Arabic Image Captioning

Submitted by:

Mariem Abdou 202000595

Alaa Hatem 202001170

Michael Hanna 202002662

Remon Gerges 19104593

Under the supervision of:

Dr. Ghada Khoriba

Eng. Aly Abdelmageed

(CSCI462)

I. Introduction

The process of image captioning aims to create accurate and meaningful descriptions of the content of an image in natural language. Accomplishing this objective requires identifying the key elements present in the image, such as objects, people, and locations, and conveying their meaning in a way that reflects the image's mood or emotion and provides contextual understanding. While the human brain can quickly identify and interpret the contents of an image, machines require computer-based models to aid them in this task [1]. These models leverage advanced technologies such as machine learning, computer vision, and natural language processing to analyze the visual content of images and generate descriptive captions. However, the complexity of the Arabic language presents a unique challenge to this process [2].

One of the main challenges in Arabic image captioning is the extensive variety of dialects and regional variations within the Arabic language. These differences can make it difficult for a machine learning model to precisely comprehend and portray the content of an image in Arabic. Another challenge in Arabic image captioning is the complex meanings of Arabic words and phrases. The Arabic language is known for its rich vocabulary and subtleties of meaning depending on their context as shown in figure 1, making it challenging to generate accurate and natural-sounding descriptions of images [1].

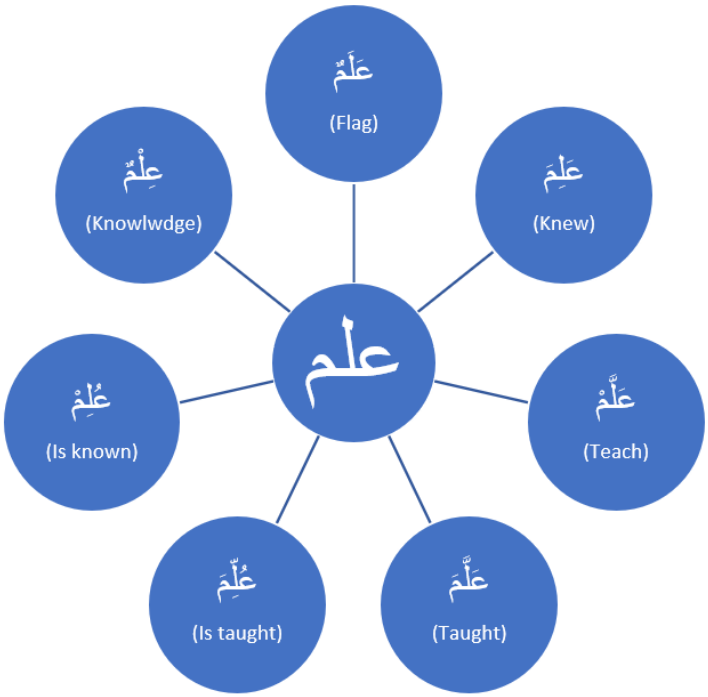

Figure 1: Ambiguity in Arabic Language [3]

The potential applications of Arabic image captioning are numerous. One of the most significant is in e-commerce. Many businesses rely on images to promote their products online, and automated image captioning can help them reach a wider audience by providing accurate and meaningful descriptions of their products in Arabic. Arabic image captions are very useful in social media. With the growing use of social media platforms there is a rising need for tools that can automatically generate captions for images in Arabic. This can help users expand their reach and interact with their followers more effectively. Moreover, it can also benefit individuals with visual impairments. By providing detailed descriptions of images, these individuals can better comprehend the visual content they encounter online [4].

The ability to automatically generate descriptive captions for images in Arabic is increasingly vital in today's digital age. As the internet continues to expand, the amount of Arabic content available online has also significantly increased. This includes images, which are an essential component of the online world. To make these images more accessible and useful, there is a need for tools that can automatically generate descriptive captions in various languages.

II. Background

This literature review aims to provide a comprehensive overview of the existing efforts in the field of image captioning. The section is divided into two subsections. English image captioning-related publications will be included in the first subsection, while Arabic image captioning-related articles will be covered in the second.

