Cairo University

Faculty of Computers and Artificial Intelligence

Artificial Intelligence Department

Pictalk Generator

Supervised by:

Dr/ Doaa Saleh

Dr/ Ghada Dahy

Alaa Ashraf	20200329
Arwa Sallam	20200067
ALMoatasim	20200865
Esraa Haytham	20200077
Mariam Mostafa	20200530
Yara Mohamed	20200630

July 2024

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

We would like to express our deepest appreciation to our advisor, Dr. Doaa Saleh, whose guidance, encouragement, and expertise have been invaluable throughout this project. Dr. Doaa Saleh consistently conveyed a spirit of enthusiasm for research and innovation, inspiring us to push the boundaries of our knowledge and skills. Their insightful feedback and unwavering support were crucial in overcoming various challenges and achieving the goals of this project. Without their support and persistent help, this project would not have been possible.

We extend our heartfelt thanks to Dr. Ghada Dahy who was our Teaching Assistant, for their invaluable support and guidance throughout this project. Their dedication and assistance were instrumental in refining our ideas and ensuring the successful completion of this work. Dr. Ghada Dahy 's enthusiasm and commitment greatly contributed to our learning experience.

We are also deeply grateful to our committee members for their insightful feedback and constructive suggestions, which greatly enhanced the quality of this work. Their dedication to academic excellence and their support were instrumental in the completion of this project. Their diverse perspectives and expertise provided a well-rounded foundation for our research.

In addition, we would like to thank the faculty and staff of the Faculty of Computer Science and Artificial Intelligence, especially Artificial Intelligence, for providing a stimulating academic environment and for their unwavering support. The resources and facilities provided by the college, including well-equipped computers and advanced computational resources, were essential for carrying out this project and running complex codes. Special thanks to Dr. Ghada Dahy for their administrative support and for facilitating various aspects of our research.

A special note of thanks goes to our fellow students and colleagues, whose camaraderie and collaborative spirit made the journey enjoyable and enriching. The discussions, brainstorming sessions, and mutual support were invaluable in refining our ideas and overcoming obstacles.

Finally, we extend our heartfelt thanks to our families and friends for their encouragement and understanding during this journey. Their continuous support kept us motivated and focused, and their belief in us was a source of strength and inspiration. This accomplishment is as much theirs as it is ours.

ABSTRACT OF THE GRADUATION PROJECT

In the realm of law enforcement and support for visually impaired individuals, our project leverages advanced artificial intelligence techniques, specifically a transformer model, to develop an innovative image captioning and image visualization system by giving text to the user.

This system improves accessibility for visually impaired individuals by interpreting images and generating descriptive sentences that enrich their understanding of the surrounding environment. Central to our approach is the integration of the Flickr 8k Dataset, a comprehensive collection of images annotated with descriptive captions, which ensures robust training and precise interpretation of visual content. This aids in their testimony in courts and hearing as witnesses in cases.In addition to leveraging the Flickr 8k Dataset, we have developed an additional dataset that complements and enhances our project's objectives. This dataset is specifically tailored to align closely with our innovative approach in image captioning. By creating this supplementary dataset, we ensure that our artificial intelligence model receives diverse and relevant data points that further refine its ability to interpret and describe images accurately.

In addition, our project expands its scope beyond image captioning to include the generation of verbal descriptions that aid law enforcement in identifying individuals. Textual descriptions are encoded and integrated with latent vectors and images from the CelebA Dataset, renowned for its detailed facial attributes. This integration facilitates the training of Generative Adversarial Networks (GANs), specifically Deep Fusion GAN architectures. These architectures employ deep fusion blocks to directly synthesize high-resolution images while seamlessly incorporating textual cues and visual features. The result is images that exhibit enhanced realism and coherence, closely aligning with textual descriptions.

This aids police investigations and the testimony of witnesses by accurately describing the features that might identify a criminal, thus supporting law enforcement in gathering evidence and hearing testimonies.

Moreover, our project incorporates specialized capabilities for the Arabic language, enhancing the system's applicability in Arabic-speaking regions. This localization not only facilitates accurate interpretation and description of images in Arabic but also ensures cultural sensitivity and relevance in diverse linguistic contexts. The integration of Arabic language proficiency further extends the usability of our AI model, empowering law enforcement personnel and visually impaired individuals alike with enhanced tools for image understanding and accessibility. This dual approach enhances the accuracy of image interpretation and provides law enforcement agencies with sophisticated tools for facial recognition and suspect identification. aiming to empower law enforcement personnel with advanced AI capabilities tailored for specific investigative needs.

CHAPTER 1: INTRODUCTION

1.1 Main problem:

The project addresses a significant issue related to the challenges visually impaired individuals face when dealing with crime scenes or participating in court trials. Since visually impaired eyewitnesses can provide valuable insights and direct testimonies, their participation in investigations is crucial. However, the main challenge lies in their ability to interpret visual evidence and interact with environments that heavily rely on visual information.

Generating images of criminals poses several challenges, particularly in law enforcement and investigative contexts. Traditional methods often rely on eyewitness descriptions, which can be subjective and prone to inaccuracies. This reliance can lead to challenges in accurately depicting suspects, especially when visual evidence is limited or of poor quality. Moreover, the need for generating lifelike images that align with textual descriptions adds another layer of complexity, requiring advanced AI techniques like Generative Adversarial Networks (GANs). These models aim to synthesize realistic facial images based on textual descriptions provided by eyewitnesses or victims, aiming to aid in quicker suspect identifications and arrests. However, ensuring the accuracy and reliability of these generated images remains a critical concern, as they play a pivotal role in criminal investigations and judicial proceedings.

1.2 Problem we suffer:

During the development of our image captioning model, several challenges were encountered that required significant troubleshooting and problem-solving efforts. One major obstacle was the integration of custom layers and the handling of complex tensor shapes, particularly with the TransformerDecoderBlock. Issues related to the dimension mismatches in masks used for attention mechanisms caused recurrent errors. Ensuring that the masks had the correct shape and type was crucial but non-trivial, as the causal and padding masks needed precise alignment for the multi-head attention layers to function correctly. Additionally, handling the data flow between the CNN encoder and Transformer-based decoder demanded careful attention to detail to maintain consistent tensor shapes throughout the training process. These issues collectively extended the development timeline but ultimately led to a deeper understanding of advanced model architectures and tensor operations in TensorFlow.

In the GAN model we suffer from finding a suitable dataset and the model is complex to training it in our laptops, so we wait to run in faculty labs.

1.3 Objective of the project:

The project represents a pioneering effort in leveraging artificial intelligence (AI) to address critical challenges in both accessibility and law enforcement.

1. **Accessibility for Visually Impaired Individuals:** One of the primaries focuses of the project is to empower visually impaired individuals by providing them with tools that enhance their ability to interact with visual information. The image captioning system developed aims to transform how visually impaired individuals perceive and interpret crime scene images. By generating detailed textual descriptions, the system enables them to independently understand their surroundings and contribute valuable insights to investigations. This capability not only supports their inclusion in legal processes but also promotes their overall independence and safety in everyday life.
2. **Advanced AI Techniques in Law Enforcement:** The project utilizes state-of-the-art AI techniques, particularly GANs, to enhance law enforcement capabilities. The Deep Fusion GAN architecture specifically enables the generation of highly realistic facial images from textual descriptions. This innovation facilitates quicker and more accurate suspect identifications, which is crucial in accelerating criminal investigations and improving operational efficiency for law enforcement agencies. By automating and enhancing the process of generating visual evidence, the project aims to reduce investigative timelines and increase the likelihood of successful case resolutions.

1.4 Proposed models:

captioning model:
The model takes an image as input, likely represented as a tensor of pixel values then A positional encoding layer is added to the extracted features. This layer injects information about the relative positions of different elements within the image, then the encoded image features are passed through a stack of Transformer encoder layers. These layers use self-attention mechanisms to allow the model to attend to different parts of the image features and capture long-range dependencies between them. This helps the model to understand the global context of the image. The decoder takes the encoded image features and an embedded start token as input. It uses a stack of Transformer decoder layers to generate the caption word by word. Each decoder layer attends to both the previous decoder outputs and the encoded image features, allowing it to generate captions that are relevant to the image content.
Image Visualization model:
Firstly, the textual descriptions of facial attributes are encoded into a numerical representation using techniques such as word embeddings or recurrent neural networks. This encoding step transforms the textual input into a format that can be processed by the subsequent stages of the algorithm. Next, them encoded text is fed into the generator network, which consists of convolutional layers That are responsible for translating the textual input into a visual representation of a facial image. Through a series of mathematical operations and transformations, the generator gradually constructs an image that corresponds to the provided textual description. Concurrently, a discriminator network is employed to evaluate the realism of the generated facial images. The discriminator is trained to distinguish between real facial images and those generated by the generator. This adversarial training process provides feedback to both the generator and discriminator networks, helping them improve their performance over time.

1.5 The main contraption of the modules:

Image to Text (Image Captioning):

English Model: Often involves Deep Fusion Generative Adversarial Networks (DF GANs), where textual descriptions serve as inputs to generate corresponding images. DF GANs consist of a generator network that synthesizes images and a discriminator network that distinguishes between real and generated images, ensuring the realism of generated outputs. The deep fusion blocks in DF GANs allow for the integration of textual cues and visual features, enhancing the coherence and detail of the generated images.

Arabic Model: In Text to Image (Image Visualization), the Arabic model uses translations of Arabic textual descriptions into English or another language compatible with English model architectures. This ensures effective generation of images from translated texts. By utilizing Generative Adversarial Networks (GANs) trained on Arabic texts and images, the model focuses on translating Arabic nuances into a language suitable for existing English models.

Text to Image (Image Visualization):

English Model: Typically utilizes advanced deep learning models, such as convolutional neural networks (CNNs) for image feature extraction and transformer models for generating textual descriptions. CNNs extract visual features from images, which are then fed into transformers to generate descriptive captions. Transformers, with their attention mechanisms, ensure the generated captions are coherent and contextually relevant.

Arabic Model: Adapts similar deep learning architectures, fine-tuned on Arabic datasets to ensure accuracy in interpreting and describing images in Arabic. This includes using CNNs for feature extraction and transformer models for generating captions. Transformers are trained directly on Arabic text data, ensuring that the generated descriptions are accurate and culturally appropriate

1.6 Summary for all chapters:

CHAPTER 2: Literature Review

- History of Criminals: Discusses the historical context and evolution of criminal records and identification methods.
- Translation from Image to Text: Reviews existing literature and methods for translating visual content (like images) into textual descriptions.
- Translation from Text to Image: Examines research and techniques for generating images based on textual descriptions.

CHAPTER 3: Proposed Model

- Overall Model (App): Introduces the comprehensive architecture of the proposed application, highlighting its components and workflow.

- Image Caption Module: Details the design and functionality of the module responsible for generating descriptive captions from images.
- Image Visualization Module: Explores the module dedicated to visualizing textual descriptions through the generation of realistic images.

CHAPTER 4: Experimental Results

- Datasets: Provides detailed insights into the datasets used, with separate sections covering each dataset's characteristics and relevance to the project.
- Results of Image Caption Module: Presents experimental outcomes and evaluations from the image captioning module.
- Results of Image Visualization Module: Discusses findings and outcomes from the image visualization module.
- Application of the Proposed Model: Demonstrates practical applications and use cases of the developed model in real-world scenarios.
- Discussion: Analyzes and interprets the experimental results obtained in Chapter 4, discussingtheir implications, strengths, and limitations.

CHAPTER 5: Conclusion

- Summarizes the key findings and contributions of the project.
- Provides concluding remarks on the overall success of the proposed model and its potential impact.

CHAPTER 2: LITERATURE REVIEW

2.1 History of using AI in crime investigations:

The History of Using AI in Crime Investigations Artificial Intelligence (AI) has rapidly evolved over the past few decades, making significant strides in various domains, including crime investigations. The application of AI in this field is not only transforming the ways in which law enforcement agencies operate but also redefining the essence of criminal investigations altogether. This essay will delve into the historical context, influential figures, and impact of AI in crime investigations while addressing its positive and negative aspects and future developments.

The use of AI in crime investigations gained mainstream attention with the advent of facial recognition technologies and data analytics in the 1990s. One of the most notable early implementations was the integration of AI-driven systems in surveillance networks to identify and track suspect movements. This period also witnessed the use of rudimentary machine learning algorithms to sift through vast amounts of data to predict crime patterns and hotspots.

Fast forward to the 21st century, and AI tools like natural language processing (NLP), machine learning, and deep learning have enhanced crime investigation methodologies. Tools like PredPol (Predictive Policing) have emerged, employing machine learning algorithms to analyze crime data and predict future criminal activities.

Key Figures and Influential Individuals several individuals have significantly contributed to the development and application of AI in crime investigations. One such pioneer is Geoffrey Hinton, often referred to as the "Godfather of Deep Learning." His research and development of neural networks have provided the foundation for many AI applications used in law enforcement today.

Andrew Ng, a prominent figure in AI, has also been instrumental in advocating for the ethical use of AI and has been a key proponent of integrating AI technologies in various sectors, including crime investigations. His work on machine learning and data science has been pivotal in refining the algorithms used in predictive policing.

Another notable figure is Timnit Gebru, an AI researcher who has highlighted the importance of ethical considerations in AI applications. Her work has been crucial in ensuring that AI systems used in crime investigations do not perpetuate existing biases or result in unfair targeting of certain demographic groups.

Impact and Perspectives The integration of AI into crime investigations has yielded significant positive outcomes. AI-driven data analytics enables law enforcement agencies to analyze vast amounts of data rapidly, identifying patterns and potentially linking seemingly unrelated incidents to uncover crime networks. Predictive policing has also shown promise by potentially enabling more efficient allocation of police resources to areas identified as high-risk zones.

2.2 Translate from image to text:

Image captioning has evolved over the years, with various techniques being explored to generate descriptive and coherent captions for images. This section provides an overview of different approaches leading up to the use of Transformers in image captioning.

Template-based approaches have fixed templates with a number of blank slots to generate captions. In these approaches, different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled. Farhadi et al. [1] use a triplet of scene elements to fill the template slots, while Li et al. [2] extract phrases related to detected objects, attributes, and their relationships. Kulkarni et al. [3] adopt a conditional random field (CRF) to infer objects, attributes, and prepositions. These methods generate grammatically correct captions but may lack flexibility.

Retrieval-based image captioning involves retrieving captions from a set of existing captions. Similar images and their captions are selected as candidate captions, from which captions for the query image are chosen. While these methods produce syntactically correct captions, they may not generate image-specific and semantically accurate captions.

Novel captions can be generated from both visual space and multimodal space. A general approach of this category is to analyze the visual content of the image first and then generate image captions from the visual content using a language model. Kiros et al. [4] apply a CNN for image feature extraction and use a multimodal space for joint representation learning and caption generation. Karpathy et al. [5] propose a deep multimodal model for bidirectional image and sentence retrieval. Chen et al. [6] introduce a multimodal space-based method for generating novel captions and restoring visual features from descriptions. These methods leverage deep learning techniques and generate more accurate and novel captions.

More recently, Transformer-based approaches have been applied to image captioning. Zhu et al. [7] employ a standard Transformer architecture, with a ResNext CNN as the encoder and the image features as keys and values in the decoder. Herdade et al. [8] introduce the object relation transformer, incorporating spatial relationship information using geometric attention. Huang et al. [9] introduce an attention-on-attention module (AoA) in the encoder, extracting feature vectors of objects and applying self-attention. Li et al. [10] propose the Entangled Attention (ETA) transformer, aiming to bridge the semantic gap by jointly exploiting semantic and visual information. He et al. [11] propose an image transformer, adapting the original transformer layer to the structure of images by increasing its width.

These advancements in image captioning techniques have paved the way for the development of our architecture, which leverages the power of Transformers and self-attention mechanisms to generate accurate and contextually relevant captions by capturing semantic relationships and spatial dependencies between images and captions.

2.3 Translate from text to image:

Throughout the early stages of GANs, the initially proposed and following models were evaluated for specific tasks, e.g., images of faces or bedrooms. Although the generated samples look promising as realistic, these models would only be trained with samples of limited interest. Next, researchers were interested in employing non-L2 loss modeling techniques. Even though these techniques are common in image-to-image regression tasks, they should be regularly taken with caution, since the non-L2 approach requires the use of approximate inference, and in most cases, model fitting is more time-consuming. Eventually, better results were produced by combining a variety of training procedures within the independent models or by end-to-end integration with learned adversarial objectives using mature architecture pairs.

Generative Adversarial Networks (GANs) are deep neural network architectures, which were first introduced by Goodfellow, et al., in the "Generative Adversarial Nets" paper, published at NIPS in 2014. Since the date of their introduction, the research that is fueled by GANs has embraced and is producing a rapidly increasing body of state-of-the-art results in generative modeling activities. More than that, GANs can be found involved in a wide variety of other tasks related to and beyond classic computer vision and image manipulation. The notable thing about GANs is their hybrid structure, where a generator and a discriminator are collaborating. Despite being a recent concept, numerous GAN versions have been developed by researchers at a jaw-dropping pace, targeting many different areas or problems.

Generative adversarial networks (GANs) are a type of generated model which targets to produce a realistic sample, it consists of two neural networks, namely the generator and the discriminator. The generator receives a random noise vector as input and aims to create fake samples that closely resemble real samples. The discriminator learns to differentiate between real samples and fake samples generated by the generator. Conditional GAN is a type of GAN that can generate data conditioned on given conditions. Control GAN is a type of GAN that contains classifier with generator and discriminator. Control GAN aims to separate the classifier from the discriminator, enabling the use of DA (Data Augmentation) without impeding GAN training by leveraging DA solely for independent classifiers. Self-Attention Generative Adversarial Networks (SAGANs) introduces a self-attention mechanism. The self-attention module helps with modeling long range, multi-level dependencies across image regions. Armed with self-attention, the generator can draw images in which fine details at every location are carefully coordinated with fine details in distant portions of the image. StackGAN generates high-resolution images by stacking multiple generators and discriminators and provides the text information to the generator by concatenating text vectors as well as the input noises. AttnGAN introduces the cross-modal attention mechanism to help the generator synthesize images with more details. MirrorGAN regenerates text descriptions from generated images for text-image semantic consistency. SD-GAN: it employs the Siamese structure to distill the semantic commons from texts for image generation consistency. DM-GAN introduces the Memory Network to refine fuzzy image contents when the initial images are not well generated in stacked architecture.

Discriminators suffer from the similar features in the facial images, such as round facial contours, two eyes, one mouth, etc. The similarities among the training data make it easier for the generator to create realistic images that can deceive the discriminator. Even if the generated faces lack accurate attributes, the discriminator may still accept them as real. This imbalance complicates further optimization of the generator. Additionally, when generating images from fMRI data, ensuring consistency is crucial, as a person can only perceive one face at a time. To enhance the generator's ability to produce high-quality and consistent images, we introduced two methods to strengthen the discriminator. To address this issue, some researchers introduced Double-Flow GAN [1].

Deep Fusion GAN (DF-GAN) is much different from previous methods. First, it generates high-resolution images directly by a one stage backbone in one generator, it avoids the entanglements between different generators. Unlike that, happen in stack GAN that have 3 backbones that have discriminators and generators. Second, it adopts a Target-Aware Discriminator to enhance text-image semantic consistency without introducing extra networks. Third, it fuses text and image features more deeply and effectively through a sequence of DFBlocks to make use of all text information. Compared with previous models, our DF-GAN is much simpler but more effective in synthesizing realistic and text-matching images, so we decided to use DFGAN as it generates high-resolution images with high consistency between image and text.

CHAPTER 3: PROPOSED MODEL

This chapter outlines the proposed model architecture for our mobile application, which integrates two core components: an image captioning model and an image generation model. The image captioning model is based on a transformer architecture, while the image generation model utilizes a Deep Fusion Generative Adversarial Network (DF-GAN). Each model is described in detail, along with the specific components and their roles.

3.1 The overall model:

The mobile application is designed to perform two primary tasks: generating descriptive captions for input images and creating images based on textual descriptions. These functionalities are enabled by two distinct models:

- **Image Captioning Model**: This model analyzes input images and generates descriptive textual captions. It leverages a transformer-based architecture to process image features and produce coherent and contextually relevant captions.
- **Image Generation Model (DF-GAN)**: This model synthesizes images from textual descriptions using a Deep Fusion Generative Adversarial Network (DF-GAN). The DF-GAN consists of a generator and a discriminator, which work together to produce high-quality images that match the given descriptions.

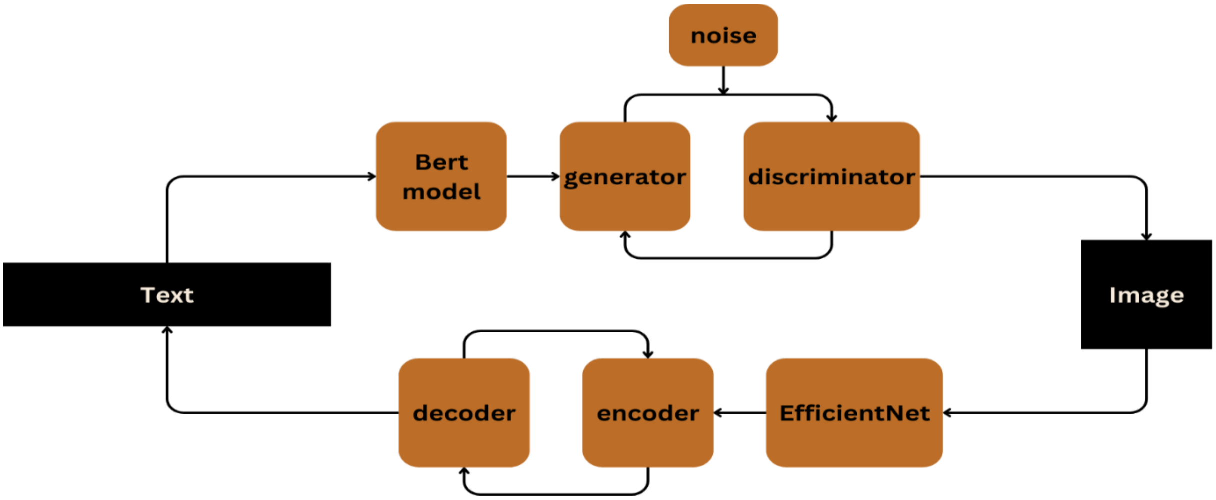


Figure 9: the overall system

System Flow and Interaction

The system is designed to handle user inputs and deliver outputs in a streamlined manner. Here is a step-by-step explanation of the data flow and interactions between the components:

- **User Input**:
  - The user can either upload an image or enter a textual description through the mobile application interface.
- **Image Captioning (If Image is Uploaded)**:
  - **Preprocessing**: The uploaded image undergoes preprocessing steps, including resizing and decoding, to prepare it for feature extraction.
  - **Feature Extraction**: The preprocessed image is passed through an EfficientNetV2B3 CNN to extract high-level features.
  - **Caption Generation**: The extracted features are fed into the transformer-based encoder-decoder architecture. The encoder processes the features, and the decoder generates the caption one token at a time until an end token is produced.
  - **Output**: The generated caption is displayed to the user.
- **Image Generation (If Text Description is Provided)**:
  - **Text Encoding**: The input text description is converted into a fixed-length sentence vector using a text encoder.
  - **Image Synthesis**: The sentence vector is passed to the DF-GAN's generator, which progressively refines the image features through multiple UPBlocks to create a high-resolution image.
  - **Discriminator Evaluation**: The synthesized image is evaluated by the discriminator to ensure it closely matches the input text description. The adversarial loss and matching-aware gradient penalty guide the generator to improve image quality.
  - **Output**: The generated image is displayed to the user.

