Neelu Verma (IITJ) , Rati Kumari (IITJ)

Abstract

Visual question answering has diverse research areas that aims to recognize the content of the input image and then answer the query related to the given image. Our cross-lingual question answering on the real-world chart dataset is also a part of the Visual Question Answering domain. Here, we focus on addressing the problem of asking a question related to different charts in the regional language and designing a system that can make practical inferences from the given charts like bar, pie, etc.

Data visualization helps in better understanding of the underlying data, but designing a model that can understand at a human level is challenging. Chart specific query require multi-step attention, deep understanding and reasoning capabilities. So, we are trying to design a new model that can handle multilingual question and answer in the same language as that of the question.

Cross-lingual question answering on any given chart image is a new problem. We are addressing only English and Hindi language, by dividing it into sub parts such that query and it's answers are in English or query and it's answers in Hindi or the questions are in English and answers are in Hindi or vice versa.

We have taken DVQA dataset and used google translation to convert English questions and answers into Hindi. Afterward, we are trying to implement a model that can handle the above problem. Also, we are planning to test the model on a real-world chart dataset where we have randomly collected the images from the internet, newspaper, magazine, etc.

Our project will help analyze the chart images in medical and scientific projects. As a result, it will help in getting a better insight into the charts.

Contents

List of Figures

List of Tables

Cross-lingual Question Answering on Chart datasets

1 Introduction and background

Data visualizations such as pie charts, bar charts, and plots consists of abundant information in a precise manner and are typically present in magazines, scientific and business reports, newspapers,etc. As, data visualization helps in better understanding of data, but to design a model to have a human level understanding is difficult problem. Understanding the visualization in the charts is itself a difficult task as the available labelled data is either limited or biased. Recent advancements in Deep Learning techniques can help to solve this problem.

Ongoing research in the field of chart visualization [1, 2, 3, 4, 5, 6, 7, 8] highlight the models that involves the question answer on the chat dataset. Most VQA problems are classification problem but new architectures are designed involving CNN and Faster RCNN to extract the features from bar and pie charts along with labels and legend details to incorporate the reasoning, structure and data understanding in the chart.

Various approach [9, 10, 11, 12, 13] has been used to have understanding of the charts. Our problem is to design a model that can understand the chart content and answer the related questions. We are implementing cross-lingual system where end user can ask question in its own language. We have subdivided our cross-lingual problem to English and Hindi language only.

Consider a scenario where the source data visualizations are in English, i.e. they contain legends, title and labels in English, and a non-English speaker asks a natural language question in his/her native language and expects the question-answering system to provide answer in language of his/her choice. The aims is to address chart-based question answering in this scenario. This is a challenging task that requires precise numerical and visual reasoning as well as has to deal with linguistic challenges. One obvious plausible approach could be to translate questions into the source language and then solve the chart QA problem. However, the success of such approach is limited for low-resource languages and is heavily depended on the quality of translation. In this work, we take a different path and address the problem without an explicit translation module.

The existing datasets for question answering on chart images such as DVQA [4], FigureQA [6], PlotQA [5] are designed for monolingual Question Answering (QA). Multi- lingual Question Answering [14] system are only limited to context based QA. To our knowledge, there does not exist any dataset that can be used to study Multi-Lingual chartQA. Therefore, in this work we introduce cross-language data visualization question answering (CL-DVQA in short) by manually translating questions of DVQA dataset [4] from English to Hindi. Our dataset contains bar charts and associated cross-lingual Question Answer pairs illustrated in figure 1.1

