Empowering the Captioning of Fashion Attributes from Asian Fashion Images

KVS Perera[1], DDA Gamini[1#]
[1]Department of Computer Science, University of Sri Jayewardenepura, Sri Lanka
[#]gamini@sjp.ac.lk

ABSTRACT Fashion image captioning, an evolving field in AI and computer vision, generates descriptive captions for fashion images. This paper addresses the prevalent bias in existing studies, which focus predominantly on Western fashion, by incorporating Asian fashion into the analysis. This paper describes developing more inclusive AI technologies for the fashion industry by bridging the gap between Western and Asian fashion in image captioning. We leverage transfer learning techniques, combining the DeepFashion dataset (primarily Western fashion) with a newly curated Asian fashion dataset. Our approach employs advanced deep learning methods for the encoder and decoder components to generate high-quality captions that capture various fashion attributes, such as style, color, and garment type, tailored specifically to Asian fashion trends. Results demonstrate the efficacy of our methods, with the model achieving accuracies of 93.63% for gender, 83.42% for article type, and 61.34% for base color on the training dataset, and 94.13%, 79.25%, and 59.71%, respectively, on the validation dataset. These findings highlight the importance of inclusivity and diversity in AI research, advancing the field of fashion image captioning.

INDEX TERMS Multi-label image captioning, deep learning, fashion image analysis, Asian fashion images, transfer learning

I. INTRODUCTION

The fashion industry is a dynamic and ever-evolving domain, with new trends and styles continually emerging. Fashion enthusiasts and industry professionals are constantly engaged in decoding these trends, understanding consumer preferences, and predicting future shifts. Traditionally, the categorization and analysis of fashion products have required extensive manual effort and expertise, a process both time-consuming and prone to human error.

The advent of computer vision and machine learning has brought transformative changes to this landscape, offering new possibilities for automating and enhancing the analysis of fashion products. These technologies enable the extraction of valuable insights from visual data, facilitating accurate categorization, classification, and prediction of various fashion attributes. This paper leverages these advancements to develop a robust system focused on gender, article type, and base color classification in fashion images.

Unlike traditional methods reliant on manual assessment or rule-based systems, our approach employs advanced machine learning algorithms to analyze and interpret fashion images. This method allows us to uncover underlying patterns, correlations and trends that may not be immediately discernible to human observers. By integrating computer vision and machine learning with fashion analysis, we aim to streamline the categorization process, providing invaluable insights to designers, retailers, and consumers. Beyond fashion analysis, our research holds potential applications in e-commerce, retail merchandising, and trend forecasting, revolutionizing the industry's approach to product analysis.

II. RELATED WORK

The domain of fashion image captioning has seen notable advancements, driven by the demand for AI solutions capable of generating accurate and descriptive captions for fashion content. Despite its growing importance, this area remains under-researched, with few studies focusing on specialized models for fashion-related content.

One significant contribution is the work by researchers who introduced "Accurate and Expressive Fashion Captioning: A Learning Framework" [1]. Their framework utilizes attribute-level semantic rewards and attribute embedding techniques to create precise and expressive descriptions for fashion items. Building on this foundation, the study integrates maximum likelihood estimation (MLE) and Reinforcement Learning (RL) to enhance caption quality further. The researchers developed the FAshion CAptioning Dataset (FACAD), a comprehensive collection of 993,000 fashion images paired with 130,000 diverse and enchanting descriptions. This dataset enables training models to effectively capture and convey fashion item attributes. Experiments on FACAD demonstrate the framework's effectiveness in generating high-quality captions, showcasing advancements in fashion captioning through innovative methodologies and robust dataset utilization. Building on this foundation, Moratelli et al. [2] proposed an approach integrating external memory retrieval with transformer-based neural networks for fashion captioning. Their method leverages transformer architectures with cross-attention mechanisms for reading and retrieving items from external textual memory, facilitated by k-nearest neighbor (kNN) searches. This design optimizes the flow of information from external sources using a novel fully attentive gate mechanism. The study achieved state-of-the-art performance on the FACAD fashion captioning dataset, demonstrating its ability to generate detailed and contextually rich descriptions of fashion articles. This approach highlights the effectiveness of incorporating external textual memory to enhance the quality and informativeness of image captions in fashion AI applications.

[3] presents an advanced approach to fashion synthesis using generative adversarial networks (GANs). The research focuses on generating new clothing designs on a wearer while preserving the wearer's body structure and pose, guided by a descriptive sentence about the desired outfit. To achieve this, the generative process is divided into two conditional stages. In the first stage, a semantic segmentation map is generated with a spatial constraint that respects the wearer's pose. This map serves as a latent spatial arrangement guiding the synthesis process. The second stage employs a novel compositional mapping layer within the GAN framework to render the final image with precise regions and textures based on the generated segmentation map. The researchers extended the DeepFashion dataset by annotating 79,000 images with descriptive sentences, enabling training and evaluation of their approach. Quantitative metrics and qualitative assessments demonstrate the effectiveness of the method in producing realistic and structurally coherent fashion images aligned with textual descriptions, highlighting advancements in generative fashion modeling.

A novel approach to clothing style detection and retrieval through fine-grained learning is introduced in [4]. They address challenges like clothing item variability and deformability by creating a detailed attribute vocabulary from human annotations on a specialized dataset. This vocabulary trains a visual recognition system capable of identifying complex stylistic attributes beyond basic features like color and pattern. The study validates its effectiveness with benchmark tests on the Women's Fashion Coat Dataset, demonstrating superior recognition and differentiation of nuanced stylistic elements. Furthermore, it explores the application of attribute-based multimedia retrieval in mobile interfaces, enhancing user experience through detailed image annotations and precise clothing item searches. This integration of human-derived attribute vocabularies with advanced visual recognition techniques provides valuable insights for enhancing the granularity and accuracy of clothing style detection and retrieval systems.

