Final Project – XAI Case Study Applied to the Fashion Industry

Emma Rey Sánchez

1 Introduction

Deep learning models have become widely used for image classification tasks, including applications in the fashion industry such as product categorisation and recommendation systems. While these models often achieve high predictive accuracy, their decision-making processes are typically opaque, which limits trust and may hide undesired or unstable behaviours.

This issue is especially relevant in fashion image classification, where different categories can share strong visual similarities. For instance, dresses and skirts may appear alike in local image regions, yet differ in their overall structure. A model that bases its predictions on narrow or spurious visual cues may perform well on a benchmark dataset while failing to generalise in real-world scenarios.

These limitations affect multiple stakeholders. End users may receive incorrect search results or recommendations when visually similar garments are misclassified, leading to a frustrating user experience. Fashion retailers and e-commerce platforms may suffer from catalogue inconsistencies, such as products appearing in the wrong categories, which can negatively impact discoverability and conversion rates. Finally, for machine learning practitioners, the lack of interpretability makes it difficult to diagnose model errors, identify spurious visual cues, and design effective improvements to the training process. Model explainability is therefore essential not only for transparency, but also for reliable deployment and iterative model development.

In this project, the behaviour of a convolutional neural network trained on a subset of the DeepFashion dataset is studied, focusing on the classification of women's clothing into five categories: *Dresses, Graphic Tees, Pants, Shorts,* and *Skirts*. Rather than limiting the analysis to performance metrics, explainable artificial intelligence (XAI) techniques are applied to understand which visual regions drive the model's predictions and how these patterns vary across classes.

Grad-CAM++ is used to produce local explanations for individual predictions and to aggregate these explanations in order to obtain global insights into the model's average behaviour. In addition, a sanity check based on model randomisation is performed to validate the reliability of the explanations. Finally, the insights obtained through XAI are used to guide a targeted intervention in the training process, demonstrating how interpretability can support actionable model improvements.

2 Data and Methods

2.1 Dataset

The experiments are conducted using a subset of the DeepFashion dataset, restricted to images of women's clothing. Five visually related categories are selected: *Dresses, Graphic Tees, Pants, Shorts,* and *Skirts*. This selection intentionally includes classes with significant visual overlap, making the classification task non-trivial and particularly suitable for explainability analysis.

The dataset is organised into training, validation, and test splits, ensuring that evaluation is performed on unseen data. All images are resized to a fixed resolution and normalised using

ImageNet statistics. Class imbalance across categories is addressed through the use of class-weighted loss functions during training.

2.2 Model Architecture and Training

A ResNet-18 convolutional neural network pretrained on ImageNet is used as the base architecture. The final fully connected layer is replaced to match the number of target classes. This architecture is selected due to its strong performance in image classification tasks and its compatibility with gradient-based explainability methods.

Training is performed using the AdamW optimizer and a weighted cross-entropy loss. Data augmentation techniques, including random horizontal flips, colour jittering, affine transformations, and random erasing, are applied during training to improve generalisation. Model selection is based on validation accuracy, and the checkpoint achieving the best validation performance is retained for final evaluation on the test set.

2.3 Explainability Methods

Model behaviour is analysed using Grad-CAM++, a gradient-based explainability technique designed for convolutional neural networks. Grad-CAM++ produces class-specific heatmaps that highlight the spatial regions of an input image that contribute most strongly to a given prediction.

Local explanations are obtained by visualising Grad-CAM++ heatmaps for individual test images, including both correct predictions and frequent misclassifications. To capture global model behaviour, these heatmaps are aggregated across multiple samples, enabling the identification of systematic attention patterns associated with each class.

The reliability of the explanations is assessed through a sanity check based on model randomisation. Grad-CAM++ explanations generated from the trained model are compared with those obtained from a randomly initialised model, verifying that meaningful explanations depend on learned parameters rather than on image artefacts.