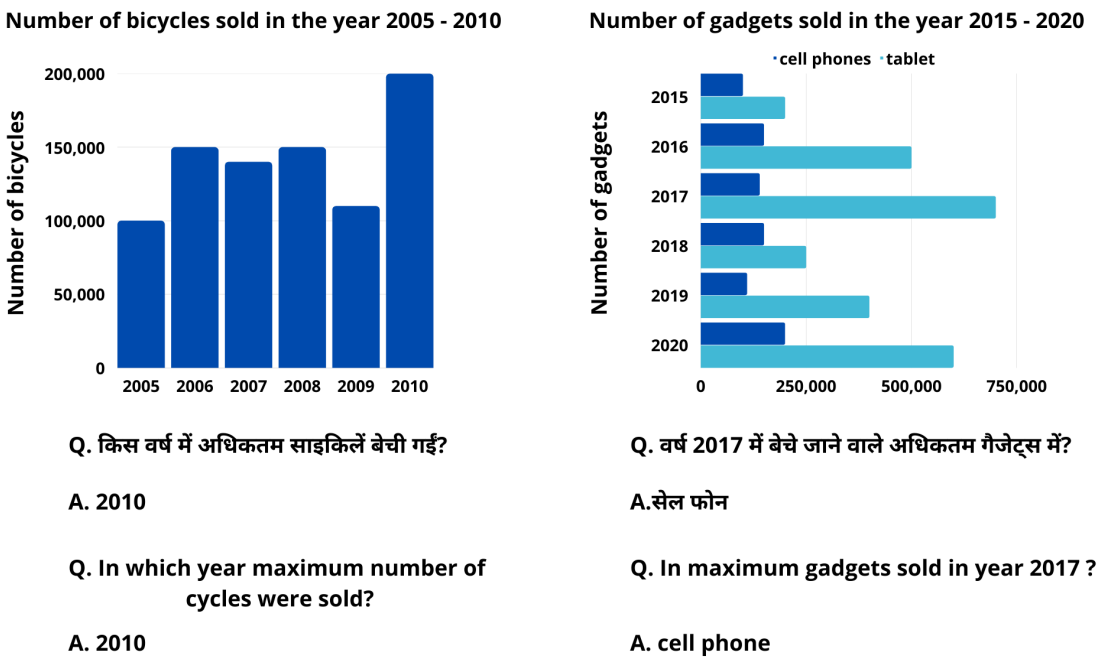


Figure 1.1: **Cross-lingual DVQA sample example**: Given a chart containing labels, title and/or legends in English and a natural language question on a non-English language say Hindi, our goal is to arrive at an answer in the same language in which question is asked.

2 Literature survey

Humans can quickly determinate the image content, understand the spatial positions, determine their attributes and relationships, and sense each object's context. We can ask random queries about the given figure and convey the extracted information. VQA involves both the domain of computer vision and natural language processing. It includes problems like object detection [15, 16], image classification [17, 18], counting, spatial relation, common sense reasoning, etc.

Image classification tasks have used CNNs [19] models. The advanced object recognition algorithms have surpassed humans in accuracy. Object recognition of the object in the given image classifies it without focusing on the spatial details. Object detection is the localization by putting a bounding box around object instances. Semantic segmentation is one stage ahead of localization by categorizing individual pixels as belonging to their semantic class. [20, 21]. Instance segmentation advance by separating between isolated occasions of the same semantic class [22, 23].

Image caption has the liberty to choose the granularity of image features. For example, 'What game is this?' needs the knowledge of the entire scene, but 'What is there behind the dining table?' requires attention to the location details. For unambiguous replies, the response is evaluated by matching the ground truth.

Visual Question Answering is an active research area in deep learning aiming to develop a system that can answer the question with human-level understanding. VQA [24] is a system designed intersecting both language and vision. VQA has used MS COCO dataset [25] and abstract scene dataset [26]. As,

data visualization helps in better understanding of data, but to design a model to have a human level understanding is difficult problem. Understanding the visualization in the charts is itself a difficult task as the available data is either limited or biased. Several models are designed to extract the content of images and questions embedding. These features are combined and by concatenation, point to point multiplication,etc. Features are used to predict the corresponding answer to the question.

Ritwick Chaudhry [7] devised LEAF-QA along with an attention network to effectively deal with the limitations of multimodal QA system. Here, the data was augmented with a set of test that was built from unavailable data sources to test in order to generate the question answering models. This LEAF-QA approach attained better performance across various question types, but failed to provide higher performance for more complicated relational questions. Monika Sharma [2] developed an effective QA system called ChartNet-based on a MAC-Network to provide answers from a dictionary of of generic answers. The primary specialty of this approach was that it predicted both vocabulary answers and out of vocabulary answers, but it failed to render a textual outline of the statistical charts. Revanth Reddy, [6] presented a deep learning model for handling the issues of reasoning task of the QA system on categorical plots. This model targeted to handle the numeric and visual reasoning tasks by exploiting modular components. The major limitation exists on this method is that it delivered low performance while processing low-median and high-median questions.

The cross-lingual concept can be implemented if we incorporate zero shot translation learning [27]. It refers to the fact that we can translate between any two languages even though we don't have that training data available. The parallel data between two languages to train the system on is available. When we used bilstm sequence to sequence models, the machine translation system was trained on English to Korean data and then English to Japanese data and it was demonstrated was that same system can also translate between Korean and Japanese even though it wasn't trained on Korean and Japanese language translation. So, this is what zero shot translation refers to there was zero data for Korean to Japanese but it could still translate between the two.

