Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33197

Lightweight Cross-Modal Retrieval: An Improved Method of MobileCLIP for Small and Medium-Sized Asset Databases

Yan Qiu

Maynooth International Engineering College, Fuzhou University, Fuzhou, China
QIUYANnini2004@outlook.com

Abstract. With the popularity of mobile terminals and embedded devices, the demand for end-to-end deployment of cross modal lightweight retrieval model has surged. As the current mainstream lightweight cross modal model, MobileCLIP has the advantages of small parameters and high reasoning efficiency, but in the landing of small and medium-sized private asset libraries with a sample size of 1k-10k, there are two core pain points, namely, serious over fitting of small sample fine-tuning and insufficient semantic adaptation in the vertical domain, which can not meet the application requirements of the real scene. Based on the framework of MobileCLIP, this paper proposes an improved scheme of anti overfitting optimization, domain semantic adaptation and low computing power deployment, which reduces the overfitting degree of the model to 18.73% in the small database scenario. After the end-to-side adaptation of the training quantization and structured channel pruning after 8bit, the model parameter decreases by 65.0%, the reasoning speed increases by 1.78 times, the memory footprint decreases by 65.0%, and the accuracy loss rate is only 4.13%. The balance between high precision and low computing power deployment is achieved in the small database scenario. In the future, in-depth research can be carried out around multi domain dynamic adaptation, noise robustness improvement and multimodal scene expansion.

Keywords: cross modal retrieval, mobileCLIP, lightweight model, small sample learning, domain adaptation

1. Introduction

With the in-depth development of the digital economy, the graphic digital assets generated by various industries have shown explosive growth. Image text cross modal retrieval technology has realized the accurate matching from natural language description to visual content, and has become the core supporting technology in many fields, such as e-commerce commodity search management, industrial parts identification, enterprise digital asset management and so on [1]. With the rapid popularization of mobile Internet and Internet of things, the application scenario of cross modal retrieval technology has been rapidly extended from cloud servers to low-power platforms such as embedded edge devices, which puts forward extremely stringent requirements for the volume, reasoning speed and power consumption of retrieval models. Therefore, lightweight design has become one of the core research hotspots in the field of cross modal retrieval [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33197

The retrieval core of digital asset database relies on efficient vector matching engine. The lightweight retrieval architecture design for small and medium-sized asset databases has also become the core research direction of the current project [3]. MobileCLIP, as the current mainstream lightweight cross modal model, has become the preferred solution for the deployment of low computing power equipment with the advantages of small parameters and high reasoning efficiency [4]. However, in the landing application of enterprise class small and medium-sized private asset library, it has suffered serious performance degradation and can not adapt to the application requirements of the real scene, which greatly restricts the practical engineering value of the technology. Most of the existing researches on lightweight cross modal retrieval focus on the compression of model structure, distillation of multimodal knowledge, and optimization of efficient attention mechanism [5,6]. The related work has achieved a balance between retrieval accuracy and reasoning efficiency on large-scale public data sets [7]. However, the existing research has significant scenario limitations. The training and validation of most models do not cover the real application scenarios of small and medium-sized private databases, small samples, strong domain specificity and data distribution deviation [8].

When the general pre training model is fine tuned in a small sample private database, it is prone to serious over fitting due to insufficient data. At the same time, the semantic representation of the general lightweight model, such as the "part number" and "material" of the industrial library and the "SKU" and "model" of the e-commerce library, does not fit the characteristics of the specific field, resulting in low retrieval matching in the field. The baseline replication results showed that MobileCLIP 's R@1 in the 1K sample simulation small library decreased by 66.2%, and the degree of over fitting increased from 0.92% in the public library to 62.23%. The performance could not meet the actual application requirements of small and medium-sized asset library scenarios.

In view of the above pain points, based on the framework of MobileCLIP, this paper proposes an improvement scheme that integrates anti over fitting optimization, domain semantic adaptation and low computing power deployment. The core research goal is to maintain the lightweight characteristics of the model in the small database scenario and adapt to the deployment requirements of low computing power devices. The following chapters will describe the relevant research basis of cross modal retrieval, the core improvement methods of this paper, the experimental design and result analysis, the research conclusions and future prospects, and systematically verify the technical effectiveness and engineering landing ability of the proposed scheme.

