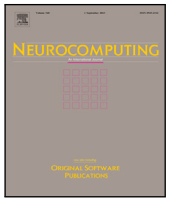
Neurocomputing 587 (2024) 127687

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Vision transformers are active learners for image copy detection

Zhentao Tan [a], Wenhao Wang [b,*], Caifeng Shan [c,d]

[a] Academy for Advanced Interdisciplinary Studies (AAIS), Peking university, Beijing, China
[b] ReLER Lab, University of Technology Sydney, Sydney, Australia
[c] College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China
[d] School of Intelligence Science and Technology, Nanjing University, Nanjing, China

ARTICLE INFO

Communicated by J. Han

Keywords:
Image copy detection
Vision transformer
Active learning

ABSTRACT

Image Copy Detection (ICD) is developed to identify and track duplicated or manipulated images. The majority of existing methods rely on Convolutional Neural Networks (CNNs) and are trained using unsupervised learning techniques, which leads to subpar performance. We discover that by carefully designing the training process, Vision Transformer (ViT) backbones yield superior results. Specifically, directly training a ViT for ICD often leads to overfitting on the training images, which in turn results in poor generalization to unseen (test) images. Consequently, we initially train a CNN (such as ResNet-50), and during the ViT training, the distances between the features of CNN and ViT are regularized. We also incorporate an active learning method to further enhance performance. Notably, due to the visual discrepancy between auto-generated transformations and those used in the query set, we incorporate a small number (approximately 0.5% of unlabeled training images) of manually produced and labeled positive pairs. Training models on these pairs results in a significant performance boost though with little cost. Experimental findings demonstrate the effectiveness of our approach, and our method achieves state-of-the-art performance. Our code is available at: https://github.com/WangWenhao0716/ViT4ICD.

1. Introduction

Image Copy Detection (ICD) is a cutting-edge technology that identifies and locates instances of image duplication, manipulation, or unauthorized reproduction, thereby safeguarding intellectual property and fostering digital integrity. A demonstration of ICD can be seen in Fig. 1. Numerous methods [1–5] rely on the CNN backbone and employ self-supervised learning techniques to achieve this goal. Given the success of Vision Transformers (ViT) [6] in various computer vision tasks, we believe that ViT holds significant potential in the ICD domain. Moreover, self-supervised learning may not be sufficient for the ICD task, and manual labeling processes can be both time-consuming and costly. Consequently, we propose incorporating active learning into ICD, aiming to enhance performance while minimizing expense.

Training a ViT for ICD is not a straightforward task due to its tendency to overfit. To address this issue, we propose a training method called "regularized training" to adapt ViT for ICD. CNNs possess an inductive bias and are easier to train, so we first train a CNN, such as ResNet-50, as the base model for ICD. Using the trained CNN, we obtain the feature distribution of the original images (those without transformations). During the ViT training process, our objective is to align the features extracted by the ViT with the feature distribution obtained from the CNN. To achieve this, we propose a loss function that regularizes the L_2 distance between image features extracted by the CNN and the ViT. After optimization, the two feature distributions become similar. With a fixed feature distribution for the original images, the ViT can freely arrange the features of the edited copies in the feature space. This regularized training process unlocks the potential of the ViT, leading to improved performance.

To overcome the limitations of self-supervised learning in a time and cost-efficient manner, we incorporate active learning into the ICD community. As depicted in Fig. 2, we observe that training images generated by auto-generated transformations display visual discrepancies compared to the query images. Consequently, there is a need to introduce manually produced and labeled positive pairs. However, considering the high costs associated with manual production and labeling, we suggest an active learning approach that enables model training on a limited amount of data. To avoid overfitting on this small dataset, we maintain the use of losses in the training of CNN and ViT. Subsequently, we minimize the feature distances between manually labeled positive pairs and maximize the feature distances between negative pairs.

* Corresponding author.
E-mail address: wangwenhao0716@gmail.com (W. Wang).

https://doi.org/10.1016/j.neucom.2024.127687
Received 22 January 2024; Received in revised form 10 March 2024; Accepted 8 April 2024
Available online 15 April 2024
0925-2312/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

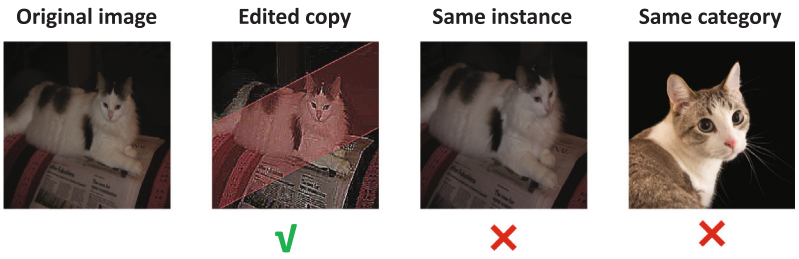


Fig. 1. The demonstration for Image Copy Detection (ICD). Our goal is to identify edited copies rather than images that belong to the same instance or category.

