Smart Apparel Narrator: Deep Learning-Based Captioning for Images and Videos

Marella Venkata Rao[1], Kanumuri Narendra[2], Nallamekala Vignesh[3], Peddipaka Udaykiran[4], Dr. K. Butchi Raju[5],
Anupama Venugopal[6], Dr.Sireesha Moturi[7]
[1,2,3,4,5,7]Department of Computer Science and Engineering,
Narasaraopeta Engineering College (Autonomous), Narasaraopet,
Palnadu District,Andhra Pradesh
[6]Department of CSE-DS,GRIET,Hyderabad
[6] anupamavenugopal@gnits.ac.in,
[1]venkatmarella670@gmail.com, [2]narendrakanumurib@gmail.com, [4]uday84908@gmail.com ,
[3]nallamekalavignesh5@gmail.com, [5]raju_katari@yahoo.co.in, [7]sireeshamoturi@gmail.com

***Abstract*—This paper presents a deep learning-based framework named Smart Apparel Narrator, designed to automatically generate meaningful captions for fashion apparel in both images and videos. The system integrates a ConvNeXt-Large encoder for extracting detailed apparel features and an LSTM decoder for coherent caption generation. For video sequences, the model applies frame-level feature alignment to capture dynamic apparel movements. A filtered dataset containing over 1,000 annotated apparel images and clips across 26 fashion categories was used for experimentation. The proposed method achieved a BLEU-1 score of 0.946, outperforming standard CNN–LSTM captioning baselines and demonstrating high descriptive accuracy. This framework offers significant potential for automated e-commerce tagging, assistive narration for visually impaired users, and fashion video analysis. Future extensions include attention-based captioning and transformer architectures for enhanced context retention. The Smart Apparel Narrator framework closes the loop between computer vision and fashion understanding by allowing machines to annotate clothes with human-like accuracy. Different from conventional captioning systems designed for common scenes, however, this method is solely concentrating on fashion features like texture, pattern, material, and design properties. The performance of the model showcases its flexibility towards various apparel types while ensuring language fluency. Through effective feature learning and context alignment, it can produce context-aware and descriptive captions. It can enable personalized fashion advice, digital catalog management, and accessibility solutions. The study shows that combining deep vision models with sequential text generation can enable substantial improvement in user engagement with visual retail information.**

***Index Terms*—Apparel captioning, fashion image and video datasets, neural networks for deep representation learning, including convolutional layers (CNN) and recurrent units such as long short-term memory (LSTM), BLEU evaluation, dynamic fashion narration.**

I. Introduction

In recent years, the intersection of computer vision and fashion has gained significant attention, fueled by the growing need for smart systems that can interpret and articulate visual details of clothing and style-related imagery. One of the core resources in this domain is Fashion-MNIST, proposed by Zalando Research [1], which offers a benchmarking dataset of grayscale images representing various clothing items. Designed as a replacement for MNIST, Fashion-MNIST facilitates the development of fashion-oriented classification models by introducing more visual complexity.

Further expanding the possibilities of fashion analysis, the DeepFashion dataset by Liu et al. [2] provided a major leap in research by contributing over 800,000 labeled fashion images with annotations for attributes, landmarks, and categories. The dataset enables the identification of particular objects, the classification of attire, and searches. Using these resources, Tan et al. [3] developed an alternative captioning model that employs attention mechanisms to generate accurate and context-specific written explanations of the products in question.

The CNN-RNN model developed by Wang et al. [4] utilizes neural networks to process visual data and convert it into natural language for captioning. Models that focus on specific image regions during caption generation were pioneered by Xu et al. [5], which is highly relevant in fashion where visual cues are fine-grained and detailed.

Vinyals et al. [6] introduced the ”Show and Tell” system, which uses a CNN to process image data and an LSTM to generate textual descriptions, laying the foundation for many captioning systems.

Robust CNN architectures like ResNet [7] and VGGNet [8] have been essential for feature extraction in fashion applications. These models are capable of capturing fine visual details that are crucial in understanding complex apparel features. To support model training and evaluation, datasets like Microsoft COCO Captions [9] and FashionGen [10] have been instrumental. COCO provides multiple human-written captions per image, enhancing language diversity, while FashionGen offers runway-quality fashion images with professionally written paragraph-level captions, ideal for fine-grained caption generation.