2.4 Actionable Use of Explanations

Insights derived from the explainability analysis are used to inform targeted modifications to the training process. In particular, Grad-CAM++ reveals that confusions between dresses and skirts are often driven by an over-reliance on lower-body visual cues. Based on this observation, a data augmentation strategy incorporating a top-biased cropping transformation is introduced to encourage the model to focus more strongly on upper-body information.

The impact of this intervention is evaluated through a before-and-after comparison, considering both overall performance metrics and the frequency of *Dresses* $\rightarrow$ *Skirts* misclassifications. This approach illustrates how explainability techniques can be leveraged not only for post-hoc interpretation, but also to guide actionable improvements in model behaviour.

3 Results

3.1 Correct Predictions and Local Explanations

Figure 1 presents one correctly classified example for each of the five garment categories, together with the corresponding Grad-CAM++ visualisations. In these cases, the attention maps highlight semantically meaningful regions of the garments, such as the torso area for *Graphic Tees*, the full vertical silhouette for *Pants*, and the lower-body region for *Skirts*. These examples illustrate that, when predictions are correct, the model generally focuses on visually coherent and class-relevant regions.

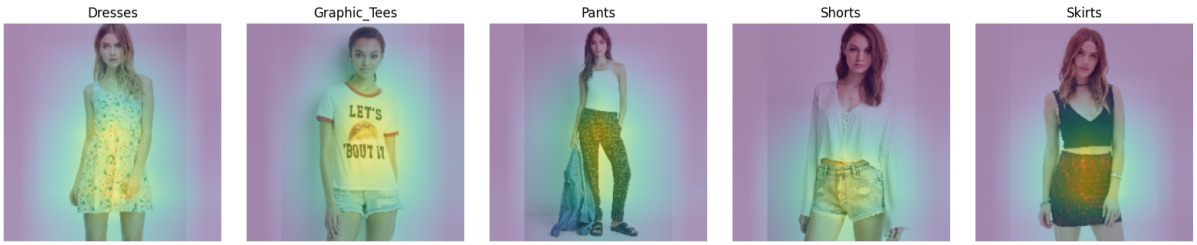

Figure 1: Correct predictions for each class with Grad-CAM++ overlays.

3.2 Confusion Matrix Analysis

Figure 2 shows the confusion matrix of the baseline model evaluated on the test set. While most classes are classified reliably, a notable confusion pattern emerges between *Dresses* and *Skirts*. A significant number of dress images are misclassified as skirts, suggesting that the model struggles to distinguish between garments that share similar lower-body visual characteristics.

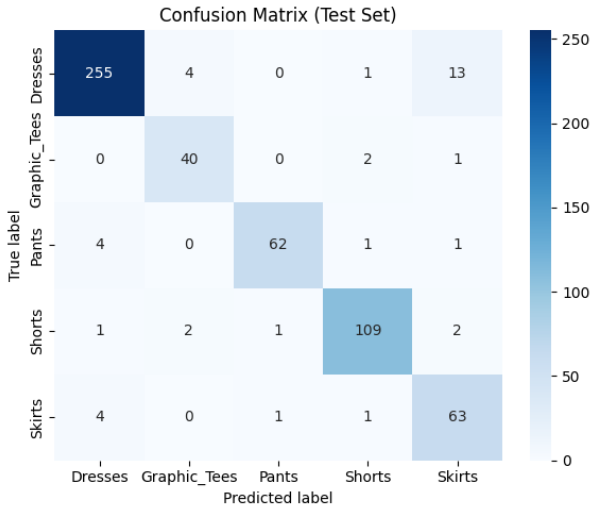

Figure 2: Confusion matrix of the baseline model.

3.3 Misclassification Analysis with Grad-CAM++

Figure 3 displays three representative examples of *Dresses* → *Skirts* misclassifications, together with their Grad-CAM++ heatmaps. In all cases, the attention is concentrated predominantly on the lower part of the image, particularly around the hemline, while upper-body regions are largely ignored. This consistent behaviour indicates that the model relies on local lower-body cues rather than on the global garment structure, leading to systematic errors.

Figure 3: Examples of Dresses misclassified as Skirts with Grad-CAM++ overlays.

