Ar-CM-ViMETA: Arabic Image Captioning based on Concept Model and Vision-based Multi-Encoder Transformer Architecture

Asmaa Osman
Department of Information Technology, Cairo University
Egypt
asmaa.a.elsayed@fci-cu.edu.eg

Mohamed Shalaby
Department of Information Technology, Cairo University
Egypt
m.wahby@fci-cu.edu.eg

Mona Soliman
Department of Information Technology, Cairo University
Egypt
mona.solyman@fci-cu.edu.eg

Khaled Elsayed
Department of Information Technology, Cairo University
Egypt
khaledms@fci-cu.edu.eg

Abstract: *Image captioning is a major artificial intelligence research field that involves visual interpretation and linguistic description of a corresponding image. Successful image captioning relies on acquiring as much information as feasible from the original image. One of these essential bits of knowledge is the topic or the concept that the image is associated with. Recently, concept modeling technique has been utilized in English image captioning for completely capturing the image contexts and make use of these contexts to produce more accurate image descriptions. In this paper, a concept-based model is proposed for Arabic Image Captioning (AIC). A novel Vision-based Multi-Encoder Transformer Architecture (ViMETA) is proposed for handling the multi-outputs result from the concept modeling technique while producing the image caption. BiLingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) standard metrics have been used to evaluate the proposed model using the Flickr8K dataset with Arabic captions. Furthermore, qualitative analysis has been conducted to compare the produced captions of the proposed model with the ground truth descriptions. Based on the experimental results, the proposed model outperformed the related works both quantitatively, using BLEU and ROUGE metrics, and qualitatively.*

Keywords: *Arabic image captioning, transformer, concept, computer vision.*

Received January 18, 2024; accepted March 19, 2024
https://doi.org/10.34028/iajit/21/3/9

1. Introduction

We see several images every day from a variety of sources, including the internet, news and documents. Images from these sources are left up to the viewers' interpretation. Although the majority of these images lack descriptions, most people can still understand them without a description. Providing a caption for an image includes giving clear and brief description of the corresponding image.

Image captioning is a necessary task because it is required by many applications. In the past 20 years, the fields of image indexing, information retrieval, robotics, and medicine have all made extensive utilization of image captioning [13]. Captioning images is challenging because it requires both an in-depth comprehension of the semantic elements of the image and the ability of expressing these elements in words that seem natural. So, the image captioning involves employing Computer Vision (CV) approaches to extract the image's features in addition to natural language processing approaches to produce the description [28].

Arabic Image Captioning (AIC) is the task of describing input image in Arabic sentence. The majority of research projects in the image captioning concentrate on producing descriptions in English language, that is primarily due to the shortage of the available public datasets for other languages, particularly Arabic. The Arabic language is incredibly complicated and, as a result, very challenging to work with. The language contains various dialects and is primarily caused by the usage of diacritics. The captioning models need to recognize the text's semantics included in the intended language [1].

Great attention should be paid for Arabic language because it is the first language in 22 countries and there are about 420M people speak it in the Arab world [8]. Arabic is additionally the fourth most often utilized language on the internet. Furthermore, it has been the most rapidly growing language in the last eight years. Subsequently, even though there have been considerable improvements in English captioning models, those models are not immediately applicable for other languages, like Arabic. Therefore, the image captioning for the Arabic language is still under developing [8].

Successful image captioning relies on acquiring as much information as feasible from the original image. One of these essential bits of knowledge is the concept

or the topic that the image is associated with. In order to produce these concepts, the concept modelling technique takes both the caption data and the images into account when determining which concepts to extract. The concept modeling technique has been utilized in image captioning for completely capturing the image contexts and make use of these contexts to produce more accurate image descriptions.

Concept modeling technique has been developed two years ago by Grootendorst [11] through customizing the topic modeling to consider images. Concept modeling technique is a multimodal technique that generates set of concept vectors based on the input images and their corresponding texts using Contrastive Language-Image Pre-Training (CLIP) model [24] and BerTopic model [10]. The idea behind the concept modeling technique is to extremely capture the image's semantic information. The concept modeling technique can work on 50+ languages [11].

In this paper, the concept modeling technique is used to propose concept-based model for AIC. Concept modeling technique is used on the dataset images and Arabic captions to extract set of new concept vectors and the images embedding. After that, the concept model's outputs are sent to the decoder.