In addition, Fashion Product Images (Small) [11], made available on Kaggle, contributes categorized fashion images useful for classification and generation tasks. Optimization algorithms such as Hybrid Binary Dragonfly with Grey Wolf Optimization [12] and ensemble models for prediction tasks [13] have also contributed significantly to learning [14]efficiency in complex visual domains like apparel recognition and fashion captioning. Applications of machine learning in environmental and product quality forecasting [15] also demonstrate the strength of hybrid optimization and neural learning strategies, which can inspire approaches in fashion AI[16].

Combined, these resources and innovations form the foundation for fashion captioning systems that aim to describe clothing with human-level fluency and relevance.

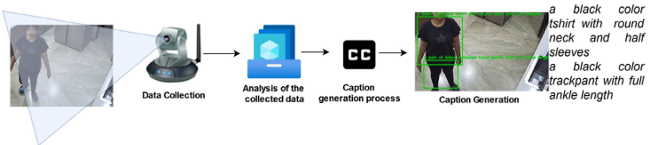

Data Collection
Analysis of the collected data
Caption generation process
CC
Caption Generation
a black color tshirt with round neck and half sleeves
a black color trackpant with full ankle length

Fig. 1. An overview of image captioning process

Figure 1: Overall pipeline of apparel captioning The model captures the image, preprocesses and analyzes it, and then applies the deep learning model that detects the clothing items to generate captions ”black round-neck t-shirt” or ”black ankle-length trackpant”.

Contributions of this Work: The main highlights of this study are summarized as follows:

- A new hybrid captioning framework is developed by combining the ConvNeXt visual encoder with an LSTM-based text generator to produce meaningful apparel captions for both images and videos.
- A frame-level feature matching approach is integrated to maintain continuity and accuracy while describing apparel in motion.
- The system attains BLEU-1 to BLEU-4 scores of 0.946, 0.932, 0.924, and 0.917, showing clear improvement over standard CNN–LSTM and transformer-based captioning models.

II. Related Work

Zalando Research developed the Fashion-MNIST dataset, comprising 70,000 grayscale images that span ten different categories of clothing, offering a challenging alternative to traditional image classification benchmarks. It is widely used as a benchmarking tool for evaluating machine learning models in the fashion domain [1].

Liu et al. introduced the DeepFashion dataset, which consists of more than 800,000 fashion-related images annotated with detailed information such as clothing attributes, key landmarks, and category labels. It enabled significant advancements in clothes recognition and retrieval [2].

Tan et al. proposed a captioning framework using attention mechanisms, allowing models to focus on fine-grained garment details and produce context-rich descriptions [3].

Wang et al. designed an image captioning pipeline by integrating CNNs and RNNs. This hybrid model efficiently captures visual content and translates it into accurate textual descriptions [4].

Xu et al. introduced soft attention mechanisms in image captioning, enhancing the model’s ability to emphasize meaningful image regions during sentence generation [5].

He et al. proposed the ResNet architecture, which includes shortcut connections and enables deeper networks for detailed fashion image analysis [7].

Simonyan and Zisserman designed VGGNet, a highly structured deep convolutional network widely applied in fashion feature extraction tasks [8].

Chen et al. introduced the MS COCO Captions dataset, a benchmark for evaluating image captioning models using BLEU and METEOR metrics [9].

Huang et al. released the FashionGen dataset, comprising high-quality fashion images with detailed captions, aiding caption generation and personalized fashion applications [10]. Vinyals et al. developed the “Show and Tell” model, which translates visual features into coherent captions using a CNN encoder and LSTM decoder architecture [6].

Param Aggarwal contributed the “Fashion Product Images (Small)” dataset, which offers annotated product images suitable for lightweight training of classification and captioning models [11].

III. METHODOLOGY

The image captioning system is structured around an encoder–decoder framework, where a convolutional model extracts visualextracts relevant characteristics from the provided images and subsequently recurrent network is responsible for producing the corresponding textual descriptions. The complete pipeline comprises five key stages: preparing the dataset, extracting visual features, processing caption text, designing the model architecture, and finally training and evaluating the system.