3.4 Global Explainability Analysis

To characterise average model behaviour, Grad-CAM++ heatmaps are aggregated across multiple samples. Figure 4 presents global heatmaps for correctly classified *Dresses*, correctly classified *Skirts*, and *Dresses* misclassified as *Skirts*.

The aggregated explanations reveal that *Skirts* predictions are dominated by attention on the lower region of the image, whereas *Dresses* exhibit a more vertically distributed attention pattern. Importantly, the global heatmap of misclassified dresses closely resembles that of correctly classified skirts, confirming that the confusion arises from consistent, class-level reliance on similar visual cues.

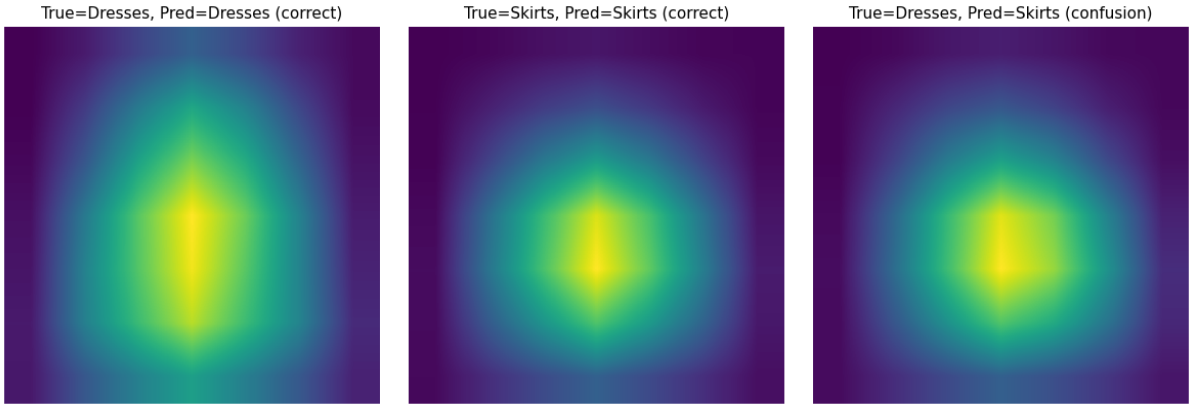

Figure 4: Aggregated Grad-CAM++ heatmaps for Dresses, Skirts, and misclassified Dresses.

3.5 Sanity Check for Explanations

Figure 5 illustrates the sanity check performed using model randomisation. Grad-CAM++ heatmaps generated from the trained model display structured and interpretable attention patterns, whereas those obtained from a randomly initialised model appear diffuse and uninformative. This contrast supports the conclusion that the explanations are driven by learned model parameters rather than by artefacts inherent to the input images.

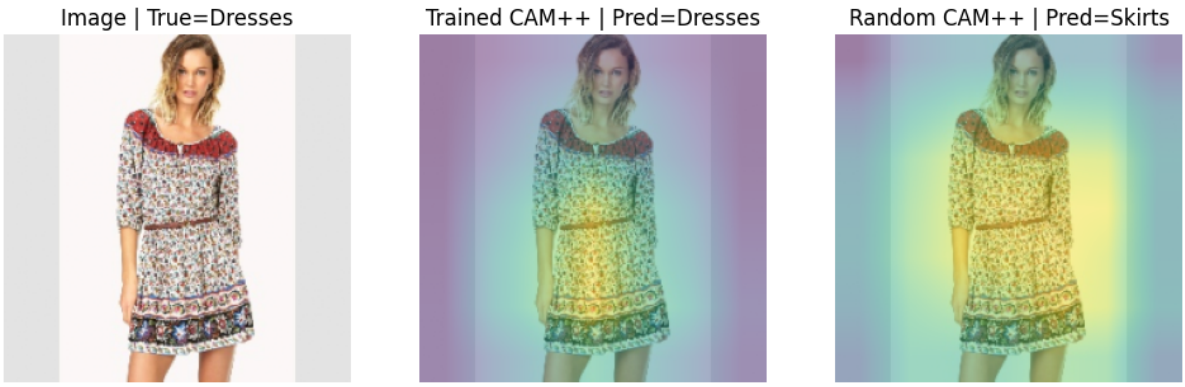

Figure 5: Sanity check comparing Grad-CAM++ explanations for trained and random models.

