Enhanced Duplicate Image Detection via Multi-Task Learning and Contrastive Features

Jin-Hyuk Song
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
song020@etri.re.kr

Yongseong Cho
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
yscho73@etri.re.kr

***Abstract*—Detecting duplicate images in large-scale databases presents unique challenges compared to video-based detection, primarily due to the limited contextual information within static images. To overcome this, we propose a novel multi-task learning (MTL)-based feature extraction model designed to enhance both the efficiency and accuracy of duplicate image detection. Our model utilizes a MobileNetV2 backbone augmented with two task-specific branches: a low-dimensional embedding branch for efficient similarity retrieval and a classification branch that leverages content-type cues to reduce false positives. Furthermore, we incorporate a contrastive learning strategy with cross-batch negative samples to significantly boost the discriminative power of the learned features. Experimental results confirm that our proposed architecture yields compact and semantically rich representations, evidenced by improved clustering in t-SNE visualizations and stable convergence during training. This approach not only reduces GPU memory requirements but also enhances robustness to various image transformations.**

***Index Terms*—Duplicate image detection, Multi-task learning (MTL), Feature embedding, Contrastive learning, Image retrieval.**

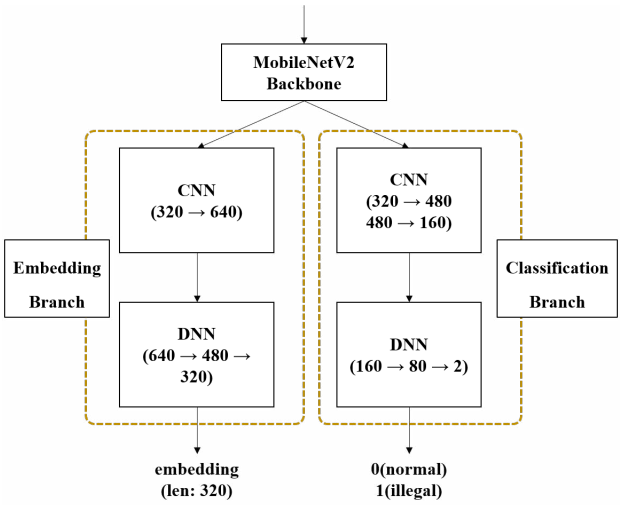

Fig. 1. Overall architecture of the proposed multi-task feature extractor based on MobileNetV2.

I. Introduction

Detecting duplicate content in image databases remains a challenging task, particularly when compared to video-based duplication detection [1]. Unlike video data, where temporal consistency and motion patterns provide additional cues, static images offer significantly less information for content characterization. This limitation necessitates the design of more sophisticated and diverse feature extraction strategies to ensure accurate duplicate detection [2]. In scenarios where the target of the search is explicitly defined — such as forensic applications or illicit content filtering — the ability to classify the type of target content can significantly enhance detection performance. Incorporating such classification information into the feature representation pipeline enables the system to distinguish between relevant and irrelevant matches more effectively. Conventional convolutional neural network (CNN) models, such as MobileNetV2, have been widely adopted for image embedding extraction due to their balance between performance and computational cost [3], [4]. However, as the number of reference images increases, the 1,280-dimensional output embedding generated by MobileNetV2 imposes significant demands on GPU memory, limiting the scalability of large-scale similarity search systems [5]–[8].

To address these challenges, we propose a novel feature extraction technique based on a multi-task learning (MTL) framework for duplicate image detection. The proposed method redesigns the original MobileNetV2 architecture to reduce embedding dimensionality while introducing an auxiliary classification branch. This approach aims to simultaneously extract compact yet discriminative embedding vectors and content-type classification information, which can contribute to reducing false positives during the retrieval phase. In this paper, we describe the design of our modified feature extraction model, explain the learning strategy including data and loss configuration, and evaluate its performance through experiments. In conclusion, we discuss how the combination of embedding and classification features can further be leveraged for enhanced duplicate image detection in future work.

II. Proposed Feature Extraction Method

In this work, we propose a redesigned feature extraction architecture based on the MobileNetV2 backbone to improve duplicate image detection performance while reducing com-

putational costs. The primary goal is to generate compact yet discriminative image embeddings and introduce semantic classification information to assist in reducing false detections. The proposed architecture, illustrated in Fig. 1, maintains the original MobileNetV2 backbone, which serves as the base feature extractor. Upon this backbone, two task-specific branches are added: the embedding branch and the classification branch.

