 http://dx.doi.org/10.14500/aro.12335

AraFashion: A Novel Fashion Captioning Dataset Leveraging Attention-Based EfficientNet and xLSTM

Shams A. Ahmed† and Ahmed T. Abdulameer

Department of Information Technology Management, Technical College of Management-Baghdad, Middle Technical University, Baghdad, Iraq

***Abstract*–The significance of creating models that can produce precise textual descriptions of photographs has become apparent, particularly in specialized domains such as fashion. Arabic suffers from a severe shortage of publicly available resources, particularly fashion picture databases, in contrast to the wealth of databases and studies about the English language. This restricts the creation of Arabic language models and impedes scholarly research in this area. By creating a hybrid model for automatically producing Arabic descriptions of fashion photos, our study seeks to close this gap. Based on the EfficientNet-B4 architecture, this model incorporates an attention mechanism to extract visual features and, for the first time in this field, links it to an xLSTM module for text creation. This study produced a new dataset with Arabic captions called AraFashion; the Arabic descriptions were translated into English through Google Translate. Using real Arabic data improves the model's accuracy and realism, as seen by the model's top BLEU-1 score of 0.7335 for Arabic descriptions. This study suggests growing Arabic databases in the fashion industry and highlights the need to support the Arabic language in AI technology.**

***Index Terms*—Arabic Image Captioning, AraFashion, Dataset, EfficientNetB4, xLSTM**

I. Introduction

Given their potential benefits for e-commerce, clothing classification (Rawate, et al., 2022) and image captioning (Moratelli, et al., 2023) have recently received extensive research attention. However, due to the focus of research on English-language datasets (Liu, et al., 2016, Xiao, et al., 2017, Rostamzadeh, et al., 2018), studies lack a robust fashion database in Arabic. We can also compare and identify the performance of existing and future algorithms for Arabic fashion commentary using Arafashion. To help train deep learning models for feature extraction and fashion description generation, we developed this database. Due to the expansion of e-commerce, the rise in fashion sales, and the general interest in this area, we are directing future research in this direction. We are also releasing this database and using it as a guide for future studies. It allows us to compare various models and algorithms that extract information from images directly, revealing their advantages, disadvantages, and optimal approaches. We can then generate captions using text generation models and compare them to actual descriptions.

The new Arabic fashion dataset used in this study includes approximately 52,000 high-resolution images of men's and women's fashion items, along with comprehensive descriptions of product type, color, gender, usage, season, length, and sleeve length. Each description is provided and written manually. Each product includes approximately 6 to 7 images taken from different perspectives against a consistent background in uniform illumination. The images were collected from an open-source database (Kaggle). This paper's primary contributions are (1) statistics of the dataset created in this research, as well as its publication and availability for future research; (2) comparison with the available Arabic fashion dataset; (3) training the database using pre-trained models to identify and develop the best; and (4) the model in this research was developed by adding an attention layer to the feature extraction model, using an advanced model xLSTM to generate texts.

II. Related Work

Before talking about related datasets, we give an overview of the models that are used to create Arabic image captions.

A. Arabic Image Captioning

The two main processes in picture captioning are the encoder and the decoder. Encoders are used to extract features or objects from images or other visual input, and the process of creating a description of those features or objects is known as decoding (Pan, et al., 2020, Al-Malla, Jafar and Ghneim, 2022). In 2021, a new transformer-based model was introduced that extracts features from pictures using MobileNetV2 and EfficientNet. To obtain the word sequence for the construction of the caption, they also employ LSTM and GRU with attention. Objects Shared in Context (COCO) and Flicker8k are two databases to which the paradigm is applied, and obtained a BLEU-1 = 44.3. By developing cutting-edge natural language processing methods to streamline the Arabic language's structure, it made clear

ARO-The Scientific Journal of Koya University
Vol. XIV, No. 1 (2026), Article ID: ARO.12335. 10 pages
DOI: 10.14500/aro.12335
Received: 05 June 2025 Accepted: 11 December 2025
Regular research paper; Published: 15 March 2026
†Corresponding author's e-mail: dac2018@mtu.edu.iq

the future optimization mechanism (Sabri, 2021). In 2022, an effective deep learning model for Arabic picture annotation was proposed. It relies on the encoder–decoder architecture, which employs RESNet-101 for the encoder and LSTM for the decoder, and was trained on the Arabic Flickr8k dataset. He saw success with the complete application of back-propagation to develop soft attention. The corresponding value of BLEU-1 is 58.708. However, detailed control of attributes for each image or bias between different clothing categories was not addressed, leaving a research gap that could be explored in future studies (Lasheen and Barakat, 2022). In 2024, an updated version of the Flickr8k database was created, adding four annotations per image instead of three, to improve the diversity of training texts. The VGG16 and VGG19 models were used to extract features from images, and LSTM and GRU models were built to process text prediction sequences, improving the quality of description and achieving BLEU-1 = 40. However, the model did not focus on the detailed features of each image or address bias in the data, leaving a research gap that could be addressed in future studies (Ibrahim, Shati and Alsewari, 2024). In 2025, a model was created to generate feature-guided annotations for fashion photographs, allowing users to choose descriptive attributes such as color, pattern, and fabric type. The FACAD database was used to power a transformer-based encoder–decoder architecture that included visual and textual elements. The model performed better than standard models, with BLEU-1 = 0.62 and BLEU-4 = 0.54. The study found that including feature cues improves the accuracy of detail in descriptions, but it did not address the diversity of fashion categories or lessen data bias, indicating a possible research gap (Cai, Yap and Wang, 2025).