The overall model architecture of the proposed mobile application successfully integrates advanced machine learning techniques to deliver dual functionalities of image captioning and image generation. The careful design and seamless integration of these models ensure high performance, versatility, and a superior user experience. The following sections will provide detailed descriptions of the image captioning and image generation models, exploring their architectures, components, and implementation in depth.

3.2 Image caption module:

The image captioning model is designed to generate descriptive captions for given images, leveraging a combination of convolutional neural networks (CNN) for feature extraction and transformer architecture for sequence modeling. The model aims to bridge the gap between visual data and natural language, providing a comprehensive understanding of the depicted scenes.

The data flow begins with the input image, passes through various stages of processing and feature extraction, and culminates in the generation of a textual description. Below is a high-level diagram illustrating the architecture:

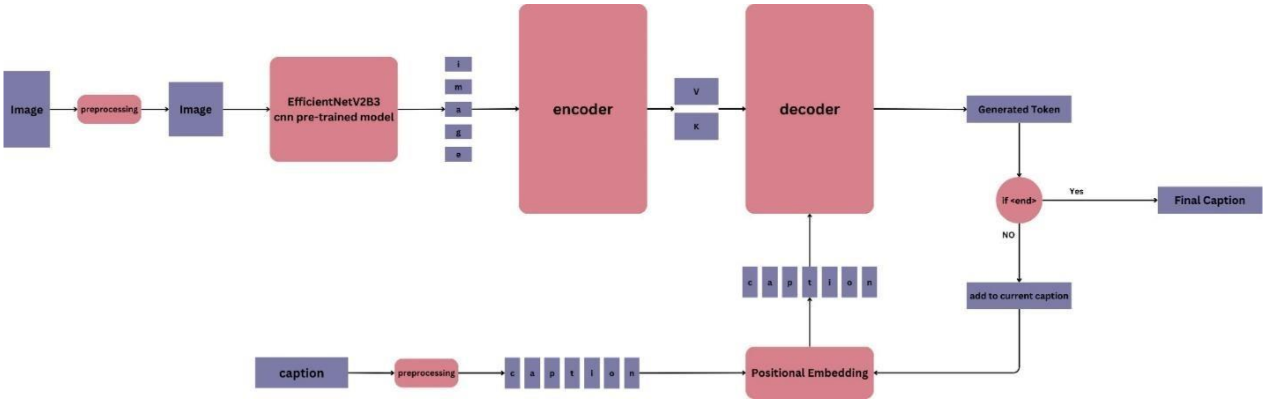


Figure 10:The image captioning model

Preprocessing

Preprocessing is a crucial initial step that ensures the input data is in a suitable format for the model.

Image Preprocessing:

- **Decoding:** Converting the input image from its original format (e.g., JPEG, PNG) into a tensor suitable for processing.
- **Resizing:** Adjusting the image dimensions to match the input size required by EfficientNetV2B3.

Text Preprocessing:

- **Removing Special Characters:** Stripping out punctuation and other non-alphanumeric characters to ensure a clean text input.
- **Lowercasing:** Converting all text to lowercase to maintain consistency and reduce vocabulary size.

- **Vectorization:** Using TensorFlow's TextVectorization layer to convert text into a sequence of integer indices based on a predefined vocabulary.

EfficientNetV2B3 CNN

Role and Selection: EfficientNetV2B3 is employed for its superior performance in feature extraction with relatively fewer parameters. It strikes a balance between accuracy and computational efficiency, making it an ideal choice for this task.

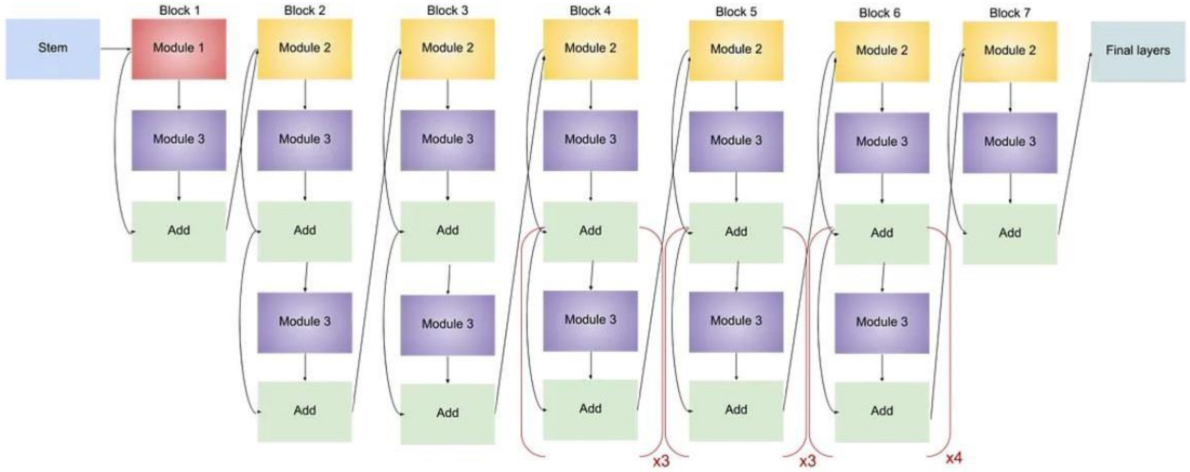


Figure 11:The pretrained model EfficientNetV2B3

Feature Extraction:

- **Convolutional Layers:** Capture spatial hierarchies in the image.
- **Batch Normalization:** Improve training stability and performance.
- **Swish Activation:** Non-linear activation function that helps in learning complex patterns.
- **Global Average Pooling:** Reduces the dimensionality of the feature map while retaining essential information.

EfficientNetV2B3 outputs a rich, high-dimensional feature vector that represents the salient aspects of the input image, which is then fed into the encoder.

Encoder

The encoder is responsible for processing the input image features and transforming them into a context-aware representation that can be utilized by the decoder. The diagram below visually represents the encoder architecture:

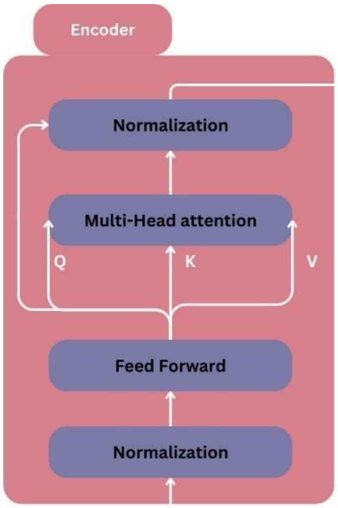


Figure 12:The model's encoder

Below is an in-depth explanation of each component in the encoder:

Normalization:

- Purpose: Stabilizes and accelerates the training process by ensuring that the input data has a consistent distribution.
- Function: Applies layer normalization to the input feature vectors.

Multi-Head Attention:

- Purpose: Allows the model to focus on different parts of the input image simultaneously, capturing various aspects of the data.
- Function:
  - **Q (Query), K (Key), V (Value) Vectors:** These vectors are computed from the input feature vector. Queries, keys, and values help in determining the relevance of different parts of the input.
  - **Scaled Dot-Product Attention:** Computes attention scores to weigh the importance of different parts of the input, allowing the model to focus on relevant features.
  - **Multi-Head Attention Mechanism:** Consists of multiple attention heads that operate in parallel, enhancing the model's ability to capture diverse patterns.

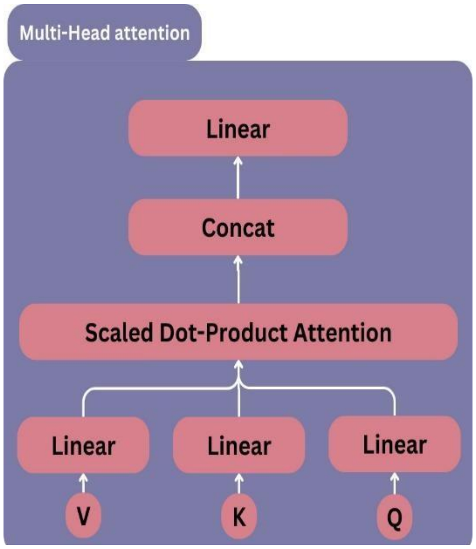


Figure 13:The Multi-Head Attention layer in the encoder

Feed-Forward Neural Network:

- Purpose: Introduces non-linearity and allows the model to learn complex patterns.
- Function: Applies a series of fully connected layers to the output of the multi-head attention mechanism.

Normalization:

- Purpose: Applied again to stabilize and normalize the output of the dense layer.
- Function: Ensures consistent data distribution before passing the output to the next stage.

Decoder

The decoder is responsible for generating the caption by predicting the next word in the sequence based on the encoded image features and previously generated words. The diagram below visually represents the decoder architecture:

Below is an in-depth explanation of each component in the decoder:

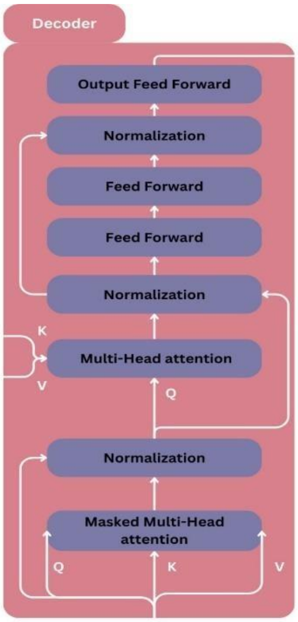


Figure 14:model's decoder

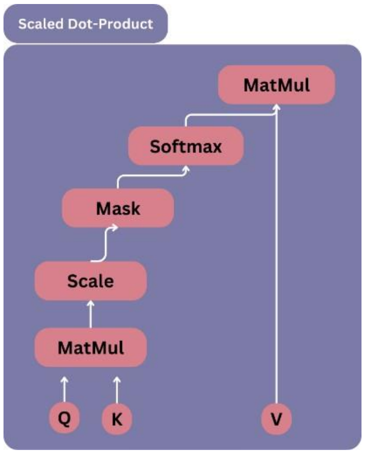


Figure 15:Scaled Dot-Product in the multi-head attention layer

1. **Masked Multi-Head Attention:**
   - Purpose: Ensures that the prediction of the current word is based only on the previously generated words, preventing the model from "cheating."
   - Function:
     - **Q, K, V Vectors:** Like the encoder, these vectors help in determining the relevance of different parts of the input sequence.
     - **Masking:** Ensures that the model cannot see future tokens in the sequence during training.
     - **Scaled Dot-Product Attention:** Computes attention scores to weigh the importance of different

parts of the input, allowing the model to focus on relevant features.

2. Normalization:

- Purpose: Stabilizes the training process by ensuring consistent data distribution.
- Function: Applies layer normalization to the output of the masked multi-head attention mechanism.

3. Multi-Head Attention:

- Purpose: Allows the decoder to focus on different parts of the encoded image features while generating each word in the caption.
- Function:
  - **Q, K, V Vectors:** Computed from the output of the previous layer and the encoded image features.
  - **Cross-Attention Mechanism:** Attends to relevant parts of the encoded image features to generate each word.

4. Feed-Forward Neural Network:

- Purpose: Introduces non-linearity and allows the model to learn complex patterns.
- Function: Applies a series of fully connected layers to the output of the multihead attention mechanism.

Positional Embeddings

Transformers do not inherently capture the order of sequences. Positional embeddings are added to the input embeddings to provide information about the position of each token in the sequence.

Token Generation:

The decoder generates tokens iteratively, predicting the next token based on the current input and previously generated tokens. This process continues until an end token is produced, signaling the completion of the caption.

Final Caption:

Once the end token is generated, the tokens are concatenated to form the final caption for the input image.

The image captioning model effectively combines state-of-the-art techniques in computer vision and natural language processing to generate descriptive captions for images. By leveraging EfficientNetV2B3 for feature extraction and a transformer architecture for sequence modeling, the model achieves high accuracy and efficiency.

3.3 Image generating module:

The Image Generation model combines the power of generative adversarial networks (GANs) with a focus on preserving deep feature representations between generated and real images, integrating textual descriptions as input to guide the image generation process.

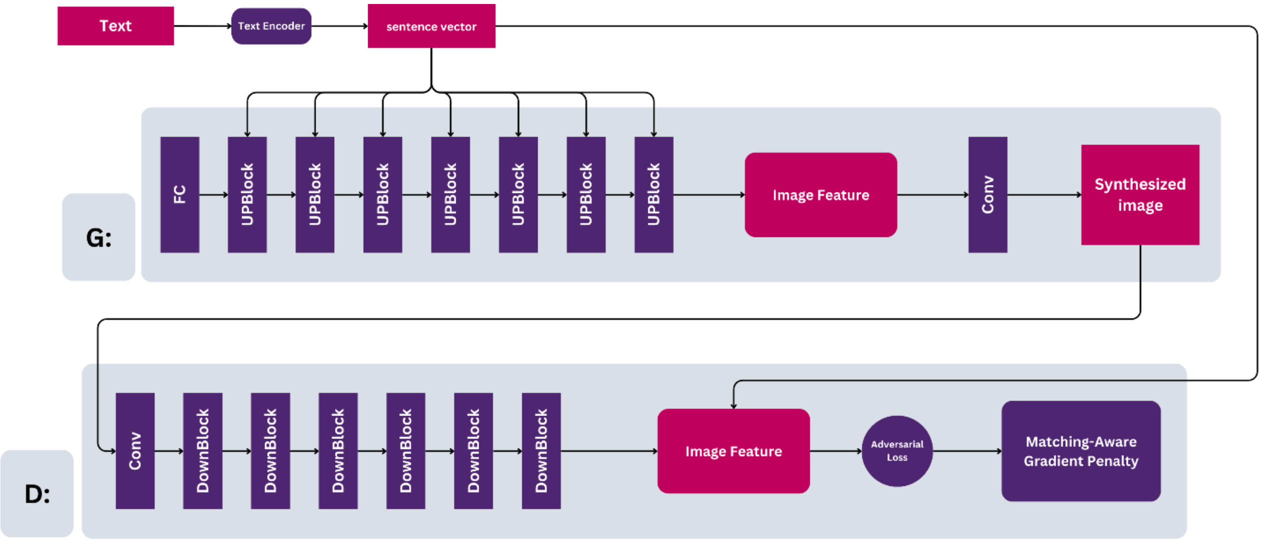


Figure 16:Image generating model

BERT (text encoder):

BERT is a state-of-the-art pre-trained language model developed by Google. It revolutionized natural language understanding tasks by pre-training on vast amounts of text data and achieving impressive results on various downstream tasks without task-specific architecture modifications.

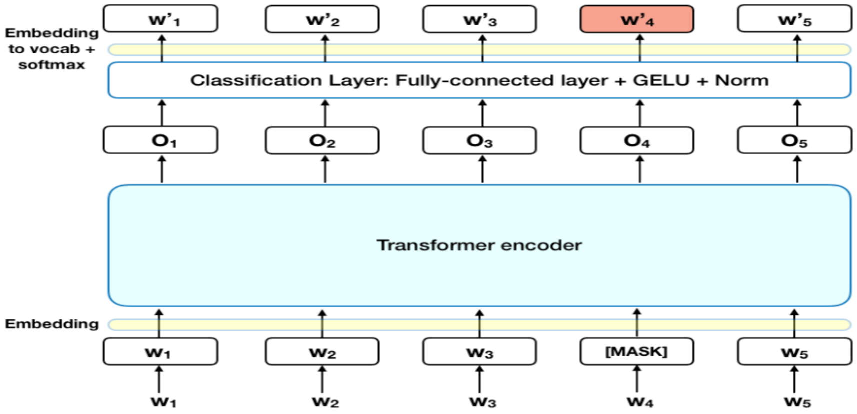


Figure 17:text encoder (Bert)

BERT captures bidirectional context in text, providing rich embeddings that encapsulate semantic meaning and relationships within the input description.

Generator

The generator in a DFGAN is responsible for synthesizing images from random noise vectors conditioned on input text embeddings (in your case, encoded text from a BERT model). Key components include:

- **Text Embedding Input**: Encoded textual descriptions are fed into the generator to guide the image synthesis process.
- **Convolutional Layers**: Initial layers typically consist of convolutional operations that process the input noise and text embeddings.
- **Upsampling Layers**: These layers progressively increase the spatial dimensions of the tensor, transforming it into a high-resolution image that resembles the real images in the training dataset.
- **Skip Connections**: Often used to connect early convolutional layers directly to later layers, aiding in the flow of gradient information and improving training stability.

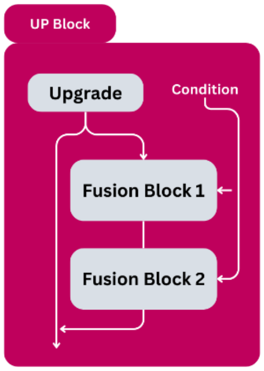


Figure 10: up block layer

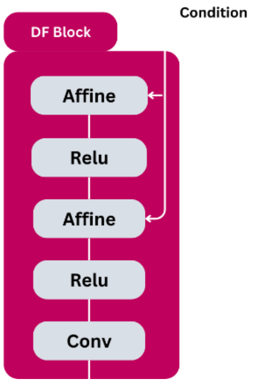


Figure 11: Fusion block of the up block.

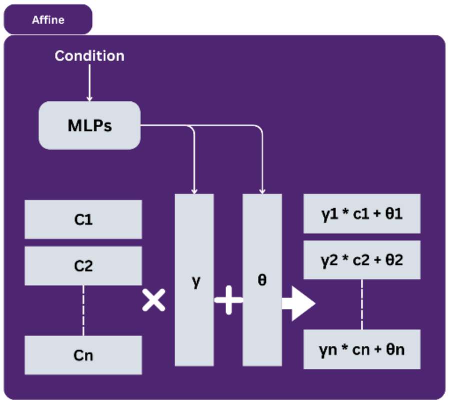


Figure 12: affine of the fusion block

Discriminator

The discriminator in a DFGAN is trained to distinguish between real images from the dataset and images synthesized by the generator. Unlike traditional GAN discriminators, the emphasis is also on ensuring that deep feature representations extracted from generated images align closely with those from real images:

- **Adversarial Loss**: Similar to standard GANs, the discriminator aims to minimize the probability of misclassifying real images as fake and vice versa.
- **Feature Matching Loss**: In addition to the adversarial loss, DFGANs introduce a feature matching loss. This loss encourages the generator to produce images whose deep features (extracted from intermediate layers of the discriminator) match closely with those of real images.
- **Architecture**: The discriminator typically consists of convolutional layers followed by dense layers, ending with a classification layer that outputs the probability of an image being real or generated.

Training Process

- **Adversarial Training**: The generator and discriminator are trained in an adversarial manner, where the generator aims to produce images that fool the discriminator, and the discriminator aims to distinguish real from fake images accurately.
- **Feature Consistency**: The feature matching loss ensures that while the discriminator learns to differentiate real and generated images, it also promotes consistency in deep feature representations between the two, enhancing the quality and realism of generated images.

DFGANs represent a sophisticated advancement in the field of generative modeling, combining the strengths of GANs with a specific emphasis on preserving deep feature representations. In your project, integrating a DFGAN for image generation alongside a Transformer-based model for image captioning can provide a robust platform for creative and interactive user experiences.

The proposed model integrates advanced techniques in image captioning and generation, leveraging transformer architectures and GANs to achieve high-quality results. Each component is carefully designed to maximize performance, efficiency, and accuracy. The following sections will delve deeper into the implementation details, training procedures, and experimental results to validate the efficacy of the proposed system.

CHAPTER 4: EXPERIMENTAL RESULT

4.1 Datasets:

Data sets used to train the image captioning model (from image to text):

Introduction to Flickr8k Dataset

The Flickr8k dataset is a widely used benchmark for image captioning models. It consists of 8,000 images sourced from the Flickr photo-sharing website, each accompanied by five different captions provided by humans. The captions are designed to describe the content and context of the images accurately.

Characteristics of Flickr8k Dataset

- **Number of Images**: 8,000
- **Number of Captions**: 40,000 (5 captions per image)
- **Source**: Flickr photo-sharing website
- **Diversity**: The images in the Flickr8k dataset cover a broad range of scenes and activities, including indoor and outdoor settings, animals, people, vehicles, and various objects. The diversity helps in training robust image captioning models capable of handling different contexts.

Dataset Annotation

The captions in the Flickr8k dataset are rich in detail, providing comprehensive descriptions of the images. This annotation quality ensures that models trained on this dataset learn to generate accurate and contextually relevant captions.

Usage in Image Captioning

The Flickr8k dataset serves as a standard benchmark for evaluating the performance of image captioning models. The diversity and richness of the dataset make it an ideal choice for training and testing models to ensure they generalize well across various scenarios.

The Crime Dataset

Introduction

The crime scene image dataset is a comprehensive collection of high-resolution images designed for the development and training of advanced image captioning models. The dataset encompasses a wide array of crime scenes, both indoors and outdoors, featuring various types of criminal activities. This diversity ensures the dataset's robustness and its capability to enhance model accuracy in identifying and describing different crime scenes. This dataset is particularly useful for AI researchers, law enforcement agencies, and forensic experts aiming to develop tools for crime detection, analysis, and investigation.

Dataset Composition

The dataset comprises over a thousand images categorized into different types of crime scenes. These images are sourced from real-life situations, synthetic simulations, and scenes from movies and TV shows, ensuring a rich and varied collection. Each image is accompanied by descriptive captions, providing contextual information and aiding in the training of image captioning models.

Categories

1. **Violent Crimes:**
   - **Murder:** Images depict crime scenes involving homicides, including victims, assailants, and the immediate environment.
   - **Assault:** Scenes showing physical altercations, with visible injuries and the act of violence captured.
   - **Domestic Violence:** Depictions of violence in home settings, often involving close-up shots of distressed individuals and perpetrators.
2. **Property Crimes:**
   - **Burglary:** Indoor and outdoor scenes showing break-ins, thefts, and the aftermath of such crimes.
   - **Robbery:** Images of individuals being robbed, often with weapons involved, and the resulting chaos.
3. **Kidnapping:**
   - **Abductions:** Scenes showing individuals being forcibly taken, often with bound and gagged victims in various locations.
4. **Torture and Abuse:**
   - **Physical Abuse:** Disturbing images depicting the infliction of severe pain or injury, often in confined spaces.
   - **Psychological Abuse:** Scenes focusing on the mental and emotional trauma inflicted on victims.
5. **Arson:**
   - **Fire Scenes:** Images of buildings and properties set on fire, capturing the intensity of flames and the destruction caused.
6. **Miscellaneous:**
   - **Drug-Related Crimes:** Depictions of illegal drug activities, including production, distribution, and use.
   - **Gang Violence:** Images capturing gang-related activities, including fights, tagging, and other forms of territorial disputes.

