Generating an Image from Arabic Text using Generative Network of Fine-Grained Visual Descriptions

Sara Maher[1*], Mohamed Loey[2]
[1]Mathematics Dept., Faculty of Science, Benha University, Benha, Egypt
[2]Computer Science Dept., Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

Abstract:
Converting text representations in natural language into pictures is a complex computer vision issue and has many practical applications. Text-image does not vary from issues with language translation. In the same way similar semantics can be encoded in two different languages, images and text are two different languages to encode related information. None the less, these problems are totally different because text-image or image-text conversions are highly multimodal problems. In this paper, we propose our model for Arabic text description that allows multi-stage, attention-driven for refinement for fine-grained Arabic text-to-image generation. With a modern attentional generative network, The Attentional model enables fine-grained information to be synthesized in different image sub-regions by paying attention to the relevant terms in the definition of the natural Arabic language. We train the model from scratch to Modified-Arabic dataset. A word level fine-grained image-text matching loss computed by our Proposed-Method is the significant term in our Network. Two key neural networks that map sub-regions of the picture and Arabic words of the sentence to a common semantic space are learned by our Proposed-Method. On the Arabic text encoder and image encoder, our model achieves good efficiency, characterized by ease and accuracy in describing the images on the Caltech-UCSD Birds 200-2011 dataset.

Keywords:
Machine Learning, Deep Learning, Generative Adversarial Networks, Recurrent Neural Network, Natural Language Processing, Text Analysis.

1. Introduction

In image comprehension, the most challenging major difficulty is to apply natural language concepts correctly to the visual content of images. Notable progress has been made in learning visual-semantic embeddings for the English and Arabic languages in recent years.

These methods have used a large image and text datasets in addition to the progress in deep neural networks for image and language modelling, which already provides powerful new applications such as the image to text [1] and text to image [2].

Manuscript received January 5, 2025
Manuscript revised January 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.1.22

Fig 1. Example results of the Proposed-Method. The second row show the top-4 most attended words in the text descriptions.

Despite these advances, the problem of relating images and text is still far from solve, especially in Arabic texts perhaps because of the scarcity of large and high-quality training data. For example, in the private bird database CUB 200 2011 [3] that we have worked on does not contain a copy of the texts in Arabic that describe image but we worked to solve this problem and we make a relation between image and our Arabic text (see Fig 1).

We assume that higher-capacity text models are required in order to close the performance gap between text embeddings and human-annotated attributes for fine-grained visual discernment. However, for each fine-grained category, more advanced text models would require more qualified data, precisely aligned images and several visual descriptions per image. These descriptions would assist to a word level fine-grained image-text matching by our Proposed-Method.

2. Related Work

Recently, one of major advances in deep convolutional networks and recurrent neural networks that have driven rapid progress in general-purpose visual recognition on large-scale benchmarks such as ImageNet [5]. Trendy image and video captioning models work on generating natural language descriptions.

The approach of Reed et al. [6] to pre-train a text encoder. It is a character level CNN-RNN model that maps text descriptions to the common feature space of images by learning a correspondence function between texts with images. The encoder places the images and the captions in a common embedding space, such as images and descriptions that fit vectors with a high internal product. For this mapping, a Convolutional Neural Network (CNN) processes the images, and a hybrid Convolutional-Recurrent Neural Network (RNN) transforms the text descriptions (Fig 2).

A combined alternative is Skip-Thought Vectors [7] which is a clear language-based model. The model puts sentences with similar syntax and semantics to similar vectors. Nevertheless, the char-CNN-RNN encoder is better than the one used before that is suited for vision tasks as it uses the corresponding images of the descriptions as well.

The embeddings are similar to the convolutional features of the images they correspond to, which makes them visually discriminative. This property reflects in a better performance when the embeddings are employed inside convolutional networks. These models use LSTMs [8] for modelling captions at word level and focus on generating general high-level visual descriptions of a scene. M. Schuster and K. K. Paliwal.[9] built the bi-directional Long Short-Term Memory (LSTM) that extracts semantic vectors from the text description which is used in our model as text encoder, the encoder maps the images and the captions to a common embedding space such that images and descriptions which match are mapped to vectors with a high inner product, and we use a Convolutional Neural Network (CNN) that maps images to semantic vectors, specifically, our image encoder is built upon the Inception-v3 model [10] pretrained on ImageNet.