English Image Captioning

Huang et. al. (2019) [5] used MS COCO dataset for image captioning in English. They applied the concept of Attention on Attention (AoA) to both the image encoder and the caption decoder of their image captioning model, AoANet. For the encoder, it extracts feature vectors of the image's objects, models their relationships with the vectors using self-attention, and then uses AoA to ascertain their interactions with one another. To maintain only the useful attention results and filter out the unnecessary or deceptive ones for the decoder, AoA is used. AoA determines the relevance between the attention result and query, whereas multi-modal fusion combines data from several modalities; AoA is a general expansion to attention processes and may be used to any of them. With AoA, there is just one "attention gate" necessary and no hidden states. In contrast, LSTM/GRU is only relevant to sequence modelling and calls for hidden states and more gates.

Al-Malla et. al. (2022) [6] tested their method using the MS COCO and Flickr30k datasets, which are typically used for picture captioning. An attention-based Encoder-Decoder architecture has been proposed in their study. It uses two feature extraction techniques for image captioning: an object detection model (YOLOv4) and an image classification CNN (Xception). They employed the Xception CNN that had been trained on ImageNet to extract spatial information during image encoding. In line with current developments in picture captioning, they extract features from the layer immediately preceding the fully linked layer. Instead of merely

concentrating on the image class, the total model may now learn more about the objects in the image and the connections among them. Moreover, they employ the YOLOv4 model for object detection since it is quick and accurate, making it suited for massive data and real-time applications. The Bahdanau soft attention methodology is utilised in their approach. Two completely connected layers follow the GRU. The first one has a length of 512, while the second one has a vocabulary size that will result in output text. The training process for the decoder is as follows. First, the encoder is run once the features have been extracted. Second, the encoder output, the hidden state, which is initialized to 0, and the decoder input, which is the start token, are sent to the decoder. Third, the decoder returns the predictions and the decoder's hidden state. Fourth, the loss is then determined using the predictions and the hidden state of the decoder, which is subsequently fed back into the model. Finally, the approach known as "teacher forcing" is used to choose the next decoder input, passing the target word as the input. They employed a collection of assessment metrics that are frequently employed in the picture captioning industry: METEOR and BLEU.

To sum up, there have been a number of intriguing attempts at image captioning in the English language, all of which varied in their accuracy.

Arabic Image Captioning

To the best of our knowledge, Arabic imagine captioning hasn't been the subject of a lot of research. However, some recent works have started exploring this area. This section will provide a quick overview of these works.

An important contribution was made by Shalabi and Obeid (2019) [7], who suggested a deep learning-based method for producing Arabic image captions by combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. In comparison to other methods used at the time, their method produced captions that were more accurate and cohesive, and it showed promise in terms of BLEU scores.

Afyouni et al. (2021) [8] aimed to improve the accuracy of Arabic image captioning by proposing three models: object-based captioner, attention-based captioner, and object detector + attention-based captioner. The authors used Microsoft COCO and Flickr30k datasets and translated the English captions to Arabic using the Google cloud Translation API. The models were evaluated using multi-lingual semantic sentence similarity techniques to assess the generated captions' accuracy against the actual ground truth captions (in English). The results showed that the similarity scores for Arabic generated captions from all three proposed models outperformed the basic captioning technique.

Lasheen and Barakat (2022) [9] aimed to improve Arabic image captioning by proposing an encoder-decoder architecture. The authors used Google Translate to translate the Arabic-Flickr8 dataset from the original English Flickr8 utilised attention in picture categorization in Arabic. Karpathy's data is separated into three groups for training, validation, and testing: 6000, 1000, and 1000 photos are used for each group. The model utilised in this study has an attention mechanism,

teacher forcing, and beam search along with an encoder-decoder architecture. They utilised The RESNet-101 in the encoder. They employed an LSTM network in the decoder. To create the image descriptions, they applied two models. The model achieved a BLEU-4 score of 27.12, which is an improvement over the previous state-of-the-art results.for both models, trained AraVec embedding was employed. However, the data was preprocessed pyArabic tokenizer in models 1 vs. using FARASA Segmenter in model 2. It was evident that models 2 outperforms model 1 in terms of BLEU scores.

