Received 17 May 2024, accepted 10 August 2024, date of publication 14 August 2024, date of current version 26 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3443422

RESEARCH ARTICLE

Image and Video Captioning for Apparels Using Deep Learning

GOVIND AGARWAL[1], KRITIKA JINDAL[1], ABISHI CHOWDHURY[1], (Member, IEEE), VISHAL K. SINGH[2], (Member, IEEE), AND AMRIT PAL[1], (Member, IEEE)

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India
[2]School of Computer Science and Electronics Engineering, University of Essex, Colchester Campus, CO4 3SQ Colchester, U.K.

Corresponding author: Amrit Pal (amrit.pal@vit.ac.in)

ABSTRACT In the rapidly evolving world of apparel, writing clear and interesting product descriptions is crucial to attract customers. In light of the importance of automated descriptions for apparel, this work explores the field of image captioning for apparel photos and expands its use to include captioning videos to enable visually impaired people to access and understand dynamic apparel content. To address the issue of diversity in datasets, we curated a collection of images that were divided into 26 classifications. With the use of Convolutional Neural Network (CNN) architectures like ConvNeXtLarge and Long Short-Term Memory (LSTM) architectures, our suggested system can automatically provide accurate and captivating captions for both still photos and moving videos that feature clothing. The LSTM network smoothly blends the visual data extracted by the CNN component from clothing photos and videos to produce captions that are both semantically and linguistically accurate. In addition, a YOLO model is included for real-time object detection, which makes it possible for the model to precisely identify and track several articles of clothing at once. The suggested architecture is evaluated using the BLEU score performance metric; research on the selected dataset yielded a BLEU-1 score of 0.983 for the ConvNeXtLarge-based model.

INDEX TERMS Apparel captioning, BLEU score, CNN, ConvNeXtLarge, deep learning, LSTM, YOLO.

I. INTRODUCTION

In the ever-expanding realm of artificial intelligence, the convergence of computer vision and natural language processing has paved the way for transformative applications, among which image captioning stands out as a potent means of enriching visual data with descriptive textual information. Image captioning is the process of automatically generating a textual description of an image. It is essentially bridging the gap between vision and language, allowing machines to "see" and understand the content of an image and then express that understanding in words. In recent years, researchers have directed their focus on the problem of image captioning [17], [18]. However, one of the areas in which image captioning has not been widely used is fashion, despite all its attractiveness for researchers in AI and specifically in CV. Indeed, in the last decade, the fashion industry has attracted many researchers in CV [7]. A lot of growth is happening in the fashion industry with the increase in population. In fact, according to Statista [16], the number of users in the fashion market are expected to amount to 2.6 billion by 2029. Within this context, the particular field of apparel description presents unique challenges and opportunities. This technology of image captioning can be used to caption images on e-commerce platforms. E-commerce platforms can leverage advanced technologies like Natural Language Processing (NLP) to automate the creation of detailed and appealing product descriptions. By analyzing product images and specifications, an automated system can generate accurate and interesting captions, reducing the time and effort required for manual content creation. Although consumers are increasingly visiting physical stores, e-commerce is

The associate editor coordinating the review of this manuscript and approving it for publication was Krishna Kant Singh.

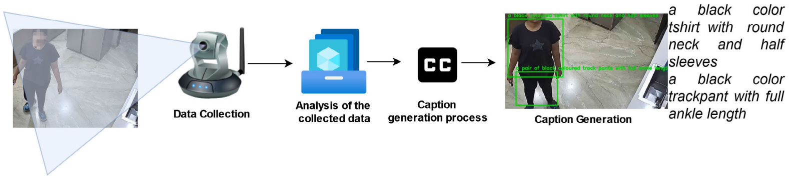

Data Collection
Analysis of the collected data
CC
Caption generation process
a black coloured tshirt with round neck and half sleeves
a pair of black coloured track pants with full ankle length
Caption Generation
a black color tshirt with round neck and half sleeves
a black color trackpant with full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

FIGURE 1. An overview of image captioning process.

projected to account for 41% of global retail sales by 2027, a substantial rise from its 18% share in 2017, as indicated by recent data published by Boston Consulting Group (BCG) [19]. E-commerce systems have the ability to utilize advanced technologies such as Natural Language Processing (NLP) to automate the process of generating comprehensive and attractive product descriptions. An automated system may generate precise and captivating captions by evaluating product photos and specs, thereby minimizing the manual labor and time needed for content production. The inclusion of image captions on e-commerce websites is an essential accessibility element that aids visually impaired users and promotes a more inclusive online purchasing experience. This not only enables those with visual impairments to make well-informed purchase decisions, but also guarantees that they have equal autonomy and involvement in the digital economy.

This work seeks to harness the capabilities of deep learning to advance the state of the art in image captioning, focusing initially on apparel imagery and extending its objectives to include the dynamic dimension of video annotation as shown in figure 1. The goal is to create a system that can understand the content of an image and express it in human-readable language. This involves the use of deep learning models, typically neural networks, to analyze the visual features of an image and generate a relevant and coherent caption. Several techniques have been developed to address the challenges inherent in image captioning. These techniques can be broadly categorized into two main approaches namely traditional computer vision-based methods and modern deep learning-based methods. The advent of deep learning has revolutionized the field of image captioning, enabling models to automatically learn hierarchical representations from data. Convolutional Neural Networks (CNNs) are commonly employed for image feature extraction, while Recurrent Neural Networks (RNNs) or Transformer-based models are used for sequential language generation. This combination allows the model to capture both visual and semantic information effectively. This research work delves into the realms of deep learning to develop a sophisticated image captioning system tailored for visually impaired individuals in the realm of apparel. The primary objective of this research is to explore and implement state-of-the-art deep learning techniques, such as convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequential data generation, to create an image captioning model that can encapsulate the essence of diverse apparel items. This research seeks to explore the development of a comprehensive deep learning model that not only generates descriptive captions for apparel images but also extends its application to the nuanced task of video captioning. The dynamic nature of videos introduces temporal dependencies, demanding a seamless integration of spatial and temporal understanding. By leveraging the foundations established in image captioning, the model will be refined and expanded to generate coherent and informative captions for videos depicting apparel items. This extension not only broadens the scope of the research but also meets the growing demand for multimedia content understanding in various industries, with a particular focus on enhancing accessibility for visually impaired individuals. Research paper objectives are:

- Gathering and creating a thorough dataset of clothing photographs accompanied by detailed textual descriptions for the purpose of training and assessing the image captioning model.
- Including some Indian dresses and styles in the dataset to increase diversity.
- Extending the system such that it works with video input.

Some of the challenges faced during this study are:

- In order to develop robust image captioning models, it is crucial to have a large amount of datasets for model training. One primary challenge is the absence of datasets with uniform and standardized captions consisting images across diverse categories of clothing.
- Images may feature various clothing items in a single photograph, consisting accessories and background elements like different architecture, resulting in visual clutter. It can be challenging to develop a system which can identify the main clothing items and describe them accurately in the midst of the clutter. Advanced computer vision methodologies, employing deep learning models which are trained on diverse and massive datasets of annotated clothing images, while incorporating contextual understanding and domain-specific knowledge of fashion, are required in order to address these challenges.
- Another challenge is generating attribute-level captions that are more accurate and descriptive in comparison to traditional captions; this requires a deeper understanding of the clothing attributes.

The subsequent sections of the paper are structured as follows: The related work outlines the existing research pertaining to the field of image captioning. The proposed architecture provides further details regarding the suggested methodology for image and video captioning of apparel, as well as an account of the dataset employed in the study. The implementation of different components of the proposed architecture and the experimental outcomes are detailed in the Results and Implementation section, which is followed by the Conclusion.