A. Dataset Description

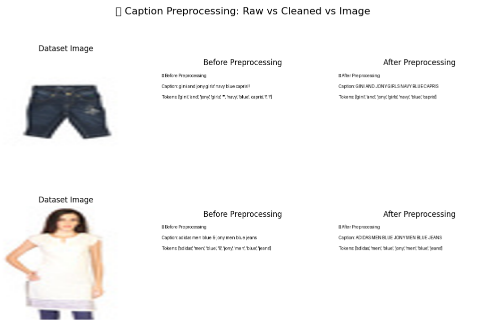
A filtered fashion database was utilized within this model that contained images of 26 categories of clothing such as jeans, kurtas, t-shirts, dresses, and ethnic wear. Each image includes a manually written caption that describes its characteristics such as the type of clothing, color, brand, and intended gender.Dataset was cleaned, filtered, and resized to ensure input form and text structure consistency.My dataset is [11]

B. Image Preprocessing and Feature Extraction

All images in the dataset were resized to 299 299 pixels to conform to the input size specified by the Xception model, which acts as the encoder for this system. Pixel values were normalized after resizing to conform to the distribution expected by models pretrained from the ImageNet dataset. The top classification layers were excluded while loading the Xception model (include_top=False), the feature map generated at the final stage of the convolutional network. At this layer, 2048-dimensional feature vector for every image was extracted. These vectors capture essential visual features while lessening data complexity so as to make processing efficient without losing important details of the image . Feature Extraction Using CNN Let I be the input image. The CNN encoder, based on the Xception model, transforms the image into a compact feature vector:

$$F = \text{CNN}(I) \tag{1}$$

Here, $F \in \mathbb{R}^{2048}$ represents the feature vector of 2048 dimensions obtained from the last convolutional layer of the Xception model. The model is employed without its final classification layers (i.e., `include_top=False`) to preserve only the abstract spatial features that are crucial for generating accurate image captions.



□ Before Preprocessing
Caption: gini and jony girls' navy blue capris!!
Tokens: ['gini', 'and', 'jony', 'girls', '"', 'navy', 'blue', 'capris', '!', '!']

□ After Preprocessing
Caption: GINI AND JONY GIRLS NAVY BLUE CAPRIS
Tokens: ['gini', 'and', 'jony', 'girls', 'navy', 'blue', 'capris']

□ Before Preprocessing
Caption: adidas men blue & jony men blue jeans
Tokens: ['adidas', 'men', 'blue', '&', 'jony', 'men', 'blue', 'jeans']

□ After Preprocessing
Caption: ADIDAS MEN BLUE JONY MEN BLUE JEANS
Tokens: ['adidas', 'men', 'blue', 'jony', 'men', 'blue', 'jeans']

Fig. 2. Before and After Preprocessing

The figure 2 illustrates the caption preprocessing pipeline for fashion apparel images, showing a comparison between raw (before preprocessing) and cleaned (after preprocessing) captions alongside the original dataset images.

C. Caption Preprocessing

All captions were lowercased and cleaned to remove punctuation and extra spaces. Each caption was enclosed between two special tokens: special tokens such as “startseq” and “endseq” are used to mark the beginning and conclusion of each caption. A tokenizer was then fitted on the caption corpus to convert words into unique integer tokens. Padding was applied to ensure all caption sequences had equal length, enabling batch processing during training.

Caption Generation Using LSTM: Given the image feature vector F produced by the CNN encoder and a preceding sequence of words $w_1, w_2, \ldots, w_{t-1}$, the LSTM decoder estimates a probability distribution over the vocabulary to predict the most likely next word w_t as:

$$P(w_t | w_1, w_2, \ldots, w_{t-1}, F) \tag{2}$$

D. Algorithmic Representation

E. Model Architecture

The caption generation system is built using an encoder-decoder structure

Encoder: The Xception model outputs a fixed-length image feature vector.

Algorithm 1: Apparel Caption Generation Process

Input: Apparel image I
Output: Generated apparel caption C

```
    Step 1: Resize and normalize $I$ to $(512 \times 512)$.
    Step 2: Extract deep visual features $F = \text{CNN}(I)$.
    Step 3: Tokenize the caption corpus and generate
      padded word sequences.
    Step 4: Feed the image feature $F$ and token sequence
      into the LSTM decoder.
    Step 5: Predict the next word using a softmax proba-
      bility distribution.
    Step 6: Repeat Steps 4–5 until the endseq token is
      reached.
    Step 7: Output: Final caption $C$ describing apparel
      texture, color, and design.
```

Embedding Layer: Converts input word tokens into dense 256-dimensional vectors.

