# MULQA: Adapting Multimodal Models to Unimodal Tasks by Ensembling FLAVA with ALBERT

**Akash Gujju**
gujju@usc.edu

**Anushka Kamath**
arkamath@usc.edu

**Trisha Mandal**
trishama@usc.edu

**Varsha Kini**
vpkini@usc.edu

## Abstract

Multimodal models are machine learning models that are capable of processing and integrating information from multiple modalities or sources, such as text, images, videos, and audio. Visual question answering (VQA) (Antol et al., 2015) is a multimodal task that requires understanding both visual and textual inputs to answer questions about an image. VQA is an excellent benchmark for evaluating and improving machine learning models in both Natural Language Processing and Image Processing since it requires an extensive understanding of both the image's visual content and the question's textual content. However, evaluating VQA models on text-only and vision-only tasks is crucial to assessing their language and context understanding. This paper compares the performance of the traditional FLAVA model (Singh et al., 2022) and the MULQA (ensembled FLAVA and ALBERT: A Lite BERT) (Lan et al., 2020) model on unimodal tasks. The goal is to evaluate the effectiveness of multimodal models on language-only and vision-only tasks to provide insights for future research. (GitHub Repo Link).

## 1 Introduction

FLAVA (A Foundation Language And Vision Alignment Model) is a powerful vision and language model that aligns language and visual information in a common embedding space. While it was primarily designed for tasks that involve both language and vision, such as image captioning and visual question answering, it can also be adapted for VQA text-only tasks. FLAVA can generate accurate answers to text-only questions by leveraging its ability to learn a joint representation that incorporates both language and visual information. In text-only VQA tasks, FLAVA can learn to represent words and phrases in a shared vector space and effectively capture their relationships. Additionally, FLAVA uses attention mechanisms to selectively focus on the most relevant parts of the input text, which can help it generate more accurate answers. Its large-scale pre-training on textual data can also help it learn a wide range of linguistic features useful for VQA text-only tasks.

VQA datasets tested on FLAVA text-only tasks are important because they can be used to evaluate the performance of VQA models on text-only inputs. By testing on these datasets, researchers can assess the ability of VQA models to accurately answer text-based questions and can compare the performance of different models. Our project aims to compare the performance of the traditional FLAVA and the ensembled FLAVA-ALBERT model, which we call MULQA, on text-only tasks. This is important because it allows us to assess the models' ability to understand language and context without visual input. If the ensembled model can perform well on text-only tasks, it suggests that it understands language well and is a better multimodal model that can perform unimodal tasks than traditional FLAVA. This model allows researchers to answer text-based questions more precisely. Moreover, we are testing both the original and ensembled models with different QA datasets not used for training in the original model to check performance on newer QA datasets.

## 2 Related Work

(Andrew et al., 2013) used the projections produced by two independent Autoencoders by the standard correlation between the features. This model makes the assumption that a vectorial space projection into the other is adequate to depict the qualities that two distinct modalities have in common. They tested the model by using MNIST and XRMB datasets (Westbury, 1994). Only the generated features' correlation was considered; the model was not tested

in a classification task.

(Ngiam et al., 2011) and (Andrew et al., 2013) have both described multimodal strategies; (Wang et al., 2016) provide a summary of these techniques, go into more detail about their goal functions, and suggest novel designs that combine these objective functions. The Deep Canonical Correlated Autoencoder (DCCAE) leverages both aims from prior efforts to enhance the correlation between derived features and fidelity to original data. The accuracy of this architecture is then tested and compared to others in unsupervised classification tasks in noisy MNIST and XRMB datasets; multimodal representations performed better in these tests.

