ITC 3/53

Information Technology and Control
Vol. 53 / No. 3/ 2024
pp. 833-845
DOI 10.5755/j01.itc.53.3.36943

Few-Shot Learning on Edge Devices Using CLIP: A Resource-Efficient Approach for Image Classification

Received 2024/04/10 | Accepted after revision 2024/06/29

HOW TO CITE: Lu, J. (2024). Few-Shot Learning on Edge Devices Using CLIP: A Resource-Efficient Approach for Image Classification. *Information Technology and Control,* 53(3), 833-845. https://doi.org/10.5755/j01.itc.53.3.36943

Few-Shot Learning on Edge Devices Using CLIP: A Resource-Efficient Approach for Image Classification

Jin Lu

Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen Polytechnic University, Shenzhen 518055, Guangdong, China

Corresponding author: lujin0808@szpu.edu.cn

In the field of deep learning, traditional image classification tasks typically require extensive annotated datasets and complex model training processes, which pose significant challenges for deployment on resource-constrained edge devices. To address these challenges, this study introduces a few-shot learning method based on OpenAI's CLIP model that significantly reduces computational demands by eliminating the need to run a text encoder at the inference stage. By pre-computing the embedding centers of classification text with a small set of image-text data, our approach enables the direct use of CLIP's image encoder and pre-calculated text embeddings for efficient image classification. This adaptation not only allows for high-precision classification tasks on edge devices with limited computing capabilities but also achieves accuracy and recall rates that closely approximate those of the pre-trained ResNet approach while using far less data. Furthermore, our method halves the memory usage compared to other large-scale visual models of similar capacity by avoiding the use of a text encoder during inference, making it particularly suitable for low-resource environments. This comparative advantage underscores the efficiency of our approach in handling few-shot image classification tasks, demonstrating both competitive accuracy and practical viability in resource-limited settings. The outcomes of this research not only highlight the potential of the CLIP model in few-shot learning scenarios but also pave a new path for efficient, low-resource deep learning applications in edge computing environments.

KEYWORDS: Few-shot learning, CLIP model, image classification, edge devices, deep learning.

1. Introduction

In current deep learning research and applications [9, 16], especially in the progress of image classification tasks, the issue of data dependency is particularly prominent. The core challenge lies in that most efficient and advanced models rely on large-scale, accurately annotated datasets for training. This dependency is reflected not only in the enhancement of model performance but also in the richness of the features that the model can learn and its generalization ability [14-15, 23]. However, acquiring these large-scale annotated datasets requires significant human and time costs [17, 25], especially in fields requiring precise annotations, such as medical image analysis [5, 18]. Furthermore, as the complexity of the model structure increases, the demand for data correspondingly rises. Complex model structures can capture finer-grained features and improve the accuracy of classification, but this also means that more training data is needed to avoid overfitting issues [35]. Overfitting refers to a model performing well on the training data but poorly generalizing to new, unseen data. Therefore, a large amount of annotated data becomes key to improving the model's generalization ability.

The reliance on large-scale annotated datasets presents several problems. First, as mentioned earlier, collecting, and annotating these datasets is time-consuming and costly [11]. Second, in some fields, such as the analysis of medical images for rare diseases, it may be difficult to obtain sufficient annotated data [10]. Finally, even if a large volume of data can be acquired, the quality of annotations may vary, affecting the reliability of the training outcomes. To address these challenges, researchers have explored various methods, including but not limited to transfer learning [21, 27], few-shot learning [23], and self-supervised learning [1]. These approaches aim to reduce the dependence on large-scale annotated datasets by utilizing pre-trained models, data augmentation generated by Generative Adversarial Networks (GANs) [3], or by designing algorithms capable of learning from unannotated data, thereby improving the training efficiency and generalization ability of the models [4]. Nevertheless, enhancing the performance and generalization ability of models under limited data conditions remains a significant issue in deep learning research.

In recent advancements, Few-Shot Learning (FSL) has emerged as a focal area of research with the goal of training efficient learning models using very limited data samples [28]. The primary challenge of this learning method is to construct a model that can deeply mine and understand the intrinsic characteristics of data while being effective in generalization with only a handful of samples [6]. In this context, the CLIP (Contrastive Language–Image Pre-training) model developed by OpenAI, through contrastive learning on a vast array of image and text pairs, has demonstrated its exceptional capability in cross-modal understanding [22]. The success of the CLIP model has introduced new insights into the domain of few-shot learning, showing how leveraging cross-modal information can enhance a model's generalization ability and learning efficiency when faced with limited data [7]. This research has developed a novel technique for few-shot image classification based on OpenAI's CLIP model, specifically designed to adapt to resource-constrained edge computing environments. Compared to traditional deep learning strategies, this method optimizes computational requirements by eliminating the need for text encoding during inference, significantly reducing computational complexity without compromising high classification accuracy. By pre-computing the embedding centers of category texts and utilizing the image encoding component of the CLIP model for similarity comparison, this technology can efficiently utilize limited image-text datasets for precise image classification. The design of this method pays special attention to the computational and storage resource limitations of edge devices, greatly simplifying the complexity and cost of deploying advanced deep learning models on such devices.