B. Related Datasets

The datasets that have already been utilized to generate fashion captions in Arabic and English are covered here:

- DeepFashion (Liu, et al., 2016) is a huge database of clothes. First off, with over 800,000 distinct fashion images, it is the largest clothes database to date. Second, DeepFashion offers comprehensive English-language information on clothing items.
- Fashion-Gen (Rostamzadeh, et al., 2018) with 293,008 photos. High-resolution photos taken in a particular studio setting are offered. Depending on the category, each fashion item is photographed from one to six different views. Descriptions of each item include the brand, designer, fashion season, and fashion designer.
- Fashion Captioning Dataset (FACAD) (Yang, et al., 2020), it contains 993K images, a clothing category might be "dress" or "T-shirt," for instance, whereas an attribute such as "pink" or "lace" offers more specific details about a specific piece. The final word from item titles is chosen to create the category list.
- Fashion-MNIST (Xiao, et al., 2017), 70,000 fashion items in 10 categories, shown in 28 by 28 grayscale photographs. Fashion-MNIST is a newly published dataset that has 7000 photos in each category. There are 10,000 photos in the test set and 60,000 in the training set. The goal of Fashion-MNIST's creation was to offer English subtitles for fashion photos.
- ArabicFashionData (Al-Malki and Al-Aama, 2023) to promote artificial intelligence research and provide automatic picture descriptions in Arabic, an area that is severely lacking in open resources. The database attempts to offer a visual linguistic resource in Arabic for characterizing fashion and clothing photographs. It is the first database in Arabic devoted to fashion image labeling. Each of the 79,115 photos has an Arabic-written description of the style.

III. AraFashion Dataset

With the help of the extensive clothes dataset AraFashion, we are giving back to the community. Among its many aspects is the size of the dataset, which includes more than 52,000 different fashion photos in 256 × 256 resolution. Colors, clothing styles, gender, season, length, usage, and sleeve length were all described in the Arabic annotations that were carefully applied to these photos. Seven primary attributes and 50 secondary attributes were used to classify each image in this dataset into 17 major categories that corresponded to the kind of clothing. The database contains 7836 images of men's fashion and 44,871 images of women's fashion, and is available to the research community. Fig. 1 represents an example of an image set with major categories for each image.

A. Image Collection

Fashion databases were searched, and then the fashion images were collected from the Kaggle database, which only included uncategorized fashion images. The 52,707 images contain a wide range of clothing categories, and the images include both men's and women's fashion. There are approximately 6 to 7 images of each item taken from different angles.

B. Image Annotation

To assist in identifying clothes and writing Arabic captions for them, the following information is the focus of the effort. The following are: (1) Massive features this information is essential for identifying elements of type, color, use, length, gender, season, and sleeve lengths; (2) availability because there are not many databases for fashion image captions in Arabic, they are being created for use in future research and scientific aspects; (3) evolution most scientific research has been focused on this area due to the continued growth and expansion of fashion e-commerce sites.

- Making Lists of Categories and Attributes: Category and attribute lists were manually created on the collected image set, based on carefully studied previous research methods. For clothing categories, clothing types were first extracted from the images (e.g., "dress"), resulting in 17 unique main category names, as shown in Fig. 2. The top 10 categories of fashion in the database (e.g., "Women's Red Formal Winter Long Dress with Long Sleeves").

http://dx.doi.org/10.14500/aro.12335

Fig. 1. Some images from the AraFashion database. Some of the main categories are: Pants, blouses, T-shirts, shirts, dresses, and shorts.

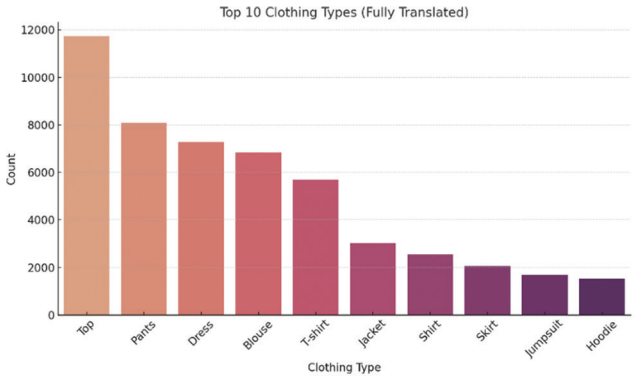

Fig. 2. The top 10 categories of fashion in the database and their distribution.

- Quality Control: The actions listed below were implemented to regulate the labeling quality. (1) Images were carefully reviewed to detect any damaged or unclear images. (2) After manually annotating the attributes, other people quickly checked to exclude or modify the written captions, which were actually modified and checked for consistency. (3) The handwritten texts were submitted to text cleaning algorithms to remove special characters, punctuation, and extra spaces, and to standardize the text format.

C. Benchmarks

The AraFashion dataset was used to create benchmarks for comparing the outputs of text generation models to human-annotated captions, and the BLEU, METEOR, ROUGE-L, CIDEr, and SPICE metrics were used to assess the generated captions.

This is the second Arab fashion challenge, and there is already an Arabic-language database of fashion photographs. Our current challenge is to generate captions for fashion photographs utilizing the existing database. Addressing data imbalance is essential, as it can lead to mild overfitting on larger classes. As Fig. 2 shows, there are significant differences in sample counts between classes, e.g., Top: 12,000 images versus Hoodie: 1800 images. To mitigate this disparity, data augmentation techniques were applied to the menswear images in the training set to enhance their representation.

IV. Comparison with Other Datasets

AraFashion statistics are as displayed in Table I. In contrast to other fashion datasets, such as (Liu, et al., 2016, Xiao, et al., 2017, Rostamzadeh, et al., 2018, Yang, et al., 2020, Al-Malki and Al-Aama, 2023), Ara Fashion is one of the first fashion databases that contains accurate descriptions in both Arabic and English. This dataset is published as a source for future research.