A novel method for multimodal fusion utilizing paired crossmodal neural networks (BiDNN) is presented by (Vukotic et al., 2016). In order to directly map one modality to another, two neural networks were built, and their hidden layer weights were connected. A central layer in both networks that maps each given modality to a common representation space between the two provided modalities is the end result of this training. This architecture is compared to others utilizing metrics discovered via studying the MediaEval 2014 dataset (Ito et al., 2018a). This architecture links audio transcripts and video segments to connect a video to a particular notion (anchor), achieving better results in this classification test. The film was translated into human visual conceptions, and the transcripts were rendered with Word2Vec embeddings.

In addition to investigating unimodal and multimodal representations, (Ito et al., 2018b) evaluated how well they performed a classification task using real-world data. They proposed that the underlying unimodal representations that were utilized to create multimodal representations have a direct impact on them. By evaluating new combinations of multimodal representations, they built on the work of (Ngiam et al., 2011).

The authors of (Hagström and Johansson, 2022) suggest a brand-new method called Unicoder-VL that extends the functionality of the current Unicoder model to handle tasks involving both language and vision. The Unicoder-VL model incorporates pre-training on massive text corpora and fine-tuning on vision-and-language tasks to develop a shared representation for text-only inputs. They ran tests utilizing text-only inputs on a number of benchmark datasets, including VQA and GQA, to gauge Unicoder-VL's performance. The outcomes revealed that Unicoder-VL outperformed cutting-edge models that incorporate both text and visual inputs, proving the viability of the suggested method for converting trained vision-and-language models to text-only inputs.

## 3 Datasets

In this research paper, we present a summary of six datasets used for our QA unimodal tasks:

1. TextVQA (Text Visual Question Answering) (Singh et al., 2019) is a dataset for answering questions based on text and images. Introduced in 2019, the dataset contains over 45,000 questions, each with an associated image and a textual question. The images are taken from the real world and are selected to be challenging, featuring a variety of backgrounds, lighting conditions, and objects. The questions in TextVQA require both visual perception and language comprehension, making it a challenging benchmark for research in the field of Text Visual Question Answering. TextVQA has been used in numerous research papers and competitions and is freely available for academic research purposes.

2. SocialIQA (Social Interaction Question Answering) (Sap et al., 2019)is a dataset containing over 38,000 questions that focus on answering questions about social situations. Introduced in 2020, the dataset includes dialogues from various sources, such as movie scripts, TV shows, and online forums, covering a wide range of social situations, from dating and job interviews to family gatherings. SocialIQA is designed to be challenging, requiring both social intelligence and language comprehension to answer the questions. It has become a benchmark for research in the field of social question answering, contributing to several research papers and competitions.

3. CommonsenseQA (Talmor et al., 2019) is a dataset designed to test common sense reasoning abilities. Introduced in 2019, it contains over 12,000 multiple-choice questions, each with five answer choices, encompassing a range of common sense reasoning abilities, such as causality, temporal reasoning, and spatial reasoning. The questions are drawn from various sources and cover diverse topics, including science, history, and everyday

life. CommonsenseQA requires understanding common sense and general knowledge, making it a challenging benchmark for evaluating natural language processing models. It has been used in numerous research papers and competitions.

4. WikiQA (Yang et al., 2015) is a dataset for open-domain question answering based on Wikipedia articles. Introduced in 2015, the dataset contains over 3,000 question-answer pairs, with questions designed to be open-ended and diverse, covering both fact-based and opinion-based topics. The answers are short passages extracted from corresponding Wikipedia articles. WikiQA requires both language understanding and knowledge retrieval, making it a popular benchmark for evaluating open-domain question-answering systems. It has been used in several research papers and competitions and is freely available for academic research purposes.

5. PubMedQA (Jin et al., 2019) is a biomedical question-answering dataset consisting of questions and corresponding answers from articles in the PubMed database, a large collection of biomedical literature. The dataset contains over 1.3 million question-answer pairs, with questions covering a wide range of topics in the biomedical domain, such as anatomy, diseases, drugs, and genetics. The answers are extracted from abstracts and full texts of PubMed articles, providing concise and accurate information. PubMedQA is designed to support research in biomedical question-answering and natural language processing, helping clinicians, researchers, and patients access and understand biomedical information more easily and efficiently. It is freely available for research purposes and can be downloaded from the official website.