Emami et al. (2022) [10] aimed to improve Arabic image captioning by proposing an approach that combines region-based visual features and object tags with a language model. The authors made use of the Microsoft COCO and Flickr8k datasets. With the use of the Advanced Google Translate API2, also known as Arabic-COCO, they used 414,113 pre-translated captions over 82,783 training photos. They employed a two-step pipeline as their methodology. First, use a convolutional neural network (CNN) encoder to extract area features and object tags from a picture. Then, using a language model, in this case a pre-trained transformer, create a sentence using the region features and object tags. For feature extraction, they used the X152-C4, and they pre-trained on 2.49 million different images, including the COCO dataset. Moreover, they used pre-trained transformers (ArabicBERT, GigaBERT, and AraBERT) to generate the final captions. The proposed models were evaluated using BLEU-1, 2, 3, 4, and found that the best performing model scored 0.39, 0.25, 0.15 and 0.092, respectively. They demonstrate that while using Arabic captions and English object tags when training image captioning models is a viable option, using a pure Arabic dataset with Arabic object tags is superior.

Despite the progress made in Arabic language processing, there remains a requirement to develop more efficient models capable of effectively handling the language's unique structural and syntactic features. Moreover, it is vital for researchers to continue collaborating to compile more comprehensive Arabic datasets that contain a wider range of visual scenes and linguistic expressions.

III. Methodology

Dataset description

The Flickr Image dataset [11] is a large collection of photos and their corresponding captions. The dataset contains over 1 million images, each with several captions describing the contents. The images cover a wide variety of topics, such as animals, nature, people, and architecture. The captions are written in English and cover a range of styles, from factual descriptions to poetic interpretations.

The Flickr30k dataset is a standard benchmark for sentence-based image description. Flickr30k Entities augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued

progress in automatic image description and grounded language understanding. They enable us to define a new benchmark for the localization of textual entities mentioned in an image.

Preprocessing

After downloading and unzipping the data, we began exploring the data. The data consisted of 30 thousand images along with a csv file that contains the captions of each image. There were five captions per image.

For the images, we first addressed their varying sizes by resizing them to (240,240) pixels to ensure uniformity. Additionally, we converted the color channels from BGR to RGB and normalized the images.

Regarding the captions, we combined the multiple captions associated with each image into a single list, simplifying the data structure. We also incorporated any comments related to the captions into our dictionary. Next, we performed several preprocessing steps on the captions:

- Eliminating extra spaces
- Converting all capital letters to lowercase
- Removing special characters, punctuation marks, and numerals
- Adding two tokens, "startseq" and "endseq," to mark the beginning and end of each description, aiding the model in learning caption generation.

Subsequently, the captions were tokenized, converting each word in the caption into an index representation. To split the data for training, validation, and testing, we used 80% of the dataset for training, while the remaining 20% was divided between validation and testing sets. By following these preprocessing steps, we prepared the dataset, ensuring consistency and compatibility for training our Arabic image captioning model.

Model architecture

The Figure below illustrates the whole model implementation architecture. The model used consists of 3 main components: a pretrained encoder, decoder with attention, and a pretrained transformer. First the preprocessed image goes through the encoder, extracting the features of the image. Next, the extracted features along with the text embeddings are passed through the decoder to generate the captions sequence. Finally, the generated captions are used by the pre trained transformer to translate the English captions to Arabic captions.

The chosen model architecture incorporates attention blocks, which allow the model to focus on important features and disregard irrelevant information. Moreover, it has been proven to lead the better results in NLP tasks. The selection of this model architecture was inspired by Xu et. al. (2015) [12] since their model could accurately tell where the model is looking. By utilizing attention blocks, our model can effectively align image features with corresponding textual information, enhancing the caption generation process. This alignment enables the model to attend to relevant visual elements while generating accurate and contextually meaningful captions.