V. Challenge

Not only are we making a rich dataset available, but we are also starting a contest that uses our fashion dataset to categorize fashion images. As far as we are aware, this is the task's first challenge. In addition, we urge participants to make use of all the data in the collection, including lengths, colors, and fashion styles. We provide a framework that enables researchers to easily assess the effectiveness of their models by contrasting generated descriptions with actual ones. We will also include scores for the enhanced model suggested in this study for researchers to use in the challenge's final assessment in Arabic captions and English. We also launch the challenge to train English descriptions and obtain high accuracy.

VI. Experiments

In this section, the processing steps for texts and images for all models that have been worked on are discussed. These models' effectiveness is evaluated using the BLEU (1/2/3/4) (Papineni et al., 2002), METEOR) (Banerjee and Lavie 2005), ROUGE1/2/3 (Lin, 2004), SPICE (Anderson, et al., 2016), Cider (Vedantam, et al., 2015) metrics.

A. Data Augmentation

This is performed only on male images in the training dataset, aiming to address the imbalance between the number

TABLE I
COMPARISON BETWEEN ARABIC AND ENGLISH FASHION IMAGE DATASETS.

Datasets	Language	Images	Image size	Category	Attribute
MNIST Dataset (Xiao, et al., 2017)	English	70,000	28*28	10	---------
FACAD (Yang, et al., 2020)	English	993,000	1560*2392	78	990
Fashion-Gen (Rostamzadeh, et al., 2018)	English	293,008	1360*1360	48	---------
DeepFashion (Liu, et al., 2016)	English	800,000	700*1000	50	1000
ArabicFashionData (Al-Malki and Al-Aama, 2023)	Arabic	79,000	100*100	15	39
Ours AraFashion Dataset	Arabic and English	52,707	256*256	17	50

of men and women. The database contains approximately 8000 images of men compared to over 44,000 images of women. 30,000 new images are generated using geometric transformation techniques such as rotation, cropping, and brightness adjustment to ensure greater diversity in visual distribution, while preserving the original descriptions associated with the original images. After splitting and augmentation, there were 66,951 photographs in the training set, 7882 in the validation set, and 7874 in the test set.

B. Text Preprocessing

For word sequences of different lengths to fit the neural network, text preprocessing is necessary. (1) Text cleaning: Standardizing the text format and eliminating excessive spaces, punctuation, and special characters. (2) Encode words by assigning a unique index number to each one and generating a vector or integer for each caption. (3) The lengths of the caption vectors were measured, with the longest being 19. The shorter captions were padded with zeros using padding (Pre). Fig. 3 depicts the distribution of annotation length, with the majority falling between 12 and 14 words and relatively few short or long remarks.

C. Preprocessing Images and Feature Extraction

256*256 RGB images can be found in the database. Due to the specific nature of the model, preprocessing procedures will be employed. 52707 RGB-formatted images of men and women are included in the package.

VII. BASELINE MODELS FOR BENCHMARKING

Several cutting-edge pre-trained architectures were chosen as baselines in order to assess the efficacy of the suggested model. The trained models EfficientNetB4 (Fudholi, et al., 2021), ResNet50 (Hoeser and Kuenzer, 2020), ResNet152 (Su, et al., 2018), VGG16 (Pal, et al., 2020), and VGG19 (Geetha, et al., 2020) extracted features in batches for each data split (train, validation, and test) and saved them in the in.npy format. The Arabic text descriptions were cleaned and tokenized using Keras' Tokenizer. In addition, repeating picture characteristics were concatenated along the time axis, and the text input was embedded in an embedding layer of size 256. It is a combined representation that was passed through a layer of LSTM (Sameer, et al., 2023) or GRU (Dey and Salem, 2017) models for text generation. To guarantee a fair and consistent comparison, the same dataset and training conditions were used for both baseline model evaluation and training (LSTM and GRU) with 128 units. This procedure is done for every test image; the results were saved in an Excel file together with the image name, the actual Arabic description, and the expected description in order to facilitate human evaluation and further metric analysis (BLEU, ROUGE, etc.).

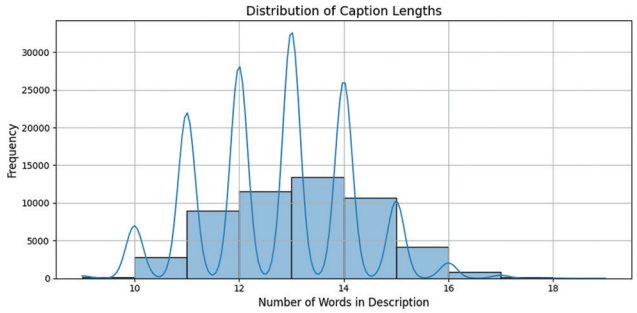

Fig. 3. Demonstrates the various caption lengths.

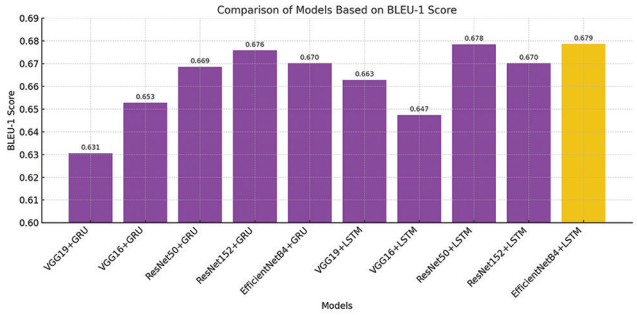

Fig. 4. A comparison of the BLEU-1 scores.

VIII. RESULTS AND COMPARISON

Several pre-trained convolutional models were assessed utilizing a consistent caption generation process in order to determine the best visual feature extractor for captioning Arabic fashion images. The performance scores of each model are shown in Table II using a variety of common assessment metrics, including BLEU, CIDEr, METEOR, ROUGE, and SPICE. Fig. 4 shows a comparison of BLEU-1 scores among the evaluated models.

