Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.DOI

Advancing Image Captioning: A Comparative Analysis of CNN-LSTM and CNN-GRU Architectures with Diverse Pre-trained Models

ABDULRAOUF MONIR[*1,2], ADHAM AHMED ABDELMAKSOUD[*1,3], ENGY AHMED HASSAN[*1,4], MOHAMED ALY MATAR[*1,5], OMAR HELMY ELBANNA[*1,6],

[1]Faculty of Engineering, Ain Shams University, Cairo, Egypt.
[2]e-mail: 19P4442@eng.asu.edu.eg
[3]e-mail: 19P1250@eng.asu.edu.eg
[4]e-mail: 19P4390@eng.asu.edu.eg
[5]e-mail: 19P5238@eng.asu.edu.eg
[6]e-mail: 19P3904@eng.asu.edu.eg

ABSTRACT

The pervasive use of social media and the surge in image-sharing platforms have led to an unprecedented volume of visual data. As a response, automated image captioning has emerged as a crucial task in the field of computer vision. This paper investigates the efficacy of deep learning techniques, specifically the combination of Convolutional Neural Networks (CNNs) with Long Short Term Memory (LSTM) in the context of image captioning. Our experiment involves introducing diversity by incorporating various pre-trained models, including ResNet and mobileNet, into the CNN architecture to enhance feature extraction and subsequently improve the quality of generated captions and then using LSTM for generating the captions. To further advance the state-of-the-art in image captioning, we extend our investigation to include a hybrid approach utilizing Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs). This combination aims to harness the strengths of CNNs in spatial feature extraction and GRUs in capturing sequential. The experimental evaluation, conducted on widely-used benchmark English and Arabic datasets, employs comprehensive performance metrics like BLEU. Results highlight the substantial impact of integrating different pre-trained models within the CNN on the overall performance of image captioning systems. The study provides a detailed comparative analysis, shedding light on the specific strengths and trade-offs associated with each pre-trained model, offering a nuanced understanding of how different pre-trained models influence the generation of high-quality captions for images in the context of both CNN with LSTM and CNN with GRU.

INDEX TERMS Image Captioning, Deep Learning, Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), Gated Recurrent Units (GRU), Pre-trained Models

I. INTRODUCTION

On a daily basis, we come across numerous images sourced from various outlets like the internet, news articles, document diagrams, and advertisements. These images often lack accompanying descriptions, relying on viewers to interpret them without detailed captions. While humans can generally understand images without explicit captions, machines require some form of image captions for automatic description generation. The significance of image captioning is manifold. It plays a crucial role in automatic image indexing, a process vital for Content-Based Image Retrieval (CBIR). This functionality extends its application to diverse fields such as biomedicine, commerce, the military, education, digital libraries, and web searching.

Image captioning (IC) is the task of automatically generating a description of an image. It is the combination of

computer vision and natural language processing. Given an input image I, the goal is to generate a caption C describing the visual contents present inside the given image, with C being a set of sentences C = {c1, c2, ..., cn} where each ci is a sentence of the generated caption C. However, unraveling the intricacies of image content and establishing coherent connections between visual elements pose significant challenges in this domain.

One of the primary challenges in image captioning lies in deciphering the nuanced visual features present in images. Traditional computer vision approaches often struggle to capture the contextual subtleties and intricate relationships among various visual elements. Moreover, generating accurate and contextually relevant captions requires a profound understanding of the semantics embedded within the visual data.

To address these challenges, our research focuses on leveraging classical deep learning techniques, specifically through the integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, building upon the work done by [1]. We also extend our investigation to include a hybrid approach utilizing Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs). This combination is chosen for its capacity to capture intricate visual patterns, long-range dependencies, and temporal dynamics within the image data, paving the way for more sophisticated image understanding.

In this study, Our Contribution is as follows:

- A wise choice is made in selecting english and arabic datasets to pursuit excellence in model learning.
- Images pre-processing including resizing to fit the pre-trained CNN input size.
- Extracting features using diverse pretrained models, including ResNet and mobileNet, incorporated into the CNN architecture.
- Integration of LSTM and GRU techniques for generating captions.
- Conducting comprehensive experimentation and evaluation on benchmark datasets.
- Detailed comparative analysis of various pretrained models to understand their specific impacts on system performance.

The rest of the paper is organized as follows: section II presents the collected related work. Section III introduces the proposed models. The experimental results are shown in section IV. Section V concludes the paper and provides some future works.

II. RELATED WORK

Many approaches of Image Captioning have been developed and can be categorized into three approaches as follows:

A. RETRIEVAL-BASED APPROACH

In the initial phases of image captioning, a frequently employed method is the retrieval-based approach. In this approach, when presented with an image for captioning, the system generates a caption by choosing one or more sentences from a pre-established set of sentences. The resulting caption may consist of an existing sentence or a newly crafted one derived from the retrieved sentences.

In [2], the author uses a discriminative approach to score a match between an image and a sentence based on the similarity of their predicted triplets. The triplets referred to in the approach are composed of three elements: object, action, and scene. These triplets provide a holistic representation of the content depicted in an image or described in a sentence. The system uses pre-trained models to detect objects and classify scenes in the image. The system then uses these features to compute node potentials -likelihood of specific objects, actions, or scenes being present in an image- and edge potentials – which capture the relationships or connections between these elements within the meaning space- which are then used to predict triplets for the image. The system predicts these triplets for both images and sentences, and then evaluates the similarity of the predicted triplets to establish a match between an image and a sentence. This matching process enables the system to generate descriptive sentences from images and find images that best correspond to given sentences.

Expanding on the retrieval-based approach, a new method has been introduced, known as the concept-based pipeline [3]. This approach enhances both image and sentence retrieval tasks by using visual concepts automatically identified by the system. Instead of relying solely on predefined sentences, the system generates descriptions based on these visual concepts, leading to more nuanced and contextually relevant image captions. This bidirectional approach improves the connection between visual content and textual information, making the overall image captioning process more flexible and adaptive.

Author of [4] introduced a two-stage method for captions retrieval, aiming to enhance the connection between visual data and natural language descriptions. It utilizes web images and their corresponding tags to develop a robust joint representation for both images and text. In the initial stage, a supervised approach is employed to learn a well-aligned representation that can be shared between visual and textual modalities. Subsequently, in the second stage, weakly annotated pairs of images and tags from the web are used to further refine the shared representation learned in the first stage. The motivation behind this approach draws from the concept of learning using privileged information and employs multi-task learning strategies within deep neural networks. Experimental results indicate a significant enhancement in the performance of the image-text retrieval task across two benchmark datasets, demonstrating the effectiveness of the proposed method. This paper added to the image-text retrieval task by addressing the challenges of limited labeled data and noisy tags associated with web

images.

While the produced outputs are typically grammatically sound and coherent, limiting image descriptions to pre-existing sentences may fail to accommodate new combinations of objects or unfamiliar scenes. In specific situations, the generated descriptions might even be unrelated to the actual contents of the image. Retrieval-based methods exhibit significant limitations in their ability to effectively describe images.

B. TEMPLATE BASED APPROACH

In the initial stages of image captioning research, another frequently employed method is template-based. This approach involves generating image captions through a process that adheres to syntactic and semantic constraints. To employ a template-based method for generating a description for an image, a predefined set of visual concepts must be initially detected. Subsequently, these detected visual concepts are interconnected using sentence templates and specific language grammar rules to construct a sentence.

Yang et al. [5] introduced a method for employing a sentence template in generating image descriptions, utilizing a quadruplet format (Nouns-Verbs-Scenes-Prepositions). In this approach, to depict an image, the authors initially utilize detection algorithms to estimate objects and scenes within the image. Subsequently, a language model, trained on the Gigaword corpus 3, is applied to predict verbs, scenes, and prepositions suitable for constructing the sentence. By computing probabilities for all elements, the optimal quadruplet is determined through Hidden Markov Model inference. Finally, the image description is generated by filling the sentence structure outlined by the selected quadruplet.

Ushiku et al.[6] introduce a novel approach called "Common Subspace for Model and Similarity" for the direct learning of phrase classifiers in the context of image captioning. The method involves the extraction of continuous words from training captions, treating them as phrases. The next step includes mapping both image features and phrase features into a shared subspace. Within this subspace, the authors integrate similarity-based and model-based classification methods to effectively train a classifier for each identified phrase. During the testing phase, when confronted with a query image, the method estimates relevant phrases, and these phrases are interconnected using a multi-stack beam search approach. This sophisticated technique enhances the generation of a coherent and contextually relevant description for the given image.

Captioning images based on templates can produce sentences that are grammatically correct, and the resulting descriptions are often more closely aligned with the content of the image compared to retrieval-based methods. However, template-based approaches also come with drawbacks. The generation of descriptions within the template framework is tightly bound to the image contents identified by visual models. With a typically limited number of available visual models, these methods may face constraints on coverage, creativity, and the complexity of the sentences they generate. Furthermore, relying on rigid templates as the primary structures for sentences can lead to descriptions that are less natural when compared to captions written by humans.

C. DEEP NEURAL NETWORKS APPROACH

Retrieval and template image captioning methods have limitations on the sentences they create. Thanks to advanced deep neural networks, there are new ways to describe images without using existing captions or following specific sentence structures. These methods can generate more expressive and flexible sentences with richer structures. One approach is to use multimodal neural networks, which learn directly to generate image captions.

1) CNN with RNN

In order to create fresh captions for images [7] Mao et al. modify a Recurrent Neural Network language model to suit multimodal scenarios, directly capturing the likelihood of generating a word based on a given image and previously generated words. In their approach, they employ a deep Convolutional Neural Network to extract visual features from images [8]. Additionally, a Recurrent Neural Network, incorporating a multimodal component, is utilized to model word distributions conditioned on both image features and context words.

The author [9] introduces a distinctive dimension to image captioning by proposing the task of dense captioning. In contrast to conventional image captioning approaches that typically generate a single description for an entire image, dense captioning requires the computer vision system not only to identify but also to describe multiple salient regions within an image using natural language. This sets it apart from traditional image captioning methods, offering a more granular and detailed narrative of the visual content. To address this novel task, the authors present the Fully Convolutional Localization Network (FCLN) architecture which combines a Convolutional Neural Network, a novel dense localization layer, and a Recurrent Neural Network language model. This integrated architecture allows the system to efficiently process images, be trained end-to-end, and generate multiple descriptions across different regions of an image.

