JAIST Repository
https://dspace.jaist.ac.jp/

Title	視覚障害者のための視覚的質問応答の研究
Author(s)	Le, Thanh Tung
Citation	
Issue Date	2021-12
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/17600
Rights	
Description	Supervisor:NGUYEN, Minh Le, 先端科学技術研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

氏名	LE, Thanh Tung		
学位の種類	博士（情報科学）		
学位記番号	博情第 461 号		
学位授与年月日	令和 3 年 12 月 24 日		
論文題目	A STUDY OF VISUAL QUESTION ANSWERING FOR BLIND PEOPLE		
論文審査委員	Nguyen Le Minh.	JAIST	Professor
	Satoshi Tojo	JAIST	Professor
	Shirai Kiyoaki	JAIST	Assoc. Professor
	Shinobu Hasegawa.	JAIST	Professor
	Tran The Truyen	Daikin University	Assoc. Professor

論文の内容の要旨

Multi-media website which contains tons of image and text data has a high demand for extracting and understanding representation and relationship of image and question si-multaneously to support users for retrieving information, answering questions, and so on.

Besides, it is essential to support blind people as well as the visually impaired community to overcome diffculties in their daily lives. The vision-language systems are promising to learn and understand the visual and textual representation together without the physical vision. Together with its potential, this task also raises some challenges due to unique characteristics of multi-modal systems as well as a specific domain for blind people in-cluding i) question may not be in well-grammar texts; ii) image is poor quality from the collecting process that requires a robust approach to extract visual features; iii) unan-swerable sample appears the question-answering task.

This study aims to take advantage of advanced Deep Learning techniques to understand and extract meaning and relationship between image and question to predict answers. To this end, the research question is how to employ deep learning architectures to repre-sent and combine the image and question effectively to obtain their hidden relationship especially in the special challenges in VQA dataset for the blind.

To answer the above research question, we propose a hierarchal VQA system including four sub-tasks as follows:

- Answerability Prediction - determines whether the content of images is answered by a question or not, which is useful to eliminate error samples in VQA systems. By taking advantage of Transformer architecture, we propose a VT-Transformer model to extract the visual and textual features delicately thanks to the strength of pre-trained models. According to the experimental results, VT-Transformer generally outperforms the existing baselines. Besides, we also achieve the significant result in VizWiz-VQA 2020 and 2021 competitions.
- Visual Question Classification - divide VQA samples into the specific kinds of ques-tions. Dealing with the difficulties on object-less images, we thus propose an Object- less Visual Question Classification model, OL-LXMERT, to generate virtual objects replacing the

dependence of Object Detection in previous Vision-Language systems. Through our experiments in our modified VizWiz-VQC 2020 dataset of blind people, our Object-less LXMERT achieves promising results in the brand-new multi-modal task in comparison to competitive approaches.

- Yes/No Visual Question Answering - solves the speci_c kind of question instead of all kinds of questions. In this task, we point out the importance of Yes/No question types and propose the BERT-RG model which combines the strength of ResNet and VGG to extract the residual and global features to obtain the visual information. By integrating the stacked attention, the relationship of question and images are intensi_ed by the regional features. Through the detail of experiment and ablation studies, our model outperforms the competitive approaches in VizWiz-VQA 2020 dataset and competition.
- General Visual Question Answering - determines the answer in all kinds of questions. In this work, we propose the novel Bi-direction Co-Attention Network to intensify the textual and visual features simultaneously. Besides, we also apply the VT-Transformer to extract meaningful image and text information. Our method Bi-direction Co-Attention VT-Transformer consistently shows strong performance in the VizWiz-VQA dataset. Besides, it also achieves a promising result in the latest competition in VizWiz-VQA 2021.

Besides the success of each sub-task in the above, our hierarchial VQA system also proves the promising performance against the independent VQA architectures in previous works, especially in VQA for blind people.

Keywords: Visual Question Answering, BERT, Vision Transformer, Co-Attention, Answerability, Yes/No Question, VizWiz-VQA, Blind People.

論文審査の結果の要旨

This thesis focuses on Visual Question Answering (VQA) for blind people. The challenge of the research is to deal with noisy and ambiguous data, which causes difficulties in obtaining a high quality of VQA performance. The candidates proposed a novel method and conducted solid experiments on the public dataset. This method shows an excellent combination between text representation and image presentation via a transformer architecture. As a result, the proposed method obtained a good performance in comparison with various published works. In addition, the candidate published a quality journal (Neurocomputing) and a top conference on image processing (ICIP).

The thesis presents three major chapters in which each chapter solves a sub-problem for VQA tasks: Answerability prediction, Classification of question type, Yes/No VQA, and General VQA. As a result, the candidate performed a suitable method for dealing with each sub-problem, which is necessary for enhancing the quality of VQA for blind people. Therefore, the quality of the thesis is sufficient to receive a Ph.D. degree. This is an excellent dissertation, and we approve awarding a doctoral degree to Mr. Le Thanh Tung.