A. Embedding Branch

The embedding branch is designed to produce compact 320-dimensional feature vectors suitable for similarity-based image retrieval. It first applies a convolutional block to expand the intermediate feature dimensionality from 320 to 640, followed by a deep neural network (DNN) that gradually reduces the vector size to 320 through intermediate layers of 480 dimensions. This compression aims to lower GPU memory usage during large-scale retrieval while preserving essential discriminative information.

B. Classification Branch

In parallel, the classification branch is introduced to enhance detection robustness by providing semantic information about the image content. This branch includes two convolutional layers (320→480, 480→160) followed by a DNN that reduces the representation to a final 2-class output. This auxiliary classification task encourages the model to learn content-aware features, which can help suppress false positives during duplicate detection by differentiating between visually similar but semantically distinct images. Through this multi-task learning structure, the network is trained to optimize both embedding similarity and content classification simultaneously, leading to more robust and accurate duplicate image detection.

C. Training Strategy

To enhance the discriminative power of the embedding space, we adopt a contrastive learning approach that leverages both within-batch and cross-batch negative samples during training. Unlike conventional contrastive losses that utilize only a single negative pair per anchor, our method expands the comparison scope to include all negative samples within the batch as well as representative negatives from other batches. This strategy enables the model to learn more robust and generalizable features by exposing it to a wider variety of dissimilar examples. The conceptual illustration of this training mechanism is depicted in Fig. 2. In the conventional method (top), the anchor (dark blue) is compared only against its corresponding positive (light blue) and a limited set of negative samples (red). In contrast, our proposed approach (bottom) expands the comparison to multiple negatives from both current and past batches, including green and yellow samples, thereby improving the separation between different classes in the feature space. This improved training strategy contributes to maximizing inter-class variance and minimizing intra-class variance, which is critical for accurate duplicate image retrieval in large-scale systems.

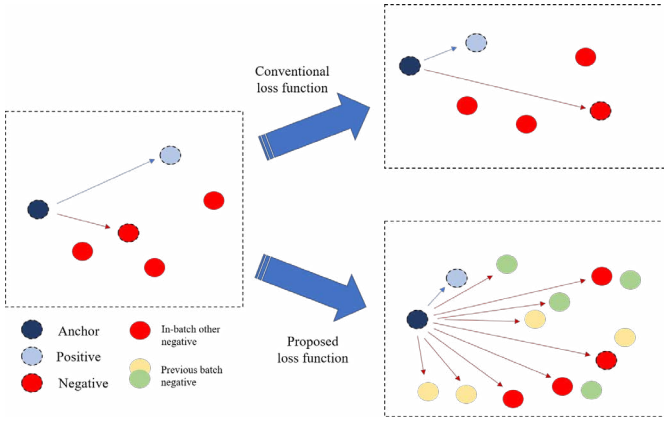

Fig. 2. Illustration of the contrastive learning strategy. Top: traditional within-batch negative sampling. Bottom: proposed strategy incorporating cross-batch negative samples to enhance contrastive signal.

III. Simulation Parameters and Results

A. Training Configuration

The proposed multi-task feature extraction model was trained in two stages using a freezing strategy. In the first stage, the MobileNetV2 backbone was frozen, and only the task-specific branches were trained to stabilize early learning. In the second stage, the entire network was fine-tuned jointly to improve embedding quality and classification accuracy. Table I summarizes the key hyperparameters used during training. The batch size was set to 64, with a learning rate of 4×10^{-6}. The margin value for the contrastive loss was configured as 0.45, and the number of negative samples used per batch reached up to 10,240, including cross-batch negatives as described in Section II.

TABLE I
Hyperparameter Configuration

Parameter	***Value***
Batch size	64
Learning rate	0.000004
Margin (α)	0.45
Max. number of negatives	10,240

B. Training and Inference Results

Figure 3 and Figure 4 show the training and validation loss curves for Stage 1 and Stage 2, respectively. In Stage 1 (frozen backbone), rapid convergence was observed in the early epochs, with both training and validation loss decreasing steadily. In Stage 2 (full model fine-tuning), further improvements were achieved, and the gap between training and validation loss was significantly reduced, indicating stable generalization.

