الجامعة البريطانية في دبي

The British University in Dubai

Arabic Image Captioning (AIC): Utilizing Deep Learning and Main Factors Comparison and Prioritization

التعليق على الصورباللغة العربية: توظيف التعلم العميق و مقارنة المعاملات الرئيسية و اولوياتها

by

HANI DAOUD HEJAZI

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

February 2022

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

____

Signature of the student

Abstract

Captioning of images has been a major concern for the last decade, with most of the efforts aimed at English captioning. Due to the lack of work done for Arabic, relying on translation as an alternative to creating Arabic captions will lead to accumulating errors during translation and caption prediction. When working with Arabic datasets, preprocessing is crucial, and handling Arabic morphological features such as Nunation requires additional steps. We tested 32 different variables combinations that affect caption generation, including preprocessing, deep learning techniques (LSTM and GRU), dropout, and features extraction (Inception V3, VGG16). Moreover, our results on the only publicly available Arabic Dataset outperform the best result with BLEU-1=36.5, BLEU-2=21.4, BLEU-3=12 and BLEU4=6.6. As a result of this study, we demonstrated that using Arabic preprocessing and VGG16 image features extraction enhanced Arabic caption quality, but we saw no measurable difference when using Dropout or LSTM instead of GRU.

Keywords: Deep learning; NLP; Arabic Image Captioning; Arabic Text Preprocessing; LSTM; VGG16; INCEPTION V3.

ملخص

ازداد الاهنمام بالتعليق الالي على الصور بالعقود الاخيرة, و معظم الجهود كانت منصبة على التعليق باللغة الانجليزية, ونظراً لندرة الجهود المبذولة لتطوير التعليق باللغة العربية اعتمدت بعض الدراسات على ترجمة التعليقات الالية الانجليزية بدلاً من اصدار تعليقات الية مبينية على اساس نص عربي. لكن هذا يؤدي الى تراكم الاخطاء في الترجمة مع الاخطاء في انتاج التعليقات, وهذا يقلل الدقة بشكل عام. عن التعامل مع النص العربي تعتبر عملية تهيئة النص خطوة فارقة و مؤثرة على سائر عملية انتاج النص الالي. تم اجراء 32 تجربة لاختبار تاثير مختلف قيم 4 معاملات مؤثرة في العملية, و تشمل المعاملات, التحضير المسبق للنص و تقنية النعلم العميق المستخدمة (LSTM , GRU) و نموذج الصور المستخدم لاستخراج خصائص الصور(VGG16, inception V3). وجدنا مجموعة بيانات واحدة باللغة العربية متاحة للاستخدام العام, و نم اجراء التجربة عليها و كانت افضل نتائج نم النوصل اليها هي: BLEU4=6.6 BLEU-1=36.5, BLEU-2=21.4, BLEU-3=12 وهذه النتائج تفوقت على افضل نتائج سابقة. و تم اثبات ان استخدام الطرق المحسة لمعالجة النص و نموذج الصور VGG16 قد حسن من النتائج بينما نموذج التعلم العميق لم يكن له تاثير كبير على نتائج التجارب.

Dedication

At the beginning and all times, I thank Allah Almighty for giving me the courage and persistence in my way to complete this dissertation.

I dedicate this dissertation work to my beloved wife Afnan for here patient and supported through all the way and during ups and downs to finish this work, Also my thanks are going to brothers, sisters, friends who never stopped giving advice and support this journey.

A special dedication is for my beloved sons Wesam and Osama and my wonderful daughter Aseel for their patient on my limited time during this research.

Acknowledgement

Throughout of this thesis accomplishment, I have received a big assistance and support.

First, I would like to thank my supervisor, Professor Khaled Shaalan, whose expertise and feedback was very valuable for this research. Your visionary feedback has brought my work to a much better level.

I would like to acknowledge my colleague in this journey Ahmed Khmees for his wonderful collaboration. The time we spend studying, driving together was main source of energy for completing this life project.

I would also like to thank Prof. Khaled and Dr. Sherief Abdalla, for their insightful guidance throughout my study. You provided me with the tools that I needed to choose the right direction and successfully complete my dissertation.

Table of Contents

List of Abbreviations

Abbreviations	**Description**
IC	Image Captioning
AIC	Arabic Image Captioning
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
BLEU	Bilingual Evaluation Understudy
AI	Artificial Intelligence
ML	Machine Learning
CNN	Convolutional Neural Network
VGG16	Visual Geometry Group 16 layers
MS-COCO	Microsoft - Common Objects in Context

List of Figures

List of Tables

1 Chapter one: Introduction

Visual information is one of the main sources of knowledge for human about the world around us. By accumulating the transferred knowledge to human cognition his world understanding is progressively increased. Moreover, social media has increased the number of images uploaded to the web. In June 2019, Facebook received 300 million photos a day, while Instagram received 95 million (Dustin Stout 2020). Additionally, the advent of smart devices and cameras in public places has created a challenge for automatic captioning of images which can help in image search by content using human language, as well as for video context description, all this can be utilized in many fields like image security scanning, image retrieval, learning by images for young ages, tourism, self-driving cars, video subtitle, and people with vision impairment assistance.

According to WHO (WHO 2021) there are more than 2.2 billion humans have distance or near vision impairment around the world with different vision levels, in USA there are 7,894,900 persons with vision disability (National Federation of the Blind 2019). many of them have problem in travel and hard to avoid on time road emergencies, applying IC can help them to move safely and cognition the surrounding better especially near roads

More over applying IC in social media can help in content filtration, where it can restrict access to undesirable content that can be offensive, extreme, inappropriate, risky, or age restricted. This will increase users' security and protect users from undesired images.

1.1 Problem Statement

Image Captioning (IC) considered as part of computer vision field, it involves a lot of work since it starts with detecting and identifying objects, then it relates these detected objects, and finally it translates them into human understandable text by using their language syntax and semantics. A lot of efforts were done to overcome these challenges and a good result was achieved using deep learning techniques.

Most of the work was based on western languages. As a result, language translation was applied to benefit from these models in different languages, but the results were not as good as the original language model. For example (ElJundi et al. 2020)and (Mualla & Alkheir 2018) show that building an image captioning model that generates Arabic captions outperforms an English based model with the aid of Arabic translation.

Many factors were studied to understand the effect of which on Image Captioning, like Preprocessing method, Deep learning technique, Dropout usage, and image classifier.

1.2 Research Questions

The main research questions that discussed and we try to answer in this thesis are:

- How to enhance Arabic image captioning quality using deep learning?
- How Preprocessing, deep learning method, image features, and dropout affect the Arabic image captioning quality?

1.3 Contribution

In this research, we have worked on the only publicly available dataset (ElJundi et al. 2020) and tried to enhance the Arabic Image Captioning (AIC) and used BLEU as a measure. Four factors were studded to find the most contributing ones. 32 experiments were conducted and results compared using paired t-test.

The contribution of this paper is to:

- Develop an Arabic Image Captioning model that outperforms the best results on the publicly available dataset and use the latest Arabic Image Captioning (AIC) dataset as input to the model. Analyze the results from the perspective of Arabic preprocessing and the model's performance.
- Build 32 models using different parameters: 2 Deep learning methods (LSTM, GRU) X 2 With / Without Dropout X 4 Preprocessing techniques X 2 image classifiers (VGG16, INCEPTION V3), and compare the results to show the most significant factors.
- Compare the four Arabic language preprocessing techniques and compare their effects to illustrate the importance of preprocessing for Arabic versus English, where all reviewed articles do not preprocess the text.