2. Literature Review

In the realm of deep learning, Few-Shot Learning (FSL) has emerged as a pivotal research direction, aimed at addressing the challenge of efficiently training models with only a minimal number of labeled samples available. The research efforts in this field primarily focus on designing algorithms and model architectures capable of swiftly adapting to new tasks. (1) Among these, the concept of Meta-Learning, or "learning to learn," occupies a central position in Few-Shot Learning. Its core idea involves training a model to rapidly adapt to new tasks upon encountering a limited number of samples. The Model-Agnostic Meta-Learning (MAML) algorithm introduced by Finn et al. [6] marks a significant milestone, optimizing the initial parameters of a model to enable rapid adaptation to new tasks through a minimal number of gradient update steps. MAML and its variants have been extensively applied across a variety of Few-Shot Learning tasks. (2) In the domain of transfer learning, models are initially pre-trained on a large-scale dataset, followed by the migration of the model to a new task characterized by a scant number of labeled samples. The efficacy of this approach is predicated on the hypothesis that features learned on extensive datasets are, to some extent, universal and can facilitate rapid learning on new tasks. The work of Yosinski et al. [34] demonstrated that lower-level features of deep neural networks are particularly beneficial for new tasks. (3) Data augmentation represents a method to enhance the model's generalization capabilities by artificially increasing the diversity of training samples. Within the context of Few-Shot Learning, data augmentation proves especially beneficial, as it can effectively expand a minimal dataset. Wang et al. [29] illustrated the potential to enhance model performance on few-shot image classification tasks through dataset augmentation by proposing a data augmentation method based on a self-attention mechanism. (4) Generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [8], have been employed to enhance Few-Shot Learning. These models can produce new, seemingly authentic samples, thereby increasing the quantity and diversity of training data. For instance, Antoniou et al. (2017) introduced a method utilizing GANs for data augmentation to support Few-Shot Learning. (5) Self-Supervised Learning represents a learning methodology that does not require explicit labeling. It trains models by constructing supervision signals automatically generated from unlabeled data. Within the context of Few-Shot Learning, Self-Supervised Learning can be utilized to pre-train models, thereby achieving improved performance with limited labeled data. He et al. [13] improved the performance of image recognition tasks through self-supervised pre-training, demonstrating the potential of Self-Supervised Learning in leveraging unlabeled data. To reduce the dependency on labeled data in video emotion recognition tasks, Sun et al. [26] proposed the HiCMAE method based on self-supervised learning, which leverages unlabeled audio-visual data to enhance the accuracy of audio-visual recognition, achieving significant effectiveness.

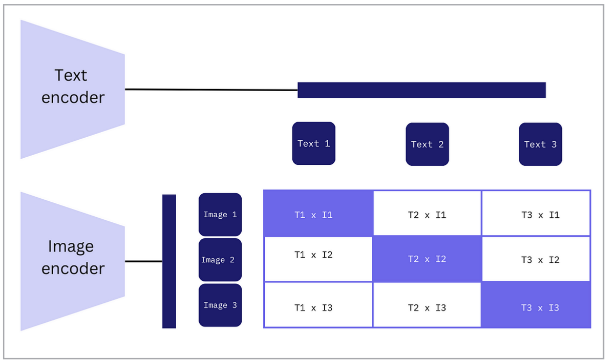
In the domain of deep learning, particularly in the study of cross-modal learning between images and text, OpenAI's CLIP (Contrastive Language–Image Pre-training) model has garnered widespread attention. By utilizing a contrastive learning approach on a vast dataset of unlabeled images and text, CLIP has successfully learned representations that capture the deep semantic relationships between images and text. This unique training strategy endows CLIP with a robust generalization capability, enabling it to exhibit exceptional performance across a variety of image and text-related tasks, such as image classification, cross-modal retrieval, and zero-shot learning. The core advantage of the CLIP model lies in its cross-modal understanding capability. By learning the correspondence between image content and natural language descriptions, CLIP as show in Figure 1 can effectively classify and comprehend images without explicit labels. This capability is particularly suited to Few-Shot Learning scenarios, where labeled data is extremely scarce. CLIP leverages pre-computed text embeddings as class identifiers, measuring the similarity between image embeddings and these class embeddings to quickly adapt to new categories, thereby significantly reducing the dependency on extensive labeled data. Furthermore, CLIP was designed with practicality and flexibility in mind. It can process a wide range of image and text inputs without being constrained by specific dataset annotation conventions, allowing for easy application in various scenarios. Moreover, as CLIP only requires the use of an image encoder during inference, eliminating the need for a text encoder at the inference stage, this further reduces the computational costs of deploying the model, especially on resource-constrained edge devices. Recent research efforts have begun to explore how to further extend and optimize the CLIP model to accommodate a broader range of application requirements. For instance, researchers have attempted to combine CLIP with other models to enhance performance on specific tasks, or to develop new training strategies to improve the model's understanding within specific domains. These efforts not only validate the potent potential of the CLIP model but also provide rich insights for future research directions and applications.

Vision Transformer (ViT) [12] signifies a pivotal shift from conventional Convolutional Neural Networks (CNNs) to Transformer-based approaches in the field of computer vision. Introduced by Google Research in 2020, ViT has garnered widespread attention due to its straightforward yet efficient architecture. ViT processes images by segmenting them into multiple fixed-size patches, linearly embedding these patches into a sequence, and subsequently utilizing the Transformer model to handle the sequence for feature extraction and classification. The principal advantage of this method lies in its ability to capture global dependencies, unlike CNNs, which are confined to local receptive fields. As the depth of the model increases, ViT demonstrates superior performance, especially when trained on large-scale datasets. Moreover, the architecture of ViT provides a new direction for subsequent research, such as the incorporation of hierarchical or sparse attention mechanisms to enhance efficiency and scalability. Currently, ViT has been extensively applied to various vision tasks, including image classification, object detection, and semantic segmentation, highlighting its significance and potential in the domain of deep learning. Yao et al. [33] proposed a novel hybrid deep model named HIRI-ViT, which integrates the features of Vision Transformer and CNN, specifically designed for high-resolution inputs. By decomposing typical CNN operations into two parallel branches, the model enhances cost-efficiency. Achieving a top-1 accuracy of 84.3% on the ImageNet dataset at a comparable computational cost, the model has shown an improvement of 0.9% over previous models, demonstrating its superior performance in high-resolution vision tasks. Xu et al. [32] introduced a deep learning approach called HCF-Net, which significantly enhances the performance of infrared small object detection through the introduction of three key modules: Parallelized Patch-Aware Attention (PPA) module, Dimension-Aware Selective Integration (DASI) module, and Multi-Dilated Channel Refiner (MDCR) module. These modules utilize multi-scale and multi-level feature extraction strategies to improve the capability of feature capture and fusion, effectively addressing the identification and localization of small targets, particularly in complex background infrared images. Liu et al. [19] developed a Hierarchical Feature Fusion Attention Network (HFANet), specifically designed to classify fluorescence intensity and distribution patterns in glomerular immunofluorescence images. By integrating the Hierarchical Feature Fusion Attention (HFA) module, the network leverages shallow texture features to enhance deep semantic features, optimizing feature extraction and information fusion efficiency. HFANet employs weighted concatenation of feature maps from different hierarchies to emphasize more discriminative regions, and, in conjunction with the Intensity Equalization (IE) algorithm, U-Net++, and Grad-CAM, it constructs a computer-aided diagnostic system that significantly enhances the classification accuracy of fluorescence features, performing comparably to senior pathologists.