Research work is going on involving the multi lingual question answering on the given passage papers like [14] has been a great work that helps resource constraint languages. We are planning to incorporate the previous work and add cross-lingual concepts on chart dataset.

3 Problem definition and Objective

Charts are everywhere, it helps humans to better understand the data but to have a model that can understand in the same way is a challenging problem. We have various deep neural network that can understand the features of natural images predict the answer specific to the query. To incorporate reasoning and deeper understanding for charts in cross-lingual language is our problem. The problem is to analyze the chart and extract the feature embedding from chart images and combine it with question embedding then predict the result on the basis of combined embedding.

We are designing one model for all of the languages and a advantage of this is that there's no need for

a new model for each language pair that we want to translate. Disadvantage could potentially be that we need a larger model overall than a separate language pair model and this could lead to either higher inference time or more computational requirements for training or inferencing or both. It is observed that when the dataset is obtained with the language pair with of low resource and abundant resource, they learn the hidden relation and understanding is improved. So, we are designing single model to handle it.

Consider a scenario where the source data visualizations are in English, i.e. they contain legends, title and labels in English, and a non-English speaker asks a natural language question in his/her native language and expects the question-answering system to provide answer in language of his/her choice. The aims is to address chart-based question answering in this scenario. This is a challenging task that requires precise numerical and visual reasoning as well as has to deal with linguistic challenges. One obvious plausible approach could be to translate questions into the source language and then solve the chart QA problem. However, the success of such approach is limited for low-resource languages and is heavily depended on the quality of translation. In this work, we take a different path and address the problem without an explicit translation module.

4 Methodology

4.1 Existing cross-lingual dataset

Several dataset that helped in exploring the cross-lingual model are context based [14, 28, 29] where answers are generated for each questions. Freestyle multilingual image question answering [28] has 310,000 question-answer pairs in both English and Chinese language for 150,000 images. In MMQA [14] 5,495 question answer pairs [14] for 250 document containing 500 articles in both English and Hindi is designed to study Cross-lingual concept in six different domains. The MCVQA dataset [29] contains training questions - 248,349 and validation questions - 121,512 for given input images in both English, Hindi or mixed of both English and Hindi.

4.2 Dataset Used

As DVQA dataset contains synthetic bar chart images of various categories like structure understanding, data retrieval and reasoning, this has reduced biasness in the dataset. There are 3,487,194 total question answer pair. The dataset is divided in three sub parts one of training which has 200,000 images and two for test each containing 50,000 images, containing familiar and new images.

We have used google translation to convert DVQA dataset questions and answers into Hindi to modify the dataset according to our problem statement Fig 1.1.

Synthetically generated charts provide the precise control over location of the visual elements. We have manually translated the questions and answers from English to Hindi, thus making bilingual dataset for question answering containing 6,974,388 question answer pairs, covering structured, data and reasoning domains. We are considering only English and Hindi languages for the experiment, but other languages

can also be included in future. Dataset is publicly available on github.

4.3 VQA Algorithms

Numerous VQA algorithms have been proposed in recent years. It mainly involves these steps with some variations:

- *Image featurization*: features are extracted using CNNs trained on VGGNet [17], ResNet [30], GoogLeNet [31], best features are extracted using ResNet-152 [32].
- *Question featurization*: features are extracted using bag of word(BOW), LSTM [33] encoders, skip thought vectors [34], grated recurrent units GRU [35]
- *Combination of both features*:
  - features are concatenated either by element-wise multiplication or element-wise addition, and then forwarded to a linear classifier [28, 36, 24, 37].
  - features are combined using bi-linear pooling in neural network [38, 39, 40],
  - spatial attention maps for features based on relative importance along with question features passed though a classifier [10, 41, 42, 43, 44]
  - either by dividing VQA task into sub task [45, 46]

The combined features are used as a classifier to predict the answer. Mostly VQA algorithms generates only those answers that are seen during training.

The attention-based model helps understand the task relevant regions of the given image. Several Visual Question Answering models have utilized spatial attention to create region specific CNN features.

4.4 Models Used

Various models are designed to evaluate the modified DVQA dataset. Parameters like **Adam optimizer** [47] learning rate of **0.001** and dropout of **0.5** used in all classification based model with vocabulary of 2031 words obtained from training set containing words of both English and Hindi, so each classification model each have 2031 output units. The predicted answer is always one of unique value the model has seen during training.

Baseline methods evaluate a dataset's difficulty and establish the threshold performance level that decent algorithms should surpass.

4.4.1 YES

YES: It is the basic baseline where answer of each question is only "Yes". For any chart input image the model will predict "Yes" for all questions.