To evaluate the effectiveness of the learned embeddings, we applied t-distributed stochastic neighbor embedding (T-SNE) to visualize the distribution of high-dimensional feature vectors in a 2D space. Figure 5 shows the T-SNE plot before training and after training with the proposed multi-task model.

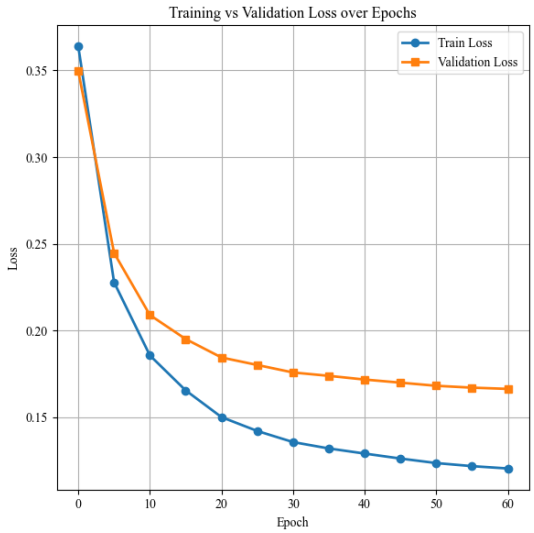

Fig. 3. Embedding Branch: Training and validation loss curves with frozen backbone.

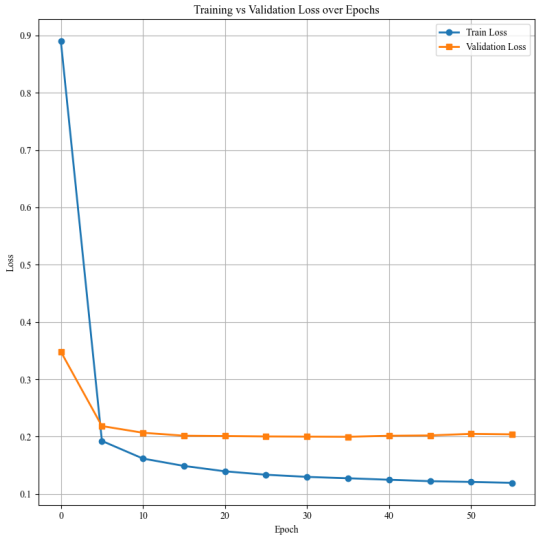

Fig. 4. Classification Branch: Training and validation loss curves with full model fine-tuning.

From Figure 5(a), we observe that the embeddings corresponding to different image classes are highly entangled with little to no discernible clustering. This indicates that the feature representations extracted by the untrained model do not sufficiently capture the semantic distinctions between image classes. After training, as shown in Figure 5(b), the embeddings exhibit a clear clustering pattern according to class labels. Notably, images derived from the same anchor—including transformed or augmented variants—are tightly grouped around their original anchor embeddings. This indicates that the learned embedding space successfully preserves semantic similarity and improves the robustness of duplicate image detection under various transformations.