II. RELATED WORK

Feng, Wen et al. employed a pre-trained BLIP model based ViT-B approach in their study [1] to extract distinct features from Dai clothing images. To create textual descriptions of Dai clothes from these attributes, they utilized BertLMHeadModel to decode feature vectors. They curated a Dai clothing dataset of image and annotation datasets to simplify model training. After performance evaluation, it was observed that the ViBeCap model surpassed Clip-Clap in readability, image extraction, and caption correctness. In the study [2], Al-Malki et al. explored the application of deep learning methods to generate descriptions for images displaying clothing in Arabic language. They curated a dataset named "ArabicFashionData" containing images of clothes coupled with corresponding single Arabic captions. They sourced the images for the dataset from the DeepFashion collection, and the Arabic translations of the captions were obtained from previously published sources. They utilized an encoder-decoder architecture with an attention layer, and their proposed models enable the decoder to select the most relevant regions of the image based on recovered attributes. The model used a pre-trained ResNet50 network as its image encoder, which has been thoroughly trained on huge datasets to achieve exceptional accuracy. With a BLEU-1 score of 88.52, the suggested approach outperformed the existing systems according to performance evaluation. In the research [3], an integration of Attribute-based Alignment Module (AAM) is introduced by Tang et al. as a means to enhance fashion captioning. In order to increase efficacy, it is suggested that well-aligned grid features be utilized rather than dense features. They present a dataset which was collected from reputable fashion and retail websites known as Fashion-Style-27k dataset. A wide variety of detailed designs and metadata were incorporated. Their technology outperforms current approaches by combining linguistic data at the sentence and attribute levels with a trained Fashion Language Model (FLM). A BLEU score of 88.52 was obtained by means of comprehensive tests and visualizations on the FACAD and FashionGen datasets.

Convolutional neural networks (CNNs) like VGG-16 and VGG-19 were used by Dwivedi and Upadhyaya in [4] to extract features from the Fashion MNIST dataset, which included 70,000 28 × 28 pixel gray-scale images. Recurrent neural networks (RNN) were used to analyze text data by converting the input sequences into output vectors of fixed length. They present a convolutional neural network with five layers (CNN-5) and evaluate its performance in comparison to transfer learning models. Text was converted into compact vector representations through word embedding, and the outputs of the text processor, feature extractor, and other processes were combined by a decoder. A densely connected layer of 256 neurons was used to determine the following phrase's vocabulary, and BLEU scores were used to assess its effectiveness. According to the findings, a CNN model with a reduced density computes faster than the VGG-16 and VGG-19 models. Khurram et al. introduced a sophisticated deep learning architecture in their study [5], which focused on region-based analysis to deliver comprehensive descriptions of image semantics. Their proposed system, Dense-CaptionNet, utilized the Visual Genome dataset for region and attribute description generation, along with the MSCOCO and IAPR TC-12 datasets for sentence generation. The three main components of the network are sentence generation modules, object attributes and descriptions of the regions and relationships. Dense-CaptionNet surpasses state-of-the-art benchmarks with guaranteed grammar correctness, outperforming current approaches. In the research [6], a multi-attribute classification model which is based on a deep neural network is introduced. A dataset comprising 30,000 images collected from shopping centers and TV shows was used for model training. The dataset consists of two types of bottoms and seven categories of tops. This approach enhanced classification accuracy by taking into account the relationships between several attribute values to recognize different categories of clothing and their associated sub-attributes. Their method improves attribute linkage based on DeepMar's classifier, achieving precise predictions on a dataset of aged clothing without using binary classification. The proposed method achieves a BLEU-1 score of 67.2, indicating precise predictions across several apparel categories.

In the research study [7], Hacheme and Sayouti introduced a dataset named "InFashAIv1", which consisted over 16,000 images of African fashion items sourced from datasets like DeepFashion and Afrikrea. They employed the Show and Tell strategy for their approach which involved pre-training a CNN encoder with ResNet152 and utilized a RNN decoder to generate captions. To enhance the caption quality after evaluating multiple parameter sets, they opted for joint training on both datasets, which indicated the potential of utilizing Western fashion data for transfer learning. Though the generated captions performed well, any gender biases should be noted. Tateno et al. developed a system in their work [8] for visually impaired individuals for selecting clothes by transforming visual apparel information into verbal descriptions using a DNN-based technique. The DeepFashion dataset contains annotated images with attributes such as color, print, and sleeves, resulting in a customized dataset. They employed VGG16 for feature extraction from the images, it was fine-tuned on a dataset

TABLE 1. **A summary of the existing approaches in the context of used models and benchmark datasets.**

Author	Dataset	Methodology	Result
Feng, Wen et al. [1]	Dai clothing dataset	ViBeCap (BertLMHeadModel) and ClipCap model	ViBeCap outperformed ClipClap in terms of generating accurate caption.
Al-Malki et al. [2]	Made Arabic-FashionData using DeepFashion	Pre-trained ResNet50 Network and an attention layer to decode input images.	BLEU-1 score of 88.52.
Tang et al. [3]	Fashion-Style-27k	Attribute-based Alignment Module and a pre-trained Fashion Language Model.	BLEU score of 88.52.
Dwivedi et al. [4]	Fashion MNIST dataset	(VGG-16+RNN), (VGG-19+RNN), (CNN-5+RNN)	BLEU -1 score of 0.5348.
Khurram et al. [5]	Visual Genome, MSCOCO and IAPR TC-12 datasets	VGG-16, RPN and LSTM	BLEU-1 score of 0.707 on MSCOCO dataset.
Chankyu Park et al. [6]	Custom dataset using photos from TV and retail mall.	DeepMar framework (CNN+RNN)	BLEU-1 score of 67.2.
Hacheme et al. [7]	InFashAIv1 and DeepFashion datasets	Pre-trained ResNet152 and LSTM	BLEU-1 score of 0.474 on InFashAIv1 dataset.
Tateno et al. [8]	Custom dataset using DeepFashion dataset.	The suggested system is based on VGG16+LSTM architecture.	Proposed system generated appropriate caption with an accuracy of 90%.
Cai et al. [9]	Built FACAD170K dataset using FACAD dataset.	The model consists ResNet101 backbone and LSTM.	BLEU-1 score of 46.5.
Wooders et al. [10]	Custom dataset using ShopStyle and Amazon.	The model consists Inceptionv3 backbone and LSTM.	The system generated image embedding of great and consistent quality.
Verma et al. [11]	Flickr8k and MSCOCO	The research uses VGG16 Hybrid Places 1365 model and employs LSTM.	BLEU-1 score of 0.6666 on the Flickr8k dataset and 0.7350 on the MSCOCO dataset.
Mamatha et al. [12]	Flickr8k	ResNet+LSTM	ResNet based architectures outperformed custom CNN and VGG architectures.
Gosh et al. [13]	Flickr8k	VGG16+LSTM	BLEU score of 0.562.
Jafar et al. [14]	Flickr8k	InceptionV3 and LSTM	BLEU-1 score of 0.4251.
Aristanbekov et al. [15]	MSCOCO	The research introduce image captioning using ExpansionNet v2.	BLEU-4 score of 41.

comprising 15,000 clothing images. An LSTM analysis revealed a significant correlation between the descriptive words and the visual qualities. This allows for the generation of appropriate labels for 80% of unlabeled photos. In the research [9] by Cai et al. an innovative approach for describing fashion images is proposed in which they combine various input methods and enable the user to define semantic criteria for personalized captions. To support their study they derive a dataset called "FACAD170K" from an existing dataset called "FACAD." There is a Transformer-based picture encoder with feature enhancement, an attribute controlling encoder, and a language decoder in the ACC model. Their approach surpasses the performance of the current strategies for extracting image inputs by utilizing Transformer based image encoder built on ResNet101 architecture.

