EXPLORING ROBUSTNESS OF MULTIMODAL DEEP LEARNING ARCHITECTURE

Lohith Prasanna Teja Kakumanu
Department of Industrial and Systems Engineering
University at Buffalo, State University of New York
Buffalo, NY, USA
lohithpr@buffalo.edu

ABSTRACT

The ever-changing landscape of deep learning involves investigating novel techniques for harnessing the wealth of information available in the real world. This research dives into multimodal deep learning, a technology that overcomes the limits of single-modality models by combining a variety of data sources such as text, audio, video, and images. Considering these models' resilience to disturbances, however, remains to be a critical challenge. In this study, we investigate the robustness of multimodal deep learning models through the lens of Visual Question Answering (VQA) tasks. I have leveraged Vision and Language Tranformer (ViLT) for this task, specifically, I subject the model to perturbations primarily focused on textual and cross modal inconsistencies. This experiment reveal insights into the vulnerabilities of these models. Furthermore, I have analyzed the models confidence levels before and after perturbations on model reliability and statistically demonstrate a significant decrease in confidence, indicating the impact of perturbations on model resilience and reliability.

1 INTRODUCTION

Machine Learning and Neural Networks were the cornerstone technology behind generating or understanding complex patterns. Deep learning emerges as a subset of AI, propelling the field toward remarkable powers. It uses neural networks with numerous layers to extract information and patterns from complex datasets[1]. In the field of health care, this capability has enormous potential to alter diagnosis, treatment, and management of medical disorders. As we dive deeper into Deep Learning, for example, in healthcare, the data could be medical images, text or tabular patient information, genomic data, or sensor readings. These various data kinds and formats were classified as Modalities (Ngiam et al., 2011).

These models can take use of complementary information sources by combining data from many modalities, which enhances understanding and reasoning abilities (Wu et al., 2017). For instance, Combining genetic data with medical imaging data, can reveal insights into the underlying molecular pathways of diseases that inform treatment choices (Stahlschmidt et al., 2022). Analogously, (Mallick et al., 2022) mentioned about utilizing the Video Data and sensor signal data of the patients to detect the risky situations for weak adults. Furthermore, (Venugopalan et al., 2021) mentioned about early detection of Alzheimer's disease stage by fusing the data from multiple modalities such as Magnetic Resonance Images, Single Nucleotide Polymorphism Genetic Data, and clinical text data to classify the patients into AD, MCI, and controls. Providing this extra information along with MR images made the model see more patterns like hippocampus, RAVLT, amygdala brain areas which significantly outperformed the conventional Deep Learning setup.

When we consider only the image and text modality into account the boundaries of the deep learning models are significantly broadened enabling the model to more effectively comprehend

[1] https://shelf.io/blog/neural-networks-and-how-they-work-with-generative-ai/

the semantic context of the text and accurately extract the relevant information of the associated image (Gao et al., 2015). The combination of these text and image modalities finds application across various domains and tasks for example Object Detection (Redmon et al., 2016) sentiment analysis (Majumder et al., 2018), Medical Diagnosis (Rajpurkar et al., 2017). Are these models reliable? Investigating the reliabilty of these models is crucial across various domains including the health care. The study by (De Fauw et al., 2018) exemplifies this by demonstrating the use of Deep learning in identification of eye illness using retinus fundus images, while the research addresses the challenges of employing deep learning model in clinical contexts highlighting the necessity of model reliability and interpretability.

Hence, considering these all scenarios into account, my experiment aims at investigating the robustness of a model by bridging the gap between both text and image modalities and perturbing the model inputs to understand the model's reliability. To achieve this, I have chosen the Visual Question Answering (VQA) task where a question about an image is given and the goal is to generate an answer from both text and image information. The initial data set were coined by Antol et al. (2015).