[5] provides large-scale clothes dataset with over 800,000 images annotated comprehensively with attributes, landmarks, and cross-scenario correspondences. They propose FashionNet, a deep learning model optimized iteratively to predict clothing attributes and landmarks simultaneously. FashionNet utilizes predicted landmarks to enhance feature learning through pooling or gating mechanisms. This work establishes DeepFashion as the largest annotated clothes dataset, facilitating advancements in clothes recognition and retrieval algorithms, with defined benchmark datasets and evaluation protocols.

An automated system [6] is designed to generate detailed semantic descriptions of clothing from images. The system extracts low-level features in a pose-adaptive manner, ensuring that attributes are accurately identified regardless of the position of the person in the image. By combining these features, the system trains attribute classifiers to recognize various clothing characteristics. It further improves attribute predictions by using a Conditional Random Field (CRF) to model the dependencies between attributes, which allows for more accurate and contextually relevant descriptions. The system's performance was validated on a challenging clothing attribute dataset, demonstrating its ability to generate precise and meaningful clothing descriptions. This research provides valuable insights into the use of pose-adaptive feature extraction and CRF for enhancing the accuracy of semantic attribute recognition in clothing.

In their study [7], an advanced approach to image captioning that goes beyond the traditional encoder-decoder models, which typically only recognize objects and their relationships in a given image is explored. The authors identify a critical gap in existing methodologies, particularly when applied to fashion images, where it's essential not only to describe items but also to capture intricate details such as texture, fabric, shape, and style. To address this gap, the researchers propose an innovative model that integrates an attention mechanism within the conventional encoder-decoder framework, enabling it to generate more comprehensive and nuanced descriptions of fashion items. This model leverages spatial attention to dynamically adjust the sentence generation context across multiple layers of feature maps, ensuring that both the items and their detailed attributes are effectively covered in the generated captions. The efficacy of this approach is validated through experiments on the Fashion-Gen dataset, a leading dataset in the field of fashion image analysis. The results demonstrate significant improvements, with the model achieving impressive scores on key metrics like CIDEr, ROUGE-L, and BLEU-4, surpassing baseline methods on the same dataset. This research is particularly relevant to our study as it highlights the importance of incorporating detailed attribute recognition and spatial attention mechanisms in fashion image captioning, offering a robust framework that enhances the descriptive quality and relevance of generated captions.

[8] addresses the challenge of maintaining captioning quality when input data diverges from the training distribution, crucial in fashion's nuanced garment descriptions. By employing a pre-training strategy with noise generation, the study enhances system generalization. It integrates GPT-2, Vision Transformer, and BERT, showing competitive results even with limited fine-tuning. Colombo emphasizes user-centric evaluation via a user study, confirming improved captioning quality. This work suggests transfer learning's efficacy in adapting captioning systems to varied data, vital for reliable outputs across applications.

Recent research has increasingly emphasized the importance of inclusivity and diversity in artificial intelligence (AI) technologies, addressing the under-representation of various cultural and regional fashion styles in existing datasets and models. Hacheme and Sayouti [9], applying the "Show and Tell" model introduced by Vinyals et al. in 2015 [10], highlighted the lack of representation of African fashion styles in current datasets. In response, a pioneering study introduced the InFashAIv1 dataset, comprising nearly 16,000 African fashion item images with titles, prices, and general descriptions. This dataset, alongside the well-known DeepFashion dataset, serves as a critical resource for training AI models in fashion image captioning. Captions are generated using the Show and Tell model, leveraging a CNN encoder and RNN decoder architecture. The research demonstrates that joint training on both datasets significantly enhances caption quality, particularly for African style fashion images, demonstrating effective transfer learning from Western style data. The release of the InFashAIv1 dataset on GitHub aims to foster research that promotes diversity and inclusion in fashion AI applications.

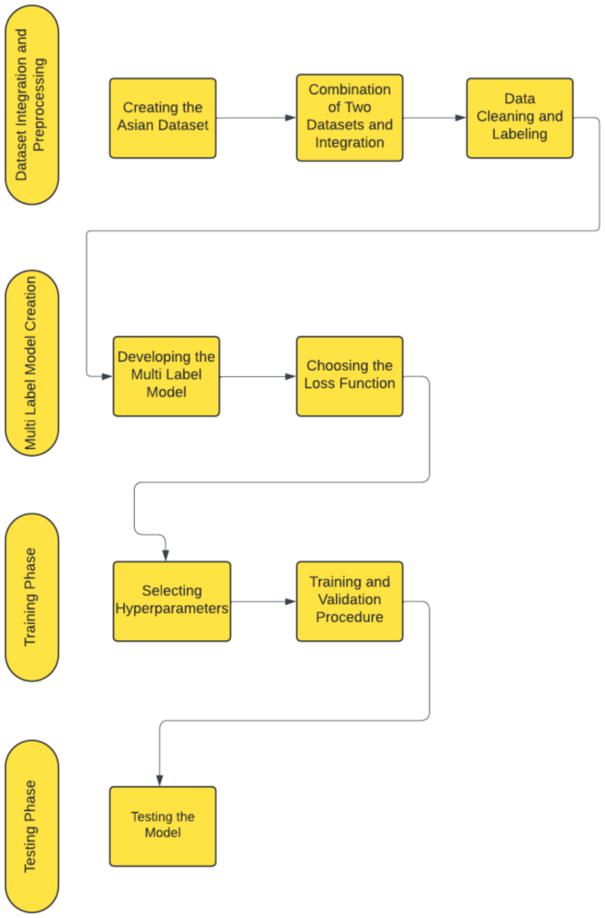
Recent advancements in fashion image captioning have explored various techniques to enhance the accuracy and richness of descriptions for fashion items, employing technologies such as deep learning models, external memory retrieval, and semantic segmentation. Methodologies encompass attribute-level semantic rewards, reinforcement learning, attention mechanisms, and generative adversarial networks (GANs). Table 1 presents a summary of key findings and contributions from recent research in fashion image captioning. Current fashion image captioning research predominantly focuses on Western fashion, resulting in a significant gap in the representation and understanding of non-Western styles, particularly Asian fashion. Existing datasets and models lack the diversity needed to accurately capture and describe the unique attributes of these underrepresented fashion styles. This limitation leads to poor performance and limited applicability of existing captioning systems for non-Western fashion. So, as the research gap highlights, there is a need for diverse datasets that include Asian fashion styles and specialized models that can effectively address this gap. Our research aims to fill this void by curating a comprehensive dataset of Asian fashion and developing models that improve captioning accuracy and inclusivity for global fashion audiences.

