PARTCLIP: HOW DOES CLIP ASSIST MECHANICAL PART IMAGE RETRIEVAL?

Shangbo Mao, Dongyun Lin, Aiyuan Guo, Yiqun Li

Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore

ABSTRACT

CLIP demonstrates impressive performance across several downstream tasks, such as zero-shot image classification. However, these tasks typically involve images from everyday scenarios, and the efficacy of CLIP in domain-specific computer vision tasks associated with the manufacturing industry remains unexplored. This paper first investigates how well CLIP understands the mechanical part images from the manufacturing industrial scenes by conducting a thorough evaluation of its performance in the mechanical part image retrieval task. It turns out that direct employment of CLIP is less effective for this task. At the same time, considering the requirement of this task for deployment on the industry platform in a factory, the large size of the CLIP model presents a practical challenge. Therefore, we explore the knowledge distillation techniques to transfer the knowledge of CLIP into a lighter Efficientnet B1. Our experimental results demonstrate that this CLIP-based knowledge distillation approach can enhance the performance of Efficientnet B1 on mechanical part image retrieval significantly.

***Index Terms*—** CLIP, image retrieval, knowledge distillation

1. INTRODUCTION

Since CLIP [1] was proposed, it has significantly benefited various downstream tasks like image retrieval [1]. However, these tasks predominantly align with everyday scenarios, which are under the similar distribution to the training data of CLIP. The effectiveness of CLIP in understanding domain-specific images, like those of mechanical part images in manufacturing and industrial applications, remains unexplored.

The motivation to explore this area stems from the reality in manufacturing factories, where hundreds of mechanical parts require ongoing maintenance and tracking. Currently, this process relies heavily on inefficient manual identification. This task will become increasingly impractical as the number of parts grows over time. Thus, there is a pressing demand for automated solutions for part maintenance to enhance efficiency and save human labor. Mechanical part image retrieval represents the initial step in such solutions for subsequent mechanical part maintenance and tracking activities.

This task presents unique challenges, including the varied quality of industrial images and the fine-grained categorization of mechanical parts. This paper aims to investigate whether CLIP can help address these challenges to improve the performance of mechanical part image retrieval. We begin with examining the effectiveness of the direct employment of CLIP on mechanical part image retrieval. The results show that CLIP itself is less effective for this task. Besides, deploying the large architecture of CLIP directly for daily part maintenance tasks on the industry platform in a factory requires substantial computational resources. Thus, this paper explores the feasibility of applying knowledge distillation to transfer the knowledge learned by CLIP into a lighter network, such as EfficientNet [2]. Our objective is to verify whether CLIP can act as an effective teacher model, thereby improving the performance of the light, distilled student network for the mechanical part image retrieval task. In summary, this paper makes two key contributions: (1) We evaluate the performances of the original CLIP model performance on mechanical part image retrieval to investigate how well CLIP represents the mechanical part images; (2) We explore distilling the knowledge in CLIP into a light model, i.e., EfficientNet [2], which achieves its significant performance increment. It demonstrates that CLIP contains the knowledge that is beneficial for mechanical part image retrieval.

2. PARTCLIP

In this section, we will present our research in detail to address the question: How does CLIP assist mechanical part image retrieval?

2.1. Original CLIP

The straightforward method to answer this question is using the original pretrained CLIP model as an off-the-shelf image feature extractor, as shown in Figure 1(b). This approach allows us to directly assess the effectiveness of CLIP on mechanical part image retrieval. Moreover, to establish a baseline for this task, we fine-tune a compact model i.e., EfficientNet B1 as illustrated in Figure 1(a).

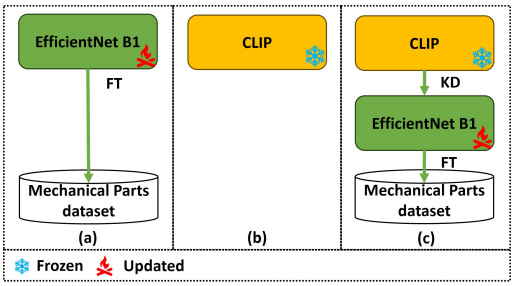

Fig. 1. (a) Fine-tune the EfficientNet B1 model as baseline; (b) Use the original CLIP; (c) Distill the original CLIP to EfficientNet B1. ('FD' refers to 'Fine-tune'; 'KD' refers to 'knowledge distillation'.)