Fig. 2. The demonstration for the visual discrepancies between the transformations utilized to generate query images and the auto-generated transformations employed for creating training images. The query image is transformed by adding emojis and applying Legofy [7], whereas the training image is generated by adding random noise.

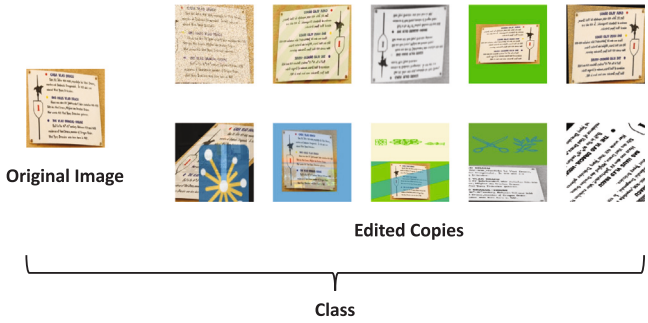


Fig. 3. The demonstration for generating edited copies. An original image and its edited copies form a training class for ICD.

In the Experiments section, we conduct a series of analyses to evaluate our method. First, we compare our method to state-of-the-art algorithms, demonstrating its superiority. Second, we establish the effectiveness of the proposed regularized learning by contrasting it with the vanilla training of ViT. Third, we carry out an ablation study for the proposed active learning method. Fourth, we visualize the matching results obtained from various models. Finally, we examine the changes in cosine similarity during the last two training stages by means of visualization.

To sum up, this paper makes the following contributions:

1. We introduce regularized training to improve ViT performance in ICD by aligning features with a CNN-based distribution.
2. We propose an active learning approach that efficiently addresses self-supervised learning limitations by leveraging a small amount of manually labeled data.
3. Extensive experimental results demonstrate the effectiveness of both the proposed regularized learning and active learning methods.

2. Related work

2.1. Image copy detection

In the past, image copy detection methods have primarily relied on unsupervised learning techniques. For instance, BoT [3] uses deep metric learning by creating a "class" by augmenting an image multiple times. SSCD [8] modifies the architecture and training objective of SimCLR [9] to explore the effectiveness of self-supervised learning methods on ICD. Another earlier work, Multigrain [10], uses joint training to generate image embeddings at multiple levels, including class, instance, and copy. While all of these methods rely solely on unsupervised learning, this paper proposes an active learning approach to achieve significant performance improvements by labeling only a few samples.

2.2. Vision transformer

Transformers [6] have achieved state-of-the-art results in various computer vision tasks, including image classification [11,12], image segmentation [13,14], object detection [15–17], video understanding [18,19], and object re-identification [20]. Despite these successes, adapting the ViT [6] to new tasks remains challenging due to (1) the difficulty in training and convergence, given the large number of parameters, and (2) the potential for overfitting on training data, owing to the lack of inductive bias in ViT. This paper tries to apply ViT to ICD tasks, demonstrating improved performance compared to CNNs [21].

2.3. Active learning

Active Learning (AL) [22–29] is a technique that concentrates on maximizing performance improvements while minimizing the number of labeled samples needed. This approach involves meticulously selecting the most valuable examples and presenting them to an expert, such as a human annotator, for labeling. The goal of active learning is to decrease labeling costs while maintaining high levels of performance. In the case of the ICD task, there are typically no labels available. However, we have discovered that by adding only a small number of manually produced and labeled positive pairs (an image and its edited copy), the ICD accuracy is significantly enhanced. Therefore, we introduce active learning to the ICD community.

3. Method

The proposed approach includes three stages, *i.e.* the ICD baseline, regularized learning for ViT, and tuning on produced and labeled pairs. These stages are elaborated in the following sections.

3.1. The ICD baseline

In this stage, we train a CNN (ResNet-50 [30]) as the baseline to extract features for ICD.

Generate edited copies. Given the original image, we use predefined transformations to generate a training dataset. Specifically, we

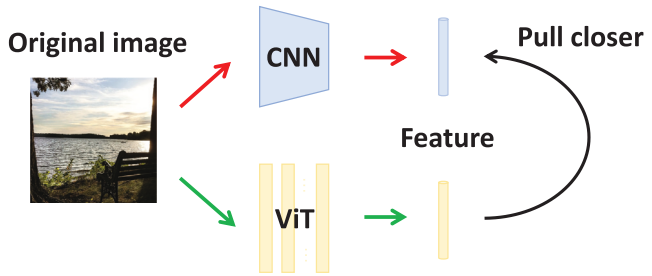


Fig. 4. The demonstration of the proposed regularized learning. During the training process, features extracted from original images by the ViT are pulled closer to those extracted by the trained CNN.

