Text Representation Enhancement using multi-modal attention mechanism for Text Visual Question Answering

ODUNAYO ESTHER ODUNTAN[1] AND JOSEPH MARIE DOMGUIA,[1]

[1]*African Masters in Machine Intelligence (AMMI, Rwanda)*
Thesis Report, 2020
eoduntan@aimsammi.org
jdomguia@aimsammi.org

1. Abstract

Text Visual Questions and Answering is an integral part of communication. It takes the respondent the ability to comprehend what the speaker is asking and relate the text to visual objects in the scene including textual information, there to be an effective communication. This cuts across divers discipline, in which machine learning has a role to perform in ensuring that questions are understood and well answered. Researchers in Natural Language Processing and Computer Vision have been used various architectures such as Bottom-up and Top-Down Visual Attention Mechanism, Visual-BERT, M4C to mention a few. In this research, we looked into the application of multi-modal attention mechanism to improve the question and answering system. We use multi-modal transformer with auto regressive answer generation to process the Text, Image and Optical Character Recognition(OCR). Three approaches were proposed (Evaluating the pre-train model, fine-tuning the model with Google OCR and training from scratch). The first two approaches were implemented while the third is intended for future work. Hence,it was observed that the quality of answer generated from the pre-trained MC4 model with an improved OCR will enhance the performance for text visual question and answering system.

2. Introduction

Communication is simply the act of transferring information from one place, person or group to another. Every communication involves (at least) one sender, a message and a recipient [1]. The ability for the three components of communication to interact effectively leads to effective communication. In addition, emotions, the cultural situation, the medium used to communicate, and even our location are very important. Communicating can be in form of questions and answers, which can be textual or visual.

To researchers in Natural Language Processing and Computer Vision [2], building a system that can answer natural language questions about any image has been considered a very ambitious goal. Although as humans we can normally perform this task without major inconveniences, the development of a system with these capabilities has always seemed closer to science fiction than to the current possibilities of Artificial Intelligence (AI). However, with the advent of Deep Learning (DL), there has been an enormous research progress in Text Visual Question Answering (VQA), in such a way that systems capable of answering these questions are emerging with promising results.

Visual Question Answering (VQA) system can be defined as an algorithm that takes as input an image and a natural language question about the image and generates a natural language answer as the output [3, 4]. This is a computer vision task where a system is given a text-based question about an image, and it must infer the answer. The main idea in Text VQA is that the search and the reasoning part must be performed over the content of an image. The system must be able to detect objects, it needs to classify a scene and needs world knowledge, commonsense reasoning and knowledge reasoning are necessary. Many of these tasks (object recognition,

object detection, scene classification, etc.) have been addressed in the field of Computer Vision (CV), with impressive results in the last few.

Attention mechanism in [5] provides a way of capturing question relevant regions and representing desired visual information accurately. Attention mechanism has become a standard configuration for VQA models. [6, 7] [8].The inputs into the VQA are the multiple regions representations and a question representation, the correlations values between them are calculated and the larger values denotes the corresponding region with more relevancy to the question. The bottom-up mechanism proposes a set of salient image regions, with each region represented by a pooled convolution feature vector. The top-down mechanism uses task-specific context to predict an attention distribution over the image regions, while Faster R-CNN was used to represent a natural expression of a bottom-up attention mechanism [7] [5] demonstrated a great success on image captioning and VQA tasks for fine grained visual representation.

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions [9]. The ability of the Multi – head attention to correctly align the questions on text within the image was a challenge. This is an essential semantic gap between them, which is often overlooked by attention-based methods.

2.1. Problem Statement

VQA requires intensive analysis and understanding of images and questions, including image recognition, object localization, attribute prediction, text tokenization, word representation, object counting and knowledge inference [10]. Visual Question-Answering has appeared as a problem where models need to be able to perform different sub-problems of the above three fields in order to succeed. To achieve acceptable accuracy close to human judgment, there is need for a model to have a deeper understanding and comprehension of the image and the input question.

Fig. 1. Predicted images from existing VQA model

It was observed that image features were detected correctly in figure 1, but the VQA model couldn't adequately answer the questions as it relates to proper understanding of the objects in the extracted image features. This could be traced to collapsing of the visual features into one attention feature vector which could have limited the VQA model ability to identify multiple occurrences of the objects.The input features of the multi-head attention models stem from different model data [11].