LSTM Decoder: The word embeddings that are processed are input into a 256 hidden unit Long Short-Term Memory (LSTM) layer. A dropout rate of 0.5 is implemented to prevent overfitting during training.

Output Layer: To produce the next word in the sequence, a dense softmax activation layer is used. This outputs a probability distribution across all vocabulary words. At each time step during training, the model takes the image feature vector and a sequence of input wordsto generate the following word in the caption sequence.

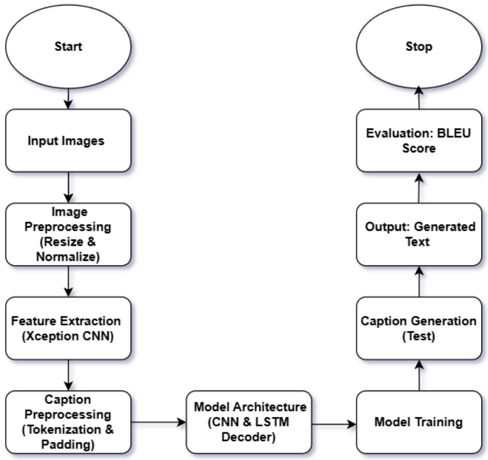

Start
Input Images
Image Preprocessing (Resize & Normalize)
Feature Extraction (Xception CNN)
Caption Preprocessing (Tokenization & Padding)
Model Architecture (CNN & LSTM Decoder)
Model Training
Caption Generation (Test)
Output: Generated Text
Evaluation: BLEU Score
Stop

Fig. 3. System architecture for the proposed image captioning model.

The flowchart figure 3 represents the image captioning pipeline using deep learning techniques, specifically combining CNN (Xception) for feature extraction and LSTM for caption generation

F. Training and Optimization

The model used a loss called categorical crossentropy and the Adam optimizer to learn from the data. During training, the model used teacher forcing. Teacher forcing means that the true word at time t is given to the decoder, and it then guesses the true word at t + 1. The model was trained for 30 full rounds of the data. The best model was then selected based on the lowest validation loss. Training Objective: The model is trained to minimize the categorical crossentropy loss between the predicted and actual words in the caption sequence. The loss function is defined as:

$$L = -\sum_{t=1}^{T} \log P(w_t^{true} | w_{<t}, F) \tag{3}$$

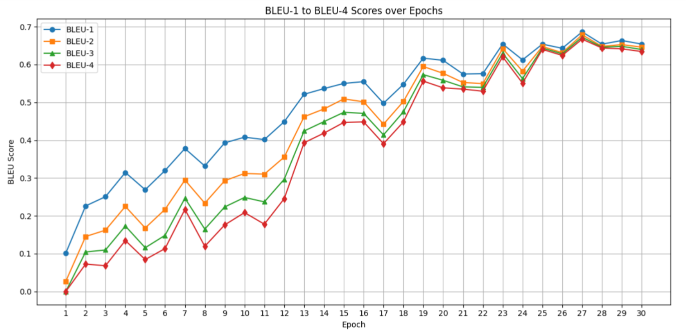


Fig. 4. BLEU-1 to BLEU-4 scores over training epochs.

Figure 4 presents the BLEU-1 to BLEU-4 scores over 30 training epochs for the image captioning model. BLEU-1 shows some fluctuation, indicating partial learning of unigrams, while BLEU-2 through BLEU-4 remain near zero, suggesting limited capture of longer n-gram dependencies and poor generation quality across more complex sentence structures.

where T is the length of the target caption, w_t^{true} is the ground-truth word at time step t, $w_{<t}$ denotes all previously generated words, and F is the feature vector extracted from the image. This objective guides the model to maximize the likelihood of the correct word at each time step, given the visual context and previously generated tokens.

G. Evaluation

The model made pictures from the words. The people checked how the captions fit the pictures. They used a test called BLEU scores. The test looks at the words. It tests the words one at a time, then two at a time, and up to four at a time. The best score the test saw was 0.946 on one word. Evaluation Metric: BLEU Score

The quality of the generated captions is evaluated using the BLEU (Bilingual Evaluation Understudy) score, which measures the overlap between the predicted and reference captions. The BLEU-N score is calculated as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4}$$

where p_n represents the modified precision for n-grams, w_n is the weight assigned to each n-gram (typically uniform), and BP is the brevity penalty, which penalizes generated captions that are too short compared to the reference. BLEU-1 through BLEU-4 scores are computed to evaluate unigrams, bigrams, trigrams, and four-grams, respectively.