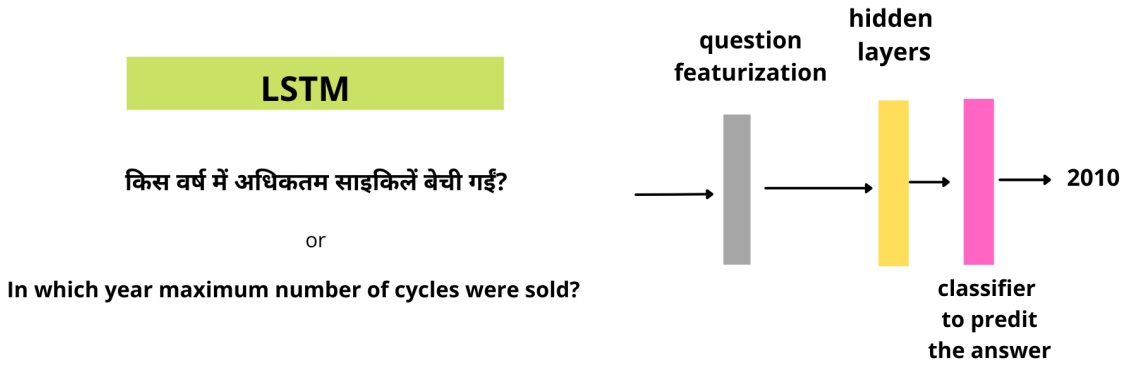


Figure 4.1: **QUES model : Chart based Cross-lingual question answering**: Given a chart containing labels, title and/or legends in English and a natural language question on a non-English language say Hindi, our goal is to arrive at an answer in the same language in which question is asked. Here, model is unaware of the input image. Model uses two layer LSTM to encode question as in VQA paper.

4.4.2 QUES

QUES: It is the basic linear softmax classifier model that classifies one of the unique answer from the global answer dictionary corresponding to each image. We have used tf-idf for creating embedding. Term frequency-inverse document frequency [48] weights are used to understand the frequency of the important word in the corpus. The importance is proportional to the frequency of word in the document. Term frequency measures the how frequently the word occurred in the document. Inverse document frequency measures the word's importance, in TF each term are consider as equally important but there are few words that occur more frequently but are of low importance, so it weigh down the frequent word and scales up the rare ones. Encoding of the question is done by a 1024 unit two layer LSTM.

It is image blind model i.e. for each question embedding the model will tell a unique answer from the vocabulary.

4.4.3 IMG

IMG: Image features are extracted using last layer of resNet-152 with input as 448x448, resulting in tensor of 14x14x2048 followed by a hidden layer of 1024 unit followed by the softmax layer to predict one of the unique answer. It is the question blind model where the answer for the image is one of the unique value from vocabulary.

4.4.4 QUES+IMG

QUES+IMG: It is the combination of QUES+IMG model where answer is one of the unique value from vocabulary. In this model question embedding is obtained from tfidf and resnet helps to extract image feature embedding. The embedding layer based on questions contains all the questions words seen in the training. Both these embedding are merged and pass though a softmax layer and one of the unique answers is predicted from the global vocabulary.

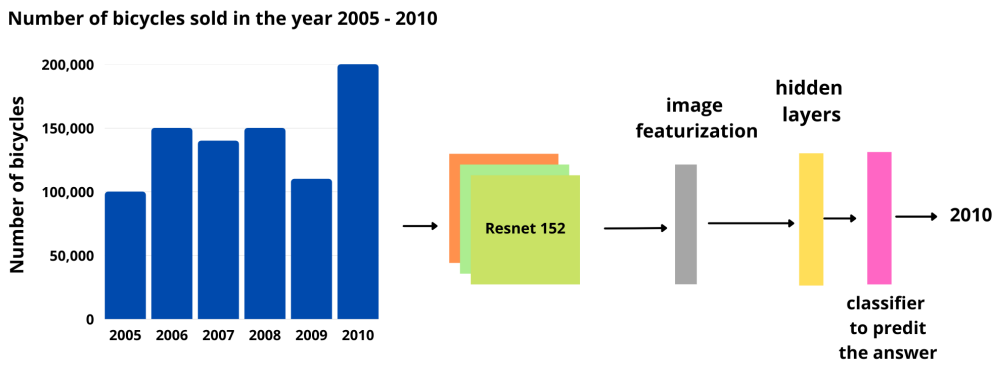


Figure 4.2: **IMG model: Chart based Cross-lingual question answering**: Given a chart containing labels, title and/or legends in English and a natural language question on a non-English language say Hindi, our goal is to arrive at an answer in the same language in which question is asked. Here, model is unaware of the question being asked. It uses ResNet-152 embedding to encode the images and passed though the hidden layer followed by softmax layer to predict the one value from global answer dictionary.