The table shows the superiority of the EfficientNetB4 + LSTM model over the rest of the models, as it achieved the highest values in most metrics, such as BLEU-1 = 0.6786, BLEU-4 = 0.395, CIDEr = 0.656, and ROUGE-L = 0.702, in addition to the value of SPICE = 0.637. This superiority

 http://dx.doi.org/10.14500/aro.12335

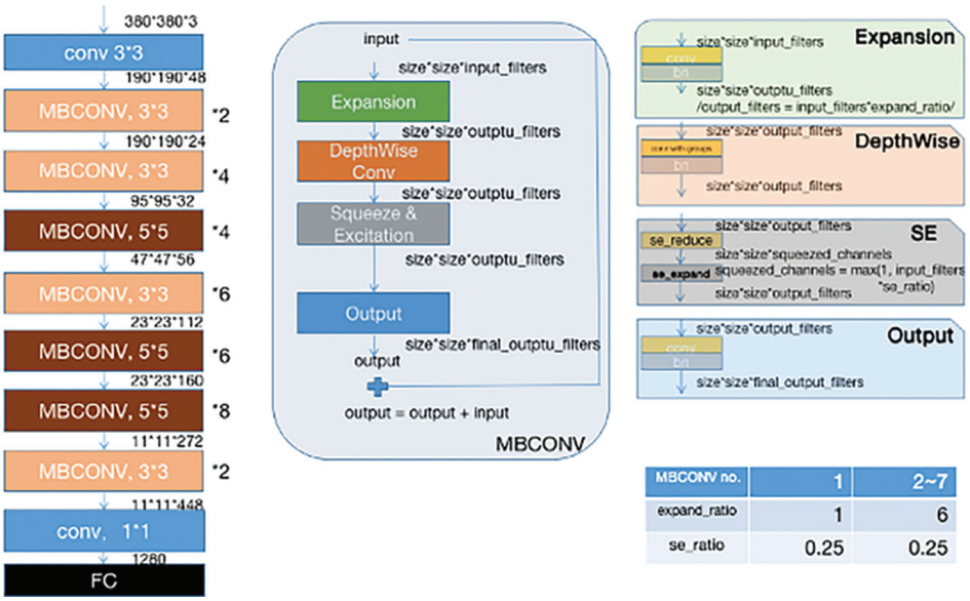

Fig. 5. EfficientNetB4 model structure (Tan and Le, 2019).

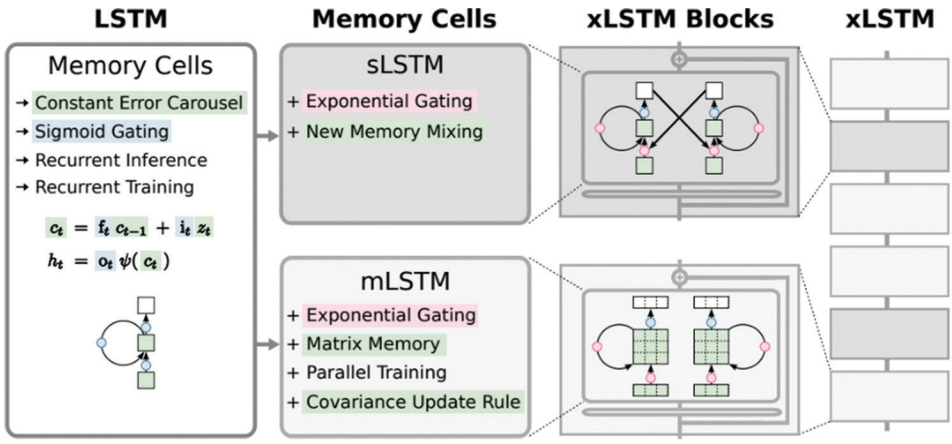

Fig. 6. Illustration of the xLSTM architecture (Beck, et al., 2024).

TABLE II
COMPARISON RESULTS BETWEEN THE BASIC TRAINED MODELS AND THE ARAFASHION DATASET.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	SPICE
VGG19+LSTM	0.662	0.571	0.473	0.378	0.642	0.632	0.693	0.525	0.689	0.621
VGG16+LSTM	0.647	0.5502	0.443	0.3362	0.6202	0.624	0.676	0.498	0.672	0.599
ResNet50+LSTM	0.678	0.585	0.486	0.3904	0.658	0.649	0.706	0.536	0.703	0.633
ResNet152+LSTM	0.6701	0.582	0.489	0.394	0.651	0.641	0.698	0.537	0.694	0.631
Efficient NetB4+LSTM	0.679	0.5906	0.493	0.395	0.656	0.648	0.706	0.544	0.702	0.637
VGG19+GRU	0.6305	0.533	0.425	0.323	0.601	0.609	0.662	0.482	0.656	0.586

is attributed to the architecture of EfficientNetB4, which relies on compound scaling to increase depth, width, and resolution in a balanced way, thus providing more accurate and representative features of the image content. For this reason, this model will be used for its development.

IX. EFFICIENTNETB4 MODEL

The input layer of EfficientNetB4 uses a 3 × 3 convolutional layer with a kernel filter of size 2 and a stride of 2 to upscale the incoming image, followed by a batch normalization layer and a swish activation function. In this step, the image's spatial dimensions are reduced, and a decent starting representation is produced. The network is made up of several MBConv blocks that adhere to a particular structure to get great efficiency: The expansion layer, which first expands the number of channels by a 1 × 1 convolution operation without changing the image size, may allow the model to enhance its representation power without significantly impacting computing performance. The depthwise convolution layer,

which comes next, applies a distinct convolution operation to each channel (depthwise separable convolution) as opposed to performing a whole convolution operation to all channels as in classical convolution. The foundation of the EfficientNetB4 design is the compound scaling principle, which strikes a balance between input resolution, depth, and width in order to provide excellent performance with great computational economy (Tan and Le, 2019). The architecture of the EfficientNetB4 model is illustrated in Fig. 5.

