Investigation on Versatile Multi-Modal Question Answering Systems

Sanam Molaee
molaee@chalmers.se

Mirarash Keshavarz Kelachayeh
mirkes@chalmers.se

Daniel Söderqvist
danisode@chalmers.se

Abhijeet Singh Dhillon
dhillon@chalmers.se

***Abstract*—Considering the diverse data sources that the companies and research centers work with these days, the need for question answering systems has become more important in nearly every industry. Traditional question answering systems which are only text-based, have limitations in handling some information, such as images. In this project, the aim is to address this challenge by investigating multi-modal systems that can handle text, table and image data. The key parts for these systems are: Data fusion, natural language processing, computer vision and answer generation. Multiple models were investigated by literature research and the final chosen models to implement were an LSTM with VGG-19 and a TAPAS model. The first one is a visual question answering system and the latter is a table question answering system. Training and validation loss were shown as the result for the LSTM model, while a table containing answers for each question was shown as the result for TAPAS. Overall, each model seemed suitable for their specific task and combining those two while giving them a larger dataset can be the next possible step for building a multi-modal question answering system in the future researches.**

Introduction

Artificial Intelligence (AI) has explosively found its way into everything from businesses to individuals. The popularity of AI is due to various reasons including the fact that it helps to calculate and process data and handle tasks that are time-consuming or complex for humans. However, this has made a growing demand on AI to be able to efficiently interact with humans to obtain useful information. Among the various types of AI systems, Multi-modal system is a system that helps AI to imitate humans and thus, interact on nearly the same level.

Multi-modal learning has extensively gained attention within the deep learning area in the recent years. This paper deals with two different type of multi-modal models called as visual question answering (VQA) and TAPAS (Table Parser) systems. In these systems, the combination of different modalities, such as text with image, text with speech, video with text and so on, depending on the application, will be investigated.

Nowadays, industries are creating data in the form of documents, images and even videos during product development or for post market data. By using multi-modal models, these data can be used to develop a question answering system. This question answering system can thereafter be used as a virtual assistant that can help the company management for decision making tasks. The Multi-Modality question answering system works based on understanding the sentiments of the inputs, fusing them and relating them to the context of the question. This can be in various forms and needs a wide range of information such as object detection, object recognition and classification that can be obtained from for example, text and image through cross-modal interaction.

Background

In this section, both concepts and functions that provide a basis for understanding this project will be presented. In this project, at first, the fundamentals of text and image question answering system is presented and at the second step, a table and question model is investigated as well.

Modalities

Modalities refer to different types or modes of data. These modalities can be image, text, tabular data, speech or video. The choice of combinations of modalities to build a multimodal system depends on the overall architecture and the information that is provided [15].

- Heterogeneous and Homogeneous modalities: Heterogeneous modalities refer to modalities that show diverse qualities, structures and representations, while homogeneous modalities refer to the modalities that have similar properties. Figure 1 displays few examples of heterogeneous and homogeneous modalities.

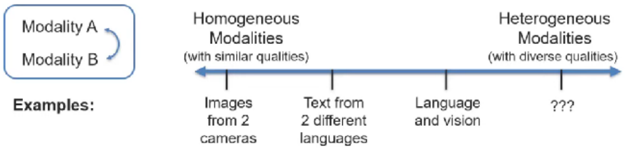


Fig. 1. Heterogenous and Homogenous Modality

- Modality element representations: Elements help to describe the different modalities. For example, objects are the elements in images, words are the elements in texts and speech is the element for audio modality. These elements have properties such as density which can be, for example, objects per image or words per minute. The elements can also have structures such as temporal, spatial, hierarchical and so on. These properties play an important role in cross modality interactions.

- Modality connections: Modalities are often related and share commonalities. These commonalities can be statistical or semantic. Statistical consists of association between elements, for example correlation and co-occurrence, as well as dependency, which means one element of modality A depends on another element of modality B. Semantic, on the other hand, means correspondence and relationship between two elements of different modalities. In Figure 2 the difference between statistical and semantic modality is shown.