Wooders et al. introduce a dataset comprising 373,521 images of dresses and shirts along with their titles, colors, and descriptions sourced from ShopStyle and Amazon in their research [10]. The dataset includes retail products from over a thousand e-commerce websites and aggregated by ShopStyle.com. 37% of the items are women's blouses, another 37% are dresses, 6% are sweaters and men's shirts are another 20%. They used a Show and Tell model which utilized an LSTM to transform the image embedding from the encoder that consists of Inception v3 into human readable descriptions. Their research shows that the image embedding made by the trained model is more effective at finding similarities between photos than standard dataset-trained models like MS COCO. Verma et al. present a methodology for generating grammatically accurate captions for images in their study [11]. They used an encoder-decoder architecture as a foundation for their framework which sequentially generates words and congregate them into sentences in order to produce captions. They employ LSTM as a decoder and VGG16 Hybrid Praces as encoder which were trained on Places 365 and ImageNet datasets. The approach was trained on annotated datasets like Flickr8k and MSCOCO Captions datasets. The model's performance was assessed using BLEU, METEOR, and GLEU scores. The experimental results indicated the following values of BLEU-1, METEOR, and GLEU 0.666, 0.5060, and 0.2469 respectively, on the Flickr8k dataset and 0.7350, 0.4768, and 0.2798 on MSCOCO Captions dataset. Mamatha et al. [12] employed an encoder-decoder system consisting convolutional neural network (CNN) as encoder and long short-term memory network (LSTM) as decoder to address the problem of generating captions for images. The CNN used for encoder was ResNet models which exhibited superior performance compared to custom CNN and VGG networks. The model was developed using FLICKR8K dataset, comprising of 8,000 preprocessed images intended for input to ResNet.

In the subsequent stages the captions of the dataset were employed to generate a vocabulary through the elimination of superfluous digits and blank spaces. LSTM was employed which progressively generated captions for inputs consisting of previously generated words by integrating outputs from ResNet and vocabulary.

Gosh et al. in their research [13], introduce an innovative approach named Next-LSTM for generating image captions. The methodology employs VGG16 for extracting features from the images and Long Short-Term Memory (LSTM) networks for generating captions with precision. They evaluated the approach on the Flickr8k dataset, and observed that the model demonstrated superior accuracy compared to the most recent methods. The proposed framework consisted of five phases including text pre-processing, feature extraction, caption construction, model and outcomes, and output predictions. Prior to feature extraction the images were preprocessed with three color channels and scaled to 224 × 224 pixels. The model's high performance was evidenced by a BLEU score of 0.562 after which the study concluded that VGG outperformed other neural network architectures. Jafar et al. [14] presented a novel deep learning apprach for generation of image captions via integrating convolutional neural networks (CNNs) like Inception model with recurrent neural networks (RNNs) like Long Short-Term Memory (LSTM). They employed a pre-trained CNN for feature extraction and used a RNN based language model for caption generation. Performance evaluations resulted in a BLEU-1 score of 0.4251 for the proposed approach surpassing the performance of prior approaches. For evaluation of the methodology they utilized Flickr8k dataset. Aristanbekov et al. in [15] developed an assistive system which integrated image captioning and text-to-speech capabilities to provide visually impaired individuals with real time audio descriptions in a local language Kazakh. Their developed software uses the MSCOCO dataset to create descriptions for over 123,000 photos. They used google translate to convert the captions to Kazakh language which were later confirmed and verified by a human translator. This novel strategy improves accessibility by providing visually challenged users with detailed audio descriptions in Kazakh, containing 18,363 words. The model's ability to produce accurate captions is demonstrated by its BLEU-4 score of 41.

Based on the related work, it can be concluded that most of the research work has focused their area of interest on generic image captioning and translation of those captions. However, some of the research that exists doesn't extend their approach to video systems and either don't have a publicly available dataset or the system is generating inconsistent captions indicative by BLEU scores. Focusing on the above discussion, this research attempts to address the challenge of implementing image and video captioning on apparel images by developing an image captioning system that also captions videos using state-of-the-art technologies like YOLO and deep learning techniques.

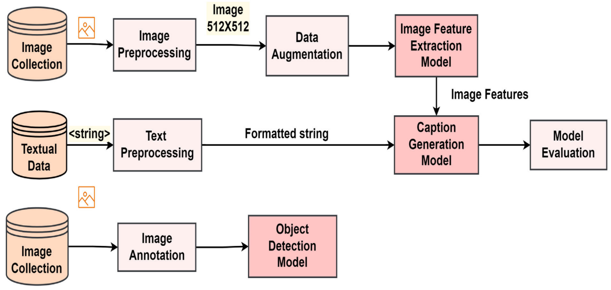


FIGURE 2. Proposed architecture for apparel based image and video caption generation system.

III. PROPOSED APPROACH

The objective of the research is to distinguish a region of interest in a video which includes clothing item. After identifying the region of interest it is classified into a category of top, bottom or a dress and is segmented using a box. Further which this a series of feature vectors is extracted based on the image in the region with the help of which a textual description of apparel in the box is generated using a CNN-LSTM based approach. The Apparel based image and video caption generation system is composed of the subsequent modules: dataset collection, image preprocessing, text preprocessing, image annotation, image feature extraction, caption generation model, object detection model and result comparison as shown in Figure 2. The notations used for the proposed work are summarized in table 2

A. DATASET COLLECTION

The study employed a systematic technique to collect data, obtaining 863 high-resolution images of clothing from copyright-free websites such as Unsplash, Pexels, Pixahive, and PxHere. There were 26 kinds of apparel in the photographs, including 14 tops, 9 bottoms, and 3 dresses, with a focus on featuring Indian attire such as sarees, kurtas, and leggings to highlight cultural diversity. The dataset's representations were enhanced by prioritizing diversity, as seen by images including models of all ages, genders, body types, and ethnicities from various areas of the world. Quality control techniques were implemented to guarantee that the dataset included visually clear and relevant representations of clothing by removing unnecessary or low-quality images. Each image was annotated with descriptive captions to provide valuable training data for the image captioning model, resulting in insightful and contextually relevant explanations. This comprehensive approach prioritizes quality, diversity, and relevance to establish the foundation for creating a robust deep learning model capable of accurately captioning images of various apparel items.

B. IMAGE PREPROCESSING

Image processing is a crucial module in the image and video caption generation system targeting the task of enhancing the

TABLE 2. Notation description.

Notation	Description
$W_{original}$	Original width of image.
$H_{original}$	Original height of image.
W_{max}	Maximum possible width of image wile resizing.
H_{max}	Maximum possible height of image wile resizing.
$Scale_{width}$	Width scale of original image w.r.t W_{max}.
$Scale_{height}$	Height scale of original image w.r.t W_{max}.
$Aspect\_ratio$	Minimum scale up or down of the original image.
W_{new}	Width for resizing image.
H_{new}	Height for resizing image.
W_{target}	Width to be achieved using padding.
H_{target}	Height to be achieved using padding.
Pad_{width}	Padding to be added horizontally.
Pad_{height}	Padding to be added vertically.
Pad_{left}	Padding to be added on left.
Pad_{right}	Padding to be added on right.
Pad_{top}	Padding to be added on top.
Pad_{bottom}	Padding to be added on bottom.
cap	Caption after text preprocessing.
$caption_{final}$	Caption after adding start and end tokens.
C_x	X coordinate of center of annotation box.
C_y	Y coordinate of center of annotation box.
B_w	Width of annotation box.
B_h	Height of annotation box.
fg_t	Forget gate of LSTM at time t.
W_{fg}	Weight for Forget gate of LSTM.
X_t	LSTM input at time t.
h_{t-1}	Hidden state of LSTM at time t-1.
b_{fg}	Bias for Forget gate of LSTM.
inp_t	Input gate of LSTM at time t.
W_{inp}	Weight for input gate of LSTM.
b_{inp}	Bias for input gate of LSTM.
$\tilde{C}_t$	Intermediate cell state of LSTM at time t.
W_c	Weight for cell state of LSTM.
b_c	Bias for cell state of LSTM.
C_t	Updated cell state of LSTM at time t.
C_{t-1}	Cell state of LSTM at time t-1.
out_t	Output gate of LSTM at time t.
W_{out}	Weight for output gate of LSTM.
b_{out}	Bias for output gate of LSTM.
h_t	Hidden state of LSTM at time t.
BP	Brevity penalty.
w_n	weight assigned to n-gram precision.
$\sum_{clipped_{c}ount}$	Cumulative clipped count of n-gram.
$\sum_{count}$	Cumulative count of n-gram.

input images prior to feeding them to the captioning model. Firstly, the images were resized to 512 X 512 pixels and converted to JPG standard format to retain their visual quality while reducing computational complexity. The images are transformed from RGBA to RGB format to assess the computation of the images in the neural network architecture. The proposed process for image processing preserves the original dimensions and compensates for the aspect ratio errors. Data augmentation was also employed comprising of the following operations like horizontal flipping, rotating and color modification to induce differences in the images. The training set contained a total of 5411 images after augmenting the original 863 images.