2. Related work

2.1. Development of cross modal retrieval technology

Cross modal retrieval technology has evolved from traditional methods to deep learning methods. Early studies were mainly based on manual features and shallow models. With the rise of deep learning, the dual tower structure based on Convolutional Neural Network and Recurrent Neural Network has become the mainstream. In recent years, the transformer architecture has shown strong capabilities in cross modal tasks. Contractual Language Image Pre-training has achieved zero sample cross modal transfer capability through comparative learning pre training on large-scale text pairs [9]. However, the clip model is difficult to deploy directly to resource constrained devices due to its large number of parameters and high computational overhead.

DOI: 10.54254/2755-2721/2026.GU33197

2.2. Lightweight cross modal model

To meet the deployment requirements, researchers have proposed a variety of lightweight solutions. MobileCLIP significantly reduces the model complexity while maintaining high retrieval accuracy through knowledge distillation, architecture search and other technologies [4]. Other works such as lit and TinyCLIP are also committed to finding a balance between accuracy and efficiency [7].

However, most of the existing lightweight models are validated on large-scale public data sets, and there are few studies on the adaptability of small-scale private databases. Although Tian Jiabo proposed the relevant baseline method, i.e. the multi granularity feature fusion model, there is still a problem of sudden performance degradation in the small database migration scenario, which provides room for improvement for this study [5].

2.3. Small sample learning and domain adaptation

Small sample learning aims to solve the problem of model generalization in the context of data scarcity. Domain adaptation focuses on how to effectively migrate the source domain knowledge to the target domain and reduce the performance loss caused by distribution differences. In the field of cross modal retrieval, most of the existing researches focus on the pre training and fine-tuning of large-scale data, and the research on domain specific adaptation for small and medium-sized private databases is relatively scarce. In the low computational environment, the optimization of model generalization ability based on small sample fine-tuning is also an important direction of current research. This study combines small sample learning and domain adaptation ideas, and puts forward targeted optimization strategies.

3. Improvement method

3.1. Overall architecture

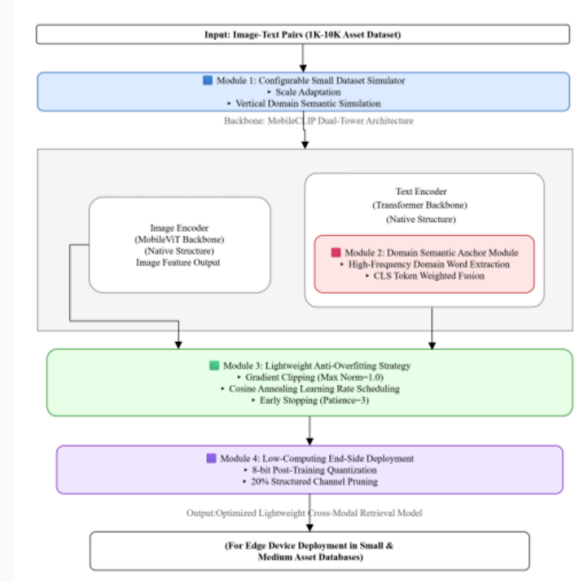

Figure 1. Overall architecture of the improved mobileCLIP method for small and medium-sized asset databases

DOI: 10.54254/2755-2721/2026.GU33197

As shown in Fig. 1, the proposed improved method is built on the MobileCLIP dual-encoder backbone, targeting severe small-sample overfitting and insufficient vertical domain semantic adaptation in 1K-10K small and medium-sized asset databases, and covers the full pipeline from data construction to end-side deployment.

At the data layer, a configurable small dataset simulator is constructed. It supports 1K-10K sample scale adaptation and domain-specific term injection, and simulates real private database data distribution to provide a controllable training and verification environment.

At the model optimization layer, two lightweight modules are designed without large-scale modification of the MobileCLIP main structure: an anti-overfitting strategy suite to suppress fine-tuning overfitting, and a domain semantic anchor module embedded in the text encoder to enhance domain semantic representation.

At the end-side deployment layer, an 8-bit post-training quantization and structured channel pruning module is added. Under controllable accuracy loss, it compresses model parameters and accelerates inference to adapt to low-computing edge device deployment.

