ADOBE IMAGE CLASSIFICATION AND ARTIFACT DETECTION

Team 67

Understanding the Problem

Digital Trust Crisis: Rapid rise of generative AI technologies like Stable Diffusion, GANs.

Challenge: Identifying AI-generated media and providing arguments to validate the detection process.

Need: Lightweight, generalisable, robust and explainable solutions.

Our Solution

TASK 1

Accurately identifying AI-generated images.

TASK 2

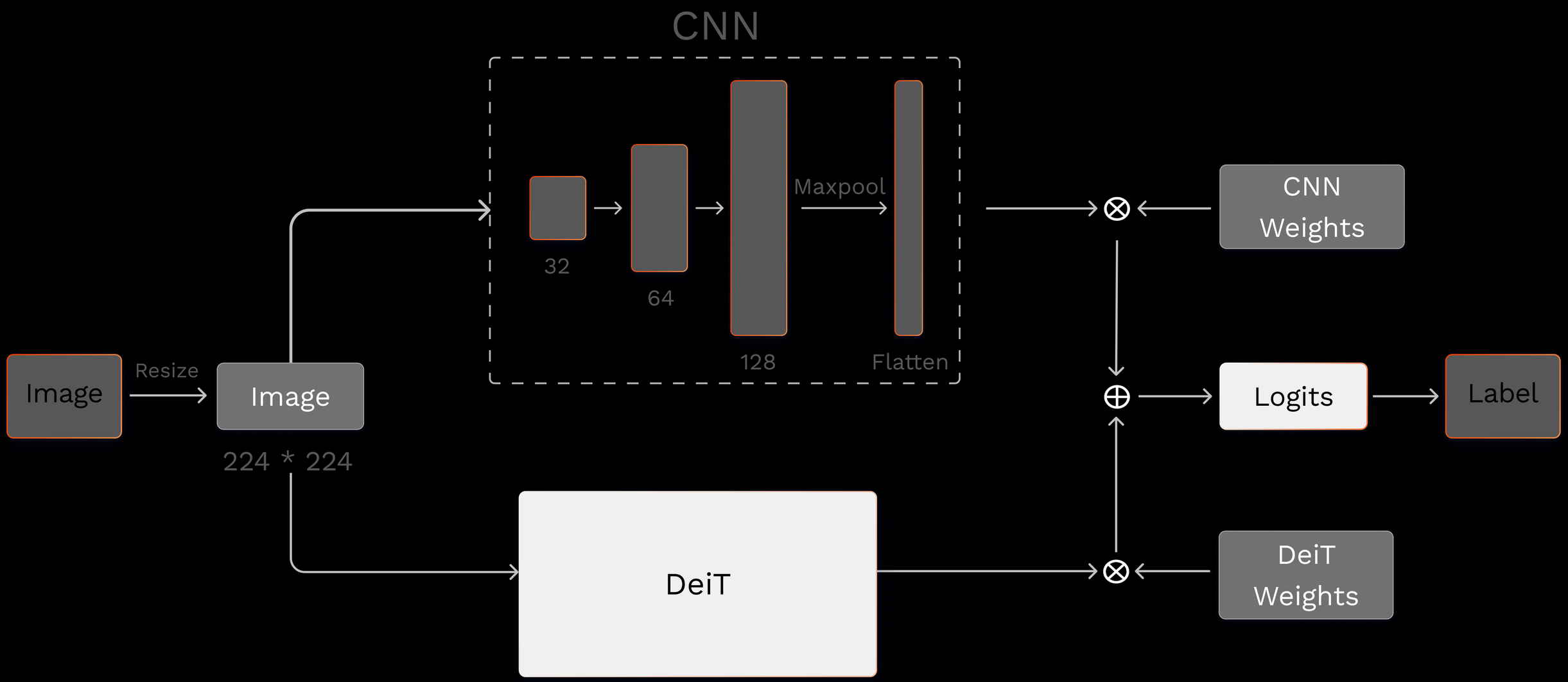
Providing clear and interpret-able explanations highlighting artifacts that led to the classification.

Task 1

IMAGE CLASSIFICATION

Detect whether or not the image is AI-generated.

Architecture



Combining CNN and DeiT for Detection

Features

- CNN : Captures local details.
- DeiT Backbone : Focuses on global context.

Key Advantages and Novelty

- Enhanced performance due to synergy.
- Combining the robustness of DeiT and the generalisability of CNN.
- Lightweight and efficient.
- Performs well with limited dataset.

