أكاديمية سدايا
SDAIA Academy

IMAGE CAPTIONS

PROJECTS REPORT -T5 BOOTCAMP

Batoul Alosaimi
Amal Altamran
Amirah Alotaibi
Alanoud Alhussain
Shoroq Almutiri
Norah AlQahtani
Asma Alsulami

Table of Contents

Introduction ***3***

Objective: ***3***

Design And Data Description: ***3***

Methodology ***4***

Algorithms and Tools ***4***

Communication ***4***

Workflow ***9***

Conclusion ***9***

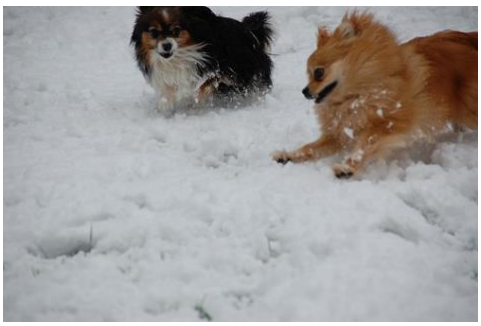
Future Work ***9***

Introduction

Image caption Generator is a popular research area of Artificial Intelligence that deals with image understanding and a language description for that image. Generating well-formed sentences requires both syntactic and semantic understanding of the language. Being able to describe the content of an image using accurately formed sentences is a very challenging task, but it could also have a great impact, by helping visually impaired people better understand the content of images.

This task is significantly harder in comparison to the image classification or object recognition tasks that have been well researched. The biggest challenge is most definitely being able to create a description that must capture not only the objects contained in an image, but also express how these objects relate to each other.

Consider the following Image from the Flickr8k dataset:-



What do you see in the above image?

You can easily say 'A black dog and a brown dog in the snow' or 'The small dogs play in the snow' or 'Two Pomeranian dogs playing in the snow'. It seems easy for us as humans to look at an image like that and describe it appropriately.

Objective:

- help disability people especially blinded society and make it easier for them to see what we are seeing in pictures by converting text caption into a voice
- improve technologies that serve the Arabic language
- Facilitate learning Arabic and English languages for early childhood

Design And Data Description:

In the Flickr8k dataset, each image is associated with five different captions that describe the entities and events depicted in the image that were collected. By associating each image with multiple, independently produced sentences, the dataset captures some of the linguistic variety that can be used to describe the same image.

Our dataset structure is as follows:-

Flick8k/

- Flick8k_Dataset/ :- contains the 8000 images
- Flick8k_Text/
- Flickr8k.token.txt:- contains the image id along with the 5 captions in the English language.
- Flickr8k.arabic.full.txt:- contains the image id along with the 5 captions in the Arabic language.

Methodology

1) Understand problem
2) load data
3) pre-processing
4) Visualization + tokenisation
5) split data
6) transfer learning
7) model
8) Evaluation

Algorithms and Tools

- Keras
- TensorFlow
- PIL
- collections, random, re
- Convolutional Neural Networks
- Pandas, Numpy, matplotlib, seaborn
- pydotplus
- googletrans
- gTTS