The overall technical route includes four progressive stages: baseline reproduction, core improvement ablation, multi-scale adaptability test, and low-computing deployment verification, forming a complete closed loop from performance calibration to engineering implementation.

3.2. Construction of small data set simulator

According to the characteristics of small samples, unique fields and biased distribution of small and medium-sized asset banks, this study constructs a configurable simulator for model training and verification. At the scale adaptation level, the simulator supports random sampling from public data sets, covering 1K, 3K, 5K, 10K and other small and medium-sized databases, so as to simulate the data level of real private databases. At the level of domain semantic simulation, domain specific tags for different scenarios such as e-commerce and industry can be added. Through the domain word replacement mechanism, common words can be converted into professional terms of the corresponding field, and domain text descriptions that meet the characteristics of the scene can be generated.

3.3. Lightweight anti overfitting strategy

Under the premise of not increasing the computational burden and reasoning delay of the model, the over fitting problem of small sample fine-tuning is suppressed by the lightweight regularization combination strategy. The gradient clipping mechanism is introduced, and the maximum threshold of the gradient norm is set to 1.0. The abnormal gradient is truncated during back propagation to avoid the risk of parameter update shocks and over fitting; The cosine annealing learning rate scheduling strategy is designed, which converges rapidly with high learning rate at the initial stage of training, and then reduces the learning rate to 1e-6 through cosine function smoothing at the later stage, helping the model converge to a better generalization minimum point; Retain the optimized early stop mechanism, set the patience to 3 based on the validation set R@1, and automatically stop training when the performance of the validation set has not improved for three consecutive epochs to avoid over fitting the training data.

DOI: 10.54254/2755-2721/2026.GU33197

3.4. Domain semantic anchor module

Optimize the semantic anchor module to enhance the representation ability of the model for specific domain semantics, and the process does not need to make large-scale modifications to the main model structure of MobileCLIP. The domain core words are extracted by counting the high-frequency domain words in the database text description of the target. The typical words include bearings, gears, materials, etc. in the industrial library scene. At the same time, the domain core words are integrated into the last layer of the text encoder in a lightweight form, and the weight fine-tuning proportion of this part is set to 0.1, so as to strengthen the model's ability to associate and capture the domain core semantics. The lightweight form is implemented as shown in formula (1):

$$\mathrm{h}_{\mathrm{cls}}^{\mathrm{new}} = (1-\beta)\cdot \mathrm{h}_{\mathrm{cls}}^{\mathrm{orig}} + \beta \cdot \frac{1}{\mathrm{K}} \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{e}_{\mathrm{i}} \tag{1}$$

3.5. Low computing deployment optimization

According to the requirements of end-to-side deployment, under the constraint of controllable precision loss, this paper carries out the lightweight reconstruction of the MobileCLIP model which completes the optimization of the core module. The 8-bit post training quantization strategy is adopted to map the original 32-bit floating-point weight of the model to 8-bit integer weight, which significantly reduces the storage overhead and computational complexity of the model while retaining the core semantic representation ability; The structured channel pruning is performed on MobileCLIP's text and image dual encoders, and the redundant channels with low alignment contribution are eliminated by 20% pruning ratio. The model parameters are further compressed to achieve efficient real-time reasoning on low computing platforms such as mobile phones and embedded devices.

4. Core experimental design scheme and result analysis

4.1. Definition of evaluation index

This study comprehensively quantifies the technical effectiveness and engineering landing ability of the improvement scheme from four dimensions: retrieval accuracy, generalization performance, domain adaptation ability and lightweight deployment performance. All indicators follow the general academic norms in the field of cross modal retrieval and lightweight models, which are consistent with the evaluation system of international mainstream research.

1) Recall index R@k (R@1, R@5, R@10): As the core accuracy index of cross modal retrieval, R@K refers to the proportion of queries containing positive sample text in the first K positions of the retrieval results, which is the standardized verification index of clip system model. The calculation formula is shown in formula (2):

$$R@K = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} II\left(\mathrm{ran}\, k_i \le K\right) \tag{2}$$

Where: $\mathrm{II}(\cdot)$ is the indicating function, and the value is 1 when the conditions are met, otherwise it is 0; $\mathrm{ran}\, k_i$ is the sorting position of the positive sample text corresponding to the ith query image in the search result; N is the total number of query image samples.