Recurrent Neural Networks (RNNs) are acknowledged to face challenges in effectively capturing long-term dependencies. To address this limitation in the context of image captioning, [10] the author proposed bi-directional model for image caption generation stands out due to its capability to grasp and retain long-term interactions and concepts within visual scenes. It achieves this through the incorporation

of a recurrent visual memory, dynamically capturing the evolving visual aspects of the scene as the caption is either generated or read. This recurrent visual memory undergoes updates to reflect the incoming information associated with each word, enabling the network to automatically prioritize and include salient concepts that have not been expressed yet. This approach differs from traditional techniques that often struggle with long-term dependencies, facing challenges in remembering concepts over multiple iterations of recurrence. This innovative design allows the network to produce more precise and meaningful captions, overcoming the limitations of forgetting crucial context over time. The effectiveness of this model has been demonstrated in various tasks, including sentence generation, sentence retrieval, and image retrieval, where it has achieved state-of-the-art results.

The author in this paper [11] emphasizes on adding multimodal embedding to the CNN and bidirectional RNN, the model aligns language and visual modalities, creating a shared space for representation. This enables the mapping of image regions and words, facilitating effective alignment and comparison of visual and language features. The model further deduces latent correspondences between sentence segments and image regions, aligning language and visual data based on content and context. This integrative approach, combining CNNs and bidirectional RNNs while employing multimodal embedding, allows the model to generate natural language descriptions of images and their regions. Notably, it achieves this without relying on predetermined templates or assumptions about specific categories, learning dynamically from the training data to reason about image content and representation in natural language.

In this work [12] , The m-RNN model is composed of two interconnected sub-networks: a language model segment and a vision segment, linked through a multimodal layer. Within the language model segment, a two-layer word embedding system and a recurrent layer work collaboratively to learn dense feature embeddings for each word in the sentence description. The recurrent layer, in particular, retains the temporal context of the sentence, enabling the generation of flexible and intricate sentence structures. On the other hand, the vision segment utilizes a convolutional neural network (CNN) like VggNet or AlexNet to extract visual features from the input image. These visual features, alongside word representations from the language model segment, are fed into the multimodal layer. The multimodal layer effectively combines the extracted visual and language features to generate the probability distribution for the next word in the sentence description. This iterative process continues until the completion of the sentence, yielding a comprehensive sentence description for the input image.

2) CNN with LSTM

In the intersection of computer vision and natural language processing, the integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks has driven substantial progress in multi-modal learning. This approach explores diverse CNN-LSTM architectures tailored for image captioning, highlighting their adeptness in harmonizing visual and sequential data for enhanced performance.

The author [13] introduces a multi-modal Recurrent Neural Network (m-RNN) model, leveraging the synergies between deep Convolutional Neural Networks (CNNs) and recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, for the task of image captioning. The model employs a deep CNN for extracting features from images and a deep LSTM for processing sequential sentence information. Through interaction in a multi-modal layer, the model generates image captions by modeling the probability distribution of words given previous words and the image. The effectiveness of this approach is demonstrated in generating innovative image descriptions and excelling in both captioning and retrieval tasks. The m-RNN model is showcased for its flexibility, with the potential to incorporate complex image representations and advanced language models, making it a promising avenue for multi-modal learning advancements. Superior performance is validated on benchmark datasets, including IAPR TC-12, Flickr 8K, Flickr 30K, and MS COCO, surpassing state-of-the-art methods. In summary, the m-RNN model represents a notable stride in connecting images and sentences for image captioning and multi-modal learning.

This paper [14] introduces Long-term Recurrent Convolutional Networks (LRCN), a robust architecture integrating CNNs for visual recognition with LSTM networks for handling sequential data. LRCN excels in processing variable-length visual inputs and generating variable-length outputs, such as comprehensive sentence descriptions. The model features a CNN for hierarchical visual feature extraction and an LSTM module to capture temporal dependencies. By combining these elements, the LRCN forms an end-to-end trainable network suitable for complex visual and sequence prediction tasks. Experimental results highlight the effectiveness of LRCN in various applications, including video activity recognition, image caption generation, and video description. The model showcases substantial improvements over state-of-the-art counterparts, especially with ample training data. In essence, LRCN stands out as a potent architecture adept at modeling temporal dependencies and generating coherent natural language descriptions from visual inputs across diverse tasks.

The author in this paper [15] introduces a method for image captioning is proposed that combines Language CNN (CNNL) with recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM). CNNL addresses traditional RNN limitations by capturing long-range depen-

dencies in word sequences through multi-layer ConvNets. By integrating CNNL with LSTM, the model aims to effectively capture and represent sequences, thereby enhancing performance in image captioning tasks. Evaluation on datasets such as Flickr30K and MS COCO demonstrates the superiority of CNNL-based models, especially when combined with LSTM, outperforming vanilla RNN-based models and achieving competitive performance compared to existing methods. These results underscore the advantage of leveraging Language CNN to model long-term dependencies in words, indicating promising prospects for advancing image captioning.

This paper [16] involves combining CNNs with RNNs, specifically LSTM networks, for the joint generation and comprehension of referring expressions in images. The CNN is utilized to represent the image, while the LSTM generates textual descriptions. The approach includes augmentation with region of interest information and location details, and the training process incorporates maximum likelihood training with dropout regularization to mitigate over-fitting. This method is tailored for generating sequences of words that precisely identify objects in images, enhancing the joint generation and comprehension of referring expressions within the image context.

An end-to-end trainable deep bidirectional LSTM model is presented for image captioning in [17]. The model integrates a CNN with two separate LSTM networks to facilitate the learning of long-term visual-language interactions. Comprising components for image encoding, text encoding, multi-modal LSTM, and deep bidirectional variant models, the proposed approach is evaluated on benchmark datasets, including Flickr8K, Flickr30K, and MSCOCO. The results demonstrate competitive performance, underscoring the effectiveness of the deep bidirectional LSTM model for image captioning. Overall, the approach effectively leverages both visual and textual information to generate descriptive sentences for images, showcasing its potential for applications requiring comprehensive visual and sequential understanding.

The author proposed a LSTM-A architecture in [18] integrates the robust image representation obtained from the 1,024-way pool5/7 × 7 layer of CNN with a Long Short-Term Memory (LSTM) network, a specialized type of recurrent neural network (RNN) adept at processing sequential data. This fusion allows for an end-to-end training process, enabling the model to simultaneously learn to extract pertinent features from the image representation and employ them in generating precise and descriptive captions for images. By leveraging the strengths of both CNN's detailed visual content representation and LSTM's sequence processing capabilities, the LSTM-A architecture enhances image captioning, providing a holistic approach that captures not only the visual content but also high-level attributes, thereby improving the overall captioning accuracy and richness of generated descriptions.

3) CNN with Attention

The showcased CNN-LSTM models, including m-RNN, LRCN, and others, excel in seamlessly combining CNNs for image feature extraction and LSTMs for sequential processing. However, early CNN-LSTM models faced challenges with fixed-size representations and struggled to handle detailed scenes. Attention mechanisms address this by dynamically focusing on specific regions of the image, allowing the model to generate captions with improved context and localization.

This paper [19] introduces an attention-based approach for generating textual descriptions of images. This model selectively focuses on parts of the input image during description generation, enhancing interpretability. SAT achieves state-of-the-art performance on benchmark datasets without ensemble methods. The attention mechanism aligns with human intuition and handles diverse image types. The approach offers interpretability, state-of-the-art performance, and fosters further development in image captioning.

A novel image captioning approach using a deep neural architecture with a CNN featuring an attention mechanism is introduced in [20]. The captioning decoder is based on a Conditional Recurrent Neural Network (C-RNN) with LSTM. Evaluation on benchmarks, including Flickr8K, Flickr30K, and MSCOCO, shows the proposed method outperforming other techniques in various metrics. It generates enriched captions with detailed salient object information and spatial relationships. The approach proves effective for producing accurate and detailed image captions.

This paper [21] added to the image captioning task using CNN and Attention mechanism a method for automatically generating attention maps from model predictions. Results demonstrate significant improvements over previous state-of-the-art models, particularly with the supervised attention mechanism. The paper includes a thorough analysis of attention maps' impact on caption quality.

The model introduced in this [22] incorporates a novel attention mechanism that models the interplay between the RNN state, image region descriptors, and word embedding vectors through three pairwise interactions. This allows for a comprehensive consideration of the direct interaction among caption words, image regions, and RNN state, leading to improved captioning performance. Unlike previous attention-based approaches, the "Areas of Attention" model enables a direct association between caption words and image regions. This association is inferred from image-level captions during training, like weakly-supervised object detector training, and aids in localizing corresponding regions during testing. The model integrates a localization

subnetwork, like spatial transformer networks, to regress a set of attention areas from the image content. This allows for the generation of image-specific attention areas, which can be trained jointly with the rest of the network, contributing to improved captioning performance. The model proposes and compares different methods for generating attention areas, including CNN activation grids, object proposals, and spatial transformer nets applied in a convolutional fashion. Through this comparison, the model identifies the most effective approach for generating attention areas, leading to state-of-the-art performance on the MSCOCO dataset.

The methodology introduced in this paper [23] involves a novel semantic attention model within the framework of recurrent neural networks, effectively integrating visual information through both top-down and bottom-up approaches. The algorithm is designed to learn the selective attention to semantic concept proposals and adeptly fuse them into the hidden states and outputs of recurrent neural networks. This process establishes a feedback mechanism, facilitating the connection between top-down and bottom-up computations. Additionally, the paper explores two cutting-edge deep learning models for attribute prediction. The first approach employs a ranking loss as an objective function, training a multi-label classifier. The second approach utilizes a Fully Convolutional Network (FCN) to extract attributes from local patches. Both methods generate a relevance score indicating the correlation between an image and a visual attribute. This relevance score is then employed to select top-ranked attributes for input into the captioning model. To assess the proposed algorithm's performance, comprehensive evaluations are conducted on two widely recognized benchmarks: Microsoft COCO and Flickr30K. The results demonstrate a significant and consistent outperformance of the proposed algorithm compared to state-of-the-art approaches across diverse evaluation metrics.