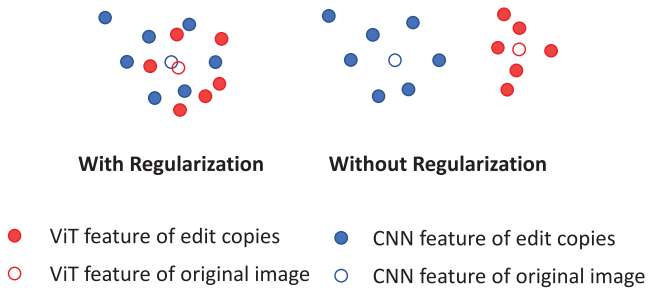


Fig. 5. The comparison between with and without regularization. With regularization, the features distribution of original images are similar between CNN and ViT.

randomly select various transformations and utilize them to convert the original image into multiple modified versions. The original image and its edited copies together comprise a training class. A demonstration is shown in Fig. 3.

Perform deep metric learning. Utilizing auto-generated training classes, we perform deep metric learning to train a CNN model. This can be achieved using pairwise training [31,32], classification training [33–35], or a combination of both methods. To simplify the process, we exclusively use CosFace [35] as our loss function, denoted as $\mathcal{L}_{mtr}$.

3.2. Regularized learning for ViT

As depicted in Fig. 4, we introduce a training approach called regularized learning, specifically designed for training ViT in the context of ICD.

Denote the original image as x_o, the trained CNN as f, and the ViT as g. The features of the original image extracted by CNN and ViT can be represented by $f\left(x_o\right)$ and $g\left(x_o\right)$, respectively. We propose the regularized loss to help the training of ViT:

$$\mathcal{L}_{reg} = \sum_{i=0}^{N} \| \frac{f\left(x_{o_i}\right)}{\|f\left(x_{o_i}\right)\|_2} - \frac{g\left(x_{o_i}\right)}{\|g\left(x_{o_i}\right)\|_2} \|_2, \tag{1}$$

where: N is the number of the original images, and $\|\cdot\|_2$ is L_2 normalization. Therefore, when using ViT backbone, the final loss is:

$$\mathcal{L}_{final} = \mathcal{L}_{mtr} + \lambda_r \cdot \mathcal{L}_{reg}, \tag{2}$$

where λ_r is the balance parameter. The advantage of this regularization approach is that, after training, the features extracted from the original images by ViT will closely resemble those extracted by the CNN. This transfers the inductive bias from the CNN to ViT, preventing ViT from learning an ungeneralizable feature distribution. A comparison of using this regularization loss versus not using it can be seen in Fig. 5. Experimental results will demonstrate the effectiveness of this design.

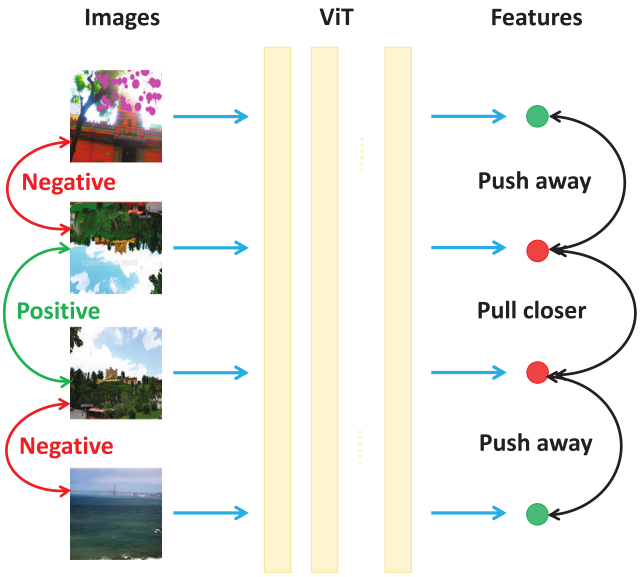


Fig. 6. The demonstration of active learning. By tuning ViT on a small number of manually produced and labeled positive pairs, significant performance improvement is observed.

3.3. Tuning on produced and labeled pairs

In this section, we introduce an active learning approach to further enhance performance. It is shown in Fig. 6. Our findings indicate that the auto-generated transformations utilized for training the CNN and ViT models exhibit visual discrepancies with query images in the test set. As a result, we aim to employ manually-generated edited images. Given that manual generation and labeling of positive pairs (original image and its edited counterpart) can be time-consuming and costly, we investigate an active learning strategy that tunes models using an extremely limited number of image pairs (approximately 0.5% of unlabeled training images). Despite the low labeling cost, the proposed tuning method proves to be effective and substantially improves performance.