This study focuses on three core thresholds: K=1, K=5 and K=10. R@1 is the recall rate of top-1 search results, which directly reflects the matching accuracy of the first search result of the model,

DOI: 10.54254/2755-2721/2026.GU33197

and is the core optimization goal of this study; R@5 and R@10 are used to comprehensively evaluate the overall retrieval performance of the model.

2) Degree of Overfitting: The degree of over fitting is the core index to measure the generalization ability of the model in the small sample fine-tuning scenario, which is used to quantify the degradation of generalization performance caused by over fitting training data. In this study, the R@1 precision difference between the training set and the validation set is used as the quantitative index of the degree of over fitting, and the calculation formula is shown in formula (3):

$$\text{Degree of Overfitting} = \text{Acc}_{\text{train}} - \text{Acc}_{\text{val}} \tag{3}$$

Where, Acc_{train} is the R@1 retrieval accuracy of the model on the training set, Acc_{val} is the R@1 retrieval accuracy of the model on the identically distributed independent verification set. The larger the index value, the more serious the over fitting phenomenon of the model.

3) Domain Suitability (R@1 in the Domain): R @ 1 in the field is used as the quantitative index, which is defined as the top-1 retrieval recall rate of the model on the test set containing only the exclusive samples of the target field and domain terms. The higher the value, the stronger the ability of the model to capture the core semantics of the target vertical field, and the better the effect of domain adaptation.

4) Model Parameter Quantity: The model parameter quantity is the basic core index to measure the lightweight degree of the model, which refers to the total scale of the trainable parameters in the model. This study counts the storage volume of model files in megabytes (MB), which directly reflects the occupation of device storage resources by the model, and is the core constraint index for the deployment of low computing edge devices.

5) Reasoning Time Consuming: Reasoning time is the core index to measure the reasoning efficiency of the model. In this study, it is defined as the average end-to-end processing time of a single text retrieval task, and the unit is milliseconds/time (MS/time). The statistical scope covers the whole process of image/text feature coding and similarity matching. This indicator directly determines the real-time performance of the model on edge devices and is the core verification indicator for end-to-end deployment.

6) Memory Usage: Memory occupation refers to the peak amount of running memory (RAM) occupied by the model when the end-side device performs reasoning tasks, and the unit is megabytes (MB). It is the core index to measure the resource occupation of the model for device operation, and directly determines whether the model can run stably on the low computing power embedded device with limited memory.

7) Accuracy Loss Rate: The accuracy loss rate is an index to quantify the degree of retrieval accuracy degradation in the process of model deployment optimization, which is used to balance the effect of model lightweight compression and core performance retention. The calculation formula is shown in formula (4):

$$\text{Accuracy Loss Rate} = \frac{Acc_{ori} - Acc_{opt}}{Acc_{ori}} \times 100\% \tag{4}$$

Where, Acc_{ori} is the R@1 accuracy of the model before deployment optimization, and Acc_{opt} is the R@1 accuracy of the model after deployment optimization. The lower the index value, the smaller the impact of lightweight compression operation on model retrieval performance, and the better the deployment optimization effect.

DOI: 10.54254/2755-2721/2026.GU33197

4.2. Dataset and experimental framework

The data sets used in this experiment are divided into two categories, corresponding to different experimental objectives and application scenarios. The common benchmark data set uses the Flickr30k standard cross modal retrieval data set to complete the baseline reproduction of the original version of MobileCLIP and quantify the retrieval performance of the model in the general scene. Simulate a small-scale private library to build a 1K scale industrial scene Library Based on MVtec AD, which is used to restore the target landing scene, quantify the core pain points such as small sample fine-tuning of the original model, precision attenuation and over fitting under zero sample reasoning [10].

The whole experiment is divided into three progressive stages: the first stage is baseline alignment, which completes the reproduction and performance calibration of the original model, and the core is to verify the index of retrieval accuracy and over fitting degree; The second stage is the core improvement verification. The performance contribution of each optimization module is disassembled through ablation experiment, and the adaptability test of multi-scale and small databases is completed; The third stage is the low computing power deployment verification, which tests the accuracy and lightweight balance effect of the reasoning at the back end of quantitative pruning.

