PLAGIARISM SCAN REPORT

Report Generation Date: June 10,2021

Words: 1262

Characters: 9479

Excluded URL :

2%
Plagiarism

98%
Unique

1
Plagiarized Sentences

59
Unique Sentences

Content Checked for Plagiarism

Visual Question Answering for Visually Impaired

1
st Devesh Chandak
IT
Pune Institute of Computer Technology
Pune, India
dschandak@gmail.com
2
nd Omkar Deshpande
IT
Pune Institute of Computer Technology
Pune, India
desphandeomkar77@gmail.com
3
rd Sanya Varghese
IT
Pune Institute of Computer Technology
Pune, India
varghese.sanya@gmail.com
4
th Shubham Mahajan
IT

Pune Institute of Computer Technology
Pune, India
svmahajan8899@gmail.com
5
th Dr. Shweta Dharmadhikari
IT
Pune Institute of Computer Technology
Pune, India
scdharmadhikari@pict.edu

Abstract—The lack of access to basic visual information like text labels, icons, and colors can cause exasperation and decrease independence for the visually impaired. We thus propose a visual question answering system to mitigate concerns about undesired consequences from today's status quo for blind people, that is, relying on able-bodied humans to answer visual questions. We wish to assist blind people to overcome their daily visual challenges and break down social accessibility barriers.
In this paper, we discuss in detail the idea of an open ended visual question answering system to assist visually impaired. The task requires an in-depth understanding of visual and language features, finally, we need to assess the relationship between the two modalities to use co-attention. The proposed model is goal-oriented, stressing on the images and questions generated by blind people.

Index Terms—Visual Question Answer(VQA), SelfAttention(SA), Guided-Attention(GA), Long short-term memory(LSTM), Modular Co-Attention(MCA), Optical Character Reader(OCR), Bottom-Up and Top-Down(BUTD)

I. INTRODUCTION

Vision and language reasoning need the knowledge of visual context, language semantics, and cross-modal arrangements and relationships. Considerable amount of work has been done in both the fields, vision and language. [1] [2]. Despite these distinguished single-modality efforts, studies for the modality-pair of vision and speech, principally, pretraining and fine-tuning are still under development. The proposed model focuses on learning vision and language interactions, specifically for representations of a single image and its explanatory sentence.
The aim of the model is to address the following two tasks: (1) provide answers to the vision based question (2) predict the answerability of the visual question. The overall solution can be divided into two parts: (1) data processing, where we reweigh the label according to the confidence and frequency of the ten answers and (2) we structure the questions with extra semantics, adding objects and attributes embedding from object detection to the language part.
In this task we will be focusing on the VizWiz-VQA dataset. It emerged from a natural visual-language question-answer arrangement where visually challenged people each recorded a spoken question concerning an image, and with each visual question along with ten crowdsourced responses. Compared to the previous version of this dataset this new version contains more train-validation images with local obfuscation which prevents privacy disclosure.

The evaluation metric followed:

Accuracy = min(humans that answered/3, 1) (1)

If minimum 3 humans answered the same answer then the prediction is successful otherwise a partial score is offered. Text, color, counting and object identification parameters were used to evaluate the model.

II. LITERATURE SURVEY

A. Ensembling Rich Image Features

In recent techniques, the visual attention mechanism has been included into VQA by adaptively learning the attended picture features for a specific question and then doing multimodal feature fusion to produce the accurate prediction. [3]. VQA necessitates a thorough comprehension of the semantics of the natural language inquiry in addition to the visual elements of the image.

The Modular Co-Attention(MCA) Layer is the focal point of this technique. The MCA layer is made up of two basic attention units: self-attention (SA) and guided-attention (GA). The model uses GloVe + LSTM to extract question features and uses Faster RCNN to extract image features. Self Attention and Guided Attention are applied to build encoder and decoder. It uses MLP to output these features, which arefused together. The loss function is called Multi label BCE loss.

To get precise convolutional features the architecture resizes an image to get FC features from ResNet 152. Bottom up attention to generate image features using ResNet 101. For question representation the paper uses the following steps. In the beginning a question is trimmed to maximum of 14 words. The vectors from pretrained GloVE which generates word embeddings are fed into an LSTM.

Lastly, by ensembling 14 models the paper achieves an accuracy of 56.20.

This model was the runner up of the VizWiz-VQA 2020 challenge

Fig. 1. SUDOKU model

B. Bottom-Up and Top-Down Attention

To achieve the best results the architecture uses the extensively used Top-Down visual attention mechanism. The architecture is well-accepted for image captioning, image description and answering open-ended visual questions(VQA) combined with Bottom-Up attention to capture more salient features [4]. The top-down method determines feature weightings while the bottom-up method (object recognition) pushes forward sections of the image, each with an associated feature vector.

The highlights of this architecture are:

1) Image Features: A Convolutional Neural Network is used to analyze the input image (CNN). Bottom-up attention results in higher performance. It has been trained to emphasis on specific elements in a particular image.

2) Image Attention: The top-down attention, concatenates image features with the question embedding [5]. Establishing a relation between the input text and image later in the

architecture.
3) Fusion: The representations of the query and the picture are mixed with a simple Hadamard product after passing through non-linear layers.
The model achieves accuracy of 72 on the VQA dataset. The model was the winner of VQA 2018 challenge
Fig. 2. Bottom-Up Top-Down Attention

III. PROPOSED MODEL

The proposed model is an extension of the the up-down model to improve training speed and accuracy. Instead of using the gated hyperbolic tangent activation, the model uses weight normalization followed by ReLU to reduce computation. To compute the question representation, we used GloVe vectors to initialize the word embeddings and then passed it to a RNN network which is the GRU and a question attention module to extract attentive text features, improving the performance of the model from 65.32% to 66.91%.

The model considerably improves the performance of the up-down model on the VQA v2.0 dataset – from 65.67% to 70.22% – by making minor but crucial adjustments to the model architecture and learning rate schedule, finetuning picture features, and providing data augmentation.

Furthermore, by using a diverse ensemble of models trained with different features and on different datasets, Pythia is able to significantly improve over the 'standard' way of ensembling (i.e. same model with different random seeds) by 1.31 %. Overall, we achieve 72.27% on the test-std split of the VQA v2.0 dataset.

Fig. 3.IV. SYSTEM DESIGN

The user is able to click an image and record their questions with taps on their screens. The entire process from clicking and image to getting and answer to your open-ended question is assisted by a voice assistant. The image and the spoken question are fed in to the model and return an answer to the user in speech form. The system has been designed to serve as an assisting technology for the blind and visually impaired. The app's back-end is designed using Django and Flask with React Native front-end.

Fig. 4. System Design

V. FUTURE SCOPE

To improve the architecture an Optical Character Reader(OCR) can be included. It would further increase the viability of the product. With a separate OCR, the app would be able to answer more critical questions and give better results with text based questions.

VI. CONCLUSION

The objective of the paper was to discuss the system design and methodology adopted to design an assisting technology for the blind and visually impaired. We compared different Deep Learning architectures and proposed a real-time application with end-to-end implementation which is user friendly and protects privacy.

Matched Sources :

[1807.09956v2] Pythia v0.1: the Winning Entry to the VQA ...

· Furthermore, by using a diverse ensemble of models trained with different features and on different datasets, we are able to significantly improve over the 'standard' way of ensembling (i.e. same model with different random seeds) by 1.31%. Overall, we ...

3%

https://arxiv.org/abs/1807.09956v2 (https://arxiv.org/abs/1807.09956v2)