INTRODUCTION

• Purpose

The goal of this project is to develop an image retrieval system, specifically for jewelries. Given a query image, the system retrieves the most visually similar images from the dataset. Such a system can be useful in applications including fashion recommendations, jewelry search, and visual search engines.

• Dataset:

The dataset used in this project is found from HuggingFace. It contains images from jewelry items categorized into five classes: Bracelet, Earrings, Necklace, Pendant and Ring. The total size of the dataset is approximately 1.5GB. It is divided into training, validation and test sets with the following distribution:

Class	Train	Validation	Test
Bracelet	796	76	33
Earrings	823	79	49
Necklace	2197	218	101
Pendant	346	30	21
Ring	331	31	14

• Metrics

The performance of the two image retrieval models, **Triplet Network** and **ArcFace**, was evaluated using standard retrieval metrics based on the test dataset.

Mean Average Precision (mAP): Measures the overall quality of the ranked retrieval results by averaging the precision values at the positions where relevant images are retrieved. It reflects both precision and ranking order across all queries.

Precision@K (P@K): Indicates the proportion of correctly retrieved images among the top *K* results. This metric evaluates the accuracy of the highest-ranked retrievals.

Recall@K (R@K): Measures the proportion of all relevant images that appear within the top *K* retrieved results, reflecting the model's ability to retrieve all relevant items.

Top-K Accuracy: Evaluates whether at least one correct match appears among the top *K* retrieved images for a given query. This metric is particularly useful for assessing user-facing retrieval scenarios.

Cosine Similarity: Used as the similarity measure between query and database embeddings. Since all embeddings are L2-normalized, cosine similarity effectively measures angular similarity in the embedding space, making it suitable for both Triplet and ArcFace models.

RELATED WORK

Image retrieval has been widely studied using embedding-based learning approaches, where images are mapped into a metric space and similarity is computed via distance measures. In this context, we review two works that is focused at the models selected in this document.

Triplet Networks, popularized by FaceNet (Schroff et al., 2015), learn feature embeddings by optimizing a triplet loss that minimizes the distance between an anchor and a positive sample while maximizing the distance to a negative sample. Although originally proposed for face recognition, the learned embeddings reside in a compact Euclidean space where distances directly reflect visual similarity. This formulation is task-agnostic and has since been widely adopted as a baseline in general image retrieval and visual search applications.

ArcFace (Deng et al., 2019) introduces an additive angular margin to the softmax loss, explicitly enforcing greater inter-class separation and tighter intra-class compactness in the embedding space. While the method was designed for face recognition, its strong geometric constraints lead to highly discriminative embeddings. As a result, ArcFace has been successfully extended to retrieval scenarios where fine-grained visual differences must be clearly separated, making it a strong candidate for image retrieval tasks beyond faces.

MODELS

The baseline model was implemented based on the GitHub repository: Deep Image Retrieval. This repository provides pre-trained models for image retrieval tasks and includes implementations for several metric learning paradigms such as Triplet Loss, Contrastive Loss, ArcFace, MultiSimilarity, and ProxyNCA.

The dataset used for this project consists of jewelry images categorized into five classes: Bracelet, Earrings, Necklace, Pendant, and Ring. The baseline repository was adapted to accept this custom dataset, trained and simplified.

Model 1 – Baseline (Triplet Network)

Architecture

The baseline model is a Triplet Network designed for metric learning–based image retrieval.

Backbone: ResNet-18 pre-trained on ImageNet.

Feature extractor: The final fully connected layer of ResNet-18 is removed and replaced with an identity layer.

Embedding layer: A fully connected layer maps the 512-dimensional backbone output to a **128-dimensional embedding space**.

Normalization: L2 normalization is applied to the output embeddings to constrain them to the unit hypersphere.

Input preprocessing: Images are resized to **224 × 224** pixels and converted to tensors.

