شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

HANAA ALY

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering

Visual Question Answering Using Deep Learning Techniques

A Thesis submitted in partial fulfillment of the requirements of
Master of Science in Electrical Engineering
(Computer and Systems Engineering)

by

Ahmed Mostafa Soliman Radwan
Bachelor of Science in Electrical Engineering
(Computer and Systems Engineering)
Faculty of Engineering, Ain Shams University, 2017

Supervised By

Prof. Dr. Hazem Abbas
Professor in Computer and Systems Engineering Department, Faculty of Engineering - Ain Shams University
Prof. Dr. Mahmoud I. Khalil
Professor in Computer and Systems Engineering Department, Faculty of Engineering - Ain Shams University

Cairo, 2021

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering Department

Visual Question Answering Using Deep Learning Techniques

by

Ahmed Mostafa Soliman Radwan
Bachelor of Science in Electrical Engineering
(Computer and Systems Engineering Department)
Faculty of Engineering, Ain Shams University, 2017

Examiners' Committee

Name and affiliation	Signature
Prof. Dr. Omar H. Karam Faculty of Informatics & Computer Science British University in Egypt.	.....
Prof. Dr. M.Watheq El-Kharashi Computer and Systems Engineering Department Faculty of Engineering, Ain Shams University.	.....
Prof. Dr. Hazem Mahmoud Abbas Computer and Systems Engineering Department Faculty of Engineering, Ain Shams University.	.....
Prof. Dr. Mahmoud Ibrahim Khalil Computer and Systems Engineering Department Faculty of Engineering, Ain Shams University.	.....

Date: 16 December 2021

Statement

This thesis is submitted as a partial fulfillment of Master of Science in Electrical Engineering, Faculty of Engineering, Ain shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Ahmed Mostafa Soliman Radwan

Signature

.....

Date: 11 October 2021

Researcher Data

Name: Ahmed Mostafa Soliman Radwan
Date of Birth: 1/7/1993
Place of Birth: Cairo, Egypt
Last academic degree: Bachelor of Science
Field of specialization: Computer and Systems Engineering
University issued the degree: Ain Shams University
Date of issued degree: 2017
Current job: Teaching assistant

Abstract

Visual Question Answering (VQA) is a recent task that challenges algorithms to reason about the visual content of an image to be able to answer a natural language question. In this work, the performance of state of the art VQA algorithms on different VQA benchmarks is evaluated. Each benchmark is more effective at testing VQA algorithms on different levels. Some datasets challenge the algorithms to perform complex reasoning steps to arrive to an answer. Other datasets might challenge algorithms to retrieve external world knowledge to answer the posed questions. The algorithms reviewed and used in our experiments are categorized by their main contributions into 4 categories. Firstly, the joint embedding approach which focuses on how to map the visual and textual data into a common embedding space. Secondly, attention based methods which focuses on relevant parts of the image or the question. Thirdly, compositional models which deal with composing a model from smaller modules. Finally, we introduce external-knowledge based algorithms which need external sources to be able to retrieve facts necessary to answer a question. Other algorithms that don't specifically belong to the aforementioned categories, but offer state of the art performance, are also included.

Our work also introduces the first Arabic dataset in VQA that testes algorithms abilities to do complex visual reasoning, AR-CLEVR. The Arabic questions are generated for synthetic scenes using algorithms that auto-generates questions based on ground truth information from the scene's graph. Results from the experiments conducted on the state of the art algorithms helps us conclude the best algorithm we should choose for our newly introduced dataset. The new dataset is integrated within the openvqa framework, to enable future researchers interested in the VQA problem to easily reproduce our results and use new algorithms on our new benchmark.

Thesis Summary

Summary

This thesis reviews state of the art visual question answering (VQA) algorithms and datasets, describes out work on constructing the first Arabic VQA dataset, and the details of our approach on that dataset.

This thesis is divided into 7 chapters as show below

Chapter 1

Gives an introduction about recent advances in integrating language and vision research to show the motivation behind working on the VQA problem. It explains the big picture of the problem, then gives an overview on the work done in this thesis.

Chapter 2

Gives an overview of the different categories of datasets used to benchmark VQA algorithms and the methodology of constructing those datasets, and state of the art approaches to deal with the problem

Chapter 3

Gives the basic theoretical foundations that our work is based on. It explains the main recent advances in computer vision, natural language processing and deep learning algorithms. Also it explains multi-modal machine learning algorithms since both the visual and linguistic modalities are integrated in our work to be able to answer visual questions about images.

Chapter 4

Shows our work with different visual question answering algorithms, categorising them into classes, each of which has a common scheme that could be best suitable for a specific type of VQA benchmark.

Chapter 5

Describes the work done in this thesis on the first Arabic VQA dataset. It describes the methodology used to construct the dataset. It gives statistics about images, questions in the dataset and also shows samples of the dataset.

Chapter 6

This chapter shows the experiments we performed on publicly available datasets, and conclusions we dervied from those experiments. It then shows how we employed the conclusions we derived from the aforementioned experiments into our experiments on the Arabic VQA dataset we created. Parameters of the models we used in the experiments, results are all included in this chapter.

Chapter 7

Concludes our work and shows possible future directions to build on the work we have done.

Key words: multi-modal machine learning, computer vision, natural language processing, visual question answering.