SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE NAME: Machine Learning Laboratory **COURSE CODE:** DJS22L602

CLASS: Third Year B.Tech **SEM:** VI

NAME: Anish Sharma

EXPERIMENT NO. 7

CO Measured:

CO2 - Identify machine learning techniques suitable for a given problem.

TITLE: Mini Project: Stage 1

AIM / OBJECTIVE: Mini Project

Step 1: Literation survey, problem identification and data collection

Step 2: Data Preprocessing and Cleaning

Step 3: Selecting the Right Machine Learning Model

DESCRIPTION OF EXPERIMENT:

In this mini project you are expected to choose any algorithm in machine learning with respect to some use case of your choice. It can be a small-scale project where you apply machine learning algorithms to a specific dataset to solve a problem, often focusing on a single concept or technique, typically used for learning purposes and usually involving data collection, cleaning, feature engineering, model training, and evaluation within a manageable scope.

Key characteristics of a mini machine learning project to consider in this experiment:

1. Literature Survey:

- **Define your research topic:**

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

The research focuses on **bias mitigation in Vision-Language Models (VLMs)**, particularly in Visual Question Answering (VQA) systems. These models often inherit societal biases from training data, leading to problematic or unfair predictions.

- **Identify relevant keywords:**
  - Vision-Language Models
  - Visual Question Answering (VQA)
  - Dataset bias
  - Debiasing techniques
  - Fairness in AI
  - VQA 2.0 dataset
- **Access relevant databases:**

To understand the scope of bias in VLMs, papers and resources were reviewed from:

- Google Scholar
- arXiv
- Springer
- IEEE Xplore
- ResearchGate
- Kaggle notebooks and datasets
  Notable papers included:
- *"Gender and Racial Bias in Visual Question Answering Datasets"*
- *"AUnified Framework and Dataset for Assessing Societal Bias in Vision-Language Models"*
- *"Explicit Bias Discovery in Visual Question Answering Models"*

- **Critically analyze sources:**

Each source was reviewed for:

- **Methodology**: whether datasets were annotated with bias labels or restructured
- **Model architecture**: use of transformers, attention layers, etc.
- **Bias mitigation strategy**: including rule-based filtering, adversarial training, or data augmentation

- **Results**: effectiveness of each method in reducing bias without harming accuracy

◇ **Identify gaps in knowledge:**

Visual-Language Models (VLMs), as dual-modal systems integrating vision and language, represent a relatively recent advancement in the field of artificial intelligence. Research focused specifically on the biases present within VLMs is even more nascent. Existing literature predominantly concentrates on analyzing model outputs to identify explicit biases, often neglecting the subtle and systemic manifestations of such biases within the models' reasoning processes.

Our work seeks to address this gap by investigating how nuanced and implicit biases emerge in VLM outputs. Furthermore, unlike many studies that stop at bias identification, we extend our contribution by implementing and evaluating mitigation strategies. This dual approach—bias detection followed by mitigation—aims to move the discourse forward from diagnostic analysis to proactive solution development.

◇ **Summarize key findings:**

- Bias in VLMs often arises due to **imbalanced datasets** (e.g., VQA 2.0 containing more male subjects).
- Fine-tuning on **neutralized text and visual features** shows promise in mitigating bias.
- Combining **quantitative evaluation (bias score)** with **qualitative analysis (e.g., attention maps)** gives better insights.
- Techniques like **association rule mining** can uncover hidden bias patterns, offering actionable insights for mitigation.

2. Problem Identification:

◇ **Formulate a Research Question:**

How can we detect and mitigate gender biases in Vision-Language Models (specifically VQA models)?

◇ **Define the Scope of the Problem:**

- Focused on the **VQA 2.0 dataset**, targeting **biases in gendered question-answer pairs**.
- Analyzed **question words, visual features (e.g., detected objects, attention maps)**, and **model predictions** to find problematic associations.
- Scope includes: preprocessing multimodal data, visual feature extraction, and finetuning text and image encoders.

◇ **Justify the Problem's Importance:**

- VLMs are used in **critical applications** like education, healthcare, surveillance, and assistive technologies.
- Biased outputs can cause **ethical issues, unfair treatment**, and **reinforcement of stereotypes**.
- our project provides a **data-centric approach** to detect and reduce such biases, promoting fairness in AI.