In this study, the comparison between the proposed method and traditional machine learning techniques such as Support Vector Machines (SVMs) [2] and fully trained deep learning approaches like Convolutional Neural Networks (CNNs) reveals significant advantages and some potential drawbacks of CLIP in addressing few-shot learning challenges. The pre-training mechanism of CLIP enables it to comprehend a vast array of visual concepts, allowing it to adapt effectively to new tasks with limited annotated data, thereby demonstrating its exceptional cross-modal capabilities and generalization performance. However, the pre-training of CLIP is relatively challenging. In contrast, SVMs are renowned for their theoretical maturity and efficiency on small-scale datasets, yet they face limitations in processing large-scale datasets and in feature engineering. Meanwhile, CNNs exhibit strong adaptability and feature extraction prowess in image classification tasks through their ability to learn hierarchical features of images automatically, albeit this depends on the availability of extensive training data and significant computational resources. This comparison not only highlights the potential of CLIP in the domain of few-shot learning but also underscores considerations regarding resource and data availability, providing a basis for selecting the most suitable approach.

Figure 1

The CLIP (Contrastive Language-Image Pre-training) model employs a compound encoding structure that simultaneously utilizes a text encoder and an image encoder to understand and correlate visual images with descriptive texts. Specifically, the text encoder processes input text data, converting it into high-dimensional feature vectors, while the image encoder transforms input images into corresponding feature vectors. At the core of the model, these feature vectors from the two different encoders interact within a contrastive learning framework to compute their similarity



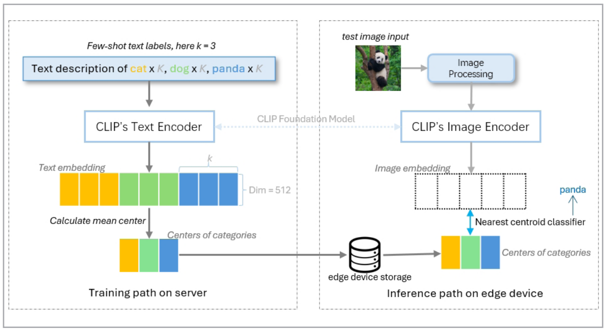
3. Method

This study employed a mixed-methods research design, integrating both quantitative and qualitative approaches, to evaluate the efficacy of the CLIP model in few-shot learning scenarios. The primary objective was to assess the model’s performance across diverse categories and to understand its learning capabilities under conditions of limited data availability.

For data collection, the study utilized three distinct datasets: “Animals with Attributes 2” (AwA2), the SUN Database, and the Wiki Art dataset. The AwA2 dataset, known for its rich attribute labels and broad representation of animal species, was particularly suitable for few-shot learning experiments. The SUN Database provided a diverse range of scene images, while the Wiki Art dataset included a comprehensive collection of artworks. The combination of these datasets allowed for extensive coverage of both visual and textual materials, thereby facilitating a thorough evaluation of the CLIP model’s performance. A subset of 10 animal species, each represented by 10 images, was selected from the AwA2 dataset. For each image, 20 relevant attributes were identified and used to generate textual descriptions. These descriptions were then encoded into embeddings using the CLIP model’s text encoder, forming the main dataset for subsequent model training.

Figure 2

This figure illustrates the workflow of the proposed method. The left side shows the training process, which is deployed on a server. After obtaining the text classification centers, these centers vector are transferred to the edge device, where they are used in conjunction with CLIP’s image encoder for classification



The proposed method set up the CLIP model to process both text and image inputs through a sophisticated embedding comparison mechanism, leveraging a few-shot learning framework to enhance the model’s classification accuracy under limited data conditions. Initially, the model employed separate but interconnected encoders to encode textual descriptions and images into a shared high-dimensional embedding space. Textual descriptions of each category—typically synthesized from multiple samples to enrich semantic understanding—were encoded to form “class center embeddings,” which served as the central reference points for each category in the dataset. To calculate these class centers, we typically averaged the embeddings of several representative text samples for each class. This process helped to reduce the variability and noise of individual samples, resulting in more stable and representative class descriptors. For images, each input image was independently encoded into the same embedding space. Once the textual class centers were established, the model entered the inference stage. The main task during this phase was to match new image samples with the established textual class centers to determine the classification of the images. The specific process was as follows: During the inference phase, for each test image, the model used the image encoder to transform it into an embedding representation. This allowed the image to be compared with the textual class centers in the same high-dimensional space, accurately reflecting the visual features of the image to facilitate effective matching with the text embeddings. Next, the model calculated the similarity between the image embedding and each class center embedding, typically using cosine similarity as the metric. The model then computed a similarity score for each category. Finally, the model determined the classification of the image based on the similarity scores, typically choosing the class center with the highest similarity to the image embedding as the predicted category of the image.

