VQA for Cartoons: Parameter Tuning to Maximize Performance for Efficient VQA

Rishant Sharma*
s222458666@deakin.edu.au
Deakin University
Melbourne, Victoria, AU

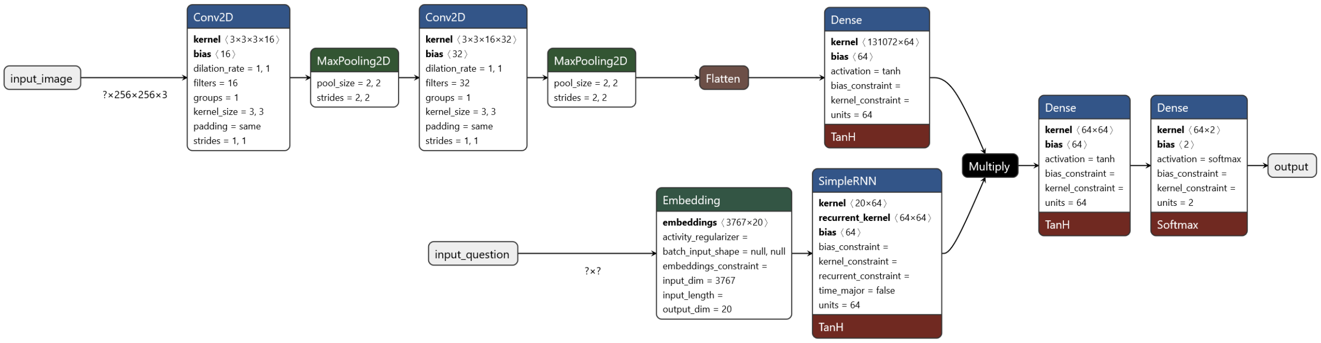


Figure 1: Model Architecture at a Glance

ABSTRACT

Visual Question Answering (VQA)[1] is a task that combines computer vision and natural language processing. It involves answering questions about an image using AI algorithms. Given an image and a related question in natural language, VQA models analyze the visual content and comprehend the question to produce a text-based answer. The process includes extracting features from the image and encoding the question, which are then combined to generate the final response. In this competition, we trained a VQA model that achieved 58% accuracy in answering questions related to images from the cartoon 'Simpsons'. This was achieved by taking a combination of Convolutional Neural Networks and Recurrent Neural Networks and maximising their performance through parameter tuning. The link to the solution canbe found here: Link to Solution

CCS CONCEPTS

• **Computing methodologies → Neural networks**;

KEYWORDS

computer vision, neural networks, natural language processing, visual question answering

Deakin Simpsons Challenge 2023, Mel,VIC,AU,
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Rishant Sharma. 2023. VQA for Cartoons: Parameter Tuning to Maximize Performance for Efficient VQA. In *Proceedings of August 2023 (Deakin Simpsons Challenge 2023).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, Visual Question Answering (VQA) has emerged as a compelling interdisciplinary research area that seeks to integrate computer vision and natural language processing. VQA aims to enable artificial intelligence systems to comprehend visual content and answer questions about images using advanced algorithms. This fusion of visual perception and language understanding holds great promise in various domains, ranging from robotics and autonomous vehicles to accessibility technologies.

The motivation behind this research stems from the need to explore and evaluate VQA in the specific context of images from the animated series 'Simpsons'. By focusing on questions related to scenes from the show, specifically narrowed down to Yes/No type questions, we aim to showcase the capabilities of VQA models.

The primary objective of this competition is to train a VQA model that achieves high accuracy in answering questions about images from 'Simpsons'. To achieve this goal, we adopt a combination of Convolutional Neural Networks (CNN) and Simple RNNs as the underlying architecture. Furthermore, we emphasize the significance of parameter tuning to realize an efficient model that maximizes the performance of the model while reducing the number of total parameters as much as possible.

This paper presents our approach, experimental setup, and results in detail. We outline our methodology, including the VQA model architecture and the techniques used for parameter optimization. Subsequently, we describe the dataset used for training and evaluation, followed by a discussion of the results.

2 APPROACH

As noted in the introduction, the task of VQA involves a combination of computer vision and natural language processing[2]. However, compared to conventional VQA models, which have focused on extracting information from real-life images to answer questions, the required model complexity isn't as high here as cartoon images are simplified and are closer to a combination of shapes rather than real objects which are more complex and hence require more parameters for a usable model. Hence, we use a simple Image Classification CNN for the Image component and a Simple Recurrent Neural Network for the Question and Answer component. The outputs of these networks were element-wise multiplied and then passed through two dense layers to output the probability of both 'yes' and 'no'.