6. The General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019) is a resource for training, evaluating, and analyzing natural language understanding systems. It includes several datasets, such as the Microsoft Research Paraphrase Corpus (MRPC), which has sentence pairs from online news sources annotated for semantic equivalence. The MRPC dataset has 5,801 training pairs and 1,725 validation pairs. The GLUE benchmark aims to provide a standardized platform for comparing models across diverse tasks, promoting the development of models that capture human language nuances, like paraphrasing and semantic equivalence. This advances the field, encouraging sophisticated and accurate NLP models for real-world applications, such as sentiment analysis, machine translation, and information extraction. The GLUE evaluation process allows researchers to refine models and explore novel techniques and architectures to enhance performance and applicability.

7. Fashion-MNIST (Xiao et al., 2017) is an image dataset that can be used for picture classification tasks. It comprises of a 60,000-example training set and a 10,000-example test set. Each sample is a 28x28 gray-scale image with a label from one of the ten classes. The dataset is designed to replace the well-known MNIST dataset but with more difficult classification tasks. Fashion-MNIST can be utilized to do picture categorization, deep learning, and computer vision tasks. It is often used for comparing and assessing the performance of several machine learning models on this dataset.

8. Street View home Numbers (SVHN) (J. et al., 2013) is a real-world picture dataset created by extracting home numbers from Google Street View photographs. The SVHN dataset contains almost 600,000 digit images, each with a size of 32 by 32 pixels, and is divided into three sets: train, test, and extra. The train and test sets each have 73,257 and 26,032 photos, while the additional set has 531,131 images. The SVHN dataset's purpose is to recognize the digit contained within each image. The SVHN dataset is commonly used to train and test machine learning and computer vision algorithms that recognize and classify digits in real-world photographs.

## 4 Architecture

### 4.1 Baseline Architecture

The FLAVA (Foundational Language And Vision Alignment) model is an advanced transformer-based architecture designed to learn a unified representation of vision and language. This allows

it to perform multimodal reasoning and unimodal language and vision understanding tasks. FLAVA comprises three main components: an image encoder, a text encoder, and a multimodal encoder, all of which are based on transformer architectures. They have the following characteristics:

1. The image encoder is founded on the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2021), a state-of-the-art deep learning model for image recognition tasks. The ViT takes an image as input, divides it into smaller patches, and processes each patch using a transformer network. This strategy allows the image encoder to capture both local and global contextual information from the input image.

2. The text encoder utilizes the BERT (Bidirectional Encoder Representations from Transformers) architecture (Devlin et al., 2018), which is a widely-used transformer model for natural language processing tasks. BERT takes tokenized text as input and processes it using a bidirectional transformer network. This enables the text encoder to effectively capture complex linguistic structures and relationships within the input text.

3. The multimodal encoder plays a vital role in combining the image and text representations by aligning and fusing them into a single representation. This is achieved using a separate transformer network that inputs the hidden state vectors from the image and text encoders. The multimodal encoder allows the model to reason across both modalities effectively by learning to align the visual and textual information.

### 4.2 MULQA Architecture

In our project, we modify the FLAVA baseline model architecture by replacing the BERT text encoder with the ALBERT text encoder. Ensembling FLAVA with ALBERT can be useful for adapting multimodal models to unimodal QA tasks. We have based our observations of the performance of ALBERT on the paper (Lan et al., 2020).

1. **Improved performance**: Ensembling FLAVA with ALBERT combines the strengths of multimodal learning with textual processing to improve performance on unimodal QA tasks.

2. **Complementary representations**: The ensembled model combines complementary representations from different pre-training methods, loss functions, and architectures to produce a more robust model.

3. **Flexibility**: Adjustable weighting of ensembled model outputs can optimize task performance depending on modality focus.