Recently, transformer [29] is utilized in image captioning as a decoder to provide more accurate captions with less complexity due to its parallelization capabilities and the recursion bypassing. The conventional transformer makes use of one encoder to represent the features of the image and one decoder to decode the partial captions by paying attention to the encoded features.

To pay attention for concept data in addition to the image's features, this research work proposes Vision-based Multi-Encoder Transformer Architecture (ViMETA). A new encoder has been added to the transformer to reflect the concept information. Moreover, the transformer decoder is modified by adding an encoder-decoder attention layer so that the attention can be paid to the concept information that is represented by the new encoder. The encoders of the proposed (ViMETA) is inspired from the vision transformer encoder [7].

ViMETA is proposed in this paper as the decoder for the captioning model. The ViMETA includes two encoders and one decoder. The first transformer encoder receives the concept vectors and the second transformer encoder receives the CLIP features. Then, the transformer decoder is utilized for generating the next words in the caption depending on the information given by the two encoders and the embeddings of the partial captions.

The proposed AIC based on concept model and Vision-based Multi-Encoder Transformer Architecture (Ar-CM-ViMETA) model has been evaluated on the Flikr8K dataset [12] using BiLingual Evaluation Understudy (BLEU) [23] and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [18].

Contributions of this research are as follow:

1. Proposed ViMETA.
2. Proposed Ar-CM-ViMETA.

The structure of this paper is as follow: in section 2, the related work is discussed; the proposed model is presented in section 3; the experimental work and results are included in section 4 and the proposed work has been concluded in section 5.

2. Related Work

The most recent state-of-the-art in image captioning tasks is represented by attention-based approaches [22]. The attention methods were first put to use into the machine translation problem [4]. Based on the significantly improved outcomes in machine translation task, the attention method was subsequently used in the captioning task in addition to many other tasks like the breast cancer classification [5]. The attention methods focus on the image's main portions and useful attributes then make use of these information to decide where to concentrate next during producing the appropriate caption. Several attention-based methods were established with the intention of improving the final captions.

Based on the literature presented by Osman *et al*. [22], the main attention categories that achieved great improvement in the image captioning field are the guided-attention and the transformer-based methods. The task of image captioning may take a guidance either from information based on the texts or information retrieved from the features of the images. The image captioning is frequently guided by the topic modeling techniques [6, 30]. The topic modeling was established for a text-based data for extracting the latent variables in a huge dataset.

Recently, the image captioning has been guided by concept modeling technique for capturing the semantic information included in the image. The concept modeling technique has been employed to extract set of concept vectors in addition to the image's embeddings. The outputs of the concept modeling technique are then sent to a multi-encoder transformer architecture to produce the output caption.

Multi-encoder transformer architectures have been developed for text-based tasks [20, 25, 26, 27]. It is proposed for the automatic-post editing task such that the transformer has been modified to include two encoders [26]. The first encoder represents the machine translation sentence (*mt*) and the other encoder represents the source sentence (*src*). The transformer decoder in [26] is also modified such that three cross-attention layers are included.

The encoders' architecture of the multi-encoder transformer architectures, previously proposed, is identical to the encoder's architecture of the standard

transformer. These architectures work well with text data but it needs to be updated to perform in a good way with images. For this purpose, a novel multi-encoder transformer architecture is proposed in this paper by updating the encoder's architecture to have the same architecture of the vision transformer encoder [7].

For Arabic Image Captioning (AIC), Al-Muzaini *et al*. [2] constructed an Arabic dataset depending on MSCOCO [19] and Flickr8K [12] datasets. In addition, they employed a merging model to produce the captions utilizing Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)-Long Short-Term Memory (LSTM).

Ultra edit smart translating system was utilized by Mualla and Alkheir [21] for producing Arabic captions. CNN have been employed for feature extraction after that the extracted features in addition to translated captions are sent to LSTM. The authors observed that just translating the observed captioning in English to Arabic is not a good idea due to the bad structure of the generated Arabic sentences.

Jindal [15] took use of the Arabic language's strong root word influence to construct root words by using images rather than captions. Instead of extracting real phrases from the images, a CNN is employed for extracting set of root words. The roots are subsequently converted to morphological inflections. The roots are subsequently converted to morphological inflections then the order of the words in the statement has been verified using a dependency tree. The findings indicate that producing Arabic descriptions in one step as opposed to translating English captions into Arabic in two steps gave better outcomes.

