Katz
Katz School
of Science and Health

SparrowVQE: Harnessing Computer Vision and Language Processing for Enhanced Visual Question Explanation in Education

Radek Holik, Ruslan Gokhman and Manish Kumar Thota,
M.S. in Artificial Intelligence

Faculty Advisor: Youshan Zhang, Ph.D.

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.

INTRODUCTION

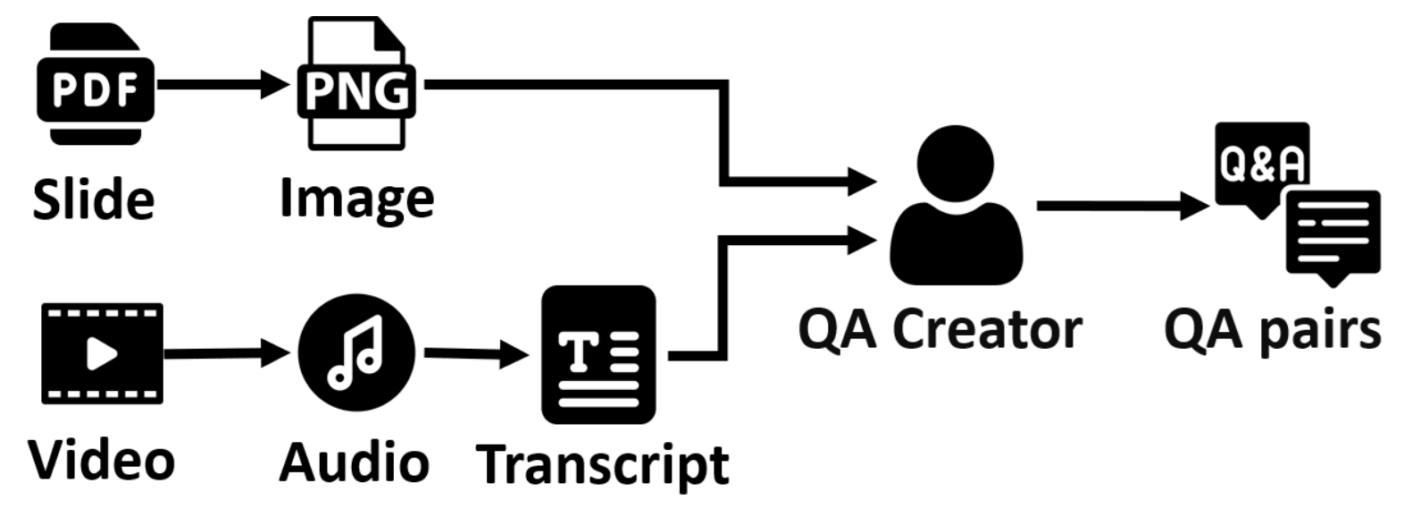
While Visual Question Answering (VQA) combines computer vision and language processing to interpret image-based queries, existing systems fall short in educational contexts, particularly when dealing with the detailed visuals and text found in machine learning lectures (Barra et al., 2021).

- **Problem Statement**: There is a significant gap in VQA research related to the educational content, where systems fail to adequately interpret and explain complex visual information in learning environments (Cheng, 2023).
- **Research Focus**: This study concentrates on enhancing VQA systems for educational purposes, specifically tailored to decipher and elucidate the rich visual and textual content found in machine learning lectures.
- **Approach**: This project introduces SparrowVQE, a Visual Question Explanation (VQE) model developed using a specialized dataset from machine learning lecture (Chen et al., 2020).
- **SparrowVQE** is a novel VQA system that offers detailed, context-aware explanations for educational content, specifically visual questions, enhancing learning in educational settings.