Aspect ratio is the proportional relation between the height and width of the image. It is necessary to resize the image according to the aspect ratio as for clothing item height and width are key identifiers. So for resizing the image according to the aspect ratio, first scale of height and width are calculated using equations 1 and 2 using which the aspect ratio is calculated as shown in equation 3.

$$Scale_{width} = W_{max}/W_{original} \tag{1}$$

$$Scale_{height} = H_{max}/H_{original} \tag{2}$$

$$Aspect\_ratio = min(Scale_{width}, Scale_{height}) \tag{3}$$

After calculation of the aspect ratio, the new height and width of the image are calculated using 4 and 5 such that the resized image is in the same aspect ratio.

$$W_{new} = Scale_{width}XAspect\_ratio \tag{4}$$

$$H_{new} = Scale_{height}XAspect\_ratio \tag{5}$$

To ensure all the images in the dataset have a uniform size a white color padding was added after resizing the image. Padding converts the image to the same dimensions while maintaining the aspect ratio of the images. For this, first the required padding for the height and width of the image was calculated using 7 and 6.

$$Pad_{width} = W_{target} - W_{new} \tag{6}$$

$$Pad_{height} = H_{target} - H_{new} \tag{7}$$

After, this the padding that is to be added on each side of the image was calculated using equations 8, 10, 11 and 9.

$$Pad_{left} = Pad_{width}/2 \tag{8}$$

$$Pad_{top} = Pad_{height}/2 \tag{9}$$

$$Pad_{right} = Pad_{width} - Pad_{left} \tag{10}$$

$$Pad_{bottom} = Pad_{height} - Pad_{top} \tag{11}$$

C. TEXT PREPROCESSING

In order to enhance the raw textual input linked to the apparel images, the text preprocessing module plays a critical role in the proposed system. All the words were initially converted to lowercase to maintain the consistency and remove the capitalization differences. Following this, the special characters and punctuation were eliminated to further simplify the text. After this, redundant words like stop words were eliminated to decrease the interference and direct the model's focus towards primary language characteristics. After this, starting and ending tokens were added to the caption as shown in equation 12. The processed text was then broken down into tokens or words using tokenization, enabling the formation of organized sequences that may be efficiently interpreted by the deep learning model. To improve the system's capacity to depict apparels various text preprocessing techniques were used to standardize and enhance the textual data to produce captions which closely align with the visual content of the apparel images, ensuring they are logical and contextually appropriate.

$$caption_{final} = 'startseq' + cap + 'endseq' \quad (12)$$

D. IMAGE ANNOTATION

The image annotation module is a crucial component of the apparel description system's architecture. This module optimizes the manual annotation procedure by utilizing the functionalities and techniques offered by the Computer Vision Annotation Tool (CVAT) website [20] for the comprehensively chosen collection of images of apparels. Annotations were performed on the original 863 images and not on the augmented dataset to avoid redundancy. The dataset is thoroughly annotated, utilizing a range of techniques to promote diversity such as bounding boxes, polygons, and key points. By using CVAT's user-friendly interface, annotation task was completed smoothly and rigorous quality control procedures were carried out to ensure every annotation is accurate and consistent. The dataset's quality is enhanced through continuous enhancement using iterative procedures. After completion of the process the annotated dataset was exported in a standardized format to facilitate its integration into the training process of the deep learning models. This process enables the generation of detailed and coherent captions of the apparel images via utilizing precise image captions models which necessitates datasets of superior quality, which can be generated through a comprehensive annotation process. Following is the format of the annotation of each object:

$$Annotation = [class, C_x, C_y, B_w, B_h] \quad (13)$$

E. IMAGE FEATURE EXTRACTION MODEL

The apparel captioning system primarily depends on the Image Feature Extraction Model, which employs convolutional neural networks (CNNs) and transfer learning approach to extract essential features from the images of apparels for with the agenda of providing accurate and contextually pertinent captions. This module enhances the models on the collected dataset of apparels using progressing transfer learning techniques, including the following architectures like VGG16, VGG19, InceptionV3, ResNet50, ResNet101, ResNet152, and ConvNeXtLarge, in order to refine them for the apparel image detection procedure. The employed CNN models are pre-trained on the ImageNet dataset. Clothes comprise a variety of distinct characteristics such as designs, colors, and garment structures. The utilized pre-trained CNNs were refined to identify these attributes. This approach accelerates the model's convergence as well as improves its understanding of various clothing items and structures by capturing domain-specific properties, which are crucial for precise clothing captioning. ResNet50, ResNet101, and ResNet152 utilize residual learning to transmit complex information at different levels. In contrast, the ConvNeXt-Large framework enhances feature representation and model expressiveness by combining group and depth-wise convolutions as shown in Figure 3.

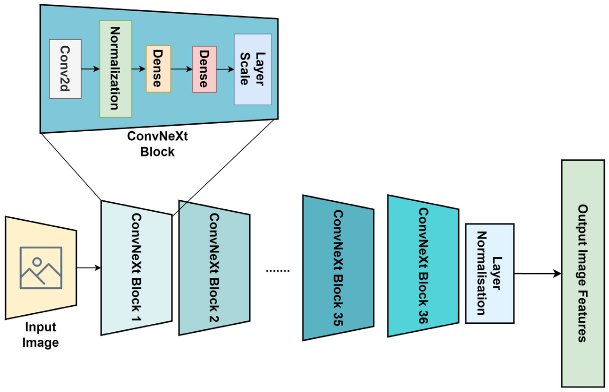


FIGURE 3. ConvNeXtLarge architecture for extraction of the features.

F. CAPTION GENERATION MODEL

The Caption Generation Model is essential for generation of sequential captions by employing Long Short-Term Memory (LSTM) networks and convolutional neural networks (CNNs) for extracting visual features as depicted in Figure 4. This model is comprised of the system's decoder, which is formed using word embedding, LSTM cells and inputs the encoded image attributes. The LSTM network is a language model, it generates words and sentences in a sequential manner to fabricate detailed captions for an image of an apparel. Along with this, word embedding were used to represent every word in the vocabulary as well as words generated by the LSTM decoder. Additionally, a vocabulary was created by compiling the unique tokens out of the corpus of texts. In practice the uncommon words can be replaced with special symbols or can be controlled using techniques like sub-word tokenization to limit the vocabulary size without sacrificing important information. This methodology was used to transform words into numerical representations that may be examined further. As shown in Figure 4 first the word is processed by an embedding which learns 256 features of the words which is further processed by a 256 neuron LSTM layer after adding a dropout to avoid over-fitting and output text related features. These text related features are then combined with the image features obtained using CNN encoder and passed through a dense layer to learn more complex feature. At last a soft max layer is applied which results in a probability distribution of the next probable word. This technique enabled the anticipation of the following word in the sequence by using this distribution.