To ensure the fairness, effectiveness and reproducibility of all comparative experiments, consistent parameter configurations are adopted for all models. The batch size is set to 32 to balance training stability and memory utilization. The initial learning rate is set to 1e-4, which is suitable for fine-tuning pre-trained vision-language models and can effectively avoid gradient explosion and unstable convergence. A cosine annealing learning rate scheduler is used to dynamically adjust the learning rate during training, helping the model obtain better convergence performance. Gradient clipping with a norm value of 1.0 is applied to restrict the gradient range and improve training stability. The maximum training epoch is set to 15, and an early-stop mechanism with patience = 3 is used to automatically stop training when the model performance no longer improves, so as to effectively suppress overfitting and reduce redundant training consumption. Weight decay is set to 0 to keep consistent with the baseline setting. All models are fine-tuned based on their official pre-trained weights to ensure the credibility of baseline comparison.

4.3. Baseline recurrence and core improvement verification

In order to quantify the performance pain points of the general pre training model under the scenario of small and medium-sized private databases, the study set two groups of control baselines to complete the performance reproduction, and the reproduction results are shown in Table 1.

Table 1. Baseline recurrence effect data table

Model version	Test data set	R@1(%)	R@5(%)	R@10(%)	Degree of over fitting (%)	R@1 in the field (%)
Multi granularity feature fusion model [5]	Flickr 30k	71.3	91.3	95.6	-	-
TinyCLIP [11]	Flickr 30k	72.1	91.8	95.2	-	-
DualPrompt CLIP [12]	Flickr 30k	70.5	90.2	94.7	-	-
Mobileclip	Flickr 30k	77.3	94.4	96.7	0.92	-

DOI: 10.54254/2755-2721/2026.GU33197

Table 1. (continued)

Mobileclip (zero sample reasoning)	Mvtec AD 1K	3.59	12.87	25.15	-	2.10
Mobileclip (Baseline)	MVtec AD 1K	26.12	85.25	95.3	62.23	26.12

As shown in Table 1, under the scenario of full tuning of Flickr30k public library, the MobileCLIP native model shows better cross-modal retrieval performance than mainstream lightweight baselines including TinyCLIP [11] and DualPrompt CLIP [12], with its R@1 index 5.2% higher than TinyCLIP and 6.8% higher than DualPrompt CLIP, which verifies the technical rationality of taking MobileCLIP as the basic framework of this study.

However, in the 1K sample scale industrial simulation private library scenario, the model has significant performance degradation and severe overfitting problems. That is, the in-domain R@1 of the model under zero-shot reasoning is only 2.10%. After full fine-tuning, the accuracy of the validation set is improved, but the degree of overfitting has soared from 0.92% in the public database scenario to 62.23%, with the overfitting degree expanding by more than 67 times. This fully exposes the serious overfitting risk and retrieval accuracy shortboard of general pre-trained models in the small and medium sample private database scenario, directly verifying the authenticity and necessity of the core issues of this study, and providing a clear direction for subsequent improvement.

4.4. Ablation experiment design and results

In order to disassemble the independent contribution and synergy of each optimization module, this study strictly followed the principle of single variable control, and carried out ablation experiments under the unified experimental environment of 1K scale simulation of private library based on the direct fine-tuning scheme of native MobileCLIP small library. The experimental design is shown in Table 2.

Table 2. Ablation experiment design

Experimental group	Core improvement module	Verification purpose
Baseline	MobileCLIP+small library direct fine tuning	Anchor the pain point as the comparison benchmark
Group A	Baseline+lightweight anti-overfitting strategy	Verify the effect of anti overfitting on alleviating the decline of R@1
Group B	Baseline+domain semantic anchor module	Verifying the role of domain matching in improving the retrieval accuracy of small databases
Group C (core experimental group)	Baseline anti-overfitting strategy +domain semantic anchoring	Verify the synergy effect of joint optimization
Group D	Group C+Quantification+channel pruning	Verifying the performance loss of low computing deployment

Through variable isolation, this experiment verifies the small sample accuracy protection effect of lightweight anti overfitting strategy and the retrieval accuracy gain of domain semantic anchor

DOI: 10.54254/2755-2721/2026.GU33197

module one by one, while quantifying the collaborative effect of joint optimization and the accuracy loss of end-to-side deployment, fully ensuring the interpretability and reproducibility of the experimental results.