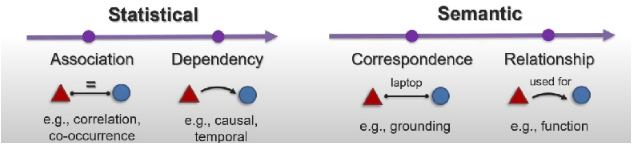


Fig. 2. Modality Connections

- Interconnected modalities: During inference, modalities often cross interact with each other. When these modalities interact, they give out a response and this response can lead to either redundancy or non redundancy. Redundancy refers to the modalities maintaining their properties while non redundancy refers to the interaction between two modalities which leads to creation of a new kind of modality. Figure 3 gives an overall visual description of the concept.

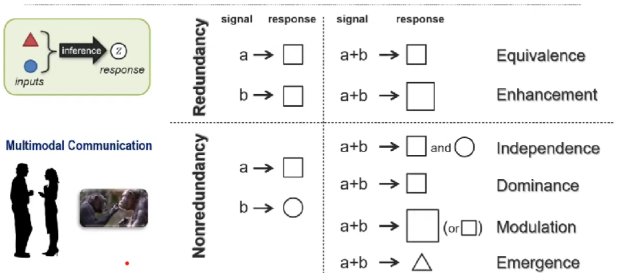


Fig. 3. Interconnected Modalities

Building Blocks of Multi-Modal Systems

Multi-Modal systems involve programs that learn and improve through the use and experience of data from multiple modalities and also demonstrate understanding, reasoning and planning. [18].

- Representation: It describes the cross-modal interactions between various modalities using fusion, coordination and fission method as shown in Figure 4.
- Alignment: This refers to identifying and modelling cross-modal connections, as shown in Figure 5, between the elements of multiple modalities.
- Reasoning: This refers to combining knowledge via multiple inferential steps for decision making process.

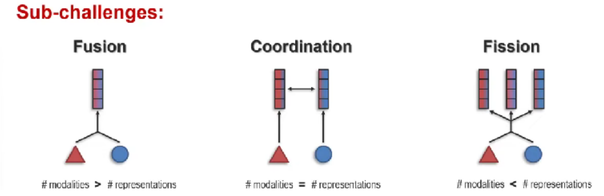

Sub-challenges:
Fusion
Coordination
Fission
# modalities > # representations
# modalities = # representations
# modalities < # representations

Fig. 4. Different representations of multi-modal systems

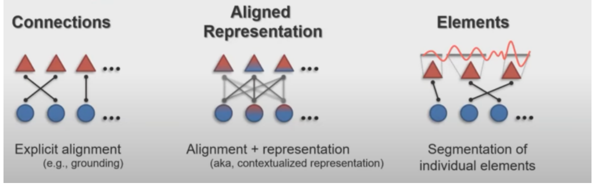

Connections
Aligned Representation
Elements
Explicit alignment
(e.g., grounding)
Alignment + representation
(aka, contextualized representation)
Segmentation of individual elements

Fig. 5. Alignment in multi-modal systems

- Generation: As shown in Figure 6, it includes the process of producing raw modalities that can be in form of text summarizing, translation and creation e.g. ChatGPT 4 is a creation process.

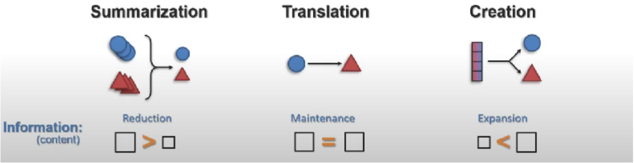


Fig. 6. Generation in multi-modal systems

- Transference: This refers to transfer of knowledge between modalities that is shown in Figure 7. It consists of transfer, co-representation and co-learning.

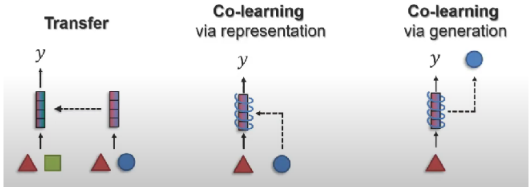


Fig. 7. Transference in multi-modal systems

- Quantification: This is the post processing part to better understand heterogeneity, cross modal interactions and the multi-modal learning processes.

The multi-modal systems use representation and alignment to make a single architecture and then the reasoning make decision depending on application which is shown in Figure 8.