2.2. Objective

To provide a solution to the stated challenge, this study focuses on the region representation in the multi-attention, in addition, the idea of text representation enhancement: introducing the optical character recognition(OCR) model to multi-head attention was examined making use of the transformer architecture. Transformer positional embedding which could handle semantically more complicated questions was integrated. Hence, the system will be able to handle questions about attributes or fine grained types by providing the answers to questions relating to caption on the image irrespective of the direction of the text.

3. Related Works

In recent times, researchers in the field of computer vision and natural language processing have looked into solving divers challenges in Visual Questions Answering System. This section delves into review of related researches in VQA.

[5] combined bottom-up and top-down visual attention mechanism. The bottom-up mechanism proposes a set of salient image regions, with each region represented by a pooled convolution feature vector. The top-down mechanism uses task-specific context to predict an attention distribution over the image regions, while Faster R-CNN was used to represent a natural expression of a bottom-up attention mechanism.

However, it was observed that image features were detected correctly, but the VQA model couldn't answer accurately the questions on each object in the extracted image features. This could be traced to collapsing of the visual features into one attention feature vector which could have limited the VQA model ability to adequately comprehend the multiple occurrences of the objects.

[12]introduce a novel model architecture that reads text in the image, reasons about it in the context of the image and the question, and predicts an answer which might be a deduction based on the text and the image or is composed of the strings found in the image. They tagged their approach, Look, Read Reason and Answer(LoRRA). According to Singh et.al(2019), the VQA model use some variant of attention to get a representation of the image that is relevant for answering the given question. The object region proposals and the associated features are generated by using a detection network which are then spatially attended to and conditioned on a question representation. Optical Character Recognition(OCR) was used to either detect text on images or provide information about anwers to questions.

However, with introduction of OCR, the LoRRA model or approach was able to reason with an image that answer relevant questions, but fails at detecting text that is rotated and was unable to count number of objects within a region of interest(ROI).

[13]came up with model based on a pointer augmented multimoldal transformer architecture with iterative answer prediction. The model projects the feature representations of questions words, detected objects and detected OCR tokens from the three modalities as vectors in learned common embedding space and apply multiple transformer layer over the projected features. It learns to predict the answer through iterative decoding accompanied by a dynamic pointer network. During decoding, it feeds in the previous output to predict the next answer component in an auto-regressive manner. At each step, it either copies an OCR token from the image, or selects a word from its fixed

4. Approach

The method used in this study, is a multi-modal technique that involves more than one input and outputs. The inputs are the questions, images and optical character reader (OCR) extracted features. The images features were extracted using the Faster R-CNN model, the contextual word embedding features from the questions were extracted using global vec(GLO Vec) and text on images were extracted using the google ocr-app. All the input features were passed into the transformer. The dataset that was used is the TextVQA dataset.

4.1. Proposed Architecture:

This architecture comprises of three(3) stages: the input, processing and output. The input involves the image features are extracted with Fast R-CNN; OCR features was the extracted from image but only the bounding box around the text; the Question was represented by a sequence of token encoded with GloVec representation and the tokenization of the text features from the questions. The processing phase involves the passing of the input features into the transformer

model, both self and multi-head attention mechanism are applied on the input features.The output vector from the begin token give a vector representation of the answer, this vector representation is passed through the projection layer that will give predicted token from the vocabulary or the OCR from the detected OCR on the image.The decoding is auto-regressive meaning that the output is given again as input to generate the output token from the vocabulary or OCR until END token was obtained.