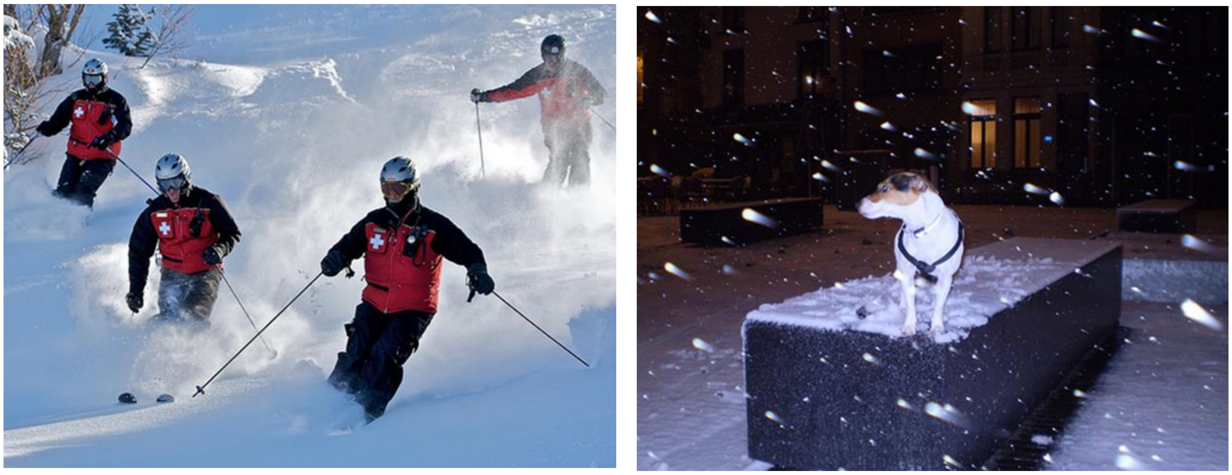
1.3.1 Dataset

One public Dataset was found for this task (ElJundi et al. 2020) based on Flickr8K each image has three Arabic captions, that translated then reviewed and edited. However, the original Flickr8K has five captions per image which can cover more human style of writing. Figure 1) illustrates two samples of this Dataset.

1.3.2 Image Features Extraction

Building CNN is common for this task, but it requires a big dataset and high processing power. An alternative way is to use a pre-trained model, as an example (ElJundi et al. 2020) used

VGG16 (Simonyan & Zisserman 2014) as a features extractor. Our work also utilized Inception V3 (Szegedy et al. 2016), which provides a well-optimized trained model that can be utilized even without pre-processing and training.



1 الناس يتزلجون على تلة مغطاة بالثلوج.

2 المتزلجين في الزي الموحد يتقدمون نزولا على منحدر ثلجي.

3 هناك أربعة متزلجين على الثلج يتزلجون على جانب التل.

1 كلب يقف على مقعد على الثلج.

2 كلب يقف على مقعد بينما الثلج يتساقط.

3 الكلب يقف على شيء بينما الثلج يسقط حوله.

Figure 1 Sample images with three captions from (ElJundi et al. 2020) *Dataset*

1.3.3 Arabic Text Preprocessing

Arabic is obviously different from English and needs preprocessing. It might have diacritic signs which affect the word’s meaning and use, but it is commonly ignored (Shoukry & Rafea 2012). Moreover, we noticed that the conjunction Waw "و" in the Arabic Dataset is attached to the next word like "ويقول" (and-he-says). As per our preprocessing rule, if the letter Waw "و" (and) appears separately, it is removed as we remove all single character occurrences. Due to this, we decided to fix the typo.

1.3.4 Models

Experiments were conducted with two deep learning algorithms (GRU and LSTM), two image classifiers, and four preprocessing methods, resulting in 32 models. They were compared based on their performance.

1.3.5 Evaluation

Bilingual evaluation understudy (BLEU) metric is used to evaluate between different language translation and image captioning accuracy. For the purpose of comparing the effects of each understudy factor, we have used BLEU-1, BLEU-2, BLUE-3, and BLUE-4.

1.4 Declaration

Part of the work was published in the article (Hejazi & Shaalan 2021) before submitting this thesis and we have the authorization to use it and build on it.

1.5 Organization of Thesis

The rest parts of the thesis are organized as the following: In the next chapter we review the related work done for both Arabic and English IC and methods used in each approach. In chapter 3, Methodology, experiment design, implementation, and Dataset are described, then in chapter 4 the results are discussed and comparisons were illustrated to show the enhancement achieved by each experimented factor, in the last chapter we give some concluding remarks based on the results achieved and some future work was proposed.**Error! Reference source not found.**

2 Chapter two:
Literature Review

Recent work on Image Captioning is reviewed for both Arabic and English. We noticed that there is a lack of Arabic image captioning datasets available for tackling this task in Arabic compared to English.

2.1 English Image Captioning

(Aneja, Deshpande & Schwing 2018) Introduced a convolution framework for image captioning consisting of four parts that begin with embedding layer for the input text, embedding for the input image, and then convolution model at the end embedding for output generation. A comparison is made against the LSTM model on the challenging MS-COCO dataset. Another experiment was done based on feed forward network that can operate over all words in parallel, and the results outperformed the baseline LSTM model.

(Wang et al. 2020) Introduced a novel method for image captioning by using visual regions relationships, graph neural network and context aware attention mechanism for caption generation, memorizing previous visual content was the competitive edge in the model. The model is trained and tested on MS-COCO and Flickr30K Dataset; the reported results showed that this model can outperform the state-of-the-art attention-based methods as per the authors.

(Tian, Zhou & Zhao 2020) Proposed new Visual Question Answering (VQA) model based on Cascading Top-Down attention (CTDA) captioning where each keyword in question is mapped to a region in the image. A good performance was demonstrated with VQA V2.0 and V1.0 datasets.

(Rennie et al. 2017) applied reinforcement learning with self-critical sequence training (SCST) with CIDEr metric as a reward. It is applied on MS-COCO dataset and the result was promising in its time.

(Anderson et al. 2018) introduced Bottom-up attention CNN by dividing the image into regions and features vector. The model was built on MS-COCO Dataset and showed a promising result.

(Wu, Hu & Mooney 2019) built a model for captioning images, which was then applied to question answering based on MS-COCO datasets.

2.2 Arabic Image Captioning

(ElJundi et al. 2020) Have built end to end model for Arabic Image Captioning (AIC) based on image features extractor VGG16 and LSTM for language model. Also introduced a new public dataset for AIC. They found that directly generating captions from an Arabic dataset yielded better results than translating captions from English datasets based on models generated from those datasets.

(Mualla & Alkheir 2018) have used a subset of Fliket8K that consists of 2000 images and their Arabic caption in Jason file. A CNN was used for image features extraction for captions using LSTM. Two models for English and Arabic captions were introduced and the results showed that Arabic based captioning from genuine Arabic dataset has better results than those derived from English-to-Arabic translation dataset.

while (Jindal & Vasu 2018) explored generating the text based on the Arabic root using CNN ImageNet and mapping each root to an image region. Then finding the best word to describe the image using root words trained on RNN. The caption is generated through a dependency tree representing the generated words and their relations. 405,000 images from newspapers with their captions as well as those provided by Fliker8K were translated by professional translators. Unfortunately, this dataset was not yet made public.

(Al-muzaini, Al-yahya & Benhidour 2018) also used two datasets: one with 5358 captions for 1176 images translated by human and the second has 150 images along with 750 captions. RNN was used. The evaluation showed promising results for a larger dataset.

1 لفتاة الصغيرة ترفس كرة القدم الزرقاء والحمراء.
2 فتاة صغيرة في ثوب لعب كرة القدم.
3 تلعب فتاة صغيرة في ثوب ملون بكرة القدم الزرقاء والحمراء.