Table 1. Summary of related works

Reference	Technology Used	Key Findings
[1]	Attribute-level semantic rewards, RL, MLE	Enhanced caption quality for fashion images
[2]	Transformer-based NNs, External memory retrieval, Cross-attention mechanisms	Achieved good performance on FACAD dataset, generating detailed and context-rich captions
[3]	GANs, Semantic segmentation	Produced realistic and structurally coherent fashion images, preserving body structure and pose
[4], [5]	Visual recognition system, Human annotations	Highlighted the capability of recognizing and retrieving fine-grained clothing styles with a visual recognition system. Demonstrated the system's capability to distinguish intricate stylistic details in clothing items
[6], [9], [10]	CNN, RNN, Semantic attributes, Neural image captioning	Generated detailed descriptions of clothing items. Enhanced model performance for under-represented fashion styles. Highlighted advancements in neural image caption generation
[7]	Attention mechanisms, Deep learning models	Improved the relevance and richness of fashion image captions
[8]	Transfer learning, NNs	Revealed the effectiveness of transfer learning for improving captioning models

III. METHODOLOGY

This section details the methodology adopted for developing our fashion image captioning model with a focus on Asian fashion styles. The process involves several key steps: dataset integration and preprocessing, creation of a multi-label model, and the subsequent training and testing phases. Initially, a specialized Asian fashion dataset is curated and integrated with an existing benchmark dataset to form a comprehensive repository. Following this, a multi-label classification model is developed, leveraging advanced neural network architectures and tailored loss functions. The model is then trained and validated using carefully selected hyperparameters, with performance metrics calculated at each stage. Finally, the model is tested on unseen data to evaluate its accuracy and robustness. Figure 1 provides a concise summary of the methodological steps and the technologies employed in this research.



Dataset Integration and Preprocessing
Creating the Asian Dataset
Combination of Two Datasets and Integration
Data Cleaning and Labeling
Multi Label Model Creation
Developing the Multi Label Model
Choosing the Loss Function
Training Phase
Selecting Hyperparameters
Training and Validation Procedure
Testing Phase
Testing the Model

Figure 1. Methodological steps

A. Dataset Integration and Preprocessing

1) Asian Dataset Creation: In this research, a specialized Asian-focused dataset comprising 2,500 meticulously curated fashion images is developed, sourced primarily from platforms like Pinterest. Careful selection criteria ensure the dataset's breadth, encompassing traditional, contemporary, casual, and formal Asian fashion styles. Preprocessing steps, including image resizing and quality enhancement, prepare the dataset for annotation and model training. Special emphasis is placed on capturing a diverse palette of colors prevalent in Asian fashion, ensuring richness and inclusivity. Additionally, the dataset covers a wide array of clothing types commonly found in Asian fashion, guaranteeing versatility and accuracy in recognizing and describing various fashion items and styles.

2) Dataset Combination and Integration: In the process of dataset combination and integration, merging the curated Asian dataset with the benchmark Fashion dataset is a critical step aimed at enhancing the model's robustness and generalization capabilities. This fusion leverages the diverse attributes of both datasets, creating a more comprehensive training environment for the multi-label image classification model. Before merging, meticulous attention is paid to ensuring consistency and compatibility between the two datasets. This involves aligning column names and preserving their order across both data frames to facilitate seamless integration and avoid discrepancies during concatenation. Attribute names such as 'gender,' 'articleType,' and 'baseColour' are standardized to ensure uniformity and coherence, maintaining data integrity. Following the alignment process, the curated Asian dataset is merged with the benchmark Fashion dataset, resulting in a unified dataset that encompasses a broader spectrum of fashion styles, trends, and attributes. This combined dataset serves as a comprehensive repository, enabling the model to learn from a diverse range of fashion data and improve its ability to recognize and describe various fashion items accurately.

3) Data Cleaning and Labeling: In the dataset cleaning and labeling phase, ensuring the quality, consistency, and reliability of the dataset is crucial. This involves removing inconsistencies, missing values, and irrelevant data entries, and annotating the dataset with detailed attributes for multi-label classification and captioning tasks. A thorough cleaning process is conducted to eliminate discrepancies, duplicates, and missing values to ensure completeness and reliability. Subsequently, label dictionaries are created and mapped for categorical attributes, transforming textual labels into a numerical format for efficient model training and inference. This includes generating unique category mappings for gender, articleType, and baseColour attributes, with special handling for NaN values in baseColour.

B. Multi Label Model Creation

1) MultiHeadResNet50: In the multi-caption model development, the cornerstone is the MultiHeadResNet50 architecture, leveraging the ResNet50 backbone pretrained on the ImageNet dataset for its feature extraction capabilities. This model features three distinct fully connected layers, each serving as a classification head for specific fashion attributes. The gender classification head outputs predictions across five gender categories, including male, female, unisex, kids, and others, enabling effective categorization of gender representation in fashion images. The article type classification head predicts across 142 categories, covering a wide range of fashion items from shirts and dresses to specialized articles, providing a comprehensive understanding of fashion ensembles. The base color classification head outputs predictions across 47 color categories, capturing a diverse palette from traditional colors like black and white to nuanced shades like turquoise blue and fluorescent green, reflecting the richness and diversity of fashion colors.