2.2. CLIP-based Knowledge distillation

Deploying the CLIP model directly to the industry platform for mechanical part image retrieval is inefficient. Utilizing it as an image feature extractor leads to an increased computational budget and slower feature extraction speed. Therefore, for deployment on the industry platform, it is common to select a lighter model like EfficientNet [2].

In this context, the way that CLIP assists mechanical part image retrieval becomes different. Instead of direct employment, we introduce two knowledge distillation techniques: logits distillation [3] and feature distillation [4], as displayed in Figure 2(a) and Figure 2(b), respectively. Within the knowledge distillation framework, the CLIP model is utilized as the teacher model and a light model, namely EfficientNet B1, as the student model, as shown in Figure 1(c).

In this paper, we assume the input image is denoted as x_i, and the corresponding ground-truth label is denoted as y_i. The features extracted by the teacher and student models are referred to as f_i^t and f_i^s, respectively. Additionally, the classifier layer of the teacher and student model generates the logits, denoted as z_i^t and z_i^s, respectively.

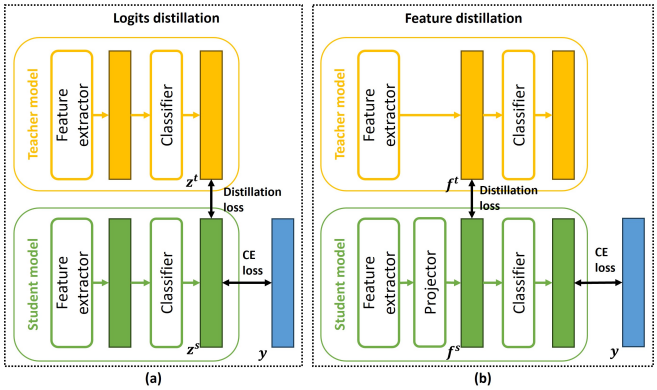

Fig. 2. The illustration of two knowledge distillation techniques: (a) Logits distillation; (b) Feature distillation.

Logits distillation loss in this paper is based on the cross entropy (CE) between the logits produced by the student and teacher model, as described by [3]. To obtain high-quality classification logits z_i^t from the teacher model, it is necessary to conduct an additional pretraining phase. This phrase involves keeping the CLIP model frozen and training an additional single-layer linear classifier. Once this pretraining is complete, the student model is trained using the logits distillation. This process involves applying 'softmax' to z_i^t and z_i^s to convert the logits to the probability p_i^t and p_i^s, respectively, as illustrated in the following equation:

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}, \tag{1}$$

where T represents the temperature, set to 2 in our experiments. The final logits distillation loss is then calculated as:

$$\mathcal{L}_{ld}(p_i^s, p_i^t) = T^2 \mathcal{L}_{CE}(p_i^s, p_i^t). \tag{2}$$

Multiplying T^2 is advised by [3]. During the training process, $\mathcal{L}_{ld}(p_i^s, p_i^t)$ is optimized alongside the CE loss between z_i^s and the ground-truth label y_i. Therefore, the total loss to be optimized is as follows:

$$\mathcal{L}_{total} = (1-\beta)\mathcal{L}_{CE}(z_i, y_i) + \beta \mathcal{L}_{ld}(p_i^s, p_i^t), \tag{3}$$

where $\beta = 0.25$.

Feature distillation loss employed here is inspired by [4], which is the Mean Squared Error (MSE) loss between the features generated by the student and teacher model:

$$\mathcal{L}_{fd}(f_i^s, f_i^t) = \lambda \mathcal{L}_{mse}(f_i^s, f_i^t), \tag{4}$$

with λ is set to 10. Due to the dimensional differences between the features produced by the teacher and student models, we introduce a single-layer linear projector after the feature extractor of the student model. This ensures that the features produced by the student model are compatible with those from the teacher model. The total loss optimized during training is similar to Equation 3. The only difference is replacing the $\mathcal{L}_{ld}(p_i^s, p_i^t)$ by $\mathcal{L}_{fd}(f_i^s, f_i^t)$.

By utilizing these two distinct knowledge distillation techniques with CLIP, we aim to provide a detailed analysis of the capability of CLIP as a teacher model to enhance the performance of the lighter model. Our analysis of both the original CLIP and its distilled EfficientNet will provide comprehensive insights into the capabilities of CLIP on this domain-specific task.