Image Quality and Resolution

The images in this dataset are high-resolution, ensuring clarity and detail necessary for accurate analysis. The high quality of the images allows for the identification of minor details that might be crucial in understanding the crime scene, such as facial expressions, weapon details, and environmental context.

Descriptive Captions

Each image is accompanied by multiple captions that vary in complexity and detail. These captions serve several purposes:

1. **Training Image Captioning Models:** The diverse and detailed captions help train models to generate accurate and contextually relevant descriptions of new images.
2. **Providing Context:** The captions offer context to the images, describing the crime, the individuals involved, and the environment.
3. **Improving Accuracy:** The variation in caption complexity helps in refining the models to cater to different levels of descriptive needs.

Applications

The primary application of this dataset is in the development and training of image captioning models. However, it has broader implications in various fields:

1. **Law Enforcement:** Assist in crime scene analysis and investigation by providing automated descriptions and analysis of crime scenes.
2. **Forensic Analysis:** Aid forensic experts in examining crime scenes and gathering evidence by offering detailed visual and contextual data.
3. **AI Research:** Provide a rich source of data for AI researchers working on image recognition, captioning, and related fields.

Ethical Considerations

Given the sensitive nature of the content, ethical considerations are paramount in the usage of this dataset:

1. **Privacy:** Ensure that images of real individuals are used with consent and are anonymized to protect identities.
2. **Usage Guidelines:** Establish clear guidelines on the appropriate use of the dataset to prevent misuse or exploitation.
3. **Content Warning:** Provide adequate warnings about the graphic nature of the content to prepare users for potentially disturbing images.

The crime dataset was meticulously collected to train an image captioning model focused on crime-related scenes. This dataset comprises 1,000 images from different sources, including movie scenes, TV series scenes, and CCTV recordings from various environments such as supermarkets, homes, streets, and different shops.

Characteristics of Crime Dataset

- **Number of Images**: 1,000
- **Source**:
    - Movie and TV series scenes depicting crimes
    - CCTV recordings from Kaggle and other sources
- **Types of Crimes Covered**:
    - Murders
    - Kidnapping
    - Robbery
    - Arson (burning houses or properties)
    - Torturing
    - Threatening

Dataset Collection and Annotation

The images were collected by taking screenshots of crime scenes in movies, TV series, and CCTV recordings. This diverse collection aims to cover various crime types, providing a rich dataset for training models to understand and describe crime scenes accurately.

Challenges

Annotating crime scenes involves a level of subjectivity and requires careful attention to context and details. Ensuring accurate and meaningful captions for such a dataset is challenging but essential for developing robust image captioning models.

Number of Images

- **Flickr8k Dataset**: 8,000 images
- **Crime Dataset**: 1,000 images

The Flickr8k dataset is significantly larger than the crime dataset. The larger size of Flickr8k allows for more extensive training and better generalization of image captioning models.

Distribution of Images

- **Flickr8k Dataset**: Diverse images including various scenes, objects, animals, and activities.

- **Crime Dataset**: Focused on crime scenes, including indoor and outdoor settings, and various crime types.

The Flickr8k dataset's diversity helps in training models on a wide range of scenarios, while the crime dataset's focus on specific crime-related scenes ensures that the model learns to describe such scenes accurately.

Visual Content

- **Flickr8k Dataset**: Images capture everyday life, nature, and human activities.
- **Crime Dataset**: Images capture crime scenes, often with a tense or violent context.

The content of the two datasets is quite different. Flickr8k images are generally benign and depict regular activities, while the crime dataset includes images with explicit crime scenes, which can be more challenging to describe accurately.

In conclusion, the process of training an image captioning model on both the Flickr8k and the crime dataset involved extensive effort to ensure the generated captions are semantically meaningful and contextually accurate.

Key Points

- **Effort in Data Collection and Annotation**: Collecting and annotating the crime dataset required meticulous attention to detail to ensure that the captions accurately described the crime scenes.
- **Model Robustness**: Training on the diverse Flickr8k dataset helped in building a robust model capable of generalizing across different scenarios. The focused training on the crime dataset further enhanced the model's ability to describe crime-related scenes accurately.
- **Challenges**: The primary challenge was to ensure that the model could handle the diverse and often complex scenes in the crime dataset while maintaining the ability to generalize well on the more diverse Flickr8k dataset.

Data sets used to train the image generating:

(CelebA)

Dataset name (CelebA: CelebFaces Attributes Dataset) The CelebA dataset is great for training and testing models for face detection, particularly for recognizing facial attributes such as finding people with brown hair, are smiling, or wearing glasses. Images cover large pose variations, background clutter, diverse people, supported by a large quantity of images and rich annotations.

It has large-scale face attributes dataset that contains over 200,000 celebrity images, each annotated with 40 different attribute labels per image. It is commonly used for tasks such as face recognition, facial attribute analysis, and facial image generation. Its applications can revolves around identifying faces for various applications from logging into your phone with your face or searching through surveillance images for a particular suspect. This data was originally collected by researchers at MMLAB, The Chinese University of Hong Kong (specific reference in Acknowledgment section).

The dataset consists of RGB images of celebrity faces, each with dimensions typically ranging from 64x64 to 178x218 pixels. These images showcase a diverse range of facial expressions, poses, and lighting conditions. Each image in the CelebA dataset is associated with attribute annotations, providing additional information about the depicted celebrity's facial characteristics. Some of the commonly annotated attributes include: Gender, Age, Presence of facial hair, Presence of eyeglasses, Presence of headwear.

Content:

- 202,599 number of face images of various celebrities
- 10,177 unique identities, but names of identities are not given
- 40 binary attribute annotations per image
- 5 landmark locations

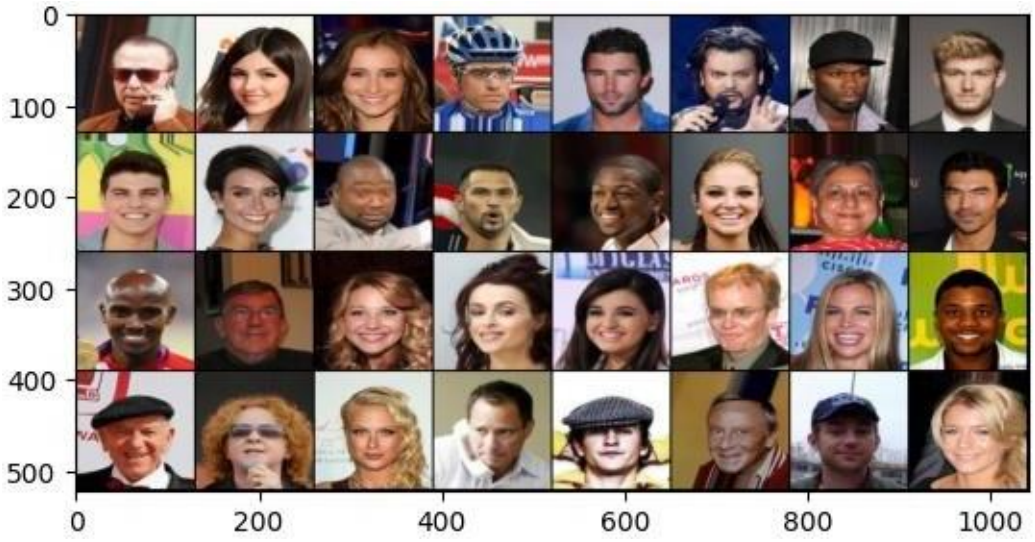


Figure 13. A batch of 32 images from the dataset.

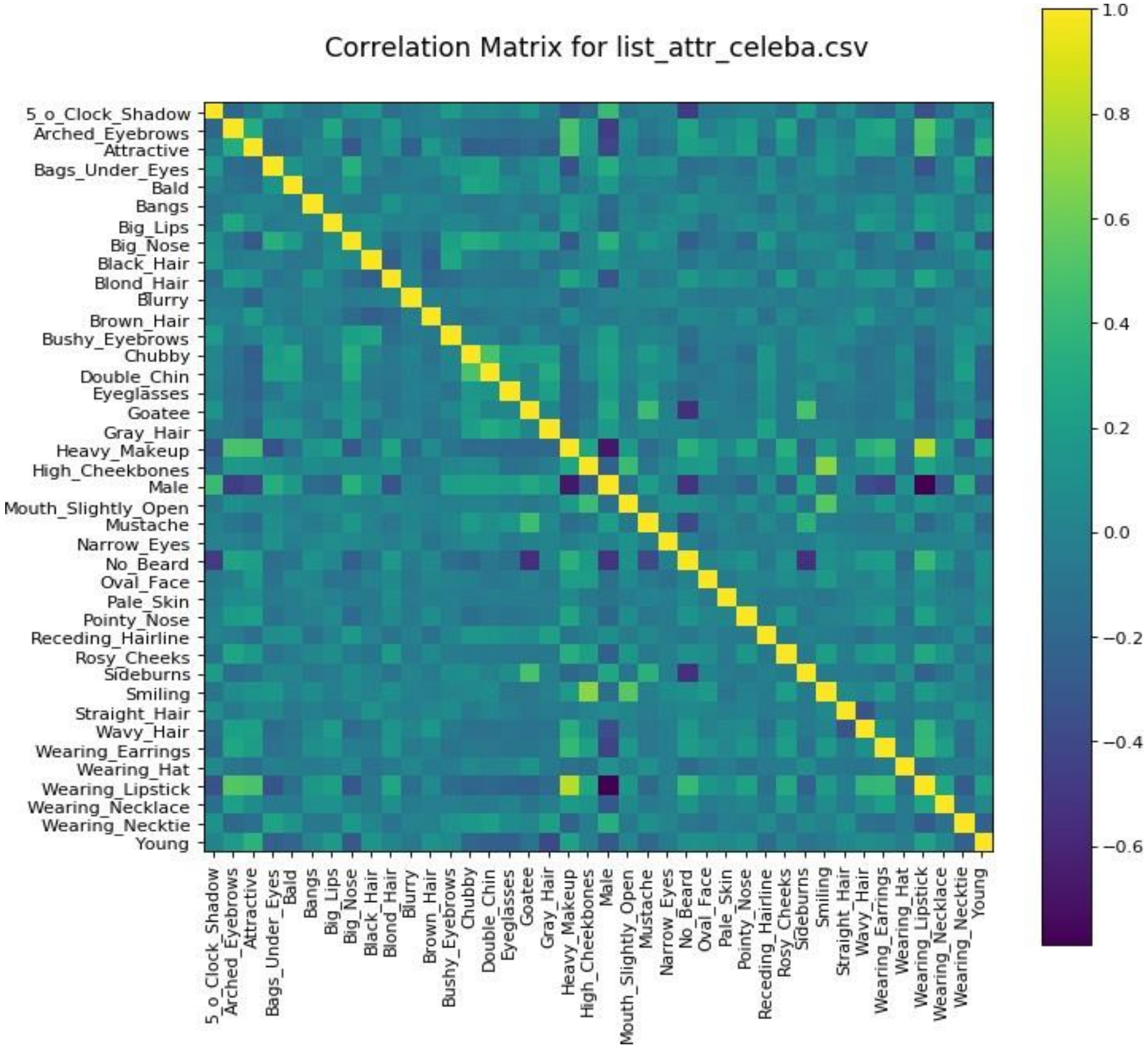


Figure 14. a correlation matrix of the attributes.

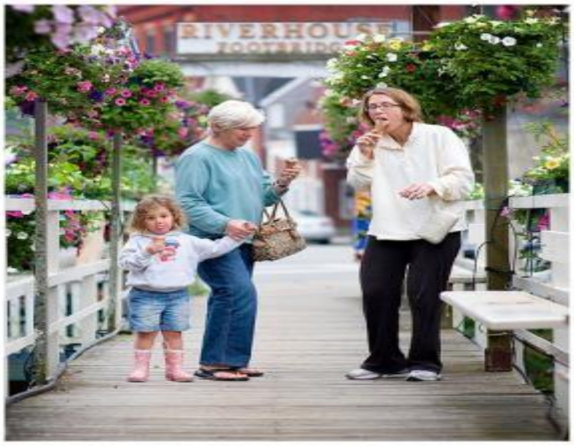
In this draft we were able to train on 25,000 image for 22 epoch which is a significant improvement from the first draft where we was only able to train on 20,000 image for 4 epoch and that of course lead to huge improvements as we will see later in the results.

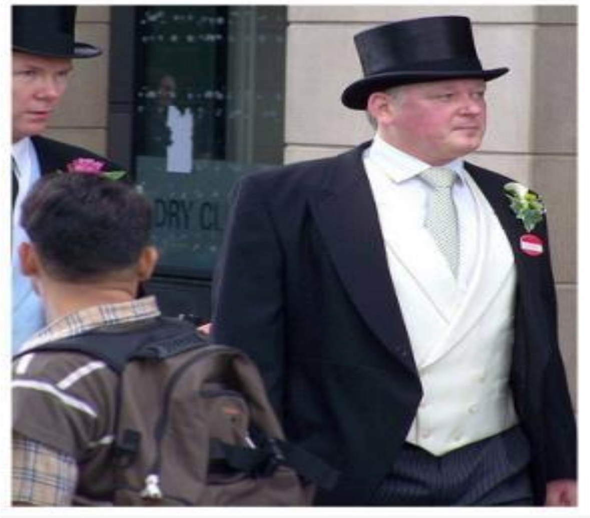
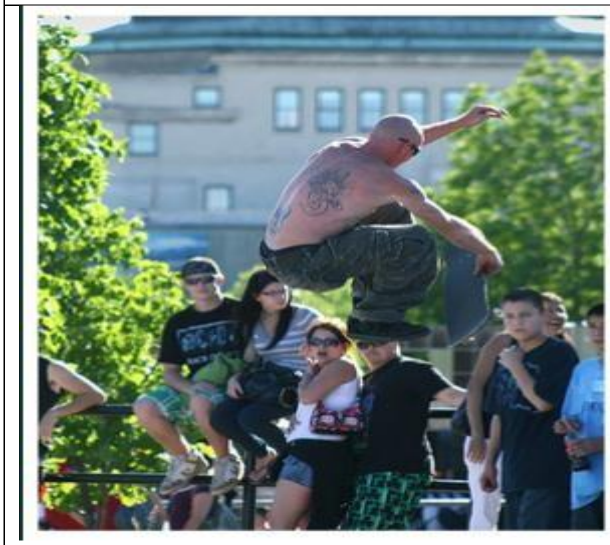
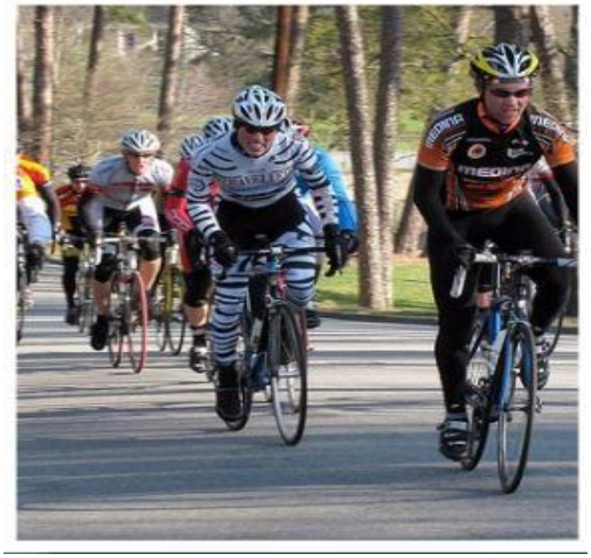
4.2 Results of image caption module:

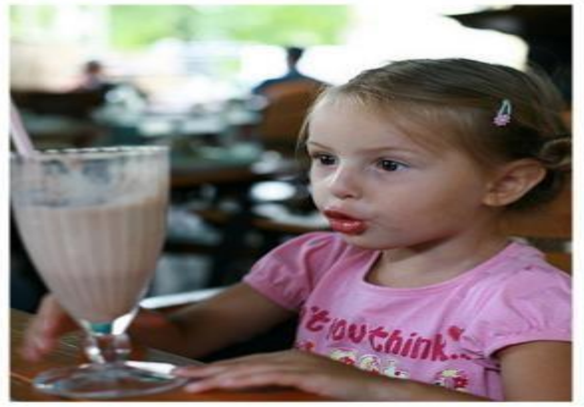
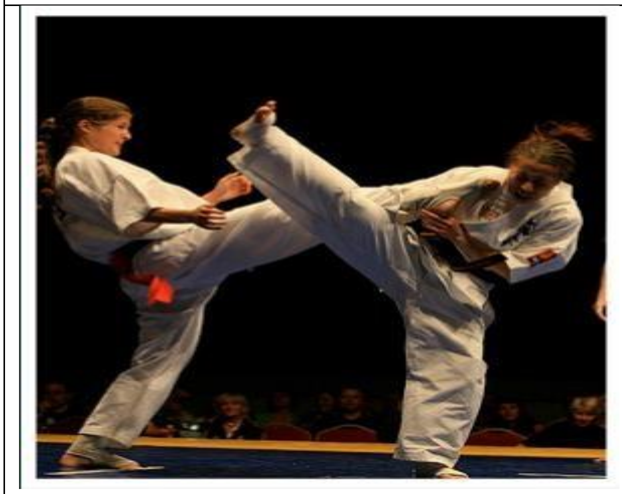
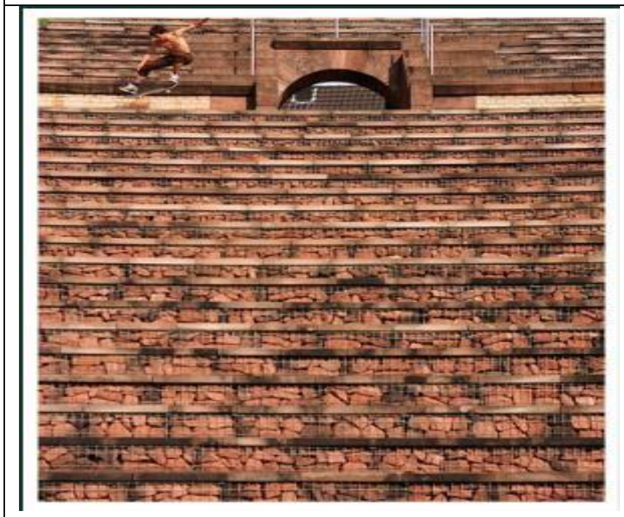
BLEU: Compares n-gram overlaps between text and references. Easy to compute, language independent, its strength lies in being Simple and widely used, making results comparable across studies, Language-independent and computationally efficient however its limitations lie in being heavily on n-gram overlap, neglecting word order and grammatical correctness, sensitive to short texts and unreliable scoring. Despite its limitations, it remains a standard benchmark for comparison and ease of interpretation.

ROUGE-L: Measures overlap in various ways (n-grams, word sequences) between text and ideal summaries. its strength lies in Offering various metrics (n-grams, word sequences, longest common subsequences) for diverse evaluation aspects, considers sentence-level structure similarity but however its weakness lies in struggling in its Performance, a sit depends on reference summaries, potentially biased with poor references. But still consideration of sentence structure (ROUGE-L) aligns well with the goal of generating grammatically correct and coherent captions.

METEOR: Compares word segments with references using word stemming and synonyms. Better correlation at sentence-level but focuses on precision more than recall. its strength lies in Incorporating stemming and synonymy matching, addressing paraphrases and word variations, also Showing good correlation at the sentence level, capturing more semantic meaning. However, its weakness lies in Primarily focusing on precision (matching generated words to references), neglecting recall (covering all relevant information). however, its ability to handle synonyms and paraphrases helps ensure captions capture the image's meaning even if using different words.Here are samples that we tested the model on:

Image	First caption result	Final caption result	Final caption in Arabic
	a group of people walk on a sidewalk near flowers	two women and a little girl are eating ice cream	امرأتان كبيرتان وفتاة صغيرة تقفان بالخارج تحت الزهور

	two men are standing in front of a building	a man wearing a hat and formal suit and hat stands next to a building	رجل يرتدي بدلة رسمية وقبعة يقف بجوار أحد المبا · ن
	a skateboarder is doing a trick in the air	a man shocks the audience with his skateboard tricks	رجل بدون قميص ووشم على ظهره · ف الهواء مع لوح ال تزلج·
	a group of people ride bikes	a group of men racing on their bicycles	مجموعة من الدراجات يركبون دراجاتهم

	a little girl is sitting at a table and holding a large hair barrette pursing her lips at a big chocolate milkshake	a little girl is staring at the camera	فتاة صغية تأكل خ. ي
	two girls are performing karate at a tournament	a man in white performing a dance studio	امرأن تمارسان رياضة الكاريته
	a man is skateboarding on a large brick stair	a man skateboards around the empty outdoor sports arena	رجل ي تزلج عل سلمكب يمن الطوب

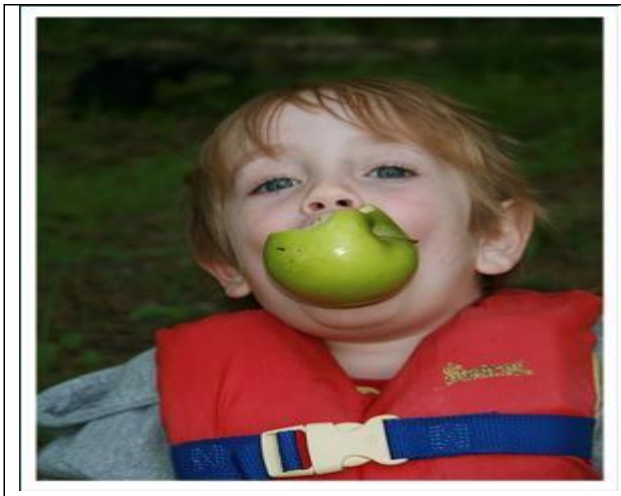
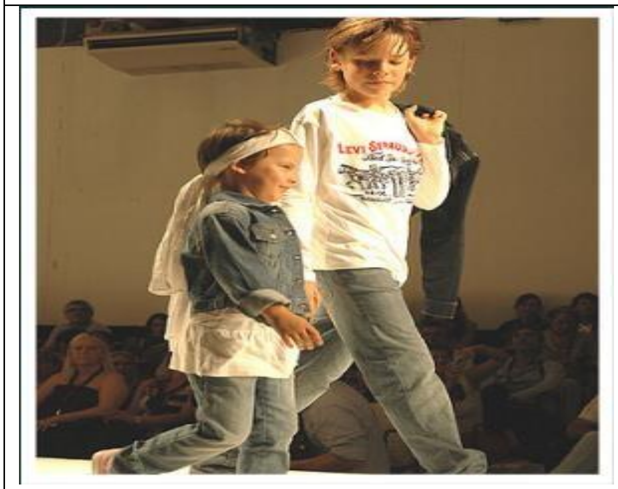
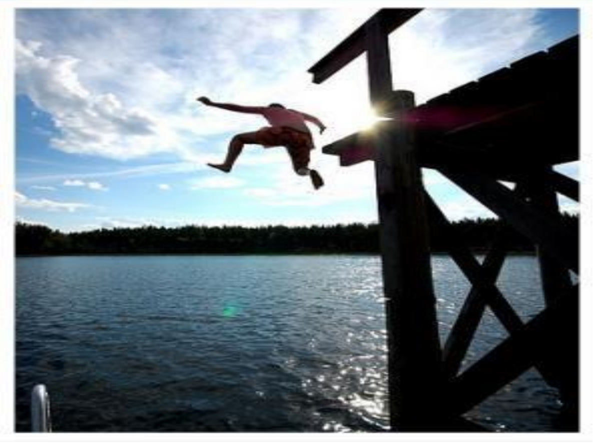
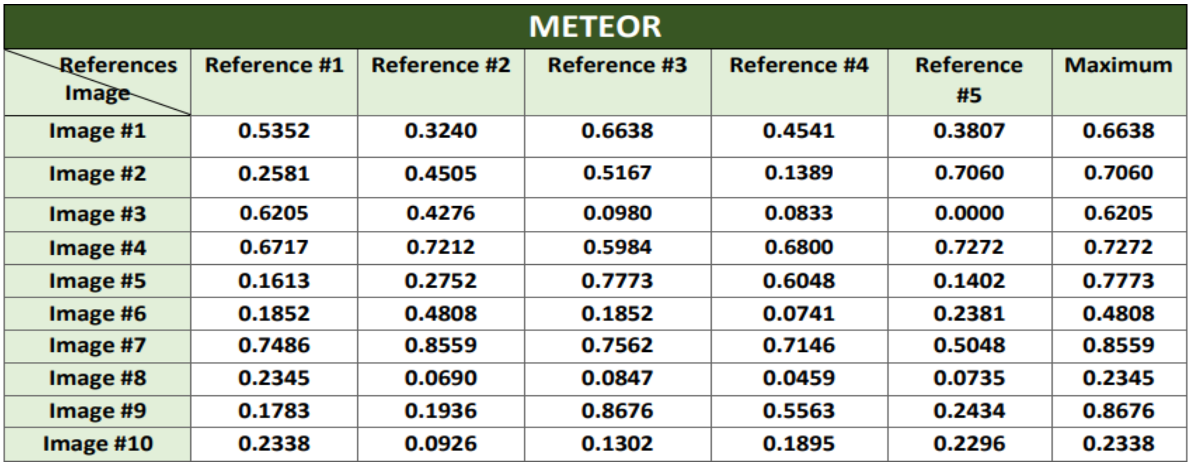
	a child wearing a life jacket has a green apple in his mouth	a child is playing with his toy	طفل صغي يرتدي س ية النجاة وهو على العشب
	two children walking on a platform	two children walking on a platform	امرأة ترتدي قمي ًص أبيض تتحدث إلى رجل يرتدي رسوا ًال أسود
	a person jumping off of a dock into the water	a person jumping off of a dock and into the water	طفل صغي يقفز من فوق سطح السفينة الدوارة على الماء