2) Loss Function: Complementing the model architecture, a bespoke loss function is formulated to efficiently optimize the multi-label classification task. This function employs the Cross-Entropy Loss mechanism, a widely adopted approach for classification tasks, to compute the loss for each classification head. Mathematically, the loss function is defined as:

$$Loss_{total} = \frac{Loss_{gender} + Loss_{articleType} + Loss_{baseColour}}{3}$$

where:

$Loss_{gender}$, $Loss_{articleType}$ & $Loss_{baseColour}$ represent the Cross-Entropy Losses computed for gender, articleType and baseColour predictions, respectively.

C. *Training Phase*

1) Selecting hyperparameters: The training phase optimizes the MultiHeadResNet50 model using key hyperparameters and configurations. The learning rate is set to 0.001, determining the step size during optimization and affecting convergence and stability. The Adam optimizer is employed for its adaptive learning rate capabilities, efficiently handling large-scale datasets. A batch size of 32 is used, specifying the number of samples processed before updating the model's weights, which impacts memory utilization and learning stability. The model undergoes 40 training epochs, indicating the number of times the entire dataset is passed forward and backward through the neural network.

2) Training and Validation Procedure: The training and validation procedures are divided into two functions: train and validate. These functions compute loss and accuracy metrics for both datasets, facilitating model evaluation. Loss is calculated using the multi-head loss function, which combines Cross-Entropy losses for gender, article type, and base color classifications, resulting in an averaged overall loss. Accuracy is determined by the percentage of correct predictions across all categories, calculated using the formula:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \times 100$$

During each epoch, the model undergoes forward and backward passes on both training and validation datasets, with metrics such as average loss and accuracy tracked. Upon completion, the model, optimizer, and loss criterion are saved, and a visualization of training and validation loss evolution is generated for further analysis.

D. *Testing Phase*

The testing phase involves evaluating the model's performance on unseen data. This begins with initializing necessary configurations and loading the model. Proper environment configuration ensures efficient image processing to optimize memory usage, while device specification determines the computation device for inference. The MultiHeadResNet50 model is instantiated and loaded with trained weights from a saved checkpoint. Image processing and model inference form the core of the testing pipeline. Input images undergo resizing and normalization to match model requirements before being fed into the model. The model generates outputs corresponding to gender, article type, and base color classifications. Predicted indices with the highest label scores are extracted, representing predicted classes for each category.

Post-processing and visualization are the final stages of the testing phase. Predicted indices are mapped back to their respective labels using pre-loaded dictionaries, providing human-readable labels for each category. These labels are then annotated onto the original image using OpenCV, offering a visual representation of the model's predictions. Annotated images are saved to disk for future reference and analysis.

IV. RESULTS

A. *Training and Validation Loss*

Over the course of 60 epochs, the training loss consistently decreases, signaling the model's effective learning from the training data. This decreasing trend indicates that the model is converging towards an optimal solution, improving its ability to minimize the difference between predicted and actual outputs. The validation loss generally decreased during the initial training epochs, indicating that the model was learning from the data. However, around epoch 20, the decrease plateaued, showing minimal fluctuations until epoch 60. This stabilization suggests that the model might not benefit significantly from further training beyond this point, potentially hinting at the onset of overfitting. Figure 2 provides visual insight into the phenomenon of the training and validation loss.

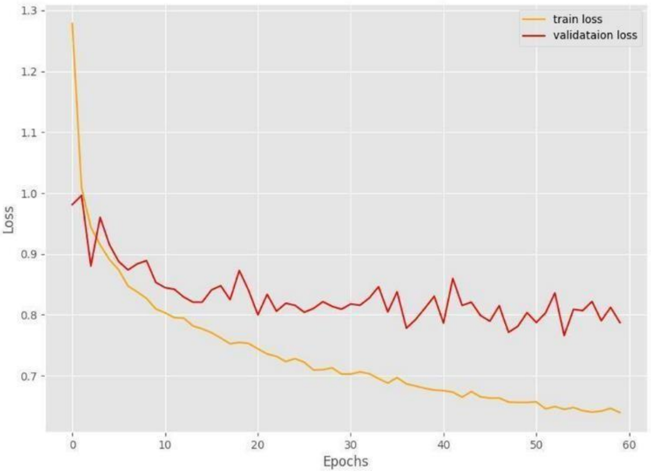


Figure 2. Training and validation loss learning curve

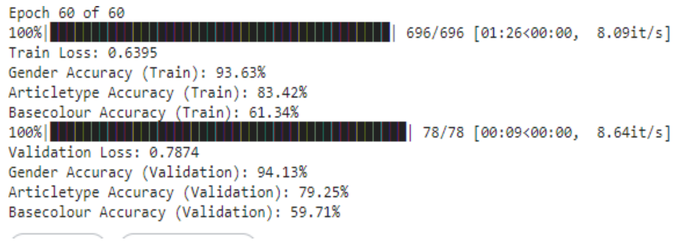
B. *Training and Validation Accuracy*

The accuracy of the model is calculated for both the training and validation datasets across 60 epochs. This metric is computed using the formula:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions\ for\ the\ particular\ label}{Total\ Number\ of\ Predictions\ for\ the\ particular\ label}$$

At the conclusion of the 60 epochs, the model exhibited impressive accuracy rates. For the training dataset, the accuracies were 93.63% for Gender, 83.42% for Article Type, and 61.34% for Base Color. Similarly, for the validation dataset, the accuracies were 94.13% for Gender, 79.25% for Article Type, and 59.71% for Base Color. Gender and Article Type labels achieved notably high accuracies, indicating the model's strong ability to classify these categories effectively.

