Explaining Embedding-Based Matching of Hand-Drawn Binary Symbols with Grad-CAM: A Case Study on Cattle Brands

Leandra Alves Soares
Institute of Physics, UFG
Goiânia, Goiás, Brazil
leandra.physicsengineering@gmail.com

Marcos Vinícius S. Medeiros
Institute of Informatics, UFG
Goiânia, Goiás, Brazil
marcosvsatil@gmail.com

Aldo A. Díaz-Salazar
Institute of Informatics, UFG
Goiânia, Goiás, Brazil
aldo.diaz@ufg.br

ABSTRACT

Cattle-brand identification poses unique challenges due to the sparse, high-contrast nature of symbolic marks and the large stylistic gap between clean reference templates and hand-drawn queries. Since a prediction layer cannot be applied reliably in this setting, weperform identification through feature extraction and similarity search using a VGG-16 encoder and a FAISS-based retrieval system. weinvestigate how the encoder represents these binary symbols by combining channel-level activation statistics, Grad-CAM saliency maps, and Top-10 retrieval results. Our analysis shows that a small group of filters, especially an outer-contour detector, dominates the embedding space, leading to systematic underrepresentation of internal strokes, multi-component patterns, and thin or rotated structures. By comparing the saliency profiles of each query and its reference template, we observe how spatial contribution mismatches propagate into ranking errors. The study provides a concise, reproducible framework for auditing CNN feature extraction in symbolic CBIR tasks and highlights structural vulnerabilities relevant to livestock identification, logo retrieval, and other domains involving sparse binary imagery.

1. INTRODUCTION

Convolutional Neural Networks (CNNs) are widely used in supervised visual recognition, in which the prediction layer provides a natural point from which decisions can be interpreted through weights, gradients, or class scores [6]. However, this paradigm does not apply to problems where no shared classes exist, as in cattle brand identification, where each symbol is unique and cannot be treated as a repeated category. In such cases, the lack of labeled samples makes it impossible to train a classifier, rendering any form of logit-based or class-based explainability infeasible.

Under these structural constraints, the problem must be formulated as a Content-Based Image Retrieval (CBIR) task. Instead of prediction, a CNN is used as a feature extractor, producing spatial embeddings that can be compared in a metric space. The search is performed using Facebook Artificial Intelligence Similarity Search (FAISS), which returns a ranked list of the most similar brands in the database [15]. This approach scales well to databases containing thousands of distinct symbols, but shifts the interpretability challenge to a more fundamental level: how can we interpret decisions that emerge solely from the embedding space?

The difficulty increases when the queries are hand-drawn sketches, whereas the registered brand images are clean and binarized. Small variations in stroke thickness, proportions, internal connections, or rotations introduce topological distortions that significantly alter the position of the corresponding embedding. As a result, seemingly minor visual differences can drastically change the returned ranking, making it hard to understand why certain brands appear in the top positions while others do not.

To address this issue, we investigate interpretability at the feature-extraction level, combining three complementary perspectives:

1. **Gradient-weighted Class Activation Mapping (Grad-CAM) applied to the convolutional extractor:** highlighting the regions that contribute most to the construction of the embedding [14];
2. **Structural difference maps between the binarized template and the corresponding hand-drawn sketch:** revealing how drawing distortions visually affect the relevant patterns;
3. **Analysis of the top-10 results returned by FAISS [5]:** identifying which geometric patterns are favored by the embedding structure and which types of errors occur systematically.

Our findings demonstrate that convolutional activations primarily concentrate on the external contours of brands, while internal details such as connections and complex structures receive limited representation in the embeddings. This selectivity precisely explains why minor internal modifications in sketches disproportionately impact the ranking: since the embedding captures only partial visual semantics (prioritizing global silhouettes), any changes to underrepresented internal elements generate substantial variations in calculated similarity. In other words, the system becomes overly sensitive to internal changes precisely because it fails to encode them adequately in its representations, focusing excessively on global features at the expense of distinctive details. The top-10 analysis corroborates this behavior, showing that brands with similar silhouettes but different internal structures are frequently grouped together, confirming the selectivity in visual representation.

