CVPR
#

Fashion Retrieval using ResNet-50 and FAISS Indexing

Ayra, Arif Hakimi, Seung Hun Lee

2023320097, 2023320329, 2021320125

Abstract

Fashion image retrieval plays a central role in e-commerce applications, allowing users to find visually similar clothing items based on example inputs. This project focuses on developing a deep learning model that performs fashion image retrieval by learning compact and discriminative embeddings. Our current model utilizes a ResNet50 backbone trained with triplet loss to optimize for visual similarity in an embedding space. The output embeddings are indexed using FAISS to support efficient similarity search. Although classification functionality is planned as an extension, the current version is solely retrieval-focused. This work serves as the foundation for a future dual-purpose fashion recognition system capable of both retrieval and category prediction.

1. Introduction

Image–based retrieval is a cornerstone of modern fashion-AI applications: given a query photo, the system must surface visually similar garments from a candidate gallery. Unlike pure classification—which assigns a single category label—retrieval demands an embedding space whose geometry captures subtle cues such as silhouette, texture, and color.

Our baseline tackles this requirement by training a deep convolutional network (ResNet-50) with *triplet loss*, which explicitly pulls visually-similar items together while pushing dissimilar ones apart. The learned embeddings are indexed by FAISS to enable sub-second top-K nearest-neighbor search.

Because project time allowed, we also **added a lightweight classification head** on the same backbone. Although its current top-1 accuracy is only $\sim 57.5\%$, the predicted category lets us filter the FAISS search to a semantically-coherent subset, yielding cleaner retrieval results in many cases. Future iterations will focus on boosting the classifier and jointly fine-tuning both objectives.

2. Problem Identification

Convolutional neural networks like ResNet50 are commonly used for image classification, producing softmax probabilities over fixed labels. However, such models are not optimized for measuring visual similarity, which is essential in fashion image retrieval. Classification outputs do not reflect how similar two items look, especially when they belong to the same category but differ in style or appearance.

3. Technical Soundness

3.1. Baseline Model and its Limitations

As a baseline, we consider a standard image classification model based on ResNet50 trained with cross-entropy loss. This model is designed to assign a single clothing category (e.g., ”t-shirt”, ”dress”, ”pants”) to each input image. It represents a simpler approach compared to our retrieval-focused model, as it focuses solely on category prediction without learning an embedding space or supporting image similarity search.

This baseline fulfills the role of a minimal benchmark:

- It is **simpler** than our proposed model, with a single classification objective.
- It performs **only one task well**, categorical prediction but does not support retrieval or visual similarity matching.
- It uses **standard components**: a ResNet50 backbone and a fully connected classification head trained with cross-entropy loss, which is a widely accepted setup in image classification.

Despite its effectiveness in predicting class labels, this model has several limitations:

- It cannot retrieve visually similar clothing items, limiting its utility in recommendation or search systems.
- The output is a discrete label, which discards nuanced visual differences between garments within the same category.
- It does not support efficient similarity indexing or ranking, as there is no learned embedding space.

These limitations highlight the need for a more versatile model that can simultaneously classify and retrieve fashion items. Our current project addresses this by focusing on metric learning and embedding-based retrieval, with plans to later integrate classification for dual-task capability.

4. Experimental Methodology

This section describes the full technical pipeline used to develop and evaluate deep neural network-based image retrieval models for fashion products. Our experiments are grounded in four trained models that vary in terms of their loss functions, preprocessing methods, architectural components, and auxiliary objectives. Each model is designed to learn a 512-dimensional embedding suitable for visual similarity search. We evaluate their performance on the DeepFashion2 dataset using cosine similarity and a top-K retrieval protocol.

4.1. Dataset and Preprocessing

We use the DeepFashion2 dataset, a widely used benchmark for image-based fashion retrieval and recognition. It contains over 800,000 images with annotations including bounding boxes, clothing categories, landmark points, segmentation masks, and most critically, pair IDs that connect images of the same garment (e.g., shop and consumer photos). These pair IDs are used to define positive matches for training and evaluation.

To study the impact of background clutter and preprocessing, we construct two versions of the dataset:

- **Uncropped:** In the TripletMarginLoss model, the full image is used without any spatial filtering, which includes background, human pose, and lighting noise.
- **Cropped:** In the TripletMarginLossWithCropping, CrossEntropyLoss, and ContrastiveLossandReLU models, each image is cropped to the ground-truth bounding box before being passed to the model, encouraging the network to focus only on garment-specific visual features.

In all settings, the images are resized to 224×224, normalized using ImageNet channel statistics, and augmented with horizontal flipping, color jittering, and random affine transformations. These augmentations help improve generalization by simulating real-world variations in lighting and pose.