Figure 5 shows the architecture of the LSTM module used in the study. Like a standard LSTM network, the reading, writing, and updating of all the cells which include the memory cell, forget cell and output cell are dependent on tanh and sigmoid gates. At a given time stamp t following are the equations used in the LSTM module:

$$fg_t = \sigma(W_{fg} \cdot X_t + W_{fg} \cdot h_{t-1} + b_{fg}) \quad (14)$$

As LSTM is used for long sentence generation there is a need to evaluate previous information or context of

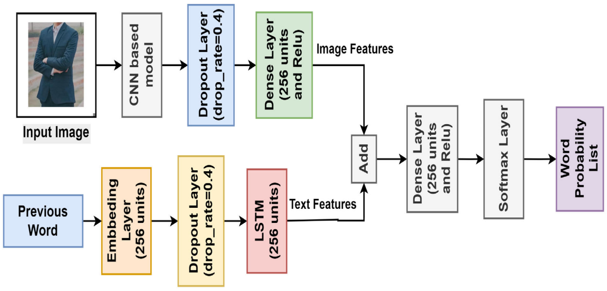

Input Image
CNN based model
Dropout Layer (drop_rate=0.4)
Dense Layer (256 units and Relu)
Image Features
Previous Word
Embbeding Layer (256 units)
Dropout Layer (drop_rate=0.4)
LSTM (256 units)
Text Features
Add
Dense Layer (256 units and Relu)
Softmax Layer
Word Probability List

FIGURE 4. Caption generation model architecture.

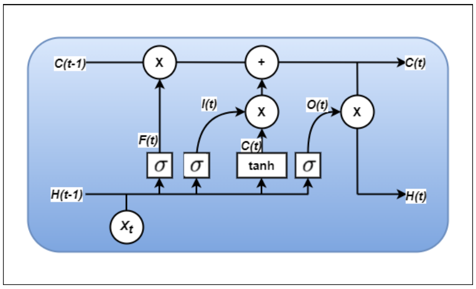


FIGURE 5. LSTM memory component architecture.

the sentence and forget unnecessary information for this equation 14 is used as it helps to forget irrelevant information using a sigmoid gate.

$$inp_t = \sigma(W_{inp} \cdot X_t + W_{inp} \cdot h_{t-1} + b_{inp}) \tag{15}$$

$$\tilde{C_t} = tanh(W_c \cdot X_t + W_c \cdot h_{t-1} + b_c) \tag{16}$$

$$C_t = fg_t \cdot C_{t-1} + inp_t \cdot \tilde{C_t} \tag{17}$$

Equation 15 has been used to evaluate which new information needs to be stored for output and also for future reference. For this current input X_t and the previous hidden state of t-1 LSTM i.e. h_{t-1} is considered. Also the new information is being stored in $\tilde{C_t}$ and calculated by equation 16. These are then being updated in the current cell state for present and future reference in C_t using equation 17.

$$out_t = \sigma(W_{out} \cdot X_t + W_{out} \cdot h_{t-1} + b_{out}) \tag{18}$$

$$h_t = out_t \cdot tanh(C_t) \tag{19}$$

The output of the LSTM layer is being calculated using equation 18 and the hidden state for the current timestamp t which will be used in the next timestamp is calculated using the output of equation 18 and 17 as shown in equation 19.

G. OBJECT DETECTION MODEL

The research utilized the state-of-the-art YOLOv8 framework to precisely and accurately locate and identify clothing items in the images for the object detection module. The systematic flow for object detection is displayed in Figure 6. YOLOv8 was trained meticulously on heterogeneous datasets which consisted a wide range of clothing categories, orientations and styles to acquire the ability to discriminate between distinct clothing objects with a high degree on accuracy. This module is essential for the functionality of the proposed system as it quickly and accurately generates bounding boxes around the identified apparel items in the images by integrating it smoothly into the image processing pipeline. This process of object detection guarantees reliable and effective detection of items by utilizing the speed and precision of YOLOv8 in conjunction with rigorous optimization methods. It establishes a strong basis for the stages of clothing description and caption generation.

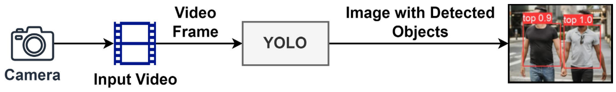


FIGURE 6. An overview of the object detection process.

H. VIDEO CAPTIONING SYSTEM

Figure 7 depicts the video captioning process, which is divided into several steps. A video is first captured with a camera, and then the input video (the captured video) is divided into individual frames. The frames are then processed in an iterative manner. The YOLO algorithm is used to detect objects in each frame. If an object is not detected, then the frame is appended to the output video without any changes, and if an object is detected, the image is cropped in accordance with the boundaries of the detected object. The image is then pre-processed by an image feature extractor module to extract features from the processed output image. Subsequently, the caption generation model receives these features and uses them to produce captions, as demonstrated in Figure 7. After the caption is generated for the detected object in the frame, it is written on it using CV2. The CV2 library's putText() function is used on the frame with a green-colored font style called Hershey Simplex and a font scale factor of 0.8. This frame is then added to the frame for the output video generation. Eventually, the output is a video that includes captions.

I. MODEL EVALUATION

The research utilized the BLEU score as a performance indicator to assess the effectiveness of the models employed for producing apparel descriptions. The research incorporated BLEU-1, BLEU-2, BLEU-3, and BLEU-4 by utilizing the corpus_bleu function from the NLTK package. The scores indicate the similarity between the caption generated by the model and the actual caption of the garment. This includes comparing the real and predicted captions using uni-grams, bi-grams, tri-grams, and four-grams.

$$BLEU - n = BP \times exp(w_n \times \frac{\sum_{clipped_count}}{\sum_{count}}) \tag{20}$$

The BLEU score, which is computed using the n-gram precision and the brevity penalty as shown in equation 20, is a value between 0 and 1 as determined by the NLTK

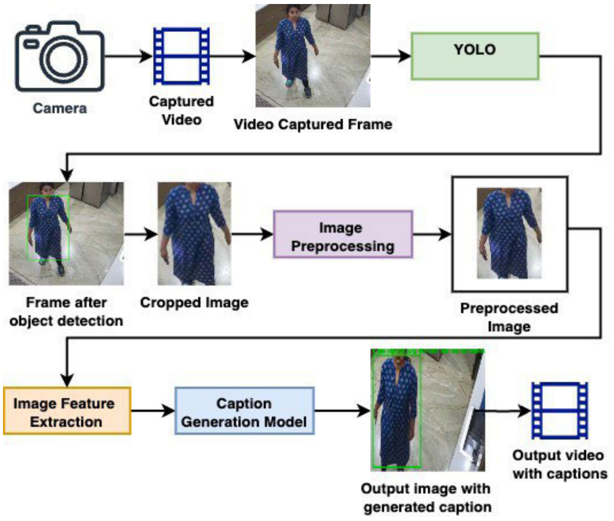


FIGURE 7. System diagram for the video captioning process.

function. A higher BLEU score, specifically a score closer to 1, indicates a stronger resemblance between the actual and anticipated caption. This evaluation technique will enable a full assessment of the model's performance, offering vital insights into its efficacy in accurately depicting apparel items in images as well as videos.

IV. IMPLEMENTATION & RESULTS

A. IMPLEMENTATION DETAILS

During the training phase, seven transfer learning models were utilized as feature extractors: VGG16, VGG19, InceptionV3, ResNet50, ResNet101, ResNet152, and ConvNeXtLarge. All of these models were pretrained on the ImageNet dataset and implemented using the TensorFlow Keras program. The basic layout of ResNet50 was altered by eliminating the top layer and incorporating a batch normalization layer. Additionally, two dense layers of 1024 and 512 neurons respectively were introduced, with relu serving as their activation function. Finally, a softmax layer was appended on top. ResNet101 and ResNet152 were enhanced by incorporating a batch normalizing layer. Additionally, a single dense layer with 512 units and a relu activation function was introduced. Finally, softmax layers were appended on top. However, VGG16, VGG19 and InceptionV3 were not altered to maintain the simplicity of their architecture. Along with this ConvNeXtLarge was also not altered since it is already a comprehensive and profound model. The ResNets, which are modified models, were then trained as classifiers on the same dataset. They were then utilized as feature extractors in the caption generating module by removing top softmax layer in order to get more complex set of features about the input image. The caption generating module utilizes an LSTM architecture with 256 units and was trained for a total of 30 epochs. In addition, a dropout layer with a rate of 0.4 is included before the LSTM. The Adam optimizer is utilized for optimization, with a consistent learning rate of 0.001 applied to all models. The YOLOv8m model is utilized for object detection in videos, improving the system's capacity to comprehend the visual environment and recognize pertinent clothing items. The YOLOv8m model undergoes training for 150 epochs, guaranteeing strong detection performance over a wide range of video datasets.

