Multilingual TextVQA: Towards generalization of TextVQA to Low-Resource Languages

Ajay Narayanan
TCS Research and Innovation
Bangalore, India
ajay.narayanan@tcs.com

***Abstract*—Comprehensive understanding of a visual scene requires understanding of the text embedded in them. Recent work has explored the TextVQA task that requires reading and understanding text in images to answer questions about them. However, current approaches are largely limited to a single language, both in terms of the text in the image and that of the questions. In this work, we propose a novel paradigm for evaluation of TextVQA model i.e. multilingualism. This proposal aims to understand the extend to which current TextVQA models can handle multiple languages and to improve on that baseline.**

***Index Terms*—vision and language, TextVQA**

I. Introduction

As a key task in bridging vision and language, the Visual Question Answering(VQA) task [1] has received wide attention from the research community, both in terms of datasets(e.g [1], [2], [3], [4], [5]) and methods(e.g [1], [6], [7], [8], [9], [10]). However, these data sets or models mostly ignored the modality of text in the image for scene understanding and reasoning. To address this drawback, a new VQA task with questions that explicitly rely on understanding the text in the image, which is more commonly referred to as TextVQA. A new set of data sets [11]–[13] was proposed with questions that rely on reasoning and understanding of the text in the image.

The TextVQA task is more complicated than general VQA task because it requires the model to see, understand and reason about the text in the image, the visual contents of the image and the input question. Several approaches [11]–[14] make use of OCR results of the image. In particular, [11] extends previous VQA models [15] with an OCR attention branch and adds OCR tokens to the answer classifier. Similar work is done in [13] ,where OCR tokens are grouped together and added to the output space.

While these approaches enables the model to read the text in the image to some extent, they are limited by the language in the text. Usually, models are trained to recognise only English based OCR tokens, which limits the capabilities of such approaches. Furthermore, the input questions are also based on the English language alone and does not offer the flexibility of multilingualism.

This proposal expands the scope of TextVQA and aims to study the effect of multilingual constraints on the current SOTA algorithms. A naive approach to deal with multilingualism in the TextVQA problems would be to use a translator module which translates both the text in the input question and the text in the image to a single language i.e English and uses this as input for further processing. Although this might be trivial, there might be serious errors in the image understanding and question understanding caused due to the inherent errors in the translation modules.

Another key aspect to be explored is on how TextVQA can be achieved on Low-Resource Languages, where it will be a challenge to even find a viable Translator module. One key aspect of Low-Resource Languages is the lack of available data to train a neural translator or a TextVQA model from scratch. Therefore, for the generalization of TextVQA to such low resource language, we will have to explore methods like meta-learning whose objective is to learn a new task with minimal amount of data.

II. Problem Statement

One of the issues with the current state of TextVQA data as well as the techniques for solving them is that they are based only on a single language, English. Due to this, SOTA models and architectures at present lack a particularly important aspect of AI, generalization. As the title suggests, I want to explore how TextVQA can be extended to Low-Resource languages through techniques like meta-learning, domain adaptation or transfer learning. One possible trivial solution would be to just translate all text to English, but I feel that there might be serious and silly errors in this method due to loss of knowledge during translation. Another technical challenge is the lack of good data for some languages and the abundance of it for other languages. How to handle this imbalance of data will be an interesting aspect of this problem.

Some of the key problems this proposal wishes to address are: 1) How effective is current SOTA models in addressing multilingualism in data? 2) How does adding simple Translator module affect the performance of TextVQA modules in a multilingual scenario? 3) How to generalize TextVQA to low-resource languages, where translation might not be possible? 4) Is it possible to do better than the trivial solution of Translation through methods like meta learning or the creating of a common embedding space? 5) Can these techniques be used to build a Universal TextVQA model?