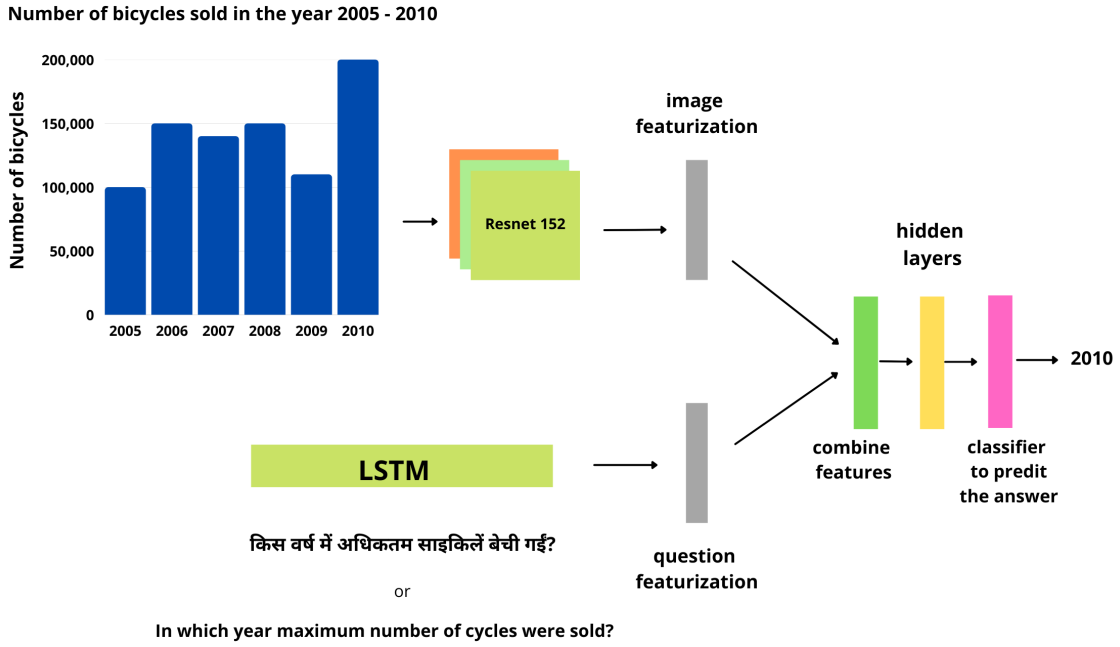


Figure 4.3: **IMG+QUES model: Chart based Cross-lingual question answering**: Given a chart containing labels, title and/or legends in English and a natural language question on a non-English language say Hindi, our goal is to arrive at an answer in the same language in which question is asked. Here, both image features obtained from last layer of resNet-152 and question embedding from two layer of LSTM are concatenated then passed through a hidden layer followed by a softmax layer to predict the output.

4.4.5 IMG+QUES+OCR

IMG+QUES+OCR: It is the combination of QUES+IMG model along with text features of chart image obtained from OCR. The image features extracted through resNet-152, question features through lstm