B. RESULTS

The analysis of the results serves three purposes. Initially, the generated features are scrutinized in order to ascertain their importance and impact on the performance of the system. Additionally, it assesses the efficacy of the implemented models in comparison to established systems. Ultimately, the efficacy of apparel item detection in video footage is evaluated. Through the execution of this analysis, our objective is to furnish valuable insights pertaining to the merits, demerits, and overall efficacy of our methodology. This will contribute to a holistic comprehension of its capacities and prospective advancements within the discipline.

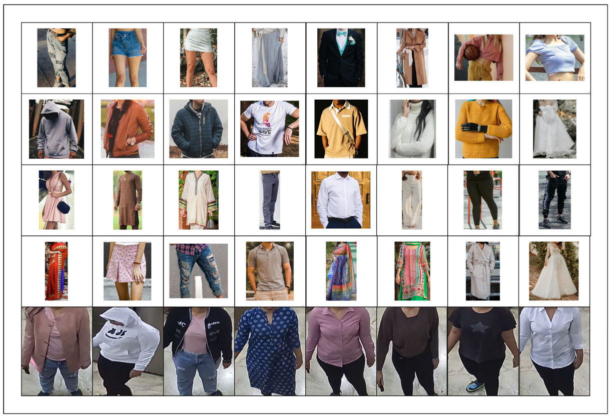


FIGURE 8. A set of sample images from the dataset used for the experimental analysis.

Figures 8, 9, 10, 11 show some of the sample images from the data set collected for the study. These samples include images from varied cultural and gender-diversity images to ensure that the model has reduced bias towards certain factors.

FIGURE 9. Sample images of the Flickr8k dataset.

Figure 12 shows the diversity in the type of clothing used in the study, and from the graph it can be inferred that there is very little imbalance between the different clothing categories. This will ensure consistent performance of the model on each type of clothing item present in the dataset.

FIGURE 10. Sample images of the MSCOCO dataset.

FIGURE 11. Sample images of the InFashAlv1 dataset.

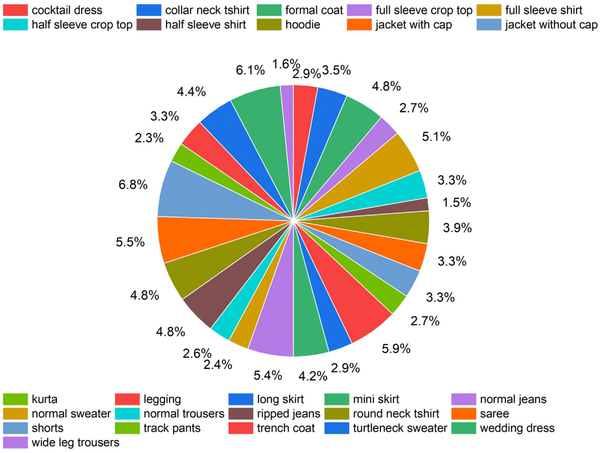


FIGURE 12. Distribution of different cloths in the dataset.

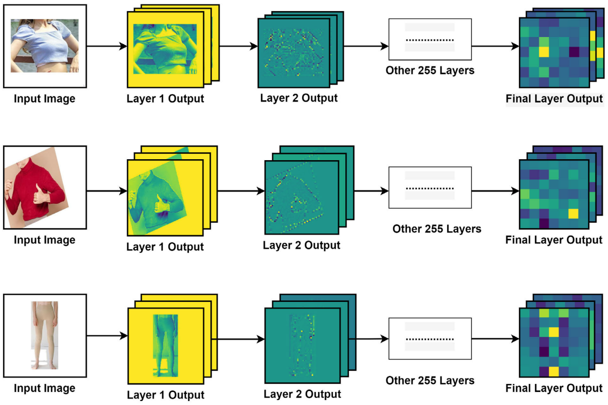


FIGURE 13. Output of intermediate layer of ConvNeXtLarge.

The input image is processed for extraction of the features. Figure 13 illustrates the manner in which the models interpret the input image and emphasize details such as edges and other relevant information with each iteration of the model.

TABLE 3. BLEU scores comparisons of applied approaches.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16	0.779	0.68	0.6	0.545
VGG19	0.808	0.719	0.65	0.602
Inceptionv3	0.861	0.8	0.752	0.709
ResNet50	0.972	0.956	0.939	0.924
ResNet101	0.981	0.97	0.959	0.947
ResNet152	0.977	0.964	0.951	0.937
ConvNeXtLarge	0.983	0.973	0.964	0.953

The results of the performance metrics of the different captioning models implemented are shown in Table 3. Corpus_bleu function of NLTK was used for model evaluation. The results indicate that the ConvNeXtLarge model outperformed the other models with a BLEU-1 score of 0.983. The model also has the highest BLEU-2, BLEU-3 and BLEU-4 scores of 0.973, 0.964, and 0.953 respectively.

TABLE 4. BLEU scores of ConvNextLarge on existing datasets.

Methods	Flickr8k [24]	MSCOCO [22] [23]	InFashAlv1 [21]
[7]	-	-	0.474
[11]	0.66	0.73	-
[13]	0.562	-	-
[14]	0.425	-	-
[15]	-	0.41	-
ConvNextLarge	0.617	0.695	0.681

Training loss analysis is presented in Figure 14 while using VGG16, VGG19, InceptionV3, ResNet50, ResNet101, ResNet152, and ConvNeXtLarge. It can be inferred that ConvNeXtLarge has the steepest training loss curve, which suggests that the model learning is better for ConvNeXtLarge. It also has the least training error out of all the proposed models.

To analyse the performance of the VGG16, VGG19, InceptionV3, ResNet50, ResNet101, ResNet152, and ConvNeXtLarge, an analysis is conducted for BLEU score as shown in the figures 15a, 15b, 16a and 16b. It can be observed that ConvNeXtLarge has the better Epoch vs. BLEU score curve for all the BLEU scores in comparison to other implemented algorithms. The BLEU score analysis of the existing approaches is presented in the table 4.

Further analysis is conducted for the performance analysis of the prediction of the area of interest by surrounding it with a box. Figure 17a shows the loss curve for the object detection box of the YOLOv8 algorithm, while Figure 17b shows the loss for object classification using the YOLOv8 algorithm. A lower box loss indicates that the model is precisely able to locate and estimate the size of the object in a given frame, and a lower classification loss indicates the model's ability to correctly identify the object in the frame. To analyze the correctness of the approach, a precision and recall analysis is conducted with respect to an increasing number of epochs. Figures 18a and 18b show epoch-wise progression of precision and recall, respectively, of the YOLOv8