By unifying Grad-CAM, structural comparison, and behavioral analysis of the ranking, we present a practical methodology for auditing and interpreting CNN-based CBIR systems in high-contrast binary input domains. The approach contributes to understanding and diagnosing of visual retrieval systems applied to cattle brands. Nevertheless, it can be extended to analogous contexts such as logos, industrial markings, and historical documents.

2. RELATED WORK

Explainable AI (XAI) for image models has historically developed around classification, where logits, class scores, and discriminative gradients provide natural entry points for interpretation. Prototype-based methods leverage this structure by learning representative visual patterns that resemble actual training samples. Neural Prototype Trees integrate differentiable prototypes with hierarchical decision paths [11], whereas hierarchical prototype systems organize learned exemplars into semantic layers that approximate human reasoning [7]. These approaches depend heavily on the visual density and redundancy found in natural images. As a result, they do not transfer well to sparse, high-contrast domains such as binary hand-drawn symbols, where semantic content is encoded in minimal geometric arrangements rather than textured regions.

A second major line of work focuses on activation-map explanations, which highlight the spatial regions that contribute most to a model's response. Class Activation Mapping (CAM) [20] first demonstrated how to combine channel activations with learned class weights to produce coarse localization maps. Grad-CAM [14] generalized this principle by using gradients flowing into the final convolutional layer to generate class-specific saliency, becoming a standard tool due to its simplicity and architectural compatibility. Comparative studies show that Grad-CAM strikes a practical balance between fidelity and interpretability in spatial reasoning tasks [3]. Domain-specific applications reinforce this trend: in medical imaging, saliency maps are aligned with anatomical regions for diagnostic validation [9], and in handwritten character recognition, saliency-based and layer-wise relevance methods reveal where models confuse shape-level invariances with noise [16, 19]. Together, these works suggest a recurring pattern: symbolic and abstract visual domains require tailored interpretability protocols, because their semantics rely on small structural cues rather than photorealistic features.

Parallel to XAI, a substantial body of research in CBIR focuses on similarity rather than classification. Classical retrieval relied on local descriptors such as SIFT [8] and SURF, which encode keypoint-based geometry and support metric comparisons. More recent systems employ deep architectures to generate embeddings, continuous vector representations intended to preserve semantic proximity in a metric space. Advances in metric learning and representation learning have shown that CNN-based embeddings, Siamese networks, and contrastive objectives yield robust retrieval performance across a range of visual tasks [13]. For large-scale applications, efficient vector indexing becomes essential. FAISS is now an industry standard, offering optimized implementations of product quantization, inverted-file indexing, Hierarchical Navigable Small World (HNSW) graphs, and GPU-accelerated k-NN search [5]. Such systems power large-scale retrieval engines across vision, recommendation, and multimodal search.

3. METHODOLOGY

This study examines the interpretability of convolutional feature extraction in binary symbolic retrieval, using cattle brand images as a focused case study. Our goal is to identify which visual structures most strongly influence the embedding space generated by a CNN and how these structures propagate through a FAISS-based similarity pipeline. The methodology consists of four components: reference-data construction, preprocessing, feature-extraction modeling, and similarity-driven retrieval with post-hoc explainability.

3.1 Similarity-Based Retrieval Pipeline

Our retrieval architecture follows a two-phase CBIR workflow based on convolutional embeddings and FAISS vector indexing, following the methodology originally introduced in [10], and summarized in Fig. 1.

3.1.1 Phase 1: Reference-Database Construction

The reference images are created through a standardized workflow consisting of:

- **Preprocessing:** Binarization, size normalization, and aspect-preserving padding are applied to remove background noise (pelage, skin texture, burn artifacts) and to enhance stroke contrast, which is crucial for accurate symbol matching;
- **Feature Extraction:** A pretrained CNN encoder generates a high-dimensional embedding that encodes the spatial structure of the brand;
- **Storage**: Both the processed image and its embedding are stored in the reference database;
- **Indexing:** All embeddings are inserted into a FAISS index using either a flat L2 structure or an approximate index such as Inverted File Index (IVF) or HNSW [5], enabling millisecond-scale nearest-neighbor retrieval.

