Design and Develop a Versatile Multi-Modal Question Answering System

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 both private persons and companies 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, like images. In this project, the aim is to address this challenge by developing a multi-modal system that can handle text, table and image data. The key parts for this system are: Data fusion, natural language processing, computer vision, and answer generation.**

Introduction

Artificial intelligence (AI) has explosively found its way into everything from businesses to individuals. The popularity of AI is for many reasons including the fact that it helps to calculate, process data and handle tasks that are time-consuming or complex for humans. However, this has put more and more demands in AI to be able to interact with humans in a good way so that everyone can understand and get as much of the processed information as possible. Multi-modal system is a system that helps AI to imitate humans, and thus, interact on the same level.

Multi-modal learning has extensively gained attention within the deep learning area in the recent years. This paper deals with a group of multi-modal models called as visual question answering (VQA) system. 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, this data can be used to develop a question answering system. This question answering system can thereafter be used as a virtual assistant that helps in the decision making for the industry in the management of the company. The Multi-Modality question answering system works based on understanding the sentiment of the input, fusing them and relating it 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. The scope of the project is narrowed down to text and image question answering systems and therefore only modalities including text and images will be considered.

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 multi-modal 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.

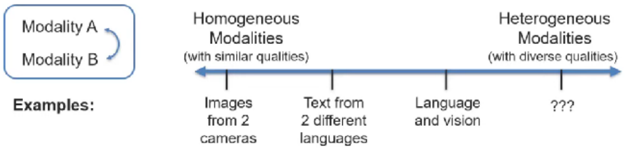


Fig. 1. Heterogenous and Homogenous Modality

- 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, 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.

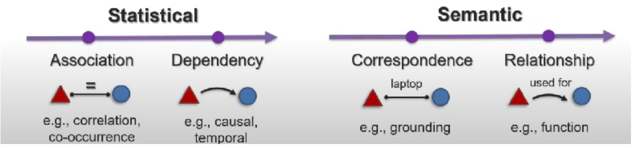


Fig. 2. Modality Connections

- 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.

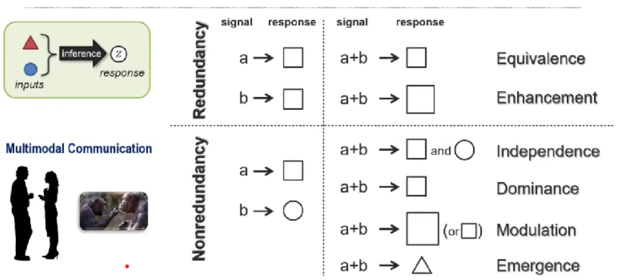


Fig. 3. Interconnected Modalities

Building Blocks of Multi-Modal Systems

Multi-Modal systems involve computer algorithms that learn and improve through the use and experience of data from multiple modalities and also demonstrate understanding, reasoning and planning. The following presented are the building blocks of multi-modal systems [18]:

- Representation: The representations describes the cross-modal interactions between individual elements, across different modalities and it further consists of fusion, coordination and fission. Representation in this context means modalities architectures are combined together to make a single architecture that represents the system. This is visually shown in figure 4.

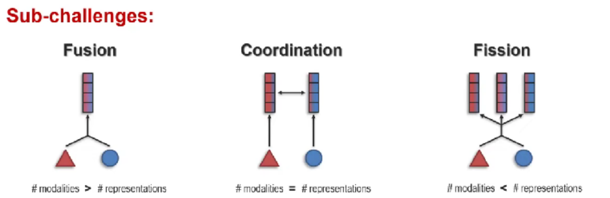
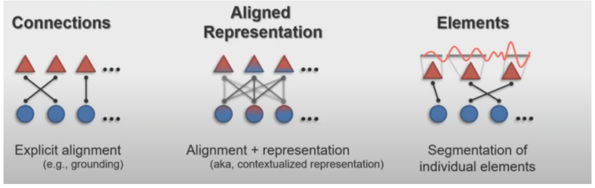

Sub-challenges:
Fusion
Coordination
Fission
# modalities > # representations
# modalities = # representations
# modalities < # representations

Fig. 4. Different representations of multi-modal systems

- Alignment: This refers to identifying and modelling cross-modal connections between all elements of multiple modalities which is built from the the data structure. These further consists of explicit alignment, alignment with representation and segmentation of individual elements. This can be seen in figure 5


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

- Reasoning: This refers to combining knowledge via multiple inferential steps, looking at multi-modal alignment and problem structure. Reasoning refers to the decision making capabilities of the system.

- Generation: This is a process to produce raw modalities that reflect cross-modalities interactions, structure and coherence. Generation can be of three types including, text summarizing, translation and creation. ChatGPT 4 is an example of creation as it can take in text and pictures to create new inference about the system. The different types is shown in figure 6.