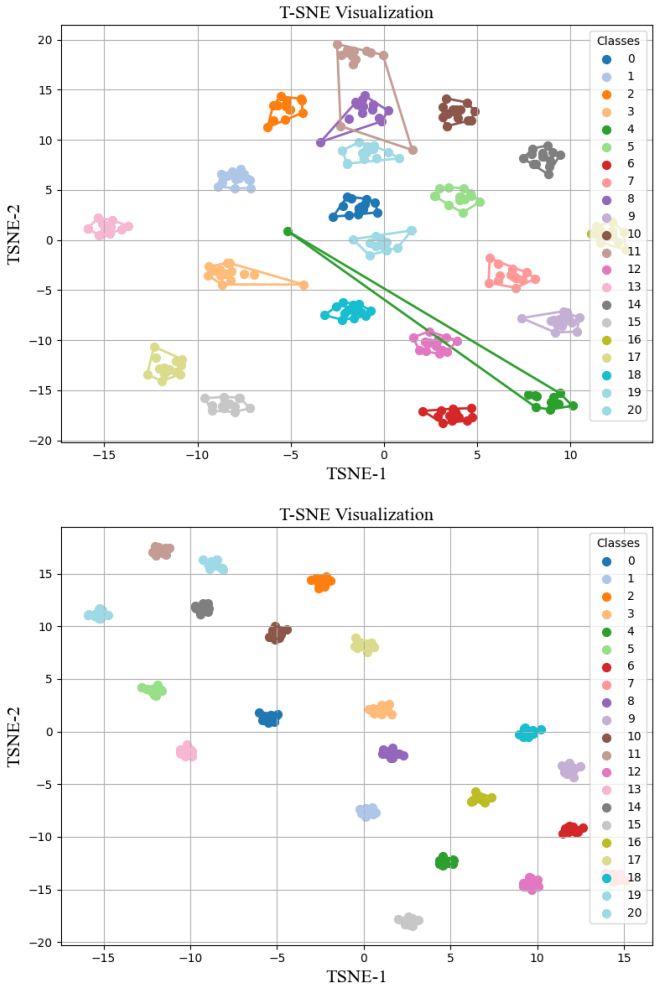

Fig. 5. T-SNE visualizations of embedding space: (a) before training, (b) after training.

These results demonstrate the effectiveness of the proposed model in producing discriminative and compact features that are well-suited for large-scale image retrieval tasks.

IV. Conclusions

In this paper, we proposed a multi-task learning-based feature extraction technique for duplicate image detection. The architecture was designed to simultaneously generate low-dimensional embeddings for similarity search and classification outputs for semantic discrimination. By extending the conventional MobileNetV2 backbone with two task-specific branches, our approach achieves both compactness and robustness in the learned representations. The proposed method offers several advantages. First, the embedding dimensionality is significantly reduced from 1,280 to 320, resulting in lower memory consumption and improved scalability in large-scale image retrieval systems. Second, the inclusion of a classification branch enhances the system's ability to suppress false positives by providing additional semantic cues. Finally, the

use of an expanded contrastive loss with cross-batch negative mining contributes to more discriminative feature learning.

As future work, we plan to develop a unified framework that jointly leverages both the embedding and classification outputs to improve duplicate detection accuracy. By integrating these complementary forms of information, we aim to enhance the system's robustness against various image transformations and improve precision in real-world retrieval scenarios.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00224740, Development of technology to prevent and track the distribution of illegally filmed content).

REFERENCES

[1] J. Zhu, S. Li, and Y. Wang, "A Review on Near-Duplicate Detection of Images using Computer Vision Techniques," *Multimedia Tools and Applications*, vol. 80, no. 10, pp. 15359–15391, 2021.

[2] S. K. Singh and S. K. Singh, "Near Duplicate Detection of Images with Area and Proposed Pixel-Based Feature Extraction," *Signal, Image and Video Processing*, vol. 14, no. 3, pp. 617–624, 2020.

[3] H. Y. Lee, J. H. Kim, and S. J. Lee, "Replication Image Detection Using Convolutional Neural Network," *IEEE Access*, vol. 8, pp. 112345–112355, 2020.

[4] X. Chen, S. Xie, and K. He, "Multi-Task Self-Training for Learning General Representations (MuST)," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 11567–11576, 2022.

[5] J. Acuna, J. Lin, and C. Hoi, "Detecting Duplication of Scientific Images with Manipulation-Invariant Image Similarity," *Nature Communications*, vol. 13, no. 1, pp. 1–10, 2022.

[6] A. Cho, W.-K. Yang, D.-S. Jeong, and W.-G. Oh, "Concentric circle-based image signature for near-duplicate detection in large databases," *ETRI J.*, vol. 32, no. 6, pp. 871–880, Dec. 2010.

[7] Y. M. Latha and B. S. Rao, "Amazon product recommendation system based on a modified convolutional neural network," *ETRI J.*, vol. 46, no. 4, pp. 633–647, Aug. 2024, doi: 10.4218/etrij.2023-0162.

[8] J. Song and Y. Cho, "Design of a duplicate image detection method based on multiple features," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Jeju, Korea, Oct. 2024, pp. 383–385, doi: 10.1109/ICTC62082.2024.10827639.