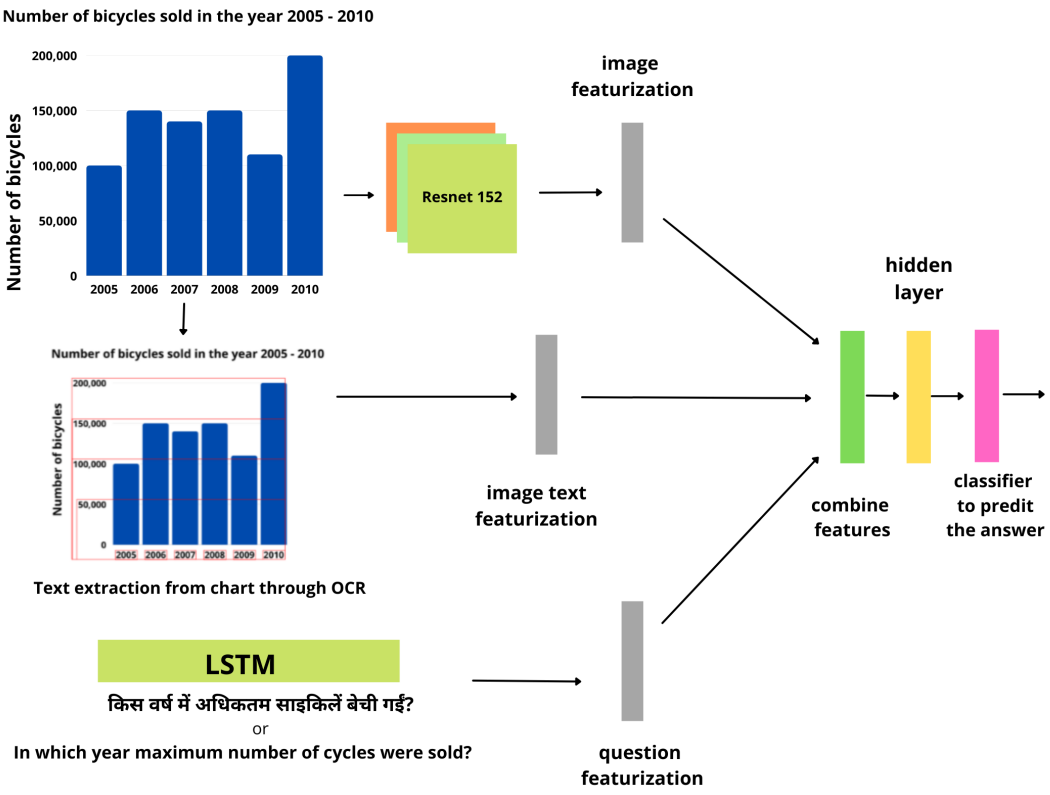


Figure 4.4: **IMG+QUES+OCR model: Chart based Cross-lingual question answering**: Given a chart containing labels, title and/or legends in English and a natural language question on a non-English language say Hindi, our goal is to arrive at an answer in the same language in which question is asked. Here, image features obtained from last layer of resNet-152, question embedding from two layer of LSTM and embedding of chart text are concatenated then passed through a hidden layer followed by a softmax layer to predict the output.

and text in chart is extracted with the help of OCR. tf-idf embedding is used in both question and extracted text of chart. The features are concatenated and passed through the hidden layer followed by a softmax layer and the answer is predicted from one of the unique answers dictionary.

All these model works as classification problem without any reasoning.

4.5 Image content extraction

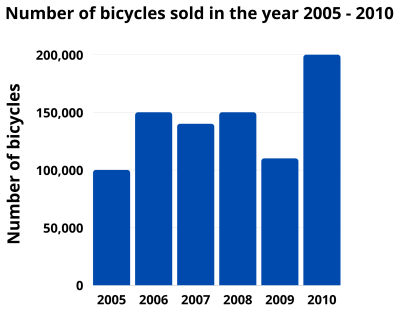
To extract the content of given input bar chart, we used Optical Character Recognition (OCR) followed by text detection and image disassembly [49]. For data extraction several steps need to be performed:

- *Figure extraction*:

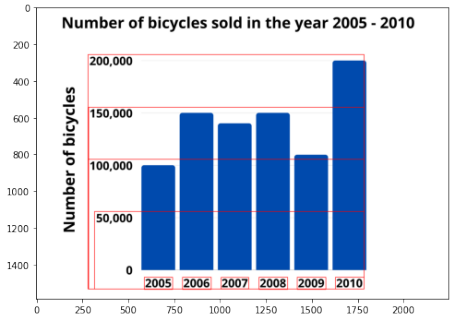
  In this the bar chart is extracted and segmentation by labeling and grouping pixels by their intensity and calculating the smallest bounding box that encapsulates each set of labels.

- *Text detection*:

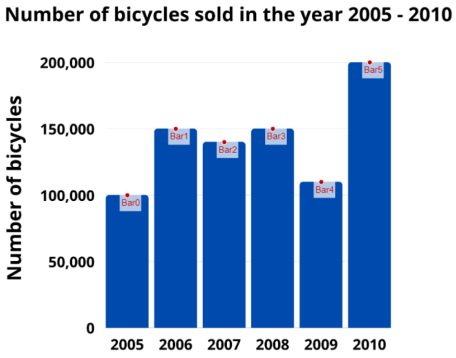
  Easy OCR is used to extract the content of image. It is an incredible tool with support for 80+ languages and popular writing scripts like Chinese, Arabic, Devanagari, Latin and etc. We can



(a) Given barchart image as input for understanding of content in the image



(b) Easy-OCR helps in extracting the content of image represented by placing the bounding box around the x-axis and y-axis labels



(c) Count of bar along with the corresponding y co-ordinate values

```
Bar0, 98795.52506876437
Bar1, 148230.27872228684
Bar2, 138342.38014854418
Bar3, 148229.33087924868
Bar4, 108682.47579946886
Bar5, 197654.6061023898
```

(d) Mapping of the bar value with the x-axis labels to have a relation table for question answering

Figure 4.5: **Image feature extraction: Optical Character Recognition helps in extracting the content**: Given a chart containing labels, our goal is to calculate the number of bars their location and the corresponding for each label from y-axis.

check Easy OCR functionality online `https://huggingface.co/spaces/tomofi/EasyOCR`.

