Benha Journal of Applied Sciences (BJAS)
Vol.(5) Issue(7) Part (2) (2020), (167-173)

print : ISSN 2356–9751
online : ISSN 2356–976x

http://bjas.bu.edu.eg

Arabic Text to Image Generation based on Generative Network of Fine-Grained Visual Descriptions

S. M. Salem[1] and M. L. Ramadan[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
E-Mail: sara.mohamed@fsc.bu.edu.eg

Abstract

Converting natural language text descriptions into images is a challenging problem in computer vision and has many practical applications. Text-image is not different from language translation problems. 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 enable to synthesize fine-grained details at different sub-regions of the image by paying attentions to the related words in the natural Arabic language description. We train the model from scratch to Modified-Arabic dataset. The important term in our Network is a word level fine-grained image-text matching loss computed by the Deep Attentional Multimodal Similarity Model (DAMSM). The DAMSM learns two main neural networks that map sub-regions of the image and Arabic words of the sentence to a common semantic space. Our model achieves strong performance on Arabic-text encoder and image encoder, it is characterized by ease and accuracy in description 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; Image Matching.

1. Introduction

The most difficult major challenge in image understanding is to correctly relate natural language concepts to the visual content of images. In recent years there has been remarkable progress in learning visual-semantic embeddings for English and Arabic language.

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].

Despite these advances, the problem of relating images and text is still far from solve, specially 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).

In order to close the performance gap between text embeddings and human-annotated attributes for fine-grained visual discernment, we suppose that higher-capacity text models are required. Nevertheless, more developed text models would require more trained data, specially aligned images and multiple visual descriptions per image for each fine-grained group.

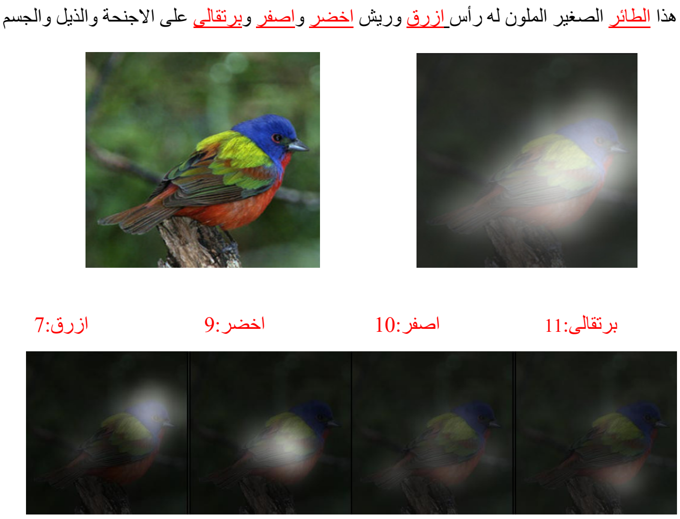

هذا الطائر الصغير الملون له رأس ازرق وريش اخضر واصفر وبرتقالي على الاجنحة والذيل والجسم
7:ازرق
9:اخضر
10:اصفر
11:برتقالى

Fig (1) Example results of the proposed DAMSM**.** The second row show the top-4 most attended words in the text descriptions.

These descriptions would assist to a word level fine-grained image-text matching by Deep Attentional Multimodal Similarity Model (DAMSM) [4].

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 puts the images and the captions to a common embedding space like that images and descriptions which match are mapped to vectors with a high inner 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.

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.

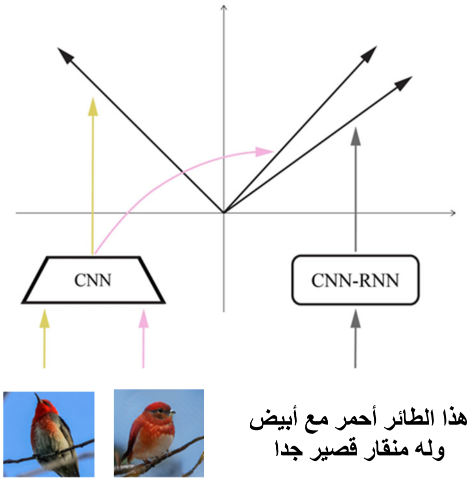


Fig (2) The char-CNN-RNN encoder maps images to a common embedding space. Images and descriptions which match are closer to each other. Here the embeddingspaceis$R^2$to make visualisation easier. In practice, the pre-processed descriptions are in R^{1024}.

3.Deep Attentional Multimodal Similarity Model

The DAMSM learns two main neural networks that map sub-regions of the image and words of the sentence to a joint semantic space, so measures the image-text similarity at the word level to calculate a fine-grained loss for image generation.

3.1. The text encoder

In this section we describe the deep neural language models that we use for representing fine-grained visual descriptions.