1 A little girl in a dress playing with a soccer ball.

2 A little girl in a colorful dress is playing with a blue and red soccer ball.

3 Girls in brightly-colred clothes plays with a blue ball.

4 The young girl is kicking a blue and red soccer

Figure 2 Sample image with translated English caption result in inaccurate Arabic sentences from (ElJundi et al. 2020) Dataset

2.3 Arabic Text Preprocessing

The objective of this section is to provide a review of the various methods used for Image Captioning and to compare them with AIC research so we can identify any gaps that need to be addressed.

It is obvious that applying machine learning approach to AIC requires big data. Our study indicates that there is less research performed in AIC and this can be due to a lack of publicly available dataset for this task. Moreover, no results yet outperformed English captioning performance. The majority of work is focused on reapplying the deep learning method used in

English image captioning without considering the Arabic language and differences. As a result, we decided to examine the factors that influence Arabic image captioning.

In addition, we found one public Arabic image captioning dataset that we can use for our experiments. Using this dataset, we will choose different factors that affect the task. The purpose is to identify factors that can outperform these studies' results.

2.4 Deep learning Method:

Recurrent Neural Network (RNN) is type of deep learning networks with help of the internal state and back propagation it accepts input of arbitrary length sequence and fixed output length such as handwriting to text transformation, speech recognition, or Text generation (Image captioning, translation, summarization...).

But it in its original architecture it suffers from the vanishing gradient where in the back propagation the values of the gradient become smaller after each cell, which reduce the effect of the earlier cells, this can appear if we try to process a paragraph using RNN the more moving forward in text the less the predictions related to the paragraph starting text, in other words the network has a short memory and lack of relating very early states.

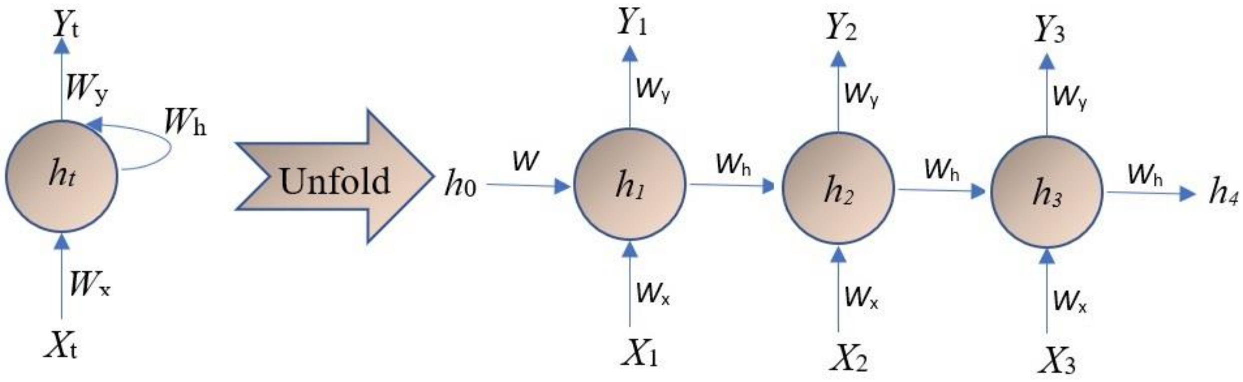

Y_t
W_y
W_h
h_t
W_x
X_t
Unfold
h_0
W
Y_1 W_y h_1 W_x X_1
W_h
Y_2 W_y h_2 W_x X_2
W_h
Y_3 W_y h_3 W_x X_3
W_h
h_4

Figure 3 RNN network and unfolding for input sequence on length three illustration

Figure 3) demonstrates RNN network where input vector is accepted and used to change update the current hidden state and both input and previous hidden state used for output prediction. In addition to three weights used for input, output, and new hidden state calculation. On the right of the unfolded states over time are shown.

RNN capable to process input sequence of any length without any increase in the model size, prediction take in consideration input history sequence (Memory), same weights are used across time for the input sequence. But it still it has a short-term memory so old memory are forgotten due to vanishing gradient over time.

2.4.1 Long Short-Term Memory (LSTM)

LSTM introduced by (Hochreiter & Schmidhuber 1997) as a variant network from RNN where a gates concept is introduced to solve the short-term memory problem in RNN, these gates control information propagation from earlier states to current and next states, as in human reading a paragraph, he will not remember every single word to understand the context or the idea, but he remembers main words and information that can help to connect the ideas.

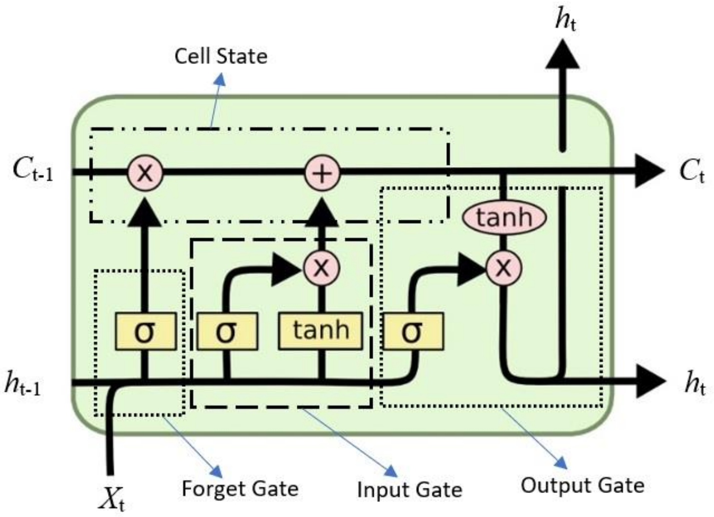


Figure 4 LSTM cell architecture, illustrates Input, Output, Forget, Cell State Gates

Figure 4) Illustrates LSTM cell architecture the main difference from traditional RNN it has gates to control the previous states (Memory) propagation, first the input and earlier hidden state are passed to the forget gate and using a sigmoid function it decides what information to keep or forget (by converting the values to zero to forget or near 1 to keep). Also input and earlier hidden state are passed to input gate where a tanh function is used to regulate the values and a sigmoid to decide which values need to be passed to next gate (Cell State).

Now cell have enough information to calculate cell state so previous state is pointwise multiplied by the output of the Forget gate to decide what data to pass, then the remaining values are pointwise added to the result of Input gate, now cell state is calculated as per the weights stored in the cell,

Lastly the Output gate passed the current state to a tanh function then pointwise multiplied with a sigmoid function (based on the cell input and earlier hidden state), now new hidden state is ready to be passed to next cell.

2.4.2 Gated Recurrent Unit (GRU)

(Cho et al. 2014) Introduced GRU as updated generation Recurrent Neural Network it is similar to LSTM network but with simpler design, where only two gates Update Gate and Reset Gate, the Cell State was dropped and the hidden state is used to transfer information to next cell.

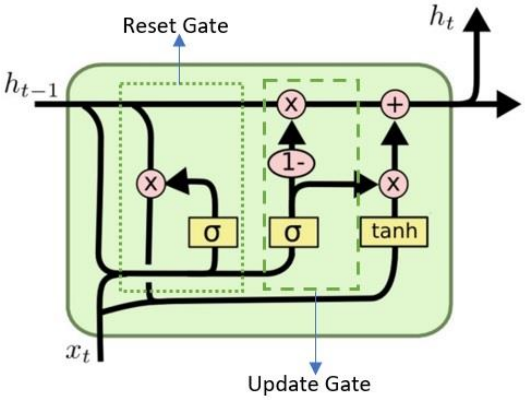


Figure 5 GRU cell architecture, illustrates Reset and Update Gates

Figure 5) Illustrates the GRU network architecture the update gate is used instead of both Forget and Input Gates in LSTM, where is decides what data points to drop (forget) and what is critical data points to keep. While Reset Gate is used to control how much data from old state information to pass.