Optimization Details

Weighted Feature Fusion

- $f_{combined} = \alpha \times f_{CNN} + (1 - \alpha) \times f_{DeiT}$

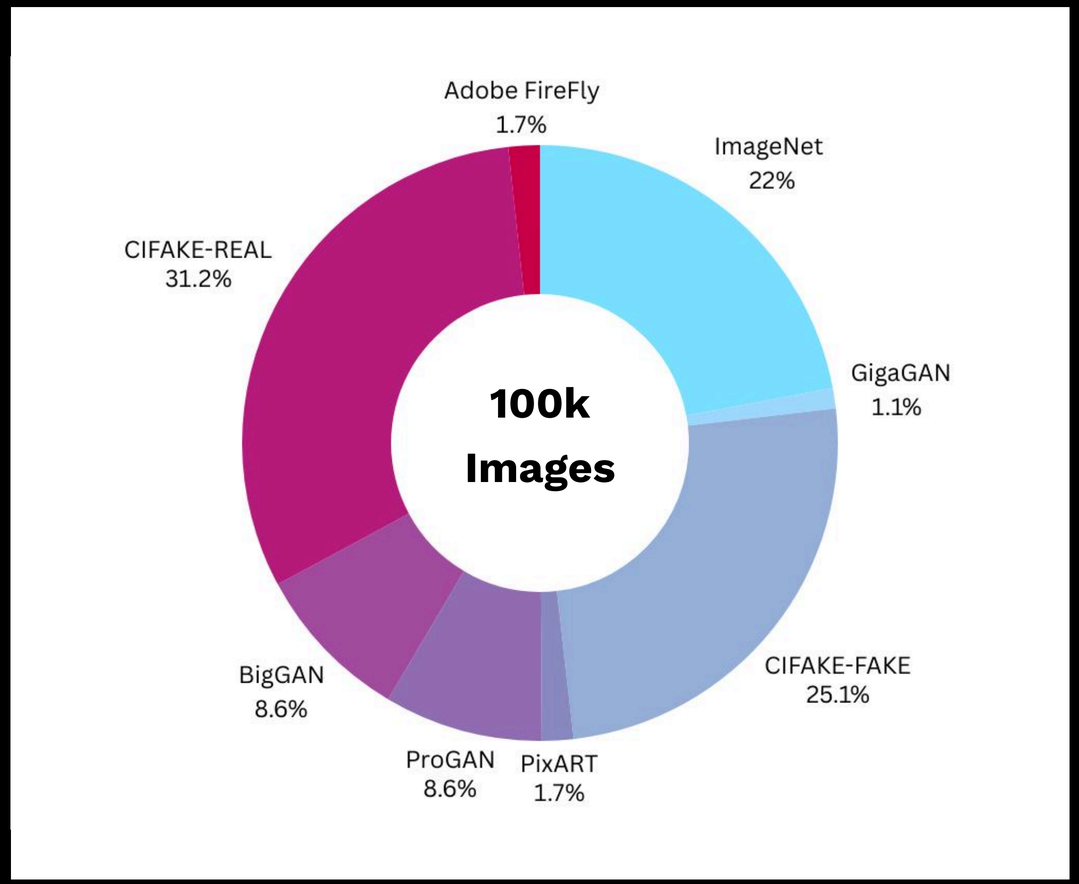
Loss Function

- Binary Cross-Entropy (BCE)

Learning Rate

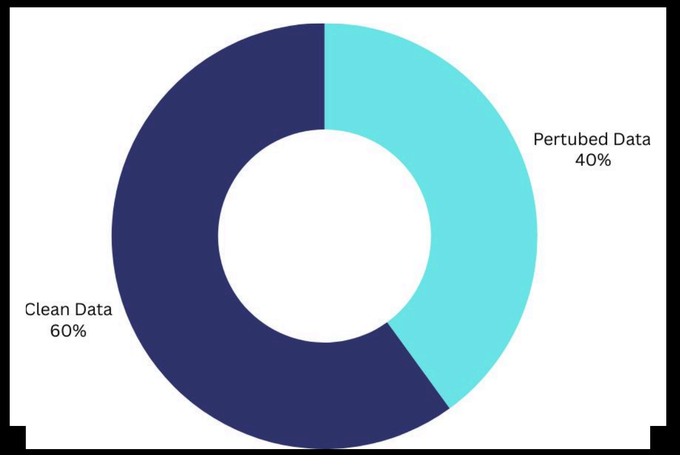
- 1e-6

Training Strategy
Dataset Preparation



Perturbations added

Gaussian Noise, Poisson Noise, Laplace Noise, Salt and Pepper Noise, Color Variations, Random Augmentations and JPEG compression