Experiment I: Animal Species Recognition Based on the "Animals with Attributes 2" (AwA2) Dataset.

The "Animals with Attributes 2" (AwA2) [30] dataset provides a rich set of attribute labels for animal species, making it suitable for demonstrating the application of few-shot learning in animal species recognition. AwA2 includes 50 animal species and 85 attributes. Image classification assisted by attribute labels can significantly enhance the understanding and protection of biodiversity. 1) Data Collection: We select 10 animal species from the AwA2 dataset and use 20 attribute columns to construct textual descriptions, employing LLM through prompt engineering to form sentences from multiple attributes. For example, we choose animals with attributes such as the ability to fly, aquatic lifestyle, hooves, nocturnal activity, stripes, spots, etc. Each animal is provided with 10 representative images. 2) Generation of Classification Centers: Using the text encoder of the CLIP model, we convert the textual descriptions of each animal into embedding vectors, serving as the classification centers for few-shot learning. This step is crucial in enabling us to perform effective classification with a minimal amount of data. 3) Experimental Procedure: After obtaining the classification centers for the 10 categories, we select 5,00 sample images as the test dataset, which includes images of the 10 animal species. During the testing phase, each image to be classified is converted into an embedding vector through CLIP's image encoder. For each image, we calculate the similarity between its embedding vector and the 10 classification centers, determining the image's category based on the nearest classification center in terms of similarity. 4) Performance Evaluation: We assess the classification accuracy of the model on the 5,00 test images to determine the model's efficacy in the task of animal species recognition.

Experiment II: Scene Recognition in the SUN Database Using Few-Shot Learning.

This experiment aims to classify scenes in the SUN dataset [31] using few-shot learning (FSL), augmented by text descriptions generated for this purpose. These descriptions encompass specific attributes present in the scenes, such as objects and lighting conditions.

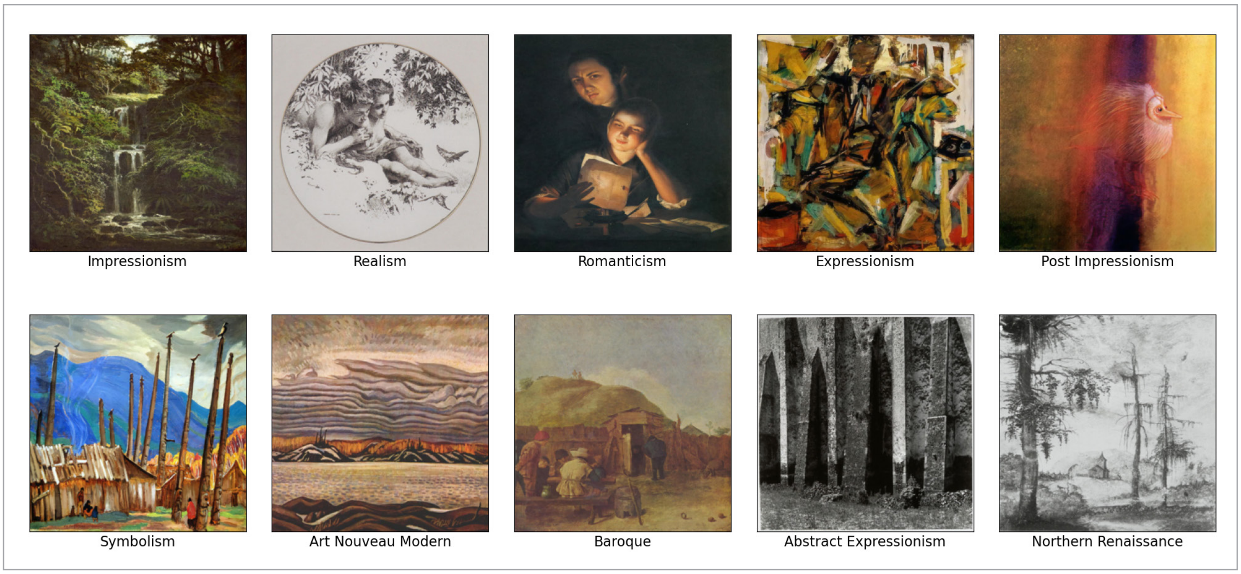
1. Data Collection and Text Description Generation: From the SUN database, we selected 10 different outdoor scene categories, such as "urban streets" and "natural lakes." For each category, we carefully chose 10 representative images. We utilized a large language model, GPT-3.5, to generate detailed text descriptions for each selected image. These descriptions not only mention the main category (e.g., "urban streets" or "natural lakes") but also include specific attributes observed in the images, such as weather conditions, main objects (cars, trees, buildings), the time of day, and lighting conditions. The goal is to create rich descriptive captions that reflect the complexity and diversity of real-world environments.
2. Generation of Classification Centers: Using the text descriptions, we apply the text encoder of CLIP to transform these detailed descriptions into high-dimensional embedding vectors.
3. Experimental Process: We collected an additional 5,00 images from the SUN database as a test dataset. For each image in the test set, we used CLIP's image encoder to generate the corresponding embedding vector. Then, we determined the category assignment by computing the cosine similarity between the embedding vectors of the test images and the two predefined classification centers. The nearest classification center determines the category allocation.
4. Performance Evaluation: The model's performance is evaluated based on its classification accuracy over 5,00 test images, focusing on how well it can generalize the representations learned from a limited set of initial examples to a broader, unseen image collection.

Experiment III: Artwork Style Classification with CLIP and Few-Shot Learning.