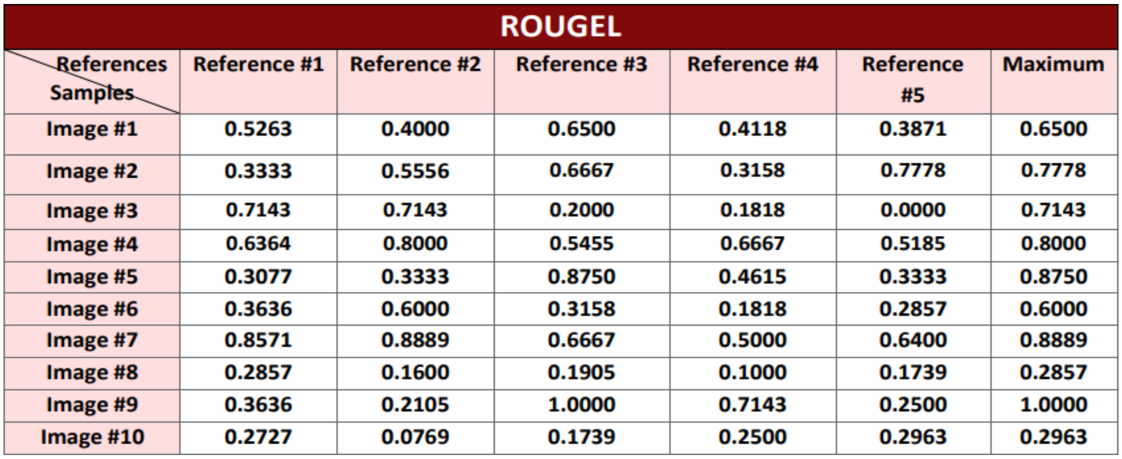
Table 3:image samples with their old and new captions en-ar

As shown above the captions got a better semantic meaning and the Arabic captions are even better in some cases.



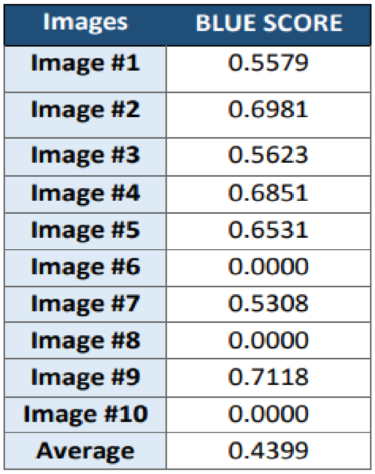
METEOR						
References Image	Reference #1	Reference #2	Reference #3	Reference #4	Reference #5	Maximum
Image #1	0.5352	0.3240	0.6638	0.4541	0.3807	0.6638
Image #2	0.2581	0.4505	0.5167	0.1389	0.7060	0.7060
Image #3	0.6205	0.4276	0.0980	0.0833	0.0000	0.6205
Image #4	0.6717	0.7212	0.5984	0.6800	0.7272	0.7272
Image #5	0.1613	0.2752	0.7773	0.6048	0.1402	0.7773
Image #6	0.1852	0.4808	0.1852	0.0741	0.2381	0.4808
Image #7	0.7486	0.8559	0.7562	0.7146	0.5048	0.8559
Image #8	0.2345	0.0690	0.0847	0.0459	0.0735	0.2345
Image #9	0.1783	0.1936	0.8676	0.5563	0.2434	0.8676
Image #10	0.2338	0.0926	0.1302	0.1895	0.2296	0.2338

Table 2. Result of METEOR evaluation matrix



ROUGEL						
References Samples	Reference #1	Reference #2	Reference #3	Reference #4	Reference #5	Maximum
Image #1	0.5263	0.4000	0.6500	0.4118	0.3871	0.6500
Image #2	0.3333	0.5556	0.6667	0.3158	0.7778	0.7778
Image #3	0.7143	0.7143	0.2000	0.1818	0.0000	0.7143
Image #4	0.6364	0.8000	0.5455	0.6667	0.5185	0.8000
Image #5	0.3077	0.3333	0.8750	0.4615	0.3333	0.8750
Image #6	0.3636	0.6000	0.3158	0.1818	0.2857	0.6000
Image #7	0.8571	0.8889	0.6667	0.5000	0.6400	0.8889
Image #8	0.2857	0.1600	0.1905	0.1000	0.1739	0.2857
Image #9	0.3636	0.2105	1.0000	0.7143	0.2500	1.0000
Image #10	0.2727	0.0769	0.1739	0.2500	0.2963	0.2963

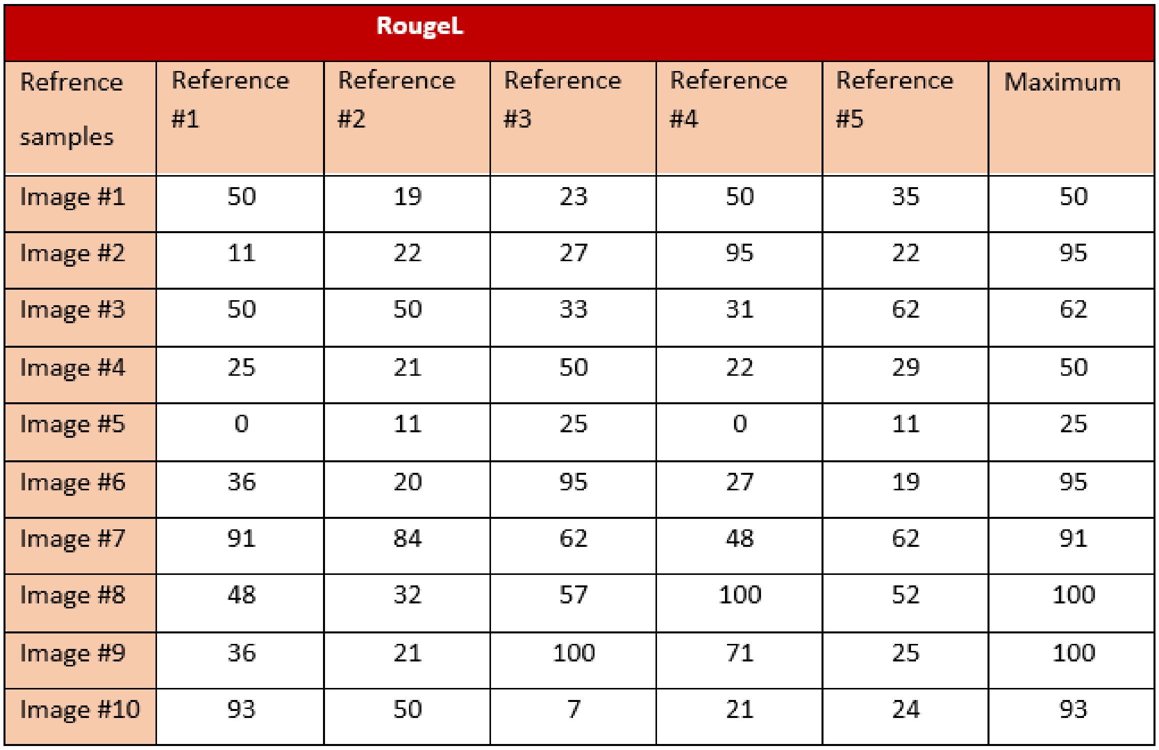
Table 3. Result of ROUGEL evaluation matrix on 10 samples



Images	BLUE SCORE
Image #1	0.5579
Image #2	0.6981
Image #3	0.5623
Image #4	0.6851
Image #5	0.6531
Image #6	0.0000
Image #7	0.5308
Image #8	0.0000
Image #9	0.7118
Image #10	0.0000
Average	0.4399

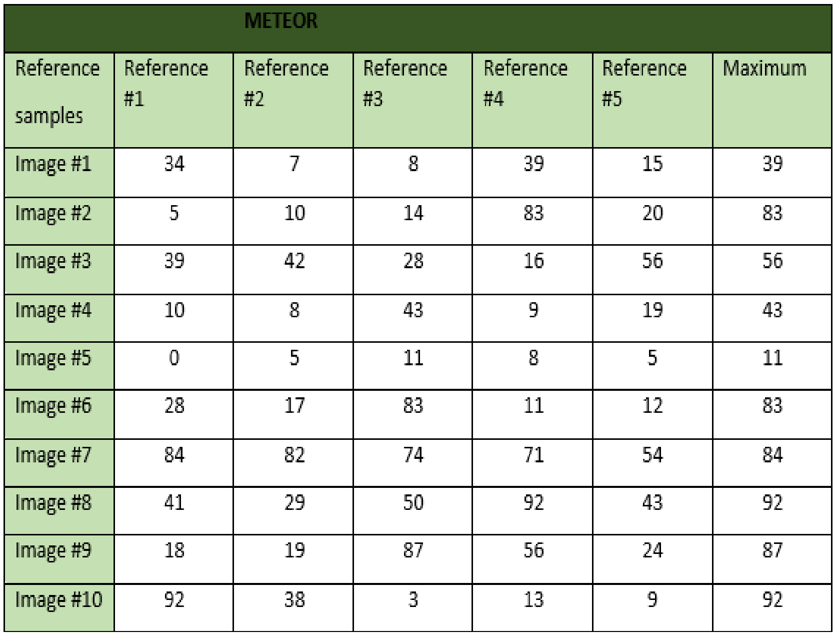
Table 4. Result of BLUE SCORE evaluation matrix on 10 samples before adjustments

We also used the similarity measures methods before that we used on the generated captions that we got before adjusting the dataset and the model, here are the results of the measures on the new generated English caption for each image:



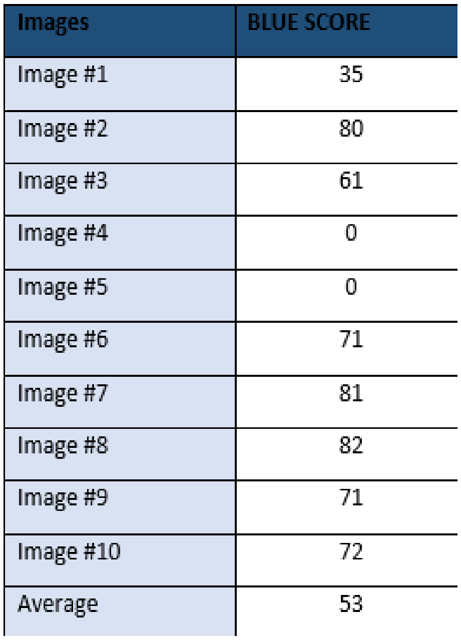
RougeL						
Refrence samples	Reference #1	Reference #2	Reference #3	Reference #4	Reference #5	Maximum
Image #1	50	19	23	50	35	50
Image #2	11	22	27	95	22	95
Image #3	50	50	33	31	62	62
Image #4	25	21	50	22	29	50
Image #5	0	11	25	0	11	25
Image #6	36	20	95	27	19	95
Image #7	91	84	62	48	62	91
Image #8	48	32	57	100	52	100
Image #9	36	21	100	71	25	100
Image #10	93	50	7	21	24	93

Table 5. Result of ROUGEL evaluation matrix on same 10 samples after adjustments



METEOR						
Reference samples	Reference #1	Reference #2	Reference #3	Reference #4	Reference #5	Maximum
Image #1	34	7	8	39	15	39
Image #2	5	10	14	83	20	83
Image #3	39	42	28	16	56	56
Image #4	10	8	43	9	19	43
Image #5	0	5	11	8	5	11
Image #6	28	17	83	11	12	83
Image #7	84	82	74	71	54	84
Image #8	41	29	50	92	43	92
Image #9	18	19	87	56	24	87
Image #10	92	38	3	13	9	92

Table 6. Result of METEOR evaluation matrix on same 10 samples after adjustments



Images	BLUE SCORE
Image #1	35
Image #2	80
Image #3	61
Image #4	0
Image #5	0
Image #6	71
Image #7	81
Image #8	82
Image #9	71
Image #10	72
Average	53

Table 7. Result of Blue score evaluation matrix on same 10 samples after adjustments

Also, we tested our model on 10 random samples from the crime dataset and got 10 generated captions in English and in Arabic as shown below:

a man in flames while a fire

رجل إطفاء يعمل بعد اشتعلت فيها النيران

A man threatens a man with a gun

رجل يرتدي بدلة سوداء يقتل رجل بمسدس

a man threatens another man with a gun

رجل يخنق رجال آخر

a man in suit kills another man

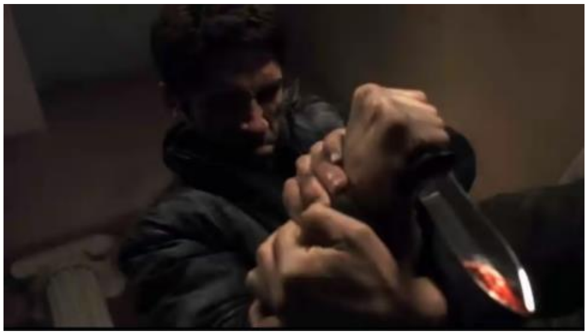
رجل يرتدي بدلة سوداء يقتل رجل بمسدس

a woman wearing a black and a blue bed

امرأة مستلقية على غطاء سرير ملطخ بالدماء

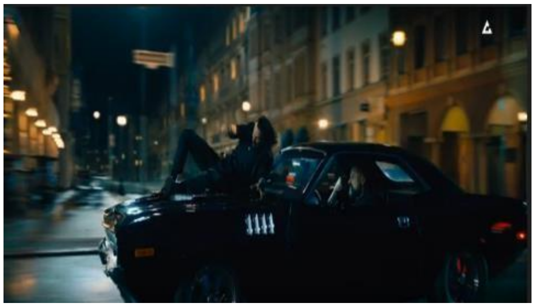
a man in black shorts is beaten by a man in blue shorts

جل يرتدي السراويل الزرقاء يقاتل رجل يرتدي السراويل السوداء



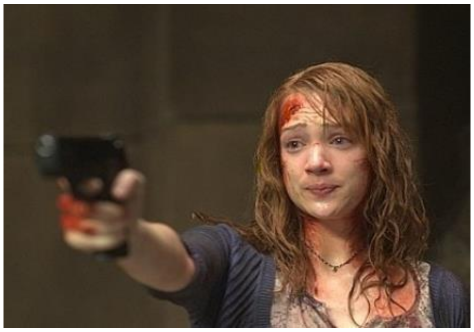
a man is holding a knife

رجل يقتل رجال بسك ي



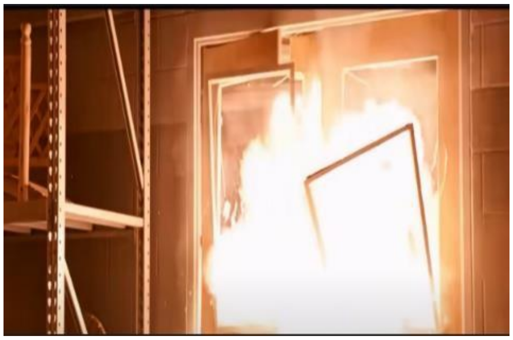
a man was hitten by a car

سيارة فضية ي فمنطقة الم يو



painful yet strong she holds the gun steady

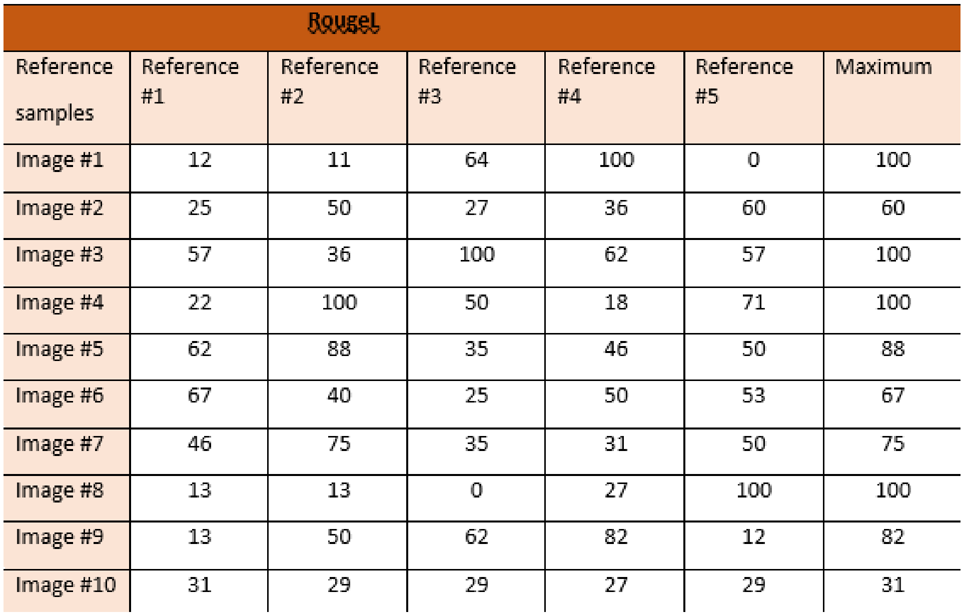
امرأة مصابة تصوب مسدسها بقوة



the fire is about to burn the kitchen

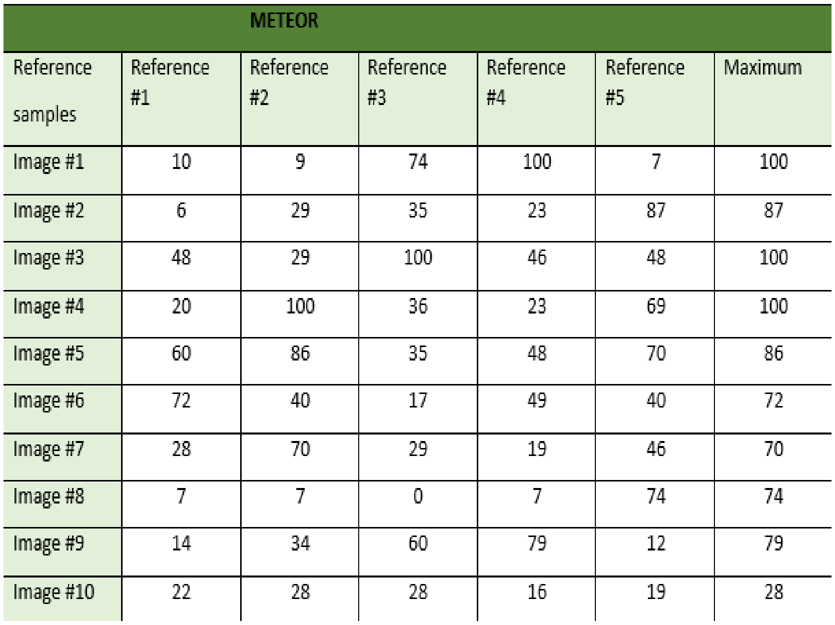
النار مشتعلة بقوة

And to test the model`s efficiency we applied the similarity measures methods on the 10 samples and here are the results:



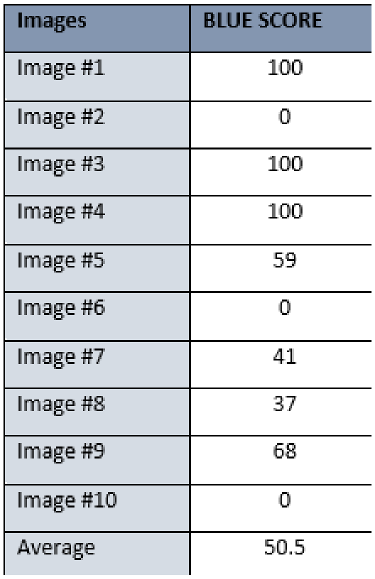
RougeL						
Reference samples	Reference #1	Reference #2	Reference #3	Reference #4	Reference #5	Maximum
Image #1	12	11	64	100	0	100
Image #2	25	50	27	36	60	60
Image #3	57	36	100	62	57	100
Image #4	22	100	50	18	71	100
Image #5	62	88	35	46	50	88
Image #6	67	40	25	50	53	67
Image #7	46	75	35	31	50	75
Image #8	13	13	0	27	100	100
Image #9	13	50	62	82	12	82
Image #10	31	29	29	27	29	31

Table 8. Result of ROUGEL evaluation matrix on same 10 samples from crime dataset



METEOR						
Reference samples	Reference #1	Reference #2	Reference #3	Reference #4	Reference #5	Maximum
Image #1	10	9	74	100	7	100
Image #2	6	29	35	23	87	87
Image #3	48	29	100	46	48	100
Image #4	20	100	36	23	69	100
Image #5	60	86	35	48	70	86
Image #6	72	40	17	49	40	72
Image #7	28	70	29	19	46	70
Image #8	7	7	0	7	74	74
Image #9	14	34	60	79	12	79
Image #10	22	28	28	16	19	28