3.1.2 Phase 2: Query Processing and Retrieval

The query images are likewise processed through a standardized workflow consisting of:

- **Preprocessing and Encoding:** The same normalization and CNN encoding pipeline is applied to the hand-drawn query, ensuring embedding consistency with the reference database;
- **Similarity Search:** FAISS performs approximate k-NN retrieval using cosine similarity or normalized Euclidean distance as the metric;
- **Ranking:** The retrieved neighbors are sorted to produce a Top-k similarity list, which serves as the basis for all explainability analyses.

This architecture allows incremental database expansion: new brands can be indexed without retraining.

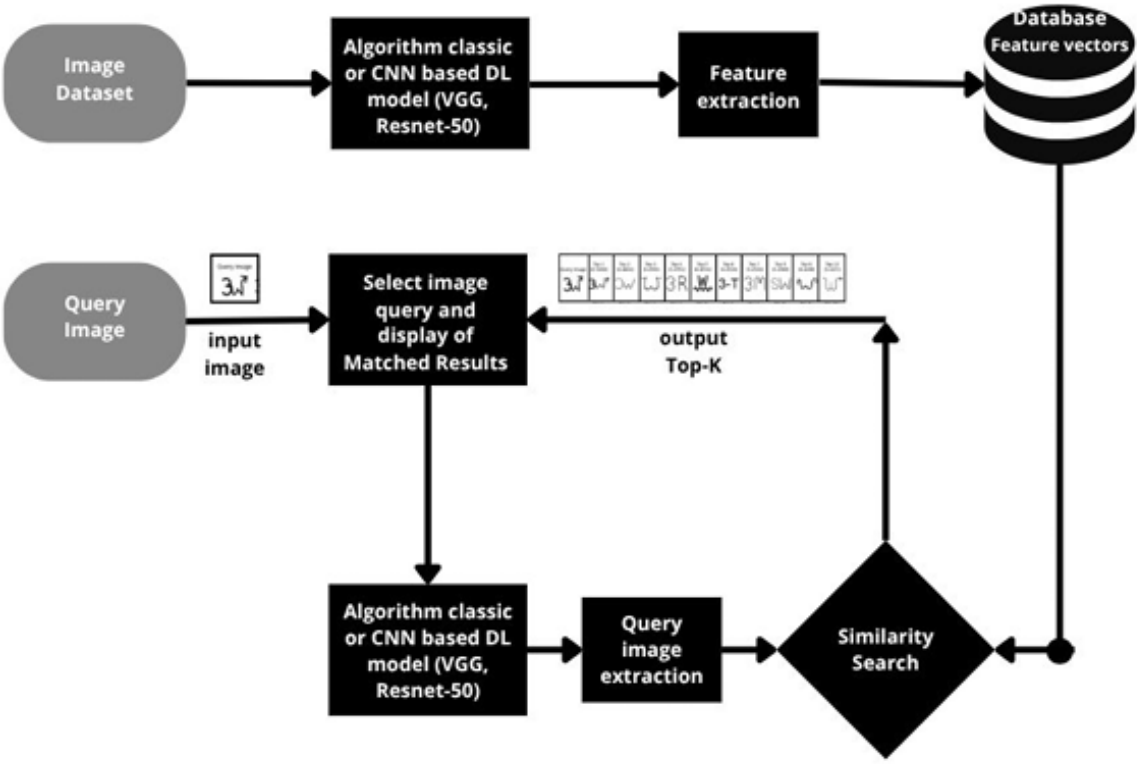


Figure 1: Workflow of the similarity-based retrieval pipeline, adapted from the methodology originally introduced in [10].

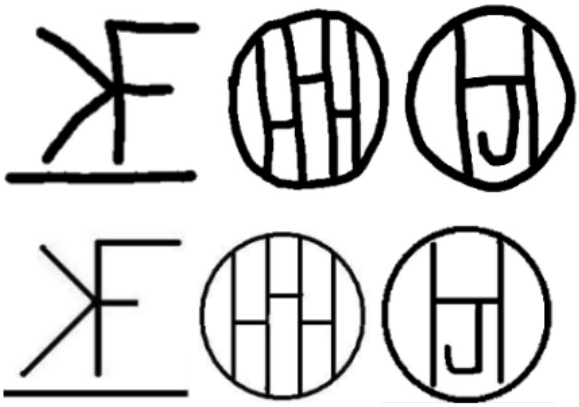


Figure 2: Comparison between the hand-drawn dataset (up) and the corresponding sample from the standard dataset (down).