Visual Image Modality

Visual modalities refer to different types of visual data including images, videos e.g. in computer vision tasks, algorithms trained on images to recognize objects, patterns, or even

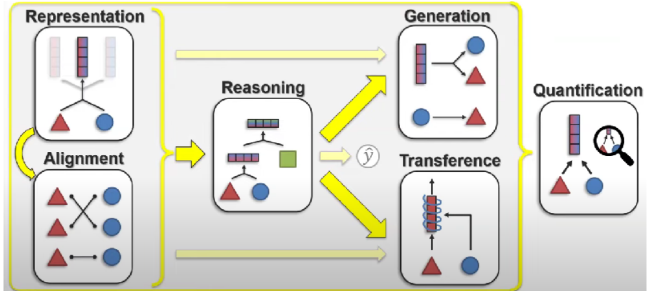


Fig. 8. Working of Multi-Modal Systems

to perform tasks like image segmentation or object detection. Convolution neural networks (CNN) are One of the most prevalent way of image data processing [17]. A typical digital image contains a series of grid-like pixels that will be processed with the several filters that are shown in Figure 9. The general structure of CNN contains several convolutional layers, pooling layers and fully connected layers.

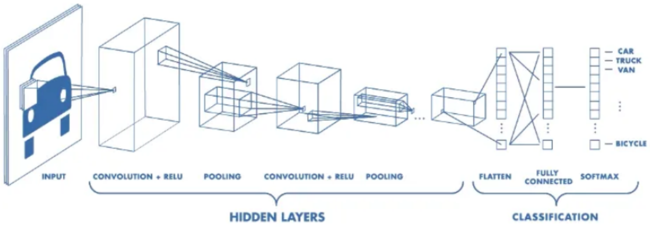


Fig. 9. Working of CNNs

CNNs helps building more abstract and hierarchical visual representations and have some advantages associated with it such as:

- Translation Invariance: It refers to the ability of CNN to recognize patterns and features in an image regardless of their locations.
- Learned Kernels/Filters: Pre-trained CNN's are models that have been trained on large datasets, such as ImageNet and can be reused for various tasks, such as image classification, object detection or face recognition. Some of the famous pre-trained models are VGG, AlexNet and GoogleNet. The VGG architecture that is used in this project, is shown in Figure 10.

Text Modality

[12] Text Modality refers to natural language processing (NLP). NLP combines computational linguistics—rule-based modeling of human language with statistical, machine learning and deep learning models.

Text representation is a critical component of NLP because human language encompasses a vast range of vocabulary, sentence structures and idiomatic expressions. Hence, to enable computers to handle the mentioned complexities, text must be transformed into a structured and numerical format. Thus, a few techniques used for representing words in the form of numbers or tensors are as follows:

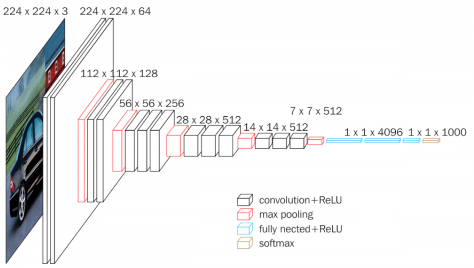
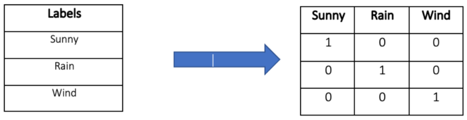


Fig. 10. Example of Pre-trained CNN Model-VGG

- One hot Encoding: [3] This method ,shown in Figure 11, converts the words to numerical values in vector format.



Labels
Sunny
Rain
Wind

Sunny	Rain	Wind
1	0	0
0	1	0
0	0	1

Fig. 11. Matrix showing the conversion from words into One Hot encoding vectors.

- Recurrent Neural Network (RNN): One of the biggest drawbacks for one hot encoding is that it cannot keep track of relation between words. To overcome this issue, recurrent neural networks can be used. [11] RNNs are a class of neural networks that are helpful in modeling sequential data inspired by human brains function to establish relationship between words by generation of weights representing the importance of previous words. RNNs can suffer from the problem of vanishing or exploding gradients, which can make it difficult to train the network effectively. 12 shows a schematic view of RNN to process text data.

