Few-Shot Logo Recognition

Oggero Paolo
s342937
address:
TODO: insert email address

Cancemi Alessia
s347156
address
TODO: insert email address

Mincigrucci Gabriele
s358987
address
TODO: insert email address

Abstract

Logo recognition in open-world scenarios is hampered by the long-tailed nature of the data. This work presents a Few-Shot Learning pipeline based on LogoDet-3K and Deep Metric Learning. Using a ResNet-50 optimized with Triplet and Contrastive Loss, we map logos into a 128-dimensional embedding space. Through a progressive freezing strategy, the model learns generalized representations that allow the retrieval of brands never seen during training. The results demonstrate that optimizing the latent space geometry significantly improves the mAP and F1-score compared to standard classification methods.

1. Introduction

Accurate logo recognition has become a fundamental pillar of media analysis and copyright protection. In uncontrolled environments, however, computer vision systems must contend with market dynamics: new brands emerge every day with unique visual identities that must be instantly identified. Traditional classification paradigms suffer from two structural limitations: the need for enormous datasets for each class and the inability to recognize categories not included in the training set. This phenomenon, known as long-tailed distribution, makes models rigid and expensive to update. Few-Shot Learning emerges as a necessary solution, shifting the focus from ”mnemonic recognition” to ”morphological comparison.” In this work, we address this challenge by implementing a Deep Metric Learning system. Instead of training the model to assign a label, we train it to ”measure” similarity. Using the LogoDet-3K dataset, we developed a pipeline that extracts deep features using a ResNet-50 and projects them into a metric space where the Euclidean distance reflects the semantic relatedness of the logos. This approach not only better manages data sparsity, but also allows the system to operate on unseen classes without the need for retraining. The approach presented here focuses on overcoming the limitations of traditional classification through various technical solutions.

Figure 1. Sample of images from LogoDet-3K

The work first introduces an embedding-based architecture, developed using a linear projection head that enables a standard CNN to extract metric features. This structure allows for an in-depth comparative analysis of losses, evaluating how Triplet and Contrastive Loss (in the Euclidean and Cosine variants) influence the topology of the latent space. To support training, a dynamic transfer learning management system based on progressive layer unlocking is implemented, ensuring an optimal balance between model stability and performance. The work concludes with an Open-Set Evaluation conducted on brands never encountered during the training phase, using retrieval metrics such as mAP to validate the system’s actual generalization capability.

2. Data

For this project, we used the LogoDet-3K dataset[1], which currently represents one of the largest and most complex benchmarks for logo recognition.

2.1. Dataset Description

LogoDet-3K comprises 3,000 logo categories, with approximately 200,000 manually annotated objects distributed across 158,652 images. The dataset is hierarchically organized into nine supercategories (Food, Clothes, Necessities, Others, Electronics, Transportation, Leisure, Sports, and Medical). The dataset is inherently long-tailed, we can see that in figure 2, some classes have a very large number of samples, while others (especially in the Medical

or Sports categories) have very few instances. This distribution faithfully reflects the challenges of real-world scenarios.

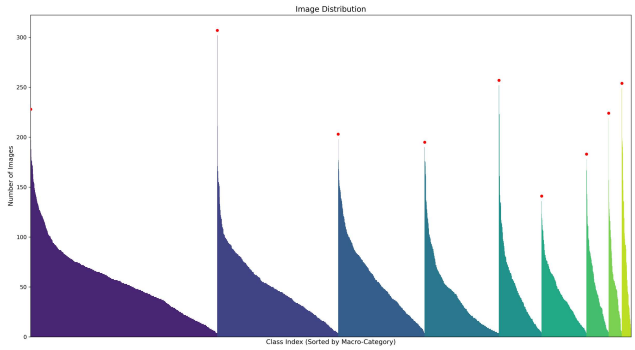

Figure 2. Distribution of the LogoDet-3K dataset by macro-category

2.2. Splitting of the Dataset (Brand-Level Split)

Unlike standard classification pipelines, where images are randomly shuffled and split, our code implements the `getPathsSetsByBrand` function. This logic applies splitting at the class (brand) level rather than at the individual image level:

- Training Set (70% of brands): The model learns the general morphological characteristics of logos.
- Validation Set (20% of brands): Used to monitor the F1 score and convergence on brands never seen during training.
- Test Set (10% of brands): Reserved for the final performance evaluation in Open-Set Retrieval mode.