2 BACKGROUND

In the case of Modalities, we consider Image, Video, Audio and Text to be the most common modalities used for the network (Ngiam et al., 2011). Convolutional neural networks (CNNs) are highly effective for processing visual data due to their inherent properties like translation invariance and locality (Krizhevsky et al., 2012). RNN is used for processing the text data, which is temporal, to capture the semantic relationships and contextual information within the text (Graves & Schmidhuber, 2005). In our research, attention mechanisms plays a major role, they employ the network to focus on specific parts of the input, emphasizing relevant information while suppressing noise or irrelevant information. For text input, the attention mechanism assigns higher weights to relevant tokens, and lower weights to the irrelevant ones. Similarly for image input, it enables the model to focus on specific regions or features within the visual data that are most relevant to the task. It dynamically adjusts the attention weights assigned to different spatial locations or channels within the image. (Vaswani et al., 2017). In the case of multiple inputs, they enable the model to focus on relevant information from both modalities while performing joint learning, thereby improving the model's interpretability, performance, and robustness (Hori et al., 2017).

Furthermore, they facilitate the fusion of information from both text and image modalities by allowing our model to attend to relevant features from both modalities simultaneously. It computes cross modal attention weights, the model can dynamically adjust the importance of text and image features based on their mutual relevance to the task (Ngiam et al., 2011).

Most widely used fusion techniques in the literature are mentioned below:

Early Fusion: Features gathered from individual modalities Pawłowski et al. (2023), (Zhou et al., 2019) are concatenated before being fed into a single deep learning model. This technique is efficient (Stahlschmidt et al., 2022), but it may not capture intricate relationships between modalities.
Late Fusion: Features are analyzed separately (Gandhi et al., 2023) by deep learning models for each modality, and the results are then fused using approaches such as attention processes or simple concatenation (Pandey et al., 2019) (Mehrish et al., 2023). This enables improved modeling of modality-specific data (Banerjee et al., 2021).
Attention-based Fusion: Attention mechanisms dynamically weigh the value of elements from many modalities, focusing on the information that is most important to the task. This has resulted in considerable increases in performance (Vaswani et al., 2017).

Most recently, Transformers are used in the vision tasks in the case of multimodal inputs. We deal with the concepts or principles in this journey of Multimodality starting with auto-encoding, descriptive learning, contrastive learning, representation Learning, translation, alignment, fusion, co-learning (Ngiam et al., 2011). Representation is the fundamental challenge in learning how to represent and summarize multimodal data, the heterogeneity of the data makes it challenging to construct such representations. For example, language is always symbolic, while audio and visual

modalities will be represented as signals. The representations could be single, joint and coordinated representations. The second challenge could be the mapping or translation of data from one modality to another. The relationship between modalities is often open-ended, where there exists a number of correct ways to describe or represent an image and one perfect translation may not exist (Baltrušaitis et al., 2018). There are divisions, like example-based, like image captioning, and generative-based, like video description.

3 Related Work

Multimodal deep learning has developed as a powerful strategy for dealing with complex tasks that require comprehending input from many sources. Multimodal deep learning models produce outstanding outcomes in a variety of tasks, including image captioning, which is the process of accurately generating written descriptions of photographs (Yu et al., 2019). Video understanding entails analyzing and extracting information from videos, such as object detection, action recognition, and event summary. Answering inquiries with information from text, graphics, and other sources. Affective computing is the recognition and comprehension of human emotions through facial expressions, speech, and text (Zheng et al., 2018).

Transitioning into investigations or experiments conducted by scholars in recent years, some of the following studies with the same design were presented below:

A Video Audio Text Transformer (VATT) (Akbari et al., 2021) takes raw signals as inputs and extracts multimodal representations that are rich enough to benefit a variety of downstream tasks. The methodology introduces a convolution-free VATT architecture for multimodal learning, featuring separate or shared backbone Transformers for each modality. Tokenization and positional encoding handle raw signals, while Drop Token reduces computational complexity. Common space projection and contrastive learning align modalities, utilizing Noise Contrastive Estimation and Multiple Instance Learning NCE loss objectives. They fine-tuned for video action recognition, image classification, Zero-shot text-to-video retrieval, and their VATT architecture outperforms the convnet. Especially their vision transformer achieves the top 1 accuracy of 82.1% on kinetics-400, 83.6% on Kinetics-600, 72.7% on Kinetics-700, and 41.1% on Moments in Time.