3. Data Collection:

◇ Data Collection Methods:

- Utilized **secondary data** from:
  - **VQA 2.0 dataset** (image-question-answer triples)
  - **COCO 2014 images** (for visual content)
  - Extracted **object detection outputs** and **attention maps** for analysis

Data Collection Instruments:

- Used tools and scripts to:
  - Collect question-words using NLP tokenization
  - Analyzed visual features by freezing attention layers
  - Generate model predictions and probabilities

Pilot Testing:

- Tested pipeline on a **small batch (e.g., 7000 samples)** to verify:
  - Feature extraction consistency
  - Token-label pairing
  - Attention region validity

Key steps involved in data pre-processing and cleaning:

To ensure compatibility with the Vision Transformer (ViT) model and to prepare the VQA 2.0 dataset for bias detection and mitigation, we performed a series of structured data preprocessing and cleaning steps:

1. Textual Preprocessing (Questions and Answers)

- **Cleaning and Tokenization**: Questions were cleaned by removing irrelevant characters and structures. Tokenization was applied to break down each question into meaningful tokens suitable for the model.

- **Gender Neutralization**: To address potential gender biases in the dataset, gendered terms (e.g., *he*, *she*, *man*, *woman*) were replaced with gender-neutral alternatives (e.g., *person*, *child*, *parent*).
- **Labeling Answers**: Each answer was assigned a categorical label to make the data suitable for classification tasks.

2. Image Preprocessing

- **Linking Images**: Each question-answer pair was linked with the corresponding image using formatted image IDs.
- **Image Transformation**: Images from the COCO 2014 dataset were resized to standard dimensions (224x224), converted into tensors, and normalized to ensure consistency across the dataset.

3. Visual Feature Extraction

- **Visual Encoder Integration**: The preprocessed images were passed through a visual encoder (based on ViT architecture) to extract deep visual embeddings.
- **Attention and ROI Identification**: Attention maps and object detection techniques were employed to identify key regions of interest (ROIs) within each image. These ROIs helped focus the model on contextually relevant visual information while avoiding biased associations.

Key steps involved in Selecting the Right Machine Learning Model:

The Vision Transformer (ViT) was selected for this project due to its advanced capabilities in capturing global contextual relationships within images, making it particularly well-suited for tasks involving nuanced visual understanding and bias detection. The following key factors motivated this choice:

1. Superior Representation Learning

ViT treats an image as a sequence of patches (like tokens in text), enabling it to model long-range dependencies and spatial relationships more effectively than traditional convolutional neural networks (CNNs). This ability is crucial when analyzing subtle patterns or biases that span across different regions of an image.

2. Alignment with Transformer-Based VLMs

Since most modern Vision-Language Models (such as ViLBERT, LXMERT, and BLIP) use transformer-based architectures, ViT provides architectural compatibility and seamless integration for multimodal learning tasks, including VQA and bias analysis.

3. Better Attention Mechanisms

ViT's self-attention mechanism allows the model to focus on semantically important regions in an image. This is especially useful when used in conjunction with textual attention mechanisms to study how biases emerge in model outputs based on specific visual regions.

4. Robust Performance on VQA and Related Tasks

ViT has shown state-of-the-art performance on image classification and transfer learning benchmarks, and its embeddings have proven effective in downstream tasks like Visual Question Answering (VQA)—the core task in this project. Its pretraining on large-scale datasets makes it adaptable to domain-specific fine-tuning like bias mitigation.

5. Flexibility and Scalability

ViT is highly scalable and flexible, supporting varying input resolutions and patch sizes. This allows experimentation with different levels of visual granularity during bias analysis without altering the core model structure.

PROCEDURE:

1. Prepare a document with detailed content about your mini project including problem identification, literation survey and data collection.
2. Perform data preprocessing and cleaning on dataset.
3. Select ML model for you mini project, explain working of that algorithm n detail.

OBSERVATIONS / DISCUSSION OF RESULT:

1. List all challenges faced during each of these steps individually.

CONCLUSION:

Base all conclusions on your actual results; describe the meaning of the experiment and the implications of your results.

REFERENCES:

(List the references as per format given below and citations to be included the document)

1. Ethem Alpaydın, “Introduction to Machine Learning”, 4th Edition, The MIT Press, 2020.

2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012.

3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.

4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1st Edition, O'reilly, 2016.

5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012.

Website References:

[1] https://www.altexsoft.com/blog/data-collection-machine-learning/

[2]https://www.researchgate.net/publication/336992097_LITERATURE_SURVEY_ON_MACHINE_LEARNING_BASED_TECHNIQUES_IN_MEDICAL_DATAANALYSIS

[3]https://www.kaggle.com/code/alirezahasannejad/data-preprocessing-in-machine-learning

[4] https://medium.com/analytics-vidhya/data-cleaning-and-preprocessing-a4b751f4066f