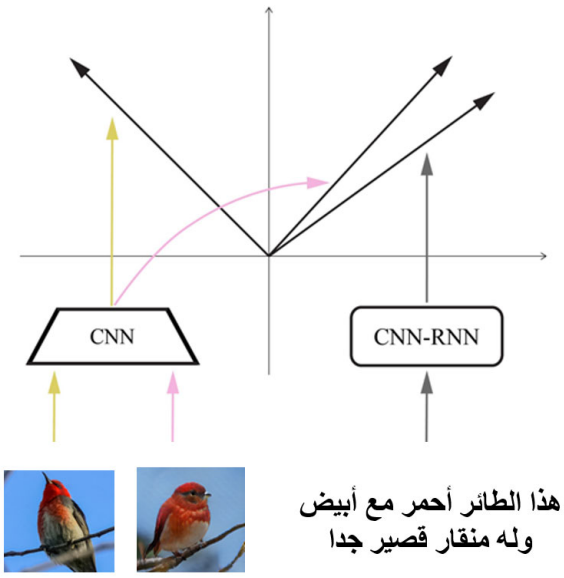

Fig 2. The char-CNN-RNN encoder maps images to a common embedding space. Images and descriptions which match are closer to each other. Here the embeddingspaceis$R^2$to make visualisation easier. In practice, the pre-processed descriptions are in R^{1024}.

The attention mechanism has recently become an integral part of sequence transduction models. It has been successfully used in image question answering [11], modelling multi-level dependencies in image captioning [12, 13], and machine translation [14]. Vaswani et al. [15] also showed that machine translation models could achieve state of-the-art results by only using an attention model.

3. Methodology

Two key neural networks that map sub-regions of the picture and phrase words to a common semantic space are learned by our Proposed-Method, so that the image-text similarity is calculated at the word level to determine a fine-grained loss for image generation.

As shown in Fig.3, we illustrate how our Proposed-Method model works. It starts from text description input until to measure the image-text similarity at the word level to calculate a fine-grained loss for image generation by the loss. We explain each step of our Proposed-Method in details in the following sections.

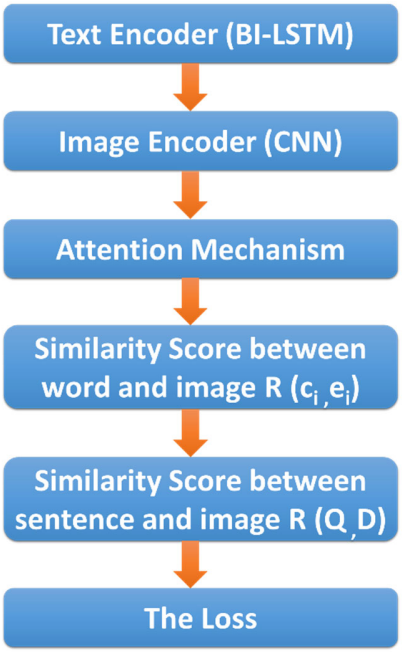

Fig 3. The Proposed-Method architecture that provides the generative network with a fine-grained image-text matching.

3.1. The text encoder

The deep neural language models that we use to depict fine-grained visual descriptions are described in this section.

A. Recurrent Neural Networks (RNN):

A group of multiple neurons in each layer is the Artificial Neural Network (ANN). ANN is also referred to as a Feed-RNN that has a recurring relation to the hidden state. This looping limit ensures that sequential data is captured in the input information.
A looping constraint on the hidden layer of ANN turns to RNN

RNN's provide a very elegant way of dealing with (time) sequential data that embodies correlations between data points that are close in the sequence.

A main RNN architecture can make use of all available input information up to the present time frame (i.e.,) to predict how much of this information is captured by a particular RNN depends on its structure and the training algorithm.

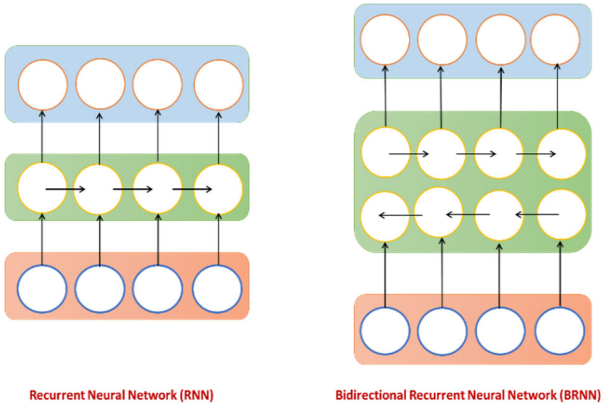

Fig 4. The difference between an RNN and an BRNN.