4.2. Model Architectures

Each retrieval model uses a ResNet-50 backbone pretrained on ImageNet. The original classification head is discarded and replaced by a new feature embedding head. This head consists of a global average pooling layer followed by a fully connected layer that maps the feature maps to a 512-dimensional vector. L2 normalization is applied to the output embeddings so that all vectors lie on the unit hypersphere, making cosine similarity a valid distance metric.

TripletMarginLoss. This model is trained with triplet margin loss and uses the full, uncropped images. It includes no classification head or other auxiliary tasks. The embeddings are directly optimized to satisfy the relative distance constraint among anchor, positive, and negative samples. Hard negative mining is applied within each batch to select the most informative negative examples for training.

TripletMarginLossWithCropping. This model applies the same architecture and loss function as TripletMarginLoss but uses images cropped to the garment bounding box. This modification reduces the variance introduced by irrelevant background information and helps the model focus more on the shape, texture, and structure of garments. It retains a pure metric learning objective with no auxiliary task.

CrossEntropyLoss. This model extends TripletMarginLossWithCropping by incorporating a secondary prediction branch. In addition to the 512-dimensional embedding output, a parallel classification head is attached to the penultimate ResNet feature layer. This head is trained to predict one of the 13 DeepFashion2 clothing categories using softmax and cross-entropy loss. The total loss is a weighted combination of triplet margin loss and classification loss:

$$\mathcal{L}_{total} = \mathcal{L}_{triplet} + \lambda \cdot \mathcal{L}_{classification} \qquad (146)$$

where λ is empirically set to 0.5. This dual-task learning scheme allows the model to learn category-discriminative features while preserving fine-grained similarity relationships between garments.

ContrastiveLossandReLU. This model differs from the others in two key aspects: it uses contrastive loss instead of triplet loss, and introduces a ReLU activation before the final embedding layer. The contrastive loss treats training samples as positive or negative pairs and directly penalizes their pairwise distance:

$$\mathcal{L}_{contrastive} = y \cdot D^2 + (1 - y) \cdot \max(0, m - D)^2$$

where y is 1 for positive pairs, 0 otherwise, and m is the margin (set to 1.0). The inclusion of ReLU restricts the embeddings to the positive orthant, reducing their representational freedom, which we later show can affect performance. This model receives cropped images and focuses purely on pairwise similarity learning without category supervision.

4.3. Training Setup

All models are trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-4}. We use a batch size of 1084 to maximize utilization of GPU memory and stabilize gradient estimates across large batches. No learning rate scheduling or early stopping is applied; each model is trained for a fixed number of epochs to maintain consistency in evaluation. All training runs are performed using a single GPU. Model checkpoints are generated during training, although intermediate checkpointing frequency is not a focus of this study.

4.4. Retrieval Pipeline

After training, we extract 512-dimensional embeddings from all query and gallery images. These embeddings are L2-normalized and indexed using FAISS with inner product similarity, which is equivalent to cosine similarity due to normalization. For each query image, we retrieve the top-K most similar gallery embeddings.

Following the DeepFashion2 evaluation setup, a query is considered successfully retrieved if at least one of the top-K retrieved gallery items shares the same pair ID and has an intersection-over-union (IoU) greater than 0.5 with the ground-truth bounding box of the query image. Where applicable, we report retrieval accuracy at Top-1 and Top-5 levels. In models where quantitative metrics are not reported, qualitative inspection of the top-5 retrieved results is used to assess embedding performance.

5. Results and Analysis

This section presents a quantitative and technical analysis of the four trained models under evaluation. Retrieval performance is assessed using top-K accuracy metrics on the DeepFashion2 dataset, where a retrieval is considered correct if any of the top-K retrieved gallery images share the same pair ID or have a low distance value. Each model differs in either its loss function, preprocessing pipeline, or auxiliary objectives. Retrieval performance is evaluated using Top-K accuracy where explicitly available. For models without quantitative accuracy output, we analyze qualitative retrieval results and failure cases based on visual inspection of the top-5 nearest neighbors.

5.1. TripletMarginLoss

This baseline model applies triplet margin loss without cropping. As such, full images are passed into the embedding model, which allows non-garment elements (e.g., background, lighting, pose, shadows) to influence training. Retrieval results show that the model frequently retrieves images with similar background settings rather than semantically similar garments. For example, in one case, a dress was recovered along with a shirt with similar lighting in the scene and wall color, but dissimilar structure. This indicates that the learned embeddings are heavily influenced by global image features and do not reliably capture garment-specific identity. As shown in the diagram below is one of the results of the TripletMarginLoss model.