III. METHODOLOGY

Figure 1 illustrates the proposed framework for Image Captioning, which consists mainly of seven significant steps:

1) Data Collection
2) Data Preprocessing
3) Images Features Extraction
4) Data Splitting
5) Captions Features Extraction
6) Image Captioning Deep Learning Algorithms
7) Prediction and Evaluation Metrics

A. DATA COLLECTION

In our work, we have utilized one dataset, Flickr8k, with two different versions, one for the English sentences, and another one for the Arabic sentences. Researchers and developers interested in computer vision and image processing will find the Flickr 8k dataset to be an invaluable resource. The dataset, which consists of eight thousand high-resolution photos with three descriptive captions for each image, is a benchmark for multimodal learning tasks and image captioning. The Flickr 8k dataset's richness and diversity of images make it a perfect option for training and testing algorithms meant to comprehend and produce descriptions of visual information that are human-like. This dataset can be utilized by researchers to improve the accuracy of image captioning models, leading to breakthroughs in domains including natural language processing, image recognition, and the convergence of vision and language. Furthermore, the availability of high-quality annotations encourages creativity and the advancement of increasingly complex. Furthermore, the presence of well-annotated images encourages creativity and the advancement of increasingly complex and contextually aware AI systems. For these reasons, the Flickr 8k dataset is an invaluable resource for researchers seeking to enhance machine learning models' capacity to comprehend and produce textual content from visual input. . **See Figure 2**

B. DATA PREPROCESSING

Data preprocessing is an indispensible step in preparing textual data for input into image captioning models. Usually, input texts in datasets, in our case captions, contain noise, stopping words, punctuation, and synonyms that can affect the model training and its performance. Data preprocessing aims to standardize textual input converting them into a consistent form that models can handle. This contributes in decreasing the vocab size that the model has to learn and simplifies the learning process of the model, hence achieving better accuracies. In our case, each of the two datasets used needed some preprocessing depending on the captions in that dataset. Therefore, in this section, we will discuss the preprocessing techniques applied to each of the two datasets. **See Figure 3**

1) English Captions Preprocessing

For the first dataset, the english captions contained inconsistencies such as unnecessary spaces, punctuation, stopping words, and combinations of upper and lower case characters. That was addressed through the following preprocessing steps which is shown in **Table 1**:

- **Step 1 : Lower-Case Conversion**: In this step, all characters are converted to lower-case to universalize all the characters preventing the model handling two identical words as different just because a single character in the first word is upper-case while all the characters of the second word are lower-case.
  e.g. "Girl" is converted to "girl".
- **Step 2 : Special Characters Removal**: In this step, all the punctuation was removed including any character other than the alphabet and the spaces, as some punctuation were detected such as periods at the end of the line.
  e.g. "." is removed.
- **Step 3 : Noise Removal**: In this step, all the extra unnecessary spaces are removed along with words of

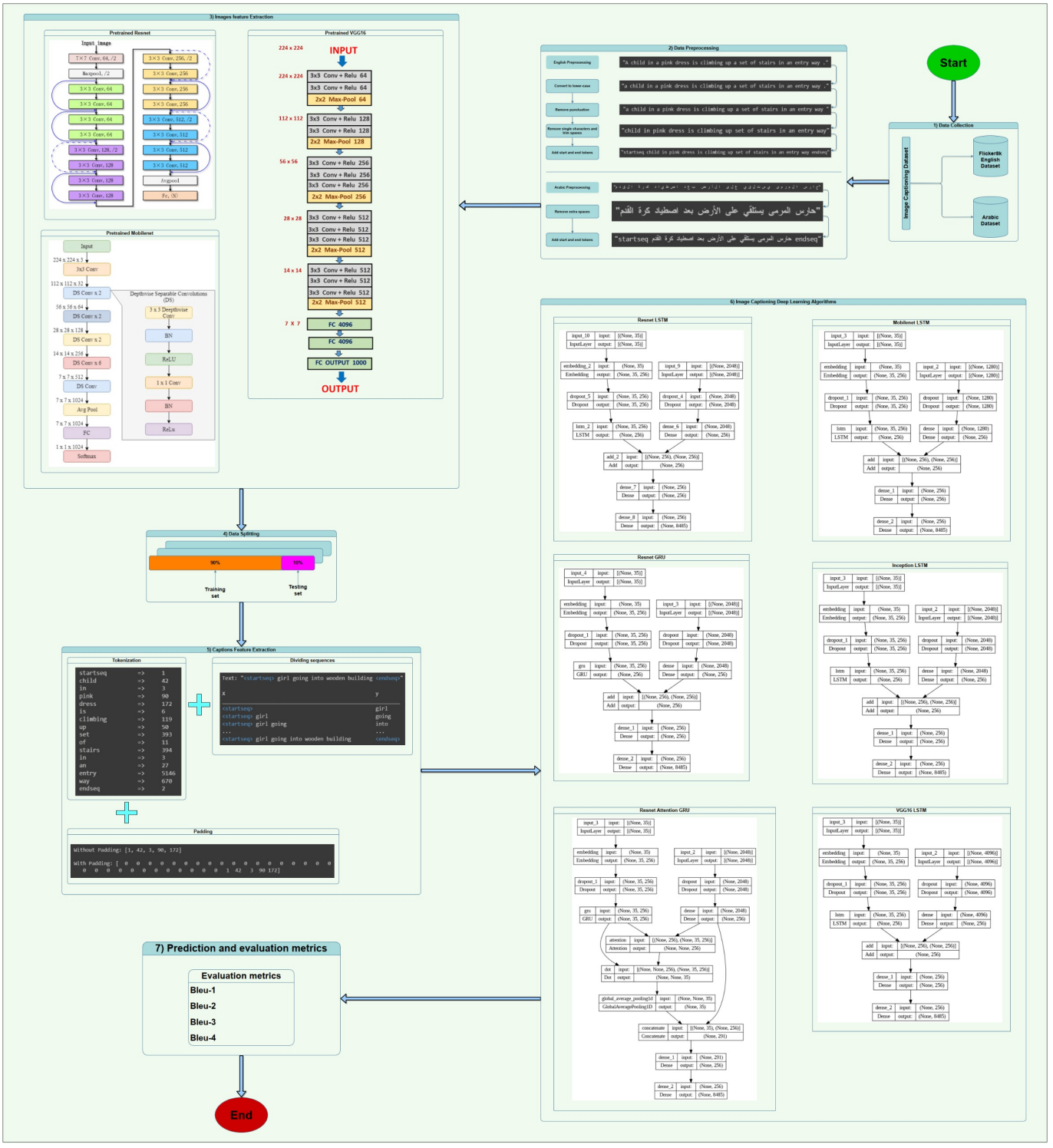


Figure 1: Image Captioning Methodology Flowchart

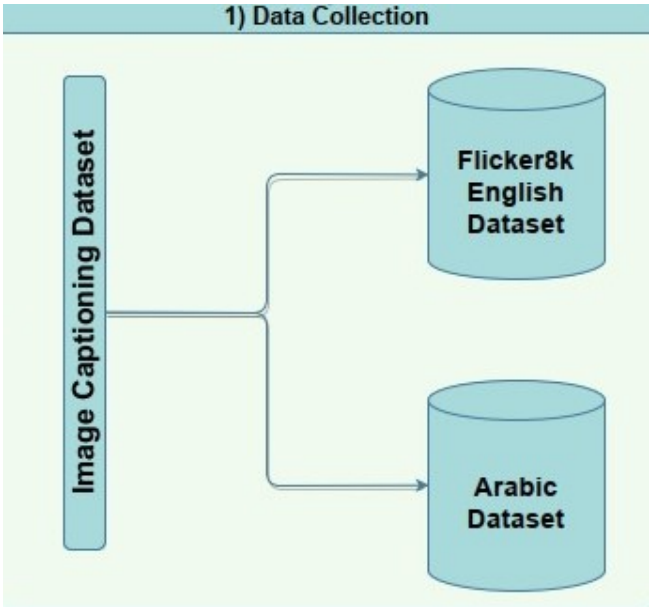


Figure 2: Data Collection

length lower than 2 characters.
e.g. "a" is removed.

- **Step 4 : Adding Start and End Tokens**: In this step, a start and end tokens are added to the beginning and the end of the sentence respectively, so that the model knows which word it begins its sentence with and which word it should stop when it is reached.
e.g. "girl sitting on a swing" is converted to "startseq girl sitting on a swing endseq"

2) Arabic Captions Preprocessing

For the second dataset, the arabic captions was very clean relative to the english captions, however it extra spaces were unexpectedly added when the captions were read from the file which needed to be handled. THe following shows the preprocessing steps undertaken to adjust the arabic captions which can be demonstrated in **Table 2**:

- **Step 1 : Extra Spaces Removal**: When python reads the arabic captions from the dataset, many unnecessary extra spaces are introduced between the arabic characters, so in this step, these extra spaces are removed to return the caption into its original form.
- **Step 2 : Adding Start and End Tokens**: In this step, a start and end tokens are added to the beginning and the end of the sentence respectively, so that the model knows which word it begins its sentence with and which word it should stop when it is reached.

C. IMAGES FEATURES EXTRACTION

In this stage, the aim is to extract features from the images in the two datasets and convert them into numerical values that can be entered as input to the image captioning model to be used along with its captions to train the model. The approach taken in this phase was to get a pretrained CNN model, provide it with the images of one of the datasets, and extract the significant features from each image. That was done by removing the last layer of the model that is responsible for the classification of the image and acquire the features from the layer before it. The features of each image would be encoded as a group of numbers that the image captioning model can handle. The proposed pretrained models for this phase were VGG16, ResNet50, and MobileNetV2. **See Figure 4**

1) VGG16

VGG16 is one of the most famous CNN models in deep learning. Although it has achieved great accuracy on ImageNet dataset, it still is one of the most highly computational CNN models. Its architecture is basically a repeating block of two to three layers 3x3 convolutional layers followed by a 2x2 max pooling layer, and finally three fully-connected layers with the final one being the output layer. The presence of the three fully-connected layers applied to the image is one of the causes of why VGG16 takes a lot of memory and time. The size of the features vector acquired from the pretrained VGG16 provided by keras is 4096 which is relatively large compared to other models.

2) ResNet50

ResNet is known to be one of the best CNN models that were ever invented, where it gave an unparalleled accuracy on the ImageNet dataset. Furthermore, it is considered a very efficient deep learning model in image classification. The reason why ResNet is highly accurate is that is that it introduced a block called the residual block, which gives the model the capability to learn the identity function, so that it can work at least as well as shallower architectures depending to avoid deep network underfitting. The size of the features vector acquired from the pretrained ResNet provided by keras is 2048 which is better than VGG16 in addition to its accuracy privilege, which makes it almost the most suitable CNN to be used for extracting images features.

3) MobileNetV2

MobileNet is characterized by being a light weight and very efficient model and it gave relatively good accuracy on ImageNet dataset. MobileNet basically contains three types of convolution: 3x3 conv, 1x1 conv, and 3x3 depthwise conv. The size of the features vector acquired from the pretrained MobileNet provided by keras is 1280 which is very low compared to the above models, and this makes the training of the image captioning model much faster than with the features of the other models.