2.1 Image Layer

A Convolutional Neural Network (CNN) is used as it is a specialized deep learning architecture designed for image and visual data processing. It consists of convolutional layers that apply filters to capture spatial patterns, followed by pooling layers for downsampling. The extracted features are then flattened and passed through fully connected layers for classification or regression tasks, enabling CNNs to excel in computer vision applications.

The CNN part in our model processes the image input through two convolutional layers with 16 and 32 filters respectively, followed by max pooling layers for downsampling. The extracted image features are then flattened and passed through a dense layer with 64 neurons activated by the hyperbolic tangent ('tanh') function.

2.2 Question Layer

A Recurrent Neural Network (RNN) is a type of neural network designed to handle sequential data, such as time series or natural language. Unlike feedforward networks, RNNs have recurrent connections that allow information to persist over time.

For the language input in our model, we use an RNN with a SimpleRNN layer of 64 units processes the sequential question data after it has been passed through the embedding layer to convert it to a feature vector of length 20.

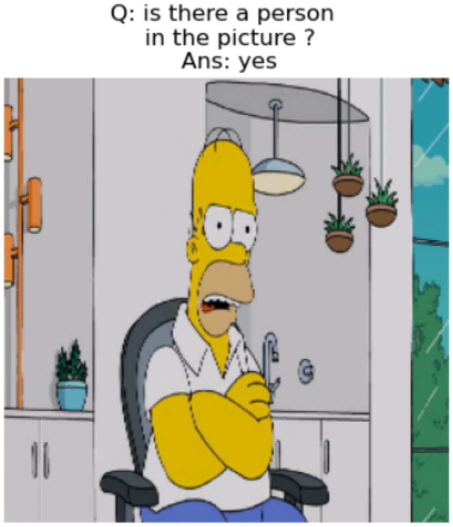
2.3 Combination and Output

The model then fuses the image and question features using element-wise multiplication and applies another dense layer with 64 neurons activated by 'tanh'. Finally, the output layer predicts the probability distribution of answers using softmax activation.

3 DATA

The training dataset consisted of 24457 image-question pairs with answer annotations. The images were resized to a 256 by 256 resolution with 3 channels and the pixel values were normalized to [0,1] range of float values.
The questions were pre-processed using standard NLP practices such as converting to lowercase and removing unnecessary punctuation marks from the string. The questions were then tokenized using the NLTK library and the tokens were stored in the Keras vocabulary for later embedding. The model had a vocabulary size of 3767.

Figure 2: Example of Image-Question-Answer sample data



4 FINE-TUNING

The original model upon which our solution is based is the model described in the Easy VQA article by Victor Zhou[3]. The original model was developed for images that contained just shapes and hence could not capture all the features of the competition dataset. We increased the Image Model's Conv2D filter sizes to 16 and 32 respectively with a kernel size of 3 and the padding as same to preserve spatial dimensions and doubled the final Dense Layer's neuron units to 64. The 'tanh' activation function was chosen as it gave best results in the test set.

The Question Model was different from the article's question model as our approach was different. We used Embedding rather than Bag of Words to create the input vector of the RNN of length 20. The RNN had 64 neurons to compliment the output of the Image Model and allow for Element-Wise Multiplication for the merging of the two models' outputs. Further experimentation on the parameters of the RNN could not be done due to lack of time and may serve as a direction for future improvement.

A further Dense layer with 64 neurons and a 'tanh' activation was added after the multiplication layer to improve the model's memory.

These parameters allowed for the model to achieve a final test accuracy of 58% while having a total parameter count of 8,478,830, keeping it in line with our efficiency-oriented approach.

Lastly, a batch size of 128 and and Adam Optimization with a learning rate of 2e-4 were found to provide the best training results.

5 CONCLUSION AND FUTURE WORK

At the beginning of the project, our goal was to develop a model that could exceed 70% in accuracy which would allow it to classified as a model usable for actual application. Due to lack of time due to concurrent research and development of other approaches such as Ensemble Methods and Pre-Trained Models, there still lies scope for further improvement through tuning of parameters like stride, kernel and bias regularizations and dropout.

In future, I would like to further explore these hyper-parameters and would like to study other approaches such as transformers, which can improve upon some of the limitations of our approach, such as the vanishing gradient problem of RNNs.

Additionally, I would use larger datasets like VQA v2 for training as the dataset provided to us was relatively small compared to other larger VQA datasets available in the public domain.

REFERENCES

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. VQA: Visual Question Answering. arXiv:1505.00468 [cs.CL]

[2] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and Answering Binary Visual Questions. arXiv:1511.05099 [cs.CL]

[3] Victor Zhou. 2020. *Easy Visual Question Answering: A gentle introduction to Visual Question Answering (VQA) using neural networks.* Retrieved July 25, 2023 from https://victorzhou.com/blog/easy-vqa/

Received 6 August 2023