Eljundi *et al*. [8] constructed an Arabic translated data of the Flickr8K descriptions and made it public. The authors additionally built an end-to-end approach which converts image into Arabic sentences and compared it to a basic approach to Arabic captioning that depends on translating texts from English image descriptions. Their proposed model achieved 33.2, 19.3, 10.5, 5.7 on BLEU1, BLEU2, BLEU3 and BLEU4, respectively [8].

None of the previously mentioned works used attention or transformer in the AIC a transformer-based AIC was presented by Emami *et al*. [9]. The authors used CNN for feature extraction then a pre-trained bidirectional transformer has been employed as a language model for generating the caption words. Their proposed AraBERT32 model achieved 39.1, 24.6, 15.0, 9.2, 33.1 on BLEU1, BLEU2, BLEU3, BLEU4 and ROUGE, respectively.

In this research work, a concept-based image captioning model has been proposed for AIC. A ViMETA has been proposed as a language model to produce the description of the image given the outputs of the concept modeling technique.

3. Proposed Arabic Image Captioning based on Concept Model and Vision-based Multi-Encoder Transformer Architecture (Ar-CM-ViMETA)

The proposed AIC model based on concept model and ViMETA is presented in Figure 1. The proposed Ar-CM-ViMETA model starts by applying the concept modeling technique on the input images and the corresponding Arabic captions.

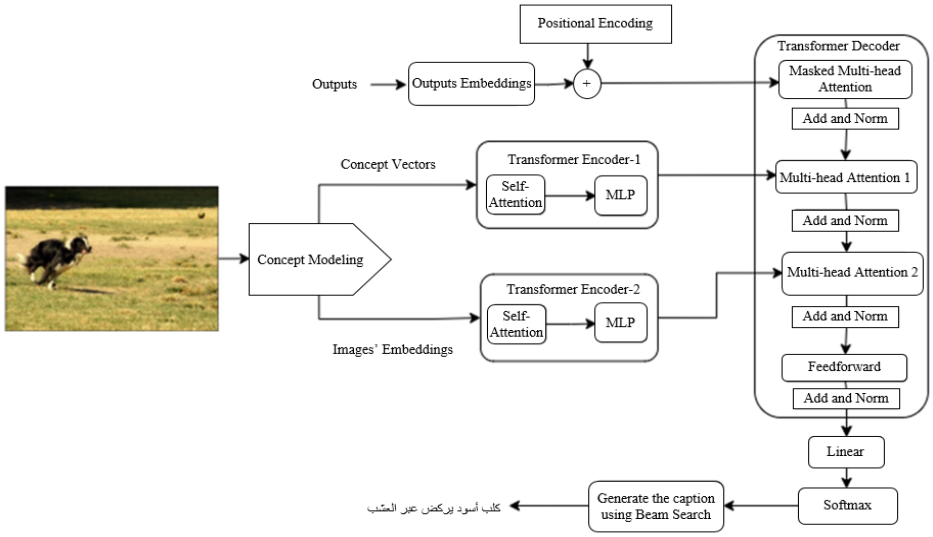


Figure 1. The proposed AIC based on concept model and ViMETA.

The resulting concepts after applying the concept modeling technique on the Arabic captions Flikr8K dataset [9] is shown in Figure 2. A set of concept vectors and CLIP image's embeddings are the output of the concept modeling technique. Each concept vector, corresponding to an image, has a length equal to number of the concepts. This vector represents the concepts distribution for the corresponding image. The outputs of the concept modeling technique are then sent to the decoder to produce the caption words.

a) Concept 25. b) Concept 22. c) Concept 15.

Figure 2. Sample of the resulting concepts after applying the concept modeling technique on the Flickr8K dataset with Arabic captions.

For the purpose of producing more expressive captions at lower computational complexity, transformer is used in this research work because of its ability to bypass the recurrence and its parallelization design. The standard transformer includes one encoder for handling the input image's features and one decoder for producing the resulting caption based on the encoded features and the partial captions.