D. DATA SPLITTING

In this stage, data is split into 90% training and 10% testing. The choice of the split was based on the fact that deep learning models require much data to reach reasonable accuracy and to overcome overfitting, so the chosen split

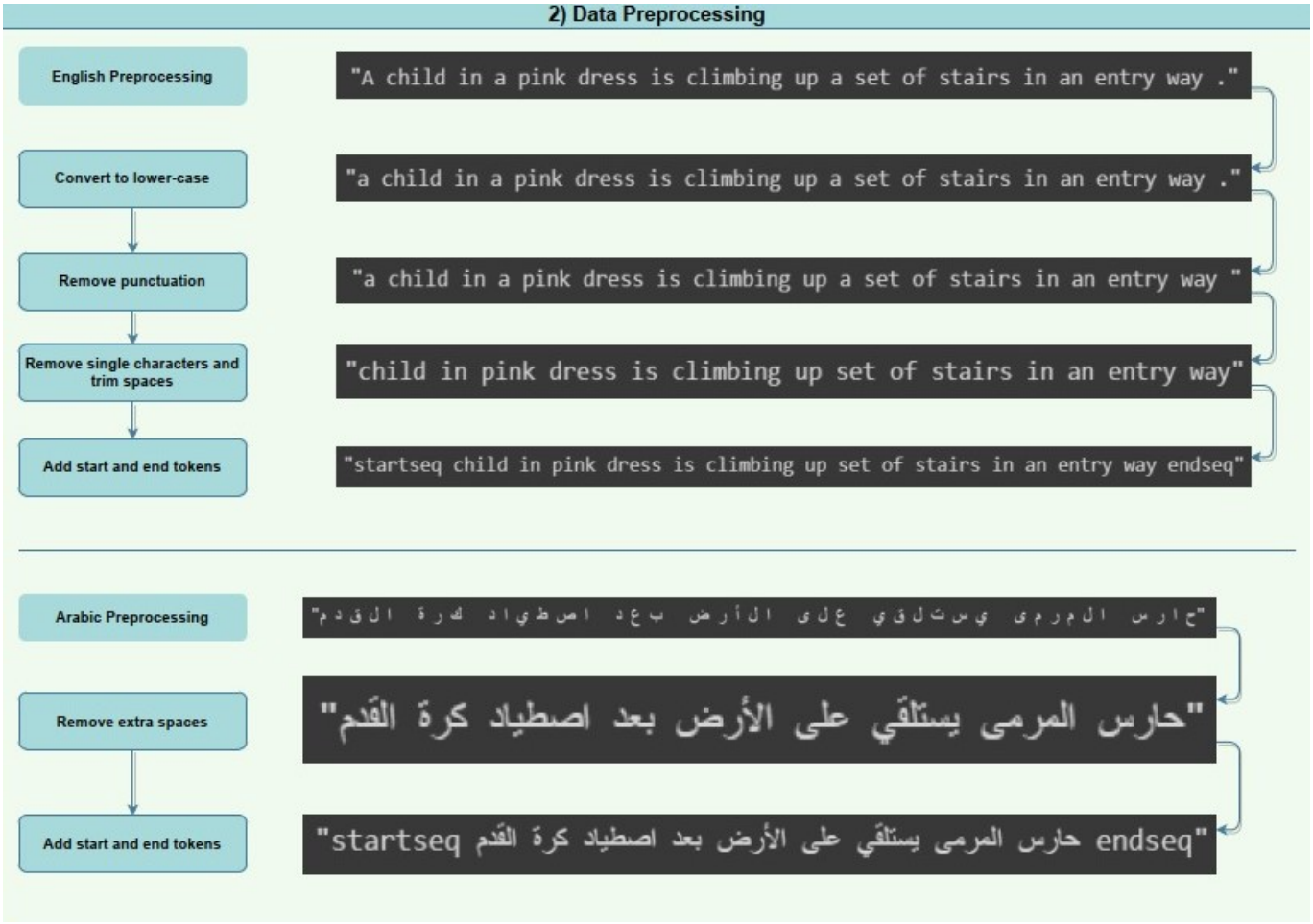


Figure 3: Data Preprocessing

Table 1: English Text Preprocessing

Original Caption	A child in a pink dress is climbing up a set of stairs in an entry way .
After Lower-Case Conversion	a child in a pink dress is climbing up a set of stairs in an entry way .
After Special Characters Removal	a child in a pink dress is climbing up a set of stairs in an entry way
After Noise Removal	child in pink dress is climbing up set of stairs in an entry way
Final Caption	startseq child in pink dress is climbing up set of stairs in an entry way endseq

Table 2: Arabic Text Preprocessing

Original Caption	ح ا ر س ا ل م ر م ى ي س ت ل ق ي ع ل ى ا ل أ ر ض ب ع د ا ص ط ي ا د ك ر ة ا ل ق د م
After Spaces Removal	حارس المرمى يستلقي على الأرض بعد اصطياد كرة القدم
Final Caption	startseq حارس المرمى يستلقي على الأرض بعد اصطياد كرة القدم endseq

was better than an 80%-20% split which would have made the model achieve poor accuracy. A small portion of the training set was used as a validation to detect overfitting and perform an early stopping, hence tune the number of epochs hyperparameter. **See Figure 5**

E. CAPTIONS FEATURES EXTRACTION

As previously explained, the features of images have been extracted through pretrained CNN models, thus similarly, features must be extracted from the captions of the images for the model to easily handle them, since both the images features and the captions features are to be inserted to the image captioning model. The captions features extraction can be summarized as tokenization, sequences dividing, and padding. **See Figure 6**

1) Tokenization

Tokenization is to process of converting words into numerical values to be understood by the model. Tokenizing captions is done by iterating over all the words vocab of all the captions and giving each of them a unique index, then map each index to its corresponding word. After that, the

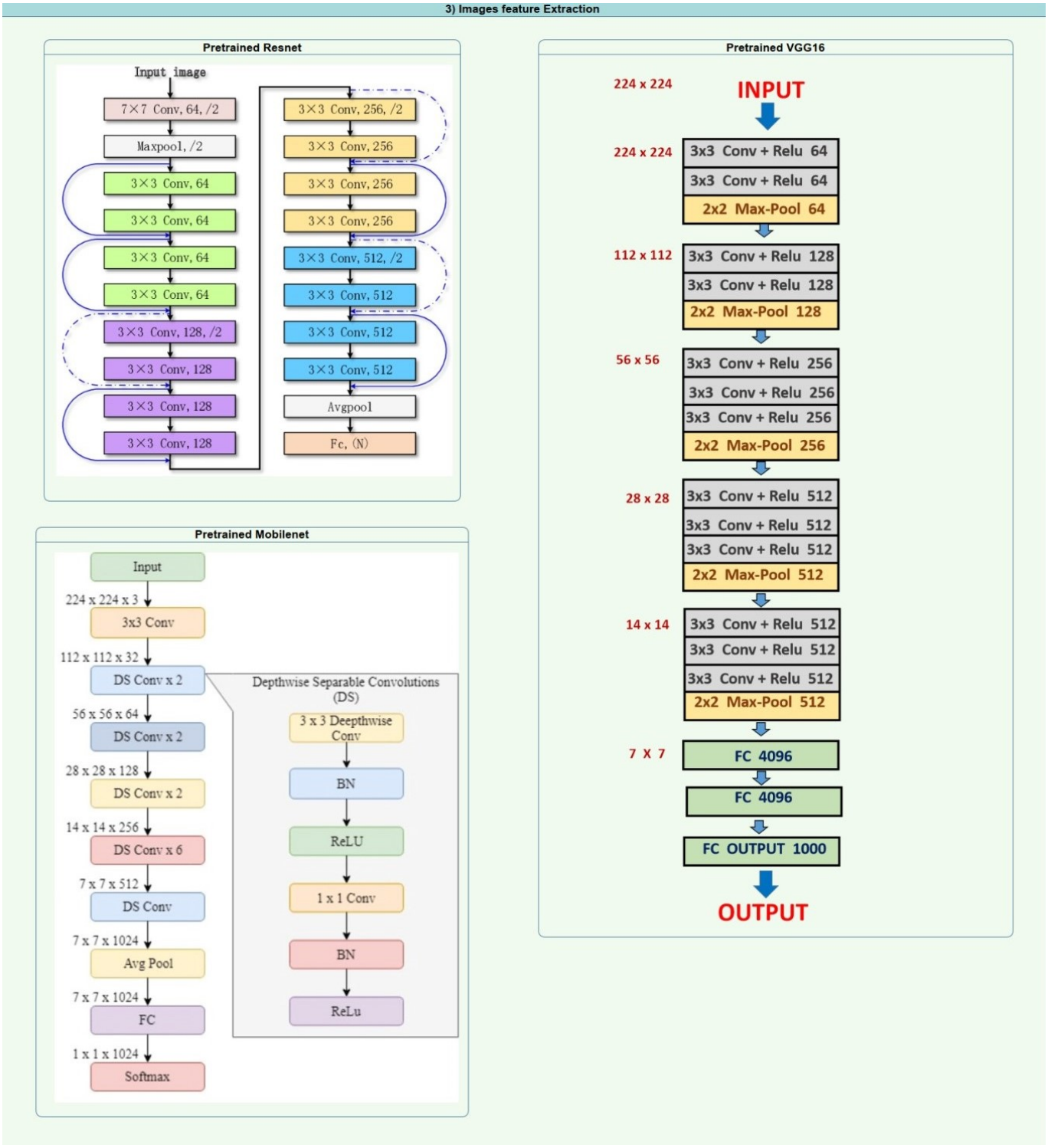


Figure 4: Images Feature Extraction

Figure 5: Data Splitting

words of each caption are replaced with their indices and these indices are used for predicting the text. **See Figure 7**

2) Sequences Dividing

Following the tokenization step, the tokenized output is divided into input and output sequences. By doing that, we are capable to train the model by giving it an input sequence of the divided input sequences and it is expected to predict the following word to this sequence, which is the output sequence. In that manner, the model would just receive the start token and then it learns to predict the first word, then the second, then the third, and so on. **See Figure 8**

3) Padding

The final step to the captions feature extraction is the padding. Captions could have variable sizes, hence input sequences lengths can vary, so a fixed input length should be determined as the model expects a constant input size. This is solved by choosing the greatest sequence length as the input length and pad the other smaller sequences with zeros to be of the same size as the maximum length. **See Figure 9**

F. IMAGE CAPTIONING DEEP LEARNING ALGORITHMS

Figures 10 and 11 illustrate the proposed deep learning models for image captioning to be explained in the following points.