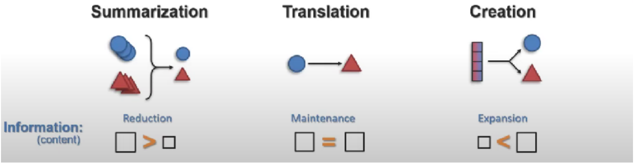


Fig. 6. Generation in multi-modal systems

- Transference: This aspect refers to transfer of knowledge between modalities. Usually to help the target modality which may be noisy or with limited resources. It consists of transfer, co-representation and co-learning. This is explained visually in figure 7.

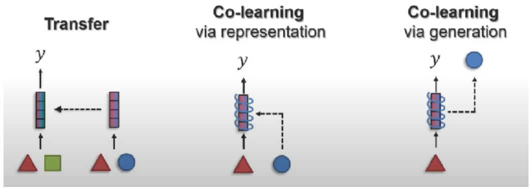


Fig. 7. Transference in multi-modal systems

- Quantification: This is the post processing part where an empirical and theoretical study are done 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, to which input is fed and reasoning is used to make a decision based on the knowledge the system has. Depending on the application, the reasoning will help in generation or transference which means it will create new text or images. The structure of both components and approaches are shown in figure 8.

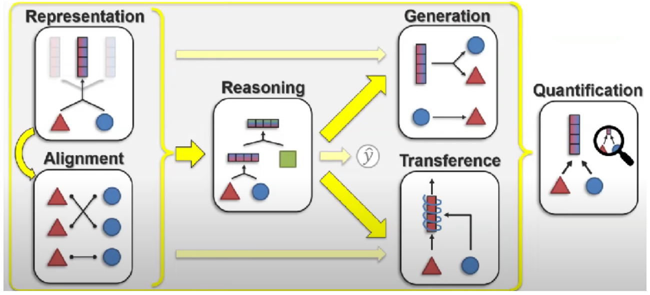


Fig. 8. Working of Multi-Modal Systems

Visual Image Modality

Visual modalities refer to different types or formats of visual data that algorithms can analyze and learn from. This could include images, videos, or any other visual input. For example, in computer vision tasks, algorithms might be trained on images to recognize objects, patterns, or even to perform tasks like image segmentation or object detection.

One of the most used ways of representing in images is convolution neural network (CNN). [17] A CNN is a class of neural networks that specializes in processing data that has a grid-like topology, such as an image. A digital image is a binary representation of visual data. It contains a series of pixels arranged in a grid-like fashion that contains pixel values to denote how bright and what color each pixel should be. These pixel representations are learnt with the help of filters. A CNN typically has three layers: a convolutional layer, a pooling layer, and a fully connected layer. An example of an CNN arcitecture is shown in figure 9.

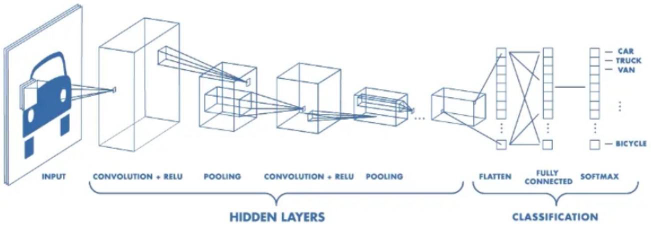


Fig. 9. Working of CNNs

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

- Translation Invariance: Translation invariance in CNN's refers to the network's ability to recognize patterns or features in an image regardless of their specific location. In other words, if a certain feature is presented in one part of the image, the network should be able to identify and recognize the same feature even if it appears in a different location.
- 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 model architecture is shown in figure 10.

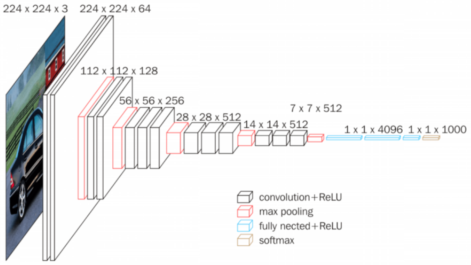


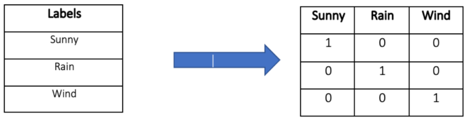
Fig. 10. Example of Pre-trained CNN Model-VGG

Text Modality

[11] 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 for several reasons. First and foremost, human language is inherently complex and diverse. It encompasses a vast range of vocabulary, sentence structures, idiomatic expressions and linguistic nuances. To enable computers to understand and work with this complexity, text must be transformed into a structured and numerical format that algorithms can process. A few techniques used for representing words in the form of numbers or tensors can be:

- One hot Encoding: [2] One hot encoding is a technique that we use to represent categorical variables as numerical values in a machine learning model. It can improve model performance by providing more information to the model about the categorical variable. It has some disadvantages such as it can lead to increased dimensionality, as a separate column is created for each category in the variable. This in turn can make the model more complex and slow to train. In figure 11 a visual representation of the transformation from words into One Hot encoding vectors are shown.
- 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. [10] RNNs are a class of neural networks that are helpful in modeling sequence data. Derived from feed-forward networks, RNNs



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.

exhibit similar behavior to how human brains function, i.e it can establish relationship between words by keeping an weight associated with the previous words. RNNs can suffer from the problem of vanishing or exploding gradients, which can make it difficult to train the network effectively. This occurs when the gradients of the loss function with respect to the parameters become very small or very large as they propagate through time. An example of how an RNN visually can be explained is shown in figure 12.