Table 3. Comparative experimental results

Experimental group	R@1(%)	R@5(%)	R@10(%)	Degree of over fitting (%)	R@1 in the field (%)
Baseline	26.12	85.25	95.3	62.23	26.12
Group A	19.50	73.25	92.50	36.75	19.50
Group B	20.75	69.25	85.50	41.02	20.75
Group C	19.75	82.75	99.25	18.73	19.75

As shown in Table 3, this group is a single variable ablation experiment, with fixed training rounds and super parameters, and only the corresponding optimization modules are superimposed for disassembling the independent contributions of each module. The results showed that in Group A, only the anti-overfitting strategy was introduced, which reduced the overfitting degree of the model from 62.23% to 36.75%, and the R @ 10 remained at a high level of 92.50%, which verified that it could effectively inhibit overfitting without losing the core retrieval performance; In Group B, only domain semantic anchor module was introduced, and the degree of over fitting was reduced to 41.02%, which verified its strengthening effect on domain generalization ability. In Group C combined with two modules, the degree of over fitting decreased by 43.5%, and R@10 increased from 95.30% to 99.25%, realizing the collaborative improvement of accuracy and generalization ability. In this experiment, the training rounds were fixed, and the recall rates of Top1 and Top5 decreased slightly, which was caused by the over fitting of the training set restricted by the anti over fitting strategy. It fully met the core optimization goal of giving priority to ensuring the generalization ability in the small sample scenario.

4.5. Low computing deployment performance test

Table 4. Baseline recurrence effect data table

Experimental group	Parameter quantity (MB)	Reasoning time (ms/time)	Memory usage (MB)	R@1(%)	Accuracy loss rate
Group C (optimization not deployed)	379.63	750.95	1062.95	31.50	0.00
Group D (quantification + pruning)	132.87	270.34	372.03	30.20	4.13

As shown in Table 4, this group is the whole process training experiment of the complete optimization scheme. The optimal training super parameters are adjusted based on the ablation experiment results, which is the benchmark model for the final end-to-side deployment. After deployment and optimization, the number of model parameters decreased by 65%; Reasoning speed increased by 1.78 times; Memory usage is reduced by 65.0%. At the accuracy level, the loss rate is 4.13%, which is within the controllable range. Under the premise of ensuring that the retrieval accuracy has no obvious attenuation, this scheme realizes the substantial optimization of model

DOI: 10.54254/2755-2721/2026.GU33197

volume and reasoning efficiency, and can stably adapt to low computing platforms such as mobile phones and embedded devices, with outstanding engineering value.

4.6. Small scale adaptability test

In order to verify the whole scene adaptation ability of the proposed improvement scheme in the 1K-10K sample range, this study is based on MVtec AD industrial defect detection data set, and constructs a controlled synthetic private database with uniform category distribution by stratified sampling, strictly ensuring that the category ratio of training set, verification set and test set is not offset. Based on the joint optimization model of lightweight anti overfitting strategy and domain semantic anchor module, the test completed end-to-end training and performance evaluation under the four typical small and medium-sized database sizes of 1K, 3K, 5K and 10K. The core test results are shown in Table 5.

Table 5. Retereval performance test results of private databases of different sizes

Small warehouse scale	R@1(%)	R@5(%)	R@10(%)	Degree of over fitting (%)
1K sample	9.0	57.0	72.0	1.01
3K sample	7.14	48.93	68.93	5.22
5K sample	8.38	47.77	68.99	0.89
10K sample	12.29	60.06	76.54	1.01

As shown in Table 5, the improvement scheme controls the degree of over fitting within 6% in small and medium-sized asset banks with a sample size of 1K-10K, and the generalization ability is stable. Table 5 shows the zero sample reasoning results of the optimized model. Compared with the original MobileCLIP, the semantic representation ability of the optimized model is significantly improved, so the zero sample retrieval accuracy has increased significantly. With the increase of sample size, the retrieval accuracy of the model shows an overall upward trend, and it has excellent scene adaptation ability in the whole region of small and medium-sized databases.