Denote the trained ViT as g_t, two images in a positive pairs as x_p^1 and x_p^2, and x_n^j as the hardest negative of x_p^j $(j = 1, 2)$. Therefore, we have the training objectives:

$$\mathcal{L}_{pos} = \sum_{i=0}^{M} \| \frac{g_t\left(x_{p_i}^1\right)}{\|g_t\left(x_{p_i}^1\right)\|_2} - \frac{g_t\left(x_{p_i}^2\right)}{\|g_t\left(x_{p_i}^2\right)\|_2} \|_2, \tag{3}$$

$$\mathcal{L}_{neg} = \frac{1}{2} \sum_{i=0}^{M} \left(\left\| \frac{g_t\left(x_{p_i}^1\right)}{\|g_t\left(x_{p_i}^1\right)\|_2} - \frac{g_t\left(x_{n_i}^1\right)}{\|g_t\left(x_{n_i}^1\right)\|_2} \right\|_2 \right. \tag{4}$$

$$\left. + \left\| \frac{g_t\left(x_{p_i}^2\right)}{\|g_t\left(x_{p_i}^2\right)\|_2} - \frac{g_t\left(x_{n_i}^2\right)}{\|g_t\left(x_{n_i}^2\right)\|_2} \right\|_2 \right), \tag{5}$$

$$\mathcal{L}_{final} = \mathcal{L}_{mtr} + \lambda_r \cdot \mathcal{L}_{reg} + \lambda_{pn} \cdot \left(\mathcal{L}_{pos} - \mathcal{L}_{neg}\right), \tag{6}$$

where M is the number of the positive pairs, and λ_{pn} is the balance parameter. This active learning approach not only allows the ViT to efficiently learn manually-produced transformations, but also prevents overfitting on a limited number of image pairs by incorporating additional loss functions. Experimental results will demonstrate changes in cosine similarity (between two images in the positive pairs) and the effectiveness of fine-tuning ViT on this small dataset of image pairs.

Table 1
Comparison with state-of-the-arts. Our method consistently demonstrates superior performance when compared to methods from research papers and those from the Facebook AI Image Similarity Challenge.

Method	μAP (%) ↑	$R@P90$ (%) ↑	$R@1$ (%) ↑	$R@10$ (%) ↑
GIST [40]	15.21	10.44	24.06	24.93
Multigrain [10]	36.49	26.83	44.96	49.31
ASMK [41]	37.16	20.35	47.85	49.31
SSCD [8]	72.81	63.76	78.18	81.90
BoT [3]	72.73	67.25	78.91	82.15
EfNet [1]	75.81	67.61	79.91	83.57
CNNCL [5]	77.43	68.97	81.89	**85.40**
Ours	**78.60**	**73.60**	**82.49**	84.27

Table 2
The ablation study for each component in our proposed approach.

Method	μAP (%) ↑	$R@P90$ (%) ↑	$R@1$ (%) ↑	$R@10$ (%) ↑
Baseline	74.14	69.27	77.96	79.83
ViT-Only	54.32	50.22	58.23	61.11
ViT-Reg	74.68	69.27	78.89	80.79
Random	74.55	69.75	78.73	80.43
AL-10%	75.30	70.55	80.07	81.87
AL-20%	76.12	70.99	80.73	82.55
AL-50%	77.52	72.40	81.29	83.25
AL	**78.60**	**73.60**	**82.49**	**84.27**

4. Experiments

4.1. Dataset and metrics

Dataset. We use DISC21 [36] for evaluation the proposed method. DISC21 [36] is a comprehensive ICD benchmark that offers many advantages for researchers and developers. First, it contains one million training images and one million gallery images, providing a rich data source for exploring deep learning algorithms. Second, it includes many complex patterns, *i.e.* queries generated by sophisticated transformations, which greatly challenge ICD algorithms. Third, the number of distractor queries is four times that of queries with true matches, providing a realistic setting for the ICD task.

Metrics. We consider 4 common-used metrics, *i.e.* μAP, $R@P90$, $R@1$, and $R@10$. μAP is the area under the precision–recall curve when all matching pairs are taken into account. $R@P90$ refers to the threshold where 90% of the relevant copies have been retrieved. $R@k$ measures the proportion of relevant items that appear within the top k results returned by an algorithm, compared to the total number of relevant items available.