X. XLSTM

A recent study published in 2024 proposes xLSTM (Beck, et al., 2024) (Extended Long Short-Term Memory), a more sophisticated version of the classic LSTM model. By improving the information flow process within a single cell, this research seeks to address some of the drawbacks of the original LSTM architecture. Three separate LSTM cells – the front, back, and side – process data at every time step in xLSTM's multipath design, in contrast to regular LSTM, which uses a single path to transfer memory state. As a result, the model can store more complex temporal patterns and better understand long-term relationships between words. Furthermore, the gating mechanism has been enhanced to be more adaptable and sensitive to modifications in time series inputs, which helps to improve the retention of crucial context and lessen the tendency to forget critical information. The architecture of the xLSTM model is illustrated in Fig. 6.

XI. Attention Mechanism

By allowing the model to selectively focus on the most informative areas of the feature maps, the attention mechanism improves the performance of convolutional neural networks. This integration improves the model's performance on tasks such as fine-grained recognition and photo captioning, while also enhancing the model's ability to highlight significant visual aspects and capture global contextual linkages. Such hybrid architectures have shown improved accuracy and resilience by combining CNNs' spatial comprehension with the adaptive focus of attention mechanisms (Ruan and Zhang, 2024).

$$\text{Attention}\left(Q,K,V\right)=\text{Softmax}\left(\frac{QK^{T}}{\sqrt{d_{k}}}\right)V \tag{1}$$

The attention mechanism highlights the most important areas within the extracted feature maps, allowing the model to focus on the visual details most relevant to the characterization process. By combining attention with EfficientNetB4 features, the representation of contextual information and the accuracy of characterizing the fashion image content are improved.

XII. Proposed Model

After the above experiments on the pre-trained models, the highest result was obtained in almost all metrics on the EfficientNetB4 model with the LSTM model, which also obtained the highest results. Therefore, work will be done to develop the EfficientNetB4+LSTM model. Although integrating EfficientNetB4 with attention and xLSTM is theoretically straightforward, applying this integration to the task of characterizing Arabic fashion images represents a significant contribution. This task suffers from a lack of dedicated Arabic databases and the poor performance of general models when handling fashion details and Arabic language formulation. Our model bridges this gap by designing an architecture specifically tailored to the Arabic language and fashion image characteristics, along with creating AraFashion, the first specialized Arabic database in this field. Thus, our work contributes to the development of a new Arabic resource and the adaptation of a modern architecture to a previously unaddressed task. The proposed approach includes architectural improvements and focusing methods to improve the quality of feature representation and captions generation using an improved LSTM model, xLSTM (Beck, et al., 2024), which was recently developed in 2024 and tested in the field of fashion labeling for the first time in this research, to the author's knowledge. The following subsections provide detailed information on the model's components and architecture. Table III summarizes the training configuration of the developed model.

A. Architecture Model

To enhance the network's capacity to fine-tune and extract task-relevant characteristics, the final 50 layers of EfficientNetB4 were unfrozen for this experiment. In particular, a four-head multihead self-attention layer with a key dimension of 128 was employed. The Attention Layer works together with the EfficientNetB4 model to extract visual information from images. The attention layer gives more weight to the most significant sections of the image, so the model focuses on important information while creating captions. This improves the correctness of the generated text and ensures that it matches the image content. The model has the ability to record long-range dependencies and contextual interactions throughout the image because each attention head learns to focus on distinct kinds of spatial or semantic links inside the feature maps. After the attention mechanism, a projection layer is used to randomly disable neurons during training to prevent overfitting. A dropout rate of 0.3 further regularizes the network. The final feature vector is extracted using a fully connected (dense) layer with 1024 units and ReLU activation, yielding a compact and semantically meaningful representation appropriate for tasks such as visual semantic retrieval or caption synthesis. An additional dropout layer with a rate of 0.2 is also applied before the output. Finally, the sequences were converted into NumPy matrices. Text input was embedded in a 512-bit embedding layer, and recurring image features were concatenated along the time axis. Two xLSTM model layers made of 128 and 256 units, dropout (0.4), recursive dropout (0.4) following batch normalization, and an extra dropout layer (0.3), were fed the embedded representation. The following word in the

http://dx.doi.org/10.14500/aro.12335

sequence was predicted using a final TimeDistributed dense layer with Softmax activation. The model was trained during 20 epochs with a batch size of 64. The model synthesizes the caption by generating words one at a time using greedy decoding, then feeds the anticipated word back into the model until the maximum sequence length is attained. To aid in human evaluation and additional metric analysis, the findings were saved in an Excel file along with the image name, the actual Arabic description, and the expected description.

Fig. 7 represents the enhanced EfficientNetP4 model used to extract features from fonts, colors, and patterns in images. Fig. 8 represents the enhanced xLSTM model for generating descriptions of fashion images, which surpasses the standard LSTM due to its three in-sentence text understanding pathways. The unfrozen layers were trained with a learning rate of 1e-4. Although slightly higher than frequently recommended values (1e-5-1e-6), this choice was empirically found to give steady training, good convergence, and no substantial overfitting.

TABLE III
Training the Developed Model on Arabic Captions.