Communication

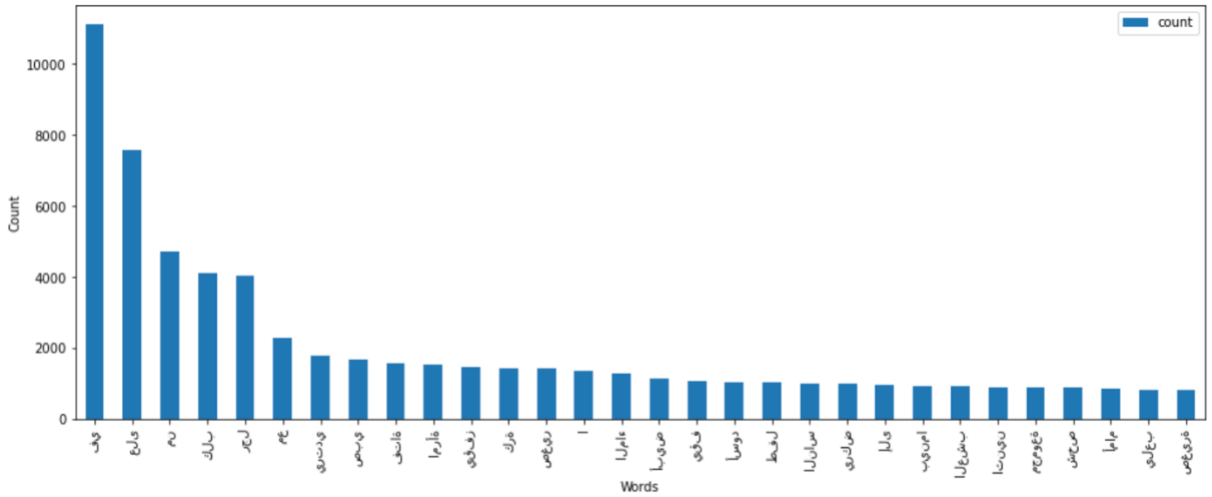


Figure1: bar chart present count of each word in the Arabic dataset

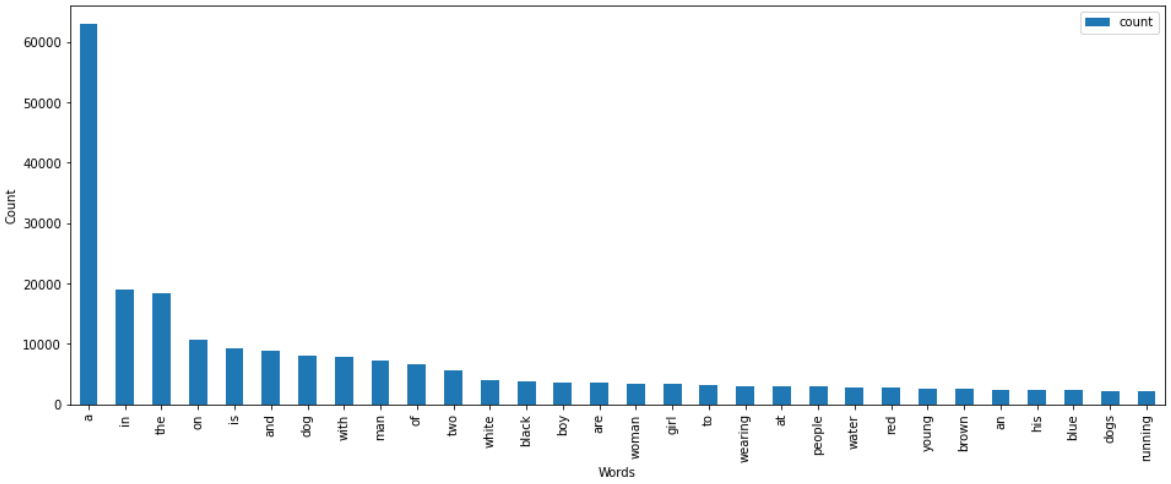


Figure2: bar chart present count of each word in the English dataset

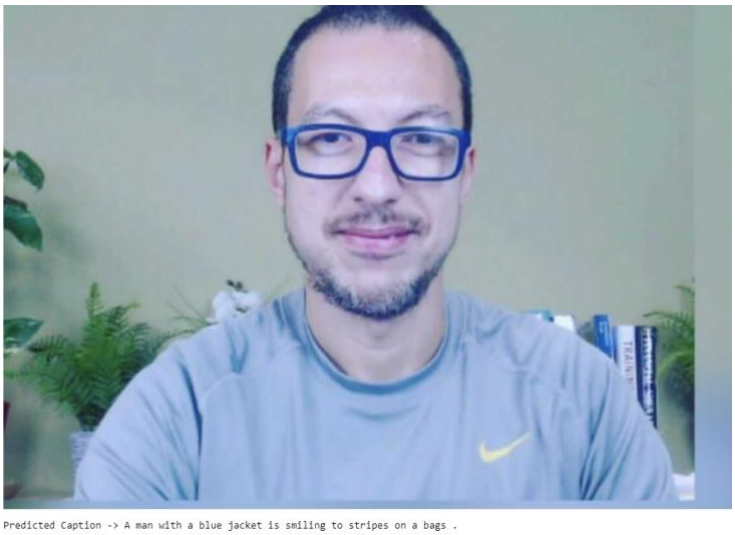


Figure3: Predicted English caption for Mr. Baddar picture

Predicted Caption -> A group of people sitting down a snowy mountain with a mountain in a mountains .