Similarly, Team from Google Deep Mind had a question: what can be learnt by looking at and listening to many unlabeled videos? To approach the solution (Arandjelovic & Zisserman, 2017) they introduced audio-visual corresponding learning task as part of which they trained visual and audio networks, without any additional supervision other than the raw, unconstrained videos themselves, the network has three focus areas, starting with vision subnetwork to deal with the color image, followed by audio subnetwork which will be converted to a log spectrogram and then treated as a greyscale image which are then passed to a fusion network where the two 512-D visual and audio features are concatenated into 1024-D to produce a two way classification output.

It has become more common to create multimodal representations by combining auditory waveforms and words from linguistic inputs. (Kiela & Clark, 2015) examined grounding semantic representations in the auditory data using evaluation like conceptual similarity and relatedness and then implemented cross model zero shot learning using technique like partial least squares regression and then projecting from one space to another and vice versa. The study concluded that multimodal representations perform better than single auditory or linguistic representations on a musical instrument clustering task.

Meng et al. (2021) presented a deep learning model on Electronic Health Record data with a transformer architecture to predict the diagnosis of depression through Bidirectional representation learning. They summed up the input embeddings from the 5 dimensions, like codes, medication lists, demographics, and topics, in a unified and temporal manner. To capture the contextual relationship in the sequence they pretrained the model using masked language modelling and fine-tuned on a specific dataset. Their model significantly reduced the false positives by 50%, with an average precision and recall of 0.94 and 0.84 respectively. The results illustrate the capacity of the two-stage transfer learning strategy for EHR modeling to overcome restrictions in the amount of accessible data, and bidirectional learning can provide higher performance versus unidirectional.

In the past few years, Researchers have been focused on applying it in all the applicable areas, Venugopalan et al. (2021) used this Multimodal approach to outperform the existing models in early

detection of Alzheimer's disease stage. The modalities are Clinical data consisting demographics, cognitive assessments, medications and the second modality includes Cross sectional MR Imaging data, and the third modality includes whole genome sequencing data. They used stacked denoising auto encoders to extract the features from the clinical and genetic data and used 3-D CNN for imaging data for the ADNI data set. The extracted features from each of the subnetworks are simply concatenated and passed to the fully connected layer for classifying the input. They demonstrated that this model outperformed the existing conventional Deep Learning models with a single modality in terms of accuracy, precision, recall and F1 scores. Furthermore, the model has identified hippocampus, and other brain areas which are considered as the top distinguished features.

Overall, multimodal deep learning is a fast-expanding discipline, with ongoing breakthroughs in architectures, fusion techniques, and training methodologies. As research advances, we should expect even more impressive outcomes and greater applications across multiple domains.

Vision Transformers (ViT) Introduced in 2020, ViTs revealed (Dosovitskiy et al., 2020) the promise of pure transformer for image classification problems, outperforming CNNs while providing more flexibility.

Hierarchical Transformers architectures (Liu et al., 2021) combine the capabilities of CNNs and transformers. CNNs thrive at local feature extraction, whereas transformers handle long-range dependencies, demonstrated improved performance in picture categorization problems. Aside from the transformer architectures, scholars pass multiple modalities on their specific networks individually and extract the features and fuse them with different strategies in the later stages as discussed above.

In terms of image modalities, natural images have different frequencies and when these are passed through the convolution layers, the output feature maps also possess this mixture of information at different frequencies. Chen et al. (2019) came up with an octave convolution operation to store and process the feature maps by factorizing the feature maps by frequency and analyze slower feature maps at lower spatial resolution to reduce the memory and computation power. The results showed that substituting convolutions with octave convolutions improved accuracy for image and video recognition.

In recent research focusing on the Visual Question Answering (VQA) task, there is a notable shift towards utilizing only image and text modalities. Specifically, these studies explore how neural networks or models can swiftly identify objects and their attributes from images, comprehend complex questions, and utilize this information to generate accurate answers. Initially introduced by Malinowski & Fritz (2014), the concept of question answering has evolved significantly. Gao et al. (2023) introduced a spatial-temporal transformer tailored for long-form VQA tasks. Their innovative architecture dissects traditional dense spatial-temporal self-attention into cascaded segment and region selection modules. These modules adaptively choose frames and image regions relevant to the given question, effectively reducing computational costs while enhancing performance. By incorporating Multimodal Multi-grained features, such as image patches and segments, their model better captures the relationships between visual concepts of varying granularities. Moreover, their approach facilitates temporal and causal reasoning by iteratively conducting selection and self-attention processes across multiple events.