Due to less tensors used in GRU, it takes less time and processing for training but both LSTM and GRU has comparable performance in Artificial Intelligence (AI) tasks, and using experiment we can decide the best one for certain task.

2.5 Image Features Extraction (IFE)

In the used Dataset (ElJundi et al. 2020), subset of Fliker8 (Hodosh, Young & Hockenmaier 2013), the average image dimension is 450 x 300 pixels and in colored images we have three values Red, Green, and blue. Therefor input size will be 450x300x3= 405,000 value per image.

This raised the need for dimensionality reduction phase where the image is transformed into features vector to be used in the deep learning network.

We have found two main methods

2.5.1 Visual Geometry Group (VGG16)

VGG16 introduced by (Simonyan & Zisserman 2014) a very deep Convolutional Neural Network (CNN), trained on ImageNet (Deng et al. 2009) which consist of more than 14M images with 1000 classes, VGG16 scored 92.7% accuracy.

The 16 in VGG16 refers to 16 layers architecture, the output of the VGG16 pre trained model can be used for the 1000 classes classifications or to use the dense layer directly with 4096 vector size, we have used this vector in our experiments as an image features representation.

2.5.2 Inception V3

(Szegedy et al. 2016) introduced Inception V3, the name is inspired by the phrase "'we need to go deeper' from Hollywood movie Inception by Christopher Nolan.

Inception v3 is a very deep convolutional neural network (CNN), its challenge was to get more convolution layer and staying with reasonable parameters to be computationally efficient, Inception v3 output is vector of 2048 values, represent image features.

3 Chapter Three: Methodology

In this section, we describe the characteristics of the AIC dataset. We show how we apply the preprocessing task to produce appropriate training datasets. Nevertheless, we describe Deep learning models that act as image classifiers which we are able to use for extracting features from the images.

3.1 Dataset

For the Image Captioning (IC) task, finding or creating a dataset is crucial in general for having better prediction results. In English, there are many benchmark IC Datasets. For example, Flickr8K (Hodosh, Young & Hockenmaier 2013) contains 8000 images with 5 English captions per image. Likewise, Flicker30K (Young et al. 2014) contains 30,000 images with 150,000 captions.

Flickr30K entities (Plummer et al. 2015) are reusable images which contain the caption text for either a specific entity or region and can be used for searches or retrieval tasks.

The largest dataset is MS-COCO (Chen et al. 2015) that contains more one and half million captions, 330,000 images with five independent captions for consistent evaluation.

For Arabic captioning, (ElJundi et al. 2020) introduced the first publicly available AIC dataset that is based on Fliker8K, with 8000 images, 6000 for training, 1000 for validation, and the remaining 1000 for testing. Figure 1) shows a sample of images and captions from this dataset. (ElJundi et al. 2020) translated Flickr8K output using Google Translate API and the best three translations is post-edited, if needed by human expert. Since the dataset was generated by machine translation, some low-quality Arabic sentences appear in **Error! Reference source not found.**).

3.2 Preprocessing Techniques

We have used four Preprocessing techniques. Each technique generates a different dataset, namely: A, B, C, and D. Bellow we provide the detailed description of each of which:

3.2.1 Original Text (Method A)

To evaluate the effect of text preparation in the experiment, we used the captions as is.

3.2.2 Base Preprocessing (Method B)

Both (ElJundi et al. 2020), (Hejazi et al. 2021) used the traditional technique proposed by (Shoukry & Rafea 2012). In this method, punctuation, diacritics, non-Arabic letters, single letter words were dropped. Also, a lexographic normalization process took place to unify similar letters, including "إأآاٱ" -> "ا" ,"ى" -> "ي" ," ؤ ئ" -> "ء" ,"ة" -> "ه" ,

"گ" -> "ك".

3.2.3 Removing the Alef with the Nunation (Method C)

We have noticed that when removing Nunation diacritic the extra Alef is not removed. So, we removed this extra Alef too, such that the word "قميصًا" (shirt-with extra nunation-) becomes "قميص" (shirt-without nunation-) instead of "قميصا" (shirt-with Alef as partial nunation-). Applying this technique would reduce the total vocabularies because in the previous method each surface form was considered a different vocabulary as illustrated in Figure 6). Moreover, we separated and removed the Waw conjunction from next word, e.g. "ويقول" (and-he-says) becomes "يقول" (he-says).

3.2.4 Full Preprocessing (Method D)

We partially followed Method C, but we kept the conjunction Waw. In all previous methods all single letter words was removed including the isolated conjunction Waw, , e.g. “ويقول” becomes “يقول” “و”. but we think this highly affect syntactic and semantic of the captions. Figure 6) shows differences in the frequency counts for preprocessing methods B, C, and D.

Table 1 Four Preprocessing methods used, with sample caption and number of detected

Preprocessing method	Sample Caption	Vocabularies		
		Total	unique	Repeated
Method A	صبي يرتدي نظارات و قميصًا أحمر	179,532	11,386	5,893
Method B	صبي يرتدي نظارات قميصا احمر	178,176	10,692	5,729
Method C	صبي يرتدي نظارات قميص احمر	178,175	9,713	5,344
Method D	صبي يرتدي نظارات و قميص احمر	183,342	9,714	5,345

3.2.5 Technique

The final caption is then surrounded by a start and end tags. The length of each caption is set to 25 words; shorter captions are padded with nulls. Table 1) shows the output of the four preprocessing methods along with their statistics. Since we dropped words with single appearance, we can notice in the third column of the table a big reduction in the repeated vocabularies count. For example, applying Method C to the dataset produces 9,713 unique vocabularies but only 5,344 of them were repeated and the remaining 4,369 should be removed.

The reason of having these words sparse in the caption dataset might be due to misspelled words or the use of rare words. If size of the dataset is small, it might make the caption not a good representative for the Arabic Language model, since many words rarely appear or do not appear at all. This raises the need for a big enough dataset for AIC.

Low frequency words affect the prediction process, so they have to be treated at the preprocessing stage since often they are typos. Figure 6) shows how the proposed methods C and D reduced the occurrences of words with just one appearance from 4963 (Method B) to 4369, a decrease of about 12%.

As per Figure 6), the number of low frequency words reduced in most cases, but we can observe an increase of the number of words with 12 and 13 frequency this might be due to the matching between words with low frequencies after applying the preprocessing task.

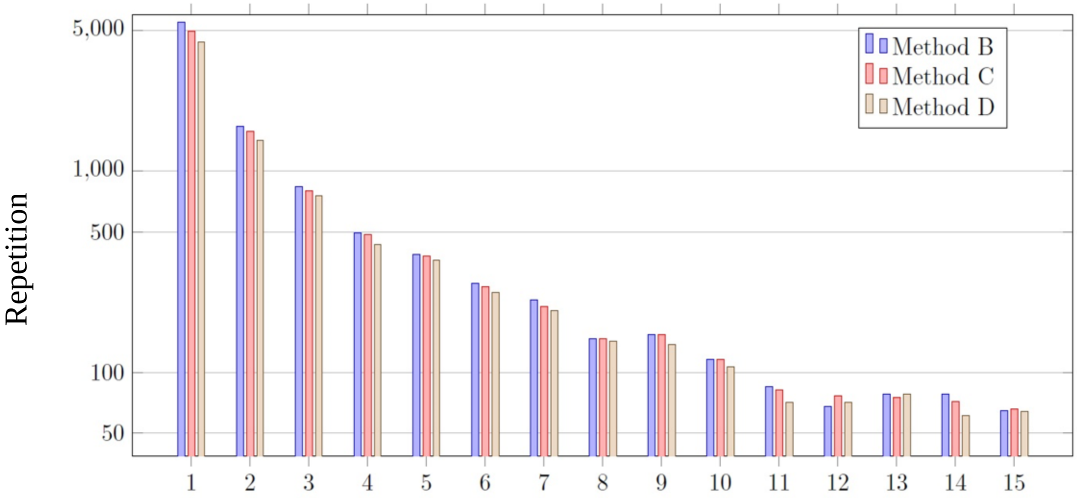


Figure 6 Variation in the frequency counts for each preprocessing method (rare counts).