The resulting concepts after applying the concept modeling technique on the Arabic captions Flikr8K dataset [9] is shown in Figure 2. A set of concept vectors and CLIP image's embeddings are the output of the concept modeling technique. Each concept vector, corresponding to an image, has a length equal to number of the concepts. This vector represents the concepts distribution for the corresponding image. The outputs of the concept modeling technique are then sent to the decoder to produce the caption words.

For the purpose of producing more expressive captions at lower computational complexity, transformer is used in this research work because of its ability to bypass the recurrence and its parallelization design. The standard transformer includes one encoder for handling the input image's features and one decoder for producing the resulting caption based on the encoded features and the partial captions.

In this paper, it is proposed to modify the traditional transformer by having multi-encoder architecture for representing the concept vectors in the captioning process. Therefore, a novel transformer architecture is proposed by changing the traditional transformer to have multiple encoders, i.e., *M* encoders. In addition, an encode-decoder attention layer is added to the decoder for each added encoder, i.e., *M* encoder-decoder attention layers. The proposed ViMETA is shown in Figure 3.

Inspired by the architecture of the Vision Transformer (ViT) [7] which was designed specifically for CV tasks, a ViMETA is proposed by using the same encoder design of the ViT. The encoder of the ViT architecture includes self-attention module and Multi-layer Perceptron (MLP) module. In addition, layernorm is used before each module and residual connection after each module.

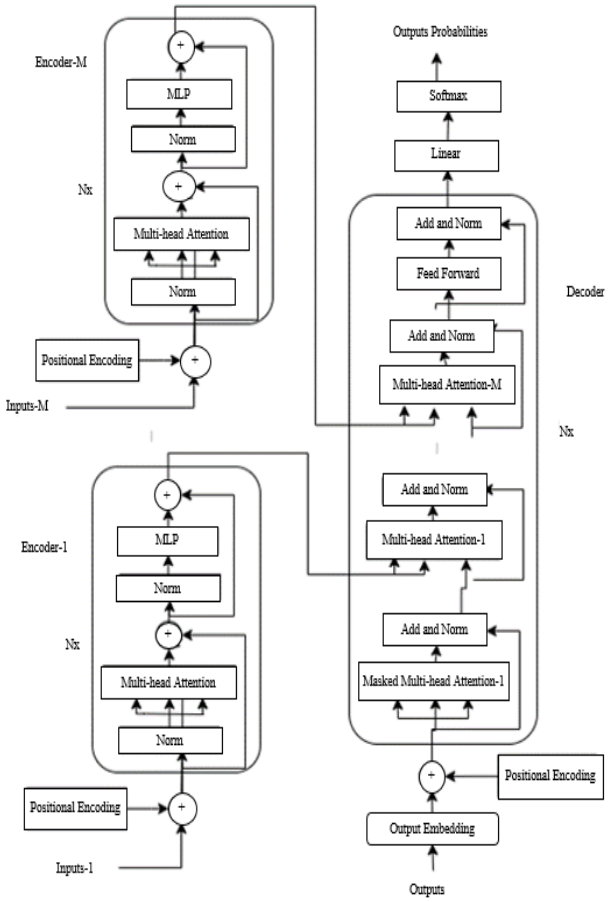


Figure 3. The proposed ViMETA.

In this paper, as a special case from the proposed ViMETA, the design of the proposed ViMETA includes 2 encoders and subsequently, 2 encoder-decoder attention layers are included in the decoder. The first encoder is equipped with a self-attention module, which derives its inputs Query Q, Keys K, and Values V from the concepts vector, and MLP block comes next. The MLP block includes two layers with a Gaussian Error Linear Units (GELU) non-linearity. However, the second encoder is equipped with a self-attention module, which derives its inputs Query Q, Keys K, and Values V from the image's feature vector, and also an MLP block comes next.

For the proposed Ar-CM-ViMETA model, the concept vectors, derived from the concept modeling, are fed to the first encoder of the proposed ViMETA and the

CLIP image's embeddings are fed to the second encoder of the proposed ViMETA. After that, the outputs of the two encoders are fed to the proposed ViMETA decoder in order to produce the description word by word based on the encoded inputs and the embeddings of the partial captions. The output of the first encoder is passed to the decoder's first multi-head attention module and the output of the second encoder is passed to the decoder's second multi-head attention module.