Element	Value/Description
Data split	Training 70%/Validation 15%/Test 15%
Loss function	Categorical Cross-entropy
Data augmentation methods	Rotate 20°, Flip horizontally, Random crop.
Batch size	64
Number of epochs	20
Learning rate	1e-4
Optimizer	Adam

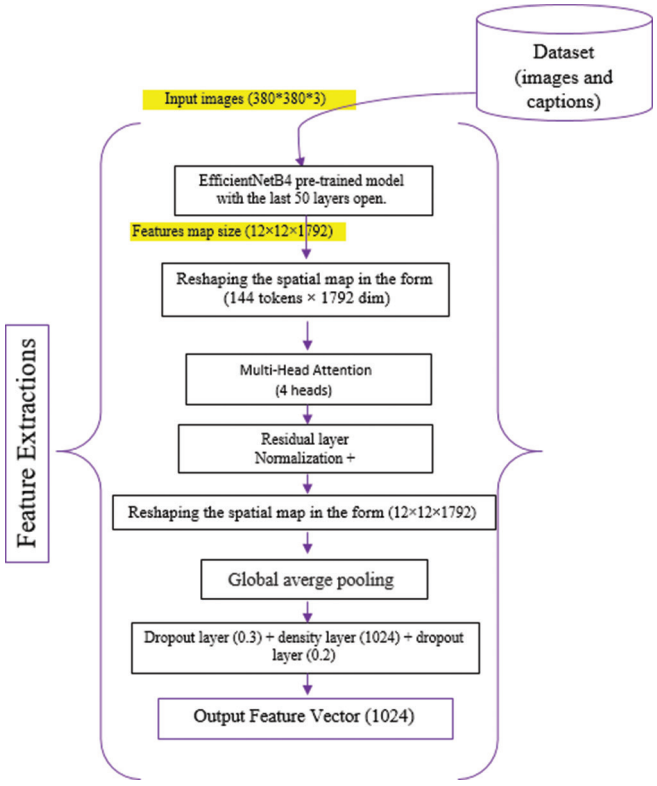

Fig. 7. The EfficientNetB4 model is a proposed developer for extracting features from images.

B. Results and Interpretation

After training the developed model on Arabic fashion captions, the results of these performance measures were obtained: The performance results of the developed model are presented in Table IV.

The Arabic model's output shows excellent performance in terms of description generation. The majority of the keywords in the projected descriptions match those in the actual descriptions, according to the BLEU-1 value of 0.73. The model's accuracy in matching longer word sequences is demonstrated by the logical evolution of the remaining BLEU values (up to BLEU-4 with 0.48). Good semantic understanding and effective language processing are also indicated by the high METEOR (0.71) and SPICE (0.68) values. This conclusion is supported by the CIDEr value (0.69), which shows that the generated descriptions and the reference are highly comparable. With respective scores of 0.75 and 0.75, ROUGE-1 and ROUGE-L demonstrate that the model accurately represents a linguistic sequence that is near the reference. Table V compares the results of the baseline designs with the developed model, demonstrating the developed model's gains in description accuracy. Fig. 9 shows the training and validation accuracy and loss curves of the developed model.

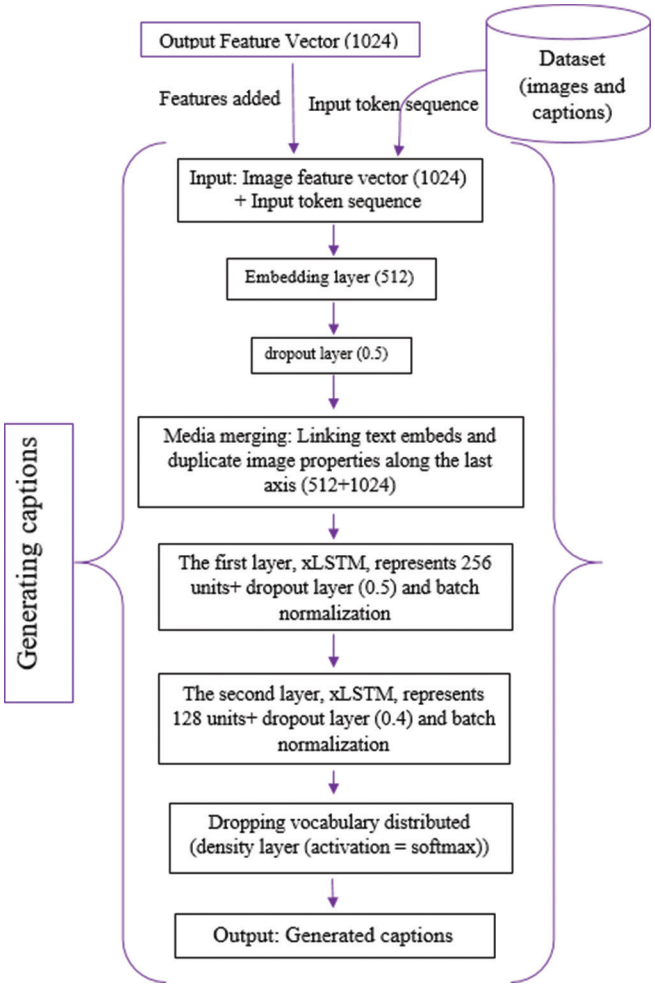

Fig. 8. The proposed improved xLSTM model for generating texts.

TABLE IV
RESULTS OF TRAINING THE DEVELOPED MODEL ON ARABIC CAPTIONS.

BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	SPICE
0.7335	0.6429	0.5605	0.4812	0.6933	0.7186	0.7553	0.5867	0.7516	0.6869

TABLE V
RESULTS OF TRAINING THE DEVELOPED MODEL ON ARABIC CAPTIONS/COMPARISON RESULTS BETWEEN THE BASIC TRAINED MODELS ON THE ARAFASHION DATASET AND THE DEVELOPED MODEL.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	SPICE
VGG19+LSTM	0.662	0.571	0.473	0.378	0.642	0.632	0.693	0.525	0.689	0.621
VGG16+LSTM	0.647	0.5502	0.443	0.3362	0.6202	0.624	0.676	0.498	0.672	0.599
ResNet50+LSTM	0.678	0.585	0.486	0.3904	0.658	0.649	0.706	0.536	0.703	0.633
ResNet152+LSTM	0.6701	0.582	0.489	0.394	0.651	0.641	0.698	0.537	0.694	0.631
Efficient NetB4+LSTM	0.679	0.5906	0.493	0.395	0.656	0.648	0.706	0.544	0.702	0.637
VGG19+GRU	0.6305	0.533	0.425	0.323	0.601	0.609	0.662	0.482	0.656	0.586
Our developed model	0.7335	0.6429	0.5605	0.4812	0.6933	0.7186	0.7553	0.5867	0.7516	0.6869

TABLE VI
RESULTS OF TRAINING THE DEVELOPED MODEL ON A DATASET (AL-MALKI AND AL-AAMA 2023).

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	SPICE
Al-Malki and Al-Aama, 2023	0.482	0.3614	0.30	0.133	-	-	-	-	-	-
Our efficient + attention with xLSTM	0.6721	0.560	0.37684	0.2272	0.5114	0.653163	0.7023	0.4924	0.7023	0.6426

TABLE VII
MODEL RESULTS BEFORE AND AFTER DATA AUGMENTATION.

Metrics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	SPICE
Result after data augmentation	0.7335	0.6429	0.5605	0.481	0.6933	0.7186	0.7553	0.5867	0.7516	0.6869
Result after data augmentation	0.7217	0.6286	0.5433	0.461	0.6768	0.707	0.7439	0.571	0.7404	0.6737

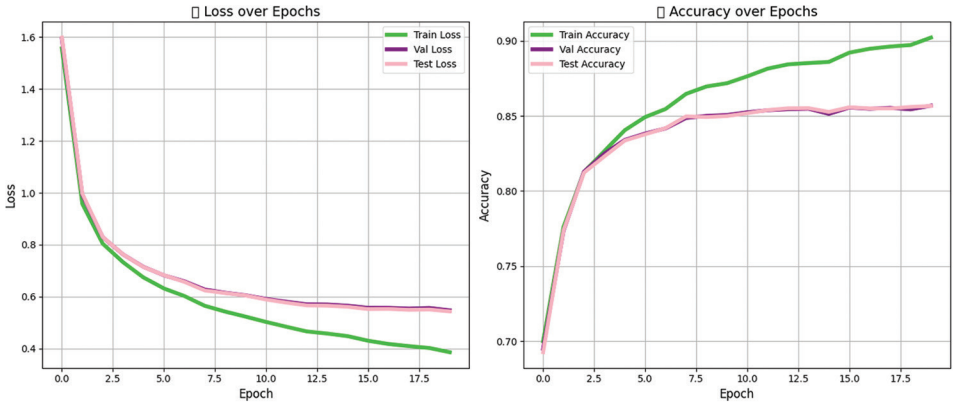

Fig. 9. Accuracy and loss curves of the developed model on Arabic translation.

The results showed a clear improvement when comparing the developed model to base models such as VGG16, VGG19, ResNet50, ResNet152, and EfficientNetB4. All performance metrics increased significantly, with BLEU-1 achieving a higher value of 0.7335 compared to the base models' highest value of 0.679. Long sentence quality also improved, as evidenced by the increase in BLEU-4 from 0.395 to 0.4812. CIDEr, METEOR, ROUGE, and SPICE metrics showed similar increases, indicating the developed model's ability to generate accurate, comprehensive, and coherent descriptions. This improvement is primarily attributed to the integration of an attention mechanism, which allowed the model to focus on the most important parts of the image when generating text, along with the use of multi-path xLSTM, which improved text sequencing and image feature integration, thus enhancing the quality of the resulting descriptions compared to traditional models.

Throughout the training phase, the suggested model's performance in Arabic demonstrates steady progress. The test, validation, and training sets' loss values all show a consistent decline, suggesting that the model is learning efficiently and not overfitting. The model's capacity to generalize well and extract precise semantic characteristics for Arabic texts is demonstrated by the accuracy curves, which again show a continuous rise, with training accuracy surpassing 90% and

validation and test accuracy stabilizing at roughly 86%, the training was 20 epochs to ensure the stability of training.

C. Comparative Analysis

The developed model was trained using the only previously available Arabic database (Al-Malki and Al-Aama 2023), which includes 79.115 images, of which 8.000 are labeled "men" and 71.115 are "women." The male-only dataset received targeted augmentation due to class disparity. Specifically, 54.000 new images were created for the training data for men using various manipulations, including flipping, brightness, rotation, and transformation. The dataset size increased to 133.115 images when added to the training set. To maintain data consistency, group-level partitioning was used to ensure that all images from the same product were kept in the same set (training, validation, or test). Training set 60.000, validation set 10.000, and test 9.115. To be a fair comparison, the developed model was used as is in all other settings. Since other evaluation scores, such as CIDEr, METEOR, ROUGE, and SPICE, were not reported in the previous work, the comparison mainly focuses on BLEU metrics. The results below were obtained: The comparison results with the previous dataset are presented in Table VI.

Our technique is clearly superior across all BLEU measures, as demonstrated by the comparison of the findings from our proposed model (EfficientNetB4 with attention and xLSTM) with the earlier work by Al-Malki and Al-Aama (2023). In particular, our model achieves a BLEU-1 of 0.6721 as opposed to 0.4820, which shows a significant improvement in word-level matching and unigram precision. Likewise, BLEU-2 increases from 0.3614 to 0.5601, demonstrating improved bigram sequence preservation ability. In addition, BLEU-3 improves from 0.30 to 0.3768, indicating improved phrase continuity and contextual understanding. Notably, BLEU-4 rises from 0.133 to 0.2273, indicating a markedly improved capacity to produce whole, cohesive sentences that closely match the reference descriptions. With a CIDEr score of 0.5114, METEOR of 0.6532, ROUGE-1 and ROUGE-L of 0.7023, ROUGE-2 of 0.4924, and SPICE of 0.6426, our model further confirms its robustness, despite the fact that the previous study does not disclose additional metrics, all of which taken together validate the model's ability to capture both lexical and semantic components of the captions.