3. EXPERIMENTS

In this section, we will first introduce the dataset for evaluation and the implementation details. Following this, we present the baseline for mechanical part image retrieval. We then provide a comprehensive evaluation of the performance of CLIP on this task. Finally, we evaluate the performance of the CLIP-distilled EfficientNet B1 model on this task.

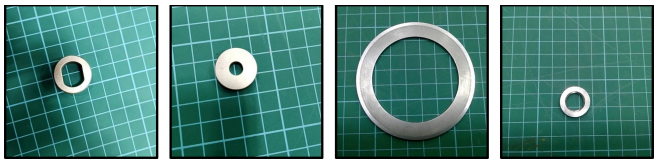

Fig. 3. Example part images from the Mechanical Parts dataset

3.1. Experiment dataset and implementation details

We evaluated the performance of mechanical part image retrieval on a new dataset, named Mechanical Parts. All images were captured during real-world manufacturing procedures. This dataset comprises 898 unique parts and a total of 35050 images. We divided the dataset into train, val, and test sets with 15236, 8945, and 10869 images, respectively. To simulate the real-world deployed scenario, we constructed the gallery based on the training set. Four example images from this dataset are presented in Figure 3.

All pretrained CLIP models used in our experiments, including those with 'ViT-B/32', 'ViT-L/14', and 'ViT-L/14@336px' encoders, were obtained from their official GitHub page [1]. In our experiments, we will only use its image encoder. The pretrained EfficientNet B1 is obtained from torchvision.models subpackage with 'EfficientNet_B1_Weights.IMAGENET1K_V1' weights. We select EfficientNet B1 as the baseline and small student model because the EfficientNet family is a representative mobile-size model specifically designed for embedded applications. Our choice, EfficientNet B1, contains 7.8M parameters [2], which is significantly fewer than those of our teacher models. These teacher models include 'ViT-B/32' with 86M parameters and 'ViT- L/14' with 307M parameters [5] image encoder from CLIP.

For fine-tuning the EfficientNet B1 model and conducting knowledge distillation, we utilize a learning rate of 0.01 with an SGD optimizer, a batch size of 64, for 60 epochs, and a step learning rate scheduler that reduces the learning rate by 0.1 every 30 epochs. All experiments are conducted using a GeForce RTX 3090 GPU. For image preprocessing, We directly applied the image preprocessing technique specified in the corresponding pretrained CLIP model. This involves resizing the image to 224×224 using the CLIP with 'ViT-B/32' and 'ViT-L/14' encoder, or 336×336 using the CLIP with the 'ViT-L/14@336px' encoder, followed by center cropping and normalizing the image.

To quantitatively evaluate the performance of mechanical part image retrieval, we used the following two metrics: (1) **Mean average precision (mAP)**, as described in[6], is computed by first sorting all registered images in descending order of relevance for each query image, then averaging the Average Precision (AP) across all query images; (2) **Recall at rank k** ($R@K$), detailed in [7], computes the percentage of queries for which the correct retrieved image appears within the top K retrieved results (In this paper, $K = \{1, 5, 10, 50\}$)

[1]CLIP: `https://github.com/OpenAI/CLIP`

Table 1. Part image retrieval performances of fine-tuning the EfficientNet B1 model (Baseline)

EfficientNet B1	mAP	R@1	R@5	R@10	R@50
EfficientNetB1(224)	0.5910	61.62	66.71	70.94	82.78
EfficientNetB1(336)	0.6055	62.86	67.73	72.30	84.39

3.2. Baseline: Fine-tune EfficientNet B1 on mechanical part image retrieval

To establish a baseline for subsequent investigations, we fine-tuned the ImageNet-pretrained EfficientNet B1 model by adding an extra single-layer linear classifier and using the CE loss. We then utilize this fine-tuned EfficientNet B1 as the feature extractor for image retrieval. We examined its performance using two different image pre-processing procedures, '224' and '336' in Table 1, to align with the image preprocessing methods of the 'ViT-B/32' or 'ViT-L/14' CLIP model and the 'ViT-L/14@336px' CLIP model, respectively.

The results in Table 1 indicate that the task of mechanical part image retrieval presents significant challenges and opportunities for improvement. This suggests that while deploying small-sized models is feasible for real-world applications, they may not deliver optimal performance. Therefore, there is a keen interest in exploring solutions that preserve the compactness of network architecture while improving retrieval effectiveness.

3.3. Original CLIP on mechanical part image retrieval

In this section, we examine the performances of three different original CLIP models as off-the-shelf feature extractors for mechanical part image retrieval. The results are presented in Table 2. Compared to the results in Table 1, there is still a performance gap between the original CLIP and the fine-tuned EfficientNet due to the differences between the images from the Mechanical Parts dataset and the training images of CLIP.