We found 1,141 Vocabularies that Starts with Waw in the original dataset the Incorrect words starts with Waw was 1,029, which means about 90.1% that starts with Waw is incorrect.

3.3 Models

Recurrent Neural Networks (RNN) is best used for time series data, but it suffers from the short term memory problem or the vanishing gradient where the earlier inputs effect starts to be exponentially smaller when we move more steps forward in the prediction. We can resolve this by using one of the following variations: Gated Recurrent Unit (GRU) or Long Short-Term

Memory (LSTM) where gates are used to control the older sequence information by saving in memory unit and propagate to next units.

Since text is considered a Time Series prediction, we propose to use GRU and LSTM network in our experiment and compare their performance and effect on the results.

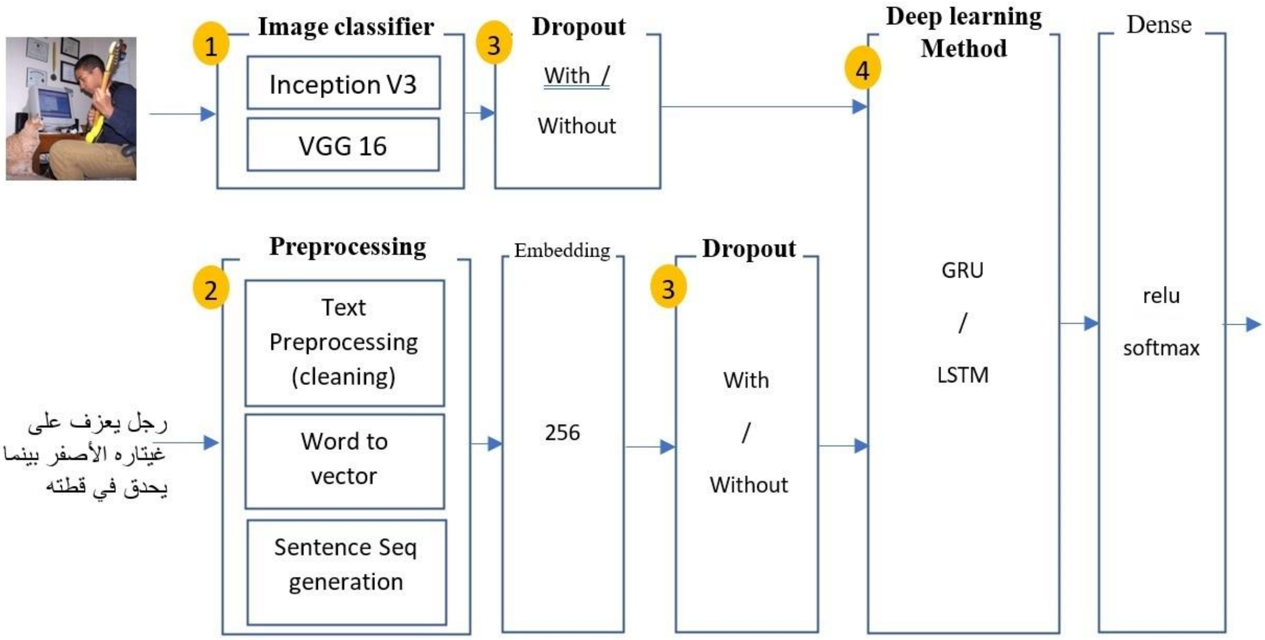


Figure 7 Experiment flow that yields a total of 32 experiments: (1) two image classifiers, (2) 4 Preprocessing methods, (3) Dropout, (4) two Deep learning techniques.

3.4 Experiment

Experiments were designed to test the impact of our independent variable on the quality and accuracy of Arabic captions. We have conducted experiments that involved 32 variable combinations: 4 Datasets, 2 image classifiers, 2 dropout usage, and 2 Deep Learning methods. Figure 7) shows the experiment design where we have indicated four labels to highlight the variant stages of the experiments. In the first stage (1) images are passed to one of two features extractors (Inception V3, VGG16). Next, a vector that contains image features is produced, captions are preprocessed using the four methods then tokenized, and then passed to embedding layer.

Afterwards, a dropout layer is used, if required by experiment, and the results are passed to either LSTM or GRU. At the end a Dense layer is used for prediction. Each model is saved, and test images are passed to it for caption prediction. All predicted captions are recorded and compared with the actual ones. BLEU-1/2/3/4 scores are calculated and stored per each experiment. Table 2) shows the recorded results which we analyze and discuss in the next sections. In each experiment one path is chosen at a time until all combinations are covered. Many experiments were repeated with lower epoch when Overfitting is detected.

The configuration of the hardware used is: Intel(R) Core(TM) i7 10th generation (6 core , 12 logical processors) with NVIDIA GeForce GTX1 1650 (4GB) for processing, 16 GB RAM Memory, total accumulated training time for latest models about 7 hours.

The collected experiment data was analyzed to find the effect of each factor. Also, a t-test is applied to find the significance of each variable.

3.5 Overfitting

since the size of Dataset is small training and testing (validation) loss value is monitored after each epoch, if the testing loss increases or stays the same while the training loss decreased, this means an overfitting is detected and we observe a lower prediction accuracy from that model.

Then lower number of epochs are made to reach the lower testing loss value and a better model accuracy (BLEU measure).

3.6 Evaluation

To evaluate each experiment result, BLEU-1/2/3/4 are used. BLEU is a precision-based metric that ranges between zero (lowest) and one (best). The number of n-grams that appears in the candidate text is compared to total n-grams in the reference text. This metric is used by (ElJundi et al. 2020), which we use to compare our results with their results.

4 Chapter Four: Results

In this section, we present results from 32 experiments. Table 2) shows the BLEU results of each experiment. Figure 8) illustrates these results.

4.1 BLEU

BLEU-1/2/3/4 was used to measure accuracy of each model prediction. Table 2) shows the results of these experiments.

We can notice that the best BLEU scores achieved from using VGG16 with GRU on the Dataset generated using the method D, and without dropout, are BLEU-1=36.5, BLEU-2=21.4, BLEU-3=12, and BLEU-4=6.6.