B. Bidirectional Recurrent Neural Networks

A bidirectional recurrent neural network (BRNN) solves the limitations of a regular RNN outlined in the previous section, that can be trained using all available input information in the past and future of a specific time frame. The difference is shown in Fig. 4.

A bi-directional Long Short-Term Memory (LSTM) that extracts semantic vectors from the description of the text is our text encoder. Each word corresponds to two hidden states in the bi-directional LSTM, one for each direction.

In order to reflect the semantic sense of a word, we concatenate its two hidden states. The feature matrix of all words is indicated by ($e \in \mathbb{R}^{D \times T}$). Its ($i^{th}$)column ($e_i$)is the feature vector for the i^{th} word. D is the dimension of the word vector and T is the number of words. Meanwhile, the last hidden states of the bi-directional LSTM are concatenated to be the global sentence vector, denoted by ($\overline{e} \in \mathbb{R}^{D}$).

3.2. The image encoder

A. convolutional neural network

CNNs are well built image processing, artificial intelligence (AI) that use deep learning to implement both generative and descriptive tasks, often using machine vison that includes image and video identification, along with recommender systems and natural language processing (NLP).

B. Inception-v3

Inception-v3 is a neural convolutional network which is trained from the ImageNet database on more than a million images. The network is 48 layers deep and can classify images, such as keyboard, mouse, table, pencil, and several animals, into 1000 object categories. Finally,

for a wide range of images, the network has learned rich performance characteristics. The network has a resolution of 299-by-299 image input.

A Convolutional Neural Network (CNN) that maps images to semantic vectors is our image encoder. CNN's middle layers learn local characteristics of different sub-regions of the image, while the following layers learn global image characteristics. More precisely, our image encoder is based on the ImageNet-pretrained Inception-v3 model. The input image is first rescaled to 299 to 299 pixels. And then, we extract the local feature matrix $f \in \mathbb{R}^{768 \times 289}$ (reshaped from 768×17×17) from the "mixed_6e" layer of Inception-v3. Each column of f is the feature vector of a sub-region of the image. 768 is the dimension of the local feature vector, and 289 is the number of sub-regions in the image. Meantime, the global feature vector $\overline{f} \in \mathbb{R}^{2048}$ is extracted from the last average pooling layer of Inception-v3. Finally, we convert the image features to a common semantic space of text features by adding a perceptron layer:

$$v = D\, f\ , \qquad \overline{v} = \overline{D}\,\overline{f} \quad \forall\, \overline{f} \in \mathbb{R}^{2048} \tag{1}$$

where $v \in \mathbb{R}^{D \times 289}$ and its i[th] column v_i is the visual feature vector for the i[th] sub-region of the image; and $\overline{v} \in \mathbb{R}^{D}$ is the global vector for the whole image. D is the dimension of the multimodal (i.e., image and text modalities) feature space. For more effectiveness, all parameters in layers built from the Inception-v3 model are fixed, and the parameters in newly added layers are jointly learned with the others of the network.

3.3. Attention mechanism

Both image and text contain richer data, but they remain in heterogeneous modalities. The designed model for cross-modal retrieval not only needs to learn the image and text characteristics to indicate their respective content, but also a test for cross-modal similarity calculation by matching information retrieval within the same modality. Attention mechanism attempts to take the correspondences between the detected visual objects and the textual items (words or phrases).

The attention-driven image-text matching score designed to estimate the matching between the image and the text of an image-sentence pair based on an attention model.

For all possible pairs of words in the sentence and sub-regions in the image, we first calculate the similarity matrix by

$$= e^{s^T v}\,, \tag{2}$$

where $s \in \mathbb{R}^{T \times 289}$ and $s_{i,j}$ is the dot-product similarity between the i[th] word of the sentence and the j[th] sub-region of the image. We find that it is beneficial to normalize the similarity matrix as follows

$$\overline{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{K=0}^{T-1} \exp(s_{k,j})} \tag{3}$$

Then we construct an attention model for each term to determine an area background vector (query). The region-context vector ci is a dynamic representation of the sub-regions of the image connected to the phrase word i[th]. It is computed as the weighted sum over all regional visual vectors i.e.,

$$c_i = \sum_{j=0}^{288} \alpha_j v_j\ , \quad where\ \alpha_j = \frac{exp(\gamma_1 \overline{s}_{i,j})}{\sum_{K=0}^{288} \exp(\gamma_1\, \overline{s}_{i,k})} \tag{4}$$

Here, γ1 is a factor deciding how much attention is paid to the characteristics of its related sub-regions when the region-context vector is computed for a name.