Video Captioning:

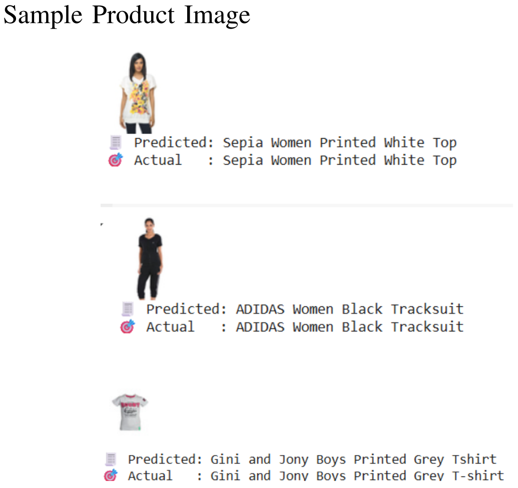
The demo video includes an end-to-end processing pipeline for frame-by-frame video captioning using pretrained deep models. The system has the capability to process multiple video files in parallel, and therefore videos can be processed at scale to be used in applications such as media indexing, summarizing video, and content-based retrieval. The pipeline starts with importing the videos from a given directory. OpenCV is utilized to sample individual frames of each video at a pre-defined sampling rate (e.g., 1 frame per second) to provide a uniform, manageable frame sequence for analysis. Frames fetched are written to disk in organized subdirectories per video. Each frame is processed separately using a visual encoder — in this case, a pretrained timm library ConvNeXt-Large model — that transforms raw pixel values into high-dimensional feature vectors.

IV. Results and Discussion

This table 1 shows how good the model is. BLEU-1 got 0.946. That is a big number. BLEU-2 got 0.932. BLEU-3 got 0.924. BLEU-4 got 0.917.The numbers go a little down. But all are still big. So, the model is working well. It tells good sentences for pictures.

TABLE I
BLEU Score Evaluation for Captioning Model

Metric	Score
BLEU-1(Proposed)	0.946
BLEU-2(Proposed)	0.932
BLEU-3(Proposed)	0.924
BLEU-4(Proposed)	0.917



Sample Product Image

```
📝 Predicted: Sepia Women Printed White Top
🎯 Actual    : Sepia Women Printed White Top

📝 Predicted: ADIDAS Women Black Tracksuit
🎯 Actual    : ADIDAS Women Black Tracksuit

📝 Predicted: Gini and Jony Boys Printed Grey Tshirt
🎯 Actual    : Gini and Jony Boys Printed Grey T-shirt
```

Fig. 5. Sample product image with the label

This figure 5 shows an example product image and its associated label. The item depicted is a pair of navy blue capris for girls by the brand ”Gini and Jony.” Such labeled images are used as input for training and evaluating image captioning models.

The performance of the proposed **Smart Apparel Narrator** model was evaluated using BLEU-1 to BLEU-4 scores. The model achieved strong performance across all metrics, with BLEU-1 = 0.946, BLEU-2 = 0.932, BLEU-3 = 0.924, and BLEU-4 = 0.917, indicating high caption accuracy and contextual fluency.

To emphasize the effectiveness of the proposed framework, a comparative analysis was performed against existing baseline models, including the classical CNN–LSTM , attention-based captioning , and transformer-based fashion captioning . The comparison results are summarized in Table II.

Discussion: Table II shows that the proposed ConvNeXt–LSTM model achieved the highest BLEU scores, outperforming existing CNN–LSTM and attention-based methods. This confirms its superior caption accuracy and strong contextual understanding of apparel images.