Artwork style classification presents a complex challenge within computer vision, aiming to identify and categorize artworks by their artistic styles. Utiliz-

Figure 3

Representative samples from the WikiArt dataset. This figure showcases a selection of artworks included in the WikiArt dataset, illustrating the diversity and breadth of the collection. The dataset encompasses a wide range of art styles, periods, and genres, providing a comprehensive overview of historical and contemporary art



ing OpenAI's CLIP model, this task is approached through few-shot learning, with a dataset comprising various art styles across 27 distinct folders.

1 Data Preparation: The dataset, sourced from WikiArt [20] (Figure 3), is organized into folders representing different art styles. Attributes such as filename, artist, genre, and a synthesized description. A balanced sample across styles ensures equitable representation.

2 Textual Feature Computation: CLIP's text encoder is deployed to convert generated artwork descriptions into high-dimensional vectors. These vectors create a centroid for each art style, correlating textual and visual characteristics of the artworks.

3 Classification Experiment: Images are processed using CLIP's image encoder, and their feature vectors are compared with text feature centroids. Artworks are classified into genres based on the highest similarity score with these centroids.

4 Evaluation and Visualization: Classification performance is measured using accuracy and precision metrics. Confusion matrices and bar charts provide visual insights into the model's performance, highlighting strengths and areas for improvement in distinguishing art styles.

4. Results

1 As shown in Table 1, in the recent experiment utilizing the "Animals with Attributes 2" (AwA2) dataset for animal species recognition through Few-Shot Learning and CLIP, the classification accuracies for ten animal species were examined. The results are as follows: Deer (0.82), Bobcat (0.43), Pig (0.89), Lion (0.32), Mouse (0.70), Zebra (0.85), Collie (0.78), Walrus (0.69), Raccoon (0.72), and Cow (0.80). These outcomes highlight a significant variance in the model's ability to accurately classify different animal species based on minimal examples and textual descriptions. The high accuracies observed for species like the Pig (0.89) and Zebra (0.85) indicate the model's strong generalization capability when distinctive attributes are well-represented within the textual descriptions and image features. Conversely, lower accuracies in species such as the Bobcat (0.43) and Lion (0.32) suggest challenges in capturing and differentiating subtle or complex attributes that distinguish these species, possibly due to similarities in their natural habitats, behaviors, or physical characteristics that are not as explicitly captured in the attributes. The experi-

Table 1

Animal recognition accuracy in AwA2

Deer	Bobcat	Pig	Lion	Mouse	Zebra	Collie	Walrus	Raccoon	cow
0.82	0.43	0.89	0.32	0.70	0.85	0.78	0.69	0.72	0.80

ment underscores the potential and limitations of Few-Shot Learning in recognizing animal species with varied accuracies across different classes. The high classification accuracies for certain species demonstrate the effectiveness of combining CLIP's image and text encoders to understand complex relationships between visual representations and textual attributes.

2 The results of Experiment 2 indicate that scenes with distinct features and those easily distinguishable through textual descriptions (such as "archways" and "beaches") demonstrated higher classification accuracy. The uniqueness and visual characteristics of these scenes were effectively captured and transformed into model-utilizable information through textual descriptions, thereby enhancing recognition precision.

Conversely, for categories with more complex visual features or higher similarity to other scenes (such as "hills" and "campuses"), the model's classification accuracy was relatively lower. This may be due to the insufficient distinguishing information contained within the textual descriptions of these scenes to support accurate model classification, or because the diversity within the scenes made it challenging for the model to learn generalizable features from a limited number of samples. Furthermore, the experimental results also highlighted the impact of the quality of generated textual descriptions on the classification task. High-quality, detailed textual descriptions can provide the model with richer semantic information, thus improving classification accuracy to a certain extent. This experiment validated the feasibility of combining generated textual descriptions with few-shot learning for the scene classification task in the SUN database, demonstrating that the CLIP model also achieved viable accuracy rates in few-shot scene classification tasks.

3 Experiment 3 initially utilized the text encoder from CLIP to obtain text features based on 30 samples for each category, showcasing the visualization of the classification centers. As illustrated in the Figure 4, it is apparent that data from different categories are effectively distinguished, with clear inter-class gaps, proving that few-shot learning based on CLIP is entirely feasible.

The accuracy bar (Figure 5) chart intuitively reflects the model's performance across various artistic style categories. The accuracy data reveals that certain styles, such as "Abstract Expressionism," "Baroque," and "Northern Renaissance," demonstrate classification accuracies as high as 93%, 100%, and 100%, respectively, indicating the model's strong recognition capabilities for these styles. This might be due to the unique visual features of these styles, which the model can effectively learn and recognize. Conversely, some styles like "Expressionism," "Realism," and "Symbolism" have relatively lower accuracies, at 67%, 52%, and 72%, respectively, suggesting the model's difficulty in distinguishing these categories, possibly because these artistic styles are more subtly varied or nuanced in their visual representation, making accurate classification challenging. The bar chart analysis reveals that the model excels in identifying distinct art genres like Baroque and Northern Renaissance with perfect accuracy, struggles with genres that have subtler distinctions like Realism, and shows variable performance across others.

Table 2

Scene recognition accuracy in sun dataset

alcove	alley	arch	archive	attic	barn	bar	beach	butte	campus
0.85	0.88	0.97	0.84	0.75	0.64	0.75	0.92	0.43	0.62

Figure 4

This graph shows the t-SNE visualization of text features from artworks, categorized by the CLIP model, with different colors indicating various artistic styles. This graph illustrates the CLIP model's generalization capabilities on the WikiArt dataset, effectively categorizing text features from artworks into distinct artistic styles as indicated by the various colors in the t-SNE visualization. The model's ability to discern and group these styles suggests a deep understanding of the nuanced differences and similarities within the dataset