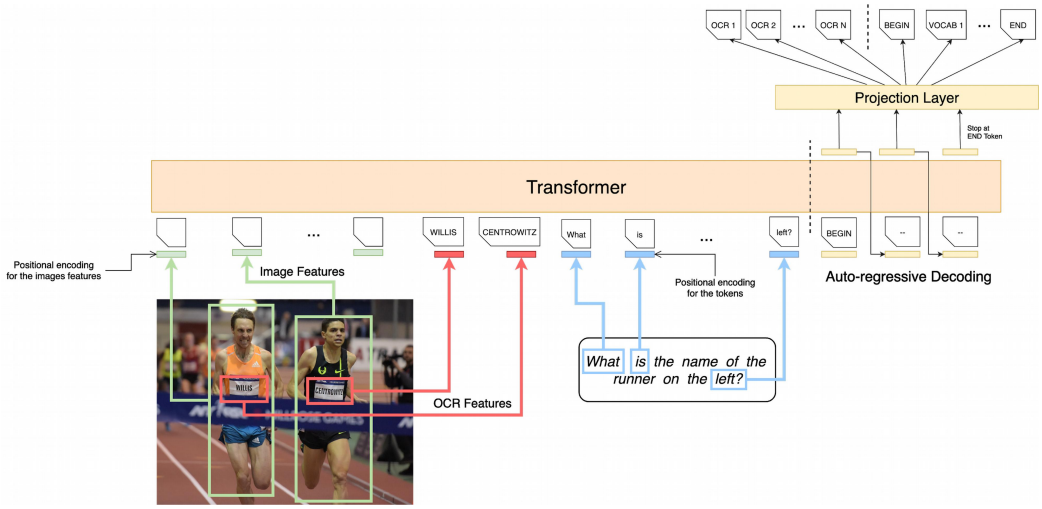

Fig. 2. The Multimodal OCR-Transformer Based VQA Architecture

In the transformer both the self attention and the multi-head attention mechanism were used. In an attention based VQA models, attention mechanism takes as input the question and every image region feature, and computes correlation values between them. After normalization, the attention weights indicates the probability of question selection on the image regions. The visual feature is the sum of all these region features weighted by their attention weights, and this is fed to later fusion with question representation. In a multi-head attention, one set of Q, K, V matrices is called an attention head, and each layer in a Transformer model has multiple attention heads. While one attention head attends to the tokens that are relevant to each token, with multiple attention heads the model can learn to do this for different definitions of "relevance".

However, the Text VQA model was implemented using co-attention mechanism based on the transformer architecture, by exchanging key-value pairs in multi-headed attention has limited reading and counting capabilities. Experiment carried out in this study shows that improvement on the ability to read text on images can increase the performance of the TextVQA model. Hence, the introduction of an improved OCR model for text representation for TextVQA Dataset.

4.2. Transformer

4.2.1. Self-Attention

In this study, three vectors were created from each of the multi-modal inputs. For the questions, image and ocr features, a query vector, key vector and a value vector were created. The vectors were created by multiplying the embedding by the matrices of extracted features. This was followed by calculating the scores of each input, which helps to determine how much focus is to placed on the various parts of the input features. The score was calculated by taking the dot product of the query vector with the key vector of the input feature to be scored. To have more stable gradients, the scores were divided and passed through a softmax operation, which nomalized the scores to sum up to 1. The input feature with the highest softmax score will be

focused. Each Value vector is multiplied by the softmax score to keep intact the values of the input features to be focused on. Then, the weighted value vectors are added up to produce the result of the self attention layer at that position. Hence, the resulting vector is feed forward to the neural network.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

4.2.2. Multi-Head Attention

To improve the performance of our architecture, a multi-head attention mechanism was performed. This was done to enable the Text VQA model to be able to focus on different positions in the input features. The Attention layer of the transformer was able to use multiple representation subspaces. In other words, multiple sets of Query, Key and Value weight matrices were used, that is our transformer used eight sets for each encoder/decoder. Each of these representations were used on the input features to project the input embedding into different representation subspace. Performing same self attention in eight different times with eight different weights matrices, resulted in eight different outputs. Hence, the matrices from the different heads were concatenated and passed into the feed-forward neural network layer.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)W^o \tag{2}$$

where $head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$
where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_K}$,$W_i^K \in \mathbb{R}^{d_{model} \times d_K}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_V}$ and $W^Q \in \mathbb{R}^{hd_v \times d_{model}}$

4.2.3. Position-wise Feed-Forward

Networks In this architecture, each of the layers in the encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between. The linear transformation are same across different position. $FFN(x) = max(0, xW1 + b1)W2 + b2$

4.2.4. Transformer Positional Encoding

This indicates the position of each word in the sequence. Since our model contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, we must inject some information about the relative or absolute position of the tokens in the sequence. To this end, we add "positional encodings" to the input embeddings at the bottoms of the encoder and decoder stacks. The positional encodings have the same dimension d model as the embeddings, so that the two can be summed. The transformer adds a vector to each input embedding, which follows a specific pattern that the VQA model learns to determine the position of each feature.

4.2.5. The Softmax Layer

The output of the decoder are vectors called the logits vector. The scores from the softmax are turned into probabilities which sums up to one. Hence, the feature with highest probability is chosen and the answer associated with it is produced as the output of this step.

4.3. Visual Question Answering Model (VQA MODEL)

VQA is a system that takes as input an image, an open ended, natural language question about images and produces a natural language answer as the output. Details of the VQA model features are discussed below:

4.3.1. Image Features Representations

Faster RCNN for image feature extraction and the Transformer Attention (Image as K, extracted question features as Query and output of the encoder as value) for aggregation of the extracted images features and question encoder output, which will be forwarded into the decoder. Image region features were generated by extracting bounding boxes and their visual features. Unlike words in text, image regions lack a natural ordering, hence spatial location was encoded by constructing a 4-d vector from region position (normalized top-left and bottom-right coordinates). This is then projected to match the dimension of the visual feature and they are summed.