Number of studies within the domain have put forth datasets specifically designed for Visual Question Answering (VQA) Krishna et al. (2017), Ren et al. (2015), Tapaswi et al. (2016), Shin et al. (2016), VQA v2 Goyal et al. (2017), VQA-CP Agrawal et al. (2018) and a few other datasets which holds comprehensive contents like flicker30K Plummer et al. (2015), Visual Genome Krishna et al. (2017) and models Xiong et al. (2016), hierarchical question-image co attention Lu et al. (2016), Lu et al. (2015),Saito et al. (2017) etc.,

Recent work has combined symbolic program execution for reasoning with deep representation learning for visual identification as part of the ongoing investigation in Visual Question Answering (VQA). This novel method, called neural-symbolic VQA (NS-VQA) (Yi et al., 2018), involves extracting a program trace from questions and a structural scene representation from images. Then, the program is run on the scene representation to extract responses. Three key benefits have been identified by NS-VQA research: greater robustness against long program traces, as evidenced by an astounding 99.8 percent accuracy on the CLEVR dataset; improved memory and data usage

efficiency, demonstrating promising performance even with a small amount of training data and less storage needs; and greater reasoning process transparency, enabling careful interpretation and diagnostic analysis of each execution step. These developments highlight how NS-VQA can help improve the functionality and interpretability of VQA systems.

Attention mechanisms were also used to the vision to language tasks. In vision-to-language activities, the issue of "where to look" (Shih et al., 2016) is addressed by the use of visual attention. In VQA we use the question to search for the relevant regions in the image.(Yu et al., 2017) proposes a stacked attention model which queries the image for multiple times to infer the answer progressively.

A novel model has been developed to enhance the interaction between vision and language modalities, thereby improving performance on tasks such as VQA. The paper (Kim et al., 2021) introduced Vision and Language Transformer (ViLT), which aims to simplify the processing of both visual and textual information within a unified framework. Unlike older methods that use complicated image analysis techniques, ViLT simplifies things by treating images more like words. This makes it much faster and still performs just as well or better on tasks like VQA. It suggests focusing future research on making ViLT even better by using larger datasets and exploring new ways to teach it about images and language. Additionally, Kim et al. (2021) highlights the importance of finding better ways to improve ViLT's understanding of images through techniques like masked modeling and smart data augmentation.

4 Proposed Methodology

Understanding the model's reliability and robustness is a major concern in my work. The proposed approach involves the use of Vision-and-Language Transformer (ViLT) as the base model for solving the problem at hand. By now, we knew that ViLT, a state-of-the-art model, is known for its ability to process both visual and linguistic inputs simultaneously, making it an ideal choice for tasks that involve understanding both images and text.

The model will be fine-tuned on a specific dataset relevant to the problem to ensure it learns the necessary features and patterns. Fine-tuning involves adjusting the pre-trained parameters of the ViLT model using our dataset, enabling it to learn task-specific features and patterns (Howard & Ruder, 2018). The fine-tuning process will involve monitoring the decrease in loss to assess the model's improvement and adaptation to the dataset, as suggested by Devlin et al. (2018) in their work on BERT.

Once the model is fine-tuned, the next step is to analyze its response to perturbations in the inputs. This will involve interchanging the question-image pairs and observing the model's responses systematically. The aim is to understand how the model reacts to unexpected or altered inputs, which can provide insights into its robustness and generalization capabilities. This approach is inspired by previous studies that have used input perturbations to test and understand deep learning models (Fong & Vedaldi, 2017), (Raffel et al., 2020).

Lastly, to provide statistical evidence that the model's confidence is affected when the inputs are perturbed, we conducted a series of statistical tests. These tests will compare the model's confidence scores, which are the predicted probabilities on the original and perturbed inputs, which were inspired by Dror et al. (2018) and followed by Hastie et al. (2009). A significant decrease in confidence scores for perturbed inputs would suggest that the model is sensitive to the changes of the inputs made, giving a numerical estimate of the model's reaction to a disturbance in the input. This approach aligns with previous research that has used statistical testing to validate observations in deep learning models (Nguyen et al., 2015).