1) Resnet LSTM
   Convolutional neural networks (CNNs), notably ResNet, are used as encoders in the first suggested DL model architecture, whereas Long Short-Term Memory (LSTM) networks are used as decoders.
   - Encoder: The encoder for extracting high-level features from input images is the ResNet50 model. A deep convolutional neural network called ResNet50, which was trained on ImageNet beforehand, can recognize intricate hierarchical patterns in pictures. The ResNet50's final layer, which is usually utilized for classifying images, is eliminated. A rich representation of the input image is now provided by the model output, which now includes the features that were extracted from the penultimate layer.
   - Encoder Processing: A number of processing stages are applied to the extracted image features. In order to introduce regularization and to avoid overfitting during training, dropout is used. The dimensionality is subsequently reduced to 256 by passing the processed features through a fully connected (Dense) layer with ReLU activation. The goal of this condensed depiction is to preserve crucial details from the picture.
   - LSTM: Additionally, a series of word indices that correspond to the captions for the photos are fed into the model. To transform the discrete word indices into continuous vector representations, an embedding layer is utilized. To regularize the embedded sequences, dropout is performed. An LSTM layer with 256 units receives the sequence data that has been processed. Because of its capacity to extract long-term context and sequential dependencies from data, LSTM is the model of choice.
   - Decoder: The encoder's and the LSTM layer's outputs are merged. This fusion, which combines the data from the sequence and the image, is frequently carried out element-by-element (for example, element-wise addition). In order to further capture linkages and patterns in the fused representation, the merged features are subsequently processed through a Dense layer with ReLU activation. Finally, a probability distribution across the vocabulary is produced using a Dense layer with softmax activation. This predicts the probability that every word in the vocabulary will appear as the subsequent word in the created caption.
   - Optimizer: The categorical cross-entropy loss, which gauges how different the real word distribution in the ground truth captions is from the predicted word distribution, is used to train the model. During training, the weights of the model are modified using the Adam optimizer.

2) MobileNet LSTM
   The second proposed DL model combines LSTM and MobileNet components in an encoder-decoder architecture for image captioning.
   - Image Encoder: To extract high-level characteristics from input photos, the image encoder uses a MobileNetV2 model that has already been trained. This trained model is good at identifying and comprehending patterns in images.
   - Encoding Image Features: A feature vector that represents the image's content is created using the MobileNetV2 output. Capturing pertinent visual information, this simplified representation forms the basis for caption generation.
   - Sequence Encoder: The sequential data from captions is processed simultaneously. To capture the semantic meaning of each word in the caption, it is integrated into a dense vector.
   - LSTM for Sequential Context: The sequential infor-

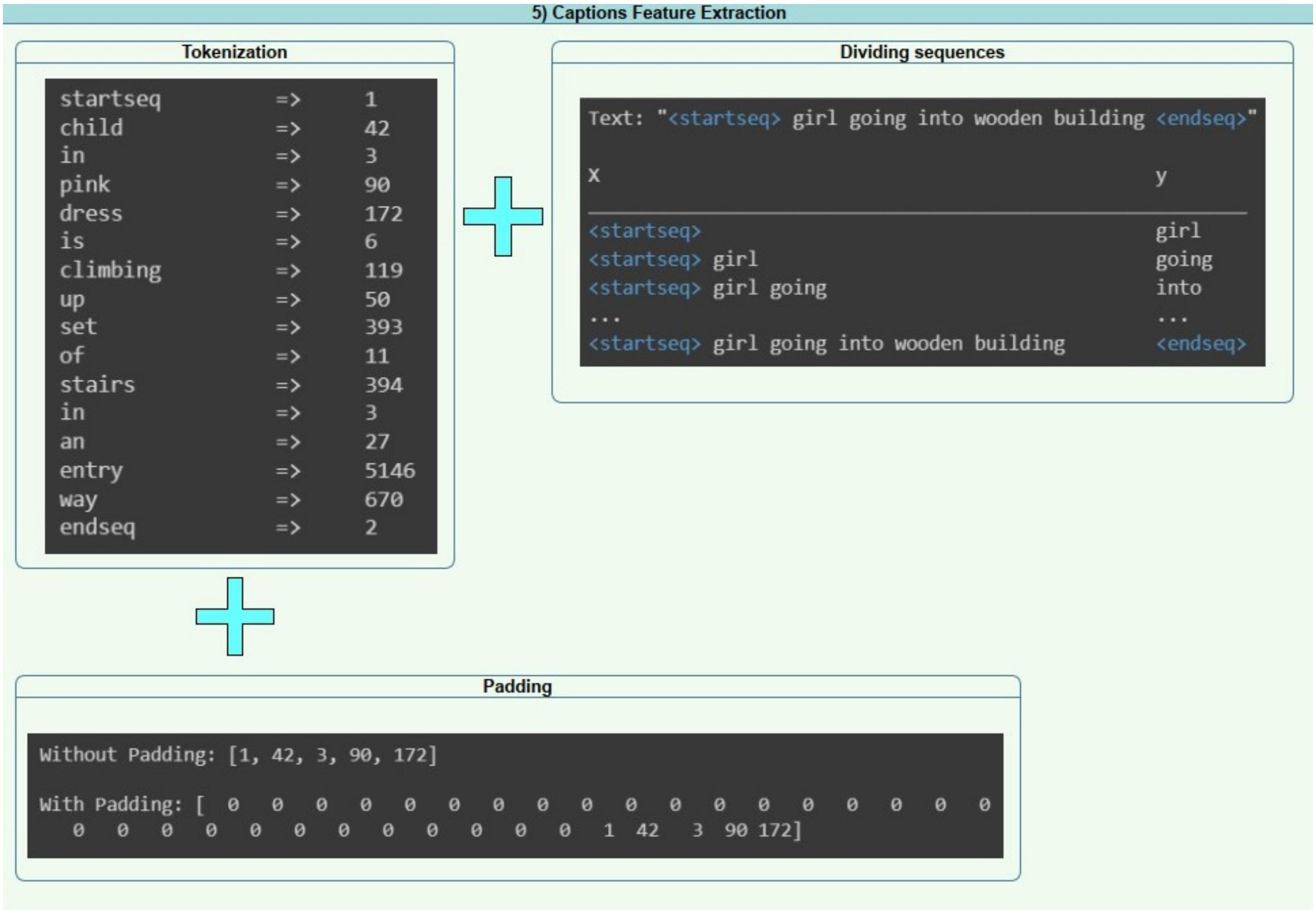


Figure 6: Captions Feature Extraction

mation that is embedded is processed using Long Short-Term Memory (LSTM). Because LSTMs are good at comprehending context in sequences, they work well for creating captions where word order is important.

- Combining Image and Sequence Information: Contextual data from the LSTM is mixed with the features of the retrieved image. With this fusion, the visual and semantic aspects will be combined to create a comprehensive representation that can be used to create captions that make sense.
- Decoding: Dense layers are used to further process the combined features. In order to help the model provide captions that are contextually appropriate for the input image, this stage improves the joint representation.
- Output Layer: A probability distribution over the vocabulary is produced by the dense layer with softmax activation that makes up the final layer. The next word in the caption is predicted using this distribution.
- Optimization: The model is optimized to minimize the loss due to categorical crossentropy during training. In order to motivate the model to produce correct and contextually relevant captions, this loss function calculates the difference between the predicted and actual word distributions.

3) Resnet GRU Attention

The proposed image captioning model is a complex architecture that combines the capabilities of an attention-based sequence decoder with an image encoder based on ResNet50. Convolutional and recurrent neural networks are combined in this technique to produce captions for input photos that are logical and pertinent to the context. The sequence decoder uses a GRU (Gated Recurrent Unit) to interpret the sequential information within captions and integrates attention mechanisms to focus on salient parts of the sequence, while the picture encoder uses the ResNet50 model to extract rich visual elements. These elements work together to provide the model the ability to comprehend both semantic and visual elements, resulting in more complex and evocative captions.

- Image Encoder: Uses the pre-trained ResNet50 model to extract features at a high level from input photos. The activations of the second-to-last layer

```
startseq    =>    1
child       =>    42
in          =>    3
pink        =>    90
dress       =>    172
is          =>    6
climbing    =>    119
up          =>    50
set         =>    393
of          =>    11
stairs      =>    394
in          =>    3
an          =>    27
entry       =>    5146
way         =>    670
endseq      =>    2
```

Figure 7: Caption Tokenization

```
Text: "<startseq> girl going into wooden building <endseq>"

X                                          y
-------------------------------------------------
<startseq>                                 girl
<startseq> girl                            going
<startseq> girl going                      into
...                                        ...
<startseq> girl going into wooden building <endseq>
```

Figure 8: Sequences Dividing

function as a feature vector that represents the content of the image when the final layer of ResNet50 is eliminated.

- Encoder: The retrieved features are represented by the input layer for the image features, which has the shape (2048,). Applying a dropout layer helps with regularization. Image features are converted into a condensed representation using a dense layer with 256 units and ReLU activation.
- Sequence Encoder: Layer of input for the data that is sequential (captions). Vocabulary encoded with integers is transformed into dense vectors of size 256 by the embedding layer. Values for padding are hidden. Applying a dropout layer helps with regularization. Sequential dependencies are captured in the captions by the GRU (Gated Recurrent Unit) layer, which has 256 units and return sequences.
- Attention Mechanism: The attention mechanism improves the model's concentration on pertinent sequence segments related to the image features. The relationship between sequential features and picture features is used to calculate attention weights. A dot product of attention weights and sequential characteristics is used to calculate the context vector. An application of global average pooling yields a context vector with a defined size.
- Decoder: A joint representation is produced by concatenating the context vector and picture features. The combined representation is further processed by a dense layer with 256 units and ReLU activation. In order to anticipate the next word in the caption, the final dense layer with softmax activation generates a probability distribution across the vocabulary.
- Optimization: The model's parameters are iteratively optimized using the Adam optimizer, which promotes effective convergence during training.

4) Inception LSTM

Utilizing the InceptionV3 architecture as its image encoder, the fourth suggested DL model for image captioning focuses on obtaining significant visual attributes from input images. The base for producing descriptive and contextually relevant captions is a sophisticated convolutional neural network (CNN) coupled with a recurrent neural network (RNN) for sequence processing.