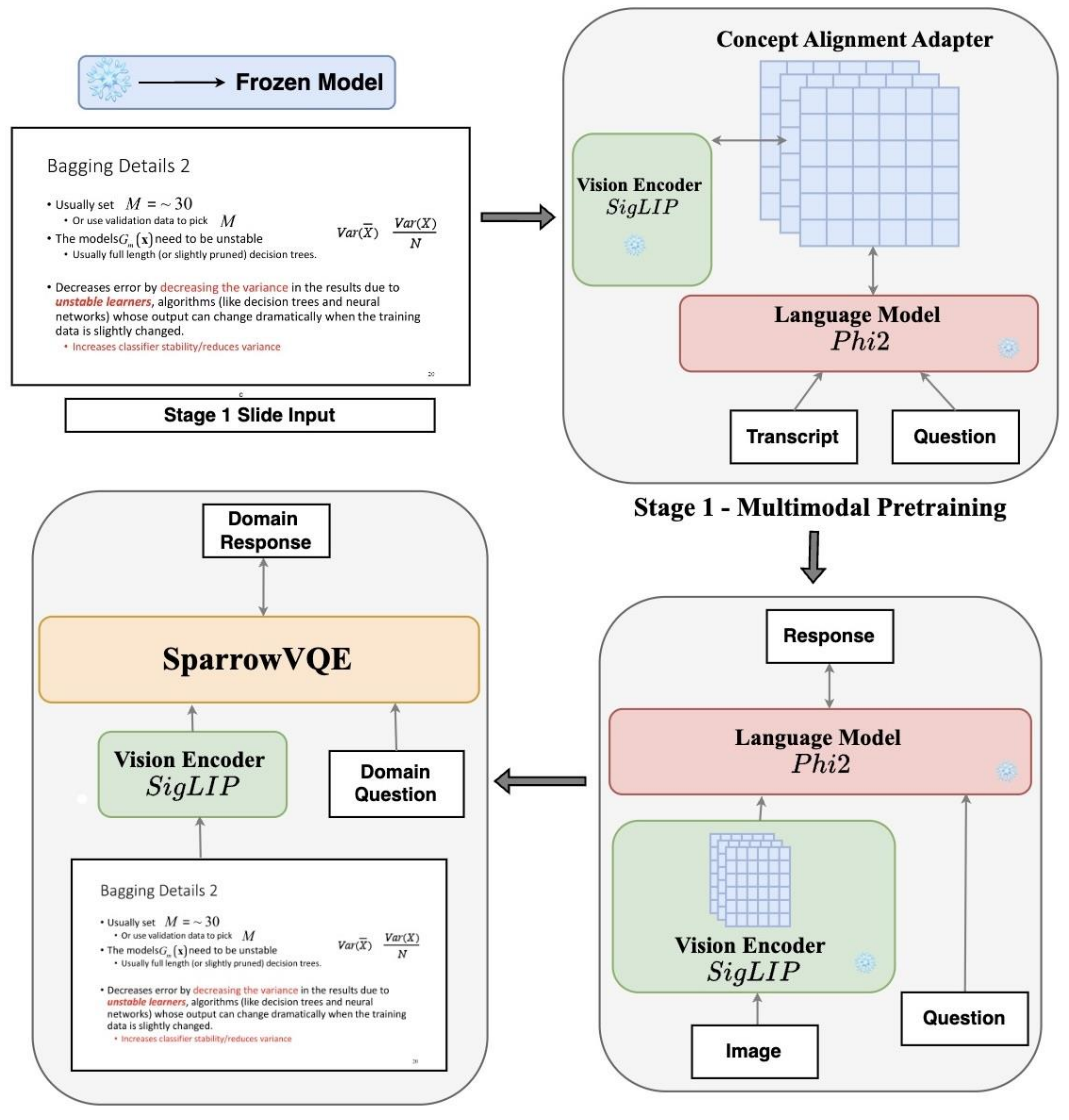
METHODOLOGY

SparrowVQE enhances educational VQA, particularly for machine learning lectures, by proficiently interpreting complex visuals.

Data Collection: Our dataset comprises 885 slides and transcripts, forming 9,416 question-answer pairs (Deepgram, 2023).



Training Process: SparrowVQE was optimized through a focused three-stage training process, tailored for educational content (Liu wt al., 2023).