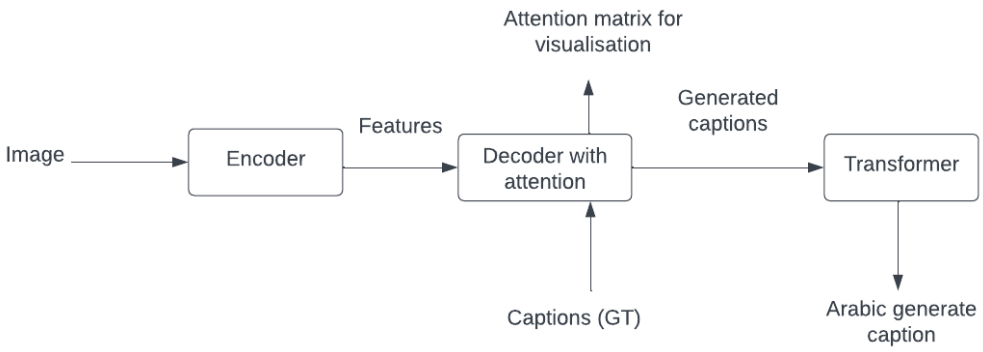

Figure 2: Whole Model Architecture

a. Encoder

The first component of our architecture is the encoder. We have used a pretrained ResNet101 model to extract the features from the preprocessed images. ResNet (Residual Network) is convolutional deep learning model that makes use of residual blocks and skip connections. The residual blocks usually consist of stacked convolution layers with varying number of filters. The residual blocks are stacked after each other and surrounded with skip connections. These skip connections allow the flow of gradient between the residual blocks, therefore enabling the training deeper models without struggling with vanishing gradients.

There are several variations if the vanilla ResNet model such as ResNet34, ResNet50, and ResNet101. ResNet101 is a variation of the vanilla ResNet model which consists of 101 layers. The ResNet101 model used was pretrained on the ImageNet dataset which makes the model experienced with extracting relevant and important features from the images.

b. Decoder with Attention

The decoder with attention is the second component of the architecture. It mainly consists of an attention block and LSTM blocks. The decoder takes in the encoded features along with the encoded captions. First, the attention block computes an alpha array and an attention weighted encoding array. More details on this will be provided in the next paragraph. After that, the attention weighted encoding matrix is multiplied by the sigmoid of the LSTM hidden state. The result is then concatenated with the embedded captions. This acts as an input to the LSTM model. The model consists of 50 units, and it outputs a hidden state and a context vector. Both the context vector and hidden state are redirected as input for the next computation. Moreover, the hidden state is also sent to the attention block, sigmoid layer, and fully connected layer. The output of the fully connected layer is our predicted caption. The decoder architecture is presented in the Figure below.

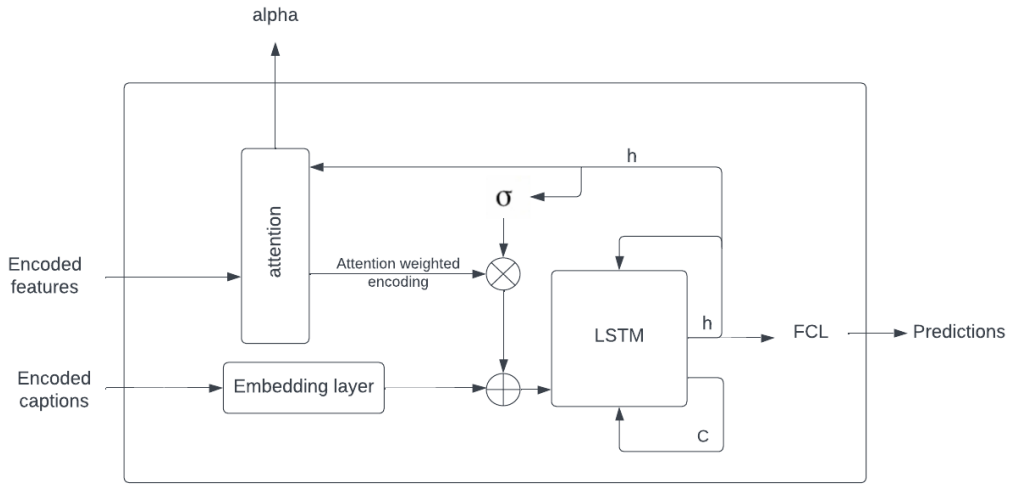

Figure 3: Decoder with Attention

The attention is a major component of this decoder architecture. Its inputs are the encoded image features as well as the LSTM hidden state which can be thought of as the text features. They are both multiplied by weight matrices and concatenated together. The concatenated matrix is then passed through a ReLu activation function which only saves the matrix's important features. The SoftMax layer is then used to transform the inputs into a probability space which will be used in attention layer visualization and the computing of the attention weighted encoding vector. The attention weighted encoding vector is calculated by concatenating the SoftMax layer output and the image features vector. The steps of the attention block are shown in the Figure below.

