ABSTRACT

In the context of educational settings, where students often struggle to understand their machine learning courses or study slides, this research represents a significant step forward with the introduction of the Machine Learning Visual Question Explanation (MLVQE) dataset, a novel enhancement in Visual Question Answering (VQA). The MLVQE dataset, derived from a machine learning course, includes 885 slide images paired with 110,407 words from transcripts, forming 9,416 question-answer pairs. The cutting-edge SparrowVQE model, which amalgamates the strengths of two distinct models, SigLIP and Phi-2, undergoes a comprehensive three-stage training regimen multimodal pre-training, instruction tuning, and domain fine-tuning. This strategic training enables SparrowVQE to adeptly blend and interpret visual and textual data, markedly elevating its explanatory prowess. Demonstrating exceptional performance on the MLVQE dataset and surpassing existing VQA benchmarks, SparrowVQE offers students in-depth, context-aware explanations, significantly enriching their interaction with and comprehension of visual course material.