4. **Transferability**: The ensembled model can enhance performance on downstream tasks involving text and image inputs, such as visual dialogue and image captioning.

ALBERT is a more efficient and lighter version of the original BERT (Bidirectional Encoder Representations from Transformers) model (**?**). It is designed to overcome some of the limitations of BERT, such as high memory consumption and slow training speed. ALBERT and BERT differ in their parameter-sharing strategy and tokenization techniques. ALBERT employs cross-layer parameter sharing, reducing the number of parameters in the model without sacrificing its expressiveness, resulting in faster training and lower memory usage. Furthermore, ALBERT uses a SentencePiece(Kudo and Richardson, 2018) tokenization technique, which supports multiple languages and improves the model's ability to handle rare and out-of-vocabulary words. Overall, ALBERT has demonstrated competitive performance compared to BERT on a variety of natural language understanding tasks while being more efficient in terms of computational resources and training time.

## 5 Experiments

### 5.1 Replacing Text Encoders

Our initial experiments aimed to enhance the FLAVA text encoder by integrating ALBERT and DistilBert embeddings. We used the text tokenizers available at hugging-face transformers (Wolf et al., 2020) package to generate the required text embeddings for our model. To do so, we utilized a PyTorch fine-tuned model for classification and developed a pipeline to externally encode our text and image embeddings before feeding them into the FLAVA multimodal encoder. To evaluate the performance of our model, we conducted a classification task on the TextVQA dataset, which involved assigning each example to one of 3,997 possible classes. To accomplish this, we incorporated a vocabulary file and used Softmax activation

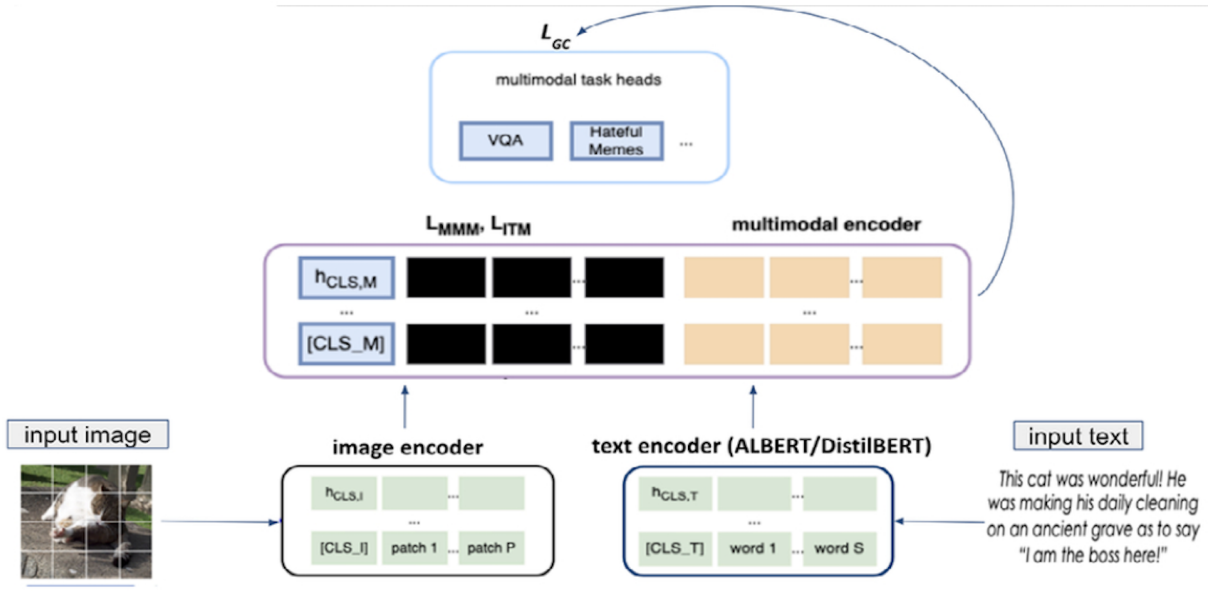


Figure 1: MULQA Architecture with ALBERT and DistilBERT Encoder

to extract the most likely class from the final layer of the output embeddings. Overall, our experiments demonstrated the potential for improving the FLAVA text encoder with advanced embeddings and highlighted the importance of developing robust pipelines for multimodal data processing.