4.3.2. Text Features Extraction:

In this section, a block of text were extracted from the images, text blocks and respective OCR text were represented as as features for VQA task. Text features were extracted using Global Vector(Glo Ve) model. This was used because text features will be extracted from the whole image features of the dataset. The GloVe model is trained on the non-zero entries of a global word-word co-occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus. Populating this matrix requires a single pass through the entire corpus to collect the statistics. For large corpora, this pass can be computationally expensive, but it is a one-time up-front cost. Subsequent training iterations are much faster because the number of non-zero matrix entries is typically much smaller than the total number of words in the corpus.

4.3.3. OCR Text Extraction Features

In recent researches, transformers have been used in natural language processing(NLP) tasks.Text features were extracted from a pre-trained (Google OCR API). It was used to solve the problem of optical character recognition(OCR). This invloves two modules: feature extraction and transformer module. The text features were represented as word embeddings and used as input to the transformer, hence the efficiency of attention mechanism of transformers are fully utilized.

In this study, the OCR was added to the existing attention mechanism based on the transformer architecture an approach which was earlier used by [11]. Key-value pairs in multi-headed attention was exchange, which enabled vision-attended language features to be incorporated into visual representations. A transformer is an encoder-decoder architecture model which uses attention mechanisms to forward a more complete picture of the whole sequence to the decoder at once rather than sequentially. TextVQA dataset comprising of images with text was used. The experiment comprises of three(3); the extracted image features, the questions and the OCR word embedding features.

5. Experiment

This section gives a description of how the VQA model was trained with the improved OCR for text representation.

5.1. Data Collection: TextVQA Dataset

TextVQA requires models to read and reason about text in an image to answer questions based on them. In order to perform well on this task, models need to first detect and read text in the images. Models then need to reason about this to answer the question. Each question-image pair has 10 ground truth answers provided by humans Study shows that some state-of-the-art models fail to answer questions in TextVQA because they do not have text reading and reasoning capabilities. See the examples in the image to compare ground truth answers and corresponding predictions by a state-of-the-art model.

TextVQA v0.5.1 contains 45,336 questions based on 28,408 images. The v0.5.1 training set contains 34,602 questions based on 21,953 images from OpenImages' training set. The v0.5.1

validation set contains 5,000 questions based on 3,166 images from OpenImages' training set while the v0.5.1 test-std set contains 5,734 questions based on 3,289 images from OpenImages' test set.

5.2. Features Extraction

Features extracted from the dataset are images, questions and ground truth answers. The image features were extracted using the Faster RCNN architecture, it comprises of images with label and others were images without labels. Labels on images were extracted using the optical character reader(OCR). Another features extracted is the questions, it was categorised into two: questions that requires a general answer such as: "What is on the table?" was grouped into a dictionary known as "VOCAB" while questions such as: "What is the plate number of the bus" was categorised as "OCR, VOCAB" dictionary. The ground truth answers provided by the TextVQA dataset were also extracted and predicted answers were generated. In this study, both datasets stated above were used for experimenting with transformer based multi-modal VQA.

5.3. Experiment Settings

In this study, three approaches were intended; and are stated thus:

5.3.1. Approach One: Evaluation of M4C Pre-train Model with Improved OCR

The pre-train model for M4C was evaluated with a improved OCR using blue-score. It was expected that there will an improvement on the accuracy without doing any training. This is because Rosetta Model used the google OCR and got better performance.

If our hypothesis is wrong and the validation is not better, we can check if we the second experiment. It can be wrong because the M4C model train on Rosetta OCR perform well only on text similar to the training data with Rosetta OCR but perform on another OCR even this one is better.

5.3.2. Approach Two: Fine tune the pre-train model for M4C with the improved OCR

This approach will help us to see if fine-tuning the model on the training with better OCR (better bounding boxes, better character recognition) will enable the model to improve its ability to understand what going on in the scene and provide more accurate result.

5.3.3. Approach 3: Train from scratch for M4C with a improved OCR

This last approach have to better than all the previous approach. The hypothesis there is to train from scratch with the new OCR, the model will learn how to use the improved OCR in the best way.