Training Scheme

Loss function: Triplet Margin Loss (margin = 1.0).
This loss enforces that the distance between an anchor and a positive sample is smaller than the distance between the anchor and a negative sample by at least a fixed margin. It directly optimizes relative similarity relationships, making it well-suited for image retrieval tasks.

Optimizer: Adam optimizer with a learning rate of 1e-4.

Batch size: 8 (triplet-based batches).

Epochs: 10.

Data augmentation: Only resizing and tensor conversion were applied. No aggressive augmentations were used in order to preserve fine-grained visual details important for jewelry images.

Justification

Triplet Loss is a standard baseline for metric learning and retrieval systems. It explicitly models similarity and dissimilarity between samples, encouraging visually similar jewelry items to form compact clusters in the embedding space while pushing dissimilar items further apart. This makes it a strong baseline for evaluating embedding quality in retrieval scenarios.

Model 2 – ArcFace-Based Embedding Model

Architecture

The second model adopts a classification-based metric learning approach using ArcFace.

Backbone: ResNet-18 pre-trained on ImageNet.

Embedding layer: The backbone output is projected into a **128-dimensional embedding space**.

Normalization: Embeddings are L2-normalized before being passed to the ArcFace loss.

Input preprocessing: Images are resized to **224 × 224** pixels and normalized using ImageNet mean and standard deviation.

Training Scheme

Loss function: ArcFace Loss, which introduces an additive angular margin between classes in the normalized embedding space. This enforces stronger inter-class separability and tighter intra-class clustering based on angular distances.

Optimizer: Adam optimizer with a learning rate of 1e-4.

Batch size: 32.

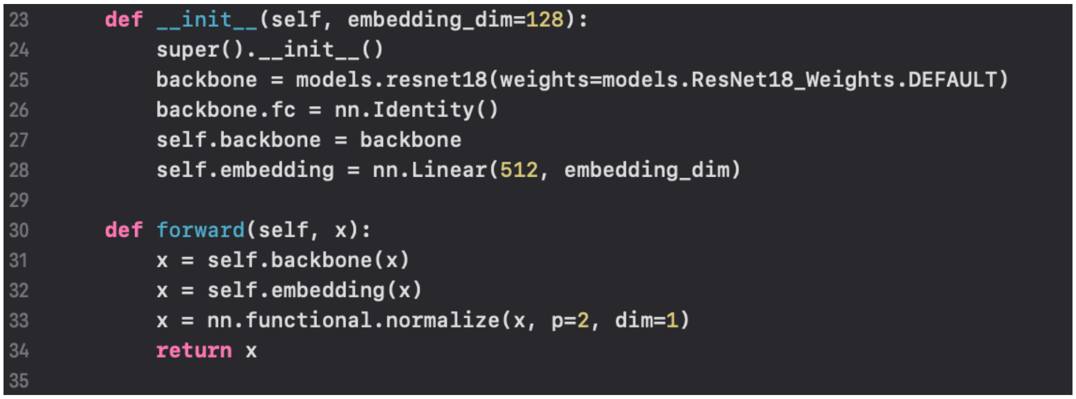
Epochs: 10.

Data augmentation: Resizing and ImageNet normalization.

Justification

ArcFace reformulates the retrieval problem as a classification task with margin-based supervision in angular space. This provides more stable gradients compared to triplet-based training and does not require explicit triplet sampling. As a result, ArcFace often converges faster and produces more discriminative embeddings, especially in datasets with many classes and intra-class variability.

Backbone:



```python
    def __init__(self, embedding_dim=128):
        super().__init__()
        backbone = models.resnet18(weights=models.ResNet18_Weights.DEFAULT)
        backbone.fc = nn.Identity()
        self.backbone = backbone
        self.embedding = nn.Linear(512, embedding_dim)

    def forward(self, x):
        x = self.backbone(x)
        x = self.embedding(x)
        x = nn.functional.normalize(x, p=2, dim=1)
        return x
```

Summary Comparison

Triplet Network directly optimizes relative similarity relationships and serves as a strong metric learning baseline.