- Image Encoder: The image encoder used is the InceptionV3 model, which has been pretrained on extensive picture classification tasks. It successfully extracts complex visual characteristics from input photos.
- Encoding Image Features: The image features with the shape of (2048,), which reflect the high-level information retrieved by InceptionV3, are sent to an input layer. Dropout is used to lessen the effects of overfitting. The visual characteristics are further refined by a Dense layer with 256 units and ReLU activation, which compresses them into a lower-dimensional representation.
- Sequence Encoder: Data is processed sequentially by an input layer. The vocabulary that has been integer-encoded is transformed into 256-size dense vectors via an embedding layer. To deal with sequences of varying lengths, padding values are hidden. For regularization, dropout is applied. The 256-unit LSTM (Long Short-Term Memory) layer examines the sequential data, identifying patterns and dependencies in the captions.
- Decoder: An element-wise addition is used to integrate the sequence information and encoded image characteristics. The combined data is processed by a Dense layer with ReLU activation, which records the joint representation of sequential and visual

```
Without Padding: [1, 42, 3, 90, 172]

With Padding: [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   1  42   3  90 172]
```

Figure 9: Sequence Padding

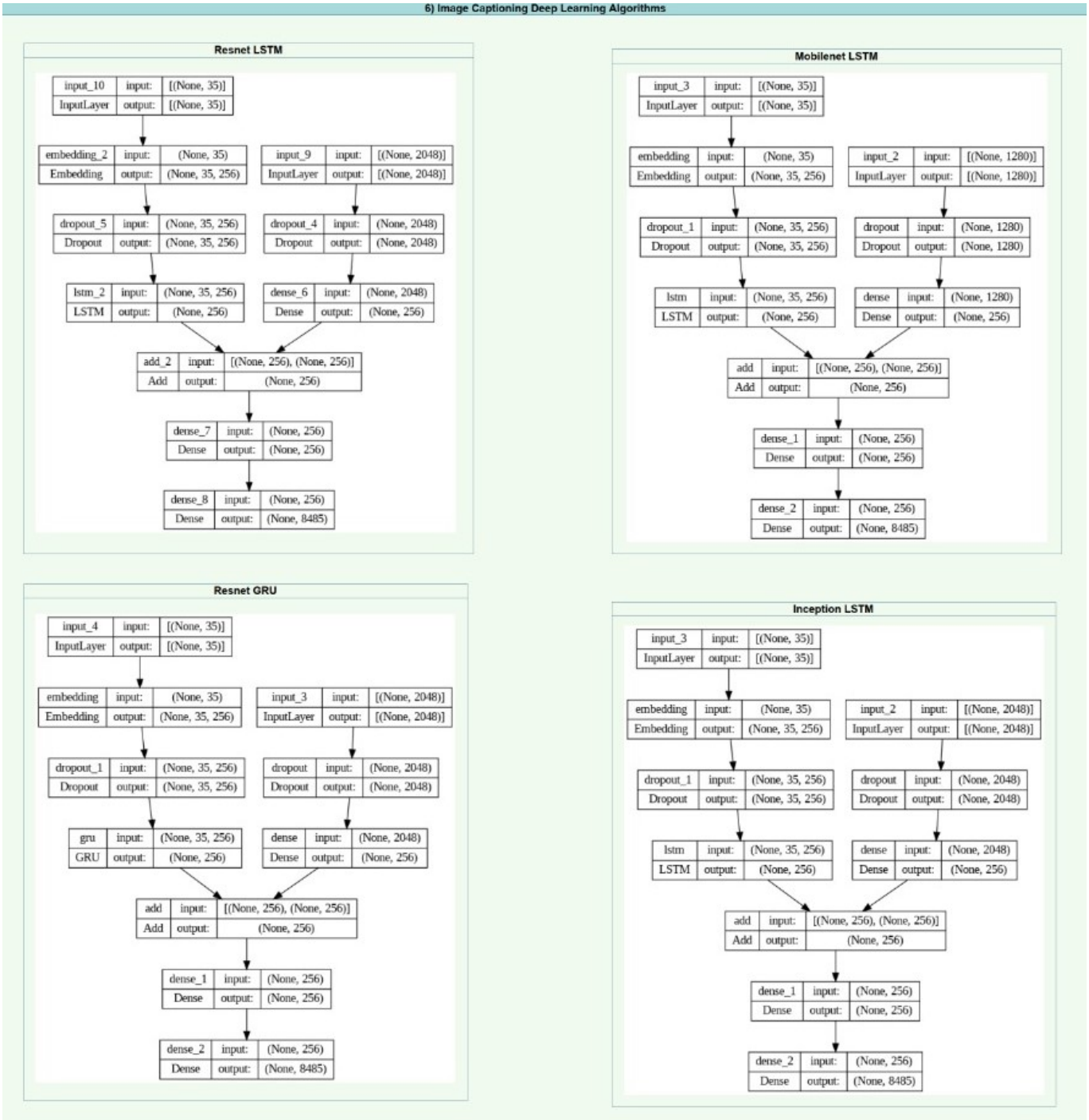


Figure 10: Image Captioning Deep Learning Algorithms Part 1

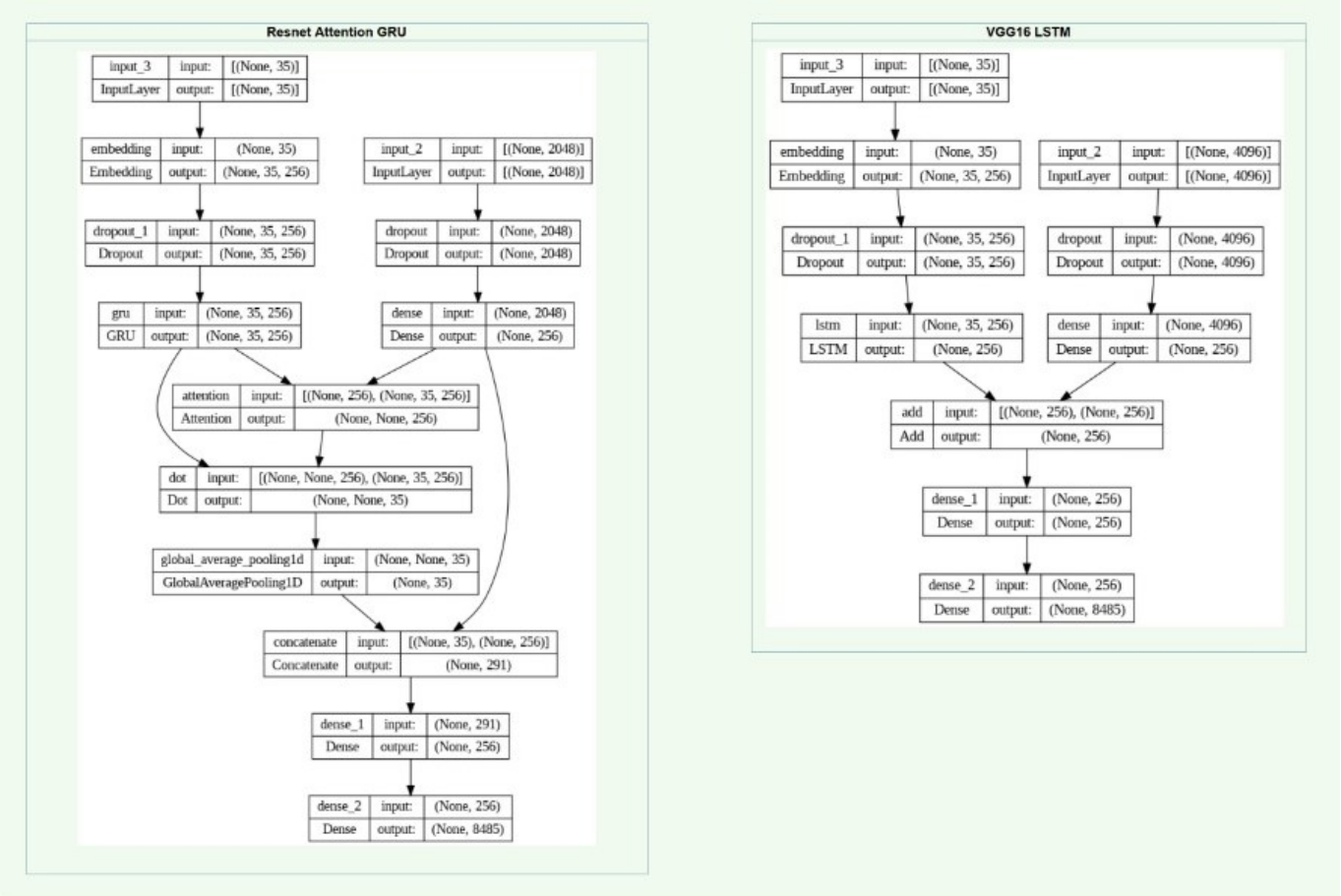


Figure 11: Image Captioning Deep Learning Algorithms Part 2

elements. Predicting the next word in the caption, the last Dense layer (outputs) with softmax activation creates a probability distribution over the vocabulary.

- Optimization: The model is optimized using the Adam optimizer. The Adam optimizer efficiently adjusts the model's parameters during training, combining elements of momentum and RMSprop to adaptively update weights. This dynamic optimization technique helps the model converge more effectively and accelerates the learning process.

5) VGG16 LSTM

The model for image captioning leverages the VGG16 architecture as the image encoder, designed to extract detailed visual features from input images. The composition includes an image encoder using VGG16, a sequence decoder comprising Dropout, Dense, Embedding, LSTM, and Dense layers, and an overall structure facilitating caption generation.

- Image Encoder: VGG16 serves as the image encoder, capturing detailed visual features from input images.
- Encoding Image Features: The encoding of image features is orchestrated through a series of operations. The image encoder, leveraging the VGG16 model, generates a feature vector of 4096 dimensions. This feature vector is fed into the model through an input layer . To prevent overfitting and enhance generalization, a Dropout layer with a rate of 0.4 is applied. The subsequent layer is a Dense layer with 256 units and a Rectified Linear Unit (ReLU) activation function, transforming the image features into a condensed representation suitable for joint processing with the sequential information.
- Sequence Encoder: The sequence encoding component of the model involves two key steps. Firstly, the input layer is responsible for processing the sequential data, representing the captions. Subsequently, an Embedding layer converts the vocabulary into dense vectors of 256 dimensions, with the inclusion of masking to handle variable-length sequences. A Dropout layer with a rate of 0.4 is employed for regularization. Finally, a Long Short-Term Memory (LSTM) layer with 256 units is incorporated to process and capture sequential dependencies within the captions.
- Decoder: The decoder component amalgamates the encoded image features and the sequence information, setting the stage for generating captions. The

element-wise addition of these encoded representations takes place in the decoder1 layer. Subsequently, a Dense layer with a Rectified Linear Unit (ReLU) activation function processes this combined information, refining the joint representation. Finally, the model's output layer utilizes a Dense layer with softmax activation to produce a probability distribution over the vocabulary, predicting the next word in the caption. The entire model is then compiled using the categorical crossentropy loss function and the Adam optimizer for efficient training.

- Optimization: The model is optimized using the Adam optimizer.

6) ResNet GRU

The model presented here is designed for image captioning and employs the ResNet50 architecture as the image encoder, known for its effectiveness in extracting detailed visual features from input images. This composition includes an image encoder utilizing ResNet50, a sequence decoder featuring Dropout, Dense, Embedding, GRU (Gated Recurrent Unit), and Dense layers. Together, these components create a holistic structure for generating contextually relevant captions.