To conclude, by harnessing the power of the ViLT model and employing a robust methodology in this study aimed at contributing to a deeper understanding of how deep learning models respond to input perturbations. The statistical evidence gained from our testing will not only corroborate our findings but will also yield significant insights into the sensitivity of our models to input variations and the underlying relationships. Thus, this knowledge could potentially catalyze the development of exceptionally robust and reliable models in the future, enhancing their efficacy across diverse scenarios and applications.

5 EXPERIMENTS

In my experimental setup, I am considering the VQA v2 dataset [2]. This dataset was introduced to address the language bias in the original VQA dataset, where the model was answering the questions correctly without even reading the images (Goyal et al., 2017).

The dataset contains a balanced set of images and corresponding questions, with each question paired with 10 human-generated answers. More than 200,000 photos from the COCO dataset [3] are included in the dataset. Three questions are linked to each image, for a total of more than 600,000 questions. A broad range of subjects were covered by the questions, including objects, actions, quantities, attributes, and other spatial relationships in the image. Since the dataset is not limited to a specific set of classes, models trained on VQA v2 will have a good understanding of both visual and linguistic modalities and be able to reason the relationships so that the models could accurately answer natural language questions about a given image, with the performance of the model evaluated based on how well its answers match the human-generated answers (Goyal et al., 2017).

For this I have utilized ViLT model, a vision and language Transformer. This Transformer model completely bridges the gap between vision and language understanding by jointly processing visual and textual information. Text information are tokenized into sequence of Tokens, each token is then embedded using a word embedding matrix, which maps each token to continous vector representation in a high dimensional space. To encode each token's position in the sequence, position embeddings are additionally appended to its embedding. This combination of word and position embeddings represents a contextualized representation of text which is essentially capturing both the semantic meaning of the tokens and their relative positions with in the sequence. Similarly, the images are initially divided into patches and then projected into embeddings using linear projection layers. Additionally position embeddings are added to each patch's embeddings to encode its spatial position within the image. Both the embeddings are then combined along the embedding dimension which means embeddings from both the modalities are stacked together side by side forming a single combined sequence of embeddings to form a single representation as shown in the below image:
.

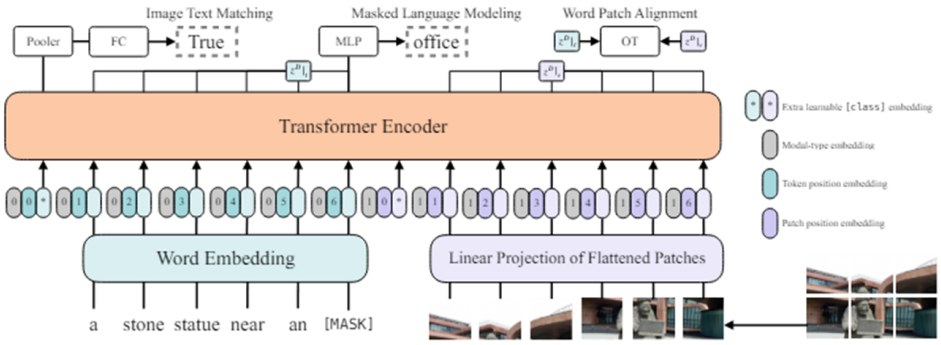

Figure 1: ViLT Architecture

The speciality of this ViLT design is that it enables the model to capture visual information in the same manner as text eleminating all the traditionally Convolutional and preprocessing steps making the process seamlessly faster and computationally efficient parallely enabling the model to learn the complex relationships between both the modalities more effectively. Since the embeddings are fused from different modalities it is termed as cross-modal fusion or multimodal fusion and since the fusion is happening at input level this is termed as Early fusion or feature-level fusion or input-level fusion (Baltrušaitis et al., 2018). This unified representation is fed through a series of

[2] https://visualqa.org/download.html
[3] https://cocodataset.org

transformer layers which utilizes the self attention to attend the relative parts of the input sequence. Finally, the output of the last transformer layer is pooled to obtain a single unified representation of the entire multimodal input.