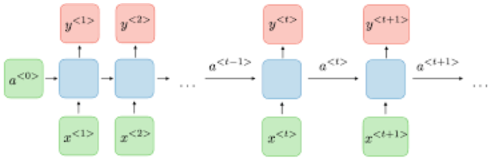


Fig. 12. Recurrent Neural Network Architecture

- Long Short Term Memory (LSTM): [21] Long Short-Term Networks is a sequential neural network that is able to handle vanishing gradient problem which is shown in Figure 13.
- Transformers: [8] The Transformer architecture that is shown in Figure 14 contains a stack of encoder and decoder that takes a text sequence as input and produce another text sequence as output. In fact, the transformers use attention methods which makes them a useful tool to process sequential data.

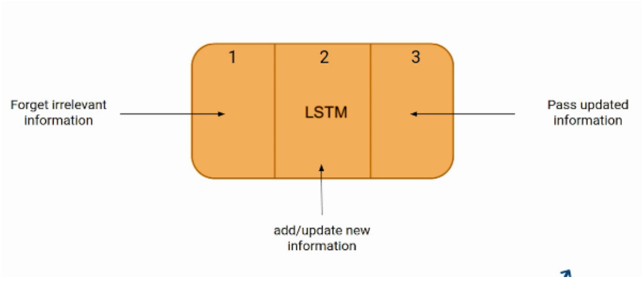


Fig. 13. LSTM Architecture

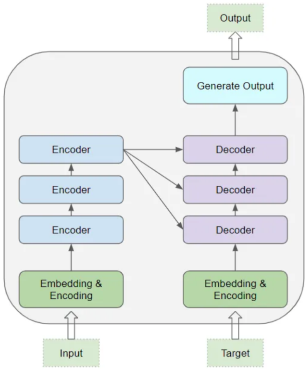


Fig. 14. Transformer Architecture

Fusing models

[19] In order to combine various modalities, multi-modal models uses different techniques such as fusion-based approach. The fusion-based approach works based on encoding the different modalities into a common representation and then combining (or fusing) them to capture the semantic information. Depending on when the fusion happens, there are two version of fusion methods which are early fusion and late fusion. The former, combines the data e.g text, image before training. The later combines the predictions from models that are trained separately. Then, the predictions are combined to build a final prediction.

The early fusion model is easy to implement but slow in training processes. On the other hand, late fusion technique is easy to train, but it cannot detect the interaction between the modalities at the early stage of processing, therefore it might need pre-trained transformers which are computationally expensive. Regarding that each fusion method has its own advantage and disadvantage, there is also another type of fusion, so called as intermediate fusion (feature level fusion) that concatenates each modality after some pre-proccesing stage in intermediate layers and before prediction stage.

Visual Question Answering System

Visual Question Answering (VQA) System is a computer vision and natural language processing task where the system takes an image and a related question and then it generates a textual answer.

[20] The working of visual question answering is as follows:

- Image featurization - converting images into their feature representations for further processing.
- Question featurization - converting natural language questions into their embeddings for further processing.
- Joint feature representation - ways of combining image features and the question features to enhance algorithmic understanding.
- Answer generation - utilizing the joint features to understand the input image and the question asked, to finally generate the correct answer.

PURPOSE

Through this project a literature survey will be done which will provide a deeper knowledge and give some resourceful understanding of different multi-modal question answering systems. Through this knowledge, two such systems will be investigated and tried considering the different fusion methods i.e. a late fusion model and a table data question answering model. Thereafter, the systems will be evaluated considering what features might be best suited to develop a virtual assistant for industrial data. Some of the key objectives include:

- Make a literature research considering different multi-modal question answering systems.
- Improving accuracy and contextual understanding of systems by implementing different fusion models.
- Compare and evaluate these models reflecting upon how well suited for a virtual assistant using industrial data they might be.