Table 9. Result of METEOR evaluation matrix on same 10 samples from crime dataset



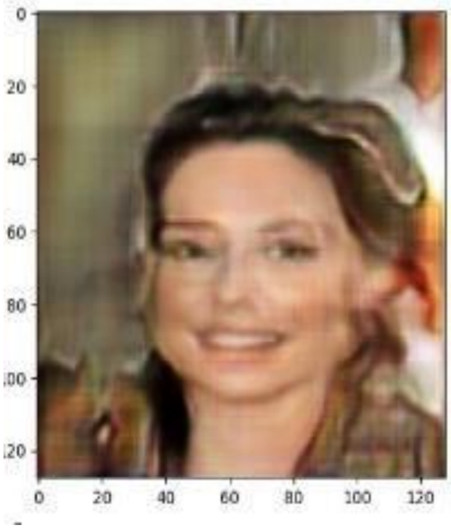
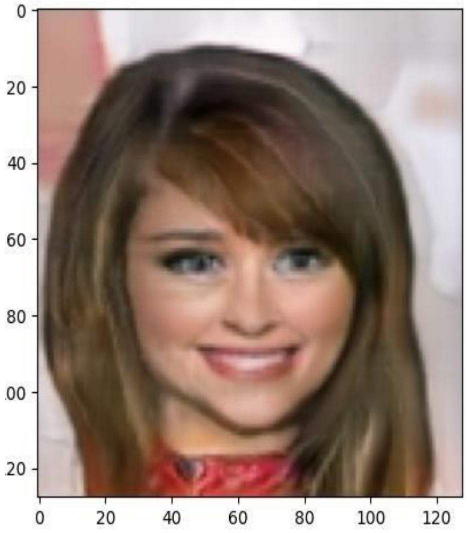
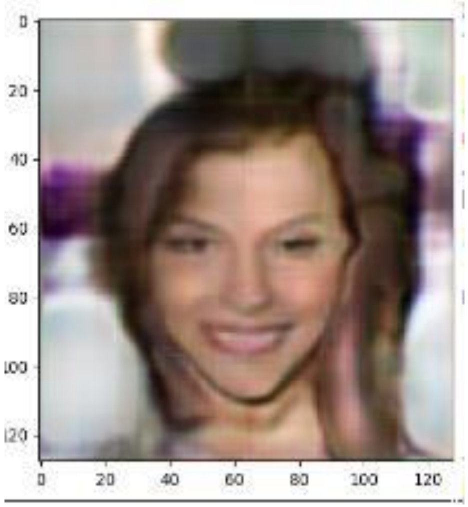
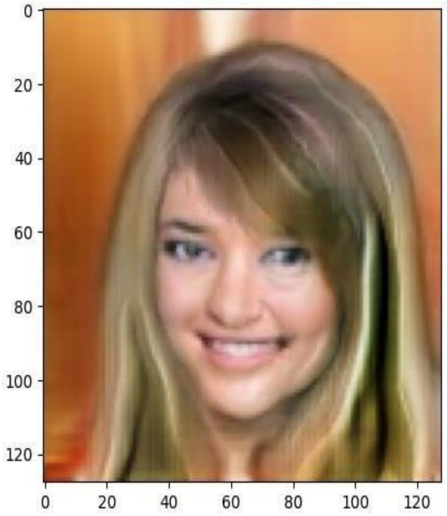
Images	BLUE SCORE
Image #1	100
Image #2	0
Image #3	100
Image #4	100
Image #5	59
Image #6	0
Image #7	41
Image #8	37
Image #9	68
Image #10	0
Average	50.5

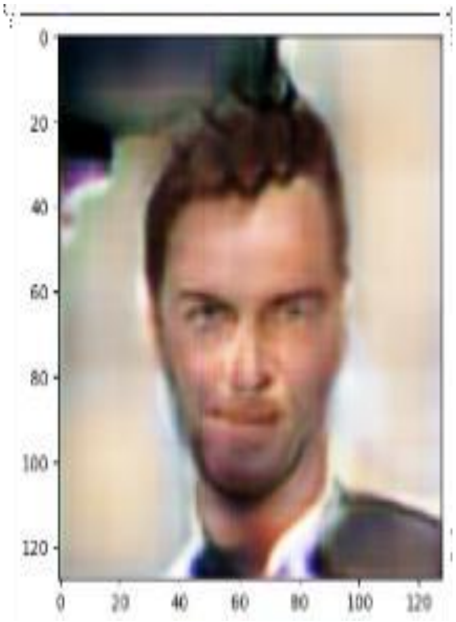
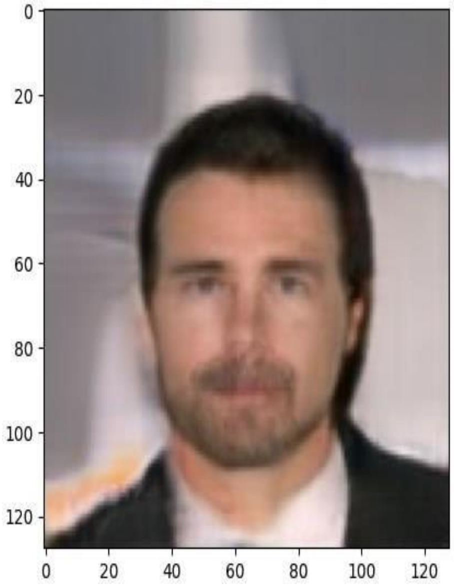
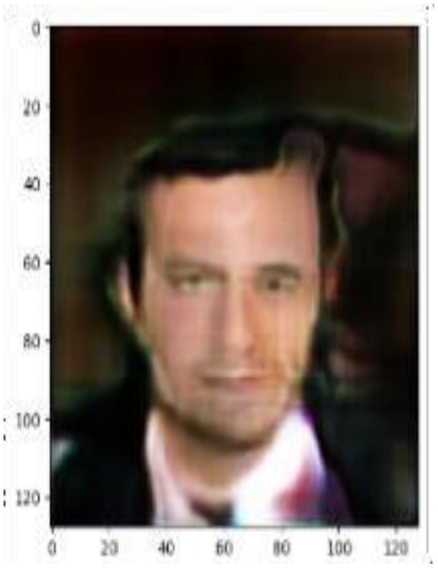
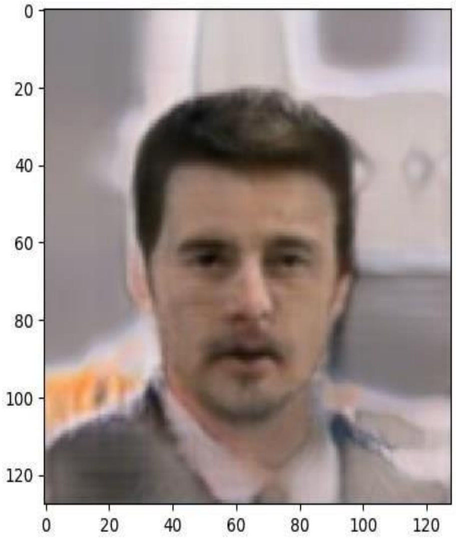
Table 10. Result of Blue score evaluation matrix on same 10 samples from crime dataset

4.3 Results of image generating module:

Score	value
Inception Score (IS)	9.127
Fréchet inception distance (FID)	38.689

Table 11: image generation scores

First draft	Captions	Final draft
 0 20 40 60 80 100 120 0 20 40 60 80 100 120	" The woman has high cheekbones. She has straight hair which is brown in colour. She has arched eyebrows and a slightly open mouth. The smiling, young attractive woman has heavy makeup. She is wearing lipstick. "	 0 20 40 60 80 100 120 0 20 40 60 80 100 120
 0 20 40 60 80 100 120 0 20 40 60 80 100 120	" The criminal is a woman. she is young and beautiful, and she has a straight blonde hair and a small nose. "	 0 20 40 60 80 100 120 0 20 40 60 80 100 120

 0 20 40 60 80 100 120 0 20 40 60 80 100 120	" The criminal is a male. He has an oval face, short brown hair, narrow eyes, and a beard. "	 0 20 40 60 80 100 120 0 20 40 60 80 100 120
 0 20 40 60 80 100 120 0 20 40 60 80 100 120	" The criminal is a male. He has an oval face, short brown hair, narrow eyes, and a beard. "	 0 20 40 60 80 100 120 0 20 40 60 80 100 120

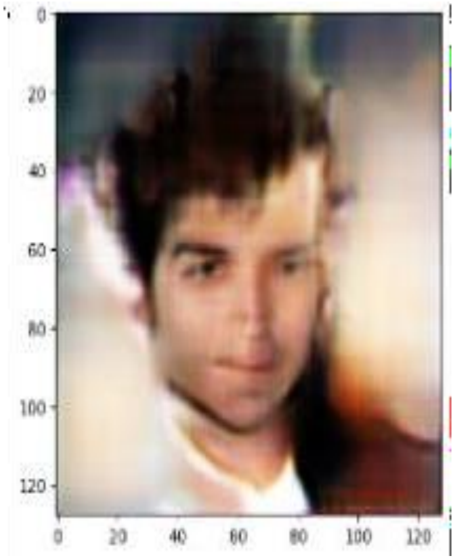
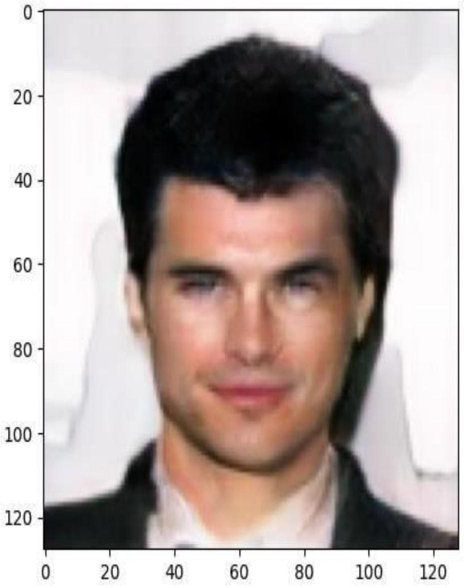
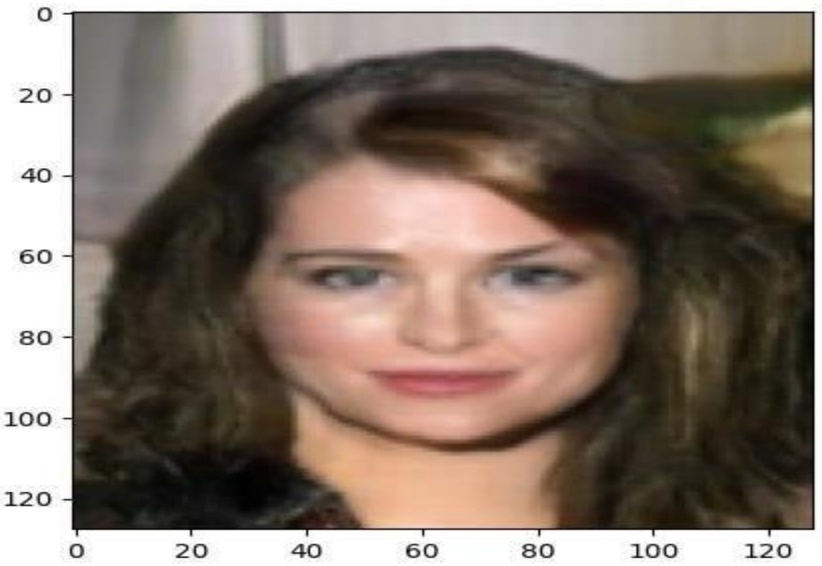
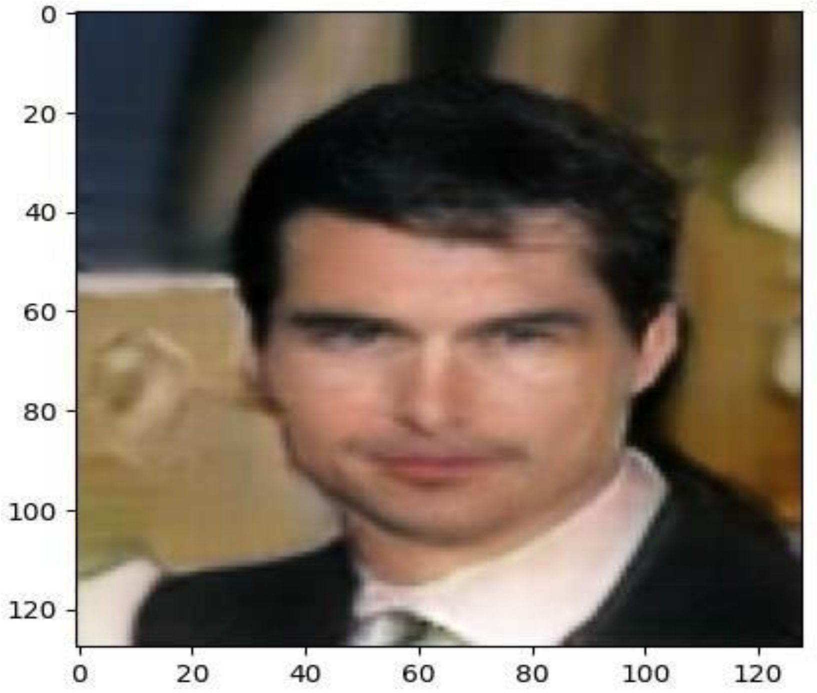
	He wears a 5 o' clock shadow. He has black and straight hair. He has big lips, a big nose, bushy eyebrows and a pointy nose. The man looks attractive and young.	

Table 14: old and new image samples of same captions

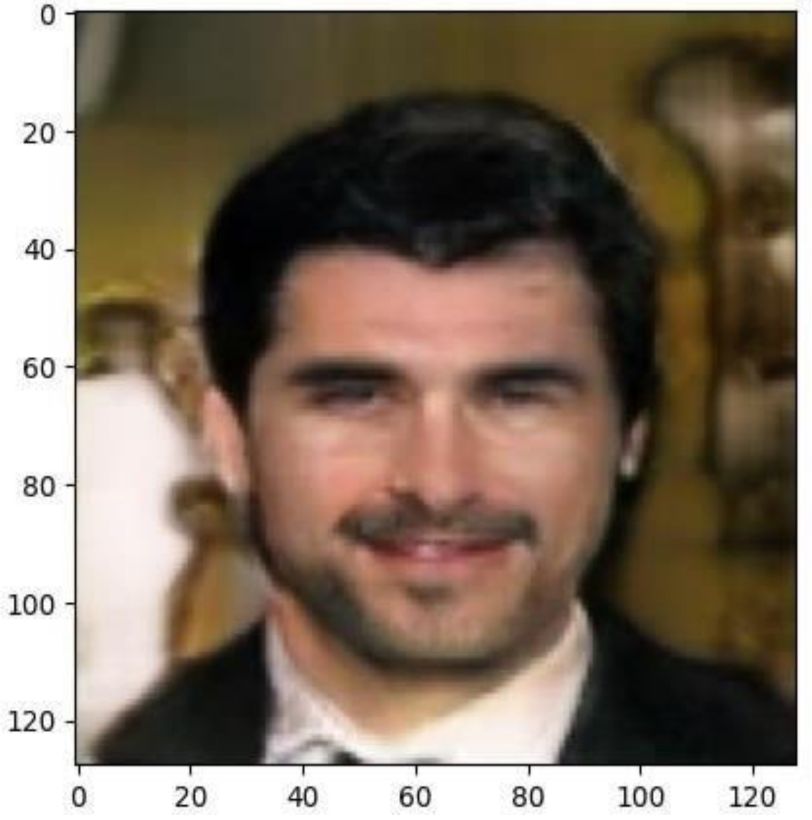
Some samples



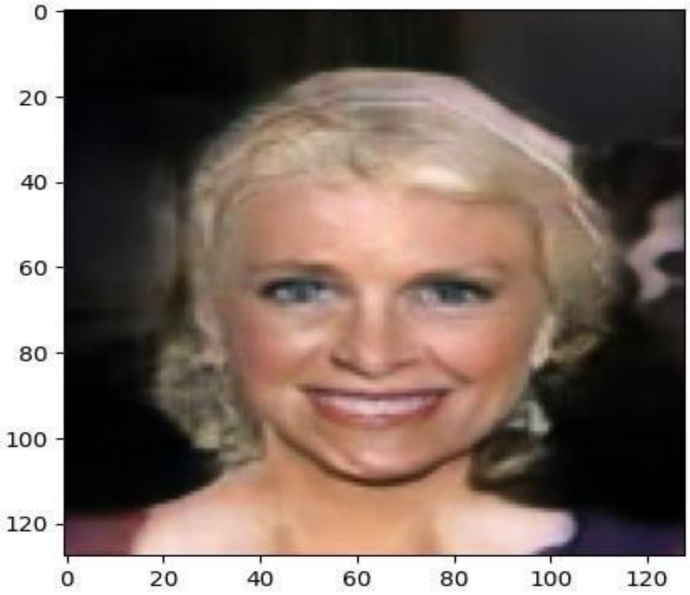
The woman has high cheekbones and an oval face. Her hair is wavy. She has bushy eyebrows. The female is smiling, is attractive, young and has heavy makeup. She is wearing lipstick.



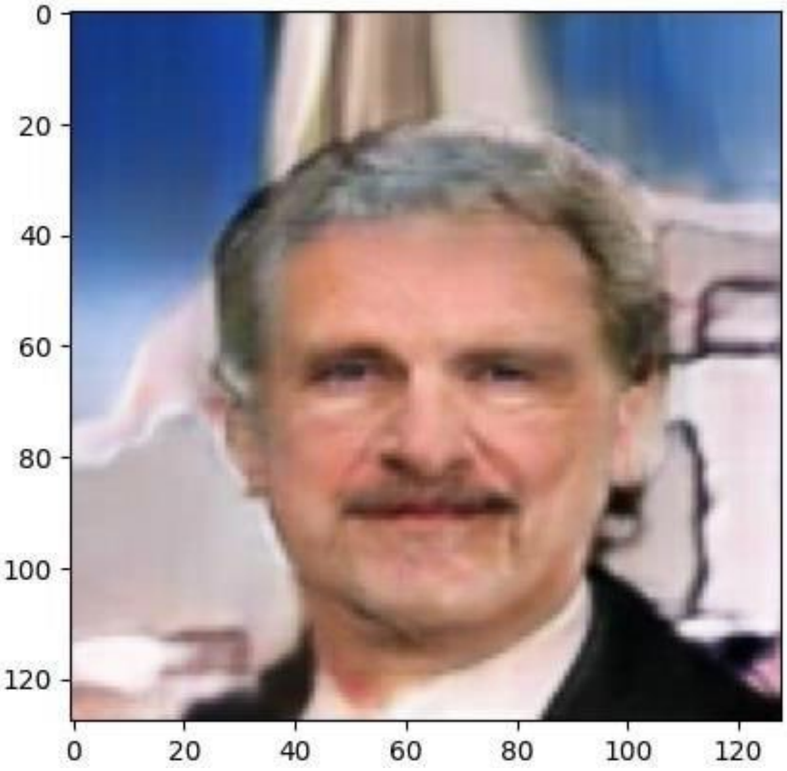
His hair is black. He has a pointy nose. The gentleman is attractive and young. He is wearing a necktie.



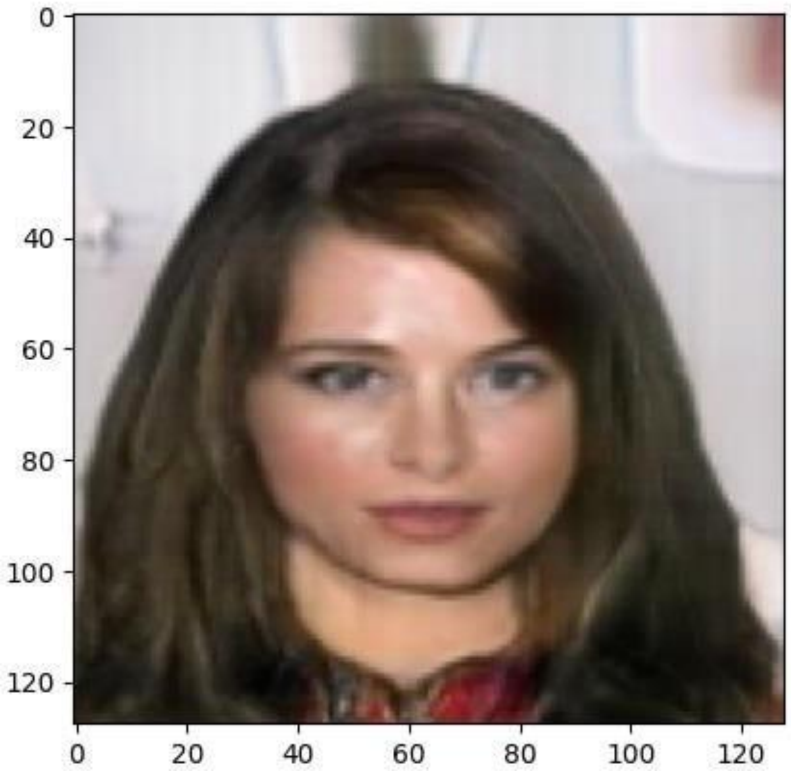
He grows a 5 o' clock shadow and has sideburns. His hair is black. He has bushy eyebrows, a slightly open mouth and a pointy nose. The male is smiling, is attractive and young. He is wearing a necktie.