### 5.2 Running Text-Only Unimodal Tasks

Our second experiment aims to evaluate the effectiveness of the modified FLAVA model architecture on text-only unimodal tasks. In this setup, the model receives only text inputs and must answer questions based solely on the contextual information provided by the text. To achieve this, we begin by passing the text inputs through the new ALBERT and DistilBERT text encoders, which generate fresh embeddings for the text. These embeddings are then sent to the multimodal encoder, which combines them with image representations to create a final representation for the input.

However, in this experiment, we block off the output from the image encoder that feeds into the multimodal encoder for which we have referenced the paper (Hagström and Johansson, 2022). This means that the model receives black-and-white images or tensors of 0's and 1's, which prevents it from accessing any visual context. As a result, the model must rely solely on text embeddings to answer the questions, as we do not learn anything from the image embeddings.

By testing the performance of this modified model architecture on text-only unimodal tasks, we can assess the efficacy of the new ALBERT and DistilBERT text encoders compared to the original BERT text encoder used in the baseline FLAVA model. By blocking off the image inputs, we can also evaluate the extent to which the model depends on visual context when answering questions.

Furthermore, by benchmarking the performance using different text encoders, we can compare their performance against each other and against the original BERT text encoder. This will enable us to determine which text encoder works best for text-only unimodal tasks.

### 5.3 Running Vision-Only Unimodal Tasks

In our third experiment, we aim to assess the performance of the modified FLAVA model on image-only tasks. To achieve this, we block off the text encoder by passing an empty or "What is this image?" string, which ensures that the model processes only image inputs or multimodal tasks on the embeddings generated by the latter string.

To evaluate the model's ability to recognize or classify images, we can provide it with a set of images and ask it to predict the label or category of each image. This can be accomplished using labeled image datasets such as SVHN and Fashion MNIST datasets, where each image is associated with a class or category.

For this experiment, we use the same image encoder as in the base FLAVA model based on the Vision Transformer architecture. This model divides the input image into smaller patches and processes

each patch using a transformer network. The output from the image encoder is then fed into the multimodal encoder, which combines it with the text blank embeddings to generate a final representation.

By blocking off the text encodings and relying only on image modalities, we can evaluate the model's performance on image-only tasks and compare it to other state-of-the-art models in the field. This will enable us to determine how well the modified FLAVA model compares to other image recognition or classification models and identify any areas where it can be further improved.

## 6 Results

We evaluated the MULQA model on the eight unimodal datasets.

| Model | Accuracy |
|---|---|
| FLAVA-BASE | 54.12 |
| FLAVA-DistilBert | 51.83 |
| FLAVA-ALBERT | 51.23 |
| **SOTA Model** | |
| LXMERT | 66.7 |

Table 1: Performance of the Ensembled FLAVA, Ensembeled Models on TextVQA.

The table presents the comparative performance of the FLAVA, FLAVA-DistilBERT, and FLAVA-ALBERT models on the TextVQA dataset. We have also compared it with SOTA models for this task which are LXMERT (Language Cross-Modal Pre-training for Vision-and-Language Representation) (Tan and Bansal, 2019), a multi-modal pre-trained language model which had been developed by researchers at the University of Illinois Urbana-Champaign, Carnegie Mellon University, and Facebook AI Research. LXMERT is an intricate model that integrates multiple components to support cross-modal understanding. LXMERT achieves cutting-edge performance on a range of multi-modal benchmarks by combining textual and visual input processing.