LITERATURE RESEARCH

Several recent researches have used various models and architectures to study visual question answering are as follows:

- Agrawal [2] et al proposed a VQA for free-form and open ended questions on MS COCO dataset. They used several methods such as Bag-of-Words, LSTM (with one hidden layer) and deeper LSTM (with two hidden layers) for question embedding and VGGNet for image embedding. Finally, they compared the accuracy of different architectures such as BOW+VGGNet, LSTM+VGGNet and deeper LSTM+VGGNet. The comparison results show that the last model outperforms other models.
- There are also other methods such as bilinear method that is used by H. Ben-yoanes [5] proposed BLOCK that works based on block-term decomposition in which image and question are merged by a fusion technique called bilinear model. The comparison of BLOCK results with models from other papers show that BLOCK performance is better than other in all kinds of questions.
- Some works such as MUREL network proposed by R. Cadene [7] is different from classical attention methods [4], [9], [14], [22], which detects the correlation between image and text using an iterative process through several

MUREL cells to detect spatial and semantic. In fact, each MUREL cell uses bilinear fusion module to merge question and regional image vectors using Tucker decomposition method that will be given to pairwise relational model to generate context embedding for image regions. The comparison of MUREL network with other models on different databases show that the overall accuracy of MUREL network is higher than state of the art models such as MUTAN [4], Pythia [13], Couneter [23] and so on whereas some models perform better in yes/no or number questions.

- Hierarchical question-image co-attention by Jiasen Lu et al, [16]presents a novel method for Visual Question Answering (VQA) using a hierarchical question-image co-attention model. This model processes both the textual question and visual image simultaneously at multiple levels: word, phrase and question. This hierarchical approach allows the system to focus on relevant parts of both the question and the image, facilitating a more detailed and accurate understanding. The method significantly enhances the performance of VQA tasks by enabling more subtle interaction between text and image data. The paper includes experimental results and ablation studies to demonstrate the efficacy of their model. The model, particularly when combined with ResNet features (Oursa+ResNet), had a very good performance.

Method

Codewise two models have been investigated: LSTM combined with VGG-19 which is a late fusion method and also TAPAS which is a table answering system.

A. LSTM combined with VGG-19

The first model that is used in this research consists of LSTM to encode questions and CNN for embedding the images. This model was trained on a part of VQA-V1(2015) abstract scenes dataset which has 20,000 training images, 10,000 validation images and 20,000 testing images. Abstract scenes means images created using art that don't have an immediate association with the physical world. Each image in this dataset has 3 questions associated with it and each question has 10 possible answers. Therefore, the dataset in total has 150,000 questions and 1,500,000 answers.

The architecture of the models is shown in Figure 15. The whole late-fusion model consists of three models that individually corresponds to different part of the multi-modal system. The model for handling the images are a VGGNet with normalization and pooling layers. The model for handling the text, a LSTM model consists of two hidden layers that are used to obtain 2048-dim embedding for the question. As shown in Figure, the question embedding is the concatenation of last state (512-dim) and last hidden state (512-dim) representations which gives 2 (hidden layers) by 2 (cell state and hidden states) by 512 (dimension of each of the cell states, as well as hidden states). At the end, the output will be given to a fully-connected layer with tanh activation function to transform the input with 2048 dimension to 1024 dimension. The data fusion process of image and LSTM embedding are done through element-wise multiplication. Finally, the fusion outputs are given to a fully connected neural network containing 2 hidden layers and 1000 hidden units with dropout 0.5 in each layer with tanh as activation function. At the end, a softmax layer evaluates answer distribution. The cross entropy loss is used as loss function in the model.

One of disadvantages of VQA models is that they have huge datasets which need very high computational power to train and test on. Therefore to tackle this problem, the huge database is trimmed such that a certain amount of data from the whole dataset is used for training and validation so that there is no need for high computational power. For convenience, 10000 images are selected for training and 5000 images are chosen for validation. The total questions in this shortened training dataset and validation dataset are 30000 and 15000 respectively. On the other hand, the answers for the same is 300,000 and 150,000 respectively. The main advantage of this methodology is its flexibility toward changing the size of the training and validation in accordance to the available computational power. This trimmed dataset was trained and validated with the existing architecture as discussed above.

.