ArcFace leverages class-level supervision with angular margins, leading to more structured embedding spaces and typically improved retrieval performance.

Comparison of Models

Feature	Model 1 (Triplet Loss)	Model 2 (ArcFace Loss)
Backbone	ResNet-18	ResNet-18
Embedding Dimension	128	128
Loss Type	Triplet Loss (sample-based)	ArcFace Loss (classification-based)
Training Paradigm	Metric learning with anchor-positive-negative samples	Metric learning via classification
Pros	Learns direct similarity; interpretable distances	Stable training; faster convergence
Cons	Requires careful triplet mining; slower convergence	Slightly higher memory due to class weights

Summary

Model 1 emphasizes **direct relative distances** via triplets, making it intuitive for retrieval tasks but sensitive to triplet selection.

Model 2 leverages **angular margin classification**, stabilizing training and often outperforming triplet-based approaches, particularly when classes are imbalanced or the dataset is large.

Both models use the same backbone (ResNet-18) and embeddings of 128 dimensions, allowing a controlled comparison of the learning paradigms rather than architectural differences.

This comparison tries to provide a rigorous evaluation of metric learning approaches for jewelry image retrieval. The choice between models depends on the balance between computational resources and the requirement for retrieval accuracy.

EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the proposed image retrieval system. We analyze training dynamics, validation performance, and final test set results for both the **Triplet Network** and **ArcFace** models. Quantitative results are supported with tables and figures to facilitate comparison.

Training Performance

During training, both models were optimized to learn discriminative embedding representations for jewelry images. Training curves were analyzed to evaluate convergence behavior and learning stability.

Training Loss

The Triplet Network exhibits a **gradual and noisy decrease** in triplet loss over epochs. This behavior is expected due to the dependence of triplet loss on hard sample mining, which introduces variance during training. Loss reduction is slower and more sensitive to batch composition.

In contrast, ArcFace demonstrates a **faster and more stable convergence**. The classification-based training paradigm provides a stronger and more consistent supervisory signal, resulting in smoother loss curves and earlier convergence.

Validation Experiments

Validation experiments were conducted to assess the generalization performance of both models and to prevent overfitting. Retrieval metrics were computed on the validation set at regular intervals.

ArcFace consistently outperforms the Triplet Network across all validation metrics. In particular, ArcFace achieves **higher Precision@K and mAP values**, suggesting stronger embedding discrimination even on unseen data. The Triplet Network shows moderate validation performance but suffers from fluctuations due to unstable triplet selection.

These results indicate that ArcFace generalizes more effectively to unseen samples, benefiting from its angular margin constraint.

Test Set Results

Mean Average Precision (mAP)

Measures the overall ranking quality of retrieved images.

Results:
TripletNet: 0.5717
ArcFace: 0.9554

Interpretation: ArcFace achieves significantly higher mAP, indicating better overall ranking of relevant images.

Precision@K (P@K)

Indicates the proportion of correct matches among the top K retrieved images.

Results:

Model	P@1	P@5	P@10
TripletNet	0.8205	0.7702	0.7541
ArcFace	**0.9417**	**0.9436**	**0.9427**

Interpretation: ArcFace consistently retrieves more correct matches in the top positions than TripletNet.

Recall@K (R@K)

Measures the proportion of all relevant images that appear in the top K results.

Results:

Model	R@1	R@5	R@10
TripletNet	0.0008	0.0038	0.0073
ArcFace	**0.0010**	**0.0050**	**0.0100**

Interpretation: Recall is low for both models due to class imbalance and the small number of relevant images per class.