3.2 Dataset and Preprocessing

To minimize confounding visual factors and ensure a consistent symbolic domain, all images are converted into binary masks that emphasize stroke geometry. We use two datasets, the standard dataset and the hand-drawn dataset, as shown in Fig. 2.

3.2.1 Reference Dataset

We collected 5,230 binary brand templates from publicly available government registries, including the British Columbia BrandBook [4], the Tennessee Agricultural Database [12], and the Oklahoma Brand Registry [1]. We removed duplicates, corrupted images, and overly complex symbols with many disconnected components or text-like structures.

3.2.2 Hand-Drawn Query Dataset

We collected 1,454 hand-drawn sketches captured on touchscreen devices at 640×640 px. Pen widths ranged from 30–60 px to simulate natural human variation and distortions such as rotation, irregular curvature, and junction imprecision.

3.2.3 Preprocessing Pipeline

All images from both datasets were processed using a standardized pipeline consisting of:

- **Binarization:** grayscale conversion and fixed thresholding;
- **Resizing:** scale-and-pad to 200×200 px while preserving aspect ratio;
- **Normalization:** rescaling to $[0, 1]$ and ImageNet-standard normalization.

3.2.4 Sample Selection

From the 1,454 sketches, we selected 304 representative samples using Cochran's formula (95% confidence, $p = 0.5$) [2]. Additionally, the 50 lowest-performing queries (by Top-10 accuracy) were selected for targeted failure analysis.

3.3 Feature-Extraction Model

3.3.1 Network Architecture

We adopt VGG-16 pretrained on ImageNet, truncated at the `block5_conv3` layer. This configuration balances high-level semantic capacity with spatial resolution. The resulting $7 \times 7 \times 512$ activation tensor is flattened and L_2-normalized into a 25,088-dimensional embedding [17].

3.3.2 Similarity Computation

Euclidean distances are computed between query and reference embeddings. We report Top-k accuracy for $k \in \{1, 5, 10, 100\}$ [18].

3.4 Explainability Analysis

Because retrieval depends exclusively on embeddings, interpretability must address how convolutional activations shape the geometry of the embedding space. We focus on identifying which image regions influence the embedding and how distortions propagate into the FAISS neighborhood.

3.4.1 Grad-CAM

We apply Grad-CAM [14] to the `block5_conv3` layer. For feature map A^k and target score y, Grad-CAM computes

$$\alpha_k = \frac{1}{Z} \sum_{i,j} \frac{\partial y}{\partial A^k_{i,j}}, \tag{1}$$

$$\text{CAM} = \text{ReLU}\left(\sum_k \alpha_k A^k \right), \tag{2}$$

where Z denotes the number of spatial locations. The ReLU operation selects positively contributing regions, revealing which contours dominate the embedding.

3.4.2 Analysis Procedure

For each query–reference pair, in both the query and the standard datasets, we apply the following methodology:

1. Grad-CAM heatmaps are generated for both images;
2. Heatmaps are aligned with semantic stroke elements (loops, intersections, endpoints);
3. Differences between query and reference activation patterns are examined to identify distortions influencing the embedding;
4. FAISS Top-10 neighborhoods are inspected to determine whether confusion arises from structural similarity, stroke omission, rotation, or filter bias.

3.4.3 Integration of Results

We obtain a coherent understanding of how the CNN prioritizes contours, and how these priorities shape the embedding space, by combining retrieval metrics, Grad-CAM activation patterns, and qualitative inspection of FAISS neighbors. This integrated perspective allows us to identify failure modes and guide robustness improvements for symbolic, high-contrast domains.

4. RESULTS OF ANALYSIS OF CHANNEL ACTIVATIONS

We systematically analyzed which channels in block5_conv3 were most frequently highlighted as the primary activations across the 304 Grad-CAM heatmaps from our test queries. Table 1 reports the filters that appeared most consistently as dominant features in these visual explanations, with 3 providing their characteristic activation patterns.

A striking concentration of activity emerged: only 10% of channels accounted for more than 90% of all above-threshold activations. This strong imbalance indicates that the embedding space is effectively shaped by a small subset of convolutional filters, each capturing coarse or highly localized geometric cues.