However, we observe that a more powerful CLIP model tends to yield better performance. Without any extra training, the original CLIP with the 'ViT-L/14@336px' encoder

Table 2. Part image retrieval performances of using the original CLIP model

CLIP	mAP	R@1	R@5	R@10	R@50
ViT-B/32	0.2821	47.24	59.98	66.07	79.08
ViT-L/14	0.3859	57.42	68.26	73.54	85.47
ViT-L/14@336px	0.4116	58.78	70.70	75.76	86.40

Table 3. Part image retrieval performances of distilling the teacher model i.e., CLIP to the student model i.e., EfficientNet B1 model

Teacher model	mAP	R@1	R@5	R@10	R@50
Logits distillation					
CLIP(ViT-B/32)	0.5814	62.95	68.23	72.78	83.96
CLIP(ViT-L/14)	0.5908	63.42	68.68	73.43	85.10
CLIP(ViT-L/14@336px)	0.6095	65.44	69.84	73.93	86.59
Feature distillation					
CLIP(ViT-B/32)	0.6390	64.21	68.69	71.84	85.03
CLIP(ViT-L/14)	0.6315	64.47	68.51	72.41	85.67
CLIP(ViT-L/14@336px)	0.6440	65.82	70.14	73.64	87.17

surpasses the baseline in some metrics ($R@5$, $R@10$, and $R@50$). It suggests that there is knowledge intrinsic in CLIP that is useful for mechanical part image retrieval. So we explore whether we could distill such useful knowledge into a light model in the next section.

3.4. CLIP-distilled EfficientNet B1 on mechanical part image retrieval

In this section, we use pretrained CLIP models as teacher models to distill their knowledge into the EfficientNet B1 model using two techniques: logits distillation and feature distillation. The results are presented in Table 3.

According to Table 3, it is evident that applying both knowledge distillation strategies enables the student model to achieve higher $R@K$ scores than the baseline performance reported in Figure 3.1. This success indicates that the knowledge learned by CLIP can assist EfficientNet B1 with improving its capabilities in mechanical part image retrieval. Of the two distillation techniques, feature distillation proves to be more effective than logits distillation. It leads to improvements across all numerical metrics, suggesting that the knowledge from the feature extracted by CLIP is more beneficial for image retrieval than the classification logits. The primary reason is that the extracted feature is crucial for matching and retrieval processes, whereas classification logits do not directly contribute to image retrieval. Besides, utilizing the knowledge from a better teacher model is more beneficial for the student model to perform the mechanical part image retrieval task. Feature distillation with the 'ViT-L/14@336px' CLIP performs best and achieves a significant performance increment compared to the baseline.

In summary, CLIP has been proven to be an effective teacher model within the knowledge distillation framework. This finding highlights the practical value of CLIP: it assists EfficientNet in enhancing mechanical part image retrieval, and there is no need for additional computational resources, at the point of deployment.

4. CONCLUSIONS

In this paper, we explored how CLIP can assist in the task of mechanical part image retrieval. We examined the performance of three different CLIP models. Our experimental results show that the original, pretrained CLIP model is less effective for such a domain-specific task, while they indicate that CLIP may contain useful knowledge for this task. At the same time, we realized that directly deploying CLIP to the industry platform in a factory is impractical. Therefore, we explored two knowledge distillation techniques to distill the knowledge learned by CLIP to a light model called EfficientNet B1, which proved to be highly successful. In summary, our investigation confirms that CLIP contains useful knowledge for mechanical part image retrieval, and demonstrates that CLIP can act as a superior teacher model in a knowledge distillation framework, helping the student model perform better on the mechanical part image retrieval task without extra computational resources during deployment. We have delved into the potential of CLIP for mechanical part image retrieval in this paper, expanding its range of applications.

5. ACKNOWLEDGMENT

This research is supported by A*STAR under its RIE2020 INDUSTRY ALIGNMENT FUND - INDUSTRY COLLABORATION PROJECTS (IAF- ICP) Grant No I2001E0073.

6. REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and Bengio Yoshua, "Fitnets: Hints for thin deep nets," *Proc. ICLR*, vol. 2, no. 3, pp. 1, 2015.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3456–3465.

[7] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays, "Composing text and image for image retrieval-an empirical odyssey," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6439–6448.