The table presents the comparative performance of the FLAVA, FLAVA-DistilBERT, and FLAVA-ALBERT models on six different QA datasets. From the results, we can derive the following insights:

1. Multimodal vs. Unimodal Tasks: BERT performs better in the TextVQA dataset, a multimodal task, than ALBERT and DistilBERT. However, in unimodal tasks, ALBERT consistently outperforms both BERT and DistilBERT across all datasets. This could indicate that ALBERT's architecture and training methodology might be better suited for unimodal tasks, while BERT may have advantages in multimodal tasks.

| Dataset | BERT | ALBERT | DistilBERT |
|---|---|---|---|
| WikiQA (black images) | 75.09 | 87.27 | 66.75 |
| WikiQA (white images) | 66.50 | 83.75 | 58.75 |
| PubMedQA | 53.10 | 55.20 | 51.80 |
| Social IQa | 33.83 | 34.24 | 32.70 |
| CommonsenseQA | 19.98 | 20.15 | 18.76 |
| GLUE MRPC dataset | 50.65 | 60.81 | 45.04 |

Table 2: Performance of Ensembled FLAVA and BERT Models on Unimodal Text Tasks.

2. Performance Variability: There is a noticeable variability in performance across different unimodal datasets. For instance, all models perform relatively well in the WikiQA datasets, while their performance is considerably lower in the CommonsenseQA dataset. This could be attributed to the complexity and nature of the tasks, with CommonsenseQA requiring a deeper understanding of commonsense knowledge, which might not be well captured by the models.

3. DistilBERT's Performance: DistilBERT is a smaller and faster version of BERT, designed to have a lower computational cost while maintaining similar performance. However, the results show that DistilBERT consistently underperforms in all datasets compared to BERT and ALBERT. This suggests that the trade-off in model size and computational cost might have reduced performance.

4. ALBERT's Strengths: ALBERT's strong performance across unimodal tasks could be attributed to its architecture, which utilizes parameter sharing and a smaller embedding size to reduce the model size and computational requirements while maintaining high performance. Additionally, ALBERT employs a different pre-training methodology compared to BERT, which could contribute to its superior performance in these tasks.

In summary, the analysis of the table highlights that the ensembled FLAVA and ALBERT model performs well on various datasets, particularly in unimodal tasks. ALBERT consistently achieves the highest accuracy compared to BERT and DistilBERT, which could be attributed to its architecture and training methodology. However, the variability in performance across different tasks suggests that there might still be room for improvement and exploration of other model architectures and pre-training strategies to achieve better performance in specific tasks.

We evaluated MULQA on two image datasets to test image classification unimodal tasks.

| Dataset | Null Question | "What is this image" |
|---|---|---|
| Fashion MNIST | 83.96 | 83.96 |
| SVHN | 36.18 | 34.48 |

Table 3: Performance of the Ensembled FLAVA on Unimodal Image Classification Tasks.

The consistent performance of the MULQA model on unimodal image classification tasks, regardless of the question type, suggests that the model is primarily relying on its visual understanding capabilities to classify images. This hypothesis implies that the textual information provided by the questions ("null string" or "What is this image?") does not significantly influence the model's decision-making process in these unimodal tasks. Further investigation and experimentation would be necessary to validate this hypothesis and explore the extent to which the model leverages its textual understanding capabilities in different contexts.

| Dataset | MULQA-Acc | SOTA-Model | SOTA-Acc |
|---|---|---|---|
| WikiQA (black) | 87.27 | BERT | 89.8 |
| WikiQA (white) | 66.50 | BERT | 89.8 |
| PubMedQA | 53.10 | BioBERT | 83.8 |
| Social IQa | 33.83 | GPT-3 | 66.3 |
| CommonsenseQA | 19.98 | GPT-3 | 90.4 |
| GLUE MRPC | 50.65 | GPT-3 | 88.9 |
| Fashion MNIST | 83.96 | CNN | 99.6 |
| SVHN | 36.18 | CNN | 98.2 |

Table 4: Performance of the MULQA Model vs. Comparable SOTA Models.