4. Experimental Work and Results

In this paper, Flickr8K dataset [12] with Arabic captions, published by Eljundi *et al*. [8], has been used for the performance evaluation of the proposed model. Arabic Flickr8K dataset includes 8000 image. Each image has 3 Arabic captions translated from English captions through google Application Programming Interface (API) translation then the translated descriptions have been validated through professionals in Arabic translation.

We follow the guidelines for pre-processing Arabic caption texts which includes adding spaces after "و", removing the identification tool "ال", removing punctuation symbols, removing single character words, and applying split on white spaces. Number of words in the vocabulary, including the "start" and "end" marks, is 10435.

The training for the proposed model is performed for the purpose of reducing the cross-entropy loss. Given the proposed captioning model of parameters θ and the sequence of the ground truth $y_{1:T}$, the cross-entropy loss is calculated as follows:

$$L(\theta) = -\sum_{t=1}^{T} \log\left(p_\theta(y_t|y_{1:t-1})\right) \quad (1)$$

The concept modeling technique is modified by selecting an embedding model that can work on 50+ languages, i.e. "clip-ViT-B-32-multilingual-v1" [14], and the resulting number of concepts was 28 concepts. The embedding dimension is 512 and the number of the transformer layers and heads are 3 and 8, respectively. Adam optimizer [16] is used with learning rate 2e^(-5). Training of the models run for 20 epochs. Beam search is used in the testing phase with a beam size of 2.

The quantitative results have been obtained using the standard metrics BLEU (B1; B2; B3; B4) [23], and ROUGE (R) [18]. Table 1 compares the quantitative results of the state-of-the-art techniques and the quantitative results of the proposed Ar-CM-ViMETA model. The table includes the results of the translated English descriptions to Arabic presented by Eljundi *et al*. [8]. Arabic captioning approach proposed by Eljundi *et al*. [8] is also included. In addition, three other models proposed by Emami *et al*. [9] have been included in the table which made use of AraBERT [3] and GigaBERT [17] with different batch sizes.

Table 1. Comparison between results of the state-of-the-art techniques and the proposed Ar-CM-ViMETA model.

Model	B1	B2	B3	B4	R
Translating English Captions [8]	26.5	11.6	4.5	1.9	-
AIC by ElJundi *et al.* [8]	33.2	19.3	10.5	5.7	-
GigaBERT32 [9]	38.6	24.1	14.4	8.27	33.1
AraBERT256 [9]	38.7	24.4	15.1	**9.3**	33.4
AraBERT32 [9]	39.1	24.6	15.0	9.2	33.1
Proposed Ar-CM-ViMETA	**40.1**	**25**	**15.26**	8.81	**34.1**

The results presented in Table 1 and Figure 4 indicates that the proposed Ar-CM-ViMETA model outperforms the state-of-the-art methods with respect to the standard evaluation metrics (B1; B2; B3; R).