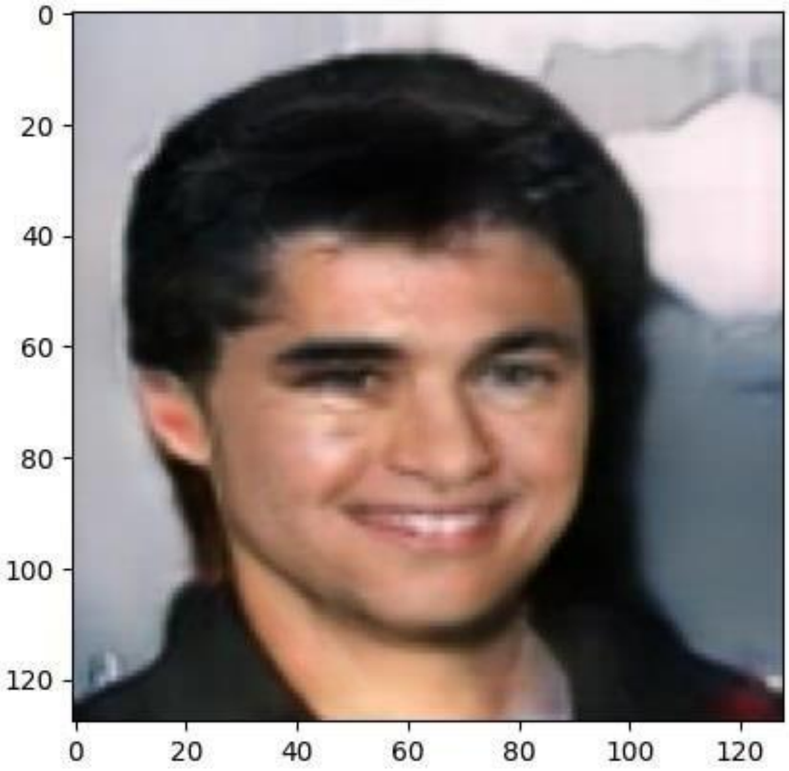
The female has high cheekbones. She has blond hair. She has arched eyebrows, big lips and a big nose. The lady is smiling, has pale skin and heavy makeup. She is wearing earrings and lipstick



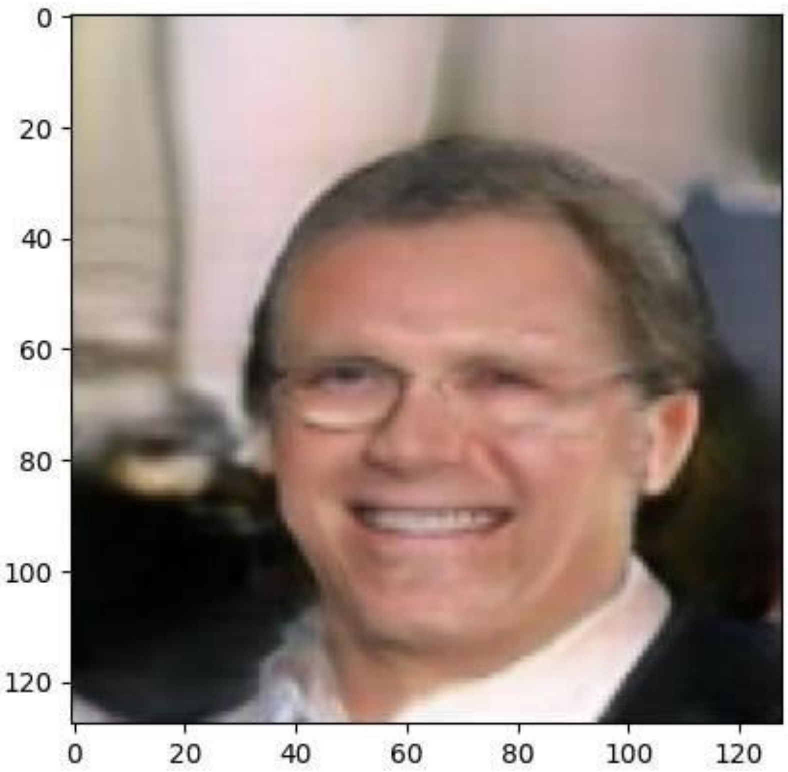
The gentleman has high cheekbones. He sports a goatee. He has gray hair.



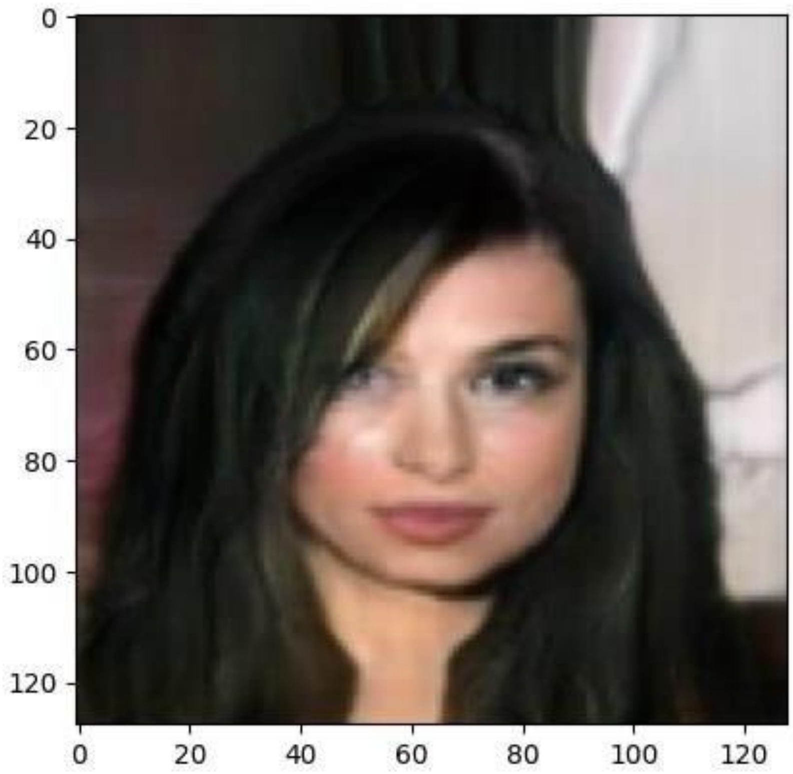
The woman has an oval face. She has brown and wavy hair. She has a pointy nose. The lady is attractive, young, is smiling and has heavy makeup. She is wearing lipstick.



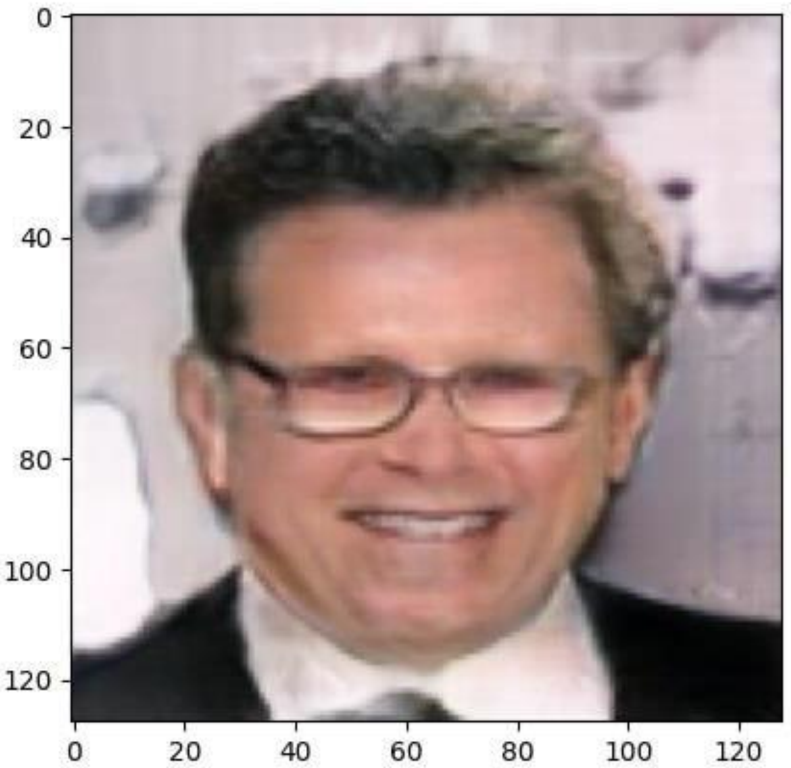
The male has pretty high cheekbones. He has black hair. He is smiling and is young.



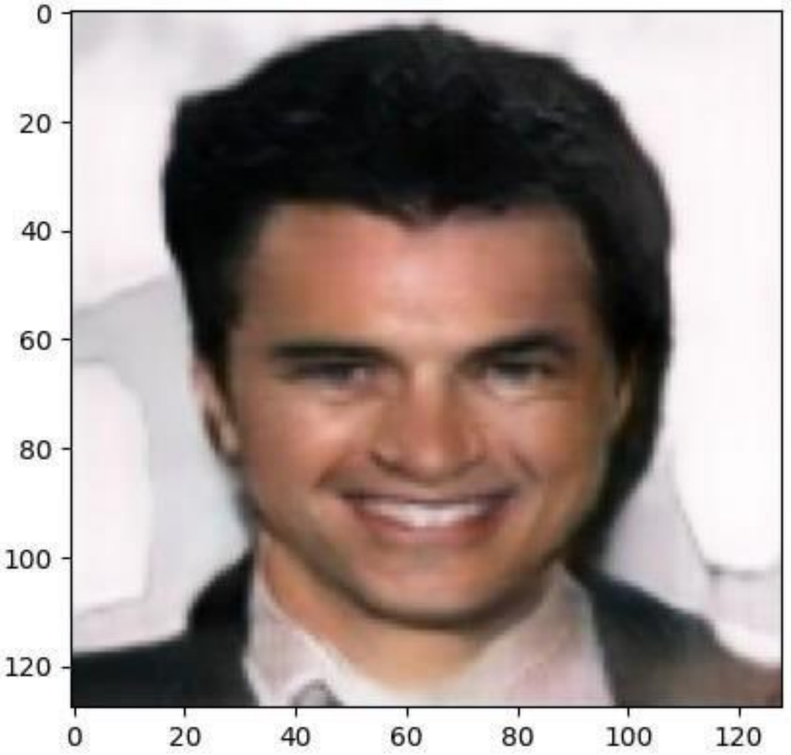
The gentleman has pretty high cheekbones. His hair is gray and straight. He has a slightly open mouth. The male is smiling. He is wearing eyeglasses.



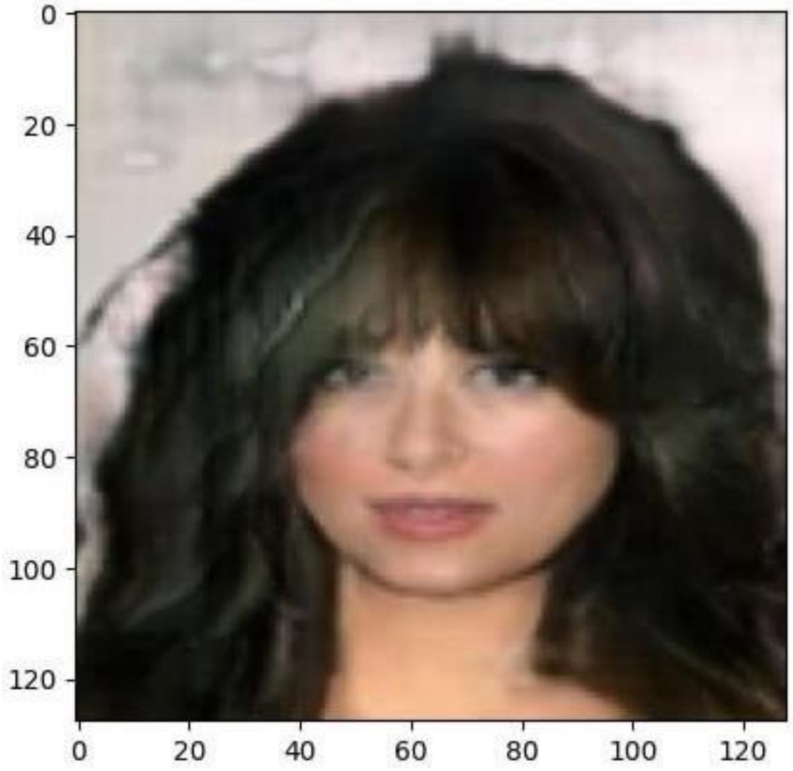
She has black, straight and receding hair. She has arched eyebrows and a big nose. She looks attractive, young and has heavy makeup. She is wearing lipstick.



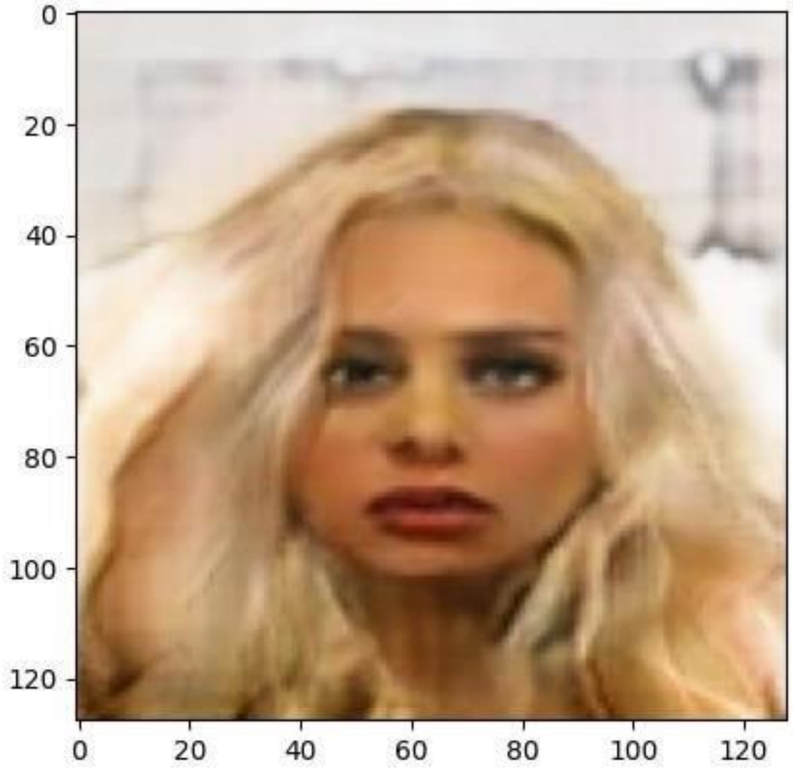
He has gray hair. He has a slightly open mouth. He is smiling. He is wearing eyeglasses and a necktie.



The man has pretty high cheekbones. He has black and straight hair. He has a big nose, bushy eyebrows and a pointy nose. He is smiling.



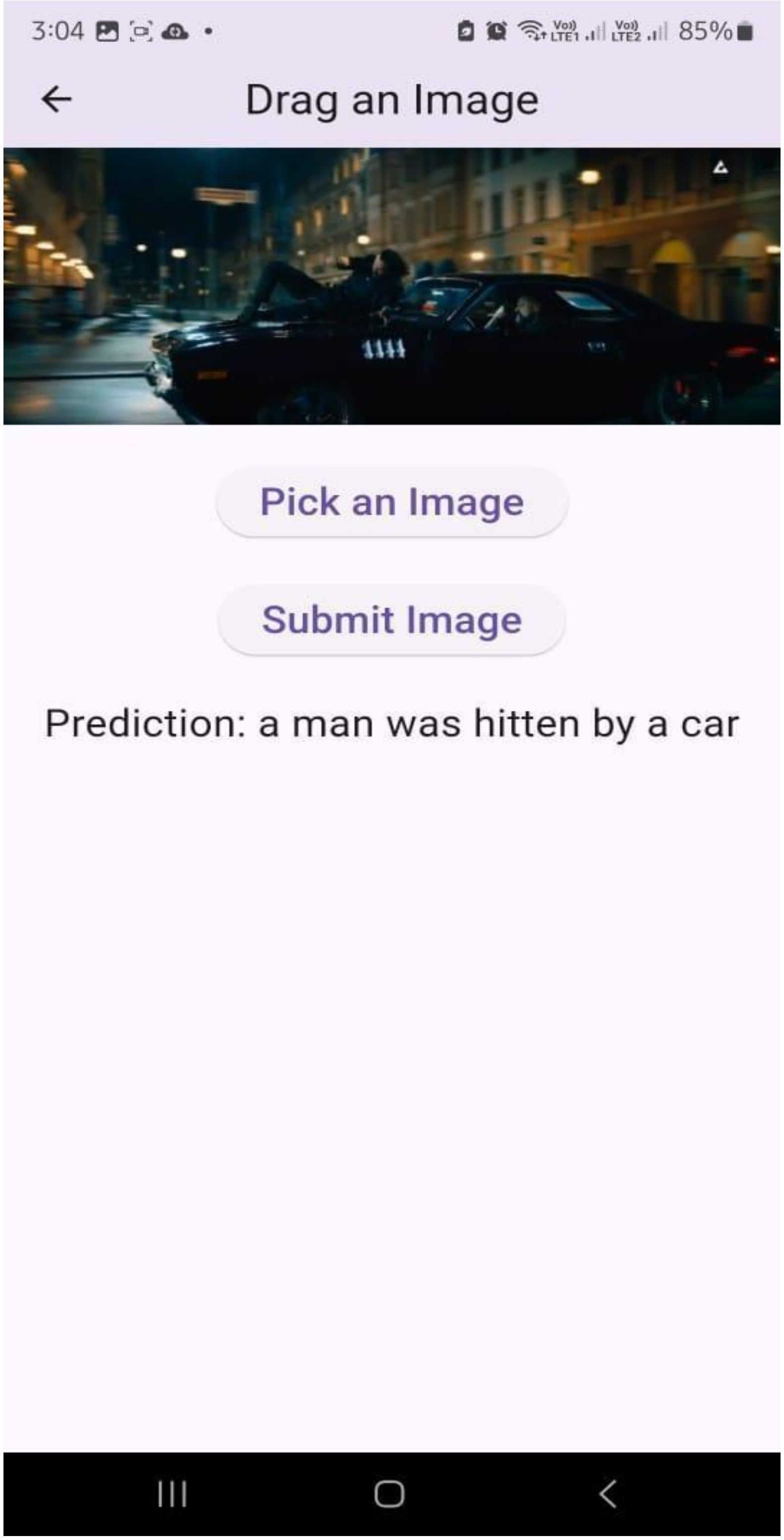
She has brownish black and wavy hair. She has a slightly open mouth and a pointy nose. The woman is attractive, young and has heavy makeup. She is wearing lipstick.



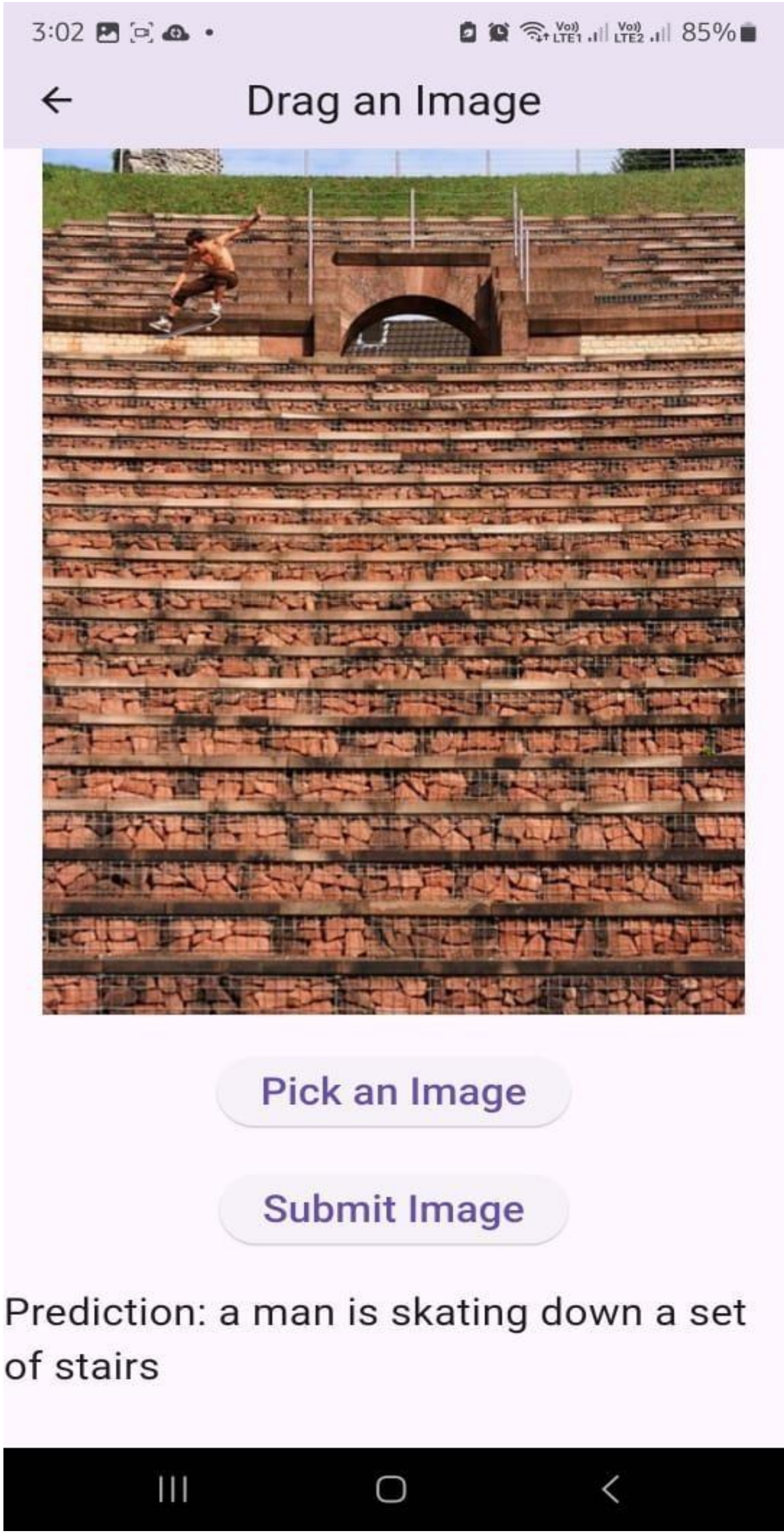
She has blond and wavy hair. The lady looks attractive, young and has heavy makeup. She is wearing lipstick.