III. Motivation

The main motivation for this problem mainly arises from the fact that I am from India, a country with more than 21 languages with each state having its own language, where most of the languages do not have much data in the format required to perform TextVQA. Like India, the variety of languages in this world is staggering. Most of these languages are not even properly documented. As discussed earlier, for any TextVQA model to be practical in the real world, I think it is crucial that the model must handle the wide variety of low-resource languages in the world. Another key reason to look at this problem at this juncture, is due to huge amount of research happening in fields or transfer learning, meta-learning, and domain adaptation. The advances made in these fields combined with the research already being done in the field of TextVQA should be able to give rise to a truly general purpose TextVQA agent.

IV. Preliminary Literature Review

A. TextVQA: Approaches and datasets

Recently, a number of datasets and approaches [11]–[14] have been proposed for the task of visual question answering based on the image in the text (commonly referred to as TextVQA task). LoRRA [11], extends the Pythia framework [15] for VQA by allowing it to copy OCR tokens to the answer, by applying single attention hop(conditioned on the questions) over the input OCR tokens. In a similar approach proposed in [14], OCR tokens are grouped into blocks and added both to the input and output space of the VQA model. Other methods [12], [14], augment existing VQA models with OCR tokens.

Current SOTA algorithms use Multimodal approaches to solving TextVQA task. For instance, in [16], a multi-modal transformer architecture is used to fuse multiple modalities to a common semantic space, where self attention is applied to infer inter- and intra-modality context. Conceptually similar work is done in [17], where spatially aware self attention is applied to the spatial graph produced by a transformer architecture. A structure multi modal attention(SMA) network is proposed in [18], which uses a structural graph representation to encode the object-object, object-text and text-text relationships appearing in the image, and then design a multimodal graph attention network to reason over it.

B. Multilingual Language models

With the advent of Transformer-based language models , language representation models such as [19] has made it possible to create multilingual language modes. Multilingual-BERT or M-BERT [19] has been studied from various angles in recent years. In [20], M-BERT is shown to be surprisingly good at zero-shot cross-lingual model transfer, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. They also found that there were systematic deficiencies in the representations affecting certain language pairs. In a similar study [20], it was shown that lexical overlap between languages plays a negligible role in the cross-lingual success, while the depth of the network is an integral part of it.

In [22], it is shown that currently available multilingual BERT model is clearly inferior to the monolingual counterparts, and cannot in many cases serve as a substitute for a well-trained monolingual model. Similar results are obtained in [23]. An unsupervised cross-lingual word-embedding is proposed in [24], where word embeddings of each language are mapped into a common latent space, making it possible to measure the similarity of words across multiple languages.

V. Challenges

The first and foremost technical challenge that this project must handle is the unavailability of multilingual TextVQA data. Since almost all the TextVQA data available today is focused on English alone, a wide and representative data curation must be the first step of this project. The 2nd important technical challenge is the imbalance of usable data from different languages. Some languages might be overrepresented, and others might be underrepresented. Unlike classic problems like classification or object detection, where data augmentation is quite trivial, how to solve the language wise imbalance of the data is a good research question.

Apart from the technical challenges involved in the data curation, there are equally tough challenges when it comes to the solution. How to incorporate the advances in the field of transfer learning, meta-learning etc. will be a research problem. Meta learning algorithms like MAML have been quite successful is simple problems like supervised learning. There are 3 main combinations possible in this:

1) Text in Image is in English, question is in Non-English Language.

2) Text in Image is Non-English, question is in English.

3) Both Text in Image and question are Non-English (both can be either in the same language or different language).

I think these different combinations of possible multilingualism in TextVQA problems will serve as the technical milestones in this problem.

VI. Milestones

1) **Data curation**: The first step in solving the multilingual TextVQA problem is curating a representative data set of images with text in multiple languages. For each image, we will need to create a set of questions both in English and Non-English languages. Here, we must address the challenges involved by the imbalance of the data and find some possible solutions or techniques for data augmentation. The main issue will be in the data augmentation of the images since the augmentation of questions in multiple languages can be easily done through translation.