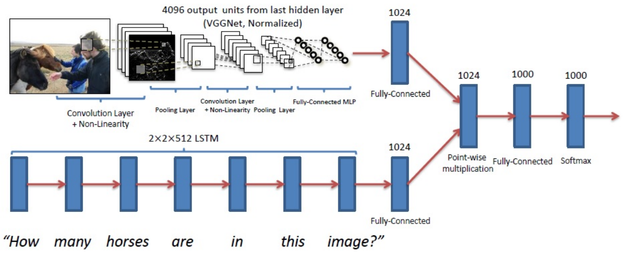

4096 output units from last hidden layer
(VGGNet, Normalized)
1024
Convolution Layer + Non-Linearity
Pooling Layer
Convolution Layer + Non-Linearity
Pooling Layer
Fully-Connected MLP
Fully-Connected
1024
1000
1000
2×2×512 LSTM
1024
Point-wise multiplication
Fully-Connected
Softmax
Fully-Connected
"How many horses are in this image?"

Fig. 15. LSTM combined with VGG-19 Architecture

B. TAPAS (TAble ParSing)

TAPAS [10] is a weakly supervised question answering model developed by google as an alternative to traditional semantic parsing for question answering over tables without generating logical forms. The model was pre-trained in Hugging-Face. Traditional approaches involve translating a question into a logical form, which can be costly and complex. The TAPAS model predicts answers directly by selecting table cells and applying aggregation operations, thereby bypassing the need for generating logical forms.

TAPAS extends the architecture of BERT by making additional embedding including position, segment,column/row,rank and previous answer IDs to encode tables. As shown in Figure16, it flattens the table into a sequence of words and then convert the words to tokens which will be concatenated to question tokens. It can be seen in Figure16 that TAPAS model includes two classification layers for cell selection which can be the final answer or the input used to find the final answer. The aggregation operator describes certain operations such as SUM, COUNT, AVERAGE or NONE that that needed to be applied on cells which will be chosen by cell selection layer. The probability of mentioned operations will be done by a linear layer with softmax that are shown on top of the CLS token in Figure16,meanwhile the subset of the cells corresponding to the probable aggregation operations are given by cells selection layers which are shown on top right of Figure16. The model is pre-trained on 6.2M horizontal tables with maximum 500 cells from wikipedia. This pre-training is crucial for its effectiveness in parsing and understanding table structures and contents. It is then fine-tuned on various semantic parsing datasets.

The TAPAS model has three different types of loss functions that are corresponded to cell selection, scalar answer and ambiguous answer. The loss function for cell selection is a summation of column selection loss, the cell of the selected column loss and aggregation loss that can be written as:

$$J_{\text{CS}} = \frac{1}{|\text{Columns}|} \sum CE(p_{\text{col}}) + \frac{1}{|\text{Cells}|} \sum CE(p_{\text{cell}}) - \alpha \log p_a(op_0) \quad (1)$$

In which α and op_0 are scaling hyperparamteter and the operation aggregation parameters e.g SUM,COUNT and NONE, respectively. In addition, p_{col} and p_{cell} represent the column and cells that are related to final answer. In fact, the component of column selection loss focuses on selecting the appropriate column from the table and uses binary cross-entropy to measure the accuracy of this selection. Once a column is selected, column cell selection loss evaluates how well the model selects specific cells within that column. It is notable that the last expression of J_{CS} represents the aggregation loss which is relevant to cases where the question requires an aggregation operation. When no aggregation is needed (as in the case of cell selection), this loss is calculated with respect to the 'NONE'.

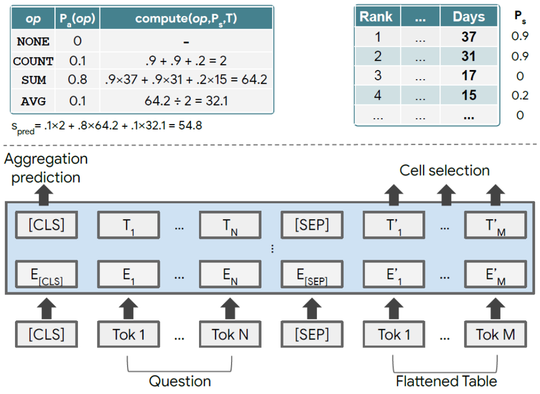
Furthermore, for the scalar answers the loss function is defined as:

$$J_{\text{SA}} = -\log \left(\sum_{i=1}^{n} p_a(op_i) \right) + \beta J_{\text{scalar}} \quad (2)$$