Quantitative Results

On CIFAKE: **97.8%**
Custom Dataset: **99.3%**

Accuracy

~47 ms on CPU

Inference Time

~31M

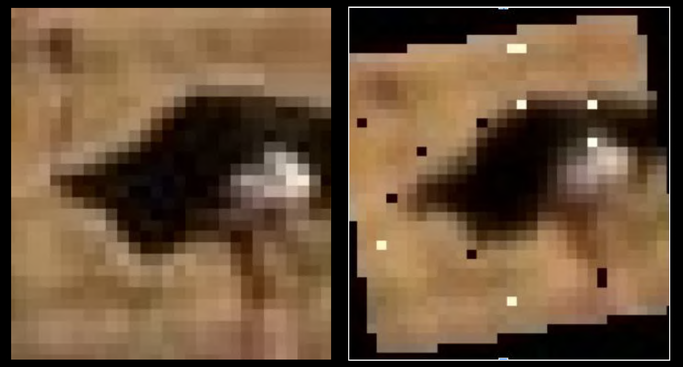
Parameters

~120MB

Model Size

Qualitative Results

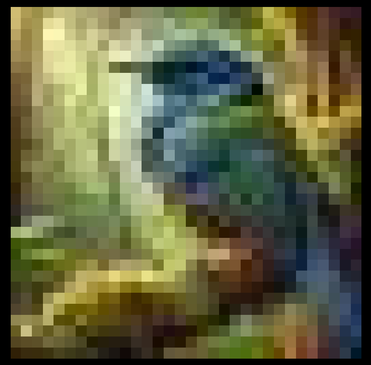
Robustness



Both the original and perturbed image predicted in the same class.

Generalisability

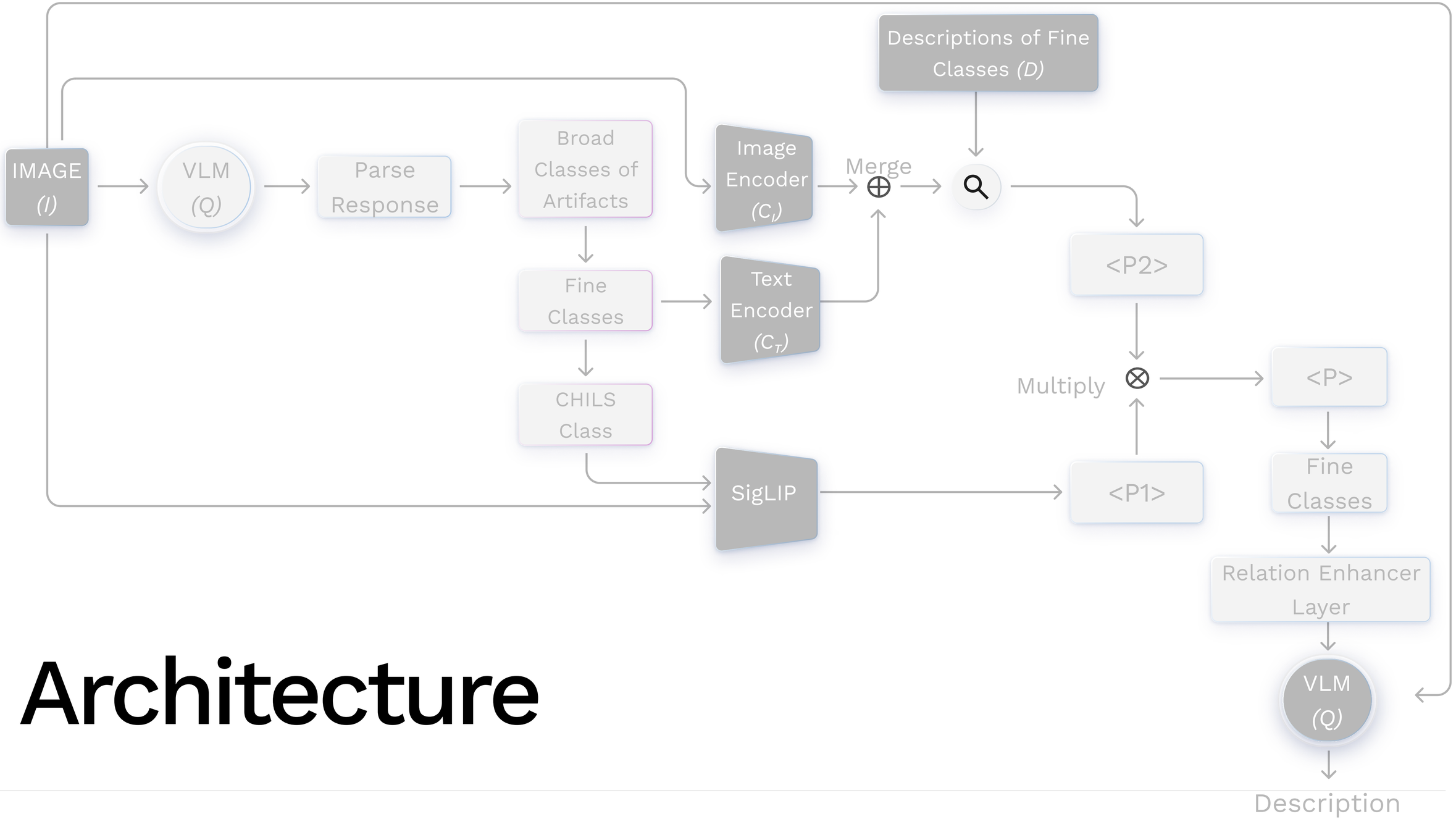
Image generated from DallE (not included in train set) predicted fake.