2) **Text in Image is English, question is Non-English**: As discussed before, the first milestone will be handling questions in multiple languages about images with English text on them. This part of the project can be done easily with the existing dataset and formulating questions in multiple languages. The Image Understanding part of the model will remain unchanged

while we make changes to the Question Encoding part. One interesting research problem that we will encounter during this milestone, the how whether we can create a universal attention model, like attention models for the English language. It is important to note here that the grammatical structure of the source language might be a key to providing the correct answer. The grammatical structure is something that is usually lost is translation, so simply translating the question to English might not be the best solution.

3) **Text in Image is Non-English, question is English**: The 3rd technical milestone in this problem will be to look at pictures whose text is in a Non-English language and the questions are in English. For this part of the project, a significantly new dataset must be formed with images with non-English text. In this part, the Question Encoding part remains unchanged while we make changes to the Image Understanding part. One interesting research problem that we must answer during this step is on how to build a Universal Image Encoder, that can 'read' multiple languages. This will the toughest part of the exercise and requires quite of lot of expertise. Like above, simple translation might lead to misinformation which in turn will affect the performance.

4) **Combining the Universal Image Encoder and Universal Question Understanding modules**: The last step will be the combination of the 2 previous steps to provide an end-to-end solution to multilingual VQA. A successful combination of the previous 2 results should be able to gracefully handle 3rd type of data points, where both the Image in Text and the Question are in different languages. Along with this, this final model or Universal TextVQA model should be easily able to handle all the other types of data points described as well.

I believe each of the technical milestones can be presented at top-tier conferences and requires quite a bit of research into meta learning, computer vision and natural language process.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, MargaretMitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh.Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, pages 2425–2433, 2015

[2] Yash Goyal, Tejas Khot, Douglas Summers Stay, Dhruv Batra, and Devi 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, pages 6904–6913, 2017

[3] Drew A. Hudson, Christopher D. Manning;GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6700-6709

[4] Justin Johnson, Bharath Hariharan, Laurens van der Maaten,Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages2901–2910, 2017.

[5] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In Proceedings of the IEEE International Conference on Computer Vision, pages 1965–1973, 2017.

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach,Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages457–468, 2016.

[7] Peter Anderson, Xiaodong He, Chris Buehler, DamienTeney, Mark Johnson, Stephen Gould, and Lei 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, pages6077–6086, 2018.

[8] Hedi Ben Younes, Remi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, pages 2612–2620,2017.

[9] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In Advances in Neural Information Processing Systems, pages 1564–1574, 2018.

[10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems, 2019.

[11] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang,Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8317–8326, 2019.

[12] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez,Marcal Rusinol, Ernest Valveny, CV Jawahar, and Dimo-thenis Karatzas. Scene text visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, 2019

[13] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In Proceedings of the International Conference on Document Analysis and Recognition,2019

[14] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez,Marcal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering.arXiv preprint arXiv:1907.00490, 2016

[15] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen,Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision language research.In SysML Workshop, NeurIPS, volume 2018, 2018.

[16] Ronghang Hu, Amanpreet Singh, Trevor Darrell, Marcus Rohrbach ;Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for TextVQA.In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9992-10002

[17] Yash Kant, , Dhruv Batra, Peter Anderson, Alex Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. "Spatially Aware Multimodal Transformers for TextVQA." (2020).

[18] Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton van den Hengel, Qi Wu. (2020). Structured Multimodal Attentions for TextVQA.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[20] Telmo Pires ,Eva Schlinger and Dan Garrette (2019). How multilingual is Multilingual BERT?CoRR, abs/1906.01502.

[21] Karthikeyan K, Zihan Wang, Stephen Mayhew and Dan Roth (2020). Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In International Conference on Learning Representations.

[22] Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski and Filip Ginter. (2019). Is Multilingual BERT Fluent in Language Generation?.

[23] Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, Sampo Pyysalo. (2019). Multilingual is not enough: BERT for Finnish.

[24] Takashi Wada, Tomoharu Iwata. (2018). Unsupervised Cross-lingual Word Embedding by Multilingual Neural Language Models.