4.2. Experimental settings

We use PyTorch [37] to implement our approach. For all training stages, the batch size is 128, the number of GPUs is 4, and iterations are 8000. The training epochs for the three stages are 25, 25, and 10, respectively. ResNet-50 [30] is pre-trained on ImageNet [38], and ViT [6] is also pre-trained on ImageNet [38] by DeiT [39]. The balance parameters in the loss functions are set as $\lambda_r = 100$ and $\lambda_{pn} = 1$. We adopt the standard PK sampling, *i.e.* in each batch, there are $P = 32$ classes and each class has $K = 4$ images. Specifically, each group of K images are generated with different transformations from a single image. In the active learning stage, there are only about 5000 positive pairs.

4.3. Comparison with state-of-the-arts

We compare the proposed method with state-of-the-art methods in Table 1: GIST [40], Multigrain [10], ASMK [41], and SSCD [8] are from research papers while BoT [3], EfNet [1], and CNNCL [5] are the winning solutions of Facebook AI Image Similarity Challenge. Our findings reveal that: (1) Traditional descriptors, such as GIST [40], struggle to effectively handle the complexities of modern ICD tasks. In comparison to other methods, GIST demonstrates significantly inferior performance. (2) Some earlier deep learning-based ICD algorithms, like Multigrain [10] and ASMK [41], can recognize simple transformations but are unable to efficiently manage more complex ones. (3) When compared to SSCD [8], our approach achieves improvements of +5.79%, +9.84%, +4.31%, and +2.37% for the four metrics, respectively. This highlights the effectiveness of our proposed ViT training and active learning approach. (4) Interestingly, our straightforward method still outperforms the more complex winning solutions. This comparison may be considered unfair, as these winning solutions employ sophisticated techniques like detection augmentation [42], multi-scale testing, multi-model ensembles, and post-processing, while our approach remains direct and uncomplicated. Additionally, we provide the performance of our method on VSC2022 [43]: our method achieves 85.1%μ AP in the VCD track and 77.1%μ AP in the VCL track, yielding a +24.6% and +33.0% performance improvement compared to the baseline method, respectively [44].

4.4. Ablation studies

Our ICD baseline is strong. We train a ResNet-50 as our baseline, detailed in Section 3.1. As observed in Table 2, the μAP reaches 74.14%, surpassing five state-of-the-art methods presented in Table 1. This demonstrates that even without incorporating specific design techniques, our baseline performs commendably. Furthermore, this strong baseline lays the foundation for achieving even higher results in subsequent experiments. Additionally, by utilizing a strong baseline instead of a weak one, we can truly identify and distinguish effective methods.

The effectiveness of regularized learning for ViT. In this section, we compare our proposed regularization training (ViT-Reg) with vanilla training (ViT-Only) as shown in Table 2. Vanilla training refers to training a ViT using the same process as for CNNs. Although we find that vanilla training can achieve convergent results, the final performance is significantly lower by −20.36%, −19.05%, −20.66%, and −19.68% when compared to regularization training, respectively. This substantial performance discrepancy demonstrates that directly training a ViT for ICD is infeasible. Moreover, we observe that ViT-Reg outperforms the baseline, with improvements of +0.54% in μAP and +0.93% in $R@1$. This indicates that by carefully designing the training procedure for ViT, the ViT backbone can achieve better performance than traditional CNNs.

The effectiveness of active learning. The effectiveness of active learning is demonstrated in Table 2. Despite using only a small number of positive pairs for active learning, we observe a significant performance improvement across the four metrics, with gains of +3.92%, +4.33%, +3.60%, and +3.48%, respectively. We attribute this to two factors: (1) the careful selection and labeling of image pairs, which minimizes labeling costs while maintaining performance; and (2) the specifically designed tuning process that enables the utilization of a limited number of image pairs. For comparison, we randomly select positive pairs from the training classes (denoted as Random in Table 2). No performance improvement is observed for the four metrics. Additionally, we investigate scenarios where the manually produced and labeled image pairs are extremely low (such as 0.05% of the original images, denoted as AL-10%). Experiments demonstrate that improvements can be achieved even with an extremely low number of labeled image pairs.