Finally, we define the relevance between the i[th] word and the image using the cosine similarity between c_i and e_i, i.e., $R(c_i, e_i) = (c_i^T e_i)/(\| c_i \| \| e_i \|)$. Inspired by the minimum classification error formulation in speech recognition (see, e.g.,[16, 17]), the attention-driven image-text matching score between the entire image (Q) and the whole text description (D) is defined as

$$R(Q,D) = \log\left(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i))\right)^{\frac{1}{\gamma_2}} \tag{5}$$

where γ2 is a factor that determines how much to magnify the importance of the most relevant word to-region-context pair.

When $\gamma_2 \longrightarrow \infty$, $R(Q,D)$ approximates to $max_{i=1}^{T-1} R(c_i, e_i)$.

3.4. The Loss

It is designed to learn the attention model in a semi-supervised way, which the only supervision is the matching between whole images and all sentences (a sequence of words).
In other word, we maximize the similarity score between the images and their corresponding text descriptions (ground truth), i.e.,

$$\mathcal{L}_S = -\sum_{i=1}^{M} \log P(D_i|Q_i) \quad (6)$$

M is the number of training pairs.

The Proposed-Method is pre-trained by minimizing $\mathcal{L}_S$ using real image-text pairs. Since the size of images for pre-training. Our Proposed-Method is not limited by the size of images that can be generated, real images of size 299×299 are utilized. in addition, the pre-trained text encoder in the Proposed-Method provides fine-grained visual word vectors learned from image-text paired data.

Conventional word vectors pre-trained on clear text data are a lot not visually-discriminative, e.g., word vectors of different colours, such as white, red, black, etc., are often grouped together in the vector space, due to the reduction of grounding them to the actual visual signals.

4. Experiments

4.1. Dataset

There are 11,788 photos of 200 bird species in the Caltech-UCSD Birds-200-2011 Dataset. A Wikipedia article is connected with each species and arranged by scientific classification (order, family, genus, species). An online field guide11 was used to acquire the list of species names. Images were collected and then filtered by showing each image to multiple users of Mechanical Turk [18] using Flickr image search. The photos are divided into 8,855 training and 2,933 groups of disjoint research. Only photos are included in these datasets, however no details. Table (1) lists the statistics of datasets.

Table (1) Statistics of datasets.

Dataset	CUB [3]	
Number of Samples	train	test
	8,855	2,933
Number of Categories	200	
Number of Total	11,788	
Number of caption/image	10	

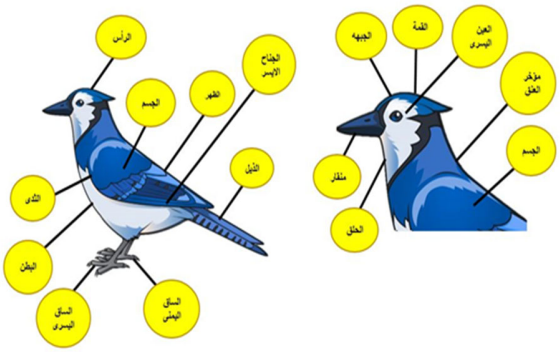

Fig 5. The 15-part location labels collected for each image in dataset.

Nevertheless, as the same way of Reed et al. [5] Using Amazon Mechanical Turk to compile captions. There are five explanations in each of the pictures. The deep network models win, but with more training captions. So, we train and test our dataset at ten Arabic sentence per image. They are at least ten words in length, the history is not identified, and the bird species are not listed. we determined a total of 15 Part Locations (see Fig. 5) to help us improve the training process.

Fig 6. Samples generated by the Proposed-Method that provides the generative network with a fine-grained image and text matching

4.2. Results

In this section, we investigated the performance of our Proposed-Method. It has made great progress in text encoder and image encoder.

Table 2. Dictionary of Arabic words and it is the first result from our Proposed-Method.