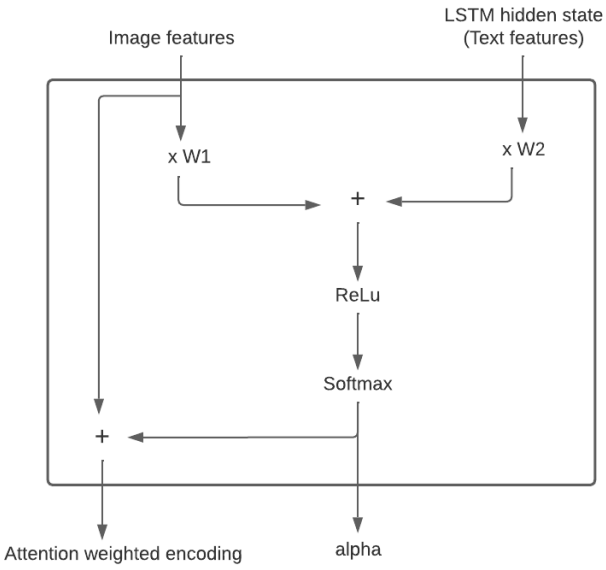

Figure 4: Attention block

c. Transformer

The final step was to translate the models' pretrained English captions to Arabic captions. We have decided to use a pretrained transformer to achieve that. We have used an English to dialect model which generates translations from English to colloquial Arabic, conditioning the translations on dialect. The model was trained by fine-tuning the opus-mt-ar-en (MSA to English) model on ~85K parallel sentences in four dialects of colloquial Arabic. We have chosen to work with the Egyptian dialect.

Architecture training

The chosen model was trained using Google Colab, a cloud-based platform that provides access to powerful computing resources, facilitating efficient training of deep learning models. During the training process, several parameters were carefully selected to optimize the model's performance. These parameters include a batch size of 64, learning rate of 4e-3 (0.004), Adam optimizer, and 35 epochs. These training parameters were chosen to optimize the model's performance in terms of convergence, accuracy, and generalization to unseen data.

IV. Results and discussion

This section presents the results of the model used. We have used 2 metrics to assess the performance of the model: BLUE-4 and top 5 accuracy.

The Top-5 accuracy metric [13] is used to assess a machine learning model's performance in multi-class classification problems. It gauges the proportion of test examples for which the correct label falls inside the model's top five projected labels. Top-5 accuracy offers a more forgiving evaluation metric than standard accuracy, which makes it particularly useful when the dataset is huge, and the number of classes is high.

The BLEU-4 (bilingual evaluation understudy) [14] is based on the idea that the closer the predicted sentence is to the human-generated target sentence, the better it is. BLEU Scores are usually between 0 and 1. A score of 0.6 or 0.7 is considered the best you can achieve. Even two humans would likely come up with different sentence variants for a problem and would rarely achieve a perfect match. For this reason, a score closer to 1 is unrealistic in practice and should raise a flag that your model is overfitting. Our training metrics for the top 5 accuracy and BLEU score were 60.851, and 0.117 respectively.

We have tested our model using the beam search algorithm. The beam search algorithm [15] is used in natural language processing (NLP) to generate a word sequence from an input. To evaluate how successfully beam search algorithms handle NLP tasks like text generation or machine translation, beam size is a measure that is employed. The beam size is the number of candidate sequences that are kept at each stage of the beam search procedure. In the testing phase, we used only BELU-4 metric which was 0.1253. A few of the captions generated are shown below.