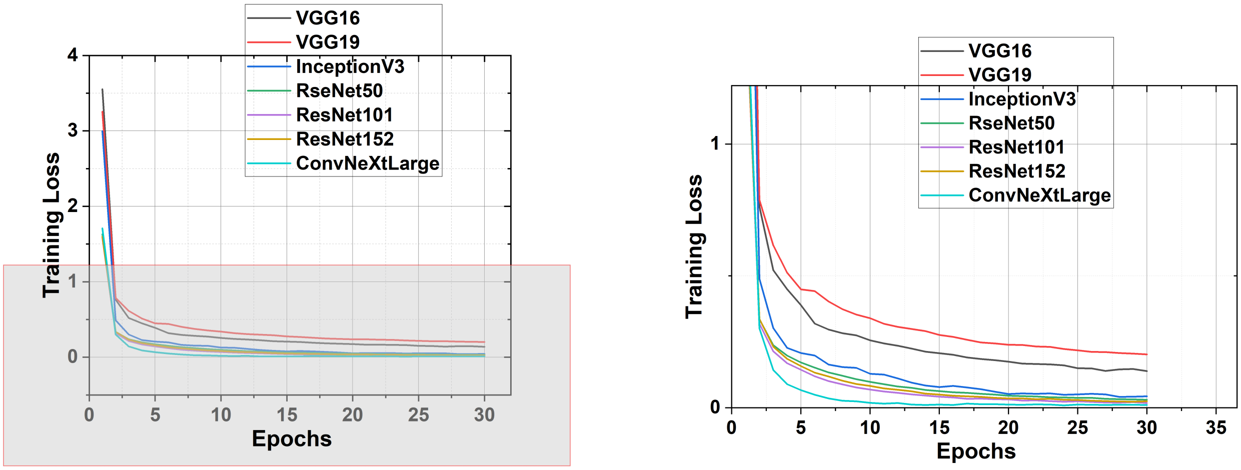
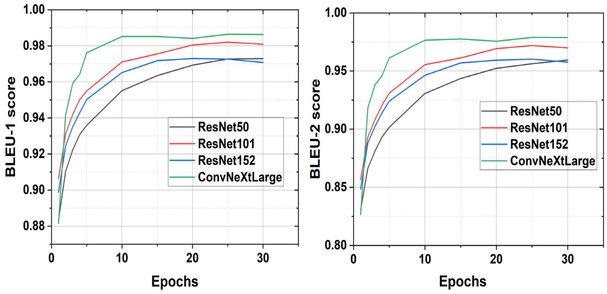
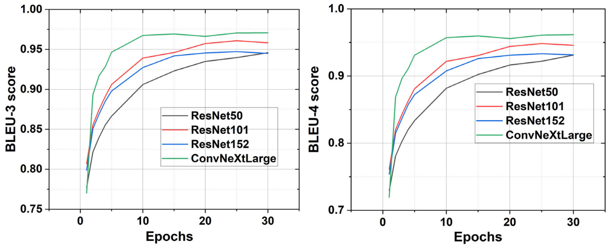


FIGURE 14. Training Loss of models used w.r.t number of epochs.



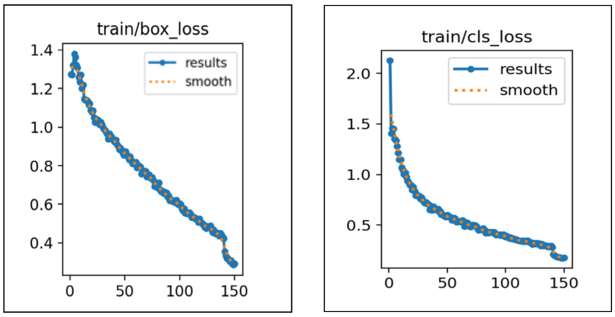
(a) BLEU-1 score vs Epochs (b) BLEU-2 score vs Epochs

FIGURE 15. Analysis of BLEU-1 and BLEU-2 score for applied models with increasing number of epochs.



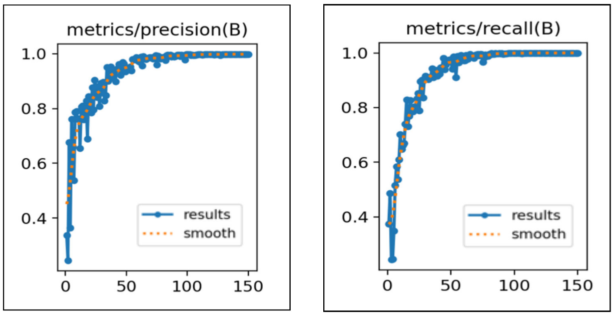
(a) BLEU-3 score vs Epochs (b) BLEU-4 score vs Epochs

FIGURE 16. Analysis of BLEU-3 and BLEU-4 score for applied models with increasing number of epochs.



(a) Box prediction Loss curve (b) Class prediction Loss curve

FIGURE 17. An analysis for prediction of area of interest and classification loss while training of the model.



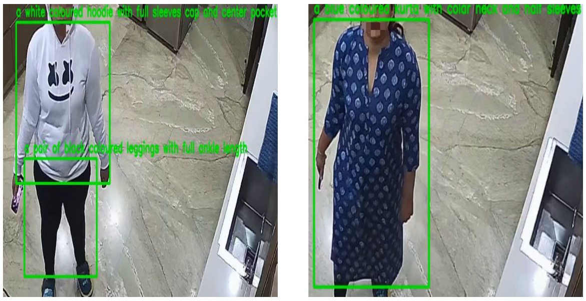
(a) Precision analysis (b) Recall analysis

FIGURE 18. Precision and recall analysis of the proposed approach with respect to increasing number of epochs.

algorithm in correctly detecting the object within the frame.

Figures 19a, 19b, 20a and 20b show the detection of clothing and the appropriate description that is generated using a video input. In addition, there is proper detection of both the boundary boxes of the top and bottom, which further proves the model's ability to correctly detect and generate captions in case there are multiple objects present in a frame, as shown in Figures 19a, 20a and 20b.

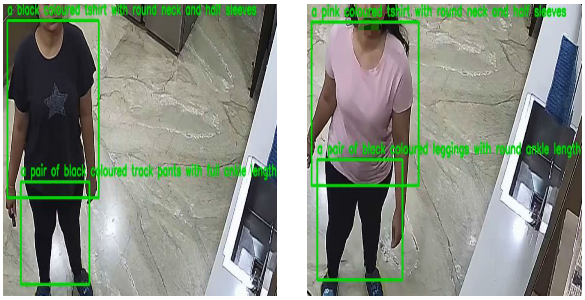
The time taken by YOLO to preprocess a single frame is around 3.0 ms. After this, YOLO takes around 10.5 ms to run the YOLO model on the frame. After this, the output is constructed for the frame, which highlights the detected objects in 1.5 ms. Following this, the frame is cropped based on the detected object's boundaries and resized according to the image feature extraction module, and this cropped object image takes about 135 ms to generate image features along with the caption of the object. Overall, it takes approximately 150 ms to caption an object in a single frame of a video.


a white coloured hoodie with full sleeves cap and center pocket
a pair of black coloured leggings with full ankle length
a blue coloured kurta with colar neck and half sleeves

(a) Video output for Hoodie & Leggings

(b) Video output for Indian dress i.e. Kurta

FIGURE 19. **A sample output of the image captioning process for two different dresses in a video containing a winter clothing and an Indian dress.**


a black coloured tshirt with round neck and half sleeves
a pair of black coloured track pants with full ankle length
a pink coloured tshirt with round neck and half sleeves
a pair of black coloured leggings with round ankle length

(a) Figure 18a: Video output for T-shirt & Track-pants

(b) Figure 18b: : Video output for T-shirt & Leggings

FIGURE 20. **A sample output of the image captioning process for two different dresses in a video containing t-shirts and leggings.**

V. CONCLUSION

This study offered a method for annotating apparel images by utilizing deep learning algorithms. The study's dataset was acquired by procuring 863 high-resolution apparel images from websites that do not require copyright, such as Unsplash, Pexels, Pixahive, and PxHere. To prepare the data for the model training phase, approaches such as image processing, data augmentation, and text processing were used. Convolutional Neural Networks (CNNs) and transfer learning were used in the proposed approach to extract reliable characteristics for precise and contextually relevant descriptions from photos of apparel. For feature extraction from the images, the system employed cutting-edge transfer learning techniques such as VGG16, VGG19, InceptionV3, ResNet50, ResNet101, ResNet152, and ConvNeXtLarge, which were pre-trained on the ImageNet dataset. Sequential captions were generated using an additional LSTM. Using the YOLOv8 architecture, an object identification module was integrated into the proposed system for the extended video captioning section in order to precisely identify and locate clothing items within images. The study employed assessment measures, namely BLEU-1, BLEU-2, BLEU-3, and BLEU-4, to assess the caliber of machine-generated translations through a comparative analysis with one or more reference translations. It was found that the ConvNeXtLarge model produced the best BLEU score values. Further study can be done in the future to enhance the models' efficacy by adding more features and looking at more sophisticated algorithms. The proposed approach lacks hardware implementation. The approach can also be extended by analyzing its hardware implementation along with its integration with wearable devices.