In which β and J_{scalar} are scaling hyperparamter and Huber loss [10], respectively. Finally, for ambiguous question, if the following condition holds, the cell selection loss is taken otherwise the scalar answers loss will be implemented.

$$p_a(op) \geq S \quad (3)$$

In which S is a threshold parameter between 0 and 1.



op	$P_a(op)$	compute(op,P_s,T)
NONE	0	-
COUNT	0.1	.9 + .9 + .2 = 2
SUM	0.8	.9×37 + .9×31 + .2×15 = 64.2
AVG	0.1	64.2 ÷ 2 = 32.1

s_{pred}= .1×2 + .8×64.2 + .1×32.1 = 54.8

Rank	...	Days	P_s
1	...	37	0.9
2	...	31	0.9
3	...	17	0
4	...	15	0.2
...	...	...	0

Fig. 16. TAPAS Architecture

Results

LSTM combined with VGG-19

The result for building the preprocess method including scaling down the dataset to a more easy to handle size consists mostly of code [1]. In this repositories the code for the used architecture and methods for training and building the data loaders can also be seen.

The loss function for both the training data and validation data from training the described architecture is shown in Figure 17.

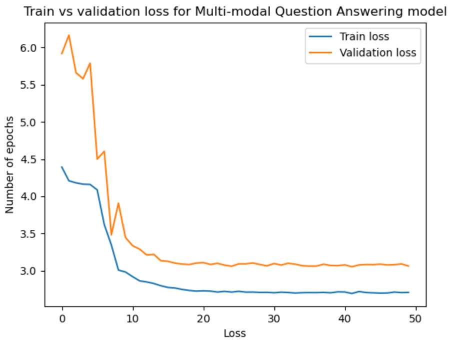
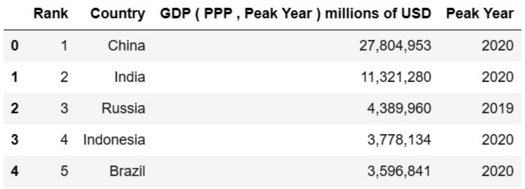


Fig. 17. Train and validation loss for training the multi-modal question answering model.

TAPAS

In order to get a sample example for question and answer for TAPAS model, a question asked to present the 5 countries with the highest GDP in 2020. The answer can be seen in Figure 18.



	Rank	Country	GDP (PPP , Peak Year) millions of USD	Peak Year
0	1	China	27,804,953	2020
1	2	India	11,321,280	2020
2	3	Russia	4,389,960	2019
3	4	Indonesia	3,778,134	2020
4	5	Brazil	3,596,841	2020

Fig. 18. TAPAS result

DISCUSSION

LSTM combined with VGG-19

Figure 17 shows from the trimmed dataset that the losses for the training and validation are relatively high. This could possibly be due to small amount of dataset of 10000 images as well as the small number of questions and answers compared to the larger dataset. The accuracy of the model depends on the amount of vocabulary that is fed into the LSTM model but in this case the number of words were only 1283 words from the questions and 2883 words from the answers which is small vocabulary set in comparison to the dataset having 150,000 questions and 1,500,000 answers. Although the losses were relatively high, both the training and validation curves shows that the model is learning and gets better converging towards a loss around 3.2 and 2.7 respectively.

The code platform built for training on the relatively small and easy to handle dataset were used on one architecture explained in the previous method section. Therefore it's hard to say whether the loss results are good or bad with no other model to compare to. The model on the other hand is easy to replace with another architecture to try if there exist a better one for the type of multi-modal question answering problem the dataset consist of. The different models to try could be other types of late-fusion models or different models including previous described architectures such as early fusion models.

TAPAS

Overall, TAPAS is proven to show improvement over traditional models in terms of accuracy and simplicity. TAPAS is suited for single tables small enough to fit in memory, but struggles with very large tables or multiple-table databases. Additionally, while TAPAS can understand basic compositional structures, it's limited to simple aggregations from table cells.

Combined discussion

By combining the table question answering model like TAPAS with the researched model (LSTM and VGG19 models)including useful information like number of people in the images for example the model might be able to output a better answer to the question asked. To get a better understanding of different modalities models like TAPAS were investigated and tried. Also investigated models like BLOCK [6] and MUREL [7] seemed to be good potential models for improving the accuracy and overall performance.