3.6 Actionable Results

Figure 6 compares the confusion matrices before and after applying the explainability-guided data augmentation strategy. Following the introduction of top-biased cropping, the number of *Dresses* → *Skirts* misclassifications is reduced from 15 to 9. Although this intervention leads to a slight decrease in overall accuracy, it effectively mitigates the specific failure mode identified through explainability analysis.

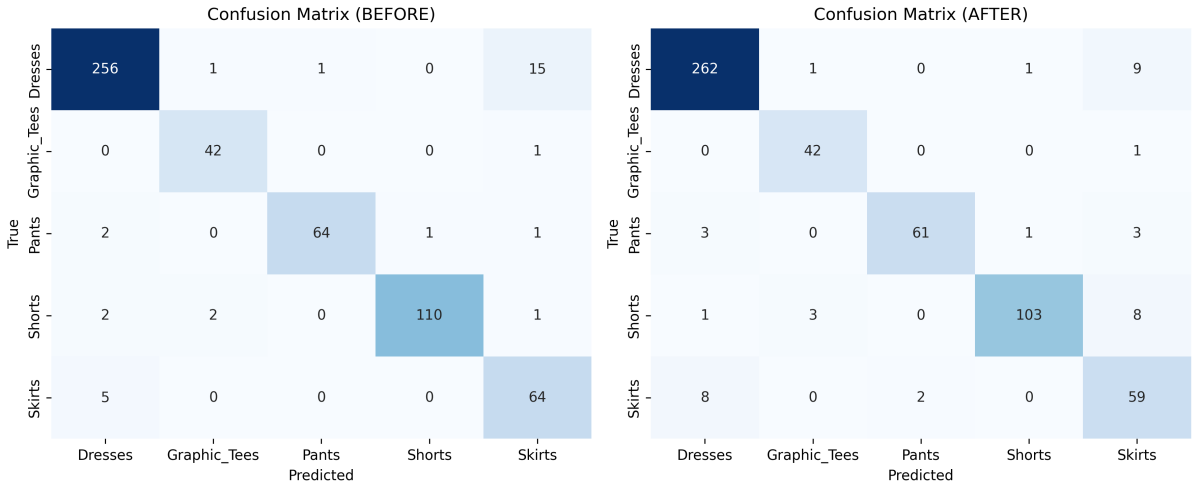

Figure 6: Confusion matrices before and after the explainability-guided intervention.

4 Discussion

The results highlight both the strengths and limitations of explainable artificial intelligence methods when applied to image classification models. Grad-CAM++ proves effective in revealing systematic model behaviours that are not immediately apparent from performance metrics alone, particularly the over-reliance on lower-body visual cues in the Dresses versus Skirts classification task.

The actionable intervention guided by these explanations demonstrates that XAI can support targeted model improvements. Although the proposed data augmentation strategy reduces a specific and interpretable error pattern, it also introduces a trade-off in terms of overall accuracy. This outcome illustrates an important limitation: mitigating a particular failure mode does not necessarily translate into global performance gains and may require careful balancing between robustness and accuracy.

Several limitations of this study should be acknowledged. First, the analysis is restricted to a subset of the DeepFashion dataset and a limited number of classes, which may constrain the generalisability of the findings. Second, Grad-CAM++ provides spatial explanations but does not capture higher-level semantic concepts or causal relationships. Finally, the evaluation of explainability remains partly qualitative, relying on visual inspection of heatmaps alongside targeted quantitative metrics.

Future work could explore complementary explainability techniques, such as concept-based or counterfactual methods, to gain a deeper understanding of model decisions. Additionally, incorporating richer garment annotations like textual descriptions could help further reduce ambiguities between visually similar classes. Extending the approach to larger datasets and more complex models would also provide insight into the scalability of actionable XAI strategies in real-world fashion applications.