Artifact Detection

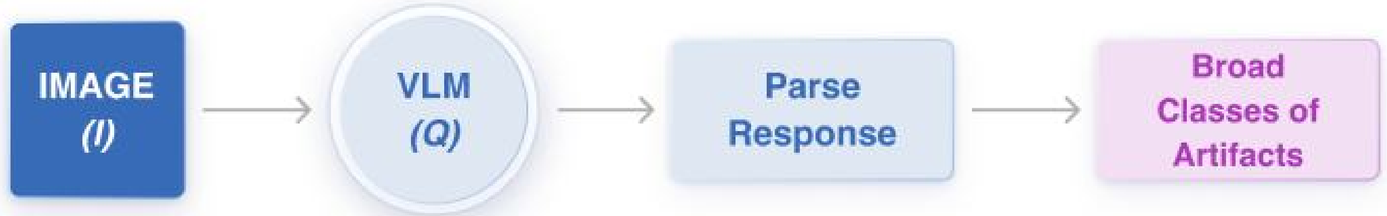
Detect artifacts present in AI generated images.

Task 2



Architecture

Broad Classification



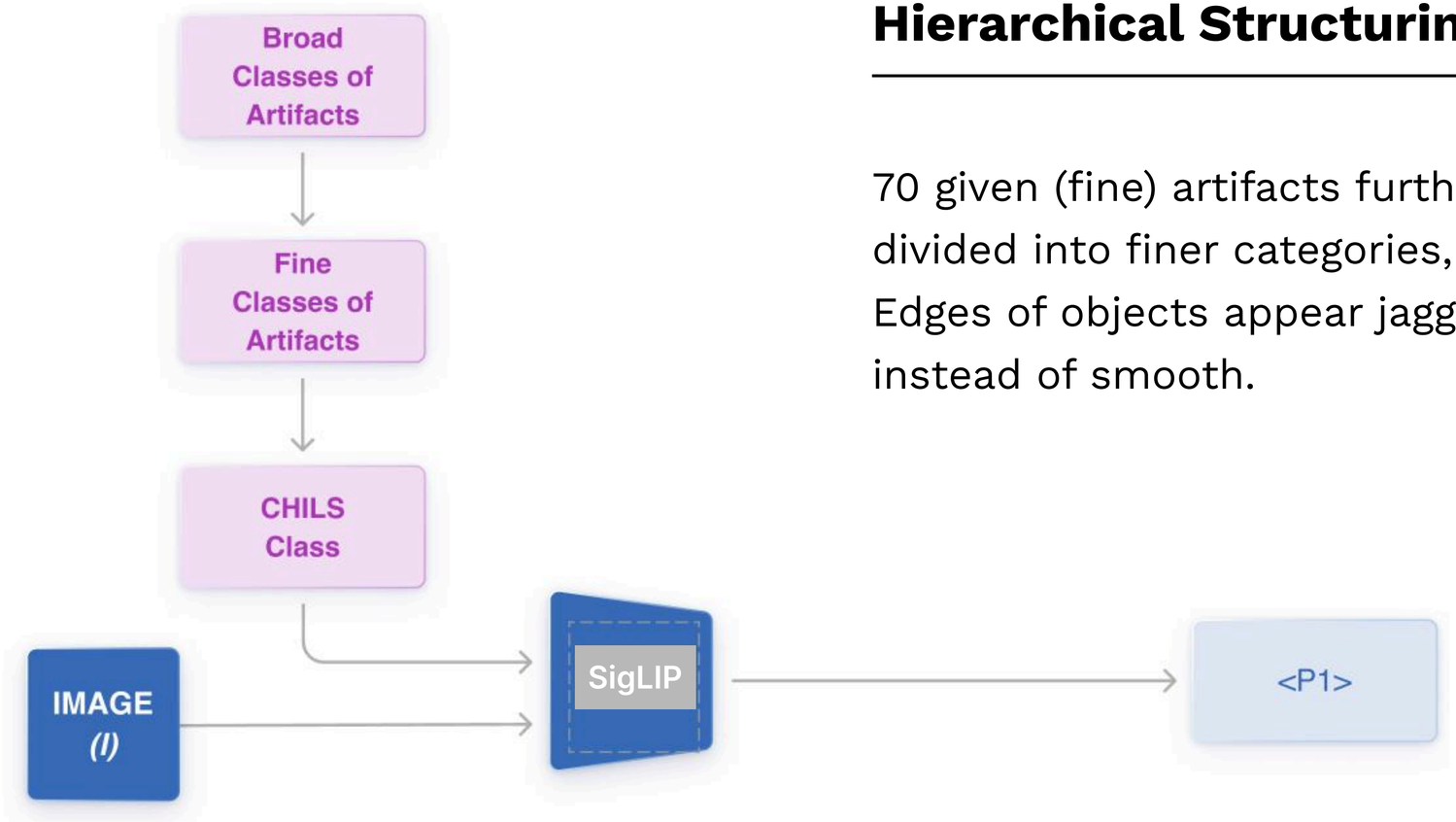
Broad Classes

We cluster similar artifacts into broad classes. For eg: Biological and Anatomical Issues, Physical and Geometric Structure etc.

Qwen2-VL-2B-Instruct

A VLM is used and it’s response is parsed to predict broad classes.