The success of our approach can be credited to the combination of active learning and the customized training process. The active learning strategy ensures that our model is able to learn from the most informative and relevant data, which contributes to the performance improvements observed. Furthermore, the tailored training process for

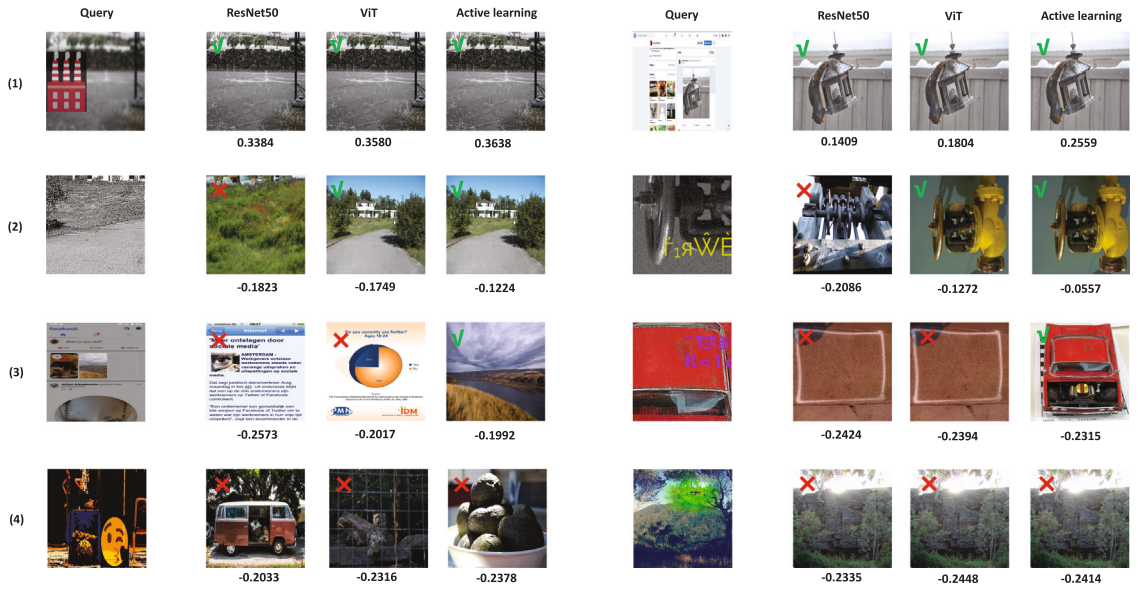


Fig. 7. The visualization of the matching results from different models. "×" and "✓" symbols indicate incorrect and correct matches, respectively. Given a query, the similarity score between a reference and the query is displayed at the bottom of the reference.

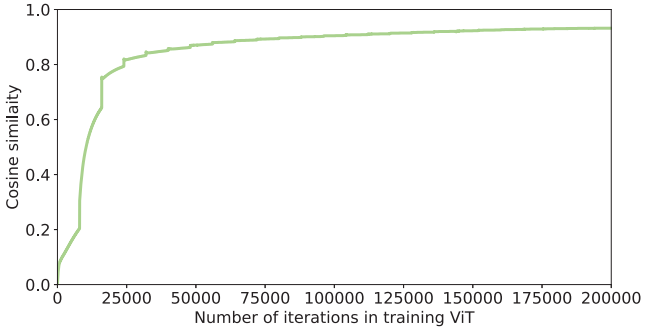


Fig. 8. The change of cosine similarities between training features of ViT and CNN in relation to iterations. A sharp increase is observed during the first few iterations.

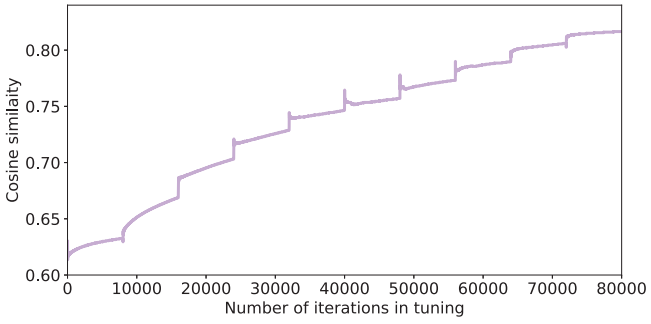


Fig. 9. The change of cosine similarities between the labeled positive pairs. The cosine similarity increases from 0.62 to 0.82 gradually.

the ViT backbone, as well as the regularization techniques, enable our model to achieve better results when compared to traditional CNNs and other state-of-the-art methods.

4.5. Discussion

Visual comparison between different trained models. We select the top-1 matches from the trained models after completing three training stages. These models are denoted as "ResNet50", "ViT", and "Active Learning" in Fig. 7. The first row illustrates that all three models successfully achieve true matching. However, the similarity scores vary, with the later models displaying higher scores, indicating more confident matching. In the second row, "ResNet50" produces an incorrect match, while "ViT" and "Active Learning" continue to show true matches. This demonstrates that a carefully trained ViT outperforms ResNet50. Additionally, the similarity score increases following the proposed active learning process. In the third row, both "ViT" and "ResNet50" generate incorrect top-1 matches, while "Active Learning" is still able to retrieve true matches, reaffirming the effectiveness of "Active Learning". Lastly, we present some failure cases in which all three models are unsuccessful. The queries in row (4) are highly unrecognizable, even to the human eye, making it understandable that our model falters. This suggests that ICD remains an unresolved issue that warrants further investigation.

