GAAC HACKATHON

PRIVATE &

ZERO-SHOT IMAGE CAPTIONING

CONFIDENTIAL

Team - Skywalkers

- Team Members: Rithika and Viswa
- LinkedIn ID : www.linkedin.com/in/rithika-bollapragada-1419b325b
- LinkedIn ID : https://www.linkedin.com/in/viswa-teja-bottu-7147831b3/
- GITHUB ID: https://github.com/Rithika2407
- GITHUB ID:https://github.com/vss-viswateja

if we dont get errors we are doing it in the wrong way

ZERO-SHOT IMAGE CAPTIONING

GENERATING CAPTIONS IN UNSEEN LANGUAGES

INTRODUCTION

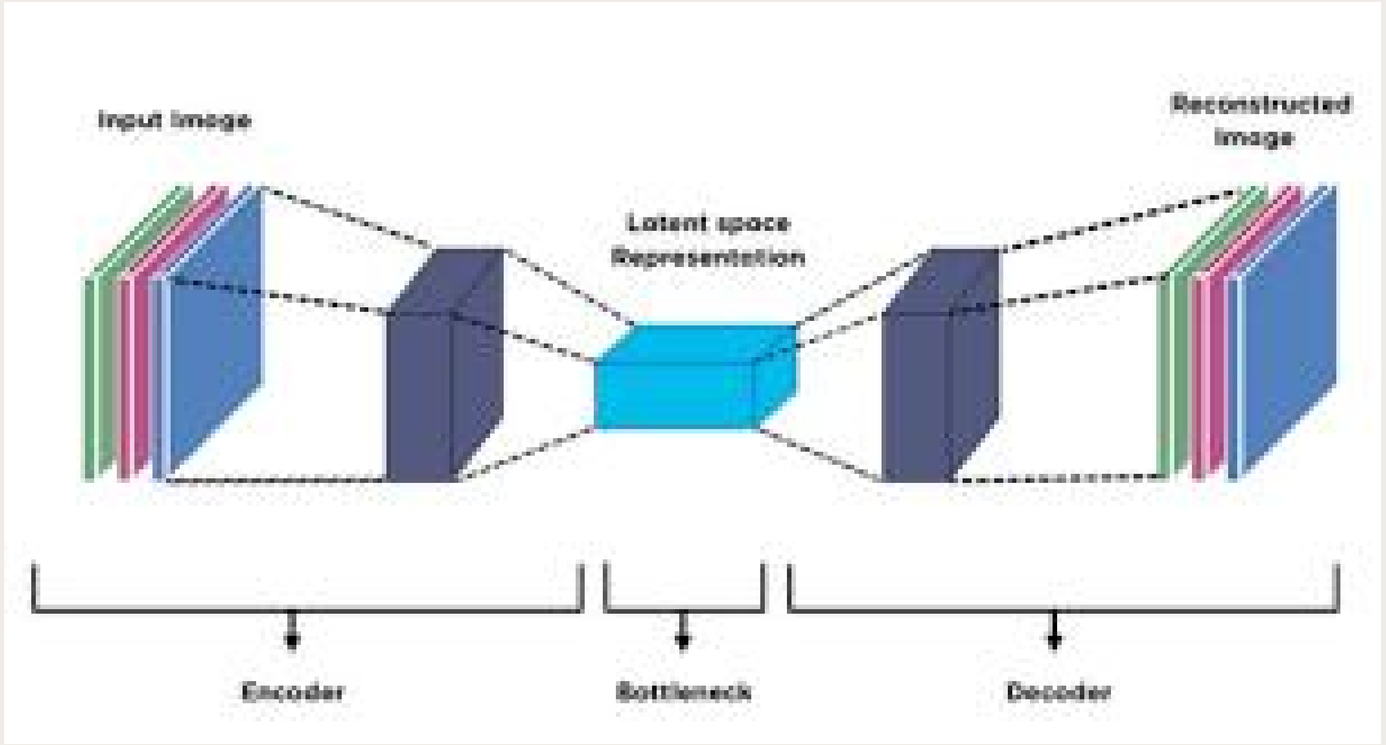
UNDERSTANDING THE PROBLEM AND ITS COMPONENTS

PROBLEM STATEMENT

- Develop a model that generates captions for images in languages it has never been trained on.
- The model should infer linguistic structure from text examples in the target language.
- Generate captions based on visual context.
- Constraints: No direct training on the target language; focus on contextual and semantic correctness.