Channel 230 dominated with activations in 59.21% of all queries. Its behavior resembles a generic contour detector that fires on the external boundaries of most symbols. Because outer silhouettes are the most consistent visual feature across the dataset, this filter disproportionately influences the final embedding, propagating its bias directly into the FAISS similarity rankings.

Channels 336, 14, and 454 behaved as specialized detectors. Their activation maps consistently emphasized (i) curved strokes, (ii) line-intersection zones, and (iii) enclosed internal regions. These structural features correspond to the portions of a mark where human variability is highest and where hand-drawn distortions more strongly affect the embedding. However, these filters were activated far less frequently; their low influence suggests that internal structures contribute minimally to the overall embedding geometry.

Channels 429 and 142 appeared exclusively in correctly matched queries. Both capture high-level structures, large-radius curves and multi-component arrangements, that tend to preserve identity even under moderate drawing variation. Their absence in rejected queries suggests that missing or distorted internal details lead the embedding to collapse onto overly generic contour shapes.

Finally, 41.2% of all channels were never activated above threshold for any sample, indicating considerable redundancy. This redundancy also means that the effective embedding dimensionality is far lower than the nominal 25,088 dimensions, which makes the retrieved neighborhoods more sensitive to the few channels that consistently fire.

Together, these findings indicate that the embedding space, and therefore the FAISS ranking, is dominated by contour-focused filters, with only minimal contribution from internal topology.

5. ANALYSIS OF PIXEL WIDTH AND ROTATION SENSITIVITY

A second set of analyses examined how geometric distortions influence embedding stability. Because FAISS retrieves neighbors based on vector distance, any transformation that shifts the embedding significantly will directly alter the Top-k neighborhood.

Table 1: Frequency of Channel Activations with 7% Rejected Queries.

Channel	Total Activations	Approved Queries	Rejected Queries	Activation Pattern	Example Patch
230	180	168	12	Heat around symbol borders	
336	40	34	6	Curved strokes (up to two separate curves)	
14	24	22	2	Line intersection zones	
454	24	22	2	Enclosed regions between elements	
429	18	18	0	Large-radius curves only in approved queries	
142	18	18	0	Complex symbols with multiple components	
Total	**304**	**283**	**21**		

5.1 Stroke Thickness

Hand-drawn strokes originally ranged from 30–60 px. After preprocessing and resizing, samples below the 30 px threshold showed severe accuracy drops: Top-10 accuracy fell from 74% to 48% for strokes near 20 px. Thin strokes weaken contour detectors such as Channel 230 and reduce the firing of specialized filters such as Channels 14 and 454, flattening the embedding into a more uniform or ambiguous region of the space. As a result, FAISS retrieves neighbors dominated by coarse silhouettes rather than fine-grained topological matches.

5.2 Rotation Sensitivity

Table 2 presents matching accuracy under systematic rotations of the reference set. Accuracy dropped from 100% at 0° to 16.67% at 45°. This symmetric degradation around 0° demonstrates that the representation learned by VGG-16 is not rotation invariant, despite the apparent simplicity of the binary domain.

From the perspective of the embedding space, rotation alters the spatial alignment of strokes relative to convolutional filters. As a result, activations shift unpredictably across channels, causing embeddings to drift toward unrelated regions. These geometric shifts propagate directly into FAISS's nearest-neighbor search, leading to mismatched returns even when symbols remain semantically identical. Thus, rotation and thinning systematically break the stability of the embedding, exposing a structural weakness of the feature extractor.

6. ANALYSIS OF COMPLEX PATTERNS

Complex cattle brands, those with tertiary elements such as internal loops, secondary connectors, or small decorative components, were rarely represented coherently in the activation maps. Even the most frequently activated filters did not cover complex marks holistically. Instead, activation maps captured only fragments of the symbol:

- Channel 230: outer contour only;
- Channel 336: single or double curves;
- Channel 14: line intersections;
- Channel 454: enclosed subregions.

As a result, marks with hierarchical or multi-component structure were reduced to incomplete embeddings dominated by the silhouette, causing FAISS to retrieve neighbors with similar outlines but different internal topology.