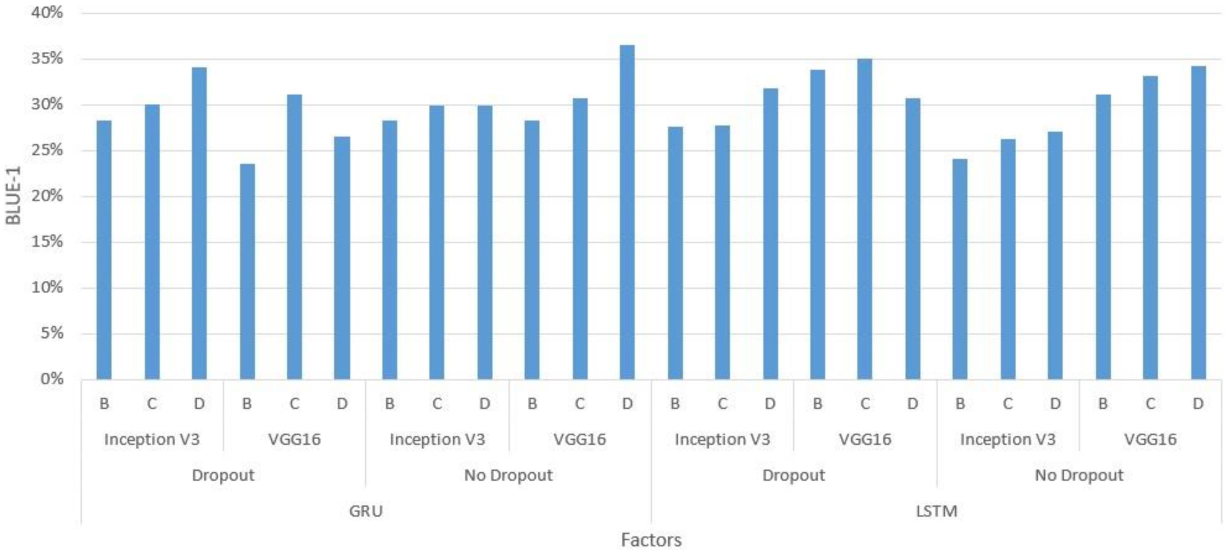


Figure 8 Experiments results for BLEU-1 upon different parameters.

Table 2 BLEU-1/2/3/4 result of the experiment per variables combinations.

Image Classifier	Model	Dataset	Dropout				No Dropout			
			BLEU%				BLEU%			
			1	2	3	4	1	2	3	4
Inception V3	GRU	A	26.6	13.4	6.8	3.6	29.5	14.9	7.8	4.2
Inception V3	GRU	B	28.3	14.7	7.4	3	28.3	13.5	6.7	3
Inception V3	GRU	C	30.1	15.8	7.9	3.9	29.9	15.7	8.3	4.6
Inception V3	GRU	D	34.1	17.7	9.5	5.3	29.9	16.6	9.4	5.1
Inception V3	LSTM	A	24.4	10.7	4.8	1.8	22.6	10.7	5.1	2
Inception V3	LSTM	B	27.6	11.7	4.7	2	24.1	11.4	5.1	2.1
Inception V3	LSTM	C	27.8	13.5	6.5	3	26.3	11.1	4.5	2.1
Inception V3	LSTM	D	31.8	15.3	8	4.6	27.1	12.2	5.7	2.9
VGG16	GRU	A	24.6	13.3	7.2	4	24	12.9	6.4	3.1
VGG16	GRU	B	23.5	13.2	7.1	3.6	28.2	15.1	8.3	4.6
VGG16	GRU	C	31.1	17.5	9	4.1	30.8	16.8	8.8	4.4
VGG16	GRU	D	26.5	15.1	8.8	5.1	36.5	21.4	12	6.6
VGG16	LSTM	A	33.6	20.1	11.2	6.4	32.3	18.5	9.8	5.3
VGG16	LSTM	B	33.9	19.5	10.5	5.7	31.2	17.9	9.7	5.5
VGG16	LSTM	C	35.1	20.9	11.5	6.3	33.1	18.9	10.1	5.2
VGG16	LSTM	D	30.7	18.2	10.1	5.4	34.2	19.9	10.8	6.1

4.2 Preprocessing Methods Comparison (Datasets)

Each Dataset is produced using a different Preprocessing method, we compared the three Datasets (B,C,D) to show the effect of Preprocessing on the results accuracy. Figure 9) illustrates the BLEU-1's result.

We can notice that the proposed new Preprocessing methods give higher BLEU measure. The reason might be due to less infrequent words that arise from consistent typo, such as concatenating Waw with the next word, or keeping the Alef of nunation, which produces a vocabulary that is irrelevant to the original word.

A paired-samples t-test was conducted to compare the Dataset C with the Dataset B. There is a significant difference in the scores from Dataset C (M=0.1482, SD=0.1045) and Dataset B (M=0.1346, SD=0.0978) under the conditions: t(31)=5.0344, p = 0.000019.

These results suggest that removing the Alef of the nunation affect the BLEU results and increases it.

Another paired-samples t-test was conducted to compare Dataset D with Dataset C. There was a significant difference in the scores for Dataset C (M=0.1482, SD=0.1045) and Dataset D (M=0.1571, SD=0.1044) under the conditions: t(31)=-2.2136, p = 0.034.

These results suggest that keeping the Waw in the preprocessing phase affect the BLEU results and increases it.

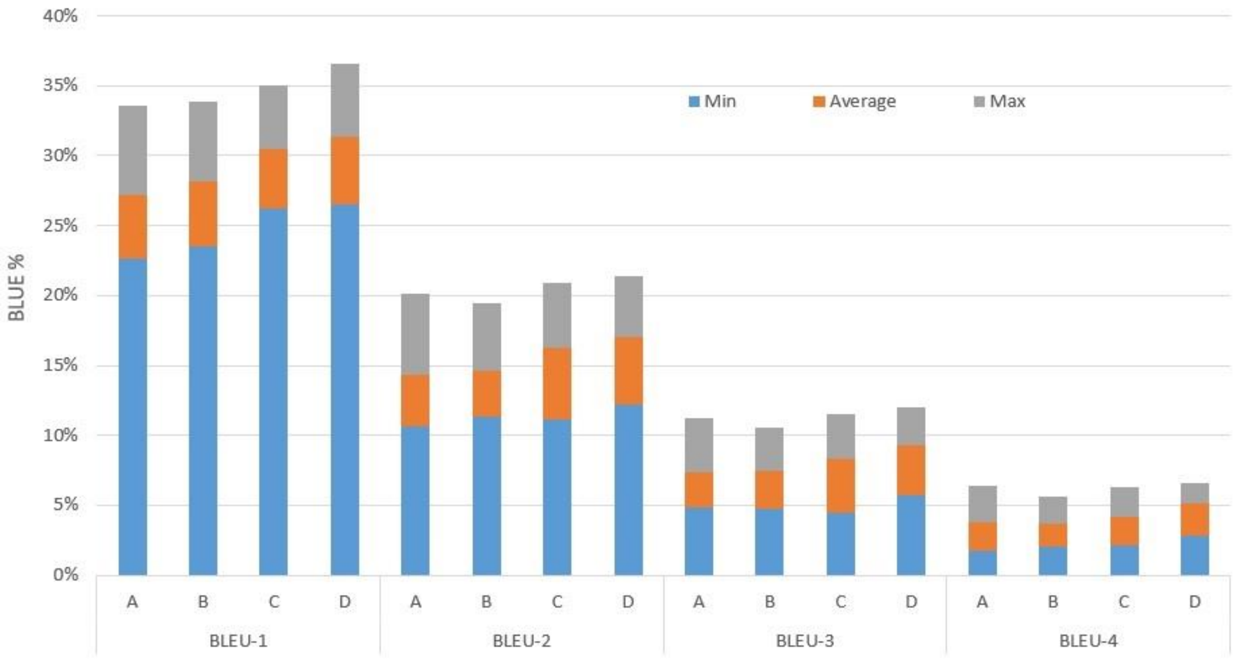


Figure 9 Average, minimum, and maximum value of BLEU-1/2/3/4 achieved per each Preprocessing method.

4.3 Image features model Comparison

We involved two image models to extract image features, VGG16 and Inception v3. Figure 10) illustrates a comparison of BLEU results of both models.

A paired-samples t-test was conducted to compare using VGG16 and Inception V3 as image features extractor.