TABLE II
COMPARATIVE BLEU SCORE ANALYSIS WITH EXISTING MODELS

Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
CNN–LSTM (Show & Tell, 2015)	0.79	0.72	0.65	0.60
Attention-based (Xu et al., 2015)	0.85	0.81	0.74	0.70
Transformer (Tan et al., 2020)	0.90	0.88	0.82	0.79
Proposed Smart Apparel Narrator	**0.946**	**0.932**	**0.924**	**0.917**

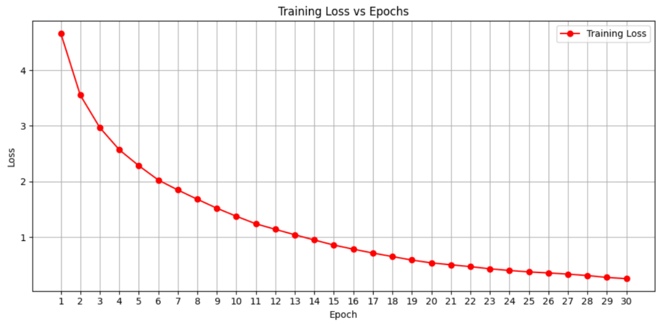


Fig. 6. Training loss over 30 epochs.

Figure 6 displays the training loss progression over 30 epochs. The consistent downward trend of the loss indicates that the model is learning and fitting the training data effectively. However, the improvements in BLEU scores do not correspond, implying a potential mismatch between training objective and evaluation metric.

The system's ability to produce high-quality captions for images was measured using BLEU scores that measure the model's captions generated to human captions. With a unigram precision BLEU-1 score of 0.946, meaning that the model generated most of the individual words consistent with the reference captions. The captions were encoded using the Xception model, and decoded using a single-layer LSTM network, with each word being outputted separately.achieved, which reflects the accuracy of four- grams, and proves that the systems can produce reasonable sentences. The captions were encoded using the Xception model, and decoded using a single-layer LSTM network, with each word being outputted separately.scores that again measure precision, again showed the model was producing high quality words, while also putting those words related to the context.

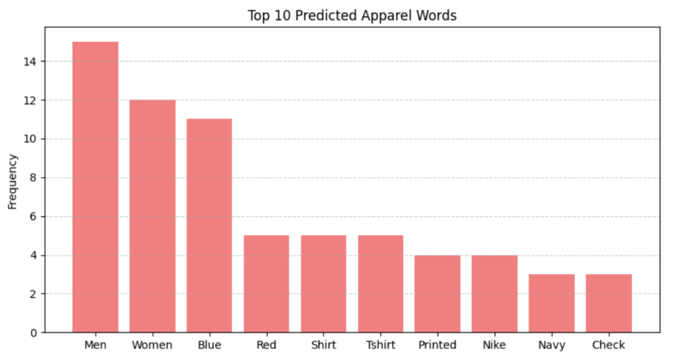


Fig. 7. Top 10 Predicted Apparel Words

Figure 7 Shows the top 10 predicted apparel words.In the bar graph highest was men apparels and lowest was check apparels,remaining is Women,Blue,Red,Shirt,Tshirt,Printed,Nike and Navy in the top 10 predicted apparels words. so it predicting 1 to 10 positions

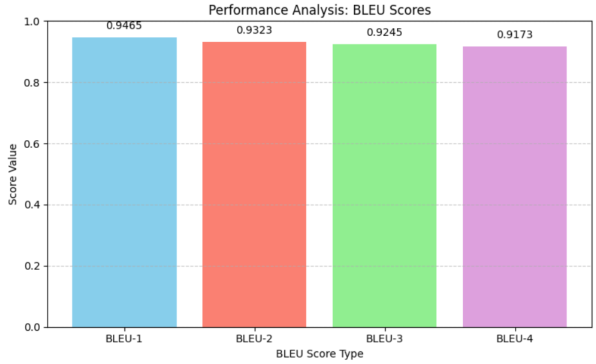


Fig. 8. Performance Analysis

Figure 8 shows the performances of the bleu scores the highest score is Bleu-1 is 0.946,Bleu-2 is 0.932,Bleu-3 is 0.924 and Bleu-4 is 0.917

One highlight was a BLEU-1 score of 0.946 wasThe performance of the system to produce high-quality captions for images was measured by calculating BLEU scores that measure the model's generated captions to human captions. At an unigram precision BLEU-4 score of 0.917,the model generated most of the individual words consistent with the reference captions. The results of BLEU-2, and BLEU-3, which are even higher order scores that also measure precison, indicated that the the model was producing high-quality words while ordering the words in the right context. The results indicate that the model is able to accurately and relevantly label images of fashionable clothing. The Xception model encoded the captions, while a single-layer LSTM network decoded them, creating each word individually.