For example, when tiny embellishments near junctions or additional letters were present (refer to 3, Channel: 454 and 454), their absence or deformation in the hand-drawn query suppressed activation in specialized channels, shifting the embedding toward coarse matches.

These observations confirm a broader pattern: the feature extractor encodes symbols through a sparse sampling of visual attributes, primarily contours, while ignoring distinctive internal elements necessary for disambiguation. This explains the recurrent FAISS errors observed for complex brands.

7. QUALITATIVE ANALYSIS

This section provides a compact overview of the qualitative differences observed across the analyzed channels as summarized in Table 3. For each channel, we include the Grad-CAM of the original cattle brand, the Grad-CAM of the hand-drawn query, and the Top-10 retrieved results used in the retrieval experiment. The textual description summarizes how structural differences between the original and hand-drawn inputs translate into shifts in activation patterns and retrieval performance.

8. IMPLICATIONS

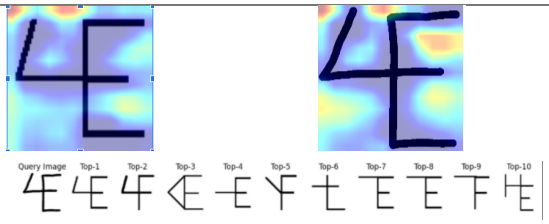
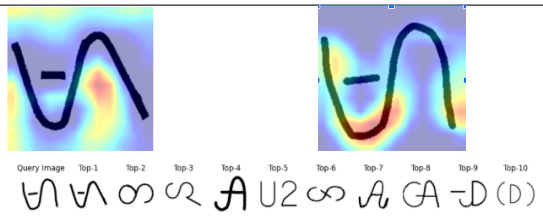
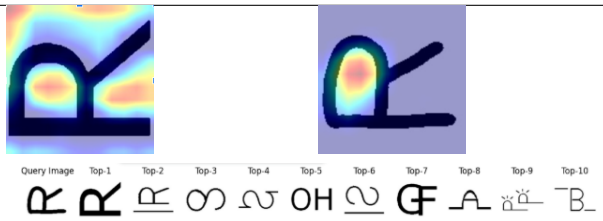
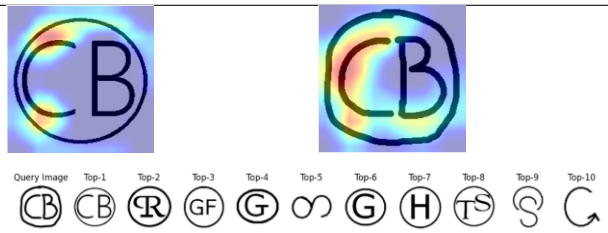
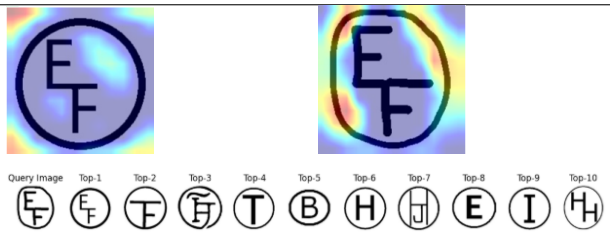
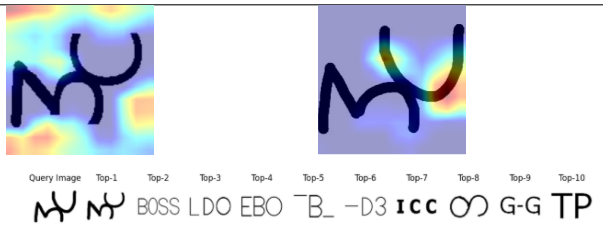
Our results reveal a structural property of the VGG-16 feature extractor: a small set of contour-dominant filters disproportionately shapes the embedding space, while detailed internal structure is systematically underrepresented. This has three major implications:

1. **Embedding geometry is coarse and contour-centric:** because most channels never activate or activate only weakly, the resulting embeddings cluster

Table 2: Matching accuracy of the CNN at different rotation increments.

Algorithm	-45º	-30º	-15º	0º	15º	30º	45º
VGG-16	18.52%	46.30%	98.15%	100.00%	99.04%	48.15%	16.67%

Table 3: Summary of qualitative outcomes across analyzed channels.