D. The Impact of Data Augmentation on Model Performance

The lack of men's clothing data was compensated for by increasing their data to expand the representational space and reduce bias, thus improving the model's ability to generalize without affecting its training stability. The original dataset (52,707 images, including 6,783 men's and 44,871 women's) was used for the training, with each image labeled. The proposed model was trained with constant parameters. The results were then compared with a model trained after increasing the men's dataset to demonstrate that expansion reduces bias and improves generalization without affecting stability. The data were divided into the same groups, resulting in 36,901 training images, 7902 validation images, and 7904 test images.

After adding more male images, all metrics showed marked improvement: BLEU-1 rose from 0.721789 to 0.734, BLEU-4 from 0.461 to 0.481, CIDEr from 0.677 to 0.693, and METEOR, ROUGE-1/2/L, and SPICE increased by approximately one to two points.

XIII. Conclusion

The field of automatic Arabic picture descriptor creation has benefited greatly from this effort. Using a hybrid architecture, it builds a model by combining EfficientNet-B4 with an attention layer to extract visual features and xLSTM memory to generate precise descriptions. In addition to performing better when trained on earlier Arabic datasets, the new model also showed enhanced performance when evaluated on actual Arabic descriptions produced in this work, indicating its capacity for generalization and attaining high linguistic and visual quality. This work emphasizes the urgent need to provide trustworthy and comprehensive Arabic datasets that support a variety of dialects, dress styles, and cultural contexts in order to improve models and broaden their range of useful applications, as there are currently few specialist Arabic databases. Future research could employ multilingual Transformer or T5 models to enhance language production quality, particularly for lengthy and intricate statements. English descriptions found in the study database might potentially be used to train it.

XIV. Data Availability Statement

The data are published on the Kaggle website and can be used for research purposes only. The AraFashion dataset, which includes pictures of fashion items with captions in Arabic and English, can only be used for academic purposes with a specific non-commercial license. All textual annotations were handwritten, while the visual images were taken from the fashion dataset on Kaggle https://www.kaggle.com/datasets/shamsaliahmed20/arafashion (Shams, 2025). The performance comparison before and after data augmentation is presented in Table VII.

References

Al-Malki, R.S., and Al-Aama, A.Y., 2023, Arabic captioning for images of clothing using deep learning. *Sensors*, 23(8), pp.3783.

Al-Malla, M.A., Jafar, A., and Ghneim, N., 2022, Pre-trained CNNs as feature-extraction modules for image Captioning: An experimental study. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 21(1), pp.1–16.

Anderson, P., Fernando, B., Johnson, M., and Gould, S., 2016. Spice: Semantic Propositional Image Caption Evaluation. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, Berlin.

Banerjee, S., and Lavie, A., 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S., 2024 *XLSTM: Extended*

Long Short-Term Memory. [arXiv Preprint]

Cai, C., Yap, K.H., and Wang, S., 2025 Toward attribute-controlled fashion image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20, 280.

Ibrahim, H.S., Shati, N.M., and Alsewari, A.A., 2024. A transfer learning approach for arabic image captions. *Al-Mustansiriyah Journal of Science*, 35, pp.81-90.

Lasheen, M.T., and Barakat, N.H., 2022. Arabic image captioning: The effect of text pre-processing on the attention weights and the BLEU-N scores. *International Journal of Advanced Computer Science and Applications*, 13, pp.413-423.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. Association for Computational Linguistics, Pennsylvania.

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X., 2016. Deepfashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., and Cucchiara, R. 2023. Fashion-oriented image captioning with external knowledge retrieval and fully attentive gates. *Sensors (Basel)*, 23(3), pp.1286.

Pan, Y., Yao, T., Li, Y., and Mei, T., 2020. X-Linear Attention Networks for Image Captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Papineni, K., Roukos, S., Ward, T., and Zhu Bleu, W.J., 2002. A Method for Automatic Evaluation of machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Rawate, S.,Vayadande, K.,Chaudhary, S., Manmode, S., Suryavanshi, R., and Chanda, K., 2022 Fashion Classification model. Techno-Societal 2016. In: *International Conference on Advanced Technologies for Societal Applications*, Springer.

Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., and Pal, C., 2018. *Fashion-Gen: The Generative Fashion Dataset and Challenge* [arXiv Preprint].

Ruan, T., and Zhang, S., 2024. *Towards Understanding How Attention Mechanism Works in Deep Learning* [arXiv Preprint].

Sabri, S.M., 2021. *Arabic Image Captioning Using Deep Learning with Attention*. University of Georgia, Georgia.

Sameer, M., Talib, A., Hussein, A., and Husni, H., 2023. Arabic speech recognition based on encoder-decoder architecture of transformer. *Journal of Techniques*, 5, pp.176-183.

Shams, 2025. *AraFashion: A New Dataset for Fashion Caption*. Kaggle, San Francisco.

Tan, M., and Le, Q., 2019. *Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks*. International Conference on Machine Learning, PMLR.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D., 2015. Cider: Consensus-Based Image Description Evaluation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Xiao, H., Rasul, K., and Vollgraf, R., 2017. *Fashion-Mnist: A Novel Image Dataset for Benchmarking Machine Learning Algorithms.* [arXiv Preprint].

Yang, X., Zhang, H., Jin, D., Liu, Y., Wu, C.H., Tan, J., Xie, D., Wang, J., and Wang, X., 2020. *Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards*. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII 16, Springer.