There is a significant difference in the scores for VGG16 (M=0.1564, SD=0.1011) and Inception V3 (M=0.1294, SD=0.0.0976 under the conditions: t(63)=5.6714, p = 0.00000038. These results suggest that using VGG16 over Inception V3 affect the BLEU results and increases it.

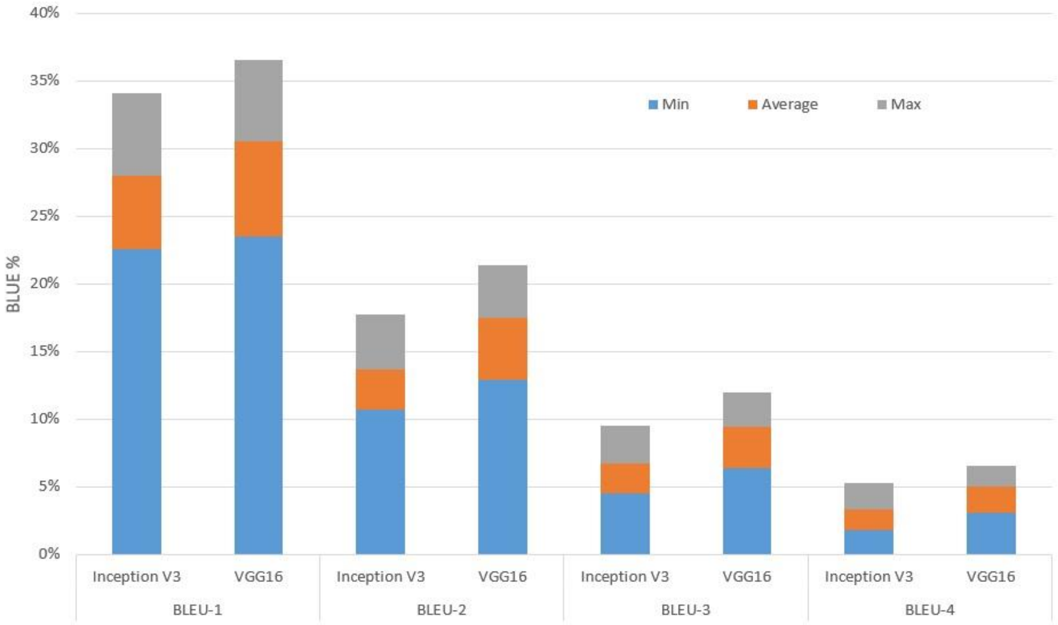


Figure 10 Average, minimum, and maximum value of BLEU-1/2/3/4 achieved per each Image features extraction Model.

4.4 DropOut Comparison

We have studied the impact of using the Dropout with Arabic image captioning process. Figure 11) illustrates the results of experiments with/without Dropout.

A paired-samples t-test was conducted to compare the results with and without Dropout. There was not a significant difference in the scores for using Dropout (M=0.1423, SD=0.1005) and not using dropout (M=0.1436, SD=0.1001 conditions; t (63) =-0.46, p = 0.647.

There is no evidence that using Dropout will affect the BLEU results of the generated captions.

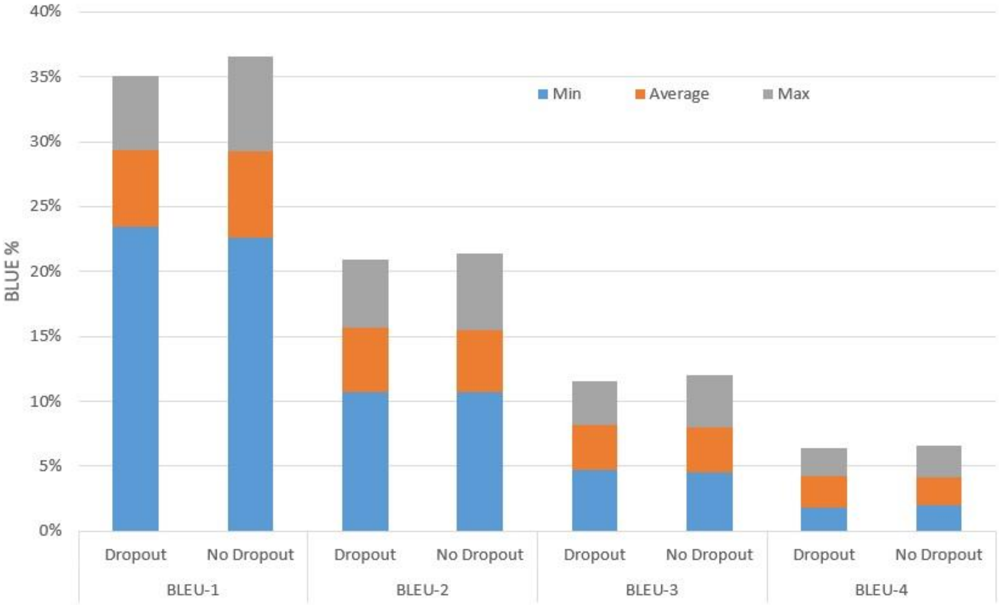


Figure 11 Average, minimum, and maximum value of BLEU-1/2/3/4 achieved per Dropout usage.

4.5 GRU vs LSTM

Two Deep Learning methods were compared (GRU,LSTM). Figure 12) illustrates the BLEU results per each method.

The use of GRU or LSTM as a text prediction model was compared using a paired-samples t-test. There is no significant difference in the scores for GRU (M=0.142, SD=0.097) and LSTM (M=0.1438, SD=0.1035) under the conditions: t(63)=0.419, p = 0.6766.

These results cannot support that using GRU instead of LSTM may affect the BLEU results of the generated captions.

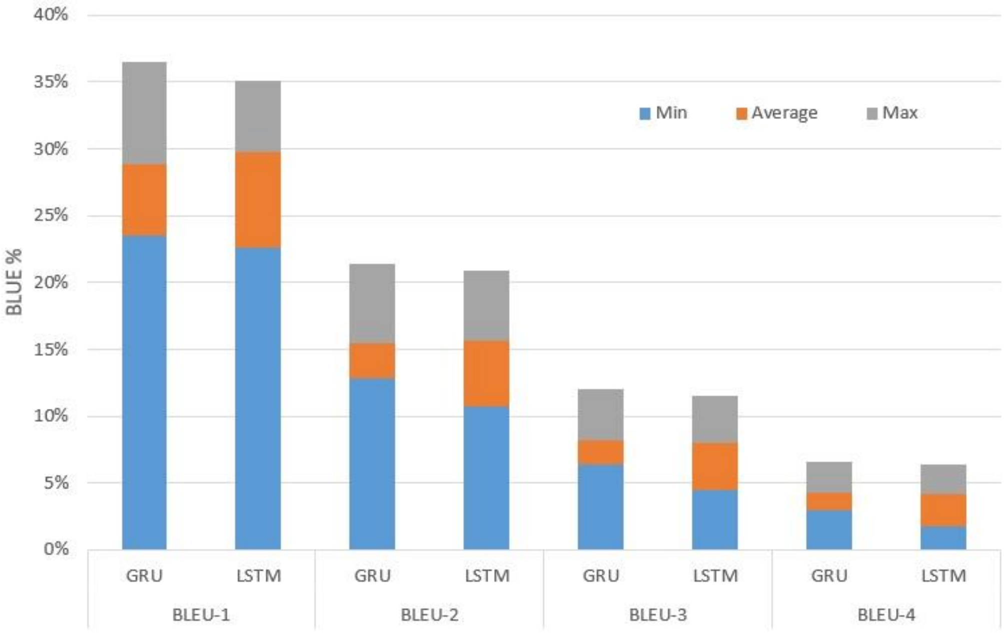
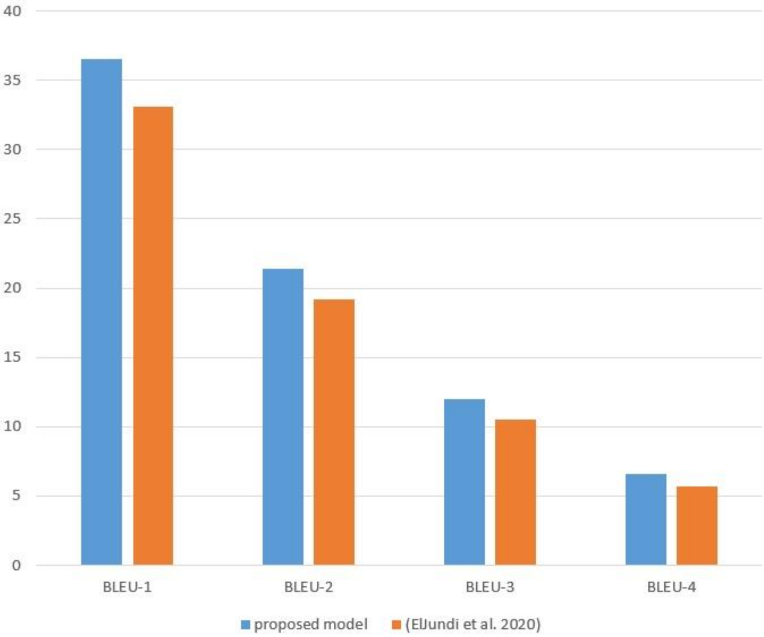