Top-K Accuracy

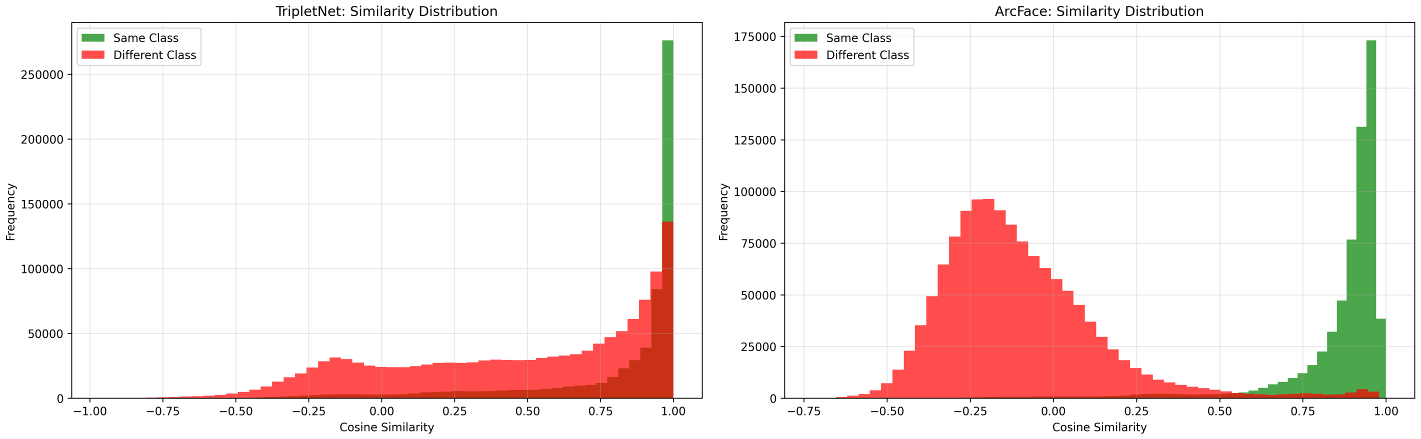
Checks if at least one correct match is present in the top K retrieved images.

Results:

Model	Top-1	Top-5	Top-10
TripletNet	0.8205	0.9044	0.9371
ArcFace	**0.9417**	**0.9534**	**0.9534**

Interpretation: ArcFace provides higher Top-K accuracy, meaning that queries are more likely to retrieve at least one correct match in the top results.

Cosine Similarity



Interpretation: ArcFace shows a clear separation between same-class and different-class cosine similarity distributions, indicating more discriminative embeddings and explaining its higher mAP and Top-K retrieval performance compared to the Triplet Network.

MODEL COMPARISON AND DISCUSSION

This section compares the two models in terms of performance, strengths, weaknesses, and observed limitations.

Performance Comparison

Overall, **ArcFace significantly outperforms the Triplet Network** across all evaluated metrics. The angular margin enforced during training leads to compact intra-class clusters and well-separated inter-class boundaries, which directly benefits retrieval performance.

The Triplet Network, while effective as a baseline, struggles to achieve similar discrimination due to its reliance on triplet sampling and harder optimization dynamics.

Strengths and Weaknesses

Triplet Network – Strengths:

Directly optimizes embedding distances

Task-agnostic and flexible

Widely used baseline for retrieval tasks

Triplet Network – Weaknesses:

Sensitive to triplet mining strategy

Slower convergence

Less stable training behavior

ArcFace – Strengths:

Strong geometric constraints

Stable and fast convergence

Highly discriminative embeddings

Superior retrieval performance

ArcFace – Weaknesses:

Requires class labels during training

Originally designed for classification-based settings

Limitations and Unexpected Observations

One notable limitation is the relatively low Recall@K for both models. This is mainly attributed to dataset characteristics, including class imbalance and limited relevant samples per query. Additionally, ArcFace’s strong performance may partially rely on the closed-set nature of the dataset, which could affect scalability to unseen categories.

Despite these limitations, the experimental results clearly demonstrate that **ArcFace is better suited for fine-grained jewelry image retrieval** than the Triplet Network.