REFERENCES

[1] Z. Feng, B. Wen, and H. Deng, "Image caption generation for Dai ethnic clothing based on ViT-B and BertLMHeadModel," in *Proc. IEEE 14th Int. Conf. Softw. Eng.*, Jul. 2023, pp. 212–216.

[2] R. S. Al-Malki and A. Y. Al-Aama, "Arabic captioning for images of clothing using deep learning," *Sensors*, vol. 23, no. 8, p. 3783, Apr. 2023, doi: 10.3390/S23083783.

[3] Y. Tang, L. Zhang, Y. Yuan, and Z. Chen, "Improving fashion captioning via attributebased alignment and multilevel language model," *Appl. Intell.*, vol. 53, no. 24, 2023, Art. no. 308030821.

[4] P. Dwivedi and A. Upadhyaya, "A novel deep learning model for accurate prediction of image captions in fashion industry," in *Proc. 12th Int. Conf.*, 2022, pp. 207–212.

[5] I. Khurram and F. M. Shahzad, "DenseCaptionNet: A sentence generation architecture for finegrained description of image semantics," *Cognit. Comput.*, vol. 13, no. 3, 2021, Art. no. 595611.

[6] C. Park, M. Jang, J. Lee, and J. Kim, "Deep multi class-wise clothing attributes recognition for the elderly care robot environment," RO-MAN, 2020.

[7] G. Hacheme and N. Sayouti, "Neural fashion image captioning : Accounting for data diversity," 2021, *arXiv:2106.12154*.

[8] K. Tateno, N. Takagi, K. Sawai, H. Masuta, and T. Motoyoshi, "Method for generating captions for clothing images to support visually impaired people," in *Proc. Joint 11th Int. Conf. Soft Comput.*, 2020, pp. 1–5, doi: 10.1109/SCISISIS50064.2020.9322767.

[9] C. Cai, K.-H. Yap, and S. Wang, "Attribute conditioned fashion image captioning," in *Proc. IEEE Int. Conf. Image Process.*, May 2022, pp. 1921–1925.

[10] S. Wooders and R. Senanayake, "FashionModel: Mapping images of clothes to an embedding space," Aug. 2018. [Online]. Available: http://rsenapps.com/fashionmodel.pdf

[11] A. Verma, A. K. Yadav, M. Kumar, and D. Yadav, "Automatic image caption generation using deep learning," *Multimedia Tools Appl.*, vol. 83, no. 2, 2024, Art. no. 5305325.

[12] S. C. Gupta, N. R. Singh, T. Sharma, A. Tyagi, and R. Majumdar, "Generating image captions using deep learning and natural language processing," in *Proc. 9th Int. Conf. Rel., INFOCOM Technol. Optim., Trends Future Directions (ICRITO)*, Sep. 2021, pp. 1–4.

[13] A. Ghosh, D. Dutta, and T. Moitra, "A neural network framework to generate caption from images," in *Emerging Technology in Modelling and Graphics* (Advances in Intelligent Systems and Computing), vol. 937, J. Mandal and D. Bhattacharya, Eds., Singapore: Springer, 2020, doi: 10.1007/978-981-13-7403-6_17.

[14] Alzubi, Jafar A, R. Jain, P. Nagrath, S. Satapathy, S. Taneja, and P. Gupta, "Deep image captioning using an ensemble of CNN and LSTM based deep neural networks," *J. Intell. Fuzzy Syst.*, vol. 40, p. 4, Aug. 2021, doi: 10.3233/JIFS189415.

[15] B. Arystanbekov, A. Kuzdeuov, S. Nurgaliyev, and H. A. Varol, "Image captioning for the visually impaired and blind: A recipe for low-resource languages," in *Proc. 45th Annu. Int. Conf. IEEE Eng.*, Jul. 2023, p. 4, doi: 10.1109/EMBC40787.2023.10340575.

[16] Statista. (2023). *Fashion-Worldwide | Statista Market Forecast*. [Online]. Available: https://www.statista.com/outlook/emo/fashion/worldwide

[17] J. Qiu, F. P.-W. Lo, X. Gu, M. L. Jobarteh, W. Jia, T. Baranowski, M. Steiner-Asiedu, A. K. Anderson, M. A. Mccrory, E. Sazonov, M. Sun, G. Frost, and B. Lo, "Egocentric image captioning for privacy-preserved passive dietary intake monitoring," *IEEE Trans. Cybern.*, vol. 1, no. 1, pp. 1–14, Mar. 2023.

[18] S. Cho and H. Oh, "Generalized image captioning for multilingual support," *Appl. Sci.*, vol. 13, no. 4, p. 2446, Feb. 2023.

[19] (2017). *E-Commerce Poised to Capture 41% of Global Retail Sales By 2027Up From Just 18% in 2017*. [Online]. Available: https://www.bcg.com/press/31october2023-ecommerce-global-retail-sales

[20] *CVAT*. Accessed: Mar. 12, 2024. [Online]. Available: https://www.cvat.ai/

[21] G. Hacheme. (2023). *Hgilles06/Infashai*. Accessed: Mar. 12, 2024. [Online]. Available: https://github.com/hgilles06/infashai

[22] (2017). *COCO 2017 Dataset*. Accessed: Mar. 13, 2024. [Online]. Available: https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset

[23] *Flickr 8k Dataset*. Accessed: Mar. 12, 2024. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k

GOVIND AGARWAL is currently pursuing the Bachelor of Technology degree in computer science and engineering with Vellore Institute of Technology. His research interests include machine learning, deep learning, and data science.

KRITIKA JINDAL is currently pursuing the Bachelor of Technology degree in computer science and engineering with Vellore Institute of Technology. She has a persistent interest and curiosity for data science, machine learning, and deep learning.

ABISHI CHOWDHURY (Member, IEEE) received the B.E. degree from the University Institute of Technology, West Bengal, India, in 2011, the M.Tech. degree from the National Institute of Technical Teachers' Training and Research, Bhopal, Madhya Pradesh, India, in 2014, and the Ph.D. degree from the Visvesvaraya National Institute of Technology, Nagpur, India, in 2020. She is currently an Assistant Professor with Vellore Institute of Technology, Chennai, India. Her research interests include cloud computing, cloud resource scheduling, machine learning, and the Internet of Things.

VISHAL K. SINGH (Member, IEEE) received the bachelor's degree in information technology, in 2010, the master's degree in computer technology and application, in 2013, and the Ph.D. degree in information technology from the Indian Institute of Information Technology, Allahabad, India, in 2018. He is currently a Lecturer and is associated with the Networks and Communications Research Group, School of Computer Science and Electronics Engineering, University of Essex, Colchester, U.K. His research interests include the Internet of Things, wireless sensor networks, in-network inference, machine learning, and data analytics.

AMRIT PAL (Member, IEEE) received the B.Tech. degree from Kurukshetra University, Kurukshetra, India, in 2011, the M.Tech. degree from the National Institute of Technical Teachers' Training and Research, Bhopal, India, in 2014, and the Ph.D. degree from the Indian Institute of Information Technology Allahabad, India, in 2020. He was an Assistant Professor with the Centre for Advanced Studies, AKTU, Lucknow, India. He is currently an Assistant Professor with Vellore Institute of Technology, Chennai, India. His research interests include big data analytics, cloud computing, machine learning, and the Internet of Things.