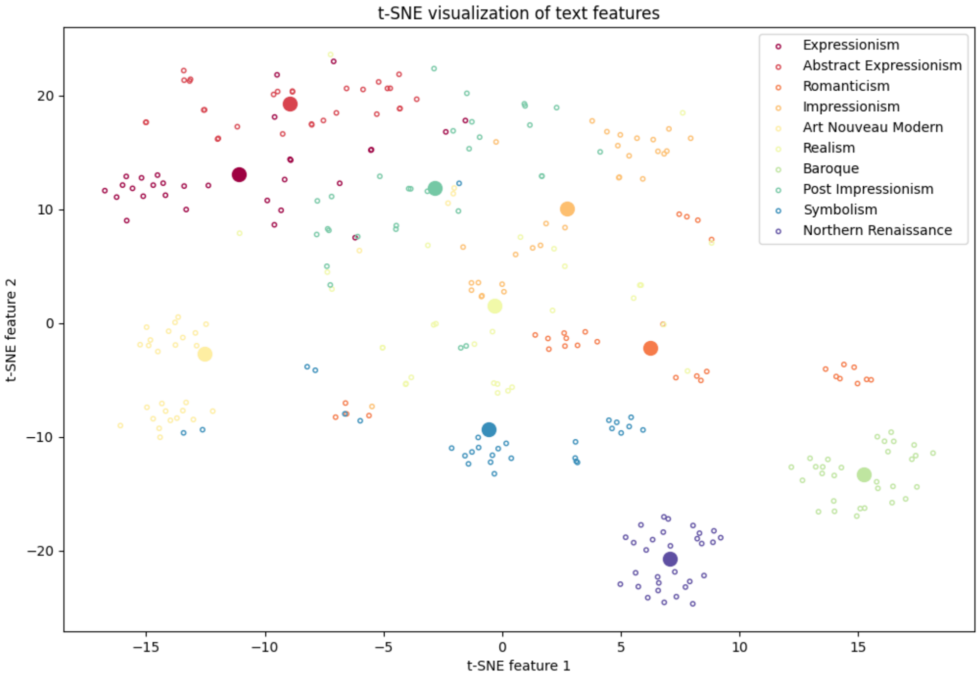


Figure 5

This chart concisely depicts the model's classification accuracy across different artistic styles. The bar graph clearly displays the clip model's performance in distinguishing between ten distinct genres of art, ranging from Abstract Expressionism to Symbolism. Notably, the model exhibits high accuracy in genres such as Abstract Expressionism, Art Nouveau Modern, and Baroque, indicating a strong capability to recognize and categorize these distinct styles

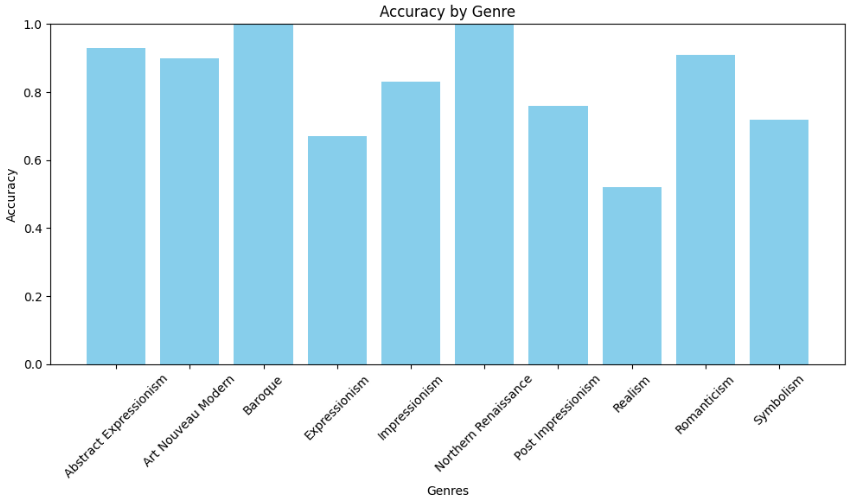
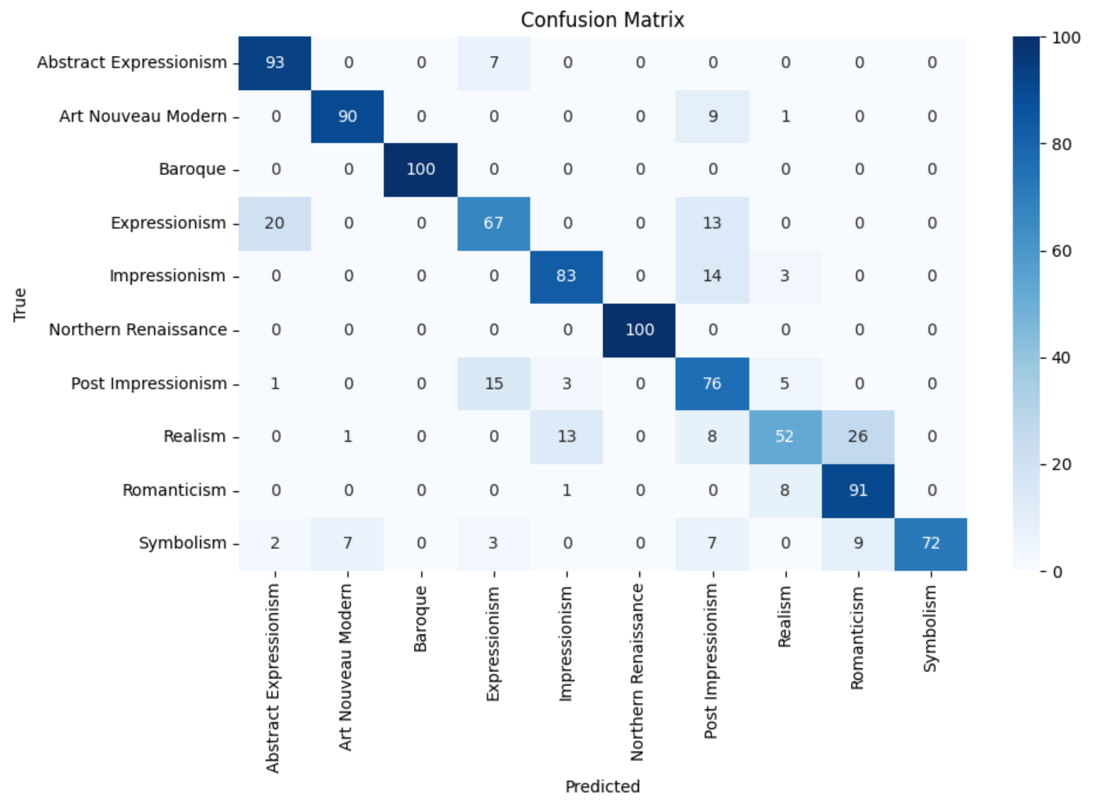