Figure 3 depicts the accuracy for each label category across both the training and validation datasets at the 60th epoch, providing insights into the model's learning progress and its ability to correctly classify each label throughout the training process.



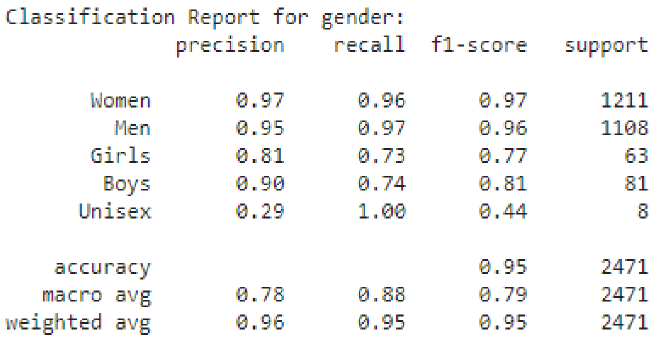
```
Epoch 60 of 60
100%|██████████████████████████████████████████| 696/696 [01:26<00:00,  8.09it/s]
Train Loss: 0.6395
Gender Accuracy (Train): 93.63%
Articletype Accuracy (Train): 83.42%
Basecolour Accuracy (Train): 61.34%
100%|████████████████████████████████████████████| 78/78 [00:09<00:00,  8.64it/s]
Validation Loss: 0.7874
Gender Accuracy (Validation): 94.13%
Articletype Accuracy (Validation): 79.25%
Basecolour Accuracy (Validation): 59.71%
```

Figure 3. Training and validation accuracies

C. *Classification Reports for Validation Dataset*

Classification reports were generated for the validation dataset to evaluate the model's performance across different label categories comprehensively. These reports offer detailed insights into the model's precision, recall, and F1-score for each label, enabling a thorough examination of its performance on a category-by-category basis. Tables 2, 3, and 4 present comprehensive evaluations of the model's performance across various classification tasks, providing a granular understanding of its capabilities within each category.

Table 2. Gender classification report

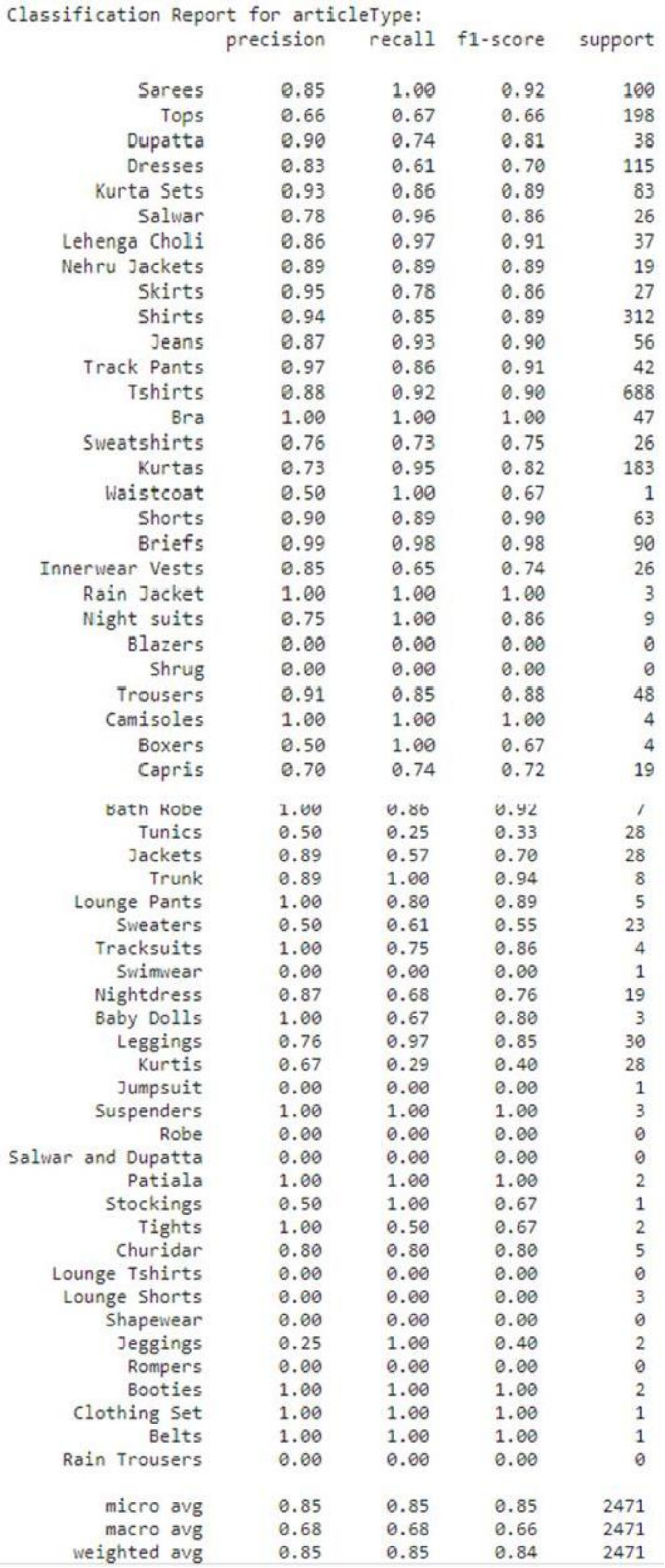


```
Classification Report for gender:
              precision    recall  f1-score   support

       Women       0.97      0.96      0.97      1211
         Men       0.95      0.97      0.96      1108
       Girls       0.81      0.73      0.77        63
        Boys       0.90      0.74      0.81        81
      Unisex       0.29      1.00      0.44         8

    accuracy                           0.95      2471
   macro avg       0.78      0.88      0.79      2471
weighted avg       0.96      0.95      0.95      2471
```