KEY COMPONENTS

- Image Encoder: Converts images into feature embeddings.
- Text Generator: Generates text using feature embeddings from the image encoder.
- Alignment Mechanism: Aligns image and text embeddings for cross-modal understanding.
- Caption Generator: Constructs coherent captions using image embeddings.



SOLUTION OVERVIEW

XX%

Exploring the technical details

Core Idea

The goal is to design a Zero-Shot Image Captioning System that can generate captions for images in languages it has never been trained on. The solution leverages a combination of vision-language models for understanding visual context and multilingual language models for linguistic transfer. The core idea is to align image representations with textual representations in a shared embedding space and then utilize the linguistic capabilities of a pretrained multilingual model to generate captions in the target language.

This approach does not rely on direct training with image-caption pairs in the target language. Instead, it uses the following techniques to bridge the gap:

Vision-Language Alignment: Create a universal understanding of images and their corresponding textual descriptions.
Cross-Lingual Transfer: Use multilingual language models trained on diverse languages to infer the structure of the target language from text-only examples.

IMAGE ENCODER

- Converts images into feature embeddings.
- Examples: Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs).
- Essential for visual context understanding.

TEXT GENERATOR

- Generates text by taking feature embeddings as input from the image encoder.
- Crucial for producing linguistically accurate captions.

ALIGNMENT MECHANISM

- Aligns image and text embeddings for effective cross-modal understanding.
- Example: Contrastive Learning in models like CLIP.
- Ensures semantic consistency between image and text.

CAPTION GENERATOR

- Uses the image embedding to construct coherent captions.
- Can use pre-trained language models fine-tuned for generative tasks.
- Important for creating meaningful and contextually relevant captions.

TECH STACK

TECHNOLOGIES AND FRAMEWORKS

- PyTorch / TensorFlow: For building, training, and fine-tuning machine learning models.
- HuggingFace Transformers: For leveraging pretrained multilingual language models and vision-language models.
- OpenAI's CLIP: To align image and text embeddings into a shared latent space.
- NLP Toolkit (spaCy, NLTK): For preprocessing text and understanding linguistic structures.
- Datasets: MS-COCO / Conceptual Captions for image-text pairs; publicly available corpora for target language text examples.

PROGRAMMING LANGUAGES AND TOOLS

- Python: Primary language for machine learning, natural language processing, and deployment pipelines.
- Weights & Biases / MLflow: For tracking experiments, hyperparameters, and model performance during training.
- Helps visualize and compare model iterations.

IMPLEMENTATION PLAN

Company

Today's Date

RESEARCH AND PLANNING

Finalize datasets for image-text pretraining and target language examples.
Decide on the pretrained models (e.g., CLIP + mBART or BLIP + mT5).
Define metrics (e.g., BLEU, CIDEr) for evaluating model performance.

DATA COLLECTION AND PREPROCESSING

Collect and preprocess image-caption pairs (e.g., resize images, tokenize captions).
Collect and preprocess monolingual text data for the target language.
Build a pipeline to standardize image embeddings and text inputs.

MODEL TRAINING & TESTING AND EVALUATION

DEPLOYMENT AND DOCUMENTATION

EXPECTED OUTCOME

Key Achievements

1. A zero-shot image captioning system capable of generating captions in languages it was not directly trained on.
2. High-quality, semantically correct captions with minimal linguistic errors in the target language.

CHALLENGES

VISION-LANGUAGE ALIGNMENT & COMPUTING RESOURCES