Figure 6

This matrix provides a detailed view of the model's performance, showing how each artistic style was classified against the others. The image reveals that the CLIP model, despite showing some variability in accurately distinguishing between similar artistic genres, generally performs well across a broad spectrum of styles. This indicates a robust capability of the model to recognize and categorize diverse artistic expressions effectively, suggesting its utility in applications that require nuanced understanding of visual art styles



The confusion matrix (Figure 6) offers a more detailed method of result analysis, showing how the model classifies each artistic style. While "Abstract Expressionism" and "Baroque" achieved high recognition accuracies, "Expressionism" was frequently misclassified as "Post Impressionism," highlighting the model's challenges in distinguishing between these similar styles. "Realism" was accurately classified in some instances but was also misidentified as "Romanticism" and "Post Impressionism," which may reflect the subtle overlaps in visual features among different artistic styles. Achieving an 83% accuracy rate in the final experiment demonstrates that the few-shot learning approach based on the CLIP model is quite effective. This outcome underscores the model's capability to leverage limited samples for reliable artistic style classification, illustrating the potential of CLIP in understanding and categorizing complex visual and textual data efficiently.

In our study, we employed the CLIP model to classify images by computing textual sample centers, adapting a unique approach that achieves comparable accuracy to traditional methods such as pre-trained ResNet models. Our method stands out by utilizing only the visual encoder during inference, like the operational setup of ResNet, significantly lower memory usage than the full CLIP model. Specifically, our adaptation requires only 1.63 GB of VRAM, considerably less than the complete CLIP setup. This efficiency demonstrates our method's capability to maintain high classification accuracy while optimizing resource usage, making it especially suitable for environments with restricted computational resources.

5. Conclusion

This study explores a resource-efficient method for image classification on edge devices by implementing few-shot learning utilizing the OpenAI CLIP model. Our experiments validate the feasibility and effectiveness of employing the CLIP model for few-shot image classification tasks, notably in practical applications such as artwork detection and scene classification.

The results of our study indicate that, despite utilizing only a minimal amount of labeled data, our method achieves accuracy and recall rates that closely approximate those of the pre-trained ResNet approach. This demonstrates the robust generalization capabilities of the CLIP model and its proficiency in decoding complex image content. A key advantage of our methodology is its efficiency in data usage, achieving high performance metrics with significantly fewer data compared to traditional methods that often rely on extensive labeled datasets.

Moreover, by leveraging pre-computed classification text embeddings, our approach eliminates the need for a text encoder during the inference phase. This strategic modification results in a substantial reduction in computational demands—specifically halving the memory usage compared to other large-scale visual models of similar capacity. This reduced memory footprint significantly enhances the feasibility of deploying our method on resource-constrained edge devices. Such an adaptation not only maintains accuracy but also offers a more practical and efficient solution for real-world applications, especially in scenarios where computational resources are limited.

This advantage renders the method particularly suitable for resource-constrained edge devices, offering new possibilities for deep learning applications within edge computing environments. Furthermore, our study reveals the significant impact of the specificity and format of textual descriptions on model performance, emphasizing the importance of optimizing textual descriptions to enhance classification efficacy. This insight is crucial not only for augmenting the performance of the CLIP model but also for understanding how to effectively utilize cross-modal information. Despite the positive outcomes achieved, there remains considerable work to enhance model performance and adaptability to a broader range of application scenarios. Future research could explore various data augmentation techniques, optimize the model training process, and develop new strategies for more finely tuning the pre-computed classification centers. Additionally, further investigation into managing the complexity of the associations between textual descriptions and image content, while maintaining efficiency, will be key to advancing few-shot learning and its application on edge devices. In summary, this research demonstrates the potential of few-shot learning methods based on the CLIP model for application in resource-limited environments, paving new pathways for the advancement of deep learning technologies in the realm of edge computing. We anticipate that this method will facilitate the deployment and utilization of deep learning technologies in a wider array of practical applications, particularly in scenarios sensitive to resource consumption.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

This work is partially supported by 2024 Guangdong Province Education Science Planning Project (Higher Education Special Project) ″Research on Smart Classroom Teaching Behavior Analysis Methods Based on Scene Semantic Understanding and Deep Learning Characteristic Representation (2024GXJK766)″ 2023 Guangdong Provincial Higher Vocational Education Teaching Reform Research and Practice Project ″Research on Online Teaching Quality Evaluation Method Based on Multimodal Affective State Analysis (2023JG277)″, 2024 Shenzhen Polytechnic University Quality Engineering Project ″Research on Classroom Scene Understanding and Behavior Analysis Method Based on Multimodal Attention Mechanisms (7024310268)", 2024 Higher Education Scientific Research Planning Project of the Chinese Society of Higher Education "Research on the Analysis of Teaching and Learning Deep Interaction Characteristics in Smart Classroom Environment Supported by Multimodal Data (24XH0407)", 2023 Shenzhen Education Science Planning Project"Research on the Evolutionary Mechanism and Intervention of Interpersonal Relationships among College Students Driven by Multimodal Data (rgzn23003)".