D. *Results on Unseen Images*

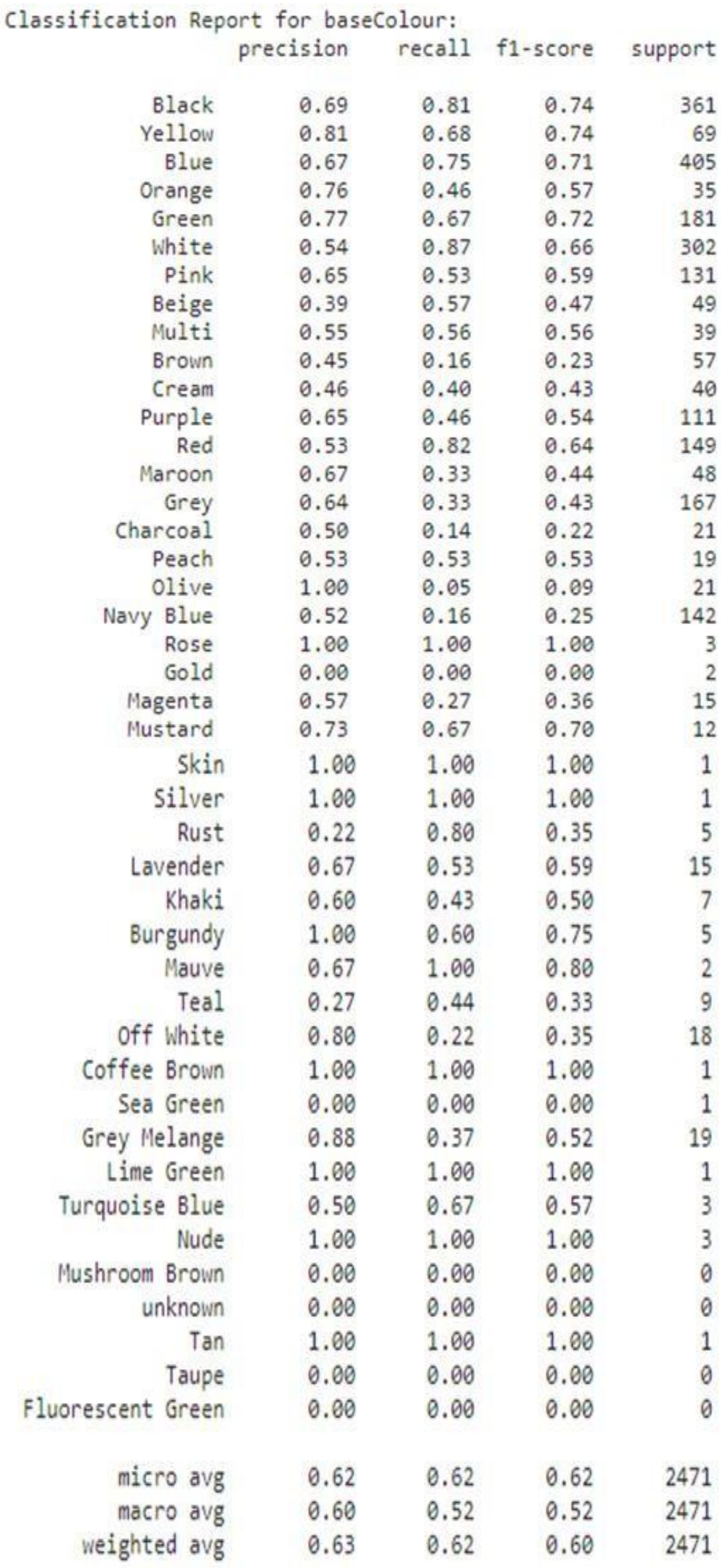
In this section, we present the results of our model's predictions on a diverse set of fashion images, as shown in Figures 4 through 9. Each image was processed by our multi-label classification model, which predicts three key attributes: gender, base color, and article type. The predicted labels for each image were then annotated directly onto the original images using OpenCV, providing a clear visual representation of the model's output.

Table 3. Article type classification report



```
Classification Report for articleType:
                   precision    recall  f1-score   support

           Sarees       0.85      1.00      0.92       100
             Tops       0.66      0.67      0.66       198
          Dupatta       0.90      0.74      0.81        38
          Dresses       0.83      0.61      0.70       115
       Kurta Sets       0.93      0.86      0.89        83
           Salwar       0.78      0.96      0.86        26
    Lehenga Choli       0.86      0.97      0.91        37
    Nehru Jackets       0.89      0.89      0.89        19
           Skirts       0.95      0.78      0.86        27
           Shirts       0.94      0.85      0.89       312
            Jeans       0.87      0.93      0.90        56
      Track Pants       0.97      0.86      0.91        42
          Tshirts       0.88      0.92      0.90       688
              Bra       1.00      1.00      1.00        47
      Sweatshirts       0.76      0.73      0.75        26
           Kurtas       0.73      0.95      0.82       183
        Waistcoat       0.50      1.00      0.67         1
           Shorts       0.90      0.89      0.90        63
           Briefs       0.99      0.98      0.98        90
  Innerwear Vests       0.85      0.65      0.74        26
      Rain Jacket       1.00      1.00      1.00         3
      Night suits       0.75      1.00      0.86         9
          Blazers       0.00      0.00      0.00         0
            Shrug       0.00      0.00      0.00         0
         Trousers       0.91      0.85      0.88        48
        Camisoles       1.00      1.00      1.00         4
           Boxers       0.50      1.00      0.67         4
           Capris       0.70      0.74      0.72        19

        Bath Robe       1.00      0.86      0.92         7
           Tunics       0.50      0.25      0.33        28
          Jackets       0.89      0.57      0.70        28
            Trunk       0.89      1.00      0.94         8
      Lounge Pants       1.00      0.80      0.89         5
         Sweaters       0.50      0.61      0.55        23
       Tracksuits       1.00      0.75      0.86         4
         Swimwear       0.00      0.00      0.00         1
       Nightdress       0.87      0.68      0.76        19
       Baby Dolls       1.00      0.67      0.80         3
         Leggings       0.76      0.97      0.85        30
           Kurtis       0.67      0.29      0.40        28
         Jumpsuit       0.00      0.00      0.00         1
       Suspenders       1.00      1.00      1.00         3
             Robe       0.00      0.00      0.00         0
Salwar and Dupatta       0.00      0.00      0.00         0
          Patiala       1.00      1.00      1.00         2
        Stockings       0.50      1.00      0.67         1
           Tights       1.00      0.50      0.67         2
         Churidar       0.80      0.80      0.80         5
    Lounge Tshirts       0.00      0.00      0.00         0
     Lounge Shorts       0.00      0.00      0.00         3
        Shapewear       0.00      0.00      0.00         0
         Jeggings       0.25      1.00      0.40         2
          Rompers       0.00      0.00      0.00         0
          Booties       1.00      1.00      1.00         2
     Clothing Set       1.00      1.00      1.00         1
            Belts       1.00      1.00      1.00         1
     Rain Trousers       0.00      0.00      0.00         0

        micro avg       0.85      0.85      0.85      2471
        macro avg       0.68      0.68      0.66      2471
     weighted avg       0.85      0.85      0.84      2471
```