Figure 1. TripletMarginLoss's result

5.2. TripletMarginLossWithCropping

This model uses the same triplet loss but includes a cropping step based on ground-truth bounding boxes during both training and inference. Qualitatively, the retrievals improve significantly. The top-5 matches for a query garment generally show strong coherence in silhouette, texture, and category. For instance, dresses are retrieved alongside other dresses of similar length and neckline, rather than being influenced by background context. These results confirm the importance of spatial preprocessing in fashion metric learning. Although no top-K accuracy metrics are printed, the visual consistency of the retrieved items supports the hypothesis that the cropping of the bounding box improves the focus of the embedding. The figure below shows the output result from TripletMarginLossWithCropping model.

Figure 2. TripletMarginLossWithCropping's result

5.3. CrossEntropyLoss

This variant extends by introducing an auxiliary classification head trained with cross-entropy loss to predict garment category. During inference, the model outputs both category predictions and retrieval results. This dual-headed structure allows the system to assign class labels to new, previously unseen fashion items (e.g., online test images). Retrievals from this model display strong category alignment: garments retrieved share not only visual features but also correct class identity. The added semantic regularization appears to constrain the embedding space more effectively. For example, a test image of a blouse is consistently matched with other blouses in the gallery. Despite the lack of numerical accuracy output, the observed retrievals are the most consistent and semantically aligned of all four models. Below is the output of combining the prediction implementation with retrieval function model.

Figure 3. CrossEntropyLoss's result

5.4. ContrastiveLossandReLU

This model uses contrastive loss instead of triplet loss and introduces a ReLU activation before the final embedding layer. Unlike triplet loss, which enforces relative distance constraints, contrastive loss focuses on absolute distances between pairs. The use of ReLU restricts the output embeddings to the nonnegative orthant, potentially limiting the model's representational power. The following is one of the results obtained from testing the model.

Figure 4. ContrastiveLossandReLU's result

Qualitative results show that retrievals are visually reasonable but less consistent than those from TripletMarginLossWithCropping or CrossEntropyLoss. Some top-5 results contain garments that differ in category or structure from the query, despite texture similarity. This suggests that the contrastive loss may be less effective at capturing fine-grained rank-sensitive relations compared to triplet-based methods.

5.5. Qualitative Insights

Across all models, we analyzed representative top-5 retrieval results for both validation and external test images. The uncropped TripletMarginLoss baseline shows the most inconsistent retrievals, with background influence evident in failure cases. Models trained with bounding box cropping consistently focus on relevant garment features, such as sleeve type, color, and cut. The CrossEntropyLoss model additionally retrieves items that belong to the same semantic category as the query, supporting the value of auxiliary classification supervision. In contrast, the ContrastiveLossandReLU model retrieves texture-similar but semantically inconsistent garments, reflecting the limitations of contrastive loss and ReLU constraint for fine-grained fashion retrieval.

6. Conclusion

This work investigates deep metric learning techniques for fashion image retrieval using the DeepFashion2 dataset. We designed and evaluated four model variants based on different combinations of loss functions, spatial preprocessing, and auxiliary supervision. Our findings emphasize the critical importance of training setup and architectural decisions in shaping retrieval behavior.

First, we demonstrate that preprocessing, specifically, the cropping of garments using ground truth bounding boxes, significantly improves the focus of the retrieval by eliminating distracting background features. Second, we show that triplet margin loss consistently outperforms contrastive loss in producing rank-sensitive embeddings suitable for top-K nearest neighbor search. Third, incorporating an auxiliary classification objective further enhances semantic alignment within the embedding space, enabling the model to generalize better to unseen examples and improving category consistency in retrievals.

Finally, we observe that architectural modifications such as restricting the embedding space via ReLU activation can negatively impact expressiveness and reduce retrieval accuracy. Our qualitative analysis across all models provides insight into how these design decisions interact and affect learned representations.

Future work will explore multi-task training strategies that jointly optimize for classification, retrieval, and attribute prediction, as well as domain adaptation methods to improve robustness to out-of-distribution fashion content.

DeepFashion2 [1] provides richer annotations such as landmarks, masks, and pair IDs, enabling fine-grained visual retrieval. Triplet loss [2], contrastive loss [3], and ProxyNCA [4] are core metric learning methods for embedding-based retrieval.

References

[1] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," *CVPR*, pp. 5337–5345, 2019. 4

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, pp. 815–823, 2015. 4

[3] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, pp. 1735–1742, 2006. 4

[4] Y. Movshovitz-Attias, A. Toshev, T. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *ICCV*, pp. 360–368, 2017. 4