Sub-Class Prediction



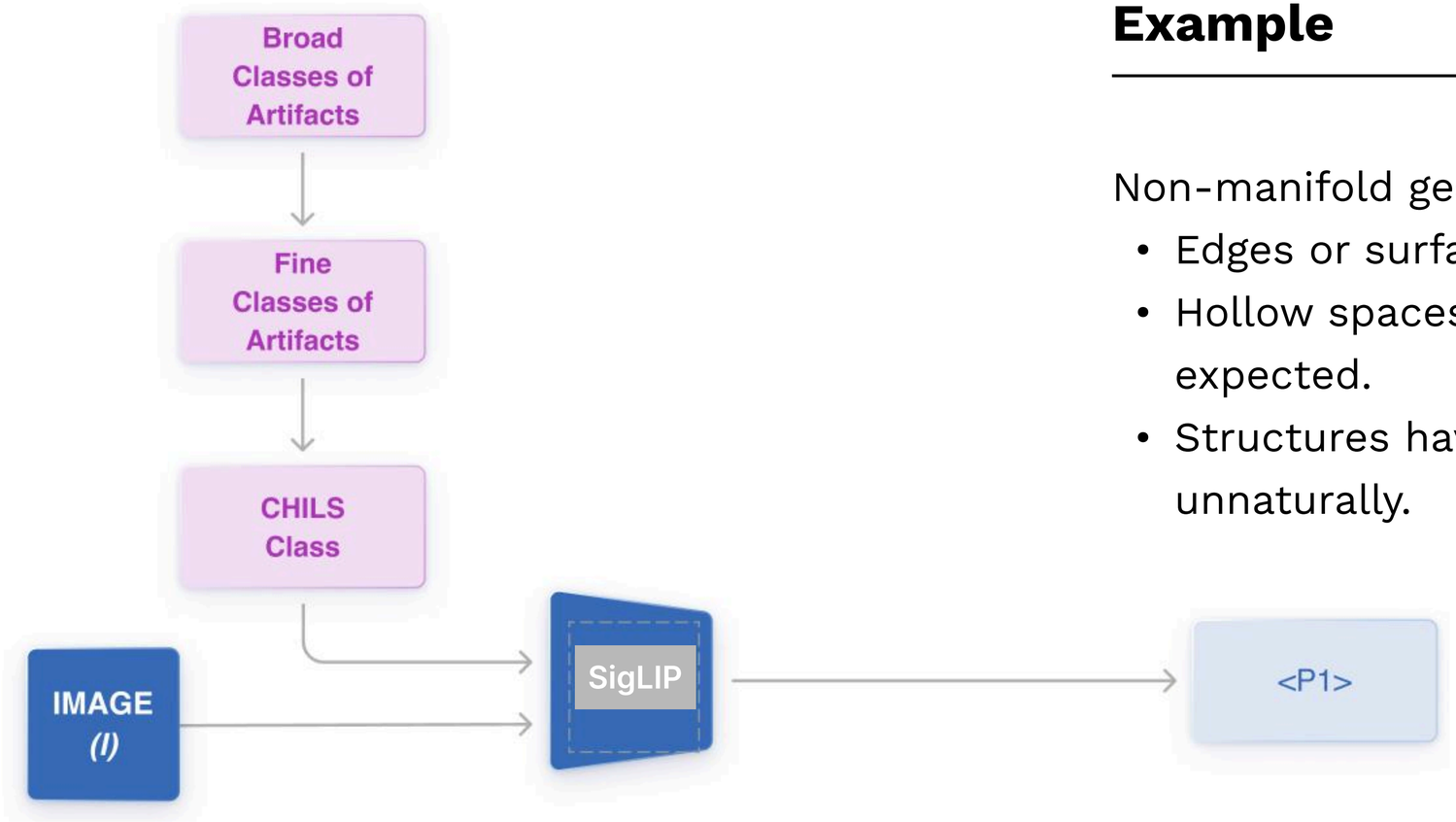
Hierarchical Structuring

70 given (fine) artifacts further divided into finer categories, eg: Edges of objects appear jagged instead of smooth.

SigLIP

SigLIP is Cross model Encoder with Sigmoid Loss function.