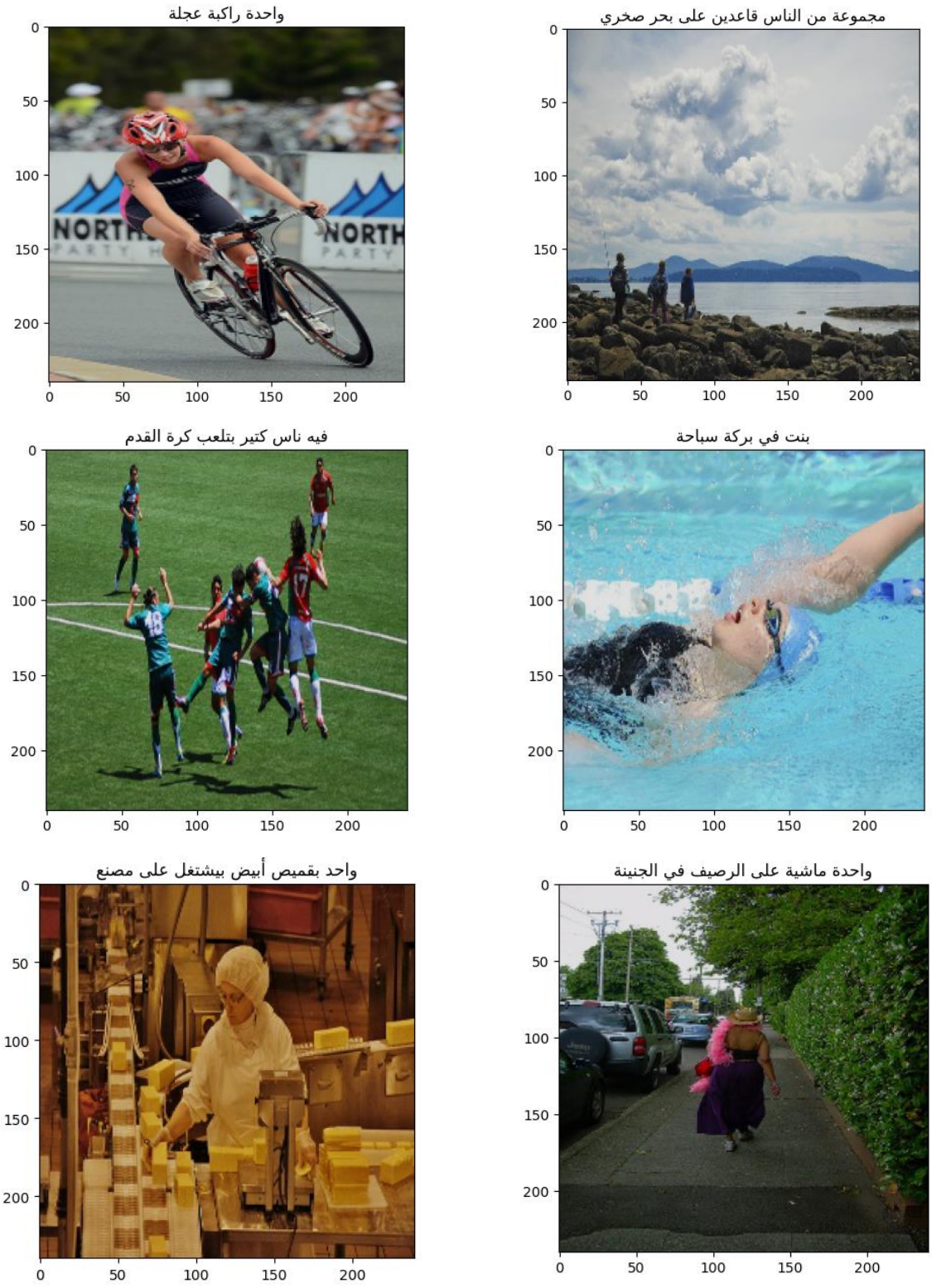

Figure 5: Generated captions sample

Visualizing the attention component of a model can provide valuable insights into its decision-making process and enhance the understandability of its output. By visualizing the attention block, we can gain a clearer understanding of why the model is producing specific captions and errors. Moreover, we can observe whether the model attends to relevant objects, regions of interest, or specific visual cues that influence its captioning decisions. Attention block visualization can also help identify potential biases or shortcomings in the model's behavior. For instance, if the model consistently attends to certain objects or regions in the image while neglecting others, it may indicate a bias towards specific visual features. This insight can be used to refine the model and improve its performance by addressing these biases. Two of the attention visualizations are shown in the Figure below.