- *Image disassembly*:

  We must first break the image down into its components to achieve deeper meanings. So, the most important thing to consider is chart image's axis because it will help to understand the x-axis and y-axis labels. To have deeper understanding of bars the Gaussian blur is applied to remove image mask from the original image. Canny edge detection and Hough transform helps in understanding lines and axis. The two longest lines that have 90-degree intersection corner are the main axis as x and y axis. Contour corners are extracted and vertical line is formed if the difference between x-coordinate is zero. If the two lines are sharing same y value and a vertex then it is grouped to form a bar. Bars with the same color or pixel values are clubbed together.

- *Data extraction*:

  The content located perpendicular to y-axis are termed as y-tick value and those parallel to the x-axis are termed as x-tick value. To have each bar value, average distance between y-tick text is calculated which approximately gives coordinate values. Here linear assumption along axis is considered.

Table 5.1: Comparative accuracy obtained on DVQA and Cross-lingual DVQA dataset

		DVQA	***CL-DVQA***
Dataset	**Baselines**	***Accuracy***	***Accuracy***
Test Familiar	Yes	11.7%	11.7%
	Img	14.8%	9.0%
	Ques	21.0%	18.7%
	Img+Ques	32.0%	29.3%
	Img+Ques+OCR	-	28.7%
Test Novel	Yes	11.7%	11.7%
	Img	14.9%	8.7%
	Ques	21.0%	18.0%
	Img+Ques	32.0%	28.8%
	IMG+Ques+OCR	-	27.4%

- *Data Analysis*:

  Image's data can be manually extracted by from the figures using WebPlotDigitizer. The same figures were then passed through automated data extraction pipeline for comparative analysis. Demo can be seen online at `https://apps.automeris.io/wpd/`

5 Experimental findings

We report the result from open ended question of multi lingual DVQA dataset. It predicts the answer with highest activation from the global answer dictionary as the baseline models are basic classification model incorporating the image and question embedding. For image feature the model is analysed and results are obtained on test familiar and test novel of modified DVQA dataset mentioned in table 5.1.

6 Summary and Future plan of work

We have use OCR to read content from the chart and extract the value of labels and content mentioned on the chart. To extract the data of the chart the architecture is designed such that one generates the answer specific to chart using ORC, question and image embedding to predict the answers of users query using existing vocabulary. The OCR model specifies the bounding box location of the label and classification model predicts one of the unique answers. So, final model should choose one of the value of the answer dictionary as classifier value.

The OCR model effectively extract horizontal content the major challenge comes in vertical contents. Also,we need to extract the precise value per label and have a temporary storage relation table for each input image. Relation table for each type of bars like stacked, vertical, horizontal including both plain as well as designed need special architecture. If we are incorporating other charts like line, pie, etc. it also need relation table and for that we need separate architecture.

Multilingual Bert [50] has also shown better language understanding on several languages. concatenating multilingual Bert question embedding with image will give better accuracy. OCR's extracted

content is stored in temporary relation table, which will predict the answer. Cross-lingual Question Answering can be practically more useful if we incorporate handwritten question answering with respect to any given image.

References

[1] Q. Wu, D. Teney, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Underst.*, vol. 163, pp. 21–40, 2017.

[2] M. Sharma, S. Gupta, A. Chowdhury, and L. Vig, "Chartnet: Visual reasoning over statistical charts using mac-networks." IEEE, 2019, pp. 1–7.

[3] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," *CoRR*, vol. abs/1612.06890, 2016.

[4] K. Kafle, B. L. Price, S. Cohen, and C. Kanan, "DVQA: understanding data visualizations via question answering," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5648–5656.

[5] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "Plotqa: Reasoning over scientific plots." IEEE, 2020, pp. 1516–1525.

[6] S. E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, and Y. Bengio, "Figureqa: An annotated figure dataset for visual reasoning," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018.

[7] R. Chaudhry, S. Shekhar, U. Gupta, P. Maneriker, P. Bansal, and A. Joshi, "LEAF-QA: locate, encode & attend for figure question answering." IEEE, 2020, pp. 3501–3510.