Table 4. Base color classification report



```
Classification Report for baseColour:
                    precision    recall  f1-score   support

           Black       0.69      0.81      0.74       361
          Yellow       0.81      0.68      0.74        69
            Blue       0.67      0.75      0.71       405
          Orange       0.76      0.46      0.57        35
           Green       0.77      0.67      0.72       181
           White       0.54      0.87      0.66       302
            Pink       0.65      0.53      0.59       131
           Beige       0.39      0.57      0.47        49
           Multi       0.55      0.56      0.56        39
           Brown       0.45      0.16      0.23        57
           Cream       0.46      0.40      0.43        40
          Purple       0.65      0.46      0.54       111
             Red       0.53      0.82      0.64       149
          Maroon       0.67      0.33      0.44        48
            Grey       0.64      0.33      0.43       167
        Charcoal       0.50      0.14      0.22        21
           Peach       0.53      0.53      0.53        19
           Olive       1.00      0.05      0.09        21
       Navy Blue       0.52      0.16      0.25       142
            Rose       1.00      1.00      1.00         3
            Gold       0.00      0.00      0.00         2
         Magenta       0.57      0.27      0.36        15
         Mustard       0.73      0.67      0.70        12
            Skin       1.00      1.00      1.00         1
          Silver       1.00      1.00      1.00         1
            Rust       0.22      0.80      0.35         5
        Lavender       0.67      0.53      0.59        15
           Khaki       0.60      0.43      0.50         7
        Burgundy       1.00      0.60      0.75         5
           Mauve       0.67      1.00      0.80         2
            Teal       0.27      0.44      0.33         9
       Off White       0.80      0.22      0.35        18
    Coffee Brown       1.00      1.00      1.00         1
       Sea Green       0.00      0.00      0.00         1
    Grey Melange       0.88      0.37      0.52        19
      Lime Green       1.00      1.00      1.00         1
  Turquoise Blue       0.50      0.67      0.57         3
            Nude       1.00      1.00      1.00         3
  Mushroom Brown       0.00      0.00      0.00         0
         unknown       0.00      0.00      0.00         0
             Tan       1.00      1.00      1.00         1
           Taupe       0.00      0.00      0.00         0
Fluorescent Green       0.00      0.00      0.00         0

       micro avg       0.62      0.62      0.62      2471
       macro avg       0.60      0.52      0.52      2471
    weighted avg       0.63      0.62      0.60      2471
```