The performance of the MULQA model, when compared to the state-of-the-art (SOTA) models on various datasets, suggests that while MULQA demonstrates competitive performance in some cases, it falls short in others. This discrepancy may indicate that MULQA's multimodal approach has not yet fully captured the nuances required for certain tasks or may not be as well-suited to some tasks as specialized models.

One possible hypothesis is that the MULQA model's performance varies due to the nature of the tasks and the extent to which they require a joint understanding of the text and visual information. In cases where the task demands more textual understanding or domain-specific knowledge, MULQA may struggle compared to models like GPT-3 or BioBERT, which have been specifically designed or fine-tuned for these tasks. On the other hand, in cases where the task requires a more balanced combination of textual and visual understanding, MULQA may perform comparably to models designed solely for text-only tasks. We can observe that for image classification tasks CNNs are still the best models and are far superior to our multimodal model.

Further research and experimentation would be necessary to validate this hypothesis and explore possible improvements to the MULQA model to enhance its performance across a broader range of tasks, leveraging its multimodal capabilities to their full potential.

## 7 Limitations

1. FLAVA's performance relies heavily on the amount and quality of training data. While the training data used to train FLAVA is relatively large, it is still small compared to the vast amount of data available on the internet.

2. We believe the traditional FLAVA model's performance can be improved by capturing additional text embeddings. Currently, ALBERT is the text encoder in FLAVA, and it does not have the integrated text and visual modalities found in some other multimodal models.

3. The FLAVA model, which combines text and image encoders, results in high computational requirements for training and inference. This can make scaling to larger datasets or deployment in resource-constrained environments difficult.

4. The adaptation methods perform differently for different models, and unimodal model

counterparts perform on par with the Visual Language models regardless of adaptation, indicating that current Visual Language models do not necessarily gain better language understanding from their multimodal training.

## 8 Future Work

While our experiments and results provide valuable insights into the performance of MULQA and the FLAVA model, there are several directions for future work to further improve and expand upon our findings. These directions are outlined below:

1. **Hyperparameter Tuning**: To improve the prediction accuracy of our experiments, we plan to perform a comprehensive hyperparameter search for MULQA. This will involve trying different optimizers, learning rates, and other hyperparameters to find the best combination that maximizes the model's performance on the target tasks.

2. **Evaluating the Performance of FLAVA on Unimodal Tasks**: As part of our ongoing investigation into the capabilities of the FLAVA model, we intend to evaluate its performance on unimodal tasks, such as audio classification. This will help to better understand the model's strengths and limitations and applicability to a broader range of tasks.

3. **Improving MULQA Multiclass Accuracies**: In the future, we plan to explore different strategies to enhance the multiclass accuracies of MULQA for unimodal tasks. This may include experimenting with different model architectures, data augmentation techniques, and training strategies.

4. **Fine-tuning MULQA for Specialized VQA and QA Datasets**: To further demonstrate the versatility of MULQA, we aim to fine-tune the model on more specialized visual question-answering and question-answering datasets. This will help assess the model's ability to adapt to different types of questions and reasoning requirements and evaluate its potential for real-world applications. By pursuing these research directions, we aim to advance our understanding of MULQA and the FLAVA model and contribute to the development of more robust and versatile models for multimodal and unimodal tasks.

By pursuing these research directions, we aim to advance our understanding of MULQA and the FLAVA model and contribute to the development of more robust and versatile models for multimodal and unimodal tasks.

## 9 Conclusion

In this research paper, we presented an extensive study of the MULQA (ensembled FLAVA and ALBERT) model for various multimodal and unimodal tasks. Our findings show that the ensembled model performs well in QA unimodal tasks, with ALBERT consistently outperforming BERT and DistilBERT. This is due to ALBERT's distinctive architecture and training methodology, which allow it to achieve high accuracy while reducing the model size and computational requirements. We also tested MULQA on image unimodal tasks, demonstrating that it is capable of handling both visual and text modalities.