Bagging Details 2

- Usually set $M = \sim 30$
  - Or use validation data to pick M
- The models $G_m(\mathbf{x})$ need to be unstable
  - Usually full length (or slightly pruned) decision trees.

$Var(\bar{X})$ $\frac{Var(X)}{N}$

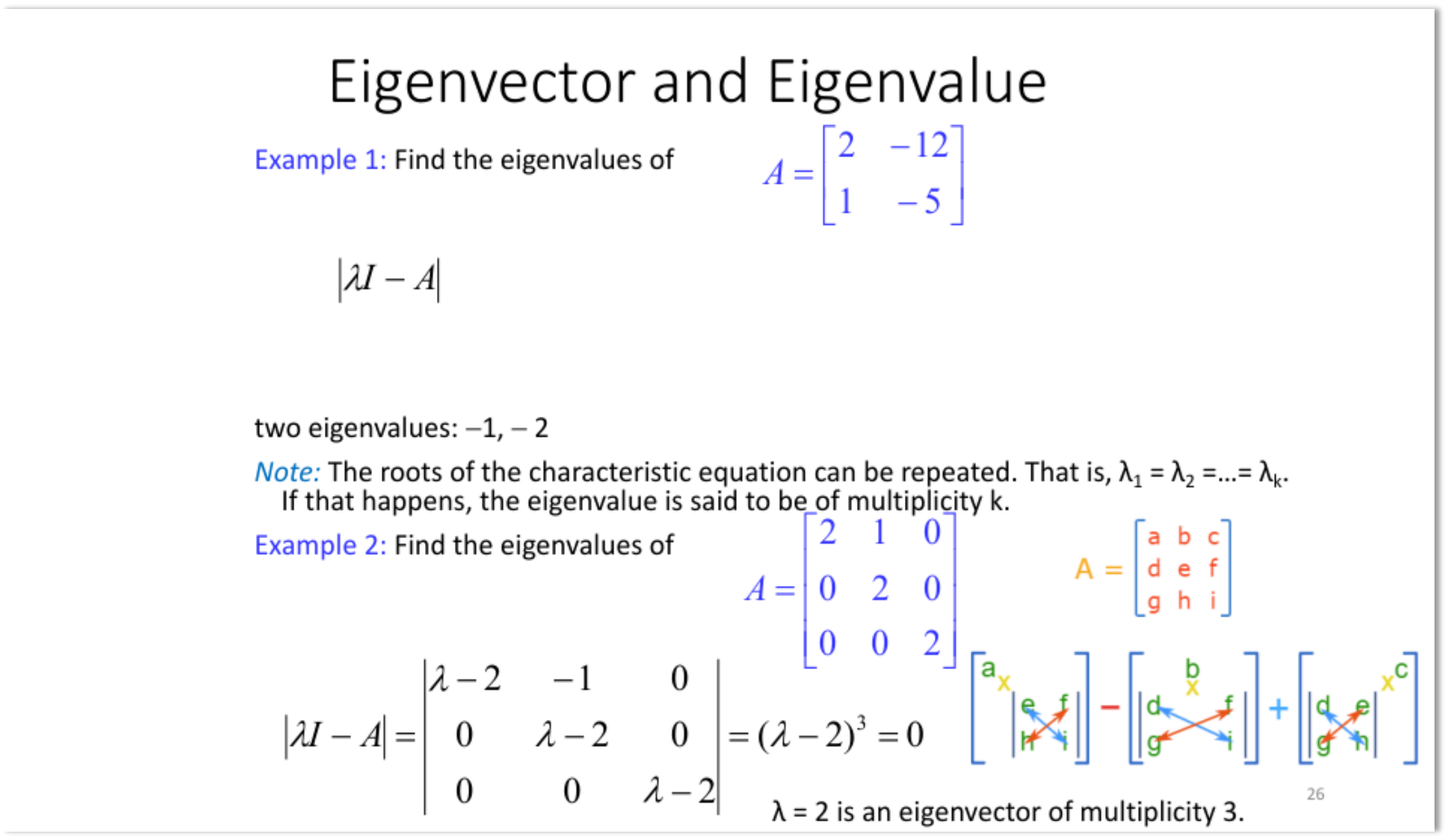
- Decreases error by decreasing the variance in the results due to ***unstable learners***, algorithms (like decision trees and neural networks) whose output can change dramatically when the training data is slightly changed.
  - Increases classifier stability/reduces variance