Sub-Class Prediction

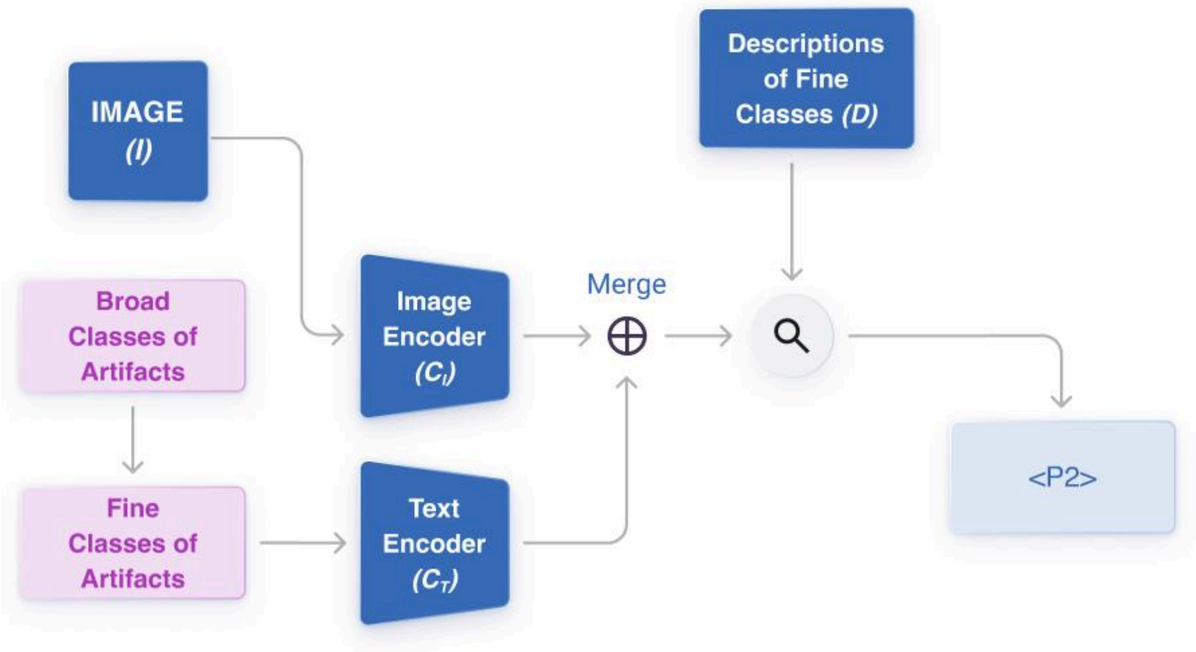


Example

Non-manifold geometries in rigid structures:

- Edges or surfaces intersect unnaturally.
- Hollow spaces exist where solid connections are expected.
- Structures have holes that expose their insides unnaturally.