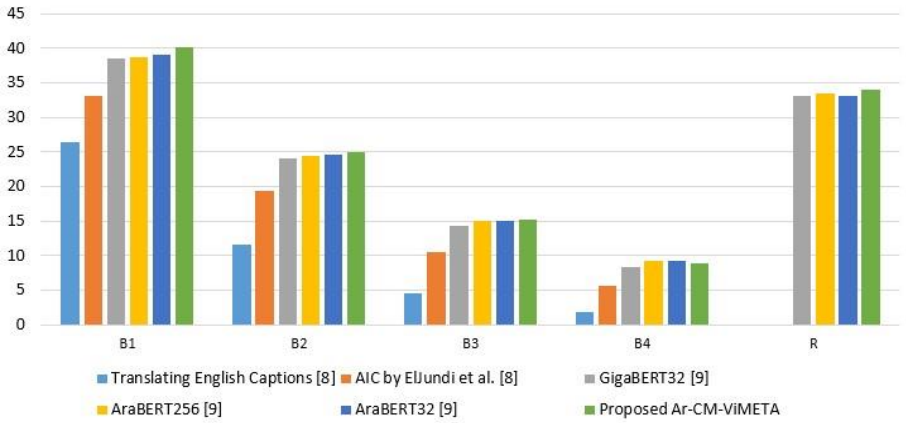


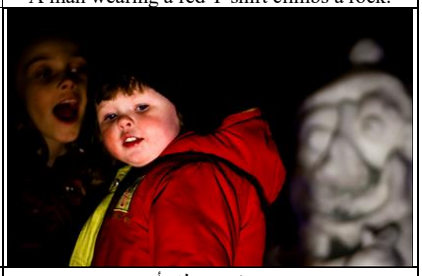
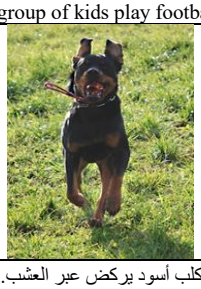
Figure 4. Bar chart for the results of the proposed Ar-CM-ViMETA model in comparison with the related work.

Table 2 includes the qualitative results of the proposed Ar-CM-ViMETA model. The proposed Ar-CM-ViMETA model shows better and more expressive descriptions in comparison with the ground truth captions. The proposed Ar-CM-ViMETA model produced captions for the images of the dog, the kid in a red jacket and the old man which were more accurate and expressive than the ground truth by depicting the image's background and the image's attributes. In addition, the proposed Ar-CM-ViMETA model was able to construct Arabic sentence which is better than the ground truth caption by explaining that the man in the

second image is "wearing" the red shirt. The proposed method results in better captions compared to the related works because of the rich semantics included in the concept vectors.

The proposed Ar-CM-ViMETA model outperforms the related work as a result of incorporating the concept modeling technique which has the ability of capturing the image contexts. These contexts have been encoded using the proposed ViMETA and fed to the decoder with the image's feature to consider it while predicting the image caption. The presented good results indicate the capability of the proposed Ar-CM-ViMETA model of being effective and expressive in terms of the produced captions.

Table 2. Qualitative results of the proposed Ar-CM-ViMETA model. Sample of the produced captions using the proposed Ar-CM-ViMETA model compared to the ground truth descriptions.

Input image		
Ground truth caption	صبى صغير يلعب فى لعبة نفخ. "Little boy playing in an inflatable game."	رجل فى قميص أحمر يحاول تسلق صخرة. "A man in a red shirt is trying to climb a rock."
Proposed Ar-CM-ViMETA	صبى صغير يلعب فى زحليقة. "Little boy playing on a slide."	رجل يرتدى قميصا احمر يتسلق صخرة. "A man wearing a red T-shirt climbs a rock."
Input image		
Ground truth caption	أطفال يلعبون كرة القدم فى حقل. "Kids playing football in a field."	صبى يرتدى معطف أحمر. "A boy in a red coat."
Proposed Ar-CM-ViMETA	مجموعة من الأطفال يلعبون كرة القدم. "A group of kids play football."	طفل يرتدى سترة حمراء ينظر الى الكاميرا. "A kid in a red jacket looks at the camera."
Input image		
Ground truth caption	كلب أسود يركض عبر العشب. "Black dog running across the grass."	رجل يرتدى قبعة سوداء يجلس على مقعد فى حديقة. "A man in a black hat sits on a park bench."
Proposed Ar-CM-ViMETA	كلب أسود يركض عبر العشب. "Black dog running across the grass."	رجل يرتدى قبعة سوداء يجلس على مقعد فى حديقة. "A man in a black hat sits on a park bench."

5. Conclusions

Novel concept-based model is proposed for AIC in this paper. Concept modeling technique is applied to extract set of concept vectors and the images' embeddings. A novel ViMETA is proposed for handling the multiple outputs of the concept modeling technique while generating the image's caption. The proposed model has been compared quantitatively with the state-of-the-art techniques. The proposed model enhanced the results, regarding the standard metrics, relative to the state-of-the-art AIC methods. In addition, the proposed model has been compared qualitatively with the ground truth captions and produced more expressive captions. The proposed model can achieve high performance in case of experimenting it on a larger dataset. So, as future research, larger dataset with Arabic descriptions can be made available for public users and so additional experiments can be implemented for the proposed model with this larger dataset. In addition, further pre-processing for the Arabic captions may help in resulting better performance.