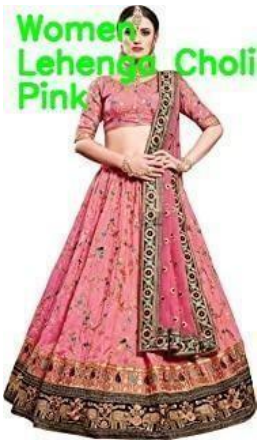


Figure 4. Image of a woman wearing a pink color lehenga choli

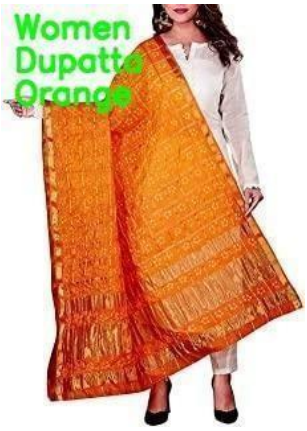


Figure 5. Image of a woman wearing an orange color dupatta

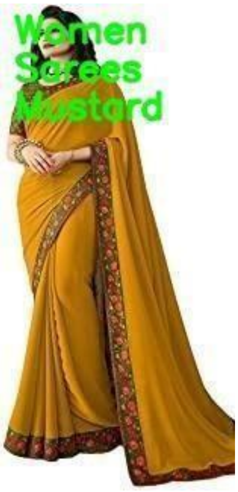


Figure 6. Image of a woman wearing a mustard color saree

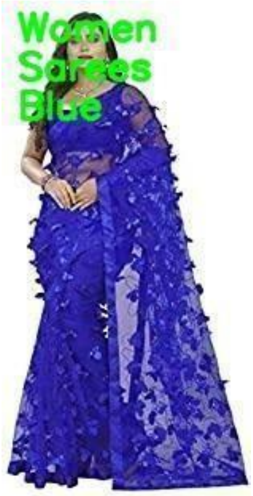


Figure 7. Image of a woman wearing a blue color saree

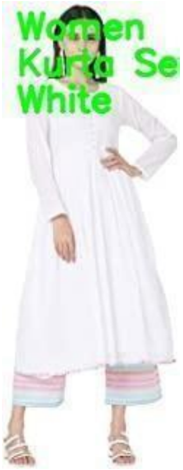


Figure 8. Image of a woman wearing a white color kurta set

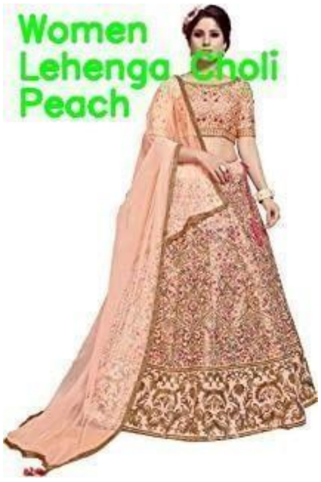


Figure 9. Image of a woman wearing a peach color lehenga choli

V. DISCUSSION AND CONCLUSION

This paper focused on developing and evaluating a Multi-Head ResNet50 model tailored for multi-label classification in the Asian fashion domain, specifically predicting gender, article type, and base color labels. The model demonstrated commendable performance in capturing the nuanced characteristics of fashion items prevalent in Asian markets, achieving high accuracy for gender and article type predictions. Specifically, our model attained an epoch-wise accuracy of 93.63% for gender and 83.42% for article type on the training dataset, which is particularly noteworthy considering the complexity and variety of the fashion items in the dataset. Validation results were similarly impressive, with accuracies of 94.13% for gender and 79.25% for article type, indicating strong generalization capabilities and minimal signs of overfitting.

However, the accuracy for base color on the training dataset was 61.34%, and 59.71% on the validation dataset, which is relatively low compared to other attributes. This discrepancy highlights a significant challenge in predicting base color accurately. The lower accuracy for base color can be attributed to several factors, including the high variability and subtlety of color shades in fashion items, especially in the context of Asian fashion where intricate and nuanced color palettes are common. Additionally, the dataset's inherent imbalance, with fewer instances of certain base colors, may have contributed to the lower performance in this category.

To address the issue of lower accuracy for base color, several strategies could be explored. First, implementing techniques to mitigate class imbalance, such as oversampling underrepresented classes, undersampling overrepresented ones, or employing weighted loss functions, could help the model learn more effectively from the available data. Additionally, enhancing the feature extraction process by incorporating more sophisticated color recognition techniques or leveraging additional data sources for more comprehensive color annotation might improve the model's accuracy. Incorporating advanced augmentation techniques specifically designed to highlight color variations could also prove beneficial.

In conclusion, we introduced and implemented a Multi-Head ResNet50 model tailored for multi-label classification within the fashion domain, specifically targeting gender, article type, and base color. The model's robust performance in predicting gender and article type underscores its potential in the fashion industry for categorizing diverse fashion items with high accuracy. However, the relatively lower accuracy for base color suggests areas for further improvement. Addressing these challenges through targeted strategies such as handling class imbalance and refining feature extraction techniques could enhance the model's overall performance and applicability. Future work will focus on these areas to develop more inclusive and accurate fashion image captioning systems, ensuring better representation and recognition of diverse fashion styles.

VI. FUTURE RESEARCH DIRECTIONS

This study highlights several avenues for future research to enhance model performance and applicability. Firstly, refining techniques for base color prediction and expanding the dataset with more diverse color samples could address the current lower accuracy in this area. Secondly, implementing methods to handle class imbalance, such as oversampling, undersampling, or weighted loss functions, can improve accuracy across all attributes. Additionally, incorporating multimodal approaches by integrating image data with textual descriptions and contextual information can enhance the richness and precision of fashion item captions. Exploring transfer learning and domain adaptation can help the model generalize better across various fashion styles, extending its applicability beyond Asian fashion. Expanding research to include underrepresented fashion categories, such as African or Latin American styles, will promote diversity and inclusivity in fashion image captioning systems. Lastly, addressing ethical implications by ensuring models are trained on diverse, representative datasets is crucial for developing fair and equitable AI systems for a global audience.

REFERENCES

[1] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion captioning: Towards generating accurate descriptions with semantic rewards," in Proceedings of the 2020 European Conference on Computer Vision (ECCV), Aug. 2020,
pp. 1-17.

[2] N. Moratelli, M. Barraco, D. Morelli, M. Cornia, L. Baraldi, and
R. Cucchiara, "Fashion-Oriented Image Captioning with External Knowledge Retrieval and Fully Attentive Gates," Sensors (Basel), vol. 23, no. 3, p. 1286, 2023, doi: 10.3390/s23031286.

[3] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "Be Your Own Prada: Fashion Synthesis with Structural Coherence," in Proceedings of the International Conference on Computer Vision (ICCV), Oct. 2017.

[4] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan, "Style Finder: Fine-Grained Clothing Style Detection and Retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 11, pp. 2670-2683, November 2013, doi: 10.1109/TPAMI.2013.78.

[5] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," International Journal of Computer Vision, vol. 124, no. 1, pp. 74-95, November 2016, doi: 10.1007/s11263-016-0932-3.

[6] H. Chen, A. Gallagher, and B. Girod, "Describing Clothing by Semantic Attributes," in Proceedings of the ACM Multimedia Conference, October 2019.

[7] B. T. Nguyen, O. Prakash, and A. H. Vo, "Attention Mechanism for Fashion Image Captioning," IEEE Transactions on Multimedia, vol. 23, no. 5, pp. 1567-1580, May 2021, doi: 10.1109/TMM.2021.3069988.

[8] X. Colombo, "Transfer Learning Analysis of Fashion Image Captioning Systems," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021.

[9] G. Hacheme and N. Sayouti, "Neural Fashion Image Captioning: Accounting for Data Diversity," arXiv preprint arXiv:2106.12154v2, Jun. 2021.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 3156–3164.