CONCLUSION

The purpose of the project and this report was to investigate and improve a deeper understanding of different multi-modal question answering systems. The first subtask was to make a literature research and study the state-of-the-art multi-modal question answering systems. This research laid a broad platform to further work on and investigate the topic.

The next subtask was to implement different multi-modal question answering models to find an optimal solution and possibly improve accuracy and contextual understanding to further improve the chosen solution. Due to the limited time frame of this project only the construction of the method to scale down and preprocess the dataset along with one such model was possible to implement. Including in this was understanding the model, its structure and then training it.

The last subtask was to evaluate the models and in the manner of evaluation, the loss for train and validation data were logged during training 17. This does not say much but it is a good starting platform for further implementation of new models to compare and investigate for which might work the best.

To better understand different types of modalities when it comes to input data for the multi-modal question answering system a table question answering model (TAPAS) was investigated. This model was effective in pre-training on large-scale data consisting of text-table pairs and was able to restore masked words and table cells. This laid a better understanding for further improvement of a multi-modal question answering system by combining more input modalities that might improve the accuracy and experience of the system.

FUTURE RESEARCH

Since this project is a starting foundation for further investigation in the problem there are many possibilities for future research. This includes implementing and comparing other different models with the researched model in this project. These models could include other types of late-fusion architectures or even compare other fusion method such as early-fusion models. Furthermore, other model that work based on bilinear fusion such as BLOCK [6] and MUREL [7] were investigated in the literature research which can be potential models for future investigation and comparison between models.

Another possibility would be to include more modalities to see if more information to the model can help improve the accuracy and result towards a more user-friendly experience. Such as table-data or a caption to the image. Let's say there is a car manufacturer that capture images of every other finished product. The images are there to see if there exist any damages or other useful information on and of the product. A reasonable question along with these images could then be something like: How many products are damaged? If then the images are provided with a table describing and listing all products with damages, could this information help the multi-modal question answering system to output a better response with higher accuracy than without the additional information?

References

[1] Group 15. *DaniSode/Multi-Modal-Question-Answering-System*. 1 2024.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.

[3] baeldung. One-hot encoding explained, 06 2023.

[4] Hédi Ben-Younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. MUTAN: multimodal tucker fusion for visual question answering. *CoRR*, abs/1705.06676, 2017.

[5] Hédi Ben-Younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. BLOCK: bilinear superdiagonal fusion for visual question answering and visual relationship detection. *CoRR*, abs/1902.00038, 2019.

[6] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. 33(01):8102–8109, 2019.

[7] Rémi Cadène, Hédi Ben-Younes, Matthieu Cord, and Nicolas Thome. MUREL: multimodal relational reasoning for visual question answering. *CoRR*, abs/1902.09487, 2019.

[8] Ketan Doshi. Transformers explained visually (part 1): Overview of functionality. 06 2021.

[9] 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. *CoRR*, abs/1606.01847, 2016.

[10] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Muller, Francesco Piccinno1, and Julian Martin Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. 2020.

[11] IBM. What are recurrent neural networks?, Accessed: 12 2023.

[12] IBM. What is natural language processing?, Accessed: 12 2023.

[13] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the VQA challenge 2018. *CoRR*, abs/1807.09956, 2018.

[14] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *CoRR*, abs/1610.04325, 2016.

[15] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions, 2023.

[16] Jiasen Lu, Jianwei Yang, Dhruv Barta, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Neural Information Processing Systems*, 2017.

[17] Mishra Mayank. Convolutional neural networks, explained, 08 2020.

[18] Louis-Philippe Morency. Lecture 1.1 - introduction (cmu multimodal machine learning course, fall 2022), 2023. YouTube.

[19] Nate Rosidi. Multimodal models explained - Unlocking the Power of Multimodal Learning: Techniques, Challenges, and Applications., 03 2023.

[20] Anuj Sable. Visual question answering: a survey, 2020.

[21] Shipra Saxena. What is lstm? introduction to long short-term memory, 10 2023.

[22] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *CoRR*, abs/1708.03619, 2017.

[23] Yan Zhang, Jonathon S. Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *CoRR*, abs/1802.05766, 2018.