References

[1] Afyouni I., Azhar I., and Elnagar A., "AraCap: A Hybrid Deep Learning Architecture for Arabic Image Captioning," *Procedia Computer Science*, vol. 189, pp. 382-389, 2021. https://doi.org/10.1016/j.procs.2021.05.108

[2] Al-Muzaini H., Al-Yahya T., and Benhidour H., "Automatic Arabic Image Captioning Using

RNN-LSTM-based Language Model and CNN," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, pp. 67-73, 2018. DOI:10.14569/IJACSA.2018.090610

[3] Antoun W., Baly F., and Hajj H., "AraBERT: Transformer-based Model for Arabic Language Understanding," *in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, pp. 9-15, 2020. https://aclanthology.org/2020.osact-1.2

[4] Bahdanau D., Cho K., and Bengio Y., "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv Preprint*, vol. arXiv:1409.0473, pp. 1-16, 2014. https://doi.org/10.48550/arXiv.1409.0473

[5] Bangalore M., Bharathi S., and Ashwin M., "Classification of Breast Cancer using Ensemble Filter Feature Selection with Triplet Attention Based Efficient Net Classifier," *The International Arab Journal of Information Technology*, vol. 21, no. 1, pp. 17-31, 2024. DOI: 10.34028/iajit/21/1/2

[6] Dash S., Acharya S., Pakray P., Das R., and Gelbukh A., "Topic-based Image Caption Generation," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3025-3034, 2020. https://link.springer.com/article/10.1007/s13369-019-04262-2

[7] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., and Dehghani M., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *in Proceedings of the 9th International Conference on Learning Representations*, Austria, pp. 1-21, 2021. https://openreview.net/forum?id=YicbFdNTTy

[8] ElJundi O., Dhaybi M., Mokadam K., Hajj H., and Asmar D., "Resources and End-to-End Neural Network Models for Arabic image Captioning," *in Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SciTePress*, Valletta, pp. 233-241, 2020. DOI:10.5220/0008881202330241

[9] Emami J., Nugues P., Elnagar A., and Afyouni I., "Arabic Image Captioning using Pre-training of Deep Bidirectional Transformers," *in Proceedings of the 15th International Conference on Natural Language Generation*, Waterville, pp. 40-51, 2022. https://aclanthology.org/2022.inlg-main

[10] Grootendorst M., "BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure," *arXiv Preprint*, vol. arXiv:2203.05794, pp. 1-10, 2022. http://arxiv.org/abs/2203.05794

[11] Grootendorst M., https://github.com/MaartenGr/Concept, Last Visited, 2024.

[12] Hodosh M., Young P., and Hockenmaier J., "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853-899, 2013. https://doi.org/10.1613/jair.3994

[13] Hossain M., Sohel F., Shiratuddin M., and Laga H., "A Comprehensive Survey of Deep Learning for Image Captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1-36, 2019. https://doi.org/10.1145/3295748

[14] HuggingFace, https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1, Last Visited, 2024.

[15] Jindal V., "Generating Image Captions in Arabic Using Root-Word Based Recurrent Neural Networks and Deep Neural Networks," *in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, New Orleans, pp. 144-151, 2018. https://aclanthology.org/N18-4020

[16] Kingma D. and Ba J., "Adam: A Method for Stochastic Optimization," *in Proceedings of the International Conference on Learning Representations*, San Diego, pp. 1-15, 2016. https://doi.org/10.48550/arXiv.1412.6980

[17] Lan W., Chen Y., Xu W., and Ritter A., "An Empirical Study of Pre-trained Transformers for Arabic Information Extraction," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Maine, pp. 4727-4734, 2020. https://aclanthology.org/2020.emnlp-main.382

[18] Lin C., "ROUGE: A Package for Automatic Evaluation of Summaries," *in Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, pp. 74-81, 2004. https://typeset.io/papers/rouge-a-package-for-automatic-evaluation-of-summaries-2tymbd14i8

[19] Lin T., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollar P., and Zitnick C., "LNCS 8693-Microsoft COCO: Common Objects in Context," *in Proceedings of the Computer Vision-ECCV 13th European Conference*, Zurich, pp. 740-755, 2014. https://doi.org/10.1007/978-3-319-10602-1_48

[20] Littell P., Lo C., Larkin S., and Stewart D., "Multi-Source Transformer for Kazakh-Russian-English Neural Machine Translation," *in Proceedings of the 4th Conference on Machine Translation*, Florence, pp. 267-274, 2019. https://aclanthology.org/W19-5326