Figure4: Predicted English caption for image from the internet

Figure 5: Predicted Arabic caption for Saudi football Player Mohammed Noor picture from the internet

فتاة صغيرة مغطاة بالطلاء تجلس أمام قوس قزح
فتاة صغيرة تجلس أمام قوس قزح ملون كبير
فتاة أمام لوحة قوس قزح

Figure 6: image from Arabic data set

A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
A little girl is sitting in front of a large painted rainbow .
A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
There is a girl with pigtails sitting in front of a rainbow painting .
Young girl with pigtails painting outside in the grass .

Figure 7: image from English data set

figure8: predicted image caption by the Arabic model

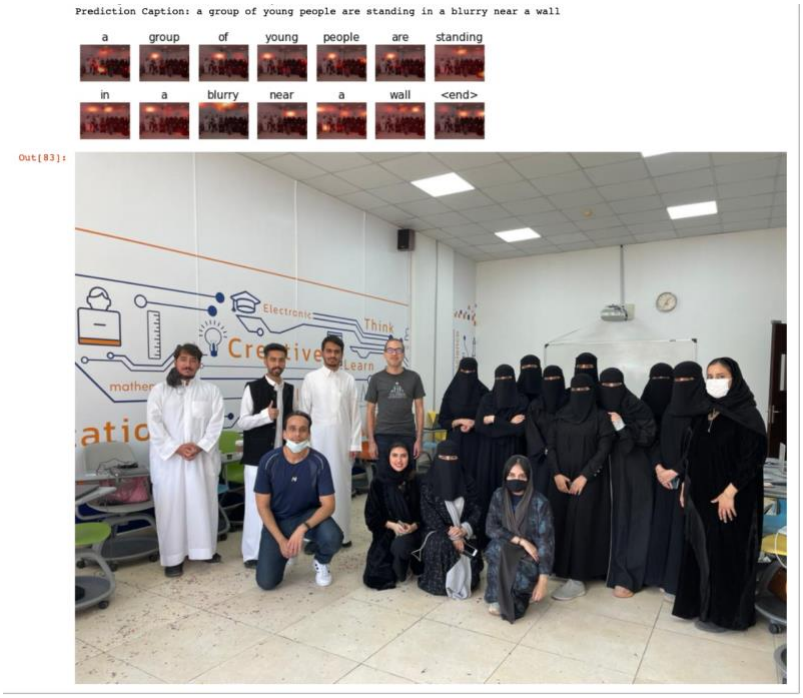


Figure 9: Predicted image caption by English model

Workflow

- we Loaded captions as values and images as key in dictionary
- then we split the data to train, test, and validation dataset
- And then doing image extract feature And Loading 50-layer Residual Network Model and getting the summary of the model.
- Then we do image process by change the images size and reshape.
- And for text process:
- Splitting each captions stored in 'sentences' and storing them in 'words' .
- Vectorization.
- Model used:
- Inception V3: Is a Convolutional Neural Network for Assisting in Image Analysis and Object Detection.
- "sequential" with patch size 512, epochs 200, the accuracy is 0.9022 and loss 0.2545
- Evaluation: BLEU – Score: (BiLingual Evaluation Understudy) evaluating machine-translated text
- and do the predict from image outside the train model like image from internet and predict caption for the image.

Conclusion

- We created a model that predict image description (caption) so people who cannot see will be able to know what inside images
- In addition, we improved a program that convert caption text into voice
- Our project support both Arabic and English languages

Future Work

- Work on a larger dataset that contain local images
- Improve Arabic voice model
- Create website