Name: Saurabh Madake

MIS: 712352023

Title:

Transformer-Driven Visual Question Answering on Complex Datasets.

Problem Statement:

Tackling the AI-complete task of visual question answering utilizing transformers to improve accuracy in processing and understanding complex visual and textual data.

Introduction to the area of work:

Visual Question Answering (VQA) is a complex AI task that combines computer vision and natural language processing to develop models capable of interpreting visual content and accurately answering related questions. Traditional methods involved separate pipelines for visual and textual data, but advancements in deep learning, particularly with transformers like BERT, have revolutionized this process. Transformers excel at integrating multimodal data, significantly improving the accuracy and robustness of VQA models. These models, fine-tuned using tools like PyTorch and Hugging Face, address the intricacies of real-world visual data, making VQA a critical yet challenging area of AI research. VQA entails a wide range of sub-problems in both CV and NLP. Thus, it is considered an Ai Complete task.

Research gaps in the current literature:

- **Handling Complex and Ambiguous Queries**:
  While recent models have made progress in answering straightforward questions, many still struggle with complex or ambiguous queries that require nuanced understanding of the visual context.

- **Bias and Fairness in VQA Models**:
  Many VQA models inherit biases from the datasets they are trained on, leading to biased or unfair outcomes.
  Mitigation:
  - Develop tools for analysing and auditing model decisions to understand how biases influence outcomes.
  - Ensure that training datasets are diverse and representative of various demographics and scenarios

- **Generalization Across Diverse Datasets:**
  VQA models often perform well on specific datasets but struggle to generalize across different or unseen datasets. This limitation highlights a gap in developing models that are robust and adaptable to various types of visual and textual data, which is crucial for real-world applications.
  Mitigation:
    - Use transfer learning techniques where models are pre-trained on large, diverse datasets and fine-tuned on specific VQA tasks.

Proposed work, Overall objectives and Expected outcomes:

Proposed Work:

Development of a Multimodal Transformer Model:

- Implement a multimodal transformer architecture that integrates both textual and visual information.

Dataset Preparation:

- Utilize a dataset like DAQUAR, COCO-QA Dataset, VQA Dataset

Training and Fine-tuning:

- This involves optimizing the model to correctly answer questions based on the visual content of images and the text of the questions.

Evaluation and Refinement:

- Evaluate the model's performance using standard VQA metrics such as accuracy, precision, and recall.

Overall Objectives:

Enhance VQA Capabilities:

- Improve the accuracy and efficiency of visual question answering systems by leveraging advanced multimodal transformer models.

Advancement of AI Techniques:

- Contribute to the advancement of AI and machine learning techniques in the realm of multimodal learning and understanding.

Integration of Visual and Textual Information:

- Develop a robust mechanism to effectively integrate visual features from images with textual features from questions to generate precise answers.

Expected Outcomes:

High-Performance VQA Model:

- A well-trained multimodal transformer model that demonstrates high accuracy in answering questions based on visual inputs.

Insights into Multimodal Learning:

- Insights and findings on how different transformer architectures handle multimodal data, potentially providing guidelines for future work in VQA and related areas.

Contributions to Research:

- specifically in multimodal learning and visual question answering, which may include publications or presentations.