[21] Mualla R. and Alkheir J., "Development of an Arabic Image Description System," *International Journal of Computer Science Trends and Technology*, vol. 6, no. 3, pp. 205-213, 2018. https://www.ijcstjournal.org/volume-6/issue-3/IJCST-V6I3P27.pdf

[22] Osman A., Shalaby M., Soliman M., and Elsayed K., "A Survey on Attention-based Models for Image Captioning," *International Journal of*

Advanced Computer Science and Applications, vol. 14, no. 2, pp. 403-412, 2023. DOI:10.14569/IJACSA.2023.0140249

[23] Papineni K., Roukos S., Ward T., and Zhu W., "BLEU: A Method for Automatic Evaluation of Machine Translation," *in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 311-318, 2002. https://aclanthology.org/P02-1040.pdf

[24] Radford A., Kim J., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., and Sutskever I., "Learning Transferable Visual Models from Natural Language Supervision," *in Proceedings of the 38th International Conference on Machine Learning*, Virtual, pp. 8748-8763, 2021. http://arxiv.org/abs/2103.00020

[25] Rikters M. and Nakazawa T., "Revisiting Context Choices for Context-aware Machine Translation," *arXiv Preprint*, vol. arXiv:2109.02995, pp. 1-6, 2021. https://doi.org/10.48550/arXiv.2109.02995

[26] Shin J. and Lee J., "Multi-Encoder Transformer Network for Automatic Post-Editing," *in Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*, Brussels, pp. 840-845, 2018. https://www.statmt.org/wmt18/pdf/WMT098.pdf

[27] Shin Y., "Multi-Encoder Transformer for Korean Abstractive Text Summarization," *IEEE Access*, vol. 11, pp. 48768-48782, 2023. DOI:10.1109/ACCESS.2023.3277754

[28] Stefanini M., Cornia M., Baraldi L., Cascianelli S., Fiameni G., and Cucchiara R., "From Show to Tell: A Survey on Deep Learning-based Image Captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 539-559, 2023. DOI:10.1109/TPAMI.2022.3148210

[29] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., and Kaiser L., "Attention is all you Need," *in Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, pp. 6000-6010, 2017. https://dl.acm.org/doi/10.5555/3295222.3295349

[30] Zhu Z., Xue Z., and Yuan Z., "Topic-Guided Attention for Image Captioning," *in Proceedings of the 25th IEEE International Conference on Image Processing*, Athens, pp. 2615-2619, 2018. DOI:10.1109/ICIP.2018.8451083

Asmaa Osman is an assistant lecturer in the information technology department at faculty of computers and artificial intelligence, Cairo University. She received the BSc. and MSc. degrees in information technology from Cairo University, Egypt. Her research interests are Computer vision, Pattern recognition, Image processing, and machine and Deep Learning.

Mohamed Shalaby is an associate professor in the Mechanical Engineering program at the School of Engineering and Applied Sciences (EAS), Nile University. He is an associate professor in the fields of artificial intelligence, information technology and robotics. He received the BSc. and MSc. degrees in computer engineering from Cairo University, Egypt, and the Ph.D. degree in electrical and computer engineering from Concordia University, Canada in 2012. Recently, He is assigned to be the EAS Quality assurance unit Manager. Dr. Shalaby has been an associate editor of the Egyptian Informatics Journal, Elsevier, since 2015. He is also an IEEE member of the computational intelligence society and IEEE SA (Standards Association). In addition, Dr. Shalaby has been a founding member of the multi-disciplinary Smart Engineering Systems Research Center (SESC) at Nile University, Giza, Egypt.

Mona Soliman is an Associate Professor at the Information Technology Department, faculty of computers and artificial intelligence, Cairo University. She received her MSc. and PhD. degree in information technology from the Faculty of Computers and information Cairo University, 2006 and 2015 respectively. She is a member of the Scientific Research Group in Egypt (SRGE). Her research interests are Pattern recognition, Image processing, machine learning, Deep Learning, Computational intelligence, medical image analysis, and Optimization techniques.

Khaled Elsayed is a Professor in the Computational Sciences and Artificial Intelligence School, Zewail City. In addition, He is a Senior IT and Business Consultant for Many software houses and enterprises. His research interests include Machine and Deep Learning, Data Science, Computer Vision, NLP, and Image Processing.