Number	Word
1	طائر
2	ابيض
3	طويل
4	المنقار
⋮	⋮
10989	الجناح

It is very necessary after we prepare our Arabic captions by pre-processing the words and then tokenizing, since no Computer Algorithm understands text, numbers are all they understand. So, with tokenization, with its unique tokens in space, we convert each unique word and select alphanumeric character sequences as tokens and drop all else. Finally, as shown in table (2), we build a dictionary of Arabic words and their numbers.

To generate realistic images with multiple levels (i.e., word level and sentence level) of conditions, the last objective function of the attentional generative network is defined as

$$\mathcal{L} = \mathcal{L}_{model} + \lambda \mathcal{L}_S \quad (7)$$

Here, λ is a hyperparameter to balance the two terms of Eq. (7). The first term is the Model loss that can be jointly approximates conditional or unconditional distributions.

4.3. The Loss

To test the proposed $\mathcal{L}_S$, we adjust the value of λ (see Eq. (7)). We observe that a larger λ leads to significantly higher performance on the dataset, the lower this value we do not get a good result. This comparison demonstrates

that properly increasing the weight of $\mathcal{L}_S$ helps to generate higher results. The reason is that the proposed fine-grained image-text matching loss $\mathcal{L}_S$ gives additional supervision (i.e., matching information of word level) for training the generator. It notes that the fine-grained image-text matching failure also helps stabilize the model's training phase with additional supervision.

In addition, with comparison of non-use of attention and use of the text encoder used in [5], the model performance will be significantly reduced. which further demonstrates the effectiveness of the proposed $\mathcal{L}_S$. We will review some results to ensure effectiveness of $\mathcal{L}_S$ (see Fig, 6).

5. Conclusion and Future Work

In this paper, we developed a deep symmetric joint embedding model using deep learning, created a modified dataset of fine-grained visual descriptions from English version to the Arabic version, and evaluated several deep neural text encoders.

Compared to attributes, our text encoders achieve a competitive retrieval result and correctly match text to image. It can be used directly to create actual applications, such as creating a visually realistic text picture and text description from an image.

In both the processing of natural language and computer vision cultures, these apps remain challenging and have become an active research field. We will focus on creating images from Arabic text by using our Proposed-Method for future work.

References

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. J. a. p. a. Lee, "Generative adversarial text to image synthesis", 2016.

[2] P. M. Manwatkar and S. H. Yadav, "Text recognition from images", In: Proc. of *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pp. 1-6: IEEE. 2015.

[3] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset", 2011.

[4] T. Xu *et al.*, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks", In: Proc. of *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316-1324. 2018.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", In: Proc. of *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255: Ieee. 2009.

[6] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions", In: Proc. of *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 49-58. 2016.

[7] R. Kiros *et al.*, "Skip-thought vectors", In: Proc. of *Advances in neural information processing systems*, pp. 3294-3302. 2015.

[8] S. Hochreiter and J. J. N. c. Schmidhuber, "Long short-term memory", Vol. 9, No. 8, pp. 1735-1780, 1997.

[9] M. Schuster and K. K. J. I. t. o. S. P. Paliwal, "Bidirectional recurrent neural networks", Vol. 45, No. 11, pp. 2673-2681, 1997.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", In: Proc. of *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826. 2016.

[11] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering", In: Proc. of *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21-29. 2016.

[12] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention", In: Proc. of *International conference on machine learning*, pp. 2048-2057. 2015.

[13] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network", In: Proc. of *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6428-6436. 2017.

[14] D. Bahdanau, K. Cho, and Y. J. a. p. a. Bengio, "Neural machine translation by jointly learning to align and translate", 2014.

[15] A. Vaswani *et al.*, "Attention is all you need", In: Proc. of *Advances in neural information processing systems*, pp. 5998-6008. 2017.

[16] B.-H. Juang, W. Hou, C.-H. J. I. T. o. S. Lee, and A. processing, "Minimum classification error rate methods for speech recognition", Vol. 5, No. 3, pp. 257-265, 1997.

[17] X. He, L. Deng, and W. J. I. S. P. M. Chou, "Discriminative learning in sequential pattern recognition", Vol. 25, No. 5, pp. 14-36, 2008.

[18] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds", In: Proc. of *Advances in neural information processing systems*, pp. 2424-2432. 2010.