The change of cosine similarities between training features of ViT and CNN. As illustrated in Fig. 8, we visualize the changes in cosine similarities between training features obtained from ViT and CNN during the regularized learning stage. We observe the following: (1) At the initial stage, the cosine similarity is close to 0, indicating that features derived from the same image by ViT and CNN are nearly orthogonal. This highlights the necessity of introducing regularization loss. (2) As we train the ViT, the cosine similarity increases to nearly 1, demonstrating the success of our regularized training approach. (3) The cosine similarity experiences a sharp increase during the first 10% of iterations, suggesting that, throughout the majority of the ViT training time, the feature distribution of original images are fixed, leading to improved performance.

The change of cosine similarities between the labeled pairs. We assess the change in cosine similarities between labeled pairs as shown in Fig. 9 and make the following observations: (1) The initial cosine similarity is approximately 0.62, indicating that training on auto-generated images is beneficial but not perfect. (2) As active learning progresses, the cosine similarity gradually increases to 0.80, signifying the success of our training procedure. (3) The upper limit of cosine similarity is about 0.82 instead of 1, illustrating that our active learning approach does not overfit to the small number of image pairs. Our findings suggest that active learning strategies can successfully improve model performance in ICD, particularly when dealing with limited labeled data.

5. Conclusion

This paper explores the process of training a Vision Transformer (ViT) and presents an innovative active learning approach for Image Copy Detection (ICD). In order to train a ViT, we initially employ a Convolutional Neural Network (CNN) as the foundation model, and subsequently introduce a regularized learning technique to constrain the feature distribution of original images. This method enables the successful training of the ViT, leading to improved performance compared to the CNN. By generating and annotating a limited number of image pairs that exhibit visual discrepancies compared to auto-generated ones, we propose an active learning strategy for ICD. Our experimental findings highlight the substantial performance enhancement achieved through this active learning approach. Additionally, although our method achieves favorable performance, it still cannot identify any edit copies, *i.e.*, some missing is expected. We plan to design a more efficient and effective method to deal with new transformations.

CRediT authorship contribution statement

Zhentao Tan: Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Wenhao Wang:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Caifeng Shan:** Resources, Project administration, Methodology, Formal analysis, Data curation, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Sergio Manuel Papadakis, Sanjay Addicam, Producing augmentation-invariant embeddings from real-life imagery, 2021, arXiv preprint arXiv:2112.03415.

[2] Wenhao Wang, Yifan Sun, Yi Yang, A benchmark and asymmetrical-similarity learning for practical image copy detection, in: AAAI Conference on Artificial Intelligence, 2023.

[3] Wenhao Wang, Weipu Zhang, Yifan Sun, Yi Yang, Bag of tricks and a strong baseline for image copy detection, 2021, arXiv preprint arXiv:2111.08004.

[4] Wenhao Wang, Yifan Sun, Weipu Zhang, Yi Yang, Dˆ 2LV: A data-driven and local-verification approach for image copy detection, 2021, arXiv preprint arXiv:2111.07090.

[5] Shuhei Yokoo, Contrastive learning with large memory bank and negative embedding subtraction for accurate copy detection, 2021, arXiv preprint arXiv:2112.04323.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR (2021).

[7] JuanPotato, Legofy, 2023, https://github.com/JuanPotato/Legofy.

[8] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, Matthijs Douze, A self-supervised descriptor for image copy detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14532–14542.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[10] Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, Matthijs Douze, Multigrain: a unified image embedding for classes and instances, 2019, arXiv preprint arXiv:1902.05509.

[11] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[12] Wenhao Wang, Yifan Sun, Wei Li, Yi Yang, Transhp: Image classification with hierarchical prompting, Adv. Neural Inf. Process. Syst. 36 (2024).

[13] Robin Strudel, Ricardo Garcia, Ivan Laptev, Cordelia Schmid, Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262–7272.

[14] Zhengdong Hu, Yifan Sun, Yi Yang, Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation, in: The Eleventh International Conference on Learning Representations, 2023.

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.

[16] Zhengdong Hu, Yifan Sun, Jingdong Wang, Yi Yang, DAC-DETR: Divide the attention layers and conquer, in: Thirty-Seventh Conference on Neural Information Processing Systems, 2023.