4.4 Application of the proposed model:

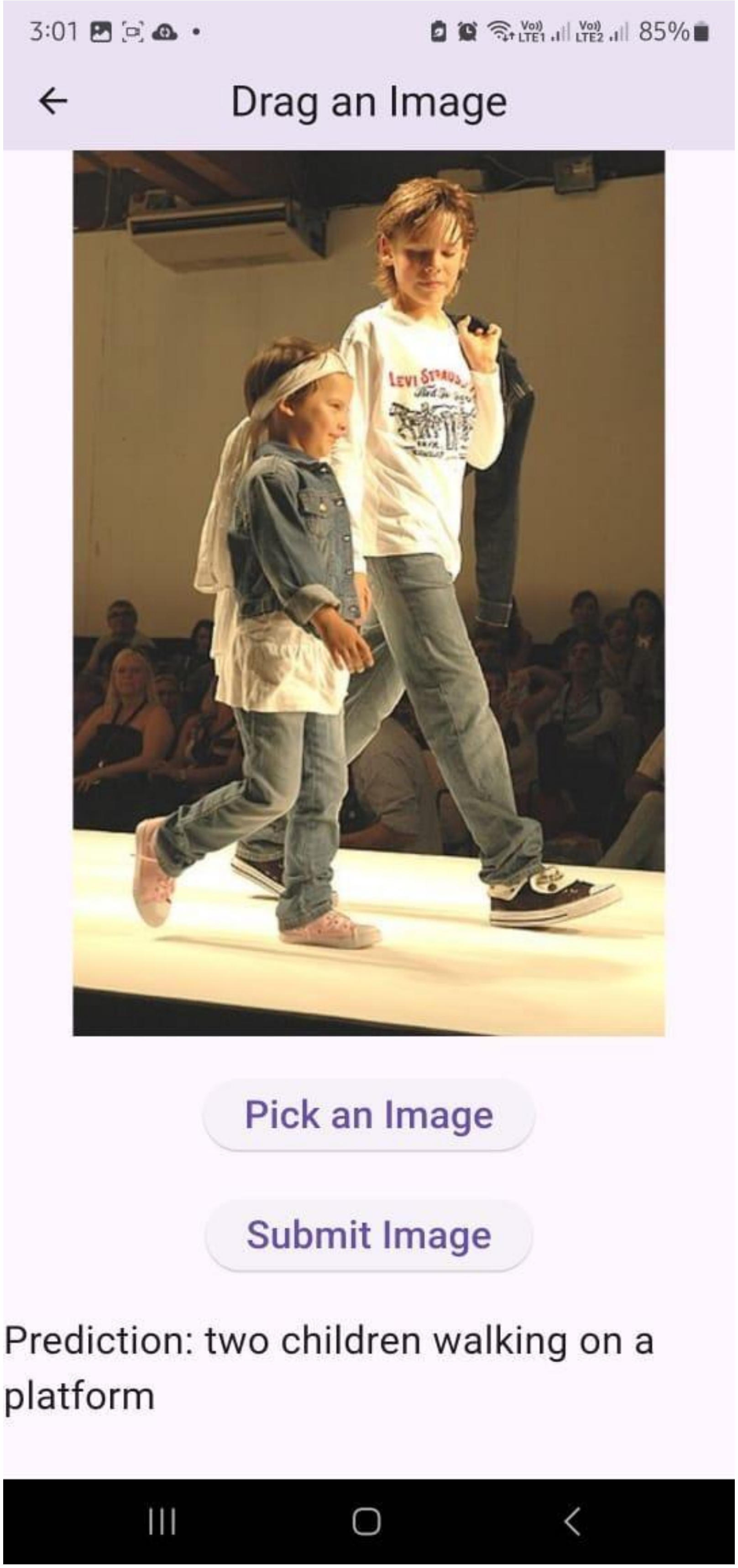
Image to text English:



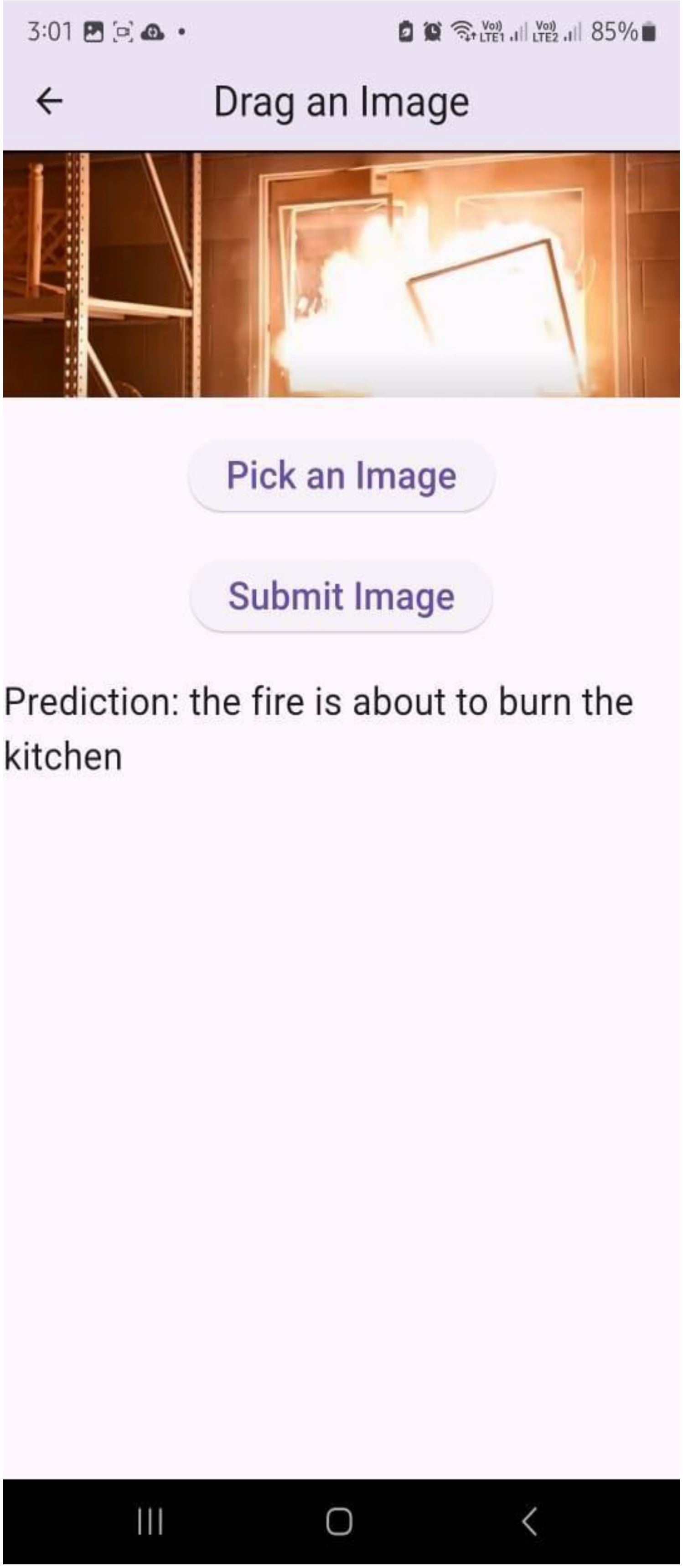
Prediction: a man was hitten by a car



Prediction: a man is skating down a set of stairs

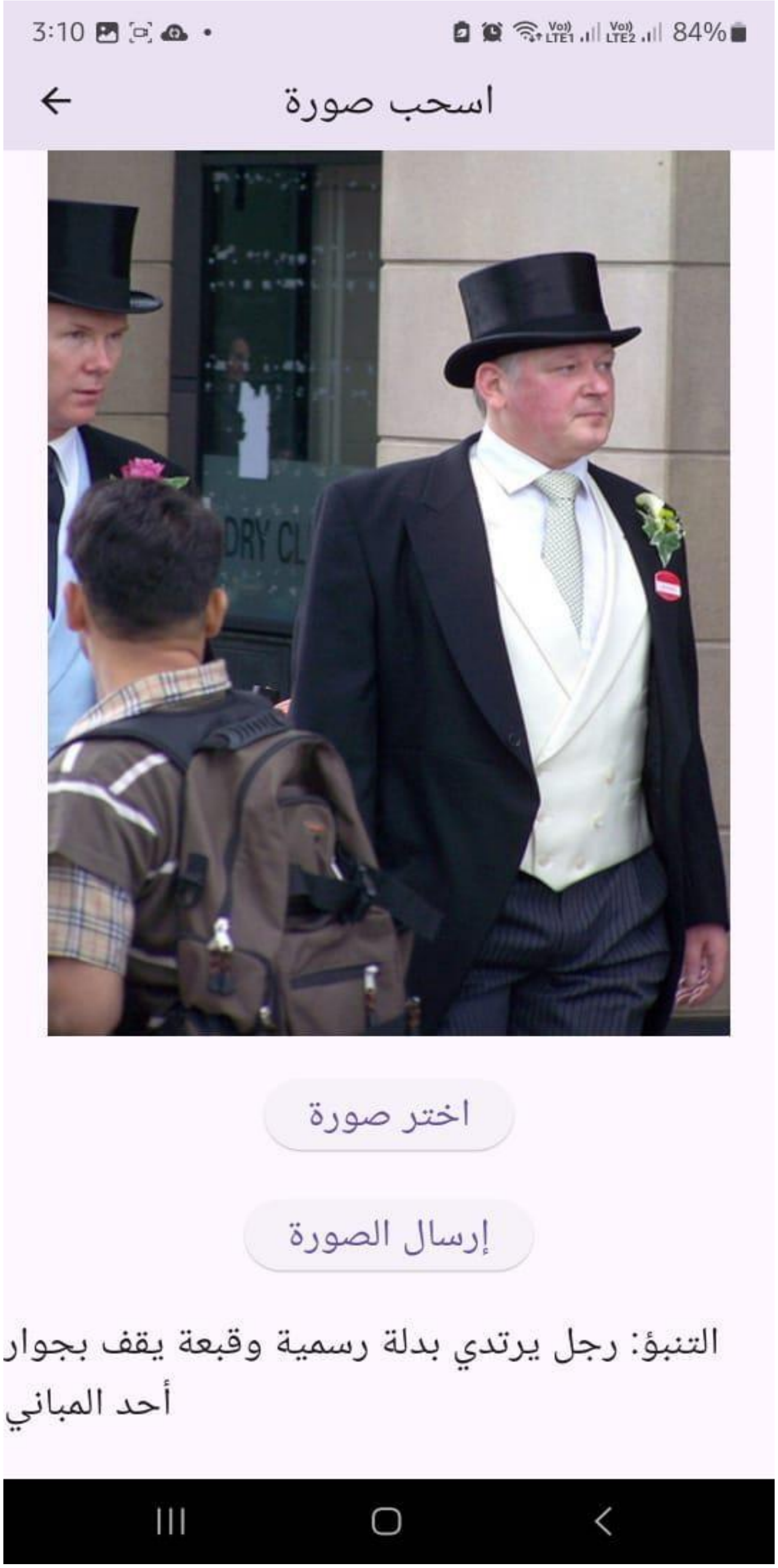


Prediction: two children walking on a platform

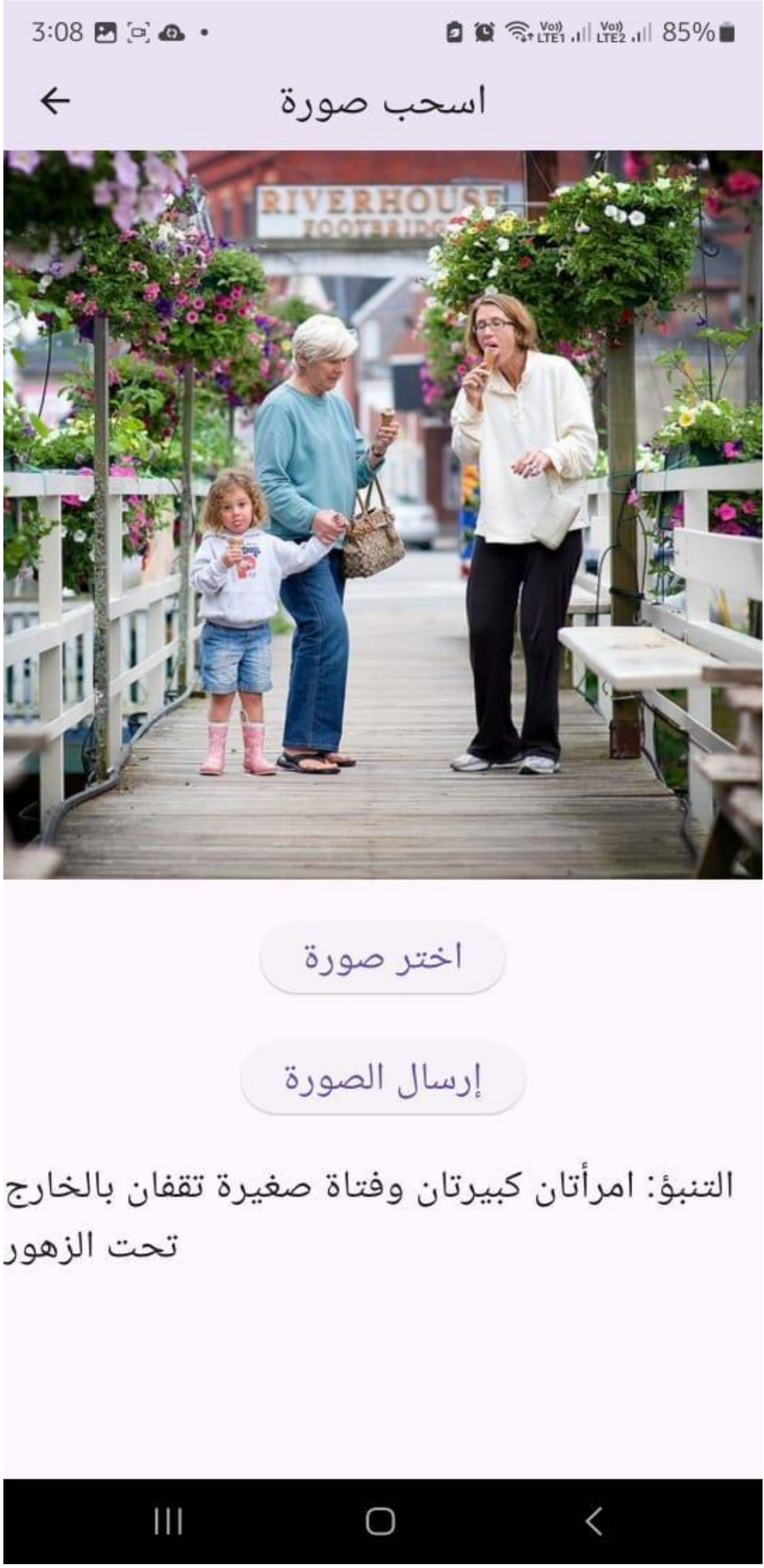


Prediction: the fire is about to burn the kitchen

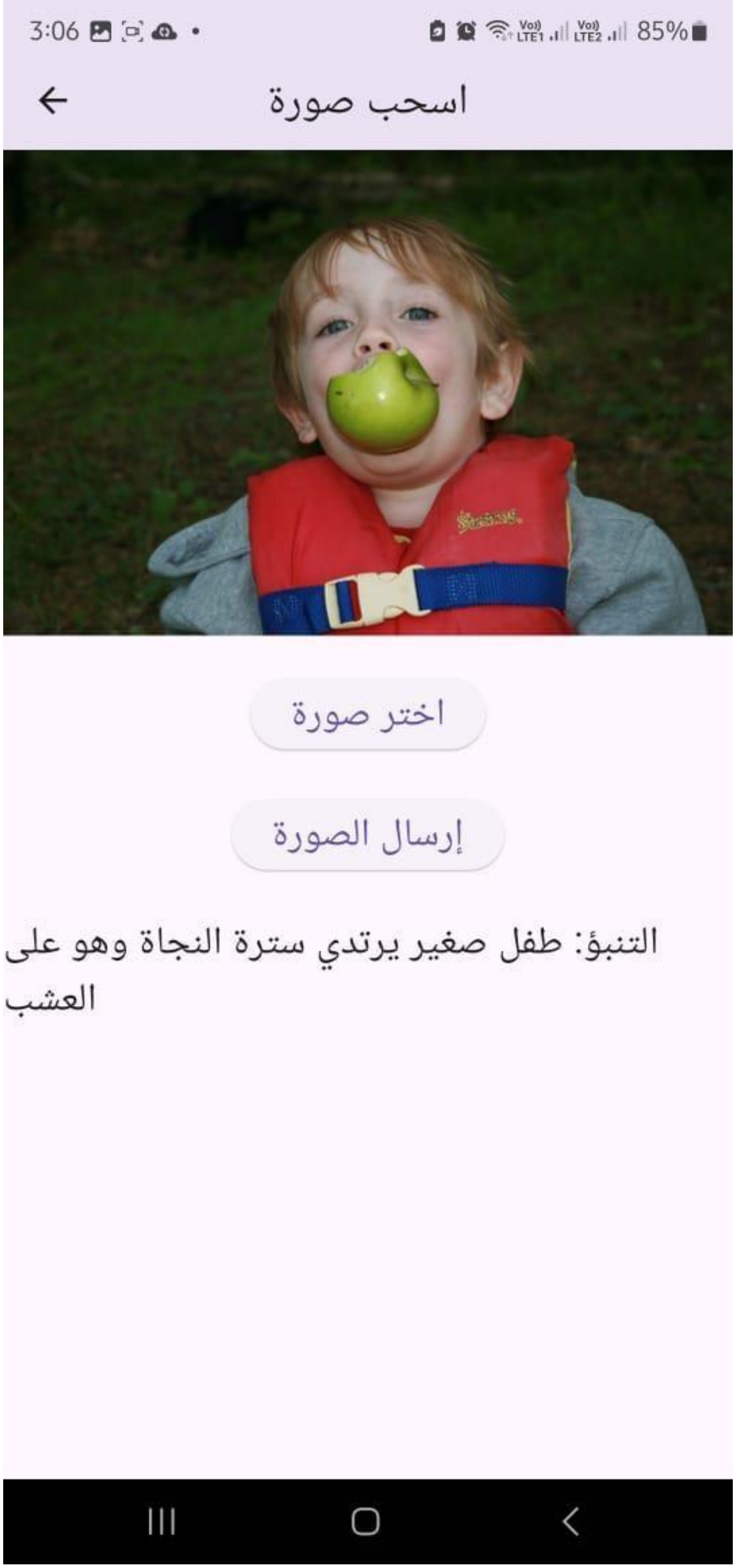
Image to text Arabic:



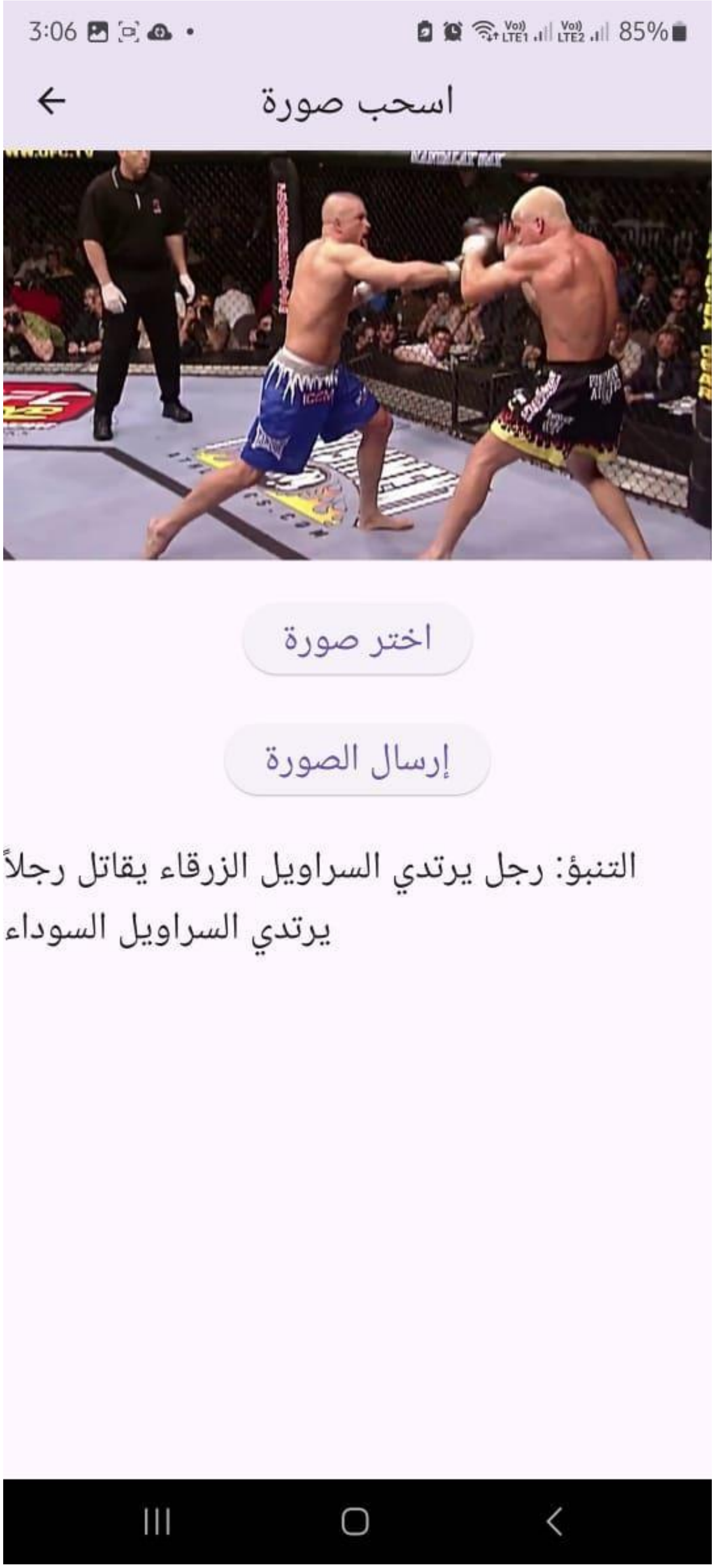
التنبؤ: رجل يرتدي بدلة رسمية وقبعة يقف بجوار
أحد المباني



التنبؤ: امرأتان كبيرتان وفتاة صغيرة تقفان بالخارج
تحت الزهور

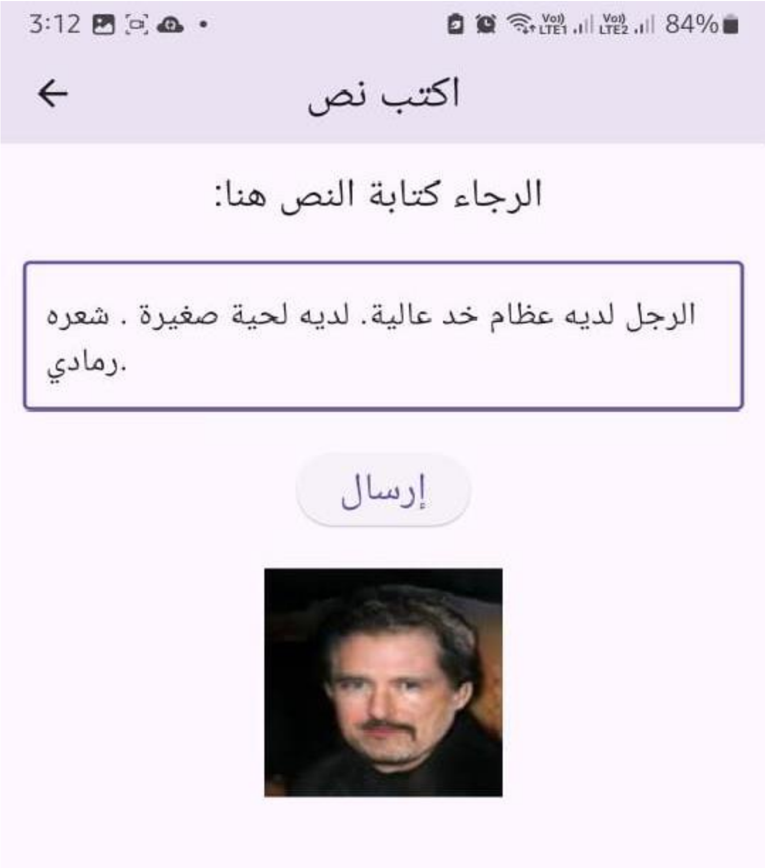


التنبؤ: طفل صغير يرتدي سترة النجاة وهو على العشب

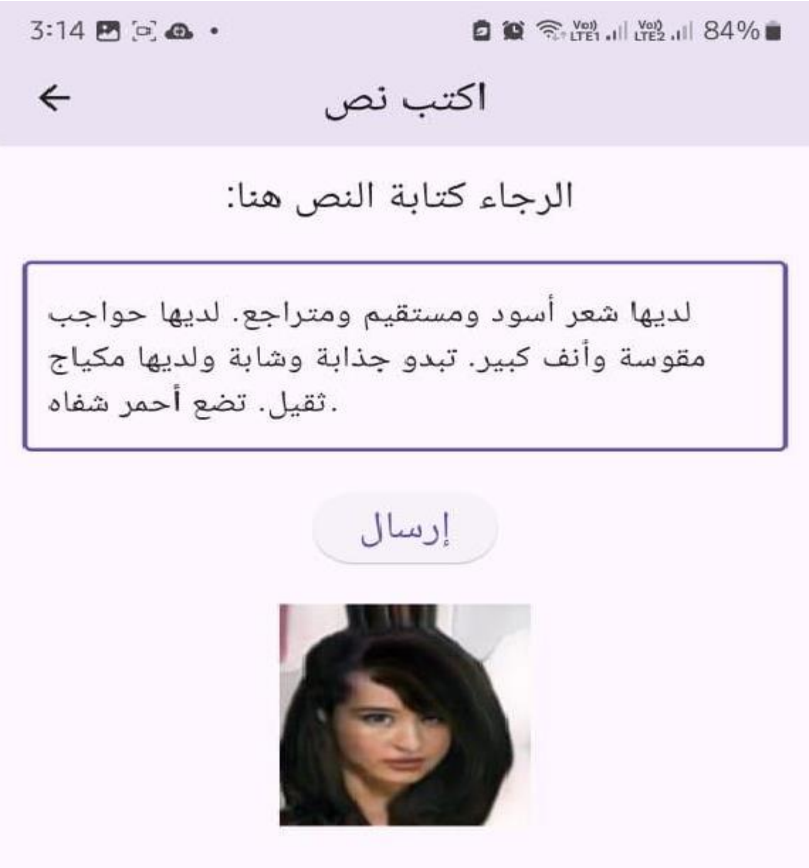


التنبؤ: رجل يرتدي السراويل الزرقاء يقاتل رجلاً يرتدي السراويل السوداء

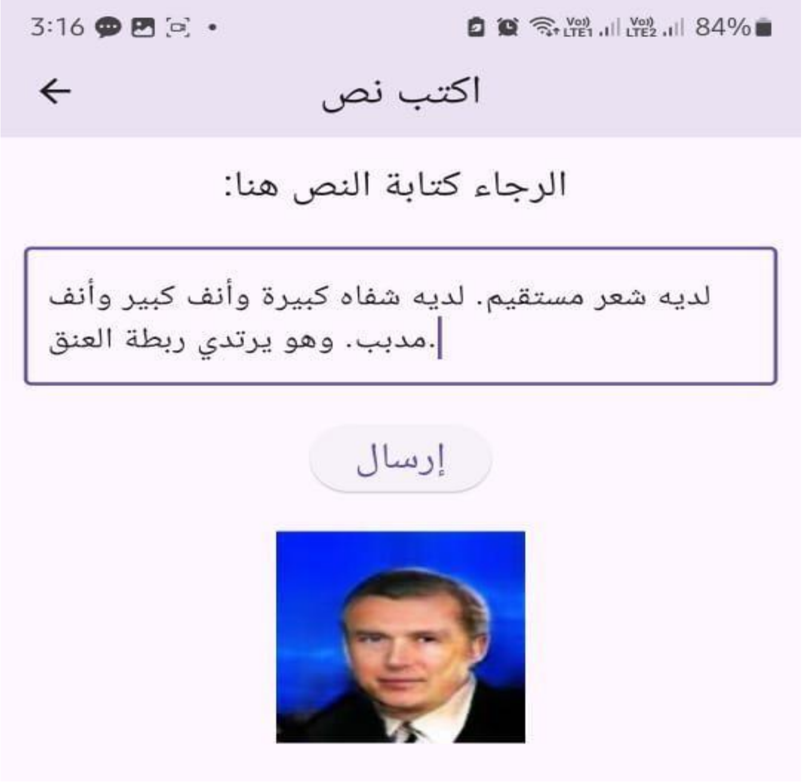
Text to image arabic:



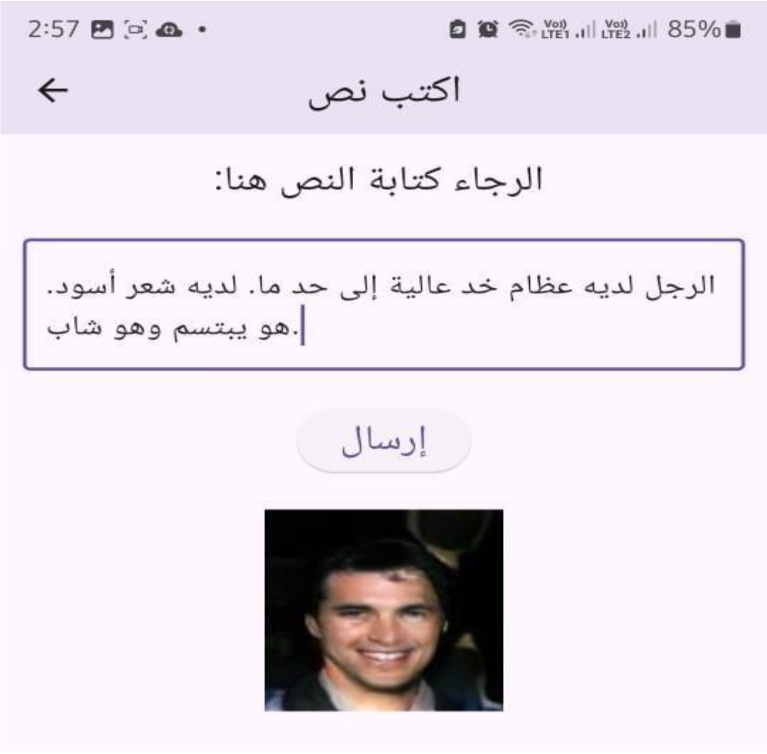
الرجل لديه عظام خد عالية. لديه لحية صغيرة . شعره
رمادي.



لديها شعر أسود ومستقيم ومتراجع. لديها حواجب
مقوسة وأنف كبير. تبدو جذابة وشابة ولديها مكياج
ثقيل. تضع أحمر شفاه

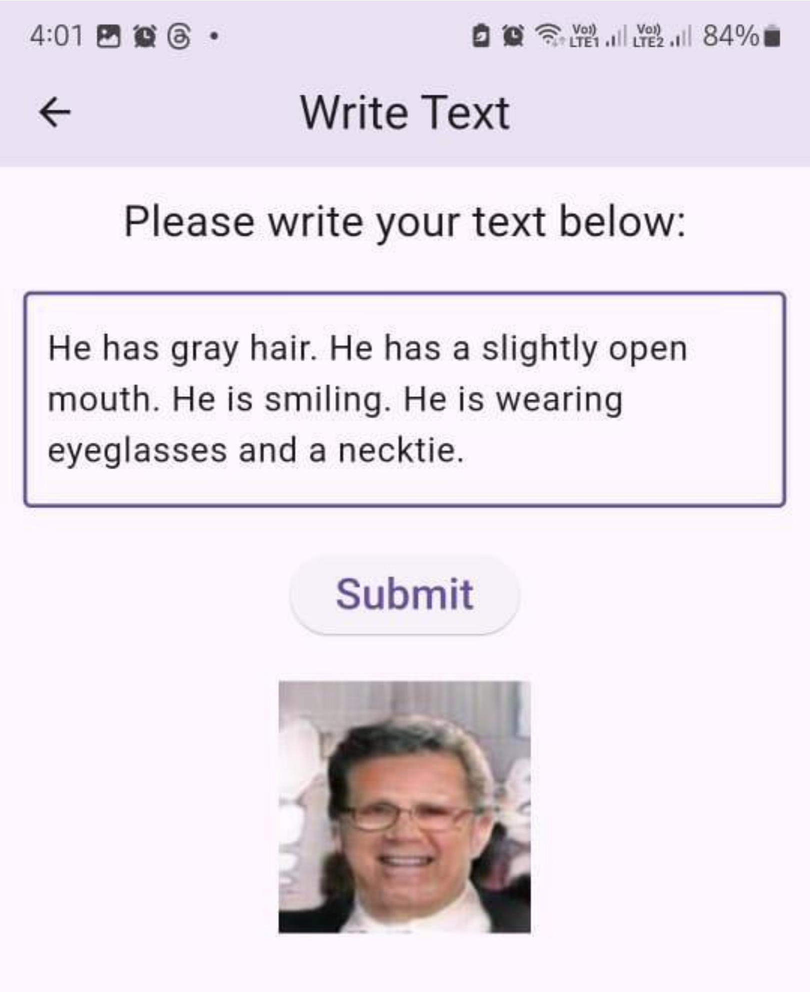


لديه شعر مستقيم. لديه شفاه كبيرة وأنف كبير وأنف
مدبب. وهو يرتدي ربطة العنق



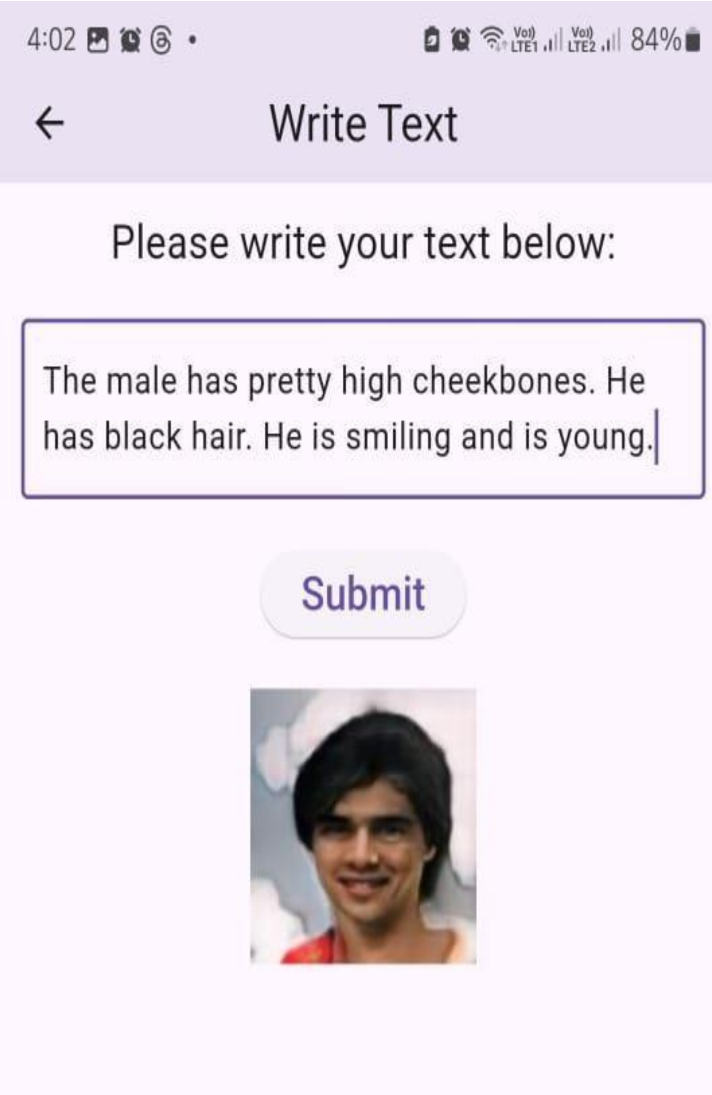
الرجل لديه عظام خد عالية إلى حد ما. لديه شعر أسود.
هو يبتسم وهو شاب.

Text to image english



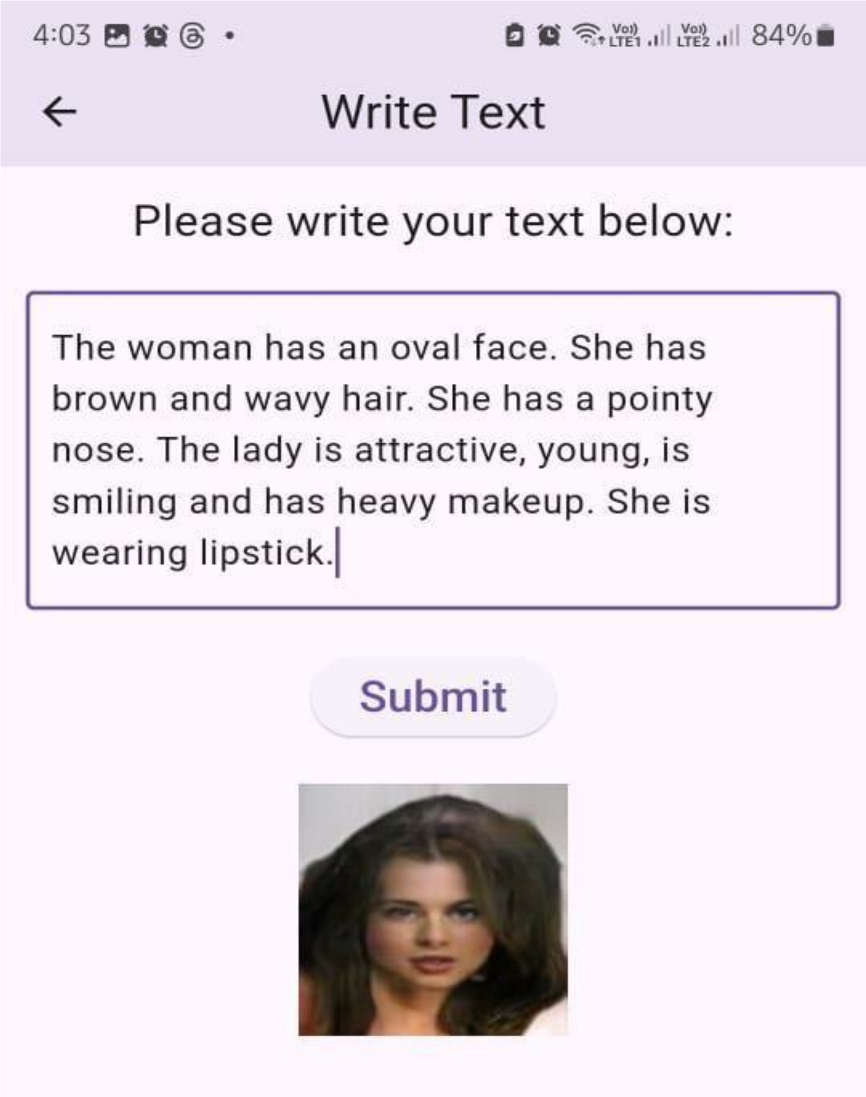
Please write your text below:

He has gray hair. He has a slightly open mouth. He is smiling. He is wearing eyeglasses and a necktie.



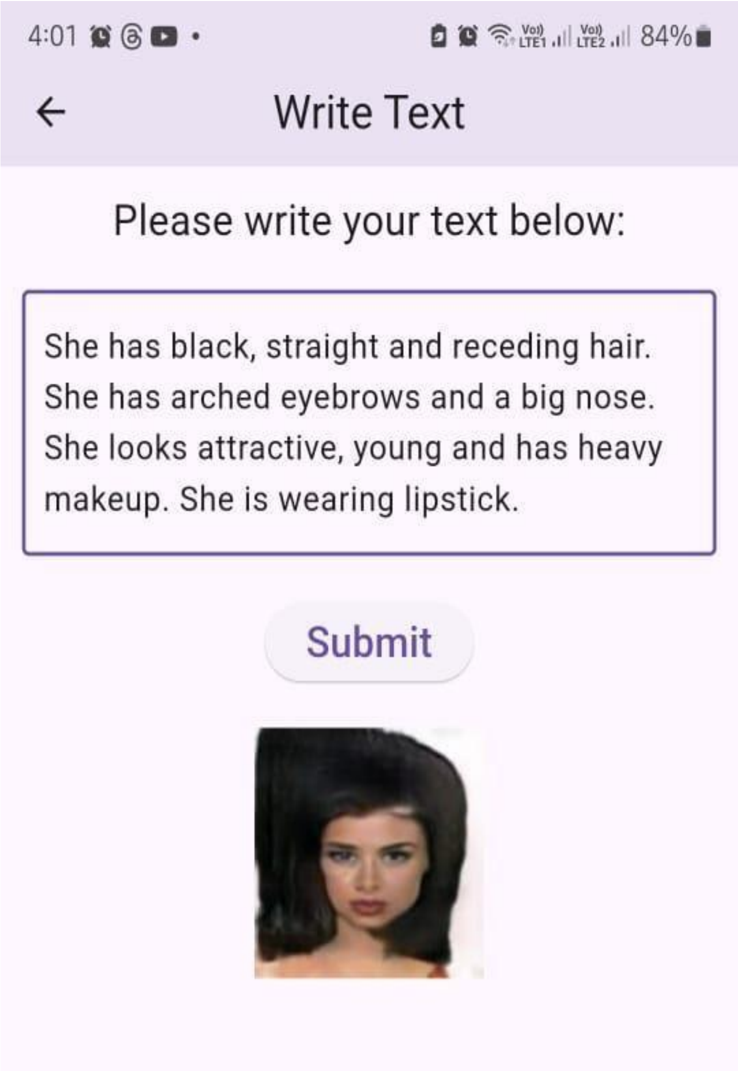
Please write your text below:

The male has pretty high cheekbones. He has black hair. He is smiling and is young.



Please write your text below:

The woman has an oval face. She has brown and wavy hair. She has a pointy nose. The lady is attractive, young, is smiling and has heavy makeup. She is wearing lipstick.



Please write your text below:

She has black, straight and receding hair. She has arched eyebrows and a big nose. She looks attractive, young and has heavy makeup. She is wearing lipstick.

4.5 DISCUSSION:

As we can see there's huge difference between the first draft and the final draft, GANS in general crave data the more data you train them on the stronger they get at generating images.

unlike the first draft where we were only able to train our DFGAN on 20,000 image for 4 epochs on this draft we were able to train on 25,000 images for 22 epoch using the deep learning lab in faculty of computer and artificial intelligence Cairo university which lead to some stunning results.

We Scored in the Inception Score (IS) (which measures the quality and the diversity of the generated images) a score of 9.127 and Scored in the Fréchet inception distance (FID) (which measures how similar the generated images are to real images) a score of 38.689. Compared to some other work on transforming from text to image, these scores are pretty descent.

As we can see in the results, we intended to use some words that wasn't included in the dataset (like the word criminal) and the model was still able to generate descent images.

Objective: Enhance an image captioning model initially trained on the Flickr8k dataset by adding 1000 crime scene images and refining the model.

Pre-adjustment Performance: The model generated captions that were often accurate but lacked detail and context, such as:

1. "Image of a Little Girl with a Milkshake"
2. "Image of Two Girls Performing Karate"
3. "Image of a Skateboarder"
4. "Image of People Riding Bicycles"

Post-adjustment Performance: After incorporating crime scene images and adjustments, the model produced more accurate and contextually relevant captions:

1. "Image of a Little Girl with a Milkshake"
2. "Image of Two Girls Performing Karate"
3. "Image of a Skateboarder"
4. "Image of People Riding Bicycles"

Impact of Crime Scene Images: The inclusion of crime scene images improved the model's ability to handle complex and varied scenes, important for security applications. Examples include:

1. "Image of a Crime Scene"
2. "Image of an Armed Robbery"

Conclusion: The adjustments significantly improved the model’s performance, making its captions more detailed and contextually accurate. The model's versatility and reliability across different scenarios were enhanced.

Arabic Translation: The final captions translated into Arabic demonstrated improved accuracy and context, reflecting the enhancements made to the model

CHAPTER5: CONCLUTION

Our project addresses the challenges faced by visually impaired individuals in understanding their surroundings, particularly in crime scenes or courtrooms, and the difficulties law enforcement agencies encounter in generating accurate images of suspects from verbal descriptions.

For image captioning, we use transformer models:

- **English Model:** Fine-tuned on large datasets to generate accurate, contextually relevant descriptions of images.
- **Arabic Model:** Adapted for Arabic, trained in Arabic datasets to handle linguistic and cultural nuances.

For image generation from text, we employ Deep Fusion GAN (DFGAN):

- **English Model:** Generates high-resolution, realistic images from textual descriptions.
- **Arabic Model:** Translates Arabic text to English and then uses the DFGAN model to generate images, ensuring the integration of Arabic language capabilities.

This dual approach enhances the participation of visually impaired individuals in legal proceedings and improves law enforcement efficiency. The project demonstrates significant advancements in AI-driven image interpretation and synthesis, applicable globally.

Our project is a reference to the power of artificial intelligence by creating a system that tells the visual world to those who can't see, in image captioning We mainly used transformers and tuned the pretrained model. Our proposed model leverages the power of self-attention mechanisms in Transformers to capture spatial dependencies and contextual relationships between different regions of the image. As we build on our promising initial results, we remain committed to our goal of transforming the world by manually collecting better data in the future that helps us get better results. Our model outperforms previous approaches in terms of caption quality, semantic accuracy, and contextual relevance.

As for generating images from text, Deep Fusion GAN (DFGAN) represents a significant advancement in image generation by seamlessly integrating textual information with visual features through deep fusion blocks. This architecture enables the generation of high-resolution images with enhanced realism, thereby producing images that closely match given text inputs. Through features like Matching-Aware-Gradient-Penalty and efficient one-way output, DFGAN demonstrates superior performance in generating realistic and text-matching images, marking a promising direction in the field of generative adversarial networks.

REFRENCES

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In European Conference on Computer Vision. Springer, 15–29.

[2] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In Proceedings of the 15th Conference on Computational Natural Language Learning. Association for Computational Linguistics, 220–228.

[3] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In Proceedings of the 24th CVPR. Citeseer.

[4] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In Proceedings of the 31st International Conference on Machine Learning (ICML'14). 595–603.

[5] Andrej Karpathy, Armand Joulin, and Fei Fei F. Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In Advances in Neural Information Processing Systems. 1889–1897.

[6] Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2422–2431.

[7] Zhu, X.; Li, L.; Liu, J.; Peng, H.; Niu, X. Captioning Transformer with Stacked Attention Modules. Appl. Sci. 2018, 8, 739.

[8] Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image Captioning: Transforming Objects into Words. In Proceedings of the Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 11135–11145.

[9] Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on Attention for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4633–4642.

[10] Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8927–8936.

[11] He, S.; Liao, W.; Tavakoli, H.R.; Yang, M.; Rosenhahn, B.; Pugeault, N. Image Captioning Through Image Transformer. In Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 30 November–4 December, 2020; Ishikawa, H., Liu, C.L., Pajdla, T., Shi, J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 153–169.

[12] Nils Reimers, Iryna Gurevych 2018. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

[13] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, Changsheng Xu 2020. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis.

[14] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena 2019. Self-Attention Generative Adversarial Networks.

[15] Shane Barratt, Rishi Sharma 2018. A Note on the Inception Score.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich 2014. Going Deeper with Convolutions

[18] Jonathon Shlens 2014. Notes on Kullback-Leibler Divergence and Likelihood

[19] A Borji - Computer Vision and Image Understanding, 2022 - Elsevier. Pros and cons of GAN evaluation measures: New developments. [PDF]

[20] E Härkönen, A Hertzmann… - Advances in neural …, 2020 - proceedings.neurips.cc. Ganspace: Discovering interpretable gan controls. neurips.cc

[21] Z Wang, G Healy, AF Smeaton, TE Ward - Cognitive Computation, 2020 - Springer. Use of neural signals to evaluate the quality of generative adversarial network perfor-mance in facial image generation. [PDF]

[22] Marr, B. (2018). How AI And Machine Learning Are Transforming Law Enforcement. Forbes.

[23] Ferguson, A. G. (2017). The Rise of Big Data Policing. New York University Press.

[24] Coppersmith, G. G. (2020). Ethical Implications of AI And Data Science In Policing. Journal of Ethics in Information Technology.

[25] Ng, A. (2017). The State of Artificial Intelligence. AI Pioneers Conference.

[26] Gebru, T. (2020). Race and Gender in AI: Ethical Implications. Proceedings of the ACM.