A. Recurrent Neural Networks(RNN):

Artificial Neural Network (ANN) is a group of multiple neurons at each layer. ANN is also known as a Feed-RNN has a recurrent connection on the hidden state.

This looping constraint ensures that sequential information is captured in the input data. The difference is shown in Fig(3).

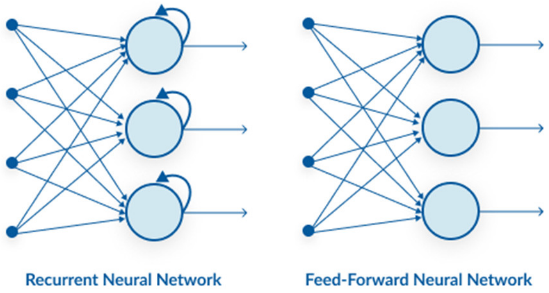


Fig (3) the difference between an RNN and an ANN.

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.

B. Bidirectional Recurrent Neural Networks

A bidirectional recurrent neural network (BRNN) solve 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.

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

Thus, we concatenate its two hidden states to represent the semantic meaning of a word. The feature matrix of all words is indicated by $e \in \mathbb{R}^{D \times T}$. Its ith column e_i is the feature vector for the ith 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 convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 48 layers deep and can classify images into 1000 object categories, such as keyboard, mouse, table, pencil, and many animals. Finally as a result, the network has learned rich feature performance for a wide range of images. The network has an image input size of 299-by-299.

Our image encoder is a Convolutional Neural Network (CNN) that maps images to semantic vectors. The middle layers of the CNN learn local features of different sub-regions of the image, whilst the following layers learn global features of the image. More specifically, our image encoder is built on the Inception-v3 model pretrained on ImageNet. We first rescale the input image to be 299×299 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:

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

where $v \in \mathbb{R}^{D \times 289}$ and its ith column v_i is the visual feature vector for the ith 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

Image and text both contain more rich information but stay in heterogeneous modalities. By matching to information retrieval within the same modality, the designed model for cross-modal retrieval not need only to learn the features for image and text to indicate their respective content but a measure for cross-modal similarity calculation.

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

We first calculate the similarity matrix for all possible pairs of words in the sentence and sub-regions in the image by

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

where $s \in \mathbb{R}^{T \times 289}$ and $s_{i,j}$ is the dot-product similarity between the ith word of the sentence and the jth 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 build an attention model to compute a region context vector for each word (query). The region-context vector ci is a dynamic representation of the image's sub-regions related to the ith word of the sentence. It is computed as the weighted sum over all regional visual vectors,
i.e.,

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

Here, γ1 is a factor that determines how much attention is paid to features of its relevant sub-regions when computing the region-context vector for a word.

Finally, we define the relevance between the ith word and the image using the cosine similarity between c_i and e_i,
i.e., $R(c_i\,, e_i) = (c_i^T\, e_i)/(\Vert\, c_i \,\Vert\Vert\, e_i\, \Vert)$. 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

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

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 $max_{i=1}^{T-1}\, R(c_i, e_i)$.

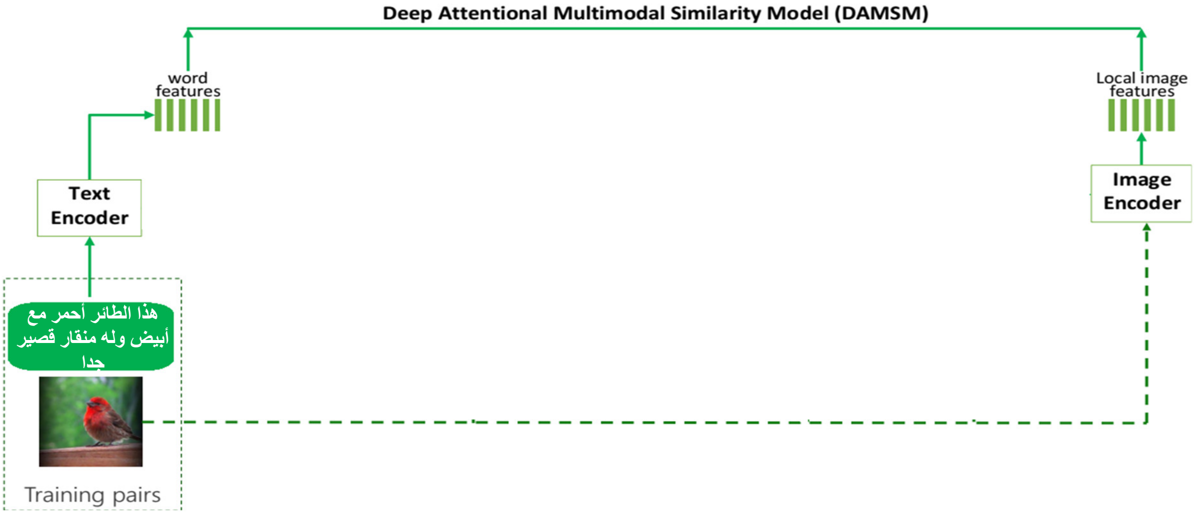
3.4. The DAMSM loss

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.,

$$\mathcal{L}_{DAMSM} = -\sum_{i=1}^{M} \log P(\,D_i|Q_i) \qquad (6)$$

□ M is the number of training pairs.

