechanisms.

The first method, all score, is to find the essential Token by adding up the attention weights of all Tokens. The second method is to randomly select another Token other than the class Token and calculate the score based on its attention weight. As shown in Figure 6, the attention weight of Class Token performs better, which means that the attention weight of Class Token is more informative for evaluating candidate Token. This is because the Class Token is used to predict the category probability in the final stage of the model. Hence, the attention weight of the Class Token indicates which Token has more influence on the last output Token for classification. Summing all the attention weights only shows the Token with the highest attention weight among all other Tokens, which is not necessarily valid for the classification of Token. Finally, we randomly select a Token and calculate the final score concerning its attention weight, which is the worst result.

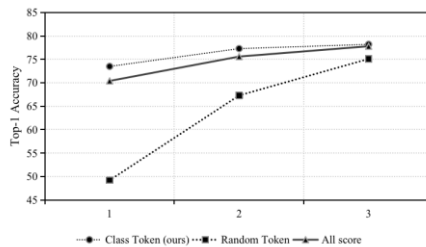


Figure 6: Comparison of different scoring methods.

Before obtaining this method, we first tried to select the top k Tokens with the highest scores directly and compare the results of the two experiments, as shown in Figure 7.

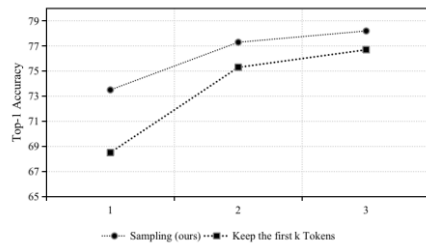


Figure 7: Comparison of different sampling methods.

It is clear that the adopted method is better than the direct selection of the top k Token, and the inverse transform sampling method based on the cumulative distribution function of the scores does not necessarily discard all the Token with lower scores, thus providing a more diverse set of Token for the later stages. In addition, the Top-K selection method will produce a fixed Token selection rate at each stage, which also limits the model's performance.

5 CONCLUSION

The combination of computer vision technology and painting creation can bring the artist and the viewer closer together and achieve friendly interaction between humans and computer. The mutual penetration of scientific and artistic thinking is of great significance in broadening human creativity and improving the artist's expressiveness and creativity. More importantly, the specific application of computer-aided technology in drawing layout feature analysis also provides an essential tool for scientifically guiding young children's drawing instruction.

Image feature extraction is extremely important in the field of computer vision and is a crucial technology for computer-aided drawing in the field of artificial intelligence, which has significant application value in helping young children learn drawing layouts. In recent years, Transformer-based image feature extraction methods have become a hot research topic, but there are still some areas for improvement in the related models. Therefore, this paper proposes two methods to accelerate the Transformer model to solve the problems of the high computational cost of the current Transformer model and the quadratic relationship between the complexity of the model and the number of input Tokens from both internal and external perspectives. First, the quadratic complexity of the self-attentive mechanism itself is reduced to linearity to improve the model's processing speed from the inside. Then, we propose a parameter-free lightweight pruning method that can sift out unimportant Token by adaptively sampling different input images to reduce the meaningless input from the outside. Finally, the two approaches are combined to obtain a new efficient attention mechanism (E-Attention). Experiments show that each method can reduce the computation of the original Transformer model by 30%-50%, and E-Attention can reduce the computation of the original Transformer model by 60%-70%. The proposed model provides a model basis for the application of computer-aided technology in the scientific guidance of young children's drawing layouts.

6 ACKNOWLEDGEMENTS

This work was supported by the 2022 general project of Henan Educational Science Planning: Research on the inheritance of Yellow River culture in children's art practice (No.2022yb0610).

Hua Xie, <https://orcid.org/0000-0001-8026-9963>

REFERENCES

- [1] Li, Z.; Han, X.; Wang, L.; Zhu, T.; Yuan, F.: Feature extraction and image retrieval of landscape images based on image processing, *Traitement du Signal*, 37(6), 2020, 1009-1018. <https://doi.org/10.18280/ts.370613>
- [2] Hassanat, A.; Prasath, V.-B.; Al-kasassbeh, M.; Tarawneh, A.-S.; Al-shamailh, A.-J.: Magnetic energy-based feature extraction for low-quality fingerprint images, *Signal, Image and Video Processing*, 12(8), 2018, 1471-1478. <https://doi.org/10.1007/s11760-018-1302-0>
- [3] Wei, Z.; Zhang, X.: Feature extraction and retrieval of ecommerce product images based on image processing, *Traitement du Signal*, 38(1), 2021, 181-190. <https://doi.org/10.18280/ts.380119>

- [4] Tu, B.; Li, N.; Fang, L.; He, D.; Ghamisi, P.: Hyperspectral image classification with multi-scale feature extraction, *Remote Sensing*, 11(5), 2019, 534. <https://doi.org/10.3390/rs11050534>
- [5] Lozano-Vázquez, L.-V.; Miura, J.; Rosales-Silva, A.-J.; Luviano-Juárez, A.; Mújica-Vargas, D.: Analysis of different image enhancement and feature extraction methods, *Mathematics*, 10(14), 2022, 2407. <https://doi.org/10.3390/math10142407>
- [6] Chen, Y.; Yang, Z.; Ma, L.; Li, P.; Pang, Y.; Zhao, X.; Yang, W.: Efficient extraction algorithm for local fuzzy features of dynamic images, *Discrete & Continuous Dynamical Systems-S*, 12(4&5), 2019, 1311. <https://doi.org/10.1080/09720529.2018.1449298>
- [7] Hernández, J.-F.-C.: Artificial intelligence and spirituality, *International Journal of Interactive Multimedia and Artificial Intelligence*, 7(1), 2021, 34-43. <https://doi.org/10.9781/ijimai.2021.07.001>
- [8] Liang, Y.; Lu, S.; Weng, R.; Han, C.; Liu, M.: Unsupervised noise-robust feature extraction for aerial image classification, *Science China Technological Sciences*, 63(8), 2020, 1406-1415. <https://doi.org/10.1007/s11431-020-1600-9>
- [9] Haenlein, M.; Kaplan, A.: A brief history of artificial intelligence: On the past, present, and future of artificial intelligence, *California Management Review*, 61(4), 2019, 5-14. <https://doi.org/10.1177/0008125619864925>
- [10] Echegaray, S.; Bakr, S.; Rubin, D.-L.; Napel, S.: Quantitative image feature engine (QIFE): An open-source, modular engine for 3D quantitative feature extraction from volumetric medical images, *Journal of digital imaging*, 31(4), 2018, 403-414. <https://doi.org/10.1007/s10278-017-0019-x>