Despite these encouraging results, our analysis reveals some limitations and areas for future development. The variability in performance across tasks suggests that more research is needed to develop models that are more robust and adaptable to the unique challenges posed by each task. Furthermore, the study identified potential future research areas, such as hyper-parameter tuning, evaluating the FLAVA model on audio tasks, and fine-tuning for more specialized VQA and QA datasets.

Finally, our findings contribute to ongoing efforts to create models that effectively use both language and vision modalities for a variety of tasks. The MULQA model, when combined, represents a promising direction for multimodal learning, providing valuable insights into the strengths and limitations of existing approaches. We anticipate that future research will build on these findings to develop even more powerful and versatile models, paving the way for innovative applications and advancements in artificial intelligence.

## 10 Project Contributions

**Dataset Preparation**: Anushka Kamath, Varsha Kini
This section focused on the collection, processing, and organization of the datasets used for the training and evaluation of the models. This ensured that the data is clean, diverse, and suitable for the experiments. We focused on the datasets TextVQA for Visual Question Answering and other QA and

image classification datasets.

**Baseline Model Creation**: Akash Gujju, Trisha Mandal

In this part, we developed a baseline model, which served as a reference point for evaluating the performance of the proposed models. It was for designing the model architecture, selecting the appropriate loss functions and optimizers, and implementing the FLAVA model in a programming framework. We used CARC to set up the model pipeline after realizing that the compute was not sufficient to run on Google Cloud and Collab-Pro.

**Baseline Model Testing**: Anushka Kamath, Varsha Kini

This section involved testing the baseline model on the prepared datasets, evaluating the model's performance, analyzing the results, and identifying areas for improvement.

**Finetuning Model Architecture**: Akash Gujju, Trisha Mandal

This part was focused on refining and optimizing the model architecture to improve performance based on the baseline model testing results. Adjustments to the model components, hyperparameters, and training techniques were made to achieve better results.

**Generating Text Embeddings**: Akash Gujju, Trisha Mandal

In this section, we were responsible for generating text embeddings using the trained models. Feature representations were extracted from the models to further analyze and evaluate text-only unimodal tasks.

**Running Text-Only Unimodal Tasks**: All

This part was for setting up the experiments, monitoring their progress, and analyzing the results to draw conclusions about the models' performance on text-only unimodal tasks.

**Running Image-Only Unimodal Tasks**: Anushka Kamath, Varsha Kini

This section was for running the experiments on image-only unimodal tasks, which means setting up the experiments, monitoring their progress, and analyzing the results to assess the models' performance in these tasks.

**Documentation**: All

The project documentation included writing and organizing the research paper, summarizing the findings, and ensuring that the work was well-presented and accessible to the broader research community.

## References


Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA. PMLR.

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

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Lovisa Hagström and Richard Johansson. 2022. How to adapt pre-trained vision-and-language models to a text-only input?

Fernando Tadao Ito, Helena de Medeiros Caseli, and Jander Moreira. 2018a. The effects of unimodal representation choices on multimodal learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Fernando Tadao Ito, Helena de Medeiros Caseli, and Jander Moreira. 2018b. The effects of unimodal representation choices on multimodal learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Goodfellow Ian J., Bulatov Yaroslav, Ibarz Julian, Arnoud Sacha, and Shet Vinay. 2013. Multi-digit number recognition from street view imagery using deep convolutional neural networks.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Ng. 2011. Multimodal deep learning. pages 689–696.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing*.

Vedran Vukotic, Christian Raymond, and Guillaume Gravier. 2016. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2016. On deep multi-view representation learning: Objectives and optimization.

J.R. Westbury. 1994. *X-ray Microbeam Speech Production Database User's Handbook: Version 1.0 (June 1994)*. Waisman Center on Mental Retardation & Human Development.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.