Refining Predictions using Similarity



Description Embeddings

Generated by SigLIP for fine classes and stored.

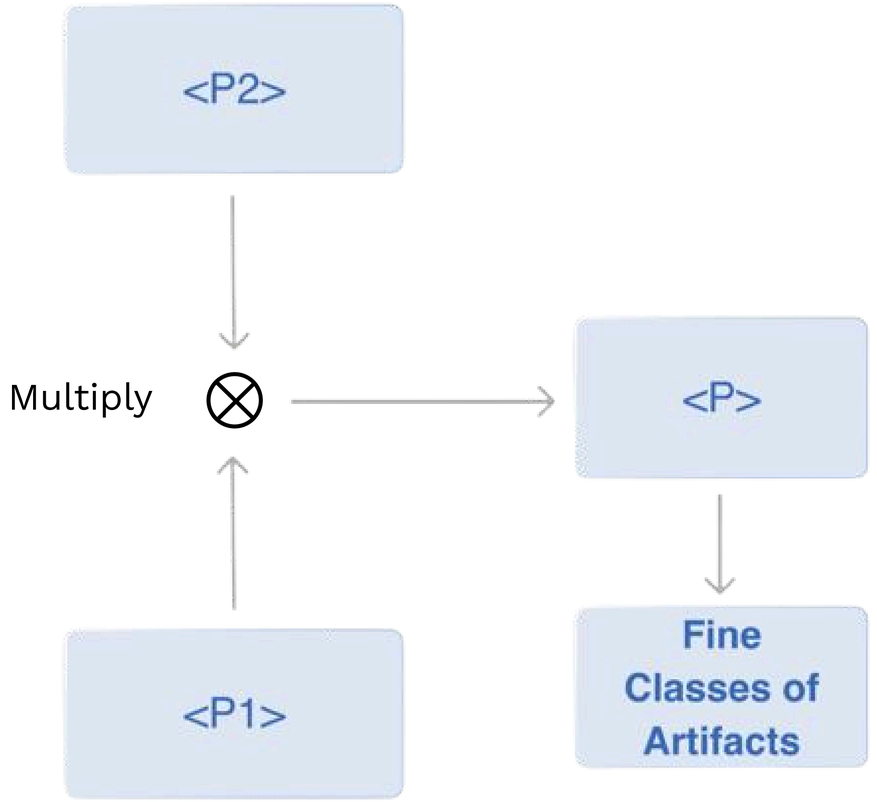
Merged embedding

Image and text embeddings for fine classes corresponding to predicted broad classes merged.

Cosine Similarity

Between merged embeddings and description embeddings to get probabilities P2 of fine classes.

Final Prediction of Fine Classes



Final Prediction

Probability P is calculated by combining P1 and P2. Top-p sampling on P to get final predictions.