Figure 6: Attention visualizations

V. Conclusion

In this paper, we have performed Arabic image captioning on the Flicker 30k dataset. We have used a model that contains 3 components: an encoder, decoder with attention, and a transformer for the translation. The results showed a top 5 accuracy of 60% and BLEU-4 score of 0.117 during the training process. During testing, however, we managed to achieve a BLEU-4 score of 0.125. We have faced limitations while visualizing the attention block outputs for the Arabic captions. The sequence mismatch between the original English captions and the translated Arabic captions posed challenges in plotting the attention weights accurately.

To address this limitation and further enhance our model, we intend to integrate the machine translation transformer to translate the text to Arabic first and train the model on Arabic embedded text. By translating the text to Arabic prior to training and using Arabic embedded text, we aim to improve the alignment between the attention weights and the generated captions. This integration would enable us to visualize and interpret the attention mechanism more effectively, providing valuable insights into the model's decision-making process. In conclusion, we hope that our research has contributed to the field of Arabic image captioning.

References

[1] B. Alazzam, “Arabic Image Captioning using ResNet50,” *Researchgate*, Mar. 2022, Accessed: Apr. 15, 2023. [Online]. Available: https://www.researchgate.net/publication/359229933_Arabic_image_captioning_using_ResNet50

[2] S. Sabri, “Arabic Image Captioning Using Deep Learning with Attention,” Aug. 2018, Accessed: Apr. 15, 2023. [Online]. Available: https://www.ai.uga.edu/sites/default/files/inline-files/theses/sabri_sabri_m_202108_ms.pdf

[3] M. Khader, A. Awajan, and A. Alkouz, “Textual Entailment for Arabic Language based on Lexical and Semantic Matching,” *International Journal of Computing and Information Sciences*, vol. 12, no. 1, pp. 67–74, Sep. 2016, doi: 10.21700/ijcis.2016.109.

[4] R. S. Al-Malki and A. Y. Al-Aama, “Arabic Captioning for Images of Clothing Using Deep Learning,” *Sensors*, vol. 23, no. 8, p. 3783, Apr. 2023, doi: 10.3390/s23083783.

[5] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, “Attention on Attention for Image Captioning,” Aug. 2019.

[6] M. A. Al-Malla, A. Jafar, and N. Ghneim, “Image captioning model using attention and object features to mimic human image understanding,” *J Big Data*, vol. 9, no. 1, p. 20, Dec. 2022, doi: 10.1186/s40537-022-00571-w.

[7] L. Shalabi and N. Obeid, “Arabic Image Captioning based on Deep Learning. ,” *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 4, pp. 455–460, 2019.

[8] I. Afyouni, I. Azhar, and A. Elnagar, “AraCap: A hybrid deep learning architecture for Arabic Image Captioning,” *Procedia Comput Sci*, vol. 189, pp. 382–389, 2021, doi: 10.1016/j.procs.2021.05.108.

[9] M. T. Lasheen and N. H. Barakat, “Arabic Image Captioning: The Effect of Text Pre-processing on the Attention Weights and the BLEU-N Scores,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022, doi: 10.14569/IJACSA.2022.0130751.

[10] J. Emami, P. Nugues, A. Elnagar, and I. Afyouni, “Arabic Image Captioning using Pre-training of Deep Bidirectional Transformers,” in *15th International Conference on Natural Language Generation*, 2022. Accessed: Apr. 15, 2023. [Online]. Available: https://aclanthology.org/2022.inlg-main.4.pdf

[11] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models,” May 2015.

[12] K. Xu *et al.*, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” Feb. 2015.

[13] A. T. Dang, “Accuracy and Loss: Things to Know about The Top 1 and Top 5 Accuracy.,” *Medium*, 2021. https://towardsdatascience.com/accuracy-and-loss-things-to-know-about-the-top-1-and-top-5-accuracy-1d6beb8f6df3 (accessed May 14, 2023).

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*, Morristown, NJ, USA: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[15] A. K. Vijayakumar *et al.*, “Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models,” Oct. 2016.