- Image Encoder: The image encoder, powered by ResNet50, is responsible for extracting high-level visual features from input images. The last layer of the ResNet50 model is removed, and the second-to-last layer's activations serve as a feature vector representing the content of the image. This feature vector undergoes further processing through an input layer, a Dropout layer with a rate of 0.4 for regularization, and a Dense layer with 256 units and ReLU activation, refining the image features into a condensed representation.
- Encoding Image Features: The encoding of image features in this model is achieved by utilizing the ResNet50 architecture as the image encoder. The model, pre-trained on image classification tasks, extracts high-level visual features, with the second-to-last layer's activations forming a feature vector representing the input image. This vector is processed through an input layer, followed by a Dropout layer (dropout rate of 0.4) for regularization, and a Dense layer with 256 units and ReLU activation. These layers collectively refine and condense the image features, creating a more expressive representation for subsequent joint processing with sequential information during the generation of descriptive captions.
- Sequence Encoder: The sequence encoding aspect involves the processing of sequential data, i.e., captions. The input layer receives the sequential information, followed by an Embedding layer that converts vocabulary into dense vectors of 256 dimensions, incorporating masking to handle variable-length sequences. For regularization, a Dropout layer with a rate of 0.4 is applied . The sequential information is then processed through a GRU layer with 256 units, capturing dependencies within the captions.
- Decoder: The decoder component combines the encoded image features and the processed sequence information to generate captions. The decoder layer performs an element-wise addition of these encoded representations. Subsequently, a Dense layer with ReLU activation processes the combined information, refining the joint representation. The final output layer employs a Dense layer with softmax activation, generating a probability distribution over the vocabulary to predict the next word in the caption.
- Optimization: The model is optimized using the Adam optimizer

G. PREDICTION AND EVALUATION METRICS

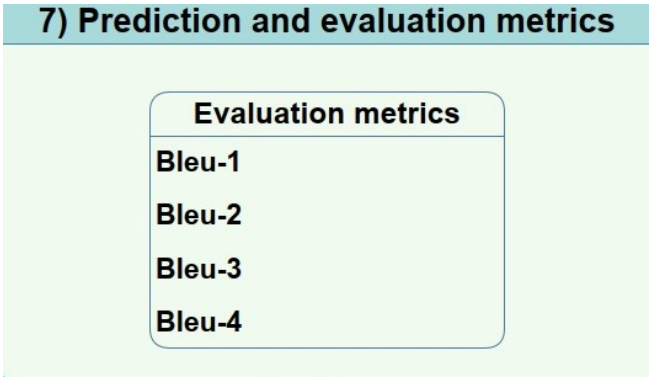


Figure 12: Prediction and Evaluation Metrics

After choosing one of the proposed deep learning models and training it on the given images and captions in the dataset, the next step would be to make predictions and evaluate the accuracy of the model. The proposed evaluation metric for the model is the Bilingual Evaluation Understudy (BLEU) score. The BLEU score is an evaluation metric ranging from 0 to 1 where the closer its value is to 1 the better accuracy it indicates. BLEU score is not like other classification performance metrics, as a BLEU score value of above 0.3 is considered a good accuracy. This evaluation metric is considered a better accuracy indication than standard precision metric, where BLEU score takes into consideration the length of the predicted text, hence gives more score to text that contains all the reference text information than that with missing information even if its output text is all correct. **Figure 12** shows the used BLEU N-Grams as an evaluation metric .**Equation 1** below represents how BLEU score is calculated.

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

- **Brevity Penalty (BP)** is a term used in BLEU score that is responsible for considering the length of the

predicted text compared to the length of the reference text. Accordingly, it puts in consideration how much information was acquired in the predicted text with respect to the reference to make sure to give a lower score to a predicted text that has less information than a predicted text that has more information even if both have accurate and correct information. This is done by introducing an exponential decay factor that is in terms of the length of the reference text and the predicted text.

e.g. Reference: "Transformers make everything quick and efficient through parallel computation of self-attention heads"
Candidate: "Transformers make everything quick and efficient"
In the above example, the candidate would receive a 100% accuracy in case of precision metric while it would receive less than that in case of BLEU score due to an exponential decay factor of BP. **See Equation 2**

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}. \quad (2)$$

- **Modified Precision (Pn)** is the main term used to measure the accuracy of the text in the BLEU score. It is similar to the precision metric except it is a stronger indicator of the accuracy of the output text. Normal precision is calculated as the number of matching words between predicted and references texts divided by the number of words of the predicted text, which ensures that the model extracted words relevant to the ground truth. However, the normal precision is not sufficient and not accurate enough, as if the predicted text contains only a single repeated word that exists in the reference text, the metric would give a 100% accuracy, which is not accurate.
  e.g. Reference: "Transformers make everything quick and efficient"
  Candidate: "Transformers Transformers Transformers Transformers"
  In the above example, the precision would be 4/4=1 in case of 1-gram which is a 100% accuracy, although the predicted text did not provide enough information.
  Here is were the role of the modified precision comes. Modified precision takes into consideration the number of times a word has occurred in the reference text, then take the true positive as the minimum of the times that this word appears in the predicted text and the times that it appeared in the reference text, which would give a more realistic accuracy. In the example above, the word "Transformers" would be counted only once because it occurred in the reference text one time and in the predicted text four times, so min(4,1)=1 which is divided by the length of the predicted text to give the accuracy 1/4=0.25 which is 25%. **See Equations 3 and 4**

$$p_n = \frac{\sum_{C \in \{\text{ Candidates }\}} \sum_{n-gram \in C} Count_{clip}(\text{n-gram})}{\sum_{C' \in \{\text{ Candidates }\}} \sum_{n-gram \in C'} Count(\text{n-gram'})} \quad (3)$$

$$Count_c lip = \min(Count, Max\_Ref\_Count) \quad (4)$$

- **N-gram Order (N)** represents the number of consecutive words that should be considered together when calculating the BLEU score. N-gram is dividing the sentence into L groups, where L is the length of the sentence, such that each group i consists of the ith word in the sentence followed by its following N-1 words. The higher the value of N, the more words in each group is taken into consideration when calculating the accuracy, which raises the restriction on the predicted text giving less accuracy than the lower values of N. In our model, we have used 1-gram, 2-gram, 3-gram, and 4-gram BLEU score, which are normally simply called 1-BLEU, 2-BLEU, 3-BLEU, and 4-BLEU respectively. The weights Wn represent a factor multiplied by each word accuracy and it is usually represented in terms of the N-gram Order
  e.g. For 1-BLEU, Wn is represented as [1, 0, 0, 0]
  For 2-BLEU, Wn is represented as [0.5, 0.5, 0, 0]
  For 3-BLEU, Wn is represented as [0.334, 0.333, 0.333, 0]
  For 4-BLEU, Wn is represented as [0.25, 0.25, 0.25, 0.25]
  See Table 3

IV. EXPERIMENTAL RESULTS AND DISCUSSION

To see how well CNN-RNN models work for captioning images, we tested them on two datasets: Flickr8k English and Flickr8k Arabic. We tried different models, like Inception LSTM, Mobilenet LSTM, Resnet LSTM, VGG16 LSTM, Resnet GRU, and Resnet GRU with attention, for the English set. For the Arabic set, we used Resnet GRU with attention, Mobilenet LSTM, Resnet LSTM, and Resnet GRU. We used BLEU-1, BLEU-2, BLEU-3, and BLEU-4 to measure how accurate the captions were. This helps us understand how these models perform with different languages and types of images.

A. CASE STUDY I (FLICKR8K ENGLISH DATASET)

This part is dedicated to the performance results of different CNN-RNN models Flickr8k English dataset.
By inspecting the results of Table. 4, we found that each model has its own strengths in different areas. Resnet GRU Attention did the best in catching single words (BLEU-1), while Mobilenet LSTM did quite well with pairs of words (BLEU-2). When it comes to longer phrases (BLEU-3 and BLEU-4), Resnet GRU Attention still did well, but the other models were really close. Even though VGG16 LSTM had slightly lower scores, the differences are pretty small, showing that all the models are quite similar in their performance. This means that the choice of the best

Table 3: N-Gram Example

Uni-Gram	Bi-Gram	Tri-Gram	Four-Gram
dog	dog is	dog is running	dog is running
is	is running	is running in	is running in the
running	running in	running in the	running in the snow
in	in the	in the snow	
the	the snow		
snow			

model might depend on specific needs or preferences for the task. There's room for exploring more and fine-tuning these models based on what's most important for the captioning job.

Table 4: Results of Models on Flickr8k English Dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16 LSTM	0.5133	0.3002	0.1856	0.1124
Mobilenet LSTM	0.5689	0.3445	0.2197	0.1381
Inception LSTM	0.5675	0.3478	0.2257	0.1430
Resnet LSTM	**0.5685**	**0.3516**	**0.2243**	**0.1388**
Resnet GRU	0.5583	0.3331	0.2057	0.1246
Resnet GRU Attention	0.5698	0.3372	0.2032	0.1182

B. CASE STUDY II (FLICKR8K ARABIC DATASET)

This part is dedicated to the performance results of different CNN-RNN models Flickr8k Arabic dataset.

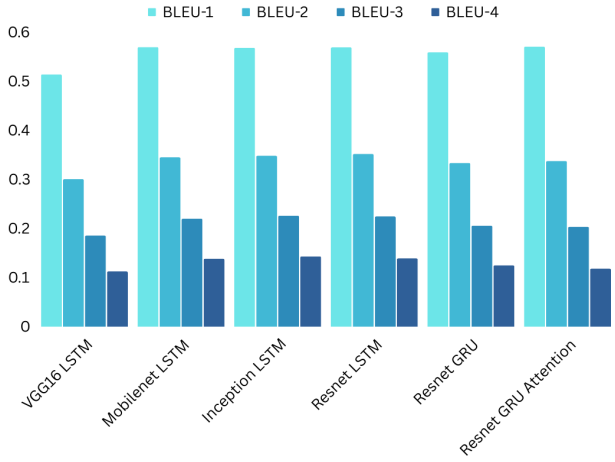
By inspecting the results of Table. 5, assessing the CNN-RNN models on the Flicker8k Arabic dataset for image captioning yields valuable insights into their respective performances. Mobilenet LSTM, while not securing the highest scores, displayed notable accuracy in capturing individual words (BLEU-1), showcasing its proficiency in single-word descriptions. Resnet LSTM and Resnet GRU Attention emerged as strong contenders, particularly excelling in the BLEU-1 metric, highlighting their effectiveness in delivering precise captions. The close competition among the models emphasizes their comparable performance, suggesting that the selection of an optimal model may hinge on specific task preferences. Further exploration and refinement offer promising avenues for enhancing these models' captioning capabilities in the context of Arabic Captions

Table 5: Results of Models on Flicker8k Arabic Dataset

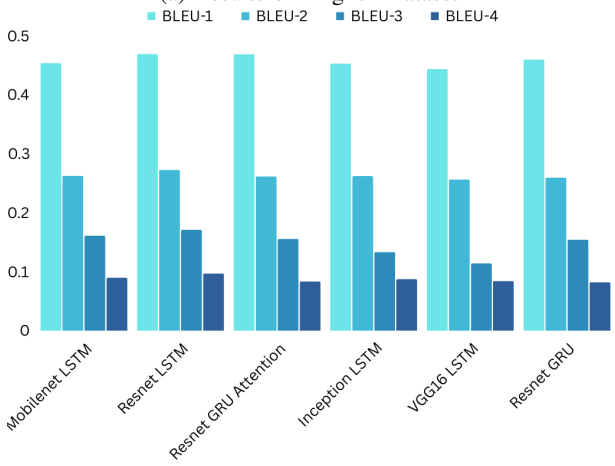
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Mobilenet LSTM	0.4536	0.2626	0.1611	0.0896
Resnet LSTM	**0.4691**	**0.2725**	**0.1711**	**0.0969**
Resnet GRU Attention	0.4686	0.2617	0.1555	0.0833
Inception LSTM	0.4532	0.2621	0.1333	0.0874
VGG16 LSTM	0.4435	0.2562	0.1143	0.0842
Resnet GRU	0.4598	0.2597	0.1542	0.0820

C. GRAPHICAL ANALYSIS

Figure 13 summarized the values of metrics in terms of BLEU-1, BLEU-2, BLEU-3 and BLEU-4, obtained by the CNN-RNN models for the English and Arabic captions datasets



(a) Results on English Dataset



(b) Results on Arabic Dataset

Figure 13: The performance metrics for Arabic and English captions using CNN-RNN models

V. CONCLUSION AND FUTURE WORK

A. CONCLUSION

In this research, we delved into the realm of image captioning, recognizing the essential role it plays in enhancing automatic image understanding. **The absence of detailed image descriptions poses a challenge for machines, necessitating the development of image captioning models**. The significance of image captioning extends to diverse domains.