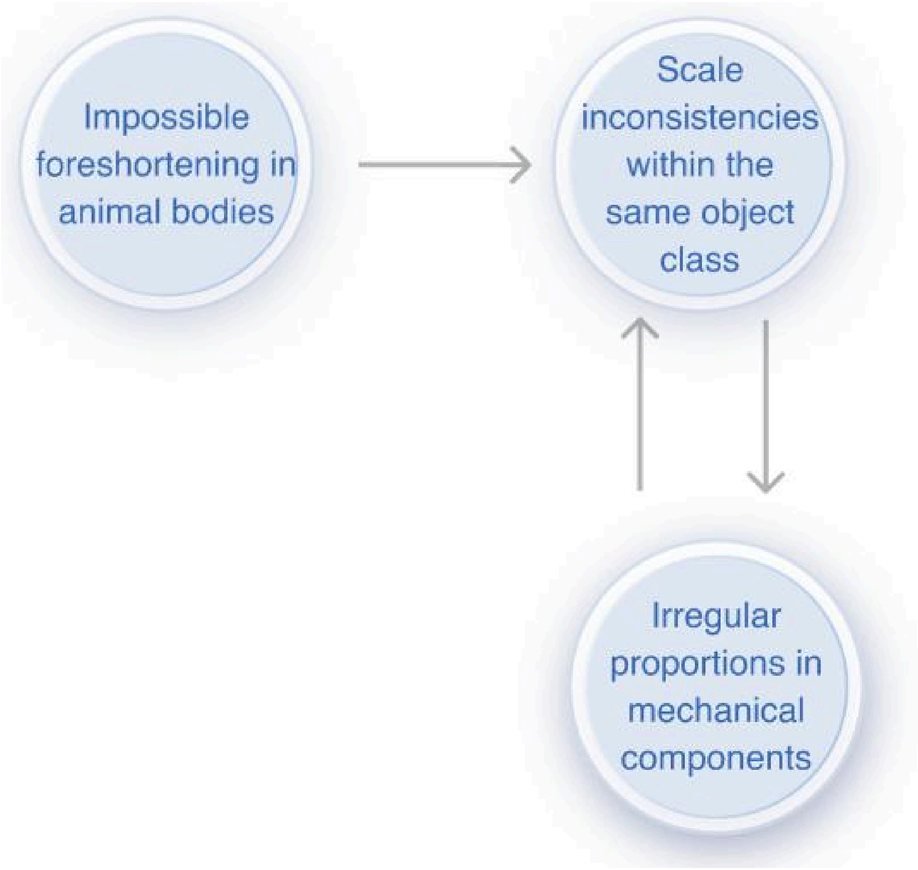
Enhancements

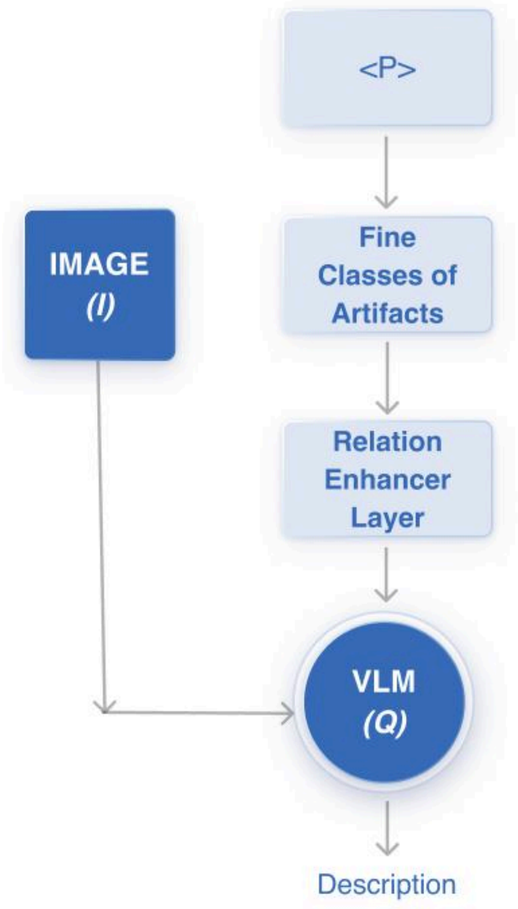
(Non)Living artifacts

Hard coded to not include some artifacts for some type of images (for example biological artifacts in images of ship, truck, etc.)

Relation Enhancement

We made an implication graph on given artifacts. If an artifact is deemed to be present, its subtree is added to predictions.





Generating Description

- Generating description using image and predicted fine classes as input.
- Instead of image can also give superimposed image+heat-map as input.
- Heat-map not implemented because creating heat- map sharply increases latency.

Fine Tuning

We hand annotated few high resolution images and created their artifacts descriptions. Using Gpt-4o, more images were annotated and their descriptions generated.

The Qwen2-VL-2B-Instruct VLM was finetuned on these images after down-scaling and later used to predict broad classes and final descriptions

Results

Predicted Classes:

- Artificial enhancement artifacts
- Regular grid-like artifacts in textures
- Synthetic material appearance
- Artificial smoothness
- Texture repetition patterns
- Artificial depth of field in object presentation
- Fake depth of field
- Over-smoothing of natural textures
- Depth perception anomalies

Future Works

Retrieval-ICL

- Images along with their artifacts and artifact description stored. Images with same artifacts retrieved and provided with their descriptions to VLM as few-shot examples.

Prototype Based Learning :

- Labelled Images with artifacts are used for training and when new images come this model tries to find the similarity of its artifacts with all the learned artifacts to find the artifacts present

THANK YOU!

APPENDIX

Experimentation

Model	Accuracy on CIFAKE(%)	Accuracy on Custom Data(%)	Inference Time(ms)
CNN	97.04	87.4	0.2
SR-Net	95.88	20	0.5
ResNet with no downsampling	93.52	10	190
DiRE	97.225	31.2	30
Aeroblade	50	-	1
Google ViT	91	-	205
Efficient Net	96.1	98.6	9
DeiT	97.5	99.2	30

Experimentation- Task 2

Using Multimodal Large Language Models for zero shot classification:

- Using VLM to generate textual context of image
- Using cross modal encoder to get fixed dimension embeddings of both image and text
- Combining both for classification.