Channel	Original	Query	Description
230			The hand-drawn query and original image share similar large-scale contour structures. Both Grad-CAM maps highlight outer-edge saliency typical of Channel 230, explaining the correct top-ranked retrieval.
336			Original emphasizes contour flow (230-like), whereas the query activates curved corners (336-like). Still, the correct match remains among the top results.
14			The query is thinner and less precise; saliency shifts from outer contours (230) to internal enclosed regions (14), especially inside the letter "R".
454			Original activates Channel 429 along circular contours; query concentrates saliency between shapes. Top-10 includes similar circle–letter structures.
429			Original dominated by diffuse 230-like patterns; query shifts toward internal activation clusters around the letter "E".
142			Both images share structure, but saliency shifts: original shows diffuse 230-like behavior, while the query emphasizes boundary activation along the letter "C".

marks based on silhouette similarity, irrespective of internal differences;

2. **FAISS neighborhoods reflect filter biases:** The nearest neighbors returned by FAISS mirror the lim-

ited visual vocabulary of the embeddings. Errors arise not from the FAISS mechanism, but from upstream representational collapse;

3. **CNNs struggle with sparse, high-contrast symbols:** VGG-16 performs well on simple datasets like MNIST, but its contour bias and lack of rotational invariance severely limit its use in complex brand patterns.

9. CONCLUSION

This work demonstrates that, when applied to binary cattle-brand recognition, the VGG-16 feature extractor collapses a rich symbolic domain into a narrow, contour-centric embedding space. Through combined analysis of channel activations, Grad-CAM visualizations, and FAISS retrieval behavior, we showed that internal structures, junctions, enclosed regions, and tertiary elements are consistently underrepresented in the embeddings. These representational gaps distort the geometry of the vector space, causing FAISS to retrieve neighbors based primarily on silhouette similarity, and performance deteriorates sharply under stroke thinning, geometric distortion, and rotation. To correct these structural weaknesses, training should incorporate controlled rotations, stroke thinning, geometric jitter preserving topology, and domain-specific contrast augmentations. Additional architectural adjustments, such as adapter blocks, attention over late-layer activations, or channel-balancing modules, could further distribute representational load across channels. Ultimately, the diagnostic methodology introduced here provides a reproducible framework for interpreting embedding-based CBIR systems and highlights the need for domain-tailored augmentations and architectural refinements. These findings underscore that explainability in retrieval must be built at the feature-extraction level, where the embedding itself is formed, rather than at the prediction layer, because in CBIR the prediction layer does not exist.

10. ADDITIONAL AUTHORS

Edmundo Hoyle, Center of Excellence in Artificial Intelligence (CEIA), edhoyle@gmail.com.

11. REFERENCES

[1] O. C. Association. Oklahoma brand registration. `https://www.okcattlemen.org/brands`. Accessed: 2025-05-30.

[2] W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 3rd edition, 1977.

[3] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, 36(1):235–249, 2020.

[4] O. I. Inc. British columbia livestock brand database (brandbook). `https://www.ownershipid.ca/brandbook`. Accessed: 2025-05-30.

[5] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with FAISS. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[7] X. Li et al. Hierarchical prototype-based explanations. *Transactions on Machine Learning Research*, 2024.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] A. S. Lundervold, A. B. Arrieta, et al. Explainable ai: A review of applications to neuroimaging data. *Frontiers in Neuroscience*, 16:906290, 2022.

[10] M. V. S. Medeiros, L. A. Soares, E. Hoyle, and A. A. Díaz-Salazar. Visual similarity search of cattle brands using deep learning on binary representations. In *2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6, Salvador, Brazil, 2025.

[11] M. Nauta, R. van Bree, and C. Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *CVPR*, pages 14933–14943, 2021.

[12] T. D. of Agriculture. Tennessee agricultural brand database. `https://agriculture.tn.gov/ListBrand.asp`. Accessed: 2025-05-30.

[13] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[15] J. Silva, B. Pereira, and L. Santos. Segmentation and detection of cattle branding images using cnn and svm. *Journal of Agricultural Informatics*, 8(2):45–53, 2017.

[16] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2013.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[18] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[19] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.