The challenges in image captioning, particularly **decoding visual features and establishing coherent connections between visual elements**, prompted our exploration into leveraging classical deep learning techniques. Specifically,

the integration of Convolutional Neural Networks (CNNs) and Long short-term memory (LSTM) networks offered a powerful solution. This combination proved effective in capturing intricate visual patterns, handling long-range dependencies, and understanding temporal dynamics within image data.

Our contributions to the field include a thoughtful selection of datasets, **namely the Flickr8k dataset in both English and Arabic versions**. The dataset's richness, consisting of **eight thousand high-resolution photos with three descriptive captions for each image**, makes it a valuable resource for training and testing image captioning algorithms. Our approach involved preprocessing steps, including **resizing images, diverse pre-trained model integration (ResNet, MobileNet), and comprehensive experimentation and evaluation** on benchmark datasets.

We **conducted a detailed comparative analysis of various pre-trained models**, such as ResNet, MobileNet, Inception, and VGG16, aiming to understand their specific impacts on system performance. Our experiments involved **ResNet-LSTM, ResNet-GRU, ResNet-Attention-GRU, MobileNet-LSTM, Inception-LSTM, and VGG16-LSTM models**. The comparison was **visualized**, providing insights into the strengths and weaknesses, and a **flow chart** presenting out system architecture flow for a better understanding

For **English** captions, we addressed inconsistencies through pre-processing steps **like lower-case conversion, special characters removal, noise removal, and the addition of start and end tokens**. For **Arabic** captions, **extra spaces were removed, and start and end tokens were added**. These pre-processing steps aimed to standardize textual data and optimize the learning process of image captioning models.

In conclusion, our research contributes to advancing the understanding and implementation of image captioning **through a systematic exploration of classical deep learning techniques, careful dataset selection, and thorough model evaluation**. The comparative analysis provides valuable insights for researchers and practitioners working on image captioning applications, opening avenues for further advancements in the intersection of computer vision and natural language processing.

B. FUTURE WORK

While the proposed model demonstrates significant advancements in image captioning, there exist several avenues for future exploration and enhancement.

1) Exploration of Additional Modalities: The model focuses on the fusion of visual and textual information through multi-modal residual learning. To further enrich the model's understanding, future research could explore the **integration of additional modalities, such as audio or sensor data**. This extension would contribute to a more comprehensive and nuanced interpretation of diverse input sources.
2) Incorporation of Attention Mechanisms: The integration of attention mechanisms can enhance the model's interpretability by allowing it to **selectively focus on relevant regions within images or specific words in the textual input**. Future work could investigate the incorporation of attention mechanisms, such as those employed in models like Resnet Attention GRU, to improve the alignment between visual and textual elements, ultimately refining the quality of generated captions.
3) Cross-Domain Transfer Learning: Cross-domain **transfer learning is an area that holds promise for improving the generalization and adaptability of image captioning models**. Future research could explore techniques to transfer knowledge learned from one domain (e.g., a specific dataset) to another with distinct characteristics. This approach could facilitate the development of models capable of performing well in diverse image captioning scenarios.
4) Fine-Tuning Architectures for Efficiency: As the field progresses, there is a growing need for more efficient models, particularly in scenarios with resource constraints. Future work could focus on **refining and fine-tuning architectures like Mobile-Net and Inception to strike a balance between computational efficiency and captioning performance**. This optimization would be particularly valuable in real-time or edge-computing applications.
5) Exploration of Hybrid Architectures: Combining the strengths of different architectural components has shown success in various deep-learning tasks. Future research could **explore hybrid architectures that combine elements of both LSTM and GRU networks, leveraging the unique advantages of each**. This exploration could lead to models that effectively capture long-term dependencies while benefiting from the computational efficiency of GRUs.
6) Evaluation on Specialized Datasets: While the model's performance has been assessed on benchmark datasets, future work could **involve evaluating its effectiveness on specialized datasets related to specific domains or industries**. This step would provide insights into the model's adaptability and performance in application-specific contexts.
7) User-Centric Evaluation Metrics: In addition to traditional evaluation metrics, future research **could delve into user-centric evaluation metrics that measure the subjective quality of generated captions**. Understanding how well the model aligns with human perceptions and preferences is crucial for ensuring the practical usability of the image captioning system.
8) Analysis of Robustness and Generalization: Robustness to diverse inputs and generalization across various scenarios are critical aspects of any image captioning model. Future work could involve **a thorough analysis of the model's robustness**, exploring its performance

under challenging conditions, and investigating strategies to enhance its generalization capabilities.

By delving into these suggested directions for future research, the image captioning field can continue to evolve, addressing new challenges and pushing the boundaries of multimodal learning.

CONFLICTS OF INTEREST

The authors have declared that there is no conflict of interest. Non-financial competing interests.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this paper, where Adham Ahmed and Omar Helmy participated in sorting the experiments, discussed and analyzed the results, and revised/edited the manuscript. Engy Ahmed : performed the experiments and analyzed the results and wrote the paper. AbdulRaouf Monir: discussed the results and wrote the paper. Mohammed Matar : discussed the results and revised the paper. All authors reads and approved the work in this paper.

References

[1] G. Sairam, M. Mandha, P. Prashanth, and P. Swetha. Image captioning using cnn and lstm. volume 2021, pages 274–277, 2021.

[2] Ali Farhadi, Mohsen Hejrati, Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. volume 6314, pages 15–29, 09 2010.

[3] Chen Sun, Chuang Gan, and Ram Nevatia. Automatic concept discovery from parallel text and visual corpora, 2015.

[4] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. pages 1856–1864, 2018.

[5] Yezhou Yang, Ching Teo, Hal III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. pages 444–454, 01 2011.

[6] Yoshitaka Ushiku, Masataka Yamaguchi, Yusuke Mukuta, and Tatsuya Harada. Common subspace for model and similarity: Phrase learning for caption generation from images. pages 2668–2676, 2015.

[7] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Explain images with multimodal recurrent neural networks, 2014.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[9] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. pages 4565–4574, 2016.

[10] Xinlei Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. pages 2422–2431, 2015.

[11] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. pages 3128–3137, 2015.

[12] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). 2014.

[13] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn), 2015.

[14] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description, 2016.

[15] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language cnn for image captioning, 2017.

[16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions, 2016.

[17] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms, 2016.

[18] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. pages 4894–4902, 2017.

[19] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[20] Reshmi Sasibhooshan, Suresh Kumaraswamy, and Santhoshkumar Sasidharan. Image caption generation using visual attention prediction and contextual spatial relation extraction. 10(1):18, 2023.

[21] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning, 2016.

[22] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. pages 1242–1250, 2017.

[23] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. pages 4651–4659, 2016.

BIOGRAPHY

ADHAM AHMED ABDELMAKSOUD is pursuing a bachelor's degree in computer engineering and software systems in Ain Shams University - Faculty of Engineering, and expected to graduate in 2024. He has worked on several projects in the fields of machine learning, deep learning, data mining, and big-data analytics, which he majors during his studies. Adham additionally has knowledge in networks, distributed systems, embedded systems, software development, and other technical fields.

ABDULRAOUF MONIR MAHMOUD is currently pursuing a bachelor's degree in Computer Engineering and Software Systems at Faculty of Engineering, Ain Shams University, and expected to graduate in 2024. Enrolling in Ain Shams University's esteemed engineering program in 2019, AbdulRaouf quickly distinguished himself as an academic luminary. Excelling in courses ranging from computer science to advanced mathematics, he laid a formidable foundation for his foray into cutting-edge technologies. Simultaneously, AbdulRaouf channeled his creative energies into crafting intuitive and user-centric mobile applications. Projects such as Car Pooling App stand testament to their dedication to enhancing user experiences through seamless app development.

MOHAMED ALY MATAR is a forward-thinking Computer Engineering and Software Systems student at Ain Shams University, expected to graduate in 2024 with a Bachelor's degree. His academic journey is characterized by a profound interest in cutting-edge technologies, particularly Machine Learning (ML), Deep Learning (DL), and Data Analysis. Mohamed's passion is reflected in the successful completion of numerous projects within these domains, showcasing his practical application of theoretical knowledge. Beyond his academic pursuits, he has actively participated in hackathons and coding competitions, where he consistently demonstrates his problem-solving acumen and innovative thinking. With a versatile skill set encompassing Flutter for mobile app development and proficiency in web development, Mohamed is a well-rounded professional poised to make significant contributions to the tech industry.

ENGY AHMED HASSAN is currently working towards a bachelor's degree in computer engineering and software systems at Ain Shams University - Faculty of Engineering, with an anticipated graduation in 2024. Throughout her academic journey, she has actively engaged in various projects within the realms of machine learning, deep learning, data mining, and big-data analytics—areas in which she specializes. Engy possesses expertise in mobile programming, embedded systems, software engineering, and various other technical fields, showcasing a well-rounded skill set cultivated during her studies.

OMAR HELMY ELBANNA is currently pursuing a bachelor's degree in Computer Engineering and Software Systems at the Faculty of Engineering, Ain Shams University, and expected to graduate in 2024. Throughout his academic journey, Omar has demonstrated a strong interest and proficiency in machine learning, deep learning, computer vision, and embedded systems. Omar's passion for technology extends beyond the classroom, as he actively engages in hands-on projects and research initiatives. His commitment to exploring cutting-edge fields is evident through his contributions to various aspects of computer engineering.