I have initialized the model with the pretrained ViLT weights, and used the same input representations with the training dataset of VQA-v2. To finetune it on this specific task, I have chosen the Adam optimizer and other hyperparameters based on the similar work done by (Gonella) and optimized the model for our specific task by minimizing the task-specific loss function and manually verifying the confidence of the model's prediction over the epochs.

Upon successful fine-tuning the ViLT model on the dataset and confirming an improvement in the model's confidence, the subsequent involved in investigating the model's response to perturbations in the inputs which entailed in systematic perturbations to the input data, such as interchanging the question-image pairs and vice versa, changing the order of the text in the question, providing irrelevant questions in all the possible ways including the question types and observing how the model's predictions and confidence scores change in response aligning with the goal of how robust the model is to different types of input attacks and identify the weakness in its performance.

To statistically validate the model's response to perturbations, I have conducted hypothesis testing to compare the model's performance on the original and perturbed inputs which includes paired t-test assuming the distribution between two groups to be Normal and the difference between the pairs are independent which was inferred from the work of (Student, 1908), and Mann-Whitney U test which is non-parametric alternative which does not assume normality as mentioned in the work of (Mann & Whitney, 1947). The p- value reflects the probability of obtaining the observed statistic where it is compared over alpha, the alpha 0.05 which is the common threshold for statistical significance (Fisher, 1970).

The Null hypothesis for the paired t-test was: "There is no difference in the means of the model's confidence scores between the original input data and the perturbed input data".
Similarly for the Mann-Whitney U test the Null Hypothesis was defined as: "There is no difference in the distributions of the model's confidence scores between the original input data and the perturbed input data".

6 RESULTS

Based on the experiment performed, it can be observed that there is a noticeable decrease in loss over epochs which could be seen from the image below. Initially started at 1600 and gradually diminishing. However, it is noted that the reduction in loss becomes somewhat variable after dipping below 100. The model was chosen at tenth epoch.

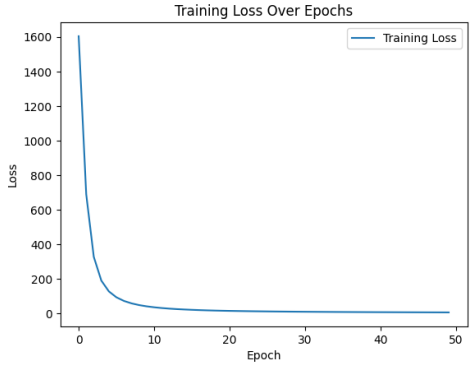

Figure 2: Loss values

We can see the performance of the model after fine-tuning it where its confidence seems to be increased as shown in the image below:

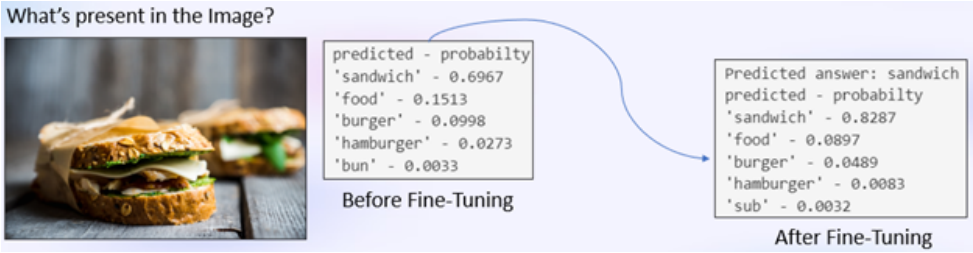

Figure 3: comparision in predictions

Moving to the Statistical test, the below tables shows the results of both paired t-test, and Mann-Whitney U test:

T-test Results	
t-statistic	-3.55928207841195
p-value	0.00447849391645557

Table 1: T-test Results

Mann-Whitney U Test Results	
U-statistic	23.0
p-value	0.005107905424995495

Table 2: Mann-Whitney U Test Results

The paired t-test (Student, 1908), the t-statistic, derived from the difference between the means of two groups divided by the standard error of this difference, is resulted as -3.559. A more negative t-statistic implies that the first mean is further below the second mean. The p-value reflects the probability of obtaining the observed t-statistic, or a more extreme value, under the assumption that there is no actual difference between the population means, which resulted to 0.0045.

