IMAGE CAPTIONING USING TRANSFORMER MODELS

Introduction

Image captioning is an important application of Artificial Intelligence that automatically generates textual descriptions for images. It combines techniques from computer vision and natural language processing to understand visual content and express it in human-readable language.

With the rise of deep learning, transformer-based models have shown exceptional performance in multimodal tasks. These models can process images and text together, making them suitable for applications such as image captioning, visual question answering, and assistive technologies.

LAB – 1: Image Captioning using Transformers

Aim

To implement an image captioning system that prepares image-caption data using transformer-based preprocessing techniques.

Theory

Transformer models use attention mechanisms to capture relationships between different parts of input data. In image captioning, visual features are extracted from images and mapped to textual descriptions using sequence-to-sequence learning.

Algorithm

1. Import required libraries.
2. Load image-caption dataset.
3. Analyze dataset samples.
4. Preprocess images and captions.
5. Convert data into tensors.

Conclusion

The dataset was successfully prepared for image captioning tasks using transformer-based preprocessing. This prepared data can be used for training deep learning models.

LAB – 2: Training and Evaluation of Image Captioning Model

Aim

To train and evaluate a transformer-based image captioning model using a prepared dataset.

Theory

Transfer learning allows models pre-trained on large datasets to be fine-tuned for specific tasks. Vision-language models use both image encoders and text decoders to generate captions based on learned visual features.

Algorithm

1. Load a pre-trained image captioning model.
2. Define training parameters.
3. Train the model on the dataset.
4. Evaluate model performance.
5. Generate captions for new images.

Conclusion

The model was trained successfully and produced meaningful captions for unseen images, demonstrating the effectiveness of transformer-based architectures.

Applications, Advantages, and Final Conclusion

Applications

• Assistive technology for visually impaired users.
• Automatic image tagging and retrieval.
• Content generation for social media and digital libraries.
• Smart surveillance systems.

Advantages

• Reduces manual effort.
• Improves accessibility.
• Scalable and efficient.
• Uses state-of-the-art AI models.

Limitations

• Requires large datasets.
• High computational cost.
• Captions may lack context sometimes.

Final Conclusion

This lab successfully demonstrated the implementation of an image captioning system using transformer models. The experiments highlight how deep learning can bridge vision and language to create intelligent systems.