Figure 12 Average, minimum, and maximum value of BLEU-1/2/3/4 achieved per each Deep learning method.

4.6 Our results vs best results on the same dataset:

We have compared our best results with best results achieved by (ElJundi et al. 2020) who is made this dataset public and published his results,

Our model have outperformed best results and according to t test this difference in due to preprocessing method, that solved more sophisticated language specific cases.



we

Figure 13 Achieved results compared with previous best result on same dataset

5 Chapter Five: Conclusion, Future Work and Limitations

5.1 Conclusion

Arabic Image Captioning resources are scarce. Fortunately, one public dataset is available. We created an AIC model with tuned factors that outperformed the best results on the publicly available dataset. According to paired t-tests conducted on the results, Arabic text preprocessing and image features extractors have a major role to play in improving the AIC results. For the purpose of comparison, two preprocessing techniques for Arabic captions were proposed and found to yield better results.

A total of 32 experiments were conducted to analyze the effects of four variables. We considered the following variables: preprocessing techniques (original text, normal preprocessing, Alef removal with nunation, and keeping conjunction Waw), Waw typo correction, Deep learning techniques (LSTM, GRU), inclusion and exclusion of Dropout, and two Image features extraction methods (Inception V3, VGG16).

As a result, BLEU1=36.5, BLEU-2=21.4, BLEU-3=12, and BLEU-4=6.6 were the best results we reached. The results were compared using paired t-tests, and the Arabic preprocessing methods exhibited an enhanced level of quality, and VGG16 significantly outperformed Inception V3 as an image features extractor. Using Dropout or LSTM instead of GRU, however, did not have a major effect so we can use GRU because of simpler design and less process power requirement.

Research questions was discussed and answered based on the results as the following:

- How to enhance Arabic image captioning quality using deep learning?

  We have outperformed the work done on (ElJundi et al. 2020) dataset using VGG16 as image features extractor and enhance the Arabic text preprocessing method. Following a more advance method for Arabic text preprocessing can enhance the results, unlike English some typo in Arabic are not affect visually reading by human, but it extremely affect the computer processing.

- How Preprocessing, deep learning method, image features, and dropout affect the Arabic image captioning quality?

  As per the results the most important factors was the preprocessing method and the image features extraction, using VGG16 with preprocessing method D achieved the best BLEU score, while there were no major contribution for Dropout or deep learning method.

  But having a larger dataset may lead to totally different results, since out dataset don't reached other English dataset size.

5.2 Future Work and Limitations

The main limitation was the relatively small Dataset size since there was only one publicly available Dataset for AIC. Other Preprocessing and Deep learning methods could be included in the comparisons but doing that will increase the number of experiments and require more resources, therefore we can consider it in the future work.

As a future work, researchers can benefit from the outcomes of this study by employing it to their future research, particularly, a larger dataset can be created and made public to avail linguistic resources research in this area.

Not to mention, having a big dataset provides several possibilities to tailor the use of extra deep learning techniques and come up with better word representation and features that can significantly improve the performance of the Arabic Image Captioning.

References

Al-muzaini, H. A., Al-yahya, T. N. & Benhidour, H. (2018). Automatic Arabic image captioning using RNN-LSTM-based language model and CNN. *International Journal of Advanced Computer Science and Applications*. Science and Information Organization, vol. 9(6), pp. 67–73.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086 [online].Available at: http://www.panderson.me/up-down-attention.

Aneja, J., Deshpande, A. & Schwing, A. G. (2018). Convolutional Image Captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5561–5570.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Doll'ar, P. & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cho, K., van Merrienboer, B., Bahdanau, D. & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *rXiv preprint arXiv:1409.1259*.

Deng, J., Dong, W., Socher, R., Li, L.-J. & Li, K. and F.-F. L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255.

Dustin Stout. (2020). *Social Media Statistics*. *https://dustinstout.com/social-media-statistics/*.

ElJundi, O., Dhaybi, M., Mokadam, K., Hajj, H. & Asmar, D. (2020). Resources and end-to-end neural network models for Arabic image captioning. *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SciTePress, pp. 233–241.

Hejazi, H. D., Khamees, A. A., Alshurideh, M. & Salloum, S. A. (2021). Arabic Text Generation: Deep Learning for Poetry Synthesis. *Advances in Intelligent Systems and Computing*. Springer Science and Business Media Deutschland GmbH, pp. 104–116.

Hejazi, H. & Shaalan, K. (2021). Deep Learning for Arabic Image Captioning: A Comparative Study of Main Factors and Preprocessing Recommendations. *International Journal of Advanced Computer Science and Applications,* pp. 37–44.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, vol. 9(8), pp. 1735–1780.

Hodosh, M., Young, P. & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899.

Jindal & Vasu. (2018). Generating image captions in Arabic using root-word based recurrent neural networks and deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for Computational Linguistics, pp. 144–151.

Mualla, R. & Alkheir, J. (2018). Development of an Arabic Image Description System. *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 8 [online].Available at: www.ijcstjournal.org.

National Federation of the Blind. (2019). *Blindness Statistics* [online]. [Accessed February 15, 2022]. Available at: https://nfb.org/resources/blindness-statistics.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J. & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J. & Goel, V. (2017). Self-critical Sequence Training for Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024.

Shoukry, A. & Rafea, A. (2012). “Preprocessing egyptian dialect tweets for sentiment mining.” , in *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, p. 47 [online].Available at: https://www.researchgate.net/publication/233746674.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

Tian, W., Zhou, R. & Zhao, Z. (2020). “Cascading top-down attention for visual question answering.” , in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7.

Wang, J., Wang, W., Wang, L., Wang, Z., Feng, D. D. & Tan, T. (2020). Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition*. Elsevier Ltd, vol. 98.

WHO. (2021). *Blindness and vision impairment*. *Blindness and vision impairment* [online]. [Accessed February 7, 2022]. Available at: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

Wu, J., Hu, Z. & Mooney, R. J. (2019). Generating Question Relevant Captions to Aid Visual Question Answering. *arXiv preprint arXiv:1906.00513* [online].Available at: http://arxiv.org/abs/1906.00513.

Young, P., Lai, A., Hodosh, M. & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78.