Conference Paper Title*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This paper develops a complete approach for Vision-Language operations through its examination of Image Captioning and Visual Question Answering (VQA). Image-based tasks use computer vision alongside natural language processing (NLP) abilities to produce descriptive text from images and identify their accurate answers through queries.**

The Image Captioning system includes a Convolutional Neural Network as image feature extractor alongside a Long Short-Term Memory network for producing sequential text output. The model operates on combined image-caption elements to create realistic captions at a high success rate. The implementation includes a Similar Image Search function accessing external API capabilities that increases the efficiency of dataset augmentation and retrieval tasks. The image quality improves through JavaScript-based image enhancement strategies just before the extraction process begins.

Visual Question Answering (VQA) uses ResNet-50 CNN to extract image features together with BERT-based transformer model for textual question encoding. A combination of image and text embedding data flows through fully connected layers to achieve accurate answer predictions. Optimization of the system occurs through the application of Cross-Entropy Loss combined with Adam optimization and Cosine Annealing Learning Rate Scheduler to improve accuracy results. The system features Similar Image Search together with JavaScript-based image enhancement which improves user experience.

Thorough research proves that Vision-Language applications benefit significantly from deep learning models as well as outside API access and algorithm optimization techniques. The developed system delivers exceptional accuracy results which make it suitable for service applications including accessibility tools and automated content generation and AI-powered image indexing solutions.

The research uses keywords from Image Captioning, Visual Question Answering, Deep Learning, Computer Vision, NLP, Transformer Models, Image Enhancement and Similar Image Search.

Identify applicable funding agency here. If none, delete this.