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. Proceedings.mlr.press; 2020. PMLR. https://proceedings.mlr.press/v119/chen20j.html
2. Cortes, C., Vapnik, V. Support-Vector Networks. Machine Learning, 1995, 20(3), 273-297. https://doi.org/10.1007/BF00994018
3. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A. A. Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, 2018, 35(1), 53-65. https://doi.org/10.1109/MSP.2017.2765202
4. DeVries, T., Taylor, G. W. Improved Regularization of Convolutional Neural Networks with Cutout. ArXiv:1708.04552 [Cs]. 2017. https://arxiv.org/abs/1708.04552
5. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. Nature, 2017, 542(7639), 115-118. https://doi.org/10.1038/nature21056
6. Finn, C., Abbeel, P., Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Proceedings.mlr.press; PMLR. 2017. https://proceedings.mlr.press/v70/finn17a.html
7. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. International Journal of Computer Vision. 2023. https://doi.org/10.1007/s11263-023-01891-x
8. Gao, R., Hou, X., Qin, J., Chen, J., Liu, L., Fang, Y., Zhang, Z., Shao, L. Zero-VAE-GAN: Generating Unseen Features for Generalized and Transductive Zero-Shot Learning. IEEE Transactions on Image Processing, 2020, 29, 3665-3680. https://doi.org/10.1109/TIP.2020.2964429
9. Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. The Mit Press. 2016.
10. Greenspan, H., van Ginneken, B., Summers, R. M. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Transactions on Medical Imaging, 2016, 35(5), 1153-1159. https://doi.org/10.1109/TMI.2016.2553401
11. Halevy, A., Norvig, P., Pereira, F. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 2009, 24(2), 8-12. https://doi.org/10.1109/MIS.2009.36
12. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D. A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 1-1. https://doi.org/10.1109/TPAMI.2022.3152247
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. ArXiv:1911.05722 [Cs]. 2020. https://doi.org/10.1109/CVPR42600.2020.00975
14. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90
15. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, 2012, 60(6), 84-90. https://doi.org/10.1145/3065386
16. LeCun, Y., Bengio, Y., Hinton, G. Deep Learning. Nature, 2015, 521(7553), 436-444. https://doi.org/10.1038/nature14539
17. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L. Microsoft COCO: Common Objects in Context. Computer Vision - ECCV 2014, 2014, 8693, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48
18. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., Sánchez, C. I. A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis, 2017, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005
19. Liu, H., Zhang, P., Xie, Y., Li, X., Bi, D., Zou, Y., Peng, L., Li, G. HFANet: hierarchical feature fusion attention network for classification of glomerular immunofluorescence images. Neural Computing & Applications, 2022, 34(24), 22565-22581. https://doi.org/10.1007/s00521-022-07676-6

20. Mohammad, S., Kiritchenko, S. WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art. ACLWeb; European Language Resources Association (ELRA). 2018. https://aclanthology.org/L18-1197/

21. Pan, S. J., Yang, Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191

22. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. Sutskever, I. Learning Transferable Visual Models from Natural Language Supervision. Proceedings.mlr.press; PMLR. 2021. https://proceedings.mlr.press/v139/radford21a.html

23. Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv.org. 2015. https://arxiv.org/abs/1409.1556

24. Snell, J., Swersky, K., Zemel, R. Prototypical Networks for Few-shot Learning. Neural Information Processing Systems; Curran Associates, Inc. 2017. https://papers.nips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html

25. Sun, C., Shrivastava, A., Singh, S., Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. 2017 IEEE International Conference on Computer Vision (ICCV). 2017. https://doi.org/10.1109/ICCV.2017.97

26. Sun, L., Lian, Z., Liu, B., Tao, J. HiCMAE: Hierarchical Contrastive Masked Autoencoder for Self-supervised Audio-Visual Emotion Recognition. Information Fusion, 2024, 108, 102382-102382. https://doi.org/10.1016/j.inffus.2024.102382

27. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C. A Survey on Deep Transfer Learning. Artificial Neural Networks and Machine Learning - ICANN 2018, 2018, 270-279. https://doi.org/10.1007/978-3-030-01424-7_27

28. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D. Matching Networks for One Shot Learning. ArXiv:1606.04080 [Cs, Stat]. 2017. https://arxiv.org/abs/1606.04080

29. Wang, X., Girshick, R., Gupta, A., He, K. Non-local Neural Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. https://doi.org/10.1109/CVPR.2018.00813

30. Xian, Y., Lampert, C. H., Schiele, B., Akata, Z. Zero-Shot Learning-A Comprehensive Evaluation of the Good, the Bad and the Ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9), 2251-2265. https://doi.org/10.1109/TPAMI.2018.2857768

31. Xiao, J., Ehinger, K. A., Hays, J., Torralba, A., Oliva, A. SUN Database: Exploring a Large Collection of Scene Categories. International Journal of Computer Vision, 2014, 119(1), 3-22. https://doi.org/10.1007/s11263-014-0748-y

32. Xu, S., Zheng, S., Xu, W., Xu, R., Wang, C., Zhang, J., Teng, X., Li, A., Guo, L. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. 2024. https://arxiv.org/pdf/2403.10778

33. Yao, T., Li, Y., Pan, Y., Mei, T. HIRI-ViT: Scaling Vision Transformer with High Resolution Inputs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 1-12. https://doi.org/10.1109/TPAMI.2024.3379457

34. Yosinski, J., Clune, J., Bengio, Y., Lipson, H. How Transferable are Features in Deep Neural Networks? NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, 2(2), 3320-3328.

35. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. Understanding Deep Learning (Still) Requires Rethinking Generalization. Communications of the ACM, 2021, 64(3), 107-115. https://doi.org/10.1145/3446776