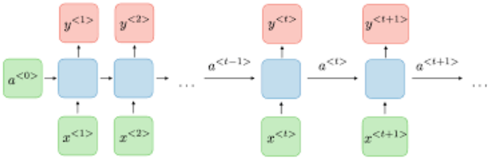


Fig. 12. Recurrent Neural Network Architecture

- Long Short Term Memory (LSTM): [21] Long Short-Term Memory Networks is a deep learning, sequential neural network that allows information to persist. It is a special type of Recurrent Neural Network which is capable of handling the vanishing gradient problem faced by RNNs which is shown in figure 13.

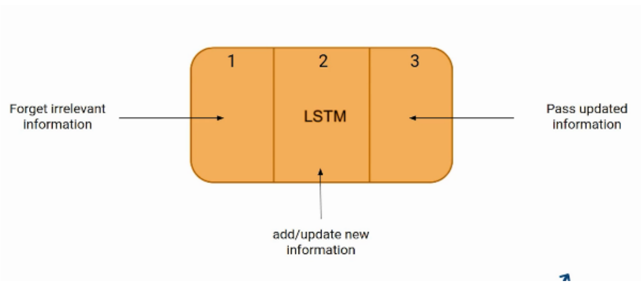


Fig. 13. LSTM Architecture

- Transformers: [7] The Transformer architecture excels at handling text data which is inherently sequential. They take a text sequence as input and produce another text sequence as output. For example when translating a sentence input from English to Spanish. At its core, it contains a stack of encoder layers and decoder layers. The Encoder stack and the decoder stack each have their corresponding embedding layers for their respective inputs. Finally, there is an output layer to generate the final output. Transformers make use of attention which will be discussed later on. In figure 14 a representation of a transformers architecture is shown.

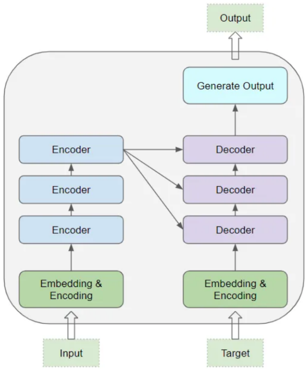


Fig. 14. Transformer Architecture

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 in an image and a related question in natural language 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 research will be made which will provide a deeper knowledge and give some resourceful understanding of different multi-model question answering systems. Through this knowledge and understanding, two such systems will be constructed and tried considering the different fusion methods i.e. early and late fusion. Thereafter, the systems will be evaluated considering which 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.

- Agrawal [1] et al propsed 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 [4] 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 [6] is different from classical attention methods [3], [9], [13], [23], 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 [3], Pythia [12], Couneter [24] 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

First model. (TBA) The other VQA model that is going to be used in this project is BLOCK that is fusion based and works based on bilinear superdiagonal tensor composition [5]. The Bottom-up image feature provided by [22] which includes a set of detected objects and their representations [8] are used for object detection and localization. The questions are embedded through a pre-trained Skip-thought encoder [14]. Thereafter, the BLOCK fusion is applied to merge question and image representations [5]. BLOCK uses a bilinear superdiagonal framework to optimize the balance between computational complexity and expressive power in processing multi-modal inputs. This makes it particularly effective for interpreting and answering questions about visual content, offering enhanced performance and parameter efficiency compared to previous models in the field.

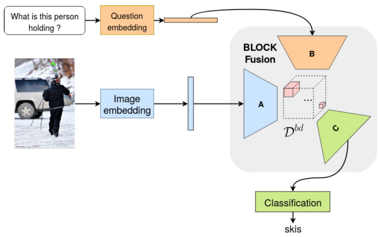


Fig. 15. Architecture for VQA embedding the BLOCK bilinear fusion

Results

Showing results of performance of the 2-3 models.

Discussion

Evaluate the results and draw parallels with the literature research and background.

Conclusion

What can we conclude? Future research?

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

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

[8] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 642–651, 2017.

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

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

[12] 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.

[13] 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.

[14] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *CoRR*, abs/1506.06726, 2015.

[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] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *CoRR*, abs/1708.02711, 2017.

[23] 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.

[24] 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.