20

RESULTS

SparrowVQE excelled in deciphering complex educational material, outshining in metrics like ROUGE, BLEU, METEOR, and CIDEr, affirming its alignment with expert-level explanations (Ganesan, 2018; Papineni, 2002).

Models	Size	R-1	R-2	R-L	COSINE	BLEU	CIDEr	METEOR
BLIP	224M	8.4	0.7	7.19	0.077	0.15	0.17	0.078
Pix2Struct	1.3B	38	20.1	35.5	0.365	0.4	0.47	0.379
LLaVA	7B	35.4	18.4	33.0	0.348	0.37	0.53	0.402
LLM Blender	124M	51.5	34.8	49	0.489	0.54	0.573	0.573
SparrowVQE	3B	68.13	51.54	63.92	0.61	0.7	0.67	0.652



Eigenvector and Eigenvalue

Example 1: Find the eigenvalues of $A = \begin{bmatrix} 2 & -12 \\ 1 & -5 \end{bmatrix}$

$|\lambda I - A|$

two eigenvalues: −1, − 2

Note: The roots of the characteristic equation can be repeated. That is, $\lambda_1 = \lambda_2 = \ldots = \lambda_k$. If that happens, the eigenvalue is said to be of multiplicity k.

Example 2: Find the eigenvalues of $A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ $A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$

$$|\lambda I - A| = \begin{vmatrix} \lambda - 2 & -1 & 0 \\ 0 & \lambda - 2 & 0 \\ 0 & 0 & \lambda - 2 \end{vmatrix} = (\lambda - 2)^3 = 0$$

λ = 2 is an eigenvector of multiplicity 3.

26

question: What is meant by eigenvalue multiplicity?

ground_truth: It refers to the number of times an eigenvalue is repeated as a root of the characteristic equation.

predicted_answer

SparrowVQE: Eigenvalue multiplicity refers to the number of times an eigenvalue is repeated as a root of the characteristic equation.

LLM Blender: Eigenvalue multiplicity refers to the number of times an eigenvalue appears in the characteristic equation.

LLaVA: Eigenvalue multiplicity refers to the number of times an eigenvalue appears in the characteristic equation.

Pix2Struct: Eigenvalue multiplicity is meant to represent the sum of the eigenvalues of a given set of values

BLIP: it ensures that it generalizes over its equal number of training examples, making it computationally intensive

CONCLUSIONS

Findings: SparrowVQE excels in Visual Question Explanation, enhancing education by making learning interactive and personalized, and improving comprehension and engagement.

Recommendations: Future efforts should focus on refining SparrowVQE's precision and adaptability across various educational contexts.

Limitations: The model's effectiveness may vary with different subjects and its real-world applicability needs further evaluation.

Impact: This study advances AI in education, with SparrowVQE heralding a new era of AI's interaction with visual content. Ongoing research is essential to overcome current limitations and fully harness AI's potential in education.

REFERENCES

Barra, S., Bisogni, C., De Marsico, M., & Ricciardi, S. (2021). Visual question answering: Which investigated applications? Pattern Recognition Letters 151, 325–331

Cheng, Y. (2023) Application of a neural network-based visual question answering system in preschool language education. IEIE Transactions on Smart Processing & Computing 12(5), 419–427.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations

Deepgram: Ai-powered speech recognition. (2023). https://www.deepgram.com/ accessed: 2023-06-30

Liu, H., Li, C., Li, Y., & Lee, Y.J. (2023). Improved baselines with visual instruction tuning.

Ganesan, K. (2018). Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 311–318. ACL '02, Association for Computational Linguistics, USA.