As shown in fig (4) which shows The architecture of the DAMSM. it is pre-trained by minimizing $\mathcal{L}_{DAMSM}$ using real image-text pairs. Since the size of images for pretraining DAMSM 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 DAMSM provides fine-grained visual word vectors learned from image-text paired data. Conventional word vectors pre-trained on clear text data are alot not visually-discriminative, e.g., word vectors of different colors, 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.



Fig(4) The architecture of the DAMSM, it provides the fine-grained image-text matching loss for the generative network.

4. Experiments

4.1. Dataset

The Caltech-UCSD Birds-200-2011 Dataset contains 11,788 images of 200 bird species. Each species is associated with a Wikipedia article and organized by scientific classification (order, family, genus, species). The list of species names was obtained using an online field guide1. Images were harvested using Flickr image search and then filtered by showing each image to multiple users of Mechanical Turk [18]. The images are split into 8,855 training and 2,933 disjoint test categories. These datasets include only images, but no descriptions. Table (1) lists the statistics of datasets.

Table (1) Statistics of datasets.

Dataset	CUB [3]	
	train	test
#samples	8,855	2,933
#categories	200	
#total	11,788	
caption/image	10	

Nevertheless, as the same way of Reed et al.[5] for collecting captions using Amazon Mechanical Turk. Each of the images has five descriptions. but with more training captions the deep network models win. So we train and test our dataset at ten Arabic sentence per image.

They are at least ten words in length, they do not describe the background, and they do not mention the species of the bird. we determined a total of 15 Part Locations (see Fig 5) to help us improve the training process.

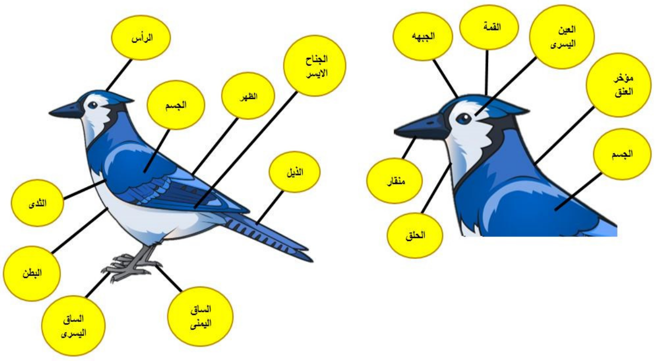


Fig (5) The 15-part location labels collected for each image in dataset.

4.2. Results

In this section, we investigated the performance of the DAMSM. It has made great progress in text encoder and image encoder. After we prepare our Arabic captions by Pre-processing the words and then tokenizing, it is very important because no Machine Algorithm understands text, all they understand is numbers. So with tokenization we convert each unique word with its unique tokens in space and picking out sequences of alphanumeric characters as tokens and drops everything else. we finally build a dictionary of Arabic words and its numbers as shown in the table (2).

Table 2. Dictionary of Arabic words and it is the first result from our model DAMSM.

Number	Word
1	طائر
2	مع
3	طول
4	الجناح
⋮	⋮
10989	مبهج

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

$$\mathcal{L} = \mathcal{L}_{model} + \lambda \mathcal{L}_{DAMSM} \tag{7}$$

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.

The DAMSM loss. To test the proposed $\mathcal{L}_{DAMSM}$, 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 $\mathcal{L}_{DAMSM}$ helps to generate higher results. The reason is that the proposed fine-grained image-text matching loss $\mathcal{L}_{DAMSM}$ gives additional supervision (i.e., matching information of word level) for training the generator. It mentions that, with extra supervision, the fine-grained image-text matching loss also helps to stabilize the training process of the model.

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 $\mathcal{L}_{DAMSM}$. We will review some results to ensure effectiveness of $\mathcal{L}_{DAMSM}$ (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.

Our text encoders achieve a competitive retrieval result compared to attributes, and correctly matching text with image. it can be directly used to build real applications such as generating a visually realistic image from text and text description generating from an image.
These applications remain challenging and also become an active research area in both natural language processing and computer vision communities.
future work, we will work on generating image from Arabic text by using DAMSM.