2.3. Data Preparation and Sampling

To meet the requirements of the implemented loss functions, two specific Dataset classes have been created:

1. `DatasetTriplet`: For each reference image (Anchor), randomly select one example of the same brand (Positive) and one of a different brand (Negative).
2. `DatasetContrastive`: Generates image pairs with a 50% probability of belonging to the same brand (label 1) or different brands (label 0).

2.4. Preprocessing and Data Augmentation

The original images have heterogeneous resolutions. In the loading module, the data is normalized and transformed to improve the model's robustness:

- **Resize and Normalization:** All images are resized to 224×224 pixels and normalized using the ImageNet mean and standard deviation, ensuring compatibility with the pre-trained weights of ResNet-50.
- **Augmentation:** During training, transformations are applied, including `RandomResizedCrop`, `RandomHorizontalFlip`, and `ColorJitter`. These techniques simulate the distortions typical of real-world logos (light variations, angled shots, etc.).

3. Methods

This section describes the system architecture and the optimization methodologies adopted to transform the logo recognition problem into a Deep Metric Learning task.

3.1. Model Architecture: LogoResNet50

The implemented architecture is based on the ResNet-50 backbone, chosen for its optimal balance between computational depth and feature extraction capability. The original model, pre-trained on ImageNet, was modified to adapt to the metric learning paradigm through two structural interventions:

- **Classifier Replacement:** The final Fully Connected layer (originally designed for 1,000 classes) was removed and replaced with a Linear Embedding Head. This module projects the high-level features extracted by the network into a 128-dimensional latent space.
- **Space Normalization:** The output embeddings are treated as geometric coordinates; the network learns to place similar logos in nearby regions of the vector space.

3.2. Loss Functions

To optimize the topology of the embedding space, three different training objectives were implemented and compared:

- **Triplet Margin Loss**: This is the primary retrieval strategy. The function works on triplets (Anchor, Positive, Negative) and minimizes the distance between Anchor and Positive, while simultaneously maximizing the distance between Anchor and Negative with a configurable margin ($\alpha = 1.0$).
- **Contrastive Loss (Euclidean)**: Optimized for image pairs, it penalizes the distance between similar pairs and forces dissimilar pairs to a distance greater than the set margin.
- **Contrastive Loss (Cosine)**: A variant of the previous one that uses cosine similarity instead of Euclidean distance. This metric is particularly effective in high-dimensional spaces because it focuses on the orientation of the vectors rather than their magnitude.

3.3. Training and Transfer Learning Strategy

Training is managed by a dynamic configuration system (Config) that allows for reproducible experiments. The main phases include:

- **Progressive Freezing**: To preserve pre-learned knowledge about ImageNet, the model starts training with frozen backbone convolutional blocks (freeze_layers).
- **Dynamic Unfreezing**: Upon reaching a predetermined epoch (`unfreeze_at_epoch`$= 5$), the code unfreezes layers, allowing for fine-tuning of logo-specific spatial features.
- **Optimization**: The Adam optimizer with a conservative learning rate $(1e-5)$ is used to ensure stable convergence and minimize oscillations in the Loss.

3.4. Evaluation Metrics

The system is evaluated not simply on accuracy, but on its ranking ability. The main metrics calculated in the validation loop are:

- **F1-Score**: Dynamically calculated during training to monitor the balance between Precision and Recall.
- **Mean Average Precision (mAP)**: Used to measure retrieval quality, i.e., how effectively the model ranks correct logos at the top of search results.

References

[1] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–23, 2022.