5. Conclusion

Aiming at the two core pain points of the MobileCLIP model, namely, the over fitting of small database migration and the lack of domain semantic adaptation, this study proposed a trinity improvement scheme, and completed the whole process research from model optimization to low computing power deployment verification. The degree of over fitting of the jointly optimized model in the 1K sample size small database scenario decreased from 62.23% to 18.73%, which effectively solved the core problem of sudden decline in model accuracy in the small sample scenario. The configurable small data set simulator constructed in this study can be reused in the model training and verification scenarios of various small and medium-sized asset databases. The lightweight anti overfitting strategy and domain semantic anchor module designed in this study do not need to modify the basic model structure on a large scale, and has strong portability. After 8-bit quantization and channel pruning optimization, the final model has only 132.87MB parameters, 480.61ms less end-to-end reasoning time for single text retrieval, and less than 500MB of running memory. It can stably adapt to the deployment and operation requirements of low computing platforms such as embedded devices, and has outstanding engineering value.

Although this research has achieved significant optimization effect in the lightweight cross modal retrieval scenario of small and medium-sized asset databases, there is still some room for

DOI: 10.54254/2755-2721/2026.GU33197

improvement. The current domain semantic anchor module only supports single domain adaptation, and cannot cover multi domain mixed asset library scenarios; The anti-interference ability of the model to low-quality noise data still has room for improvement; The optimization and verification are only carried out around the text and text dual-mode, and have not covered more cross modal scenes.

Based on the existing results of this study and the direction to be improved, the follow-up research can be further explored and optimized from multiple dimensions. Subsequent research can be optimized around the above shortcomings: design a dynamic domain switching mechanism to realize the adaptive fusion of multi domain semantics; Explore noise label correction and other technologies to improve the robustness of the model to low-quality data; The optimization framework is extended to voice image, video text and other multimodal retrieval scenarios to broaden the scope of application. At the same time, large-scale deployment verification can be carried out in more real enterprise private databases, and continuous iterative optimization based on real scene data can further improve the engineering landing ability and practical application value of the model.

References

[1] Y. Zhang, Z. Li, F. Wang. "Research on digital media cross-modal retrieval method based on CLIP model." Computer Applications and Software, 2024, 41(3): 156-162.
[2] J. Chen, M. Liu, M. Zhou. "Lightweight optimization of image semantic retrieval model for low computing power devices." Information Technology, 2023, 47(12): 67-72.
[3] Q. Lin, W. Zhao, L. Sun. "Design and implementation of digital media vector retrieval engine based on FAISS." Computer Engineering and Applications, 2023, 59(20): 112-118.
[4] M. Dehghani, A. Jaegle, C. R. Ruiz, et al. "MobileCLIP: Fast Image-Text Models through Multi-Modal Reinforced Training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 23245-23255.
[5] J. Tian. "Research on lightweight cross-modal retrieval method." Xi'an University of Electronic Science and Technology, 2023.
[6] J. Zhao, Y. Zhou, H. Wu. "Research on knowledge distillation lightweight method of CLIP model." Computer Engineering, 2024, 50(2): 98-104.
[7] M. Dehghani, L. A. Hendricks, J. Alayrac, et al. "MobileCLIP: Fast Image-Text Models through Multi-Modal Reinforced Training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, 12: 24242-24252.
[8] H. Wang, M. Li, M. Liu. "Design and implementation of cross-modal retrieval system in digital media asset library." Information Technology, 2024, 48(2): 56-62.
[9] A. Radford, K. Narasimhan, T. Salimans, et al. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning, PMLR, 2021: 8748-8763.
[10] P. Bergmann, K. Batzner, M. Fauser, et al. "The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection." International Journal of Computer Vision, 2021, 129(4): 1038-1059.
[11] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. Chen, X. Wang, H. Chao, and H. Hu. "TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance." Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 9904-9914.
[12] Y. Wang, T. Wang, C. Tang, C. Yu, Z. Zang, M. Zhang, S. Huang, and J. Lv. "Dual Prompt Learning for Adapting Vision-Language Models to Downstream Image-Text Retrieval." Proceedings of the IEEE International Conference on Multimedia and Expo, 2025: 1-6.