[8] X. Liu, D. Klabjan, and P. N. Bless, "Data extraction from charts via single deep neural network," *CoRR*, vol. abs/1906.11906, 2019.

[9] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," *CoRR*, vol. abs/1704.03162, 2017.

[10] ——, "Show, ask, attend, and answer: A strong baseline for visual question answering," *CoRR*, vol. abs/1704.03162, 2017.

[11] H. Wang, X. Zhang, S. Ma, X. Sun, H. Wang, and M. Wang, "A neural question answering model based on semi-structured tables," in *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics, 2018, pp. 1941–1951.

[12] K. Kafle, R. Shrestha, B. L. Price, S. Cohen, and C. Kanan, "Answering questions about data visualizations using efficient bimodal fusion," *CoRR*, vol. abs/1908.01801, 2019.

[13] J. Poco and J. Heer, "Reverse-engineering visualizations: Recovering visual encodings from chart images," *Comput. Graph. Forum*, vol. 36, no. 3, pp. 353–363, 2017.

[14] D. Gupta, S. Kumari, A. Ekbal, and P. Bhattacharyya, "MMQA: A multi-domain multi-lingual question-answering framework for english and hindi." European Language Resources Association (ELRA), 2018.

[15] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 779–788.

[16] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.

[19] M. A. J. Morán, F. J. L. Aligué, M. M. Macías, and M. I. A. Sotoca, "A CNN model for grey scale image processing," in *From Natural to Artificial Neural Computation, International Workshop on Artificial Neural Networks, IWANN '95, Malaga-Torremolinos, Spain, June 7-9, 1995, Proceedings*, vol. 930. Springer, 1995, pp. 882–889.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3431–3440.

[21] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 1520–1528.

[22] N. Silberman, D. A. Sontag, and R. Fergus, "Instance segmentation of indoor scenes using a coverage loss," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 8689. Springer, 2014, pp. 616–631.

[23] Z. Zhang, A. G. Schwing, S. Fidler, and R. Urtasun, "Monocular object instance segmentation and depth ordering with cnns," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 2614–2622.

[24] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "VQA: visual question answering - www.visualqa.org," *Int. J. Comput. Vis.*, vol. 123, 2017.

[25] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," vol. abs/1504.00325, 2015.

[26] R. Vedantam, X. Lin, T. Batra, C. L. Zitnick, and D. Parikh, "Learning common sense through visual abstraction," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 2542–2550.

[27] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," vol. abs/1611.04558, 2016.

[28] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question answering, (nips)," *CoRR*, vol. abs/1505.05612, 2015.

[29] H. R. Khan, D. Gupta, and A. Ekbal, "Towards developing a multilingual and code-mixed visual question answering system by knowledge distillation," in *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*. Association for Computational Linguistics, 2021, pp. 1753–1767.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 1–9.

[32] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Comput. Vis. Image Underst.*, vol. 163, pp. 3–20, 2017.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 1997.

[34] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 3294–3302.

[35] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*

2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 2014, pp. 1724–1734.

[36] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *CoRR*, vol. abs/1512.02167, 2015.

[37] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4976–4984.

[38] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. 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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 457–468.

[39] K. Saito, A. Shin, Y. Ushiku, and T. Harada, "Dualnet: Domain-invariant network for visual question answering," in *2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017*. IEEE Computer Society, 2017, pp. 829–834.

[40] J. Kim, S. Lee, D. Kwak, M. Heo, J. Kim, J. Ha, and B. Zhang, "Multimodal residual learning for visual QA," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 361–369.

[41] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4613–4621.

[42] I. Ilievski, S. Yan, and J. Feng, "A focused dynamic attention model for visual question answering," *CoRR*, vol. abs/1604.01485, 2016.

[43] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 21–29.

[44] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, ser. Lecture Notes in Computer Science, vol. 9911. Springer, 2016, pp. 451–466.

[45] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," *CoRR*, vol. abs/1511.02799, 2015.

[46] ——, "Learning to compose neural networks for question answering," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, San Diego California, USA, June 12-17, 2016. The Association for Computational Linguistics, 2016, pp. 1545–1554.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[48] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 60, no. 5, pp. 493–502, 2004.

[49] R. A. Al-Zaidy and C. L. Giles, "Automatic extraction of data from bar charts," in *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015, Palisades, NY, USA, October 7-10, 2015*, K. Barker and J. M. Gómez-Pérez, Eds. ACM, 2015, pp. 30:1–30:4.

[50] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding." Association for Computational Linguistics, 2019, pp. 4171–4186.