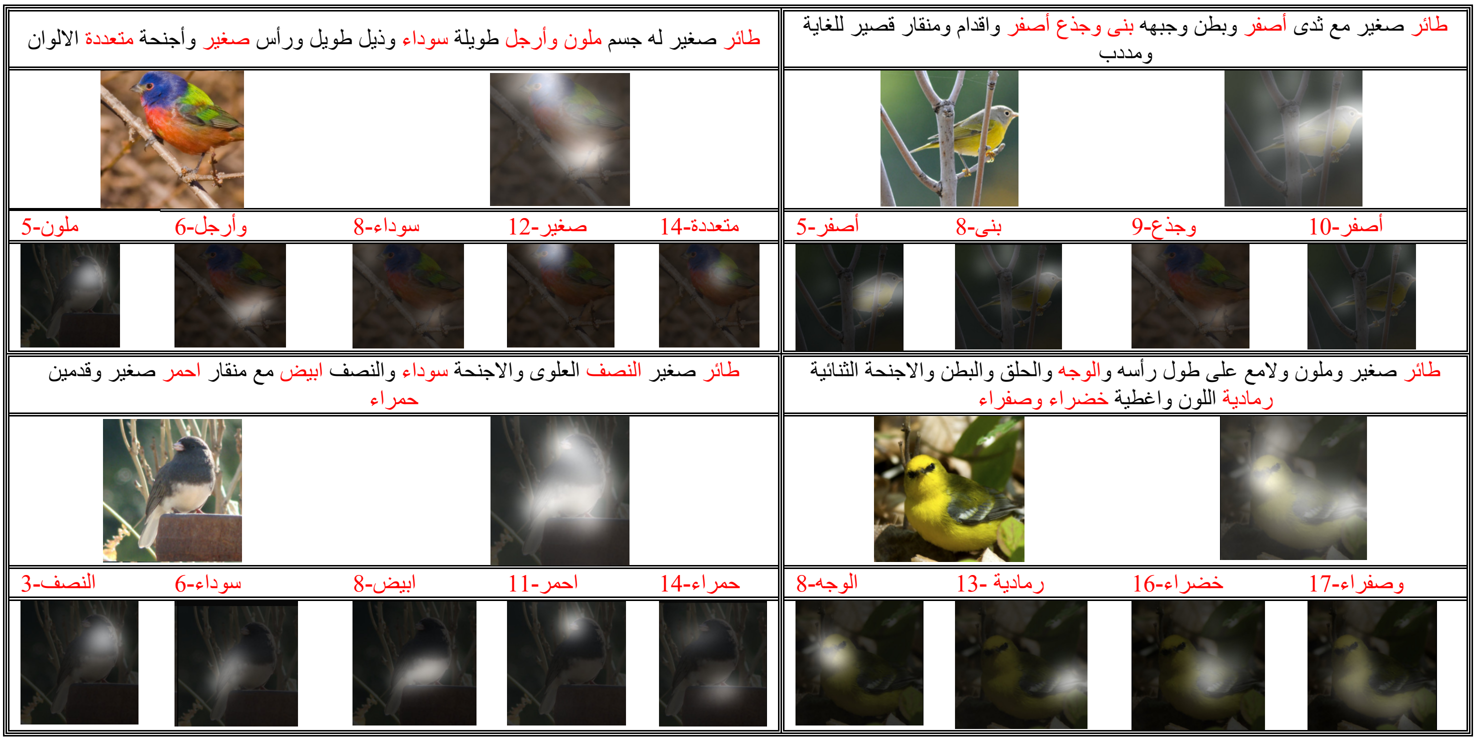

طائر صغير له جسم ملون وأرجل طويلة سوداء وذيل طويل ورأس صغير وأجنحة متعددة الالوان
5-ملون
6-وأرجل
8-سوداء
12-صغير
14-متعددة
طائر صغير مع ثدى أصفر وبطن وجبهه بنى وجذع أصفر واقدام ومنقار قصير للغاية ومدبب
5-أصفر
8-بنى
9-وجذع
10-أصفر
طائر صغير النصف العلوى والاجنحة سوداء والنصف ابيض مع منقار احمر صغير وقدمين حمراء
3-النصف
6-سوداء
8-ابيض
11-احمر
14-حمراء
طائر صغير وملون ولامع على طول رأسه والوجه والحلق والبطن والاجنحة الثنائية رمادية اللون واغطية خضراء وصفراء
8-الوجه
13- رمادية
16-خضراء
17-وصفراء

Fig (6) Samples generated by Deep Attentional Multimodal Similarity Mode

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 *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pp. 1-6, 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 *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 *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255, 2009.

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

[7] R. Kiros *et al.*, Skip-thought vectors, in *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(8), pp. 1735-1780, 1997.

[9] M. Schuster and K. K. J. I. t. o. S. P. Paliwal, Bidirectional recurrent neural networks, Vol.45(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 *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 *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 *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 *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 *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(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(5), pp. 14-36, 2008.

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