For the Mann-Whitney U Test (Mann & Whitney, 1947), the U-statistic serves as an indicator of the discrepancy between the ranks of two groups, yielding a value of 23.0 in this instance. A smaller U-statistic suggests that the ranks within one group trend towards being lower than those within the other group. The p-value signifies the likelihood of observing the given U-statistic, or a more extreme one, which is 0.0051 for this test.

7 Discussion

Based on the results of my experiment, the loss curve indicates that the model is progressively learning to better fit the training data. However, fine-tuning it across numerous epochs proved to be computationally demanding, even the machine is still equipped with an Nvidia RTX 4060 GPU, 16GB of memory, with 3072 CUDA Cores. Additionally, manually perturbing the inputs and recording them in the dictionary posed considerable challenges throughout the process.
The statistical test results shown above, with p-values of 0.004 and 0.005 are less than the alpha, indicating the observed difference in means is unlikely to have occured by chance in the case of paired t-statistic and observed difference in ranks is unlikely to have occurred by chance in the case of Mann-Whitney U test. Establishing that they are statistically significant leading to the rejection of null hypothesis (Fisher, 1970), (Neyman & Pearson, 1933). These results prove that the model is senstive to changes in the inputs and its confidence decreases when its inputs are perturbed.

When observed practically, upon examination of the images and corresponding predictions from the below:

Image ID	Question ID	Question Asked to the Model	Truth	Prediction	Perturbed	Probability
294	294007	What color is the man's shirt?	gray	gray	No	0.9988
294	1398000	What type of pants are those?	No Pants	jeans	Yes	0.5868
1398	294001	What does the black machine next to the man produce?	Nothing	Nothing	No	0.2748
1398	294003	What color is the wall?	No wall	red	Yes	0.3584

An interesting observation can be made. The image depicted a man wearing a gray shirt and holding utensils in a kitchen. When the model was asked "what color is the man's shirt?" without any

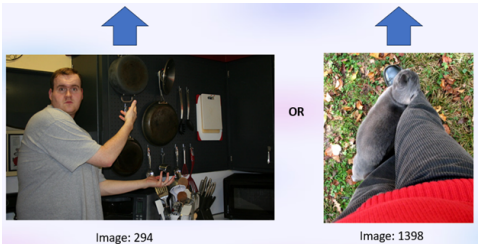

Figure 4: Sample of Stored Predictions 1

perturbation, it correctly predicted ”gray” with high confidence 99%. However, when the question was perturbed with ”What type of pants are those?” which was the image-question pair from the second image, the model predicted ”jeans” with only 58% confidence, despite the fact that the image does not show the man’s pants at all. This tells that the model’s confidence decreases when presented with perturbed inputs, and that the model may struggle to make accurate predictions when the relevant information is not present in the input which tells how robust the model was.

Similarly in the other Figure as well:

Image ID	Question ID	Question Asked to the Model	Truth	Prediction	Perturbed	Probability
1014	1014006	What are the boys doing?	posing	posing	No	0.7877
1014	1025001	What is the heart drawn in?	sand	shirt	Yes	0.4529
1025	1025000	What is the shape around the dog?	heart	heart	No	0.7503
1025	1014006	What are the boys doing?	posing	sitting	Yes	0.3584

Figure 5: Sample of Stored Predictions 2

The image depicts a dog sitting in the beach around a hand drawn heart symbol on the sand around it. When the model was asked, ”what’s the shape around the dog” it predicted ”heart” with 75% confidence, but when it is subjected to perturbation, this time the question from the other image, ”What are the boys doing” was asked, it predicted ”sitting” but with 38% confidence, which is significantly less.

These really show that the model is reacting to changes in the subject of the inputs, yet it is interesting to note that even when the model’s confidence decreased significantly under perturbation, it still maintained a certain level of confidence in its predictions. This indicates that the model may still be able to extract some information from the image or query, even if they are not immediately relevant to each other?. However, it is unclear what specific features or cues the model is relying on to make these predictions, and further analysis may be necessary to gain a better understanding of its decision-making process.