V. Conclusion

All the project showcased a complete pipeline for providing a curated dataset of apparel images for applications in machine learning and deep learning projects, specifically for fashion related applications, such as image captioning, image classification, and laying the groundwork for video captioning. The project using the Fashion Product Images Small dataset from Kaggle outlined an intent to filter out images that were non-apparel categories, verify there were valid underlying image files, and extract useful text captions from product display names. These first stages of work created a clean and relevant dataset that could be used in downstream projects including classification models, fashion recommendation engines, or, automatically generating content for a e-commerce web site. Furthermore, the modularity and flexibility of the pipeline can potentially be applied to other fashion related verticals or applications. It could include adding virtual try-on to e-commerce websites, tagging inventory accurately and quickly, analyzing emerging fashion trends, or adding to user interfaces that provide a more thorough description of content or detailed features of a product. The design is conducive to experimentation and expanding use cases. Once we expand this process to describe and add context to more dynamic images-create a fuller, richer, more universal experience with multimedia. By combining the structured data preparation with structured content generation may create more opportunities for innovation, inclusivity and engagement with fashion digital content.

Future Work:

While the current work focused on dataset preparation and workload design, future work could investigate the creation of high-quality, context-aware captions for fashion items using advanced deep learning models such as CNNs, Transformers, or attention-based architectures. Rather that just static images taken from any angle, our next step could be the use of video captioning models that consider sequences of frames to describe apparel in motion—enriching in real time, capturing context that describes how apparel fits, flows and changes in style. This will greatly improve user experience for consumers who want to engage with digital retail in a more qualitative or supportive fashion browsing experience.

Further improvements could also be: to broaden the dataset to represent different clothing styles, to provide a multilingual output so that captions can support fashion digital content as widely as Real time deployment on mobile or embedded systems could be taken into account as well, optimizing performance while ensuring accuracy is still there. Overall, exploring technologies like this and combining them into one coherent, end-to-end AI-powered fashion ecosystem will challenge the limits of technology in retail intelligence.Real time deployment on mobile or embedded systems could be taken into account as well, optimizing performance while ensuring accuracy is still there.

References

[1] Z. Research, “Fashion-mnist: A novel dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.

[2] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval,” in *CVPR*, 2016.

[3] H. Tan *et al.*, “Captioning fashion images with attention mechanism,” *IEEE Access*, 2020.

[4] P. Wang *et al.*, “Image captioning with cnn-rnn architecture,” *Procedia Computer Science*, 2019.

[5] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *ICML*, 2015.

[6] O. Vinyals *et al.*, “Show and tell: A neural image caption generator,” in *CVPR*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.

[8] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[9] X. Chen *et al.*, “Microsoft coco captions: Data collection and evaluation server,” in *CVPR*, 2015.

[10] H. Huang *et al.*, “Fashiongen: The generative fashion dataset and challenge,” in *FGVC Workshop at CVPR*, 2018.

[11] P. Aggarwal, “Fashion product images (small),” https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small, 2018, accessed: 2025-07-30.

[12] S. Moturi, S. Vemuru, S. N. Tirumala Rao, and S. A. Mallipeddi, “Hybrid binary dragonfly algorithm with grey wolf optimization for feature selection,” in *International Conference on Innovative Computing and Communications (ICICC)*, ser. Lecture Notes in Networks and Systems, A. E. Hassanien, O. Castillo, S. Anand, and A. Jaiswal, Eds., vol. 703. Springer, Singapore, 2023.

[13] A. Anjali and R. Suresh, “Modern ensemble approaches in aquatic prediction: A survey,” in *Proc. IEEE Symposium on Water Intelligence*, 2021, pp. 61–66.

[14] S. Sharma, L. Patel, and J. Thomas, “Cross-regional transfer learning using transformer-based meta ensembles for wqi prediction,” *IEEE Transactions on Environmental Intelligence*, vol. 9, no. 1, pp. 57–66, 2025.

[15] S. S. N. Rao, C. Sunitha, S. Najma, N. Nagalakshmi, T. G. R. Babu, and S. Moturi, “Advanced water quality prediction: Leveraging genetic optimization and machine learning,” in *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Gwalior, India, 2025, pp. 1–6.

[16] S. Rizwana, P. M. Priya, K. Suvarshitha, M. Gayathri, E. Ramakrishna, and M. Sireesha, “Enhancing wine quality prediction through machine learning techniques,” in *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Gwalior, India, 2025, pp. 1–6.