5.4. Training

Approach one of evaluating the MC4 pre-trained model with google OCR was attempted, because of compute resource we were able to undertake this approach,the other two(2) approaches are reserved for future experiment due to insufficient computational resources. This is how we planned to run the experiment. The Attention model is big with 90 millions parameters. We attempted Losses function binary cross entropy with mask (reference m4c) For the text, we use the Bert tokenizer with 10 possible answers according to the data. For the image, the features extracted from the image using a customized version of Faster-RCNN (https://github.com/ronghanghu/vqa-maskrcnn-benchmark-m4c from the m4c author) We can use Adam optimizer and a batch size of 128 ideally. But during our trial we were able to experiment with a batch size of 96, with a Nvidia Tesla K80.

6. Results and Evaluation

6.1. Baselines M4C:

The three input features comprising of question words, detected visual objects, and detected OCR tokens were projected into a common d-dimensional semantic space through domain-specific embedding approaches and apply multiple transformer layers over Based on the transformer outputs, the predicted answers were generated through iterative auto-regressive decoding, where at each step the model either selects an OCR token through a dynamic pointer network, or a word from its fixed answer vocabulary.

During training, the multimodal transformer at each decoding step was supervised and similar to sequence prediction tasks such as machine translation, was performed using ground-truth inputs to the decoder to train the multi-step answer decoder, where each ground truth answer was tokenized into a sequence of words. Given that an answer, word can appear in both fixed answer vocabulary and OCR tokens, a multi-label sigmoid loss (instead of softmax loss) over the concatenated scores were applied.

6.1.1. Baseline Results and Evaluation:

The output of the prediction were evaluated using blue score and accuracy metric. Table 1, show the first five predicted results from the TextVQA dataset, categorizing the predicted source as [VOCAB], [VOCAB OCR], [OCR OCR], [OCR], [VOCAB VOCAB] for each predicted answer. Figure 3, 4 shows the visualization of some of the prediction. The Blue Score gave 85.29913857414708,The Accuracy Score gave 42.72;

6.2. Visualization of Google OCR API

From the implementation, it could be observed that integrating the google OCR into MC4 model will enhance the representation of text; hence the visual question and answering system will be enhanced.

Figure 4 shows the visualization result of the integration of google OCR into MC4 model. it was observed that the system was able to read text in diverse directions, which was difficult to achieve with the MC4 model on TextVQA dataset.

The evaluation result shows using blue score and accuracy metrics, Google OCR implemented on MC4 shows a better representation of the text.

6.3. Evaluation Result of the pre-train model for M4C with the improved Google OCR and Rosetta OCR

The output of the prediction were evaluated using blue score and accuracy metric. The goal is to evaluate if the train(fine-tune) with Google OCR will overcome the over-fitting of Rosetta model. The fine-tuning the model on the training with better OCR (better bounding boxes, better character recognition) will enable the model to improve its ability to understand what going on in the scene and provide more accurate result. Though, the computation was faced with resources challenges which possibly might have effect on the evaluation result and accuracy. The Blue Score gave 39.43 and 30.42,The Accuracy Score gave 42.72 and 31.58 respectively for pre-trained M4C fine-tuned with Rosetta OCR and Google OCR respectively.

7. Conclusion

This project has been able to implement a multi-modal attention mechanism for visual question answering system, by integrating Google OCR and Rosetta OCR on an M4C pre-trained model. Out of the three approaches intended to use in enhancing text representation, two of the approaches were implemented, while the third approach of training from scratch and fine-tuning the pre-trained M4C model will be examined for future work.

Fig. 3. The result of VQA model using M4C Architecture baseline

Fig. 4. The google-ocr API was able to correctly read "Grain" though it is rotated as shown in the diagram

References

1. S. Coleman, “Book Reviews,” J. Commun. **67**, E7–E8 (2017).

2. Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (2017), pp. 6904–6913.
3. D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual turing test for computer vision systems," Proc. Natl. Acad. Sci. **112**, 3618–3623 (2015).
4. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision,* (2015), pp. 2425–2433.
5. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2018), pp. 6077–6086.
6. J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in neural information processing systems,* (2016), pp. 289–297.
7. J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," arXiv preprint arXiv:1610.04325 (2016).
8. K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2016), pp. 4613–4621.
9. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems,* (2017), pp. 5998–6008.
10. Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2016), pp. 4622–4630.
11. J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems,* (2019), pp. 13–23.
12. L. Gómez, A. F. Biten, R. Tito, A. Mafla, and D. Karatzas, "Multimodal grid features and cell pointers for scene text visual question answering," arXiv preprint arXiv:2006.00923 (2020).
13. R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* (2020), pp. 9992–10002.