Furthermore, it is critical to assess the model’s potential consequences in real-world applications, where inaccurate predictions with even low confidence can have serious effects. For example, considering a medical diagnosis application where the model is trained to predict a disease based on a patient’s symptoms and medical history, if the model predicts a mild condition with low confidence, but the actual condition is severe, it could lead to a delay in providing appropriate treatment, delayed diagnosis, or mismanagement of the patient’s condition, resulting in serious

health implications for the patient. Therefore, it is crucial to thoroughly evaluate the model's performance and potential consequences.

8 LIMITATIONS

The performance of the model may be limited by the biases present in the dataset (Mehrabi et al., 2021), where the dataset doesn't cover all the real world scenarios.

In my study I have considered only few types of perturbations. However, there are many other types in the real world scenarios including the blurred images, lighting conditions, occlusions, adversarial attacks (Kurakin et al., 2018), (Szegedy et al., 2013).

The results obtained in this study were specific to the ViLT model and the dataset used for fine-tuning. The findings may not generalize to other models or datasets (Sculley et al., 2015). It might vary based on the training, the datasets, and other architectural changes!

The most crucial aspect, majority of the Deep Learning models are black box, which are not interpretable and explainable (Fong & Vedaldi, 2017), (Samek et al., 2017). Especially in our study, the model is trying to predict something with a certain confidence level even when subjected to perturbation, as seen above, but it is difficult to understand what features the model is using to make a specific decision or prediction!

9 FUTURE WORK

It would be interesting to explore how changes to the ViLT model's architecture, such as adding attention mechanisms or increasing the number of layers, affect its performance and robustness to perturbations. Identify modality interactions and come up with better fusion strategies like projecting image embeddings to text token space (Kiela et al., 2019).

Based on the findings of the study, it is evident that understanding the model in its prediction with a certain confidence is of utmost importance. Therefore, a potential direction would be developing the techniques for improving the interpretability (Montavon et al., 2018) and explainability (Guidotti et al., 2018) of multimodal models including identifying and visualizing the modality interactions, visualizing the attention mechanisms. This would not only enhance our understanding on the model's behaviour but fairly increases the chance of robustness and reliability in its predictions!

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Debapriya Banerjee, Fotios Lygerakis, and Fillia Makedon. Sequential late fusion technique for multi-modal sentiment analysis. In *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*, pp. 264–265, 2021.

Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3435–3444, 2019.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392, 2018.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70. Springer, 1970.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14773–14783, 2023.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015.

Giacomo Gonella. Visual language models: an in-depth exploration of vilt.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pp. 4193–4202, 2017.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Douwe Kiela and Stephen Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2461–2470, 2015.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. Deeper lstm and normalized cnn visual question answering model. *GitHub repository*, 6, 2015.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133, 2018.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014.

Rupayan Mallick, Thinhinane Yebda, Jenny Benois-Pineau, Akka Zemmari, Marion Pech, and Helene Amieva. Detection of risky situations for frail adults with hybrid neural networks on multimodal health data. *IEEE MultiMedia*, 29(1):7–17, 2022.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. A review of deep learning techniques for speech processing. *Information Fusion*, pp. 101869, 2023.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE journal of biomedical and health informatics*, 25(8):3121–3129, 2021.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Sandeep Kumar Pandey, Hanumant Singh Shekhawat, and SR Mahadeva Prasanna. Deep learning techniques for speech emotion recognition: A review. In *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, pp. 1–6. IEEE, 2019.

Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors*, 23(5):2381, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015.

Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Dualnet: Domain-invariant network for visual question answering. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 829–834. IEEE, 2017.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4613–4621, 2016.

Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. The color of the cat is gray: 1 million full-sentences visual question answering (fsvqa). *arXiv preprint arXiv:1609.06657*, 2016.

Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

Student. The probable error of a mean. *Biometrika*, pp. 1–25, 1908.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4631–4640, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pp. 2397–2406. PMLR, 2016.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4709–4717, 2017.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49 (3):1110–1122, 2018.

Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004, 2019.