[17] Chuyang Zhao, Yifan Sun, Wenhao Wang, Qiang Chen, Errui Ding, Yi Yang, Jingdong Wang, MS-DETR: Efficient DETR training with mixed supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[18] Zongxin Yang, Yunchao Wei, Yi Yang, Associating objects with transformers for video object segmentation, in: Advances in Neural Information Processing Systems (NeurIPS), 2021.

[19] Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, Yi Yang, Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision, IEEE Trans. Multimed. (2022).

[20] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, Wei Jiang, TransReID: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 15013–15022.

[21] Yann LeCun, Yoshua Bengio, et al., Convolutional networks for images, speech, and time series, Handb. Brain Theory Neural Netw. 3361 (10) (1995) 1995.

[22] Soumya Roy, Asim Unmesh, Vinay P. Namboodiri, Deep active learning for object detection, in: BMVC, Vol. 362, 2018, p. 91.

[23] Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, Sethuraman Panchanathan, Deep active learning for image classification, in: 2017 IEEE International Conference on Image Processing, ICIP, IEEE, 2017, pp. 3934–3938.

[24] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, Alexander G Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, Int. J. Comput. Vis. 113 (2015) 113–127.

[25] Wenhe Liu, Xiaojun Chang, Ling Chen, Dinh Phung, Xiaoqin Zhang, Yi Yang, Alexander G Hauptmann, Pair-based uncertainty and diversity promoting early active learning for person re-identification, ACM Trans. Intell. Syst. Technol. 11 (2) (2020) 1–15.

[26] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, Xin Wang, A survey of deep active learning, ACM Comput. Surv. (CSUR) 54 (9) (2021) 1–40.

[27] Yan Yan, Feiping Nie, Wen Li, Chenqiang Gao, Yi Yang, Dong Xu, Image classification by cross-media active learning with privileged information, IEEE Trans. Multimed. 18 (12) (2016) 2494–2502.

[28] Wenhe Liu, Xiaojun Chang, Ling Chen, Yi Yang, Early active learning with pairwise constraint for person re-identification, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10, Springer, 2017, pp. 103–118.

[29] Kayo Matsushita, Kayo Matsushita, Hasebe, Deep Active Learning, Springer, 2018.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[31] Alexander Hermans, Lucas Beyer, Bastian Leibe, In defense of the triplet loss for person re-identification, 2017, arXiv preprint arXiv:1703.07737.

[32] Kihyuk Sohn, Improved deep metric learning with multi-class n-pair loss objective, Adv. Neural Inf. Process. Syst. 29 (2016).

[33] Weiyang Liu, Yandong Wen, Zhiding Yu, Meng Yang, Large-margin softmax loss for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2016, pp. 507–516.

[34] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, Yichen Wei, Circle loss: A unified perspective of pair similarity optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6398–6407.

[35] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, Wei Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.

[36] Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoë Papakipos, Lowik Chanussot, Filip Radenovic, Tomas Jenicek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, et al., The 2021 image similarity dataset and challenge, 2021, arXiv preprint arXiv:2106.09672.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Herve Jegou, Training data-efficient image transformers and distillation through attention, in: International Conference on Machine Learning, Vol. 139, 2021, pp. 10347–10357.

[40] Aude Oliva, Antonio Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.

[41] Giorgos Tolias, Tomas Jenicek, Ondřej Chum, Learning and aggregating deep local descriptors for instance-level recognition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 460–477.

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[43] Ed Pizzi, Giorgos Kordopatis-Zilos, Hiral Patel, Gheorghe Postelnicu, Sugosh Nagavara Ravindra, Akshay Gupta, Symeon Papadopoulos, Giorgos Tolias, Matthijs Douze, The 2023 video similarity dataset and challenge, 2023, arXiv preprint arXiv:2306.09489.

[44] Wenhao Wang, Yifan Sun, Yi Yang, Feature-compatible progressive learning for video copy detection, 2023, arXiv preprint arXiv:2304.10305.

Zhentao Tan is currently a Master student in Academy for Advanced Interdisciplinary Studies, Center for Big Data Research, Peking University. His research interest is in computer vision, combinatorial optimization, and machine learning.

Wenhao Wang is a Ph.D. student in ReLER, AAII, University of Technology Sydney, supervised by Yi Yang. His research interests are visual copy detection, deep metric learning, and computer vision.

Caifeng Shan received the B.Eng. degree from the University of Science and Technology of China, the M.Eng. degree from the Institute of Automaton, Chinese Academy of Sciences, and the Ph.D. degree from Queen Mary, University of London. His research interests include computer vision, pattern recognition, medical image analysis, and related applications. He has co-authored more than 150 papers and 80 patent applications. He has served as Associate Editor for journals including IEEE Journal of Biomedical and Health Informatics and IEEE Transactions on Circuits and Systems for Video Technology. He is a Senior Member of IEEE.